The genome sequence of the filamentous fungus *Neurospora crassa*

James E. Galagan[1], Sarah E. Calvo[1], Katherine A. Borkovich[2], Eric U. Selker[3], Nick D. Read[4], David Jaffe[1], William FitzHugh[5], Li-Jun Ma[1], Serge Smirnov[1], Seth Purcell[1], Bushra Rehman[1], Timothy Elkins[1], Reinhard Engels[1], Shunguang Wang[1], Cydney B. Nielsen[1], Jonathan Butler[1], Matthew Endrizzi[1], Dayong Qui[1], Peter Ianakiev[1], Deborah Bell-Pedersen[6], Mary Anne Nelson[7], Margaret Werner-Washburne[7], Claude P. Selitrennikoff[8], John A. Kinsey[9], Edward L. Braun[10], Alex Zelter[4,11], Ulrich Schulte[12], Gregory O. Kothe[3], Gregory Jedd[13], Werner Mewes[14,15], Chuck Staben[16], Edward Marcotte[17], David Greenberg[18], Alice Roy[1], Karen Foley[1], Jerome Naylor[1], Nicole Stange-Thomann[1], Robert Barrett[1], Sante Gnerre[1], Michael Kamal[1], Manolis Kamvysselis[1], Evan Mauceli[1], Cord Bielke[14], Stephen Rudd[15], Dmitrij Frishman[15], Svetlana Krystofova[2], Carolyn Rasmussen[19], Robert L. Metzenberg[20], David D. Perkins[20], Scott Kroken[19], Carlo Cogoni[21], Giuseppe Macino[21], David Catcheside[22], Weixi Li[16], Robert J. Pratt[6], Stephen A. Osmani[23], Colin P. C. DeSouza[24], Louise Glass[19], Marc J. Orbach[25], J. Andrew Berglund[3], Rodger Voelker[3], Oded Yarden[11], Michael Plamann[26], Stephan Seiler[26], Jay Dunlap[27], Alan Radford[28], Rodolfo Aramayo[6], Donald O. Natvig[7], Lisa A. Alex[29], Gertrud Mannhaupt[14], Daniel J. Ebbole[30], Michael Freitag[3], Ian Paulsen[18], Matthew S. Sachs[31], Eric S. Lander[1,32], Chad Nusbaum[1] & Bruce Birren[1]

A list of author affiliations appears at the end of the paper

Neurospora crassa **is a central organism in the history of twentieth-century genetics, biochemistry and molecular biology. Here, we report a high-quality draft sequence of the *N. crassa* genome. The approximately 40-megabase genome encodes about 10,000 protein-coding genes—more than twice as many as in the fission yeast *Schizosaccharomyces pombe* and only about 25% fewer than in the fruitfly *Drosophila melanogaster*. Analysis of the gene set yields insights into unexpected aspects of *Neurospora* biology including the identification of genes potentially associated with red light photobiology, genes implicated in secondary metabolism, and important differences in Ca^{2+} signalling as compared with plants and animals. *Neurospora* possesses the widest array of genome defence mechanisms known for any eukaryotic organism, including a process unique to fungi called repeat-induced point mutation (RIP). Genome analysis suggests that RIP has had a profound impact on genome evolution, greatly slowing the creation of new genes through genomic duplication and resulting in a genome with an unusually low proportion of closely related genes.**

Research on *Neurospora* in the early part of the twentieth century paved the way for modern genetics and molecular biology. First documented in 1843 as a contaminant of bakeries in Paris[1], *Neurospora* was developed as an experimental organism in the 1920s[2,3]. Subsequent work on *Neurospora* by Beadle and Tatum[4] in the 1940s established the relationship between genes and proteins, summarized in the 'one-gene-one-enzyme' hypothesis. In the latter half of the century, *Neurospora* had a central role as a model organism, contributing to the fundamental understanding of genome defence systems, DNA methylation, mitochondrial protein import, circadian rhythms, post-transcriptional gene silencing and DNA repair[5]. Because *Neurospora* is a multicellular filamentous fungus, it has also provided a system to study cellular differentiation and development as well as other aspects of eukaryotic biology[6].

The legacy of over 70 years of research[7], coupled with the availability of molecular and genetic tools, offers enormous potential for continued discovery. The sequencing of the *N. crassa* genome was undertaken to maximize this potential. Here, we report an initial sequence and analysis of the *Neurospora* genome.

Neurospora genome sequence

The *Neurospora* genome is much larger (greater than 40 megabases (Mb)) than that of *S. pombe* and *Saccharomyces cerevisiae* (both about 12 Mb). Accordingly, first we sought to produce and analyse a high-quality draft sequence *en route* to a finished sequence.

The genome sequence was assembled from deep whole-genome shotgun (WGS) coverage obtained by paired-end sequencing from a variety of clone types (Supplementary Information). In all, the data provided an average of >20-fold sequence coverage and >98-fold physical coverage of the genome. The Arachne package[8] was used to assemble the draft genome sequence. The resulting assembly consists of 958 sequence contigs with a total length of 38.6 Mb (Table 1) and an N50 length of 114.5 kilobases (kb) (that is, 50% of all bases are contained in contigs of at least 114.5 kb). Contigs were assembled into 163 scaffolds with a total length of 39.9 Mb (including gaps between contigs) and an N50 length of 1.56 Mb.

Most of the assembly (97%) is contained in the 44 largest scaffolds, and there are 38 tiny scaffolds with lengths <4 kb. Forty-two of the large scaffolds (and one of the smaller ones) could be anchored readily to the *Neurospora* genetic map[7] by virtue of their containing genetic markers with sequence.

The assembly has long-range continuity, with the N50 scaffold size being nearly 1,000-fold larger than the average gene size. The assembly represents the vast majority of the genome, as assessed by comparison with available finished sequence and genetic markers. It contains 99.13% of available finished sequence (17 Mb from linkage groups II and V[9]) and all of the 252 genetic markers with sequence. This estimate, however, does not account for unusual genomic regions such as the ribosomal DNA repeats, centromeres and telomeres; such regions may contain about 1.7 Mb of additional sequence[10], corresponding to 2–3% of the genome that cannot be assembled readily with available techniques. The long-range continuity of the assembly was also confirmed by comparisons with previously described bacterial artificial chromosome (BAC) physical maps for linkage groups II and V[11], as only one discrepancy was noted.

The assembly also has high accuracy, with 99.5% of the sequence having Arachne quality scores ≥30. Comparison with the 17 Mb of finished sequence confirms the sequence accuracy, with a discrepancy rate for this subset of less than 10^{-5}. The comparison also largely confirms the assembly, as only 12 minor discrepancies were identified (Supplementary Information).

© 2003 Nature Publishing Group

Genes

Gene count and basic characteristics

A total of 10,082 protein-coding genes (9,200 longer than 100 amino acids) were predicted (Table 1 and Supplementary S0). This constitutes nearly twice as many genes as in *S. pombe* (about 4,800) and *S. cerevisiae* (about 6,300), and nearly as many as in *D. melanogaster* (about 14,300). Genes cover at least 44% of the genome sequence with an average gene density of one gene per 3.7 kb. The average gene length of 1.67 kb is slightly longer than the 1.4-kb average gene length for both *S. cerevisiae* and *S. pombe*. The difference in gene length is due to the greater number of introns in *Neurospora* genes—an average of 1.7 introns per gene with an average intron size of 134 nucleotides. Notably, most predicted *Neurospora* introns lack a polypyrimidine tract, which is common in other eukaryotic introns, but do contain a strong branchpoint sequence (Supplementary Information).

Comparative analysis

A total of 4,140 (41%) *Neurospora* proteins lack significant matches to known proteins from public databases (Table 1), reflecting the early stage of fungal genome exploration and the diversity of fungal genes remaining to be described. Furthermore, 5,805 (57%) *Neurospora* proteins do not have significant matches to genes in either of the sequenced yeast species (Supplementary Information). When compared to sequenced eukaryotes, a total of 1,421 (14%) *Neurospora* genes display best BLASTP matches to proteins in either plants or animals (Supplementary Information). Of these, 584 lack high-scoring hits to either sequenced yeast species. These data reflect the biology shared by filamentous fungi and higher eukaryotes, which in a number of cases is absent in the yeasts.

Epigenetics, genome defence and genome evolution

Neurospora is an important model for the study of epigenetic phenomena, possessing a wide variety of epigenetic mechanisms and related genome defence mechanisms. The most remarkable of these mechanisms is repeat-induced point mutation (RIP), a process unique to fungi.

Repeat-induced point mutation

First discovered in *Neurospora*[12,13], RIP is a process that efficiently detects and mutates both copies of a sequence duplication. RIP acts during the haploid dikaryotic stage of the *Neurospora* sexual reproductive cycle, causing numerous C•G to T•A mutations within duplicated sequences. In a single passage through the sexual cycle, up to 30% of the C•G pairs in duplicated sequences can be mutated, with a strong preference for C to T mutations occurring at CpA dinucleotides[14]. The pattern of mutations produces a characteristic skewing of dinucleotide frequencies that allows RIP-mutated sequences to be detected accurately[15]. RIP requires a minimal duplicated sequence length of about 400 base pairs (bp)[16] and greater than roughly 80% sequence identity between duplicates[17]. In addition to suffering mutations, RIP-mutated sequences are frequent targets for DNA methylation. As with mammals, DNA methylation has been shown to cause gene silencing in *Neurospora*[18]. RIP thus mutates and epigenetically silences repetitive DNA.

RIP has been proposed to act as a defence against selfish or mobile DNA[13]. However, because RIP mutation and methylation can extend beyond the bounds of duplicated sequences[19], RIP can have both mutational and epigenetic effects on neighbouring unique sequences. Furthermore, RIP acts on all duplicated sequences, including those arising from large-scale chromosomal duplications as well as gene duplications[20]. The presence of RIP thus has profound consequences for the evolution of the *Neurospora* genome. Indeed, it has been proposed that RIP might prevent gene innovation through gene duplication[13,21]. With the availability of the *Neurospora* genome sequence, we were able to address this hypothesis.

Multigene families

To investigate the impact of RIP on protein families in *Neurospora*, genes were clustered into 'multigene families' on the basis of an all versus all comparison of protein sequences (see Methods). As shown in Fig. 1, the percentage of genes in multigene families in selected sequenced eukaryotes is correlated with genome size. However, in marked contrast to the other analysed organisms, *Neurospora* possesses many fewer genes in multigene families than expected. When the analysis is expanded to include an additional 17 sequenced prokaryotes (Supplementary Information), only *Mycoplasma genitalium*, *Mycoplasma pulminus*, *Ureaplasma urealyticum* and *Vibrio cholerae* display a correspondingly small proportion of genes in families. This is noteworthy considering that the *Mycoplasma* genus is thought to have undergone reductive evolution and represent minimal life forms[22].

Our analysis reveals another characteristic of *Neurospora* gene families. Unlike other sequenced eukaryotes, *Neurospora* possesses only a handful of highly similar gene pairs. Figure 2 displays histograms of amino acid and nucleotide similarities between each gene in the six organisms analysed and the best-matching gene in that organism. A significant proportion of genes have best matches with greater than 80% amino acid and nucleotide identity in all the organisms considered except *Neurospora*. *Neurospora* contains only eight genes with top matches of greater than 80% amino acid or coding sequence identity. This value is significant because, as described above, RIP mutates duplicated sequences that display greater than about 80% nucleotide similarity. Thus, the small proportion of genes in multigene families and the near absence of highly similar genes are consistent with the actions of RIP.

An example of the lack of highly similar genes in multigene families is revealed in an analysis of predicted major facilitator superfamily (MFS) sugar transporters (Fig. 3). *Neurospora* has about the same number of predicted MFS sugar transporters as *S. cerevisiae*. However, a phylogenetic analysis of fungal sugar transporters indicates that the *Neurospora* proteins are substantially more divergent than those of *S. cerevisiae* as well as those of *S. pombe*. Furthermore, the *Neurospora* transporters contain no apparent instances of recent duplication. In contrast, most of the *S. cerevisiae* HXT hexose and *S. pombe* GHT transporters represent two relatively recent and independent expansions and include very recently duplicated genes. Thus, despite a diversity of MFS sugar transporters, *Neurospora* seems to lack close paralogues in this gene family, consistent with the results of the genome-wide multigene family analysis.

Table 1 ***Neurospora crassa*** **genome features**

Feature	Value
General	
Size (bp) (assembly 5)	38,639,769
Chromosomes	7
G + C content (%)	50
Protein-coding genes	10,082
Protein-coding genes >100 amino acids	9,200
tRNA genes	424
5S rRNA genes	74
Per cent coding	44
Per cent intronic	6
Average gene size (bp)	1,673 (481 amino acids)
Average intergenic distance (bp)	1,953
Predicted protein-coding sequences	
Identified by similarity to known sequences	1,336 (13%)
Conserved hypothetical proteins	4,606 (46%)
Predicted proteins (no similarity to known sequences)	4,140 (41%)

© 2003 Nature Publishing Group

Analyses of other gene families yielded similar results (data not shown). Furthermore, the paucity of closely related sequences is evident not only at the level of complete genes, but even at the level of individual exons, protein domains and protein architectures (Supplementary S4).

Gene evolution through gene duplication

The above results suggest that RIP has had a powerful impact in suppressing the creation of new genes or partial genes through genomic duplication. This is consistent with the large number of mutations induced in duplicated sequences by RIP. Computer simulations (see Methods) indicate that after a gene duplication, each copy has an 80% probability of acquiring an in-frame stop codon after only a single round of RIP and a 99.5% probability by the point that RIP has mutated the copies to less than 85% nucleotide similarity. The high frequency of stop codons reflects the preference of RIP for mutating CpA to TpA, increasing the prevalence of the stop codons TAA and TAG.

These results raise the critical question of whether any significant gene duplication has occurred in *Neurospora* subsequent to the acquisition of RIP. We searched for empirical evidence of duplicated genes that have survived RIP by analysing the set of *Neurospora* coding sequences using two different measures[15] for detecting RIP-mutated sequences (see Methods). These measures use the characteristic skewing of dinucleotides produced by RIP to detect mutated sequences. According to these measures, only 59 of the 9,200 predicted genes encoding proteins ≥100 amino acids show evidence of mutation by RIP. Of these, only eight consist of pairs of predicted duplicated genes (genes in the same multigene family) in which both copies are predicted to be RIP-mutated. Thus, few pairs of duplicated genes display evidence of having both survived RIP (Supplementary Information).

Gene duplication is thought to have a primary role in the innovation of new genes[23]. However, taken together, our data support the conclusion that most, if not all, paralogous genes in *Neurospora* duplicated and diverged before the emergence of RIP, and since that point the evolution of new genes through gene duplication has been virtually arrested. This conclusion raises the question of whether and how *Neurospora* is able to evolve new genes. A number of mechanisms that do not involve gene duplication are conceivable, although ultimately a conclusive analysis may be possible only by comparing the genome of *Neurospora* with the genomes of closely related species to illuminate recent evolutionary history. Nonetheless, our results indicate that the cost to *Neurospora* of increased genome security through RIP is a significant impact on the evolution of new gene functions through gene duplication.

Repetitive DNA

An analysis of repeat sequences longer than 200 bp and with greater than 65% similarity (see Methods) revealed that 10% of the *Neurospora* assembly consists of repeat sequences, consistent with previously reported estimates[21].

The repeat sequence of *Neurospora* provides a testament to the efficiency of RIP. Applying the measures of RIP mentioned above to the *Neurospora* genome revealed that most of the repetitive sequences (81%) in *Neurospora* have been mutated by RIP. Conversely, only 18% of predicted RIP-mutated sequence is non-repetitive, potentially reflecting loss of the corresponding duplicated sequence. As described above, duplications greater than about 400 bp are susceptible to RIP[16]. In keeping with this, we observe that over 97% of genomic repeats greater than 400 bp in length are RIP-mutated. Moreover, repeats longer than 400 bp clustered by sequence similarity display an average sequence identity within clusters of 78%, with 93% of clusters displaying an average identity of less than 85%. This corresponds to previous estimates indicating

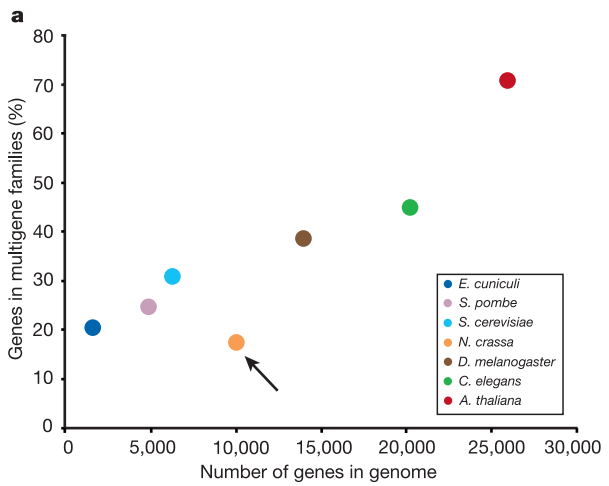


Figure 1 *Neurospora* has a low proportion of genes in multigene families. The graph displays the proportion of genes in multigene families (see Methods) as a function of the number of genes in the genomes of selected sequenced eukaryotic organisms. The arrow indicates *Neurospora*. See text for more details.

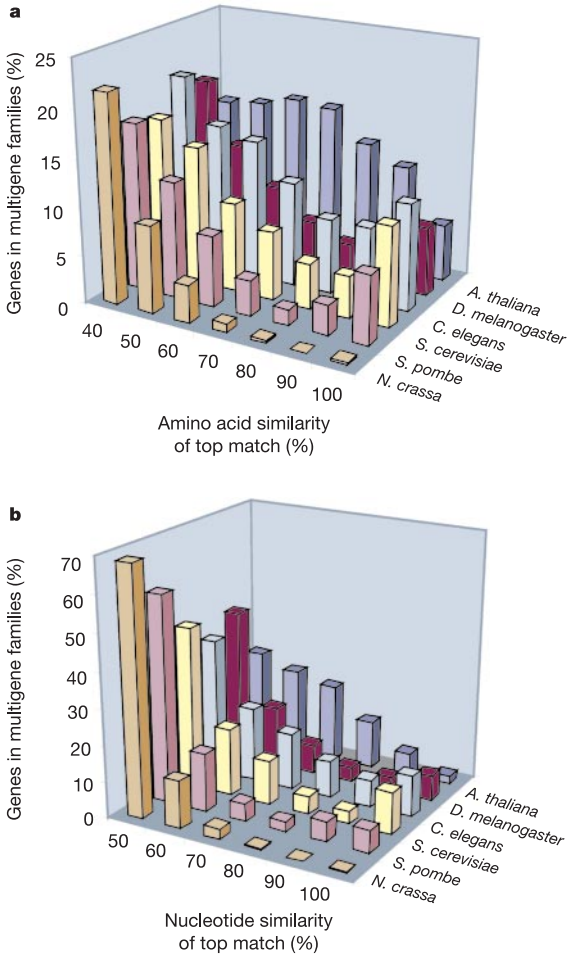


Figure 2 *Neurospora* possesses few highly similar genes. **a**, **b**, Histogram of amino acid (**a**) and nucleotide (**b**) per cent identity of top-scoring self-matches for genes in selected sequenced eukaryotic genomes. For each organism, the protein and coding regions for each gene (not including pseudogenes) were compared to those of every other gene in the same genome using BLASTX. Top-scoring matches were aligned using ClustalW and per cent identities calculated. In contrast to other eukaryotes, *Neurospora* possesses only eight genes with a top match of greater than 80% amino acid or nucleotide identity.

© 2003 Nature Publishing Group

that RIP requires greater than about 80% sequence identity to detect duplicated sequences.

Consistent with the hypothesis that RIP acts as a defence mechanism against selfish DNA[13], no intact mobile elements were identified. Furthermore, a significant proportion of the *Neurospora* RIP-mutated sequence (46% of repetitive nucleotides) can be identified as relics of mobile elements (Supplementary Information).

Ribosomal RNA

The only large repetitive sequences known to have survived RIP in *Neurospora* are the approximately 175–200 copies[10] of the large rDNA tandem repeat containing the 17S, 5.8S and 25S rRNA genes. As in higher eukaryotes, these tandem repeats occur within the nucleolar organizer region (NOR), and their resistance to RIP seems to stem from this localization[13]. Within the genome sequence we found several copies of the rDNA repeat outside the NOR. In every case, they display evidence of mutation by RIP, consistent with previous observations[13]. Thus, the sequence of the rDNA repeat does not in itself seem to confer resistance to RIP.

The 5S rRNA genes in *Neurospora* have survived RIP in a different manner. In contrast to most higher eukaryotes in which the 5S rRNA genes form tandem repeats, the 5S genes are dispersed throughout the genome in *Neurospora*[24]. A total of 74 copies comprising several different subtypes of 5S rDNA are dispersed through all seven chromosomes. This dispersal coupled with their small size (approximately 120 nucleotides) ensures that they are not recognized by RIP.

DNA methylation

Neurospora has been used extensively as a model for studying DNA methylation in eukaryotes[25]. The *Neurospora* genome includes two potential cytosine DNA methyltransferase genes. One, called *dim-2*, is required for all known DNA methylation[26]. The other, called *rid*, is

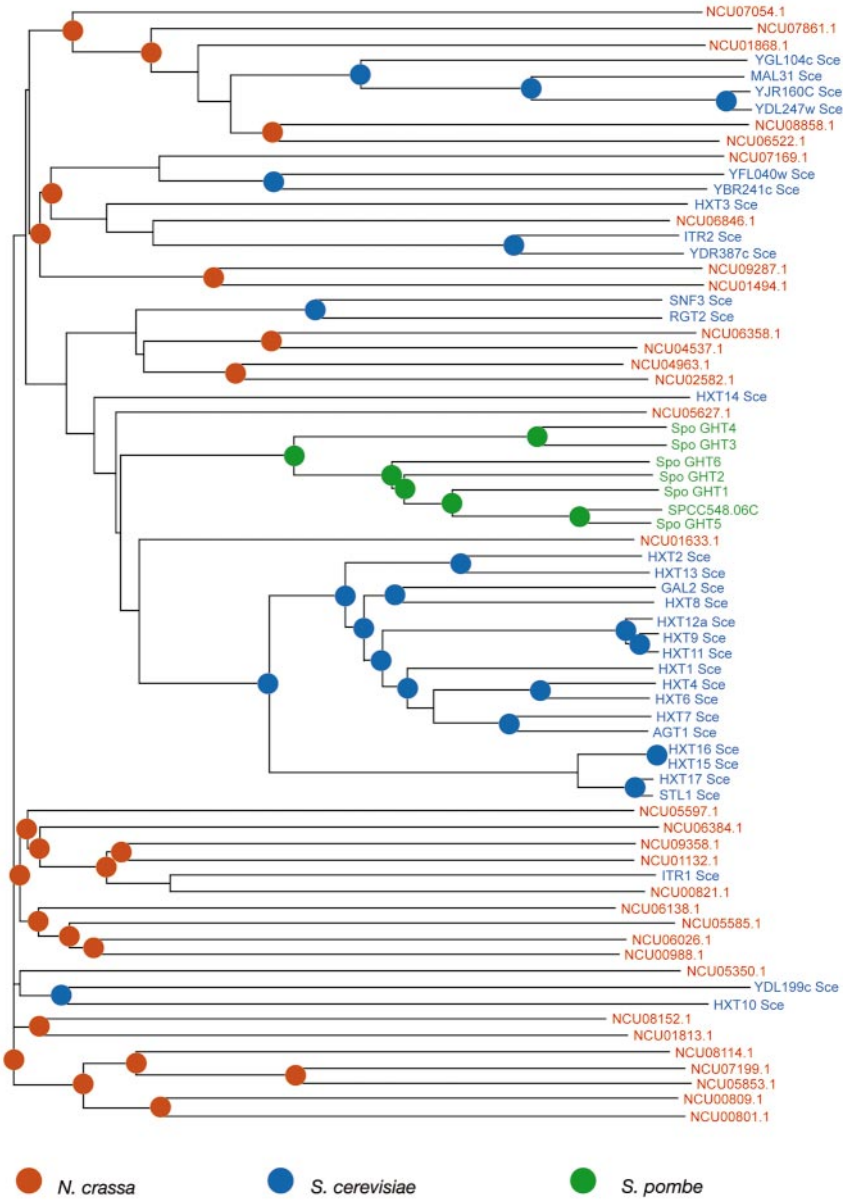


Figure 3 Example of lack of recent duplications in a *Neurospora* gene family. Phylogenetic tree of major facilitator superfamily (MFS) sugar transporters from *S. cerevisiae*, *S. pombe* and *Neurospora*. Coloured dots represent branching points between predicted paralogous genes in *Neurospora* (red), *S. cerevisiae* (blue) and *S. pombe* (green). In contrast to both yeast species, *Neurospora* transporters contain no predicted instances of recent paralogous duplication.

© 2003 Nature Publishing Group

required for RIP and is a member of a family found thus far only in filamentous fungi[27]. In *Neurospora*, an estimated 1.5% of cytosines are methylated[28,29], and it has been suggested that nearly all DNA methylation is a result of RIP[1,30,31].

Plasmid reads for *Neurospora* were sequenced from libraries cloned separately in methylation-tolerant and methylation-intolerant strains of *Escherichia coli*. Although not intended for this purpose, these libraries provided a basis for predicting DNA methylation by comparing the representation of regions in sequence obtained from each library (see Methods). Testing the accuracy of such predictions, we found that 8 of 10 regions predicted to be methylated were experimentally confirmed as such. The predictions thus have good specificity—although they lack sensitivity (see Methods).

The specificity of the predictions provides insight into the pattern of methylation in the *Neurospora* genome. Regions predicted to be methylated show a marked correspondence to regions predicted to be repetitive and RIP-mutated (Fig. 4). Fully 85% correspond to predicted RIP-mutated sequences. However, a small proportion (10%) corresponds to predicted non-repetitive and non-RIP-mutated sequence. In two out of ten such cases, both the methylation and the non-repetitive nature of these sequences were experimentally verified. This raises the possibility that methylation in *Neurospora* may also have non-defence roles, as proposed for higher organisms.

RNA silencing

Post-transcriptional gene silencing (PTGS), or RNA silencing, is widespread among organisms and is increasingly being recognized as a principal switch for controlling eukaryotic gene expression[32]. RNA-silencing pathways are thought to be derived from ancestral natural defence systems directed against invading nucleic acids[33]. Consistent with this, all known PTGS mechanisms share similar components[34].

Neurospora possesses two RNA-silencing pathways. The first, called quelling, acts during vegetative growth. This pathway was uncovered through the study of three genes, *qde-1*, *qde-2* and *qde-3*, coding respectively for an RNA-dependent RNA polymerase (RdRP), an argonaute and a RecQ helicase[35]. The second pathway, called meiotic silencing, acts during sexual reproduction[36,37]. Before our analysis, a gene called *sad-1*, encoding an RdRP, had been identified for this pathway[38].

Our analysis of the *Neurospora* genome sequence uncovered several additional genes implicated in RNA silencing (Table 2). These include one RdRP, one argonaute-like protein and one RecQ-like helicase, as well as two dicer-like ribonucleases. A phylogenetic analysis (Supplementary S7) of the predicted RdRPs, argonaute-like proteins and dicer-like proteins indicates that the *Neurospora* genes comprise two paralogous sets. One set includes the three *qde* genes and is thus predicted to correspond to the quelling pathway. The other set includes *sad-1*, and in phylogenetic trees these genes branch consistently with those of the single pathway observed in *S. pombe*[37,39]. On the basis of this analysis, we predict that one of the identified dicers, *Sms-3*, belongs phylogenetically to the meiotic silencing pathway, whereas the other, *dcl-2*, belongs to the quelling pathway (Table 2). In addition, we predicted that the identified argonaute, *Sms-2*, also belongs phylogenetically to the meiotic silencing pathway. Subsequent experimental work has supported roles for *Sms-2* (ref. 40) and *Sms-3* (M. McLaughlin, D. W. Lee, R. Pratt and R. Aramayo, manuscript in preparation) in meiotic silencing. Taken together, these results suggest that meiotic silencing and quelling represent two phylogenetically distinct RNA-dependent silencing pathways. We further hypothesize that both might have evolved from a single ancestral RNA-silencing pathway.

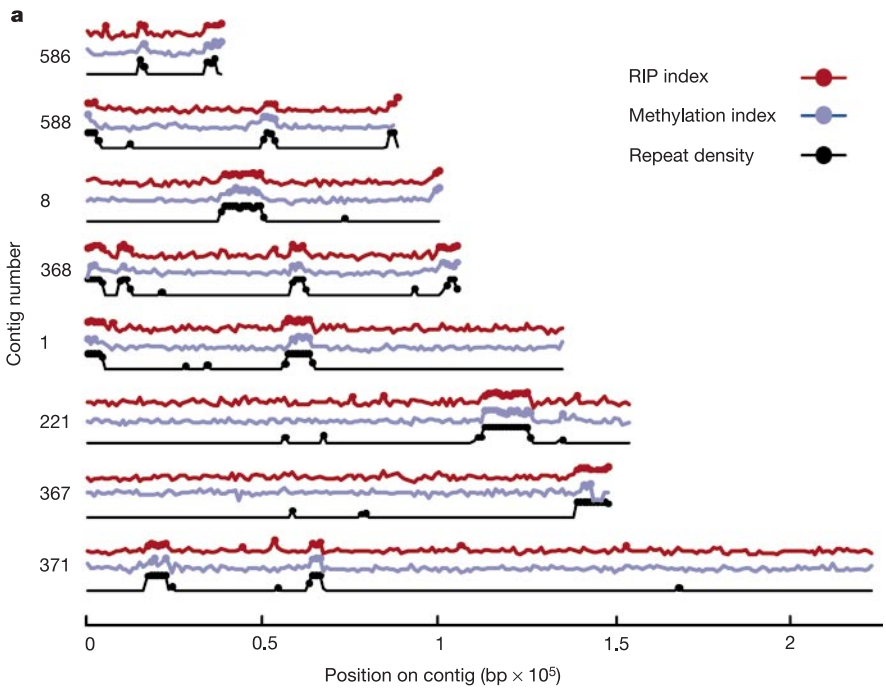


Figure 4 Correspondence between predicted RIP, methylation and repetitive DNA. Prediction of RIP, methylation and repeat sequence in 1-kb windows for selected contigs. Red lines plot the TpA/ApT RIP index (see Methods); red dots indicate windows predicted to be RIP-mutated (TpA/ApT > 1.2). Blue lines plot the proportion of reads from the methyl-tolerant library; blue dots indicate windows predicted to be methylated based on >70% methyl-tolerant reads (see Methods). Black lines plot repeat content as a fraction of nucleotides in each window that is in repetitive sequence; black dots indicate windows with >50% repeat sequence. Contigs were selected to illustrate regions predicted as methylated.

© 2003 Nature Publishing Group

Fungal biology and evolution

The *Neurospora* genome sequence provides an opportunity to study the genetic basis underlying the extraordinary biochemical and metabolic diversity exhibited by a filamentous fungus. Our analysis of the genome sequence has resulted in a number of surprising insights into the biology and evolution of *Neurospora* and other filamentous fungi.

Cell signalling and environmental responses

Discovery of putative red-light-sensing genes

Blue light is an important regulator of *Neurospora* growth and development, affecting the circadian rhythm of conidiation, carotenogenesis of hyphae and numerous facets of sexual development[41]. Although *Neurospora* photobiology has been studied intensively for more than two decades, the genome sequence has nonetheless revealed a number of previously uncharacterized sequences with similarity to blue-light-sensing genes, including both a cryptochrome homologue and a gene whose product contains a single PAS/LOV-type domain associated with light sensing.

Furthermore, *Neurospora* possesses two putative phytochrome homologues most similar to bacteriophytochromes—genes known for their role in red light sensing in prokaryotes—and a putative homologue of the *Aspergillus nidulans velvet* gene implicated in the regulation of both red and blue light responses. The presence of these genes is unexpected given that no red light photobiology has been described for *Neurospora* so far. It has been shown recently that in addition to red light sensing, some *Arabidopsis* phytochromes associate with cryptochromes to have a role in blue light sensing and signalling[42]. Therefore, the two phytochromes and the velvet homologue may also regulate this aspect of *Neurospora* photobiology.

Importance of two-component signalling in filamentous fungi

Mitogen-activated protein kinase (MAPK) pathways integrate signals from multiple receptor pathways including two-component signalling systems[43]. The basic two-component system consists of a histidine kinase and a cognate response regulator. The nine MAPK pathway proteins identified in the *Neurospora* genome sequence (Fig. 5) correspond to those found in *S. pombe* and *S. cerevisiae*, indicating that the basic MAPK machinery is conserved between these species. In contrast, *Neurospora* has a significantly expanded complement of 11 histidine kinases, as compared with one in *S. cerevisiae* and three in *S. pombe*. Two of the 11 genes have been characterized previously in *Neurospora*[44], whereas a third is similar to proteins in *Aspergillus fumigatus* and *A. nidulans* that affect conidiation (L. A. Alex and M. I. Simon, unpublished observations; see also ref. 44). Functions for the remaining genes are unknown, although seven (including the two phytochromes discussed above) contain PAS/PAC domains, implicating them in oxygen and light responses. This number of histidine kinases suggests a larger role than previously expected, and reveals filamentous fungi to be more similar in this regard to plants, where two-component systems are abundant, than to animals, where these systems are absent.

A new family of G-protein-coupled receptors

Eukaryotic cells sense many environmental stimuli through seven-transmembrane-helix, G-protein-coupled receptors (GPCRs)[45]. Our analysis indicates that *Neurospora* possesses ten predicted seven-transmembrane-helix proteins (Fig. 5), three of which belong to a new class not previously identified in any fungus. These three genes encode proteins similar to cyclic AMP GPCRs from the protists *Dictyostelium discoideum*[46] and *Polysphodylium pallidum*, and also to predicted proteins from *Arabidopsis thaliana*[47] and *Caenorhabditis elegans*. The *D. discoideum* proteins sense cAMP levels during chemotaxis and multicellular development[48]. This suggests a possible analogous function in *Neurospora*. The existence of an extracellular cAMP signalling pathway has never been demonstrated previously in any fungal system.

In support of this hypothesis, along with the presence of putative cAMP receptors, *Neurospora* was found to possess the full complement of proteins required for the synthesis and degradation of cAMP. Furthermore, *Neurospora* wild-type strains accumulate cAMP in the extracellular medium[49], although a role in extracellular signalling has not been established. Taken together, these data suggest the possibility that cAMP or a related molecule may serve as an extracellular signal in *Neurospora*.

Ca^{2+} sensory transduction in filamentous fungi

A considerable body of evidence, primarily from pharmacological studies, indicates that Ca^{2+} signalling regulates numerous processes in filamentous fungi[50]. However, the identification of the main components of even one Ca^{2+}-mediated response pathway in filamentous fungi has remained elusive. The genome sequence of *Neurospora* has provided over 25 of the proteins likely to be necessary for Ca^{2+} signalling in filamentous fungi (Fig. 5).

A notable difference between Ca^{2+} signalling in *Neurospora* as compared with plants and animals was revealed by the genome sequence. An important aspect of Ca^{2+} signalling in plant and animal cells involves Ca^{2+} release from internal stores. This is commonly mediated by the second messengers inositol-1,4,5-trisphosphate ($InsP_3$) and cADP ribose, or by Ca^{2+}-induced Ca^{2+} release[51]. $InsP_3$ is present within *Neurospora* hyphae[52], and physiological evidence including intracellular membrane-associated, $InsP_3$-activated Ca^{2+} channel activity supports a role in Ca^{2+} signalling[53,54]. In spite of this, *Neurospora* (and *S. cerevisiae*) lacks recognizable $InsP_3$ receptors. In addition, neither ADP ribosyl cyclase nor ryanodine receptor proteins, principal components of Ca^{2+} release mechanisms in plant and animal cells, are found in *Neurospora*. These observations raise the question of whether other

Table 2 ***Neurospora* has two RNA-silencing pathways**

Predicted protein	*Neurospora*	*A. fumigatus**	*S. pombe*†	Pathway‡
RNA-directed RNA polymerases	*qde-1* (NCU07534.1)	*rrpA* (contig 158)	–	Quelling
	sad-1 (NCU02178.1)	*rrpB* (contig 472)	*rdp1*$^+$ (SPAC6F12.09)	Meiotic silencing
	rrp-3 (NCU08435.1)	–	–	Unknown
Argonaute-like, related to translation initiation factors	*qde-2* (NCU04730.1)	*ppdA* (contig 720)	–	Quelling
	Sms-2 (NCU09434.1)	*ppdB* (contig 196)	*ago1*$^+$ (SPCC736.11)	Meiotic silencing
Dicer-like, related to SFII-RNAse III ribonucleases of the carpel factory	*dcl-2* (NCU06766.1)	*dclB* (contig 618)	–	Quelling
	Sms-3 (NCU08270.1)	*dclA* (contig 310)	*dcr1*$^+$ (SPCC584.10C)	Meiotic silencing
RecQ helicase-like, related to Bloom's and Werner syndrome helicases	*qde-3* (NCU08598.1)	*rqhA* (contig 443)§	–	Quelling
	RecQ-2 (NCU03337.1)[73]	*rqhB* (contig 58)§	*hus2*$^+$ (SPAC2G11.12)	Unknown

*Unfinished *A. fumigatus* genome project (http://www.tigr.org).
†*Schizosaccharomyces pombe* genome project (http://www.genedb.org).
‡Pathway assigned on the basis of either known experimental data for *qde-1*, *qde-2* and *qde-3* (quelling pathway); or *sad-1*, *Sms-2* and *Sms-3* (meiotic silencing pathway); or predicted on the basis of phylogenetic analysis.
§RecQ helicase-like (*rqh*).

© 2003 Nature Publishing Group

second messenger systems responsible for Ca^{2+} release from internal stores remain to be discovered in filamentous fungi.

Growth and development

Hyphal growth

True hyphae produced by filamentous fungi are tubular structures consisting of cellular compartments that are usually delineated by incomplete septa[5]. In contrast, the pseudohyphae produced by yeasts consist of chains of uninucleate elongated cells[55] with no apparent cytoplasmic continuity. The molecular mechanisms underlying these two modes of growth are not well understood.

The two signalling pathways that regulate pseudohyphal growth in *S. cerevisiae*—the MAPK and cAMP modules—are both conserved in the *Neurospora* genome. In *Candida albicans*, capable of pseudohyphal, true hyphal and budding growth, both pathways are required for true hyphal production, suggesting a similar role in *Neurospora*[56]. However, at least three transcription factors—Tec1p, Flo8p and Sfl1p—specifically required for regulating pseudohyphal development in *S. cerevisiae*[56] were not found in *Neurospora*. Conversely, *Neurospora* possesses a gene with similarity to a transcription factor necessary for hyphal growth in *C. albicans*[36] (Efg1). This transcription factor is not required for pseudohyphal growth in *C. albicans*, nor is the homologous protein in *S. cerevisiae* (Phd1p)[56]. More study of the complex pathways underlying these modes of growth is required. Nonetheless, these data clarify in part the distinctions and similarities between the signalling pathways and regulatory components of hyphal and pseudohyphal growth.

The macroconidiation pathway differs from that in *A. nidulans*

Macroconidia are asexual spores common to filamentous fungi but absent from yeast[5,57]. Components of the macroconidiation pathway have been identified in both *Neurospora* and the filamentous fungus *A. nidulans*, and known upstream signalling proteins seem to be conserved in both species[58]. In contrast, there is little conservation of downstream components between the two fungi. In *Neurospora*, the *acon-2*, *acon-3*, *fld* and *fl* genes are essential for conidiation[5], whereas in *A. nidulans*, the FlbC, FlbD, BrlA, AbaA and WetA gene products are required. Our analysis of the genome sequence revealed that *Neurospora* possesses no FlbC, BrlA or AbaA homologues, and a protein with only very weak similarity to approximately 100 amino acids at the carboxy terminus of WetA. These data make clear that the molecular machinery underlying macroconidiation in *Neurospora* differs significantly from that in *A. nidulans*.

Secondary metabolism

The fungal kingdom produces a vast array of small, bioactive compounds termed secondary metabolites that are best known for their roles as pigments, antibiotics and mycotoxins. With the exception of carotenoid and melanin pigment synthesis, *Neurospora* has not been shown to possess secondary metabolism. It was thus surprising that the *Neurospora* genome sequence revealed a number of putative genes for secondary metabolite production.

Non-ribosomal peptide synthetases

Three predicted non-ribosomal peptide synthetase (NRPS) genes and one NRPS-related gene were identified in the *Neurospora* genome sequence (Fig. 6). By phylogenomic analysis, one NRPS gene is orthologous to an *Aureobasidium pullulans* NRPS. The most closely related NRPS of known function is *sid2* of *Ustilago maydis*, which is responsible for production of hydroxamate siderophores[59]. The remaining two are of unknown function, although one is orthologous to an NRPS in *Magnaporthe grisea*, and the other is

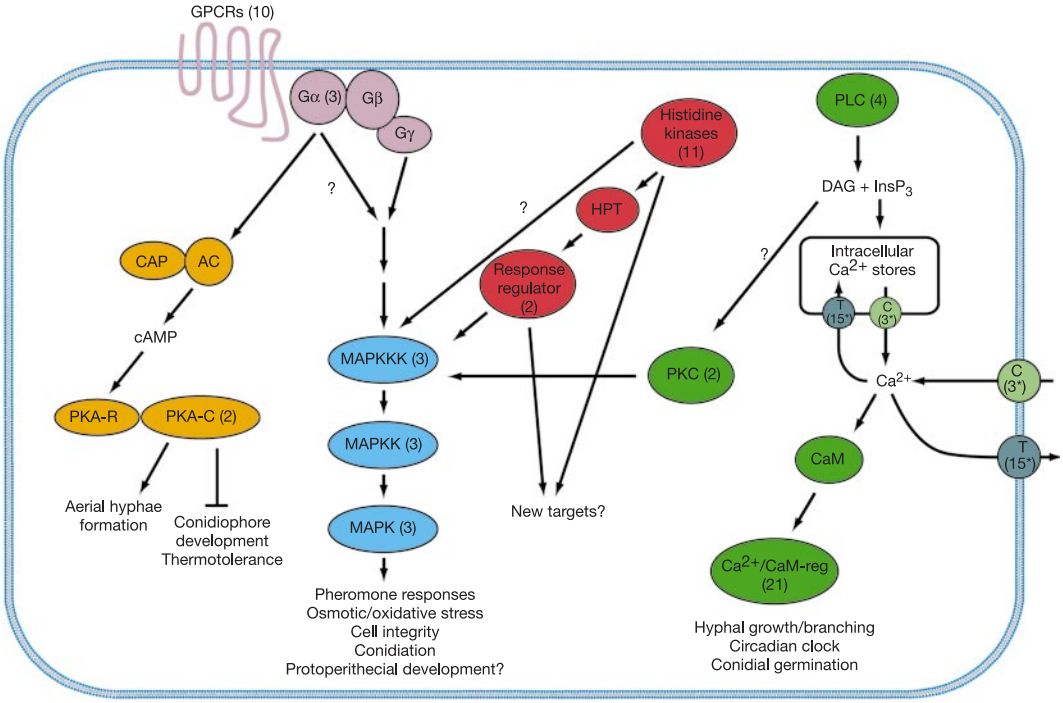


Figure 5 Overview of major intracellular signalling pathways in *Neurospora*. The numbers identified for each gene are in parentheses. An asterisk indicates that the location in the plasma membrane and/or organelle membranes is not determined. AC, adenylyl cyclase; C, Ca^{2+} channel protein; CaM, calmodulin; Ca^{2+}/CaM-reg, calcium- and calmodulin-regulated protein; CAP, cyclase-associated protein; *DAG* diacylglycerol; GPCR, G-protein-coupled receptor; Gα, G protein α-subunit; Gβ, G protein β-subunit; Gγ, G protein γ-subunit; HPT, histidine-containing phosphotransfer domain protein; MAPK, MAP kinase; MAPKK, MAPK kinase; MAPKKK, MAPKK kinase; PKA-C, protein kinase A catalytic subunit; PKA-R, protein kinase A regulatory subunit; PLC, phospholipase C; PKC, protein kinase C; T, Ca^{2+} transport protein (P-type Ca^{2+} ATPase, H^+/Ca^{2+} exchanger, or Na^+/Ca^{2+} exchanger).

© 2003 Nature Publishing Group

orthologous to an NRPS found in all other filamentous ascomycetes with genome sequence (see Methods). The NRPS-related gene shares 66% amino acid identity with the *CPS1* gene product that contributes to the virulence of *Cochliobolus heterostrophus*, *C. victoriae* and *Gibberella zeae*[60].

Polyketide synthases

Seven polyketide synthase (PKS) genes were identified in the *Neurospora* genome, which could be classified into three groups on the basis of domain structure (Fig. 6). The first class contains genes similar to DHN-melanin PKS genes of the fungi *Exophila dermatitidis*[61], *Colletotrichum lagenarium*[62] and *Alternaria alternata*[63]. Sequence identity to numerous expressed sequence tag (EST) sequences from sexual and perithecial libraries suggest a role in melanin pigment synthesis during sexual development[64]. The genes in the second class are similar in structure to several fungal PKSs, including the *Aspergillus terreus lovF* gene required for lovastatin synthesis. The genes in the third class resemble other fungal genes, including the *A. terreus lovB* gene, which is also required for lovastatin synthesis.

Diterpene metabolism

Diterpenes comprise a diverse group of compounds, primarily in plants and fungi, with roles in defence, pathogenicity and regulation of plant growth. The genome sequence revealed several genes associated with diterpene biosynthesis in other organisms, including a terpene synthase, several genes related to gibberellin oxidases, and a member of the cytochrome P450 mono-oxygenase gene family. These genes include at least one member of each of the three enzyme classes required for the biosynthesis of gibberellic acid. Gibberellic acid, a normal growth regulator in plants, was first identified as a metabolic product of the plant pathogen *Gibberella fujikuroi*, a relative of *Neurospora* that causes 'foolish seedling' disease in rice[65]. The presence of these genes in *Neurospora* suggests that many components necessary for gibberellic acid production were present in the ancestors of *Neurospora* and *G. fujikuroi*.

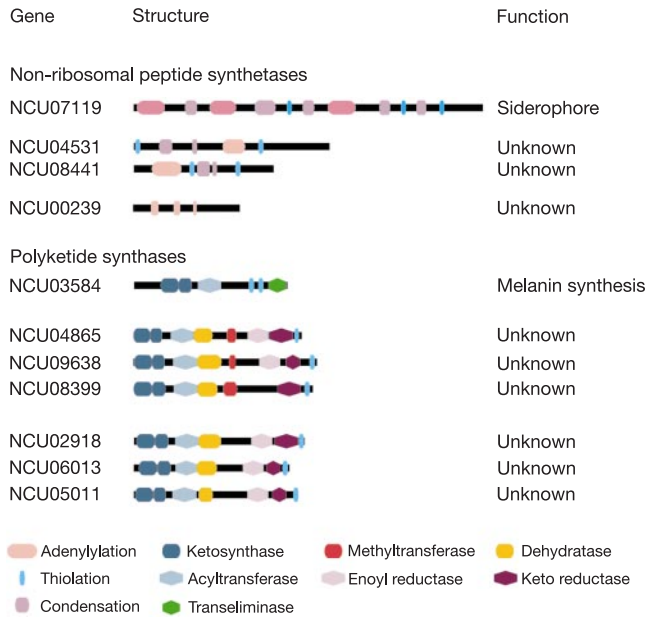


Figure 6 Domain structures of predicted *Neurospora* non-ribosomal peptide synthetase (NRPS) and polyketide synthase (PKS) genes. Domains were predicted using a combination of PFAM searches using HMMER, protein alignments and manual inspection.

We speculate that the secondary metabolism genes identified may have roles in morphogenesis and chemotropism[66], interspecies communication and possibly even chemical defence. The identification of these genes in *Neurospora* suggests that apparent major differences in lifestyles among related fungi, such as pathogenicity, may derive in part from minor modifications of gene function and expression.

Plant pathogenicity and *Neurospora*

The ability to parasitize living plants is widespread throughout the fungal kingdom. Although *Neurospora* is a saprotroph (that is, it feeds on dead or decaying matter), the genome sequence contains numerous genes similar to those required for plant pathogenesis identified in fungal pathogens. In particular, a number of genes were identified that have no known function in other organisms except in pathogenesis (Supplementary S8). *Neurospora* also possesses a wide range of extracellular enzymes capable of digesting plant cell wall polymers, although there is no clear cutinase homologue. Cutin is one of the main layers protecting the epidermis of the leaves of plants, and many, but not all, plant pathogens have cutinase activity. *Neurospora* has a wide range of cytochrome P450 enzymes that are important in some host–pathogen systems for detoxification of plant anti-fungal compounds. In addition, a large number of identified ABC (ATP-binding cassette) and MFS drug efflux systems could have a role in combating toxic plant compounds. The capability to form secondary metabolite members of the PKS, NRPS and terpenoid families, as described above, is present. Also, *Neurospora* contains all signal transduction components implicated in ascomycete pathogenesis that have been described so far. Thus, although *Neurospora* is not known to be a pathogen, the genome sequence has revealed many genes with similarity to those required for pathogenesis.

Discussion

Although *Neurospora* has been studied intensely for over 70 years, the analysis of the genome sequence has provided many new insights into a variety of cellular processes, including cell signalling, growth and differentiation, secondary metabolism and genome defence. The analysis has also uncovered surprising similarities between the saprotrophic *Neurospora* and pathogenic fungi, providing a new perspective on the molecular underpinnings of these lifestyles. Finally, the genome sequence has revealed the remarkable impact of RIP on the evolution of genes in *Neurospora*. Recent reports indicating the apparent presence of RIP in other fungi[67,68] broaden the implications of our findings. The apparent lack of functional gene duplication in *Neurospora* provides a unique opportunity to study other modes of evolution in this experimentally tractable organism.

The genome sequence of *Neurospora* provides only a first glimpse into the genomic basis of the biological diversity of the filamentous fungi. Fungal genome sequences from the many ongoing[69] and planned[70] projects will expand this view as well as provide extraordinary opportunities for comparative analyses. This new era in fungal biology promises to yield insight into this important group of organisms, as well as to provide a deeper understanding of the fundamental cellular processes common to all eukaryotes. □

Methods

Strain and growth conditions

Twenty 5-ml cultures of *N. crassa* wild-type strain N150 (74-OR23-1VA; Fungal Genetics Stock Center 2489) were grown on a shaker in Vogel's minimal medium[5] for 3 days at 32 °C. Tissues were collected, freeze-dried overnight and DNA was extracted as previously described[71]. DNA from the 20 samples was mixed and used for library construction.

Sequencing and assembly

The genome was sequenced by the WGS method. Plasmid (4-kb inserts) and fosmid (40-kb inserts) libraries were generated as described at http://www-genome.wi.mit.edu/. Jumping clone (subclone) libraries with 50-kb inserts were generated as described

© 2003 Nature Publishing Group

elsewhere[72]. *Neurospora* cosmid and BAC clones were obtained from previously constructed libraries[11,21]. Sequencing methods for all clone types are described at http://www-genome.wi.mit.edu/. All inserts were sequenced from both ends to generate paired reads. The sequence coverage generated is shown in Supplementary Information. The sequence was assembled using Arachne[8]. Finished sequence from linkage groups II and V was provided by MIPS and is available at http://mips.gsf.de/proj/neurospora/.

Annotation and analysis

We annotated the *Neurospora* genome using the Calhoun annotation system. The genome sequence was searched against the public protein databases using BLASTX with a threshold value of $E \leq 10^{-5}$. Genes were predicted using a combination of FGENESH, FGENESH+ and Genewise (Supplementary Information). The gene calling programs were validated against a test set of 191 previously characterized *Neurospora* proteins. Predicted genes were validated against ESTs aligned to the genome using SIM4. All predicted genes were searched against the PFAM set of hidden Markov models using the HMMER program and the public protein databases using BLASTP. Transfer RNAs were identified using the tRNAScan-SE program. Multigene families were constructed by searching each annotated gene against every other gene using BLASTP, requiring matches with $E \leq 10^{-5}$ over 60% of the longer gene length, and clustering genes based on single linkage transitive closure. Repeat sequences were detected by searching the genome sequence against itself using CrossMatch, filtering for alignments longer than 200 bp in length, and clustering pairs based on region overlap. Relics of RIP-mutated mobile elements were annotated by manual inspection.

The tree of MFS sugar transporters was created by aligning amino acid sequences using ClustalW, manually trimming to remove ambiguously aligned regions, and applying a neighbour-joining algorithm using PAUP*. RIP-mutated regions were detected by calculating one or both of two different dinucleotide ratios for sequence regions[15]. Regions with TpA/ApT $>$ 2 or (CpA + TpG)/(ApC + GpT) $<$ 0.7 were predicted as RIP-mutated. Prediction of RIP sequence across the genome used only the TpA/ApT ratio, whereas the analysis of coding sequences used both (with a positive prediction by either measure taken as a prediction of RIP). RIP simulations were implemented in Matlab and were based on parameters derived from Table 2 of ref. 16. The simulation code is available on request. During each round of simulated RIP, every cytosine-containing dinucleotide on one strand (selected with equal probabilities) was mutated according to the probabilities: (CA = 0.3, CT = 0.05, CG = 0.01, CC = 0.009). DNA methylation was predicted by calculating the proportion of plasmid reads overlapping 1-kb windows from both the methylation-tolerant and methylation-intolerant libraries. Regions with greater than 70% of reads derived from the methylation-tolerant library were predicted to be methylated. Specificity was estimated as described in the text. Methylation was experimentally assessed using Southern analyses as described elsewhere[30]. Sensitivity was estimated by testing 19 repetitive and RIP-mutated 1-kb regions that were not predicted to be methylated. Of the 19 regions, 14 were in fact methylated. Thus the data provide good specificity but poor sensitivity.

Predicted RNA-silencing genes were aligned with homologues from plants, animals and other fungi using T-Coffee v1.37. C-terminal and amino-terminal regions of low homology were removed and the sequences realigned until alignments started and stopped at regions of unambiguous similarity. Both neighbour-joining trees, using ClustalX, and maximum posterior probability trees, using MrBayes 2.01, were generated and analysed.

Analysis of predicted non-ribosomal peptide synthetases and polyketide synthases made use of genome data for *C. heterostrophus*, *Botryotinia fuckeliana*, *G. verticillioides* and *G. zeae* provided by the Torrey Mesa Research Institute/Syngenta. Additional details, analysis results and the genome sequence are available at http://www-genome.wi.mit.edu/.

Received 24 December 2002; accepted 14 March 2003; doi:10.1038/nature01554.

1. Payen, A. (rapporteur) Extrait d'un rapport adressé à M. Le Maréchal Duc de Dalmatie, Ministre de la Guerre, Président du Conseil, sur une altération extraordinaire du pain de munition. *Ann. Chim. Phys. 3 Ser.* **9,** 5–21 (1843).
2. Shear, C. L. & Dodge, B. O. Life histories and heterothallism of the red bread-mold fungi of the *Monilia sitophila* group. *J. Agric. Res.* **34,** 1019–1042 (1927).
3. Lindegren, C. C. A six-point map of the sex chromosome of *Neurospora crassa*. *J. Genet.* **32,** 243–256 (1936).
4. Beadle, G. W. & Tatum, E. L. Genetic control of biochemical reactions in *Neurospora*. *Proc. Natl Acad. Sci. USA* **27,** 499–506 (1941).
5. Davis, R. H. *Neurospora: Contributions of a Model Organism* (Oxford Univ. Press, New York, 2000).
6. Davis, R. H. & Perkins, D. D. *Neurospora*: a model of model microbes. *Nature Rev. Genet.* **3,** 397–403 (2002).
7. Perkins, D. D., Radford, A. & Sachs, M. S. *The Neurospora Compendium: Chromosomal Loci* (Academic, San Diego, California, 2001).
8. Batzoglou, S. *et al.* ARACHNE: a whole-genome shotgun assembler. *Genome Res.* **12,** 177–189 (2002).
9. Schulte, U., Becker, I., Mewes, H. W. & Mannhaupt, G. Large scale analysis of sequences from *Neurospora crassa*. *J. Biotechnol.* **94,** 3–13 (2002).
10. Free, S. J., Rice, P. W. & Metzenberg, R. L. Arrangement of the genes coding for ribosomal ribonucleic acids in *Neurospora crassa*. *J. Bacteriol.* **137,** 1219–1226 (1979).
11. Aign, V., Schulte, U. & Hoheisel, J. D. Hybridization-based mapping of *Neurospora crassa* linkage groups II and V. *Genetics* **157,** 1015–1020 (2001).
12. Selker, E. U., Cambareri, E. B., Jensen, B. C. & Haack, K. R. Rearrangement of duplicated DNA in specialized cells of *Neurospora*. *Cell* **51,** 741–752 (1987).
13. Selker, E. U. Premeiotic instability of repeated sequences in *Neurospora crassa*. *Annu. Rev. Genet.* **24,** 579–613 (1990).
14. Cambareri, E. B., Jensen, B. C., Schabtach, E. & Selker, E. U. Repeat-induced G-C to A-T mutations in *Neurospora*. *Science* **244,** 1571–1575 (1989).
15. Margolin, B. S. *et al.* A methylated *Neurospora* 5S rRNA pseudogene contains a transposable element inactivated by repeat-induced point mutation. *Genetics* **149,** 1787–1797 (1998).
16. Watters, M. K., Randall, T. A., Margolin, B. S., Selker, E. U. & Stadler, D. R. Action of repeat-induced point mutation on both strands of a duplex and on tandem duplications of various sizes in *Neurospora*. *Genetics* **153,** 705–714 (1999).
17. Cambareri, E. B., Singer, M. J. & Selker, E. U. Recurrence of repeat-induced point mutation (RIP) in *Neurospora crassa*. *Genetics* **127,** 699–710 (1991).
18. Rountree, M. R. & Selker, E. U. DNA methylation inhibits elongation but not initiation of transcription in *Neurospora crassa*. *Genes Dev.* **11,** 2383–2395 (1997).
19. Irelan, J. T., Hagemann, A. T. & Selker, E. U. High frequency repeat-induced point mutation (RIP) is not associated with efficient recombination in *Neurospora*. *Genetics* **138,** 1093–1103 (1994).
20. Perkins, D. D., Margolin, B. S., Selker, E. U. & Haedo, S. D. Occurrence of repeat induced point mutation in long segmental duplications of *Neurospora*. *Genetics* **147,** 125–136 (1997).
21. Kelkar, H. S. *et al.* The *Neurospora crassa* genome: cosmid libraries sorted by chromosome. *Genetics* **157,** 979–990 (2001).
22. Hutchison, C. A. *et al.* Global transposon mutagenesis and a minimal Mycoplasma genome. *Science* **286,** 2165–2169 (1999).
23. Ohno, S. *Evolution by Gene Duplication* (Springer, New York, 1970).
24. Selker, E. U. *et al.* Dispersed 5S RNA genes in *N. crassa*: structure, expression and evolution. *Cell* **24,** 819–828 (1981).
25. Selker, E. U. *et al.* The methylated component of the *Neurospora crassa* genome. *Nature* **422,** 893–897 (2003).
26. Kouzminova, E. & Selker, E. U. dim-2 encodes a DNA methyltransferase responsible for all known cytosine methylation in *Neurospora*. *EMBO J.* **20,** 4309–4323 (2001).
27. Freitag, M., Williams, R. L., Kothe, G. O. & Selker, E. U. A cytosine methyltransferase homologue is essential for repeat-induced point mutation in *Neurospora crassa*. *Proc. Natl Acad. Sci. USA* **99,** 8802–8807 (2002).
28. Russell, P. J., Rodland, K. D., Rachlin, E. M. & McCloskey, J. A. Differential DNA methylation during the vegetative life cycle of *Neurospora crassa*. *J. Bacteriol.* **169,** 2902–2905 (1987).
29. Foss, H. M., Roberts, C. J., Claeys, K. M. & Selker, E. U. Abnormal chromosome behavior in *Neurospora* mutants defective in DNA methylation. *Science* **262,** 1737–1741 (1993).
30. Selker, E. U., Fritz, D. Y. & Singer, M. J. Dense nonsymmetrical DNA methylation resulting from repeat-induced point mutation in *Neurospora*. *Science* **262,** 1724–1728 (1993).
31. Singer, M. J., Marcotte, B. A. & Selker, E. U. DNA methylation associated with repeat-induced point mutation in *Neurospora crassa*. *Mol. Cell. Biol.* **15,** 5586–5597 (1995).
32. Wolffe, A. P. & Matzke, M. A. Epigenetics: regulation through repression. *Science* **286,** 481–486 (1999).
33. Zamore, P. D. Ancient pathways programmed by small RNAs. *Science* **296,** 1265–1269 (2002).
34. Hutvagner, G. & Zamore, P. D. RNAi: nature abhors a double-strand. *Curr. Opin. Genet. Dev.* **12,** 225–232 (2002).
35. Catalanotto, C., Azzalin, G., Macino, G. & Cogoni, C. Involvement of small RNAs and role of the *qde* genes in the gene silencing pathway in *Neurospora*. *Genes Dev.* **16,** 790–795 (2002).
36. Aramayo, R. & Metzenberg, R. L. Meiotic transvection in fungi. *Cell* **86,** 103–113 (1996).
37. Shiu, P. K., Raju, N. B., Zickler, D. & Metzenberg, R. L. Meiotic silencing by unpaired DNA. *Cell* **107,** 905–916 (2001).
38. Shiu, P. K. & Metzenberg, R. L. Meiotic silencing by unpaired DNA. Properties, regulation and suppression. *Genetics* **161,** 1483–1495 (2002).
39. Volpe, T. A. *et al.* Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297,** 1833–1837 (2002).
40. Lee, D., Pratt, R. J., McLaughlin, M., Aramayo, R. An Argonaute-like protein is required for meiotic silencing. *Genetics* (in the press).
41. Linden, H., Ballario, P., Arpaia, G. & Macino, G. Seeing the light: news in *Neurospora* blue light signal transduction. *Adv. Genet.* **41,** 35–54 (1999).
42. Devlin, P. F. & Kay, S. A. Circadian photoperception. *Annu. Rev. Physiol.* **63,** 677–694 (2001).
43. Gustin, M. C., Albertyn, J., Alexander, M. & Davenport, K. MAP kinase pathways in the yeast *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.* **62,** 1264–1300 (1998).
44. Pott, G. B., Miller, T. K., Bartlett, J. A., Palas, J. S. & Selitrennikoff, C. P. The isolation of *FOS-1*, a gene encoding a putative two-component histidine kinase from *Aspergillus fumigatus*. *Fungal Genet. Biol.* **31,** 55–67 (2000).
45. Dohlman, H. G., Thorner, J., Caron, M. G. & Lefkowitz, R. J. Model systems for the study of seven-transmembrane-segment receptors. *Annu. Rev. Biochem.* **60,** 653–688 (1991).
46. Klein, P. S. *et al.* A chemoattractant receptor controls development in *Dictyostelium discoideum*. *Science* **241,** 1467–1472 (1988).
47. Josefsson, L. G. & Rask, L. Cloning of a putative G-protein-coupled receptor from *Arabidopsis thaliana*. *Eur. J. Biochem.* **249,** 415–420 (1997).
48. Aubry, L. & Firtel, R. Integration of signalling networks that regulate *Dictyostelium* differentiation. *Annu. Rev. Cell Dev. Biol.* **15,** 469–517 (1999).
49. Ivey, F. D., Yang, Q. & Borkovich, K. A. Positive regulation of adenylyl cyclase activity by a Gαi homolog in *Neurospora crassa*. *Fungal Genet. Biol.* **26,** 48–61 (1999).
50. Gadd, G. M. in *The Growing Fungus* (eds Gow, N. A. R. & Gadd, G. M.) 183–210 (Chapman & Hall, London, 1994).
51. Bootman, M. D. *et al.* Calcium signalling—an overview. *Semin. Cell Dev. Biol.* **12,** 3–10 (2001).
52. Lakin-Thomas, P. L. Effects of inositol starvation on the levels of inositol phosphates and inositol lipids in *Neurospora crassa*. *Biochem. J.* **292,** 805–811 (1993).
53. Silverman-Gavrila, L. B. & Lew, R. R. Regulation of the tip-high [Ca^{2+}] gradient in growing hyphae of the fungus *Neurospora crassa*. *Eur. J. Cell Biol.* **80,** 379–390 (2001).
54. Cornelius, G., Gebauer, G. & Techel, D. Inositol trisphosphate induces calcium release from *Neurospora crassa* vacuoles. *Biochem. Biophys. Res. Commun.* **162,** 852–856 (1989).
55. Gimeno, C. J., Ljungdahl, P. O., Styles, C. A. & Fink, G. R. Unipolar cell divisions in the yeast *S. cerevisiae* lead to filamentous growth: regulation by starvation and RAS. *Cell* **68,** 1077–1090 (1992).
56. Lengeler, K. B. *et al.* Signal transduction cascades regulating fungal development and virulence. *Microbiol. Mol. Biol. Rev.* **64,** 746–785 (2000).
57. Adams, T. H., Wieser, J. K. & Yu, J. H. Asexual sporulation in *Aspergillus nidulans*. *Microbiol. Mol. Biol. Rev.* **62,** 35–54 (1998).
58. Ivey, F. D., Kays, A. M. & Borkovich, K. A. Shared and independent roles for a $G\alpha_i$ protein and adenylyl cyclase in regulating development and stress responses in *Neurospora crassa*. *Eukaryot. Cell* **1,** 634–642 (2002).

© 2003 Nature Publishing Group

59. Yuan, W., Gentil, G., Budde, A. & Leong, S. Characterization of the *Ustilago maydis sid2* gene encoding a multidomain peptide synthetase in the ferrichrome biosynthetic gene cluster. *J. Bacteriol.* **183,** 4040–4051 (2001).
60. Lu, S.-W., Yoder, O. & Turgeon, B. G. Highly conserved virulence-related CPS1 homologs from plant and human pathogenic fungi. *Fungal Genet. News.* **48,** A246 (2001).
61. Feng, B. *et al.* Molecular cloning and characterization of *WdPKS1*, a gene involved in dihydroxynaphthalene melanin biosynthesis and virulence in *Wangiella* (*Exophiala*) *dermatitidis*. *Infect. Immun.* **69,** 1781–1794 (2001).
62. Takano, Y. *et al.* Structural analysis of PKS1, a polyketide synthase gene involved in melanin biosynthesis in *Colletotrichum lagenarium*. *Mol. Gen. Genet.* **249,** 162–167 (1995).
63. Takano, Y., Kubo, Y., Kawamura, C., Tsuge, T. & Furusawa, I. The *Alternaria alternata* melanin biosynthesis gene restores appressorial melanization and penetration of cellulose membranes in the melanin-deficient albino mutant of *Colletotrichum lagenarium*. *Fungal Genet. Biol.* **21,** 131–140 (1997).
64. Nelson, M. A. *et al.* Expressed sequences from conidial, mycelial, and sexual stages of *Neurospora crassa*. *Fungal Genet. Biol.* **21,** 348–363 (1997).
65. Kurosawa, E. Experimental studies on the nature of the substance excreted by the 'bakanae' fungus. *Trans. Nat. Hist. Soc. Formosa* **16,** 213–227 (1926) (in Japanese).
66. Vigfusson, N. & Cano, R. Artificial induction of the sexual cycle of *Neurospora crassa*. *Nature* **249,** 383–385 (1974).
67. Graia, F. *et al.* Genome quality control: RIP (repeat-induced point mutation) comes to Podospora. *Mol. Microbiol.* **40,** 586–595 (2001).
68. Ikeda, K. *et al.* Repeat-induced point mutation (RIP) in *Magnaporthe grisea*: implications for its sexual cycle in the natural field context. *Mol. Microbiol.* **45,** 1355–1364 (2002).
69. TIGR Microbial Database ⟨http://www.tigr.org/tdb/mdb/mdbinprogress.html⟩ (2003).
70. Fungal Genome Initiative ⟨http://www-genome.wi.mit.edu/seq/fgi/⟩ (2002).
71. Luo, Z., Freitag, M. & Sachs, M. S. Translational regulation in response to changes in amino acid availability in *Neurospora crassa*. *Mol. Cell. Biol.* **15,** 5235–5245 (1995).
72. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
73. Pickford, A., Braccini, L., Macino, G. & Cogoni, C. The QDE-3 homologue RecQ-2 co-operates with QDE-3 in DNA repair in *Neurospora crassa*. *Curr. Genet.* **42,** 220–227 (2003).

Supplementary Information accompanies the paper on *Nature*'s website (**http://www.nature.com/nature**).

Acknowledgements The authors would like to thank J. Arnold, H. Inoue, G. Turner, B. Bowman, P. Harriman and P. Youderian for their support. We also thank Lion Bioscience, V. Aign and J. Hoheisel for making available the BAC libraries used during sequencing; M. Karasz for developing the web pages hosting the *Neurospora* Interpro analysis at MIPS; the Torrey Mesa Research Institute/Syngenta for providing genome data for *C. heterostrophus*, *B. fuckeliana*, *G. verticillioides* and *G. zeae*; and all members of the Whitehead Institute/MIT Center for Genome Research sequencing group. Funding for the *Neurospora* genome project was provided by the National Science Foundation. Additional funding was provided by the Deutsche Forschungsgemeinschaft, The Israel Science Foundation and the National Institutes of Health.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.E.G (e-mail: jgalag@mit.edu) or B.B. (e-mail: bwb@genome.wi.mit.edu). The whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the project accession AABX00000000. The version described in this paper is the first version, AABX01000000.

Affiliations for authors: 1, Whitehead Institute Center for Genome Research, 320 Charles Street, Cambridge, Massachusetts 02141, USA; 2, Department of Plant Pathology, University of California, Riverside, California 92521, USA; 3, Institute of Molecular Biology, University of Oregon, Eugene, Oregon 97403, USA; 4, Institute of Cell and Molecular Biology, University of Edinburgh, Edinburgh EH9 3JH, UK; 5, Celera Genomics, Rockville, Maryland 20850, USA; 6, Department of Biology, Texas A & M University, College Station, Texas 77843, USA; 7, Department of Biology, University of New Mexico, Albuquerque, New Mexico 87131, USA; 8, Department of Cellular and Structural Biology, University of Colorado Health Sciences Center, Denver, Colorado 80262, USA; 9, Department of Microbiology, University of Kansas Medical School, Kansas City, Kansas 66160, USA; 10, Department of Zoology, University of Florida, Gainesville, Florida 32611-8525, USA; 11, Department of Plant Pathology and Microbiology, Faculty of Agricultural, Food and Environmental Quality Sciences, The Hebrew University of Jerusalem, Rehovot 76100, Israel; 12, Institute of Biochemistry, Heinrich Heine University, 40225 Düsseldorf, Germany; 13, Laboratory of Plant Molecular Biology, Rockefeller University, New York, New York 10021, USA; 14, Department of Genome Oriented Bioinformatics, Technical University of Munich, Wissenschaftzentrum Weihenstephan, 85350 Freising-Weihenstephan, Germany; 15, Institute for Bioinformatics (MIPS), GSF-National Research Center for Environment and Health, Ingolstaedter Landstrasse 1, 85764 Neuherberg, Germany; 16, T.H. Morgan School of Biological Sciences, University of Kentucky, Lexington, Kentucky 40506, USA; 17, The Institute for Cellular and Molecular Biology, The University of Texas at Austin, Austin, Texas 78712, USA; 18, The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20878, USA; 19, Department of Plant and Microbial Biology, University of California, Berkeley, California 94720, USA; 20, Department of Chemistry and Biochemistry, University of California, Los Angeles, California 90095, USA; 21, Dipartimento Biotecnologie Cellulari ed Ematologia, Universita' di Roma La Sapienza, Rome, Italy; 22, School of Biological Sciences, Flinders University, P.O. Box 2100, Adelaide, South Australia 5001; 23, Department of Molecular Genetics, The Ohio State University, Columbus, Ohio 43210, USA; 24, Department of Molecular Genetics, The Ohio State University, Columbus, Ohio 43211, USA; 25, Department of Plant Pathology, University of Arizona, Tucson, Arizona 85721, USA; 26, School of Biological Sciences, University of Missouri-Kansas City, Kansas City, Missouri 64110, USA; 27, Department of Genetics, Dartmouth Medical School, Hanover, New Hampshire 03755, USA; 28, School of Biology, Leeds University, Leeds LS2 9JT, UK; 29, Department of Chemistry, California State Polytechnic University Pomona, Pomona, California 91768, USA; 30, Department of Plant Pathology and Microbiology, Texas A & M University, College Station, Texas 77843, USA; 31, Department of Environmental and Biomolecular Systems, OGI School of Science and Engineering, Oregon Health and Science University, Beaverton, Oregon 97006, USA; 32, Department of Biology, MIT, Cambridge, Massachusetts 02139, USA

 © 2003 Nature Publishing Group

Sequencing and comparison of yeast species to identify genes and regulatory elements

Manolis Kellis*†, Nick Patterson*, Matthew Endrizzi*, Bruce Birren* & Eric S. Lander*‡

** Whitehead/MIT Center for Genome Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA*
† Department of Computer Science and ‡ Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Identifying the functional elements encoded in a genome is one of the principal challenges in modern biology. Comparative genomics should offer a powerful, general approach. Here, we present a comparative analysis of the yeast *Saccharomyces cerevisiae* based on high-quality draft sequences of three related species (*S. paradoxus*, *S. mikatae* and *S. bayanus*). We first aligned the genomes and characterized their evolution, defining the regions and mechanisms of change. We then developed methods for direct identification of genes and regulatory motifs. The gene analysis yielded a major revision to the yeast gene catalogue, affecting approximately 15% of all genes and reducing the total count by about 500 genes. The motif analysis automatically identified 72 genome-wide elements, including most known regulatory motifs and numerous new motifs. We inferred a putative function for most of these motifs, and provided insights into their combinatorial interactions. The results have implications for genome analysis of diverse organisms, including the human.

Extracting the complete functional information encoded in a genome—including the genic, regulatory and structural elements—is a central challenge in biological research. Ideally, one would be able to extract this information directly from the DNA sequence itself without recourse to extensive experimentation. At present, however, our ability to directly interpret genomes is rudimentary.

De novo identification of the complete set of protein-coding sequences remains imperfect, even in well-studied organisms with compact genomes. The yeast *Saccharomyces cerevisiae*, for example, has enjoyed a complete genome sequence since 1996 (ref. 1). However, the number of biologically significant open reading frames (ORFs) has been the subject of considerable debate[2–6]—with proposed counts ranging from roughly 4,800 to 6,400. The situation is worse for organisms with large, complex genomes, such as mammals. Because the results of *de novo* gene prediction are not sufficiently reliable, gene identification is typically based on a comparison of genomic sequence to known messenger RNA transcripts or to genes in other organisms[7].

Direct identification of the repertoire of regulatory elements in a genome is even more challenging. The best approach so far has relied on clustering genes into functionally related subsets (for example, genes with a common biochemical function or coordinated transcription) and then searching for common sequence motifs in the general vicinity of the genes (for example, using computer programs such as MEME[8] or AlignACE[9]; reviewed in ref. 10). However, the approach has notable drawbacks. It requires extensive prior knowledge about gene function, which is often not available and never comprehensive, and is inherently limited in its ability to extract information from a single genome.

Comparative genome analysis of related species should provide a powerful and general approach for identifying functional elements without previous knowledge of function. Because evolution relentlessly tinkers with genome sequence and tests the results by natural selection, such elements should stand out by virtue of having a greater degree of conservation across related species. The approach has the advantage that one can increase its power by increasing the number of species studied.

Recent studies have demonstrated the potential power of comparative genomic comparison. Cross-species conservation has previously been used to identify putative genes or regulatory elements in small genomic regions[11–14]. Light sampling of whole-genome sequence has been studied as a way to improve genome annotation[5,15]. Complete microbial genomes have been compared to identify pathogenic and other genes[16–19]. Genome-wide comparison has been used to estimate the proportion of the mammalian genome under selection[7].

The goal of this paper is to develop and apply general approaches for systematic analysis of protein-coding and regulatory elements within any genome by means of whole-genome comparisons with several related species. We focused on *S. cerevisiae*, because it is the best-studied eukaryote and thus provides the best setting to test the approach. High-quality draft genome sequences from three related *Saccharomyces* species were produced, aligned and analysed to identify protein-coding genes and regulatory elements.

The comparative analysis indicates that *S. cerevisiae* contains only 5,538 genes encoding ≥100 amino acids. It proposes the elimination of about 500 previously annotated ORFs; the merger of 33 pairs of consecutive genes; the redefinition of start or stop codons in at least ~300 cases; and the existence of ~60 new introns. It also identifies 188 genes encoding between 50 and 99 amino acids, including 43 that had previously escaped notice.

Analysis of intergenic regions reveals 72 well-conserved sequence motifs that occur frequently throughout the yeast genome. The set includes most known regulatory motifs, together with a comparable number of previously uncharacterized motifs. We also identified combinatorial relationships among these motifs.

The results demonstrate that comparative genomic analysis of multiple related species has substantial power to identify key functional elements without previous biological knowledge.

Genome structure and evolution

Genome sequencing

Saccharomyces paradoxus, *S. mikatae* and *S. bayanus*, which are separated from *S. cerevisiae* by an estimated 5–20 million years of evolution, were selected for sequencing on the basis of their phylogeny. The three species were found to have sufficient sequence

© 2003 Nature Publishing Group

similarity to *S. cerevisiae* to allow orthologous regions to be aligned reliably, but sufficient sequence divergence to allow many functional elements to be recognized by their greater degree of conservation by a four-way species comparison. All three are members of the *Saccharomyces sensu stricto* group.

For each species, we generated about sevenfold redundant coverage in paired end sequences from a whole-genome shotgun plasmid library (see Methods). The information was then assembled with the Arachne computer program[20] into contigs (continuous blocks of uninterrupted sequence) and scaffolds (contigs linked by paired forward–reverse reads from the same plasmid) to yield a draft sequence (Table 1).

The draft genome sequence of each species has long-range continuity (N50 scaffold length of 230–500 kilobases (kb), as compared with 942 kb for the finished sequence of *S. cerevisiae*), relatively short sequence gaps (0.6–0.8 kb, which is small compared with a typical gene), and contains most of the genome (~95%). The data are freely available through public sequence databases and through the *Saccharomyces* Genome Database (SGD) maintained at Stanford (http://genome-www.stanford.edu/Saccharomyces/).

Genome alignment

We sought to align the *S. cerevisiae* genome with that of each of the other three species. The first step was to produce a large-scale alignment of genomic regions. Each ORF in *S. cerevisiae* (a total of 6,235 ORFs in the current SGD annotation; see below) was analysed relative to each of the three related species to determine whether it had a clear one-to-one match, multiple ambiguous matches or no match (see Methods). The one-to-one matches were used as orthologous landmarks to define the large-scale alignment, which showed strong conservation of synteny across the species (see below). The second step involved generating local nucleotide-level alignments around each orthologous ORF (see Methods).

Genome evolution at large scale

Most ORFs have clear one-to-one orthologous matches in each species, providing a dense set of landmarks (average spacing ~2 kb) to define blocks of conserved synteny covering essentially the entire genome (Fig. 1). For a small number of ORFs (211), the correspondence is ambiguous, however. These ambiguous matches almost all reflect local gene-family expansion or contraction.

Most of the ambiguities are markedly clustered in telomeric regions (Fig. 2). More than 80% fall into one of 32 clusters of two or more genes (average size ~18 kb, together comprising ~4% of the genome), which correspond nearly perfectly to the 32 telomeric regions of the 16 chromosomes of *S. cerevisiae*. Only one telomeric region lacks a cluster (chromosome VI-R). And only one cluster does not lie in telomeric regions in *S. cerevisiae*: it is a recent insertion of a segment that is telomeric in the other three species.

The rapid structural evolution in the telomeric regions can also be observed at the gene level. The gene families contained within these regions (including the HXT, FLO, PAU, COS, THI and YRF families) show significant changes in number, order and orientation. The regions also harbour many new sequences, including protein-coding sequences, in the four yeasts (see below). Moreover, the telomeric regions have undergone 11 reciprocal translocations across the species.

Together, these features define relatively clear boundaries for the telomeric regions on all 32 chromosome arms, with sizes ranging from ~7 kb to ~52 kb on chromosome I-R (Fig. 2). The extraordinary genomic churning occurring in these regions probably has a key role in rapidly creating phenotypic diversity over evolutionary time. A high degree of variation in telomeric gene families has also been reported in *Plasmodium falciparum*[21], the parasite responsible for malaria, and is related to antigenic variation.

Outside of the telomeric regions, few genomic rearrangements are found relative to *S. cerevisiae* (see Methods). We enumerated all non-telomeric rearrangements affecting multiple consecutive genes (Fig. 2). *Saccharomyces paradoxus* shows no reciprocal translocations, four inversions and three segmental duplications; *S. mikatae* shows four reciprocal translocations and 13 inversions; *S. bayanus* has five reciprocal translocations and three inversions. The results confirmed four recently reported reciprocal translocations in these species, identified by pulsed-field gel electrophoresis[22], and identified four additional reciprocal translocations that had been missed.

The sequence at the chromosomal breakpoints suggested the possible mechanism that underlies the rearrangements. Notably, the 20 inversions are all flanked by transfer RNA genes in opposite transcriptional orientation and usually of the same isoacceptor type; the origins of inversions in recombination between tRNA genes has not been noted previously. The reciprocal translocations occurred between Ty elements in seven cases and between highly similar pairs of ribosomal protein genes in two cases; the implication of Ty elements in reciprocal translocation is consistent with previous reports[22–25]. One segmental duplication involves 'donor' and 'recipient' regions that are descendants of an ancient duplication in the yeast genome[26]. Differential gene loss of anciently duplicated genes has been reported previously[27], but this is the first observation of a recent re-duplication event within anciently duplicated regions.

Genome evolution at the nucleotide level

With sequence alignments at millions of positions across the four species, it is possible to obtain a precise estimate of the rate of evolutionary change, including substitutions and insertions or deletions (indels), in the phylogenetic tree connecting the species (Fig. 3; see also Supplementary Information and http://www-genome.wi.mit.edu/seq/Saccharomyces/). Using *S. bayanus* as an outgroup, the substitution rate is similar in *S. cerevisiae* and *S. mikatae*, but is ~67% lower in the lineage leading to *S. paradoxus*. We compared the rate of sequence change at aligned sites across the four species in intergenic and genic (defined here as protein-coding) regions. The proportion of sites corresponding to a different nucleotide in at least one of the three species is 58% in intergenic regions but only 30% in genic regions—a difference of about twofold. The proportion of sites corresponding to an insertion or deletion (indel) is 14% in intergenic regions, but only 1.3% in genic regions. The contrast is even sharper when one considers only indels with length not a multiple of three (which would disrupt a reading frame): 10.2% in intergenic regions versus 0.14% in genic regions—a difference of about 75-fold.

The four genomes show tremendous conservation of synteny and they can be well aligned at the nucleotide level. The overall rate of sequence divergence across the species seems to be high enough to facilitate recognition of functional elements, and the marked contrast between intergenic and genic regions should allow for greatly improved gene identification.

Table 1 **Genome assemblies of four *Saccharomyces* species**

	S. cerevisiae	*S. paradoxus*	*S. mikatae*	*S. bayanus*
Sequence coverage (fold)	Finished	7.7	5.9	6.4
Genome sequence in contigs (Mb)*	12.16	11.57	11.22	11.32
Genome length, including gaps (Mb)†	12.16	11.75	12.12	11.54
Percentage of genome in contigs	100	98	93	98
Number of scaffolds	16	51	90	100
N50 scaffold length (kb)	942	509	334	234
N50 contig length (kb)	942	51	20	25
Gaps per 100 kb	0	3.2	10.3	4.4
Average gap length (bp)	0	583	847	679

*Sum of contig lengths.
†Sum of scaffold lengths, including contig lengths and estimated gap sizes between consecutive contigs.

© 2003 Nature Publishing Group

Identification of genes

De novo identification of protein-coding genes from genomic sequence is extremely important, yet surprisingly challenging. The basic approach is to identify ORFs that are too long to have occurred by chance. However, stop codons occur at a frequency of only about 1 in 20 in random sequence. Thus, ORFs of ≥60 amino acids will occur frequently by chance (~5% under a simple Poisson model) and even ORFs of ≥150 amino acids (or 450 base pairs (bp)) will appear by chance in a large genome (approximately 0.05%). This poses a huge challenge for higher eukaryotes, in which genes are typically broken into many small exons (typical size ~125 bp for internal coding exons in mammals[28]). Yeast genes typically contain fewer and larger exons, but nonetheless there are many true exons encoding ≤150 amino acids and it can be difficult to discriminate these from randomly occurring ORFs.

Comparative genomic analysis provides a simple solution: test the ORFs seen in one species by observing whether the orthologous sequence in related species also encodes an ORF. True protein-coding ORFs will typically be under strong selective pressure to preserve the open reading frame, whereas spurious ORFs will accumulate frameshifts and stop codons. We applied this idea to perform *de novo* gene identification in *S. cerevisiae*. We started with the list of all ORFs greater than a given size, compared the list to the three related species and generated a yeast gene catalogue.

Public yeast gene catalogue

When the yeast genome sequence was completed[1], the authors identified 6,275 ORFs in the nuclear genome that could theoretically encode proteins of ≥100 amino acids and that do not overlap a longer ORF by more than half of their length. SGD has since updated the catalogue on the basis of complete resequencing and re-annotation of chromosome III, re-analysis of other chromosomes and reports in the scientific literature. This resulted in a current version (as of May 2002) with 6,062 ORFs encoding ≥100 amino acids, consisting of 3,966 'named' genes (described in at least one publication) and 2,096 'uncharacterized' ORFs. SGD also includes a small collection of ORFs encoding <100 amino acids (see below).

Reading frame conservation test

We developed a reading frame conservation (RFC) test to classify each ORF in *S. cerevisiae* as biologically meaningful or meaningless, on the basis of the proportion of the ORF over which reading frame is locally conserved in each of the other three species (see Methods). For overlapping ORFs in the *S. cerevisiae* genome ($n = 948$), the RFC was calculated only for the portion unique to each overlapping ORF. For spliced genes ($n = 240$), the RFC was calculated only on the largest exon.

Figure 4 illustrates the case of an ORF of 333 bp that is clearly biologically meaningless. The orthologous sequence in all four species is laden with frameshifts (as well as stop codons).

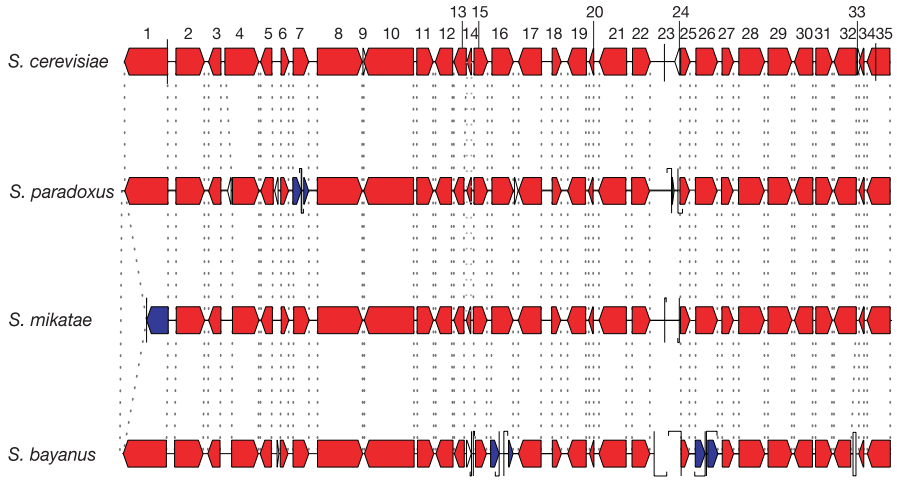
The RFC scores show a clearly bimodal distribution for each of the three species, allowing a decision threshold to be defined for each species to accept an ORF as biologically meaningful or reject an ORF as spurious; that is, occurring by chance. Each species thus 'votes' on the validity of each ORF (or 'abstains' due to lack of orthologous sequence resulting from either incomplete sequence coverage or genomic rearrangement in the species). The votes from the species are then tallied to reach a decision for each ORF (see Methods).

To investigate the power of the approach to reject spurious ORFs, we applied it to a set of control sequences consisting of 340 intergenic sequences in *S. cerevisiae* with lengths similar to the ORFs tested. About 96% were rejected as having conservation properties incompatible with a biologically meaningful ORF, showing that the test has high sensitivity. Of the remaining 4% that were not rejected, close inspection shows that three-quarters seem to contain true ORFs. Some define short ORFs with conserved start and stop codons in all four species, and others extend *S. cerevisiae* ORFs in the 5′- or 3′-direction in each of the other three species (see below). Thus, at most, 1% of true intergenic regions failed to be rejected by the RFC test.

Evaluating ORFs

We sought to apply the RFC test to all 6,062 ORFs in SGD. A total of 117 could not be analysed because they were almost completely contained within an overlapping ORF (99 cases, with average non-overlapping portion of 12 bp) or because an orthologous region could not be unambiguously defined in any of the species (18 cases).

Of the 5,945 ORFs tested, the analysis strongly validated 5,550 ORFs. The vote was unanimous in 5,458 cases (~98%). In the remaining cases, a valid gene appears to have degenerated in one of the four species. A total of 367 ORFs were strongly rejected. These rejections were unanimous in 63% of cases. In most of the remaining cases, *S. paradoxus* was too closely related to *S. cerevisiae*



1	YGL140C	19	YGL122C (*NAB2*)
2	YGL139W	20	YGL121C
3	YGL138C	21	YGL120C (*PRP43*)
4	YGL137W (*SEC27*)	22	YGL119W (*ABC1*)
5	YGL136C	23	tW(CCA)G1
6	YGL135W (*RPL1B*)	24	YGL118C
7	YGL134W (*PCL10*)	25	YGL117W
8	YGL133W	26	YGL116W (*CDC20*)
9	YGL132W	27	YGL115W (*SNF4*)
10	YGL131C	28	YGL114W
11	YGL130W (*CEG1*)	29	YGL113W
12	YGL129C	30	YGL112C (*TAF60*)
13	YGL128C	31	YGL111W
14	YGL127C (*SOH1*)	32	YGL110C
15	YGL126W (*SCS3*)	33	YGL109W
16	YGL125W (*MET13*)	34	YGL108C
17	YGL124C	35	YGL107C
18	YGL123W (*RPS2*)		

Figure 1 Aligned ORFs across four species. A 50-kb segment of *S. cerevisiae* chromosome VII aligned with orthologous contigs from each of the other three species. Predicted ORFs are shown as arrows pointing in the direction of transcription. Orthologous ORFs are connected by dotted lines and are coloured by the type of correspondence: red for 1-to-1 matches, blue for 1-to-2 matches and white for unmatched ORFs. Sequence gaps are indicated by vertical lines at the ends of contigs, with the estimated size of each gap shown by the length of the hook. See Supplementary Information for 250 such figures tiling the complete *S. cerevisiae* genome.

© 2003 Nature Publishing Group

to have accumulated enough frameshifts to allow definitive rejection.

Only 15 of these rejections involve one of the 3,966 named ORFs. We carefully inspected these 15 ORFs (*KRE20*, *KRE21*, *KRE23*, *KRE24*, *VPS61*, *VPS65*, *VPS69*, *BUD19*, *FYV1*, *FYV2*, *FYV12*, *API2*, *AUA1*, *ICS3*, *UTR5* and *YIM2*) and concluded that all were indeed likely to be spurious. Most lack experimental evidence. For the remainder, reported phenotypes associated with deletion of the ORF seem likely to be explained by the fact that the ORF overlaps the promoters of other known genes. The method thus rejects few known genes, indicating that it has high specificity.

Most of the rejections (352, or 96%) involve uncharacterized ORFs. Most of these (209) overlap another well-conserved ORF, but show many insertions and deletions in the non-overlapping portion. The remainder (147) tend to be small (median of 111 amino acids, with 93% ≤150 amino acids) and show atypical codon usage[1,29,30]. With only one exception, SGD reports no compelling biological evidence (such as changes in mRNA expression) to suggest that these ORFs encode a true gene. The one exception is YBR184W, which appears to represent a true gene that fails the RFC test because it is evolving very rapidly (see below).

The analysis deadlocked (one confirmation, one rejection, one abstention) for 28 ORFs (0.5%). Together with the 117 cases that could not be analysed, the RFC analysis thus produced clear results for all but 145 ORFs. We inspected these cases individually and found convincing evidence (on the basis of conservation of amino

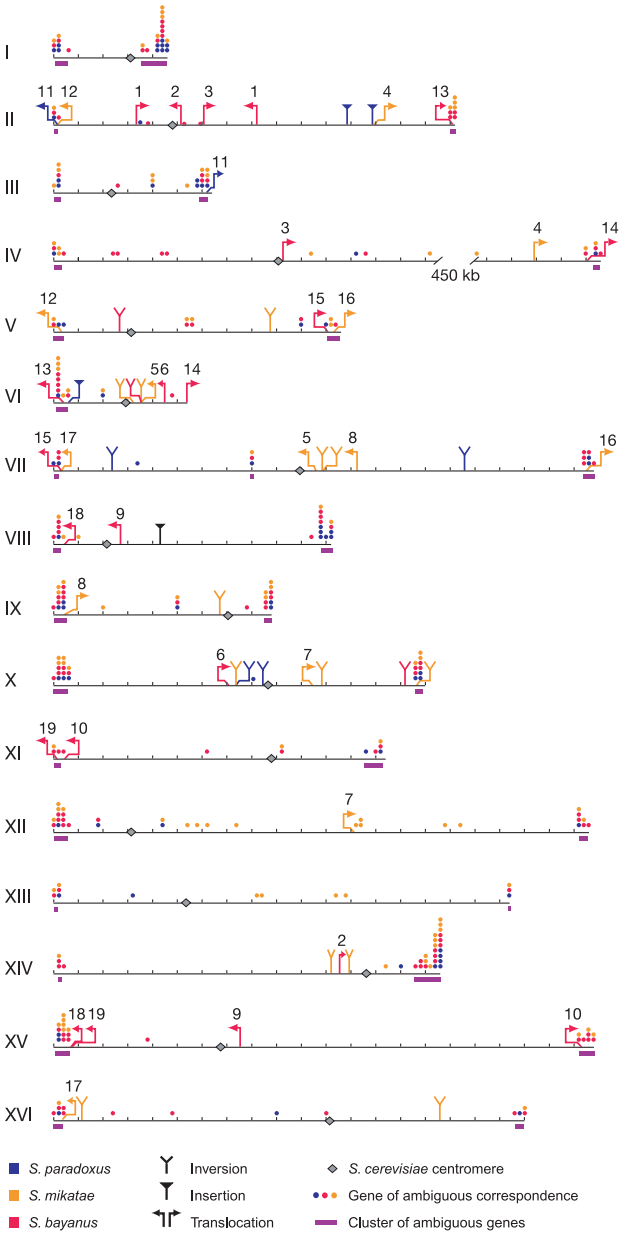


Figure 2 Genome evolution. Locations in the *S. cerevisiae* genome of chromosomal rearrangements in each of the three species are shown by vertical bars with the symbol at the top indicating the type of rearrangement (inversions, insertions and translocations) and colour indicating the species. Tick marks are shown every 50 kb. Inverted segments are small and their endpoints are indistinguishable at this scale. All inversions are flanked by tRNA genes in opposite transcriptional orientation. Reciprocal translocations involve exchange between two breakpoints; the matching pairs are indicated by their common numbering and the orientation of the arrowheads. All non-telomeric translocations are mediated by Ty elements (2, 4, 5, 6, 7, 8, 9) or nearly identical copies of ribosomal protein genes (1, 3). Location in the *S. cerevisiae* genome of genes whose correspondence in the other species is ambiguous is shown by dots, with colour indicating the species in which the ambiguity occurs. A gene whose correspondence is ambiguous in all three species thus appears as three dots of different colours. Clusters of ambiguous genes are denoted by purple horizontal bars; 31 of 32 clusters occur at telomeres.

© 2003 Nature Publishing Group

acids, start and stop codons, and the presence of indels) that 20 are valid protein-coding genes and 105 are spurious. We were unable to reach a judgment in the remaining 20 cases.

Finally, we performed certain consistency checks on the accepted ORFs. In 32 cases, two adjacent ORFs in *S. cerevisiae* are joined into a single ORF in all three other species. In every case, a single nucleotide change would suffice to join the ORFs in *S. cerevisiae* (either a substitution altering a stop codon or an indel altering the reading frame). In principle, these cases could represent errors in the genome sequence, mutations private to the sequenced strain S288C, or substitutions fixed in *S. cerevisiae*. We examined 19 cases by resequencing the relevant region in S288C. Our results revealed an error in the published sequence in 11 cases (establishing that there is a single ORF in S288C) and confirmed the published sequence in the remaining seven cases. Sequencing of additional strains will be required to determine whether these remaining cases represent differences in the S288C strain alone or in *S. cerevisiae* in general.

We also found two named ORFs (*FYV5* and *CWH36*) that pass the RFC test and cause phenotypes when deleted, but show no significant protein similarity across the four species. In both cases, inspection reveals that the opposite strand encodes a protein that shows strong amino-acid conservation. (The latter gene has two introns, increasing the count of doubly spliced genes to eight.) In each case, we postulate that the protein responsible for the reported deletion phenotype is encoded on the opposite strand.

Revised yeast gene catalogue

On the basis of the analysis above, we propose a revised yeast gene catalogue consisting of 5,538 ORFs encoding proteins of ≥100 amino acids. This reflects the proposed elimination of 503 ORFs (366 from the RFC test, 105 by manual inspection and 32 through merger). A total of 20 ORFs in SGD remain unresolved. Complete information about the gene catalogue is provided in Supplementary Information and will be discussed more fully in a subsequent manuscript in collaboration with SGD and other yeast investigators. The revised gene count is consistent with at least two recent predictions based on light shotgun coverage of related species[5,6].

We believe that this represents a reasonably accurate description of the yeast gene set, because the analysis examines all ORFs encoding ≥100 amino acids, the methodology has high sensitivity and specificity, and the evidence is unambiguous for most of the ORFs. Nonetheless, some errors are likely to remain. The results could be confirmed and remaining uncertainties resolved by sequencing of additional related yeast species, as well as by other experimental methods.

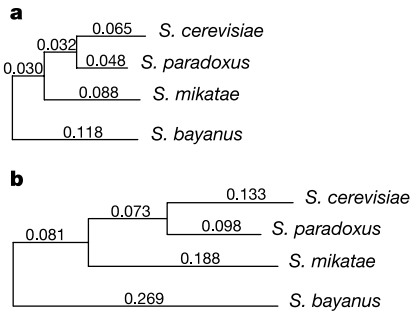


Figure 3 Evolutionary tree of the four yeast species. Branch lengths denote number of substitutions per site in genic (**a**) and intergenic (**b**) sequence. Pairwise distances were calculated using the Kimura two-parameter model and trees were constructed using the UPGMA method. The relative rate of the closest species was computed using the neighbour-joining method. On the basis of single-species changes in four-way alignments, we found *S. paradoxus* to evolve 67% slower than *S. cerevisiae* in coding regions and 70% slower in intergenic regions. In coding regions, the average nucleotide per cent identity to *S. cerevisiae* is 90% for *S. paradoxus*, 84% for *S. mikatae* and 80% for *S. bayanus*. In intergenic regions, the nucleotide per cent identity is, respectively, 80%, 70% and 62% for the three species. Additional information can be found in Supplementary Information. The nucleotide substitution level in aligned intergenic nucleotides between *S. cerevisiae* and *S. bayanus* corresponds roughly to that between human and mouse. Ribosomal DNA sequence analysis suggests divergence times of 5–20 million years for the four species.

Analysis of small ORFs

SGD also lists 141 ORFs encoding 50–99 amino acids for which some biological evidence has been published. Applying the RFC test and inspecting the results, we conclude that 120 seem to be true genes, 18 seem to be spurious ORFs and three remain unresolved.

We systematically searched the remainder of the *S. cerevisiae* genome and evaluated all ORFs in this size range. Control experiments demonstrated that the RFC test has high power to discriminate reliably between valid and spurious ORFs in this size range (see Supplementary Information). The genome contains 3,161 such ORFs; nearly all are readily rejected by the RFC test. However, 43 new genes were identified. These ORFs not only pass the RFC test, but they also have orthologous start and stop codons (see Methods). Five of these have been reported in the literature subsequent to the SGD release studied here.

SGD also lists 32 ORFs encoding <50 amino acids. We did not undertake a systematic search for all such ORFs, because control experiments showed that the RFC test lacked sufficient power to prove the validity of such small ORFs (see Supplementary Information). However, it is able to reject seven of the 32 ORFs as probably spurious. Our yeast gene catalogue thus contains 188 short genes (encoding <100 amino acids), of which 43 are new.

Defining gene structure

Comparative genome analysis not only improves the recognition of true ORFs, it also yields much more accurate definitions of gene structure—including translation start, translation stop and intron boundaries.

Previous annotation of *S. cerevisiae* has defined the start of translation as the first in-frame ATG codon. However, the actual start of translation could lie 3′ to this point or (if sequencing errors or mutations have obscured an earlier in-frame ATG codon) 5′ to this point. We identified 210 cases in which the presumed translational start in *S. cerevisiae* does not correspond to the first in-frame start codon in at least two of the three other species (see Methods). In most of these cases, inspection of the sequence alignments provides strong evidence for an alternative conserved position for the translational start, either 3′ or 5′ to the previous annotation (Fig. 5a, b). We examined four cases in which the comparative data suggested an earlier start codon and found, by resequencing, that all correspond to errors in the published sequence of S288C.

Similarly, we identified 330 cases in which the presumed translational stop codon in *S. cerevisiae* does not correspond to the first in-frame stop codon in at least two of the three species. In about 25% of these cases, the other three species share a common stop

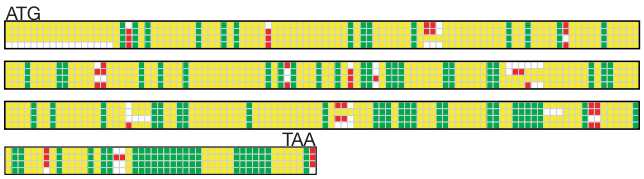


Figure 4 Spurious ORF rejected by RFC test. Schematic representation of the multiple sequence alignment of ORF YDR102C. Aligned nucleotides across the four species are shown as stacked squares (*S. cerevisiae*, S. *paradoxus*, *S. mikatae* and *S. bayanus*, respectively) coloured by their conservation: green for conserved positions, yellow otherwise. Alignment gaps are shown in white and frame-shifting insertions (length not a multiple of 3) are shown in red. In addition to the abundance of frame-shift indels shown here, numerous in-frame stop codons are observed in the other three species. Full nucleotide sequence alignment is provided in Supplementary Information.

© 2003 Nature Publishing Group

codon, and a single base change to the *S. cerevisiae* sequence would result in a stop codon in the corresponding location (Fig. 5c, d). We examined 17 such cases and found that 15 are explained by errors in the published sequence of S288C. The remaining 75% of cases seem to represent true differences in the location of the translational stop across the species. Thus, stop codons seem to show more evolutionary variability in position than start codons.

New introns

We also examined the conservation of introns in the yeast genome. We studied 218 of the 240 ORFs reported in SGD to contain at least one intron (omitting the rest primarily due to lack of an orthologous alignment).

In 92% of cases, the donor, branchpoint and acceptor sites were all strongly conserved with respect to both location and sequence. Moreover, exon boundaries closely demarcated the domains of sequence conservation as measured by both nucleotide identity and absence of indels.

Discrepancies were found in 17 cases, of which at least nine strongly suggest that the previous annotation is incorrect. Five identify a new first exon (Fig. 5e) and four predict that a previously annotated intron is spurious.

We then sought to identify previously unrecognized introns by searching the *S. cerevisiae* genome for conserved splicing signals (see Methods) and then visually inspecting each case. We predict 58 new introns. Fifty cases affect the structure of known genes (defining new 5′ exons in 42 cases, 3′ exons in seven cases and an internal splice in one case) and two indicate the presence of new genes. The relationship of the apparent splice signals to existing genes is unclear for the remaining six cases.

We compared our predictions to the results of experimental studies by Ares and colleagues to identify new introns using techniques such as microarray hybridization[31]. Of our 58 predicted introns, 20 were independently discovered by this group. Of the four annotated introns predicted to be spurious, all four show no experimental evidence of splicing. Our remaining predictions are currently being tested in collaboration with Ares and colleagues.

Despite the intensive study of *S. cerevisiae* so far, comparative genome analysis points to the need for a major revision of the yeast gene catalogue, affecting more than 15% of all ORFs. Moreover, the results suggest that comparative analysis of a modest collection of species can permit accurate definition of genes and their structure.

Rapid and slow evolution of genes

Comparative genomic approaches for gene identification require that orthologues can be recognized and aligned in most of the related species. Difficulties could arise in cases of rapid evolution, including the acquisition or loss of entire genes, the rapid divergence of nucleotide sequence or the presence of large insertions. We characterized such events in the four yeast species.

Species-specific genes

We noted above that *S. cerevisiae* contains 18 genes for which we could not identify orthologues in any of the other species, of which seven encode ≥200 amino acids. These may be species-specific genes in *S. cerevisiae*, but alternatively could simply reflect gaps in the available draft genome sequences.

This uncertainty does not arise, however, in the reverse direction in identifying genes in the related species that lack an orthologue in *S. cerevisiae*. We found a total of 35 such ORFs encoding ≥200 amino acids (with the minimum length chosen to ensure that these are likely to represent valid genes). The list includes five genes unique to *S. paradoxus*, eight genes unique to *S. mikatae* (two of which are 99% identical) and 19 genes unique to *S. bayanus* (three of which form a gene family with ≥90% pairwise identity). There is also one gene represented by orthologous ORFs found in the latter

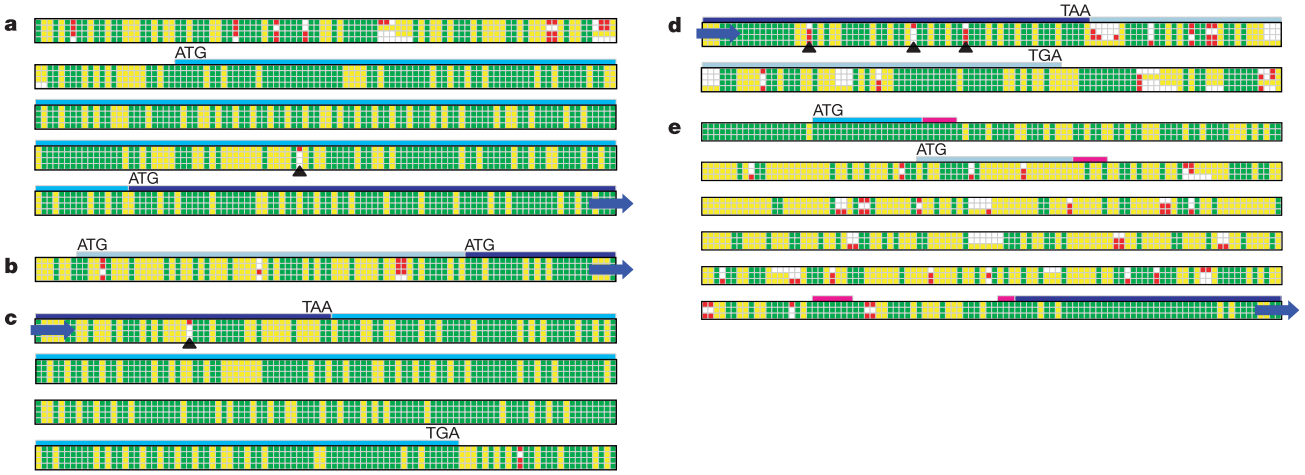


Figure 5 Examples of proposed changes in gene structure. Representation of alignments as in Fig. 4 (see Supplementary Information for full nucleotide alignments). Exon locations are indicated by overhead bars. Dark blue bars indicate portions of annotated exons unchanged in our analysis, and dark blue arrows indicate the continuation of ORFs beyond the alignment shown; light blue bars indicate new 5′ or 3′ extensions of exons (in **a** and **c**) and a new first exon (in **e**); grey bars indicate rejected portions of previously annotated exons (in **b** and **d**). Black triangles indicate the predicted errors in the published sequence of S288C; these predictions were confirmed by resequencing. In all five cases the proposed change leads to perfect agreement in ORF boundaries across all four species. **a**, Start moved upstream for *SAP155*. Elimination of a single base in the *S. cerevisiae* sequence extends the ORF 74 amino acids in the 5′ direction, with a conserved start codon. Resequencing confirmed that the published sequence of S288C is in error. **b**, Start moved downstream for *YBP14*. No change to the *S. cerevisiae* sequence is required. The comparative data suggest that the true translation start is not until the second in-frame ATG: the first ATG is not conserved and multiple frame-shifting indels are found in the rejected portion. **c**, Stop moved downstream for YJL160C. Elimination of a single base in the *S. cerevisiae* sequence extends the ORF by 108 amino acids in the 3′ direction, with an orthologous stop codon in all four species. Resequencing confirmed that the published sequence of S288C is in error. **d**, Stop moved upstream for *YIP4*. Comparative data suggest a shorter ORF with an orthologous stop codon in all four species. The previously annotated stop codon is not conserved, and multiple frame-shifting indels are found in the rejected portion. All three predicted errors in the published sequence of S288C were confirmed by resequencing. **e**, Changed first exon for *RPL26B*. No change to the *S. cerevisiae* sequence is required. *S. paradoxus* sequence not shown. Conservation of splice donor and branchpoint suggests an alternative to the previously annotated exon 1 that shows no conservation of start codon or splice donor and that has accumulated numerous mutations and frame-shifting indels.

© 2003 Nature Publishing Group

two species only, and one represented by orthologous ORFS in all three related species.

These species-specific ORFs are notable with respect to both function and location. Most (63%) can also be assigned biological function on the basis of strong protein-sequence similarity with genes in other organisms. Most involve sugar metabolism and gene regulation (including one encoding a silencer protein). The majority (69%) are found in telomeric regions and an additional set (17%) are immediately adjacent to Ty elements; these locations are consistent with rapid genome evolution.

A coincidence was noted in the region between YFL014W and YFL016W in *S. cerevisiae*. In the orthologous regions in all four species, we find a species-specific ORF in every case (165, 111, 136 and 228 amino acids), but these four ORFs show little similarity at the protein level. The amino-acid sequence has been disrupted by frame-shifting indels, but a long ORF has been maintained in each case. The explanation for this phenomenon is unclear, but may prove interesting.

Rapidly evolving genes

The gene analysis described above rejected only a single ORF (YBR184W) that is clearly biologically meaningful. The explanation seems to be that the gene is evolving rapidly. The region containing YBR184W corresponds to a large ORF in all four species (524, 558, 554 and 556 amino acids), but the alignment shows unusually low sequence conservation. The sequence has only 32% nucleotide identity and 13% amino-acid identity across the four species. Pairwise alignments across the species show numerous insertions and deletions, explaining why the gene failed the RFC test. (Notably, multiple alignment of all four species simultaneously improves the alignment sufficiently to allow the gene to pass the RFC test; this suggests a way to improve the test.)

The rapid divergence is suggestive of a gene under strong positive selection. We tested this hypothesis by calculating the K_a/K_s ratio (the normalized ratio of amino-acid-altering substitutions to silent substitutions), a traditional test for positive selection[32]. Whereas typical genes in *S. cerevisiae* show a K_a/K_s ratio of 0.11 ± 0.02, YBR184W has a ratio of 0.689. This ratio ranks as the third highest observed among all yeast genes (if three small domains with high conservation are excluded, the ratio rises to 0.774). The two genes with higher K_a/K_s ratio are YAR068W, a putative membrane protein, and YER121W, whose expression changes under stress.

The protein encoded by YBR184W has not been studied extensively, but expression studies show that the gene is induced during sporulation[33], and sequence analysis shows that it is similar to the gene *YSW1*, which encodes a spore-specific protein. This is consistent with the observation that many of the best-studied examples of positive selection in other organisms are genes related to gamete function.

Evidence of rapid protein change

Most of the nucleotide changes in protein-coding regions are silent or affect individual amino acids. However, a small number of events suggest additional mechanisms of rapid protein change. These events include closely spaced compensatory indels that affect the translation of small contiguous amino-acid stretches. They also include the loss and gain of stop codons (by a nucleotide substitution or a frame-shifting indel), which may result in the rapid change of protein segments or the translation of previously non-coding regions[34]. Such events are observed more frequently near telomeric regions and may affect silenced genes or recently inactivated pseudogenes.

Furthermore, we found a small number of differences in the length of orthologous proteins. These typically involve changes in the copy number of tri-nucleotide repeats, such as (CAA), which encodes hydrophobic stretches often involved in protein–protein interactions. The most marked example is seen for the *TFP1* gene, which encodes a vacuolar ATPase. The *S. cerevisiae* gene contains an insertion of 1,400 bp that is absent in the three related species. The insertion corresponds to the recent horizontal transfer of a known post-translationally self-splicing intein, Vma1 (ref. 35).

Slowly evolving genes

To complement the analysis of rapidly changing genes, we scanned the genome for genes with an unusually slow rate of evolution. One case stands out as an extreme outlier: the mating-type gene *MATa2*. The gene shows perfect 100% conservation at the amino-acid level over its entire length (119 amino acids) across all four species. More notably, the gene shows perfect 100% conservation at the nucleotide level as well (357 bp). This differs sharply for the typical pattern seen for protein-coding genes, which show relaxed constraint in third positions of codons.

Notably, the *MATa2* gene is the only one of the four mating-type genes (the others being *MATα1*, *MATα2* and *MATa1*) whose biochemical function remains unknown despite two decades of research[36]. An important clue may be that the sequence of *MATa2* is identical in all four species to the 3′ end of the *MATα2* gene. Perfect conservation at the nucleotide level and identity to the terminus of *MATα2* suggests that *MATa2* may function not by encoding a protein, but rather by encoding an antisense RNA or a DNA site.

Variation in the rate of genome evolution seems not to pose a serious problem for gene identification for the species studied here. Various examples of extremely rapid or slow gene evolution can be identified, however, and each is likely to merit further study.

Genome-wide identification of regulatory elements

Direct identification of regulatory elements is more challenging than for genes. Such elements are typically short (6–15 bp), tolerate some degree of sequence variation and follow few known rules. So far, most have been found by experimental manipulation, such as systematic mutation of individual promoter regions; however, the process is laborious and unsuited for genome-scale analysis.

Computational analysis of single genomes has been successfully used to identify regulatory elements associated with known sets of related genes[8–10], but these approaches do not have sufficient power to permit comprehensive direct identification of regulatory elements[37].

Comparative genomics offers various approaches for finding regulatory elements. The simplest is to perform cross-species sequence alignment to find 'phylogenetic footprints'—regions of unusually high conservation. This approach has long been used to study promoters of specific genes in many organisms[11,13,24,38,39], and recently was applied across the entire human and mouse genomes[7]. The genome alignments of the four *Saccharomyces* species can similarly be used to study each yeast gene, to help define promoters and other islands of intergenic conservation (Fig. 6).

Our interest was to go beyond inspection of individual islands of conservation to construct a comprehensive dictionary of regulatory elements used throughout the genome. We investigated the conservation properties of known regulatory motifs and used the insights gained to design an approach for *de novo* discovery of regulatory motifs directly from the genome. We then developed an approach for inferring a candidate function of these motifs, making use of biological knowledge about genes.

Conservation of the Gal4-binding site

We first studied the binding site for one of the best-studied transcription factors, Gal4, whose sequence motif is $CGGn_{(11)}CCG$ (which contains 11 unspecified bases). Gal4 regulates genes involved in galactose metabolism, including the *GAL1* and *GAL10* genes, which are divergently transcribed from a common intergenic region (Fig. 6). The Gal4 motif occurs three times in this intergenic region,

© 2003 Nature Publishing Group

and all three instances show perfect conservation across the four species. In addition, there is a fourth experimentally validated binding site[40] for Gal4 that differs from the consensus by one nucleotide in *S. cerevisiae*. This variant site is also perfectly preserved across the species.

We then examined the frequency and conservation of Gal4-binding sites across the aligned genomes. In *S. cerevisiae*, the Gal4 motif occurs 96 times in intergenic regions and 415 times in genic (protein-coding) regions. The motif displays certain marked conservation properties (see Methods). First, occurrences of the Gal4 motif in intergenic regions have a conservation rate (proportion conserved across all four species) that is about fivefold higher than for equivalent random motifs (12.5% compared with 2.4%). Second, intergenic occurrences of the Gal4 motif are more frequently conserved than genic occurrences (12.5% compared with 3%). By contrast, random motifs are less frequently conserved in intergenic regions than in genic regions (3.1% compared with 7.0%), reflecting the lower overall level of conservation in intergenic regions. Thus, the relative conservation rate in intergenic compared with genic regions is about 11-fold higher for Gal4 than for random motifs. Third, the Gal4 motif shows a higher conservation rate in divergent compared with convergent intergenic regions (those that lie upstream compared with downstream of both flanking genes); no such preferences are seen for control motifs. These three observations suggest various ways to discover motifs based on their conservation properties (see conservation criteria below).

Catalogue of known motifs

We extended these observations by assembling a catalogue of 55 known regulatory sequence motifs (Table 2), by starting with two public databases (SCPD[41] and YTFD[42]) and curating the entries to select those with the best support in the literature. Nearly all of these sequence motifs are binding sites of known transcription factors.

We defined a motif conservation score (MCS) on the basis of the conservation rate of the motif in intergenic regions: the MCS is measured in standard deviations above the rate for comparable control motifs (see Methods). Most of the known motifs show strong conservation, with 60% having MCS $\geq$4 (which is substantially higher than expected by chance). Some of the motifs, however, show relatively modest MCS. These motifs may be incorrect, suboptimal or not well conserved.

Methodology for genome-wide motif discovery

Our approach involves first identifying conserved 'mini-motifs' and then using them to construct full motifs (see Methods). Mini-motifs are sequences of the form $XYZn_{(0-21)}UVW$, consisting of two triplets of specified bases interrupted by a fixed number (from 0 to 21) of unspecified bases. Examples are TAGGAT, ATAnnGGC, or the Gal4 motif itself. The total number of distinct mini-motifs is 45,760, if reverse complements are grouped together.

Conserved mini-motifs are then defined according to three conservation criteria (CC1–3), based on our observations about

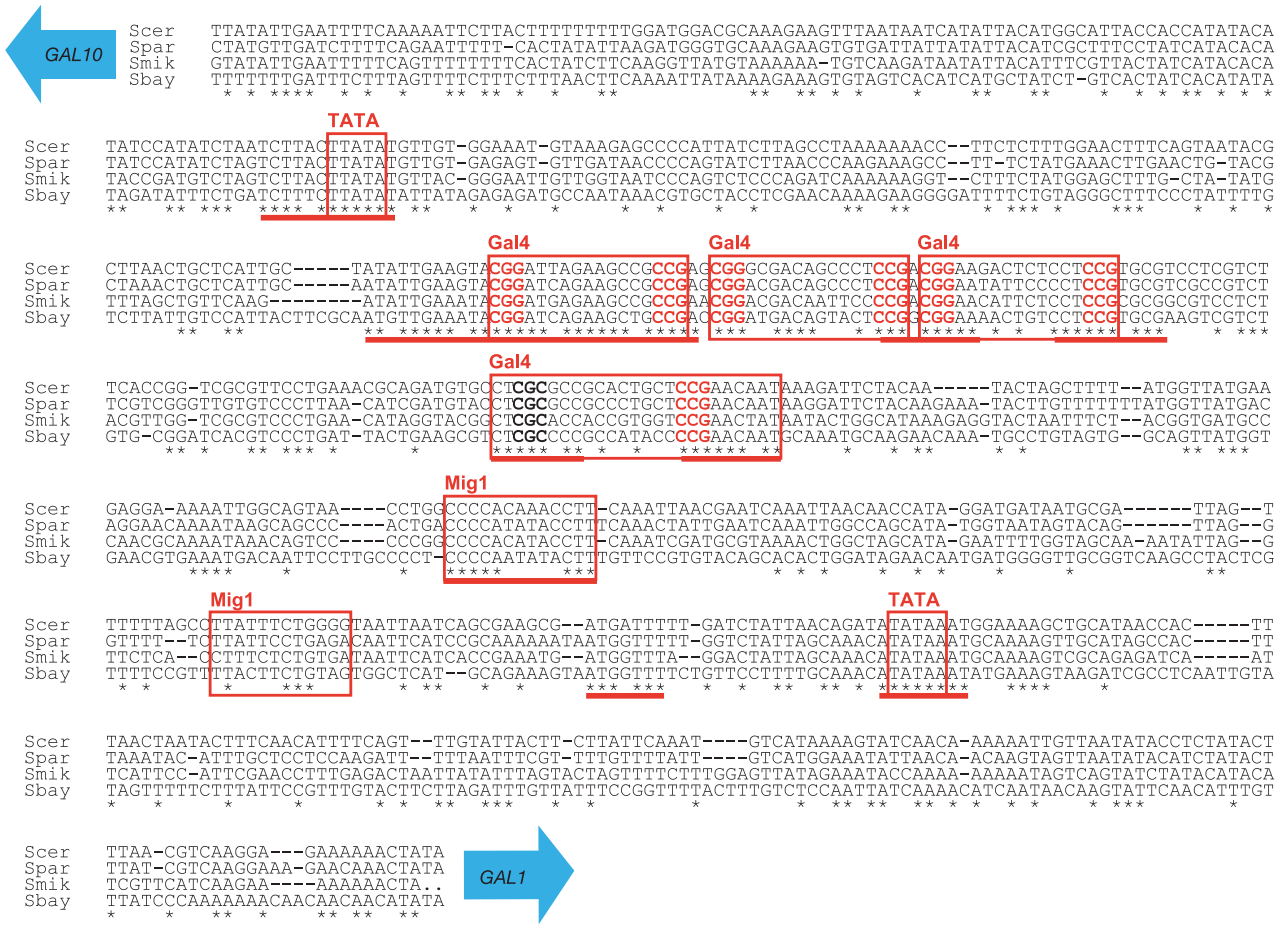


Figure 6 Conservation in the *GAL1–GAL10* intergenic region. Multiple alignment of the four species shows a strong overlap between functional nucleotides and stretches of conservation. Asterisks denote conserved positions in the multiple alignment. Blue arrows denote the start and transcriptional orientation of the flanking ORFs. Experimentally validated factor-binding footprints are boxed and labelled according to the bound factor. Stretches of conserved nucleotides are underlined. Nucleotides matching the published Gal4 motif are shown in red. The fourth experimentally validated site differs: it shows a longer footprint and a non-standard consensus motif (bold). This variant motif is also conserved across all four species. Scer, *S. cerevisiae*; Spar, *S. paradoxus*; Smik, *S. mikatae*; Sbay, *S. bayanus*.

© 2003 Nature Publishing Group

the properties of the Gal4 motif. In each case, conservation rates are normalized to appropriate random controls. The conservation criteria are: (1) intergenic conservation (CC1), the mini-motif shows a significantly high conservation rate in intergenic regions; (2) intergenic–genic conservation (CC2), the mini-motif shows significantly higher conservation in intergenic regions than in genic regions; (3) upstream–downstream conservation (CC3), the mini-motif shows significantly different conservation rates when it occurs upstream compared with downstream of a gene.

The conserved mini-motifs are then used to construct full motifs (Fig. 7). They are first extended by searching for nearby sequence positions showing significant correlation with a mini-motif. The extended motifs are then clustered, merging those with substantially overlapping sequences and those that tend to occur in the same intergenic regions. Finally, a full motif is created by deriving a consensus sequence (which may be degenerate).

Each full motif is assessed for genome-wide conservation by calculating its MCS, and those motifs with MCS ≥4 are retained. Each full motif was also tested for enrichment in upstream compared with downstream regions, by comparing its conservation rate in divergent versus convergent intergenic regions.

Results of genome-wide motif discovery

Most of the 45,760 possible mini-motifs show no distinctive conservation pattern. However, approximately 2,400 mini-motifs show high scores by one or more of these criteria (Fig. 7a–c). There

Table 2 **Known motifs and related discovered motifs**

	Known motif		Discovered motif*			
Factor	Motif	MCS†	Motif	Genome-wide‡	Category- based§	MCS‖
ABF1	**RTCRYnnnnnACG**	50.0	**RTCRY**knnnn**ACG**R	S	S	36.2
UME6	**TCGGCGGCTA**	20.9	**TSGGCGGCTA**WW	S	NC	23.4
CBF1	**RTCACRTG**	19.0	**RTCACGTG**V	S	S	17.6
NDT80	**TCGGCGGCTDW**	18.6	**TSGGCGGCTAW**W	S	NC	23.4
REB1	**TTACCCGG**	17.8	R**TTACCCGR**M	S	S	34.3
MCM1a	T**TWCC**Cn**WWW**R**GGAAA**	16.5	**TTCC**na**AttnGGAAA**	S	S	13.8
SWI6	**ACGCGT**	16.4	**WCGCGT**CGCGt	S	S	10.2
PHO4	**CACGTG**	16.1	RT**CACGTG**V	S	S	17.6
MBP1	**ACGCGTnA**	14.8	**WCGCGT**CGCGt	S	S	10.2
SWI4	**TTTTCGCG**	12.4	**WTTTCGCG**TT	S	S	12.0
DAL81	GATAAG	12.1	–	–	NE	–
RPN4	**TTTTGCCACC**	11.5	**TTTTGCCACC**G	S	NC	11.0
MSN2	**CCCCT**	11.3	hR**CCCYT**WDt	S	NE	7.8
MSN4	**CCCCT**	11.3	hR**CCCYT**WDt	S	NE	7.8
PDR1	**CCGCGG**	9.3	Y**CCGSGG**S	S	NE	6.7
ESR2	**AAAAWTTTT**	8.9	GR**RAAAWTTTT**CACT	S	NC	15.6
MIG1	**CCCCRS**WWWW	8.7	D**CCCCGC**GH	S	NE	8.2
MIG1b	**CCCCGC**	8.4	D**CCCCGC**GH	S	NE	8.2
BAS1	**TGACTC**	8.3	A**TGACTC**WT	S	S	6.1
GCN4	**ATGACTCAT**	8.2	**ATGACTCWT**	S	S	6.1
GAL4	**CGGnnnnnnnnnnCCG**	8.0	**CGG**CnnMGnnnnnnn**CG**C	S	S	5.0
HSF1b	**TTCTAGAA**	7.8	**TTCTMGAA**GA	S	S	7.0
ESR1	**GATGAG**	7.7	gc**GATGAG**mtgaraw	S	NC	24.7
MET31	AAA**CTGTGGC**	6.8	**SKGTGGS**Gc	S	S	8.1
AFT1	Y**RCACCC**R	6.8	**RVACCC**TD	S	NC	10.3
TEA1	CGGnCGG	6.8	–	–	NC	–
PUT3	**CGG**nnnnnnnnn**CCG**	6.2	**C**C**G**Mnnnnnnnnn**mSG**R	W	NE	5.4
HAP2	**TGATTGG**C	5.7	**TGATTGG**T	–	S	[6.4]
RAP1	**ACACCCA**T**ACAT**TT	5.2	**ACACCCA**C**ACATnn**C	S	S	9.9
LEU3	**CCGGnnCCGG**	4.9	**CCSG**TAn**CGG**	S	S	6.5
MCM1b	Y**TTCC**T**AATT**W**G**nnCn	4.8	**TTCCnaAttnG**GAAA	S	S	13.8
INO4	**CATGTGAA**AT	4.1	Gnnn**CATGTGAA**	–	S	[6.8]
INO2	**CATGTG**AAAT	4.1	**CATGTG**	–	S	[4.4]
GLN3	GATAAK	3.8	–	–	NE	–
ADR1	GGAGA	3.7	–	–	NE	–
FKH2	**TTGTTTAC**S**T**	3.6	t**TTGTTTACnT**TT	S	S	10.8
FKH1	**TTGTTTAC**S**T**	3.6	t**TTGTTTACnT**TT	S	S	10.8
RLM1	**CTAWW**W**WT**AG	3.6	**CTAnnTTTAG**	S	S	[4.7]
SWI5	**KGCTGR**	3.4	**TGCTGG**	–	S	[6.1]
HAP1	C**G**GnnnT**AnCGG**	2.5	**G**CnnT**TAnCGG**	S	NC	4.8
XBP1	M**CTCGARRR**nR	2.5	T**CTCGARRA**	S	NC	12.5
MAC1	TT**TGCTCA**	2.3	**TGCTCA**	–	S	[5.4]
TBF1	**TTAGGG**	2.3	G**KBAGGG**T	S	NC	4.8
MSE	**TTTTGTG**	1.4	**TTTTGTG**TCRC	S	NC	9.9
STE12	R**TGAAACA**	0.7	Y**TGAAACA**	–	S	[12.2]
DIG1	R**TGAAACA**	0.7	Y**TGAAACA**	–	S	[12.2]
MET4	**TGGCAAA**TG	0.7	CGG**TGGCAAA**A	S	NE	–
HAP4	**TnRTTGGT**	0.5	**T**G**ATTGGT**	–	S	[6.4]
SMP1	ACTACTAWWWWTAG	0.4	–	–	NE	–
ACE2	**GCTGGT**	−0.6	T**GCTGGT**	–	S	[7.4]
YAP1	TTACTAA	−1.1	–	–	NE	–
CIN5	TTACTAA	−1.1	–	–	NE	–
RME1	GAACCTCAA	−1.4	–	–	NE	–
HAC1	CAGCGTG	−1.4	–	–	NC	–
GCR1	**GGAAG**	−18.5	**GGAAG**C	–	S	[4.4]

Lower-case characters denote lower stringency. Degenerate nucleotides as follows: S = CG, W = AT, R = AG, Y = CT, K = GT, M = AC, B = CGT, D = AGT, H = ACT, V = ACG, N = ACGT. Characters in bold correspond to matches between known motif and discovered motif.

*Discovered motif is the best genome-wide discovered motif, if one was found. Otherwise, it is the category-based motif found using the gene category corresponding to chromatin immunoprecipitation with the corresponding transcription factor, if this category was available.

†Negative MCS values indicate that the motif showed a weaker conservation than the random controls. Known motifs with low MCS values (<1.0) were only found in the category-based search.

‡S indicates that a strong genome-wide match was found to the known motif; W, a weak match to the known motif was found.

§S indicates that a strong category-based motif was found using the gene category for ChIP with the corresponding transcription factor; NE indicates that the known motif was not enriched in this gene category; NC indicates that the gene category was available. Note that a strong category-based motif was found for every known motif for which the appropriate gene category was available and the known motif was actually enriched in the category.

‖MCS value of the genome-wide motif, if one was found. Category-based score shown in brackets otherwise.

© 2003 Nature Publishing Group

is substantial overlap among the mini-motifs produced by the three criteria, with about 50% of those found by one criterion also found by another.

The conserved mini-motifs give rise to a list of 72 full motifs having MCS ≥4 (Table 3). Most of the motifs show preferential enrichment upstream of genes, but six are enriched downstream of genes. These 72 discovered motifs, found with no previous biological knowledge, show strong overlap with 28 of the 33 known motifs having MCS ≥4. They include 27 strong matches and one weaker match.

The 72 discovered motifs also contain matches to eight of the 22 known motifs with MCS <4. In these cases, the comparative analysis identified closely related motifs that have higher conservation scores than the known motifs and occur largely at the same genes; these may represent a better description of the true regulatory element.

Comparative genomic analysis thus automatically discovered 36 motifs with matches to most of the known motifs (65% of the full set, 85% of those with high conservation). It also identified 42 additional 'new' motifs not found in our list of known motifs.

Inferring function of genome-wide motifs

We next developed ways to assign candidate functions to these discovered motifs by the genes adjacent to conserved occurrences of the motif with known gene categories. For inspiration, we again used the Gal4 motif. Given the biological role of Gal4, we considered the set of genes annotated to be involved in carbohydrate metabolism (126 genes according to the Gene Ontology (GO)[43] classification) with the set of genes that have a Gal4-binding motif upstream. The intergenic regions adjacent to carbohydrate metabolism genes comprise only 2% of all intergenic regions, but 7% of the occurrences of the Gal4 motif in *S. cerevisiae* (3.5-fold enrichment) and 29% of the conserved occurrences across the four species (15-fold enrichment).

These results suggest that a function of the Gal4 motif could be inferred from the function of the genes adjacent to its conserved occurrences. Such putative functional assignments can be useful in directing experimentation for understanding the precise function of a motif. We therefore assembled a collection of 318 yeast gene categories based on functional and experimental data. These categories consist of 120 sets of genes defined with a common GO classification in SGD[43]; 106 sets of genes whose upstream region was identified as binding a given transcription factor in genome-wide chromatin immunoprecipitation (ChIP) experiments[44]; and 92 sets of genes showing coordinate regulation in RNA expression studies[45]. To measure how strongly the conserved occurrences correlated with the regions upstream (or downstream) of a

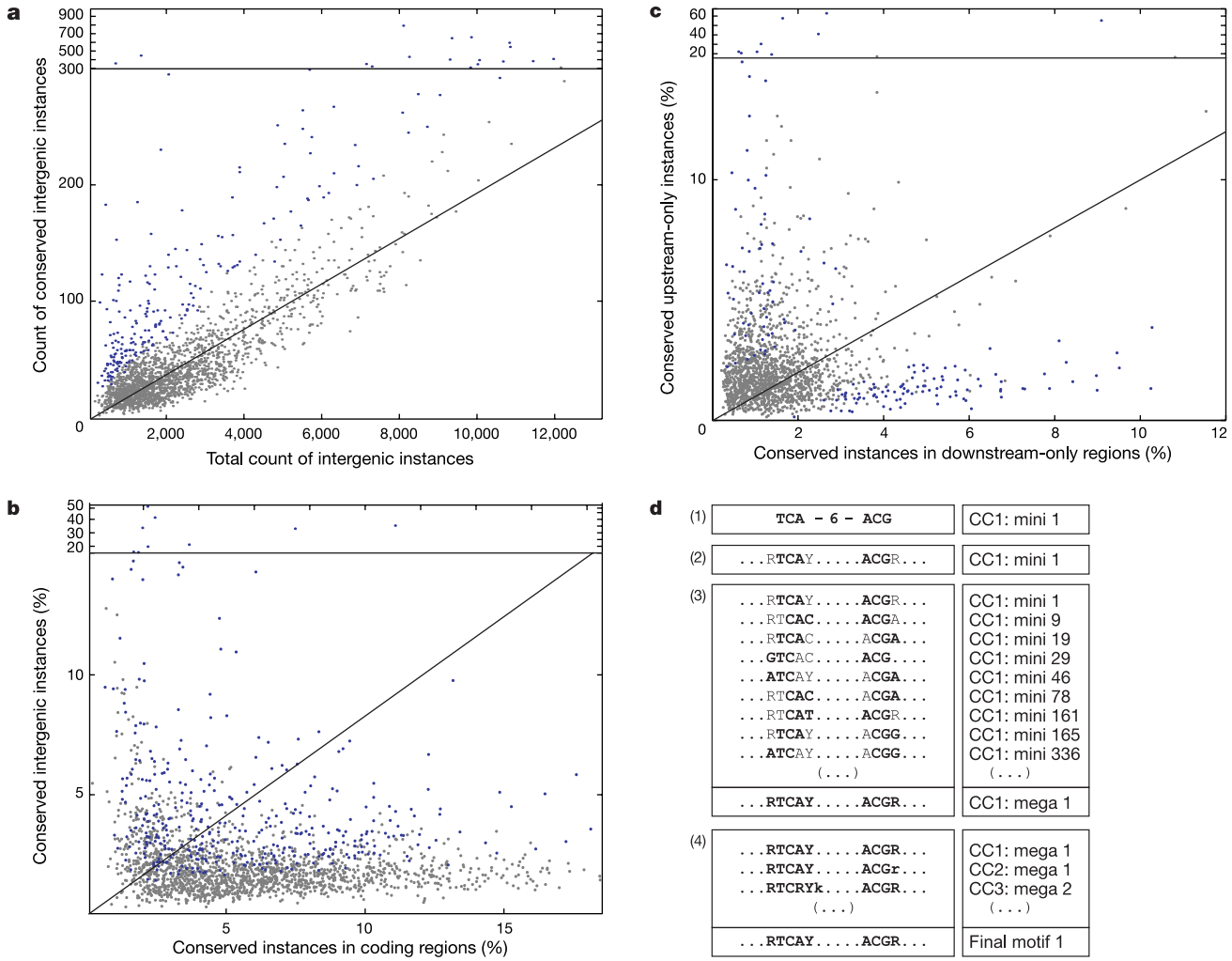


Figure 7 Distribution of motifs by conservation score. **a**–**c**, Points denote individual mini-motifs analysed with respect to the three conservation properties described in the text (CC1, intergenic conservation (**a**); CC2, intergenic–genic conservation (**b**); CC3, upstream–downstream conservation (**c**)). Blue points exceed the significance threshold used; grey points are non-significant. **d**, The generation of a full motif from mini-motifs: (1) we selected mini-motifs using each conservation criterion; (2) we extended mini-motifs with additional conserved bases; (3) we generated mega-motifs by merging mini-motifs with similar extension; (4) we merged mega-motifs that frequently co-occur in the same intergenic regions.

© 2003 Nature Publishing Group

particular gene category, we defined a category correlation score (CCS) (see Methods). A CCS ≥ 5 (nominal P-value $< 10^{-5}$) is significant after accounting for testing of multiple categories.

Most of the 36 discovered motifs that correspond to known motifs showed strong category correlation. Categories with the strongest correlation included those identified by ChIP with the transcription factor known to bind the motif, although many other relevant categories were identified. Of the 42 new motifs, 25 show

Table 3 **Discovered motifs**

No.	Discovered motif	Location*	MCS†	Best category‡	CCS§	Interpretation‖
1	YCGTnnnnmRYGAY	5′	36.2	ChIP: Abf1	90	Known: Abf1
2	RTTACCCGRM	5′	34.3	ChIP: Reb1	38	Known: Reb1
3	gcGATGAGmtgaraw	5′	24.7	Exp.: cluster 74	62	Known: Esr1 GATGAG
4	TSGGCGGCTAWW	5′	23.4	GO: meiosis	10	Known: Ume6/Ndt80
5	RTCACGTGV	5′	17.6	ChIP: Cbf1	27	Known: Cbf1/Pho4
6	WTATWTACADG	3′	17.4	Exp.: cluster 16 downstream	25	New: mitochondrial downstream
7	GRRAAAWTTTTCACT	5′	15.6	Exp.: cluster 74	37	Known: Esr2
8	TTCCnaAttnGGAAA	5′	13.8	ChIP: Mcm1	29	Known: Mcm1
9	CGTTTCTTTTTCY.	5′	13.5	GO: filamentation	7	New: filamentation
10	TYYTCGAGA.	5′	12.5	Exp.: cluster 86	5	Known: Xbp1 (Hsf1-co-ocuring)
11	TTTTCGCG	5′	12.0	ChIP: Swi4	21	Known: Swi4 fixed gap
11a	TTTT = CGCG¶	5′	12.0	ChIP: Swi4	–	New: Swi4 variable gap
12	TKACGCGTT	5′	12.0	ChIP: Mbp1	18	Known: Mbp1/Swi6
13	STGCGGnnnttTCTnnG	5′	11.8	GO: filamentation	11	New: filamentation
14	YCTATTGTT	5′	11.5	ChIP: Fkh2	6	New: Rlm1-like
15	TTTTGCCACCG	5′	11.0	GO: proteolysis	25	Known: Rpn4/Met4
16	tTTGTTTACnTTT	5′	10.8	ChIP: Fkh2	28	Known: Fkh1/2
17	RVACCCTD	5′	10.3	–	–	Known: Aft1
18	WCGCGTCGCGt	5′	10.2	ChIP: Mbp1	17	New: double Mbp1
19	GGGTnACCC	5′	10.0	ChIP: Reb1	8	New: Reb1 palindrome
20	GnnATGTGTGGGTGT	5′	9.9	ChIP: Fhl1	5	Known: Rap1
21	TTTTGTGTCRC	5′	9.9	ChIP: Sum1	14	Known: Mse
22	TTTCAnCGCGC	5′	9.8	–	–	New: no category
23	TATTAWTATTATtMtnatta	3′	9.5	–	–	New: no category
24	SCGnHGGS	5′	8.8	GO: filamentation	6	New: filamentation
25	ACAGCCGCRY	5′	8.6	Exp.: cluster 37	6	New: expression cluster 37
26	DCGCGGGGH	5′	8.1	Exp.: cluster 46	8	Known: Mig1b
27	SKGTGGSGc	5′	8.1	ChIP: Met31	5	Known: Met31
28	TTTTn(19)GCKCG	5′	7.8	–	–	Known: no category
29	HRCCCYTWDt	5′	7.8	Exp.: cluster 8	22	Known: Msn2/4
30	TKCCCnnnnGGG	5′	7.3	ChIP: Mcm1	15	Known: Mcm1 (hits tRNA)
31	GTGTCAGTAAt	5′	7.1	ChIP: Sum1	15	New: Sum1
32	RGTTTTTCCG	5′	7.1	ChIP: Rgt1	7	New: Rgt1
33	TTCTMGAAGA	5′	7.0	ChIP: Hsf1	10	Known: Hsf1
34	YCCGSGGS	5′	6.7	GO: filamentation	9	New: filamentation
35	CnCCTTTTATAC	5′	6.5	–	–	New: no category
36	CCSGTAnCGG	5′	6.5	ChIP: Leu3	8	Known: Leu3
37	SKTKCCTT	5′	6.4	GO: filamentation	7	New: filamentation
38	CTCCCCTTAT	5′	6.4	Exp.: cluster 8	11	Known: Msn2/4
39	GCCCGG	5′	6.3	GO: filamentation	10	New: filamentation
40	SGCGCGRB	5′	6.3	–	–	New: no category
41	CTCSGCS	5′	6.2	–	–	New: no category
42	TGnKAGCGCCG	5′	6.2	–	–	–
43	ATGACTCWT	5′	6.1	ChIP: Gcn4	44	Known: Gcn4/Bas1
44	CCGAnnnTCGG	5′	6.1	Exp.: cluster 46	6	New: facilitators palindrome
45	SCGMnnnnnnKCG	5′	6.0	–	–	New: no category
46	CnCCGCGCnnTTTs	5′	6.0	–	–	New: no category
47	TTTTnnnnnnnnnnngGGGT	5′	5.8	–	–	New: no category
48	TGTRnCAW	3′	5.5	–	–	New: no category
49	YCSknnnnnnnnnKCGG	5′	5.4	Exp: cluster 46	6	Known: Put3
50	CGGnnnnnnnnnnnnnKCGV	5′	5.4	–	–	New: no category
51	WGTGACg	5′	5.3	ChIP: Sum1	14	New: Sum1
52	RTCCCTV	5′	5.3	–	–	New: no category
53	YTCGTTTAGG	5′	5.2	GO: lipid metabolism	5	New: lipid metabolism
54	TYCGKRM	5′	5.2	GO: filamentation	7	New: filamentation
55	CGCnnnnnnnnnnnnnBCGB	5′	5.1	–	–	New: no category
56	TWCCCCM	5′	5.0	Exp.: cluster 46	7	Known: Mig1 + facilitators
57	CGGCnnMGnnnnnnnCGC	5′	5.0	ChIP: Gal4	7	Known: Gal4
58	CCGSnnnnnGVC	5′	5.0	–	–	New: no category
59	TRTAMATAKWT	3′	4.8	ChIP: Dig1	7	New: Ste12 (hits tRNA)
60	TtTATAnTATATAnA	3′	4.8	Exp: cluster 74 downstream	6	New: downstream cluster 74
61	GKBAGGGT	5′	4.8	GO: glycolysis	6	Known: Tbf1/new: glycolysis
62	GCnnTTAnCGG	5′	4.8	–	–	Known: Hap1
63	GGCSnnnnnGnnnCGCG	5′	4.7	ChIP: Mbp1	6	Known: Mbp1-like
64	TTCTCnnnnnnnCGC	5′	4.7	GO: filamentation	6	New: filamentation
65	SCGKnnnnKCGD	5′	4.5	–	–	New: no category
66	AATATTCTT	3′	4.4	Exp.: cluster 46 downstream	5	New: downstream facilitators
67	CGCGTnnnnnnnnACG	5′	4.4	ChIP: Swi4	8	New: Swi4-vary gap
68	CCGHVGGM	5′	4.3	–	–	New: no category
69	CGCG = TTTT	5′	4.3	–	–	New: no category
70	CGCGnnnnnGGGS	5′	4.2	Exp.: cluster 46	6	New: expression cluster 46
71	CTGCAGGGR	5′	4.2	GO: filamentation	6	New: filamentation

Lower-case characters and degenerate nucleotides as in Table 2.
*Location indicates whether the motif tends to occur 5′ or 3′ of genes, based on preferential enrichment in divergent or convergent intergenic regions.
†Motif conservation score for discovered motif.
‡The gene category having the highest category correlation score (CCS) with discovered motif. Category types are: GO, Gene Ontology; ChIP, chromatin IP experiment; Exp., expression cluster.
§CCS for best category.
‖Interpretation of discovered motif, based on similarity to known motifs (Table 2) and category correlation. No category denotes that no category was enriched in the motif.
¶Motif 11a is closely related to motif 11 but shows a variable gap.

© 2003 Nature Publishing Group

strong correlation with at least one category and thus can be assigned a suggestive biological function (Table 3).

Examples of new motifs

We describe a few of the new motifs below.

- Some motifs appear to define previously unknown binding sites associated with known transcription factors. Motif 32 is probably the binding site for Rgt1, which regulates genes involved in glucose transport[46]; the motif occurs upstream of many such genes, including appearing five times upstream of *HXT1*, which encodes a high-affinity glucose transporter. Motifs 21, 31 and 51 are all associated with genes whose upstream regions are bound by Sum1, a transcriptional repressor of genes involved in meiosis. The first motif has been reported previously (Mse)[47], but the latter two are new and occur near genes whose products are involved in chromatin silencing and transcriptional repression.
- Some motifs do not match regions bound by known transcription factors, but show strong correlation with functional categories. Motif 9 occurs upstream of genes involved in nitrogen metabolism, including amino-acid and urea metabolism, nitrogen transport, glutamine metabolism and carbamoyl phosphate synthesis. Motif 25 is enriched among co-expressed genes (expression cluster 37) whose products function in vesicular transport and secretion, including GDP/GTP exchange factors essential for the secretory machinery, clathrin assembly factors and many vesicle and plasma membrane proteins. Motifs 9, 13, 24, 34, 37 may have a role in filamentation. They are all enriched in genes co-regulated during environmental changes, involved in signalling and budding, and bound by transcription factors involved in filamentation, such as Phd1.
- Six motifs show higher conservation downstream of ORFs. Some of these may be transcribed in the 3′ untranslated region and have a regulatory role in mRNA localization or stability. The strongest (motif 6; see ref. 48) is found at genes whose product localizes to the cytosolic translational machinery, the mitochondrial DNA translational machinery or the mitochondrial outer membrane. Downstream motifs are also found enriched in a group of genes repressed during environmental stress (motif 60 with expression cluster 74) and a group of genes involved in energy production (motif 66 with expression cluster 46).
- Two motifs (motif 11a and motif 69) show variable gap spacing, suggesting a new type of degeneracy within the recognition site for a transcription factor complex. Motif 11a corresponds closely to the known motif for Swi4 (motif 11) but is interrupted by a central gap of 5, 7 or 9 bases; these variant motifs all show strong correlation with genes bound by Swi4 in ChIP experiments.

Category-based identification of regulatory elements

The motifs above were identified solely on the basis of overall conservation across the genome. We next explored whether additional motifs could be found by searching specifically for conservation within individual gene categories. We used the same procedure for constructing full motifs as above, but used a modified conservation criterion (CC4) that defines conserved mini-motifs as those enriched in the intergenic regions of genes in the category (see Methods).

Known motifs

We first considered the 43 known motifs for which ChIP experiments had been performed with the transcription factor that binds the motifs[44]. For each category defined by the ChIP experiment, we undertook category-based motif discovery.

Strong category-based motifs were found in 29 cases and these invariably corresponded closely to the known motifs (Table 2). These include 11 cases in which the motif had not been found by genome-wide motif discovery, suggesting that a category-based approach can be more sensitive in some cases.

No strong category-based motifs were found for the remaining 14 known cases, including seven cases in which genome-wide analysis yielded the known motif. Analysis of these 14 known motifs showed that none were, in fact, enriched in the ChIP-based category. This may reflect errors in the known motifs in some cases, and imperfect ChIP data in others. Genome-wide analysis may simply be more powerful than category-based analysis in some instances.

In all, 46 of the 55 known motifs were found by either genome-wide or category-based analysis. The remaining nine cases may reflect true failures of the comparative genomic analysis or errors in the known motifs.

New motifs

We then applied the approach to all of the gene categories. A total of 181 well-conserved motifs were identified, with many of these being equivalent motifs arising from multiple categories. Merging such motifs resulted in 52 distinct motifs, of which 43 were already found by the analyses described above. The remaining nine motifs represent new category-based motifs (Table 4), including the following.

- Three new motifs are associated with genes that are bound by the transcription factors Rap1, Ste12 and Cin5, respectively. Rap1 is known to bind incomplete or degenerate instances of the published motif (Table 2), and the new motif may confer additional specificity. The motif associated with Ste12 is the known binding site for the partner transcription factor Tec1, suggesting that Ste12 binding is strongly associated with its partner under the conditions examined (see below). Similarly, the new motif associated with Cin5 may be that of a partner transcription factor.
- Three new motifs are associated with the GO category for carbohydrate transport, fatty-acid oxidation and glycolysis–glycogenesis, respectively.
- Three new motifs are associated with an expression cluster (cluster 37) that includes many genes involved in energy metabolism and stress response. The cluster is also associated with a new genome-wide motif (motif 25).

Category-based motif discovery thus contributes a modest number of additional motifs beyond those found by genome-wide analysis.

Table 4 **Additional new motifs discovered by category-based analysis**

No.	Category*	Category-based motif†	Interpretation	Score‡
1	Exp.: cluster 37	YCCCTTAAA	New: cluster 37 (Msn2/4-like)	[8.5]
2	ChIP: *FHL1* in YPD	ATGTACGGATG	New: Rap1 alternate	[7.6]
3	GO: carbohydrate transport	GTTTTTCCG	New: carbohydrate transport	[7.2]
4	GO: fatty acid beta-oxidation	TTAnnnCCG	New: fatty acid oxidation	[6.3]
5	GO: glycolysis/glyconeogenesis	TAGTGGAAGC	New: glycolysis/glycogenesis	[6.0]
6	Exp.: cluster 37	TCAGCC	New: cluster 37	[5.9]
7	Exp.: cluster 37	CGGnnnnnCGG	New: cluster 37	[5.7]
8	ChIP: *CIN5* in YPD	GnTTAnnTnAGC	New: Cin5 alternate	[5.6]
9	ChIP: *STE12* in butanol	CATTCT	Known: Tec1	[5.4]

*Category used in search.
†Symbols as in Tables 2 and 3.
‡Category-based enrichment score.

© 2003 Nature Publishing Group

Combinatorial control

Given a limited number of genome-wide regulatory motifs and a large number of transcriptionally regulated processes, fine-grained regulation may depend on combinatorial control. We sought to explore such combinatorial control by searching for motifs that occur in the same intergenic regions much more frequently than would be expected by chance.

In a single genome, few significant correlations are found. This is because functional instances of the motif are overwhelmed by a much larger number of random occurrences. Cross-species conservation greatly decreases this random noise and reveals biologically meaningful correlations. We describe a few examples here.

The Ste12 and Tec1 motifs show clear correlation, with about 20% of regions having a conserved occurrence of one also having a conserved occurrence of the other. This enrichment is not apparent when considering *S. cerevisiae* alone. We found that the genes that contain only the conserved Ste12 motif are enriched for those involved in mating and pheromone response, whereas those that contain conserved occurrences of both the Ste12 and Tec1 motifs are enriched for those involved in filamentous growth. These computational observations are consistent with recent work showing genome-wide evidence that Ste12 and Tec1 cooperate during starvation to induce filamentation-specific genes[49]. We also found that genes that contain only conserved occurrences of the Tec1 motif are enriched for those involved in budding and cell polarity, suggesting that Tec1 has functions that do not require cooperative binding with Ste12.

About 60% of regions containing conserved motifs for the transcription factor Leu3 (which regulates branched-chain amino-acid biosynthesis) also contain conserved motifs for Gcn4 (a general factor regulating amino-acid biosynthesis, as well as many other processes). About 46% of regions containing conserved motifs for the transcription factor Met31 also contain conserved occurrences of Cbf1. In fact, Cbf1 (which is involves in DNA bending) is known to physically interact and cooperate with the MET regulatory complex. About 34% of regions containing a conserved Gal4 motif also contain a conserved Mig1 motif. In this case, the correlation reflects antagonistic interaction. Gal4 induces galactose metabolism genes in the presence of galactose, but Mig1 represses galactose metabolism in the presence of glucose.

Pairwise co-occurrence connects a group of five motifs: Msn2/4 (general stress response), Rlm1 (response to cell-wall stresses), Pdr1 (pleiotropic drug resistance), Tea1 (Ty element activator) and Tbf1 (Telomere-binding factor). This suggests a possible link between various stress responses and adaptive changes at the genome level[50].

Many additional correlations are seen among known and new motifs and can be used to construct comprehensive co-occurrence networks. These can provide information valuable in deciphering biological pathways in yeast.

Discussion

The goal of this paper is to explore the ability to extract a wide range of biological information from genome comparisons between related organisms. For *S. cerevisiae*, our results show that comparative genome analysis of a handful of related species has substantial power to identify genes, define gene structure, highlight rapid and slow evolutionary change, recognize regulatory elements and reveal combinatorial control of gene regulation. The power is comparable or superior to experimental analysis, in terms of sensitivity and precision.

In principle, the approach could be applied to any organism by selecting a suitable set of related species. The optimal choice of species depends on multiple considerations, largely related to the evolutionary tree connecting the species. These include the following: first, the branch length t between species should be short enough to permit orthologous sequence to be readily aligned. The yeasts studied here differ by $t = 0.23$–0.55 substitutions per site and are readily aligned. The strong conservation of synteny allowed the unambiguous correspondence of the vast majority of genes.

Second, the total branch length of the tree should be large enough that non-functional sites will have undergone substantially more drift than functional sites, thereby providing an adequate degree of signal-to-noise enrichment. For this analysis, the multiple species studied provide a total branch length of 0.83 and a probability of nucleotide identity across all four species in non-coding regions of 49%. The signal-to-noise enrichment is thus about twofold (that is, 1/0.49) for highly constrained nucleotides and correspondingly higher for composite features involving many nucleotides.

Third, the species should represent as narrow a taxon as possible, subject to the considerations above. Because the comparative analysis above seeks to identify genomic elements common to the species, it can explain only aspects of biology shared across the taxon. In the present case, the analysis identifies elements shared across *Saccharomyces sensu stricto*, a closely related set of species in which most of the genes and regulatory elements are shared.

What are the implications for the understanding of the human genome? The present study provides a good model for evolutionary distances (substitutions per site in intergenic regions) relevant to the study of the human. The sequence divergence between *S. cerevisiae* and the most distant relative *S. bayanus* (11% indels and 62% nucleotide identity in aligned positions) is similar to that between human and mouse (12% indels and 66% nucleotide identity in aligned positions).

An important difference between yeast and human is the inherent signal-to-noise ratio in the genome. Yeast has a high signal-to-noise ratio, with protein-coding regions comprising approximately 70% of the genome, and regulatory elements comprising perhaps about 15% of the intergenic regions. The human has a much lower signal-to-noise ratio, with the corresponding figures being perhaps ~2% and ~3%[7]. A lower signal-to-noise ratio must be offset by a higher signal-to-noise enrichment. Some enrichment can also be obtained by filtering out the repeat sequences that comprise half of the human genome. Greater enrichment can be accomplished by increasing the number of species studied, taking advantage both of nucleotide-level divergence and frequently occurring genomic deletion[7].

Such considerations indicate that it should be possible to use comparative analysis, such as explored here for yeast, to identify directly many functional elements in the human genome that are common to mammals. More generally, comparative analysis offers a powerful and precise initial tool for interpreting genomes. □

Methods

Strain information

The sequence and annotation for *S. cerevisiae* reference strain S288C was obtained from the SGD website (http://www-genome.stanford.edu/Saccharomyces/) in May 2002. *Saccharomyces paradoxus* strain NRRL Y-17217, *S. mikatae* strain IFO1815, and *S. bayanus* strain MCYC623 were provided by E. Louis (University of Leicester). All sequenced isolates were diploid.

Determining one-to-one orthologous ORFs

For each species, we gathered the complete set of all predicted ORFs starting with a methionine and having a length of at least 50 amino acids. On the basis of blast hits, we connected predicted ORFs to *S. cerevisiae* ORFs in a bipartite graph, which we automatically separated into subgraphs by eliminating suboptimal edges in a series of steps. We first eliminated all edges that are less than 80% of the maximum-weight edge both in amino-acid identity and in length. On the basis of the resulting unambiguous matches, we built blocks of conserved gene order (synteny). We then preferentially kept matches within synteny blocks and resolved additional matches. We finally separated gene families into subfamilies by searching for best unambiguous subgraphs. We reported the connected components of the resulting graph as homology groups of unresolved genes.

Motif conservation score

To evaluate the motif conservation score (MCS) of a motif m of given length and degeneracy, we compared its conservation ratio to that of random patterns of the same length and degeneracy. We first computed the table F containing the relative frequencies of twofold and threefold degenerate bases, given the *S. cerevisiae* nucleotide frequencies. We then selected 20 random intergenic loci in *S. cerevisiae* and used the order of nucleotides at

© 2003 Nature Publishing Group

each locus together with the order of degeneracy levels in m to construct 20 random motifs that were used as controls. We counted conserved and non-conserved instances of each control and computed r, the log-average of their conservation rates. We also counted the number of conserved and non-conserved intergenic instances of m, and evaluated the binomial probability p of observing the two counts, given r. We finally reported the MCS of the motif as a z-score corresponding to p, the number of standard deviations away from the mean of a normal distribution that corresponds to tail area p.

Conservation criteria

For CC1, we searched for mini-motifs that show a significant conservation in intergenic regions. For every mini-motif, we count i_c, the number of perfectly conserved intergenic instances in all four species, and i, the total number of intergenic instances in *S. cerevisiae*. We found that the two counts seem linearly related for most of the patterns (Fig. 7a), which can be attributed to a basal level of conservation r given the total evolutionary distance that separates the four species compared. We estimated the ratio r as the log-average of non-outlier instances of i_c/i within a control set of all motifs at a given gap size. We then calculated for every motif the binomial probability p of observing i_c successes out of i trials, given parameter r. We assigned a z-score S to every motif corresponding to probability p. Similar methods were used for computing CC2 and CC3 (see Supplementary Information).

Constructing full motifs

We extended each mini-motif selected by searching for surrounding bases that are preferentially conserved when the motif is conserved. We used an iterative approach adding at every iteration one base that maximally discriminates the neighbourhood of conserved motif instances from the neighbourhood of non-conserved motif instances. The added base was selected from 14 degenerate symbols of the IUB code. When no such symbol separated the conserved and non-conserved instances with significance above 3σ, we terminated the extension. Figure 7d shows the top-scoring mini-motif found in CC1 (panel 1), and the corresponding extension (panel 2). We found that many mini-motifs have the same or similar extensions, and we clustered the extended motifs hierarchically based on sequence similarity. We then computed a consensus sequence for every cluster of extended motifs, resulting in a smaller number of mega-motifs for each test (332 for CC1, 269 for CC2, and 285 for CC3). Panel 3 shows the first nine members of the top cluster in CC1, and the resulting mega-motif. Finally, we merged mega-motifs based on their co-occurrence in the same intergenic regions (panel 4). We computed a hypergeometric co-occurrence score between the intergenic regions containing each mega-motif and again collapsed these hierarchically. We computed a consensus for every cluster, and iterated the co-occurrence-based collapsing step (results not shown). We obtained fewer than 200 distinct genome-wide motifs, of which 72 have MCS ≥ 4 (Table 3).

Additional methods

See Supplementary Information for a complete description of methods for sequencing, assembly, annotation, nucleotide alignments, genomic rearrangements, ambiguous ORFs, reading frame conservation, gene boundaries, resequencing, intron finding, genome-wide motif discovery, motif conservation score, the three mini-motif conservation criteria, motif extension and collapsing, and category-based motif discovery.

Received 27 February; accepted 1 April 2003; doi:10.1038/nature01644.

1. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274,** 546, 563–567 (1996).
2. Kowalczuk, M., Mackiewicz, P., Gierlik, A., Dudek, M. R. & Cebrat, S. Total number of coding open reading frames in the yeast genome. *Yeast* **15,** 1031–1034 (1999).
3. Harrison, P. M., Kumar, A., Lang, N., Snyder, M. & Gerstein, M. A question of size: the eukaryotic proteome and the problems in defining it. *Nucleic Acids Res.* **30,** 1083–1090 (2002).
4. Velculescu, V. E. *et al.* Characterization of the yeast transcriptome. *Cell* **88,** 243–251 (1997).
5. Blandin, G. *et al.* Genomic exploration of the hemiascomycetous yeasts: 4. The genome of *Saccharomyces cerevisiae* revisited. *FEBS Lett.* **487,** 31–36 (2000).
6. Wood, V., Rutherford, K. M., Ivens, A., Rajandream, M.-A. & Barrell, B. A Re-annotation of the *Saccaromyces cerevisiae* genome. *Comp. Funct. Genomics* **2,** 143–154 (2001).
7. International Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
8. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2,** 28–36 (1994).
9. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. Systematic determination of genetic network architecture. *Nature Genet.* **22,** 281–285 (1999).
10. Stormo, G. D. DNA binding sites: representation and discovery. *Bioinformatics* **16,** 16–23 (2000).
11. McGuire, A. M., Hughes, J. D. & Church, G. M. Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res.* **10,** 744–757 (2000).
12. Loots, G. G. *et al.* Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288,** 136–140 (2000).
13. Pennacchio, L. A. & Rubin, E. M. Genomic strategies to identify mammalian regulatory sequences. *Nature Rev. Genet.* **2,** 100–109 (2001).
14. Oeltjen, J. C. *et al.* Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res.* **7,** 315–329 (1997).
15. Cliften, P. F. *et al.* Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.* **11,** 1175–1186 (2001).
16. Alm, R. A. *et al.* Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* **397,** 176–180 (1999).
17. Carlton, J. M. *et al.* Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature* **419,** 512–519 (2002).
18. Perrin, A. *et al.* Comparative genomics identifies the genetic islands that distinguish *Neisseria meningitidis*, the agent of cerebrospinal meningitis, from other *Neisseria* species. *Infect. Immun.* **70,** 7063–7072 (2002).
19. McClelland, M. *et al.* Comparison of the *Escherichia coli* K-12 genome with sampled genomes of a *Klebsiella pneumoniae* and three *salmonella enterica* serovars, *Typhimurium*, *Typhi* and *Paratyphi*. *Nucleic Acids Res.* **28,** 4974–4986 (2000).
20. Batzoglou, S. *et al.* ARACHNE: a whole-genome shotgun assembler. *Genome Res.* **12,** 177–189 (2002).
21. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419,** 498–511 (2002).
22. Fischer, G., James, S. A., Roberts, I. N., Oliver, S. G. & Louis, E. J. Chromosomal evolution in *Saccharomyces*. *Nature* **405,** 451–454 (2000).
23. Dunham, M. J. *et al.* Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **99,** 16144–16149 (2002).
24. Blanchette, M. & Tompa, M. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.* **12,** 739–748 (2002).
25. Fischer, G., Neuveglise, C., Durrens, P., Gaillardin, C. & Dujon, B. Evolution of gene order in the genomes of two related yeast species. *Genome Res.* **11,** 2009–2019 (2001).
26. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387,** 708–713 (1997).
27. Bon, E. *et al.* Genomic exploration of the hemiascomycetous yeasts: 5. *Saccharomyces bayanus var. uvarum*. *FEBS Lett.* **487,** 37–41 (2000).
28. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
29. Dujon, B. *et al.* Complete DNA sequence of yeast chromosome XI. *Nature* **369,** 371–378 (1994).
30. Sharp, P. M. & Li, W. H. The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15,** 1281–1295 (1987).
31. Clark, T. A., Sugnet, C. W. & Ares, M. Jr Genome-wide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science* **296,** 907–910 (2002).
32. Hurst, L. D. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* **18,** 486 (2002).
33. Chu, S. *et al.* The transcriptional program of sporulation in budding yeast. *Science* **282,** 699–705 (1998).
34. True, H. L. & Lindquist, S. L. A yeast prion provides a mechanism for genetic variation and phenotypic diversity. *Nature* **407,** 477–483 (2000).
35. Koufopanou, V., Goddard, M. R. & Burt, A. Adaptation for horizontal transfer in a homing endonuclease. *Mol. Biol. Evol.* **19,** 239–246 (2002).
36. Haber, J. E. Mating-type gene switching in *Saccharomyces cerevisiae*. *Annu. Rev. Genet.* **32,** 561–599 (1998).
37. Hampson, S., Kibler, D. & Baldi, P. Distribution patterns of over-represented k-mers in non-coding yeast DNA. *Bioinformatics* **18,** 513–528 (2002).
38. McCue, L. *et al.* Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res.* **29,** 774–782 (2001).
39. Gelfand, M. S., Koonin, E. V. & Mironov, A. A. Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. *Nucleic Acids Res.* **28,** 695–705 (2000).
40. Keegan, L., Gill, G. & Ptashne, M. Separation of DNA binding from the transcription-activating function of a eukaryotic regulatory protein. *Science* **231,** 699–704 (1986).
41. Zhu, J. & Zhang, M. Q. SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics* **15,** 607–611 (1999).
42. Mewes, H. W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **27,** 44–48 (1999).
43. Dwight, S. S. *et al. Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.* **30,** 69–72 (2002).
44. Lee, T. I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298,** 799–804 (2002).
45. Gasch, A. P. & Eisen, M. B. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol.* **3** RESEARCH0059 (2002).
46. Mosley, A. L., Lakshmanan, J., Aryal, B. K. & Ozcan, S. Glucose-mediated phosphorylation converts the transcription factor Rgt1 from a repressor to an activator. *J. Biol. Chem.* **278,** 10322–10327 (2003).
47. Lindgren, A. *et al.* The pachytene checkpoint in *Saccharomyces cerevisiae* requires the Sum1 transcriptional repressor. *EMBO J.* **19,** 6489–6497 (2000).
48. Jacobs Anderson, J. S. & Parker, R. Computational identification of *cis*-acting elements affecting post-transcriptional control of gene expression in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **28,** 1604–1617 (2000).
49. Zeitlinger, J. *et al.* Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell* **113,** 395–404 (2003).
50. Morillon, A., Springer, M. & Lesage, P. Activation of the Kss1 invasive-filamentous growth pathway induces Ty1 transcription and retrotransposition in *Saccharomyces cerevisiae*. *Mol. Cell Biol.* **20,** 5766–5776 (2000).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We thank D. Botstein, M. Cherry, K. Dolinski, D. Fisk, S. Weng and other members of the *Saccharomyces* Genome Database staff for assistance with SGD, for making our data available to the community through SGD, and for discussions; J. Butler, S. Calvo, J. Galagan, D. Jaffe, J. Lehar and L. Jun Ma for technical advice and discussions; the staff of the Whitehead/MIT Center for Genome Research Sequencing Center who generated the shotgun sequence from the three yeast species; T. Lee, N. Rinaldi, R. Young and J. Zeitlinger for sharing data about chromatin immunoprecipitation experiments and for discussions; M. Eisen and A. Gasch for sharing information about gene expression clusters and for discussions; E. Louis and I. Roberts for providing yeast strains and discussions; B. Berger, G. Fink, D. Gifford, S. Lindquist and H. True-Krobb for discussions; and L. Gaffney for assistance with figures.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.K. (manoli@mit.edu) or E.S.L. (lander@wi.mit.edu). GenBank accession numbers are AABY01000000 for *S. paradoxus*, AABZ01000000 for *S. mikatae*, and AACA01000000 for *S. bayanus*.

 © 2003 Nature Publishing Group

Research

Open Access

Phylogenetically and spatially conserved word pairs associated with gene-expression changes in yeasts

Derek Y Chiang[*], Alan M Moses[†], Manolis Kellis[‡], Eric S Lander[§] and Michael B Eisen[¶¥]

Addresses: [*]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA. [†]Graduate Group in Biophysics, University of California, Berkeley, CA 94720, USA. [‡]Whitehead/MIT Center for Genome Research, Department of Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [§]Whitehead/MIT Center for Genome Research, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [¶]Department of Genome Sciences, Life Sciences Division, Ernest Orlando Lawrence Berkeley National Lab, 1 Cyclotron Road, Berkeley, CA 94720, USA. [¥]Center for Integrative Genomics and Division of Genetics and Development, Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA.

Correspondence: Michael B Eisen. E-mail: mbeisen@lbl.gov

Published: 26 June 2003

Genome ***Biology*** 2003, **4:**R43

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/7/R43

Received: 28 January 2003
Revised: 28 April 2003
Accepted: 15 May 2003

© 2003 Chiang *et al.*; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article's original URL.

Abstract

Background: Transcriptional regulation in eukaryotes often involves multiple transcription factors binding to the same transcription control region, and to understand the regulatory content of eukaryotic genomes it is necessary to consider the co-occurrence and spatial relationships of individual binding sites. The determination of conserved sequences (often known as phylogenetic footprinting) has identified individual transcription factor binding sites. We extend this concept of functional conservation to higher-order features of transcription control regions.

Results: We used the genome sequences of four yeast species of the genus *Saccharomyces* to identify sequences potentially involved in multifactorial control of gene expression. We found 989 potential regulatory 'templates': pairs of hexameric sequences that are jointly conserved in transcription regulatory regions and also exhibit non-random relative spacing. Many of the individual sequences in these templates correspond to known transcription factor binding sites, and the sets of genes containing a particular template in their transcription control regions tend to be differentially expressed in conditions where the corresponding transcription factors are known to be active. The incorporation of word pairs to define sequence features yields more specific predictions of average expression profiles and more informative regression models for genome-wide expression data than considering sequence conservation alone.

Conclusions: The incorporation of both joint conservation and spacing constraints of sequence pairs predicts groups of target genes that are specific for common patterns of gene expression. Our work suggests that positional information, especially the relative spacing between transcription factor binding sites, may represent a common organizing principle of transcription control regions.

Background

All organisms have evolved intricate signaling networks that sense and respond to their environment. At a cellular level, the activation of one or more signaling networks often leads to coordinated changes in gene expression, via the regulated activity and binding of transcription factors to transcription control regions (TCRs) of genes (for example, enhancers and upstream activating sequences). In yeast and most other eukaryotes, the transcriptional regulation of individual genes is often multifactorial, as multiple transcription factors may bind to a single TCR [1,2]. Multifactorial regulation encompasses several distinct biochemical mechanisms. In some cases, transcription factors may bind cooperatively to adjacent DNA sites via direct physical interaction [3,4]. In other cases, multiple transcription factors that bind independently may recruit a common co-activator [5], or may act independently of one another to alter gene expression in response to distinct cellular cues [6]. Recent studies have also suggested that nearby transcription factors may collaboratively compete with nucleosomes, thus enhancing the binding of individual transcription factors [7,8]. Many experiments in yeast have shown that specific pairs of factors must be bound near each other for multifactorial regulation to occur [7–10], and it is on these spatial constraints that we focus here.

The challenges in understanding how regulatory information is encoded in genomes include both the identification of regulatory sequences in TCRs and the elucidation of sequence constraints on productive multifactorial regulation. Previous computational work has been devoted to identifying putative binding sites for transcription factors. A plethora of computational methods has been developed to find over-represented sequences in a subset of genes believed to contain a common transcription factor binding site (reviewed in [11]). The rapid pace of genome sequencing has enabled a complementary approach - phylogenetic footprinting (reviewed in [12,13]) - which recognizes that the conservation of sequences across related organisms often reflects evolutionary selection for their presence in TCRs. Several algorithms have been developed to perform systematic phylogenetic footprinting analyses [14–16].

After compiling a collection of putative binding sites, associations can be made between various binding site assortments and gene expression. Some recent approaches include Boolean logic [17], regression methods [18–21], spatial clustering [22], and multiple binding site matrix classifiers [23–25]. Spatial information on the relative locations of binding sites is ignored in all but the last two classes of approaches.

The primary aim of this work was to incorporate positional information and phylogenetic footprinting in methods of identifying sequence motifs that may regulate gene expression. Consequently, we expanded the focus of phylogenetic footprinting from the conservation of contiguous sequences to higher-order features of TCRs, namely the spatial organization of individual binding sites. Because transcription factors participating in multifactorial regulation may require binding sites in physical proximity to each other, we searched for groups of conserved sequences that were more closely spaced in TCRs than expected. We refer to these spatially organized sequences as conserved 'word templates'. As a proof of principle, we started with the simplest example of such templates: pairs of conserved words of 6 base-pairs (bp). Conservation was assessed using the genome sequences of *Saccharomyces cerevisiae* and three closely related *Saccharomyces* species, which had been sequenced to identify conserved regulatory sequences [26]. To exploit this comparative genomic data, we have devised a method that systematically tested sequence pairs for joint conservation across genomes and close spacing within individual TCRs. As genes regulated by the same set of transcription factors often display similar gene-expression patterns in certain experimental conditions, we identified pairs of conserved word templates whose gene targets were associated with common changes in gene expression. We adopted a group-by-sequence approach to first identify genes that contained the word-pair templates and then to test for significant associations with expression levels of the identified genes [27]. Significant associations between conserved word-pair templates and specific gene-expression changes and the prevalence of known transcription factor binding sites suggest that conserved word-pair templates comprise sequences important for multifactorial regulation in yeast. In addition, conserved word-pair templates represent more specific predictors of gene expression than individual words or word pairs in *S. cerevisiae*.

Results

Identification of conserved word-pair templates

Multiple genome sequences provide additional power to studies of gene regulation. Because of natural selection, mutations accumulate more rapidly in non-functional DNA regions than in functionally constrained bases. Given a multiple sequence alignment of orthologous sequences from closely related species, the aligned and invariant regions should be enriched for functionally important residues [12,13]. Additional *Saccharomyces* genomes were sequenced to ensure sufficient sequence similarity to *S. cerevisiae* such that orthologous regions could be reliably aligned, yet enough sequence divergence that functional sequences would be much more conserved than nonfunctional sequences [26,28]. To confirm that regulatory sequences were found in conserved regions, we tested a database of 47 known, nonredundant regulatory motifs and found that 35 show conservation ratios that were more than three standard deviations above that expected by chance [26,29].

We present a method to find conserved higher-order sequence templates from related *Saccharomyces* genomes (Figure 1). Our method incorporates sequential statistical tests, with each step focusing on a distinct property of

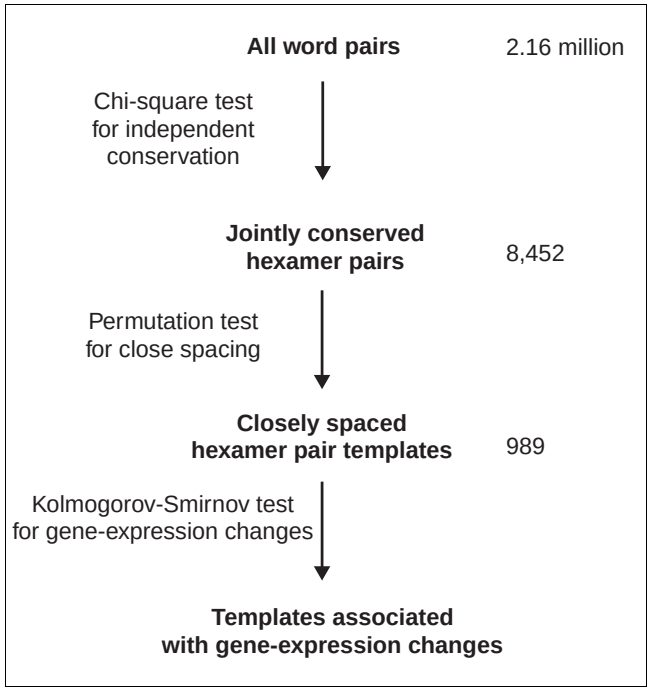


Figure 1
Overview of the method used to discover conserved word-pair templates in yeast. For the templates associated with gene-expression changes see Figure 3.

conserved sequence templates. The simplest instances of sequence templates involve word pairs and their relative spacing. As described in detail below, pairs of words that were conserved in the same intergenic regions of four *Saccharomyces* genomes were identified using a chi-square test for independence. Next, a permutation test was used to select word pairs whose physical proximity was closer than that expected by chance. Finally, to evaluate the transcriptional information contained in conserved word pairs with close spacing, the expression of genes containing TCR templates was compared to the rest of the genome. We initialized our word list using all 2,080 words of length 6, treating a given word and its reverse complement as identical. For each TCR (consisting of up to 600 bp upstream of an open reading frame), a word was labeled conserved if all six bases were identical in at least three of the four *Saccharomyces* genomes, on the basis of the CLUSTALW alignment of that TCR.

To test systematically whether words were conserved more often in the same intergenic regions of the *Saccharomyces* genomes than expected by independent conservation, a chi-square test was performed on all possible pairwise combinations of words (see Materials and methods). Pairs of words that overlapped each other by more than three nucleotides were excluded. A significant proportion of word pairs showed dependent conservation: among the 2.16 million word pairs tested, 8,452 of them (approximately 0.4%) had conservation χ^2 scores greater than 31.1. This threshold corresponds to a probability of 0.05 for obtaining one or more false positives after a Bonferroni correction for multiple testing.

Next, we selected word pairs that displayed closer physical spacing in intergenic regions than expected by chance. The choice of a statistical test to evaluate close distances must consider the local fluctuations of A+T nucleotide content in genome sequences. Previous work used a Poisson distribution to evaluate proximity between binding sites [22]. However, variability in base composition can skew occurrences of arbitrary sequences away from their expected distributions. Indeed, this statistical test was confounded by large fluctuations in the Poisson parameter estimates, which varied up to twofold within a single chromosome [22].

The effects of base-composition fluctuations, as well as varying lengths of TCRs, motivated our nonparametric statistical test for close spacing. We used the median, denoted by $\overline{D}$, to summarize a distribution of minimum distances between two words in *S. cerevisiae*. This distribution was calculated on the basis of the genes whose TCRs conserved both words, and is independent of the relative word order. If two nonoverlapping words were closely spaced in all TCRs, we should find $\overline{D}$ to be smaller than expected by chance. The statistical significance of this spacing was assessed using a permutation test by selecting the set of genes that contained a conserved word pair and then randomizing the assignment of one of the words to the genes containing that word (see Materials and methods). By permuting the TCR labels for one of the words, but not the word positions themselves, we retained the positional biases of individual words within intergenic regions.

After correcting for multiple testing by controlling the false discovery rate (FDR), a total of 989 out of 8,452 word pairs (around 12%) had significantly small values (FDR $q < 0.05$) for $\overline{D}$ (Figure 2). For a list of these closely spaced and jointly conserved word pairs see Additional data files. As a negative control, we also assayed a sample of word pairs that did not show dependent conservation (conservation $\chi^2 < 1$), yet were jointly conserved in at least 10 TCRs. No word pairs in a random sample of 42,718 pairs with nondependent conservation ($\chi^2 < 1$) showed significantly small values for $\overline{D}$, after correction for multiple testing. Figure 2 illustrates the distributions of $\overline{D}$ for conserved word-pair templates, jointly conserved word pairs, and randomly conserved word pairs. The medians of these distance distributions were 54 nucleotides, 73 nucleotides and 89 nucleotides, respectively. Notably, the median $\overline{D}$ for template pairs was significantly smaller ($p < 0.05$) than the median $\overline{D}$ for randomly conserved pairs. These results indicate that many of the word pairs that were conserved in the same intergenic regions of multiple *Saccharomyces* genomes also exhibited closer spacing in TCRs.

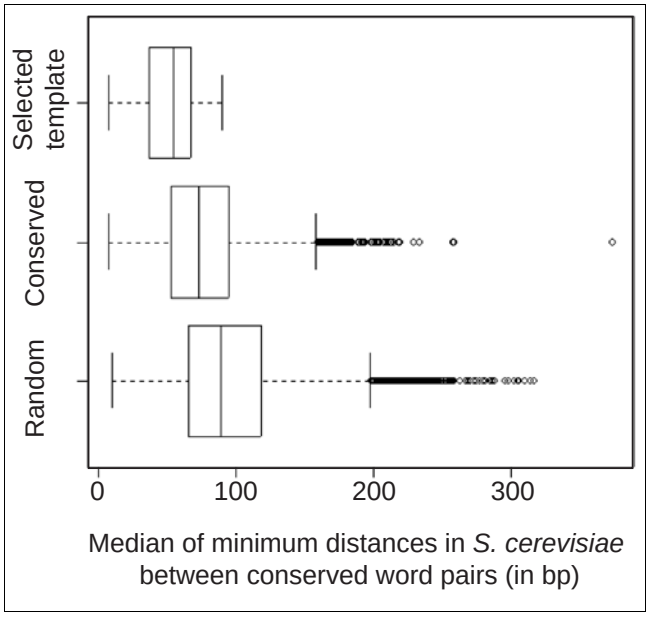


Figure 2
Word pairs in conserved word-pair templates are closely spaced in *S. cerevisiae* A comparison of the median of minimum distances is shown for $\overline{D}$ three categories of word pairs. For each category, the distribution of median of minimum distances is represented by a box-and-whisker plot, which was generated using the statistical software package R [51]; the box extends from the 25th percentile to the 75th percentile, and the vertical line within the box denotes the median of the distribution. Dashed lines extend for 1.5 times the range of the box, and circles indicate extreme values. 'Selected template' denotes closely spaced and jointly conserved word pairs ($\chi^2 > 31.1$, spacing $q < 0.05$, $N = 989$). 'Conserved' denotes dependently conserved word pairs that occur in at least 10 intergenic regions ($\chi^2 > 31.1$, $N = 3,726$) and includes all of the word pairs in the 'selected template' category. 'Random' denotes a sample of randomly conserved word pairs that occur in at least 10 intergenic regions ($\chi^2 < 1$, $N = 42,718$).

Conserved word-pair templates were significantly associated with gene expression

Our method identified conserved word-pair templates that were statistically significant with respect to both co-conservation in multiple genomes and close spacing in *S. cerevisiae* TCRs. To evaluate the regulatory information in these templates, we assessed the statistical association between gene groups that shared a template and changes in gene expression. Similarly to other group-by-sequence approaches for finding regulatory sequences, we expect that gene subsets defined by common TCR sequence features should have gene-expression patterns that are similar under conditions where the transcription factors are active, yet are different from the average expression of genes in the genome [27].

To assess the association between conserved word-pair templates and differentially expressed genes, we identified gene subsets whose TCRs contained both conserved words in the template and observed their expression patterns in *S. cerevisiae* in publicly available datasets ([30–35], see Materials and methods for details). We then conducted Kolmogorov-Smirnov (K-S) tests to evaluate for differential gene expression between each gene subset and the whole genome. K-S tests provide a nonparametric, sensitive and robust way to compare two distributions. Similar results were obtained using other statistical tests, such as *t*-tests and likelihood ratio tests (A.M. Moses, unpublished data). A $P \times C$ matrix was computed: each conserved word pair in P was assigned a K-S p-value for each experimental condition observed in C (see Materials and methods). Entries in this matrix (K-S p-values) were filtered out if the K-S p-value: did not meet the threshold for multiple testing or was less than 10 times more significant than the K-S p-value for a gene subset associated with either word alone (see Materials and methods). The latter criterion minimizes gene-expression changes that can be explained by the presence of a single conserved word.

Figure 3a shows the number of conserved word-pair templates that were significantly associated with gene-expression changes, for varying significance levels of the K-S test, which have been corrected for multiple testing (see Materials and methods). Each line indicates the number of gene subsets that were significant in a different minimum number of experimental conditions. Several hundred closely spaced word pairs were significantly associated with differential gene expression. For example, 314 word pairs met an FDR-corrected significance threshold of $p < 10^{-3}$ for five or more experimental conditions, which represented approximately 32% of all closely spaced word pairs.

The proportion of conserved word-pair templates showing significant associations with gene expression was compared to two sets of negative controls, comprising word pairs that failed either the first (co-conservation) or second (close spacing) statistical test. For the first control, we used a sample of 624 word pairs that failed the joint conservation test (conservation $\chi^2 < 1$) found in at least 25 TCRs, but also showed modest constraints on word pair spacing ($p < 0.15$). Only eight of these word pairs (approximately 1.3%) were significantly associated with gene-expression changes at an FDR-corrected threshold of $p < 10^{-3}$ for five or more experimental conditions. To assess the relative enrichment for significant associations with gene-expression changes at a variety of multiple testing thresholds, we computed an odds ratio: the proportion of significant associations among the template pairs, divided by the proportion of significant associations among the random pairs. For the above threshold, the odds ratio was about 22. In other words, gene groups that contain a common conserved word-pair template in their TCRs were about 22 times more likely to be associated with significant gene-expression changes, compared with gene groups selected using randomly conserved word pairs. As shown in Figure 3b, the odds ratios for association with gene expression changes in multiple conditions varied between 10 and 35. This analysis was repeated for a sample of 2,737 co-conserved (conservation $\chi^2 > 31.1$) word pairs that failed the close spacing test (permutation $p > 0.05$ after multiple testing), yet

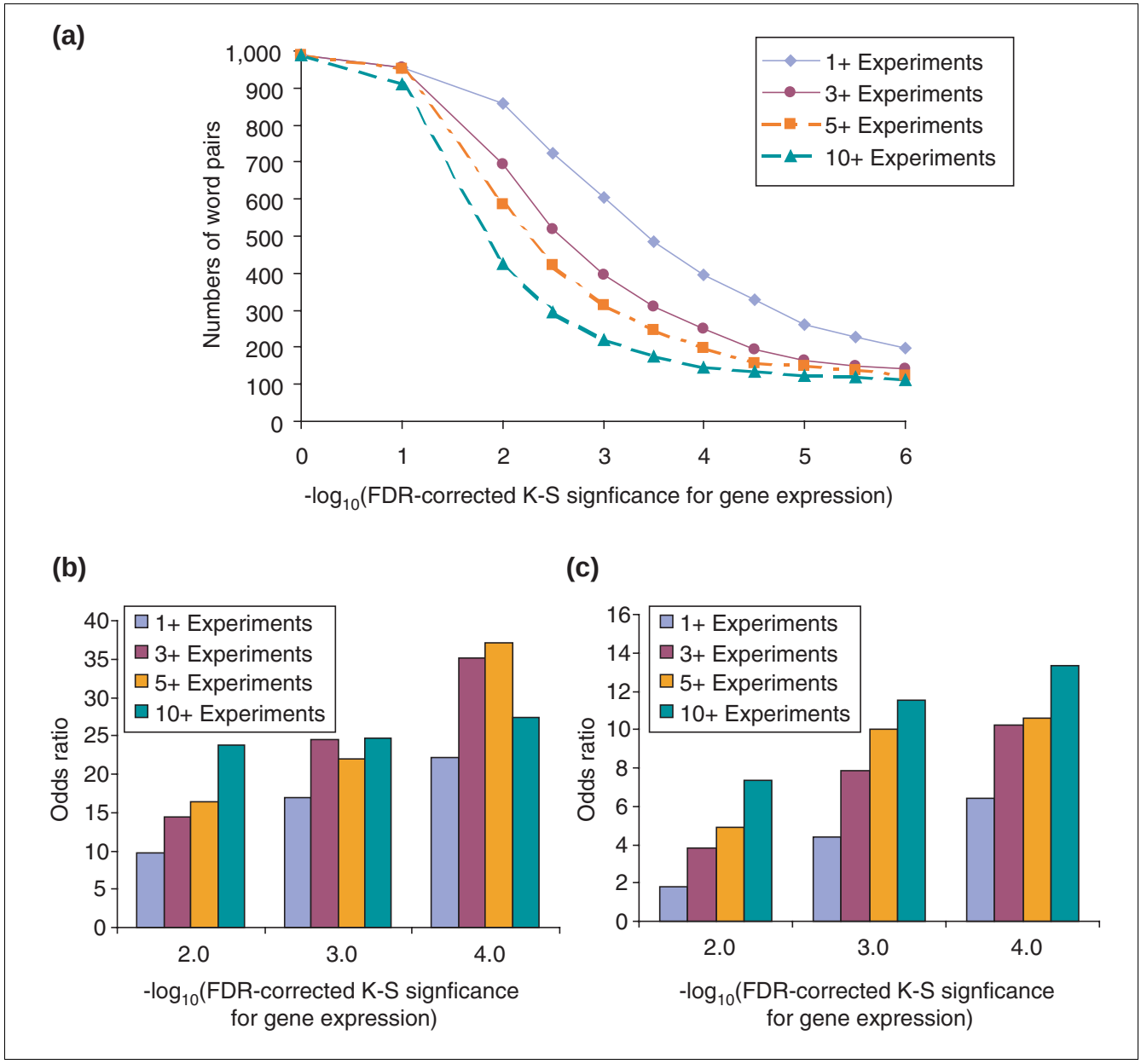


Figure 3
Conserved word-pair templates are associated with significant changes in gene expression. **(a)** Total number of conserved word-pair template associations at different K-S significance values. The horizontal axis shows different FDR-corrected significance levels for the Kolmogorov-Smirnov test (see Materials and methods). The number of closely spaced word pairs meeting this cutoff for different minimum numbers of expression conditions is shown on the vertical axis. Word pairs were also filtered for an improvement of 10× over the K-S significance from any single word. **(b)** Relative enrichment for significant gene-expression associations compared to independently conserved words. Relative enrichment was computed as an odds ratio: the fraction of gene groups selected by conserved word-pair templates associated with significant gene-expression changes, divided by the fraction of gene groups selected by randomly conserved word pairs associated with significant gene-expression changes. Templates were chosen as the set of 989 word pairs showing dependent conservation and close spacing ($\chi^2 > 31.1$, spacing $q < 0.05$); the random word pairs included 624 pairs showing independent conservation and modest spacing constraints ($\chi^2 < 31.1$, spacing $q < 0.15$). The odds ratio is shown on the vertical axis; various FDR-corrected significance thresholds for gene-expression associations are shown on the horizontal axis. Word pairs were filtered for an improvement of 10× over the K-S significance from any single word. **(c)** Relative enrichment for significant gene-expression associations compared to co-conserved words that failed the close spacing test. Templates were chosen as the set of 989 word pairs showing dependent conservation and close spacing ($\chi^2 > 31.1$, spacing $q < 0.05$); the background word pairs included 2,737 pairs showing co-conservation, but no significant close spacing constraints ($\chi^2 > 31.1$, spacing $q > 0.05$). The odds ratio is shown on the vertical axis; various FDR-corrected significance thresholds for gene-expression associations are shown on the horizontal axis. Word pairs were filtered for an improvement of 10× over the K-S significance from any single word.

occurred in at least 10 intergenic regions. The relative enrichment for gene expression associations in closely spaced words is displayed in Figure 3c. Among co-conserved word pairs, those pairs that were closely spaced than expected were still about 2 to 12 times more likely to be significantly associated with gene expression changes, compared to word pairs that were not found to have significantly close spacing. We confirmed that gene groups associated with significant gene expression changes did not have statistically significant differences in their TCR sizes, as assessed by a permutation test (data not shown). Thus, gene groups that contained co-conserved and spatially close word pairs are more significantly associated with gene-expression changes than expected by chance.

Many identified sequences represented known transcription factor binding sites

In addition to their statistical significance, many conserved word-pair templates that were most strongly associated with gene-expression changes were consistent with biological information on the transcription factors known to bind those sites [36]. In all analyses described below, we used a set of 314 word pairs that had significant associations with gene-expression changes at an FDR-corrected multiple testing threshold of $p < 10^{-3}$ for five or more experiments. For visualization purposes, we organized the $P \times C$ matrix by hierarchically clustering the K-S p-values for the 314 word pairs (see Materials and methods for details).

Hierarchical clustering of this output matrix identified groups of word pairs with similar K-S p-values in specific subsets of experimental conditions (Figure 4). For these clustered output matrices, see Additional data files. In many cases, the word pairs that clustered together also comprised overlapping hexamer sequences, suggesting that some of the hexamers in different pairs may represent a larger, somewhat variable sequence (Table 1). For example, group 9 in Figure 4 included six word pairs. In each of these word pairs, one of the component words - such as TCACGT, CACGTG, or ACGTGC - matched part of the consensus binding site for Cbf1p (TCACGTG). The other component word in each pair - such as ACTGTG, CTGTGG, TGTGGC or GTGGCT - represented part of the known Met31/32p binding site (AAACTGTGG). Therefore, genes whose TCRs contained any word pair within this group probably contained a conserved Cbf1p-binding site, along with a conserved Met31/32p-binding site, and the distances between the conserved sites in these genes were also smaller than expected by chance. These results agree with the known interaction of Cbf1p and Met31/32p for the regulation of genes involved in sulfur utilization (see Discussion).

Table 1 shows a partial list of the 14 most significant groups of consensus sequences, which were assembled by joining adjacent word pairs in the clustered output matrix with overlapping sequences. Many of these consensus sequences matched transcription factor binding sites that had been biochemically verified. Several pairs of transcription factors, denoted by stars in Table 1, were not previously known to act on the same sets of target genes.

Conditions with significant gene-expression changes coincided with transcription factor activity

Further support that templates contain transcriptional regulatory information was obtained from a key observation: the experimental conditions with significant gene-expression changes often corresponded to conditions in which the cognate transcription factors are known to be active (Table 2). In addition, many gene subsets that shared an individual word-pair template in their TCRs were highly enriched for gene-expression changes. We will survey examples of word-pair templates associated with gene-expression changes, focusing our attention on several environmental stress conditions. The environmental stress response represents a paradigm for multifactorial control of transcription regulation. Genome-wide expression studies found that around 300 genes were induced and around 600 genes were repressed in response to a wide variety of stressful environmental transitions [30,37]. Many of these genes also showed subtly different expression patterns in response to specific stimuli, suggesting that the common environmental stress response may be modulated by the activity of condition-specific transcription factors [30].

Over a third of the conserved word-pair templates were associated with gene-expression changes in multiple environmental stress conditions (Figure 4a, Tables 1, 2). The largest group of overlapping word pairs contained matches to the PAC and RRPE motifs, which were associated with genes that were repressed in multiple stresses [30,38]. These motifs were discovered by their enrichment among the approximately 600 genes that were commonly repressed in stress, yet the

Figure 4 *(see following page)*
Specific patterns of gene-expression changes are associated with templates. Conserved word-pair templates are shown with significant associations with gene-expression changes in **(a)** multiple environmental stress conditions or **(b)** distinct subsets of environmental conditions. The $P \times C$ matrix of K-S p-values was hierarchically clustered by rows and visualized with TreeView [52]. Each row corresponds to a conserved word-pair template, and each column represents a single gene-expression experiment. The experimental conditions are indicated by the color bar above each panel, according to the key. The value in each cell corresponds to the K-S p-value of gene-expression changes in each condition (column) for a group of genes that contain the conserved word-pair template (row) in their TCRs. Orange denotes a K-S p-value below the FDR critical value of 0.001 for multiple testing, while gray represents values that were not significant. Word pairs that failed to meet a FDR critical value of 0.001 for multiple testing in five or more experiments are not shown. Some of the most significant conserved word-pair associations are labeled and annotated in Tables 1 and 2.

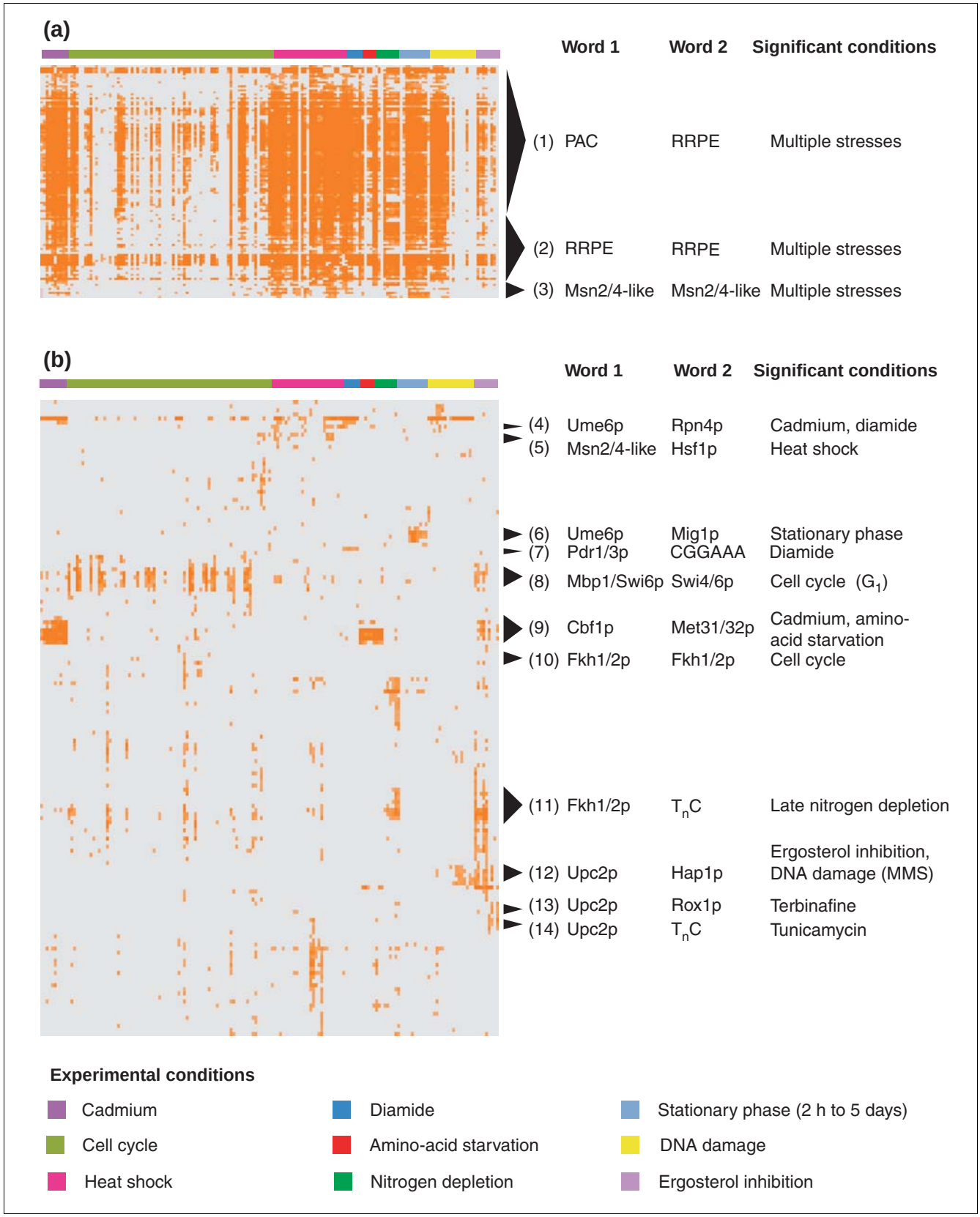


Figure 4 *(see legend on previous page)*

Table 1

Consensus sequences for the most significant groups of word pairs

	Hexamer list for word 1	Compiled sequence 1	TF for consensus 1	Hexamer list for word 2	Compiled sequence 2	TF for consensus 2	Number of word pairs
1	GAGATG GCGATG AGATGA CGATGA GATGAG ATGAGA ATGAGC TGAGAT TGAGCT GAGATG AGATGA AGCTCA	GMGATGAGMTSA	Unknown (PAC motif [38])	TGAAAA GAAAAA AAAAAT AAAATT AAATTT	TGAAAATTT	Unknown (RRPE motif [38])	75
2	AAGTGA AATGAA AGTGAA ATGAAA CTGAAA TGAAAA	ANTGAAAAA	Unknown (RRPE motif [38])	GAAAAA GAAAAT AAAATT AAATTT	GAAAAWTT	Unknown (RRPE motif [38])	40
3	GTTCCC CTCCCC ACCCCT TCCCCT	GYWCCCCT	(motif 38 [26])	CCCTTT CCTTTT CCTTAT	CCCTTWT	(motif 38 [26])	5
4*	GGCGGC GCGGCT	GGCGGCT	Ume6p	GTGGCA GGCAAA	GTGGCAAA	Rpn4p	2
5	CCCTTT CCTTTT	CCCTTTT	Msn2/4p-like	GGAGAA GGGAAA	GGRGAAA	Hsf1p	2
6	CGGCGG	CGGCGG	Ume6p	TACCCC ACCCCA CCCCAA	TACCCCAA	Mig1p	3
7*	CCGCGG	CCGCGG	Pdr1/3p	CGGAAA	CGGAAA	Unknown	1
8	AACGC GACGCG AACGCG ACGCGT ACGCGA TCGCGT CGCGTC	ARWCGCGW	Mbp1p	CGCGAA ACGAAA GCGAAA CGAAAC CGAAAA	CRCGAAAM	Swi4/6p	9
9	TCACGT CACGTG ACGTGC	TCACGTGC	Cbf1p	ACTGTG CTGTGG TGTGGC GTGGCT	ACTGTGGCT	Met31/32p	6
10	TATTTT TTTTGT TTTGTT ATTGTT	TWTTGTT	Fkh1/2p	TGTTTA GTTTAC	TGTTTAC	Fkh1/2p	4
11	TTTGTT TTGTTT	TTTGTTT	Fkh1/2p	TTTTTC TTTTTT	TTTTTY	T_nC	4
12*	TCGTTT CGTTTA	TCGTTTA	Ecm22p | Upc2p	CCGATA CGATAA	CCGATAA	Hap1p	4
13	TCGTTT CGTTTA	TCGTTTA	Ecm22p | Upc2p	TATTGT ATTGTT	TATTGTT	Rox1p	2
14	CGTTTC GTTTCT	CGTTTCT	Ecm22p | Upc2p	TTCTTT TCTTTT CTTTTT	TTCTTTTT	T_nC	5

See legend on next page

Table 1 *(see table on previous page)*

The output ***P*** × ***C*** matrix of word pairs (***P***) that were significantly associated ($p < 0.001$) with at least five or more environmental conditions (***C***) was ordered using hierarchical clustering. Numbers correspond to groups of overlapping word pairs indicated in Figure 4. Asterisks denote sequence pairs whose involvement in multifactorial regulation has not been previously reported. Compiled sequences were assembled from groups of word pairs that were found in adjacent rows in the ordering of K-S *p*-values. As individual words must have passed all three statistical tests to be included in the output matrix, these consensus sequences may not reflect the actual biological specificities of conserved transcription factor binding sites (refer to [26,36] for a more complete list). Residues are shown in bold if they are invariant in at least two hexamers. Numbers denote the groups that are indicated in Figure 4. Multiple transcription factors that may bind the same sequence motif are separated by |. IUPAC codes used: K (G or T); M (A or C); R (A or G); S (C or G); W (A or T).

putative transcription factors that bind these sequences have yet to be determined. The second largest group of overlapping word pairs corresponded to the RRPE core, which is 10 nucleotides long, along with some flanking conserved bases. These repressed genes were enriched for rRNA processing genes, the group of genes in which this motif was originally identified [38]. Nine conserved word-pair templates contained sequences that matched most of the stress response element (STRE), the consensus site for the general stress transcription factors Msn2p/Msn4p. Genes that conserved both these words in their TCRs were significantly associated with gene-expression induction in multiple environmental stresses, including cadmium, heat shock, amino-acid starvation, nitrogen depletion and stationary phase. In most cases, the sequences comprising the word pairs were mutually overlapping. We interpret these sequences as representing different halves of the same binding site. Because our test for close spacing required non-overlapping sequences, the two words must have appeared over 6 bp away in TCRs. Thus, these genes have probably conserved at least two Msn2/4p-like consensus sequences in their TCRs.

Several groups of conserved word-pair templates only showed significant associations with gene expression changes in different subsets of stress conditions (Figure 4b, Tables 1, 2). For example, binding sites for Cbf1p and Met31/32p were found to co-occur in several conserved word-pair templates. Genes that contained conserved binding sites for these transcription factors in their TCRs were strongly induced in cadmium, amino-acid starvation and early nitrogen depletion. These conditions are consistent with the biological activity of these transcription factors, which induce transcription of sulfur-assimilation genes in response to the demand of sulfur-containing metabolites [5,39,40]. In another example, several word pairs comprising binding sites for the transcriptional repressors Mig1p and Ume6p were associated with induced gene expression in stationary phase. The observed derepression of Mig1p and Ume6p targets in stationary phase is consistent with the nuclear export of the Mig1p repressor under glucose limitation, as well as recent findings that carbon source genes can be Ume6p targets [41]. In addition, genes containing a conserved sequence similar to the consensus for Msn2/4p, an inducer of the environmental stress response, and the heat-shock transcription factor Hsf1p were significantly induced under heat shock. Once again, the conditions with most significant gene-expression changes corresponded to the known activities of the transcription factors involved.

Enrichment for known transcription factor targets among individual gene groups

Some groups of genes with shared word-pair templates were enriched for known targets of transcription factors. The vast majority of genes with conserved sites for both the Cbf1p and Met31/32p transcription factors were induced more than fourfold in cadmium, amino-acid starvation and early nitrogen depletion (Figure 5a). Half of these genes have confirmed roles in sulfur-utilization processes, such as methionine metabolism, sulfate assimilation, sulfate transport and sulfur amino acid metabolism [42]. Compared to the rest of the genome, the group of genes that conserved both of these words within their TCRs was highly enriched for sulfur-utilization genes (hypergeometric p-value $< 1 \times 10^{-16}$, after Bonferroni correction for multiple testing). In addition, we found three genes in this group (*GSH1*, *RAD59* and *BNA3*) with highly correlated expression under the above conditions, and thus may be commonly regulated by Cbf1p and Met31/32p, despite their lack of direct annotation as sulfur-utilization genes. The shared conservation of both the Cbf1p and Met31/32p sites provides further evidence that these genes comprise part of the cellular response to the demand for products of this pathway.

Genes with a conserved half-site for the Hap1p transcription factor, as well as a conserved Ecm22p/Upc2p binding site in their TCRs, were significantly associated with induction in the presence of drugs that inhibited ergosterol biosynthesis (Figure 5b). This group of 30 genes contained eight ergosterol biosynthesis genes; this proportion represented an enrichment compared to the rest of the genome (hypergeometric p-value $< 6 \times 10^{-6}$ after Bonferroni correction for multiple testing). The transcription factors Ecm22p and Upc2p have been shown to induce the expression of ergosterol biosynthesis genes in response to low intracellular concentrations of ergosterol, whereas Hap1p is known to regulate the expression of these genes according to the availability of heme and oxygen, which are required for the pathway [43,44].

Table 2

Summary for most significant groups of conserved word pairs

			Most significant word pair in consensus group			
	Conserved word pairs (compilation of overlapping words)	Known transcription factors or motifs	(χ^2, *p*-value via Bonferroni)	Median of min distance $\overline{D}$	Number of TCRs	Expression conditions with significant gene subsets (FDR significance)
1	G[AC]GATGAG, TGAAAATTT	PAC, RRPE	240.6 (10^{-49})	19±0.5	162	Repressed in multiple environmental stresses (10^{-6})
2	ANTGAAA, GAAAAWT	RRPE (Overlap)	96.9 (2×10^{-16})	43±11	68	Repressed in multiple environmental stresses (10^{-6})
3	CTCCCC, CCCTTA	Msn2/4p-like, (Overlap)	53.8 (5×10^{-7})	28±3.7	15	Induced in multiple environmental stresses (10^{-6})
4	GGCGGGC, GTGGCA	Ume6p, Rpn4p	43.7 (9×10^{-5})	48±16	25	Cadmium, diamide (10^{-4}) MMS, heat shock (10^{-3})
5	CCTTTT, GAGAAA	Msn2/4p, Hsf1p	56.2 (2×10^{-7})	54±5.4	69	Heat shock (10^{-4})
6	CCGCCG, ACCCCA	Ume6p, Mig1p	41.9 (2×10^{-4})	17±1.5	14	Stationary phase (10^{-6})
7	CCGCGG, CGGAAA	Pdr1/3p, Unknown	111 (2×10^{-19})	44±12	21	Diamide (10^{-3})
8	RACGCG, RCGAAA	Swi6p/Mbp1p, Swi4/6p,	83.0 (7×10^{-13})	33±5.0	33	Cell cycle, G1 phase (10^{-6})
9	GCACGTGC, ACTGTGGC	Cbf1p | Pho4p, Met31/32p	37.4 (2×10^{-3})	22±2.5	22	Cadmium (10^{-6})
10	T[AT]TTGTT, TGTTTA	Fkh1/2p (Overlap)	51.1 (2×10^{-6})	57±6.9	48	Cell cycle (10^{-3})
11	TTTGTT, TTTTTY	Fkh1/2p, T_nC	37.6 (2×10^{-3})	49±4.4	267	Late nitrogen depletion (10^{-3})
12	CCGATA, TCGTTT	Hap1p, Ecm22p | Upc2p	36.2 (4×10^{-3})	41±5.9	28	Ergosterol inhibition (10^{-4}) MMS (DNA damage) (10^{-3})
13	TCGTTT, TATTGTT	Rox1p, Ecm22p | Upc2p	58.8 (4×10^{-8})	55±0.5	69	Early menadione (10^{-3})
14	TGACTC, TCTTTT	Gcn4, T_nC	35.6	59±9.1	63	Amino-acid starvation (10^{-5})

Statistics are listed for one representative word pair for each group of overlapping word pairs, numbered as in Figure 4. Multiple transcription factors that may bind the same sequence motif are separated by |. To summarize the close spacing ($\overline{D}$) between conserved word pairs, we report the median of the distribution of minimum distances in *S. cerevisiae* ± standard deviation of the medians of the distribution of minimum distances in all four *Saccharomyces* genomes.

Conserved word pairs were more informative than sequence features derived from single words or single species

The above results from the K-S test strongly suggested that sequence features based on the co-conservation and close spacing of word pairs identified examples of multifactorial regulation. Two other statistical tests were also performed to examine how information contained in conserved word-pair templates compared to other sequence features derived from *S. cerevisiae*, or from single conserved words. Informative sequence features should be both highly specific (a high proportion of genes that share the feature should show gene-

Figure 5 *(see following page)*
Enrichment for known transcription factor targets among individual gene groups. Gene-expression patterns are shown for genes whose TCRs contain the known binding sites for: **(a)** Cbf1p (CACGTG) and Met31/32p (TGTGGC); or **(b)** Hap1p (CCGATA) and Ecm22p/Upc2p (TCGTTT). The genes are listed in ascending order of minimum distance between the two conserved words in the corresponding TCRs of *S. cerevisiae*. Each row in these diagrams represents a given gene's expression pattern under the conditions shown in each column: exposure to increasing concentrations of cadmium chloride (from 0.05 mM to 0.4 mM); an amino-acid starvation timecourse; a nitrogen-source depletion timecourse [30]; and growth in the presence of drugs or genetic alterations that inhibit ergosterol biosynthesis (*erg3*Δ, itraconazole, *erg28*Δ, overexpressed *ERG11*, *erg2*Δ, tunicamycin, terbinafine, *erg6*Δ, overexpressed *HMG2*) [35]. A red color indicates that the gene's expression was induced under those conditions, while a green color indicates that the gene was repressed under those conditions; black indicates no detectable change in expression, and gray indicates missing data. Gene names in purple correspond to genes with confirmed roles in (a) sulfur utilization or (b) ergosterol biosynthesis; gene names in orange show highly correlated expression patterns, despite their lack of annotation as sulfur-utilization genes. Arrows above the columns indicate conditions in which the displayed gene groups show significant gene-expression changes according to the K-S test, FDR correction for multiple testing at a *p*-value of 0.001.

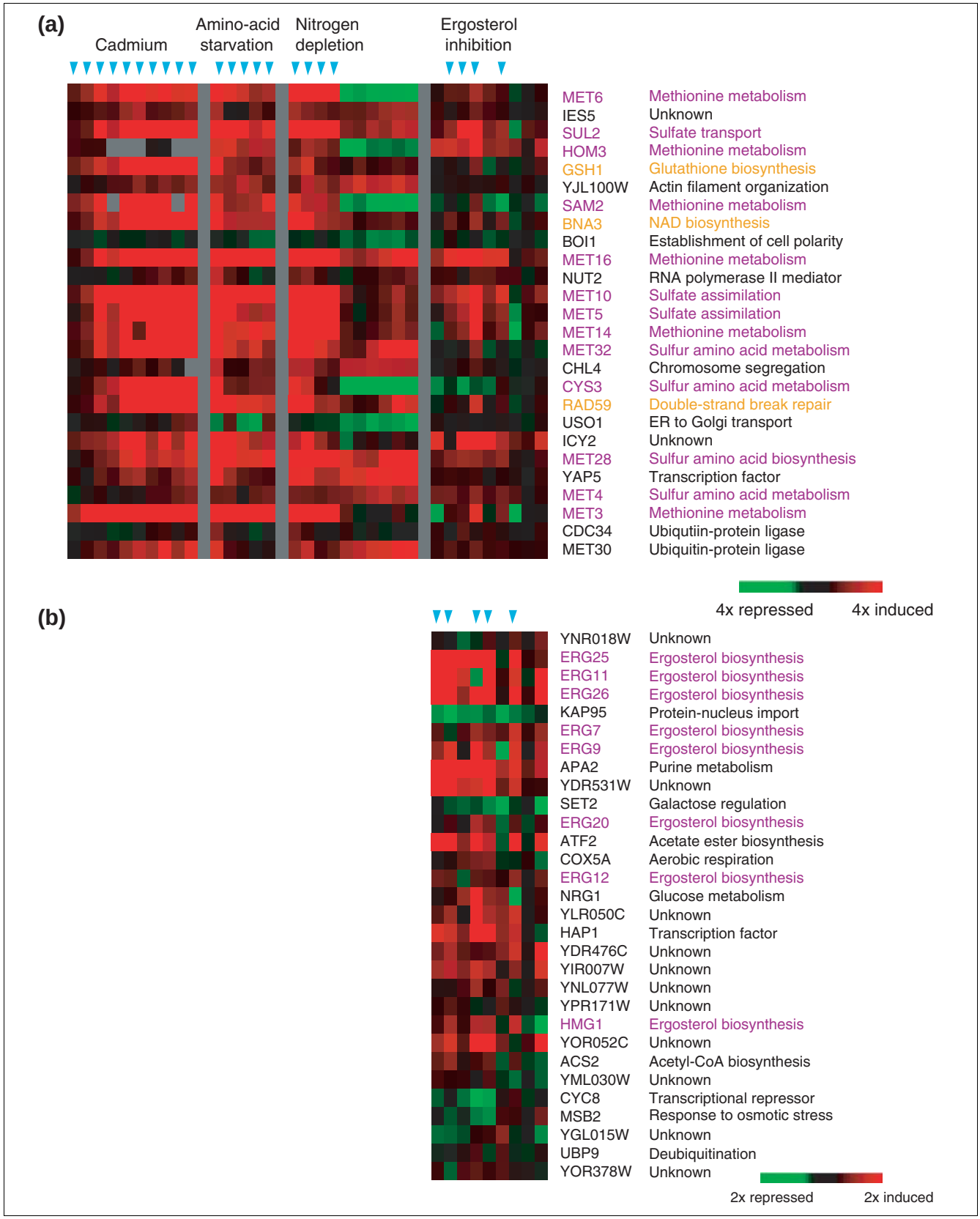


Figure 5 *(see legend on previous page)*

expression changes) and highly sensitive (most of the genes that show gene-expression changes should also share the feature).

To assess the specificity of a sequence feature in explaining gene expression, we computed the average expression profile for all genes that shared that feature. We expect that if a sequence feature represented a transcription factor binding site, genes containing that feature in their TCRs would be induced (or repressed), on average, compared to all the genes in the genome [27]. By comparing the average expression profile derived from conserved word-pair templates with average expression profiles derived from simpler sequence features, we assessed how much information was obtained by incorporating both the conservation and pairwise combination of sequences. For representative word pairs associated with significant gene-expression changes in environmental stress conditions, we compared the average expression profile for: gene subsets that shared single words in *S. cerevisiae*; gene subsets that conserved single words among multiple genomes; and gene subsets that shared both words in *S. cerevisiae* (Figure 6 and see Additional data files). In general, the average gene-expression profiles for conserved word pairs were more significantly different from the average expression of genes in the genome when either conservation or word pairs was used as an additional criterion for gene selection. In Figure 6, the last two rows for each word pair indicate the average expression profiles for genes that shared both words in *S. cerevisiae*, as well as the average expression profile for genes that conserved both words in multiple genomes, respectively. Strikingly, the consideration of word-pair conservation yielded further increases in average gene-expression profiles compared to word pairs in *S. cerevisiae* alone. Similar effects were observed for the PAC-RRPE and the overlapping RRPE pairs associated with genes repressed in the environmental stress response (see Additional data files). Thus, conserved word-pair templates contained more specific predictors of gene expression than comparable sequence templates derived from *S. cerevisiae* alone.

To evaluate how well sequence features can explain gene expression changes across the whole genome, several groups have constructed linear regression models using various choices for features [18–21]. The *R*-square statistic of a regression model indicates the percent of global variance that can be explained using the sequence features in the model. Models with better fits to the genome-wide expression data would thus have greater *R*-square values. To assess the sensitivity of individual word pairs in explaining global gene expression, we first constructed regression models using individual word pairs (see Materials and methods). We chose three representative environmental conditions: amino-acid starvation (30 min) [30]; stationary phase (10 h in YPD) [30]; and ergosterol inhibition (terbinafine) [35]. We constructed regression models using counts of individual words in *S. cerevisiae* TCRs, or using counts of words that were conserved among *Saccharomyces* TCRs. Sequence conservation improved the fit of regression models based on individual word pairs (ΔR^2 = 0.3-1.3%) (see Additional data files). These results clearly show that sequences conserved in multiple *Saccharomyces* species were more likely to be associated with gene-expression changes.

To assess the joint contribution of word pairs on gene expression, we also included interaction terms between the individual words only if their coefficients were statistically significant (see Materials and methods). Pairwise interaction terms, expressed as the product of scores for two features, represent a standard way to assess whether two features contribute nonadditively to gene expression [19]. Indeed, the inclusion of significant pairwise interaction terms improved the fits for both the *S. cerevisiae* sequence model and the conserved sequence model, increasing the *R*-square by a further 0.2% to 0.9% (see Additional data files). Whereas the interaction terms only comprise a small proportion of the global variance, they can be interpreted as statistical evidence of dependence between sequence features [19]. Therefore, the nonadditive contributions of conserved word-pair templates further suggest their involvement in multifactorial regulation.

We expanded these models to include multiple conserved word-pair templates using a stepwise linear regression procedure. The set of potential sequence features was expanded to

Figure 6 *(see following page)*
Incorporation of conservation and word pairs provided more informative average expression profiles. Groups of genes whose TCRs contained various sequence features were summarized by the average of their gene-expression profiles. Each row in these diagrams represents a given gene group's average expression pattern under the conditions shown in each column: exposure to increasing concentrations of cadmium chloride (from 0.05 mM to 0.4 mM); 20 min after heat shock to 37°C (from 17°C, 21°C, 25°C, 29°C, and 33°C); an amino-acid starvation time course; a nitrogen-source depletion time course; progression into stationary phase (2 h, 4 h, 6 h, 8 h, 10 h, 12 h, 1 day, 2 days, 3 days, 5 days of growth) [30]. Representative conserved word-pair templates were chosen for analysis, corresponding to: **(a)** Msn2/4p-like sequences (CTCCCC and CCCTTA); **(b)** Cbf1p (CACGTG) and Met31/32p (TGTGGC) binding sites; **(c)** Mig1p (ACCCCA) and Ume6 (CCGCCG) binding sites; **(d)** Msn2/4p-like (CCCCTT) and Hsf1p-like (GAGAAA) sequences. For each of the panels (a-d) each row represents the average expression profile for gene groups chosen by different sequence features in their TCRs: single words found in *S. cerevisiae* (rows 1 and 2); single words conserved in three or more *Saccharomyces* genomes (rows 3 and 4); word pairs found in *S. cerevisiae* (row 5); word pairs conserved in three or more *Saccharomyces* genomes (row 6). Arrows above the columns indicate conditions under which gene groups sharing the conserved word-pair template (row 6) were significantly associated with gene-expression changes, at a *p*-value of 0.001 (K-S test after FDR correction for multiple testing).

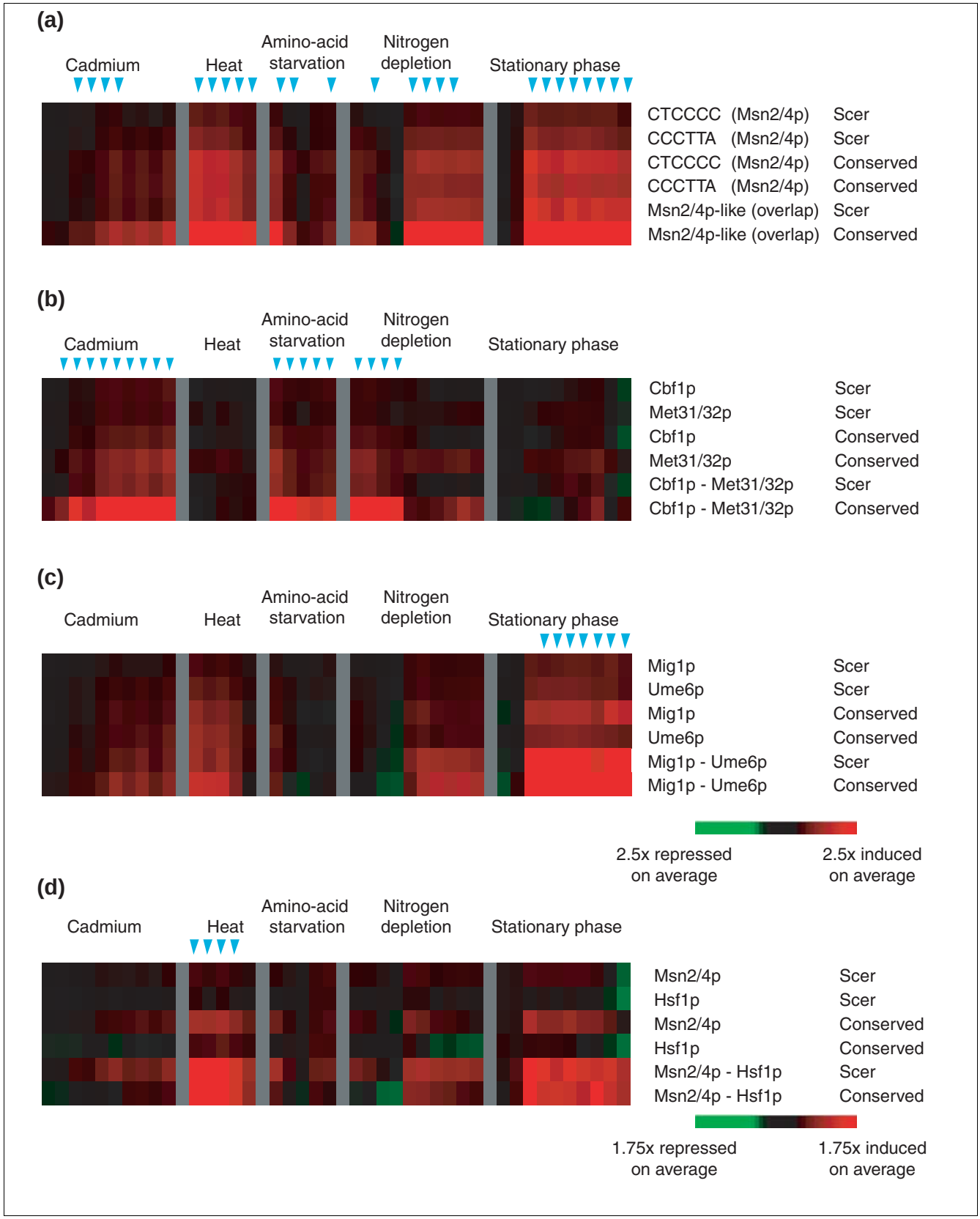


Figure 6 *(see legend on previous page)*

Table 3

Stepwise linear regression statistics

		S. cerevisiae			Three or more genomes				
	Word	M_{fc}	*p*-value	ΔR^2	M_{fc}	*p*-value	ΔR^2	*p*-value	*p*-value
Amino-acid	AAATTT	-0.165	< 2.0e-16	3.1%	-0.293	< 2.0e-16	6.6%	1.3e-37	2.0e-07
starvation	GATGAG	-	-	-	-0.333	6.7e-16	4.1%	1.6e-30	8.5e-03
0.5 h	AAGGGG	0.209	3.2e-14	1.8%	0.455	< 2.0e-16	3.5%	1.6e-22	7.9e-05
	TGTGGC	0.094	1.1e-03	0.6%	0.283	2.9e-07	1.6%	5.0e-07	8.3e-13
	CCCTTA	0.300	**2.0e-16**	**1.7%**	0.363	**< 2.0e-16**	**1.4%**	**3.2e-06**	**3.5e-03**
	TGACTC	0.229	**4.6e-10**	**0.8%**	0.311	**2.2e-11**	**1.0%**	**4.6e-01**	**1.1e-03**
	AAATTT • GATGAG	-	-	-	-0.266	**9.1e-09**	**0.5%**	-	-
	CACGTG	0.045	3.5e-01	0.5%	0.146	8.6e-03	0.5%	**1.9e-07**	**1.2e-08**
	CACGTG • TGTGGC	0.443	**3.8e-10**	**0.5%**	0.749	**1.0e-12**	**0.9%**	-	-
	GTGAAA	-0.066	**1.1e-03**	**0.3%**	-0.082	**4.6e-03**	**0.1%**	-	-
	TCTTTT	-0.022	**2.3e-02**	**0.1%**	-	-	-	-	-
	Total ΔR^2			**9.6%**			**20.2%**		
Stationary	AAATTT	-0.218	< 2.0e-16	3.2%	-0.377	< 2.0e-16	5.8%	5.5e-39	N/R
phase	AAGGGG	0.233	1.7e-11	0.9%	0.591	< 2.0e-16	4.0%	4.5e-26	N/R
YPD 10 h	**CCCTTA**	0.460	**< 2.0e-16**	**3.7%**	0.579	**< 2.0e-16**	**2.2%**	3.0e-07	N/R
	GATGAG	-	-	-	-0.242	4.1e-06	1.8%	4.4e-18	N/R
	ACCCCA	0.224	**3.0e-03**	**0.3%**	0.459	**1.5e-06**	**1.0%**	-	N/R
	AAATTT • GATGAG	-	-	-	-0.287	**1.7e-06**	**0.4%**	-	N/R
	CCGCCG	0.333	**5.1e-07**	**0.8%**	0.208	**1.5e-02**	**0.3%**	-	N/R
	ACCCCA • CCGCCG	0.294	**1.8e-02**	**0.1%**	0.807	**5.6e-05**	**0.3%**	-	N/R
	GTGAAA	-0.090	**4.2e-04**	**0.2%**	-0.122	**1.0e-03**	**0.2%**	-	N/R
	Total ΔR^2			**9.4%**			16.0%		
Terbin-	**TGACTC**	0.162	**< 2.0e-16**	**3.5%**	0.261	**< 2.0e-16**	**5.1%**	1.3e-14	N/R
afine 3 h	TCGTTT	0.071	< 2.0e-16	2.0%	0.132	< 2.0e-16	3.3%	2.5e-24	N/R
	TGAAAC	-0.055	**1.3e-12**	**1.1%**	-0.077	**9.50e-11**	**0.9%**	4.0e-03	N/R
	GATGAG	-0.029	**1.7e-03**	**0.3%**	-0.047	**6.70e-06**	**0.4%**	-	N/R
	AAGGGG	0.025	**1.1e-02**	**0.1%**	0.050	**5.40e-04**	**0.3%**	2.4e-01	N/R
	CCGATA	-0.008	**6.5e-01**	**0.1%**	0.004	**8.6e-01**	**0.1%**	-	N/R
	CCGATA • TCGTTT	0.080	**9.4e-06**	**0.3%**	0.146	**3.2e-07**	**0.5%**	-	N/R
	CCCTTA	-0.021	**5.0e-02**	**0.1%**	-0.038	**1.1e-02**	**0.1%**	-	N/R
	Total ΔR^2			**7.6%**			10.8%		

Words and pairwise interaction terms are reported in the order of selection by the stepwise linear regression procedure performed on conserved words. The influence terms (M_f), associated *p*-values, and increase in *R*-square values were computed using the statistical package R [51]. Wang *et al.* [20] and Conlon *et al.* [21] previously fit regression models using sequence features derived from *S. cerevisiae*. The *p*-values of the most similar sequences features in their regression models were reported where available; sequence features that were more significant in this analysis are indicated in bold. Dashes indicate sequence features that were insignificant in the Wang *et al.* [20] or Conlon *et al.* [21] analyses. 'N/R' indicates gene-expression data that were not analyzed by Conlon *et al.* [21]

include all words found in templates associated with significant gene expression changes in that condition, as assessed previously by the K-S test (see Materials and methods). The final *R*-square values for regression models based on occurrences of multiple words in *S. cerevisiae* ranged from 7.2% to 9.3% (Table 3 and see Additional data files). Once again, the use of conserved instances of individual words yielded better model fits, with improvements in *R*-square values from 3.1% to 9.5%. Further improvements in the model fit (ΔR^2 = 0.5% to 1.4%) were obtained using pairwise interaction terms between individual words found in the same word-pair template. The total *R*-square values for the regression models

based on conserved word-pair templates with interaction terms thus ranged from 10.8% to 20.2% (Table 3). Thus, sequence features based on conserved word-pair templates could explain more of the global gene expression changes than features based on *S. cerevisiae* alone.

Discussion

This work describes two principles for analyzing combinations of regulatory sequences. First, sequence conservation among closely related yeast species was used to find sequences that were more likely to be functionally important. Second, a template approach that considered joint positional distributions of word pairs increased the specificity of gene-expression predictions using sequence-based rules. We have demonstrated that higher-order sequence features within TCRs were conserved across multiple *Saccharomyces* genomes. Closely spaced and jointly conserved word pairs were also more likely to be associated with specific gene-expression changes. A large proportion of words contained in templates matched known transcription factor binding sites. In many cases, associations between templates and gene-expression changes were significant in conditions when the corresponding transcription factors are known to be active. In addition, groups of genes that co-conserved both words in a template often were enriched for common functional roles. These results suggest that conserved word-pair templates, which were discovered strictly on the basis of higher-order properties of sequence conservation, also carry biological relevance.

Conserved word-pair templates may be consistent with several underlying biochemical mechanisms. One possible interpretation of templates is that closely spaced sequence pairs may promote direct or indirect interactions between transcription factors by increasing the local concentrations of the individual factors. For example, the proximity of Cbf1p and Met31/32p binding sites may promote interaction between these factors in recruiting their common transcriptional activators, Met4p and Met28p. Experimental studies on the TCRs of *MET3* and *MET28* have demonstrated that the binding of Cbf1p enhances the DNA-binding affinity of Met31/32p [5]. Indeed, biochemical experiments suggest that all these proteins may interact at the TCRs of some sulfur-utilization genes [5].

Another possible regulatory scheme for conserved, closely spaced word pairs is that individual sequences found in templates may correspond to binding sites for transcription factors that bind independently under the same or separate conditions. The Msn2/4p and Hsf1p transcription factors, whose binding sites were similar to words identified in a template, represent an example of multifactorial regulation in response to distinct environmental stimuli [6]. Spacing constraints between their binding sites could nevertheless be important under conditions when both factors are active. Recent experiments have suggested that transcription factors that do not physically interact may still co-activate gene expression as long as their binding sites are spaced within a nucleosome length (approximately 150 bp), due to collaborative competition of the bound transcription factors with core histones [8].

Close spacing between word pairs may be important for reasons other than the promotion of transcription factor interactions. Different regions of TCRs at varying windows away from translation start sites may be more competent at recruiting or inhibiting RNA polymerase. These differences may be influenced by nucleosome accessibility, chromatin structure or DNA physical properties, which can be correlated with local A+T content (see [45] for references). Notably, we have also found that the relative proportions of A and T nucleotides vary considerably within the 200 bp closest to translation start sites (A.M. Moses, M.B.Eisen and Audrey Gasch, unpublished results). Low-complexity words that contained four or more As or Ts could be found in many templates (denoted by T_nC in Figure 4 and Table 1); these words may serve as surrogates for a distance window from the translation start. Binding sites that are close to these low-complexity words may be found in more transcriptionally competent regions of TCRs. Alternatively, the possibility that each word in an identified pair may be found at similar distances from a third conserved sequence element in all TCRs cannot be discounted.

Direct biological models of binding-site organization in TCRs, as exemplified by conserved word-pair templates, provided several advantages over naive statistical models based on sequence combinations in *S. cerevisiae*. Average gene-expression profiles showed that conserved word pairs were more specific predictors of gene expression (giving fewer false positives) than single or pairwise sequences derived from *S. cerevisiae*, indicating that conserved regions among these closely related *Saccharomyces* species were enriched for functional sequences. The consideration of distance constraints between pairs of conserved sequences found many more examples than a previous study of binding-site clustering for multiple transcription factors in *S. cerevisiae* [22]. In addition, we discovered new sequences and pairwise interaction terms using regression models similar to those reported in [20] and [21] (Table 3). Conserved word-pair templates accounted for similar changes in genome-wide expression (*R*-square from approximately 11% to approximately 20%) using only 8 to 10 features, compared with dozens of overlapping features used by other methods [20,21]. Therefore, individual features from our methods were more predictive of genome-wide expression changes.

A key limitation of our approach is the use of hexamers, which may fail to capture known binding sites. For example, the binding sites for Mcm1p and Rap1p are poorly modeled by exact words, in that these transcription factors bind sequences with relaxed specificity at certain positions [11].

Our method missed examples of multifactorial regulation involving Mcm1p or Rap1p that were suggested by previous work using position-weight matrices [17]. In addition, our method required sequence identity for a word to be labeled as conserved. This strict requirement omitted binding sites that may have retained their function, despite mutations in degenerate positions that may have little impact on transcription factor binding. This tradeoff between enumerating all possible words and capturing degenerate positions in binding sites was compounded by the very large number of pairwise word combinations that were enumerated. Further work should incorporate more complicated sequence models, as well as optimization methods that restrict the search space of sequence combinations.

The consideration of joint conservation and close spacing has provided insights into how TCR organization may influence the multifactorial regulation of gene expression in *Saccharomyces cerevisiae*. These criteria were motivated by experimental studies on the positional organization of individual binding sites within TCRs, with the hypothesis that this underlying architecture would be functionally conserved. Even more complicated higher-order sequence rules are apparent in the organization of *cis*-regulatory modules in *Drosophila melanogaster* [46]. Nevertheless, a common organizational theme of the TCRs in both of these organisms is the importance of relative spacing between transcription factor binding sites. The discovery of additional principles for TCR organization will further advance our understanding of how regulatory information is encoded in genome sequences.

Materials and methods

Datasets

Whole-genome shotgun sequencing of *Saccharomyces bayanus*, *S. mikatae*, and *S. paradoxus* has been previously described [26]. All are highly related to *S. cerevisiae*, as they are grouped within the *sensu stricto* branch of the *Saccharomyces* genus [28]. Intergenic regions were aligned using CLUSTALW as described [26] and are available from the Yeast Comparative Genomics website [42]. A total of 4,101 CLUSTALW alignments were analyzed. These alignments were filtered for orthologs in at least three genomes.

Gene-expression measurements were obtained from the Stanford Microarray Database [47] and Rosetta [35]. The main experimental types among the 342 conditions examined include cell cycle [31,32], environmental stress response [30], DNA damage [33,34], cadmium (N. Ogawa and P.O. Brown, unpublished data), and inhibition of ergosterol biosynthesis [35]. This data has been log-transformed (base 2), and each experimental condition has been median normalized.

Dependent conservation of word pairs

To assess whether two words were co-conserved in the same intergenic regions, a chi-square test of independence was systematically conducted for all possible words of length 6. We defined a word to include a 6-bp sequence and its reverse complement. Each transcriptional control region (TCR) for a gene was defined as the 600 bp upstream of its translation start site. TCRs shared between divergently transcribed genes less than 600 bp long were only counted once. A word was labeled conserved in a TCR if all six bases were identical among at least three of the four genomes in the CLUSTALW alignment. For each word pair (*W*, *V*) whose overlap was less than 4, a contingency table $\boldsymbol{C}_{wv}$ was constructed. In this table, $\boldsymbol{C}_{wv}$ = number of TCR($\boldsymbol{I}_{w} \cap \boldsymbol{I}_{v}$), where $\boldsymbol{I}_{w}$, $\boldsymbol{I}_{v}$ are indicator variables for the presence of each conserved word in a TCR, summed over all TCRs. TCRs shared between divergently transcribed genes less than 600 bp long were only counted once. The expected counts $\boldsymbol{E}_{wv}$ were obtained from an independence assumption; that is, the product of the individual word conservation probabilities, multiplied by the total number of TCRs. The chi-square statistic with Yates continuity correction was computed according to the definition:

$$\chi^2_{wv} = \sum_{I_w=0}^{1} \sum_{I_v=0}^{1} \frac{\left(\left| C_{wv} - E_{wv} \right| - \frac{1}{2} \right)^2}{E_{wv}}$$

Spatial proximity of constrained word pairs

The second requirement for a conserved sequence template involved constraints on spatial arrangements between individual words. Any method that evaluates spacing distributions between word pairs must take into account positional biases that may be present for individual words (A.M. Moses, unpublished results). We used a permutation test to evaluate the significance of the median of minimum distances, excluding overlaps, between conserved word pairs. By permuting the TCR labels for one of the words, but not the word positions themselves, we retained the positional biases of individual words within intergenic regions. Within any given TCR $\boldsymbol{t}$, define $\boldsymbol{p}_t(\boldsymbol{W}) = \{\boldsymbol{p}_t^1(\boldsymbol{W}), \ldots, \boldsymbol{p}_t^j(\boldsymbol{W})\}$ as a vector of positions in *S. cerevisiae* where the j occurrences of word $\boldsymbol{W}$ are conserved. Suppose that words $\boldsymbol{W}$ and $\boldsymbol{V}$ were jointly conserved in TCRs $\boldsymbol{T_1 \ldots T_N}$. For each TCR $\boldsymbol{t} \in \{\boldsymbol{T_1 \ldots T_N}\}$, the minimum distance between words $\boldsymbol{W}$ and $\boldsymbol{V}$ was computed as

$$\boldsymbol{m}_t = \min_{j,k} \left| \boldsymbol{p}_t^j(W) - \boldsymbol{p}_t^k(V) \right|.$$

The median of minimum distances, $\overline{D}$, was simply the median of the ordered distribution $\{\boldsymbol{m_1, \ldots, m_t}\}$.

We used a permutation test to generate an empirical null distribution of $\overline{D}$ for all word pairs with $N \geq 10$. After randomly permuting the labels $\boldsymbol{t}$ for the position vectors of word *V*, a permutation test statistic, $\overline{D}^*$, can be calculated as above. By repeating this resampling procedure R times, an empirical null distribution $\overline{D}_{null} = \{\overline{D}^{*1}, \ldots, \overline{D}^{*R}\}$ can be obtained. The significance of the observed median of minimum distances, $\overline{D}$, in

the N promoters was calculated as its quantile in the empirical null distribution $\bar{D}_{null}$. We set an upper bound of $R = 10^6$, but stopped permutations early if 20 or more values in $\bar{D}_{null}$ were found less than $\bar{D}$.

Correction for multiple testing involved control of the proportion of false positives using an FDR method [48]. This method has increased power over Bonferroni-type methods. Permutation quantiles for all N word pairs tested were sorted in non-decreasing order: $\boldsymbol{q_1} \leq \ldots \leq \boldsymbol{q_N}$. Let

$$\boldsymbol{k} = \max\left(\boldsymbol{i} : q_i < \frac{0.05\boldsymbol{i}}{N} \right)$$

Then the first $\boldsymbol{k}$ word pairs in the ordering had a corrected significance level of $q < 0.05$; that is, the rate of false positives is approximately 5%.

Association between template-specified gene groups and gene-expression changes

For each gene-expression condition $\boldsymbol{c}$ in our dataset, $\boldsymbol{c} \in \{1, \ldots, 342\}$, we tested the null hypothesis that a gene subset $\boldsymbol{G}_{wv} \subseteq \boldsymbol{G}$ selected by a conserved word pair (w, v) had the same distribution of gene-expression ratios ($\boldsymbol{E}_{wv}{}^c$) as the entire genome ($\boldsymbol{E}^c$). The alternate hypothesis stated that the two gene-expression distributions were significantly different. Any gene was an element of $\boldsymbol{G}_o$ if its corresponding TCR conserved both sequences in the word pair. Since the size N_o of gene subsets may be small and the distributions may not be normally distributed, we used the nonparametric K-S test. The test statistic $\boldsymbol{K}$ compares the cumulative distribution functions $\boldsymbol{F}_{wv}{}^c$ and $\boldsymbol{F}^c$ corresponding to $\boldsymbol{E}_{wv}{}^c$ and $\boldsymbol{E}^c$ by the formula

$$K = \max_x \left| F_{wv}{}^c(x) - F^c(x) \right|$$

The significance level of an observed value $\boldsymbol{K}^*$ can be obtained using a numerical approximation [49].

A gene subset determined by a word pair was deemed to have significantly different expression if its K-S p-value was less than a certain threshold. To correct for multiple testing, this threshold was established by controlling the FDR. The significance levels $\boldsymbol{p_i}$ from each K-S test were ordered in ascending order. Let $\boldsymbol{N}$ represent the total number of K-S tests performed; that is, the number of jointly conserved, closely spaced word pairs times the number of gene-expression experiments). If $\boldsymbol{k}$ was the largest $\boldsymbol{i}$ such that

$$p_i < \frac{i\alpha}{N}$$

then the first $\boldsymbol{k}$ word pairs in the ordering were deemed to have a significance level of $p < \alpha$.

We ensured that the K-S p-value for the conserved word-pair subset $\boldsymbol{G}_o$ was more significant than subsets $\boldsymbol{G}_w$ or $\boldsymbol{G}_v$ comprised of only one conserved word by computing $\boldsymbol{K}$ for $\boldsymbol{E}_w{}^c$ vs $\boldsymbol{E}_v{}^c$, as well as for $\boldsymbol{E}_w{}^c$ vs $\boldsymbol{E}^c$. The marginal improvement of the joint word pair was defined as: $\boldsymbol{K}(\boldsymbol{F}_o{}^c$ vs $\boldsymbol{F}^c)$ - max($\boldsymbol{K}(\boldsymbol{F}_w{}^c$ vs $\boldsymbol{F}^c)$, $\boldsymbol{K}(\boldsymbol{F}_v{}^c$ vs $\boldsymbol{F}^c)$).

Hierarchical clustering of word pair associations

The $P \times C$ matrix of K-S p-values was log-transformed (base 10), and the word pairs contained in P were clustered by average-linkage hierarchical clustering using the program Cluster [50]. As the log-transformed K-S p-values were all negative, a centered Pearson correlation was used as the similarity metric.

Stepwise linear regression of gene expression

Regression analyses assume that a log-transformed gene-expression measurement, $\boldsymbol{E}_{gc}$ for gene $\boldsymbol{g}$ in condition $\boldsymbol{c}$ can be modeled by a linear equation:

$$E_{gc} = \sum_f M_{fc} \times S_{gf} + \varepsilon_g$$

where $\boldsymbol{S_{gf}}$ represents the score of a sequence feature $\boldsymbol{f}$ in gene $\boldsymbol{g}$, $\boldsymbol{M_{fc}}$ represents the influence term of the feature $\boldsymbol{f}$ on gene expression in condition $\boldsymbol{c}$, and ε_g is the gene-specific error term. Genome-wide expression data was filtered for a set of 4,703 genes with TCRs conserved in three or more *Saccharomyces* genomes. For a certain experimental condition, the list of features was restricted to either two words found in a single word-pair template, or to all words found in conserved word-pair templates that were significantly associated with gene-expression changes in that condition. The score $\boldsymbol{S_{gf}}$ for feature $\boldsymbol{f}$ in a TCR corresponding to gene $\boldsymbol{g}$ was taken as either the number of occurrences in *S. cerevisiae*, or the number of occurrences conserved in three or more *Saccharomyces* genomes. Stepwise linear regression models were fit to genome-wide expression data using the statistical package R [51]. At each iteration, the sequence feature with the largest increase in the R-square goodness-of-fit score was added to the model, where

$$R^2 = \sum_f \left(M_{fc} \times S_{gf} \right)^2$$

Pairwise interaction terms between sequence features $\boldsymbol{f1}$ and $\boldsymbol{f2}$ already selected in the model, expressed as $\boldsymbol{S_{gf1}} \bullet \boldsymbol{S_{gf2}}$, could also be added to the model at each iteration if the features were found in the same conserved word-pair template. Sequence features were added to the regression model as long as the p-values for their associated influence terms ($\boldsymbol{M_{fc}}$) were less than 0.05.

Additional data files

Additional files are available with the online version of this paper. They comprise a tab-delimited text list of 989 identified conserved word-pair templates (Additional data file 1), a figure that shows that incorporation of conservation and word pairs provided more informative average expression profiles (Additional data file 2), a figure that shows that regression models using conserved word pairs represented better fits to genome-wide expression data (Additional data file 3), and the underlying data for Figure 4 that show associations between gene groups that share a conserved word-pair template with gene-expression changes (available as a zip.file, Additional data file 4; all files in TreeView, text tab-delimited, format). When visualized in TreeView (available from [52]), these files correspond to the data underlying Figure 4a and Figure 4b. See Figure 4 legend for details.

Acknowledgements

We are indebted to Audrey Gasch for her insightful advice and suggestions. We thank Justin Fay for a critical reading of the manuscript and help with the permutation test; Peter Bickel and John Storey for their statistical advice; and Nabuo Ogawa and Patrick Brown for sharing gene expression data on cadmium conditions before publication. D.Y.C. is a Howard Hughes Medical Institute Predoctoral Fellow, and M.B.E. is a Pew Scholar in the Biomedical Sciences. This work was carried out under the US Department of Energy contract ED-AC03-76SF00098.

References

1. **Composite regulatory elements: structure, function and classification** [http://www.gene-regulation.com/pub/databases/transcompel/compel.html]
2. Wolberger C: **Multiprotein-DNA complexes in transcriptional regulation.** *Annu Rev Biophys Biomol Struct* 1999, **28:**29-56.
3. Mead J, Bruning AR, Gill MK, Steiner AM, Acton TB, Vershon AK: **Interactions of the Mcm1 MADS box protein with cofactors that regulate mating in yeast.** *Mol Cell Biol* 2002, **22:**4607-4621.
4. Bhoite LT, Allen JM, Garcia E, Thomas LR, Gregory ID, Voth WP, Whelihan K, Rolfes RJ, Stillman DJ: **Mutations in the Pho2 (Bas2) transcription factor that differentially affect activation with its partner proteins Bas1, Pho4, and Swi5.** *J Biol Chem* 2002, **277:**37612-37618.
5. Blaiseau PL, Thomas D: **Multiple transcriptional activation complexes tether the yeast activator Met4 to DNA.** *EMBO J* 1998, **17:**6327-6336.
6. Gasch AP: **The environmental stress response: a common yeast response to diverse environmental stresses.** In *Yeast Stress Responses Volume 1*. Edited by: Hohmann S, Mager WH. Berlin: Springer; 200311-70.
7. Vashee S, Melcher K, Ding WV, Johnston SA, Kodadek T: **Evidence for two modes of cooperative DNA binding in vivo that do not involve direct protein-protein interactions.** *Curr Biol* 1998, **8:**452-458.
8. Miller JA, Widom J: **Collaborative competition mechanism for gene activation in vivo.** *Mol Cell Biol* 2003, **23:**1623-1632.
9. Smith DL, Johnson AD: **A molecular mechanism for combinatorial control in yeast: MCM1 protein sets the spacing and orientation of the homeodomains of an alpha 2 dimer.** *Cell* 1992, **68:**133-142.
10. Drazinic CM, Smerage JB, Lopez MC, Baker HV: **Activation mechanism of the multifunctional transcription factor repressor-activator protein 1 (Rap1p).** *Mol Cell Biol* 1996, **16:**3187-3196.
11. Stormo GD: **DNA binding sites: representation and discovery.** *Bioinformatics* 2000, **16:**16-23.
12. Duret L, Bucher P: **Searching for regulatory elements in human noncoding sequences.** *Curr Opin Struct Biol* 1997, **7:**399-406.
13. Pennacchio LA, Rubin EM: **Genomic strategies to identify mammalian regulatory sequences.** *Nat Rev Genet* 2001, **2:**100-109.
14. Blanchette M, Tompa M: **Discovery of regulatory elements by a computational method for phylogenetic footprinting.** *Genome Res* 2002, **12:**739-748.
15. Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM: **rVista for comparative sequence-based discovery of functional transcription factor binding sites.** *Genome Res* 2002, **12:**832-839.
16. Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W: **PipMaker - a web server for aligning two genomic DNA sequences.** *Genome Res* 2000, **10:**577-586.
17. Pilpel Y, Sudarsanam P, Church GM: **Identifying regulatory networks by combinatorial analysis of promoter elements.** *Nat Genet* 2001, **29:**153-159.
18. Bussemaker HJ, Li H, Siggia ED: **Regulatory element detection using correlation with expression.** *Nat Genet* 2001, **27:**167-171.
19. Keles S, van der Laan M, Eisen MB: **Identification of regulatory elements using a feature selection method.** *Bioinformatics* 2002, **18:**1167-1175.
20. Wang W, Cherry JM, Botstein D, Li H: **A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae*.** *Proc Natl Acad Sci USA* 2002, **99:**16893-16898.
21. Conlon EM, Liu XS, Lieb JD, Liu JS: **Integrating regulatory motif discovery and genome-wide expression analysis.** *Proc Natl Acad Sci USA* 2003, **100:**3339-3344.
22. Wagner A: **Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes.** *Bioinformatics* 1999, **15:**776-784.
23. Klingenhoff A, Frech K, Quandt K, Werner T: **Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity.** *Bioinformatics* 1999, **15:**180-186.
24. Pavlidis P, Furey TS, Liberto M, Haussler D, Grundy WN: **Promoter region-based classification of genes.** *Pac Symp Biocomput* 2001, **6:**151-164.
25. Kel-Margoulis OV, Ivanova TG, Wingender E, Kel AE: **Automatic annotation of genomic regulatory sequences by searching for composite clusters.** *Pac Symp Biocomput* 2002, **7:**187-198.
26. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: **Sequencing and comparison of yeast species to identify genes and regulatory elements.** *Nature* 2003, **423:**241-254.
27. Chiang DY, Brown PO, Eisen MB: **Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles.** *Bioinformatics* 2001, **17:**S49-S55.
28. Cliften PF, Hillier LW, Fulton L, Graves T, Miner T, Gish WR, Waterston RH, Johnston M: **Surveying *Saccharomyces* xs genomes to identify functional elements by comparative DNA sequence analysis.** *Genome Res* 2001, **11:**1175-1186.
29. Zhu J, Zhang MQ: **SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*.** *Bioinformatics* 1999, **15:**607-611.
30. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: **Genomic expression programs in the response of yeast cells to environmental changes.** *Mol Biol Cell* 2000, **11:**4241-4257.
31. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization.** *Mol Biol Cell* 1998, **9:**3273-3297.
32. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, *et al.*: **A genome-wide transcriptional analysis of the mitotic cell cycle.** *Mol Cell* 1998, **2:**65-73.
33. Gasch AP, Huang M, Metzner S, Botstein D, Elledge SJ, Brown PO: **Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p.** *Mol Biol Cell* 2001, **12:**2987-3003.
34. Lee SE, Pellecioli A, Demeter J, Vaze MP, Gasch AP, Malkova A, Brown PO, Botstein D, Stearns T, Foiani M, *et al.*: In *Biological Responses to DNA Damage Volume 65*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2000303-314.
35. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai HY, He YDD, *et al.*: **Functional discovery via a compendium of expression profiles.** *Cell* 2000, **102:**109-126.
36. Costanzo MC, Crawford ME, Hirschman JE, Kranz JE, Olsen P, Robertson LS, Skrzypek MS, Braun BR, Hopkins KL, Kondu P, *et al.*:

YPD™, PombePD™, and WormPD™: model organism volumes of the BioKnowledge® library, an integrated resource for protein information. *Nucleic Acids Res* 2001, **29:**75-79.
37. Causton HC, Ren B, Koh SS, Harbison CT, Kanin E, Jennings EG, Lee TI, True HL, Lander ES, Young RA: **Remodeling of yeast genome expression in response to environmental changes.** *Mol Biol Cell* 2001, **12:**323-337.
38. Hughes JD, Estep PW, Tavazoie S, Church GM: **Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*.** *J Mol Biol* 2000, **296:**1205-1214.
39. Thomas D, Surdin-Kerjan Y: **Metabolism of sulfur amino acids in *Saccharomyces cerevisiae*.** *Microbiol Mol Biol Rev* 1997, **61:**503-532.
40. Fauchon M, Lagniel G, Aude JC, Lombardia L, Soularue P, Petat C, Marguerie G, Esentenac A, Werner M, Labarre J: **Sulfur sparing in the yeast proteome in response to sulfur demand.** *Mol Cell* 2002, **9:**713-723.
41. Williams RM, Primig M, Washburn BK, Winzeler EA, Bellis M, Sarrauste de Menthiere C, Davis RW, Esposito RE: **The Ume6 regulon coordinates metabolic and meiotic gene expression in yeast.** *Proc Natl Acad Sci USA* 2002, **99:**13431-13436.
42. **Whitehead Institute Center for Genome Research - Yeast Comparative Genomics** [http://www-genome.wi.mit.edu/seq/Saccharomyces]
43. Vik A, Rine J: **Upc2p and Ecm22p, dual regulators of sterol biosynthesis in *Saccharomyces cerevisiae*.** *Mol Cell Biol* 2001, **21:**6395-6405.
44. Kwast KE, Burke PV, Poyton RO: **Oxygen sensing and the transcriptional regulation of oxygen-responsive genes in yeast.** *J Exp Biol* 1998, **201:**1177-1195.
45. Liao GC, Rehm EJ, Rubin GM: **Insertion site preferences of the P transposable element in *Drosophila melanogaster*.** *Proc Natl Acad Sci USA* 2000, **97:**3347-3351.
46. Berman BP, Nibu Y, Pfeiffer BD, Tomancek P, Celniker SE, Levine M, Rubin GM, Eisen MB: **Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome.** *Proc Natl Acad Sci USA* 2002, **99:**757-762.
47. Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA, *et al.*: **The Stanford Microarray Database.** *Nucleic Acids Res* 2001, **29:**152-155.
48. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J R Statist Soc B* 1995, **57:**289-300.
49. Press WH, Teukolsky SA, Vertterling WT, Flannery BP: *Numerical Recipes in C* 2nd edition. Cambridge: Cambridge University Press; 1992.
50. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95:**14863-14868.
51. **The R Project for Statistical Computing** [http://www.r-project.org]
52. **Eisen Lab** [http://rana.lbl.gov]

BMC Evolutionary Biology

BioMed Central

Research article

Open Access

Position specific variation in the rate of evolution in transcription factor binding sites

Alan M Moses[1], Derek Y Chiang[2], Manolis Kellis[3,5], Eric S Lander[4,5] and Michael B Eisen*[1,2,6]

Address: [1]Graduate Group in Biophysics, University of California, Berkeley, CA 94720, USA, [2]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA, [3]Department of Computer Science, Massachusetts Institute of Technology M.I.T., Cambridge, MA 02139, USA, [4]Department of Biology, M.I.T., Cambridge, MA 02139, USA, [5]Whitehead/MIT Center for Genome Research, Cambridge, MA 02139, USA and [6]Department of Genome Sciences, Life Sciences Division, Ernest Orlando Lawrence Berkeley National Lab Berkeley, CA 94720, USA

Email: Alan M Moses - amoses@lbl.gov; Derek Y Chiang - dchiang@ocf.berkeley.edu; Manolis Kellis - manoli@mit.edu; Eric S Lander - lander@genome.wi.mit.edu; Michael B Eisen* - mbeisen@lbl.gov

* Corresponding author

Published: 28 August 2003

BMC Evolutionary Biology 2003, **3**:19

Received: 08 May 2003
Accepted: 28 August 2003

This article is available from: http://www.biomedcentral.com/1471-2148/3/19

© 2003 Moses et al; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article's original URL.

Abstract

Background: The binding sites of sequence specific transcription factors are an important and relatively well-understood class of functional non-coding DNAs. Although a wide variety of experimental and computational methods have been developed to characterize transcription factor binding sites, they remain difficult to identify. Comparison of non-coding DNA from related species has shown considerable promise in identifying these functional non-coding sequences, even though relatively little is known about their evolution.

Results: Here we analyse the genome sequences of the budding yeasts *Saccharomyces cerevisiae, S. bayanus, S. paradoxus* and *S. mikatae* to study the evolution of transcription factor binding sites. As expected, we find that both experimentally characterized and computationally predicted binding sites evolve slower than surrounding sequence, consistent with the hypothesis that they are under purifying selection. We also observe position-specific variation in the rate of evolution within binding sites. We find that the position-specific rate of evolution is positively correlated with degeneracy among binding sites within *S. cerevisiae*. We test theoretical predictions for the rate of evolution at positions where the base frequencies deviate from background due to purifying selection and find reasonable agreement with the observed rates of evolution. Finally, we show how the evolutionary characteristics of real binding motifs can be used to distinguish them from artefacts of computational motif finding algorithms.

Conclusion: As has been observed for protein sequences, the rate of evolution in transcription factor binding sites varies with position, suggesting that some regions are under stronger functional constraint than others. This variation likely reflects the varying importance of different positions in the formation of the protein-DNA complex. The characterization of the pattern of evolution in known binding sites will likely contribute to the effective use of comparative sequence data in the identification of transcription factor binding sites and is an important step toward understanding the evolution of functional non-coding DNA.

Background

Although non-coding DNA makes up the majority of most eukaryotic genomes, relatively little is known about its function or the nature of the constraints on its evolution. Here we focus on the evolution of an important and relatively well-understood class of functional non-coding sequences, the binding sites of sequence-specific transcription factors.

Transcription factors recognize degenerate families of short sequences (5–25 base pairs). The binding specifities of transcription factors are typically represented as consensus sequences or position weight matrices [1] that summarize their position-specific sequence preferences. In some cases, such 'motif' models of transcription factor binding sites can be inferred from genome sequences using computational methods [2–7].

Despite the absence of a detailed understanding of the evolution of transcription factor binding sites, the comparison of sequences from related species has been used to identify transcription factor binding sites *en masse*, with the guiding hypothesis that functional regulatory sequences will be more conserved than the surrounding DNA. Several methods [8–12] have been developed to identify conserved non-coding sequences that, when tested, often function as regulatory sequences *in vivo* (reviewed in [13]).

Here we characterize the evolution of known transcription factor binding sites using the complete genome sequences of the closely related budding yeasts *Saccharomyces mikatae*, *S. bayanus*, *S. paradoxus* [12] and *S. cerevisiae* [14]. We limit our focus to the conservation of binding sites due to purifying selection [15–18], though binding site turnover [15,16] (the loss and reappearance of binding sites) and other processes also occur. Preferential conservation of transcription factor binding sites has been observed previously in the genomes of organisms from bacteria [10,11] to mammals [16–18], and we expect the same to be true of yeast. In addition to the availability of complete genome sequences, the budding yeasts are a particularly appealing system in which to test these hypotheses because of the relative wealth and easy accessibility of biochemical and genetic information [e.g., [20]].

Characterizing the pattern of evolution within transcription factor binding sites allows us to explore the nature of functional constraints on these sequences. As is well known for protein sequences [21–23], we expect the pattern of evolution in transcription factor binding sites to reflect the particular patterns of constraint under which they function; important regions or residues should be constrained, while unimportant positions may show fixed changes. Unlike protein sequences, where the relationship of the amino acid sequence to the functional constraint is often difficult to discern, in the case of transcription factor binding sites, we suggest that the evolutionary constraints can be interpreted directly with respect to the physical constraints imposed by the DNA-binding protein.

Protein-DNA interactions are of much interest (e.g., [24–27]) and an understanding of the evolution of the binding motifs may provide insight into these interactions. In particular, it has recently been shown that there is a relationship between the pattern of degeneracy in certain binding motifs and regions of contact between the DNA and the binding protein: positions with fewer points of contact in the structures of protein-DNA complexes show greater variability among binding sites within a single genome [28]. If these degenerate positions are less important for the formation of the protein-DNA complex, they might be expected to show less constrained evolution, as changes at these positions have a smaller effect on the relative fitness of the organism, and therefore may become fixed in the population by drift with greater probability. Conversely, changes at positions in the motif that disrupt the recognition of the binding site by the binding-protein are likely to be deleterious, and therefore removed from the population by purifying selection. This intuition leads to a theoretical prediction that the rate of evolution at each position is a function of the frequencies in the position weight matrix (analogous to the predictions for protein sequences found in [29]).

Results

Characterized binding sites show fewer substitutions than background DNA

We first sought to verify that functional non-coding regions evolve more slowly than 'background sequences.' To do so, we selected several transcription factors for which there were multiple experimentally validated binding sites in the *S. cerevisiae* genome listed in the Promoter database of *Saccharomyces cerevisiae* (SCPD[20]), and compared the rate of evolution within these binding sites to that of the promoter regions in which they were found. We measured the rate of evolution in substitutions (i.e., inferred nucleotide changes) per site, where 'site' refers to a single nucleotide position, not the multi-basepair 'binding sites' of transcription factors. We first looked at Gal4p, a very well studied Zn[2]Cys[6] binuclear cluster domain transcriptional activator [30]. The average rate of evolution within known Gal4p binding sites is 0.32 (+0.12, -0.09, n = 119) substitutions per site, substantially slower than the 0.75 (± 0.03, n = 2760) substitutions per site observed in the promoters in which these Gal4p binding sites are found (fig. 1A, 1B compare Gal4 'motif' and 'background.')

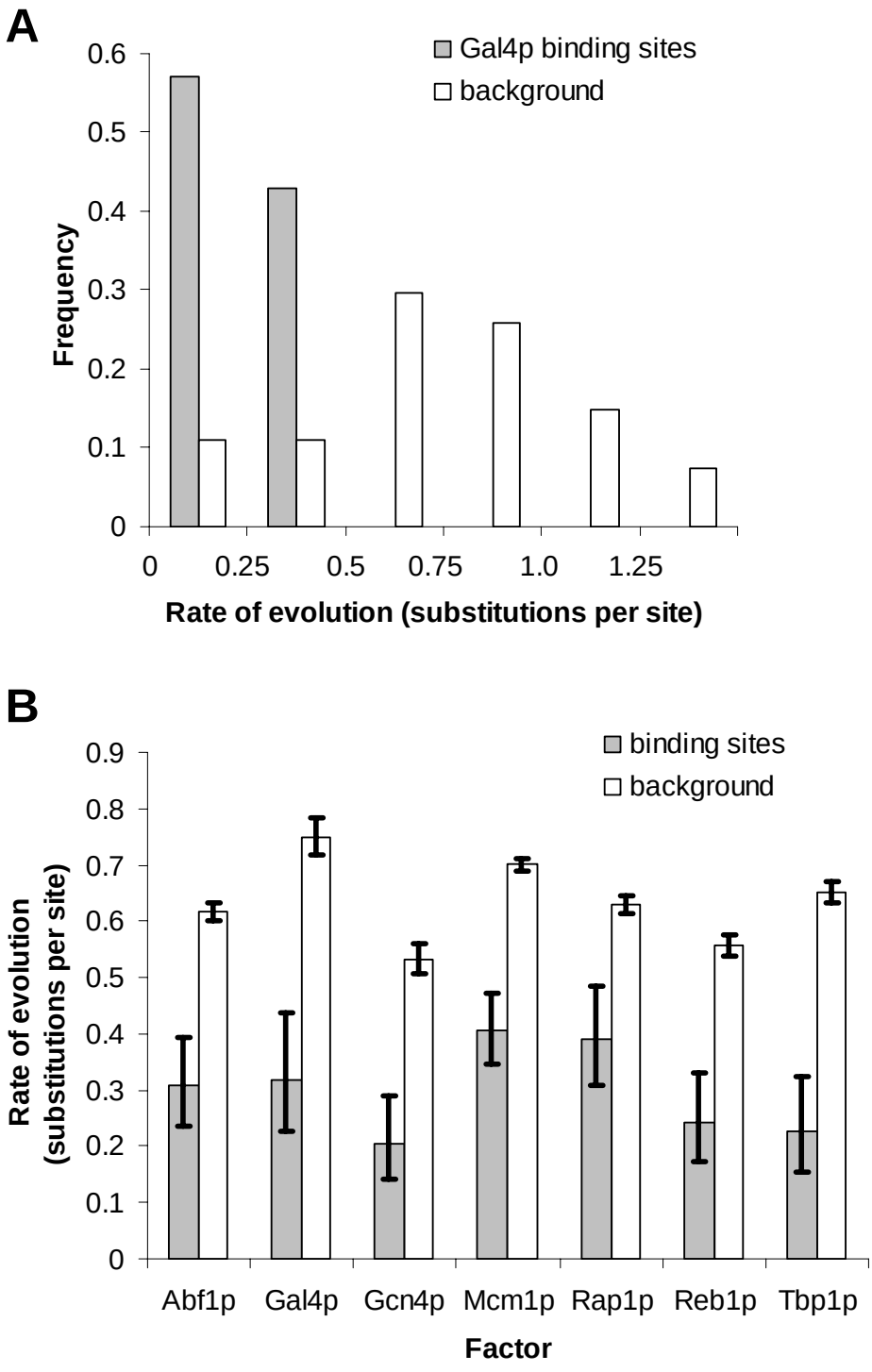


Figure 1

Characterized binding sites evolve more slowly than the promoters in which they are found. A. Histogram of the rate of evolution (estimated by maximum parsimony) in characterized Gal4p binding sites and randomly chosen sequences of the same length (17 basepairs) from the same promoters. B. Differences in the mean rate of evolution in motifs and the mean rate in the promoters in which they are found. Grey boxes represent the average in binding sites; unfilled boxes represent the average over the promoters in which the motifs are found (see methods). Error bars represent exact 95 % confidence intervals for a Poisson distribution.

Table 1: Correlation between information content and substitutions per site for the experimentally characterized binding sites in the SCPD database.

Factor	Type of DNA-binding domain (YPD)	Number of binding sites (SCPD)	Width of motif	Spearman's rank correlation	p-value
Gcn4p	BZIP	15	12	-0.84	<0.001 **
Gal4p	Zn[2]Cys[6] zinc finger	10	17	-0.83	<0.001 **
Abf1p	Atypical CHC2-type zinc finger	16	12	-0.70	0.005 *
Mcm1p	MADS box	35	14	-0.70	0.002 *
Rap1p	Myb-like	17	15	-0.72	<0.001 **
Reb1p	Myb-like	18	10	-0.81	0.002 *
Tbp1p	TATA-binding	15	9	-0.46	0.106

P-values refer to the significance of the Spearman Rank correlation coefficient. * Indicates significance at a per-factor error rate < 0.05. ** Indicates significance after Bonferoni correction to keep the global error rate < 0.05, assuming 50 tests were done in total.

To test the generality of this observation, we chose six other transcription factors representing different types of DNA-binding domains (see table 1) with relatively many characterized binding sites in the SCPD database. In each case there are significantly fewer substitutions ($p < 0.05$, 1000 bootstraps) in the characterized binding sites than in the promoters in which they lie (figure 1B), suggesting that, in general, characterized transcription factor binding sites evolve more slowly than the surrounding intergenic sequences. This is consistent with the hypothesis that these sequences are under functional constraint and their evolution reflects purifying selection.

Functionally important positions are preferentially conserved

In order to further explore the functional constraints on transcription factor binding site evolution, we computed the rate of evolution at each position within the motif and observed that the rate of evolution is not constant over the binding sites. Some positions in the motif show fewer substitutions than background, while others do not. For example, in the Gal4p binding sites positions 1, 2, 3, 15, 16, and 17 show fewer substitutions than do positions 4–14 (fig. 2, right panel).

Functionally important positions are expected to be under stronger purifying selection and therefore show stronger conservation. Indeed, the conserved positions in the Gal4p binding sites correspond to the points of contact in the crystal structure of the protein-DNA complex (fig. 2, right panel) that are required for the recognition of the target sequence [30].

Another particularly interesting example is the case of Mcm1p. Although there is no specific base in the consensus at positions 8, 9 and 10, there is a strong A/T bias in the matrix at these positions and mutagenesis studies [31] of the binding site have suggested that this is needed to allow the high degree of bending known to be necessary for the formation of Mcm1p-DNA complex [32–34]. The relative paucity of substitutions at positions 8, 9 and 10 (0.37, 0.22 and 0.5 respectively, compared to 0.70 over the entire promoters) further supports the notion that the constraint on functionally important positions slows their evolution.

Positional variation within one genome is correlated to variation between genomes

Noting that positions with fewer substitutions seem to coincide with the positions that are non-degenerate in the consensus, we constructed position weight matrices using the characterized binding sites from *S. cerevisiae* and, in order to quantify the degeneracy, computed the information content at each position. The information content of a position within a binding site has been shown to correlate with the importance of that position in the formation of the protein-DNA complex [28]. For the transcription factors used above (fig. 1B), we observe that positions of high information correspond to positions with fewer substitutions (e.g., Fig. 3). In 6 of 7 cases we found this correlation (Spearman's rank of -0.70 to -0.84) statistically significant ($p < 0.01$), the lone exception being Tbp1p, where a negative correlation was observed (-0.46), but was not significant ($p = 0.11$). (Table 1 & see discussion.) Thus the sequence variation in characterized transcription factor binding sites within one genome is directly related to the sequence variation at individual sites between genomes.

Site-specific substitution rates are consistent with the proportionality of Halpern and Bruno

If the nucleotide frequencies at each position of a position weight matrix accurately reflect the allowed sequence specificity for the formation of a functional protein-DNA complex, it is possible, under several assumptions, to predict the rates of evolution based on these frequencies, as has been done using the frequencies of residues in protein sequences [29]. The underlying intuition is that if, for

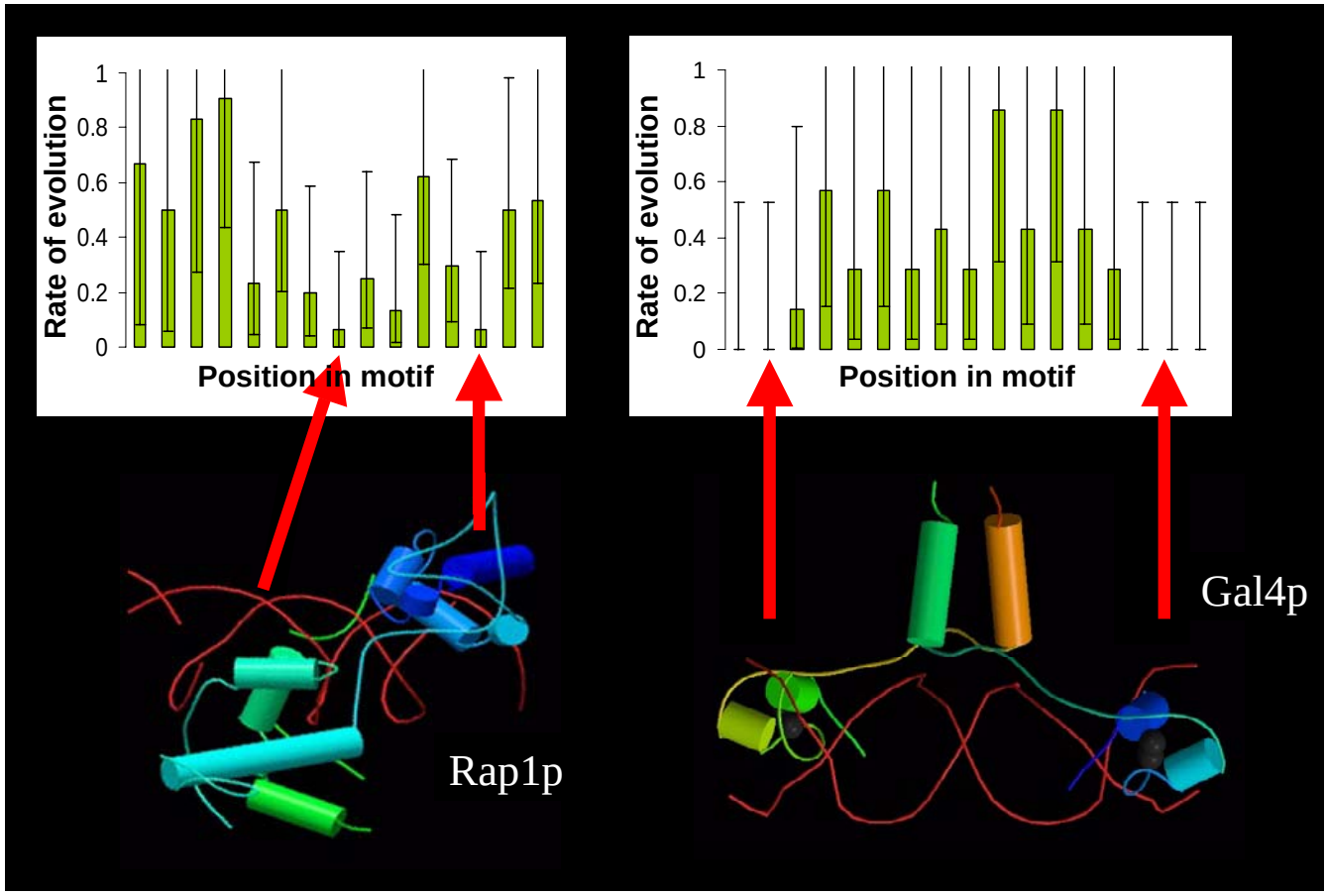


Figure 2
Comparison of rates of evolution to structures of protein-DNA complexes implies a model for the variation in the rate of evolution across binding motifs. The DNA backbone appears as a red helix; proteins appear as linked coloured cylinders. We propose that the formation of the protein-DNA complex is the functional constraint that leads to purifying selection, and therefore fewer substitutions at certain positions in the binding motif. Images of protein-DNA complex structures are from the Protein Data Bank [47]. Rate of evolution is in substitutions per site (estimated by maximum parsimony) and error bars represent exact 95 % confidence intervals for a Poisson distribution.

example, at a given position in the motif, a transcription factor recognizes only guanine, i.e., $(f_A,f_C,f_G,f_T) = (0,0,1,0)$, a mutation to any other nucleotide should prohibit formation of the protein-DNA complex, and therefore be deleterious. Such mutations should be removed from the population and therefore the number of observed substitutions at such a position is expected to be very small. Similarly, if the binding protein requires, say, A or T at a given position with no preference, i.e., $(f_A,f_C,f_G,f_T) = (1/2,0,0,1/2)$, we expect changes between A and T to persist in the population, but changes to C or G to be removed; we should therefore observe somewhat more substitutions, but still fewer than at positions where there is no preference at all, i.e., $(f_A,f_C,f_G,f_T) = (1/4,1/4,1/4,1/4)$, and all types of substitutions are permitted. Under several assumptions, it is possible to write the following proportionality for the rates of substitution between various residues and a function of their frequencies ([29] equation 10 & see methods).

$$R_{abp} \propto P_{ab} \times \frac{\ln\left(\frac{f_{bp}P_{ba}}{f_{ap}P_{ab}}\right)}{1-\frac{f_{ap}P_{ab}}{f_{bp}P_{ba}}},$$

where R_{abp} is the observed rate of substitution from residue a to residue b at position p, P_{ab} and P_{ba} are the (position independent) underlying rates of mutation from residue a to residue b and b to a, respectively, and f_{ap}

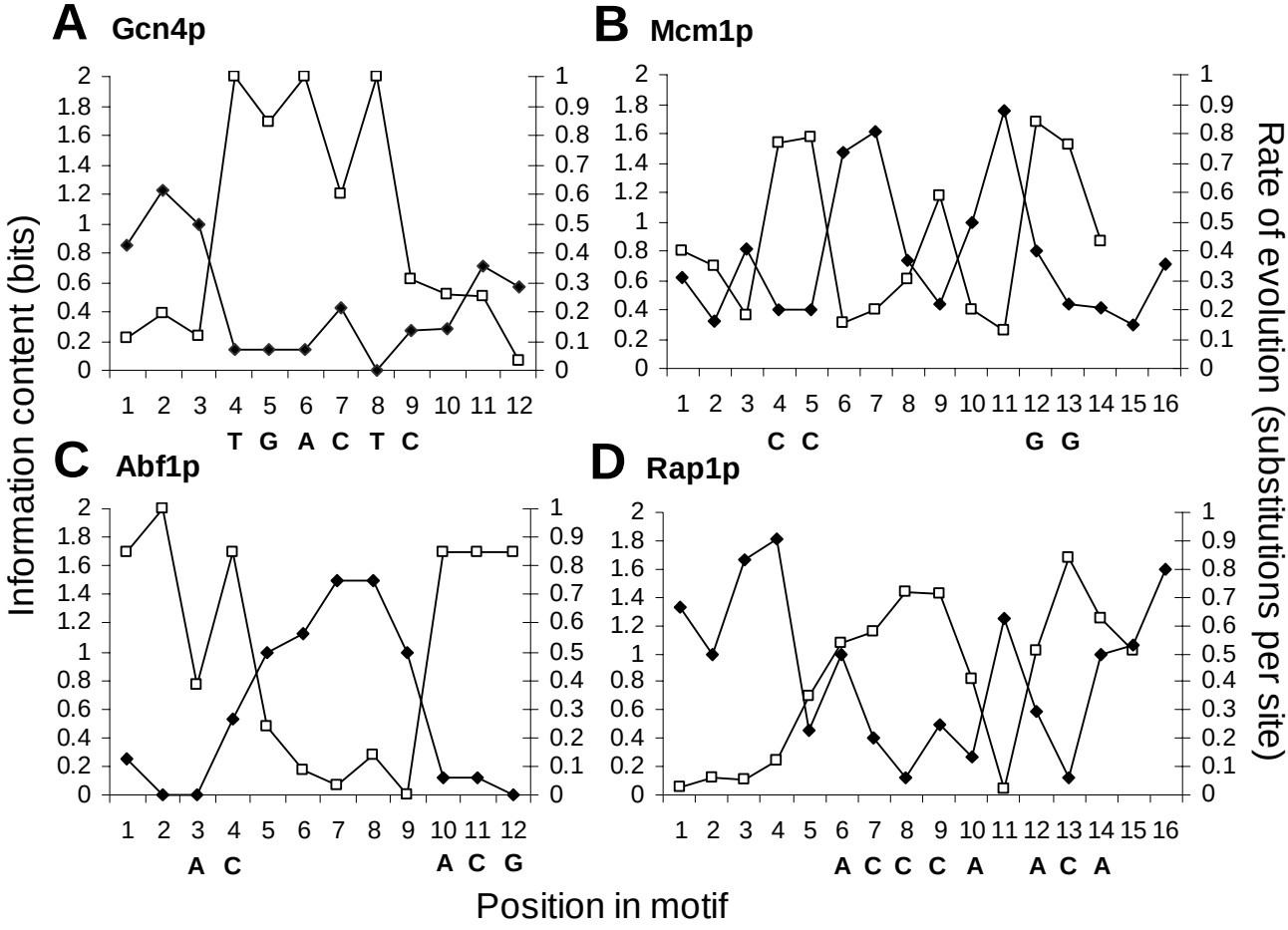


Figure 3

Association between information profile and rate of evolution in characterized binding sites from SCPD. A–D. Representative plots of information content and substitutions per site reveal a correspondence between positions of high information content and slower rates of evolution. Open symbols represent information content and filled symbols the number of substitutions per site (estimated by maximum parsimony). Consensus letters are included below the appropriate positions in the motif.

and f_{bp} are the frequencies of residue a and b at position p in the position weight matrix. The predicted rate of evolution at each position, K_p, is just the sum of the R_{abp} times the probability that that base was observed, i.e.,

$$K_p \propto \sum_a \sum_{b \neq a} f_{ap} R_{abp}.$$

In order to test these predictions, we estimated a background mutation model (P_{ab}) by fitting the HKY85 model [34] to entire promoter sequences using the PAML package [35], treating all positions independently (see methods). Using the seven position weight matrices (f_{ap}) trained on the characterized binding sites (all from *S. cerevisiae*,) and scaling the proportionality by the total number of changes observed in the motif, we compared the predicted rates to the observed rates and the results are shown in figure 4. Although there is quite a bit of variability, the observed rates of evolution seem to agree with the predictions ($R^2 = 0.67$).

Computationally predicted binding sites show similar evolutionary properties

There are relatively few transcription factors for which the number of experimentally characterized binding sites was sufficient to reliably estimate the information profile and rate of evolution at each position. To further establish the generality of these observations we extended the analysis

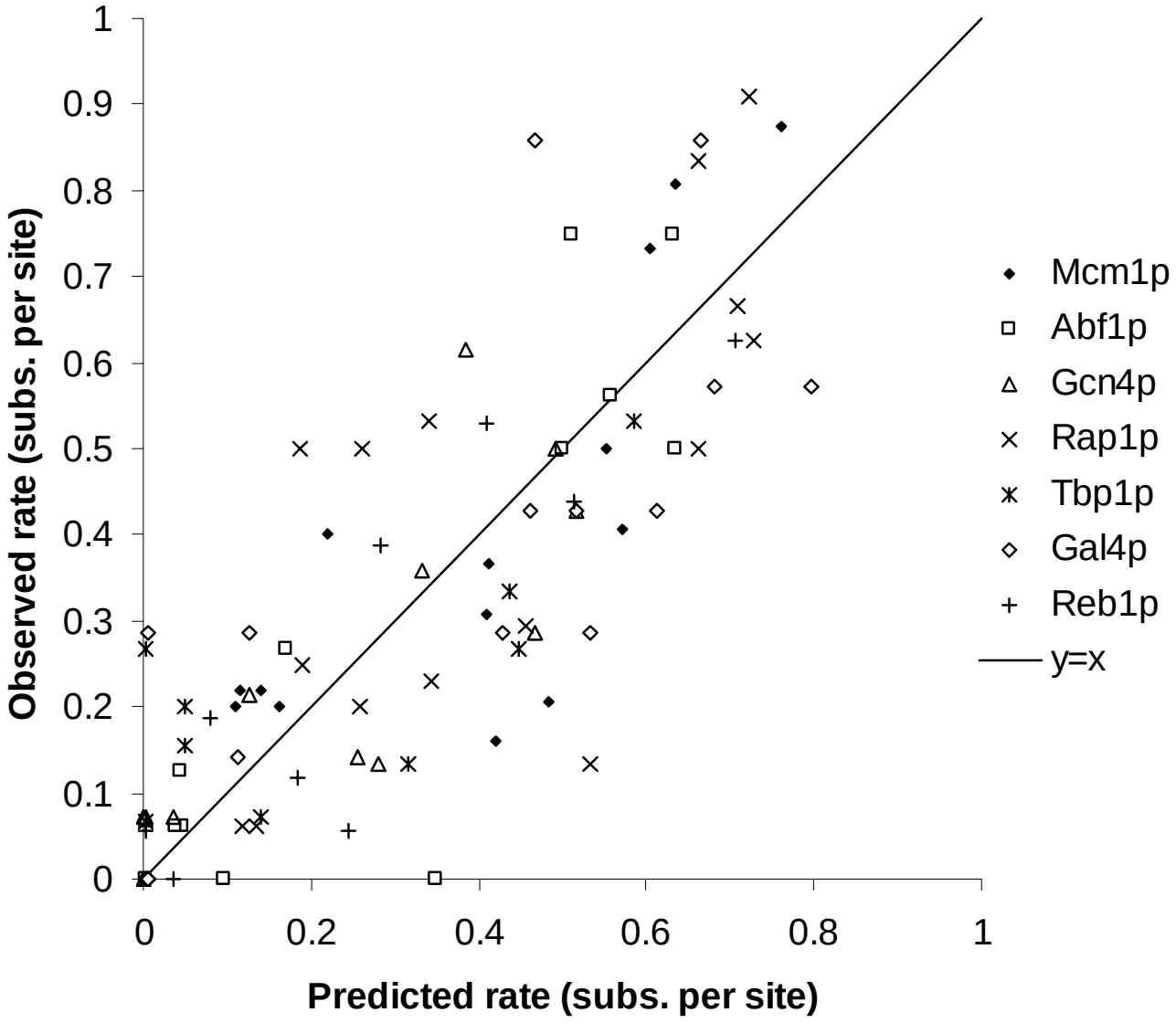


Figure 4
Test of the Halpern-Bruno proportionality. Observed rate of evolution versus the predictions based on the nucleotide frequencies in the binding motif in *S. cerevisiae*. Each point represents the predicted and observed rates at a given position in a motif. For each factor the proportionality has been normalized by the total number of substitutions observed in the corresponding binding sites. See text for details.

to include additional factors where some information regarding the consensus or target genes was available. We ran the MEME motif-finding program [3] on the promoter regions of groups of genes that showed similar expression patterns to the known targets of these factors in microarray experiments to derive models of their binding specificity and identify putative binding sites. As in the experimentally characterized cases, the rate of evolution in these binding sites was slower than that of the promoters in which the sequences were found (table 2.) Furthermore, most of the motifs showed the characteristic correlation between the information content at each position and the number of substitutions per site (table 2).

The pattern of evolution may be useful in distinguishing real motifs from computational artifacts

A challenge in computational motif detection is that algorithms often identify sequence motifs that do not

Table 2: Evolution of motifs with known consensus, but binding sites identified by MEME

Cluster	Factor	Consensus identified by MEME	Motif subs.	Bg subs.	Corr.	p-value
Protein folding +	Hsf1p	TTTTCTAGAAAGTTC	0.14	0.68	-0.42	0.060
Glycolysis +	Gcr1p	AAATAGAGGAAGCCCA	0.23	0.63	-0.80	<0.001 **
Nitrogen +	Gln3p/ Dal80p	TCTTATCA	0.39	0.74	-0.78	0.010 *
Gluco-neogenesis +	Sip4p (?) CSRE	CCGTTTGTCCG	0.33	0.57	-0.84	<0.001 **
G1 phase +	Mbp1p Swi6	TTACGCGTTTT	0.22	0.64	-0.67	0.011 *
Respiration +	Hap2/3/4p	TGATTGGTCCA	0.20	0.67	-0.53	0.048 *
Methionine +	Cbf1p	ATGTCACGTG	0.13	0.75	-0.49	0.078
Proteasome +	Rpn4p	ATTTTGCCACCG	0.20	0.73	-0.75	0.002 *
M/G1 transition +	Swi5p/ Ace2p	AACCAGCA	0.26	0.61	-0.57	0.074
Repressed in Stress ++	(?) PAC	ATGCGATGAGCTGAG	0.24	0.71	-0.69	0.006 *
leu/ilv bio-synthesis++	Leu3p	GCCGTTTCCGG	0.31	0.70	-0.54	0.044 *
Phosphate +++	Pho4p	CCCACGTGCG	0.29	0.65	-0.74	0.005 *
119 positions in all computationally identified motifs					-0.60	4e-14 **

Here binding sites are identified by running the MEME program [3] on genes that clustered with targets in micro-array gene expression data. Expected consensus sequences (from [20,40] or [7]) are underlined. 'Motif subs.' and 'bg subs.' are the substitutions per site in the binding sites and the promoters in which they are found respectively. 'Corr.' and 'p-value' are the Spearman's rank correlation coefficient and the associated p-value between the rate of evolution at each position and the information content at each position. * Indicates significance at a per factor error rate of < 0.05. ** Indicates significance after Bonferoni correction for a global error rate < 0.05, assuming 50 tests were done in total. (?) indicates uncertainty as to the identity of the binding protein. + indicates clusters taken from hierarchical clustering [40] of yeast data from the Stanford Microarray database [42], ++ indicates clusters taken from hierarchical clustering of 300 genetic perturbations [43] and +++ indicates clusters taken from hierarchical clustering of 64 control experiments [43]

Table 3: Motifs identified by MEME that do not correspond to the expected consensus sequences for the transcription factors thought to be regulating the cluster.

Cluster	Factor (expected)	Consensus identified by MEME	Motif subs.	Bg. subs.	Corr.	p-val
TRX2 +	Yap1p	AAAAAGAGGAAAAAA	0.80	0.78	-0.21	0.23
CTT1 +	Msn2/4p	GAAAAAAAAAAAAAA	0.51	0.67	0.13	0.67
Transport ++	Pdr1/3p	AAAGAGAGAAAAAAA	0.57	0.69	0.20	0.76
Ergosterol bio-synthesis++	Upc2p/ Ecm22p	ATCTTTTTTTTTTTT	0.81	0.55	0.06	0.58
60 positions in background sequence					0.08	0.73

These motifs do not show the characteristic correlation with rate of substitution or the substantial decrease in substitution rate observed for the computationally identified motifs with the expected consensus. + indicates clusters taken from hierarchical clustering of yeast data from the Stanford Microarray database [42], ++ indicates clusters taken from hierarchical clustering of 300 genetic perturbations [43].

represent real transcription factor binding sites. For example, in addition to the cases described above (table 2), there were several cases where the motif identified by MEME was not the binding motif for the factor known to regulate these genes. We computed the number of substitutions per site as well as the correlation between the number of substitutions and the information content for these motifs as well, and found no significant correlations (Table 3), suggesting that the reported motifs in these cases may be computational artefacts. It is possible that a reduction in the average number of substitutions per site, and a correlation between the information profile and the substitutions across the motif will prove to be useful heuristics in assessing the support from comparative sequence data for computationally identified motifs.

In order to further test this idea we ran MEME on the promoters of a group of proposed Crz1p target genes identified in a recent microarray study [36]. We found that the resulting motif (figure 5) was on average more conserved (0.38 subs. per site, n = 297) than the promoters of the genes in the group (0.65 subs. per site, n = 11832). In addition, it showed the characteristic correlation between the information profile and rate of evolution across the motif (Spearman's rank = -0.78, p = 0.001). Thus, in this case, the comparative sequence data support the hypothesis that this is a functional binding motif in these genes.

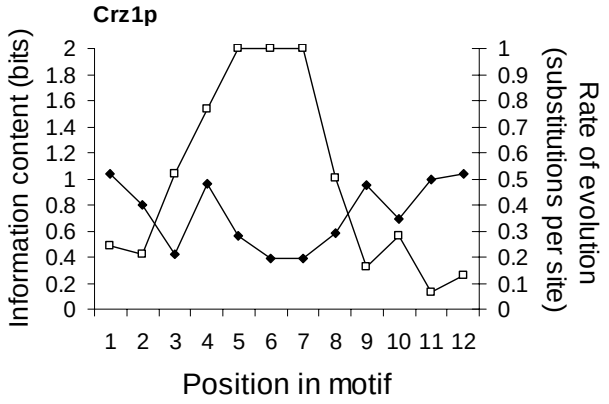


Figure 5
Information and rate of evolution for the recently reported Crz1p motif. This motif shows the characteristic pattern of evolution observed for real motifs. Open symbols represent information content and filled symbols, the number of substitutions per site (estimated by maximum parsimony.) Consensus letters are included below the appropriate positions in the motif.

Discussion

Motifs are conserved on average, but individual binding sites are not perfectly conserved

We confirm an important motivating assumption of comparative sequencing projects: the rate of evolution within functional non-coding sequences elements is slower than the surrounding intergenic DNA (fig. 1). While this means that on average binding sites are conserved, it is important to note, however, that in no case was the average number of substitutions over the motif reduced to zero. Since substitutions do occur in characterized binding sites, simply searching through alignments for perfectly conserved segments would not have revealed all the real binding sites used in this study. Nevertheless, binding sites do show characteristic patterns of evolution, and it should be possible to take these into account in attempting to distinguish the functional instances of the motif.

Position specific variation in the rate of evolution is consistent with models of functional constraint

The observation that the rate of evolution is not constant over functional non-coding DNA sequences mirrors similar observations of regional variation in the number of substitutions per site in peptide sequences; residues that are more important to the function or structure of the protein change much less rapidly, presumably because mutations at these positions are likely to be deleterious, and therefore do not drift to fixation [21–23]. By analogy to peptide sequences, the observation that the positions in functional non-coding DNA with high information content evolve more slowly is consistent with these positions being more important for the formation of the protein-DNA complex, and therefore under more functional constraint. Unlike peptide sequences, however, the purifying selection and accompanying reduction in the rate of substitution in transcription-factor binding sites seems to be a relatively straightforward mapping from the physical interaction of the DNA with the binding protein (as in fig. 2). Since the information content has been shown to correlate with the physical constraints imposed by transcription factors on their motifs [28] it is consistent that we observe significant correlations between the information profiles and the rate of evolution as well.

The binding sites of sequence specific transcription factors afford a rare opportunity to test theoretical predictions of the effects of purifying selection on site-specific rates of evolution. By assuming the nucleotide frequencies from position specific weight matrices are the equilibrium frequencies under the purifying selection imposed on these sequences, we could make seemingly reasonable predictions for the rate of evolution at each position (figure 4). Although we do not have sufficient data to reliably estimate the rates for each type of substitution (e.g., A→T vs. A→G,) the results presented here are promising. The same intuition that allows us to construct position weight matrices (i.e., that we may average over all the binding sites to learn the average sequence specificity) allows us to compute the rate of evolution across the motif by averaging the changes observed in the individual binding sites.

Improved understanding of binding site evolution can guide the use of comparative data

An accurate understanding of the evolution of functional regulatory sequences is critical to the optimal use of comparative sequence data in the analysis of transcriptional regulation. Without such an understanding, it remains difficult to distinguish sequences under functional constraint from sequences that are similar because of shared descent, or to differentiate among the various classes of conserved non-coding sequences. We believe our observations linking position-specific variation in the rate of evolution within transcription factor binding sites to position-specific sequence variation within genomes (and to structural features of the protein-DNA complex) will be useful in comparative sequence analysis.

For example, comparative sequence data can be used to verify the predictions of *de novo* motif finding algorithms that have been applied to single genomes, by allowing us to ascribe increased confidence to predicted motifs that are also conserved. However, simply assessing whether motifs are 'present' in other species can be ineffective as similar sequences are expected to be present in closely

related species because they have had insufficient time to diverge or as the result of other functional constraints. We propose that the patterns of evolution we observe for known motifs – their conservation relative to flanking sequences and the correlation between position-specific rate of evolution and intragenomic degeneracy – can more accurately distinguish motif models that correspond to *bona fide* transcription factor binding sites from computational artefacts (compare tables 2 and 3). As a demonstration we show that comparative sequence supports the motif reported in [36] (fig. 5). Verification of computationally predicted motifs may be an immediate practical application of our observations and computational methods that incorporate models of binding site evolution should take more effective advantage of comparative sequence data.

More generally, just as faster evolution at synonymous sites is an evolutionary signature of protein coding regions [21–23], the pattern of position-specific variation in evolutionary rates within binding sites can be thought of as an evolutionary signature of transcription factors. We have shown here how these evolutionary signatures might be used to identify sequences or motifs that collectively have the properties of transcription factor binding sites. With sufficient sequence data, it should ultimately be possible to estimate the rate of evolution at every base in a genome, and to identify individual short sequences with the evolutionary characteristics of functional transcription factor binding sites.

Conclusions

We show that the rate of evolution in characterized and predicted transcription factor binding sites is slower than that of the intergenic regions in which they are found. In addition we show that there is position specific variation in the rate of evolution across these binding sites. We show that this variation is correlated to the variability in the sequence specificity for that factor and can be modelled by assuming that purifying selection acts to maintain these specificities. Together this suggests that the variation in the rate of evolution is a direct reflection of differences in the strength of purifying selection due to differing physical constraints on the DNA imposed by the interaction with the binding protein.

The characterization of the pattern of conservation over known binding sites is an important step in understanding the evolution of functional non-coding DNA, and perhaps also towards the general understanding of protein-DNA interactions. Our observations should contribute to the effectiveness of comparative non-coding sequence analysis.

Methods

Rates of binding-site and intergenic evolution

Global alignments of intergenic regions from *S. mikatae*, *S. paradoxus* and *S. bayanus* were computed using clustalw (as described in [12]). Using the accepted species tree (Sbay, Smik,(Spar, Scer)) [8,12], we computed the minimal number of changes needed for each column of the alignment (the so called cost) using the classical parsimony algorithm (as described in [37]). We included only alignments where sequence from all four species was available; regions of ambiguity or missing sequence in the alignment, were treated as gaps. The average rate of evolution within a binding site (in fig. 1A) is the sum of the cost at each position in the binding site divided by its length. The average rate of evolution for a motif (in fig. 1B) is the sum of the cost in the binding sites divided by the total number of ungapped positions in the binding sites. Although gaps are not expected in alignments of functional binding sites, we allow for them so that we can apply the same metrics to binding sites as the surrounding sequences. The background histogram in figure 1A was made by calculating the average rate of evolution in randomly drawn 17-mers from the promoters of the genes containing the binding sites. The rate of background evolution (in fig. 1B) is the sum of the cost over the entire alignment divided by the total number of ungapped positions. The average rate of evolution at each position is the sum (over all the binding sites) of the cost at that position divided by the total number of binding sites that have no gap at that position. Although a maximum likelihood estimator for the number of substitutions per site in DNA sequences has been constructed [38] its performance is expected to be similar to parsimony methods for short evolutionary distances as are considered here.

We note that the rate of background evolution differed significantly among the groups of genes examined (characterized targets, expressions clusters.) We address this variation, and examine possible explanations in another manuscript (Hunter B Fraser, AMM and MBE in preparation).

Statistics

A Poisson distribution for the number of substitutions was used when reporting confidence intervals, and for error bars in figures 1 and 2, because this is thought to be a reasonable model for the underlying distribution for neutral substitution events.

The significance of difference of means between motifs and background was estimated by bootstrapping. We randomly selected sequences the same length as the motif (with replacement) from the upstream regions in which they were found, until we had the same number as we had characterized binding sites. We then calculated for these

samples the mean number of substitutions exactly as for the characterized binding sites. Finally, we repeated this process 1000 times, and asked how often we observed an evolutionary rate smaller than for the characterized sites. Both the rate of evolution in promoter sequences and the locations of yeast transcription factor binding sites are known to show positional preferences ([39], AMM and MBE, unpublished data). To control for possible effects of these biases, we also calculated the number of substitutions in sequences of the same size as the motif 5 basepairs away on either side of the binding sites, for each of the factors shown in fig. 1B. To be as conservative as possible, we simply computed the probability of observing the number of changes in the binding sites out of the total number of changes in the binding sites and the flanking regions with no assumptions about the underlying distribution of the changes (using the hypergeometric distribution) and found that there were fewer substitutions in the binding sites than in the flanking regions (data not shown).

Identification of binding sites and construction of position weight matrices

Characterized binding sites were taken from SCPD [20] for Gal4p (n = 10), Mcm1p (n = 35), Abf1p (n = 16), and Rap1p (n = 17). For some of the short Gcn4p (n = 15), Reb1p (n = 18) and Tbp1p (n = 15) sites up to 5 flanking base pairs were included. Although SCPD lists additional binding sites for many of these factors, we excluded many of these because they were redundant listings of binding sites that have been characterized multiple times (e.g., *STE6* has 4 Mcm1p binding sites listed, but these are actually the same two listed twice) or they were found in divergently transcribed genes, and were listed independently for both genes (e.g., *GAL1* and *GAL10* both have 4 Gal4p binding sites listed, but in fact they share these sites). For each factor, the sequences were aligned using the MEME program [3] and the 'letter-probability-matrix' from its output was used as the position weight matrix.

SCPD lists binding sites for many other regulatory elements and transcription factors, most of which have few sites, or have sites from a small number of target genes. For each transcription factor, we attempted to identify groups of genes with similar expression patterns as the known target genes, as well as known target genes of other transcription factors (from [40]). These groups were then chosen by hand from hierarchical clustering [41] of expression data from various experimental treatments and over the cell-cycle, downloaded from the Stanford Microarray Database [42] or from 300 publicly available deletion and drug treatment experiments or 64 control experiments [43]. We ran MEME on the putative promoter regions of genes in expression clusters with the following parameters: motif width was allowed to range between 8 and 16, 'zoops' and 'tcm' models were both tried for each case, and both strands of the promoter were searched. When the 'tcm' model was used, we specified between 0.5 n and 2 n for the number of occurrences where n is the number of genes in the cluster. For MEME runs, promoter regions were taken to be the 600 basepairs upstream of the translation start (basepairs in other coding regions were excluded), except in the case of the proteasome and the repressed stress genes where 300 basepairs were used because of a positional bias in the location of those binding sites (AMM and MBE unpublished results.) For computationally predicted binding-sites, occurrences were taken to be those listed in the MEME output, and the 'letter-probability-matrix' was used as position weight matrix. In the case of Crz1p we used the starting consensus NNNNGGCNCNN, which was reported in [36].

Correlation with information profiles

Information at each position was calculated as

$$I_p = 2 + \sum_b f_{bp} \log_2 f_{bp},$$

where f_{bp} is the frequency of base b at position p in the motif, with $b \in \{A, C, G, T\}$, and $p \in [1, W]$ where W is the width of the motif. Spearman's rank-order correlation (the linear correlation of the ranks) was computed and the significance of the correlation coefficient was assigned as described in [44].

Predictions of the rate of evolution

We follow exactly the derivation for protein sequences found in [29]. Briefly, if we assume that sites are independent, evolution is reversible, and underlying probabilities of mutation are invariant across sites, we can write the rate of evolution at each position as

$$R_{abp} \propto P_{ab} \times F_{abp},$$

where R_{abp} is the rate of substitution from residue a to residue b at position p, P_{ab} is the rate of mutation from residue a to residue b and F_{abp} is the probability of fixation of a mutation from residue a to residue b at position p. If we assume that the time of fixation is small relative to the time between fixations, a so-called weak-mutation model [45], we can use Kimura's equations [46] and write the following.

$$F_{abp} = \frac{1 - e^{-2s_p}}{1 - e^{-2Ns_p}} \approx \frac{2s_p}{1 - e^{-2Ns_p}}$$

$$F_{bap} \approx \frac{1 - e^{2s_p}}{1 - e^{2Ns_p}} \approx \frac{-2s_p}{1 - e^{2Ns_p}},$$

where N is the effective population size and s_p is the coefficient of selection at position p. As was noted in [29], if equilibrium has been reached, i.e., there has been sufficient time for all the possible mutations at that position to occur and either be fixed or removed, then

$$\frac{f_{bp}P_{ba}}{f_{ap}P_{ab}} = \frac{F_{abp}}{F_{bap}} \approx e^{2Ns_p},$$

where f_{ap} is the equilibrium frequency of residue a at position p, in our case the frequency in the position weight matrix. This implies

$$2Ns_p = \ln\left(\frac{f_{bp}P_{ba}}{f_{ap}P_{ab}}\right)$$

and therefore

$$F_{abp} \propto \frac{\ln\left(\frac{f_{bp}P_{ba}}{f_{ap}P_{ab}}\right)}{1-\frac{f_{bp}P_{ba}}{f_{ap}P_{ab}}},$$

which can be substituted to give the proportionality used in results.

To fit a background mutation model, we used PAML [35] to fit the HKY model [34] to the promoters that contain the characterized binding sites for each factor. We fixed the alpha parameter at 0 to use a constant rate across sites. The HKY model accounts for equilibrium frequencies of nucleotides as well as transition-transversion mutation bias. The equilibrium frequencies differed from ($^1/_4$,$^1/_4$,$^1/_4$,$^1/_4$), and transitions were more probable than transversions (kappa between 3 and 4.) We also tested the site-specific predictions for the rate of evolution assuming that all types of substitutions were equally likely and qualitatively the results were very similar (data not shown).

Authors' contributions

MK and ESL generously provided sequence alignments. DYC provided assistance with the organization, calculations, statistics and writing of the manuscript. MBE oversaw the project and provided much of the conceptual framework. AMM performed the computations and prepared the manuscript.

Acknowledgements

We thank Dr. Justin Fay, Dr. Audrey Gasch, Dr. Paul Spellman and Dr. Casey Bergman for comments on the manuscript. In addition, we thank Dr. Audrey Gasch for suggesting the extension to computationally predicted motifs, as well as Dr. Justin Fay for stimulating discussions. Thanks to Dr. William Bruno for pointing us to his work. MBE is a Pew Scholar in the Biomedical Sciences. This work was conducted under the US Department of Energy contract No. ED-AC03-76SF00098.

References

1. Stormo GD: **DNA binding sites: representation and discovery.** *Bioinformatics* 2000, **16(1):**16-23.
2. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF and Wootton JC: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.** *Science* 1993, **262(5131):**208-214.
3. Bailey TL and Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology AAAI Press, Menlo Park, California*; 1994:28-36.
4. Eskin E and Pevzner PA: **Finding composite regulatory patterns in DNA sequences.** *Bioinformatics* 2002, **18(Suppl 1):**S354-363.
5. Liu XS, Brutlag DL and Liu JS: **An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments.** *Nat Biotechnol* 2002, **20(8):**835-839.
6. Marsan L and Sagot MF: **Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification.** *J Comput Biol* 2000, **7(3–4):**345-362.
7. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ and Church GM: **Systematic determination of genetic network architecture.** *Nat Genet* 1999, **22(3):**281-285.
8. Cliften PF, Hillier LW, Fulton L, Graves T, Miner T, Gish WR, Waterston RH and Johnston M: **Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis.** *Genome Res* 2001, **11(7):**1175-1186.
9. Blanchette M, Schwikowski B and Tompa : **Algorithms for phylogenetic footprinting.** *J Comput Biol* 2002, **9(2):**211-223.
10. McCue L, Thompson W, Carmack C, Ryan MP, Liu JS, Derbyshire V and Lawrence CE: **Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes.** *Nucleic Acids Res* 2001, **29(3):**774-782.
11. Rajewsky N, Socci ND, Zapotocky M and Siggia ED: **The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons.** *Genome Res* 2002, **12(2):**298-308.
12. Kellis M., Patterson N, Endrizzi M, Birren B and Lander ES: **Sequencing and Comparison of Yeast Species to Identify Genes and Regulatory Elements.** *Nature* 2003, **423(6937):**241-254.
13. Hardison RC: **Conserved noncoding sequences are reliable guides to regulatory elements.** *Trends in Genetics* 2000, **16(9):**369-372.
14. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H and Oliver SG: **Life with 6000 genes.** *Science* 1996, **274(5287):**563-567.
15. Ludwig MZ, Patel NH and Kreitman M: **Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change.** *Development* 1998, **125(5):**949-958.
16. Dermitzakis ET and Clark AG: **Evolution of Transcription Factor Binding Sites in Mammalian Gene Regulatory Regions: Conservation and Turnover.** *Mol Biol Evol* 2002, **19(7):**1114-1121.
17. Elnitski L, Hardison RC, Li J, Yang S, Kolbe D, Eswara P, O'Connor MJ, Schwartz S, Miller W and Chiaromonte F: **Distinguishing regulatory DNA from neutral sites.** *Genome Res* 2003, **13(1):**64-72.
18. Wasserman WW, Palumbo M, Thompson W, Fickett JW and Lawrence CE: **Related Human-mouse genome comparisons to locate regulatory sites.** *Nat Genet* 2000, **26(2):**225-228.
19. Levy S, Hannenhalli S and Workman C: **Enrichment of regulatory signals in conserved non-coding genomic sequence.** *Bioinformatics* 2001, **17(10):**871-877.
20. Zhu J and Zhang MQ: **SCPD: a promoter database of the yeast Saccharomyces cerevisiae.** *Bioinformatics* 1999, **15(7–8):**871-877.
21. Kimura M: *The Neutral Theory of Molecular Evolution Cambridge University Press, Cambridge*; 1983.
22. Li WH: *Molecular Evolution Sinauer Associates, Sunderland MA*; 1997.
23. Nei M: *Molecular Evolutionary Genetics Columbia University Press, New York*; 1987.
24. Matthews BW: **Protein-DNA interaction. No code for recognition.** *Nature* 1988, **335(6188):**294-295.

25. Suzuki M, Brenner SE, Gerstein M and Yagi N: **DNA recognition code of transcription factors.** *Protein Eng* 1995, **8(4):**319-328.
26. Kono H and Sarai A: **Structure-based prediction of DNA target sites by regulatory proteins.** *Proteins* 1999, **35(1):**114-131.
27. Benos PV, Lapedes AS and Stormo GD: **Is there a code for protein-DNA recognition? Probab(ilistical)ly.** *Bioessays* 2002, **24(5):**466-475.
28. Mirny LA and Gelfand MS: **Structural analysis of conserved base pairs in protein-DNA complexes.** *Nucleic Acids Res* 2002, **30(7):**1704-1711.
29. Halpern AL and Bruno WJ: **Evolutionary distances for protein-coding sequences: modelling site-specific residue frequencies.** *Mol Biol Evol* 1998, **15(7):**910-917.
30. Marmorstein R, Carey M, Ptashne M and Harrison SC: **DNA recognition by GAL4: structure of a protein-DNA complex.** *Nature* 1992, **356(6368):**408-414.
31. Acton TB, Zhong H and Vershon AK: **DNA-binding specificity of Mcm1: operator mutations that alter DNA-bending and transcriptional activities by a MADS box protein.** *Mol Cell Biol* 1997, **17(4):**1881-1889.
32. Kerppola TK: **Transcriptional cooperativity: bending over backwards and doing the flip.** *Structure* 1998, **6(5):**549-554.
33. Tan S and Richmond TJ: **Crystal structure of the yeast MATalpha2/MCM1/DNA ternary complex.** *Nature* 1998, **391(6668):**660-666.
34. Yang Z, Goldman N and Friday AE: **Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation.** *Mol Biol Evol* 1994, **11(2):**316-324.
35. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **13(5):**555-556.
36. Yoshimoto H, Saltsman K, Gasch AP, Li HX, Ogawa N, Botstein D, Brown PO and Cyert MS: **Genome-wide Analysis of Gene Expression Regulated by the Calcineurin/Crz1p Signalling Pathway in Saccharomyces cerevisiae.** *J Biol Chem* 2002, **277(34):**31079-31088.
37. Durbin R, Eddy S, Krogh A and Mitchison G: *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids Cambridge University Press, Cambridge, UK*; 1998.
38. Nielsen R: **Site-by-site estimation of the rate of substitution and the correlation of rates in mitochondrial DNA.** *Syst Biol* 1997, **46(2):**346-353.
39. Hampson S, Kibler D and Baldi P: **Distribution patterns of over-represented k-mers in non-coding yeast DNA.** *Bioinformatics* 2002, **18(4):**513-528.
40. Hodges PE, Payne WE and Garrels JI: **The Yeast Protein Database (YPD): a curated proteome database for Saccharomyces cerevisiae.** *Nucleic Acids Res* 1998, **26(1):**68-72.
41. Eisen MB, Spellman PT, Brown PO and Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A* 1998, **95(25):**14863-14868.
42. Gollub J, Ball CA, Binkley G, Demeter J, Finkelstein DB, Hebert JM, Hernandez-Boussard T, Jin H, Kaloper M, Matese JC, Schroeder M, Brown PO, Botstein D and Sherlock G: **The Stanford Microarray Database: data access and quality assessment tools.** *Nucleic Acids Res* 2003, **31(1):**94-96.
43. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M and Friend SH: **Functional discovery via a compendium of expression profiles.** *Cell* 2000, **102(1):**109-126.
44. Press WH, Teukolsky ST, Vetterling WT and Flannery BP: *Numerical Recipes in C* 2nd edition. *Cambridge University Press, Cambridge, UK*; 1992.
45. Golding B and Felsenstein J: **A maximum likelihood approach to the detection of selection from a phylogeny.** *J Mol Evol* 1990, **31:**511-523.
46. Kimura M: **On the probability of fixation of mutant genes in a population.** *Genetics* 1962, **4:**713-719.
47. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28(1):**235-242.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 11, Numbers 2–3, 2004
© Mary Ann Liebert, Inc.
Pp. 319–355

Methods in Comparative Genomics: Genome Correspondence, Gene Identification and Regulatory Motif Discovery

MANOLIS KELLIS,[1,2] NICK PATTERSON,[1] BRUCE BIRREN,[1]
BONNIE BERGER,[2,3,5] and ERIC S. LANDER[1,4,5]

ABSTRACT

In Kellis *et al.* (2003), we reported the genome sequences of *S. paradoxus*, *S. mikatae*, and *S. bayanus* and compared these three yeast species to their close relative, *S. cerevisiae*. Genomewide comparative analysis allowed the identification of functionally important sequences, both coding and noncoding. In this companion paper we describe the mathematical and algorithmic results underpinning the analysis of these genomes. (1) We present methods for the automatic determination of genome correspondence. The algorithms enabled the automatic identification of orthologs for more than 90% of genes and intergenic regions across the four species despite the large number of duplicated genes in the yeast genome. The remaining ambiguities in the gene correspondence revealed recent gene family expansions in regions of rapid genomic change. (2) We present methods for the identification of protein-coding genes based on their patterns of nucleotide conservation across related species. We observed the pressure to conserve the reading frame of functional proteins and developed a test for gene identification with high sensitivity and specificity. We used this test to revisit the genome of *S. cerevisiae*, reducing the overall gene count by 500 genes (10% of previously annotated genes) and refining the gene structure of hundreds of genes. (3) We present novel methods for the systematic *de novo* identification of regulatory motifs. The methods do not rely on previous knowledge of gene function and in that way differ from the current literature on computational motif discovery. Based on genomewide conservation patterns of known motifs, we developed three conservation criteria that we used to discover novel motifs. We used an enumeration approach to select strongly conserved motif cores, which we extended and collapsed into a small number of candidate regulatory motifs. These include most previously known regulatory motifs as well as several noteworthy novel motifs. The majority of discovered motifs are enriched in functionally related genes, allowing us to infer a candidate function for novel motifs. Our results demonstrate the power of comparative

[1]Whitehead Institute Center for Genome Research, MIT, Cambridge, MA 02139.
[2]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA 02139.
[3]Department of Mathematics, MIT, Cambridge, MA 02139.
[4]Department of Biology, MIT, Cambridge, MA 02139.
[5]Corresponding author.

genomics to further our understanding of any species. Our methods are validated by the extensive experimental knowledge in yeast and will be invaluable in the study of complex genomes like that of the human.

Key words: comparative genomics, computational biology, yeast, *Saccharomyces cerevisiae*, genome alignment, gene finding, gene identification, gene regulation, regulatory motifs, motif discovery.

INTRODUCTION

With the availability of complete sequences for a number of organisms, comparative analysis becomes an invaluable tool for understanding genomes. Complete genomes allow for global views, and multiple genomes increase predictive power.

In Kellis *et al.* (2003) we used a comparative genomics approach to systematically discover the full set of conserved genes and regulatory elements in yeast. We sequenced and assembled three novel yeast species, *S. paradoxus*, *S. mikatae*, and *S. bayanus* and compared them to their close relative, *S. cerevisiae*. The work represented the first genomewide comparison of four complete eukaryotic genomes. This paper focuses on the computational algorithms or techniques or methods underpinning the work.

We first describe our methods for resolving the genome correspondence between each of the newly sequenced species and *S. cerevisiae* to identify orthologous regions. The method presented here was used for the automatic alignment of the three newly sequenced species and correctly identified unambiguous orthologs for more than 90% of known protein coding genes. It also identified the evolutionary events that separate the four species, discerning segmental duplications and gene loss, while correctly resolving genes that duplicated before the divergence of the species compared.

We then describe our methods for gene identification. We observed that in protein-coding genes, insertions and deletions locally compensate for each other, thus preserving the reading frame of amino acid translation. We quantified these properties by evaluating the reading frame conservation (RFC) in nucleotide alignments across related species. Based on the metric for frame conservation, we developed a test to accept or reject open reading frames (ORFs) as biologically meaningful or not. The test showed strong sensitivity and specificity and allowed us to revisit the genome of *S. cerevisiae*, reducing the overall gene count by 500 genes (10% of previously annotated genes) and refining the gene structure of hundreds of genes.

We finally describe our methods to identify regulatory elements based on their conservation patterns across all intergenic regions and the complete genome. The regulatory motif discovery methods presented here do not rely on previous knowledge of functionally related gene sets and in that way differ from the current literature on computational motif discovery. Based on genomewide conservation patterns of known motifs, we developed three conservation criteria that we used to discover novel motifs. We used an enumeration approach to select strongly conserved motif cores, which we extended and collapsed into a small number of candidate regulatory motifs. The motifs discovered include most previously published regulatory motifs and a number of noteworthy novel motifs. These show enrichment in functionally related gene sets, enabling us to assign candidate functions to novel motifs.

The extensive knowledge of gene function in yeast has enabled us to validate the power of our methods for discovering biological signals in closely related genomes by virtue of their conservation. The methods presented are general and should be applicable to any species, given an appropriate set of related genomes for comparative analysis. In particular, applying such comparative methods will be invaluable in the understanding of the human genome, but presents numerous challenges in extracting signal from noise, given the increased genome size and complexity, both in gene content and regulatory complexity.

1. GENOME CORRESPONDENCE

The first step in comparative genomics is determining the correct correspondence of chromosomal segments and functional elements across the species compared. This involves determining *orthologous*

segments of DNA that descend from the same region in the common ancestor of the species compared and *paralogous* regions that arose by duplication events prior to the divergence of the species compared. The mapping of regions across two genomes can be *one-to-one* in absence of duplication events, *one-to-many* if a region has undergone duplication or loss in one of the species, or *many-to-many* if duplication/loss events have occurred in both lineages.

Understanding the ancestry of the functional elements compared is central to our understanding and the applications of genome comparison. Most comparative methods have focused on one-to-one orthologous regions, but it is equally important to recognize which segments have undergone duplication events and which segments were lost since the divergence of the species. Comparing segments that arose before the divergence of the species may result in the wrong interpretations of sequence conservation and divergence. Further, in the presence of gene duplication, some of the evolutionary constraints that a region is under are relieved, and uniform models of evolution no longer capture the underlying selection for these sites. Thus, our methods for determining gene correspondence should account for duplication and loss events and ensure that the segments we compare are orthologous.

We used genes as discrete genomic anchors in order to align and compare the species. We constructed a bipartite graph connecting annotated protein-coding genes in *S. cerevisiae* with predicted protein-coding genes in each of the other species based on sequence similarity at the amino acid level. This bipartite graph contains the orthologous matches but also contains spurious matches due to shared domains between proteins of similar functions and gene duplication events that precede the divergence of the species.

We developed an algorithm for resolving the correspondence of genes across the four species and recognizing orthologous and paralogous genes. The algorithm presented has a number of attractive features. It uses a simple and intuitive graph theoretic framework that makes it easy to incorporate additional heuristics or knowledge about the genes at hand. It represents matches between sets of genes instead of only one-to-one matches, thus dealing with duplication and loss events in a straightforward way. It uses the additional information of the chromosomal positions of the compared genes, detecting stretches of conserved gene order and using these to resolve additional orthologous matches. It accounts for all genes compared, resolving *unambiguous* matches instead of simply *best* matches, thus ensuring that all genes with one-to-one correspondence are true orthologs. It works at a wide range of evolutionary distances and can cope with unfinished genomes containing sequence gaps even within genes.

1.1. Establishing gene correspondence

Previously described algorithms for comparing gene sets have been widely used for various purposes, but they are not applicable to the problem at hand.

Best Bidirectional Hits (BBH) (Fitch, 1970, 1995) identifies gene pairs that are best matches of each other and marks them as orthologous. In the case of a recent gene duplication, however, only one of the duplicated genes will be marked as the ortholog without signaling the presence of additional homologs. Thus, no guarantees are given that best bidirectional matches will represent orthologous relations and incorrect one-to-one matches may be established.

Clusters of Orthologous Genes (COG) (Tatusov *et al.*, 1997, 2001) goes a step further and matches groups of genes to groups of genes. Unfortunately, the grouping is too coarse, and clusters of orthologous genes typically correspond to gene families that may have expanded before the divergence of the species compared. This inability to distinguish recent duplication events from more ancient duplication events makes it inapplicable in this case, since the genome of *S. cerevisiae* contains hundreds of gene pairs that were anciently duplicated before the divergence of the species at hand (Wolfe and Shields, 1997). The COG method does not distinguish between copies of anciently duplicated genes, and hence many orthologous matches would not be detected.

We introduce the concept of a best unambiguous subset (BUS), namely, a group of genes such that all best matches of any gene within the set are contained within the set and no best match of a gene outside the set is contained within the set. A BUS builds on both BBHs and COGs to resolve the correspondence of genes across the species. The algorithm, at its core, represents the best match of every gene as a set of genes instead of a single best hit, which makes it more robust to slight differences in sequence similarity. A BUS can be isolated from the remainder of the bipartite gene correspondence graph while preserving all potentially orthologous matches. The BUS algorithm also allows a recursive application grouping the

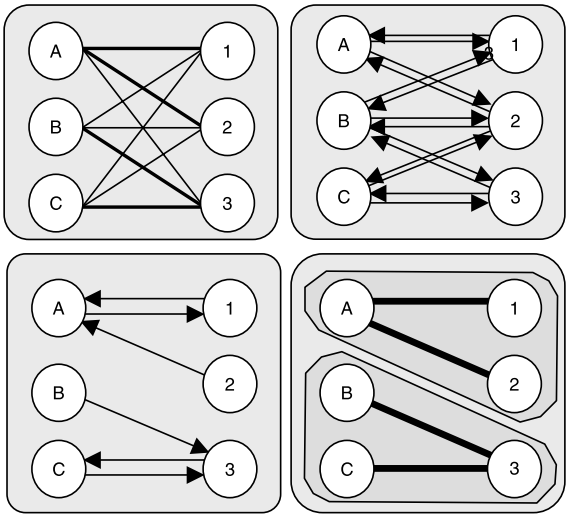


FIG. 1. Overview of graph separation. We constructed a bipartite graph based on the BLAST hits between genes A,B,C in one genome and genes 1,2,3 in a related genome. We considered both forward and reverse matches for near-optimality based on synteny and sequence similarity. Suboptimal matches were progressively eliminated simplifying the graph. We returned the connected components of the undirected simplified graph.

genes into progressively smaller subsets and retaining ambiguities until later in the pipeline when more information becomes available. Such information includes the conserved gene order (synteny) between consecutive one-to-one genes that allows the resolving of additional neighboring genes.

1.2. Overview of the algorithm

We formulated the problem of genomewide gene correspondence in a graph-theoretic framework. We represented the similarities between the genes as a bipartite graph connecting genes between two species (Fig. 1). We weighted every edge connecting two genes by both the amino acid sequence similarity between the two genes and the overall length of the match.

We separated this graph into progressively smaller subgraphs until the only remaining matches connected true orthologs. To achieve this separation, we eliminated edges that are suboptimal in a series of steps. As a preprocessing step, we eliminated all edges that are less than 80% of the maximum-weight edge both in amino acid identity and in length (1.4). Based on the unambiguous matches that resulted from this step, we built blocks of conserved gene order (synteny blocks) when neighboring genes in one species had one-to-one matches to neighboring genes in the other species; we used these blocks of conserved synteny to resolve additional ambiguities by preferentially keeping matches within synteny blocks (1.5). We finally searched for best unambiguous subsets (BUS), namely, subsets of genes that are locally optimal, such that all best matches of genes within the group are contained within the group and no genes outside the group have matches within the group (1.6). These ensured that the bipartite graph is maximally separable, while maintaining all possibly orthologous relationships (1.7).

When no further separation was possible, we returned the connected components of the final graph. These contain the one-to-one orthologous pairs resolved, as well as a small number of homology groups containing sets of paralogous genes whose correspondence remained ambiguous.

1.3. Automatic annotation and graph construction

In this section, we describe the construction of the weighted bipartite graph **G**, representing the gene correspondence across the species compared. We started with the genomic sequence of the species and the publicly available gene catalogue of *S. cerevisiae*, namely, the start and stop coordinates of known and predicted genes. We then predicted protein-coding genes for each newly sequenced genome. Finally, we connected across each pair of species the genes that shared amino acid sequence similarity.

The input to the algorithm is based on the complete genome for each species compared. For *S. cerevisiae*, we used the public sequence available from the Saccharomyces Genome Database (SGD) at

www.yeastgenome.org. SGD posts sixteen uninterrupted sequences, one for each chromosome. The sequence was obtained by an international sequencing consortium and published in 1996. It was completed by a clone-based sequencing approach and directed sequence finishing to close all gaps. Subsequent to the publication, updates to the original sequence have been incorporated in SGD based on resequencing of regions studied in labs around the world.

The genome sequences of *S. paradoxus*, *S. mikatae*, and *S. bayanus* were obtained at the MIT/Whitehead Institute Center for Genome Research. We used a whole-genome shotgun sequencing approach with paired-end sequence reads of 4 kb plasmid clones, with lab protocols as described at *www.genome.wi.mit.edu*. We used ~seven-fold redundant coverage; namely, every nucleotide in the genome was contained on average in at least seven different reads. The information was then assembled with the Arachne computer program (Batzoglou *et al.*, 2002; Jaffe *et al.*, 2003) into a draft sequence for each genome.

With the genome sequences at hand, we determined the set of protein-coding genes for each species. For *S. cerevisiae*, we used the public gene catalogue at SGD. It was constructed by including all predicted protein coding genes of at least 100 amino acids that do not overlap longer genes by more than 50% of their length. It was subsequently updated to include additional short genes supported by experimental evidence and to reflect changes in the underlying sequence when resequencing revealed errors. For the three newly sequenced species, we predicted all uninterrupted open reading frames (ORFs) starting with a methionine (start codon ATG) and containing at least 50 amino acids.

We then constructed the bipartite graph connecting all predicted protein coding genes that share amino acid sequence similarities across any two species. For this purpose, we first used protein BLAST (Altschul *et al.*, 1990) to find all protein hits between the two protein sets (we used WU-BLAST BlastP with parameters $W = 4$ for the hit size in amino acids, hitdist $= 60$ for the distance between two hits, and $E = 10^{-9}$ for the significance of the matches reported). Since the similarity between query protein x in one genome and target protein y in another genome is sometimes split in multiple BLAST hits, we grouped all BLAST hits between x and y into a single *match*, weighted by the average amino acid percent identity across all hits between x and y and by the total protein length aligned in BLAST hits. These matches form the edges of the bipartite graph **G**, described in the following section.

1.4. Initial pruning of suboptimal matches

Let $\mathbf{G} = (\mathbf{X} + \mathbf{Y}, \mathbf{E})$ be a weighted bipartite graph describing the similarities between two sets of genes **X** and **Y** in the two species compared (Fig. 1, top left panel). Every edge $e = (x, y)$ in **E** that connects nodes $x \in \mathbf{X}$ and $y \in \mathbf{Y}$ was weighted by the overall amino acid similarity and length in BLAST hits between genes x and y. When multiple BLAST hits connected x to y, we summed the nonoverlapping portions of these hits to obtain the total weight of the corresponding edge.

We constructed graph **D** as the directed version of **G** by replacing every undirected edge $e = (x, y)$ by two directed edges (x, y) and (y, x) with the same weight as **e** in the undirected graph (Fig. 1, top right panel). This allowed us to rank edges incident from a node and construct subsets of **D** that contain only the top matches out of every node.

We then used a relative cutoff to eliminate suboptimal matches that were significantly lower than the best match for either incident node and hence unlikely to represent orthologous relationships. We used a relative cutoff, based on the divergence relative to the best match, rather than an absolute cutoff based on overall divergence to ensure that potentially orthologous matches would be kept at any divergence and also that the algorithm would work at a range of evolutionary distances.

We defined $\mathbf{D_{80}}$ as the subset of **D** containing for every node only the outgoing edges that are at least 80% of the best outgoing edge. The cutoff was chosen based on the bimodality of relative score distributions (Fig. 2a). Matches below 80% relative divergence are likely to represent paralogous relationships ancestral to the divergence of the species and hence can be safely eliminated in this preprocessing step. These matches are significantly lower than the best match for either incident node and hence unlikely to represent orthologous relationships.

This preprocessing step resulted in a drastic reduction of the overall connectivity of the graph (Figs. 2b and 2c). The bulk of the matches, whose scores were centered around 30% amino acid identity, have been eliminated. These are likely to represent protein domain similarity between distantly related proteins of the same superfamily or proteins of similar function but whose separation well precedes the divergence of the species.

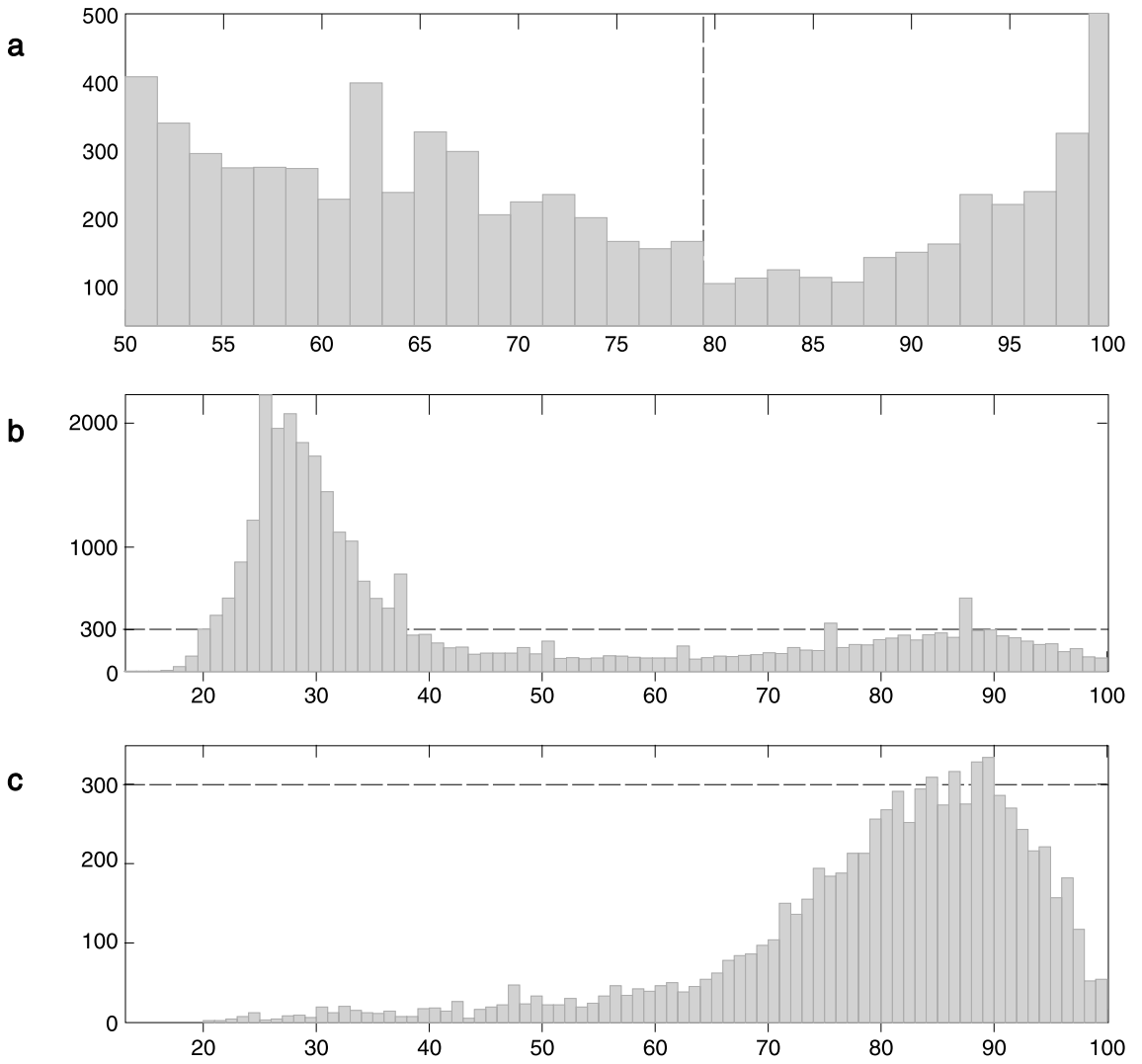


FIG. 2. Pruning of suboptimal matches. We used a relative threshold for filtering out BLAST hits that are unlikely to represent orthologous relationships. (**a**) Histogram of amino-acid percent identity of secondary matches as compared to the strongest amino-acid percent identity found amidst matches shows a bimodal distribution. Dotted line indicates relative threshold chosen (80%), separating the two modes of the distribution. (**b**) Histogram of absolute percent identity found in all matches between *S. cerevisiae* and *S. bayanus*, before applying the relative threshold. (**c**) Histogram of absolute percent identity found in matches above the relative threshold. Dotted horizontal line is indicative of scaling in the second histogram that contains fewer connections.

The distribution of the remaining matches peaks between 70% and 95% amino acid identity (Fig. 2c), which likely reflects the average evolutionary divergence of proteins at the evolutionary distance between the species compared. However, since we used a relative threshold rather than an absolute one, matches have been kept at percent identities as low as 20% and 30%. These may represent rapidly evolving proteins, possibly under positive selection with roles in speciation or environment adaptation. For example, the YBR184W gene encodes a protein longer than 500 amino acids in each species and is likely to be important in spore formation, but shows only 29% amino acid identity between *S. cerevisiae* and *S. bayanus* (and 15% across the four species). Its matches would have been lost had we used an absolute cutoff for eliminating matches, but they were kept using our relative cutoff.

In particular, this preprocessing step increased the number of proteins with one-to-one correspondence by 60% (from 2,790 to 4,648 for *S. paradoxus*; from 2,547 to 4,113 for *S. mikatae*; from 2,523 to 4,052 for *S. bayanus*). This enabled us to create a relatively dense set of unique anchors, which we next used to define blocks of conserved gene order (synteny) and resolve additional orthologs.

1.5. Blocks of conserved synteny

The initial pruning step created numerous unambiguous one-to-one matches (two-cycle directed subgraphs in the bipartite graph) between proteins that do not have other closely related paralogs. We used these to construct blocks of conserved synteny based on the physical distance between consecutive matched genes and preferentially kept edges that connect additional genes within the block of conserved gene order (Fig. 3). Edges connecting these genes to genes outside the blocks were then ignored, as unlikely to represent orthologous relationships. Without imposing an ordering on the scaffolds or the chromosomes, we associated every gene x with a fixed position (scaffold, start) within the assembly and every gene y with a fixed position (chromosome, start) within *S. cerevisiae*. If three one-to-one unambiguous matches (x1, y1), (x2, y2), and (x4, y4) were such that x1, x2, x4 were physically near each other in one genome and y1, y2, y4 were physically near each other in the other genome, we constructed a synteny block $B = (\{x1, x2, x4\}, \{y1, y2, y4\})$. Thereafter, for a gene x3 that was proximal to {x1, x2, x4}, if an outgoing edge (x3, y3) existed such that y3 was proximal to {y1, y2, y4}, we ignored other outgoing edges (x3, y') if y' was not proximal to {y1, y2, y4}.

We considered only synteny blocks that had a minimum of three genes before using them for resolving ambiguities, to prevent being misled by sequence similarity of isolated genes. We set the maximum

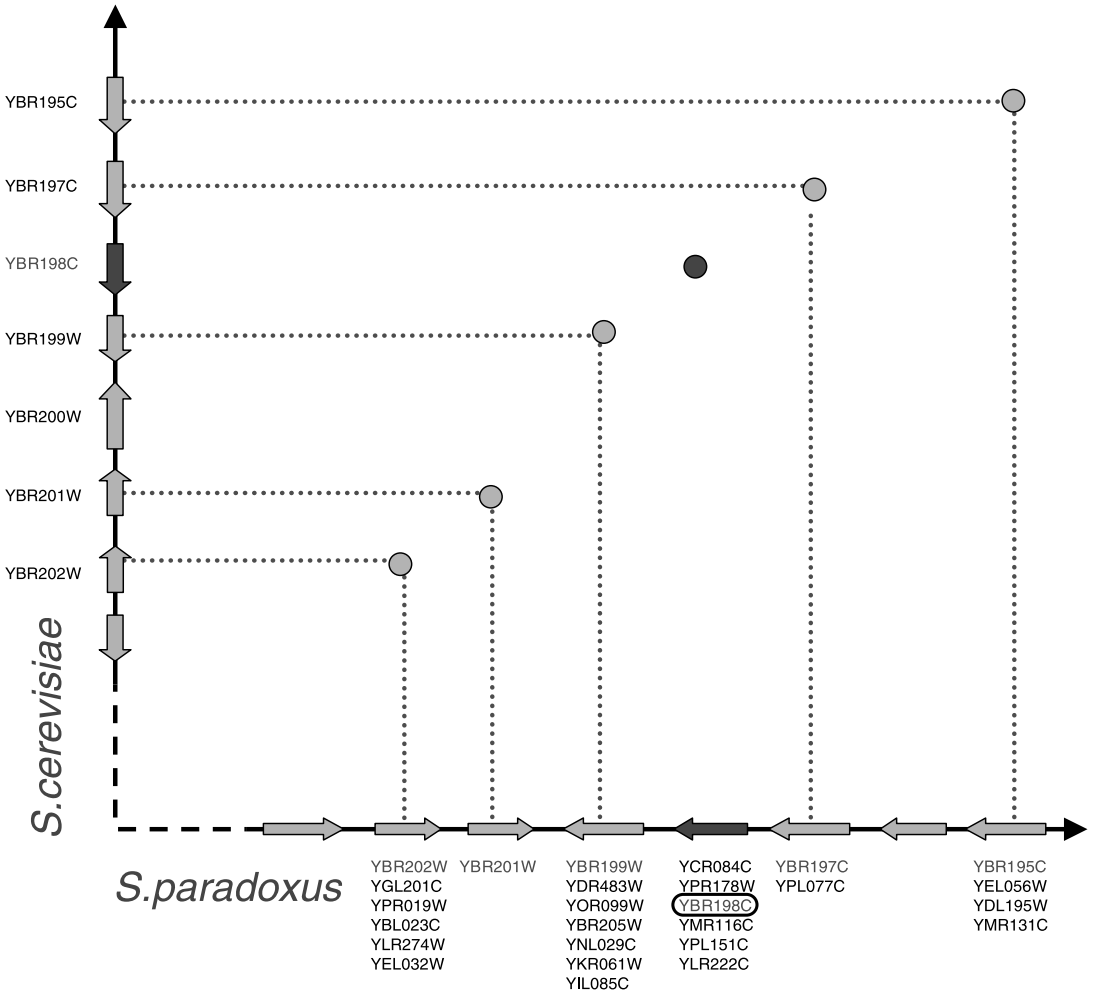


FIG. 3. Using synteny information. Within blocks of conserved gene order (synteny), we preferentially keep those matches that conserve gene order. Annotated *S. cerevisiae* ORFs and standard names are shown along the y axis. Predicted *S. paradoxus* ORFs and their matches are shown along the x axis, listed by decreasing scores. Unambiguous matches are shown as grey dots and ambiguous matches are shown in black. In this example, YBR198C is selected as the orthologous match based on synteny information, and matches to YCR084C and YPR178W are treated as paralogous.

distance ***d*** for considering two neighboring genes as proximal to 20 kb. This parameter was set to match the estimated density of unique syntenic anchors. Since *S. cerevisiae* genes are on average 2 kb apart, and roughly two thirds of the genes have unique correspondence after the pruning step (roughly one third before the pruning step, see Section 1.4), we should expect a sufficient number of anchors within such an interval. Further, transposable elements in yeast are relatively short, and this cutoff should tolerate even two tandem transposition events without interrupting a synteny block.

If many genomic rearrangements have occurred since the separation of the species, or if the scaffolds of the assembly are short, the syntenic segments will be shorter, and setting ***d*** to larger values might hurt the performance. On the other hand, if the number of unambiguous genes is too small at the beginning of this step, the genes used as anchors will be sparse, and no synteny blocks will be possible for small values of ***d***. We found this step to play a greater role as evolutionary distances between the species compared became larger, and sequence similarity was no longer sufficient to resolve all the ambiguities.

This step helped resolve genes that duplicated prior to the divergence of the species compared. Many of these paralogs have preserved high sequence identity to each other and thus were not resolved by the initial pruning step that considered only sequence identity. In particular, more than 300 pairs of paralogs were grouped in two-to-two homology groups before this step and were resolved based on synteny information.

1.6. Best unambiguous subsets

We finally separated out subgraphs that were connected to the remaining edges in the graph by solely nonmaximal edges. These subgraphs are such that the best match of any node within the subset is contained within the subset, and no node outside the subset has its best match within the subset. These two properties ensure that the subsets are both best and unambiguous.

We defined a *best unambiguous subset* (BUS) of the nodes of $\mathbf{X} \cup \mathbf{S}$ to be a subset $\mathbf{S}$ of genes, such that $\forall x$: $x \in \mathbf{S} \Leftrightarrow \text{best}(x) \subseteq \mathbf{S}$, where best(x) is the set of nodes incident to the maximum weight edges from x. To construct a BUS, we first constructed $\mathbf{D_{100}}$, following the notation above, namely, the subset of $\mathbf{D}$ that contains only best matches out of a node (weights within 100% of the best edge). Note that multiple best matches were possible based on our definition. In this graph-theoretic framework, a BUS is defined as a connected component of the undirected version of $\mathbf{D_{100}}$ (Fig. 1, bottom panels).

Intuitively, a BUS is a set of genes that starts as a best bidirectional hit (BBH) and increases in size to include all best edges incident to any gene in the set. Starting with the subset of nodes in any cycle in $\mathbf{D_{100}}$, we augmented the subset by following forward and reverse best edges, namely, including additional nodes if their best match was within the subset or if they were the best match of a node in the subset. This ensured that separating a subset did not leave any node orphaned and did not remove the strictly best match of any node. When no additional nodes needed to be included to make the component separable, the BUS condition was met and the connected component was obtained.

Figure 4 shows a toy example of a similarity matrix. Genes A, B, and C in one genome are connected in a complete bipartite graph to genes 1, 2, and 3 in another genome (ignoring for now synteny information). The sequence similarity between each pair is given in the matrix and corresponds to the edge weight connecting the two genes in the bipartite graph. The set (A,1,2) forms a BUS, since the best matches of A, 1, and 2 are all within the set and none of them represents the best match of a gene outside the set. Hence, the edges connecting (A,1,2) can be isolated as a subgraph without removing any orthologous relationships, and edges (B,1), (B,2), (C,1), (C,2), (A,3) can be ignored as nonorthologous. Similarly, (B,C,3) forms a BUS. The resulting bipartite graph is shown.

This part of the algorithm allowed us to resolve the remaining orthologs, mostly due to subtelomeric gene family expansions, small duplications, and other genes that did not benefit from synteny information. In genomes with many rearrangements or assemblies with low sequence coverage, which do not allow long-range synteny to be established, this part of the algorithm will play a crucial role.

1.7. Performance of the algorithm

We applied this algorithm to automatically annotate the assemblies of the three species of yeast. Our Python implementation terminated within minutes for any of the pairwise comparisons. We successfully resolved the graph of sequence similarities between the four species and found important biological implications in the resulting graph structure.

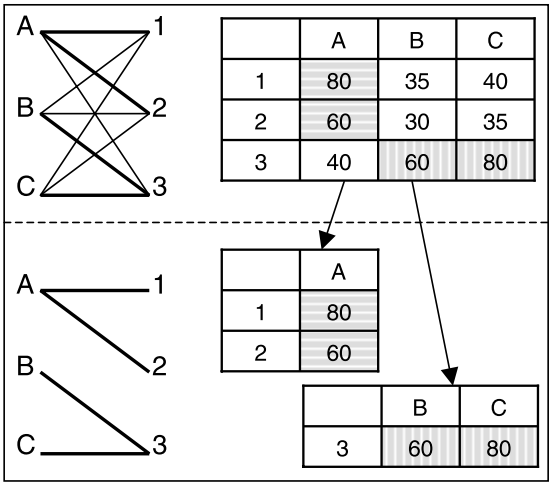


FIG. 4. Example of best unambiguous subsets (BUS). A BUS is a connected component of the graph constructed using only best matches, both forward and reverse. It defines a set of genes that can be isolated from the bipartite graph while preserving all potentially orthologous matches. Given the pairwise similarity matrix in this example, and no additional knowledge of gene order, the best unambiguous subsets of this graph are (A,1,2) and (B,C,3).

Figure 5 illustrates the performance of the algorithm for the 6,235 annotated ORFs in *S. cerevisiae* and all predicted ORFs in *S. paradoxus*. The graph was originally very dense (panel A), the vast majority of edges representing nonorthologous matches, mostly due to protein domain similarities, ancient duplications that precede the time of the common ancestor of the species compared, and transposable elements. After applying the initial pruning step, many of the spurious matches were eliminated (panel B), but a large amount of gene redundancy remained. We used the one-to-one matches to build blocks of conserved gene order and used these to resolve additional matches using the BUS algorithm (panel C). The unambiguous one-to-one matches found were mostly syntenic for *S. paradoxus* (panel D), increasing our confidence that we are comparing orthologous regions. Similar plots for the other two species revealed a small number of rearrangements.

More than 90% of genes have clear one-to-one orthologous matches in each species, providing a dense set of landmarks (average spacing ∼2 kb) to construct multiple sequence alignments covering essentially the entire genome. Amidst unambiguous matches, not surprisingly, transposon proteins formed the largest homology groups. The remaining matches were isolated in small subgraphs. These represent gene family expansions that are often found in rapidly recombining regions near the ends of chromosomes and in genes involved in environmental adaptation, such as sugar transport and cell surface adhesion (Kellis *et al.*, 2003).

We have additionally experimented running only the BUS algorithm without the original pruning and synteny steps. The algorithm resolved more than 80% of ambiguities, and the remaining matches corresponded to duplicated ribosomal proteins and other gene pairs that have remained virtually unchanged since their duplication. These results imply that amino acid similarity may provide sufficient information to resolve at least a large part of gene ambiguities at the evolutionary distances compared. The algorithm was slower, due to the large initial connectivity of the graph, but a large overall separation was obtained. Figure 6 compares the dot plot of *S. bayanus* and *S. cerevisiae* with and without the use of synteny. Every point represents a match, the x-coordinate denoting the position in the *S. bayanus* assembly, and the y-coordinate denoting the position in the *S. cerevisiae* genome, with all chromosomes put end-to-end. Lighter dots represent homology groups containing more than 15 genes (typically transposable elements), and circles represent smaller homology groups (rapidly changing protein families that are often found near the ends of chromosomes). The darker dots represent unambiguous one-to-one matches, and the boxes represent synteny blocks.

This algorithm has also been applied to species at much larger evolutionary distances, with very successful results (Kellis and Lander, manuscript in preparation). Despite hundreds of rearrangements and duplicated genes separating *S. cerevisiae* and *K. waltii*, it successfully uncovered the correct gene correspondence between the two species that are more than 100 million years apart.

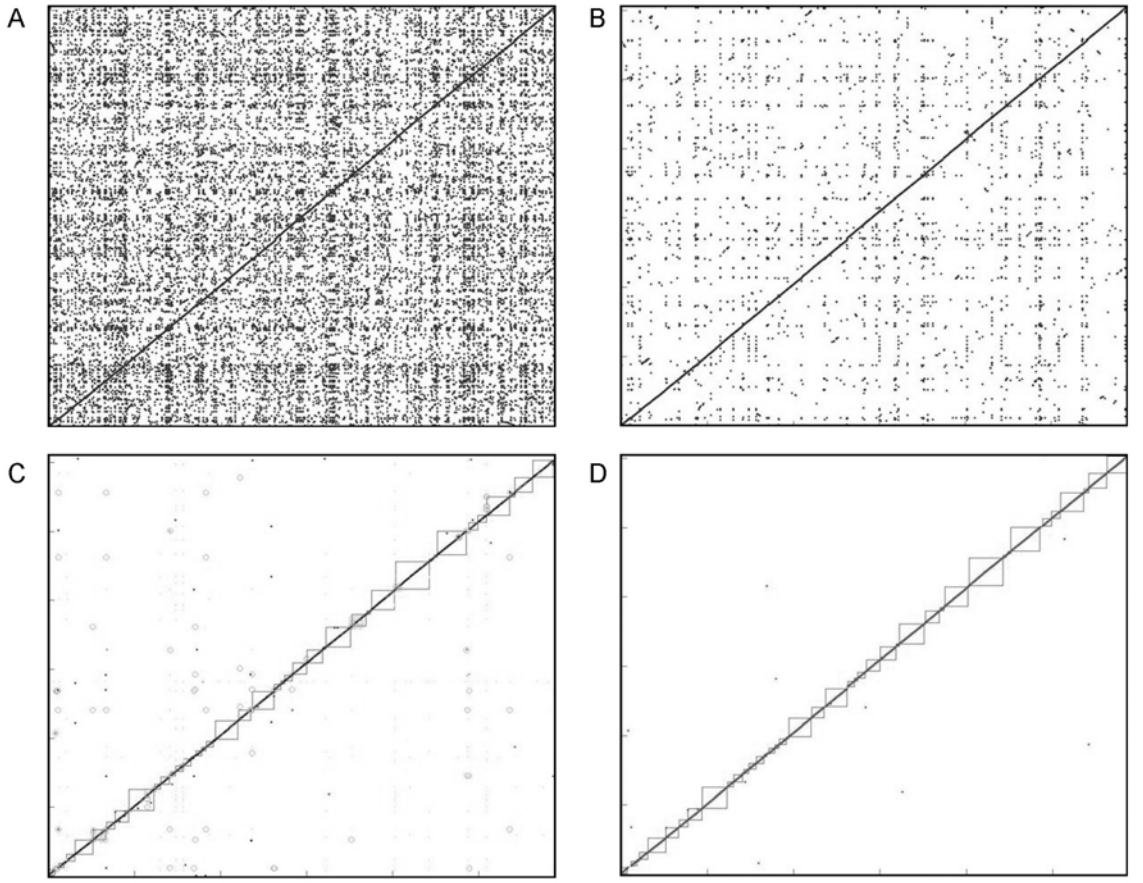


FIG. 5. Performance of the algorithm. Dot-plot representation of the bipartite graph. The 16 chromosomes of *S. cerevisiae* are stacked end to end along the y-axis and the scaffolds of *S. paradoxus* are shown along the x-axis. Every point (x, y) represents an edge between a *S. paradoxus* gene at position y and a *S. cerevisiae* gene at position x. (**A**) Full bipartite graph. (**B**) Graph resulting from initial pruning step. (**C**) Graph resulting from the use of BUS and synteny information. (**D**) Unambiguous one-to-one matches fall largely within syntenic segments, ensuring that we are comparing orthologous regions.

Additionally, the algorithm worked well with unfinished genomes, without requiring directed sequencing for closing gaps. By working with sets of genes instead of one-to-one matches, this algorithm correctly grouped in the same orthologous set all portions of genes that are split in two or multiple contigs interrupted by sequence gaps, whereas a best bidirectional hit would have instead matched only the longest portion and left part of a gene unmatched. Finally, since synteny blocks are only built on one-to-one unambiguous matches, the algorithm was robust to sequence contamination, when reads from other species were erroneously incorporated in the assembly. Foreign reads, even from more closely related species, were all marked as nonorthologous, since all features in these reads were never unique and hence not used in synteny blocks. On the contrary, the same features in genuine contigs were surrounded by unique hits that helped resolve them as orthologous and mark the contaminating sequences as paralogous. Overall, the algorithm provided a good solution to determining genome correspondence, worked well at a range of evolutionary distances, and was robust to sequencing artifacts of unfinished genomes.

1.8. Conclusion

We unambiguously resolved the one-to-one correspondence of more than 90% of *S. cerevisiae* genes. This provided us with a unique dataset to align and compare the evolutionary pressure of nearly every region in a complete eukaryotic genome across four closely related relatives. By ensuring that the regions

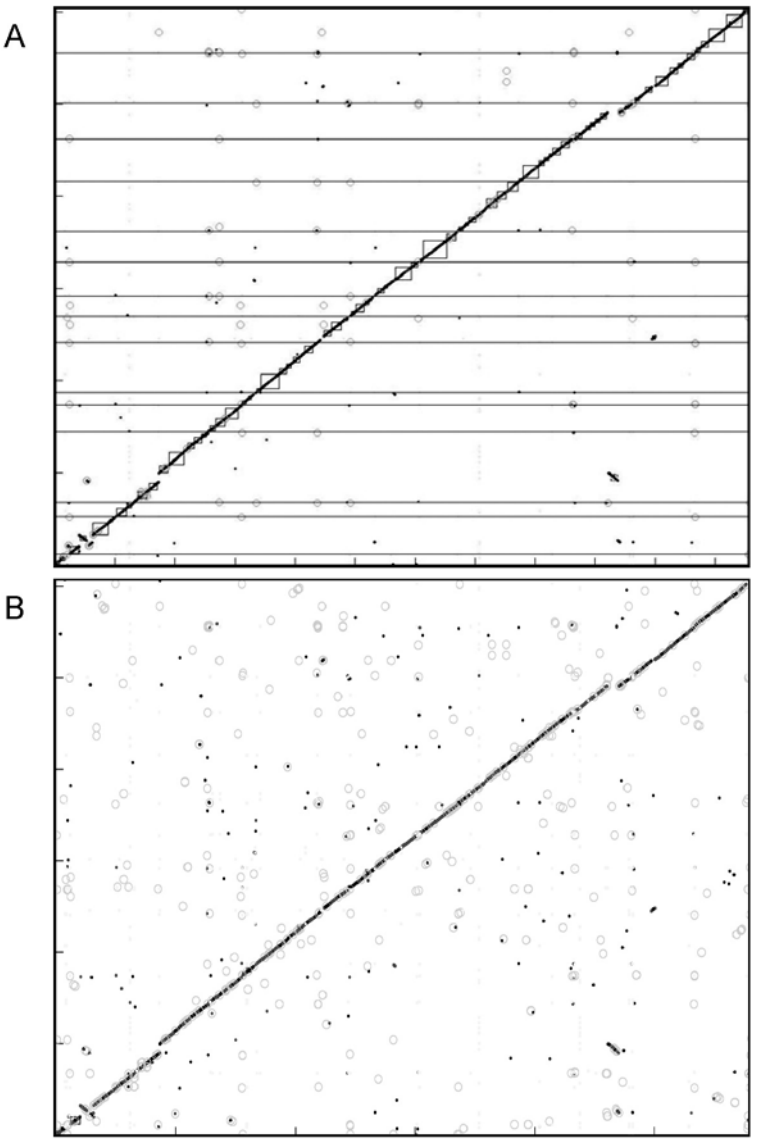


FIG. 6. The effect of using synteny. Dot-plot of gene correspondence between *S. cerevisiae* (y-axis) and *S. bayanus* (x-axis) with and without using synteny information. Blocks of conserved gene order (blue squares) help resolve additional ambiguities. Ambiguities remaining when the use of synteny is omitted are largely due to pairs of anciently duplicated genes that have not diverged significantly from each other.

compared are orthologous, we can make assumptions about the rate of change of different regions and apply statistical models to interpret the significance of strong or weak conservation in discovering biological signals.

2. GENE IDENTIFICATION

The genome of a species encodes genes and other functional elements, interspersed with nonfunctional nucleotides in a single uninterrupted string of DNA. Recognizing protein-coding genes typically relies on finding stretches of nucleotides free of stop codons (called open reading frames, or ORFs) that are too long to have likely occurred by chance. Since stop codons occur at a frequency of roughly 1 in 20 in random sequence, ORFs of at least 60 amino acids will occur frequently by chance (5% under a simple Poisson

model), and even ORFs of 150 amino acids will appear by chance in a large genome (0.05%). This poses a huge challenge for higher eukaryotes in which genes are typically broken into many, small exons (on average 125 nucleotides long for internal exons in mammals (International Human Genome Sequencing Consortium, 2001)).

The basic problem is distinguishing *real genes*—those ORFs encoding a translated protein product—from *spurious ORFs*—the remaining ORFs whose presence is simply due to chance. The current public catalogue of yeast genes lists 6,062 predicted ORFs that could theoretically encode proteins of at least 100 amino acids. Only two-thirds of these have been experimentally validated (*known*), and the remaining ~2000 ORFs are currently annotated as *hypothetical*. The total number of real protein-coding genes has been a subject of considerable debate, with estimates ranging from 4,800 to 6,400 genes (Harrison *et al.*, 2002; Kowalczuk *et al.*, 1999; Velculescu *et al.*, 1997). In mammalian genomes, estimates have ranged from 28,000 to more than 120,000 genes (Dunham, 2000; International Mouse Genome Consortium, 2002; Liang *et al.*, 2000).

In this section, we use the comparative information to recognize real genes based on their patterns of nucleotide conservation across evolutionary time. With the availability of genomewide alignments across the four species, we first examined the different ways by which sequences change in known genes and in intergenic regions. The alignments of known genes revealed a clear pressure to preserve the reading frame of protein translation. We constructed a computational test to evaluate reading frame conservation (RFC) and showed that the method has high sensitivity and specificity in identifying protein-coding genes. We used the RFC test to revisit the annotation of yeast, showed that more than 500 previously annotated ORFs are not meaningful, and discovered 43 novel ORFs that were previously overlooked. We additionally refined the gene structure of hundreds of genes, including translation start, stop, and exon boundaries. Overall, the suggested changes affect nearly 15% of yeast genes.

2.1. Different conservation of genes and intergenic regions

We constructed genomewide nucleotide alignments across the four species, then examined the different properties of nucleotide conservation in genes and intergenic regions. Based on these properties, we constructed a conservation-based test to identify biologically meaningful protein-coding genes.

We used the one-to-one orthologous anchors to construct a nucleotide-level alignment of the genomes. The strong conservation of local gene order and spacing allowed us to construct genome-wide multiple alignments. We aligned each gene together with its flanking intergenic regions using CLUSTALW (Thompson *et al.*, 1994) for the multiple alignments across the four species. When sequence gaps were present in one or more species, we constructed the alignment in multiple steps. We first aligned the gapless species creating a base alignment. Then we aligned each ungapped portion of a species containing a gap onto the base alignment, and constructed a consensus for that species based on the individually aligned portions. We marked sequence gaps between contigs by a dot and conflicts in independently aligned overlapping portions by N. Finally, we constructed a multiple alignment of the four species by merging the piecewise alignments, adding gaps where appropriate such that the sequences in common between these alignments contain identical sets of gaps.

With sequence alignments at millions of positions across the four species, it is possible to obtain a precise estimate of the rate of evolutionary change, including substitutions and insertion–deletions (indels), in the tree connecting the species. We counted transitions, transversions, insertions, and deletions within these alignments and used these to estimate the rate of evolutionary change between the species. We counted the rate of synonymous and nonsynonymous substitutions for every protein coding gene to find evidence of positive selection. The detailed results will be described in a subsequent publication.

We specifically compared the properties of nucleotide change in intergenic and genic (protein-coding) regions (Fig. 7). We found radically different types of conservation. Intergenic regions typically showed short stretches between 8 and 10 bases of near-perfect conservation, surrounded by nonconserved bases, rich in isolated gaps. Protein-coding genes, on the other hand, were much more uniform in their conservation and typically differed in the largely degenerate third-codon position. The proportion of sites corresponding to a different nucleotide in at least one of the three species is 58% in intergenic regions but only 30% in genic regions—a difference of ~two-fold. The difference becomes much greater when one considers the gapped positions in alignments, representing insertion and deletion events (indels). The proportion of indels

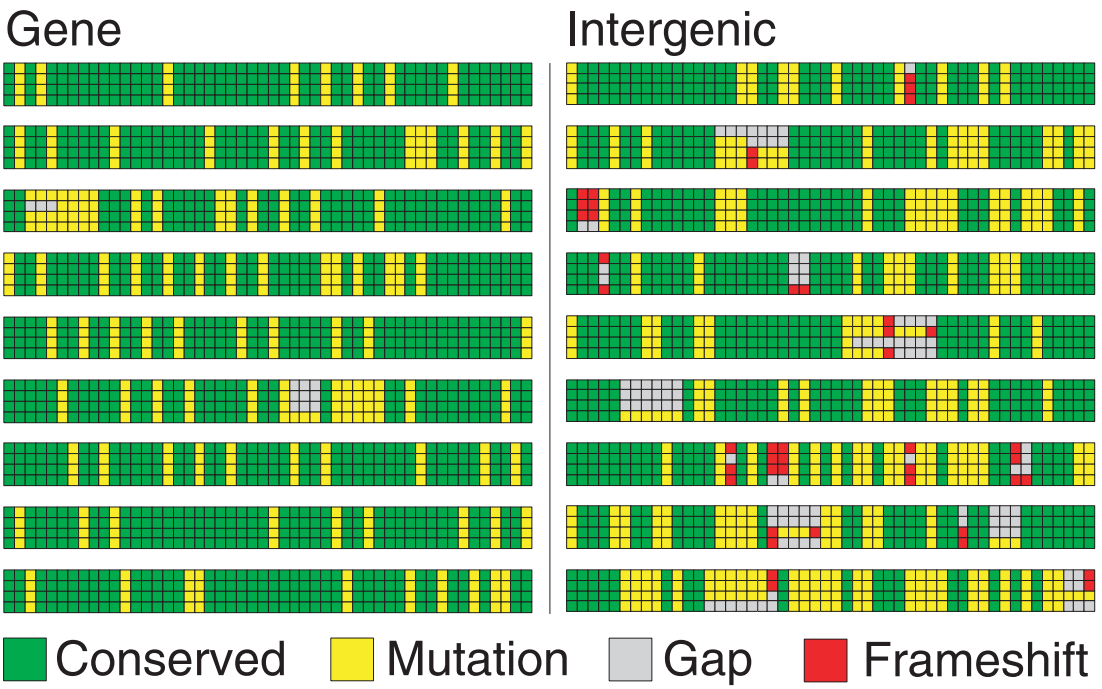


FIG. 7. Patterns of change in genes and intergenic regions. Schematic representation of multiple sequence alignments in ORF YMR017W (**left**) and neighboring intergenic region (**right**). Aligned nucleotides across the four species are shown as stacked squares, colored by their conservation: green for conserved positions, yellow otherwise. Alignment gaps (representing insertions and deletions) are shown in white. Frame-shifting insertions (length not a multiple of 3) are shown in red.

is 14% in intergenic regions, but only 1.3% in genic regions. The contrast is even sharper for indels whose length is not a multiple of three. These would disrupt the reading frame of a functional protein-coding gene and are detrimental when they occur in real genes, unless they are compensated by a nearby indel that restores the reading frame. Frame-shifting gaps are found in 10.2% of aligned positions in intergenic regions, but only in 0.14% of aligned positions in genic regions, a 75-fold strong separation. We used these alignment properties to recognize real genes.

2.2. Reading frame conservation test

We developed a reading frame conservation (RFC) test to classify each ORF in *S. cerevisiae* as biologically meaningful or not, based on the proportion of the ORF over which the reading frame is locally conserved in each of the other three species. Each species with an orthologous alignment cast a vote for accepting or rejecting the ORF, and the votes were tallied to reach a decision for that ORF.

We evaluated the percent of nucleotides that are in the same frame within overlapping windows of the alignment. For every such window, we labeled each nucleotide of the first sequence by its position within a codon as 1, 2, or 3 in order, starting at codon offset 1 (Fig. 8). We similarly labeled the nucleotides of the second sequence, but once for every start offset (1, 2, or 3). We then counted the percentage of gapless positions in the alignment that contained the same label in both aligned species and selected the maximum percentage found in each of the three offsets of the second sequence. Since only one frame will be translated, that maximum percentage gives an upper bound on the pressure to preserve frame conservation on a given genomic segment. The final RFC value for the ORF was calculated by averaging the percentages obtained at overlapping windows of 100 nucleotides starting every 50 nucleotides. For overlapping ORFs in the *S. cerevisiae* genome (948 cases), the RFC was calculated only for the portion unique to each overlapping ORF. For spliced genes (240 cases), the RFC was calculated only on the largest exon.

We found that the distribution of frame conservation within each species is bimodal, and we chose a simple cutoff for each species based on the separation point in the bimodal distribution of RFC scores:

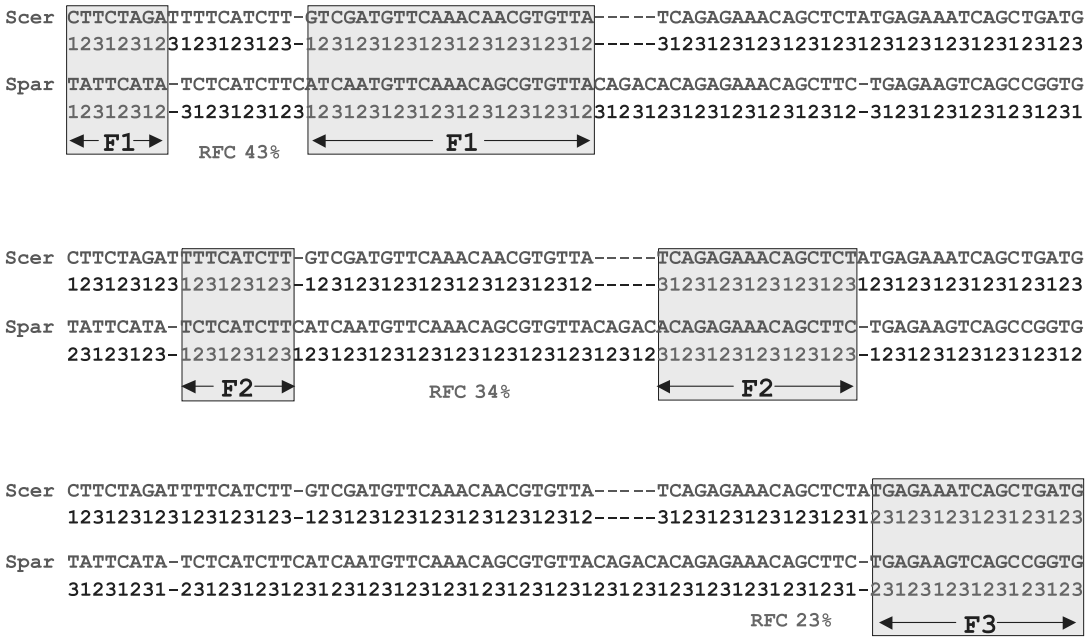


FIG. 8. Reading frame conservation (RFC). Sample alignment between *S. cerevisiae* and *S. paradoxus* and frame numbering for the three possible frame numbering offsets. For each offset, matching frame indices are shown in red, and the portion of the alignment with agreeing reading frame is boxed. Gaps in the alignment change the correspondence of reading frame from frame F1 to frame F2 and so on. The RFC score for this window is given as the largest percentage of nucleotides in the same frame, here F1. The overall RFC score for an ORF is calculated as the average of the RFC scores in overlapping windows of 100 nucleotides.

85% for *S. paradoxus*, 80% for *S. mikatae*, and 75% for *S. bayanus* (Fig. 9). If the RFC of the best hit was above the cutoff, a species voted for keeping the ORF tested. If the RFC was below the cutoff and the hit was trusted as orthologous, the species voted for rejecting the tested ORF. Finally, if no orthologous hit could be found due to coverage or rearrangements, a species abstained from voting. Since the presence of a genomic segment in each genome was largely independent (due to independent sequencing gaps and gene gain or loss events), we treated the information provided by each species as an independent vote towards keeping or rejecting an ORF, when information was available, or as the absence of a vote, when the information available was insufficient to reach a conclusion.

For every ORF, we tallied the votes from the different species by calculating a score between -3 and $+3$ as the number of species that accepted it ($+1$) minus the number of species that rejected it (-1). Since the RFC cutoffs had previously accounted for differences in species divergence, we treated the votes equally across the three species, as independent assessments of the validity of a given ORF. We kept all ORFs with a score of 1 or greater and rejected all ORFs with a score of -1 or smaller. We manually inspected the remaining ORFs for in-frame stop codons and overall protein conservation (Kellis *et al.*, 2003).

2.3. RFC test shows high sensitivity and specificity

To evaluate the power of the method to distinguish real genes from other regions, we benchmarked it on 3,966 known genes and 340 intergenic regions of similar lengths. We found that the test has at least 99.6% sensitivity and 95% specificity (Table 1). We then applied it systematically to all hypothetical ORFs to distinguish those that represent real genes. We found that more than 500 previously annotated ORFs are not real.

To evaluate the sensitivity of the test to accept real genes, we applied it to 3,966 annotated ORFs with associated gene names. These have been studied and named in at least one peer-reviewed publication and are likely to be represent real genes. Only 15 of these (0.38%) were rejected (KRE20, KRE21, KRE23, KRE24, VPS61, VPS65, VPS69, BUD19, FYV1, FYV2, FYV12, API2, AUA1, ICS3, UTR5, YIM2). We

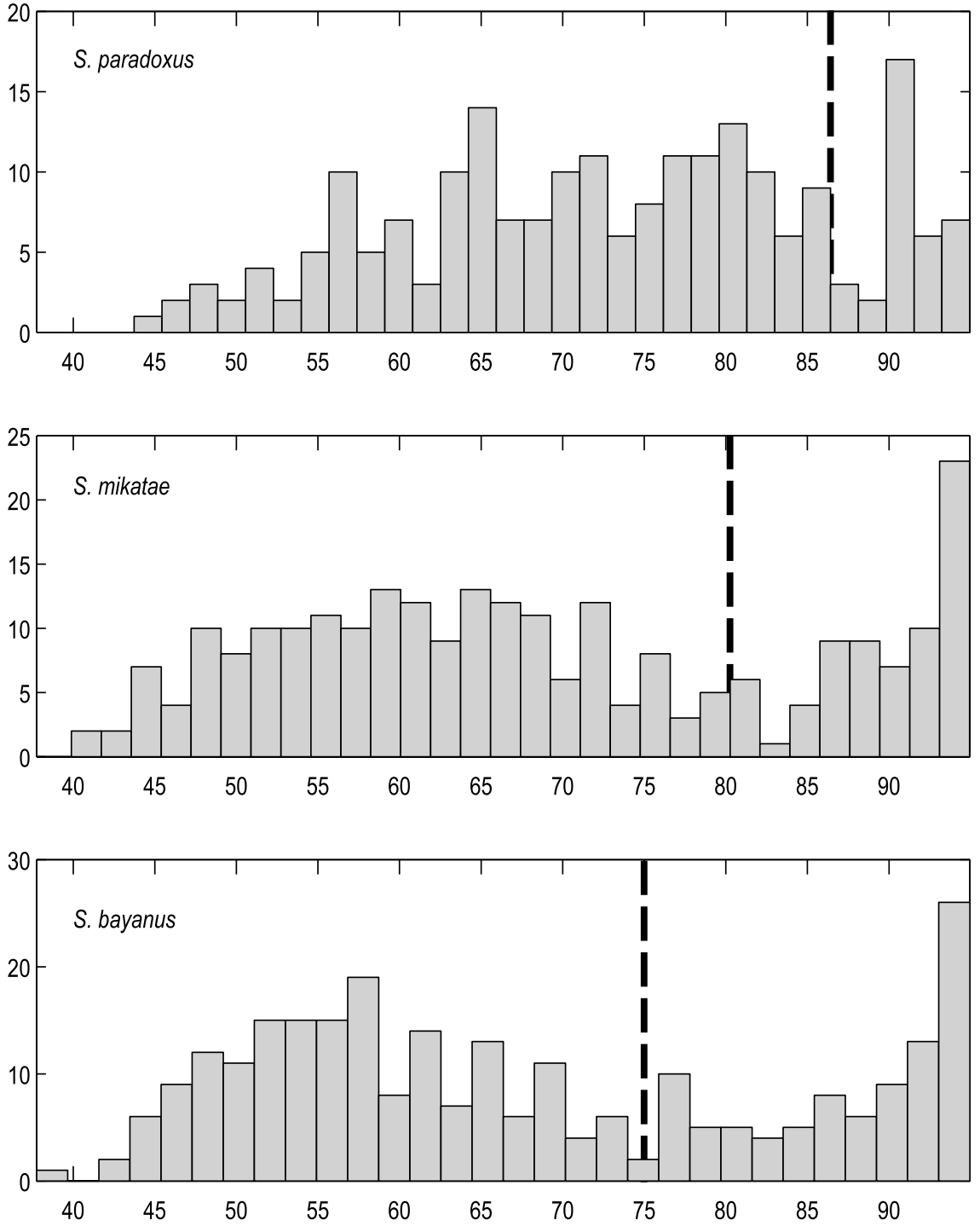


FIG. 9. Bimodality of RFC test for hypothetical ORFs. Distribution of RFC scores for hypothetical ORFs in each of the species. The bimodality of the score distribution allows us to select a cutoff for each species above which an ORF is accepted as biologically meaningful and below which it is rejected and marked dubious.

TABLE 1. PERFORMANCE OF THE RFC TEST[a]

RFC test applied to:	*Accept*	*Reject*
3966 named genes	99.6%	0.4%
340 intergenic regions	1%	99%
2056 hypothetical genes	1528	528

[a]The RFC test showed strong sensitivity and specificity, correctly accepting 99.6% of experimentally verified genes (named genes) and correctly rejecting 99% of intergenic regions tested. We further applied this test to all hypothetical genes and showed that more than 500 currently annotated genes are not real.

inspected these manually and concluded that all were indeed likely to be spurious. Most lack experimental evidence. For the remainder, reported phenotypes associated with deletion of the ORF seem likely to be explained by the fact that the ORF overlaps the promoters of other known genes.

To investigate the specificity of the approach to reject spurious ORFs, we also applied it to a set of control sequences consisting of 340 intergenic sequences in *S. cerevisiae* with lengths similar to currently annotated ORFs (Table 1). We found that 96% of intergenic regions were rejected as having conservation properties incompatible with a biologically meaningful ORF, showing that the test has high sensitivity. Of the remaining 4% that were not rejected, close inspection shows that three-quarters appear to contain portions of real protein-coding genes. Some define short ORFs with conserved start and stop codons in all four species, and others extend *S. cerevisiae* ORFs in the 5′- or 3′-direction in each of the other three species. We conclude that at most 1% of true intergenic regions failed to be rejected by the RFC test.

The conservation-based gene identification algorithm we proposed has thus high sensitivity and specificity for annotated ORFs. We next applied it systematically to revisit the gene annotation of *S. cerevisiae*.

2.4. Results: Hundreds of previously annotated genes are not real

We first used the RFC test to systematically classify all annotated ORFs as biologically meaningful (transcribed and translated into protein) or spurious (occurring by chance). Our analysis resulted in a drastic reduction of the yeast gene count, rejecting nearly 500 ORFs. The rejected ORFs tend to be small (median = 111 aa, with 93% = 150 aa) and thus are likely to be occurring by chance. Additionally, they show atypical codon usage (mean CAI = 0.105, with 65% having CAI < 0.11), providing additional support that they may be occurring by chance (Dujon *et al.*, 1994; Goffeau *et al.*, 1996; Sharp and Li, 1987). The details of the gene analysis are described in Kellis *et al.* (2003), and a revised gene catalogue will be described in a subsequent publication in collaboration with SGD and other yeast researchers.

We further examined the power of the RFC test to discover novel short genes. The presence of frame-shifting indels provided strong evidence that a predicted ORF is spuriously occurring. However, the absence of frame-shifting indels in a short ORF may be due to lack of divergence time. To test the validity of our RFC-based predictions for small *S. cerevisiae* ORFs, we additionally examined the presence or absence of in-frame stop codons in the other species. When a small ORF in *S. cerevisiae* showed a strong overall frame conservation, we measured the length of the longest ORF in the same orientation in each orthologous locus. We measured the percentage of the *S. cerevisiae* length that was open in each species (no stop codons) and defined OPEN to be the minimum of the three percentages across the three additional species. When the reading frame was open in each of the other species, the lengths found were identical to that of *S. cerevisiae*, and OPEN was 100%. When OPEN was below 80%, we concluded that stop codons appeared in the orthologous sequence and therefore that the RFC test falsely accepted a segment that did not correspond to a true gene.

OPEN and RFC values provide two independent measurements of the pressure to preserve protein coding potential in a given genomic segment. Their comparison was instrumental in benchmarking the power of the RFC test for small ORFs.

We first observed the distribution of OPEN for ORFs with high RFC score (above 90%) and for ORFs with low RFC score (between 50% and 80%). For *S. cerevisiae* ORFs between 50 and 100 amino acids, selecting for high RFC automatically selected for high OPEN (Fig. 10a), and we inferred that the test has

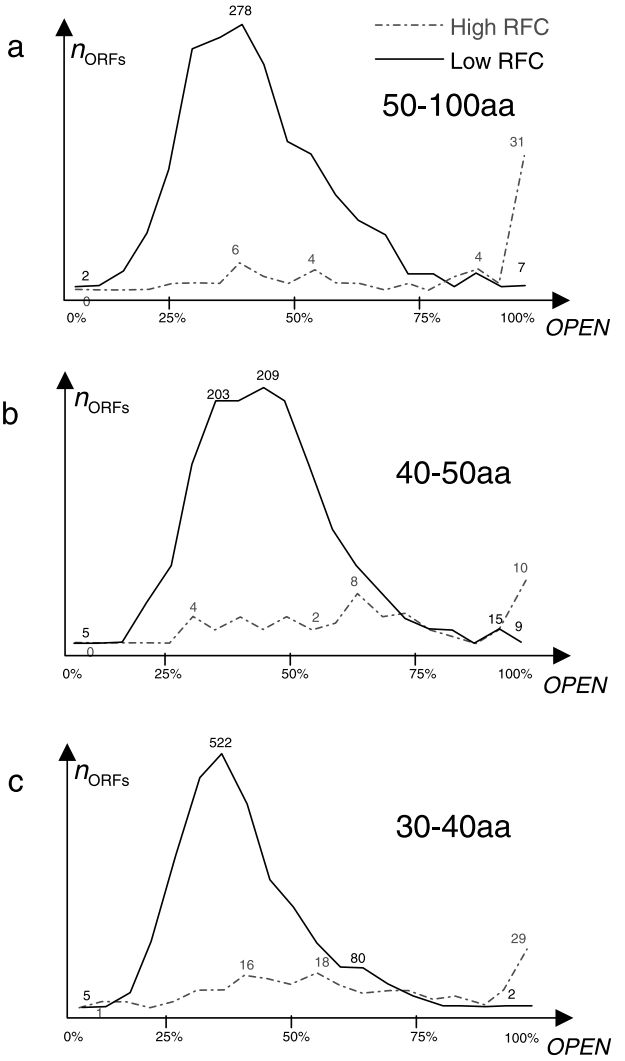


FIG. 10. Power of RFC test to identify small ORFs. ORFs of decreasing lengths are tested using the RFC test. To evaluate the validity of the tested ORFs, we measure *OPEN*, namely the minimum percentage of the segment that is free of stop codons across the four species. The histogram of *OPEN* values is shown for ORFs with high RFC scores (dotted red lines) and low RFC scores (solid black lines). (**a**) For lengths between 50 and 100 amino acids, ORFs with high RFC scores consistently show high *OPEN* scores and conversely high *OPEN* scores consistently correspond to ORFs with high RFC scores. Hence, the test has high sensitivity and specificity for that length distribution. (**b, c**) For lengths smaller than 50 amino acids, high *OPEN* scores typically come from high RFC.51 scores, revealing high sensitivity. However, high RFC scores do not always show high *OPEN* scores, and we conclude that the specificity of the RFC test is limited for ORFs of this size.

high specificity. We reported 43 novel genes at this size range. For ORFs between 30 and 50 amino acids, however, only a small portion of the ORFs with high RFC show a high OPEN (Figs. 10b and 10c), and we concluded that the lack of indels within the small interval considered is not due to selective pressure, but instead to lack of evolutionary distance between the species aligned. In all cases, however, ORFs with high values of OPEN were selected by the RFC test.

We concluded that the RFC test has high sensitivity in identifying real protein-coding genes, recognizing their pressure to conserve the reading frame of amino acid translation. Additionally, for ORFs longer than 50 amino acids, the test showed high specificity in rejecting spurious ORFs. However, the specificity of our method was limited for small ORFs. Smaller regions may indeed show lack of indels due to chance, and hence a high reading frame conservation score may not be instructive. Additional methods should be

developed to reliably discover small protein-coding genes with high sensitivity and high specificity. Such methods may rely on Ka/Ks-like (Hurst, 2002) metrics of amino acid conservation in addition to additional properties of the nucleotide conservation.

2.5. Refining gene structure

Comparative genome analysis not only improves the recognition of true ORFs, it also yields a much more accurate definition of gene structure—including translation start, translation stop, and intron boundaries. We used the comparative data to refine the boundaries of true genes and identify possible sequencing errors.

Previous annotation of *S. cerevisiae* has defined the start of translation as the first in-frame ATG codon. However, the actual start of translation could lie 3′ to this point, and the earlier in-frame ATG may be due to chance. Alternatively, if sequencing errors or mutations have obscured an earlier in-frame ATG codon, the true translation start could lie 5′ to this point. Similarly, the annotated stop codon could be erroneously annotated, due to sequencing errors. Identifying the correct gene boundaries is important for many reasons, both experimental (for example to construct gene probes), as well as computational (for example to search for regulatory motifs using the correct promoter region boundaries). Additionally, ORF extensions may reveal functionally important protein domains and lead to a candidate function for previously uncharacterized genes. For example, our reannotation of CWH36 revealed a doubly spliced gene on the opposite strand than previously annotated that showed sequence similarity to a vacuolar ATPase protein in species as distant as insects, flies, cow, and rat.

We examined the multiple alignment of unambiguous ORFs to identify discrepancies in the predicted start and stop codons across the four species. We searched for the first in-frame ATG in each species and compared its position to that of the annotated ATG in *S. cerevisiae*. If the *S. cerevisiae* start was not conserved and a subsequent in-frame ATG was conserved in all the species, then we automatically suggested a changed translation start. Similarly, we suggested changes in stop codons when a common stop in all other species disagreed with the *S. cerevisiae* annotation. We manually inspected the alignments to confirm that the suggested start and stop boundary changes agreed with conservation boundaries. We also identified merges of consecutive *S. cerevisiae* ORFs, when they unambiguously matched a single ORF in at least one other species, and when their lengths added up to the length of the matching ORF.

We identified 210 cases in which the previously annotated translation start in *S. cerevisiae* does not correspond to the first in-frame start codon in all three other species and 330 cases in which the presumed translational stop codon in *S. cerevisiae* does not correspond to the first in-frame stop codon in at least two of the three species (Kellis *et al.*, 2003). In the vast majority of start codon discrepancies, a different ATG was conserved across the four species, and we proposed a change in the true translation start. However, for stop codon discrepancies, a common stop was found in only ∼25% of cases, and we believe the remaining 75% represent true differences in the location of the translational stop across the species. Stop codons thus appeared to show more evolutionary variability in position than start codons.

We also developed methods for the automatic detection of frame-shifting sequencing errors within protein-coding genes. We applied the RFC test in a window-based fashion, first determining windows of strong frame conservation, then extending these windows until a change of frame was detected. The regions of the multiple alignment that shifted from one well-conserved reading frame to another well-conserved reading frame revealed potential sequencing errors in each of the species. A number of these were detected in the reference sequence of *S. cerevisiae*. We tested 31 of these predictions by resequencing and found that in each case the published sequence was in error and an experimentally confirmed sequence change was always within a few bases from the predicted erroneous nucleotide. The sequence corrections frequently resulted in longer ORFs and sometimes merged consecutive genes. We predicted a total of 34 cases where two adjacent ORFs in *S. cerevisiae* are joined into a single ORF in all three other species. The majority of these were confirmed by resequencing.

We then sought to identify previously unrecognized introns by searching the *S. cerevisiae* genome for conserved splicing signals. We conducted our search using ten variants of splice donor signals (6–7 bp) and eight variants of branch site signals (7 bp) that are found in experimentally validated *S. cerevisiae* introns (Clark *et al.*, 2002). We did not require that the splice signals be fully conserved at the nucleotide level in the multiple alignment, but only that splice signals appear in each species within 10 bp of each other. We additionally required that branch and donor signals be no more than 600 bp apart, which is the case for 90% of known *S. cerevisiae* introns. We then evaluated the multiple alignment surrounding the conserved

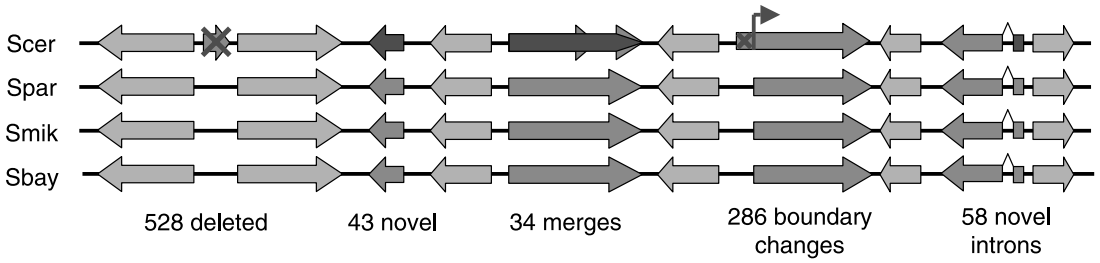


FIG. 11. Revised gene catalogue. Our comparative analysis has revised the yeast gene annotation, affecting nearly 15% of all currently annotated genes. We found that 528 previously annotated ORFs are not real genes, identified 43 novel short ORFs that were previously overlooked, proposed merges of 34 pairs of consecutive ORFs, proposed 286 boundary changes (210 start changes and 76 stop changes), and identified 58 novel introns.

signals for three properties: (1) a conserved acceptor signal, [CT]AG, 3′ of the branch site, (2) high RFC 5′ of the donor signal and 3′ of the acceptor signal, and (3) low RFC within the intron. Roughly half of the conserved donor/branch pairs met our additional requirements, yielding a total of 58 novel introns (Kellis *et al.*, 2003). Of these, 20 were independently discovered by Ares and colleagues using techniques such as microarray hybridization (Clark *et al.*, 2002). Our remaining predictions are currently being tested in collaboration with Ares and colleagues.

2.6. Conclusion: Revised yeast gene catalog

Based on the analysis above, we proposed a revised yeast gene catalog consisting of 5,538 ORFs greater than 100 amino acids in length. This reflects the proposed elimination of 503 ORFs (366 from the RFC test, 105 by manual inspection and 32 through merging). A total of 20 ORFs in SGD remain unresolved. Complete information about the gene catalog is provided in Kellis *et al.* (2003), and will be discussed more fully in a subsequent manuscript in collaboration with SGD and other yeast investigators. The revised gene count is consistent with at least two recent predictions based on light shotgun coverage of related species (Blandin *et al.*, 2000; Wood *et al.*, 2001). We believe that this represents a reasonably accurate description of the yeast gene set, because the methodology has high sensitivity and specificity and the evidence is unambiguous for the vast majority of ORFs. Nonetheless, some errors are likely to remain. The results could be confirmed and remaining uncertainties resolved by sequencing of additional related yeast species, as well as by other experimental methods.

Despite the intensive study of *S. cerevisiae* to date, comparative genome analysis resulted in a major revision of the yeast gene catalog affecting more than 15% of all ORFs (Fig. 11). The results suggest that comparative analysis of a modest collection of species can permit accurate definition of genes and their structure. Comparative analysis can complement the primary sequence of a species and provide general rules for gene discovery that do not rely solely on known splicing signals for gene discovery. Previous studies have shown that such methods are also applicable to the understanding of mammalian genes (Batzoglou *et al.*, 2000), and they will be invaluable in identifying all human genes. The ability to observe the evolutionary pressures that nucleotide sequences are subjected to radically changes our power for signal discovery.

3. REGULATORY MOTIF DISCOVERY

Regulatory motifs are short DNA sequences that are used to control the expression of genes, dictating the conditions under which a gene will be turned on or off. Each motif is typically recognized by a specific DNA-binding protein called a transcription factor (TF). A transcription factor binds precise sites in the promoter region of target genes in a sequence-specific way, but this contact can tolerate some degree of sequence variation. Thus, different binding sites may contain slight variations of the same underlying motif, and the definition of a regulatory motif should capture these variations while remaining as specific as possible.

The direct identification of regulatory motifs presents numerous challenges. By their nature, they are very short (6 to 15 bp), frequently degenerate, and can appear at varying distances and orientations upstream of target genes. Unlike genes that contain clear start and stop codons, as well as well-defined splicing signals, motifs have no detectable sequence features, and they are indistinguishable from random sequences of the same length. Their identification has thus relied heavily on experimental intervention, such as mutational analysis of promoter regions or genomewide gene expression studies under various environmental cell perturbations.

3.1. Comparison with previous work

Computationally, discovering regulatory motifs amounts to extracting signal from noise. When the motifs searched are expected to be more frequent than other patterns of the same length, one can apply discovery algorithms such as expectation maximization (EM) or Gibbs sampling and others reviewed by Stormo (2000). These were pioneered by Lawrence and coworkers (Lawrence *et al.*, 1993) and made popular in software programs such as MEME (Bailey and Elkan, 1994; Grundy *et al.*, 1997), AlignACE (Hughes *et al.*, 2000; Roth *et al.*, 1998; Tavazoie *et al.*, 1999) or BioProspector (Liu *et al.*, 2001).

These methods have typically been applied to the upstream sequences of small sets of genes, but are not applicable to a genomewide discovery. Instead, k-mer counting methods have been used to find short sequences that occur more frequently in intergenic regions, as compared to coding regions in a genome-wide fashion (Hampson *et al.*, 2002). However, these typically find very degenerate sequences (such as poly-A or poly-T) and have shown limited power to separate regulatory motifs from the mostly nonfunctional intergenic regions. This is largely due to the small number of functional instances of regulatory motifs, as compared to the large number of nonfunctional nucleotides. The discovery of regulatory motifs still relies heavily on extensive experimentation.

Comparative genomics provides a powerful way to distinguish regulatory motifs from nonfunctional patterns based on their conservation. Over evolutionary time, mutations accumulate in nonfunctional nucleotides whereas changes in functional nucleotides are detrimental and eliminated by natural selection. Hence, by comparing related genomes, we can increase our ability to separate signal from noise based on evolutionary conservation. Phylogenetic footprinting methods have traditionally applied this idea to recognize islands of conservation in individual promoter regions (Blanchette *et al.*, 2002; Blanchette and Tompa, 2002; Jiao *et al.*, 2002; McCue *et al.*, 2001; McGuire *et al.*, 2000; Oeltjen *et al.*, 1997; Pennacchio and Rubin, 2001; Tompa, 2001). Similarly, conservation has been used to distinguish possibly functional instances of previously known regulatory motifs (Gelfand *et al.*, 2000; Levy and Hannenhalli, 2002; Loots *et al.*, 2002).

In this paper, we address the genomewide discovery of the dictionary of regulatory motifs in an organism. Namely, we are interested in going beyond the individual islands of conservation and discovering the subtle signals that cross-cut these islands. Regulatory motifs may appear in slight variations in different intergenic regions, and our methods should be able to discover positions that tolerate sequence degeneracy, while capturing the full sequence specificity in constrained positions. Additionally, we are interested in the ability to discover regulatory motifs directly from genome sequence, relying solely on conservation information and without use of biological knowledge of gene function, expression, or transcription factor binding.

We studied the conservation patterns of known regulatory motifs to derive conservation criteria that would allow us to discover new motifs. We evaluated motif conservation at the genomewide level, simultaneously observing all conserved and nonconserved instances of a motif throughout the genome. By observing multiple conserved instances of a motif, we increased our predictive power over traditional methods that work with individual islands of conservation. Genomewide metrics enabled increased specificity by eliminating bases that may be conserved in individual sites by chance alone but not in most sites. Similarly, they enabled increased sensitivity in discovering degenerate motif positions that may be weakly conserved at any one site but show a conservation preference across the genome.

Our motif discovery strategy consisted of an exhaustive enumeration and testing of short sequence patterns (minimotifs) to find strongly conserved motif cores, followed by a motif refinement and collapsing step that ultimately produced a small number of full motifs (Fig. 12). We introduce three different conservation criteria (CC) for testing candidate motifs: overall genomewide intergenic conservation (CC1), preference for intergenic conservation over genic conservation (CC2), and differential conservation in upstream-only versus downstream-only regions (CC3). Using these three criteria, we tested a large set of

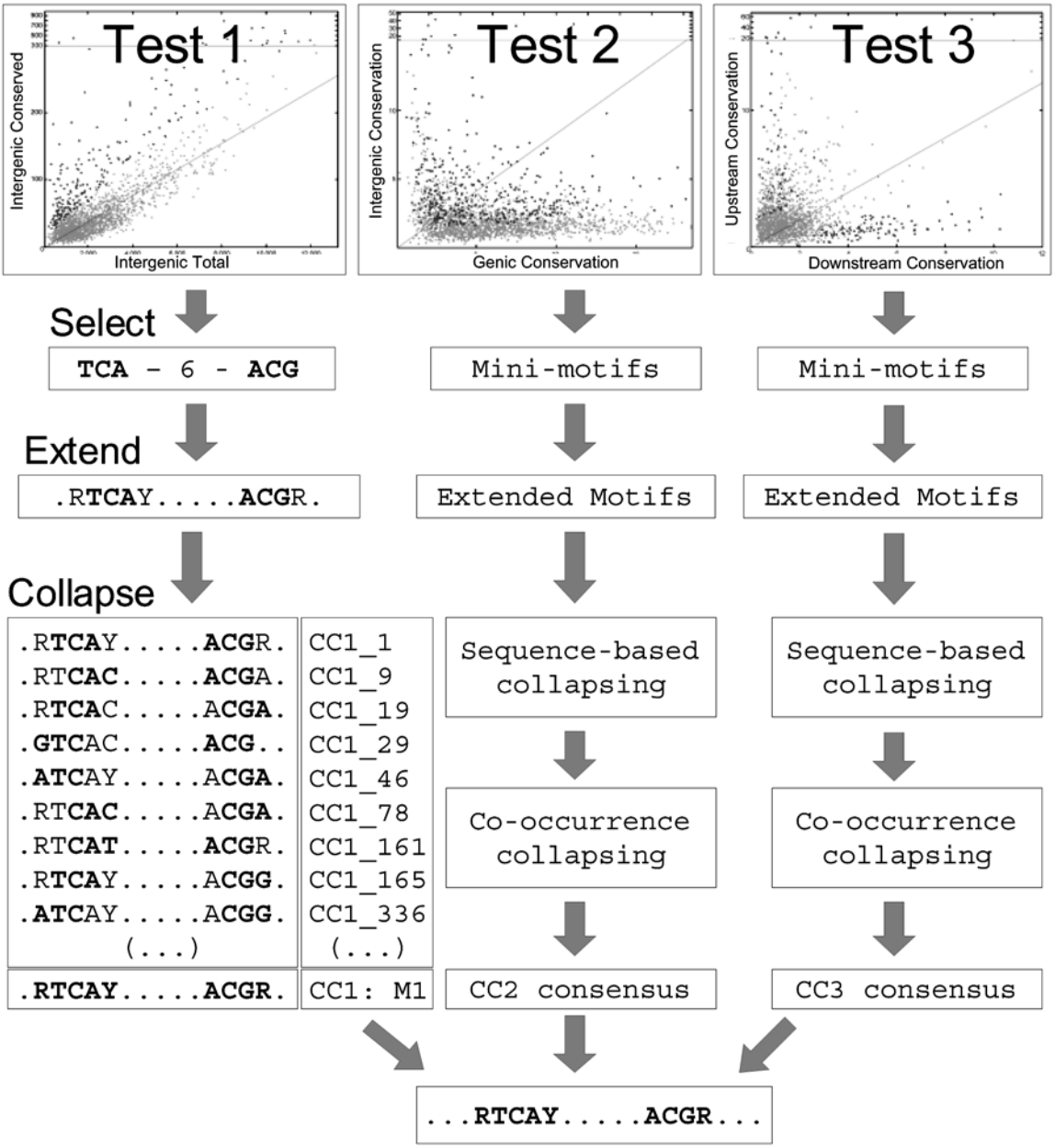


FIG. 12. Genome-wide motif discovery. We select minimotifs with significant nonrandom conservation according to three conservation criteria. We extended these by searching for neighboring bases that are preferentially conserved when a minimotif is conserved. We collapsed minimotifs with similar extensions within and across tests based on sequence similarity and genome-wide cooccurrence in the same intergenic regions. This analysis yielded 72 full motifs with significantly strong genome-wide conservation.

short sequence patterns (minimotifs) and selected those with nonrandom conservation. We extended these minimotifs, searching for neighboring positions with significant nonrandom conservation and collapsed the extended motifs hierarchically based on sequence similarity and genomewide cooccurrence. This resulted in a dictionary of regulatory elements, discovered solely based on their genomewide conservation.

The final list of genomewide motifs includes most previously published regulatory motifs, as well as additional motifs that correlate strongly with experimental data. Our results provide a global view of functionally important regulatory motifs obtained independently of knowledge of gene function, gene expression clusters, protein interactions, or transcription factor binding. Hence, they can provide an important

link between these data sources of cell state towards understanding the dynamic nature of the cell and the complexity of regulatory interactions.

3.2. Genomewide discovery of motif cores

Regulatory motif discovery requires the ability to explore the space of possible sequences and a scoring scheme that allows the selection of relevant motifs. Gibbs sampling and expectation-maximization discovery methods (Stormo, 2000; Tavazoie *et al.*, 1999) explore motif space by directly sampling the genome, the performance of which can be very dependent on initialization. On the other hand, enumeration approaches (Hampson *et al.*, 2002) have had the disadvantage that the motifs discovered are frequently too simplistic to adequately capture the sequence specificity of transcription factors. We used a mixed approach, first exploring sequence space by the exhaustive enumeration of short sequence patterns, then refining the patterns with nonrandom conservation to construct full motifs that include degenerate motif positions. We chose to use a finite alphabet representation rather than the more general position weight matrices for ease of manipulation of motifs, but also for the ability to use enumeration approaches. We recognize that more general representations might better capture the versatility of transcription factor binding and additional species might be necessary to fully specify the additional parameters, but these will be the subject of future work.

To select a form of short sequence compact enough to allow enumeration, but also general enough to resemble biologically important regulatory motifs, we studied the types of recognition sequences of transcription factors in yeast. These regulatory motifs depend on the three-dimensional structure that a transcription factor assumes at its protein-DNA contact. Two main types of contact account for nearly all known regulatory motifs in yeast: either a single stretch of nucleotides typically six to eight base pairs long, or two smaller stretches separated by a gap of unspecified nucleotides that are not involved in the recognition. For example, the Gal4 transcription factor binds DNA as a dimer, each part contacting three base pairs, the two halves separated by a gap of 11 unspecified nucleotides (one full turn of the double helix). The associated regulatory motif is $CGG(N)_{11}CCG$, where N can be any of the four bases. On the other hand, factors like Mbp1 and Cbf1 contact DNA at a contiguous stretch of base pairs, recognizing respectively the motifs ACGCGT and RTCACRTG, where R = [AG].

We defined a minimotif to be a short sequence pattern of the type XYZ-*m*-UVW, consisting of two stretches of specified nucleotides, each stretch exactly three base pairs long, and the two stretches separated by a fixed gap of ***m*** unspecified nucleotides. These minimotifs will serve as seeds in sequence space for constructing full motifs in the following section. Both Gal4 and Mbp1 can be expressed directly as minimotifs, respectively, CGG-11-CCG and ACG-0-CGT, by allowing for a zero-length gap. However, it is only through the motif optimization stage (see Section 3.2) that a motif like Cbf1 can be constructed. Multiple seeds may lead to the construction of Cbf1, such as TCA-0-CGT, TCA-1-GTG, etc. Hence, although insufficient to directly capture the exact sequence of each regulatory motif, minimotifs allow us to explore motif space and follow strongly conserved seeds.

To explore sequence space, we enumerated all minimotifs with gap size m between 0 and 21 nucleotides. For a given gap size, there are $4^6 = 4{,}096$ possible sequences, but only 2,080 unique minimotifs, considering a motif and its reverse complement as the same motif. This resulted in a total of 45,760 distinct minimotifs. We assumed that the large majority of these show a conservation typical of random sequences, and thus minimotifs provided an internal control for estimating a basal rate of conservation (the average rate of conservation at the evolutionary distance separating the species in absence of selective pressure). We then compared the conservation rate of each candidate minimotif to this basal conservation rate in order to evaluate the nonrandomness of its conservation.

In order to derive conservation signatures that allow us to distinguish real motifs from nonfunctional sequences, we studied the conservation patterns of known motifs and compared these to random motifs. For example, in the *S. cerevisiae* genome, the Gal4 motif occurs a total of 96 times in intergenic regions and 415 times in genic (protein-coding) regions. The motif displays certain striking conservation properties (Table 2). First, occurrences of the Gal4 motif in intergenic regions have a conservation rate (proportion conserved across all four species) that is ~five-fold higher than for random minimotifs (12.5% versus 2.4%). Second, the Gal4 motif is preferentially conserved in intergenic regions rather than in genic regions (12.5% versus 3%). By contrast, random motifs are less frequently conserved in intergenic regions than genic regions (3.1% versus 7.0%), reflecting the higher overall level of conservation in genic regions. Thus, the relative conservation rate in intergenic versus genic regions is ~11-fold higher for Gal4 than

Table 2. Conservation Properties of Gal4 Motif Inspire Three Tests for Motif Discovery[a]

Evaluate conservation within:	*Gal4*	*Controls*	*Test*
(1) All intergenic regions	12.5%	2.4%	INT
(2) Intergenic: coding	12:3	3:7	GEN
(3) Upstream: downstream	12:0	1:1	UPD

[a]We compared the conservation of Gal4 to that of a population of random control motifs of similar form (three nucleotides spaced from another three nucleotides by a fixed gap of unspecified nucleotides). We found that 12.5% of all intergenic instances of Gal4 were conserved, but only 2.4% for control motifs (a 6-fold enrichment). Moreover, coding instances of Gal4 diverged more frequently than for control motifs (3% versus 7%), providing an 11-fold enrichment for Gal4 when comparing intergenic to coding conservation. Finally, Gal4 showed a preference for conservation in intergenic regions that are upstream of both flanking genes, but the control motif show no such preference. We select motif cores (minimotifs) based on three conservation criteria inspired by these conservation properties of known motifs.

for random motifs. Third, the Gal4 motif shows a higher conservation rate in divergent versus convergent intergenic regions (those that lie upstream versus downstream of both flanking genes); no such preference is seen for control motifs. These three observations suggest various ways to discover motifs based on their conservation properties (see conservation criteria below).

Based on these observations, we developed three computational tests that distinguish those patterns that are under selective pressure for conservation. We then searched for candidate regulatory motifs whose conservation is significantly higher according to the three conservation criteria below.

3.2.1. Conservation criterion 1 (CC1): Intergenic conservation. The first test evaluates the overall conservation throughout intergenic regions. The inspiration came from the conservation properties of the Gal4 motif, namely, its strong overall intergenic conservation (Table 2). We searched for minimotifs that show similar conservation patterns by enumerating all 45,760 minimotifs.

For every minimotif, we obtained the total count of intergenic instances $\boldsymbol{i}$ and the count of perfectly conserved intergenic instances $\boldsymbol{ic}$. We found that the two counts are linearly related for the vast majority of motifs at a given gap size. This allows us to estimate the basal conservation rate of minimotifs $\boldsymbol{r}$ and use it to evaluate the nonrandomness in the conservation of any one minimotif. If we assume that the chance conservation of any one minimotif instance is independent of the conservation of any other motif instance, we can model the total number of conserved instances as a binomial process with success probability $\boldsymbol{r}$, the basal minimotif conservation rate. The probability of observing $\boldsymbol{ic}$ successes out of $\boldsymbol{i}$ trials by chance alone is given by $p(k) = \binom{i}{ic} \cdot r^{ic} \cdot (1-r)^{i-ic}$.

For example, the Gal4 motif showed $\boldsymbol{ic} = 12$ conserved instances out of $\boldsymbol{i} = 96$ total instances (12.5%). However, the 2,080 minimotifs at gap size 11 show an average conservation rate of 2.4%, and hence we estimate the probability that any intergenic instance be conserved to be $\boldsymbol{r} = 0.024$. Under a binomial model, the probability that the Gal4 conservation is due to chance alone is less than $3 \cdot 10^{-6}$.

We applied this test systematically to all 45,760 minimotifs, to discover novel sequences with conservation patterns similar to Gal4. For each gap size, we estimated the basal conservation rate $\boldsymbol{r}$ in two steps. In the first step, we selected nonoutlier minimotifs as those within three standard deviations of the log-average ratio $\boldsymbol{ic}/\boldsymbol{i}$. We then calculated for every motif the binomial probability $\boldsymbol{p}$ of observing $\boldsymbol{ic}$ successes out of $\boldsymbol{i}$ trials, given parameter $\boldsymbol{r}$. We finally assigned a z-score $\boldsymbol{S}$ to every minimotif as the number of standard deviations away from the mean of a normal distribution that correspond to tail area $\boldsymbol{p}$. This score is positive if the motif is conserved more frequently than random and negative if the motif is diverged more frequently than random.

We found a number of minimotifs with striking conservation scores. For example, the top-ranked minimotif TCA-6-ACG was perfectly conserved in 397 intergenic regions (33% of all instances) and showed a z-score of 37 standard deviations. This minimotif corresponds to the central portion of RTCRYnnnnnACG, the motif recognized by the Abf1 transcription factor. Minimotifs ranked 2, 3, and 5 (CGG-0-GTA, GGG-0-TAA, CGG-1-TAA); all had more than 300 conserved intergenic instances (33% of the total, z-scores

above 32 standard deviations). Their sequences heavily overlap to form a longer sequence, CGGGTAA, that is part of the recognition site for the transcription factor Reb1. Hence, our test was able to discover real regulatory motifs, simply based on their genomewide conservation. Additionally, a number of novel motifs were discovered (see below).

To estimate the total number of significant minimotifs, we studied the tails of the z-score distribution (Fig. 13, top panel). We found that the right tail of the distribution (strongly conserved minimotifs with positive z-scores) was much longer than the left tail (weakly conserved minimotifs with negative z-scores). For a cutoff of 5 sigma, as many as 1,190 minimotifs showed z-scores of +5 sigma or greater, as compared to only 25 minimotifs with a z-score below −5 sigma. Comparing the left and the right tail of the distribution provides an internal control as to how many of the significant motifs are real and how many are due to chance. We estimate the specificity at a given cutoff as (R − L)/R, where L is the count of minimotifs in the left tail, and R is the count of minimotifs at the right tail of the distribution. At a cutoff of 5 standard deviations, we therefore estimate that 97.5% of the 1,190 minimotifs are nonrandom in their conservation enrichment.

3.2.2. Conservation criterion 2: Intergenic versus coding conservation. The second test evaluates the relative conservation in intergenic regions as compared to the conservation in coding regions. We were again inspired by the conservation properties of Gal4, namely, that the Gal4 motif was preferentially conserved in intergenic regions rather than in genic regions (Table 2). We scored this preference for each minimotif by comparing its conservation to the conservation of a control population of random minimotifs.

We quantified the conservation preference for intergenic regions over coding regions based on counts of conserved and nonconserved instances as in the previous section. In addition to ic and i (the conserved and total number of intergenic instances defined in the previous section), we counted the number of conserved coding instances gc and the number of total coding instances g, for every minimotif. We searched for minimotifs for which $\frac{ic}{i} \gg \frac{gc}{g}$, which can alternatively be written as $\frac{ic}{gc} \gg \frac{i}{g}$. Intuitively, these are minimotifs for which the proportion of conserved occurrences that are intergenic $b = ic/(ic+gc)$ is higher than expected from the proportion of total occurrences that are intergenic $a = i/(i+g)$. We observed that on average, a = 25% of minimotif instances are found in intergenic regions (as expected, since intergenic regions account for roughly $\frac{1}{4}$ of the yeast genome). Additionally, we found that motifs with similar GC content have similar b/a ratios. Hence, b_x/a_x for minimotif x is similar to b_{avg}/a_{avg}, the average ratio for minimotifs containing the same number of C or G bases as motif x. Thus, for minimotif x, we can approximate $b_x \sim f_{avg} \cdot a_x$, where f_{avg} is the log-average ratio b/a for minimotifs of identical GC content at a given gap size. The estimated b_x thus gives us the proportion of conserved instances that we expect to be intergenic for a given minimotif x.

To quantify the observed imbalance between ic and gc, as compared to the expected ratio b_x for a given minimotif, we used a binomial test as previously. We considered b_x to be the probability that any conserved minimotif instance would be intergenic and evaluated the probability of observing ic conserved intergenic instances out of $(ic + gc)$ conserved instances overall. In absence of selective bias for intergenic versus genic conservation, the probability of observing ic successes out of $(ic + gc)$ trials given success rate b_x is given by the binomial random variable $p(ic) = \binom{ic+gc}{ic} \cdot b_x^{ic} \cdot (1 - b_x)^{gc}$. We associated a z-score S to this probability as previously, corresponding to the standard deviations away from the mean of a normal distribution that correspond to tail area p. This score was positive when for motifs preferentially conserved in intergenic regions, and negative for motifs preferentially conserved in genic regions.

We found a number of high-scoring minimotifs that showed a preference for intergenic conservation. The top-scoring minimotifs again contained a number of matches to known motifs. For example, the subcomponents of the Reb1 motif CGG-0-GTA and CGG-1-TAA showed only 2% conservation within genic regions, as compared to 51% and 41% in intergenic regions, and the Abf1-like minimotifs CGT-6-TGA and CGT-7-GAT showed respectively 5% and 4% genic conservation as compared to 38% and 30% intergenic conservation. Overall, we found an anticorrelation between genic and intergenic conservation: minimotifs with the strongest intergenic conservation (average $ic/i = 20\%$ for the top 50 minimotifs, and average $ic/i = 8\%$ for the top 500 minimotifs) show weak genic conservation (average $gc/g = 4\%$ for the two sets, as compared to 7% for random minimotifs).

To estimate the number of significant motifs, we used again the shape of the distribution of z-scores. We found that z-scores are again centered around zero for the large majority of minimotifs, but show a longer right tail (Fig. 13, bottom panel). We compared the right and left tails to estimate the specificity of

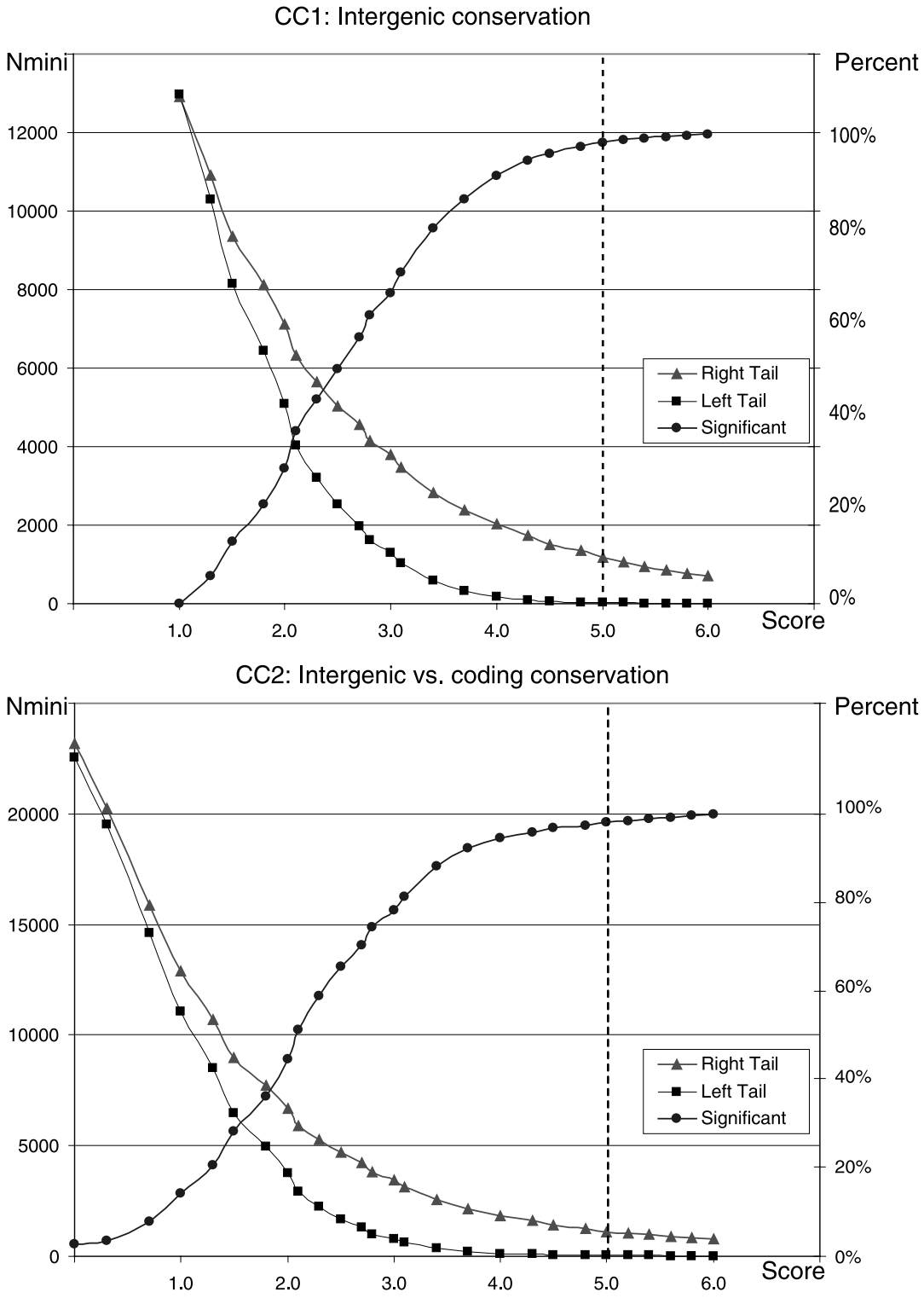


FIG. 13. Estimating number of significant minimotifs. We compare the right tail (R) and the left tail (L) of the distribution of minimotif conservation scores. The ratio (R − L)/R allows us to the estimate the number of significant points at a given score cutoff. We find that 94% of scores 5 sigma away from the mean are on the right tail for intergenic conservation score *CC1* and 97% for intergenic versus coding conservation score *CC2*. The vast majority of minimotifs selected at this cutoff are nonrandom.

this conservation test at a given cutoff. We found that for a cutoff z-score of 5 sigma, the right tail of the distribution (positive z-scores) contains 1,110 motifs, as compared to 39 motifs for the left tail (negative z-scores). Hence, at this cutoff, we expect 97% of the selected 1,110 minimotifs to be nonrandom.

3.2.3. Conservation criterion 3: Upstream versus downstream conservation. The third conservation test evaluates a preference for either upstream or downstream conservation. We classified intergenic regions as *upstream-only* if the flanking genes are divergently transcribed from a common promoter, or *downstream-only* if the flanking genes are convergently transcribed towards a common terminator. Motifs in upstream-only regions are upstream (5′) of both flanking genes, whereas motifs in downstream-only regions are downstream (3′) of both flanking genes. When the flanking genes were in the same transcriptional orientation, we did not classify the intergenic region, since promoter elements within it could play either upstream or downstream roles.

The upstream-only and downstream-only regions allow us to recognize motifs with specific roles acting upstream of genes or specific roles acting downstream of genes. To detect a specificity in the upstream versus downstream conservation of a minimotif, we counted $\boldsymbol{uc}$ and $\boldsymbol{u}$, the conserved and total counts in upstream-only regions, and similarly $\boldsymbol{dc}$ and $\boldsymbol{d}$ in downstream-only regions. Although upstream-only regions account for twice the total length of downstream-only regions, they show the same rate of conservation. Thus, we found that the two ratios $\boldsymbol{uc/u}$ and $\boldsymbol{dc/d}$ are both similar to $\boldsymbol{ic/i}$ for the large majority of motifs. To find variations from a balanced conservation between upstream-only and downstream-only regions, we used a chi-square contingency test on the four counts [(uc,u),(dc,d)].

We found both known and novel motifs that showed significant evidence for either upstream-specific or downstream-specific conservation. A number of minimotifs corresponding to transcription factors that are known to bind upstream of genes showed upstream-only conservation. For example, the Abf1 and Reb1 minimotifs showed a strong conservation in upstream-only regions ($\boldsymbol{uc/u} = 50\%$) and weak conservation in downstream-only regions ($\boldsymbol{dc/d} = 2\%$).

Strikingly, we also discovered minimotifs that showed a strong preference for downstream-only conservation rather than upstream-only. For example, the minimotif TGT-1-AAT showed a strong conservation in downstream regions ($\boldsymbol{dc/d} = 14\%$) and a weak conservation in upstream-only regions ($\boldsymbol{uc/u} = 2\%$), suggesting that it acts at the 3′ end of genes. Indeed, this sequence is found enriched in the 3′ untranslated region of genes that are targeted to the mitochondrion and may serve as an mRNA localization signal.

We found 1,089 minimotifs with significant difference between upstream and downstream conservation (chi-square value of 10.83 or greater; p-value less than .001). Correcting for multiple testing, we thus expect to see roughly 46 of the 45,760 tested motifs with such a score by chance alone and hence estimate that 96% of the 1,089 chosen minimotifs to be nonrandom.

3.3. Constructing full motifs

Our results in the previous section demonstrated the power of using the lens of evolutionary conservation to discover biological signals. We were able to discover sequences important in gene regulation by virtue of their conservation alone and without any knowledge of gene function. We next refined the minimotifs discovered and created full motifs that better represent the full sequence specificity of regulatory motifs and transcription factors.

We used the minimotifs discovered according to the three genomewide conservation criteria as seeds in sequence space, which we then used to discover full motifs. We first extended each minimotif by searching for surrounding positions that show nonrandom conservation. We found that different minimotifs sometimes showed similar extensions, and we developed methods for merging similar extended motifs based on their sequence similarity and their genomewide cooccurrence. We finally developed methods for testing the genomewide conservation of complex motifs, allowing for degenerate bases and varying lengths or gaps in their sequence definition. We used these methods to select those full motifs with significant genomewide conservation.

3.3.1. Motif extension. We saw previously that the minimotifs discovered were frequently only a subset of the known binding site for a particular factor (for example, the minimotif CCG-1-GTA is only part of **CCGGGTA**A, the recognition sequence for Reb1). Since the corresponding transcription factors recognize the longer sequences, we expect additional bases to be under selective pressure for conservation. We

developed a motif extension algorithm to recognize these bases, and we extended minimotifs by including surrounding bases that show additional nonrandom conservation.

We used an iterative approach, at each iteration specifying one additional position, including it in the motif, and iteratively extending the longer motif until no further extension was significant. At each iteration, we enumerated all possible one-nucleotide extensions for any position within the gap of a minimotif or up to four nucleotides away on either side of the minimotif. For each position, we selected the best character extension amidst the 14 degenerate characters of the IUB code A, C, G, T, S, W, R, Y, M, K, B, D, H, V, where S = [CG], W = [AT], R = [AG], Y = [CT], M = [AC], K = [GT], B = [ACG], D = [AGT], H = [ACT], V = [ACG]. By using a larger motif alphabet, we allowed degeneracies in transcription factor binding, enabling us to recognize more subtle signals in the motif extension. For a minimotif with gap size m, we considered $14 \cdot (4 + m + 4)$ possible extensions at the first iteration (for example, we tested 154 one-base extensions for minimotifs with gap size $m = 3$ and 266 extensions for $m = 11$).

We evaluated the nonrandomness of specifying an additional position by its ability to discriminate between the neighborhood of conserved motif instances (the *causal* set) and the neighborhood of a control set of intergenic sequences (the *control* set). The *causal set* contained the alignments of every perfectly conserved minimotif and its neighborhood as defined previously. The *control set* contained the alignments and neighborhood of all "one-off" instances of the minimotif, namely, all instances where one base was not conserved on each half of the minimotif. For minimotif XYZ-*m*-UVW, the control set contained all instances of (notX)YZ-*m*-(notU)VW, (notX)YZ-*m*-U(notV)W, and so on. By using a control close to the motif tested, we ensured that the extension does not simply reflect the background sequence composition of intergenic regions.

We then quantified the discriminative power of each one-base extension. We considered $\boldsymbol{a}$, the number of alignments matching the extension in the causal set, and $\boldsymbol{b}$, the number of alignments matching the same extension in the control set. We compared these two counts to $\boldsymbol{A}$ and $\boldsymbol{B}$, the cardinalities of the causal and control sets without the extension. We searched for those extensions for which $\boldsymbol{a}/\boldsymbol{A} \gg \boldsymbol{b}/\boldsymbol{B}$. To evaluate the significance of the enrichment, we used a chi-square contingency test on the four counts $[(\boldsymbol{a}, \boldsymbol{A}), (\boldsymbol{b}, \boldsymbol{B})]$. We then selected the one-base extension with the highest chi-square score. We terminated the extension when the chi-square score did not exceed a z-score of 3 standard deviations and reported the extended motif.

A special case arose in extending minimotifs that are reverse palindromes. These are minimotifs XYZ-*m*-UVW that contain the same sequence on both strands, namely, for which XYZ is the reverse complement of UVW (X = W′, Y = V′, Z = U′, where x′ denotes the complement of x). When constructing the neighborhood of such motifs, one can place the motif in either orientation, but the resulting neighborhood alignments are different if the surrounding sequence is not palindromic too. Namely, if a palindromic minimotif XYZ-2-UVW occurs in the sequence abcdXYZefUVWghij, then it also occurs in the sequence j′i′h′g′XYZf′e′UVWd′c′b′a′ at the same locus, since XYZ = W′V′U′. For each conserved locus, we therefore considered both neighborhoods in our counts. This has the implication that only palindromic extensions of palindromic minimotifs were considered and the symmetry was not resolved until subsequent motif collapsing steps. Note that to resolve the symmetry, we could include the first neighborhood in only one orientation and include subsequent neighborhoods in the orientation that best matches previously included neighborhoods.

The extension strategy successfully completed the recognition sequences of known motifs, starting from their minimotif seeds. Minimotif CGT-6-TGA was extended to Y**CGT**NNNNNR**TGA**Y, successfully discovering the additional degenerate bases of the Abf1 motif. The minimotifs CGG-0-GTA, GGG-0-TAA, and CGG-1-TAA were all three extended to the motif CGGGTAAY, the binding site for Reb1. However, the minimotif CGG-11-CCG remained as it is, identical to the known Gal4 motif, showing that it was already optimal. Hence, by comparing the sequence neighborhoods of conserved minimotif instances to the appropriate controls, we were able to extend motifs to local maxima in sequence conservation space. The ability to effectively climb the search space of possible motifs has implications regarding the smoothness of genomewide conservation scores, whereby simple motifs can be used as a proxy for discovering similar complex motif, and iterative refinement algorithms are possible for exploring sequence space.

3.3.2. Motif collapsing by sequence similarity. Minimotifs that are subsequences of the same regulatory element will frequently have the same or similar extensions. To construct a dictionary of unique

regulatory elements, we need to recognize these similarities and group together motif variations that all stem from the same motif. We developed methods for collapsing a large list of motif candidates into a set of unique regulatory elements. We used two different metrics for recognizing that two motifs should be grouped together: sequence similarity and genomewide cooccurrence. We grouped motifs using a hierarchical clustering approach and constructed a consensus sequence for each group.

We measured the sequence similarity between two profiles as the number of bits they have in common. We represented each sequence profile as a weight matrix in a probabilistic framework. We associated a probability vector $p[j] = [p_A, p_C, p_G, p_T]$ to each motif position $m[j]$, representing the probability of occurrence of each of the four bases at that position (for example, W = [A|T] = $[\frac{1}{2}, 0, 0, \frac{1}{2}]$, and D = [A|G|T] = $[\frac{1}{3}, 0, \frac{1}{3}, \frac{1}{3}]$). We then assigned a similarity score between two motifs m_1 and m_2, based on the number of bits they had in common in the best ungapped alignment between them. We scanned the two motifs past each other, comparing corresponding positions $m_1[j]$ and $m_2[k]$ at every offset of each orientation. At a given offset, we summed the bits in common at each position, corresponding to the probability $p_1[j] \cdot p_2[k]$ that the same base would be selected from the two vectors at each position when sampling from the probability distribution. If two motifs had no base in common in a given position, we did not carry over negative scores, but instead ignored the mismatches at that position. Since the theoretical maximum of bits in common between two motifs is given by the number of bits in the shorter motif, we scored each similarity as the percentage of that maximum obtained in their best alignment. We used a percentage-based score to avoid that longer motifs be grouped together simply for having more bits in common.

Based on the pairwise motif similarity matrix, we clustered the motifs hierarchically. We ordered the joins by decreasing similarity scores, such that the most similar motifs would be merged first, forming clusters in motif space, where more distant motifs would subsequently be added. We computed the similarity between two motif clusters $c1$ and $c2$ based on the precomputed similarity scores of motifs across the two clusters. For every motif m_i in cluster $c1$, we scored $S_{1\to2}(m_i)$, the similarity to the closest point in $c2$. Similarly, for every motif m_j in cluster $c2$, we scored $S_{2\to1}(m_j)$, the similarity to the closest point in $c1$. We then averaged all forward and reverse similarities between the two clusters to obtain a similarity score between the two clusters $S_{12} = \frac{1}{(i+j)}[\sum_{m_i \in c1} S_{1\to2}(m_i) + \sum_{m_j \in c2} S_{2\to1}(m_j)]$. We merged two clusters $c1$ and $c2$ if the similarity score S_{12} between them was above 70%, and we left them as separate clusters otherwise. The order of merges of motif clusters was dictated by the order of similarity scores of individual motifs; namely, we evaluated collapsing clusters $c1$ and $c2$ only when the next strongest untested similarity score was between $m_i \in c1$ and $m_j \in c2$. A high-scoring similarity between motifs m_i and m_j may not result in a merge if the clusters that contain the two motifs are too dissimilar, and conversely a low-scoring similarity between m_i and m_j may actually result in a merge if the remaining motifs in the two clusters show sufficient similarity.

We applied this algorithm within each test independently to collapse the minimotifs discovered. The 1,190 extended motifs discovered in test 1 (CC1) clustered into 332 unique patterns, the 1,110 motifs from test 2 (CC2) into 269, and the 1,089 motifs from test 3 (CC3) into 285. The first nine members of a cluster containing ABF1-like motifs from test 1 are shown in Fig. 12, with minimotif cores shown in bold and the corresponding consensus CC1_M1. The clusters of motifs for each of the tests are shown in Supplementary Tables S8b (*http://www.broad.mit.edu/seq/Saccharomyces/*).

3.3.3. Motif collapsing by genome-wide cooccurrence. Finally, we merged motifs from the three conservation tests together when they cooccurred in the same intergenic regions. This step collapsed motifs that were part of the same regulatory element, even if their sequences did not necessarily overlap. For example, two halves of a transcription factor binding site that shared only a few central bases would not cluster based on their sequences alone, but they will be conserved in largely the same intergenic regions.

We computed a motif cooccurrence score C_{12} between motifs m_1 and m_2 based on the respective counts of intergenic regions r_1 and r_2 containing conserved instances of m_1 and m_2, respectively, and the count of intergenic regions $r_{1\&2}$ that contain conserved instances of both motifs. Given a total count n of all intergenic regions and r_1 intergenic regions containing motif m_1, we evaluated the probability $p_{1\&2}$ that a random set of r_2 intergenic regions contains at least $r_{1\&2}$ items in common with r_1. This chance probability of observing at least such an overlap is given by the hypergeometric sum $p_{1\&2} = \sum_{k \geq r_{1\&2}} \binom{r_1}{k}\binom{n-r_1}{r_2-k}/\binom{n}{r_2}$. Intuitively, this probability corresponds to the ways of selecting k = $r_{1\&2}$ or more items at random from a

set of r_1 objects (and thus $r_2 - k$ items amidst the remaining $n - r_1$ objects), given all the ways to select r_2 objects amidst a total of n.

Given this similarity metric, we used the same hierarchical clustering approach as previously to group motifs based on the matrix of pairwise cooccurrence scores. We obtained as few as 400 distinct groups across the three categories, 160 of which contained motifs discovered in at least two tests (Supplementary Information S8c: *http://www.broad.mit.edu/seq/Saccharomyces/*). We then constructed consensus motifs for each of these and evaluated their genomewide conservation to select a final list of strongly conserved genomewide motifs.

3.3.4. Genomewide conservation score for complex motifs. We previously developed methods for evaluating the genomewide conservation of a minimotif by comparing its conservation to that of similar minimotifs. In this section, we develop a general motif conservation score (MCS) that we use to evaluate the conservation of complex motifs of any length and degeneracy type. Similarly to scoring minimotif conservation, for every motif $\boldsymbol{M}$, we first constructed a control set of motifs, then used these to estimate a basal rate of conservation $\boldsymbol{r}$, and finally evaluated the non-randomness of the conservation of $\boldsymbol{M}$, given the basal rate $\boldsymbol{r}$.

For a given motif $\boldsymbol{M}$ of length $\boldsymbol{L}$, we constructed random control motifs $\boldsymbol{R}$ of the same length and degeneracy. For each position $\boldsymbol{R[i]}$, we selected a unique, two-fold, three-fold, or four-fold degenerate base, depending on the degeneracy level of $\boldsymbol{M[i]}$. For example, if the motif base $\boldsymbol{M[i]}$ is W = [A|T], then the degeneracy level for $\boldsymbol{R[i]}$ will be 2, and we will select a two-fold degenerate base for $\boldsymbol{R[i]}$.

The actual sequence of the random motifs, however, was dictated by the k-mer distribution in *S. cerevisiae*. Namely, for each control motif $\boldsymbol{R}$, we selected a random intergenic sequence $\boldsymbol{S}$ of length $\boldsymbol{L}$ in the *S. cerevisiae* genome, regardless of conservation. The nucleotide $\boldsymbol{S[i]}$ was used to select the degenerate base at $\boldsymbol{R[i]}$, such that $\boldsymbol{S[i]}$ appears with nonzero probability in $\boldsymbol{R[i]}$. For example, if we seek to select a two-fold degenerate base for $\boldsymbol{R[i]}$ and $\boldsymbol{S}[i] = G$, then $\boldsymbol{R}[i]$ can be one of R = [G|A], S = [G|C], K = [G|T], all of which are two-fold degenerate bases containing G as one of the possible bases. When multiple possibilities existed for the degenerate base, we selected one at random, using the nucleotide distribution in *S. cerevisiae* as weights in our choice ($p_A = p_T = .32$ and $p_C = p_G = .18$). In the example above, R and K would be selected with probability $p_A/(p_A + p_T + p_C) = .39$ each, and S would be selected with probability $p_C/(p_A + p_T + p_C) = .22$.

Using this method, we constructed 20 random motifs $\boldsymbol{R_1 - R_{20}}$ for a given motif $\boldsymbol{M}$, which allowed us to obtain a typical conservation rate for random motif-like sequences of the specified degeneracy. For each random motif, we counted the number of conserved intergenic instances across the four species and the number of total intergenic instances in *S. cerevisiae*, regardless of conservation. Since these motifs contain degeneracies, counting conserved and total instances was nontrivial, since we could no longer count exact matches and require perfect conservation as for minimotifs.

We instead used a probabilistic framework to detect conserved motif instances. We interpreted every genomewide motif $\boldsymbol{M}$ of length $\boldsymbol{L}$ as a probabilistic model, generator of sequences of length $\boldsymbol{L}$ over the alphabet {A,C,G,T}. We then evaluated, for every genome position, the probability that the sequence was generated by motif $\boldsymbol{M}$ and compared this to the probability that the sequence was generated at random, given the ratio of A,C,G,T in the genome. We evaluated each species in turn, expressing this likelihood ratio in bits and summing across the four species to obtain a total number of bits in the alignment. Since gaps may exist in the alignment, we did not evaluate the motif match directly on the alignment. Instead, we evaluated the motif in the ungapped sequence of each species in turn and translated the motif start coordinates based on the alignment. To avoid evaluating each of 12 million start positions in the yeast genome against the motif, we first hashed the four genomes for rapid lookup and subsequently only searched those intergenic regions that contain k-mers present in the motif searched. To allow for degenerate matches, we also looked up sites containing k-mers with one or two discrepancies from the query motif. We then used a simple threshold $\boldsymbol{t}$ and obtained the count of all intergenic regions containing conserved instances of the motif with score at least $\boldsymbol{t}$. We similarly obtained the count of all intergenic regions within *S. cerevisiae* containing instances of the motif above threshold $\boldsymbol{t}$.

Depending on the threshold $\boldsymbol{t}$, the count of conserved and total instances of motifs $\boldsymbol{M}$ and $\boldsymbol{R_i}$ will vary. For a given threshold, we can evaluate the enrichment of genomewide conservation of $\boldsymbol{M}$ as compared to the controls, but we had to choose a sensitive threshold. If the threshold $\boldsymbol{t}$ is too low, then too many instances will be found, and the signal for $\boldsymbol{M}$ will be lost amidst the noise of false positives. If the threshold

t is too high, then the motif counts will be too small and carry little significance. To avoid both extremes, we picked the highest threshold ***t*** such that random motifs showed on average at least 100 instances.

We then counted conserved and total intergenic instances of each of the 20 generated control patterns and computed ***r***, the log average of their conservation rates. We additionally counted the number of conserved (***ic***) and total (***i***) intergenic instances of ***M*** and computed the binomial probability ***p*** of observing the two counts, given ***r***. We finally reported the MCS of the motif as a z-score corresponding to ***p***.

We then used the MCS score to select a dictionary of full genomewide motifs with significant conservation. We found that 72 of the consensus motifs showed an MCS score of 4.0 or higher, and we reported these as the dictionary of strongly conserved genomewide motifs (Kellis *et al.*, 2003, Table 3).

3.4. Results of genomewide motif discovery

We evaluated the performance of our methods by comparing our dictionary of 72 genomewide motifs to a catalog of 55 previously known motifs and to 358 categories of functionally-related gene sets.

3.4.1. Most previously known motifs rediscovered. We first compared the discovered motifs against previously known motifs. We assembled a catalog of 55 known regulatory sequence motifs (Kellis *et al.*, 2003, Table 2), based on public databases SCPD (Zhang, 1999; Zhu and Zhang, 1999) and YTFD (Mewes *et al.*, 1999), as well as literature support. We found that most of these known motifs indeed show extremely strong conservation, with 33 motifs (60%) having MCS $\geq$ 4. Thus, comparative genomics should allow their discovery by virtue of their stronger conservation, the sensitivity and specificity of course depending on the methods used and the signal-to-noise ratio in the comparison. Some of the motifs, however, show relatively modest conservation; these may be incorrect, suboptimal, or not well conserved.

We compared our dictionary of genomewide motifs to these known motifs. We found that our list includes 28 of the 33 known motifs with strong conservation (MCS $\geq$ 4). Strikingly, the correct pattern of degeneracies was uncovered in the vast majority of these cases, our motifs containing all strongly specified bases, but also all degenerate bases. We also found matches to 8 of the 22 known motifs with MCS $<$ 4. In these cases, our motif discovery methods identified closely related motifs that have higher conservation scores than do the previously described motifs and occur largely at the same genes. These may represent a better description of the true regulatory element for these factors.

In summary, without use of any biological knowledge of gene function, our comparative genomics methods were able to directly identify a dictionary of regulatory motifs containing matches to most of the known motifs (65% of the full set, 85% of those with high conservation).

3.4.2. Most new motifs show candidate function. We also identified 42 additional "new" motifs, previously not described in the literature and public databases. To assign candidate functions to novel motifs, we compared their conserved intergenic instances to lists of functionally related gene sets and looked for enrichment either upstream or downstream of such a list.

We assembled a collection of 318 yeast gene categories based on functional and experimental data described earlier. These categories consist of 120 sets of genes defined with a common GO classification in SGD (Ashburner *et al.*, 2000; Dwight *et al.*, 2002), 106 sets of genes whose upstream region was identified as binding a given transcription factor in genomewide chromatin immunoprecipitation (ChIP) experiments (Lee *et al.*, 2002), and 92 sets of genes showing coordinate regulation in RNA expression studies (Gasch and Eisen, 2002).

To measure how strongly the conserved occurrences correlated with the regions upstream (or downstream) of a particular gene category, we used a hypergeometric distribution (Tavazoie *et al.*, 1999). To select a meaningful cutoff for the hypergeometric score, we compared top-ranked minimotifs from each conservation score and randomly sorted minimotifs. For a hypergeometric score of at least 10^{-5}, Fig. 14 compares the number of minimotifs that show category enrichment for increasingly larger sets of top-ranked minimotifs in each test (CC1, CC2, CC3), as compared to randomly sorted minimotifs (RND). From the top 100 minimotifs of each test, 71, 80, and 89 are enriched in at least one category, as compared to only 1 for random motifs. This trend continues for the top 200, 500, and 1,000 minimotifs. Hence, for a cutoff of 10^{-5}, our search showed a 90-fold enrichment in enriched minimotifs as compared to random sequences.

We used this cutoff to assign a candidate function to the 72 genomewide motifs. Most of the 36 genomewide motifs that correspond to known motifs showed strong category correlation. Categories with

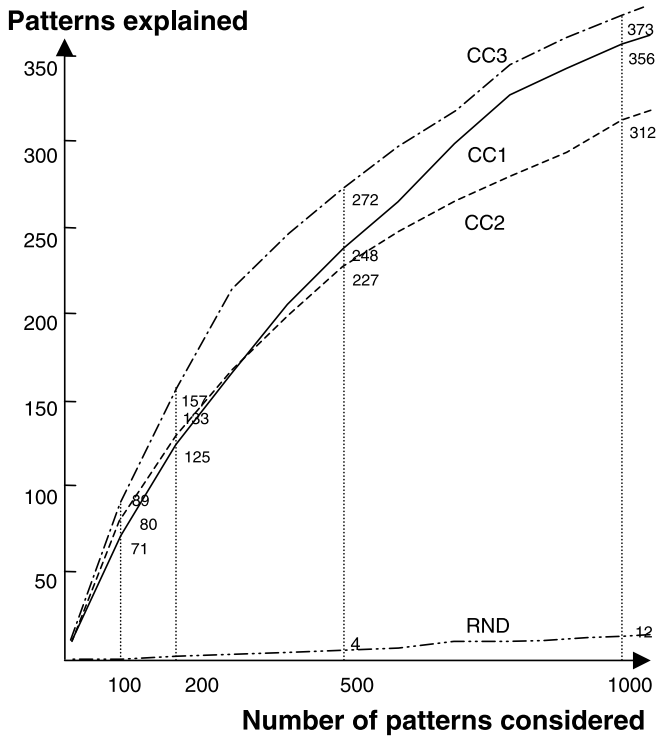


FIG. 14. Selected minimotifs show functional enrichment. Mini-motifs selected by any of the three tests (CC1,CC2, CC3) correlate with function 90 times more frequently than randomly chosen minimotifs (RND).

the strongest correlation included those identified by ChIP with the transcription factor known to bind the motif, although many other relevant categories were identified. Of the 42 novel motifs, 25 (60%) show strong correlation with at least one category and thus can be assigned a suggestive biological function (Kellis *et al.*, 2003, Table 3).

Some motifs appear to define previously unknown binding sites associated with known transcription factors, including Rgt1 involved in glucose transport and Sum1 involved in meiosis. Other motifs do not match regions bound by known transcription factors, but show strong correlation with functional categories, including genes involved in nitrogen metabolism, vesicular traffic and secretion, clathrin assembly factors, and membrane proteins. Similarly, other motifs show enrichment in genes of coordinated expression levels that are likely to be involved in energy metabolism, environmental changes, and filamentation.

Six motifs show strong conservation downstream of ORFs and are likely to act posttranscriptionally. Some of these may be in the 3′ untranslated region of a transript and play a regulatory role in mRNA localization or stability. The strongest of these is found downstream of genes whose product localizes in the mitochondrial outer membrane (Jacobs *et al.*, 2000). Other downstream motifs are found in genes repressed during environmental stress and in genes involved in energy production.

Two motifs showed variable gap spacing, suggesting a new type of degeneracy within the recognition site for a transcription factor. One corresponds closely to the known motif for Swi4 but is interrupted by a central gap of 5, 7, or 9 bases; these variant motifs all show strong correlation with genes bound by Swi4 in ChIP experiments.

In addition to discovering most previously known motifs, our analysis has thus discovered a number of novel motifs with candidate functions.

3.5. Category-based motif discovery

We next explored whether additional motifs could be found by using conservation data in combination with knowledge of functionally related genes sets. We turned to the same 318 categories of functionally re-

lated genes that we previously used to assign candidate functions to genomewide motifs (see Section 3.4.2). This time, we used these categories to discover motifs by searching specifically for conservation within individual categories.

We first evaluated our motif discovery method for ChIP experiments of factors with known motifs, and we found high sensitivity and specificity. We then searched for novel motifs in all 318 functional categories to discover novel motifs.

3.5.1. Category-based motif discovery algorithm. For each category, we used a discovery strategy similar to our genomewide search. We first enumerated minimotifs, selecting the ones that show conservation specifically within the category. We then extended these minimotifs within each category, to discover additional positions that increase specificity to the category. Finally, we collapsed similar motifs based on sequence similarity and generated a motif consensus for each group, reporting all significant full motifs for that category.

We tested each minimotif by counting the conserved instances within the category (IN) and the conserved instances outside the category (OUT). We estimated the ratio $r = \text{IN}/(\text{IN} + \text{OUT})$ that we should expect for the category, based on the entire population of minimotifs. We then calculated the significance of an observed enrichment as the binomial probability of observing IN successes out of IN + OUT trials given the probability of success r. We assigned a z-score to each minimotif, as described in the genomewide search (Section 3.2.1).

We extended those minimotifs with z-score at least 5 sigma. We used the extension method described in Section 3.3.1, but our selection criterion for scoring neighboring bases was aimed at discriminating conserved minimotif instances within the category from conserved minimotif instances outside the category, thus increasing our specificity.

We finally collapsed motifs of similar extension based on their sequence similarity. We did not use genomewide cooccurrence as a collapsing criterion, since the motifs discovered were already biased to significantly cooccur in regions within the functional category.

3.5.2. Results for transcription factors with known motifs. We first evaluated our ability to detect known regulatory motifs based on the list of regions bound by the corresponding factor in chromatin IP experiments. We assembled 43 factors for which the motif was previously known in the literature and for which ChIP experiments had been performed by Lee *et al.* (2002).

For each category defined by the ChIP experiment, we undertook category-based motif discovery. Strong category-based motifs were found in 29 cases, and these invariably corresponded closely to the known motifs. These include 11 cases in which the motif had not been found by genomewide motif discovery, suggesting that a category-based approach can be more sensitive in some cases. No strong category-based motifs were found for the remaining 14 known cases, including 7 cases in which genomewide analysis yielded the known motif. Analysis of these 14 known motifs showed that none were, in fact, enriched in the ChIP-based category. This may reflect errors in the known motifs in some cases and imperfect ChIP data in others. Genomewide analysis may simply be more powerful than category-based analysis in some instances. In all, 46 of the 55 known motifs were found by either genomewide or category-based analysis. The remaining nine cases may reflect true failures of the comparative genomic analysis or errors in the known motifs.

We compared our comparison-based results to the motifs discovered by MEME working in a single species without conservation data (Lee *et al.*, 2002). The comparison-based method showed increased sensitivity in discovering all motifs for which the ChIP experiment indeed contained the correct motif (Table 3). Additionally, the method showed strong specificity in the motifs discovered: the motifs were short and concise and closely matched the published consensus. On the contrary, MEME failed to find the true motif in a number of cases, and when a motif was found, it was generally obscured by a number of surrounding spurious bases that are not reported in the known motifs. Thus, we successfully used the additional information that comes from the multiple alignment to improve category-based motif discovery with very satisfactory results. By comparing multiple species, we increased the signal-to-noise ratio. The evolutionary information allowed the search to focus on the conserved bases, eliminating most of the noise. Table 3 summarizes the results. For each factor, we show the published motif, the negative log of the hypergeometric enrichment score of the motif within the category, the motif discovered by MEME and

TABLE 3. CATEGORY-BASED MOTIF DISCOVERY[a]

Factor	Known motif	Hyper	MEME motif (Lee *et al.*)		Category-based motif		Comparison
Abf1	RTCRYnnnnnACG	91.4	T**RTCAY**T-Y--**ACG**RA	Good	**RTCACnnnnnACG**A	Good	Same
Gcn4	ATGACTCAT	47.8	**TGAGTCA**Y	Good	**RTGACTCA**	Good	Same
Reb1	CCGGGTAA	44.7	**SCGGGTAA**Y	Good	**CCGGGTAA**C	Good	Same
Mcm1	TTWCCcnwwwrGGAAA	35.9	**TTTCC-AAW-RGGAA**A	Good	**TCCnnnnnnGGA**	Good	Same
Rap1	ACACCCATACATTT	30.0	TTW**ACAYCCRTACAY-Y**	Good	**ACCCCA.ACA**	Good	Same
Cbf1	RTCACRTG	24.2	TR**GTCACGTG**	Good	**GTCACGTG**	Good	Same
Fkh2	TTGTTTACST	20.7	**TTGTTTAC-T**WTT	Good	**TGTTTTAC..TT**	Good	Same
Swi4	CRCGAAAA	19.9	CSMRR**CGCGAAAA**	Good	CAA**CRCGAAAA**	Good	Same
Mbp1	ACGCGT	19.6	G-RR-A-**ACGCGT-R**		A**ACGCGT**CG	Good	Better (+)
Ste12	RTGAAACA	17.8	GSAASRR-**TGA**TRAWGYA		**YTGAAACA**	Good	Better (+)
Gal4	CGGnnnnnnnnnnnCCG	16.1	**CGG**M---CW-Y--C**CCG**		**CGGnnnnnnnnnnnCCGA**	Good	Better (+)
Swi6	ACGCGT	15.6	**WCGCGT**CGCGTY-C	Good	**ACGCGT**	Good	Same
Pho4	CACGTG	14.2	TTGTA**CACTT**YGTTT		CG**CACGTG**	Good	Better (+)
Hsf1	TTCTAGAA	14.1	TY**TTCYAGAA--TTC**Y	Good	G**TTCTAGAA**nn**TTC**nnG	Good	Same
Dig1	RTGAAACA	13.6	CCY**TG-AY**TTCW-CTTC		**TGAAACR**	Good	Better (+)
Ino4	CATGTGAAat	13.4	G..G**CATGTGAAAA**	Good	G...**CATGTGAA**	Good	Same
Fkh1	TTGTTTACST	13.2	C**YTRTTTAY**-WTT	Good	**TGTTTAC**	Good	Same
Leu3	CCGGNNCCGG	13.1	G**CCGG**TMM**CGS**YC--	Good	**CCGGnnnCGG**	Good	Same
Bas1	TGACTC	10.2	CS-CCAATGK--CS		**TGACTCT**A	Good	Better (+)
Swi5	KGCTGR	9.2	CACACACACACACACACA		T**GCTGG**	Good	Better (+)
Hap4	TnRTTGGT	8.5	YC**T-ATTSG-**C-GS		**TGATTGGT**	Good	Better (+)
Rlm1	CTAWWWWTAG	8.4	A-CTSGAAGAAATGCGGT		**CTA..TTTAG**	Good	Better (+)
Ino2	CATGTGAAat	7.4	G**CATGTGRAAA**	Good	**CATGTG**	Good	Same
Met31	AAACTGTGGC	7.0	GCAC**GTG**ATS		**TGTGGC**	Good	Same
Ace2	GCTGGT	5.2	GTGTGTGTGTGTGTG		T**GCTGGT**	Good	Better (+)

[a]Comparison of category-based motifs found for known transcription factors. For every factor whose motif is known, we show the published consensus sequence, the enrichment of that sequence in bound regions, and the motifs discovered by MEME and our category-based motif discovery method. With the additional use of comparative information, we increase the sensitivity and specificity for category-based regulatory motif discovery. The experiments are ordered by the enrichment of the previously known motif in the intergenic regions bound. For strongly enriched motifs, both methods find the correct sequence. However, as the motifs become hidden in increasing noise, the conservation-based method outperforms methods that do not use conservation information.

a quality assessment, the motif discovered by our method, as well as the corresponding category-based score and a quality assessment, and finally, the comparison of our method to MEME. The performance of MEME degraded for less enriched motifs, but our comparison-based method consistently found the correct motif.

3.5.3. New category-based motifs found. We then applied the approach to all 318 gene categories. A total of 181 enriched motifs were identified, with similar motifs arising frequently from different categories. Merging such motifs resulted in 52 distinct motifs, of which 43 were already found by the analyses described above. The remaining nine motifs represent new category-based motifs (Kellis *et al.*, 2003, Table 4.3), including the following.

Three novel motifs are associated with genes that are bound by the transcription factors Rap1, Ste12, and Cin5, respectively. Rap1 is known to bind incomplete or degenerate instances of the published motif, and the new motif may confer additional specificity. The motif associated with Ste12 is the known binding site for the partner transcription factor Tec1, suggesting that Ste12 binding is strongly associated with its partner under the conditions examined. Similarly, the novel motif associated with Cin5 may be that of a partner transcription factor. Three novel motifs are associated with the GO category for carbohydrate transport, fatty-acid oxidation, and glycolysis–glycogenesis, respectively. Three novel motifs are associated with an expression cluster (cluster 37) that includes many genes involved in energy metabolism and stress response.

3.5.4. Motif reuse across multiple categories. Category-based motif discovery contributes only a modest number of additional motifs beyond those found by genomewide analysis, confirming the relatively small number of regulatory motifs in yeast. Such a limited count is surprising, given the large number of coordinately transcribed processes in yeast.

Additionally, the motifs discovered across different categories largely overlap. Each motif was discovered on average in three different categories. This overlap is certainly to be expected between functionally related categories, such as the chromatin IP experiment for Gcn4 and the expression cluster of genes involved in amino acid biosynthesis, as well as the GO annotations for amino acid biosynthesis, all of which are enriched in the Gcn4 motif, the master regulator of amino acid metabolism.

More surprisingly, however, different transcription factors are often enriched in the same motif (which may be due to cooperative binding), and the same motif appears enriched in multiple expression clusters and functional categories. For example, Cbf1, Met4, and Met31 share a motif, and so do Hsf1, Msn2 and Msn4; Fkh1 and Fkh2; Fhl1 and Rap1; Ste12 and Dig1; Swi5 and Ace2; Swi6, Swi4, Ash1 and Mbp1. Also, a single motif involved in environmental stress response is discovered repeatedly in numerous expression clusters and in functional categories ranging from secretion, cell organization and biogenesis, transcription, ribosome biogenesis, and rRNA processing.

Hence, the set of regulatory motifs that are specific to exactly one functional category seems limited. This can hamper category-based motif discovery methods, when a category will rarely be enriched in only one motif and a motif will rarely be enriched in only one category.

Additionally, a category-based approach faces a number of inherent experimental limitations. On one hand, the expression clusters we have used, although constructed over an impressive array of experiments, are still limited to the relatively few experimental conditions characterized in the lab, as compared to the versatility of responses yeast cells undergo in nature. Similarly, the functional categories we used are limited to the few well-characterized processes in yeast, and the molecular function of nearly two thousand genes remains unknown.

A genomewide approach presents a new and powerful paradigm to understanding the dictionary of regulatory motifs. By discovering in an unbiased way the complete set of conserved sequence elements, we now have the building blocks for subsequent analyses of regulation.

An important future direction will be the understanding of the versatility of fine-grain gene regulation given a small number of regulatory motifs. Such a versatility may be rooted in a combinatorial control of gene expression, and understanding motif combinations rather than individual motifs will be crucial in this exploration. Additionally, as chromatin state information becomes available, understanding the large-scale effects of histone modifications and chromatin structure in yeast will prove invaluable in our study of higher eukaryotes.

4. CONCLUSION

Our results show that comparative analysis with closely related species can be invaluable in understanding a genome. Comparisons reveal the way different regions change and the constraints they face, providing clues as to their use. Even in a genome as compact and well studied as that of *S. cerevisiae*, much remains to be learned about the gene content. We found that nearly a tenth of all annotated genes are spurious, adjusted the boundaries of hundreds of genes, and discovered 43 new short genes and 58 novel introns. Moreover, our comparisons have enabled a glimpse into the dynamic nature of the cell, by discovering a complete set of regulatory motifs in a *de novo* way, without knowledge of gene function. Our methods rediscovered most previously known regulatory motifs as well as a number of novel motifs with candidate functions. Such regulatory signals are present within the primary sequence of *S. cerevisiae*, but represent only a small fraction of all intergenic regions. Under the lens of evolutionary conservation, these signals stand out from the nonconserved noise. Hence, in studying any genome, comparative analysis of closely related species can provide the basis for a global understanding of all functional elements.

5. ACKNOWLEDGMENTS

We would like to acknowledge the continuous help of the SGD curators and in particular Mike Cherry, Kara Dolinski, and Dianna Fisk. We thank Tony Lee, Nicola Rinaldi, Rick Young, and Julia Zeitlinger for discussions and for providing prepublication chromatin immunoprecipitation data. We thank Mike Eisen and Audrey Gasch for discussions and for providing prepublication clusters of expression data. We thank Jon Butler, Sarah Calvo, Matt Endrizzi, James Galagan, David Jaffe, Joseph Lehar, Li-Jun Ma, and all the people at the MIT/Whitehead Institute Center for Genome Research for their help and discussions. We thank Ziv Bar-Joseph, John Barnett, Tim Danford, David Gifford, and Tommi Jaakkola in the MIT Computer Science and Artificial Intelligence Laboratory, and Gerry Fink, Ernest Fraenkel, Ben Gordon, Trey Ideker, Sue Lindquist, and Owen Ozier in the Whitehead Institute for their help and discussions.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and Sherlock, G. 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* 25, 25–29.

Bailey, T.L., and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2, 28–36.

Batzoglou, S., Jaffe, D.B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J.P., and Lander, E.S. 2002. ARACHNE: A whole-genome shotgun assembler. *Genome Res.* 12, 177–189.

Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B., and Lander, E.S. 2000. Human and mouse gene structure: Comparative analysis and application to exon prediction. *Genome Res.* 10, 950–958.

Blanchette, M., Schwikowski, B., and Tompa, M. 2002. Algorithms for phylogenetic footprinting. *J. Comp. Biol.* 9, 211–223.

Blanchette, M., and Tompa, M. 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.* 12, 739–748.

Blandin, G., Durrens, P., Tekaia, F., Aigle, M., Bolotin-Fukuhara, M., Bon, E., Casaregola, S., de Montigny, J., Gaillardin, C., Lepingle, A., Llorente, B., Malpertuy, A., Neuveglise, C., Ozier-Kalogeropoulos, O., Perrin, A., Potier, S., Souciet, J., Talla, E., Toffano-Nioche, C., Wesolowski-Louvel, M., Marck, C., and Dujon, B. 2000. Genomic exploration of the hemiascomycetous yeasts: 4. The genome of Saccharomyces cerevisiae revisited. *FEBS Lett.* 487, 31–36.

Clark, T.A., Sugnet, C.W., and Ares Jr., M. 2002. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science* 296, 907–910.

Dujon, B., Alexandraki, D., Andre, B., Ansorge, W., Baladron, V., Ballesta, J.P., Banrevi, A., Bolle, P.A., Bolotin-Fukuhara, M., Bossier, P., *et al.* 1994. Complete DNA sequence of yeast chromosome XI. *Nature* 369, 371–378.

Dunham, I. 2000. The gene guessing game. *Yeast* 17, 218–224.

Dwight, S.S., Harris, M.A., Dolinski, K., Ball, C.A., Binkley, G., Christie, K.R., Fisk, D.G., Issel-Tarver, L., Schroeder, M., Sherlock, G., Sethuraman, A., Weng, S., Botstein, D., and Cherry, J.M. 2002. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucl. Acids Res.* 30, 69–72.

Fitch, W.M. 1970. Distinguishing homologous from analogous proteins. *Syst. Zool.* 19, 99–113.

Fitch, W.M. 1995. Uses for evolutionary trees. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 349, 93–102.

Gasch, A.P., and Eisen, M.B. 2002. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol.* 3, RESEARCH0059.

Gelfand, M.S., Koonin, E.V., and Mironov, A.A. 2000. Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. *Nucl. Acids Res.* 28, 695–705.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S.G. 1996. Life with 6000 genes. *Science* 274, 546, 563–547.

Grundy, W.N., Bailey. T.L., Elkan, C.P., and Baker, M.E. 1997. Meta-MEME: Motif-based hidden Markov models of protein families. *Comput. Appl. Biosci.* 13, 397–406.

Hampson, S., Kibler, D., and Baldi, P. 2002. Distribution patterns of over-represented k-mers in non-coding yeast DNA. *Bioinformatics* 18, 513–528.

Harrison, P.M., Kumar, A., Lang, N., Snyder, M., and Gerstein, M. 2002. A question of size: the eukaryotic proteome and the problems in defining it. *Nucl. Acids Res.* 30, 1083–1090.

Hughes, J.D., Estep, P.W., Tavazoie, S., and Church, G.M. 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J. Mol. Biol.* 296, 1205–1214.

Hurst, L.D. 2002. The Ka/Ks ratio: Diagnosing the form of sequence evolution. *Trends Genet.* 18, 486.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature*, 860–921.

International Mouse Genome Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562.

Jacobs Anderson, J.S., and Parker, R. 2000. Computational identification of cis-acting elements affecting post-transcriptional control of gene expression in Saccharomyces cerevisiae. *Nucl. Acids Res.* 28, 1604–1617.

Jaffe, D.B., Butler, J., Gnerre, S., Mauceli, E., Lindblad-Toh, K., Mesirov, J.P., Zody, M.C., and Lander, E.S. 2003. Whole-genome sequence assembly for Mammalian genomes: Arachne 2. *Genome Res.* 13, 91–96.

Jiao, K., Nau, J.J., Cool, M., Gray, W.M., Fassler, J.S., and Malone, R.E. 2002. Phylogenetic footprinting reveals multiple regulatory elements involved in control of the meiotic recombination gene, REC102. *Yeast* 19, 99–114.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423, 241–254.

Kowalczuk, M., Mackiewicz, P., Gierlik, A., Dudek, M.R., and Cebrat, S. 1999. Total number of coding open reading frames in the yeast genome. *Yeast* 15, 1031–1034.

Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wootton, J.C. 1993. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science* 262, 208–214.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K., and Young, R.A. 2002. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 298, 799–804.

Levy, S., and Hannenhalli, S. 2002. Identification of transcription factor binding sites in the human genome sequence. *Mamm. Genome* 13, 510–514.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S.L., and Quackenbush, J. 2000. Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat. Genet.* 25, 239–240.

Liu, X., Brutlag, D.L., and Liu, J.S. 2001. BioProspector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, 127–138.

Loots, G.G., Ovcharenko, I., Pachter, L., Dubchak, I., and Rubin, E.M. 2002. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.* 12, 832–839.

McCue, L., Thompson, W., Carmack, C., Ryan, M.P., Liu, J.S., Derbyshire, V., and Lawrence, C.E. 2001. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucl. Acids Res.* 29, 774–782.

McGuire, A.M., Hughes, J.D., and Church, G.M. 2000. Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res.* 10, 744–757.

Mewes, H.W., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S., and Frishman, D. 1999. MIPS: A database for genomes and protein sequences. *Nucl. Acids Res.* 27, 44–48.

Oeltjen, J.C., Malley, T.M., Muzny, D.M., Miller, W., Gibbs, R.A., and Belmont, J.W. 1997. Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res.* 7, 315–329.

Pennacchio, L.A., and Rubin, E.M. 2001. Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.* 2, 100–109.

Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M. 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* 16, 939–945.

Sharp, P.M., and Li, W.H. 1987. The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucl. Acids Res.* 15, 1281–1295.

Stormo, G.D. 2000. DNA binding sites: Representation and discovery. *Bioinformatics* 16, 16–23.

Tatusov, R.L., Koonin, E.V., and Lipman, D.J. 1997. A genomic perspective on protein families. *Science* 278, 631–637.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V. 2001. The COG database: New developments in phylogenetic classification of proteins from complete genomes. *Nucl. Acids Res.* 29, 22–28.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., and Church, G.M. 1999. Systematic determination of genetic network architecture. *Nat. Genet.* 22, 281–285.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* 22, 4673–4680.

Tompa, M. 2001. Identifying functional elements by comparative DNA sequence analysis. *Genome Res.* 11, 1143–1144.

Velculescu, V.E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M.A., Bassett Jr., D.E., Hieter, P., Vogelstein, B., and Kinzler, K.W. 1997. Characterization of the yeast transcriptome. *Cell* 88, 243–251.

Wolfe, K.H., and Shields, D.C. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387, 708–713.

Wood, V., Rutherford. K.M., Ivens, A., Rajandream, M.-A., and Barrell, B. 2001. A re-annotation of the *Saccaromyces cerevisiae* genome. *Comparative and Functional Genomics* 2, 143–154.

Zhang, M.Q. 1999. Promoter analysis of co-regulated genes in the yeast genome. *Comput. Chem.* 23, 233–250.

Zhu, J., and Zhang, M.Q. 1999. SCPD: A promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics* 15, 607–611.

Address correspondence to:
Eric S. Lander
Department of Biology
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139

and

Bonnie Berger
Room 2-373
Mathematics Department
MIT
Cambridge, MA 02139

E-mail: lander@wi.mit.edu **and** bab@mit.edu

Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*

Manolis Kellis[1,2], **Bruce W. Birren**[1] **& Eric S. Lander**[1,3]

[1]*The Broad Institute, Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02138, USA*
[2]*MIT Computer Science and Artificial Intelligence Laboratory, and* [3]*Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02139, USA*

Whole-genome duplication followed by massive gene loss and specialization has long been postulated as a powerful mechanism of evolutionary innovation. Recently, it has become possible to test this notion by searching complete genome sequence for signs of ancient duplication. Here, we show that the yeast *Saccharomyces cerevisiae* arose from ancient whole-genome duplication, by sequencing and analysing *Kluyveromyces waltii*, a related yeast species that diverged before the duplication. The two genomes are related by a 1:2 mapping, with each region of *K. waltii* corresponding to two regions of *S. cerevisiae*, as expected for whole-genome duplication. This resolves the long-standing controversy on the ancestry of the yeast genome, and makes it possible to study the fate of duplicated genes directly. Strikingly, 95% of cases of accelerated evolution involve only one member of a gene pair, providing strong support for a specific model of evolution, and allowing us to distinguish ancestral and derived functions.

Genomic duplication has been proposed as an advantageous path to evolutionary innovation[1], because duplicated genes can supply genetic raw material for the emergence of new functions through the forces of mutation and natural selection. Such duplication can involve individual genes, genomic segments or whole genomes. Whole-genome duplication (WGD) is a particularly intriguing but poorly understood situation. In principle, coordinate duplication of an entire genome may allow for large-scale adaptation to new environments. However, polyploidy comes at the cost of major genomic instability[2], which persists until the genome returns to functionally normal ploidy through mutation, gene loss and genomic rearrangements. Direct study of such a cataclysmic genomic event may provide major insights into the dynamics of genome evolution and the emergence of new functions.

In the yeast *S. cerevisiae*, subtle analysis of the genomic locations of paralogues revealed the existence of ancestral duplication blocks, but their origin has remained controversial[3–17]. Wolfe and colleagues interpreted the presence and distribution of such regions in the *S. cerevisiae* genome as supporting a model of WGD[3,4]. Others have challenged this conclusion, noting that the WGD model is based on a small set of yeast genes (~8%), explains only a minority of the gene redundancy in the genome and defines putative sister regions covering less than 50% of the genome. An alternative model that has been proposed is that the duplicated segments arose via independent local duplication events[8–15]. Recent reports assert that low-coverage sequence data from related yeast species supports this alternative model[14,15]; Wolfe and colleagues have in turn challenged this analysis[16]. In addition, Wolfe and colleagues postulated that WGD occurred after the divergence of *S. cerevisiae* and *K. lactis*, while Langkjaer *et al.* have recently presented evidence that any putative WGD would necessarily predate this divergence[17].

Here we provide direct evidence of WGD in yeast, by sequencing and analysing a related species whose divergence precedes the duplication event. We show that *S. cerevisiae* arose from complete duplication of eight ancestral chromosomes, and subsequently returned to functionally normal ploidy by massive loss of nearly 90% of duplicated genes in small deletions. These were balanced and complementary in paired regions, preserving at least one copy of virtually each gene in the ancestral gene set. We identify 145 paired regions in *S. cerevisiae*, tiling 88% of the genome and containing 457 duplicated gene pairs.

We then analyse the post-duplication divergence of gene pairs, and show evidence of accelerated evolution in many cases. Strikingly, 95% of cases of accelerated evolution involve only one member of a gene pair, providing strong support for a specific model of evolution[1], and allowing ancestral and derived functions to be distinguished. We find that derived genes tend to be specialized in function, expression and localization, and lose essential aspects of their ancestral function. In addition, we find striking examples of neofunctionalization, including the emergence of silencing from origin-of-replication binding, and the emergence of viral defence mechanisms from translation elongation.

Evidence of whole-genome duplication

The expected signature of genome duplication is illustrated in Fig. 1. Following duplication, sister regions would undergo progressive gene loss by random local deletion: one or the other of the two paralogous copies of each gene would be lost in most cases, with both paralogues being retained if they happen to acquire distinct functions. Eventually, the only residual signature to show that two regions arose from ancestral duplication is the presence of a few paralogous genes in the same order and orientation scattered in a sea of otherwise unrelated genes. Paired regions containing such signatures will be relatively short, because chromosomal rearrangements would disrupt global gene order over time, leaving only weak evidence of ancestral duplication.

The clearest way to prove the existence of an ancient WGD would be to find a yeast species, denoted as Y, that descends directly from a common ancestor along a lineage that diverged before the duplication (Fig. 1). In principle, species Y and *S. cerevisiae* would be related by a 1:2 mapping satisfying several properties: (1) nearly every region in species Y would correspond to two sister regions in *S. cerevisiae*; (2) the two sister regions in *S. cerevisiae* would each contain an ordered subsequence of the genes in the corresponding region of species Y, with each containing roughly half of the genes and the two subsequences interleaving to account for nearly all of the genes; and (3) nearly every region of *S. cerevisiae* would correspond to one region of species Y, and thus be paired to a sister

©2004 Nature Publishing Group

region in *S. cerevisiae*. Below, we show that *K. waltii* matches these criteria.

Genome sequencing and alignment

We sequenced the genome of *K. waltii*, as part of a larger fungal genome programme. We note that *Kluyveromyces* is a polyphyletic genus, and in particular, *K. waltii* is closer to *S. cerevisiae* than the previously studied *K. lactis*. In light of this, *K. waltii* has been proposed to be renamed *Lachancea waltii*[18], although we use the more familiar name here. We used whole-genome shotgun sequencing to obtain eightfold sequence coverage and 42-fold physical coverage, and assembled the information into essentially eight complete chromosomes of total length 10.7 Mb (see Methods).

We annotated the draft genome sequence to identify 5,230 likely protein-coding genes (see Methods), notably fewer than the 5,714 genes in *S. cerevisiae*[19]. The genome also contains 240 transfer RNA genes (36 fewer than *Saccharomyces* species) and 60 Ty elements (primarily TyB). It also contains several telomeric gene families that are probably involved in adaptation to changing environments, including families of membrane proteins (51 members), hexose transporters (20 members), and flocculins (50 members, a notable expansion relative to *S. cerevisiae*). Finally, 7% of genes show no detectable protein similarity to *S. cerevisiae*. The genome sequence has been deposited in public databases (GenBank/DDBJ/EMBL), and the sequence and annotation are available from our website (http://www.broad.mit.edu/seq/YeastDuplication) and as Supplementary Information.

We compared the genomes of *K. waltii* and *S. cerevisiae* to identify orthologous regions. We first identified closely matching genes between the species and then identified blocks of conserved synteny—that is, regions containing numerous matching genes in the same order across the two species (see Methods). In contrast to the 1:1 mapping seen for close relatives[19], most local regions in *K. waltii* mapped to two regions in *S. cerevisiae*, with each containing matches to only a subset of the *K. waltii* genes.

To quantify this observation, we developed a computational algorithm to parse *K. waltii* into blocks of doubly conserved synteny (DCS)—defined as maximal regions that map across their entire length to two distinct regions in *S. cerevisiae* (see Methods). Briefly, we labelled the genes within each synteny block, and used the genes that belonged to multiple synteny blocks to detect regions of duplicate synteny. We then used the duplicate mapping in DCS blocks to define sister regions in the *S. cerevisiae* genome (see Methods).

Complete duplication

We identified a total of 253 DCS blocks, containing 75% of *K. waltii* genes and 81% of *S. cerevisiae* genes. The cross-species mapping of DCS blocks is shown in Fig. 2a, with a detailed view of a region of chromosome 1. DCS blocks tile 85% of each *K. waltii* chromosome in the pattern expected for WGD. Detailed views of all regions are shown in Supplementary Figs S5.

The typical gene falls in a DCS block containing 27 genes, and the

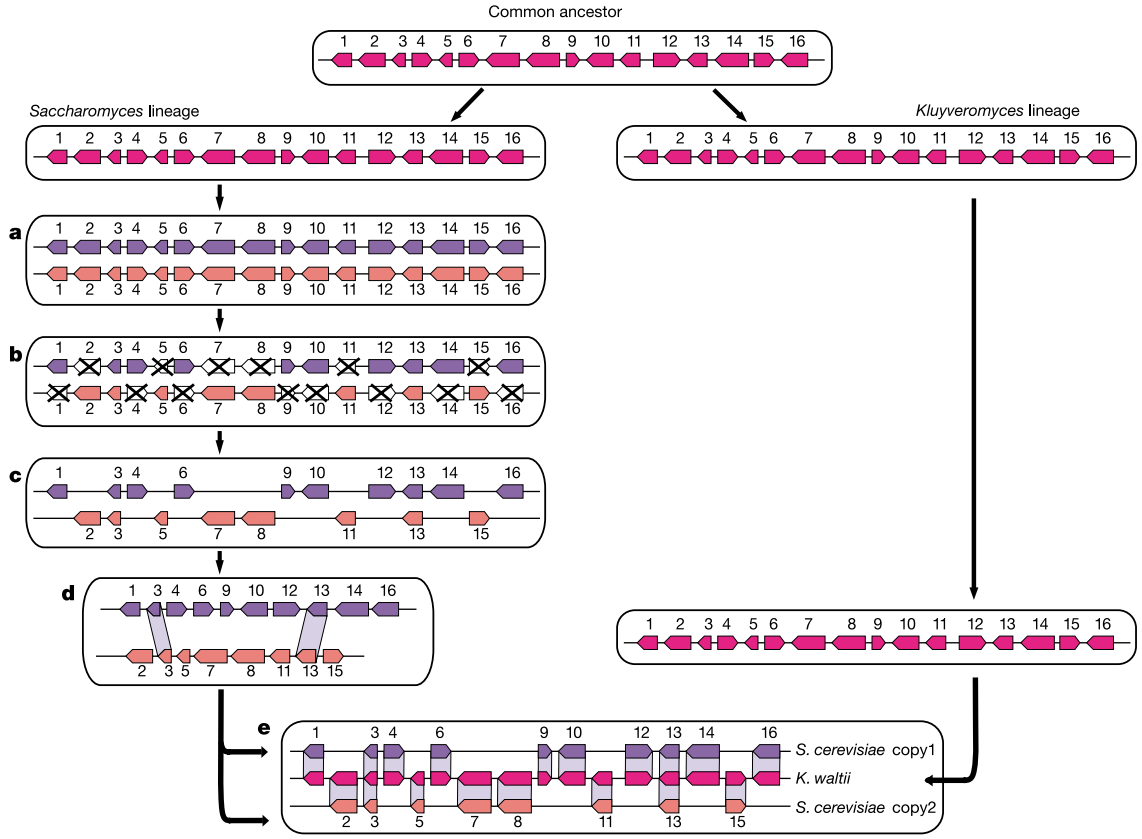


Figure 1 Model of WGD followed by massive gene loss predicts gene interleaving in sister regions. **a**, After divergence from *K. waltii*, the *Saccharomyces* lineage underwent a genome duplication event, creating two copies of every gene and chromosome. **b**, The vast majority of duplicated genes underwent mutation and gene loss. **c**, Sister segments retained different subsets of the original gene set, keeping two copies for only a small minority of duplicated genes, which were retained for functional purposes. **d**, Within *S. cerevisiae*, the only evidence comes from the conserved order of duplicated genes (numbered 3 and 13) across different chromosomal segments; the intervening genes are unrelated. **e**, Comparison with *K. waltii* reveals the duplicated nature of the *S. cerevisiae* genome, interleaving genes from sister segments on the basis of the ancestral gene order.

©2004 Nature Publishing Group

largest such block contains 81 genes. DCS blocks are separated by small segments (~3 genes on average) that appear to match only one conserved region in *S. cerevisiae*; it is likely that the second paralogous segment cannot be confidently recognized because it has lost too many genes or undergone rearrangement (see Methods). Finally, ~1% of the *K. waltii* genome lies in segments that match three or more regions in *S. cerevisiae*; the additional matching regions result mostly from slight local rearrangements (see Methods).

Complete interleaving of genes

Within DCS blocks, the interleaving of genes shows essentially

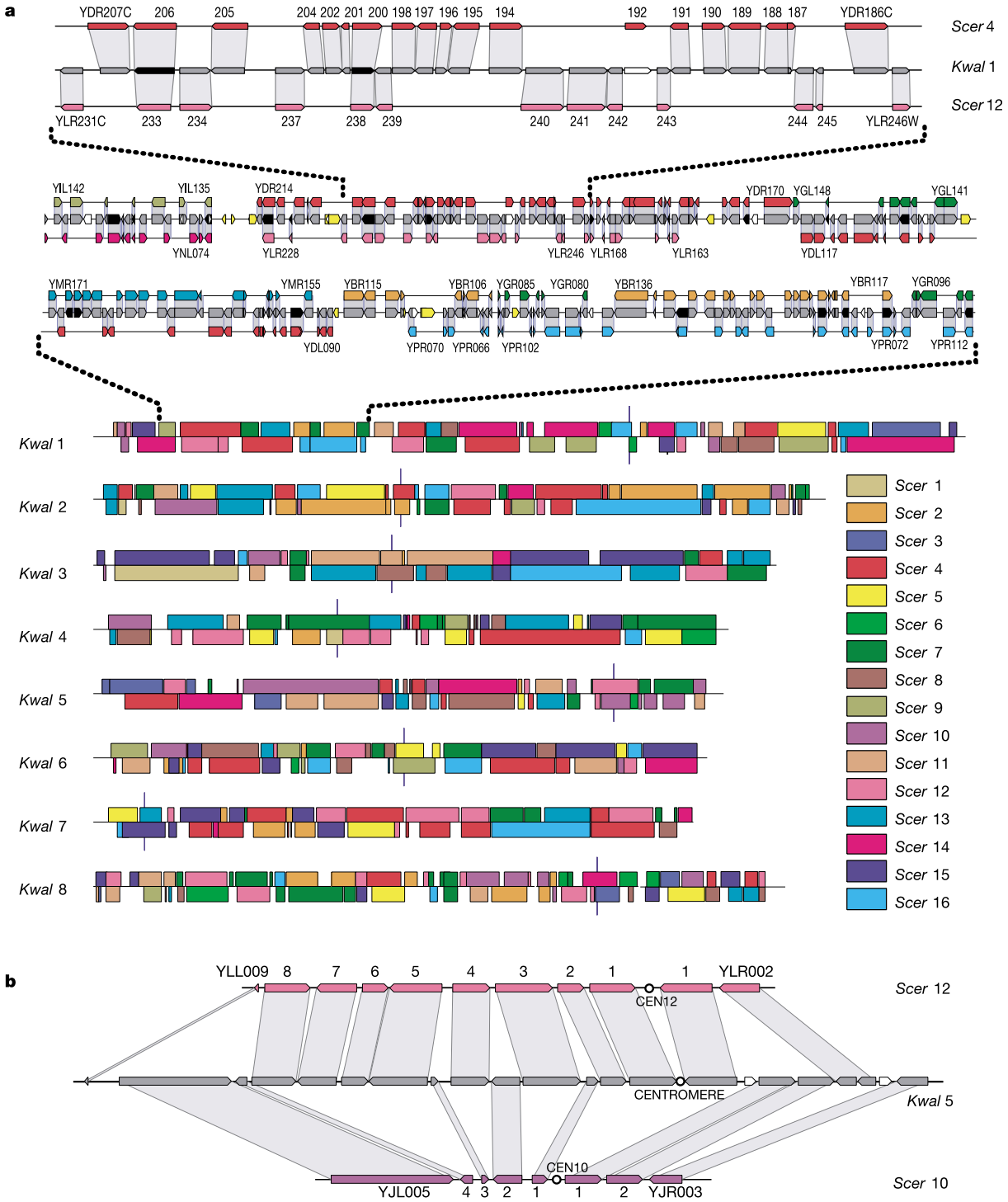


Figure 2 Gene and region correspondence with *K. waltii* reveals WGD. **a**, Each region of *K. waltii* (coloured by number of matches: white, 0; grey, 1; black, 2; yellow, > = 3) shows conserved gene order with two regions of *S. cerevisiae* (coloured by chromosome number). Spacing between *S. cerevisiae* genes is set to match *K. waltii* chromosomal positions. Vertical blue bars denote centromeres. **b**, DCS region showing duplicate mapping of centromeres (black circles). All sixteen *S. cerevisiae* centromeres show such duplicate mapping with *K. waltii* centromeres. This DCS region also illustrates that we can reliably recognize anciently duplicated segments even in the absence of any remaining two-copy genes. Evidence of WGD comes from gene interleaving and 2:1 mapping with orthologous *K. waltii* segments. The segments containing intervening genes are deleted, resulting in condensed sister regions. *Kwal*, *K. waltii*; *Scer*, *S. cerevisiae*.

©2004 Nature Publishing Group

complete duplicate mapping (Fig. 2b and Supplementary Information). In a typical DCS block, 90% of *K. waltii* genes have matches in at least one of the two *S. cerevisiae* regions (and on average ~12% have matches in both regions). Similarly, in each DCS block, 90% of genes from both *S. cerevisiae* regions have matches within the single corresponding region of *K. waltii*, and they are interleaved onto *K. waltii* while preserving gene order and orientation. The mapping fits the pattern expected for WGD followed by massive gene loss.

Complete pairing of *S. cerevisiae*

Based on the cross-species correspondence, we identified sister regions within *S. cerevisiae* (Fig. 3). We used the DCS blocks to define 253 initial sister regions for which the ancestral gene order was preserved in all three matching segments (*K. waltii* and the two sister regions in *S. cerevisiae*). We then merged consecutive sister regions when both copies in *S. cerevisiae* preserved the ancestral gene order, but were apparently separated owing to a rearrangement in *K. waltii*. This resulted in 145 sister regions in *S. cerevisiae* covering 88% of the genome.

Many of the paired regions could not be recognized without the relationship to *K. waltii*. For example, 47 DCS regions contain no duplicated genes whatsoever and the evidence of duplication comes solely from the interleaving of genes (Fig. 2b).

The chromosome structure of the two genomes provides further evidence of WGD. Indeed, the sixteen centromeres of *S. cerevisiae* are paired in a 2:1 manner with the eight centromeres of *K. waltii* (Fig. 2b and Methods).

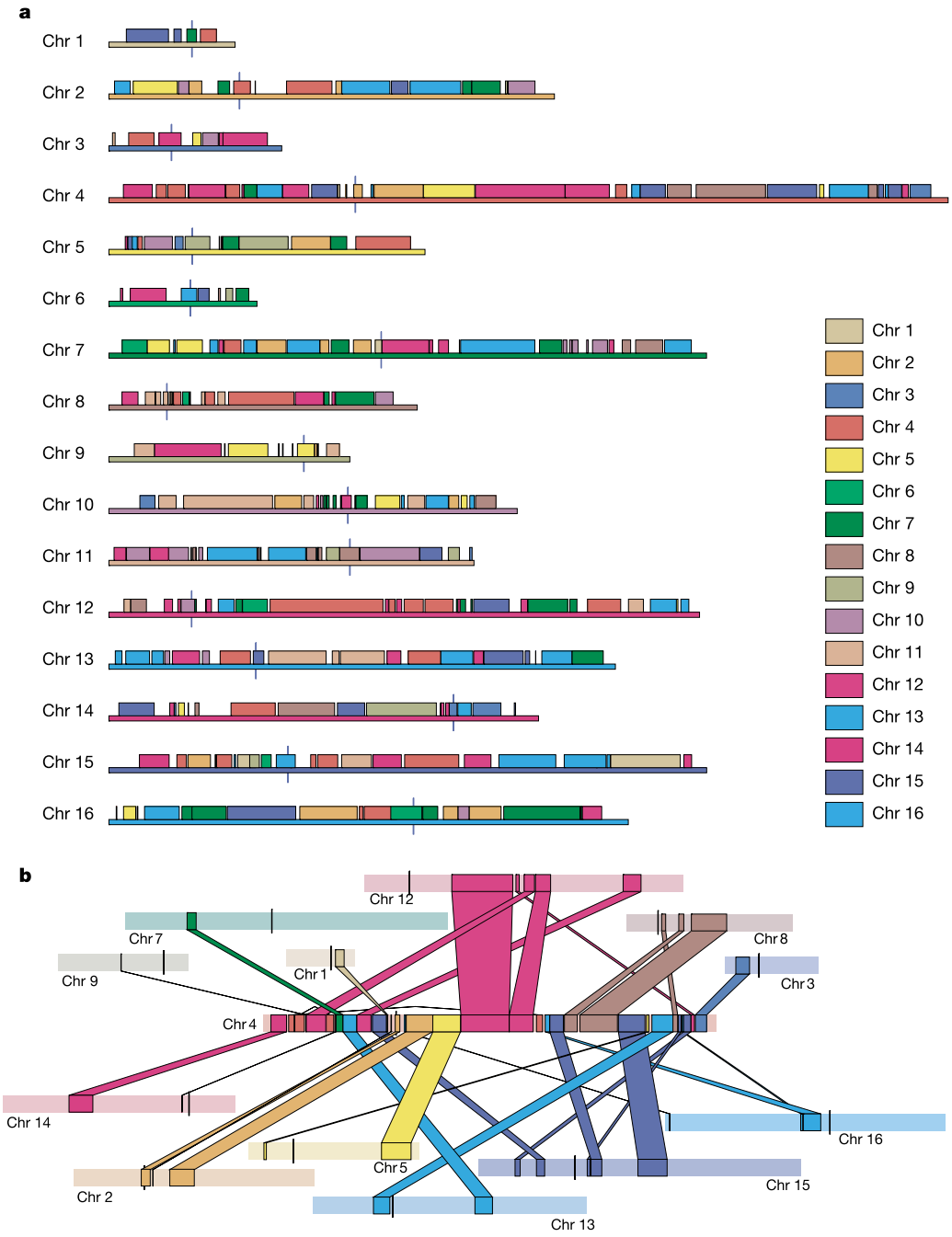


Figure 3 Duplicated blocks in *S. cerevisiae*. **a**, The duplicate mapping in sister regions extends to tile each *S. cerevisiae* chromosome, revealing complete duplication. Blue vertical bars denote centromeres. **b**, Detailed mapping of chromosome 4 with sister regions in other chromosomes.

©2004 Nature Publishing Group

Summary

We conclude that a WGD event occurred in the *Saccharomyces* lineage after the divergence from *K. waltii* and that *S. cerevisiae* is therefore a degenerate tetraploid. WGD may have occurred either at the haploid or the diploid stage and either by endo-duplication (auto-polyploidy) or fusion of two close relatives (allo-polyploidy).

Evolutionary analysis

The relationship between *K. waltii* and *S. cerevisiae* provides the first comparison across an ancient WGD event and offers the opportunity to study the long-term fate of a genome after duplication.

Pattern of gene loss

The WGD event doubled the number of chromosomes in the *Saccharomyces* lineage, but subsequent gene-loss events led to the current *S. cerevisiae* genome, which is only 13% larger than *K. waltii* and contains only 10% more genes. The polyploid genome returned to functional normal ploidy, not by meiosis or chromosomal loss, but instead by a large number of deletion events. We can infer that 12% of the paralogous gene pairs were retained in each DCS block, and the remaining 88% of paralogous genes were lost.

In principle, gene loss can occur by large segmental deletions or individual gene deletions, and can be balanced between the two sisters or act primarily on one of them. Analysis of DCS blocks revealed that gene loss occurred by many small deletions (the average size of a lost segment is two genes), and was typically balanced between the two sister regions (average balance 57% to 43%). Additionally, a large number of rearrangement events have shuffled the duplicated chromosomes; the intergenic regions at the breakpoints seem to be associated with tRNA genes.

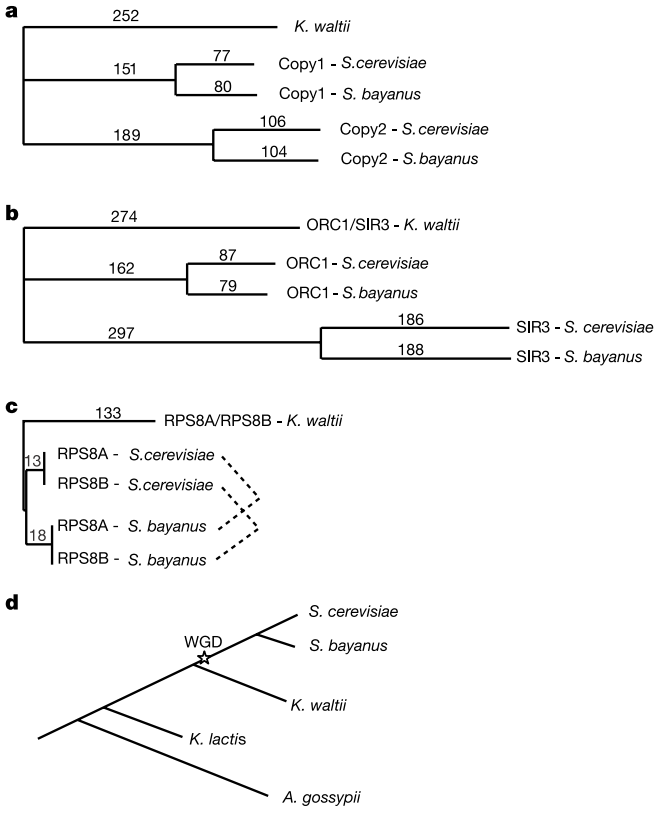


Figure 4 Divergence of duplicated gene pairs. Branches show number of substitutions per thousand amino acids. Evolutionary trees are rooted at the divergence of three species. **a**, Average protein divergence of all 457 gene pairs that arose by WGD. The faster-evolving paralogue is arbitrarily designated as copy 2 for each pair. **b**, Example of a gene showing accelerated protein divergence (1 of 72 cases). Ancestral and derived gene function can be inferred by comparison to *K. waltii*. In this case, the origin-of-replication recognition complex protein Orc1 is inferred to be ancestral and the silencing protein Sir3 is inferred to be derived. **c**, Example of duplicated gene pairs that have undergone recent gene conversion (1 of 60 cases). Comparison with *S. bayanus* shows that recent gene conversion events occurred both in *S. cerevisiae* and in *S. bayanus* lineages. Dotted lines connect orthologous genes in *S. cerevisiae* and in *S. bayanus*. **d**, Phylogeny and relative time of WGD. Estimated tree lengths are as reported in ref. 45.

Accelerated protein divergence

Most interestingly, we can study the evolution of the 457 gene pairs that arose by WGD. The synteny information allows us to unambiguously establish the ancestry for these genes pairs and distinguish them from many gene pairs in the yeast genome that arose by local duplication events (see Methods). This allows us to study their evolution with respect to each other and the non-duplicated *K. waltii* orthologue.

Two alternatives have been proposed for post-duplication divergence of duplicated gene pairs. Ohno has hypothesized that after duplication, one copy would preserve the original function and the other copy would be free to diverge[1]. However, this model has been contested by others, who argued that both copies would diverge more rapidly and acquire new functions[20–22]. We calculated divergence rates for the 457 pairs, using sequence information from *K. waltii*, *S. cerevisiae* and the related yeast *S. bayanus* (Fig. 4d). We searched for instances of unusually rapid or slow evolution.

We found that 76 of the 457 gene pairs (17%) show accelerated protein evolution relative to *K. waltii*, defined as instances in which the amino acid substitution rate along one or both of the *S. cerevisiae* branches was at least 50% faster than the rate along the *K. waltii* branch (Tables 1–3). We note that this calculation is conservative and may miss some cases of accelerated evolution (see Methods). These genes were heavily biased towards protein kinases (13 pairs, *P*-value 10^{-8}) and regulatory proteins (8 pairs, including the cell-cycle transcription factors Swi5 and Ace2, and the filamentation factors Phd1 and Sok2), and were generally involved in metabolism (38 pairs, *P*-value 10^{-18}) and cell growth (32 pairs, *P*-value 10^{-10}).

Ancestral and derived functions

Strikingly, in nearly every case (95%), accelerated evolution was confined to only one of the two paralogues. This strongly supports the model in which one of the paralogues retained an ancestral function while the other, relieved of this selective constraint, was free to evolve more rapidly[1].

For these 76 gene pairs showing accelerated evolution with respect to *K. waltii*, we can infer that the slowly evolving paralogue has probably retained the ancestral gene function and that the rapidly evolving paralogue probably acquired a derived function after duplication. In fact, the inference of ancestral and derived status can be extended to a total of 115 gene pairs consisting of those for which one *S. cerevisiae* paralogue has evolved at least 50% faster than the other *S. cerevisiae* paralogue.

This analysis thus identifies newer functions that evolved from older ones. We can infer, for example, that the silencing function of Sir3 (which acts in telomeres and mating type cassettes)[23] derived from the origin-of-replication binding function of Orc1 (ref. 24). Similarly, the anti-viral function of Ski7 that recognizes and represses non-polyadenylated messenger RNA was probably derived from the translation-elongation function of Hbs1 (ref. 25).

In many cases, the derived paralogues are specialized in their cellular localization or temporal expression. Acc1 (acetyl-CoA carboxylase) gave rise to Hfa1, which now functions in the mitochondrion[26]. Skt5, involved in cell-wall biosynthesis, gave rise to Shc1, whose function is restricted to spore formation[27]. And the pleiotropic transcription factor Skn7, involved in a number of stress responses, gave rise to the transcription factor Hms2, involved in pseudohyphal growth under nitrogen depletion[28].

©2004 Nature Publishing Group

Table 1 **Rapid protein divergence**

Slowly evolving gene (ancestral function)		Rapidly evolving gene (derived function)		Kwal rate	Relative rates	
Gene 1	Description 1	Gene 2	Description 2		Rate1 (%)	Rate2 (%)
RNR2	Small subunit of ribonucleotide reductase	*RNR4*	Ribonucleotide reductase	0.069	146	519
MCK1	Serine/threonine/tyrosine protein kinase	*YGK3*	Yeast homologue of mammalian glycogen synthase kinase 3	0.102	110	479
SEC14	Phosphatidylinositol transfer protein	*SFH1*	Sec14p homologue with highest sequence similarity but lowest functional redundancy	0.061	149	462
YSH1	Subunit of polyadenylation factor I (PF I)	*SYC1*	C-terminal portion of YSH1	0.130	112	380
GCS1	ADP-ribosylation factor GTPase-activating protein (ARF GAP)	*SPS18*	Transcription factor	0.172	112	309
FEN1	Probable subunit of 1,3-beta-glucan synthase	*ELO1*	Elongation enzyme 1	0.111	106	293
CET1	mRNA capping enzyme beta subunit (80 kDa)	*CTL1*	RNA triphosphatase, no capping function	0.231	76	263
EGD1	Regulator of pol II-transcribed genes	*BTT1*	Negative regulator of RNA polymerase II	0.146	109	240
SKN7	Transcription factor involved in oxidative stress	*HMS2*	Heat-shock transcription factor homologue	0.243	103	229
HBS1	Translation elongation factor	*SKI7*	Anti-viral protein	0.241	116	223
SEC24	Vesicle coat component	*SFB2*	Putative zinc finger protein	0.130	114	203
IFH1	Overexpression interferes with silencing at telomeres / HM, Interacts with FHL1	YDR223W	Hypothetical ORF	0.269	91	189
SKT5	Probable Ca^{2+} binding membrane protein (prenylated)	*SHC1*	Sporulation-specific homologue of csd4	0.242	93	188
ORC1	Largest subunit of origin recognition complex	*SIR3*	Silencing regulator HML, HMR, telomeres	0.274	91	177
BUB1	Serine/threonine protein kinase required for arrest in loss of microtubule function	*MAD3*	Spindle checkpoint complex subunit	0.249	133	172
ECM18	Involved in cell wall biogenesis	*ICT1*	Increased copper tolerance	0.224	98	161

Accelerated protein divergence of duplicated gene pairs (16 of 76 pairs). Kwal, amino acid substitution rate along *K. waltii* lineage. Rate1 and Rate2, substitution rates for Gene1 and Gene2 as a percentage of Kwal.

The functional distinction between the derived and ancestral paralogue is confirmed by the phenotype of deletion mutants when grown in rich medium conditions. For each of the 115 gene pairs, we examined the reported results of systematic deletion experiments[29,30]. Strikingly, deletion of the ancestral paralogue was lethal in 18% of cases (similar to the overall average for all *S. cerevisiae* genes), whereas deletion of the derived paralogue was never lethal. The derived paralogue is thus not essential under these conditions, either because it does not function in rich medium or because the ancestral paralogue can complement its function. Moreover, we infer that along with possibly gaining a new function, the derived copy has lost some essential aspect of its function, and cannot typically complement deletion of the ancestral gene.

Changes in codon usage

We also found that 32 of the 457 gene pairs show accelerated nucleotide, but not protein, evolution (Table 2). For example, the transcription elongation genes *YEF3* and *HEF3* encode very similar proteins and show a rate of protein change similar to that of *K. waltii*. However, *HEF3* shows twofold acceleration in degenerate third-codon positions (relative to both *YEF3* and *K. waltii*), an unusual pattern of codon usage (codon adaptation index[31] is 0.17 as compared to 0.78 for *YEF3*) and a much lower level of expression[32]. The same phenomenon is seen for the pyruvate kinase genes *CDC19* and *PYK2*. The accelerated divergence appears to reflect a relief from selection in the codon usage of the accelerated copy, and the presence of both gene copies may allow the cell to sense changes in metabolic state. Interestingly, *PYK2* is expressed preferentially in conditions of low glycolytic flux[33].

Gene conversion

At the other end of the spectrum, 60 of the 457 pairs show decelerated protein evolution, defined as instances where one or both copies evolved at least 50% slower than *K. waltii* (Table 3). These cases often involve proteins known to be highly constrained, such as ribosomal proteins (25 pairs), histone proteins (2 pairs) and translation initiation/elongation factors (4 pairs). In the vast majority of cases (90%), the paralogues both show decelerated evolution and tend to be very similar (98% amino acid identity versus 55% for all pairs), suggesting that they may be subject to periodic gene conversion. Indeed, the strong nucleotide similarity in fourfold degenerate codon positions that are not under the same selective pressures for conservation provides evidence supporting gene conversion (>90% nucleotide identity versus 41% for all pairs). Furthermore, in about half of these cases, the two paralogues in *S. cerevisiae* are closer in sequence to each other than either is to its syntenic orthologue in *S. bayanus*, showing that gene conversion has occurred after the relatively recent divergence of the two *Saccharomyces* lineages (Fig. 4c). For strongly conserved gene pairs, the two copies may confer an evolutionary benefit by serving as a repository for gene conversion and thus buffering deleterious changes in either copy.

Pairs with similar rates

The remaining 321 of the 457 gene pairs did not meet our stringent criteria for accelerated or decelerated evolution. In some cases, the functional changes may be similar to those above but subtler. In other cases, the primary functional divergence may be in regulatory sequences. This seems a likely possibility because the extensive gene

Table 2 **Rapid nucleotide divergence**

Gene 1	Gene 2	Description	Codon usage		Protein identity (%)	Rate in fourfold degenerate sites		
			CAI1	CAI2		Rate1	Rate2	Acceleration (%)
CDC19	*PYK2*	Pyruvate kinase	0.89	0.13	93	0.22	0.63	283
ADH1	*ADH5*	Alcohol dehydrogenase	0.81	0.25	91	0.35	0.57	164
YEF3	*HEF3*	Transcription elongation factor	0.78	0.17	92	0.34	0.62	181
GND1	*GND2*	Phosphogluconate dehydrogenase	0.62	0.20	96	0.39	0.64	166
GDH1	*GDH3*	NADP-specific glutamate dehydrogenase	0.59	0.16	95	0.41	0.62	150

Accelerated nucleotide divergence of duplicated gene pairs. CAI, codon adaptation index[31], a measure of typical codon usage between 0 and 1. Protein identity was measured between Gene1 and Gene2. Rate1 and Rate2, nucleotide substitution rate in fourfold degenerate sites for conserved amino acid residues. Acceleration, relative acceleration of Rate2 compared to Rate1.

©2004 Nature Publishing Group

Table 3 **Slowly evolving genes**

Gene 1	Gene 2	Description	Within *S. cerevisiae*		To *K. waltii*	
			All (%)	Fourfold (%)	All (%)	Fourfold (%)
TEF2	*TEF1*	Translational elongation factor EF-1 alpha	100	100	87	74
TIF2	*TIF1*	Translation initiation factor eIF4A	100	99	83	68
RPL1B	*RPL1A*	Ribosomal protein L1B	99	97	88	67
YCL073C	YKR106W	Hypothetical ORF	99	97	51	25
HHF1	*HHF2*	Histone H4 (HHF1 and HHF2 code for identical proteins)	97	92	89	71

Gene conversion events (5 of 60 cases). Nucleotide identity is shown within *S. cerevisiae* (between Gene1 and Gene2) and to *K. waltii* (between Gene2 and *K. waltii*). All, overall nucleotide divergence. Fourfold, nucleotide divergence in fourfold degenerate sites encoding conserved amino-acid residues. Supplementary Information contains complete divergence information for all gene pairs.

loss following WGD would probably juxtapose new combinations of intergenic and genic regions. In still other cases, the gene pairs may have been retained primarily to increase gene dosage, as with the slowly evolving genes.

Discussion

WGD followed by massive gene loss and gene specialization offers an important path for large-scale evolutionary innovation. Compared to multiple independent duplications and divergence of individual genes or segments, WGD may be more efficient and may offer great opportunities for coordinated evolution. Although organisms clearly can undergo WGD resulting in complete polyploids (as evidenced by existing tetraploid species of plants and animals[34–37]), it has been unclear whether WGD can then be followed by massive genome reshaping to yield diploids with expanded gene content.

Our analysis of *K. waltii* definitively proves that *S. cerevisiae* is the descendant of an ancient WGD, as originally proposed by Wolfe[3] on the basis of subtle genomic patterns. Comparison with *K. waltii* allows rigorous study of the many evolutionary innovations that arose in this dramatic event. We note that a similar analysis of the *Ashbya gossypii* genome has been carried out (P. Philippsen, personal communication). The results here suggest that it may also be fruitful to search for similar genomic signatures of ancient WGD in other organisms[38–40]. It will be interesting to see just how far such distant echoes of genomic upheaval may be traced. □

Methods

See the Supplementary Information for assembly contigs and scaffolds (Supplementary Table S1), predicted protein-coding genes (Supplementary Information S2), complete gene correspondence (Supplementary Information S3), a table of genes in syntenic segments (Supplementary Tables S4), gene-by-gene mapping for each chromosome (Supplementary Figs S5), detailed visualization of 253 DCS regions (Supplementary Figs S6), 145 paired regions in *S. cerevisiae* by chromosome (Supplementary Figs S7), alignments of duplicated gene pairs (Supplementary Information S8), protein and nucleotide divergence rates (Supplementary Tables S9) and mapping of centromeres (Supplementary Figs S10), also available from http://www.broad.mit.edu/seq/YeastDuplication.

Strains, sequencing and assembly

The sequence and annotation for *S. cerevisiae* reference strain S288C was obtained from http://www.yeastgenome.org, and updated according to recent comparative work[19]. *K. waltii* strain NCYC2644 was provided by I. Roberts (Norwich Research Park, UK). The WGS sequence was obtained at the Whitehead Institute/MIT Center for Genome Research. We used paired-end sequencing with 7.4-fold coverage in 4-kb plasmid clones and 0.6-fold coverage in 40-kb fosmid clones, providing eightfold sequence coverage and 42-fold physical coverage, with laboratory protocols as described at http://www.broad.mit.edu. We used the Arachne[41,42] program to assemble the resulting sequence information into contigs (continuous blocks of uninterrupted sequence) and scaffolds (contigs linked by paired forward-reverse reads from the same clone). The resulting draft genome sequence has a total length of 10.7 megabases, with ~96% contained in ten large scaffolds containing relatively short sequence gaps (average size 144 base pairs, together comprising only 0.5% of total scaffold length). Seven of the scaffolds appear to correspond to complete chromosomes, with recognizable centromeric and telomeric regions, and the remaining three scaffolds appear to belong to a single chromosome, with the order and orientation of the scaffolds readily inferred from the sequence (see below).

Centromeres and telomeres

The conservation of gene order in syntenic regions surrounding the centromeres allowed us to pinpoint candidate centromere-containing intergenic intervals within *K. waltii*. Sequence analysis of these intervals revealed a centromere structure similar to that of *S. cerevisiae*, containing three elements CDEI–CDEII–CDEIII, where CDEI is TCACCTG|TCANGTG, CDEII is 100–115 nucleotides with AT > 78%, and CDEIII is TTCCGA. This sequence occurs specifically only at the intervals we determined. Each *K. waltii* centromere maps to exactly two *S. cerevisiae* centromeres, as follows: Cen1:Cen7, Cen2:Cen4, Cen3:Cen14, Cen5:Cen9, Cen6:Cen16, Cen8:Cen11, Cen10:Cen12 and Cen13:Cen15. This pairing agrees with previous reports[16]. Telomeric repeat elements were identified in the ends of all scaffolds, except where joins are proposed. On the basis of centromere and telomere locations, paired fosmid ends, and synteny with *S. cerevisiae*, we infer that the three scaffolds together form the eighth *K. waltii* chromosome.

Annotation and gene correspondence

We determined the complete set of all predicted open reading frames (pORFs) in *K. waltii* that start with a methionine and have the potential to encode at least 50 amino acids. We found 6,753 non-overlapping ORFs, of which 5,230 are likely to represent functional protein-coding genes. These contain 4,711 ORFs of at least 150 amino acids, of which 92% show protein similarity to *S. cerevisiae*. From the remaining 2,042 ORFs between 50 and 150 amino acids, 519 show protein similarity to *S. cerevisiae*, and we expect these to encode functional proteins. We estimate that only an additional 45 ORFs at this range have been missed (0.8% of the total), on the basis of the percentage of longer ORFs that show similarity. We then constructed a bipartite graph of protein similarities across the two species, and resolved the gene correspondence based on protein similarity and gene order information using the BUS algorithm[43].

Doubly conserved synteny regions

We established blocks of conserved synteny by clustering matches of neighbouring genes that are less than 20 kb apart in each genome. We found 353 synteny blocks containing at least three conserved genes each, and matching a total of 4,590 *S. cerevisiae* genes and 4,139 *K. waltii* ORFs. We treated every synteny block as defining an interval on the *K. waltii* genome, and we marked all genes whose coordinates were contained within that interval. We then searched for *K. waltii* genes that were marked as belonging to multiple blocks, thus defining regions of overlap between synteny blocks (DCS regions).

Establishing ancestry of duplicated genes

We determined the ancestry of duplicated gene pairs solely on the basis of gene order information in DCS regions. The difficulty of reliably establishing ancestry based on amino acid divergence, especially in the presence of accelerated evolution, is illustrated by previous reports that have erroneously established WGD as preceding divergence from *K. lactis*[17] (the correct topology is shown in Fig. 4d), or erroneously inferred that GRS2 would represent the ancestral form of the GRS1/GRS2 pair[44] (in fact, GRS2 is the derived copy, and shows fourfold amino acid acceleration). In both cases, variation in protein divergence rates has misrepresented the true phylogenetic relationships.

Establishing accelerated divergence

Having established gene ancestry based on gene order information, we used amino acid and nucleotide divergence rates to detect cases of accelerated evolution. In seven of the 457 gene pairs, one or both copies in *S. cerevisiae* are split into multiple ORFs and were not used. We used CLUSTALW to construct protein, nucleotide, and codon-aware nucleotide alignments. We rooted three-way trees at the central node, whereas the true root lies on the *K. waltii* branch, and hence our criteria for accelerated divergence are conservative because they overestimate the *K. waltii* branch length. This is especially true in the case of gene-conversion events, and additional pre-duplication relatives will be necessary to estimate divergence between duplication and gene conversion.

Received 17 December 2003; accepted 19 January 2004; doi:10.1038/nature02424.
Published online 7 March 2004.

1. Ohno, S. *Evolution by Gene Duplication* (Allen and Unwin, London, 1970).
2. Mayer, V. W. & Aguilera, A. High levels of chromosome instability in polyploids of *Saccharomyces cerevisiae*. *Mutat. Res.* **231,** 177–186 (1990).
3. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387,** 708–713 (1997).
4. Seoighe, C. & Wolfe, K. H. Updated map of duplicated regions in the yeast genome. *Gene* **238,** 253–261 (1999).
5. Melnick, L. & Sherman, F. The gene clusters ARC and COR on chromosomes 5 and 10, respectively, of *Saccharomyces cerevisiae* share a common ancestry. *J. Mol. Biol.* **233,** 372–388 (1993).
6. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274,** 546, 563–567 (1996).

©2004 Nature Publishing Group

7. Philippsen, P. *et al.* The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XIV and its evolutionary implications. *Nature* **387,** 93–98 (1997).
8. Mewes, H. W. *et al.* Overview of the yeast genome. *Nature* **387,** 7–65 (1997).
9. Coissac, E., Maillier, E. & Netter, P. A comparative study of duplications in bacteria and eukaryotes: the importance of telomeres. *Mol. Biol. Evol.* **14,** 1062–1074 (1997).
10. Friedman, R. & Hughes, A. L. Gene duplication and the structure of eukaryotic genomes. *Genome Res.* **11,** 373–381 (2001).
11. Hughes, T. R. *et al.* Widespread aneuploidy revealed by DNA microarray expression profiling. *Nature Genet.* **25,** 333–337 (2000).
12. Piskur, J. Origin of the duplicated regions in the yeast genomes. *Trends Genet.* **17,** 302–303 (2001).
13. Koszul, R., Caburet, S., Dujon, B. & Fischer, G. Eukaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *EMBO J.* **23,** 234–243 (2004).
14. Llorente, B. *et al.* Genomic exploration of the hemiascomycetous yeasts: 18. Comparative analysis of chromosome maps and synteny with *Saccharomyces cerevisiae*. *FEBS Lett.* **487,** 101–112 (2000).
15. Llorente, B. *et al.* Genomic exploration of the hemiascomycetous yeasts: 20. Evolution of gene redundancy compared to *Saccharomyces cerevisiae*. *FEBS Lett.* **487,** 122–133 (2000).
16. Wong, S., Butler, G. & Wolfe, K. H. Gene order evolution and paleopolyploidy in hemiascomycete yeasts. *Proc. Natl Acad. Sci. USA* **99,** 9272–9277 (2002).
17. Langkjaer, R. B., Cliften, P. F., Johnston, M. & Piskur, J. Yeast genome duplication was followed by asynchronous differentiation of duplicated genes. *Nature* **421,** 848–852 (2003).
18. Kurtzman, C. P., Lynch, M. & Force, A. Phylogenetic circumscription of *Saccharomyces, Kluyveromyces* and other members of the Saccharomycetaceae, and the proposal of the new genera *Lachancea*, *Nakaseomyces*, *Naumovia*, *Vanderwaltozyma* and *Zygotorulaspora*. *FEMS Yeast Res.* **4,** 233–245 (2003).
19. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423,** 241–254 (2003).
20. Lynch, M. & Force, A. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154,** 459–473 (2000).
21. Force, A. *et al.* Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151,** 1531–1545 (1999).
22. Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. Selection in the evolution of gene duplications. *Genome Biol.* **3,** RESEARCH0008.1–0008.9 (2002).
23. Shore, D., Squire, M. & Nasmyth, K. A. Characterization of two genes required for the position-effect control of yeast mating-type genes. *EMBO J.* **3,** 2817–2823 (1984).
24. Bell, S. P., Kobayashi, R. & Stillman, B. Yeast origin recognition complex functions in transcription silencing and DNA replication. *Science* **262,** 1844–1849 (1993).
25. Benard, L., Carroll, K., Valle, R. C., Masison, D. C. & Wickner, R. B. The ski7 antiviral protein is an EF1-alpha homolog that blocks expression of non-Poly(A) mRNA in *Saccharomyces cerevisiae*. *J. Virol.* **73,** 2893–2900 (1999).
26. Huh, W. K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425,** 686–691 (2003).
27. Sanz, M., Trilla, J. A., Duran, A. & Roncero, C. Control of chitin synthesis through Shc1p, a functional homologue of Chs4p specifically induced during sporulation. *Mol. Microbiol.* **43,** 1183–1195 (2002).
28. Lorenz, M. C. & Heitman, J. Regulators of pseudohyphal differentiation in *Saccharomyces cerevisiae* identified through multicopy suppressor analysis in ammonium permease mutant strains. *Genetics* **150,** 1443–1457 (1998).
29. Winzeler, E. A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285,** 901–906 (1999).
30. Gu, Z. *et al.* Role of duplicate genes in genetic robustness against null mutations. *Nature* **421,** 63–66 (2003).
31. Sharp, P. M. & Li, W. H. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15,** 1281–1295 (1987).
32. Sarthy, A. V. *et al.* Identification and kinetic analysis of a functional homolog of elongation factor 3, YEF3 in *Saccharomyces cerevisiae*. *Yeast* **14,** 239–253 (1998).
33. Boles, E. *et al.* Characterization of a glucose-repressed pyruvate kinase (Pyk2p) in *Saccharomyces cerevisiae* that is catalytically insensitive to fructose-1,6-bisphosphate. *J. Bacteriol.* **179,** 2987–2993 (1997).
34. Hughes, M. K. & Hughes, A. L. Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. *Mol. Biol. Evol.* **10,** 1360–1369 (1993).
35. Gallardo, M. H., Bickham, J. W., Honeycutt, R. L., Ojeda, R. A. & Kohler, N. Discovery of tetraploidy in a mammal. *Nature* **401,** 341 (1999).
36. Bailey, G. S., Poulter, R. T. & Stockwell, P. A. Gene duplication in tetraploid fish: model for gene silencing at unlinked duplicated loci. *Proc. Natl Acad. Sci. USA* **75,** 5575–5579 (1978).
37. Otto, S. P. & Whitton, J. Polyploid incidence and evolution. *Annu. Rev. Genet.* **34,** 401–437 (2000).
38. Blanc, G., Barakat, A., Guyot, R., Cooke, R. & Delseny, M. Extensive duplication and reshuffling in the *Arabidopsis* genome. *Plant Cell* **12,** 1093–1101 (2000).
39. Sidow, A. Gen(om)e duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Dev.* **6,** 715–722 (1996).
40. Pebusque, M. J., Coulier, F., Birnbaum, D. & Pontarotti, P. Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *Mol. Biol. Evol.* **15,** 1145–1159 (1998).
41. Batzoglou, S. *et al.* ARACHNE: a whole-genome shotgun assembler. *Genome Res.* **12,** 177–189 (2002).
42. Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: arachne 2. *Genome Res.* **13,** 91–96 (2003).
43. Kellis, M., Patterson, N., Birren, B., Berger, B. & Lander, E. S. Methods in comparative genomics: genome correspondence, gene identification, regulatory motif discovery. *J. Comput. Biol.* (in the press).
44. Turner, R. J., Lovato, M. & Schimmel, P. One of two genes encoding glycyl-tRNA synthetase in *Saccharomyces cerevisiae* provides mitochondrial and cytoplasmic functions. *J. Biol. Chem.* **275,** 27681–27688 (2000).
45. Kurtzman, C. P. & Robnett, C. J. Phylogenetic relationships among yeasts of the 'Saccharomyces complex' determined from multigene sequence analyses. *FEMS Yeast Res.* **3,** 417–432 (2003).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We thank I. Roberts for providing strains; M. Endrizzi, L.-J. Ma, D. Qi and the staff of the MIT/Whitehead Institute Center for Genome Research who generated the shotgun sequence from *Kluyveromyces waltii*; J. Butler and the Arachne assembly team who generated the genome assembly; D. Bartel, S. Calvo, J. Galagan, D. Jaffe and J. Vinson for discussions and comments on the manuscript; and G. Fink and A. Murray for discussions and advice.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.K. (manoli@mit.edu) and E.S.L. (lander@broad.mit.edu). The GenBank accession number for *K. waltii* is AADM01000000.

©2004 Nature Publishing Group

Meeting report

The changing face of genomics

Manolis Kellis

Address: The Broad Institute, Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02138, and Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: manoli@mit.edu

Published: 30 April 2004

Genome ***Biology*** 2004, **5:**324

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/5/324

© 2004 BioMed Central Ltd

A report on the 5th annual Advances in Genome Biology and Technology (AGBT) and Automation in DNA Mapping and Sequencing (AMS) meeting, Marco Island, USA, 4-7 February 2004.

The annual meeting on Advances in Genome Biology and Technology was very different this year - in contrast to previous years, only a handful of talks covered the latest large-scale sequencing projects and the next species to be sequenced. This meeting took for granted that we can sequence, assemble and align complete genomes - achievements that only a few years ago seemed daunting, if not unthinkable. The focus of the meeting has instead shifted towards the new challenges in genomics, particularly in the areas of gene regulation, cell dynamics and genome evolution.

Cell regulation and organism development

Given the primary sequence of a species, a major goal of current genomics efforts is to understand the regulatory mechanisms and control circuitry of the cell. Towards this goal, Rick Young (Massachusetts Institute of Technology and Whitehead Institute, Cambridge, USA) presented the completion of the yeast protein-DNA interaction map. Using chromatin immuno-precipitation (ChIP) technology in combination with microarray 'chips' containing all intergenic regions, his group has undertaken a genome-wide study of the targets of all the roughly 200 transcriptional regulators in yeast under multiple environmental conditions. They characterized the sequence-specificity of these regulators using numerous motif-discovery tools, together with evolutionary sequence conservation and protein structure information. The resulting regulatory map revealed general principles of regulation in yeast, including the organizational architectures of promoter regions (single motif, multiple sites, multiple regulators or factor combinations), and the different types of regulatory response to environmental changes (off/on, invariant, expanded or altered). The Young group also studied 20 chromatin regulators that do not directly recognize DNA sequences, but instead rely on their association with transcription-factor partners for binding and can keep a record of transcription by maintaining chromatin state. The Young lab is now moving this technology into studying transcriptional-regulator binding in the human genome, which will have applications to understanding diseases from diabetes to cancer.

Michael Levine (University of California, Berkeley, USA) presented work aiming to understand the *cis*-regulatory circuitry of promoter regions in the fruitfly. He proposed that enhancer complexity might be a better measure of organismal complexity than overall gene count, and noted that the majority of enhancer elements act cooperatively in higher eukaryotes - autonomously-acting elements would in fact be the exceptions. Levine's group has studied the promoter architecture of *Drosophila* developmental genes that respond to different levels of the concentration gradient of the developmental protein Dorsal. Five activation levels, created by the combinatorial action of a three-response-level activator and a repressor, were detected for this set of genes. Levine and colleagues searched for conserved sequence elements in the promoter regions of genes belonging to the same activation level and discovered a common 'grammar' in the organization of three basic enhancer elements. Searching for a similar regulatory grammar in the mosquito genome revealed ten genes with similar putative regulatory clusters, two of which contain the same architecture, despite 230 million years of divergence.

Both Young and Levine have benefited from earlier work - classical studies of gene function in yeast and known developmental genes in the fly - that have set the foundations for future studies. In this spirit, Nancy Hopkins (Massachusetts Institute of Technology, Cambridge, USA) presented her research program that aims to systematically

identify all developmental genes in the zebrafish. Zebrafish is an appealing model for studying early vertebrate development because of its transparent body and the short time - a mere four days - between fertilization and free-swimming larval stages. As reaching the adult stage takes another four months, following multiple generations can, however, be prohibitively slow. To face this challenge, Hopkins and colleagues have used insertional mutagenesis to create mosaic parents whose germ cells each contain a different mutation. Following the fish with developmental defects and classifying each mutation has allowed her group to screen 32,000 founder fish, identify 550 mutants and 390 loci, 298 of which have human homologs. As many as 20% of these genes have no previously known biochemical function, providing a great starting point for experimentation and new biological discoveries. Additionally, the systematic approach allows one to estimate the total number of developmental genes, which Hopkins sets at 1,600, of which 25% have already been isolated. The cost of identifying additional genes increases as the study approaches saturation, however, and her group is not planning to pursue the systematic discovery phase of the work. They are currently working on understanding the genes identified and have revealed important new insights about genes involved in kidney, jaw, liver and myeloid cell formation.

Protein interactions and network evolution

Beyond the identification of genes involved in a developmental process lies the major challenge of understanding the genes' dynamic patterns of behavior during development. Josh LaBaer (Harvard Medical School, Boston, USA) presented his lab's proteomics work enabling such a pursuit, which they have approached by developing a 'protein-expression clone repository' that contains full-length protein-coding sequences for every gene in a number of model organisms, including yeast, bacteria and human. These protein-coding sequences are inserted into 'master' clone vectors and can be easily transferred to specialized vectors for expression studies, tagging with green fluorescent protein (GFP) to study localization, two-hybrid assays to determine protein interactions, or expression in constructs to detect protein modification states by mass spectroscopy. The system architecture is designed to be flexible, modular, reliable, comprehensive and catalogued. Mark Vidal (Harvard Medical School, Boston, USA) described a similar system for understanding the worm proteome. His lab is now mapping the localization of all 19,000 worm gene products across development, and so building 'chronograms' of gene expression from the head to the tail. The patterns of protein localization can then be clustered across space and time, thus constructing the dynamic aspect of the protein-interaction network of the worm. Vidal's group has currently completed 10% of the interactome matrix, and is moving towards completeness with the goal of understanding not only the components of the interactome but also protein behaviors during complex organismal tasks.

The protein-interaction network of a species provides the foundation for understanding the organism's responses to environmental changes and developmental signals. Across evolutionary time, these responses change and the regulatory circuits shift towards new ones. Lisa Stubbs (Lawrence Livermore National Laboratory, Livermore, USA) presented work aimed at understanding the evolution of such regulatory mechanisms across species. Her group studied the KRAB-ZNF family of chromatin-interacting zinc-finger transcriptional regulators. These proteins arose 400 million years ago, after divergence of land vertebrates from fish, and have recently undergone lineage-specific expansions in the human and mouse lineages via tandem duplications and deletions. Researchers shy away from using traditional comparative genomics tools for the analysis of such expanding gene families, as orthologous pairs are hard to determine. In the absence of orthology information, Stubbs constructed multiple alignments of all paralogous gene copies within each species in turn, rather than across species. This clever methodology, although atypical, yielded biologically meaningful intergenic sequence elements that are highly conserved across paralogs, and which can be shown, on the basis of reporter assays, to indeed act as enhancers. This approach also allowed the identification of species-specific elements that have arisen since the divergence of the mouse and human lineages. Stubbs' group is now working out the protein-level differences between paralogs, with the goal of understanding how structural changes affect function, in particular with respect to tissue-specific regulation and parental imprinting.

Human divergence and diversity

Differences between more closely related species can be much more subtle than the protein family expansions observed between human and mouse. For example, divergence between human and chimp shows as few as 12 changes every 1,000 nucleotides, which makes biological signal discovery a real challenge. Mike Zody (Massachusetts Institute of Technology and Broad Institute, Cambridge, USA) presented a very interesting way to use such a close relative to reveal insights into recent evolutionary history. Assuming that neutral divergence between species and neutral diversity within species are driven by the same underlying mutational mechanisms, Zody and colleagues used human-chimp divergence information to model the background mutation rate for each region of the genome. Using this information, they were able to distinguish regions of low within-human diversity that were due to selective sweeps (where a favorable mutation becomes incorporated into the genome very quickly), rather than simply to a lower mutation rate. These regions of low diversity were confirmed by genotyping nucleotide polymorphisms in humans of different ethnicity and building allele-frequency profiles. *FOXP2*, a well-known gene involved in language development, showed only a moderate signal for selection, whereas some regions that are devoid of annotated genes showed very strong evidence of

selection, raising new questions about what type of genetic elements may be under selection in the human genome.

With the aim of understanding human variation and how genetic differences may relate to phenotypic differences, Joanna Mountain (Stanford University, USA) presented her work relating language similarities to population polymorphisms. In particular, she has studied the genetic relationships of two geographically isolated populations in Africa, both of which speak with clicking languages (Khoisan). Using a multitude of metrics across various loci and populations, she found that the genetic data clearly support an evolutionarily distinct group for each of the two languages. To explain the results, Mountain put forward three possibilities: that the two languages arose independently, that they traveled across populations, or that instead the ancestral language was click-based and all intervening groups lost the click sounds. One advantage to such click languages might be the ability to communicate while hunting without alerting the prey. The use of modern genomics tools in the population genetics of language evolution illustrates how diverse the genomics field has become, and that, in fact, the maturity of genomics as a science has moved its applications well beyond the boundaries of the modern laboratory.

Overall, the meeting showcased a wide range of innovative talks, combining ground-breaking technological inventions with important scientific applications. The attendance was unusually low this year, especially on the industry side, witnessing the tight economic situation in the USA and internationally. At the same time, the changing focus of the meeting is evidence of a maturing field. Genomics has mastered its initial challenges, and is now extending its arms to embrace a growing number of fundamental questions in the study of life.

breaks induced at ectopic transcription units[21] activated by demethylation.

The data presented here show that Dnmt3L is required for heritable silencing of retrotransposons in male germ cells. Temporary deprivation of Dnmt3L through drug-induced conditional alleles of the *Dnmt3L* gene might allow the controlled mobilization of endogenous retrotransposons without meiotic catastrophe and could form the basis of a system of retrotransposon insertional mutagenesis that would be useful in forward genetic screens in the mouse. □

Received 22 May; accepted 26 July 2004; doi:10.1038/nature02886.
Published online 18 August 2004.

1. Aapola, U. *et al.* Isolation and initial characterization of a novel zinc finger gene, DNMT3L, on 21q22.3, related to the cytosine-5-methyltransferase 3 gene family. *Genomics* **65,** 293–298 (2000).
2. Hata, K., Okano, M., Lei, H. & Li, E. Dnmt3L cooperates with the Dnmt3 family of *de novo* DNA methyltransferases to establish maternal imprints in mice. *Development* **129,** 1983–1993 (2002).
3. Bourc'his, D., Xu, G. L., Lin, C. S., Bollman, B. & Bestor, T. H. Dnmt3L and the establishment of maternal genomic imprints. *Science* **294,** 2536–2539 (2001).
4. La Salle, S. *et al.* Windows for sex-specific methylation marked by DNA methyltransferase expression profiles in mouse germ cells. *Dev. Biol.* **268,** 403–415 (2004).
5. Dobson, M. J., Pearlman, R. E., Karaiskakis, A., Spyropoulos, B. & Moens, P. B. Synaptonemal complex proteins: occurrence, epitope mapping and chromosome disjunction. *J. Cell Sci.* **107,** 2749–2760 (1994).
6. Odorisio, T., Rodriguez, T. A., Evans, E. P., Clarke, A. R. & Burgoyne, P. S. Meiotic checkpoint monitoring synapsis eliminates spermatocytes via p53-independent apoptosis. *Nature Genet.* **18,** 257–261 (1998).
7. Enders, G. C. & May, J. J. Developmentally regulated expression of a mouse germ cell nuclear antigen examined from embryonic day 11 to adult in male and female mice. *Dev. Biol.* **163,** 331–340 (1994).
8. Yoder, J. A., Walsh, C. P. & Bestor, T. H. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* **13,** 335–340 (1997).
9. Lin, S. P. *et al.* Asymmetric regulation of imprinting on the maternal and paternal chromosomes at the *Dlk1–Gtl2* imprinted cluster on mouse chromosome 12. *Nature Genet.* **35,** 97–102 (2003).
10. Davis, T. L., Trasler, J. M., Moss, S. B., Yang, G. J. & Bartolomei, M. S. Acquisition of the *H19* methylation imprint occurs differentially on the parental alleles during spermatogenesis. *Genomics* **58,** 18–28 (1999).
11. Ueda, T. *et al.* The paternal methylation imprint of the mouse H19 locus is acquired in the gonocyte stage during foetal testis development. *Genes Cells* **5,** 649–659 (2000).
12. Bestor, T. H. Cytosine methylation mediates sexual conflict. *Trends Genet.* **19,** 185–190 (2003).
13. Walsh, C. P., Chaillet, J. R. & Bestor, T. H. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nature Genet.* **20,** 116–117 (1998).
14. Kuff, E. L. & Lueders, K. K. The intracisternal A-particle gene family: structure and functional aspects. *Adv. Cancer Res.* **51,** 183–276 (1988).
15. Ostertag, E. M. & Kazazian, H. H. Jr Biology of mammalian L1 retrotransposons. *Annu. Rev. Genet.* **35,** 501–538 (2001).
16. Kaneda, M. *et al.* Essential role for *de novo* methyltransferase 3a in paternal and maternal imprinting. *Nature* **429,** 900–903 (2004).
17. Bestor, T. H. & Tycko, B. Creation of genomic methylation patterns. *Nature Genet.* **12,** 363–367 (1996).
18. Chedin, F., Lieber, M. R. & Hsieh, C. L. The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a. *Proc. Natl Acad. Sci. USA* **99,** 16916–16921 (2002).
19. Cambareri, E. B., Jensen, B. C., Schabtach, E. & Selker, E. U. Repeat-induced G-C to A-T mutations in *Neurospora*. *Science* **244,** 1571–1575 (1989).
20. Maloisel, L. & Rossignol, J.-L. Suppression of crossing-over by DNA methylation in *Ascobolus*. *Genes Dev.* **12,** 1381–1389 (1998).
21. Petes, T. D. Meiotic recombination hot spots and cold spots. *Nature Rev. Genet.* **2,** 360–369 (2001).
22. Peters, A. H., Plug, A. W., van Vugt, M. J. & de Boer, P. A drying-down technique for the spreading of mammalian meiocytes from the male and female germline. *Chromosome Res.* **5,** 66–68 (1997).
23. Ponzetto-Zimmerman, C. & Wolgemuth, D. J. Methylation of satellite sequences in mouse spermatogenic and somatic DNAs. *Nucleic Acids Res.* **12,** 2807–2822 (1984).
24. Sanford, J., Forrester, L. & Chapman, V. Methylation patterns of repetitive DNA sequences in germ cells of *Mus musculus*. *Nucleic Acids Res.* **12,** 2822–2835 (1984).
25. Pietras, D. F. *et al.* Construction of a small *Mus musculus* repetitive DNA library: identification of a new satellite sequence in *Mus musculus*. *Nucleic Acids Res.* **11,** 6965–6983 (1983).

Acknowledgements We thank B. Tycko for suggestions, P. Moens for antibodies against synaptonemal complex proteins and H1t and for discussions, G. Enders for antibodies against GCNA1, J. Walonoski for advice on *in situ* hybridization, and K. Anderson, M. Damelin and M. Goll for criticism of the manuscript. This work was supported by grants from the National Institutes of Health to T.H.B. and by a fellowship from the Rett Syndrome Research Foundation to D.B.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.H.B (thb12@columbia.edu). *Dnmt3L*^{tm1Bes} mutant mice are available from the Jackson Labs (http://jaxmice.jax.org/).

Transcriptional regulatory code of a eukaryotic genome

Christopher T. Harbison[1,2]*, D. Benjamin Gordon[1]*, Tong Ihn Lee[1], Nicola J. Rinaldi[1,2], Kenzie D. Macisaac[3], Timothy W. Danford[3], Nancy M. Hannett[1], Jean-Bosco Tagne[1], David B. Reynolds[1], Jane Yoo[1], Ezra G. Jennings[1], Julia Zeitlinger[1], Dmitry K. Pokholok[1], Manolis Kellis[1,3,4], P. Alex Rolfe[3], Ken T. Takusagawa[3], Eric S. Lander[1,2,4], David K. Gifford[3,4], Ernest Fraenkel[1,3] & Richard A. Young[1,2,4]

[1]*Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA*
[2]*Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[3]*MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar Street, Cambridge, Massachusetts 02139, USA*
[4]*Broad Institute, One Kendall Square, Building 300, Cambridge, Massachusetts 02139, USA*

* These authors contributed equally to this work

DNA-binding transcriptional regulators interpret the genome's regulatory code by binding to specific sequences to induce or repress gene expression[1]. Comparative genomics has recently been used to identify potential *cis*-regulatory sequences within the yeast genome on the basis of phylogenetic conservation[2–6], but this information alone does not reveal if or when transcriptional regulators occupy these binding sites. We have constructed an initial map of yeast's transcriptional regulatory code by identifying the sequence elements that are bound by regulators under various conditions and that are conserved among *Saccharomyces* species. The organization of regulatory elements in promoters and the environment-dependent use of these elements by regulators are discussed. We find that environment-specific use of regulatory elements predicts mechanistic models for the function of a large population of yeast's transcriptional regulators.

We used genome-wide location analysis[7–10] to determine the genomic occupancy of 203 DNA-binding transcriptional regulators in rich media conditions and, for 84 of these regulators, in at least 1 of 12 other environmental conditions (Supplementary Table 1, Supplementary Fig. 1; http://web.wi.mit.edu/young/regulatory_code). These 203 proteins are likely to include nearly all of the DNA-binding transcriptional regulators encoded in the yeast genome. Regulators were selected for profiling in an additional environment if they were essential for growth in that environment or if there was other evidence implicating them in the regulation of gene expression in that environment. The genome-wide location data identified 11,000 unique interactions between regulators and promoter regions at high confidence ($P \leq 0.001$).

To identify the *cis*-regulatory sequences that are likely to serve as recognition sites for transcriptional regulators, we merged information from genome-wide location data, phylogenetically conserved sequences, and prior knowledge (Fig. 1a). We used six motif discovery methods[11–13] to discover 68,279 DNA sequence motifs for the 147 regulators that bound more than ten probes (Supplementary Methods, Supplementary Fig. 2). From these motifs we derived the most likely specificity for each regulator through clustering and stringent statistical tests. This motif discovery process identified highly significant ($P \leq 0.001$) motifs for each of 116 regulators. We determined a single high-confidence motif for 65 of these regulators by using additional criteria including the requirement for conservation across three of four related yeast species. Examples of discovered and rediscovered motifs are depicted in Fig. 1b, and comparisons of the discovered motifs with those described previously are shown in Supplementary Table 2.

©2004 Nature Publishing Group

The discovered motifs provide significantly more information than was previously available; for 21 of the regulators there was no prior specificity information in the literature, and detailed probability matrices had previously been determined for only 17 regulators for which we report motifs[14]. For Cin5, which showed the largest difference between the computationally derived motif (TTACRTAA) and the previously reported site (TTACTAA; Supplementary Table 2), we found that the motif we report is also the preferred target *in vitro* (Supplementary Fig. 3). We supplemented the discovered motifs with additional motifs from the literature that also passed conservation tests, and we used this compendium of sequence motifs for 102 regulators (Supplementary Table 3) in all subsequent analysis.

We constructed an initial version of the transcriptional regulatory code by mapping on the yeast genome sequence the motifs that are bound by regulators at high confidence ($P \leq 0.001$) and that are conserved among *sensu stricto Saccharomyces* species (Fig. 2; http://web.wi.mit.edu/fraenkel/regulatory_map). This map includes 3,353 interactions within 1,296 promoter regions. Maps of regulatory sites encompassing larger numbers of promoters, constructed with lower-confidence information, can also be viewed on the authors' website. Because the information used to construct the map includes binding data from multiple growth environments, the map describes transcriptional regulatory potential within the genome. During growth in any one environment, only subsets of the binding sites identified in the map are occupied by transcriptional regulators, as we describe in more detail below.

Where the functions of specific transcriptional regulators were established previously, the functions of the genes they bind in the regulatory map are highly consistent with this prior information. For example, the amino-acid biosynthetic regulators Gcn4 and Leu3 bind to sites in the promoter of *BAP2* (chromosome II), which encodes an amino-acid transporter (Fig. 2a). Six well-studied cell cycle transcriptional regulators bind to the promoter for *YHP1* (chromosome IV), which has been implicated in the regulation of the G1 phase of the cell cycle. The regulator of respiration Hap5 binds upstream of *COX4* (chromosome VII), which encodes a component of the respiratory electron transport chain. Where regulators with established functions bind to genes of unknown function, these target genes are newly implicated in such functional processes.

The utility of combining regulator binding data and sequence conservation data is illustrated in Fig. 2b. All sequences matching the regulator DNA binding specificities described in this study (Supplementary Table 3) that occur within the 884-base-pair intergenic region upstream of the gene *BAP2* are shown in the upper panel. The subset of these sequences that have been conserved in multiple yeast species, and are therefore likely candidates for

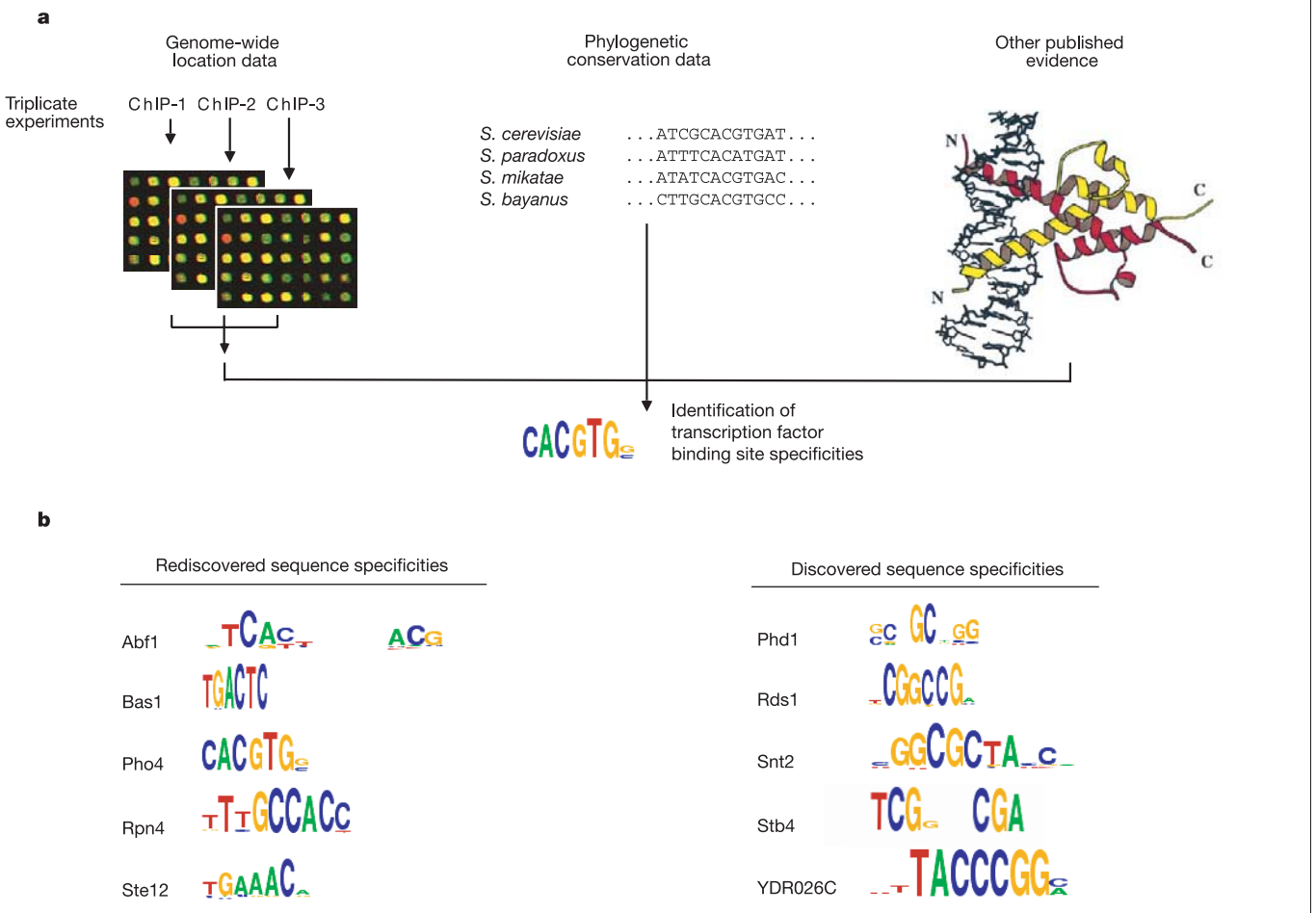


Figure 1 Discovering binding-site specificities for yeast transcriptional regulators. **a**, *Cis*-regulatory sequences likely to serve as recognition sites for transcriptional regulators were identified by combining information from genome-wide location data, phylogenetically conserved sequences and previously published evidence, as described in Supplementary Methods. The compendium of regulatory sequence motifs can be found in Supplementary Table 3. **b**, Selected sequence specificities that were rediscovered and were newly discovered are shown. The total height of the column is proportional to the information content of the position, and the individual letters have a height proportional to the product of their frequency and the information content[30].

©2004 Nature Publishing Group

regulator interactions, is shown in the middle panel of Fig. 2b. The presence of these conserved regulatory sites indicates the potential for regulation through this sequence but does not indicate whether the site is actually bound by a regulator under some growth condition. The incorporation of binding information (Fig. 2b, bottom panel) identifies those conserved sequences that are used by regulators in cells grown under the conditions examined.

The distribution of binding sites for transcriptional regulators reveals constraints on the organization of these sites in yeast promoters (Fig. 2c). Binding sites are not uniformly distributed over the promoter regions but instead show a sharply peaked distribution. Very few sites are located in the region 100 base pairs (bp) upstream of protein-coding sequences. This region typically includes the transcription start site and is bound by the transcription initiation apparatus. The vast majority (74%) of the transcriptional regulator binding sites lie between 100 and 500 bp upstream of the protein-coding sequence, far more than would be expected at random (53%). Regions further than 500 bp upstream contain fewer binding sites than would be expected at random. It seems that yeast transcriptional regulators function at short distances along the linear DNA, a property that reduces the potential for inappropriate activation of nearby genes.

We note that specific arrangements of DNA binding sites occur within promoters, and we suggest that these promoter architectures provide clues to regulatory mechanisms (Fig. 3). For example, the presence of a DNA binding site for a single regulator is the simplest promoter architecture and, as might be expected, we found that sets of genes with this feature are often involved in a common biological

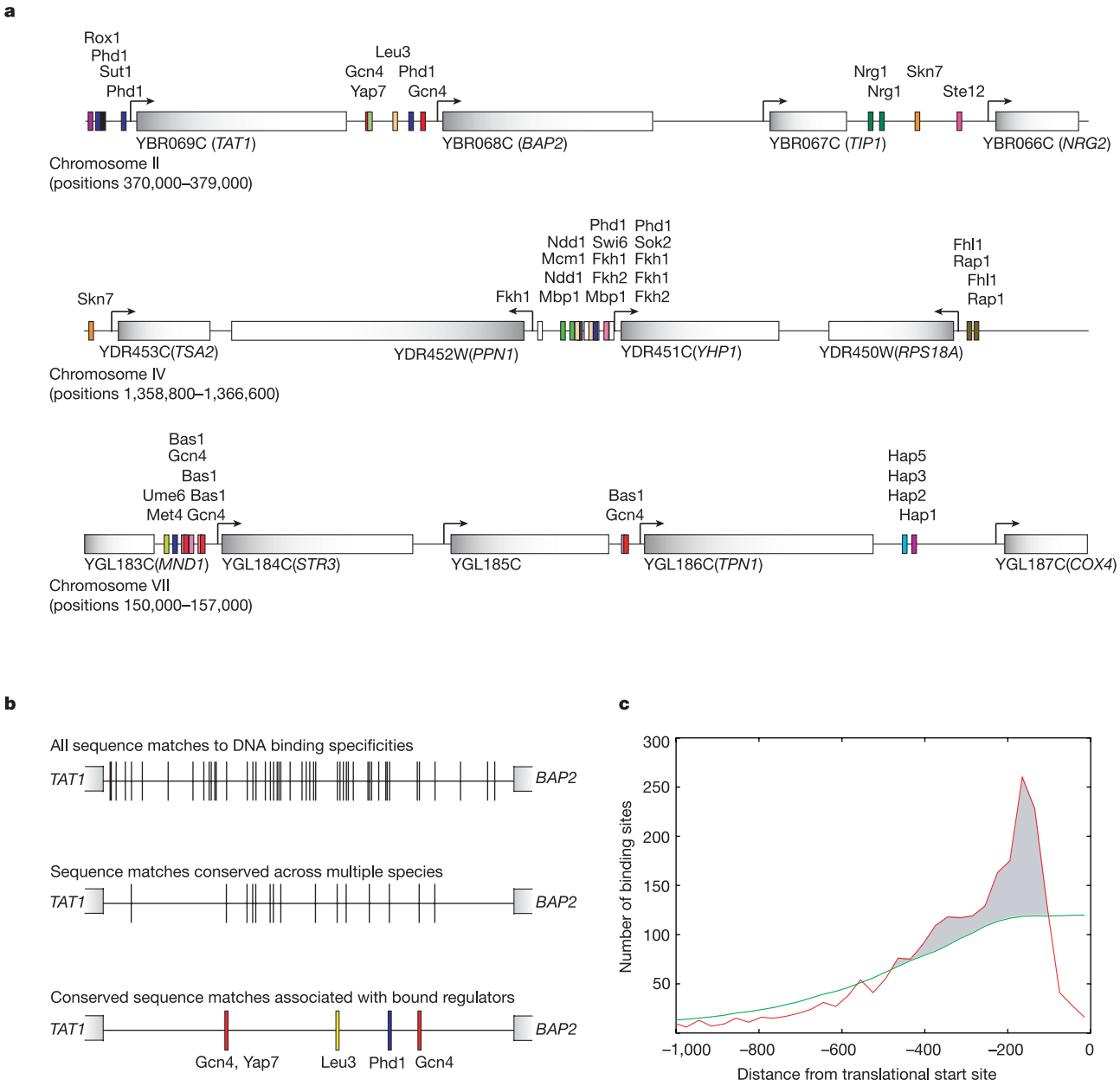


Figure 2 Drafting the yeast transcriptional regulatory map. **a**, Portions of chromosomes illustrating locations of genes (grey rectangles) and conserved DNA sequences (coloured boxes) bound *in vivo* by transcriptional regulators. **b**, Combining binding data and sequence conservation data. The diagram depicts all sequences matching a motif from our compendium (top), all such conserved sequences (middle) and all such conserved sequences bound by a regulator (bottom). **c**, Regulator binding site distribution. The red line shows the distribution of distances from the start codon of open reading frames to binding sites in the adjacent upstream region. The green line represents a randomized distribution.

©2004 Nature Publishing Group

function (Supplementary Table 4). A second type of promoter architecture consists of repeats of a particular binding site sequence. Repeated binding sites have been shown to be necessary for stable binding by the regulator Dal80 (ref. 15). This repetitive promoter architecture can also permit a graded transcriptional response, as has been observed for the *HIS4* gene[16]. Several regulators, including Dig1, Mbp1 and Swi6, show a statistically significant preference for repetitive motifs (Supplementary Table 5). A third class of promoter contains binding sites for multiple different regulators. This promoter arrangement implies that the gene might be subject to combinatorial regulation, and we expect that in many cases the various regulators can be used to execute differential responses to varied growth conditions. Indeed, we note that many of the genes in this category encode products that are required for multiple metabolic pathways and are regulated in an environment-specific fashion. In the fourth type of promoter architecture we discuss here, binding sites for specific pairs of regulators occur more frequently within the same promoter regions than would be expected by chance (Supplementary Table 6). This 'co-occurring' motif architecture implies that the two regulators interact physically or have related functions at multiple genes.

By conducting genome-wide binding experiments for some regulators under multiple cell-growth conditions, we learned that regulator binding to a subset of the regulatory sequences is highly dependent on the environmental conditions of the cell (Supplementary Fig. 4). We observed four common patterns of regulator binding behaviour (Fig. 4, Supplementary Table 7). Prior information about the regulatory mechanisms employed by well-studied regulators in each of the four groups suggests hypotheses to account for the environment-dependent binding behaviour of the other regulators.

'Condition-invariant' regulators bind essentially the same set of promoters (within the limitations of noise) in two different growth environments (Fig. 4). Leu3, which is known to regulate genes involved in amino-acid biosynthesis, is among the best studied of the regulators in this group. Binding of Leu3 *in vivo* has been shown to be necessary but not sufficient for the activation of Leu3-regulated genes[17]. Rather, regulatory control of these genes requires the association of a leucine metabolic precursor with Leu3 to convert it from a negative to a positive regulator. We note that other zinc cluster type regulators that show 'condition-invariant' behaviour are known to be regulated in a similar manner[18,19]. It is therefore reasonable to propose that the activation or repression functions of some of the other regulators in this class have requirements in addition to DNA binding.

'Condition-enabled' regulators do not bind the genome detectably under one condition, but bind a substantial number of promoters with a change in environment. Msn2 is among the best-studied regulators in this class, and the mechanisms involved in Msn2-dependent transcription provide clues to how the other regulators in that class might operate. Msn2 is excluded from the nucleus when cells grow in the absence of stresses but accumulates rapidly in the nucleus when cells are subjected to stress[20,21]. This condition-enabled behaviour was also observed for the thiamine biosynthetic regulator Thi2, the nitrogen regulator Gat1 and the developmental regulator Rim101. We suggest that many of these transcriptional regulators are regulated by nuclear exclusion or by another mechanism that would cause this extreme version of condition-specific binding.

'Condition-expanded' regulators bind to a core set of target promoters under one condition but bind an expanded set of promoters under another condition. Gcn4 is the best-studied of the regulators that fall into this 'expanded' class. The levels of Gcn4 are reported to increase sixfold when yeast cells are introduced into media with limiting nutrients[22], owing largely to increased nuclear protein stability[21,23], and under this condition we find that Gcn4 binds to an expanded set of genes. The probes bound when Gcn4 levels are low contain better matches to the known Gcn4-binding site than probes that are bound exclusively at higher protein concentrations, which is consistent with a simple model for specificity based on intrinsic protein affinity and protein concentration (Supplementary Fig. 5). The expansion of binding sites by many of the regulators in this class might reflect increased levels of the regulator available for DNA binding.

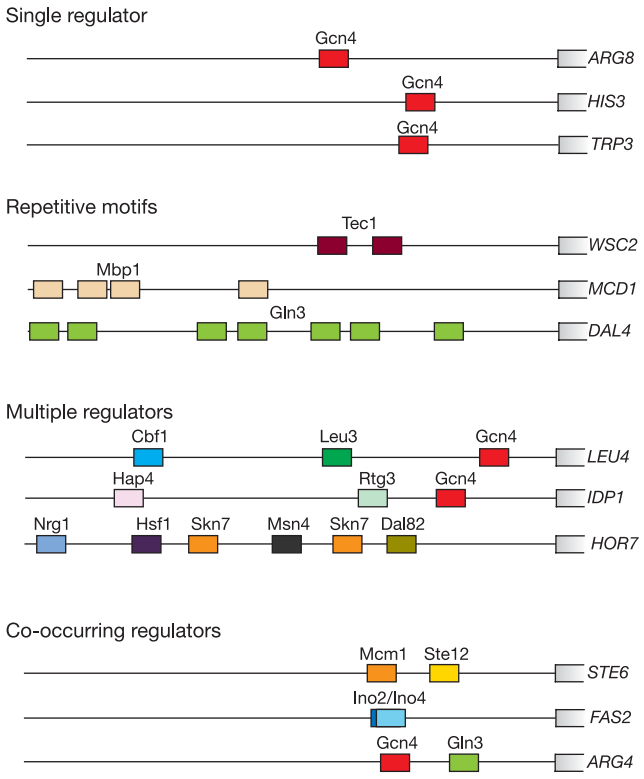


Figure 3 Yeast promoter architectures: single regulator architecture, promoter regions that contain one or more copies of the binding site sequence for a single regulator; repetitive motif architecture, promoter regions that contain multiple copies of a binding site sequence of a regulator; multiple regulator architecture, promoter regions that contain one or more copies of the binding site sequences for more than one regulator; co-occurring regulator architecture, promoters that contain binding site sequences for recurrent pairs of regulators. For the purposes of illustration, not all sites are shown and the scale is approximate. Additional information can be found in Supplementary Tables 4–6.

'Condition-altered' regulators exhibit an altered preference for the set of promoters bound in two different conditions. Ste12 is the best-studied of the regulators whose binding behaviour falls into this 'altered' class. Depending on the interactions with other regulators, the specificity of Ste12 can change and alter its cellular function[24]. For example, under filamentous growth conditions, Ste12 interacts with Tec1, which has its own DNA-binding specificity[25]. This condition-altered behaviour was also observed for the transcriptional regulators Aft2, Skn7 and Ume6. We propose that the binding specificity of many of the transcriptional regulators might be altered through interactions with other regulators or through modifications (such as chemical) that are dependent on environment.

Substantial portions of eukaryotic genome sequence are believed to be regulatory[2,3,26], but the DNA sequences that actually contribute to regulation of genome expression have been ill-defined. By mapping the DNA sequences bound by specific regulators in various environments, we identify the regulatory potential embedded in the genome and provide a framework for modelling the mechanisms

©2004 Nature Publishing Group

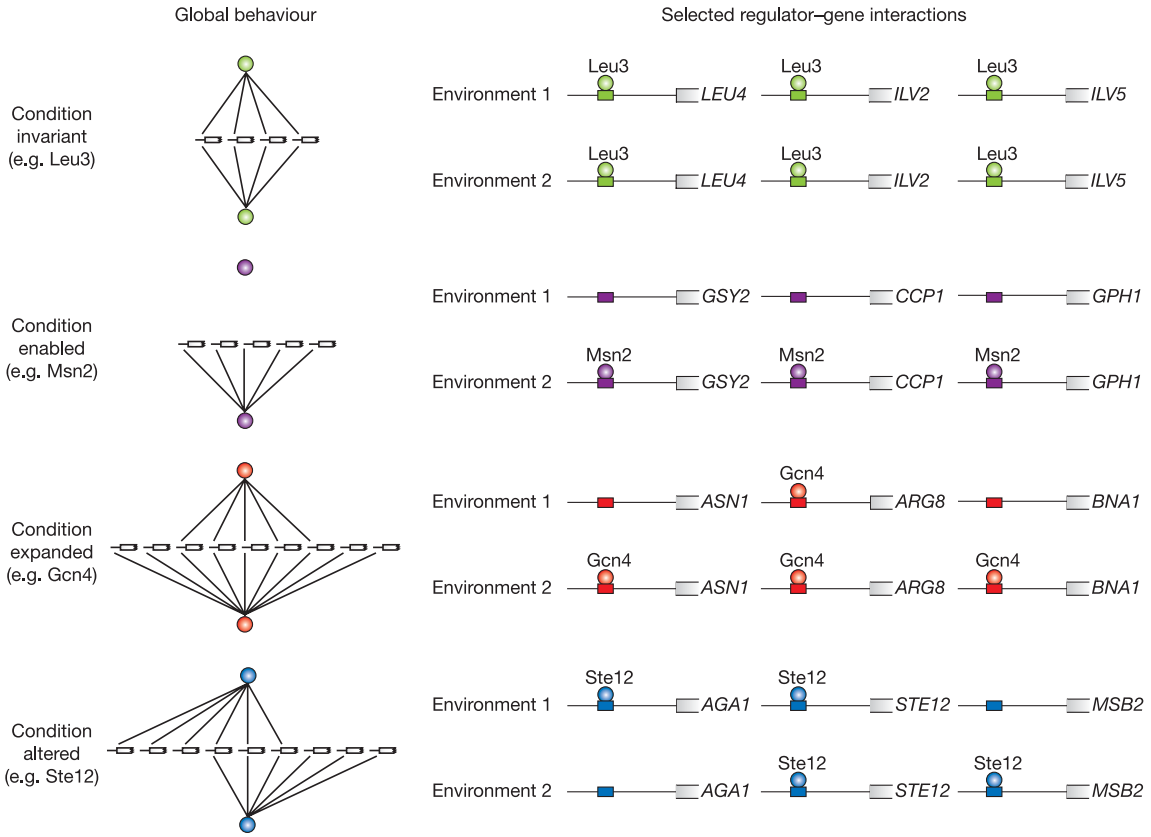


Figure 4 Environment-specific use of the transcriptional regulatory code. Four patterns of genome-wide binding behaviour are depicted on the left, where transcriptional regulators are represented by coloured circles and are placed above and below a set of target genes/promoters. The lines between the regulators and the target genes/promoters represent binding events. Specific examples of the environment-dependent behaviours are depicted on the right. Coloured circles represent regulators and coloured boxes represent their DNA binding sequences within specific promoter regions. We note that regulators might exhibit different behaviours when different pairs of conditions are compared.

that contribute to global gene expression. We expect that the approaches used here to map regulatory sequences in yeast can also be used to map the sequences that control genome expression in higher eukaryotes. □

Methods

Strain information

For each of the 203 regulators, strains were generated in which a repeated Myc-epitope-coding sequence was integrated into the endogenous gene encoding the regulator. Polymerase chain reaction (PCR) constructs containing the Myc-epitope-coding sequence and a selectable marker flanked by regions of homology to either the 5′ or 3′ end of the targeted gene were transformed into the W303 yeast strain Z1256 (refs 8, 9). Genomic integration and expression of the epitope-tagged protein were confirmed by PCR and western blotting, respectively.

Genome-wide location analysis

Genome-wide location analysis was performed as described previously[8,9]. Bound proteins were crosslinked by formaldehyde to DNA *in vivo*, followed by cell lysis and sonication to shear DNA. Crosslinked material was immunoprecipitated with an anti-Myc antibody, followed by reversal of the crosslinks to separate DNA from protein. Immunoprecipitated DNA and DNA from an unenriched sample were amplified and differentially fluorescently labelled by ligation-mediated PCR. These samples were hybridized to a microarray consisting of spotted PCR products representing the intergenic regions of the *Saccharomyces cerevisiae* genome. Relative intensities of spots were used as the basis for an error model that assigns a probability score (P) to binding interactions. All microarray data are available from ArrayExpress and from the authors' website (http://web.wi.mit.edu/young/regulatory_code).

Growth environments

We profiled all 203 regulators in rich medium. In addition, we profiled 84 regulators in at least one other environmental condition. The list of regulators is given in Supplementary Table 1.

Regulator binding specificity

The putative specificities of regulators were identified by applying a suite of motif discovery programs to the intergenic sequences identified by the binding data. The resulting specificity predictions were filtered for significance with uniform metrics and then clustered to yield representative motifs (Supplementary Fig. 2).

We used six methods to identify the specific sequences bound by regulators: AlignACE[11], MEME[13], MDscan[12], the method in ref. 2 and two additional new methods that incorporate conservation data: MEME_c and CONVERGE. MEME_c uses the existing MEME program without change but applies it to a modified set of sequences in which bases that are not conserved in the *sensu stricto Saccharomyces* species were replaced with the letter 'N'. CONVERGE is a novel expectation-maximization (EM)-based algorithm for discovering specificities by using sequence information from multiple genomes. Rather than searching for sites that are identical across the *sensu stricto* species, as occurs with MEME_c, CONVERGE searches for loci at which all aligned sequences are consistent with the same specificity model. See Supplementary Methods for runtime parameters and additional details for all of these methods.

Each of the programs we used attempts to measure the significance of its results with one or more statistical scores. However, we observed that these programs report results with high scores even when applied to random selections of intergenic regions. To distinguish the true motifs we chose a set of statistical measures that are described in Supplementary Methods and converted these scores into the empirical probability that a motif with a similar score could be found by the same program in randomly selected sequences. To estimate these P values we ran each program 50 times on randomly selected sets of sequences of various sizes. We accepted only those motifs that were judged to be significant by these scores ($P \leq 0.001$).

Significant motifs from all programs were pooled and clustered with the use of a k-medoids algorithm. Aligned motifs within each cluster were averaged to produce consensus motifs and filtered according to their conservation. This procedure typically produced several distinct consensus motifs for each regulator. To choose a single specificity for each regulator we compared the results with information in the TRANSFAC[27], YPD[28] and SCPD[29] databases. When no prior information was available we chose the specificity with the most significant statistical score.

Regulatory code

Potential binding sites were included in the map of the regulatory code if they satisfied two criteria. First, a locus had to match the specificity model for a regulator in the *S. cerevisiae* genome and at least two other *sensu stricto Saccharomyces* genomes with a score of at least 60% of the maximum possible. Second, the locus had to lie in an intergenic region that also contained a probe bound by the corresponding regulator in any condition ($P \leq 0.001$). All

©2004 Nature Publishing Group

analyses of promoter architecture and environment-specific binding were based on this map and can be found in Supplementary Information.

Received 11 March; accepted 1 July 2004; doi:10.1038/nature02800.

1. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3,** 318–356 (1961).
2. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423,** 241–254 (2003).
3. Cliften, P. *et al.* Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301,** 71–76 (2003).
4. Pritsker, M., Liu, Y. C., Beer, M. A. & Tavazoie, S. Whole-genome discovery of transcription factor binding sites by network-level conservation. *Genome Res.* **14,** 99–108 (2004).
5. Wang, T. & Stormo, G. D. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* **19,** 2369–2380 (2003).
6. Blanchette, M. & Tompa, M. FootPrinter: A program designed for phylogenetic footprinting. *Nucleic Acids Res.* **31,** 3840–3842 (2003).
7. Iyer, V. R. *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409,** 533–538 (2001).
8. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290,** 2306–2309 (2000).
9. Lee, T. I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298,** 799–804 (2002).
10. Lieb, J. D., Liu, X., Botstein, D. & Brown, P. O. Promoter-specific binding of Rap1 revealed by genome–wide maps of protein-DNA association. *Nature Genet.* **28,** 327–334 (2001).
11. Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnol.* **16,** 939–945 (1998).
12. Liu, X. S., Brutlag, D. L. & Liu, J. S. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnol.* **20,** 835–839 (2002).
13. Bailey, T. L. & Elkan, C. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **Vol. 3** 21–29 (AAAI Press, Menlo Park, California, 1995).
14. Knuppel, R., Dietze, P., Lehnberg, W., Frech, K. & Wingender, E. TRANSFAC retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins. *J. Comput. Biol.* **1,** 191–198 (1994).
15. Cunningham, T. S. & Cooper, T. G. The *Saccharomyces cerevisiae* DAL80 repressor protein binds to multiple copies of GATAA-containing sequences (URSGATA). *J. Bacteriol.* **175,** 5851–5861 (1993).
16. Donahue, T. F., Daves, R. S., Lucchini, G. & Fink, G. R. A short nucleotide sequence required for regulation of HIS4 by the general control system of yeast. *Cell* **32,** 89–98 (1983).
17. Kirkpatrick, C. R. & Schimmel, P. Detection of leucine-independent DNA site occupancy of the yeast Leu3p transcriptional activator *in vivo*. *Mol. Cell. Biol.* **15,** 4021–4030 (1995).
18. Axelrod, J. D., Majors, J. & Brandriss, M. C. Proline-independent binding of PUT3 transcriptional activator protein detected by footprinting *in vivo*. *Mol. Cell. Biol.* **11,** 564–567 (1991).
19. Ma, J. & Ptashne, M. The carboxy-terminal 30 amino acids of GAL4 are recognized by GAL80. *Cell* **50,** 137–142 (1987).
20. Beck, T. & Hall, M. N. The TOR signalling pathway controls nuclear localization of nutrient-regulated transcription factors. *Nature* **402,** 689–692 (1999).
21. Chi, Y. *et al.* Negative regulation of Gcn4 and Msn2 transcription factors by Srb10 cyclin-dependent kinase. *Genes Dev.* **15,** 1078–1092 (2001).
22. Albrecht, G., Mosch, H. U., Hoffmann, B., Reusser, U. & Braus, G. H. Monitoring the Gcn4 protein-mediated response in the yeast *Saccharomyces cerevisiae*. *J. Biol. Chem.* **273,** 12696–12702 (1998).
23. Kornitzer, D., Raboy, B., Kulka, R. G. & Fink, G. R. Regulated degradation of the transcription factor Gcn4. *EMBO J.* **13,** 6021–6030 (1994).
24. Zeitlinger, J. *et al.* Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell* **113,** 395–404 (2003).
25. Baur, M., Esch, R. K. & Errede, B. Cooperative binding interactions required for function of the Ty1 sterile responsive element. *Mol. Cell. Biol.* **17,** 4330–4337 (1997).
26. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
27. Matys, V. *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31,** 374–378 (2003).
28. Hodges, P. E., McKee, A. H., Davis, B. P., Payne, W. E. & Garrels, J. I. The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res.* **27,** 69–73 (1999).
29. Zhu, J. & Zhang, M. Q. SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics* **15,** 607–611 (1999).
30. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18,** 6097–6100 (1990).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We thank T. Ideker and S. McCuine for help in selecting regulators to study in environmental conditions; E. Herbolsheimer, G. Bell, R. Latek and F. Lewitter for computational assistance; and E. McReynolds for technical assistance. E.F. is a Whitehead Fellow and was funded in part by Pfizer. D.B.G. was supported by a NIH/NIGMS NRSA award. This work was supported by an NIH grant.

Competing interests statement The authors declare competing financial interests: details accompany the paper on **www.nature.com/nature**.

Correspondence and requests for materials should be addressed to R.A.Y. (young@wi.mit.edu) or E.F. (efraenkel@wi.mit.edu). ArrayExpress number E-WMIT-10 has been given for microarray data.

Structure of the acrosomal bundle

Michael F. Schmid[1], Michael B. Sherman[1]*, Paul Matsudaira[2] & Wah Chiu[1]

[1]*National Center for Macromolecular Imaging, Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, Texas 77030, USA*
[2]*Whitehead Institute, Department of Biology and Division of Biological Engineering MIT, Cambridge, Massachusetts 02142, USA*

* Present address: Purdue University, Department of Biological Sciences, West Lafayette, Indiana 47907-139, USA

In the unactivated *Limulus* sperm, a 60-μm-long bundle of actin filaments crosslinked by the protein scruin is bent and twisted into a coil around the base of the nucleus. At fertilization, the bundle uncoils and fully extends in five seconds to support a finger of membrane known as the acrosomal process. This biological spring is powered by stored elastic energy and does not require the action of motor proteins or actin polymerization[1]. In a 9.5-Å electron cryomicroscopic structure of the extended bundle, we show that twist, tilt and rotation of actin–scruin subunits deviate widely from a 'standard' F-actin filament. This variability in structural organization allows filaments to pack into a highly ordered and rigid bundle in the extended state and suggests a mechanism for storing and releasing energy between coiled and extended states without disassembly.

The actin bundle of the *Limulus* acrosomal process is an unusual actin-based engine that is highly ordered in its structure[2]. The coiled state has a difference in average twist of 0.23° per subunit relative to the extended form. Change in the twist accompanies uncoiling of the bundle into the final extended state[3]. Mechanical measurements of the extended bundle yield a Young's modulus (a measure of stiffness) of 2 GPa (refs 4–6). This stiffness enables the bundle to penetrate the jelly layer surrounding the egg membrane. A 40-Å tomographic reconstruction[7] revealed periodic and dense packing of the scruin-crosslinked actin filaments. However, to understand all these functions, a higher resolution structure is necessary.

An electron micrograph of the native extended bundle shows its crystallinity (Fig. 1a). We merged data from 153 such bundles at various orientations[8] to produce a three-dimensional map at a resolution of 9.5 Å (Fig. 1b). As far as we know, this is the highest resolution of any electron cryomicroscopic reconstruction containing an F-actin filament. We can identify each actin subunit and both domains of scruin, referred to as elongated (E) and spheroidal (S) on the basis of their shape[9], and secondary structure elements of both molecules.

We used correlation analysis and non-crystallographic symmetry (NCS) averaging within the asymmetric unit to determine the azimuthal and axial positions of the 14 actin–scruin subunits. The azimuthal rotation between adjacent subunits ranges from −176.5° to −142.5° (Fig. 1c, d), and thus deviates from the canonical actin helix by −10.4° under-twist to +23.6° over-twist. Nevertheless, the rise per subunit in the axial direction had less than one pixel (1.33 Å) deviation. This amount of individual azimuthal distortion is larger than the distortions found in averaged reconstructions from F-actin filaments[10–15], but such studies cannot reveal individual subunit variations.

On the basis of this two-dimensional cross-correlation, we created an average actin monomer from the 14 subunits in the asymmetric unit. In this map (Fig. 2a), three α-helices in subdomain 1 of actin were detected by an α-helix search (using the program 'helixhunter')[16]. Their locations match α-helices in the actin crystal structure (amino acids 79–91, 113–125 and 137–144)[17]. Because we obtained this map without referring to the actin crystal structure during either the reconstruction process[8] or the NCS

©2004 Nature Publishing Group

Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype

**Olivier Jaillon[1], Jean-Marc Aury[1], Frédéric Brunet[2], Jean-Louis Petit[1], Nicole Stange-Thomann[3], Evan Mauceli[3], Laurence Bouneau[1], Cécile Fischer[1], Catherine Ozouf-Costaz[4], Alain Bernot[1], Sophie Nicaud[1], David Jaffe[3], Sheila Fisher[3], Georges Lutfalla[5], Carole Dossat[1], Béatrice Segurens[1], Corinne Dasilva[1], Marcel Salanoubat[1], Michael Levy[1], Nathalie Boudet[1], Sergi Castellano[6], Véronique Anthouard[1], Claire Jubin[1], Vanina Castelli[1], Michael Katinka[1], Benoît Vacherie[1], Christian Biémont[7], Zineb Skalli[1], Laurence Cattolico[1], Julie Poulain[1], Véronique de Berardinis[1], Corinne Cruaud[1], Simone Duprat[1], Philippe Brottier[1], Jean-Pierre Coutanceau[4], Jérôme Gouzy[8], Genis Parra[6], Guillaume Lardier[1], Charles Chapple[6], Kevin J. McKernan[9], Paul McEwan[9], Stephanie Bosak[9], Manolis Kellis[3], Jean-Nicolas Volff[10], Roderic Guigó[6], Michael C. Zody[3], Jill Mesirov[3], Kerstin Lindblad-Toh[3], Bruce Birren[3], Chad Nusbaum[3], Daniel Kahn[8], Marc Robinson-Rechavi[2], Vincent Laudet[2], Vincent Schachter[1], Francis Quétier[1], William Saurin[1], Claude Scarpelli[1], Patrick Wincker[1], Eric S. Lander[3,11], Jean Weissenbach[1] & Hugues Roest Crollius[1]*

[1]*UMR 8030 Genoscope, CNRS and Université d'Evry, 2 rue Gaston Crémieux, 91057 Evry Cedex, France*
[2]*Laboratoire de Biologie Moléculaire de la Cellule, CNRS UMR 5161, INRA UMR 1237, Ecole Normale Supérieure de Lyon, 46 allée d'Italie, 69364 Lyon Cedex 07, France*
[3]*Broad Institute of MIT and Harvard, 320 Charles Street, Cambridge, Massachusetts 02141, USA*
[4]*Muséum National d'Histoire Naturelle, Département Systématique et Evolution, Service de Systématique Moléculaire, CNRS IFR 101, 43 rue Cuvier, 75231 Paris, France*
[5]*Défenses Antivirales et Antitumorales, CNRS UMR 5124, 1919 route de Mende, 34293 Montpellier Cedex 5, France*
[6]*Grup de Recerca en Informàtica Biomèdica, IMIM-UPF and Programa de Bioinformàtica i Genòmica (CRG), Barcelona, Catalonia, Spain*
[7]*CNRS UMR 5558 Biométrie et Biologie Evolutive, Université Lyon 1, 69622 Villeurbanne, France*
[8]*INRA-CNRS Laboratoire des Interactions Plantes Micro-organismes, 31326 Castanet Tolosan Cedex, France*
[9]*Agencourt Bioscience Corporation, Massachusetts 01915, USA*
[10]*Biofuture Research Group, Evolutionary Fish Genomics, Physiologische Chemie I, Biozentrum, University of Wuerzburg, Am Hubland, D-97074 Wuerzburg, Germany*
[11]*Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA*

* Present address: CNRS UMR8541, Ecole Normale Supérieure, 46 rue d'Ulm, 75005 Paris, France

Tetraodon nigroviridis **is a freshwater puffer fish with the smallest known vertebrate genome. Here, we report a draft genome sequence with long-range linkage and substantial anchoring to the 21** ***Tetraodon*** **chromosomes. Genome analysis provides a greatly improved fish gene catalogue, including identifying key genes previously thought to be absent in fish. Comparison with other vertebrates and a urochordate indicates that fish proteins have diverged markedly faster than their mammalian homologues. Comparison with the human genome suggests ~900 previously unannotated human genes. Analysis of the** ***Tetraodon*** **and human genomes shows that whole-genome duplication occurred in the teleost fish lineage, subsequent to its divergence from mammals. The analysis also makes it possible to infer the basic structure of the ancestral bony vertebrate genome, which was composed of 12 chromosomes, and to reconstruct much of the evolutionary history of ancient and recent chromosome rearrangements leading to the modern human karyotype.**

Access to entire genome sequences is revolutionizing our understanding of how genetic information is stored and organized in DNA, and how it has evolved over time. The sequence of a genome provides exquisite detail of the gene catalogue within a species, and the recent analysis of near-complete genome sequences of three mammals (human[1], mouse[2] and rat[3]) shows the acceleration in the search for causal links between genotype and phenotype, which can then be related to physiological, ecological and evolutionary observations. The partial sequence of the compact puffer fish *Takifugu rubripes* genome was obtained recently and this survey provided a preliminary catalogue of fish genes[4]. However, the *Takifugu* assembly is highly fragmented and as a result important questions could not be addressed.

Here, we describe and analyse the genome sequence of the freshwater puffer fish *Tetraodon nigroviridis* with long-range linkage and extensive anchoring to chromosomes. *Tetraodon* resembles *Takifugu* in that it possesses one of the smallest known vertebrate genomes, but as a popular aquarium fish it is readily available and is easily maintained in tap water (see Supplementary Notes for naming conventions, natural habitat and phylogeny). The two puffer fish diverged from a common ancestor between 18–30 million years (Myr) ago and from the common ancestor with mammals about 450 Myr ago[5]. This long evolutionary distance provides a good contrast to distinguish conserved features from neutrally evolving DNA by sequence comparison. *Tetraodon* sequences in fact had an important role in providing a reliable estimate of the number of genes in the human genome[6].

There has been a vigorous and unresolved debate as to whether a whole-genome duplication (WGD) occurred in the ray-finned fish (actinopterygians) lineage after its separation from tetrapods[7–9]. By exploiting the extensive anchoring of the *Tetraodon* sequence to chromosomes, we provide a definitive answer to this question. The distribution of duplicated genes in the genome reveals a striking pattern of chromosome pairing, and the correspondence of orthologues with the human genome show precisely the signatures expected from an ancient WGD followed by a massive loss of duplicated genes.

Moreover, we find that relatively few interchromosomal

©2004 Nature Publishing Group

Table 1 **Assembly statistics**

Parameter	Number	N50 length (kb)	Size with gaps included (Mb)	Size with gaps excluded (Mb)	Longest (kb)	Percentage of the genome with gaps included
All contigs	49,609	16	312.4	312.4	258	91.9
All scaffolds	25,773	984	342.4	312.4	7,612	100.7
All ultracontigs	128	7,622	276.4	247.0	12,035	81.3
Mapped contigs	16,083	26	197.7	197.7	258	58.1
Mapped scaffolds	1,588	608	218.4	197.7	7,612	64.2
Mapped ultracontigs	39	8,701	219.7	197.7	12,035	64.6

rearrangements occurred in the *Tetraodon* lineage over several hundred million years after the WGD. This allows us to propose a karyotype of the ancestral bony vertebrate (Osteichthyes) composed of 12 chromosomes, and to uncover many unknown evolutionary breakpoints that occurred in the human genome in the past 450 Myr.

The *Tetraodon* genome sequence

Sequencing and assembly

The *Tetraodon* genome was sequenced using the whole-genome shotgun (WGS) approach. Random paired-end sequences providing 8.3-fold redundant coverage were produced at Genoscope (GSC) and the Broad Institute of MIT and Harvard (see Supplementary Table SI1). From this, the assembly program Arachne[10,11] constructed 49,609 contigs for a total of 312 megabases (Mb; Table 1), which it then connected into 25,773 scaffolds (or supercontigs) covering 342 Mb (including gaps; see Supplementary Information). Half of the assembly is in 102 scaffolds larger than 731 kilobases (kb; the N50 length) and the largest scaffold measures 7.6 Mb, the typical length of a *Tetraodon* chromosome arm.

We produced additional data to physically link scaffolds and anchor them to chromosomes. These data include probe hybridizations to arrayed bacterial artificial chromosome (BAC) libraries, restriction digest fingerprints of BAC clones, additional linking clone sequence, alignment to available *Takifugu* sequence and two-colour fluorescence *in situ* hybridization (FISH) (see Supplementary Information). The impact of these additional mapping data was twofold: first, we could join 2,563 scaffolds in 128 'ultracontigs' that cover 81.3% of the assembly, and second, we were able to anchor the 39 ultracontigs among the largest (covering 64.6% of the assembly, with an N50 size of 8.7 Mb) to *Tetraodon* chromosomes (Fig. 1; see also Supplementary Table SI2 and Supplementary Notes).

The accuracy of the assembly was experimentally tested and the inter-contig links found to be correct in >99% of cases. On the basis of a re-sequencing experiment, we estimate that the assembly covers >90% of the euchromatin of the *Tetraodon* genome (Supplementary Information). Finally, the overall genome size was directly measured by flow cytometry experiments on several fish; an average value of 340 Mb was obtained, consistent with the sequence assembly and smaller than the previously reported estimate of 350–400 Mb.

The *Tetraodon* draft sequence has roughly 60-fold greater con-

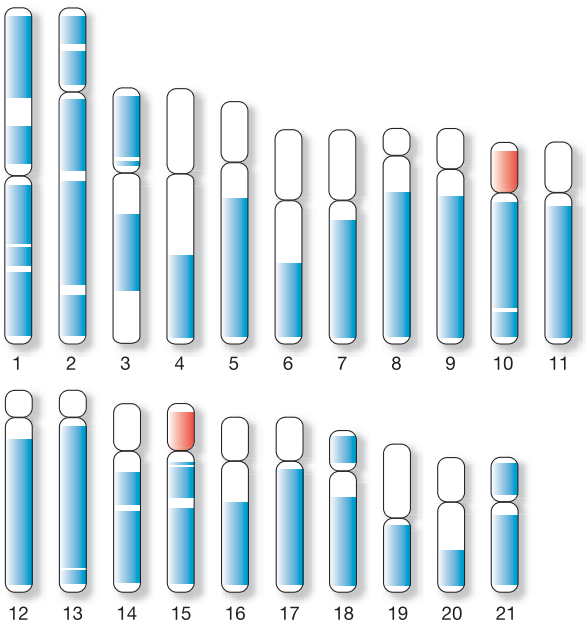


Figure 1 The *Tetraodon* genome is composed of 21 chromosomes. Red areas indicate the location of 5S and 28S ribosomal RNA gene arrays on chromosome 10 and chromosome 15, respectively. Many chromosomes are subtelocentric; that is, they only possess a very short heterochromatic arm. The extent of 39 sequence-based ultracontigs that cover about 64% of their length is shown in blue. In addition, approximately 16% of the genome is contained in another 89 ultracontigs that are not yet anchored on chromosomes, and the remaining 20% of the genome is in 23,210 smaller scaffolds.

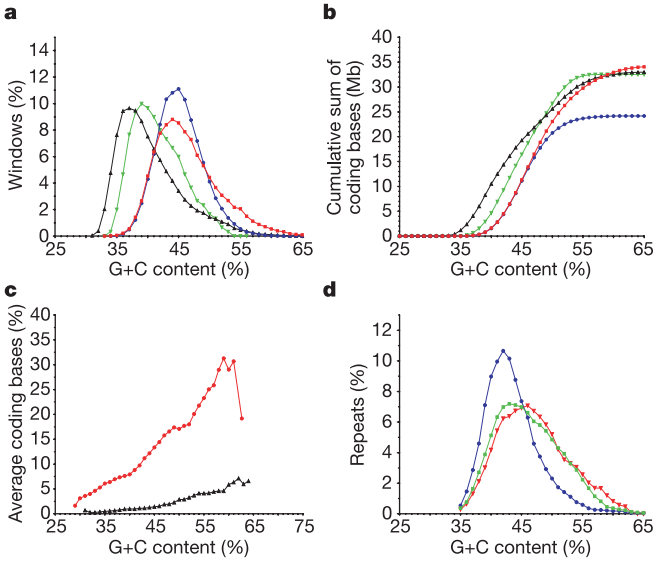


Figure 2 Distribution of the G + C content. **a**, Distribution in 5-kb non-overlapping windows across *Tetraodon* (red squares) and *Takifugu* (blue circles) scaffolds, and in 50-kb windows in human (black triangles) and mouse (green inverted triangles) chromosomes. Windows containing more than 25% ambiguous or unknown nucleotides (gaps) were excluded from the analysis. **b**, Cumulative sum of annotated coding bases in *Tetraodon* and *Takifugu* (5-kb non-overlapping windows) and human and mouse (50-kb windows) as a function of G + C content. **c**, In sharp contrast to *Takifugu*[4] the density of genes increases with the G + C content (%) in *Tetraodon* (red circles) much more than in human (black triangles). **d**, The three major families of repeats in *Tetraodon* are not distributed uniformly in the genome: long terminal repeat (LTR) and LINE elements (red diamonds and green squares, respectively) concentrate in (G + C)-rich regions and SINE elements (blue circles) concentrate in (A + T)-rich regions. In contrast, the distribution of these elements is much more uniform in *Takifugu* (Supplementary Fig. S4).

©2004 Nature Publishing Group

Table 2 **Comparison between *Tetraodon* and *Takifugu* annotations**

Parameter	*Tetraodon*	*Takifugu**	*Takifugu*†
Annotated genes	27,918	35,180	20,796
Annotated transcripts	27,918	38,510	33,003
Average number of coding exons per gene	6.9	4.3	8.6
Average number of UTR exons per gene	0.4	0‡	0.07
Average gene size (bp)	4,778	2,754	6,547
Average CDS size (bp)	1,230	745	1,397
Average exon size (bp)	178	171	163
Number of annotated bases (Mb)			
Coding	33.9	26.1	29.1
UTR	2.4	0‡	0.02

**Takifugu* annotations are from Ensembl version 18.2.1.
†*Takifugu* annotations are from Ensembl version 23.2.1.
‡*Takifugu* annotations from Ensembl version 18.2.1 do not include UTRs.

tinuity at the level of N50 ultracontig size than the *Takifugu* draft sequence (7.62 Mb versus 125 kb). Critically, the anchoring of the assembly provides a comprehensive view of a fish genome sequence organized in individual chromosomes.

Genome landscape

A consequence of the remarkably compact nature of the *Tetraodon* genome is that its G+C content is much higher than in the larger genomes of mammals. Although the G+C content is shifted markedly, it still shows the same asymmetric bell-shaped distribution with an excess of higher values as seen in human and mouse (Fig. 2a). (G+C)-rich regions tend to be gene-rich in mammals, and analysis of our data shows that this is also true for *Tetraodon* (Fig. 2b, c). The *Tetraodon* genome thus cannot be considered as a single homogeneous component but, as in mammals, it is a mosaic of relatively gene-rich and gene-poor regions.

Transposable elements are very rare in the *Tetraodon* genome[12,13]: we estimate here that they do not exceed 4,000 copies; however, with 73 different types, they are richly represented (Supplementary Notes and Supplementary Table SI3). In sharp contrast, the human and mouse genomes contain only ~20 different types but are riddled with millions of transposable element copies. One of the intriguing features of the human genome is that the distribution of short interspersed nucleotide elements (SINEs) is biased towards (G+C)-rich regions, whereas long interspersed nucleotide elements (LINEs) favour (A+T)-rich regions. In *Tetraodon*, these preferences are precisely reverse: LINEs occur preferentially in (G+C)-rich regions and SINEs in (A+T)-rich regions (Fig. 2d). The reason for these differences is not clear.

The *Tetraodon* genome shows certain striking differences from the previously reported *Takifugu* genome sequence. *Takifugu* contains eightfold more copies of transposable elements[4] than *Tetraodon*, which may contribute to its slightly larger genome size (approximately 370 Mb; see Supplementary Information). More surprisingly, the G+C content of *Takifugu* does not show the characteristic asymmetry seen in mammals and in *Tetraodon* (Fig. 2a) nor the biases in SINE and LINE distribution (Supplementary Fig. S4). Why would the (G+C)-rich component be lacking in the *Takifugu* sequence, when this fraction is gene dense in mammals and in *Tetraodon*? This cannot be ascribed to transposable elements, which represent less than 5% of the assembly in both of these puffer fish species. One possible explanation is that the (G+C)-rich fraction exists in *Takifugu*, but was markedly under-represented as a result of aspects of the cloning, sequencing or assembly process. The fact that *Tetraodon* (G+C)-rich regions contain an excess of genes with no apparent orthologues in the *Takifugu* genome supports this hypothesis. Indeed, the *Tetraodon* genome appears to contain ~16.5% more coding exons than *Takifugu* (see below).

Tetraodon genes

Gene catalogue

The most prevalent features of the *Tetraodon* genome are protein-coding genes, which span 40% of the assembly. We constructed a catalogue of genes by adapting the GAZE[14] computational framework (Supplementary Fig. S5) in order to combine three types of data: *Tetraodon* complementary DNA mapping, similarities to human, mouse and *Takifugu* proteins and genomes, and *ab initio* gene models (Supplementary Notes and Supplementary Tables SI4 and SI5).

The current *Tetraodon* catalogue is composed of 27,918 gene models, with 6.9 coding exons per gene on average (7.3 including untranslated regions (UTRs); Table 2). Assuming that fish and mammal genes possess similar gene structures, this suggests that some *Tetraodon* annotated genes are partial or fragmented because human and mouse genes respectively show 8.7 and 8.4 coding exons per gene[2]. Adjusting the gene count for such fragmentation (by multiplying by 6.9/8.6) would yield an estimated gene count of 22,400 genes, whereas accounting for unsequenced regions of the genome might increase the estimate slightly further. Although such

Table 3 **Comparative InterPro analysis of fish, mammal and urochordate proteomes**

Tetraodon	*Takifugu*	Human	Mouse	*Ciona*	InterPro description
Actinopterygian-enriched					
61	78	22	21	48	Sodium:neurotransmitter symporter
33	29	11	13	33	Na^+/solute symporter
21	16	8	7	6	Sodium/calcium exchanger membrane region
141	191	86	97	52	Collagen triple helix repeat
15	28	6	4	19	HAT dimerization
17	15	5	4	27	Peptidase M12A, astacin
3	4	0	0	1	Inosine/uridine-preferring nucleoside hydrolase
Sarcopterygian-enriched					
0	0	275	173	0	KRAB box
0	0	14	8	0	KRAB-related
3	0	25	29	0	High mobility group protein HMG14 and HMG17
0	0	9	95	0	Vomeronasal receptor, type 1
0	0	13	21	0	Keratin, high sulphur B2 protein
0	0	3	3	0	Keratin, high-sulphur matrix protein
0	0	22	11	0	Mammalian taste receptor
0	0	11	9	0	Pancreatic RNase
0	0	7	8	0	β-Defensin
Vertebrate-enriched					
52	40	82	102	9	Histone core
252	253	240	228	88	Homeobox
62	56	80	55	9	Zn finger, B box
94	83	75	74	19	Zn-binding protein, LIM
65	56	70	135	17	HMG1/2 (high mobility group) box

Supplementary Table SI7 contains the top 100 InterPro domains in *Tetraodon*.

©2004 Nature Publishing Group

Table 4 **Evolutionarily conserved regions between mammals and fish**

Query genome	Target genome			
	Tetraodon nigroviridis	*Takifugu rubripes*	*Homo sapiens*	*Mus musculus*
Tetraodon nigroviridis	NA	ND	139,316	133,091
Takifugu rubripes	ND	NA	139,932	131,835
Combined fish	NA	NA	151,708	142,804
Homo sapiens	142,820	133,239	NA	ND
Mus musculus	140,407	129,996	ND	NA
Combined mammals	151,668	140,965	NA	NA

NA, not applicable; ND, not determined.

estimates are somewhat imprecise, it seems likely that *Tetraodon* has between 20,000–25,000 protein coding genes.

The *Tetraodon* gene catalogue appears to be the most complete so far for a fish, with coding exons and UTRs totalling ~36 Mb (~11% of the genome; Table 2). The *Takifugu* paper[4] reported an estimate of 35,180 genes, but it did not account for a high degree of fragmentation (~4.3 exons per gene model). More recent, unpublished analyses have revised this number sharply downward (Table 2). The human and *Tetraodon* genomes have a similar distribution of exon sizes but markedly different distributions of intron size (Supplementary Fig. S6a). Although neither genome seems to tolerate introns below approximately 50–60 base pairs, *Tetraodon* has accumulated a much higher frequency of introns at this lower limit. Interestingly, this phenomenon is not uniform across the genome: there is an excess of genes with many small introns (Supplementary Fig. S6b), suggesting that intron sizes fluctuate in a regional fashion.

Proteome comparison between vertebrates

We examined in detail two gene families with unusual properties that represent challenges for automatic annotation procedures and have particular biological interest. The first is the family of selenoproteins, where the UGA codon encodes a rare cysteine analogue named selenocysteine (Sec) instead of signalling the end of translation as in all other genes[15]. We annotated 18 distinct families in *Tetraodon* based on similarities with the 19 protein families known in eukaryotes, and discovered a new selenoprotein that seems to be restricted to the actinopterygians among vertebrates and does not have a Cys counterpart in mammals. We also catalogued type I helical cytokines and their receptors (HCRI), a group of genes that were not found in the *Takifugu* genome[4] because of their poor sequence conservation, leading to the hypothesis that fish may not possess this large family that includes hormones and interleukins. *Tetraodon,* in fact, contains 30 genes encoding HCRIs with a typical D200 domain (Supplementary Fig. S7) and represents all families previously described in mammals[16].

InterPro[17] domains were annotated in protein sequences predicted in the *Tetraodon*, *Takifugu*, human, mouse and the urochordate *Ciona intestinalis*[18] genome using InterProScan[19]. We did not identify major differences between fish and mammal InterPro families, except for a few striking cases (Table 3): (1) collagen molecules are much more diverse in fish than in mammals, with one *Tetraodon* gene containing 20 von Willebrand type A domains, the largest number found so far in a single protein. (2) Some domains associated with sodium transport are noticeably enriched in fishes and *Ciona*, perhaps a reflection of their adaptation to saline aquatic environments that was lost in land vertebrates. (3) Purine nucleosidases usually involved in the recovery of purine nucleosides are more abundant in fish, including an allantoin pathway for purine degradation that is present in *Tetraodon* and absent in human. (4) Several hundred KRAB box transcriptional repressors involved in chromatin-mediated gene regulation exist in mammals and are totally absent in fish. (5) Proteins involved in general gene regulation are more abundant in vertebrates than in *Ciona*.

Protein annotation with gene ontology (GO) classifications[20] shows only subtle differences between fish and mammals, as was already observed between human and mouse[2]. The largest differences between species are seen with the GO classification in molecular functions (Supplementary Fig. S9). Interestingly, the two puffer fish and *Ciona* often vary together, showing for instance a higher frequency of enzymatic and transporter functions, and a lower frequency of signal transducer and structural molecules than both mammals (human and mouse). These global observations are difficult to relate to evolutionary or physiological mechanisms but provide a framework to understand the emergence or decline of molecular functions in vertebrates.

Number of genes in mammals and teleosts

The total amount of coding sequence conserved between the two fish and the two mammalian genomes provides a measure of their respective coding capacity. The Exofish method[6] is well suited to measure this, because it translates entire genomes in all six frames and identifies conserved coding regions (ecores) with a high specificity and independently of prior genome annotation (Table 4; see also Supplementary Information). The four vertebrate genomes contain remarkably similar numbers of ecores, apart from minor differences attributable to varying degrees of sequence completion. This suggests that they possess fairly similar numbers of genes. In fact, the gene count may be slightly less in mammals than in fish because the proportion of ecores corresponding to pseudogenes is higher in mammals[21].

The human ecores can be used to search for previously unrecognized human genes. The discovery of new human genes is becoming an increasingly rare event, given the scale and intensity of international efforts to annotate the genome by systematic annotation pipelines and by human experts. Roughly 14,500 human ecores

Table 5 **Rates of DNA evolution in vertebrates**

Species	Total number of orthologues	Number of orthologues used	Average per cent identity (without gaps)	Observed number of substitutions per 4D site	Estimated amount of neutral evolution	Estimated rate of neutral evolution (sites per Myr)	K_a
Human–mouse	14,889	5,802	91.76	0.32	0.43	0.0057	0.05
Tetraodon–Takifugu	12,909	5,802	90.51	0.27	0.35	0.0146	0.06
Tetraodon–human	9,975	5,802	69.90	0.63	1.54*	–	0.24
Tetraodon–mouse	9,666	5,802	69.46	0.63	1.53*	–	0.25
Takifugu–human	9,143	5,802	70.05	0.63	1.52*	–	0.24
Takifugu–mouse	8,956	5,802	69.67	0.63	1.52*	–	0.25

*These values are saturated and cannot be considered reliable estimates.

©2004 Nature Publishing Group

conserved with *Tetraodon* sequences do not overlap any 'known' features (genes or pseudogenes) in the human genome. Using these as anchors for local gene identification using the GAZE program, we identified 904 novel human gene predictions. Of these, 63% are also supported by expressed sequence tag (EST) data (from human or other species) and 50% contain predicted InterPro protein domains (Supplementary Table SI9). The most convincing evidence supporting these gene predictions is that they are strongly enriched on chromosomes that have not yet been annotated by human experts (Supplementary Table SI10). The novel gene predictions have relatively small size (average coding sequence (CDS) of 469 bp), which may have caused them to be eliminated by systematic annotation procedures. They provide a rich resource to help complete the human gene catalogue.

Tetraodon gene evolution

We measured rates of sequence divergence between fish and mammals to estimate the relative speed with which functional and non-functional sequences evolve in these lineages. We used fourfold degenerate (4D) site substitutions in orthologous proteins as a proxy for neutral nucleotide mutations, an approach that has been shown to be robust across entire genomes[2]. To optimize further the selection of sites used for comparison, we only considered the 5,802 proteins that are identified as orthologues in all pairwise comparisons between human, mouse, *Tetraodon* and *Takifugu*. The average neutral nucleotide substitution rate, inferred using the REV model[22,23], shows that the divergence between *Tetraodon* and *Takifugu* is about twice as fast per year as between human and mouse (Table 5), or between mouse and rat[3].

We were interested to see whether this higher mutation rate is also seen in protein sequences. Pairwise comparison of all possible combinations of the 5,802 four-way orthologous proteins clearly indicates that proteins between the two puffer fish are more divergent than between the two mammals, despite the shorter evolutionary time that has elapsed (Fig. 3). This is confirmed by the fact that the average frequency of non-synonymous mutations (leading to an amino acid change, K_a) between *C. intestinalis* and human proteins is lower than between *Ciona* and *Tetraodon* (see Methods).

Independent of the overall rate of change, the ratio of non-synonymous to synonymous changes (K_a/K_s ratio) is much higher between the two puffer fish than between human and mouse (Supplementary Table SI11 and Supplementary Information), suggesting that protein evolution is proceeding more rapidly along the puffer fish lineage. The reasons for this faster tempo of protein change are unknown, although it is likely to be positively correlated with the higher rate of neutral mutation.

Genome evolution

Genome-wide sequence provides a rare opportunity to address key evolutionary questions in a global fashion, circumventing biases due to small sequence and gene samples. In this respect, the combination of long-range linkage in the *Tetraodon* sequence and its evolutionary divergence from the mammalian lineage at 450 Myr ago makes it possible to explore overall genome evolution in the vertebrate clade.

Evidence for whole-genome duplication

The occurrence of WGD in the ray-finned fish lineage is a hotly debated question due both to the cataclysmic nature of such an event and to the difficulty in establishing that it actually occurred[24–26].

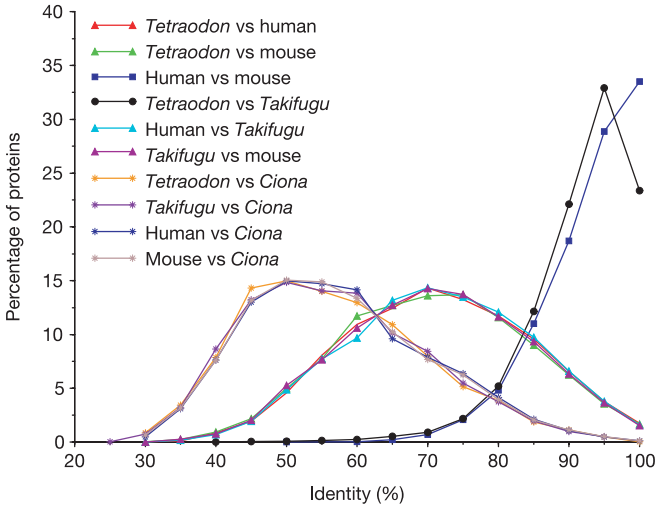


Figure 3 Distribution of the per cent identity between pairs of orthologous protein sets. Comparisons were performed with 2,289 proteins that are orthologous between the chordate *C. intestinalis* and all four vertebrates—*Tetraodon*, *Takifugu*, human and mouse (asterisks)—and with 5,802 proteins orthologous between all four vertebrates only, between fish and mammals (triangles) or between the two fish (circles), and between the two mammals (squares). As expected, all vertebrates show the same distribution profile compared to *Ciona* and both fish show the same distribution profile compared to mammals. Surprisingly, the distribution profile of the comparison between the two fish and between the two mammals is also very similar, despite the much shorter evolutionary time since the tetraodontiform radiation.

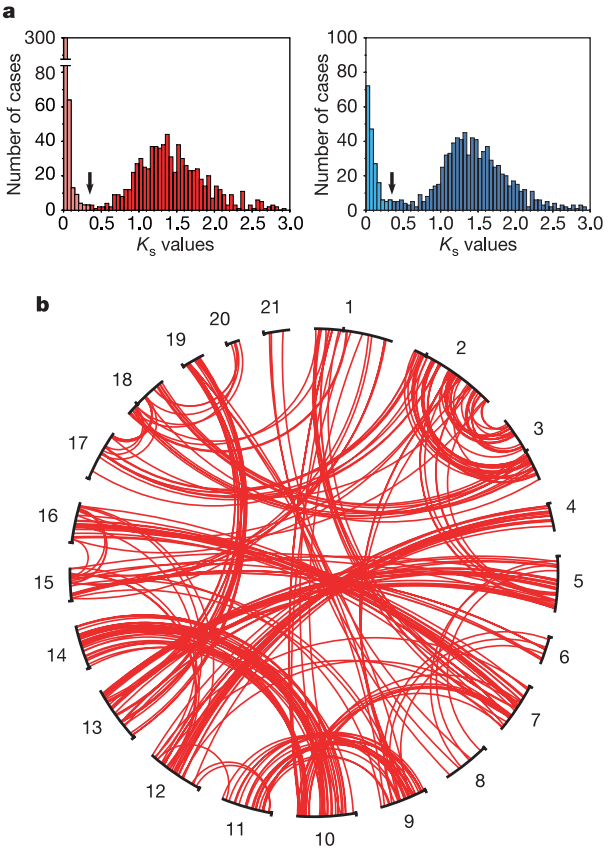


Figure 4 Genome duplication. **a**, Distribution of K_s values of duplicated genes in *Tetraodon* (left) and *Takifugu* (right) genomes. Duplicated genes broadly belong to two categories, depending on their K_s value being below or higher than 0.35 substitutions per site since the divergence between the two puffer fish (arrows). **b**, Global distribution of ancient duplicated genes ($K_s > 0.35$) in the *Tetraodon* genome. The 21 *Tetraodon* chromosomes are represented in a circle in numerical order and each line joins duplicated genes at their respective position on a given pair of chromosomes.

©2004 Nature Publishing Group

Definitive proof of WGD requires identifying certain distinctive signatures in long-range genome organization, which has previously been impossible to address with the data available.

It is expected that after WGD the resulting polyploid genome gradually returns to a diploid state through extensive gene deletion, with only a small proportion of duplicated copies ultimately retained as sources of functional innovation[26]. Paralogous chromosomes will thus each retain only a small subset of their initially common gene complement and then will be broken into smaller segments by genomic rearrangements. WGD will thus leave two distinctive signs for considerable periods before eventually fading.

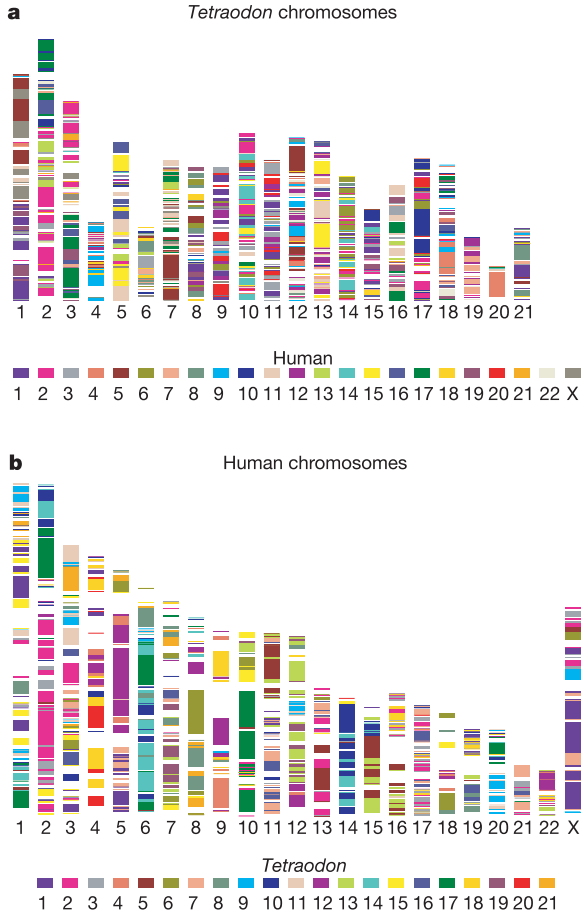


Figure 5 Synteny maps. **a**, For each *Tetraodon* chromosome, coloured segments represent conserved synteny with a particular human chromosome. Synteny is defined as groups of two or more *Tetraodon* genes that possess an orthologue on the same human chromosome, irrespective of orientation or order. *Tetraodon* chromosomes are not in descending order by size because of unequal sequence coverage. The entire map includes 5,518 orthologues in 900 syntenic segments. **b**, On the human genome the map is composed of 905 syntenic segments. See Supplementary Information for the synteny map between *Tetraodon* and mouse (Supplementary Fig. S11).

The first distinctive sign is duplicated genes on paralogous chromosomes. In the absence of chromosomal rearrangement it would be simple to recognize two paralogous chromosomes arising from a WGD from the genome-wide distribution of duplicate genes: the chromosomes would each contain one member from many duplicated gene pairs occurring in the same order along their length. The difficulty is that this neat picture will eventually be blurred by interchromosomal rearrangement, which will disrupt the 1:1 correspondence between chromosomes, and intrachromosomal rearrangement, which will disrupt gene ordering along chromosomes.

We analysed the genome-wide distribution of duplicated gene pairs to see whether a strong correspondence between chromosomes could be detected. We identified 1,078 and 995 pairs of duplicated genes in the *Tetraodon* and *Takifugu* genomes, respectively, using conservative criteria (see Supplementary Information). On the basis of the frequencies of silent mutations (K_s) between copies, ~75% are 'ancient' duplications that arose before the *Tetraodon–Takifugu* speciation (Fig. 4a).

The chromosomal distribution of these ancient duplicates follows a striking pattern characteristic of a WGD. Genes on one chromosome segment have a strong tendency to possess duplicate copies on a single other chromosome (Fig. 4b). The correspondence is not a perfect 1:1 match owing to interchromosomal exchange, but it is vastly stronger than expected by chance (Supplementary Table SI12). As expected from a WGD, all chromosomes are involved. Remarkably, some duplicate chromosome pairs such as *Tetraodon* chromosome 9 (Tni9) and Tni11 have remained largely undisturbed by chromosome translocations since the duplication event. In other cases, one chromosome has links to two or three others, suggestive of either fusion or fragmentation (for example, Tni13 matches Tni5 and Tni19).

The second distinctive sign, which is an even more powerful signature of genome duplication, comes from comparison with a related species carrying a genome that did not undergo the WGD. Such a comparison was recently used to prove the existence of an ancient WGD in the yeast *Saccharomyces cerevisiae* based on comparison with a second yeast species *Kluyveromyces waltii* that diverged before the WGD[27,28]. Although two ancient paralogous regions typically retained only a few genes in common, they could be readily recognized because they showed a characteristic 2:1 mapping with interleaving; that is, they both showed conserved synteny and local order to the same region of the *K. waltii* genome with the *S. cerevisiae* genes interleaving in alternating stretches. Such regions were called blocks of DCS (doubly conserved synteny). Whereas the first distinctive sign of WGD depends only on a

Table 6 **Distribution of human orthologues on *Tetraodon* chromosomes listed by their ancestral chromosome of origin**

	Ancestral chromosome											
	A	B	C	D	E	F	G	H	I	J	K	L
Tetraodon chromosome (copy 1)	4	17	2	2	5	13	7	1	1	10	9	6
Number of orthologues on copy 1	141	30	130	318	187	145	136	143	151	262	214	111
Percentage of orthologues on copy 1*	32.0	19.2	31.4	62.1	52.1	58.5	58.1	58.8	61.6	52.5	45.2	36.4
Tetraodon chromosome (copy 2)	12	18	3	3	13	19	16	7	15	14	11	8
Number of orthologues on copy 2	299	94	166	97	172	103	98	100	94	237	259	129
Percentage of orthologues on copy 2*	68.0	60.26	40.1	18.9	47.9	41.5	41.9	41.2	38.4	47.5	54.8	42.3
Tetraodon chromosome (copy 3)	–	20	18	17	–	–	–	–	–	–	–	21
Number of orthologues on copy 3	–	32	118	97	–	–	–	–	–	–	–	65
Percentage of orthologues on copy 3*	–	20.5	28.50	18.9	–	–	–	–	–	–	–	21.31

*Only orthologues that belong to syntenic groups are indicated here. For instance, ancestral chromosome A could be reconstructed with 141 *Tetraodon*–human orthologues belonging to *Tetraodon* chromosome 4 and 299 to chromosome 12.

©2004 Nature Publishing Group

minority of duplicated genes, the DCS signature considers all genes for which orthologues can be found in the related species.

We used 6,684 *Tetraodon* genes localized on individual chromosomes that possess an orthologue in either human or mouse to create a high-resolution synteny map (Fig. 5 and Supplementary Fig. S11, respectively). The map contains 900 syntenic groups composed of at least two consecutive genes (average 6.1; maximum 55) having orthologues on the same human chromosome; the syntenic groups include 76% of *Tetraodon*–human orthologues. The synteny map with mouse contains 1,011 syntenic groups, probably reflecting the higher degree of chromosomal rearrangement in the rodent lineage[2].

The synteny map typically associates two regions in *Tetraodon* with one region in human. Using precise criteria (see Methods) we defined DCS blocks for *Tetraodon* relative to human; in contrast to the yeast study, strict conservation of gene order within DCSs was not required. Notably, most (79.6%) orthologous genes in syntenic groups can be assigned to 90 DCS blocks (Fig. 6). As in *S. cerevisiae*[27], we see the distinctive interleaving pattern expected from WGD followed by massive gene loss. Analysis of the interleaving pattern shows that the gene loss occurred through many small deletions in a balanced fashion over the two *Tetraodon* sister chromosomes (average balance 42% and 58% of retention; Supplementary Information); this is consistent with the results in yeast.

These two analyses provide definitive evidence that the *Tetraodon* genome underwent a WGD sometime after its divergence from the mammalian lineage. The first test used only the ~3% of genes that represent duplicated gene pairs retained from the WGD. The second test used the pattern of 2:1 mapping with interleaving involving ~80% of orthologues between *Tetraodon* and human.

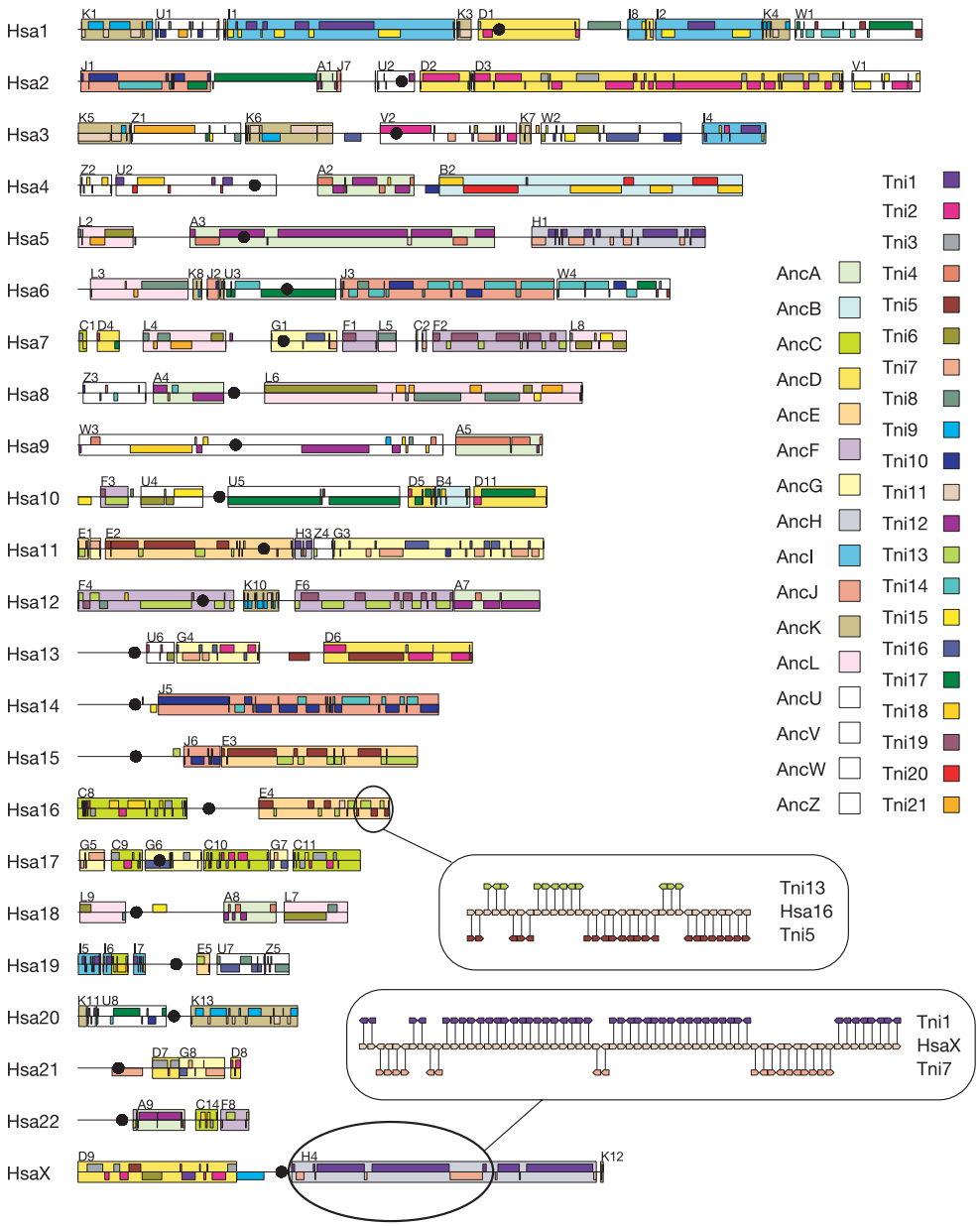


Figure 6 Duplicate mapping of human chromosomes reveals a whole-genome duplication in *Tetraodon*. Blocks of synteny along human chromosomes map to two (or three) *Tetraodon* chromosomes in an interleaving pattern. Small boxes represent groups of syntenic orthologous genes enclosed in larger boxes that define the boundaries of 110 DCS blocks. Black circles indicate human centromeres. A region of human chromosomes Xq and 16q are shown in detail with individual *Tetraodon* orthologous genes depicted on either side.

©2004 Nature Publishing Group

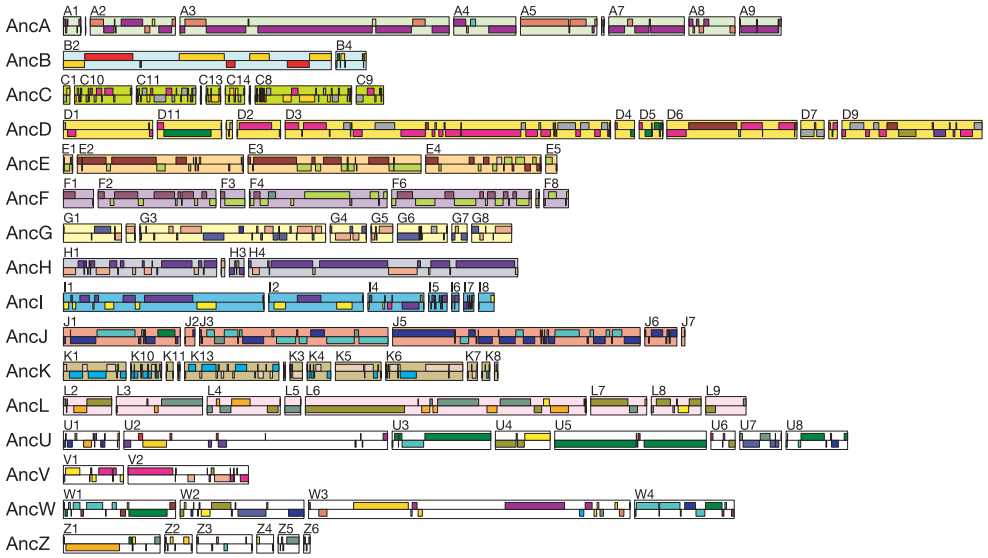


Figure 7 Composition of the ancestral osteichthyan genome. The 110 DCS blocks identified on the human genome are grouped according to their composition in terms of *Tetraodon* chromosomes, thus delineating 12 ancestral chromosomes containing 90 DCS blocks. The order of DCSs within an ancestral chromosome is arbitrary. The 20 blocks denoted by the letters U, V, W and Z (Supplementary Information) could not be assigned to an ancestral chromosome because each has a unique composition, probably due to rearrangements in the human or *Tetraodon* genome. Colour codes are as in Fig. 6.

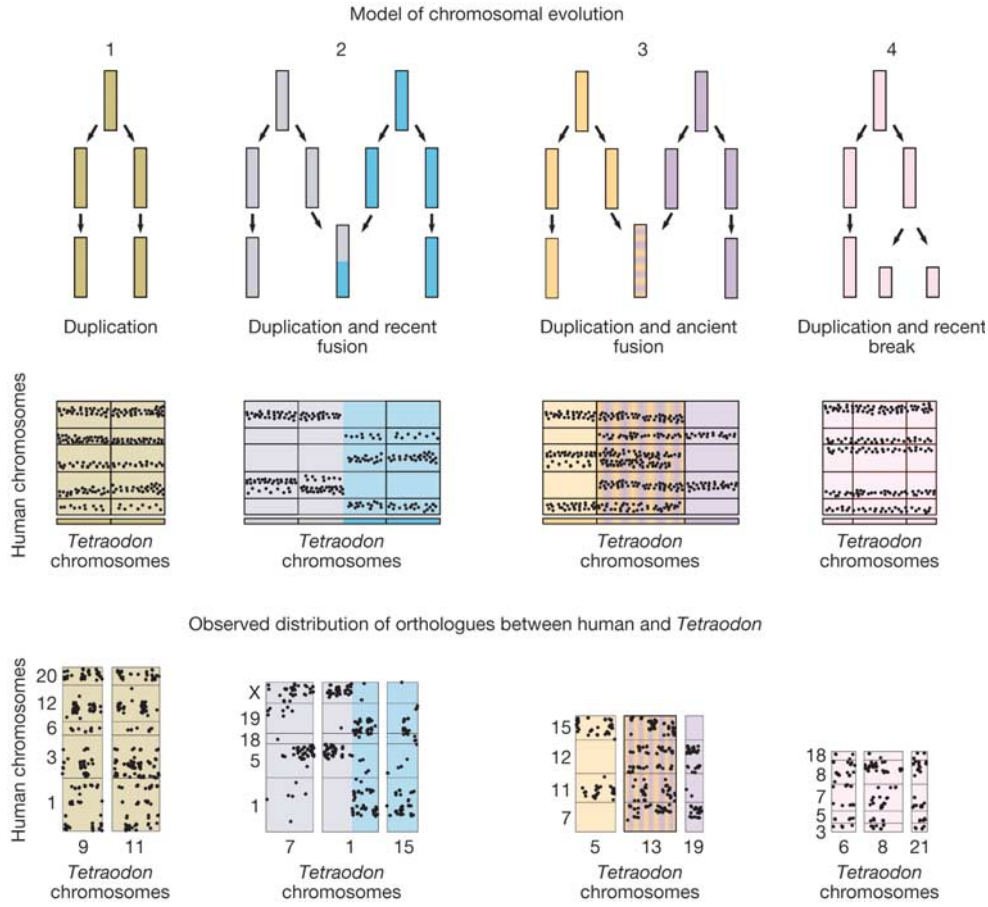


Figure 8 Reconstructing ancient genome rearrangements. Model of chromosome duplication followed by the four simplest chromosome rearrangements: (1) no rearrangement; (2) two different duplicate copies fused recently; (3) two different duplicate copies fused early after the duplication; (4) a duplicate chromosome fragmented very recently. In each model, the distribution of human orthologues from a given chromosomal region on two or three duplicate *Tetraodon* chromosomal regions is expected to be different (each dot is an orthologue, positioned in the human genome on the vertical axis and in the *Tetraodon* genome on the horizontal axis). The distinction between early or late events follows the assumption that intrachromosomal shuffling progressively redistributes genes over a given chromosome. A recent fusion would thus bring together two sets of genes that appear compartmented on their respective segments, whereas an ancient fusion shows the same pattern except that genes have been redistributed over the length of the fused chromosome. It should be noted that a fifth case exists, consisting of a chromosome break early after duplication but it is not represented here. The lower panel shows excerpts of data illustrating the four types of event. The complete Oxford grid is shown in Supplementary Fig. SI12.

©2004 Nature Publishing Group

The presence of supernumerary HOX clusters in zebrafish[7], *Tetraodon* (Fig. S8) and many other percomorphs[29] but not in the bichir *Polypterus senegalus*[30] indicates that the event has affected most teleosts but not all actinopterygians. This timing early in the teleost lineage is in agreement with recent evolutionary analyses in *Takifugu* that estimated the divergence time for most duplicated gene pairs at ~320–350 Myr ago[31,32].

The analyses above also shed light on the rate of intra- and interchromosomal exchange. The synteny analysis shows extensive syntenic segments in which gene content has been well preserved but gene order has been extensively scrambled (striking examples include conserved synteny of Tni20 with human chromosome 4q (Hsa4q) and Tni1 with HsaXq); this is consistent with observations in zebrafish[33]. The duplication analysis within *Tetraodon* also shows that the chromosomal correspondence of duplicated gene pairs has been extensively preserved, whereas local gene order has been largely scrambled. Both analyses thus indicate that a relatively high degree of intrachromosomal rearrangement and a relatively low degree of interchromosomal exchange have taken place in the *Tetraodon* lineage.

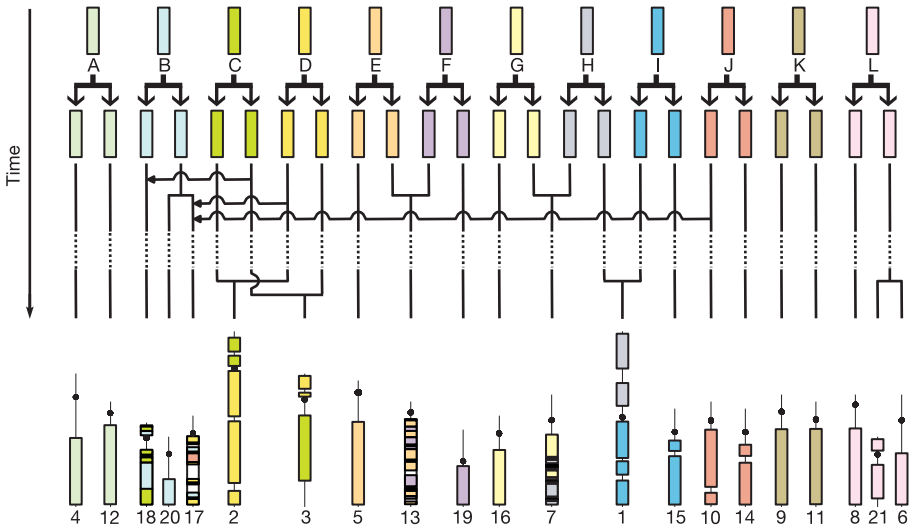


Figure 9 Model for the reconstruction of an ancestral bony vertebrate karyotype comprising 12 chromosomes, based on the pairing information provided by duplicated *Tetraodon* chromosomes showing interleaved patterns on human chromosomes. The ten major rearrangements (two ancient fusions, three recent fusions, one ancient and one recent fission, and three ancient translocations) are deduced by fitting the distribution of orthologues to the four simple theoretical models of chromosome evolution. The order between events is arbitrary although the approximate timeline differentiates between ancient and recent events respectively before and after the dashed line. Arrowheads point to the direction of three ancient translocations.

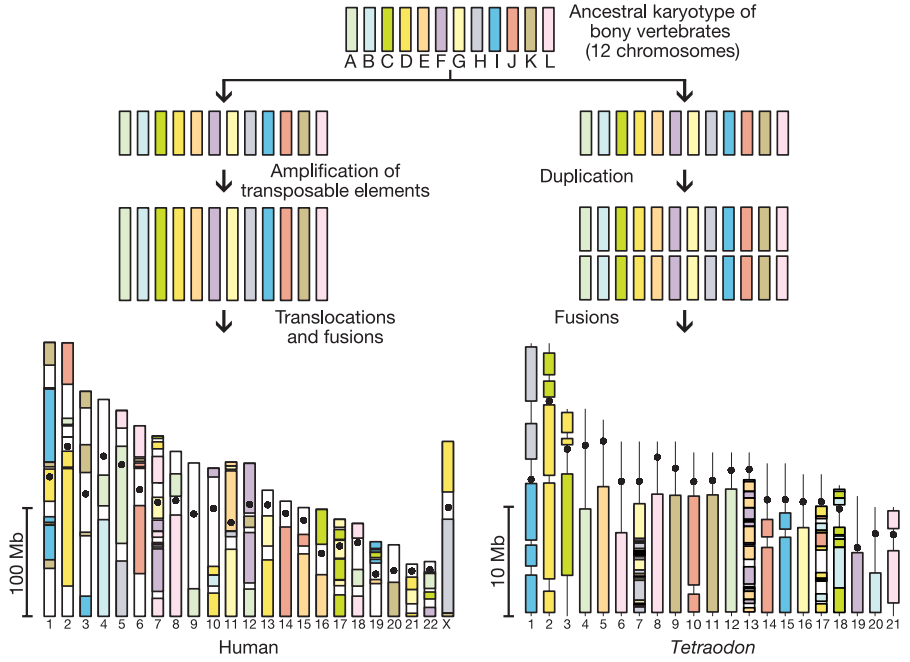


Figure 10 Proposed model for the distribution of ancestral chromosome segments in the human and the *Tetraodon* genomes. The composition of *Tetraodon* chromosomes is based on their duplication pattern (Fig. 9), whereas the composition of human chromosomes is based on the distribution of orthologues of *Tetraodon* genes (Fig. 6). A vertical line in *Tetraodon* chromosomes denotes regions where sequence has not yet been assigned. With 90 blocks in human compared with 44 in *Tetraodon*, the complexity of the mosaic of ancestral segments in human chromosomes underlines the higher frequency of rearrangements to which they were submitted during the same evolutionary period.

©2004 Nature Publishing Group

Ancestral genome of bony vertebrates

We then sought to use the correspondence between the *Tetraodon* and human genomes to attempt to reconstruct the karyotype of their osteichthyan (bony vertebrate) ancestor. The DCS blocks define *Tetraodon* regions that arose from duplication of a common ancestral region. Notably, the DCS blocks largely fall into 12 simple patterns: eight cases involving the interleaving of two current *Tetraodon* chromosomes and four cases involving three current *Tetraodon* chromosomes (Fig. 7 and Table 6). The first group represents cases in which the ancestral chromosomes have remained largely untouched by interchromosomal exchange; the second group represents cases in which one major translocation has occurred.

The distribution of *Tetraodon* orthologues in the human genome (shown as an Oxford grid in Supplementary Fig. S12) provides a detailed record that can be used to partially reconstruct the history of rearrangements in both lineages. We considered the expected distribution resulting from various types of interchromosomal rearrangements, assuming a relatively high degree of intrachromosomal shuffling (Fig. 8; see also Supplementary Information). We found that only ten large-scale interchromosomal events suffice to largely explain the data, connecting an ancestral vertebrate karyotype of 12 chromosomes to the modern *Tetraodon* genome of 21 chromosomes (Fig. 9). Eleven of the *Tetraodon* chromosomes appear to have undergone no major interchromosomal rearrangement. For example, 13 DCS blocks in human are composed of interleaved syntenic groups mapping to Tni9 and Tni11, which are presumed to be derived from a common ancestral chromosome denoted chromosome K (AncK; Fig. 7). The orthologue distribution between the two chromosomes (Fig. 8) confirms that they derive by duplication from AncK (Fig. 9). In a more complex case, Tni13 is systematically interleaved with Tni5 (AncE) or Tni19 (AncF), but Tni5 and Tni19 are never interleaved together; the orthologue distribution among the three chromosomes (Fig. 8) implies that the duplication partners of Tni5 and Tni19 fused soon after the WGD to give rise to Tni13 (Fig. 9). The overall model is consistent with a complete WGD, in that it accounts for all *Tetraodon* chromosomes.

Several lines of evidence support the historical reconstitution presented here. First, the pairing of *Tetraodon* chromosomes agrees with the independently derived distribution of duplicated genes in the genome (Fig. 4b). Second, centric fusions of the three largest chromosomes are consistent with cytogenetic studies[34], and the recent timing of the fusion leading to Tni1 is supported by cytogenetic studies showing its absence in *Takifugu*[35]. Third, the modal value for the haploid number of chromosomes in teleosts is 24 (refs 36–38), consistent with a WGD of an ancestral genome composed of 12 chromosomes.

The analysis also sheds light on genome evolution in the human lineage, with the interleaving patterns on human chromosomes delineating the mosaic of ancestral segments in the human genome (Figs 6 and 10). The results are consistent with and extend several known cases of rearrangements in the human lineage. The model correctly shows the recent fusion of two primate chromosomes leading to Hsa2 (ref. 39) occurring at the junction between two ancestral segments (D2 and D3; Fig. 6) in 2q13.2-2q14.1. It shows HsaXp and HsaXq to be of different origins (corresponding to AncD and AncH, respectively), consistent with the fact that HsaXp is known to be absent in non-placental mammals[40]. The map indicates that most of HsaXq and Hsa5q were once part of the same chromosome, but that the tip of HsaXq (Xq28) originates from a different ancestral segment and is thus a later addition. Some pairs of human chromosomes show similar or identical compositions, suggesting that they derived by fission from the same ancestral chromosome, with examples being Hsa13–Hsa21 and Hsa12–Hsa22; the latter case is consistent with cytogenetic studies showing that a fission occurred in the primate lineage[41].

The results show a major difference in the evolutionary forces shaping the *Tetraodon* and the human genomes (Fig. 10). Whereas 11 *Tetraodon* chromosomes did not undergo interchromosomal exchange over 450 Myr, only one human chromosome (Hsa14) was similarly undisturbed. Hsa7 is an extreme case, with contributions from six ancestral chromosomes. A possible explanation for the difference may be the massive integration of transposable elements in the human genome. The presence of transposable elements may increase the overall frequency of chromosome breaks, as well as the likelihood that a chromosome break fails to disrupt a gene (by increasing the size of intergenic intervals). It will be interesting to see whether teleosts that carry many more transposable elements (such as zebrafish) show a higher frequency of interchromosomal exchanges.

Conclusion

The purpose of sequencing the *Tetraodon* genome was to use comparative analysis to illuminate the human genome in particular and vertebrate genomes in general. The *Tetraodon* sequence, which has been made freely available during the course of this project, has already had a major impact on human gene annotation. It has provided the first clear evidence of a sharply lower human gene count[6] and has been used in the annotation of several human chromosomes[42–45]. Here, we show that it suggests an additional ~900 predicted genes in the human genome. Given its compact size, the *Tetraodon* genome will probably also prove valuable in identifying key conserved regulatory features in intergenic and intronic regions.

In addition, the *Tetraodon* genome provides fundamental insight into genome evolution in the vertebrate lineage. First, the analysis here shows that *Tetraodon* is the descendant of an ancient WGD that most probably affected all teleosts. Together with the recent demonstration of an ancient WGD in the yeast lineage, this suggests that WGD followed by massive gene loss may be an extremely important mechanism for eukaryote genome evolution—perhaps because it allows for the neofunctionalization of entire pathways rather than simply individual genes. There remains a fierce debate about whether one or more earlier WGD events occurred in early vertebrate evolution[25,46–50], with no direct and conclusive evidence found so far[51,52]. The examples of yeast and *Tetraodon* show that ultimate proof will probably best come from the sequence of a related non-duplicated species. An obvious candidate is amphioxus, as its non-duplicated status is supported by the presence of many single-copy genes (including one HOX cluster[53]) instead of two or more in vertebrates, and it is among our closest non-vertebrate relatives based on anatomical and evolutionary observations.

Second, the remarkable preservation of the *Tetraodon* genome after WGD makes it possible to infer the history of vertebrate chromosome evolution. The model suggests that the ancestral vertebrate genome was comprised of 12 chromosomes, was compact, and contained not significantly fewer genes than modern vertebrates (inasmuch as the WGD and subsequent massive gene loss resulted in only a tiny fraction of duplicate genes being retained). The explosion of transposable elements in the mammalian lineage, subsequent to divergence from the teleost lineage, may have provided the conditions for increased interchromosomal rearrangements in mammals; in contrast, the *Tetraodon* genome underwent much less interchromosomal rearrangement.

With the availability of additional vertebrate genomes (dog, marsupial, chicken, medaka, zebrafish and frog are underway), it will be possible to explore intermediate nodes such as the last common ancestor of amniotes, of sarcopterygians and of actinopterygians, and to gain an increasingly clearer picture of the early vertebrate ancestor. Because the early vertebrate genome is 'closer' to current invertebrates, this should in turn facilitate comparison between vertebrate and invertebrate evolution. □

©2004 Nature Publishing Group

Methods

Sequencing, assembly and data access

Sequencing was performed as described previously for Genoscope[54] and the Broad Institute[1,2]. Approximately 4.2 million plasmid reads were cloned and sequenced from DNA extracted from two wild *Tetraodon* fish and passed extensive checks for quality and source, representing approximately 8.3-fold sequence coverage of the *Tetraodon* genome. To alleviate problems due to polymorphism, the assembly proceeded in four stages: (1) reads from a single fish were assembled by Arachne as described previously[10,11]; (2) reads from the second individual were added to increase sequencing depth; (3) scaffolds were constructed using plasmid and BAC paired reads; and (4) contigs from a separate assembly combining both individuals were added if they did not overlap with the first assembly. The final assembly can be downloaded from the EMBL/GenBank/DDBJ databases under accession number CAAE01000000. Full-length *Tetraodon* cDNAs have been submitted under accession numbers CR631133–CR735083. Ultracontigs organized in chromosomes are available from http://www.genoscope.org/tetraodon. This site also contains an annotation browser and further information on the project.

Gene annotation

Protein-coding genes were predicted by combining three types of information: alignments with proteins and genomic DNA from other species, *Tetraodon* cDNAs, and *ab initio* models. All alignments with genomic DNA from human and mouse were performed with Exofish as described previously[6], whereas a new Exofish method was developed to align *Takifugu* genomic DNA. Proteins predicted from human and mouse were also matched using Exofish and a selected subset was then aligned using Genewise. The integration of these data sources was performed with GAZE[14]. A specific GAZE automaton was designed, and parameters were adjusted on a training set of 184 manually annotated *Tetraodon* genes. See Supplementary Information for details.

Evolution of coding and non-coding DNA

To identify orthologous genes between human, mouse, *Tetraodon*, *Takifugu* and *Ciona*, their predicted proteomes were compared using the Smith–Waterman algorithm and reciprocal best matches were considered as orthologous genes between two species. However, only those genes that were reciprocal best matches between four or five species, and only sites that were aligned between the four or five genes, were further considered to compute the percentage identity, K_a, K_s and fourfold degenerate sites by the PBL method applying Kimura's two-parameter model[55–57]. See Supplementary Information for details.

Genome duplication

A core set of *Tetraodon* duplicated genes was identified by an all-against-all comparison of *Tetraodon* predicted protein using Exofish. Only proteins that matched a single other protein by reciprocal best match were considered further and realigned by the Smith–Waterman algorithm to compute K_a and K_s values. Duplicates with a $K_s > 0.35$ (the amount of neutral substitution since the *Tetraodon–Takifugu* divergence) were considered 'ancient' and used to calculate *P*-values for chromosome pairing (Supplementary Table SI12). Rules for classifying alternating patterns of syntenic groups along human chromosomes in DCS blocks included the following criteria: number of genes in syntenic groups, number of syntenic groups in the DCS region, number of *Tetraodon* chromosomes that alternate, and number of times the same combination of *Tetraodon* chromosomes occur in the human genome. See Supplementary Information for details.

Ancestral genome reconstruction

One category of DCS with the following definition encompassed most orthologues: "alternating series of i syntenic groups that belong to two ($i >= 2$) or three ($i >= 3$) *Tetraodon* chromosomes. The series may only be interrupted by groups from categories 'unassigned singletons' or 'background singletons'. A given combination of two or three *Tetraodon* chromosomes must appear at least twice in the human genome". These DCS blocks showed 12 recurring combinations of *Tetraodon* chromosomes, and were thus further classified in 12 groups labelled A to L. Each of the 12 groups, consisting of at least two DCS blocks with the same combination of alternating *Tetraodon* chromosomes, represents a proto-chromosome from the ancestral bony vertebrate (Osteichthyes). A model was then designed to account for the possible fates of chromosomes after duplication of the ancestral genome in the teleost lineage (Fig. 8). The model only deals with orthologous gene distribution between two genomes. It is simply based on the postulate that interchromosomal shuffling of genes within a genome increases with time, which is a measure to distinguish between ancient and recent events (for example, chromosome fusions or fissions). The two-dimensional distribution of 7,903 *Tetraodon*–human orthologues (Oxford Grid, Supplementary Fig. S12) was then confronted to the model and all 21 *Tetraodon* chromosomes could be grouped in pairs or triplets and assigned to a given type of event. See Supplementary Information for details.

Received 14 July; accepted 8 September 2004; doi:10.1038/nature03025.

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
2. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
3. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–521 (2004).
4. Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297,** 1301–1310 (2002).
5. Hedges, S. B. The origin and evolution of model organisms. *Nature Rev. Genet.* **3,** 838–849 (2002).
6. Roest Crollius, H. *et al.* Human gene number estimate provided by genome wide analysis using *Tetraodon nigroviridis* genomic DNA. *Nature Genet.* **25,** 235–238 (2000).
7. Amores, A. *et al.* Zebrafish hox clusters and vertebrate genome evolution. *Science* **282,** 1711–1714 (1998).
8. Robinson-Rechavi, M., Marchand, O., Escriva, H. & Laudet, V. An ancestral whole-genome duplication may not have been responsible for the abundance of duplicated fish genes. *Curr. Biol.* **11,** R458–R459 (2001).
9. Taylor, J. S., Braasch, I., Frickey, T., Meyer, A. & Van de Peer, Y. Genome duplication, a trait shared by 22000 species of ray-finned fish. *Genome Res.* **13,** 382–390 (2003).
10. Batzoglou, S. *et al.* ARACHNE: a whole-genome shotgun assembler. *Genome Res.* **12,** 177–189 (2002).
11. Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13,** 91–96 (2003).
12. Roest Crollius, H. *et al.* Characterization and repeat analysis of the compact genome of the freshwater pufferfish *Tetraodon nigroviridis*. *Genome Res.* **10,** 939–949 (2000).
13. Bouneau, L. *et al.* An active non-LTR retrotransposon with tandem structure in the compact genome of the pufferfish *Tetraodon nigroviridis*. *Genome Res.* **13,** 1686–1695 (2003).
14. Howe, K. L., Chothia, T. & Durbin, R. GAZE: a generic framework for the integration of gene-prediction data by dynamic programming. *Genome Res.* **12,** 1418–1427 (2002).
15. Hatfield, D. L. *Selenium: Its Molecular Biology and Role in Human Health* (Kluwer, Dordrecht, 2001).
16. Boulay, J. L., O'Shea, J. J. & Paul, W. E. Molecular phylogeny within type I cytokines and their cognate receptors. *Immunity* **19,** 159–163 (2003).
17. Mulder, N. J. *et al.* InterPro: an integrated documentation resource for protein families, domains and functional sites. *Brief. Bioinform.* **3,** 225–235 (2002).
18. Dehal, P. *et al.* The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298,** 2157–2167 (2002).
19. Zdobnov, E. M. & Apweiler, R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17,** 847–848 (2001).
20. Harris, M. A. *et al.* The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32** (Database issue), D258–D261 (2004).
21. Torrents, D., Suyama, M., Zdobnov, E. & Bork, P. A genome-wide survey of human pseudogenes. *Genome Res.* **13,** 2559–2567 (2003).
22. Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci.* **17,** 57–86 (1986).
23. Gu, X. & Li, W. H. A general additive distance with time-reversibility and rate variation among nucleotide sites. *Proc. Natl Acad. Sci. USA* **93,** 4671–4676 (1996).
24. Holland, P. W. H. Introduction: gene duplication in development and evolution. *Semin. Cell Dev. Biol.* **10,** 515–516 (1999).
25. Martin, A. Is tetralogy true? Lack of support for the "one-to-four" rule. *Mol. Biol. Evol.* **18,** 89–93 (2001).
26. Wolfe, K. H. Yesterday's polyploids and the mystery of diploidization. *Nature Rev. Genet.* **2,** 333–341 (2001).
27. Kellis, M., Birren, B. W. & Lander, E. S. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428,** 617–624 (2004).
28. Dietrich, F. S. *et al.* The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* **304,** 304–307 (2004).
29. Prohaska, S. J. & Stadler, P. F. The duplication of the Hox gene clusters in teleost fishes. *Theor. Biosci.* **123,** 89–110 (2004).
30. Chiu, C. H. *et al.* Bichir HoxA cluster sequence reveals surprising trends in ray-finned fish genomic evolution. *Genome Res.* **14,** 11–17 (2004).
31. Vandepoele, K., De Vos, W., Taylor, J. S., Meyer, A. & Van de Peer, Y. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc. Natl Acad. Sci. USA* **101,** 1638–1643 (2004).
32. Christoffels, A. *et al.* Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol. Biol. Evol.* **21,** 1146–1151 (2004).
33. Woods, I. G. *et al.* A comparative map of the zebrafish genome. *Genome Res.* **10,** 1903–1914 (2000).
34. Fischer, C. *et al.* Karyotype and chromosomal localization of characteristic tandem repeats in the pufferfish *Tetraodon nigroviridis*. *Cytogenet. Cell Genet.* **88,** 50–55 (2000).
35. Grutzner, F. *et al.* Classical and molecular cytogenetics of the pufferfish *Tetraodon nigroviridis*. *Chromosome Res.* **7,** 655–662 (1999).
36. Ohno, S., Wolf, U. & Atkin, N. B. Evolution from fish to mammals by gene duplication. *Hereditas* **59,** 169–187 (1968).
37. Ojima, Y. in *Chromosomes in Evolution of Eukaryotic Groups* (eds Sharma, A. K. & Sharma, A.) 111–145 (CRC Press, Boca Raton, 1983).
38. Naruse, K. *et al.* A medaka gene map: the trace of ancestral vertebrate proto-chromosomes revealed by comparative gene mapping. *Genome Res.* **14,** 820–828 (2004).
39. Yunis, J. J. & Prakash, O. The origin of man: a chromosomal pictorial legacy. *Science* **215,** 1525–1530 (1982).
40. Graves, J. A., Gecz, J. & Hameister, H. Evolution of the human X—a smart and sexy chromosome that controls speciation and development. *Cytogenet. Genome Res.* **99,** 141–145 (2002).
41. Richard, F., Lombard, M. & Dutrillaux, B. Reconstruction of the ancestral karyotype of eutherian mammals. *Chromosome Res.* **11,** 605–618 (2003).
42. The chromosome 21 mapping and sequencing consortium, The DNA sequence of human chromosome 21. *Nature* **405,** 311–319 (2000).
43. Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414,** 865–871 (2001).
44. Collins, J. E. *et al.* Reevaluating human gene annotation: a second-generation analysis of chromosome 22. *Genome Res.* **13,** 27–36 (2003).
45. Heilig, R. *et al.* The DNA sequence and analysis of human chromosome 14. *Nature* **421,** 601–607 (2003).
46. Holland, P. W., Garcia-Fernandez, J., Williams, N. A. & Sidow, A. Gene duplications and the

©2004 Nature Publishing Group

origins of vertebrate development. *Development* (suppl.), 125–133 (1994).
47. Spring, J. Vertebrate evolution by interspecific hybridisation–are we polyploid? *FEBS Lett.* **400,** 2–8 (1997).
48. Friedman, R. & Hughes, A. L. Pattern and timing of gene duplication in animal genomes. *Genome Res.* **11,** 1842–1847 (2001).
49. Hughes, A. L., da Silva, J. & Friedman, R. Ancient genome duplications did not structure the human Hox-bearing chromosomes. *Genome Res.* **11,** 771–780 (2001).
50. Thornton, J. W. Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions. *Proc. Natl Acad. Sci. USA* **98,** 5671–5676 (2001).
51. McLysaght, A., Hokamp, K. & Wolfe, K. H. Extensive genomic duplication during early chordate evolution. *Nature Genet.* **31,** 200–204 (2002).
52. Panopoulou, G. *et al.* New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. *Genome Res.* **13,** 1056–1066 (2003).
53. Garcia-Fernandez, J. & Holland, P. W. Archetypal organization of the amphioxus Hox gene cluster. *Nature* **370,** 563–566 (1994).
54. Artiguenave, F. *et al.* Genomic exploration of the hemiascomycetous yeasts: 2. Data generation and processing. *FEBS Lett.* **487,** 13–16 (2000).
55. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16,** 111–120 (1980).
56. Li, W. H., Wu, C. I. & Luo, C. C. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* **2,** 150–174 (1985).
57. Pamilo, P. & Bianchi, N. O. Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. *Mol. Biol. Evol.* **10,** 271–281 (1993).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements This work was supported by Consortium National de Recherche en Génomique. We thank T. Itami and S. Watabe for their gift of *Takifugu* blood samples; C. Nardon and M. Weiss for help with flow cytometry experiments; K. Howe for discussions regarding GAZE; R. Heilig for help with the annotation; the Centre Informatique National de l'Enseignement Supérieur for computer resources; and Gene-IT for assistance with the Biofacet software package.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.W. (jsbach@genoscope.cns.fr). The final assembly is available at EMBL/GenBank/DDBJ under accession number CAAE01000000. Full-length *Tetraodon* cDNAs have been deposited under accession numbers CR631133–CR735083; ultracontigs organized in chromosomes are available from http://www.genoscope.org/tetraodon.

©2004 Nature Publishing Group

Meeting report

Large-scale discovery and validation of functional elements in the human genome

Bradley E Bernstein*† and Manolis Kellis*‡

Addresses: *Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA 02139, USA. †Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA. ‡MIT Computer Science and Artificial Intelligence Laboratory, The Stata Center, 32 Vassar Street, Cambridge, MA 02139, USA.

Correspondence: Bradley E Bernstein. E-mail: bbernst@fas.harvard.edu. Manolis Kellis. E-mail: manoli@mit.edu

Published: 1 March 2005

Genome ***Biology*** 2005, **6:**312

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2005/6/3/312

© 2005 BioMed Central Ltd

A report on the genomics workshop 'Identification of Functional Elements in Mammalian Genomes', Cold Spring Harbor, New York, 11-13 November 2004.

Computational and experimental genomics researchers convened at Cold Spring Harbor Laboratory at the end of 2004 to address the ambitious goal of identifying all the functional elements in the human genome. The functional elements discussed at the meeting included protein-coding genes, regulatory elements, RNA genes and DNA sequences that dictate chromosome structure or replication. The presentations described diverse approaches to the problem, ranging from innovative comparative genomic methods to high-throughput functional assays designed to identify and validate such elements.

The meeting followed a gathering of the ENCODE consortium, which aims to identify a comprehensive 'Encyclopedia of DNA elements' in the human genome [http://www.genome.gov/10005107]. The consortium, organized and funded by the National Human Genome Research Institute, is initially focusing on designated regions comprising approximately 1% of the human genome, and it strongly emphasizes technology development. Participating laboratories are developing computational techniques for sequence assembly, gene identification and regulatory motif discovery, as well as experimental methods for identifying transcripts, chromatin structures and regulatory regions. Against this backdrop, most speakers described experimental and computational approaches that generated vast numbers of candidate functional elements, from comprehensive transcript catalogs to lists of highly conserved sequence elements. They were complemented by a smaller number of presentations dealing with the daunting task of systematically validating these elements.

Genes and transcripts

Mike Snyder (Yale University, New Haven, USA) and Tom Gingeras (Affymetrix, Santa Clara, USA) covered genome-scale technologies for transcript identification and the large numbers of new candidate elements emerging from these studies. Snyder described a complete tiling of the non-repetitive human genome, with 134 arrays containing 52 million oligonucleotide probes. In an effort to identify transcriptionally active regions systematically, his group hybridized RNA extracted from liver against these arrays. Gingeras described similar arrays covering one third of the genome that have been used to screen several human cell lines for transcribed sequences. Both studies identified surprisingly large and diverse collections of transcripts, a high proportion of which do not correspond to existing gene annotations. The presenters pointed out that a major challenge is ahead to define the functional significance of the thousands of novel transcripts identified.

Gustavo Glusman (Institute for Systems Biology, Seattle, USA) presented an orthogonal computational approach to gene identification, which does not observe transcription directly but instead relies on the marks it leaves behind in a genome sequence. He described four computational signatures of transcribed sequences, each relying on a different side-effect of transcription. One signature is the increased frequency of G and T nucleotides observed in the coding strand of genes, which is attributable to a mutational bias introduced during transcription-coupled DNA repair. Another signature is a bias in the orientation of transposable elements within genes, which is attributable to the fact that

polyadenylation signals in the transposon are rejected if they occur early in the coding strand but can be tolerated on the reverse strand. Taken together, the four tests provide a new tool for gene identification, which performs best where other tools fail. For example, the signals observed are strongest for long genes with very small exons, where traditional tools based on hidden Markov models fail. Using these methods on the human genome, Glusman has found evidence for thousands of genes that have not previously been annotated, some of which he has validated experimentally.

Regulatory mechanisms

Tim Hubbard (Wellcome Trust Sanger Institute, Hinxton, UK) presented a new computational approach to the discovery of regulatory motifs in promoter regions. Instead of searching for single motifs, the technique looks for multiple query sequences simultaneously and effectively explores motif space in bulk. The strength of the approach comes from the parallel exploration, which makes it well suited for distinguishing promoters densely populated with interacting regulatory motifs. Although initial results in yeast were promising, Hubbard pointed out a number of computational and implementation challenges in scaling this approach to mammalian genomes. Once these challenges have been overcome, this holistic approach to motif discovery will offer an additional way of obtaining a global understanding of regulatory elements.

John Stamatoyannopoulos (Regulome, Seattle, USA) and Gregory Crawford (National Human Genome Research Institute (NHGRI), Bethesda, USA) presented methodologies for the systematic identification of regions of open chromatin in the human genome. The methods involve mapping DNase I hypersensitive sites, which are known to correlate with many types of functional elements, including promoters, enhancers and insulators (sites at boundaries between open chromatin and inactive heterochromatin). Both investigators have cloned sites cut by DNase I, carried out highly parallel sequencing to map them onto the genome, and integrated the resulting maps with the University of California at Santa Cruz (UCSC) genome browser [http://genome.ucsc.edu] for visualization. They found that hypersensitive sites correlate with transcription starts, CpG islands and regions of high sequence conservation at the genome-wide level. They are following up on their findings by screening hypersensitive sites in multiple cell types (such as the ones listed on the ENCODE website [http://www.genome.gov/10005107]) to assess their tissue specificities, and to discover additional candidate elements.

Large-scale validation

A smaller number of presentations dealt with the validation problem. Nathan Trinklein (Stanford University, USA) focused on human promoters and described high-throughput methods for validating computationally predicted regulatory regions. Predicted promoters were cloned upstream of reporter genes and their activity tested in various cell lines using transient transfection. Gabriela Loots (Lawrence Livermore National Laboratory, Livermore, USA) described functional assays for high-throughput validation of genes and regulatory sequences in the tropical frog *Xenopus tropicalis*. Transgenic techniques in frog embryos are used to test the influence of regulatory sequences on gene-expression patterns previously analyzed by *in situ* hybridization.

Greg Elgar (MRC Rosalind Franklin Centre for Genomics Research, Cambridge, UK) described methods for validating noncoding functional elements predicted through comparative genomic analysis. His group has identified more than 1,000 noncoding sequences that are highly conserved between human and pufferfish (*Fugu*) genomes. These candidate elements tended to reside near genes that act as developmental regulators, and were not found in invertebrate genomes. Zebrafish embryos were used to test a subset of the conserved elements for their regulatory potential. Candidate regions were amplified by PCR and co-injected with green fluorescent protein (GFP) reporter constructs. A remarkably high proportion of the highly conserved noncoding sequences tested (23 of 25) were found to enhance GFP expression in a tissue-specific manner.

The many state-of-the-art technologies being applied to the identification of functional elements in genomes are producing huge numbers of candidate regions. High-throughput assays in human cells and model organisms for validating and functionally characterizing these candidates are critical to the overall goal of cataloging functional elements. Given the huge numbers of candidates, however, alternative approaches are also needed. A particularly promising technique for validating, characterizing and prioritizing candidate regions is cross-validation by simultaneous analysis of complementary experimental and computational datasets, and the ENCODE consortium is seeking to maximize its potential by focusing on a well defined subset of the human genome and incorporating computational tools for correlating multiple datasets. Beyond simply cataloging functional elements, this integration should also lead to a description of their complex interactions within the regulatory network of the cell.

Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals

Xiaohui Xie[1], Jun Lu[1], E. J. Kulbokas[1], Todd R. Golub[1], Vamsi Mootha[1], Kerstin Lindblad-Toh[1], Eric S. Lander[1,2]* & Manolis Kellis[1,3]*

[1]*Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02141, USA*
[2]*Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02139, USA*
[3]*Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

* These authors contributed equally to this work

Comprehensive identification of all functional elements encoded in the human genome is a fundamental need in biomedical research. Here, we present a comparative analysis of the human, mouse, rat and dog genomes to create a systematic catalogue of common regulatory motifs in promoters and 3′ untranslated regions (3′ UTRs). The promoter analysis yields 174 candidate motifs, including most previously known transcription-factor binding sites and 105 new motifs. The 3′-UTR analysis yields 106 motifs likely to be involved in post-transcriptional regulation. Nearly one-half are associated with microRNAs (miRNAs), leading to the discovery of many new miRNA genes and their likely target genes. Our results suggest that previous estimates of the number of human miRNA genes were low, and that miRNAs regulate at least 20% of human genes. The overall results provide a systematic view of gene regulation in the human, which will be refined as additional mammalian genomes become available.

Comparative genomics provides a powerful approach for the systematic discovery of functional elements in the human genome[1–8], by virtue of their evolutionary conservation across related species. Recent studies have revealed that most functional elements are non-protein-coding; these are likely to include regulatory signals, RNA genes and structural elements. Although the largest and most conserved elements are readily identified[9], the vast majority of non-coding functional elements remain unknown. It has been particularly difficult to recognize short regulatory sequences, such as the DNA binding sites for transcription factors.

Our goal here is to create a catalogue of 'common regulatory motifs', by which we mean short, functional sequences (typically, 6–10 bases) that are used many times in a genome. Although reliably recognizing each individual occurrence of a motif requires evolutionary comparison with many species[10], it should be possible to define the motifs themselves with many fewer species based on their multiple conserved occurrences in the genome. Recent studies of four related yeast species have shown the potential power of this approach in a relatively small genome[11,12].

To apply this approach to the vastly larger human genome, we focused on two limited subsets that are likely to be highly enriched for regulatory elements: promoter regions and 3′ UTRs of protein-coding genes. We aligned these regions across the human, mouse, rat and dog genomes[13–15], and searched for highly conserved, frequently occurring sequence patterns. In promoter regions, we discovered 174 candidate motifs, including most previously known transcription-factor binding sites and 105 new motifs that seem to be functional based on multiple criteria. In 3′ UTRs, we discovered 106 motifs likely to be involved in post-transcriptional regulation. Nearly one-half of these are associated with miRNAs, demonstrating the extraordinary importance of this recently discovered regulatory mechanism[16]. Our results suggest that previous estimates of human miRNA number were low[17], and that at least 20% of human genes are probably regulated by miRNAs. More broadly, our results suggest that it will be possible to construct a comprehensive catalogue of common regulatory motifs in the human genome as more species become available.

Alignment of promoters and 3′ UTRs

We first constructed genome-wide alignments for the four mammalian species in promoter regions and 3′ UTRs. We studied a total of ~17,700 well-annotated genes from the RefSeq database[18], with a single reference messenger RNA selected for each gene. Each promoter region was defined as the non-coding sequence contained within a 4-kilobase (kb) window centred at the annotated transcriptional start site (TSS), and each 3′ UTR was defined based on the annotation of the reference mRNA (see Methods). The promoter set contains ~68 megabases (Mb) of sequence, whereas the 3′-UTR set comprises ~15 Mb. In addition, we defined a control set comprising ~123 Mb of intronic sequence consisting of the last two introns from the genes; these sequences were used as a control because terminal introns are thought to contain relatively fewer regulatory elements.

The human sequence in each set was aligned to the other three mammalian genomes[19,20]. Some segments could not be aligned in all four species, primarily because they represent new sequence recently inserted by transposons, ancestral sequence deleted in one of the other species, or regions still missing from the draft genome sequences (see Supplementary Information). The proportion of bases that could be aligned across all four species was ~51% for promoters (44% upstream and 58% downstream of the TSS) and ~73% for the 3′UTRs. These proportions are much higher than for the control intronic regions (34%) or for the whole genome (28%), presumably reflecting the presence of important conserved elements (see Supplementary Information).

The total number of aligned bases is ~35 Mb for the promoters and ~11 Mb for the 3′ UTRs, which is comparable to the size of yeast genomes. Additionally, the evolutionary tree (Fig. 1a) of the four mammals has a total branch length similar to that used for comparative analysis of *Saccharomyces* species[11] (0.68 substitutions per base in human promoters versus 0.83 in yeast intergenic regions; see Supplementary Information).

Conservation properties of regulatory motifs

Our comparative analysis can be illustrated by considering a known

© 2005 Nature Publishing Group

regulatory motif. The 8-mer TGACCTTG is known to be a binding site of the Err-α protein and to occur in the promoters of many genes induced during mitochondrial biogenesis[21,22]. The promoter of the *GABPA* gene contains a well-studied Err-α-binding site, which is conserved across all four species, and stands out from the non-conserved flanking sequence (Fig. 1b). More generally, the Err-α motif occurs 434 times in human promoter regions and 162 of these occurrences are conserved across all four species; the conservation rate is thus 37%. By contrast, a random 8-mer motif shows a markedly lower conservation rate in promoters (6.8%). Moreover, the high conservation of the Err-α motif is specific to promoter regions; it shows a much lower conservation rate in introns (6.2%).

We quantified the extent of excess conservation by using a motif conservation score (MCS), which essentially represents the number of standard deviations (s.d.) by which the observed conservation rate of a motif exceeds the expected conservation rate for comparable random motifs (see Methods). A 'conserved occurrence' of a (possibly degenerate) motif is an instance in which an exact match to the motif is found in all four species; the 'conservation rate' of a motif is defined to be the ratio of conserved occurrences to total occurrences in the human genome. Comparable random motifs are generated by sampling human genomic sequence in promoters (divided according to high or low CpG content) or 3′ UTRs in order to ensure equivalent sequence composition (see Methods). For the Err-α-binding motif, the MCS is 25.2 s.d. (the binomial probability of observing 162 conserved instances out of 434, given an expected rate of conservation of 6.8%).

We examined the conservation properties of known motifs in mammalian promoters using the TRANSFAC database of reported transcription-factor binding sites[23]. The 446 motifs in TRANSFAC were clustered on the basis of sequence similarity, resulting in 123 motif clusters (see Methods). Most of these 123 motifs had high MCS, with 63% having MCS > 3 and nearly one-half having MCS > 5. By contrast, a control set of random motifs has only 1.6% with MCS > 3 and none with MCS above 5. The TRANSFAC motifs with low conservation scores have three likely explanations: they may be erroneous or over-specified (there is no uniform standard of evidence for inclusion in the database), they may have diverged between the species compared, or they may have too few biologically functional occurrences for the conserved instances to stand out with sufficient statistical significance.

For 3′ UTRs, few common motifs have been defined and there is no database analogous to TRANSFAC. One well-known example is the polyadenylation signal (AATAAA). This motif indeed shows a high conservation rate, with 6,617 out of 14,266 (46%) occurrences in 3′ UTRs being conserved (conservation rate = 46%, MCS = 135). Control random motifs show a conservation rate of only 10% in 3′ UTRs, and the polyadenylation signal shows a conservation rate of only 6% in the intronic controls.

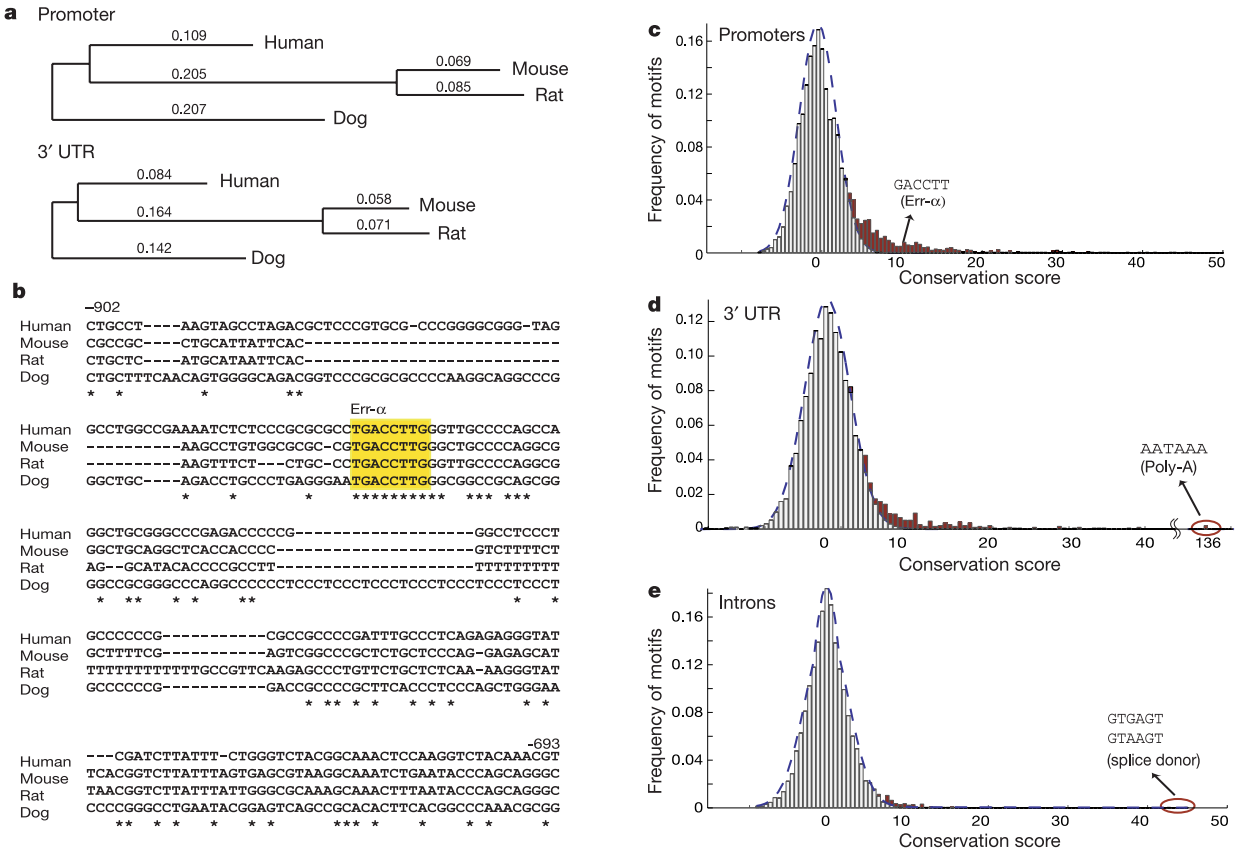


Figure 1 Conservation properties in human promoter regions and 3′ UTRs. **a**, Evolutionary tree relating the four mammalian species. Branch lengths denote number of substitutions per site. Average nucleotide per cent identity to human is 62% for mouse, 60% for rat and 69% for dog in promoter regions, and respectively 68%, 67% and 76% in 3′ UTRs. **b**, Conservation in GABPA promoter region reveals functional Err-α motif. Asterisks denote conserved bases. The yellow box marks the experimentally validated Err-α-binding site. **c**–**e**, Excess conservation in promoter and 3′-UTR regions reveals short sequences under evolutionary selection. Motif conservation score (MCS) distribution is shown for all 6-mer motifs in aligned promoters (**c**), 3′-UTR regions (**d**) and introns (**e**). The dashed curve shows fit to gaussian distribution. Excess conservation relative to this distribution is shown in red.

© 2005 Nature Publishing Group
NATURE | doi:10.1038/nature03441 | www.nature.com/nature

Discovery of new motifs

We next explored the general conservation properties of short sequences within promoters and 3′ UTRs. Specifically, we measured the conservation rate of all 6-mer sequences. For both regions, there is a large excess of 6-mers with high MCS scores (Fig. 1c, d). The distribution of MCS scores suggests that a significant proportion of 6-mers are under purifying selection (13% in promoters and 8.6% in 3′ UTRs, indicated by the shaded regions in Fig. 1c, d). By contrast, intronic regions have a largely symmetric distribution (Fig. 1e) with only a modest number of outliers. (The two most extreme cases correspond to known mRNA splice donor signals, GTGAGT and GTAAGT.)

On the basis of these properties, we set out to systematically discover common regulatory motifs in promoters and 3′ UTRs. We evaluated motifs containing 6–18 bases, with degeneracy allowed at each position; a computational algorithm was developed to make it feasible to screen $\sim 10^{12}$ possible motifs across the genomic regions (see Methods). Motifs with high conservation scores (MCS $>$ 6) were identified, similar motifs were clustered together and the motif with the highest MCS in each cluster was selected (see Methods). This resulted in a small number of motifs with MCS $>$ 6, which we refer to as 'highly conserved motifs'

Motifs in promoters

In promoters, we found 174 highly conserved motifs. Among these, some could be immediately recognized as known regulatory elements. For example, the three strongest motifs correspond to binding sites for the oncogene regulator ELK-1, the cell-cycle regulator Myc and the mitochondrial respiratory chain regulator NRF-1. Overall, 59 discovered motifs showed strong matches to known motifs and 10 showed weaker matches (see Supplementary Information), together accounting for 72% of the 123 previously known motifs that we assembled from the TRANSFAC database (Supplementary Information). (By contrast, a comparable collection of random motifs would match only ~5% of the TRANSFAC motifs; see Supplementary Information.)

The remaining 105 discovered motifs represent potentially new regulatory elements. The list includes numerous notable examples, some with conservation scores much higher than for most known motifs. For example, the newly discovered motif M_4 (ACTAYRNNNCCCR) occurs 520 times in human promoter regions, of which 317 (61%) are conserved, and the new near-palindromic motif M_8 (TMTCGCGANR) occurs 368 times, of which 236 (64%) are conserved.

In the absence of specific information about the role of these

Table 1 **Top 50 of 174 discovered motifs in human promoters**

No.	Discovered motif	MCS	Known factor*	Conservation rate†	Tissue enrichment‡	Position bias§
1	RCGCAnGCGY	107.8	NRF-1	0.49	15.0	−62
2	CACGTG	85.3	MYC	0.47	8.8	−62
3	SCGGAAGY	80.4	ELK-1	0.44	22.4	−24
4	ACTAYRnnnCCCR	69.5	–	0.61	8.1	−89
5	GATTGGY	64.6	NF-Y	0.51	9.8	−63
6	GGGCGGR	63.9	SP1	0.21	11.4	−63
7	TGAnTCA	62.8	AP-1	0.38	6.5	–
8	TMTCGCGAnR	55.7	–	0.64	9.4	−62
9	TGAYRTCA	55.7	ATF3	0.50	6.1	−66
10	GCCATnTTG	54.7	YY1	0.72	12.2	–
11	MGGAAGTG	51.6	GABP	0.43	13.9	−23
12	CAGGTG	47.6	E12	0.26	9.9	–
13	CTTTGT	46.0	LEF1	0.42	13.6	–
14	TGACGTCA	44.8	ATF3	0.44	4.2	−22
15	CAGCTG	43.9	AP-4	0.27	8.9	–
16	RYTTCCTG	43.0	C-ETS-2	0.32	7.4	−24
17	AACTTT	42.1	*IRF1*	0.43	11.1	–
18	TCAnnTGAY	40.4	SREBP-1	0.47	4.9	−64
19	GKCGCn(7)TGAYG	40.1	–	0.35	5.6	−62
20	GTGACGY	38.4	E4F1	0.34	6.6	−56
21	GGAAnCGGAAnY	37.7	–	0.68	7.0	−33
22	TGCGCAnK	37.4	–	0.24	8.2	−17
23	TAATTA	37.3	CHX10	0.29	7.1	–
24	GGGAGGRR	33.5	MAZ	0.16	9.4	–
25	TGACCTY	33.4	ESRRA	0.30	7.7	–
26	TTAYRTAA	32.6	E4BP4	0.34	6.1	–
27	TGGn(6)KCCAR	32.3	–	0.27	4.5	–
28	CTAWWWATA	32.3	RSRFC4	0.36	7.6	–
29	CTTTAAR	30.8	–	0.43	5.4	–
30	YGCGYRCGC	30.5	–	0.19	5.2	−31
31	GGGYGTGnY	30.0	–	0.24	5.4	−63
32	TGASTMAGC	27.2	NF-E2	0.39	5.4	−66
33	YTATTTTnR	26.4	MEF-2	0.21	7.1	–
34	CYTAGCAAY	26.1	–	0.50	5.2	−142
35	GCAnCTGnY	25.7	MYOD	0.25	8.2	–
36	RTAAACA	25.6	FREAC-2	0.46	7.0	–
37	GTTRYCATRR	25.3	–	0.54	7.6	−56
38	TGACCTTG	25.2	ERR-α	0.37	8.1	–
39	TCCCRnnRTGC	24.3	–	0.30	6.8	−60
40	TTCYnRGAA	24.3	STAT5A	0.19	–	–
41	TGACAGnY	24.1	MEIS1	0.27	6.9	–
42	TGACATY	23.8	–	0.23	5.8	–
43	GTTGnYnnRGnAAC	23.7	–	0.47	4.7	−57
44	YATGnWAAT	23.5	OCT-X	0.53	6.9	–
45	CCAnnAGRKGGC	23.4	–	0.47	–	−101
46	WTTGKCTG	23.0	–	0.25	5.0	−63
47	TGCCAAR	22.9	NF-1	0.25	7.0	–
48	GCGnnAnTTCC	22.8	*C-REL*	0.30	6.0	−12
49	CATTGTYY	22.5	SOX-9	0.43	5.8	–
50	RGAGGAARY	22.4	PU.1	0.22	4.0	–

*Name of the best-matching motif in TRANSFAC database, if any. Weak matches are indicated in italics.
†The percentage of human motif occurrences that match the motif consensus across all four species.
‡A measure of the maximum enrichment of conserved motif occurrences upstream of genes expressed in a compendium of 75 human tissues.
§For motifs with strong positional bias (score $>$ 5 s.d.), the mode of the distance distribution upstream of the TSS.

© 2005 Nature Publishing Group

putative motifs, we used two approaches to demonstrate that most of the new motifs are likely to be biologically meaningful. First, we correlated the presence of motifs with the tissue specificity of gene expression. We reasoned that the genes controlled by a common regulator would often (although not always) show enriched expression in specific sets of tissues. For each motif, we defined the set S_1 of genes with conserved occurrences of the motif and an equal-sized control set S_2 of genes in which the motif occurs in the human genome but is not conserved (see Methods). Using gene expression data from 75 human tissues[24], significant enrichment in one or more tissues (z-score > 4.0) was seen for 59 of the 69 (86%) known motifs and 53 of the 105 (50%) new motifs (Table 1 and Fig. 2). In contrast, the control sets show little or no enrichment across the same tissues (Supplementary Fig. S2). For example, the best-conserved new motifs M_4 and M_8 show enrichment in haematopoietic cells, motif M_{27} (TGGN$_6$KCCAR) is enriched in trachea and lung, and motif M_{29} (CTTTAAR) is enriched in brain-related tissues such as pons, parietal lobe and cingular cortex.

Second, we examined positional bias of the motifs relative to the TSS. Although the analysis covered a 4-kb region surrounding the TSS, the discovered motifs preferentially occur in the human genome within ~100 bases of the TSS, and their conserved occurrences across all four species show an even more marked enrichment in this region (Fig. 3a, b), consistent with the motifs being involved in transcription initiation. Within this overall trend, certain motifs show distinctive positional preferences. Approximately 28% of the known motifs and 35% of the new motifs show significant positional preference (see Methods). For example, the two strongest new motifs (M_4 and M_8) tend to occur at distances centred around −89 and −62 bp upstream of the TSS, respectively (Table 1 and Fig. 3c, d). Overall, 89% of the known motifs and 69% of the new motifs show tissue specificity, positional bias or both. Taken together, these results strongly suggest that the new promoter motifs are likely to be biologically meaningful.

In addition, several of the motifs tend to appear in multiple copies in the promoter. For example, 17.5% of genes containing motif M_4 have more than one copy of M_4 within 200 bp of each other (27-fold increase compared with 0.66% expected at random), and 10% of genes containing motif M_8 have multiple copies of this motif within 200 bp (compared with 0.26% expected by chance). Such clustering is a common feature of several known regulatory motifs.

Motifs in 3′ UTRs

We next turned to motif discovery in 3′-UTR regions[25]. Using the same approach as for promoter regions, we discovered 106 highly

Figure 2 Tissue specificity of expression for genes containing discovered motifs. For each of the 174 motifs, we defined the set of genes whose promoters contain conserved occurrences of the motif, and tested for enriched expression in 75 human tissues. The enrichment score (see Methods) is represented in pseudo-colour, with only scores greater than 4 shown. Motifs are ordered by MCS; asterisks denote new motifs (left); factor names are shown for known motifs (right). Only the top 50 motifs are shown. The maximum enrichment score across all tissues is reported in Table 1. Control gene sets were also constructed for each motif, consisting of an equal-sized set of genes in which the motif occurs but is not conserved; these control sets show little or no enrichment across the same tissues (see Supplementary Fig. S2).

© 2005 Nature Publishing Group
NATURE | doi:10.1038/nature03441 | www.nature.com/nature

conserved motifs in 3′ UTRs (MCS > 6). Because 3′-UTR motifs have not been extensively studied, we could not compare the discovered motifs to a large collection of previously known motifs; however, the motifs show two unusual properties that give insight into their nature.

First, the 3′-UTR motifs show a strong directional bias with respect to DNA strand. Whereas promoter motifs tend to have similar conservation rates on the coding and non-coding strands, the 3′-UTR motifs are preferentially conserved in only one strand (Fig. 4a). The strand specificity is consistent with the 3′-UTR motifs acting at the level of RNA rather than DNA and thus having a role in post-transcriptional regulation.

Second, the 3′-UTR motifs show an unusual length distribution. They have a strong peak at an 8-base length, whereas no such bias was found for promoter motifs (Fig. 4b). Moreover, the motifs of length 8 have a strong tendency to end with the nucleotide 'A' (Fig. 4c). These properties are reminiscent of a feature of miRNA[16]: many mature miRNAs start with a 'U' base followed by a 7-base 'seed' complementary to a site in the 3′ UTR of target mRNAs[17,26,27]; binding at such sites can guide degradation or repress translation of the mRNA. We reasoned that many of the highly conserved 8-mer motifs discovered by our unbiased procedure might be binding sites for conserved miRNAs.

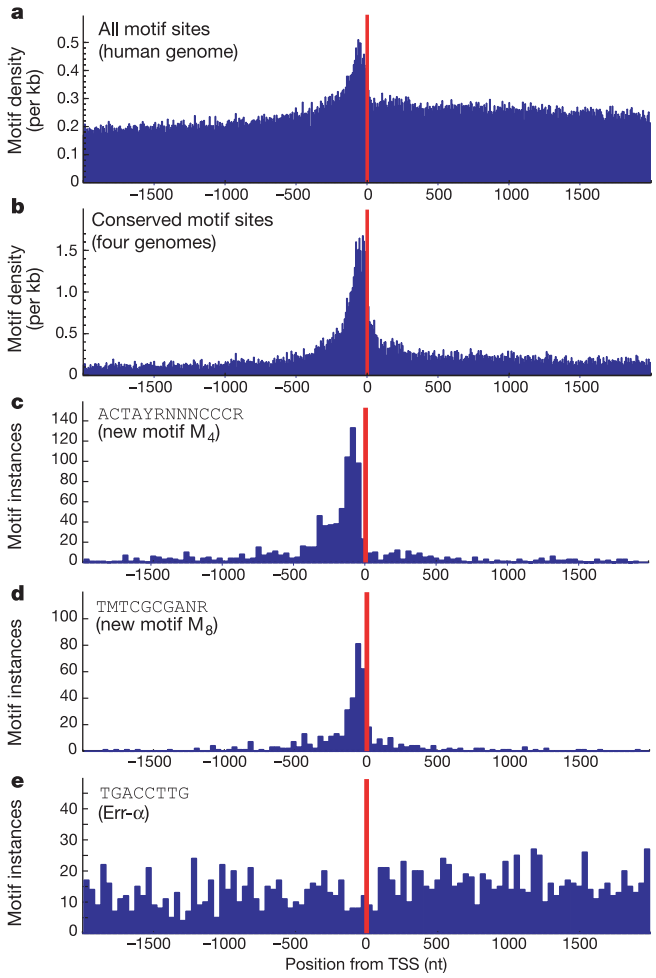


Figure 3 Discovered promoter motifs show positional bias with respect to transcriptional start site (TSS). **a**, Distribution of distance from TSS for all occurrences in human genome peaks within 100 bp before TSS. **b**, Distribution for conserved occurrences shows an even stronger peak. **c**, **d**, New motifs M_4 and M_8 peak at −81 and −69 respectively. **e**, Some motifs do not show specific peaks, including the known Err-α motif. nt, nucleotides.

Relationship with miRNAs

To investigate the relationship with miRNAs, the motif discovery procedure was repeated using only contiguous, non-degenerate 8-mers. We identified the subset of 8-mers with conservation rate > 18% (compared with the rate for a random 8-mer of 7.6%) and clustered these 8-mers into motifs defined as sets of similar 8-mers (see Supplementary Information). We refer to these as 'highly conserved 8-mer motifs'. We obtained 72 highly conserved 8-mer motifs, which match 46% of the full set of 3′-UTR motifs (Fig. 4b).

We then searched for complementary matches of the 8-mer motifs to the 207 distinct human miRNAs (encoded by 222 miRNA genes) listed in the current registry[28]. Roughly 43.5% of the known miRNAs can match through Watson–Crick pairing to the highly conserved 8-mer motifs (versus 2% for an equal number of random 8-mers). When we relax the Watson–Crick base pairing to allow one mismatch, an additional 27 miRNAs can be identified (including four miRNAs containing T–G pairing and ten miRNAs whose first 5′ base was paired unselectively to 'A' in the conserved 8-mers). Moreover, the matches begin at nucleotide 1 or 2 of the miRNA gene in more than 95% of cases (Fig. 4d). These results strongly suggest that these 8-mer motifs represent target sites for miRNAs.

The remainder of the known miRNA genes do not contain complementary matches to any of the highly conserved 8-mer motifs. These miRNAs may bind relatively few targets, form only partial complements with their targets, tend to bind regions outside the 3′ UTR, or have been diverged between the species we are comparing. It is worth noting that, among the known miRNAs, those that do not match highly conserved 8-mers are evolving much more rapidly; they show a rate of disruptive mutation that is fivefold higher than for known miRNAs that match the highly conserved 8-mers. This suggests that the miRNA genes that match the highly conserved 8-mers have many more targets and, as a result, are much more constrained in their evolution.

New miRNA genes

We then sought to use the 8-mer motifs to discover new miRNA genes. Towards this end, we searched the four aligned genomes for conserved sequences complementary to any of the 72 discovered 8-mer motifs, extracted the sequence flanking each conserved site, and used the published RNAfold[29,30] program to test for a conserved RNA-folding pattern characteristic of miRNA genes (stem-loop structures with calculated folding free energy of at least 25 kcal mol^{-1} in all four species).

We identified 242 conserved and stable stem-loop sequences, containing conserved instances of most (52 of 72) of the highly conserved 8-mer motifs (see Supplementary Information). These include 113 sequences encoding known miRNAs genes (52% of 222 known miRNA genes) and 129 sequences encoding predicted new miRNAs (Table 2 and Fig. 4e; see also Supplementary Table S8). The predicted miRNA genes show relatively few mutations that would disrupt an inferred base pair in the stem structure; the frequency of such mutations is similar to that seen for the 222 known miRNA genes.

A representative set of 12 predicted new miRNA genes was selected for validation according to accepted criteria[16,31] (involving purification of small RNAs, ligation of adaptors, polymerase chain reaction (PCR) amplification, cloning and DNA sequencing to verify the precise sequence and junction; see Methods). Lacking prior information about tissue or temporal specificity of expression, we used pooled small RNAs prepared from ten human tissues (breast, pancreas, prostate, colon, stomach, uterus, lung, brain, liver and kidney) (see Methods). This may miss many miRNAs expressed primarily during development or at low levels in the adult, or not expressed in the tissues we tested. Nonetheless, 6 of the

© 2005 Nature Publishing Group

12 (50%) predicted miRNAs genes were found to be clearly expressed in the pooled adult tissues.

We conclude that many of the 129 candidates are likely to be bona fide miRNA genes. Moreover, there are likely to be many additional miRNAs inasmuch as our approach was only designed to find those containing a conserved 8-mer and thus only detects one-half of known miRNAs. This challenges a recent conclusion, based on computational analysis, that nearly all human miRNA genes have already been discovered, and that fewer than 40 remain to be found[27].

Targets of miRNAs

The properties of the conserved motifs also allow inferences about the prevalence of regulation by miRNAs. Roughly 40% of human 3′

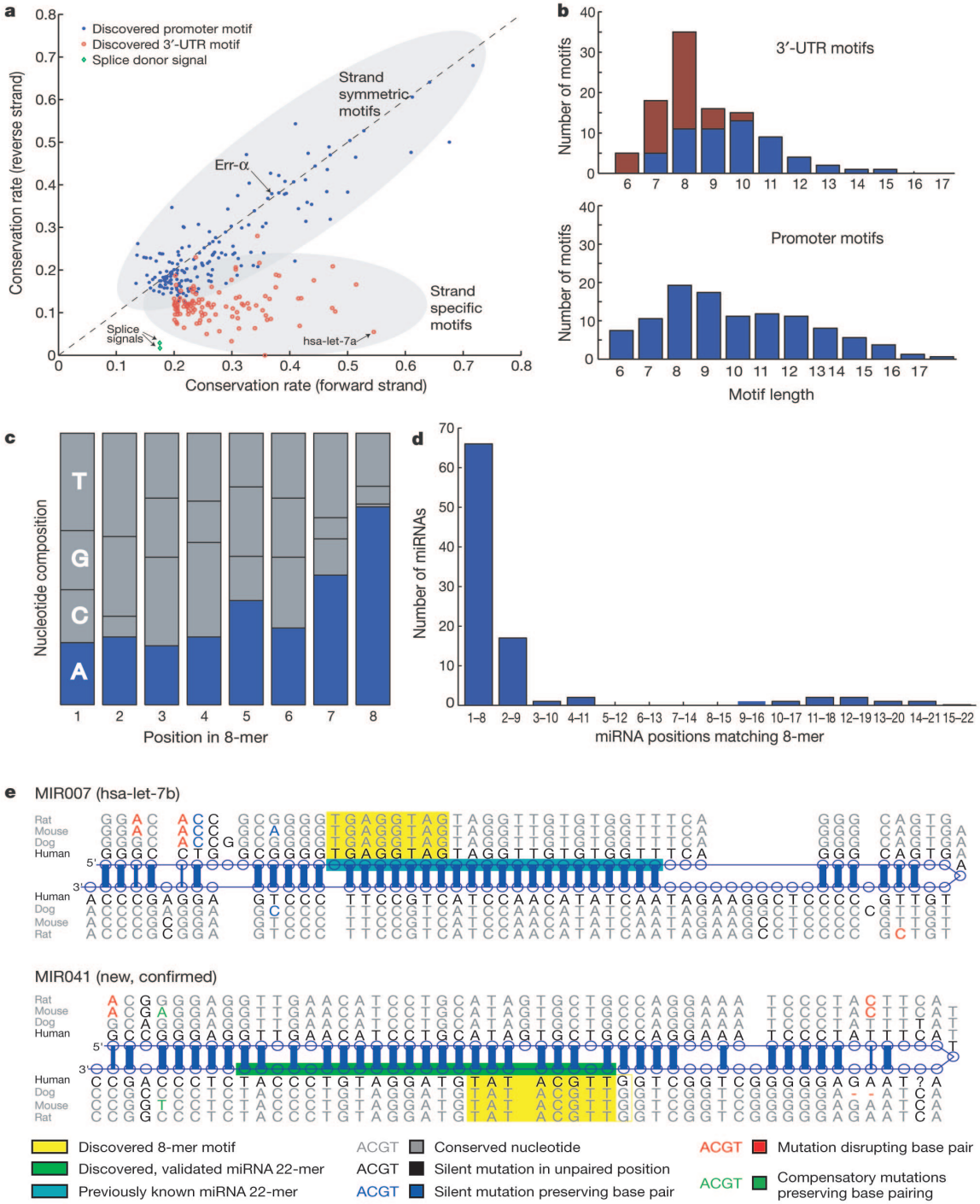


Figure 4 Properties of discovered 3′-UTR motifs and corresponding miRNA genes. **a**, Directionality of 3′-UTR motifs revealed by comparing conservation on forward and reverse strands. Strand preference is also seen for splice signals, but conservation of promoter motifs is largely symmetric. hsa-let-7a, *Homo sapiens* let-7a. **b**, Length distribution of 3′-UTR motifs shows abundance of motifs of length 8, but no such preference is seen for promoter motifs. Motifs overlapping conserved 8-mers (red) account for the length bias. **c**, A total of 72% of 8-mer motifs end with nucleotide A, suggesting complementarity with mature miRNA genes, frequently starting with U. **d**, Ninety-five per cent of discovered 8-mers match known miRNA genes at position 1–8 or 2–9. **e**, Alignments of known and new miRNA stem-loop structures, identified by their complementarity to a discovered 8-mer motif. New ~22-bp mature miRNA products predicted based on the observation shown in **d** and validated experimentally.

© 2005 Nature Publishing Group

UTRs contain a conserved occurrence of one of the miRNA-associated 8-mer motifs, whereas only ~25% contain conserved occurrences of a comparable control set. This suggests that at least 20% of 3′ UTRs may be targets for conserved miRNA-based regulation at the 8-mer motifs (see Methods). This greatly expands previous estimates of the number of targets of miRNA genes[17]. With sequence from more mammalian genomes, it should be possible to distinguish the conserved target sites with high specificity and sensitivity[10]. There are likely to be additional miRNA target sites in the human genome that are not conserved across all mammals; it should be possible to find most of these by genomic comparison with closer relatives such as primates.

The results thus provide an unbiased assessment of the relative importance of miRNA-based regulation in the human genome, consistent with recent independent studies[32–34]. Overall, ~45% of the 108 highly conserved motifs in 3′ UTRs appear to be related to miRNA regulation and ~5,000 human genes (~20% of the genome) are likely to be regulated by miRNAs through these conserved motifs.

Table 2 **Top 50 conserved 8-mers in 3′ UTRs and corresponding miRNAs**

No.	Motif	Conservation rate	MiRNA*
1	GTGCAATA	0.55	miR-92, miR-32, miR-137, miR-367, miR-25, miR-217(†), new(12)
2	GTGCCTTA	0.54	miR-124a, miR-224(†), miR-208(†), miR-34b(†), miR-9*(†), miR-34c(†), miR-330(†), new(6)
3	CTACCTCA	0.53	miR-98, let-7i, let-7 g, let-7f, let-7e, let-7c, let-7b, let-7a, let-7d, miR-196b, miR-196a, new(4)
4	ACCAAAGA	0.49	miR-9, new(11)
5	TGTTTACA	0.48	miR-30e-5p, miR-30d, miR-30c, miR-30b, miR-30a-5p, new(4)
6	GCACTTTA	0.48	miR-20, miR-106b, miR-18(†), miR-93, miR-372, miR-17-5p, miR-106a, miR-302d, miR-302c, miR-302b, miR-302a, miR-373, new(4)
7	TGGTGCTA	0.43	miR-29c, miR-29b, miR-29a, miR-107(†), miR-103(†), new(6)
8	CTATGCAA	0.42	miR-153, new(9)
9	TACTTGAA	0.42	miR-26b, miR-26a, new(4)
10	CGCAAAAA	0.42	New(2)
11	GTGCCAAA	0.41	miR-96, miR-182, miR-183, new(16)
12	GTACTGTA	0.40	miR-101, miR-199a‡, miR-144, new(2)
13	ATACGGGT	0.40	miR-99a, miR-100, miR-99b(†)
14	AAGCACAA	0.40	miR-218, new(8)
15	TTTGCACT	0.37	miR-19b, miR-19a, miR-301, miR-130b, miR-130a, miR-152, miR-148b, miR-148a, miR-139, new(10)
16	TGTACATA	0.36	–
17	AAGCCATA	0.35	miR-135b, miR-135a
18	ACTGTGAA	0.35	miR-27b, miR-27a, miR-128b, miR-128a, miR-23b(†), miR-23a(†), new(5)
19	AGACAATC	0.33	miR-219, new(2)
20	TGCTGCTA	0.33	miR-195, miR-16, miR-15b, miR-15a, miR-338(†), miR-424, new(5)
21	TTTTGTAC	0.32	New(1)
22	ACATTCCA	0.32	miR-206, miR-1, miR-122a(†)
23	TGAATGTA	0.31	miR-181b(†), miR-181c, miR-181a
24	ACGGTACA	0.30	–
25	CAGTATTA	0.30	miR-200c, miR-200b, new(1)
26	TTGCATGT	0.29	New(7)
27	TCGCATGA	0.29	New(1)
28	CTCAGGGA	0.29	miR-125b, miR-125a, new(6)
29	CAAGTGCC	0.28	New(2)
30	ACTACTGA	0.28	–
31	TGGACCAA	0.28	miR-133b, miR-133a, new(3)
32	GTAAATAG	0.28	New(1)
33	TGTAGATA	0.28	–
34	ACACTACA	0.27	miR-142-3p, new(3)
35	GTACAGTT	0.26	New(1)
36	CACCAGCA	0.26	miR-138(†), new(4)
37	GGTACGAA	0.25	miR-126(†)
38	TGTATAGT	0.24	miR-381
39	AAGGGCTA	0.24	New(1)
40	AGCTTTAA	0.24	New(1)
41	ATTTATCG	0.23	–
42	GGCAGCTA	0.23	miR-22(†), new(1)
43	GCTGTAAA	0.23	New(4)
44	GCACTAAT	0.22	–
45	AAAGGTGC	0.22	–
46	ATGTAGCA	0.22	miR-221(†), miR-222(†)
47	ACACTGGA	0.21	miR-199b(†), miR-199a(†), miR-145(†), new(2)
48	GTATATAG	0.21	–
49	TTTGATAA	0.21	miR-361(†)
50	AAGCATGC	0.21	New(1)

Top 50 of 72 highly conserved 8-mer motifs in 3′ UTRs and their corresponding miRNAs.
*Known human miRNAs matching complements of each 8-mer and its variants grouped in the same cluster. The dagger symbol (in parentheses) indicates miRNAs with one mismatch. When multiple new miRNAs are discovered using the 8-mer motifs as seeds, their number is indicated in parentheses. (See Supplementary Information for full alignments of known and discovered miRNAs.)
‡The miRNA mature product comes from the 3′ arm of the stem loop.

Remaining motifs in 3′ UTRs

When the miRNA-related motifs are removed, there remains a list of 60 3′-UTR motifs with a length distribution similar to that seen for promoters (Fig. 4b). Although there is no systematic procedure for evaluating these motifs, the list contains several interesting features. It includes the known polyadenylation signal and various (A + T)-rich elements[35], which are believed to be involved in controlling mRNA stability and degradation. It also contains 11 motifs with a TGTA core sequence, flanked by different variants of an (A + T)-rich element (the most conserved being TGTANATA). Recent work in yeast has identified such motifs as binding sites for the Puf family of RNA-binding proteins; they may represent binding sites for the homologous mammalian proteins (such as PUM1 and PUM2 in human[36]).

Conclusion

Our comparative genomic analysis of four mammalian species has provided an initial systematic catalogue of human regulatory motifs in promoters and 3′ UTRs. In promoters, the approach automatically rediscovered many known motifs and discovered many new ones that appear to be functional based on multiple criteria; many of these show tissue-specific expression and distance constraints with respect to the TSS. In 3′ UTRs, the approach provided a first view of common regulatory motifs. Notably, roughly one-half of the discovered motifs in 3′ UTRs appear to be related to miRNAs, showing the extraordinary importance of this recently discovered regulatory mechanism. Moreover, these motifs also led to the identification of numerous new miRNA genes. The remaining motifs may be targets of RNA-binding proteins, with a few matching known motifs. The next challenge will be to develop systematic methods to discern the specific functions of these motifs in a genome-wide fashion.

The results here are, of course, only a step towards a comprehensive inventory of human regulatory motifs to serve as a foundation for understanding cellular circuitry and its role in health and disease. Our analysis used stringent thresholds to focus on the most abundant and most conserved motifs. With sequence from a few dozen additional mammals[37], it should be possible to create a complete dictionary of such common functional elements. □

Methods

Motif conservation score

Regulatory motifs are represented as consensus sequences (profiles), over an alphabet of 11 characters, consisting of A, C, G, T and N, and the six twofold degenerate characters. A conserved occurrence of a motif m matches the motif consensus in each of the four species. The observed conservation rate p of a motif is the proportion of human occurrences that are conserved across all four species. The expected conservation rate p_0 for a motif of given length and redundancy is computed as the average observed conservation rate of 1,000 random motifs of the same length and redundancy, obtained by sampling the human genome in 1,000 loci. The MCS of a motif m is evaluated by comparing its observed conservation rate p to the expected conservation rate p_0; it corresponds to the binomial probability of observing k conserved instances out of total n instances given probability p_0 of conservation for any one instance.

Motif discovery overview

The MCS is evaluated separately for each type of region (promoters, introns, 3′ UTR), thus matching the specific nucleotide composition of each type of region and eliminating biases from the scoring scheme. Additionally, for promoter motifs, the MCS is evaluated separately for sequences inside CpG islands and those outside CpG islands, thus accounting for their radically different nucleotide compositions. In each type of region

© 2005 Nature Publishing Group

motifs are enumerated by hashing and refinement of 6-mer seeds possibly with a central gap. All motifs above the cutoff of MCS > 6 are selected, corresponding to specificity values higher than 98%. The selected motifs are grouped into clusters using genome-wide co-occurrence and sequence similarity: we first merge motif pairs which overlap by more than 80% of their sites, keeping only the motif with the highest MCS score; we then cluster the remaining motifs based on their pairwise sequence similarity, defined as the Pearson correlation of their equivalent position weight matrices.

Motif gene-set enrichment score (MGES)

The enrichment of a motif *m* in a given tissue is evaluated as the enrichment of its gene set *S* in the ranked list *L* for that tissue. The non-randomness of the ranks of *S* in the list *L* is evaluated using the Mann–Whitney rank sum statistic; the MGES is reported as the standard deviations corresponding to this statistic. For each motif *m*, three gene sets are generated: a target gene set S_1, and two control gene sets, S_2 and S_3, with the same number of genes. The target gene set S_1 of 'conserved instances', consists of all genes whose promoters contain at least one conserved instance of the motif *m*. The control gene set S_2 consists of genes with 'non-conserved instances' randomly sampled from genes containing instances of the motif in the human genome, regardless of conservation. The control gene set S_3 consists of genes with 'shuffled conserved instances', randomly sampled from the union of all conserved gene sets S_1, across all motifs.

Identification of new miRNAs

Conserved occurrences of the 3′ UTR 8-mer motifs are identified in the entire human genome by searching both strands for motifs reverse complementary to each 8-mer, and excluding previously annotated functional elements. The neighbourhoods of these alignments are then searched for stable stem-loops over a sliding window of 110 bp every 3 bp, and those with a folding free energy of at least 25 kcal mol^{-1} in each aligned species are selected using the program RNAfold. These were further evaluated based on several observed features of known miRNAs: higher conservation in the ~22 bp stem, lower conservation in the loop and surrounding regions, and appropriate base-pairing and bulges in the stem region.

Experimental validation of new miRNA genes

A set of 12 new miRNA genes was selected for experimental validation, such that their conservation properties and folding free energy values is representative of the set of all 258 predicted miRNAs. The experimental procedure was carefully designed to ensure that the exact ~22 bp predicted product is specifically expressed. First, three steps of rigorous PAGE purification of small RNA fractions exclude large RNA and genomic DNA. Second, PCR is done with one primer specific to the gene, and one complementary to the artificial adaptors used in miRNA ligation. Finally, the resulting clones are sequenced and the precise expected sequence and junction are verified.

Additional methods

Supplementary Information includes a complete description of the methods for whole-genome alignment, definition of promoter and 3′ UTR databases, rapid motif enumeration and scoring, motif clustering, comparison with TRANSFAC motifs, motif gene-set enrichment analysis, motif positional bias, 3′-UTR 8-mer discovery, identification of new miRNA genes, estimating number of miRNA targets, and experimental verification of miRNA genes.

Received 23 November 2004; accepted 9 February 2005; doi:10.1038/nature03441.
Published online 27 February 2005.

1. Gumucio, D. L. *et al.* Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes. *Mol. Cell. Biol.* **12,** 4919–4929 (1992).
2. Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W. & Lawrence, C. E. Human-mouse genome comparisons to locate regulatory sites. *Nature Genet.* **26,** 225–228 (2000).
3. Dubchak, I. *et al.* Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res.* **10,** 1304–1306 (2000).
4. Pennacchio, L. A. *et al.* An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science* **294,** 169–173 (2001).
5. Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299,** 1391–1394 (2003).
6. Sandelin, A., Wasserman, W. W. & Lenhard, B. ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res.* **32,** W249–W252 (2004).
7. Sinha, S., Blanchette, M. & Tompa, M. PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics* **5,** 170 (2004).
8. Dermitzakis, E. T. *et al.* Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res.* **14,** 852–859 (2004).
9. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304,** 1321–1325 (2004).
10. Eddy, S. R. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.* **3,** e10 (2005).
11. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423,** 241–254 (2003).
12. Cliften, P. *et al.* Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301,** 71–76 (2003).
13. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
14. International Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
15. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–521 (2004).
16. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116,** 281–297 (2004).
17. Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. Prediction of mammalian microRNA targets. *Cell* **115,** 787–798 (2003).
18. Maglott, D. R., Katz, K. S., Sicotte, H. & Pruitt, K. D. NCBI's LocusLink and RefSeq. *Nucleic Acids Res.* **28,** 126–128 (2000).
19. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14,** 708–715 (2004).
20. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13,** 103–107 (2003).
21. Mootha, V. K. *et al.* Errα and Gabpa/b specify PGC-1α-dependent oxidative phosphorylation gene expression that is altered in diabetic muscle. *Proc. Natl Acad. Sci. USA* **101,** 6570–6575 (2004).
22. Johnston, S. D. *et al.* Estrogen-related receptor alpha 1 functionally binds as a monomer to extended half-site sequences including ones contained within estrogen-response elements. *Mol. Endocrinol.* **11,** 342–352 (1997).
23. Matys, V. *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31,** 374–378 (2003).
24. Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA* **101,** 6062–6067 (2004).
25. Kuersten, S. & Goodwin, E. B. The power of the 3′ UTR: translational control and development. *Nature Rev. Genet.* **4,** 626–637 (2003).
26. Lai, E. C. Micro RNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation. *Nature Genet.* **30,** 363–364 (2002).
27. Lim, L. P., Glasner, M. E., Yekta, S., Burge, C. B. & Bartel, D. P. Vertebrate microRNA genes. *Science* **299,** 1540 (2003).
28. Griffiths-Jones, S. The microRNA Registry. *Nucleic Acids Res.* **32** (Database issue), D109–D111 (2004).
29. Fontana, W. *et al.* RNA folding and combinatory landscapes. *Phys. Rev. E* **47,** 2083–2099 (1993).
30. Hofacker, I. L. Vienna RNA secondary structure server. *Nucleic Acids Res.* **31,** 3429–3431 (2003).
31. Ambros, V. *et al.* A uniform system for microRNA annotation. *RNA* **9,** 277–279 (2003).
32. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120,** 15–20 (2005).
33. Lim, L. P. *et al.* Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* advance online publication, 30 January (2005) (doi:10.1038/nature03315).
34. Berezikov, E. *et al.* Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120,** 21–24 (2005).
35. Chen, C. Y. & Shyu, A. B. AU-rich elements: characterization and importance in mRNA degradation. *Trends Biochem. Sci.* **20,** 465–470 (1995).
36. Spassov, D. S. & Jurecic, R. The PUF family of RNA-binding proteins: does evolutionarily conserved structure equal conserved function? *IUBMB Life* **55,** 359–366 (2003).
37. Margulies, E. H. *et al.* An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl Acad. Sci. USA* (in the press).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We thank B. Birren, M. Kamal, K. O'Neill and A. Subramanian for advice and discussions. This work was supported in part by grants from the National Human Genome Research Institute.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.K. (manoli@mit.edu) or E.S.L. (lander@broad.mit.edu).

© 2005 Nature Publishing Group

ARTICLES

Genome sequence, comparative analysis and haplotype structure of the domestic dog

Kerstin Lindblad-Toh[1], Claire M Wade[1,2], Tarjei S. Mikkelsen[1,3], Elinor K. Karlsson[1,4], David B. Jaffe[1], Michael Kamal[1], Michele Clamp[1], Jean L. Chang[1], Edward J. Kulbokas III[1], Michael C. Zody[1], Evan Mauceli[1], Xiaohui Xie[1], Matthew Breen[5], Robert K. Wayne[6], Elaine A. Ostrander[7], Chris P. Ponting[8], Francis Galibert[9], Douglas R. Smith[10], Pieter J. deJong[11], Ewen Kirkness[12], Pablo Alvarez[1], Tara Biagi[1], William Brockman[1], Jonathan Butler[1], Chee-Wye Chin[1], April Cook[1], James Cuff[1], Mark J. Daly[1,2], David DeCaprio[1], Sante Gnerre[1], Manfred Grabherr[1], Manolis Kellis[1,13], Michael Kleber[1], Carolyne Bardeleben[6], Leo Goodstadt[8], Andreas Heger[8], Christophe Hitte[9], Lisa Kim[7], Klaus-Peter Koepfli[6], Heidi G. Parker[7], John P. Pollinger[6], Stephen M. J. Searle[14], Nathan B. Sutter[7], Rachael Thomas[5], Caleb Webber[8], Broad Institute Genome Sequencing Platform* & Eric S. Lander[1,15]

Here we report a high-quality draft genome sequence of the domestic dog (*Canis familiaris*), together with a dense map of single nucleotide polymorphisms (SNPs) across breeds. The dog is of particular interest because it provides important evolutionary information and because existing breeds show great phenotypic diversity for morphological, physiological and behavioural traits. We use sequence comparison with the primate and rodent lineages to shed light on the structure and evolution of genomes and genes. Notably, the majority of the most highly conserved non-coding sequences in mammalian genomes are clustered near a small subset of genes with important roles in development. Analysis of SNPs reveals long-range haplotypes across the entire dog genome, and defines the nature of genetic diversity within and across breeds. The current SNP map now makes it possible for genome-wide association studies to identify genes responsible for diseases and traits, with important consequences for human and companion animal health.

Man's best friend, *Canis familiaris*, occupies a special niche in genomics. The unique breeding history of the domestic dog provides an unparalleled opportunity to explore the genetic basis of disease susceptibility, morphological variation and behavioural traits. The position of the dog within the mammalian evolutionary tree also makes it an important guide for comparative analysis of the human genome.

The history of the domestic dog traces back at least 15,000 years, and possibly as far back as 100,000 years, to its original domestication from the grey wolf in East Asia[1–4]. Dogs evolved through a mutually beneficial relationship with humans, sharing living space and food sources. In recent centuries, humans have selectively bred dogs that excel at herding, hunting and obedience, and in this process have created breeds rich in behaviours that both mimic human behaviours and support our needs. Dogs have also been bred for desired physical characteristics such as size, skull shape, coat colour and texture[5], producing breeds with closely delineated morphologies. This evolutionary experiment has produced diverse domestic species, harbouring more morphological diversity than exists within the remainder of the family Canidae[6].

As a consequence of these stringent breeding programmes and periodic population bottlenecks (for example, during the World Wars), many of the ~400 modern dog breeds also show a high prevalence of specific diseases, including cancers, blindness, heart disease, cataracts, epilepsy, hip dysplasia and deafness[7,8]. Most of these diseases are also commonly seen in the human population, and clinical manifestations in the two species are often similar[9]. The high prevalence of specific diseases within certain breeds suggests that a limited number of loci underlie each disease, making their genetic dissection potentially more tractable in dogs than in humans[10].

Genetic analysis of traits in dogs is enhanced by the close relationship between humans and canines in modern society.

[1]Broad Institute of Harvard and MIT, 320 Charles Street, Cambridge, Massachusetts 02141, USA. [2]Center for Human Genetic Research, Massachusetts General Hospital, 185 Cambridge Street, Boston, Massachusetts 02114, USA. [3]Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [4]Program in Bioinformatics, Boston University, 44 Cummington Street, Boston, Massachusetts 02215, USA. [5]Department of Molecular Biomedical Sciences, College of Veterinary Medicine, North Carolina State University, 4700 Hillsborough Street, Raleigh, North Carolina 27606, USA. [6]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, California 90095, USA. [7]National Human Genome Research Institute, National Institutes of Health, 50 South Drive, MSC 8000, Building 50, Bethesda, Maryland 20892-8000, USA. [8]MRC Functional Genetics, University of Oxford, Department of Human Anatomy and Genetics, South Parks Road, Oxford OX1 3QX, UK. [9]UMR 6061 Genetique et Developpement, CNRS—Université de Rennes 1, Faculté de Médecine, 2, Avenue Leon Bernard, 35043 Rennes Cedex, France. [10]Agencourt Bioscience Corporation, 500 Cummings Center, Suite 2450, Beverly, Massachusetts 01915, USA. [11]Children's Hospital Oakland Research Institute, 5700 Martin Luther King Jr Way, Oakland, California 94609, USA. [12]The Institute for Genomic Research, Rockville, Maryland 20850, USA. [13]Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts 02139, USA. [14]The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [15]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA.
*A list of participants and affiliations appears at the end of the paper.

© 2005 Nature Publishing Group

Through the efforts of the American Kennel Club (AKC) and similar organizations worldwide, extensive genealogies are easily accessible for most purebred dogs. With the exception of human, dog is the most intensely studied animal in medical practice, with detailed family history and pathology data often available[8]. Using genetic resources developed over the past 15 years[11–16], researchers have already identified mutations in genes underlying ~25 mendelian diseases[17,18]. There are also growing efforts to understand the genetic basis of phenotypic variation such as skeletal morphology[10,19].

The dog is similarly important for the comparative analysis of mammalian genome biology and evolution. The four mammalian genomes that have been intensely analysed to date (human[20–22], chimpanzee[23], mouse[24] and rat[25]) represent only one clade (Euarchontoglires) out of the four clades of placental mammals. The dog represents the neighbouring clade, Laurasiatheria[26]. It thus serves as an outgroup to the Euarchontoglires and increases the total branch length of the current tree of fully sequenced mammalian genomes, thereby providing additional statistical power to search for conserved functional elements in the human genome[24,27–33]. It also helps us to draw inferences about the common ancestor of the two clades, called the boreoeutherian ancestor, and provides a bridge to the two remaining clades (Afrotheria and Xenarthra) that should be helpful for anchoring low-coverage genome sequence currently being produced from species such as elephant and armadillo[28].

Here we report a high-quality draft sequence of the dog genome covering ~99% of the euchromatic genome. The completeness, nucleotide accuracy, sequence continuity and long-range connectivity are extremely high, exceeding the values calculated for the recent draft sequence of the mouse genome[24] and reflecting improved algorithms, higher-quality data, deeper coverage and intrinsic genome properties. We have also created a tool for the formal assessment of assembly accuracy, and estimate that >99% of the draft sequence is correctly assembled.

We also report an initial compendium of SNPs for the dog population, containing >2.5 million SNPs derived primarily from partial sequence comparison of 11 dog breeds to a reference sequence. We characterized the polymorphism rate of the SNPs across breeds and the long-range linkage disequilibrium (LD) of the SNPs within and across breeds.

We have analysed these data to study genome structure, gene evolution, haplotype structure and phylogenetics of the dog. Our key findings include:

- The evolutionary forces molding the mammalian genome differ among lineages, with the average transposon insertion rate being lowest in dog, the deletion rate being highest in mouse and the nucleotide substitution rate being lowest in human.
- Comparison between human and dog shows that ~5.3% of the human genome contains functional elements that have been under purifying selection in both lineages. Nearly all of these elements are confined to regions that have been retained in mouse, indicating that they represent a common set of functional elements across mammals.
- Fifty per cent of the most highly conserved non-coding sequence in the genome shows striking clustering in ~200 gene-poor regions, most of which contain genes with key roles in establishing or maintaining cellular identity, such as transcription factors or axon guidance receptors.
- Sets of functionally related genes show highly similar patterns of evolution in the human and dog lineages. This suggests that we should be careful about interpreting accelerated evolution in human relative to mouse as representing human-specific innovations (for example, in genes involved in brain development), because comparable acceleration is often seen in the dog lineage.
- Analysis across the entire genome of the sequenced boxer and across 6% of the genome in ten additional breeds shows that linkage disequilibrium (LD) within breeds extends over distances of several megabases, but LD across breeds only extends over tens of kilobases. These LD patterns reflect two principal bottlenecks in dog history: early domestication and recent breed creation.
- Haplotypes within breeds extend over long distances, with ~3–5 alleles at each locus. Portions of these haplotypes, as large as 100 kilobases (kb), are shared across multiple breeds, although they are present at widely varying frequencies. The haplotype structure suggests that genetic risk factors may be shared across breeds.
- The current SNP map has sufficient density and an adequate within-breed polymorphism rate (~1/900 base pairs (bp) between breeds and ~1/1,500 bp within breeds) to enable systematic association studies to map genes affecting traits of interest. Genotyping of ~10,000 SNPs should suffice for most purposes.
- The genome sequence can be used to select a small collection of rapidly evolving sequences, which allows nearly complete resolution of the evolutionary tree of nearly all living species of Canidae.

Generating a draft genome sequence

We sequenced the genome of a female boxer using the whole-genome shotgun (WGS) approach[22,24] (see Methods and Supplementary Table S1). A total of 31.5 million sequence reads, providing ~7.5-fold sequence redundancy, were assembled with an improved version of the ARACHNE program[34], resulting in an initial assembly (CanFam1.0) used for much of the analysis below, and an updated assembly (CanFam2.0) containing minor improvements (Table 1 and Supplementary Table S2).

Genome assembly. The recent genome assembly spans a total distance of 2.41 Gb, consisting of 2.38 Gb of nucleotide sequence with the remaining 1% in captured gaps. The assembly has extremely high continuity. The N50 contig size is 180 kb (that is, half of all bases reside in a contiguous sequence of 180 kb or more) and the N50 supercontig size is 45.0 Mb (Table 1). In particular, this means that most genes should contain no sequence gaps and that most canine chromosomes (mean size 61 Mb) have nearly all of their sequence ordered and oriented within one or two supercontigs (Supplementary Table S2). Notably, the sequence contigs are ~50-fold larger than the earlier survey sequence of the standard poodle[16].

The assembly was anchored to the canine chromosomes using data from both radiation hybrid and cytogenetic maps[11,13,14]. Roughly 97% of the assembled sequence was ordered and oriented on the chromosomes, showing an excellent agreement with the two maps. There were only three discrepancies, which were resolved by obtaining additional fluorescence *in situ* hybridization (FISH) data from the sequenced boxer. The 3% of the assembly that could not be anchored consists largely of highly repetitive sequence, including eight supercontigs of 0.5–1.0 Mb composed almost entirely of satellite sequence.

The nucleotide accuracy and genome coverage of the assembly is high (Supplementary Table S3). Of the bases in the assembly, 98% have quality scores exceeding 40, corresponding to an error rate of less than 10^{-4} and comparable to the standard for the finished human sequence[35]. When we directly compared the assembly to 760 kb of finished sequence (in regions where the boxer is

Table 1 | Assembly statistics for CanFam1.0 and 2.0

	CanFam1.0	CanFam2.0
N50 contig size	123 kb	180 kb
N50 supercontig size	41.2 Mb	45.0 Mb
Assembly size (total bases)	2.360 Gb	2.385 Gb
Number of anchored supercontigs	86	87
Percentage of genome in anchored supercontigs	96	97
Sequence in anchored bases	2.290 Gb	2.309 Gb
Percentage of assembly in gaps	0.9	0.8
Estimated genome size*	2.411 Gb	2.445 Gb
Percentage of assembly in 'certified regions', without assembly inconsistency	99.3	99.6

*Includes anchored bases, spanned gaps (21 Mb in CanFam1.0, 18 Mb in CanFam2.0) and centromeric sequence (3 Mb for each chromosome).

© 2005 Nature Publishing Group

homozygous, to eliminate differences attributable to polymorphisms; see below), we found that the draft genome sequence covers 99.8% of the finished sequence and that bases with quality scores exceeding 40 have an empirical error rate of 2×10^{-5} (Supplementary Table S3).

Explaining the high sequence continuity. The dog genome assembly has superior sequence continuity (180 kb) than the WGS assembly of the mouse genome (25 kb) obtained several years ago[24]. At least three factors contribute to the higher connectivity of the dog assembly (see Supplementary Information). First, we used a new version of ARACHNE with improved algorithms. Assembling the dog genome with the previous software version decreased N50 contig size from 180 kb to 61 kb, and assembling the mouse genome with the new version increased N50 contig size from 25 kb to 35 kb. Second, the amount of recently duplicated sequence is roughly twofold lower in dog than mouse (Supplementary Table S4); this improves contiguity because sequence gaps in both organisms tend to occur in recently duplicated sequence. Third, the dog sequence data has both higher redundancy (7.5-fold versus 6.5-fold) and higher quality (in terms of read length, pairing rate and tight distribution of insert sizes) compared with mouse. The contig size for the dog genome drops by about 32% when the data redundancy is decreased from 7.5-fold to 6.5-fold. A countervailing influence is that the dog genome contains polymorphism, whereas the laboratory mouse is completely inbred.

Assembly certification. Although 'quality scores' have been developed to indicate the nucleotide accuracy of a draft genome sequence[36], no analogous measures have been developed to reflect the long-range assembly accuracy. We therefore sought to develop such a measure on the basis of two types of internal inconsistencies (see Supplementary Information). The first is haplotype inconsistency, involving clear evidence of three or more distinct haplotypes within an assembled region from a single diploid individual. The second is linkage inconsistency, involving a cluster of reads for which the placement of the paired-end reads is illogical. This includes cases in which: (1) one end cannot be mapped to the region, (2) the linkage relationships are inconsistent with the sequence within contigs, or (3) distance constraints imply overlap between non-overlapping sequence contigs. The linkage inconsistency tests are most powerful when read pairs are derived from clone libraries with tight constraints on insert size. A region of assembly is defined as 'certified' if it is free of inconsistencies, and is otherwise 'questionable'.

Approximately 99.6% of the assembly resides in certified regions, with the N50 size of certified regions being ~12 Mb or about one-fifth of a chromosome. The remaining questionable regions are typically small (most are less than 40 kb), although there are a handful of regions of several hundred kilobases (Supplementary Fig. S1 and Supplementary Tables S5, S6). The questionable regions typically contain many inconsistencies, probably reflecting mis-assembly or overcollapse owing to segmental duplication. Chromosomes 2, 11 and 16 have 1.0–2.0% of their sequence in questionable regions. The certified and questionable regions are annotated in the public release of the dog genome assembly. With the concept of assembly certification, the scientific community can have appropriate levels of confidence in the draft genome sequence.

Genome landscape and evolution

Our understanding of the evolutionary processes that shape mammalian genomes has greatly benefited from the comparative analysis of sequenced primate[21,23] and rodent[24,25] genomes. However, the rodent genome is highly derived relative to that of the common ancestor of the eutherian mammals. As the first extensive sequence from an outgroup to the clade that includes primates and rodents, the dog genome offers a fresh perspective on mammalian genome evolution. Accordingly, we examined the rates and correlations of large-scale rearrangement, transposon insertion, deletion and nucleotide divergence across three major mammalian orders (primates, rodents and carnivores).

Conserved synteny and large-scale rearrangements. We created multi-species synteny maps from anchors of unique, unambiguously aligned sequences (see Supplementary Information), showing regions of conserved synteny among dog, human, mouse and rat genomes. Approximately 94% of the dog genome lies in regions of conserved synteny with the three other species (Supplementary Figs S2–S4 and Supplementary Table S7).

Given a pair of genomes, we refer to a 'syntenic segment' as a region that runs continuously without alterations of order and orientation, and a 'syntenic block' as a region that is contiguous in two genomes but may have undergone internal rearrangements. Syntenic breakpoints between blocks reflect primarily interchromosomal exchanges, and breakpoints between syntenic segments reflect intrachromosomal rearrangements. In the analysis below, we focus on syntenic segments of at least 500 kb.

We identified a total of 391 syntenic breakpoints across dog, human, mouse and rat genomes (Fig. 1 and Supplementary Figs S2, S5). With data for multiple species, it is possible to assign events to specific lineages (Fig. 1 and Supplementary Table S8). We counted the total number of breakpoints along the human, dog, mouse and rat lineages, with the values for each rodent lineage reflecting all breakpoints since the common ancestor with human (Fig. 1). The total number of breakpoints in the human lineage is substantially smaller than in the dog, mouse or rat lineages (83 versus 100, 161 or 176, respectively). However, there are more intrachromosomal breakpoints in the human lineage than in dog (52 versus 33).

Although the overall level of genomic rearrangement has been much higher in rodent than in human, comparison with dog shows that there are regions where the opposite is true. In particular, of the many intrachromosomal rearrangements previously observed between human chromosome 17 and the orthologous mouse

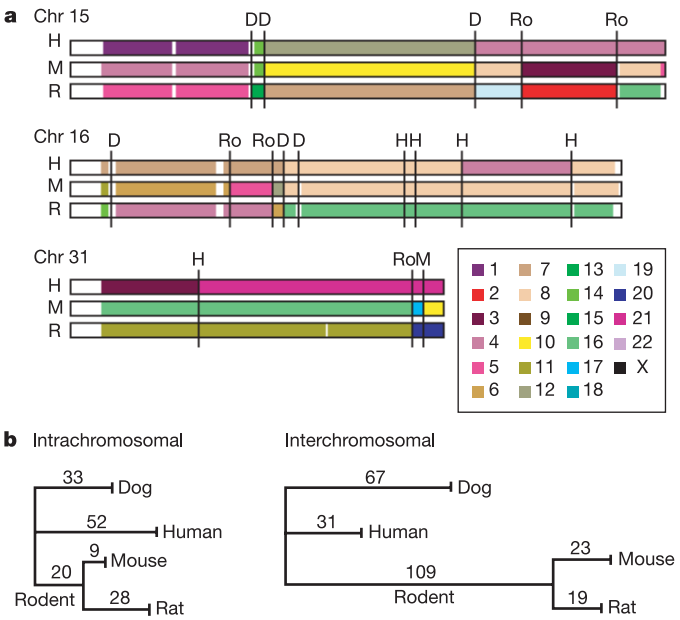


Figure 1 | Conserved synteny among the human, dog, mouse and rat genomes. a, Diagram of syntenic blocks (>500 kb) along dog chromosomes (Chr) 15, 16 and 31, with colours indicating the chromosome containing the syntenic region in other species. Synteny breakpoints were assigned to one of five lineages: dog (D), human (H), mouse (M), rat (R) or the common rodent ancestor (Ro). **b**, Lineage-specific intrachromosomal and interchromosomal breaks displayed on phylogenetic trees. Intrachromosomal breaks are seen more frequently in the human lineage than in mouse and rat, whereas interchromosomal breaks are somewhat more common in dog and considerably more common in rodents than in humans.

© 2005 Nature Publishing Group

sequence[24], most have occurred in the human lineage (see Supplementary Information). Human chromosome 17 is rich in segmental duplications and gene families[21], which may contribute to its genomic fragility[37,38].

Genomic insertion and deletion. The euchromatic genome of the dog is ~150 Mb smaller than in mouse, and ~500 Mb smaller than in human. The smaller total size is reflected at the local level, with 100-kb blocks of conserved synteny in dog corresponding to regions for which the median size is ~3% larger in mouse and ~15% larger in human.

To understand the balance of forces that determine genome size, we studied the alignments of the human, mouse and dog genomes (Fig. 2). In particular, we identified the lineage-specific interspersed repeats within each genome, which consist of particular families of short interspersed elements (SINEs), long interspersed elements (LINEs) and other transposable elements that are readily recognized by sequence analysis (Supplementary Tables S9, S10). The remaining sequence was annotated as 'ancestral', consisting of both ancestral unique sequence and ancestral repeat sequence; these two categories were combined because the power to recognize ancient transposon-derived sequences degrades with repeat age, particularly in the rapidly diverging mouse lineage[24].

This comparative analysis indicates that different forces account for the smaller genome sizes in dog and mouse relative to human. The smaller size of the dog genome is primarily due to the presence of substantially less lineage-specific repeat sequence in dog (334 Mb) than in human (609 Mb) or mouse (954 Mb). This reflects a lower activity of endogenous retroviral and DNA transposons (~26,000 extant copies in dog versus ~183,000 in human), as well as the fact that the SINE element in dog is smaller than in human (although of similar length to that in mouse). As a consequence, the total proportion of repetitive elements (both lineage-specific and ancestral) recognizable in the genome is lower for dog (34%) than for mouse (40%) or human (46%). In contrast, the smaller size of the mouse genome is primarily due to a higher deletion rate. Specifically, the amount of extant 'ancestral sequence' is much lower in mouse (1,474 Mb) than in human (2,216 Mb) or dog (1,997 Mb). Assuming an ancestral genome size of 2.8 Gb (ref. 24) and also that deletions occur continuously, we suggest that the rate of genomic deletion in the rodent lineage has been approximately 2.5-fold higher than in the dog and human lineages (see Supplementary Information). As a consequence, the human genome shares ~650 Mb more ancestral sequence with dog than with mouse, despite our more recent common ancestor with the latter.

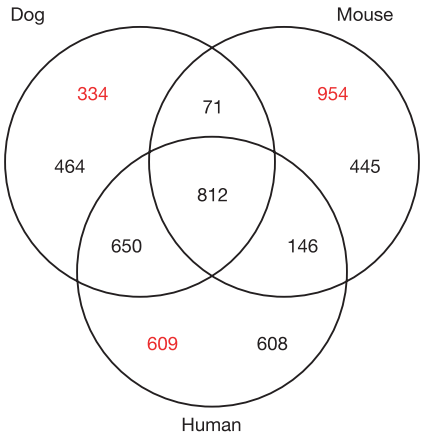


Figure 2 | Venn diagram showing the total lengths of aligned and unique sequences in the euchromatic portions of the dog, human and mouse genomes. Lengths shown in Mb, as inferred from genome-wide BLASTZ alignments (see Methods and Supplementary information). Overlapping partitions represent orthologous ancestral sequences. Each lineage-specific partition is further split into the total length of sequence classified as either lineage-specific interspersed repeats (red) or ancestral sequence (black). The latter is assumed to primarily represent ancestral sequences deleted in the two other species.

Active SINE family. Despite its relatively low proportion of transposable element-derived sequence, the dog genome contains a highly active carnivore-specific SINE family (defined as SINEC_Cf; RepBase release 7.11)[16]. The element is so active that many insertion sites are still segregating polymorphisms that have not yet reached fixation. Of ~87,000 young SINEC_Cf elements (defined by low divergence from the consensus sequence), nearly 8% are heterozygous within the draft genome sequence of the boxer. Moreover, comparison of the boxer and standard poodle genome sequences reveals more than 10,000 insertion sites that are bimorphic, with thousands more certain to be segregating in the dog population[16,39]. In contrast, the number of polymorphic SINE insertions in the human genome is estimated to be fewer than 1,000 (ref. 40).

The biological effect of these segregating SINE insertions is unknown. SINE insertions can be mutagenic through direct disruption of coding regions or through indirect effects on regulation and processing of messenger RNAs[39]. Such SINE insertions have already been shown to be responsible for two diseases in dog: narcolepsy and centronuclear myopathy[41,42]. It is conceivable that the genetic variation resulting from these segregating SINE elements has provided important raw material for the selective breeding programmes that have produced the wide phenotypic variations among modern dog breeds[16,43].

Sequence composition. The human and mouse genomes differ markedly in sequence composition, with the human genome having slightly lower average G+C content (41% versus 42% in mouse) but much greater variation across the genome. The dog genome closely resembles the human genome in its distribution of G+C content (Fig. 3a; Spearman's *rho* = 0.85 for dog–human and 0.76 for dog–mouse comparisons), even if we consider only nucleotides that can be aligned across all three species (Supplementary Fig. S6). The wider distribution of G+C content in human and dog is thus likely to reflect the boreoeutherian ancestor[44,45], with the more homogeneous composition in rodents having arisen primarily through lineage-specific changes in substitution patterns[46,47] rather than deletion of sequences with high G+C content.

Rate of nucleotide divergence. We estimated the mean nucleotide divergence rates in 1-Mb windows along the dog, human and mouse lineages on the basis of alignments of all ancestral repeats, using the consensus sequence for the repeats as a surrogate outgroup (Fig. 3b; see also Supplementary Information).

The dog lineage has diverged more rapidly than the human lineage (median relative divergence rate of 1.18, longer branch length in 95% of windows), but at only half the rate of the mouse lineage (median relative rate of 0.48, shorter branch length in 100% of windows). The absolute divergence rates are somewhat sensitive to the evolutionary model used and the filtering of alignment artefacts (data not shown), but the relative rates appear to be robust and are consistent with estimates from smaller sequence samples with multiple outgroups[28,48,49]. The lineage-specific divergence rates (human < dog < mouse) are probably explained by differences in metabolic rates[50,51] or generation times[52,53], but the relative contributions of these factors remain unclear[49].

Correlation in nucleotide divergence. As seen in other mammalian genomes[23–25], the average nucleotide divergence rate across 1-Mb windows varies significantly across the dog genome (coefficient of variation 0.11, compared with 0.024 expected under a uniform distribution). This regional variation shows significant correlation in orthologous windows across the dog, human and mouse genomes, but the strength of the correlation seems to decrease with total branch length (pair-wise correlation for orthologous 1-Mb windows: Spearman's *rho* = 0.49 for dog–human and 0.24 for dog–mouse comparisons). Lineage-specific variation in the regional divergence

© 2005 Nature Publishing Group

rates may be coupled with changes in factors such as sequence composition or chromosomal position[23,54]. Consistent with this, the ratios of lineage-specific divergence rates in orthologous windows are positively correlated with the ratios of current G+C content in the same windows (Spearman's *rho* = 0.16 for dog–human, 0.24 for dog–mouse).

Male mutation bias. Comparison of autosomal and X chromosome substitution rates can be used to estimate the relative mutation rates in the male and female germ lines (α), because the X chromosome is present in females twice as often as in males. Using the lineage-specific rates from ancestral repeats, we estimate α as 4.8 for the lineage leading to human, and 2.8 for the lineages leading to both mouse and dog. These values fall between recent estimates from murids[24,25] and from hominids[23], and suggest that male mutation bias may have increased in the lineage leading to humans.

Mutational hotspots and chromosomal fission. Genome comparisons of human with both chicken[55] and chimpanzee[23] have previously revealed that sequences close to a telomere tend to have increased divergence rates and G + C content relative to interstitial sequences. It has been unclear whether these increases are inherent characteristics of the subtelomeric sequence itself or derived characteristics causally connected with its chromosomal position. We find a similar increase in both divergence (median increase 15%, $P < 10^{-5}$; Mann-Whitney *U*-test) and G+C content (median increase 9%, $P < 10^{-9}$) for subtelomeric regions along the dog lineage, with a sharp increase towards the telomeres (Supplementary Fig. S7).

This phenomenon is manifested at other synteny breaks, not only those at telomeres. We also observed a significant increase in divergence and G+C content in interstitial regions that are sites of syntenic breakpoints[54,56] (Supplementary Fig. S7). These properties therefore seem correlated with the susceptibility of regions to chromosomal breakage.

Proportion of genome under purifying selection

One of the striking discoveries to emerge from the comparison of the human and mouse genomes[21,24] was the inference that ~5.2% of the human genome shows greater-than-expected evolutionary conservation (compared with the background rate seen in ancestral repeat elements, which are presumed to be nonfunctional). This proportion greatly exceeds the 1–2% that can be explained by protein-coding regions alone. The extent and function of the large fraction of non-coding conserved sequence remain unclear[57], but this sequence is likely to include regulatory elements, structural elements and RNA genes.

Low turnover of conserved elements. We repeated the analysis of conserved elements using the human and dog genomes. Briefly, the analysis involves calculating a conservation score S_{HD}, normalized by the regional divergence rate, for every 50-bp window in the human genome that can be aligned to dog. The distribution of conservation scores for all genomic sequences is compared to the distribution in ancestral repeat sequences (which are presumed to diverge at the local neutral rate), showing a clear excess of sequences with high conservation scores. By subtracting a scaled neutral distribution from the total distribution, one can estimate the distribution of conservation scores for sequences under purifying selection. Moreover, for a given sequence with conservation score S_{HD}, one can also assign a probability $P_{selection}(S_{HD})$ that the sequence is under purifying selection (see ref. 24 and Supplementary Information).

The human–dog genome comparison indicates that ~5.3% of the human genome is under purifying selection (Fig. 4a), which is equivalent to the proportion estimated from human–rodent analysis. The obvious question is whether the bases conserved between human and dog coincide with the bases conserved between humans and rodents[25,58]. Because the conservation scores do not unambiguously assign sequences as either selected or neutral (but instead only assign probability scores for selection), we cannot directly compare the conserved bases. We therefore devised the following alternative approach.

We repeated the human–dog analysis, dividing the 1462 Mb of orthologous sequence between human and dog into those regions with (812 Mb) or without (650 Mb) orthologous sequence in mouse (Fig. 2). The first set shows a clear excess of conservation relative to background, corresponding to ~5.2% of the human genome (Fig. 4b). In contrast, the second set shows little or no excess conservation, corresponding to at most 0.1% of the human genome (Fig. 4c). This implies that hardly any of the functional elements conserved between human and dog have been deleted in the mouse lineage (see also Supplementary Information).

The results strongly suggest that there is a common set of functional elements across all three mammalian species, corresponding to ~5% of the human genome (~150 Mb). These functional elements reside largely within the 812 Mb of ancestral sequence common to human, mouse and dog. If we eliminate ancestral repeat elements within this shared sequence as largely non-functional, most functional elements can be localized to 634 Mb, and constitute approximately 24% of this sequence.

It should be noted that the estimate of ~5% pertains to conserved elements across distantly related mammals. It is possible that there are additional weakly constrained or recently evolved elements within narrow clades (for example, primates) that can only be detected by genomic sequencing of more closely related species[29].

Clustering of highly conserved non-coding elements. We next

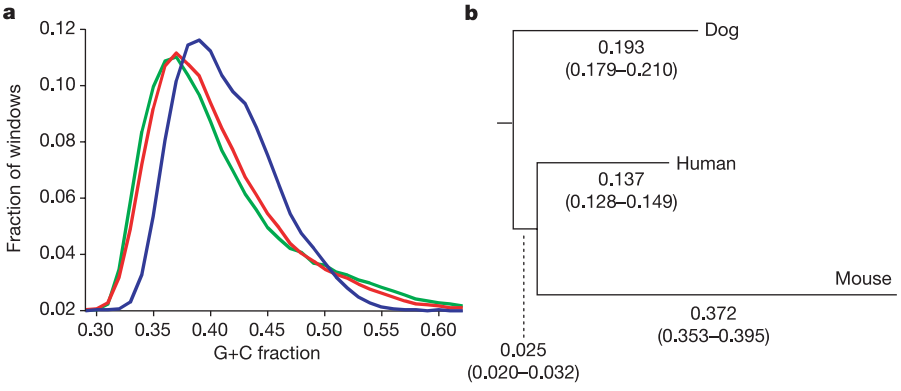


Figure 3 | Sequence composition and divergence rates. a, Distribution of G + C content in 10-kb windows across the genome in dog (green), human (red) and mouse (blue). **b**, Median lineage-specific substitution rates based on analysis of ancestral repeats aligning across all three genomes. Analysis was performed in non-overlapping 1-Mb windows across the dog genome that contained at least 2 kb of aligned ancestral repeat sequence (median 8.8 kb). The tree was rooted with the consensus sequences from the ancestral repeats. Numbers in parentheses give the 20–80th percentile range across the windows studied.

© 2005 Nature Publishing Group

explored the distribution of conserved non-coding elements (CNEs) across mammalian genomes. For this purpose, we calculated a conservation score S_{HMD} based on simultaneous conservation across all three species (see Methods). We defined highly conserved non-coding elements (HCNEs) to be 50-bp windows that do not overlap coding regions and for which $P_{selection}(S_{HMD})$, the probability of being under purifying selection given the conservation score, is at least 95%. We identified ~140,000 such windows (6.5 Mb total sequence), comprising ~0.2% of the human genome and representing the most conserved ~5% of all mammalian CNEs.

The density of HCNEs shows striking peaks when plotted in 1-Mb windows across the genome (Fig. 4d and Supplementary Figs S8 and S9), with 50% lying in 204 regions that span less than 14% of the human genome (Supplementary Table S11). These regions are generally gene-poor, together containing only ~6% of all protein-coding sequence.

The genes contained within these gene-poor regions are of particular interest. At least 182 of the 204 regions contain genes with key roles in establishing or maintaining cellular 'state'. At least 156 of the regions contain one or, in a few cases, several transcription factors involved in differentiation and development[59]. Another 26 regions contain a gene important for neuronal specialization and growth, including several axon guidance receptors. The proportion of developmental regulators is far greater than expected by chance ($P < 10^{-31}$; see Supplementary Information).

We then tested whether the HCNEs within these regions tend to cluster around the genes encoding regulators of development. Analysis of the density of HCNEs in the intronic and intergenic sequences flanking every gene in the 204 regions revealed that the 197 genes encoding developmental regulators show an average of ~10-fold enrichment for HCNEs relative to the full set of 1,285 genes in the regions (Fig. 4e and Supplementary Fig. S10). The enrichment sometimes extends into the immediately flanking genes.

We note that the 204 regions include nearly all of the recently identified clusters of conserved elements between distantly related vertebrates such as chicken and pufferfish[55,59–62]. For example, they overlap 56 of the 57 large intervals containing conserved non-coding sequence identified between human and chicken[55]. The mammalian analysis, however, detects vastly more CNEs (>100-fold more sequence than with pufferfish[59] and 2–3-fold more than with chicken) and identifies many more clusters. The limited sensitivity of these more distant vertebrate comparisons may reflect the difficulty of aligning short orthologous elements across such large evolutionary distances or the emergence of mammal-specific regulatory elements. In any case, mammalian comparative analysis may be a more powerful tool for elucidating the regulatory controls across these important regions.

Although the function of conserved non-coding elements is unknown, on the basis of recent studies[59,63–66] it seems likely that many regulate gene expression. If so, the above results suggest that ~50% of all mammalian HCNEs may be devoted to regulating ~1% of all genes. In fact, the distribution may be even more skewed, as there are additional genomic regions with only slightly lower HCNE density than the 204 studied above (Supplementary Fig. S8). All of these regions clearly merit intensive investigation to assess indicators of regulatory function. We speculate that these regions may harbour characteristic chromatin structure and modifications that are potentially involved in the establishment or maintenance of cellular state.

Genes

Accurate identification of the protein-coding genes in mammalian genomes is essential for understanding the human genome, including its cellular components, regulatory controls and evolutionary

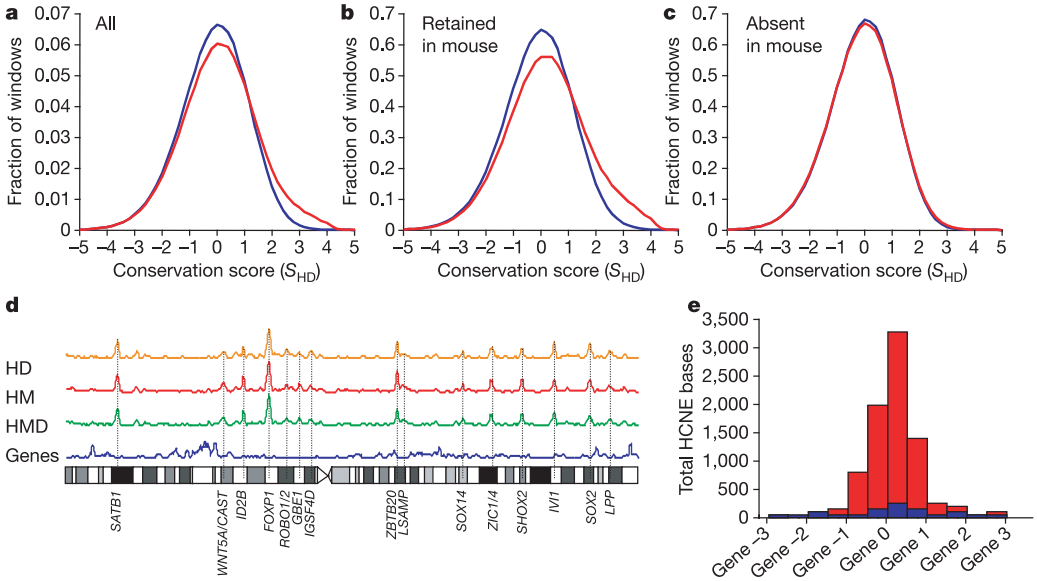


Figure 4 | Conservation of orthologous sequence between human and dog. **a**, Histogram of conservation scores, *S*, for all 50-bp windows across the human genome with at least 20 bases of orthologous sequence aligning to the dog genome, for all aligning sequences (red) and for ancestral repeat sequence only (blue). **b**, Conservation scores for the subset of windows that also have at least 20 bases of orthologous sequence aligning to the mouse genome. **c**, Conservation scores of the complementary subset of windows lacking such orthologous sequence in mouse. **d**, Density of 50-bp windows not overlapping known coding regions, for which $P_{selection}(S) > 95\%$, based on comparisons between human and dog (HD), human and mouse (HM), or between human, mouse and dog (HMD), and the density of known genes, all in 1-Mb sliding windows across human chromosome 3. **e**, Enrichment of HCNEs in the immediate neighbourhood of genes encoding developmental regulators in the 204 highly conserved regions. The histogram shows the median number of HCNE bases in the intronic and surrounding intergenic sequence, for the 197 known or putative development regulators (indicated by top of red bar) and for all of the 1,285 genes (blue bar). The histogram is centred at the 5′-end of the gene (marked 0) and each bin corresponds to half of the normalized distance to the flanking consecutive upstream genes (marked −1, −2 and −3) or consecutive downstream genes (1, 2 and 3) as indicated. The sequences surrounding the developmental genes are typically longer, have more HCNE sequence and have a higher density of HNCE sequence than other genes in the regions (see Supplementary Information).

© 2005 Nature Publishing Group

constraints. The number of protein-coding genes in human has been a topic of considerable debate, with estimates steadily falling from ~100,000 to 20,000–25,000 over the past decade[21,22,67–70]. We analysed the dog genome in order to refine the human gene catalogue and to assess the evolutionary forces shaping mammals. (In the Genes section, 'gene' refers only to a protein-coding gene.)

Gene predictions in dog and human. We generated gene predictions for the dog genome using an evidence-based method (see Supplementary Information). The resulting collection contains 19,300 dog gene predictions, with nearly all being clear homologues of known human genes.

The dog gene count is substantially lower than the ~22,000-gene models in the current human gene catalogue (EnsEMBL build 26). For many predicted human genes, we find no convincing evidence of a corresponding dog gene. Much of the excess in the human gene count is attributable to spurious gene predictions in the human genome (M. Clamp, personal communication).

Gene duplications. Gene duplication is thought to contribute substantially to functional innovation[69,71]. We identified 216 gene duplications that are specific to the dog lineage and 574 that are specific to the human lineage, using the synonymous substitution rate K_S as a distance metric and taking care to discard likely pseudogenes. (The CanFam 2.0 assembly contains approximately 24 additional gene duplications, mostly olfactory receptors.) Human genes are thus 2.7-fold more likely to have undergone duplication than are dog genes over the same time period. This may reflect increased repeat-mediated segmental duplication in the human lineage[72].

Although gene duplication has been less frequent in dog than human, the affected gene classes are very similar. Prominent among the lineage-specific duplicated genes are genes that function in adaptive immunity, innate immunity, chemosensation and reproduction, as has been seen for other mammalian genomes[24,25,69,71]. Reproductive competition within the species and competition against parasites have thus been major driving forces in gene family expansion.

The two gene families with the largest numbers of dog-specific genes are the histone H2B family and the α-interferons, which cluster in monophyletic clades when compared to their human homologues. This is particularly notable for the α-interferons, for which the gene families within the six species (human, mouse, rat, dog, cat and horse) are apparently monophyletic. This may be due either to coincidental independent gene duplication in each of the six lineages or to ongoing gene conversion events that have homogenized ancestral gene duplicates[73].

Evolution of orthologous genes across three species. The dog genome sequence allows us for the first time to characterize the large-scale patterns of evolution in protein-coding genes across three major mammalian orders. We focused on a subset of 13,816 human, mouse and dog genes with 1:1:1 orthology. For each, we inferred the number of lineage-specific synonymous (K_S) and non-synonymous (K_A) substitutions along each lineage and calculated the K_A/K_S ratio (Table 2 and Supplementary Information), a traditional measure of the strength of selection (both purifying and directional) on proteins[74].

The median K_A/K_S ratio differs sharply across the three lineages ($P < 10^{-44}$, Mann-Whitney *U*-test), with the dog lineage falling between mouse and human. Population genetic theory predicts[75] that the strength of purifying selection should increase with effective population size (N_e). The observed relationship (mouse < dog < human) is thus consistent with the evolutionary prediction, given the expectation that smaller mammals tend to have larger effective population sizes[76].

We next searched for particular classes of genes showing deviations from the expected rate of evolution for a species. Such variation in rate (heterotachy) may point to lineage-specific positive selection or relaxation of evolutionary constraints[77]. We developed a statistical method similar to the recently described Gene Set Enrichment Analysis (GSEA)[78–80] to detect evidence of heterotachy for sets of functionally related genes (see Supplementary Information). Briefly, the approach involves ranking all genes by K_A/K_S ratio, testing whether the set is randomly distributed along the list and assessing the significance of the observed deviations by comparison with randomly permuted gene sets. In contrast to previous studies, which focused on small numbers of genes with prior hypotheses of selection, this approach detects signals of lineage-specific evolution in a relatively unbiased manner and can provide context to the results of more limited studies.

A total of 4,950 overlapping gene sets were studied, defined by such criteria as biological function, cellular location or co-expression (see Supplementary Information). Overall, the deviations between the three lineages are small, and median K_A/K_S ratios for particular gene sets are highly correlated for each pair of species (Supplementary Fig. S11). However, there is greater relative variation in human–mouse and dog–mouse comparisons than in human–dog comparisons (Supplementary Fig. S12).

This suggests that observed heterotachy between human and mouse must be interpreted with caution. For example, there is a great interest in the identification of genetic changes underlying the unique evolution of the human brain. A recent study[81] highlighted 24 genes involved in brain development and physiology that show signs of accelerated evolution in the lineage leading from ancestral primates to humans when compared to their rodent orthologues. We observe the same trend for the 18 human genes that overlap with the genes studied here, but find at least as many genes with higher relative acceleration in the dog lineage (see Supplementary Information). Heterotachy relative to mouse therefore does not appear to be a distinctive feature of the human lineage. It may reflect decelerated evolution in the rodent lineage, or possibly independent adaptive evolution in the human and dog lineages[82].

A small number of gene sets show evidence of significantly accelerated evolution in the human lineage, relative to both mouse and dog (32 sets at $z \geq 5.0$ versus zero sets expected by chance, $P < 10^{-4}$; Fig. 5a). These sets fall into two categories: genes expressed exclusively in testis, and (nuclear) genes encoding subunits of the mitochondrial electron transport chain (ETC) complexes. The former are believed to undergo rapid evolution as a consequence of sperm competition across a wide range of species[83–85], and lineage-specific acceleration suggests that sexual selection may have been a particularly strong force in primate evolution. The selective forces acting on the latter category are less obvious. Because of the importance of mitochondrial ATP generation for sperm motility[86], and the potentially antagonistic co-evolution of these genes with maternally inherited mitochondrial DNA-encoded subunits[87], we

Table 2 | Evolutionary rates for 1:1:1 orthologues among dog, mouse and human

	Median (20–80th percentile range)			Spearman's *rho*		
	Dog*	Mouse	Human	Dog–human	Dog–mouse	Human–mouse
K_S	0.210 (0.138–0.322)	0.416 (0.310–0.558)	0.139 (0.0928–0.214)	0.47	0.50	0.52
K_A	0.021 (0.006–0.051)	0.038 (0.013–0.087)	0.017 (0.005–0.040)	0.87	0.87	0.86
K_A/K_S	0.095 (0.030–0.221)	0.088 (0.031–0.197)	0.112 (0.034–0.272)	0.80	0.85	0.82

*Estimates are based on unrooted tree. The dog branch thus includes the branch from the boreoeutherian ancestor to the primate–rodent split.

© 2005 Nature Publishing Group

propose that sexual selection may also be the primary force behind the rapid evolution of the primate ETC genes. Given the ubiquitous role of mitochondrial function, however, such sexual selection may have led to profound secondary effects on physiology[88].

We found no gene sets with comparably strong evidence for dog-specific accelerated evolution. There is, however, a small excess of sets with moderately high acceleration scores (19 sets at $z \geq 3.0$ versus 5 sets expected by chance, $P < 0.02$; Fig. 5b). These sets, which are primarily related to metabolism, may contain promising candidates for follow-up studies of molecular adaptation in carnivores.

Polymorphism and haplotype structure in the domestic dog

The modern dog has a distinct population structure with hundreds of genetically isolated breeds, widely varying disease incidence and distinctive morphological and behavioural traits[89,90]. Unlocking the full potential of the dog genome for genetic analysis requires a dense SNP map and an understanding of the structure of genetic variation both within and among breeds.

Generating a SNP map. We generated a SNP map of the dog genome containing >2.5 million distinct SNPs mapped to the draft genome sequence, corresponding to an average density of approximately one SNP per kb (Table 3). The SNPs were discovered in three complementary ways (see Supplementary Information). (1) We identified SNPs within the sequenced boxer genome (set 1; ~770,000 SNPs) by searching for sites at which alternative alleles are supported by at least two independent reads each. We tested a subset ($n = 40$ SNPs) by genotyping and confirmed all as heterozygous sites. (2) We compared the 1.5× sequence from the standard poodle[16] with the draft genome sequence from the boxer (set 2; ~1,460,000 SNPs). (3) We generated shotgun sequence data from nine diverse dog breeds (~100,000 reads each, 0.02× coverage), four grey wolves and one coyote (~22,000 reads each, 0.004× coverage) and compared it to the boxer (set 3; ~440,000 SNPs). We tested a subset ($n = 1,283$ SNPs) by genotyping and confirmed 96% as true polymorphisms.

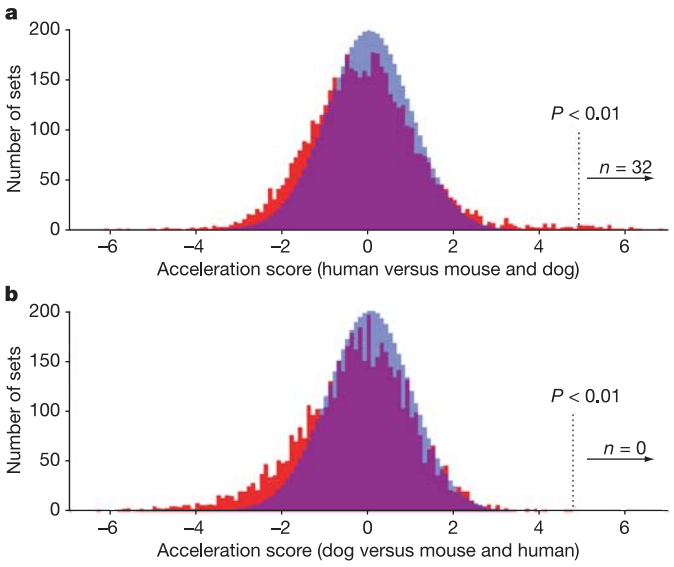


Figure 5 | Gene sets showing accelerated evolution along the human and dog lineages. a, Distribution of acceleration scores along the human lineage relative to both mouse and dog, observed for 4,950 gene sets (red). The expected distribution based on 10,000 randomized trials is shown in blue. The dotted line shows the acceleration score for which the probability of observing even a single set by random chance (out of the 4,950 sets tested) is less than 1%. In fact, 32 sets show acceleration scores on the human lineage exceeding this threshold. **b**, The observed (red) and expected (blue) distribution of acceleration scores for the dog lineage, relative to both human and mouse.

Table 3 | SNPs discovered in dogs, wolves and coyotes compared to the boxer assembly

Set number	Breed or species	Number of SNPs	SNP rate (one per *x* bases)
1	Boxer versus boxer	**768,948**	3,004 (observed) **1,637** (corrected)
2	Boxer versus poodle	**1,455,007**	**894**
3a	Boxer versus breeds*		
	German shepherd	45,271	900
	Rottweiler	44,097	917
	Bedlington terrier	44,168	913
	Beagle	42,572	903
	Labrador retriever	40,730	926
	English shepherd	40,935	907
	Italian greyhound	39,390	954
	Alaskan malamute	45,103	787
	Portuguese water dog	45,457	896
	Total distinct SNPs	**373,382**	**900**
3b	Boxer versus Canids†		
	China grey wolf	12,182	580
	Alaska grey wolf	13,888	572
	India grey wolf	14,510	573
	Spanish grey wolf	10,349	587
	California coyote	20,270	417
	Total distinct SNPs	**71,381**	
3	Set 3 total distinct SNPs	**441,441**	
Total	Total distinct SNPs	**2,559,519**	

*Based on ~100,000 sequence reads per breed.
†Based on ~20,000 sequence reads per wolf.

The SNP rate between the boxer and any of the different breeds is one SNP per ~900 bp, with little variation among breeds (Table 3). The only outlier (~1/790 bp) is the Alaskan malamute, which is the only breed studied that belongs to the Asian breed cluster[91]. The grey wolf (~1/580 bp) and coyote (~1/420 bp) show greater variation when compared with the boxer, supporting previous evidence of a bottleneck during dog domestication, whereas that the SNP rate is lower in the grey wolf than in the coyote reflects the closer relationship of the grey wolf to the domestic dog[1–3,92] (see section 'Resolving canid phylogeny').

The observed SNP rate within the sequenced boxer assembly is ~1/3,000 bp. This underestimates the true heterozygosity owing to the conservative criterion used for identifying SNPs within the boxer assembly (requiring two reads containing each allele); correcting for this leads to an estimate of ~1/1,600 bp (see Supplementary Information). This low rate reflects reduced polymorphism within a breed, compared with the greater variation of ~1/900 bp between breeds.

To assess the utility of the SNPs for dog genetics, we genotyped a subset from set 3a ($n = 1,283$) in 20 dogs from each of ten breeds (Supplementary Table S16). Within a typical breed, ~73% of the SNPs were polymorphic. The polymorphic SNPs have minor allele frequencies that are approximately evenly distributed between 5% and 50% (allele frequencies less than 5% are not reliable with only 40 chromosomes sampled). In addition, the SNPs from sets 2 and 3 have a roughly uniform distribution across the genome (Fig. 6a, see below concerning set 1). The SNP map thus has high density, even distribution and high cross-breed polymorphism, indicating that it should be valuable for genetic studies.

Expectations for linkage disequilibrium and haplotype structure. Modern dog breeds are the product of at least two population bottlenecks, the first associated with domestication from wolves (~7,000–50,000 generations ago) and the second resulting from intensive selection to create the breed (~50–100 generations ago). This population history should leave distinctive signatures on the patterns of genetic variation both within and across breeds. We might expect aspects of both the long-range LD seen in inbred mouse strains, with strain-specific haplotypes extending over multiple megabases, and the short-range LD seen in humans, with ancestral haplotype blocks typically extending over tens of kilobases. Specifically,

© 2005 Nature Publishing Group

long-range LD would be expected within dog breeds and short-range LD across breeds.

Preliminary evidence of long-range LD within breeds has been reported[90]. Five genome regions were examined (~1% of the genome) in five breeds using ~200 SNPs with high minor allele frequency. LD seemed to extend 10–100-fold further in dog than in human, with relatively few haplotypes per breed.

With the availability of a genome sequence and a SNP map, we sought to undertake a systematic analysis of LD and haplotype structure in the dog genome.

Haplotype structure within the boxer assembly. We first analysed the structure of genetic variation within the sequenced boxer genome by examining the distribution of the ~770,000 SNPs detected between homologous chromosomes. Strikingly, the genome is a mosaic of long, alternating regions of near-total homozygosity and high heterozygosity (Fig. 6b, c), with observed SNP rates of ~14 per Mb and ~850 per Mb, respectively. (The latter is close to that seen within breeds and is indistinguishable when one corrects for the conservative criterion used to identify SNPs within the boxer assembly; see Supplementary Information.) The homozygous regions have an N50 size of 6.9 Mb and cover 62% of the genome, and the heterozygous regions have an N50 size of 1.1 Mb and cover 38% of the genome. The results imply that the boxer genome is largely comprised of vast haplotype blocks. The long stretches of homozygosity indicate regions in which the sequenced boxer genome carries the same haplotype on both chromosomes. The proportion of homozygosity (~62%) reflects the limited haplotype diversity within breeds.

Long-range haplotypes in different breeds. We sought to determine whether the striking haplotype structure seen in the boxer genome is representative of most dog breeds. To this end, we randomly selected ten regions of 15 Mb each (~6% of the genome) and examined linkage disequilibrium in these regions in a collection of 224 dogs, consisting of 20 dogs from each of ten breeds and one dog from each of 24 additional breeds (see Supplementary Tables S17–S19).

The ten breeds were chosen to represent all four clusters described in ref. 91. The selected breeds have diverse histories, with varying population size and bottleneck severity. For example, the Basenji is an ancient breed from Africa that has a small breeding population in the United States descending from dogs imported in the 1930s–1940s (refs 93, 94). The Irish wolfhound suffered a severe bottleneck two centuries ago, with most dogs today being descendents of a single dog in the early 1800s (refs 5, 94). In contrast, the Labrador retriever and golden retriever have long been, and remain, extremely popular dogs

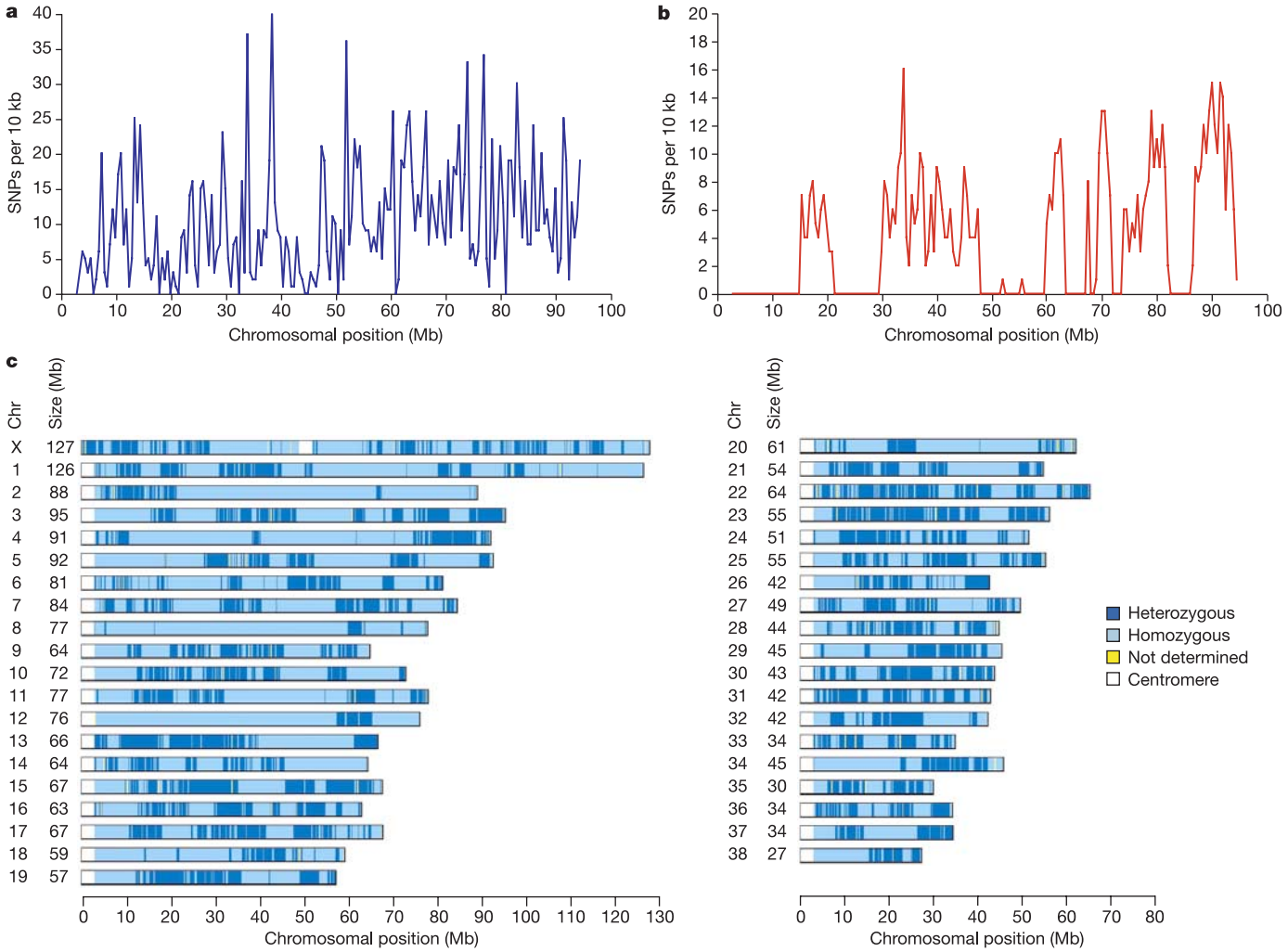


Figure 6 | The distribution of SNPs is fairly uniform across breeds, but non-uniform within the sequenced boxer assembly. **a**, SNPs across chromosome 3, generated by comparing the boxer assembly with WGS reads from nine breeds. **b**, The SNPs on chromosome 3 of the boxer assembly show an uneven distribution (plotted in 500-kb windows). Note that boxer SNPs were identified using a more conservative method, lowering the observed SNP rate by roughly twofold. **c**, An alternating pattern of large homozygous (light blue, ~62% of genome; N50 size 6.9 Mb) and large heterozygous (dark blue ~38% of genome; N50 size 1.1 Mb) blocks indicates large identical or divergent haplotypes across the boxer genome. White indicates centromeric sequence.

© 2005 Nature Publishing Group

(with ~150,000 and ~50,000 new puppies registered annually, respectively). They have not undergone such recent severe bottlenecks, but some lines have lost diversity because of the repeated use of popular sires[89]. The Glen of Imaal terrier represents the opposite end of the popularity spectrum, with fewer than 100 new puppies registered with the American kennel Club each year.

The 224 dogs were genotyped for SNPs across each of the ten regions, providing 2,240 cases in which to assess long-range LD. The SNPs ($n = 1{,}219$; Supplementary Table S19) were distributed along the regions to measure the fall-off of genetic correlation, with higher density at the start of the region and lower densities at further distances (Fig. 7a). In 645 cases, we also examined the first 10 kb in

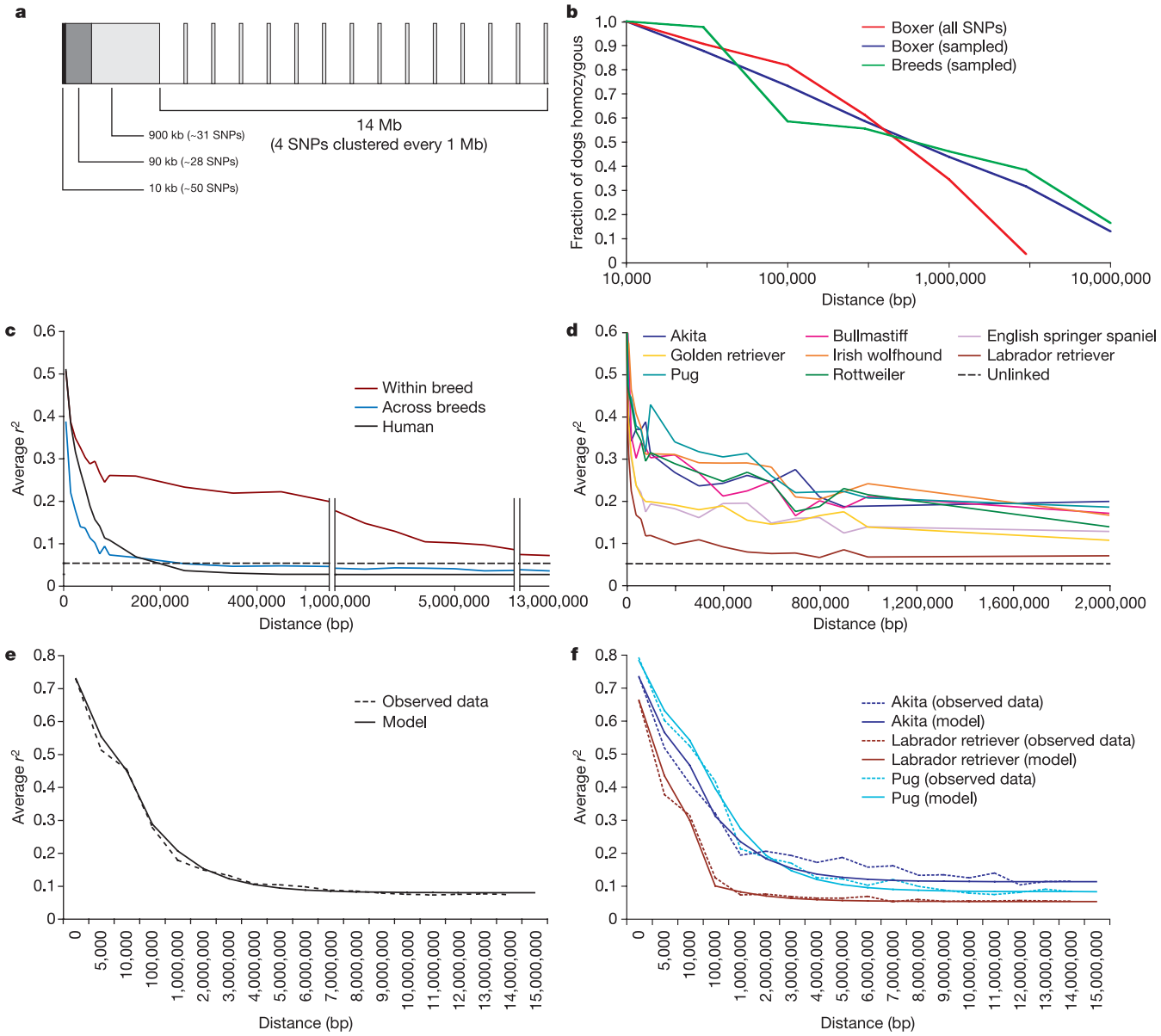


Figure 7 | Homozygous regions and linkage disequilibrium are nearly 100-fold longer within dog breeds than across the dog or human populations. **a**, Sampling design for ten random regions of 15 Mb each, used to assess the haplotype structure of ~6% of the genome (see Supplementary Information). For each region, we examined the first 10 kb through resequencing and dense genotyping. To detect long haplotypes, we genotyped SNPs distributed throughout the next 1 Mb and sampled SNPs at intervals of 1 Mb for the next 14 Mb. In total we genotyped 1,219 SNPs across the ten regions in a collection of 224 dogs (20 dogs from each of 10 breeds and one dog from each of 24 breeds). **b**, Conditional on a dog being homozygous for the initial 10-kb region ($n = 245$), we assessed the probability that the dog was homozygous for all SNPs within a given distance. The average proportion remaining homozygous is compared for the various breeds (green), for the boxer when sampled in the same ways as the breeds (blue) and for the boxer using all SNPs found in the genome sequence (red). About 50% of the individuals seem to be homozygous throughout 1 Mb both in the boxer and other breeds, indicating that other breeds have comparable long-range homozygosity. **c**, Linkage disequilibrium (LD) as a function of distance is shown as the r^2 statistic within individual breeds (red), across various breeds (blue) and a human population (black) taken from the CEPH collection genotyped as part of the ENCODE component of the International HapMap Project[118]. For the overall dog and human populations, LD falls rapidly, reaching the baseline level seen for unlinked loci by ~200 kb. In contrast, LD for individual breeds falls initially but then stays at a moderately high level across several megabases. **d**, The LD curves are broadly similar for most breeds, but the proportion of long-range LD is correlated with known breed history. **e**, The observed within-breed LD curve (averaged across breeds) is well fitted by a simple model with a domestication bottleneck 10,500 generations ago and a breed-creation bottleneck occurring 50 generations ago (see Supplementary Information). **f**, LD curves for individual dog breeds can be fitted by models with different breed-creation bottlenecks. The poorest fit is obtained for the akita, the breeding history of which is known to involve two separate breed-creation bottlenecks.

© 2005 Nature Publishing Group

greater detail by denser genotyping (with ~2 SNPs per kb) in 405 cases and complete resequencing in 240 cases. The resequencing data yielded a heterozygosity rate of ~1 SNP per 1,500 bp, essentially equivalent to the rate seen in the sequenced boxer genome.

On the basis of examining the first 10 kb, we found that ~38% of instances seem to be completely homozygous and that all dogs seem to be homozygous for at least one of the ten regions. We then measured the distance over which homozygosity persisted. Of instances homozygous in the initial 10-kb segment, 46% were homozygous across 1 Mb and 17% were still homozygous across 10 Mb (Fig. 7b). The fall-off in homozygosity is essentially identical to that seen in the boxer genome, provided that the boxer data are sampled in an equivalent manner (see Supplementary Information). This indicates that the long-range haplotype structure seen in the boxer is typical of most dog breeds, although the precise haplotypes vary with breed and the locations of homozygous regions vary between individuals.

We also assessed long-range correlations by calculating r^2, a traditional measure of LD, across the 15-Mb regions. The r^2 curve representing the overall dog population (one dog from each of 24 breeds) drops rapidly to background levels. This is in sharp contrast to the r^2 curves within each breed. Within breeds, LD is biphasic, showing a sharp initial drop within ~90 kb followed by an extended shoulder that gradually declines to the background (unlinked) level by 5–15 Mb in most breeds (Fig. 7c). The basic pattern is similar in all ten regions (Supplementary Fig. S13) and in all breeds (Fig. 7d). (Labrador retrievers show the shortest LD, probably due to their mixed aetiology and large population size.)

The biphasic r^2 curves within each breed thus consist of two components (Fig. 7e), at scales differing by ~100-fold. The first component matches the fall-off in the general dog population and is likely to represent the short-range de-correlation of local haplotype blocks in the ancestral dog population. The second component represents long-range breed-specific haplotypes (Fig. 8a). Notably, the first component falls off nearly twice as quickly as the LD in the human population (~200 kb), and the second component falls off slightly slower than seen in laboratory mouse strains[95].

Modelling the effects of population history. We tested this interpretation by performing mathematical simulations on a dog population that underwent an ancient bottleneck and recent breed-creation bottlenecks, using the coalescent approach[96] (see Supplementary Information). Our experimental results were well fitted by models assuming an ancient bottleneck (effective domesticated population size 13,000, inbreeding coefficient $F = 0.12$) occurring ~9,000 generations ago (corresponding to ~27,000 years) and subsequent breed-creation bottlenecks of varying intensities occurring 30–90 generations ago[97] (Supplementary Fig. S14). The model closely reproduces the observed r^2 curves and the observed polymorphism rates within breeds, among breeds and between dog and grey wolf. The model also yields estimates of breed-specific bottlenecks that are broadly consistent with known breed histories. For example, Labrador retrievers, and to a lesser extent golden retrievers and English springer spaniels, show less severe bottlenecks.

Deterministically modelled results (Fig. 7e, f) indicate that a simple, two-bottleneck model provides a close fit to the data for the breeds. They do not rule out a more complex population history, such as multiple domestication events, low levels of continuing gene flow between domestic dog and grey wolf[97,98] or multiple bottlenecks within breeds. Notably, the akita yields the poorest fit to the model, with an r^2 curve that appears to be triphasic. This may reflect the initial creation of the breed as a hunting dog in Japan ~450 generations ago, and a consecutive bottleneck associated with its introduction into the United States during the 1940s (ref. 99).

Haplotype diversity. We next studied haplotype diversity within and among breeds, using the dense genotypes from the 10-kb regions. Across the 645 cases examined, there is an average of ~10 distinct haplotypes per region. Within a breed, we typically see four of these haplotypes, with the average frequency of the most common haplotype being 55% and the average frequency of the two most common being 80% (Fig. 8c and Supplementary Fig. S18). The haplotypes and their frequencies differ sharply across breeds. Nonetheless, 80% of the haplotypes seen with a frequency of at least 5% in one breed are found in other breeds as well (Supplementary Table S26). This extends previous observations of haplotype sharing across breeds[90]. In particular, the inclusion of all SNPs with a minor allele frequency ≥5% across all breeds provides a more accurate picture of haplotype sharing, because the analysis includes haplotypes that are rare within a single breed but more common across the population.

We then inferred the ancestral haplotype block structure in the ancestral dog population (before the creation of modern breeds) by combining the data across breeds and applying methods similar to those used for haplotype analysis in the human genome[100] (see Supplementary Information). In the 10-kb regions studied, one or two haplotype blocks were typically observed. Additional data across 100-kb regions suggest that the ancestral blocks have an average size of ~10 kb. The blocks typically have ~4–5 distinct haplotypes across the entire dog population (Fig. 8b). The overall situation closely resembles the structure for the human genome, although with slightly smaller block size (Supplementary Figs S15–S19 and Supplementary Table S24–26).

Ancestral and breed-specific haplotypes. A clear picture of the population genetic history of dogs emerges from the results detailed above:

- The ancestral dog population had short-range LD. The haplotype blocks were somewhat shorter than in modern humans (~10 kb versus ~20 kb in human), consistent with the dog population being somewhat older than the human population (~9,000 generations versus ~4,000 generations). Haplotype blocks at large distances were essentially uncorrelated (Fig. 8a).
- Breed creation introduced tight breed-specific bottlenecks, at least for the breeds examined. From the great diversity of long-range haplotype combinations carried in the ancestral population, the founding chromosomes emerging from the bottleneck represented only a small subset. These became long-range breed-specific haplotypes (Fig. 8a).
- Although the breed-specific bottlenecks were tight, they did not cause massive random fixation of individual haplotypes. Only 13% of the small ancestral haplotypes are monomorphic within a typical breed, consistent with the estimated inbreeding coefficient of ~12%. Across larger regions (≥100 kb), we observed no cases of complete fixation within a breed (Supplementary Fig. S20).
- There is notable sharing of 100-kb haplotypes across breeds, with ~60% seen in multiple breeds although with different frequencies. On average, the probability of sampling the same haplotype on two chromosomes chosen from different breeds is roughly twofold lower than for chromosomes chosen within a single breed (Supplementary Fig. S21).

Implications for genetic mapping. These results have important implications for the design of dog genetic studies. Although early efforts focused on cross-breeding of dogs for linkage analysis[101–103], it is now clear that within-breed association studies offer specific advantages in the study of both monogenic and polygenic diseases. First, they use existing dogs coming to medical attention and do not require the sampling of families with large numbers of affected individuals. Such studies should be highly informative, because dog breeds have retained substantial genetic diversity. Moreover, they will require a much lower density of SNPs than comparable human association studies, because the long-range LD within breeds extends ~50-fold further than in humans[90,104,105].

Whereas human association studies require >300,000 evenly spaced SNPs[100,106,107], the fact that LD extends over at least 50-fold greater distances in dog suggests that dog association studies would require perhaps ~10,000 evenly spaced SNPs. To estimate the

© 2005 Nature Publishing Group

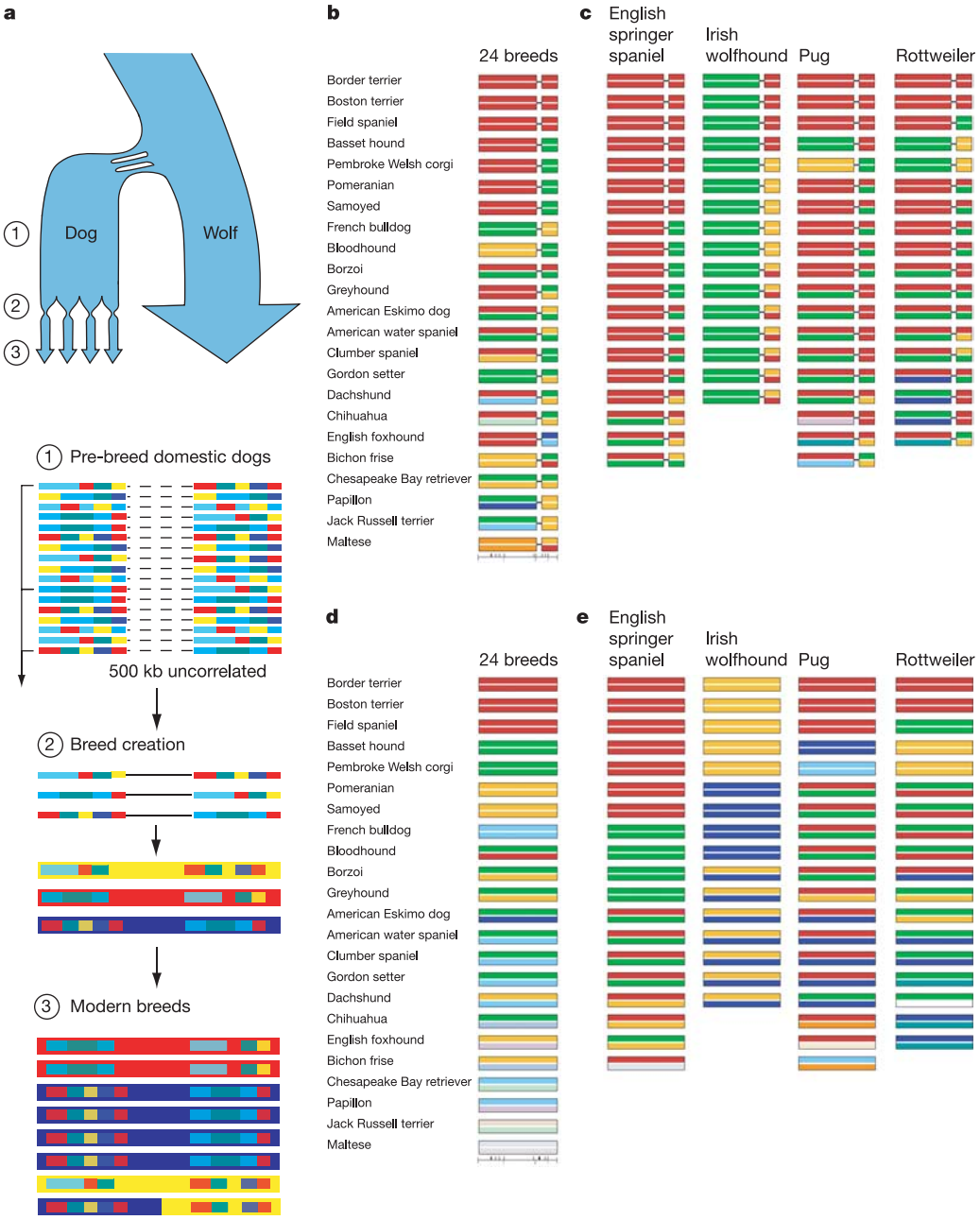


Figure 8 | Two bottlenecks, one old and one recent, have shaped the haplotype structure and linkage disequilibrium of canine breeds. **a**, Modern haplotype structure arose from key events in dog breeding history. The domestic dog diverged from wolves 15,000–100,000 years ago[97,119], probably through multiple domestication events[98]. Recent dog breeds have been created within the past few hundred years. Both bottlenecks have influenced the haplotype pattern and LD of current breeds. (1) Before the creation of modern breeds, the dog population had the short-range LD expected on the basis of its large size and time since the domestication bottleneck. (2) In the creation of modern breeds, a small subset of chromosomes was selected from the pool of domestic dogs. The long-range patterns that happened to be carried on these chromosomes became common within the breed, thereby creating long-range LD. (3) In the short time since breed creation, these long-range patterns have not yet been substantially broken down by recombination. Long-range haplotypes, however, still retain the underlying short-range ancestral haplotype blocks from the domestic dog population, and these are revealed when one examines chromosomes across many breeds. **b**, **c**, Distribution of ancestral haplotype blocks in a 10-kb window on chromosome 6 at ~31.4 Mb across 24 breeds (**b**) and within four breeds (**c**). Ancestral haplotype blocks are 5–15 kb in size (which is shorter than the ~25-kb blocks seen in humans) and are shared across breeds. Typical blocks show a spectrum of ~5 haplotypes, with one common major haplotype. Blocks were defined using the modified four-gamete rule (see Supplementary Information) and each haplotype (minor allele frequency (maf) > 3%) within a block was given a unique colour. **d**, **e**, Distribution of breed-derived haplotypes across a 10-kb window on chromosome 6 at ~31.4 Mb across 24 breeds (**d**) and within four breeds (**e**). Each colour denotes a distinct haplotype (maf > 3%) across 11 SNPs in the 10-kb window for each of the analysed dogs. Pairs of haplotypes have an average of 3.7 differences. Most haplotypes can be definitively identified on the basis of homozygosity within individual dogs. Grey denotes haplotypes that cannot be unambiguously phased owing to rare alleles or missing data. Within each of the four breeds shown, there are 2–5 haplotypes, with one or two major haplotypes accounting for the majority of the chromosomes. Across the 24 breeds, there are a total of seven haplotypes. All but three are seen in multiple breeds, although at varying frequencies.

© 2005 Nature Publishing Group

number of SNPs required, we generated SNP sets from ten 1-Mb regions by coalescent simulations using the bottleneck parameters that generate SNP rates and LD curves equivalent to the actual data (Supplementary Fig. S14 and Supplementary Table S20). We then selected individual SNPs as 'disease alleles' and tested our ability to map them by association analysis with various marker densities (Fig. 9a).

For disease alleles causing a simple mendelian dominant trait with high penetrance and no phenocopies, there is overwhelming power to map the locus (Fig. 9a). Using ~15,000 evenly spaced SNPs and a log likelihood odds ratio (LOD score) score threshold of 5, the probability of detecting the locus is over 99% given a collection of 100 affected and 100 unaffected dogs. (The LOD score threshold corresponds to a false positive rate of 3% loci per genome.)

For a multigenic trait, the power to detect disease alleles depends on several factors, including the relative risk conferred by the allele, the allele frequency and the interaction with other alleles. We investigated a simple model of an allele that increases risk by a multiplicative factor (λ) of 2 or 5 (see Supplementary Information). Using the above SNP density and LOD score threshold, the power to detect a locus with a sample of 100 affected and 100 unaffected dogs is 97% for $\lambda = 5$ and 50% for $\lambda = 2$ (Fig. 9b, c). Although initial mapping will be best done by association within breeds, subsequent fine-structure mapping to pinpoint the disease gene will probably benefit from cross-breed comparison. Given the genetic relationships across breeds described above, it is likely that the same risk allele will be carried in multiple breeds. By comparing risk-associated haplotypes in multiple breeds, it should be possible to substantially narrow the region containing the gene.

Resolving canid phylogeny

The dog family, Canidae, contains 34 closely related species that diverged within the last ~10 million years[1]. Resolving the evolutionary relationships of such closely related taxa has been difficult because a great quantity of genomic sequence is typically required to yield enough informative nucleotide sites for the unambiguous reconstruction of phylogenetic trees. We sought to streamline the process of evolutionary reconstruction by exploiting our knowledge of the dog genome to select genomic regions that would maximize the amount of phylogenetic signal per sequenced base. Specifically, we sought regions of rapidly evolving, unique sequence.

We first compared the coding regions of 13,816 dog genes with human–dog–mouse 1:1:1 orthologues to find those with high neutral evolutionary divergence (comparing K_S and K_A/K_S). We selected 12 exons (8,080 bp) for sequencing, based on the criteria that their sequences (1) are consistent with the known phylogeny of human, dog, mouse and rat, (2) have a high percentage of bases (≥15%) that are informative for phylogenetic reconstruction in the human, dog, mouse and rat phylogenies, and (3) could be successfully amplified in all canids. The chosen exons contain 3.3-fold more substitutions than random exonic sequence. Using our SNP database, we also evaluated introns to identify those with high variation between dog and coyote. We selected four introns (3,029 bp) that contained ~5-fold more SNPs than the background frequency. We sequenced these exons and introns (11,109 bp) in 30 out of 34 living wild canids, and we combined the data with additional sequences (3,839 bp) from recent studies[3,92].

The resulting evolutionary tree has a high degree of statistical support (Fig. 10), and uniquely resolves the topology of the dog's closest relatives. Grey wolf and dog are most closely related (0.04% and 0.21% sequence divergence in nuclear exon and intron sequences, respectively), followed by a close affiliation with coyote, golden jackal and Ethiopian wolf, three species that can hybridize with dogs in the wild (Fig. 10). Closest to this group are the dhole and African wild dog, two species with a uniquely structured meat-slicing tooth, suggesting that this adaptation was later lost. The molecular tree supports an African origin for the wolf-like canids, as the two African jackals are the most basal members of this clade. The two other large groupings of canids are (1) the South American canids, which are clearly rooted by the two most morphologically divergent canids, the maned wolf and bush dog; and (2) the red fox-like canids, which are rooted by the fennec fox and Blanford's fox, but now also include the raccoon dog and bat-eared fox with higher support. Together, these three clades contain 93% of all living canids. The grey fox lineage seems to be the most primitive and suggests a North American origin of the living canids about 10 million years ago[1].

These results demonstrate the close kinship of canids. Their limited sequence divergence suggests that many molecular tools developed for the dog (for example, expression microarrays) will be useful for exploring adaptation and evolutionary divergence in other canids as well.

Conclusions

Genome comparison is a powerful tool for discovery. It can reveal unknown—and even unsuspected—biological functions, by sifting the records of evolutionary experiments that have occurred over 100 years or over 100 million years. The dog genome sequence illustrates the range of information that can be gleaned from such studies.

Mammalian genome analysis is helping to develop a global picture of gene regulation in the human genome. Initial comparison with rodents revealed that ~5% of the human genome is under purifying selection, and that the majority of this sequence is not protein-

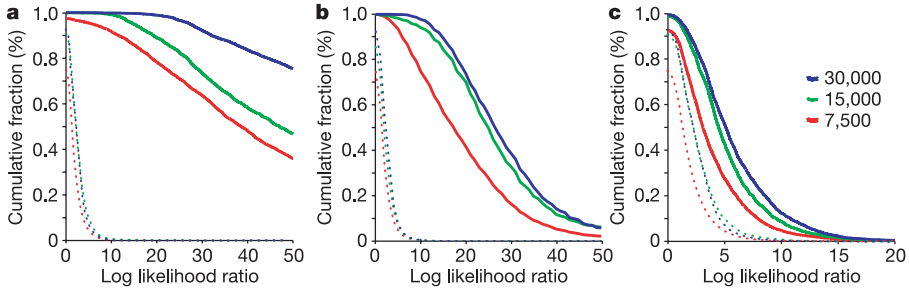


Figure 9 | Power to detect a disease locus by association mapping. One SNP was designated as a disease allele under one of three genetic models: (**a**) simple mendelian dominant, (**b**) fivefold multiplicative increase in risk and (**c**) twofold multiplicative increase in risk. SNP genotypes across surrounding chromosomal regions of 1 Mb were simulated, using the coalescent model corresponding to observed within-breed variation (see text). Diploid genotypes across the chromosomal region were then generated for 100 affected and 100 unaffected dogs, based on the disease model, and association analysis was performed to detect the presence of the disease allele. The distribution of the maximum LOD score across the 1-Mb region is shown for analyses based on multi-SNP haplotypes (solid lines) with SNP densities equivalent to a genome-wide map with a total of 7,500 (red), 15,000 (green) or 30,000 (blue) SNPs. Dotted curves show the null distribution for a genome-wide search in which no disease locus is present (see Supplementary Information). A LOD score of 5 corresponds to <3% chance of a false positive across the genome. For this threshold, the power to detect a disease allele that increases risk by twofold using haplotype analysis and a map with 15,000 SNPs is ~50%.

© 2005 Nature Publishing Group

coding. The dog genome is now further clarifying this picture, as our data suggest that this ~5% represents functional elements common to all mammals. The distribution of these elements relative to genes is highly heterogeneous, with roughly half of the most highly conserved non-coding elements apparently devoted to regulating ~1% of human genes; these genes have important roles in development, and understanding the regulatory clusters that surround them may reveal how cellular states are established and maintained. In recent papers[32,108], the dog genome sequence has been used to greatly expand the catalogue of mammalian regulatory motifs in promoters and 3′-untranslated regions. The dog genome sequence is also being used to substantially revise the human gene catalogue. Despite these advances, it is clear that mammalian comparative genomics is still in its early stages. Progress will be markedly accelerated by the availability of many additional mammalian genome sequences, initially with light coverage[28] but eventually with near-complete coverage.

In addition to its role in studies of mammalian evolution, the dog has a special role in genomic studies because of the unparalleled phenotypic diversity among closely related breeds. The dog is a testament to the power of breeding programmes to select naturally occurring genetic variants with the ability to shape morphology, physiology and behaviour. Genome comparison within and across breeds can reveal the genes that underlie such traits, informing basic research on development and neurobiology. It can also identify disease genes that were carried along in breeding programmes. Potential benefits include insights into disease mechanism, and the possibility of clinical trials in disease-affected dogs to accelerate new therapeutics that would improve health in both dogs and humans. The SNP map of the dog genome confirms that dog breeds show the long-range haplotype structure expected from recent intensive breeding. Moreover, our analysis shows that the current collection of >2.5 million SNPs should be sufficient to allow association studies of

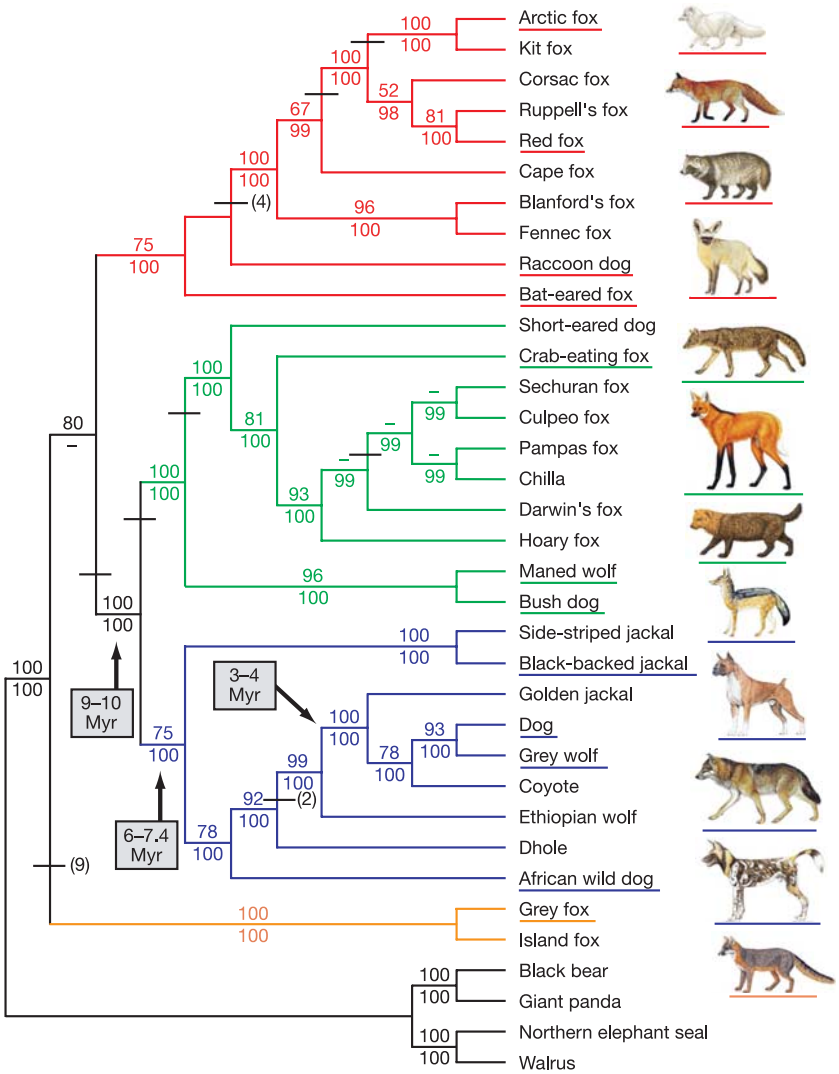


Figure 10 | Phylogeny of canid species. The phylogenetic tree is based on ~15 kb of exon and intron sequence (see text). Branch colours identify the red-fox-like clade (red), the South American clade (green), the wolf-like clade (blue) and the grey and island fox clade (orange). The tree shown was constructed using maximum parsimony as the optimality criterion and is the single most parsimonious tree. Bootstrap values and bayesian posterior probability values are listed above and below the internodes, respectively; dashes indicate bootstrap values below 50% or bayesian posterior probability values below 95%. Horizontal bars indicate indels, with the number of indels shown in parentheses if greater than one. Underlined species names are represented with corresponding illustrations. (Copyright permissions for illustrations are listed in the Supplementary Information.) Divergence time, in millions of years (Myr), is indicated for three nodes as discussed in ref. 1. For scientific names and species descriptions of canids, see ref. 119. A tree based on bayesian inference differs from the tree shown in two respects: it groups the raccoon dog and bat-eared fox as sister taxa, and groups the grey fox and island fox as basal to the clade containing these sister taxa. However, neither of these topological differences is strongly supported (see text and Supplementary Information).

© 2005 Nature Publishing Group

nearly any trait in any breed. Realizing the full power of dog genetics now awaits the development of appropriate genotyping tools, such as multiplex 'SNP chips'[109]—this is already underway. For millennia, dogs have accompanied humans on their travels. It is only fitting that the dog should also be a valued companion on our journeys of scientific discovery.

METHODS

Detailed descriptions of all methods are provided in the Supplementary Information. Links to all of the data can be obtained via the Broad Institute website (http://www.broad.mit.edu/tools/data.html).

WGS sequencing and assembly. Approximately 31.5 million sequence reads were derived from both ends of inserts (paired-end reads) from 4-, 10-, 40- and 200-kb clones, all prepared from primary blood lymphocyte DNA from a single female boxer. This particular animal was chosen for sequencing because it had the lowest heterozygosity rate among ~120 dogs tested at a limited set of loci; subsequent analysis showed that the genome-wide heterozygosity rate in this boxer is not substantially different from other breeds[91]. The assembly was carried out using an interim version of ARACHNE2+ (http://www.broad.mit.edu/wga/).

Genome alignment and comparison. Synteny maps were generated using standard methods[24] from pair-wise alignments of repeat masked assemblies using PatternHunter[110] on CanFam2.0. All other comparative analyses were performed on BLASTZ/MULTIZ[111,112] genome-wide alignments obtained from the UCSC genome browser (http://genome.ucsc.edu), based on CanFam1.0. Known interspersed repeats were identified and dated using RepeatMasker and DateRepeats[113]. The numbers of orthologous nucleotides were counted directly from the alignments using human (hg17) as the reference sequence for all overlaps except the dog–mouse overlap, for which pair-wise (CanFam1.0, mm5) alignments were used.

Divergence rate estimates. Orthologous ancestral repeats were excised from the genome alignment and realigned with the corresponding RepBase consensus using ClustalW. Nucleotide divergence rates were estimated from concatenated repeat alignments using baseml with the REV substitution model[114]. Orthologous coding regions were excised from the genome alignments using the annotated human coding sequences (CDS) from Ensembl and the UCSC browser Known Genes track (October 2004) as reference. K_A and K_S were estimated for each orthologue triplet using codeml with the F3 × 4 codon frequency model and no additional constraints.

Detection and clustering of sequence conservation. Pair-wise conservation scores and the fraction of orthologous sequences under purifying selection were estimated as in ref. 24. The three-way conservation score S_{HMD} was defined as $S_{HMD} = (p - u)/\sqrt{(u(1-u)/n}$, where n is the number of nucleotides aligned across all three genomes (human, mouse, dog) for each non-overlapping 50-bp window with more than 20 aligned bases, p is the fraction of nucleotides identical across all three genomes, and u is the mean identity of ancestral repeats within 500 kb of the window. HCNEs were defined as windows with $S_{HMD} > 5.4$ that did not overlap a coding exon, as defined by the UCSC Known Genes track, and HCNE clusters were defined as all runs of overlapping 1-Mb intervals (50-kb step size) across the human genome with HCNE densities in the 90th percentile.

Gene set acceleration scores. Gene annotation was performed on CanFam1.0. A set of 13,816 orthologous human, mouse and dog genes were identified and compiled into 4,950 gene sets containing genes related by functional annotations or microarray gene expression data. For each gene set S, the acceleration score $A(S)$ along a lineage is defined by (1) ranking all genes based on K_A/K_S within a lineage, (2) calculating the rank-sum statistic for the set along each lineage (denoted $a_{dog}(S)$, $a_{mouse}(S)$, $a_{human}(S)$), (3) calculating the rank-sum for the lineage minus the maximum rank-sum the other lineages, for example, $a_{human}(S)$–max($a_{dog}(S)$, $a_{mouse}(S)$) and (4) converting this rank-sum difference to a z-score by comparing it to the mean and standard deviation observed in 10,000 random sets of the same size. The expected number of sets at a given z-score threshold was estimated by repeating steps (1)–(4) 10,000 times for groups of 4,950 randomly permuted gene sets.

SNP discovery. The SNP discovery was performed on CanFam2.0. Set 1 SNPs were discovered by comparison of the two haplotypes derived from the boxer assembly using only high-quality discrepancies supported by two reads. SNPs in sets 2 and 3 were discovered by aligning reads or contigs to the boxer assembly and using the SSAHA SNP algorithm[115].

Haplotype structure. The SNPs within the sequenced boxer genome (CanFam2.0) were assigned to homozygous or heterozygous regions using a Viterbi algorithm[116]. To determine whether the haplotype structure seen in the boxer is representative of most dog breeds, we randomly selected ten regions of 15 Mb each (~6% of the CanFam2.0 genome) and examined the extent of homozygosity and linkage disequilibrium in these regions in a collection of 224 dogs, consisting of 20 dogs from each of 10 breeds (akita, basenji, bullmastiff, English springer spaniel, Glen of Imaal terrier, golden retriever, Irish wolfhound, Labrador retriever, pug and rottweiler) and one dog from each of 24 additional breeds (see Supplementary Information). For each instance in which a dog was homozygous in a particular 10-kb region, we measured the distance from the beginning of the 10-kb region to the first heterozygous SNP in the adjoining 100-kb, 1-Mb and 15-Mb data. This distance was used as the extent of homozygosity. The boxer sequence was sampled in an identical manner to the actual breed data. Linkage disequilibrium (represented by r^2) across the ten 15-Mb regions was assessed using Haploview[117].

Received 9 August; accepted 11 October 2005.

1. Wayne, R. K. *et al.* Molecular systematics of the Canidae. *Syst. Biol.* **46,** 622–653 (1997).
2. Vila, C. *et al.* Multiple and ancient origins of the domestic dog. *Science* **276,** 1687–1689 (1997).
3. Bardeleben, C., Moore, R. L. & Wayne, R. K. Isolation and molecular evolution of the selenocysteine tRNA (*Cf TRSP*) and RNase P RNA (*Cf RPPH1*) genes in the dog family, Canidae. *Mol. Biol. Evol.* **22,** 347–359 (2005).
4. Savolainen, P., Zhang, Y. P., Luo, J., Lundeberg, J. & Leitner, T. Genetic evidence for an East Asian origin of domestic dogs. *Science* **298,** 1610–1613 (2002).
5. American Kennel Club. *The Complete Dog Book* (eds Crowley, J. & Adelman, B.) (Howell Book House, New York, 1998).
6. Wayne, R. K. Limb morphology of domestic and wild canids: the influence of development on morphologic change. *J. Morphol.* **187,** 301–319 (1986).
7. Ostrander, E. A., Galibert, F. & Patterson, D. F. Canine genetics comes of age. *Trends Genet.* **16,** 117–123 (2000).
8. Patterson, D. Companion animal medicine in the age of medical genetics. *J. Vet. Intern. Med.* **14,** 1–9 (2000).
9. Sargan, D. R. IDID: inherited diseases in dogs: web-based information for canine inherited disease genetics. *Mamm. Genome* **15,** 503–506 (2004).
10. Chase, K. *et al.* Genetic basis for systems of skeletal quantitative traits: principal component analysis of the canid skeleton. *Proc. Natl Acad. Sci. USA* **99,** 9930–9935 (2002).
11. Breen, M. *et al.* Chromosome-specific single-locus FISH probes allow anchorage of an 1800-marker integrated radiation-hybrid/linkage map of the domestic dog genome to all chromosomes. *Genome Res.* **11,** 1784–1795 (2001).
12. Breen, M., Bullerdiek, J. & Langford, C. F. The DAPI banded karyotype of the domestic dog (Canis familiaris) generated using chromosome-specific paint probes. *Chromosome Res.* **7,** 401–406 (1999).
13. Breen, M. *et al.* An integrated 4249 marker FISH/RH map of the canine genome. *BMC Genomics* **5,** 65 (2004).
14. Hitte, C. *et al.* Facilitating genome navigation: survey sequencing and dense radiation-hybrid gene mapping. *Nature Rev. Genet.* **6,** 643–648 (2005).
15. Li, R. *et al.* Construction and characterization of an eightfold redundant dog genomic bacterial artificial chromosome library. *Genomics* **58,** 9–17 (1999).
16. Kirkness, E. F. *et al.* The dog genome: survey sequencing and comparative analysis. *Science* **301,** 1898–1903 (2003).
17. Sutter, N. & Ostrander, E. Dog star rising: The canine genetic system. *Nature Rev. Genet.* **5,** 900–910 (2004).
18. Galibert, F., Andre, C. & Hitte, C. Dog as a mammalian genetic model [in French]. *Med. Sci. (Paris)* **20,** 761–766 (2004).
19. Pollinger, J. P. *et al.* Selective sweep mapping of genes with large phenotypic effects. *Genome Res.* doi:10.1101/gr.4374505 (in the press).
20. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409,** 928–933 (2001).
21. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
22. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
23. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69–87 (2005).
24. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
25. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–521 (2004).
26. Murphy, W. J. *et al.* Molecular phylogenetics and the origins of placental mammals. *Nature* **409,** 614–618 (2001).
27. Thomas, J. W. *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424,** 788–793 (2003).
28. Margulies, E. H. *et al.* An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl Acad. Sci. USA* **102,** 4795–4800 (2005).
29. Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299,** 1391–1394 (2003).

© 2005 Nature Publishing Group

30. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304,** 1321–1325 (2004).
31. Eddy, S. R. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.* **3,** e10 (2005).
32. Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434,** 338–345 (2005).
33. Dermitzakis, E. T. *et al.* Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res.* **14,** 852–859 (2004).
34. Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13,** 91–96 (2003).
35. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
36. Richterich, P. Estimation of errors in "raw" DNA sequences: a validation study. *Genome Res.* **8,** 251–259 (1998).
37. Bailey, J. A., Baertsch, R., Kent, W. J., Haussler, D. & Eichler, E. E. Hotspots of mammalian chromosomal evolution. *Genome Biol.* **5,** R23 (2004).
38. Andelfinger, G. *et al.* Detailed four-way comparative mapping and gene order analysis of the canine *ctvm* locus reveals evolutionary chromosome rearrangements. *Genomics* **83,** 1053–1062 (2004).
39. Wang, W. & Kirkness, E. F. Short interspersed elements (SINEs) are a major source of canine genomic diversity. *Genome Res.* doi:10.1101/gr.3765505 (in the press).
40. Mamedov, I. Z., Arzumanyan, E. S., Amosova, A. L., Lebedev, Y. B. & Sverdlov, E. D. Whole-genome experimental identification of insertion/deletion polymorphisms of interspersed repeats by a new general approach. *Nucleic Acids Res.* **33,** e16 (2005).
41. Lin, L. *et al.* The sleep disorder canine narcolepsy is caused by a mutation in the *hypocretin* (*orexin*) *receptor 2* gene. *Cell* **98,** 365–376 (1999).
42. Pele, M., Tiret, L., Kessler, J. L., Blot, S. & Panthier, J. J. SINE exonic insertion in the *PTPLA* gene leads to multiple splicing defects and segregates with the autosomal recessive centronuclear myopathy in dogs. *Hum. Mol. Genet.* **14,** 1417–1427 (2005).
43. Fondon, J. W. III & Garner, H. R. Molecular origins of rapid and continuous morphological evolution. *Proc. Natl Acad. Sci. USA* **101,** 18058–18063 (2004).
44. Galtier, N. & Mouchiroud, D. Isochore evolution in mammals: a human-like ancestral structure. *Genetics* **150,** 1577–1584 (1998).
45. Belle, E. M., Duret, L., Galtier, N. & Eyre-Walker, A. The decline of isochores in mammals: an assessment of the GC content variation along the mammalian phylogeny. *J. Mol. Evol.* **58,** 653–660 (2004).
46. Bird, A. P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8,** 1499–1504 (1980).
47. Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci. USA* **90,** 11995–11999 (1993).
48. Cooper, G. M., Brudno, M., Green, E. D., Batzoglou, S. & Sidow, A. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res.* **13,** 813–820 (2003).
49. Hwang, D. G. & Green, P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl Acad. Sci. USA* **101,** 13994–14001 (2004).
50. Martin, A. P. & Palumbi, S. R. Body size, metabolic rate, generation time, and the molecular clock. *Proc. Natl Acad. Sci. USA* **90,** 4087–4091 (1993).
51. Gillooly, J. F., Allen, A. P., West, G. B. & Brown, J. H. The rate of DNA evolution: effects of body size and temperature on the molecular clock. *Proc. Natl Acad. Sci. USA* **102,** 140–145 (2005).
52. Laird, C. D., McConaughy, B. L. & McCarthy, B. J. Rate of fixation of nucleotide substitutions in evolution. *Nature* **224,** 149–154 (1969).
53. Li, W. H., Tanimura, M. & Sharp, P. M. An evaluation of the molecular clock hypothesis using mammalian DNA sequences. *J. Mol. Evol.* **25,** 330–342 (1987).
54. Webber, C. & Ponting, C. P. Hot spots of mutation and breakage in dog and human chromosomes. *Genome Res.* doi:10.1101/gr.3896805 (in the press).
55. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432,** 695–716 (2004).
56. Marques-Bonet, T. & Navarro, A. Chromosomal rearrangements are associated with higher rates of molecular evolution in mammals. *Gene* **353,** 147–154 (2005).
57. Miller, W., Makova, K. D., Nekrutenko, A. & Hardison, R. C. Comparative genomics. *Annu. Rev. Genomics Hum. Genet.* **5,** 15–56 (2004).
58. Smith, N. G., Brandstrom, M. & Ellegren, H. Evidence for turnover of functional noncoding DNA in mammalian genome evolution. *Genomics* **84,** 806–813 (2004).
59. Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3,** e7 (2005).
60. Ovcharenko, I. *et al.* Evolution and functional classification of vertebrate gene deserts. *Genome Res.* **15,** 137–145 (2005).
61. Walter, K., Abnizova, I., Elgar, G. & Gilks, W. R. Striking nucleotide frequency pattern at the borders of highly conserved vertebrate non-coding sequences. *Trends Genet.* **21,** 436–440 (2005).
62. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).
63. Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. Scanning human gene deserts for long-range enhancers. *Science* **302,** 413 (2003).
64. Kimura-Yoshida, C. *et al.* Characterization of the pufferfish *Otx2* cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification. *Development* **131,** 57–71 (2004).
65. Uchikawa, M., Ishida, Y., Takemoto, T., Kamachi, Y. & Kondoh, H. Functional analysis of chicken *Sox2* enhancers highlights an array of diverse regulatory elements that are conserved in mammals. *Dev. Cell* **4,** 509–519 (2003).
66. de la Calle-Mustienes, E. *et al.* A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate *Iroquois* cluster gene deserts. *Genome Res.* **15,** 1061–1072 (2005).
67. Daly, M. J. Estimating the human gene count. *Cell* **109,** 283–284 (2002).
68. Hogenesch, J. B. *et al.* A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* **106,** 413–415 (2001).
69. Emes, R. D., Goodstadt, L., Winter, E. E. & Ponting, C. P. Comparison of the genomes of human and mouse lays the foundation of genome zoology. *Hum. Mol. Genet.* **12,** 701–709 (2003).
70. Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet.* **25,** 232–234 (2000).
71. Wolfe, K. H. & Li, W. H. Molecular evolution meets the genomics revolution. *Nature Genet.* **33** (suppl.), 255–265 (2003).
72. Bailey, J. A., Liu, G. & Eichler, E. E. An Alu transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* **73,** 823–834 (2003).
73. Hughes, A. L. The evolution of the type I interferon gene family in mammals. *J. Mol. Evol.* **41,** 539–548 (1995).
74. Hurst, L. D. The *Ka/Ks* ratio: diagnosing the form of sequence evolution. *Trends Genet.* **18,** 486 (2002).
75. Ohta, T. Near-neutrality in evolution of genes and gene regulation. *Proc. Natl Acad. Sci. USA* **99,** 16134–16137 (2002).
76. Demetrius, L. Directionality theory and the evolution of body size. *Proc. Biol. Sci.* **267,** 2385–2391 (2000).
77. Fay, J. C. & Wu, C. I. Sequence divergence, functional constraint, and selection in protein evolution. *Annu. Rev. Genomics Hum. Genet.* **4,** 213–235 (2003).
78. Brunet, J. P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA* **101,** 4164–4169 (2004).
79. Mootha, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genet.* **34,** 267–273 (2003).
80. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102,** 15545–15550 (2005).
81. Dorus, S. *et al.* Accelerated evolution of nervous system genes in the origin of *Homo sapiens*. *Cell* **119,** 1027–1040 (2004).
82. Saetre, P. *et al.* From wild wolf to domestic dog: gene expression changes in the brain. *Brain Res. Mol. Brain Res.* **126,** 198–206 (2004).
83. Wyckoff, G. J., Wang, W. & Wu, C. I. Rapid evolution of male reproductive genes in the descent of man. *Nature* **403,** 304–309 (2000).
84. Birkhead, T. R. & Pizzari, T. Postcopulatory sexual selection. *Nature Rev. Genet.* **3,** 262–273 (2002).
85. Dorus, S., Evans, P. D., Wyckoff, G. J., Choi, S. S. & Lahn, B. T. Rate of molecular evolution of the seminal protein gene *SEMG2* correlates with levels of female promiscuity. *Nature Genet.* **36,** 1326–1329 (2004).
86. Ruiz-Pesini, E. *et al.* Correlation of sperm motility with mitochondrial enzymatic activities. *Clin. Chem.* **44,** 1616–1620 (1998).
87. Zeh, J. A. & Zeh, D. W. Maternal inheritance, sexual conflict and the maladapted male. *Trends Genet.* **21,** 281–286 (2005).
88. Grossman, L. I., Wildman, D. E., Schmidt, T. R. & Goodman, M. Accelerated evolution of the electron transport chain in anthropoid primates. *Trends Genet.* **20,** 578–585 (2004).
89. Ostrander, E. A. & Kruglyak, L. Unleashing the canine genome. *Genome Res.* **10,** 1271–1274 (2000).
90. Sutter, N. B. *et al.* Extensive and breed-specific linkage disequilibrium in *Canis familiaris*. *Genome Res.* **12,** 2388–2396 (2004).
91. Parker, H. G. *et al.* Genetic structure of the purebred domestic dog. *Science* **304,** 1160–1164 (2004).
92. Bardeleben, C., Moore, R. L. & Wayne, R. K. A molecular phylogeny of the Canidae based on six nuclear loci. *Mol. Phylogenet. Evol.* **37,** 815–831 (2005).
93. Fogel, B. *The Encyclopedia of the Dog* (D.K. Publishing, New York, 1995).
94. Wilcox, B. & Walkowicz, C. *The Atlas of Dog Breeds of the World* (T.H.F. Publications, Neptune City, New York, 1995).
95. Frazer, K. A. *et al.* Segmental phylogenetic relationships of inbred mouse strains revealed by fine-scale analysis of sequence variation across 4.6 mb of mouse genome. *Genome Res.* **14,** 1493–1500 (2004).
96. Hudson, R. R. in *Oxford Surveys in Evolutionary Biology* Vol. 7 (eds Futuyma, D. & Antonovics, J.) 1–44 (Oxford Univ. Press, Oxford, 1990).
97. Vila, C., Seddon, J. & Ellegren, H. Genes of domestic mammals augmented by backcrossing with wild ancestors. *Trends Genet.* **21,** 214–218 (2005).

© 2005 Nature Publishing Group

98. Leonard, J. A. *et al.* Ancient DNA evidence for Old World origin of New World dogs. *Science* **298,** 1613–1616 (2002).
99. Kajiwara, N. & Japanese Kennel Club in *Akita* (eds Kariyabu, T. & Kaluzniacki, S.) 1–103 (Japan Kennel Club, Tokyo, 1998).
100. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296,** 2225–2229 (2002).
101. Werner, P., Raducha, M. G., Prociuk, U., Henthorn, P. S. & Patterson, D. F. Physical and linkage mapping of human chromosome 17 loci to dog chromosomes 9 and 5. *Genomics* **42,** 74–82 (1997).
102. Todhunter, R. J. *et al.* Power of a Labrador Retriever-Greyhound pedigree for linkage analysis of hip dysplasia and osteoarthritis. *Am. J. Vet. Res.* **64,** 418–424 (2003).
103. Sidjanin, D. J. *et al.* Canine *CNGB3* mutations establish cone degeneration as orthologous to the human achromatopsia locus *ACHM3*. *Hum. Mol. Genet.* **11,** 1823–1833 (2002).
104. Lou, X. Y. *et al.* The extent and distribution of linkage disequilibrium in a multi-hierarchic outbred canine pedigree. *Mamm. Genome* **14,** 555–564 (2003).
105. Hyun, C. *et al.* Prospects for whole genome linkage disequilibrium mapping in domestic dog breeds. *Mamm. Genome* **14,** 640–649 (2003).
106. Cardon, L. R. & Abecasis, G. R. Using haplotype blocks to map human complex trait loci. *Trends Genet.* **19,** 135–140 (2003).
107. Tsui, C. *et al.* Single nucleotide polymorphisms (SNPs) that map to gaps in the human SNP map. *Nucleic Acids Res.* **31,** 4910–4916 (2003).
108. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120,** 15–20 (2005).
109. Syvanen, A. C. Toward genome-wide SNP genotyping. *Nature Genet.* **37** (suppl.), S5–10 (2005).
110. Ma, B., Tromp, J. & Li, M. PatternHunter: faster and more sensitive homology search. *Bioinformatics* **18,** 440–445 (2002).
111. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13,** 103–107 (2003).
112. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14,** 708–715 (2004).
113. Smit, A. F. A. & Green, P. RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html).
114. Yang, Z., Goldman, N. & Friday, A. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* **11,** 316–324 (1994).
115. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11,** 1725–1729 (2001).
116. Viterbi, A. J. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Trans. Inform. Process.* **13,** 260–269 (1967).
117. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21,** 263–265 (2005).
118. The International HapMap Consortium. The International HapMap Project. *Nature* **426,** 789–796 (2003).
119. Macdonald, D. W. & Sillero-Zubiri, C. in *Biology and Conservation of Canids* (eds Macdonald, D. W. & Sillero-Zubiri, C.) 1–30 (Oxford Univ. Press, Oxford, 2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are indebted to the canine research community, and in particular D. Patterson, G. Acland and K. G. Lark, whose vision and research convinced the NIH of the importance of generating a canine genome sequence. We also thank all those who shared insights at the Dog Genome Community meetings, including G. Acland, G. D. Aguirre, M. Binns, U. Giger, P. Henthorn, F. Lingaas, K. Murphy and P. Werner. We thank our many colleagues (G. Acland, G. D. Aguirre, C. Andre, N. Fretwell, G. Johnson, K. G. Lark and J. Modiano), as well as the dog owners and breeders who provided us with samples. We thank colleagues at the UCSC browser for providing data (such as BLASTZ alignments), A. Smit for providing the RepeatMasker annotations used in our analyses and N. Manoukis for providing Unix machines for the phylogenetic analyses. Finally, we thank L. Gaffney and K. Siang Toh for editorial and graphical assistance. The genome sequence and analysis was supported in part by the National Human Genome Research Institute. The radiation hybrid map was supported in part by the Canine Health Foundation. Sample collection was supported in part by the Intramural Research Program of the National Human Genome Research Institute and the Canine Health Foundation.

Author Information The draft genome sequence has been deposited in public databases under NCBI accession codes AAEX01000000 (CanFam1.0) and AAEX02000000 (CanFam2.0). SNPs have been deposited in the dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/). Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to K.L.T. (kersli@broad.mit.edu) or E.S.L. (lander@broad.mit.edu).

Broad Sequencing Platform members Jennifer Baldwin[1], Adal Abebe[1], Amr Abouelleil[1], Lynne Aftuck[1], Mostafa Ait-zahra[1], Tyler Aldredge[1], Nicole Allen[1], Peter An[1], Scott Anderson[1], Claudel Antoine[1], Harindra Arachchi[1], Ali Aslam[1], Laura Ayotte[1], Pasang Bachantsang[1], Andrew Barry[1], Tashi Bayul[1], Mostafa Benamara[1], Aaron Berlin[1], Daniel Bessette[1], Berta Blitshteyn[1], Toby Bloom[1], Jason Blye[1], Leonid Boguslavskiy[1], Claude Bonnet[1], Boris Boukhgalter[1], Adam Brown[1], Patrick Cahill[1], Nadia Calixte[1], Jody Camarata[1], Yama Cheshatsang[1], Jeffrey Chu[1], Mieke Citroen[1], Alville Collymore[1], Patrick Cooke[1], Tenzin Dawoe[1], Riza Daza[1], Karin Decktor[1], Stuart DeGray[1], Norbu Dhargay[1], Kimberly Dooley[1], Kathleen Dooley[1], Passang Dorje[1], Kunsang Dorjee[1], Lester Dorris[1], Noah Duffey[1], Alan Dupes[1], Osebhajajeme Egbiremolen[1], Richard Elong[1], Jill Falk[1], Abderrahim Farina[1], Susan Faro[1], Diallo Ferguson[1], Patricia Ferreira[1], Sheila Fisher[1], Mike FitzGerald[1], Karen Foley[1], Chelsea Foley[1], Alicia Franke[1], Dennis Friedrich[1], Diane Gage[1], Manuel Garber[1], Gary Gearin[1], Georgia Giannoukos[1], Tina Goode[1], Audra Goyette[1], Joseph Graham[1], Edward Grandbois[1], Kunsang Gyaltsen[1], Nabil Hafez[1], Daniel Hagopian[1], Birhane Hagos[1], Jennifer Hall[1], Claire Healy[1], Ryan Hegarty[1], Tracey Honan[1], Andrea Horn[1], Nathan Houde[1], Leanne Hughes[1], Leigh Hunnicutt[1], M. Husby[1], Benjamin Jester[1], Charlien Jones[1], Asha Kamat[1], Ben Kanga[1], Cristyn Kells[1], Dmitry Khazanovich[1], Alix Chinh Kieu[1], Peter Kisner[1], Mayank Kumar[1], Krista Lance[1], Thomas Landers[1], Marcia Lara[1], William Lee[1], Jean-Pierre Leger[1], Niall Lennon[1], Lisa Leuper[1], Sarah LeVine[1], Jinlei Liu[1], Xiaohong Liu[1], Yeshi Lokyitsang[1], Tashi Lokyitsang[1], Annie Lui[1], Jan Macdonald[1], John Major[1], Richard Marabella[1], Kebede Maru[1], Charles Matthews[1], Susan McDonough[1], Teena Mehta[1], James Meldrim[1], Alexandre Melnikov[1], Louis Meneus[1], Atanas Mihalev[1], Tanya Mihova[1], Karen Miller[1], Rachel Mittelman[1], Valentine Mlenga[1], Leonidas Mulrain[1], Glen Munson[1], Adam Navidi[1], Jerome Naylor[1], Tuyen Nguyen[1], Nga Nguyen[1], Cindy Nguyen[1], Thu Nguyen[1], Robert Nicol[1], Nyima Norbu[1], Choe Norbu[1], Nathaniel Novod[1], Tenchoe Nyima[1], Peter Olandt[1], Barry O'Neill[1], Keith O'Neill[1], Sahal Osman[1], Lucien Oyono[1], Christopher Patti[1], Danielle Perrin[1], Pema Phunkhang[1], Fritz Pierre[1], Margaret Priest[1], Anthony Rachupka[1], Sujaa Raghuraman[1], Rayale Rameau[1], Verneda Ray[1], Christina Raymond[1], Filip Rege[1], Cecil Rise[1], Julie Rogers[1], Peter Rogov[1], Julie Sahalie[1], Sampath Settipalli[1], Theodore Sharpe[1], Terrance Shea[1], Mechele Sheehan[1], Ngawang Sherpa[1], Jianying Shi[1], Diana Shih[1], Jessie Sloan[1], Cherylyn Smith[1], Todd Sparrow[1], John Stalker[1], Nicole Stange-Thomann[1], Sharon Stavropoulos[1], Catherine Stone[1], Sabrina Stone[1], Sean Sykes[1], Pierre Tchuinga[1], Pema Tenzing[1], Senait Tesfaye[1], Dawa Thoulutsang[1], Yama Thoulutsang[1], Kerri Topham[1], Ira Topping[1], Tsamla Tsamla[1], Helen Vassiliev[1], Vijay Venkataraman[1], Andy Vo[1], Tsering Wangchuk[1], Tsering Wangdi[1], Michael Weiand[1], Jane Wilkinson[1], Adam Wilson[1], Shailendra Yadav[1], Shuli Yang[1], Xiaoping Yang[1], Geneva Young[1], Qing Yu[1], Joanne Zainoun[1], Lisa Zembek[1] & Andrew Zimmer[1]

Affiliation for participants: [1]Broad Institute of Harvard and MIT, 320 Charles Street, Cambridge, Massachusetts 02141, USA.

© 2005 Nature Publishing Group

Network Motif Discovery Using Subgraph Enumeration and Symmetry-Breaking

Joshua A. Grochow and Manolis Kellis

Computer Science and AI Laboratory, M.I.T.
Broad Institute of M.I.T. and Harvard
joshuag@cs.uchicago.edu, manoli@mit.edu

Abstract. The study of biological networks and network motifs can yield significant new insights into systems biology. Previous methods of discovering network motifs – network-centric subgraph enumeration and sampling – have been limited to motifs of 6 to 8 nodes, revealing only the smallest network components. New methods are necessary to identify larger network sub-structures and functional motifs.

Here we present a novel algorithm for discovering large network motifs that achieves these goals, based on a novel symmetry-breaking technique, which eliminates repeated isomorphism testing, leading to an exponential speed-up over previous methods. This technique is made possible by reversing the traditional network-based search at the heart of the algorithm to a motif-based search, which also eliminates the need to store all motifs of a given size and enables parallelization and scaling. Additionally, our method enables us to study the clustering properties of discovered motifs, revealing even larger network elements.

We apply this algorithm to the protein-protein interaction network and transcription regulatory network of *S. cerevisiae*, and discover several large network motifs, which were previously inaccessible to existing methods, including a 29-node cluster of 15-node motifs corresponding to the key transcription machinery of *S. cerevisiae*.

1 Introduction

1.1 Network Motifs

In the past decade, new technologies have enabled the observation and study of networks of thousands and millions of nodes, such as social networks, computer networks, and, notably, *biological networks*, including protein-protein interaction networks [4,5,6], genetic regulatory networks [12,18], and metabolic networks [7]. In order to extract meaningful information from these vast and sometimes noisy datasets, it is necessary to develop methods of computational analysis that are both efficient and robust to errors in the underlying data.

Network motifs – patterns of connectivity that occur significantly more frequently than expected – were introduced by Milo *et al.* [18] and provide one such robust property of biological networks. Network motifs also provide an important tool for understanding the modularity and the large-scale structure of networks

T. Speed and H. Huang (Eds.): RECOMB 2007, LNBI 4453, pp. 92–106, 2007.
© Springer-Verlag Berlin Heidelberg 2007

[8,13,20,25]. The importance of network motifs as information-processing modules has been modeled theoretically [12,21] and verified experimentally [8,13,20,25]. Network motifs also have numerous other applications: they have been used to classify networks into "superfamilies" [17], they have been used in combination with machine learning techniques to determine the most appropriate network model for a given real-world network [16], and they have been used to determine which properties to use in parsimony models of phylogeny [19].

Unfortunately, all of these applications are hampered by the limited size of motifs discoverable by current methods. Exact counting methods have only been reported to find motifs up to 4 nodes [18] and motif generalizations up to 6 nodes [10]. Subgraph sampling methods have found motifs up to 7 [9] and 8 nodes [1,16]. The statistical measures developed by Ziv *et al.* [26] are an important step towards larger network structures, but unfortunately lack a one-to-one correspondence with subgraphs, making them potentially difficult to interpret. Motif generalizations [10] are another important step towards these goals, although current methods are still limited to finding motif generalizations of only 6 nodes.

This current size limitation leaves many fundamental questions unanswered, and significant additional insight could be gained by exploring larger subgraphs and finding larger motifs. [1,10]. We should not expect *a priori* that the building blocks of complex networks are as small as 4 nodes, or that the largest significant structures and pathways contain only 8 nodes. What are the fundamental building blocks? How do they combine to form larger structures? [1,10] Do networks which share the same building blocks also share the same combinations of these blocks? [10] How can larger structures be used to distinguish between networks of different types, or between proposed models for a given network? [1]

In this paper, we present a new approach for discovering network motifs. The heart of our algorithm exhaustively assesses the significance of **a single query subgraph** as a potential motif. This can then be applied to all subgraphs of a given size to emulate the behavior of previous exhaustive algorithms, but with an exponential speed-up due to a **novel symmetry-breaking technique** (which is not feasible with previous methods). The symmetry-breaking technique also allows us to write instances of a subgraph to disk as they are found, **further eliminating limitations due to memory usage**. We are thus able to find motifs of up to 15 nodes, to find all instances of subgraphs of 31 nodes, and potentially even larger subgraphs. Although this work is motivated by biological networks, and this paper focuses on the protein-protein interaction (PPI) network and the transcription network of *S. cerevisiae*, our methods are applicable to any network – directed or undirected – and thus to many different fields, even outside the realm of biology.

In this section, we review previous work and give an overview of our algorithm, outlining several novel techniques which apply both to our approach and to previous approaches. In Sec. 2 we present our algorithm in detail. In Sec. 3 we present benchmarks comparing our approach to previous approaches. Additionally, we present data as to the effectiveness of the resulting improvements as

applied to both the transcription and PPI networks of *S. cerevisiae*. In Sec. 4, we present some of the larger subgraphs we have discovered. Finally, in Sec. 5 we discuss the significance of these contributions for the understanding of networks in general.

1.2 Limitations of Network-Centric Approaches for Motif Discovery

Two basic techniques have been proposed for identifying network motifs: exact counting [18] and subgraph sampling [1,9,16]. These methods attempt to determine the significance of all or many subgraphs of a given size by comparing their frequency in a given network to their frequency in a random ensemble of networks with similar properties to the original. To determine which subgraphs are motifs, subgraph sampling [1] is an effective and efficient approach, and has been used to evaluate the significance of larger subgraphs than can be evaluated by the exact counting method.

Most methods for finding DNA *sequence motifs* scan or sample a sequence pattern from a genome. Similarly, previous techniques for finding network motifs scan or sample subgraphs from a network, and count the number of occurrences of each subgraph encountered. (This process is then repeated for each network in a random ensemble resembling the initial network, and the counts are compared.) For the discovery of DNA sequence motifs, this general methodology is very efficient, because sequence motifs can be efficiently hashed based on their content. Thus a single linear scan of the genome suffices to exhaustively count all possible substrings of a given size, *regardless of the size of the substrings.*

In contrast, for the discovery of network motifs, enumerating all subgraphs of a given size is in general *exponential in the number of nodes of the subgraphs.* Additionally, there is no known efficient algorithm that correctly identifies two graphs as isomorphic or not. (The *graph isomorphism problem* is not known to be either in P or to be NP-complete.) This intrinsic difference in complexity between discovering *sequence* motifs and discovering *network* motifs makes traditional network-scanning methodologies inefficient for network motif discovery.

1.3 Distinguishing Features of the New Algorithm

To avoid these limitations of the traditional network-centric approaches, we have taken a motif-centric approach which has several attractive features, outlined here. Features 1-3 are specific to motif-centric methods, while features 4 and 5 can also benefit traditional network-centric methods.

(1) Searching for a single query graph. To avoid the increased complexity of subgraph enumeration (in the absence of an appropriate hashing scheme) our algorithm works by exhaustively searching for the instances of a *single query graph* in a network. (To find all motifs of a given size we couple this search with subgraph enumeration, using McKay's `geng` and `directg` tools [15]). Even though the *subgraph isomorphism problem* – finding a given graph as a subgraph of a larger network – is known to be NP-complete, several algorithmic improvements

enable this search to be carried out effectively in practice, even for subgraphs up to 31 nodes (and potentially even more).

(2) Mapping instead of enumerating. Rather than enumerating all connected subgraphs of a given size and testing to see whether each is isomorphic to the query graph, our algorithm attempts to map the query graph onto the network in all possible ways. We developed this technique for subgraph isomorphism independently, but subsequently identified a prior use [23].

(3) Taking advantage of subgraph symmetries. We introduce a technique that *avoids spending time finding a subgraph more than once* due to its symmetries. This technique improves the speed of our method by a factor exponential in the size of the query subgraph (Table 1). Moreover, since each instance is discovered exactly once, our algorithm can write instances to disk as they are found, greatly improving memory usage.

(4) Improved isomorphism testing. Our isomorphism test takes into account the degree of each node, and the degrees of each node's neighbors, leading to marked improvements over current motif-finding algorithms, which use exhaustive graph isomorphism tests.

(5) Subgraph hashing. When examining all subgraphs of a given size we hash the graphs based on their degree sequences, which leads to a significant improvement in the number of isomorphism tests needed. In a *directed* network, we group the query graphs based on their *undirected* isomorphism types, find all instances, and then go back to the directed network and divide these instances into their directed isomorphism types.

2 Description of the Algorithm

For clarity, we first present the basic mapping algorithm for subgraph isomorphism, without taking into account the symmetries of the query graph. In Sec. 2.2 we incorporate our symmetry-breaking technique into the algorithm. In the pseudo-code, we identify statements used solely for symmetry-breaking by enclosing them in square brackets. Finally, In Sec. 2.3 we incorporate our technique into two new methods of finding motifs.

Throughout this section, G will denote the network being searched and H will denote the query subgraph. We say that a node g of G can *support* a node h of H if we cannot rule out a subgraph isomorphism from H into G which maps h to g based on the degrees of h and g and the degrees of their neighbors. (Other constraints could also be used here, but these two proved effective and simple to implement.) This notion of support is used to exclude inconsistent candidate maps during the backtracking search.

2.1 Finding a Given Subgraph (Subgraph Isomorphism)

We start by presenting the algorithm without symmetry-breaking. Note that symmetry-breaking is not required for correctness of the algorithm.

FindSubgraphInstances(H,G):
Finds all instances of query graph H in network G

Start with an empty set of instances.
[Find Aut(H). Let H_E be the equivalence representatives of H.]
[Find symmetry-breaking conditions C for H given H_E and Aut(H).]
Order the nodes of G by increasing degree and
 then by increasing neighbor degree sequence.
For each node g of G
 For each node h of H [H_E] such that g can support h
  Let f be the partial map associating $f(h) = g$.
  Find all isomorphic extensions of f [up to symmetry]
   i.e. call IsomorphicExtensions(f,H,G[,$C(h)$]).
  Add the images of these maps to the set of all instances.
 Remove g from G.
Return the set of all instances.

FindSubgraphInstances includes the *images* of the maps in the list of instances, thus merging all maps which differ only by a symmetry of H. (Without symmetry-breaking, the algorithm spends additional time finding several distinct maps to a single subgraph.)

IsomorphicExtensions is a backtracking search to find all isomorphisms from H into G. As is standard in backtracking searches, the algorithm uses the most constrained neighbor to eliminate maps that cannot be isomorphisms: that is, the neighbor of the already-mapped nodes which is likely to have the fewest possible nodes it can be mapped to. First we select the nodes with the most already-mapped neighbors, and amongst those we select the nodes with the highest degree and largest neighbor degree sequence.

For each call to IsomorphicExtensions, f is extended by a single node. Each time an extension is made, the algorithm ensures that the newly mapped node is appropriately connected to the already-mapped nodes. Thus when IsomorphicExtensions returns a map, it is guaranteed to be an isomorphism.

We have effectively pushed the isomorphism testing of previous exhaustive methods into IsomorphicExtensions, which allows the isomorphism test to abort early. The ability to abort early when finding instances of a particular query graph presents significant savings over previous methods.

2.2 Exploiting Subgraph Symmetries to Speed Up the Search

Due to symmetries, a given subgraph of G may be mapped to a given query graph H multiple times. For example, the subgraph in Fig. 1 can be mapped to the same 6 nodes in 8 different ways. Thus a simple mapping-based search for a query graph will find each instance of the query graph as many times as the graph has symmetries. To avoid this, we compute and enforce several symmetry-breaking conditions, which ensure that there is a *unique* map from the query graph H to each instance of H in G, so that our search only spends time finding each instance once.

IsomorphicExtensions**(f,H,G[,C(h)]:**
Finds all isomorphic extensions of partial map $f : H \to G$ [satisfying C(h)]
Start with an empty list of isomorphisms.
Let D be the domain of f.
If $D = H$, return a list consisting solely of f. (Or write to disk.)
Let m be the most constrained neighbor of any $d \in D$
(constrained by degree, neighbors mapped, etc.)
For each neighbor n of $f(D)$
If there is a neighbor $d \in D$ of m such that n is *not* neighbors with $f(d)$,
or if there is a *non*-neighbor $d \in D$ of m such that n *is* neighbors with $f(d)$
[or if assigning $f(m) = n$ would violate a symmetry-breaking condition in $C(h)$],
then continue with the next n.
Otherwise, let $f' = f$ on D, and $f'(m) = n$.
Find all isomorphic extensions of f'.
Append these maps to the list of isomorphisms.
Return the list of isomorphisms.

The symmetries of a graph H are known as automorphisms (self-isomorphisms), and the group of automorphisms of H is denoted $\mathrm{Aut}(H)$. For a set A of automorphisms, two nodes are said to be "A-equivalent" if there is some automorphism in A which maps one to the other, or simply "equivalent" if $A = \mathrm{Aut}(H)$. We denote the A-equivalence of two nodes n_1, n_2 by $n_1 \sim_A n_2$. This equivalence relation partitions the nodes of H into equivalence classes. Since starting a map from two equivalent nodes is unnecessary and wasteful, FindSubgraphInstances uses a set consisting of one representative from each equivalence class of H.

The symmetry-breaking conditions are based on labellings of the nodes of H by integers, represented as maps from $H \to \mathbf{Z}$. Let $\ell : G \to \mathbf{Z}$ be a labelling of the nodes of G by *distinct* integers. Then each map $f : H \to G$ generates a labelling $L : H \to \mathbf{Z}$, given by $L(n) = \ell(f(n))$ for nodes $n \in H$. Thus, conditions on labellings of H translate into restraints on maps from H into G.

Given a set of conditions C, we say an automorphism α *preserves the conditions* C if, given a labelling L_1 of H which satisfies C, the corresponding labelling $L_2 : H \to \mathbf{Z}$ given by $L_2(n) = L_1(\alpha(n))$ also satisfies C. We are thus searching for conditions C such that the only automorphism which preserves C is the identity. This ensures there will be exactly one map from H onto each of its instances in G which satisfies the conditions.

To find these conditions, we pick an $\mathrm{Aut}(H)$-equivalence class $\{n_0, \ldots, n_k\}$ of nodes of H, and we impose the condition $L(n_0) < \mathrm{MIN}(L(n_1), \ldots, L(n_k))$. Any automorphism must send n_0 to one of the n_i, since these are all of the nodes equivalent to n_0. But to preserve this condition, an automorphism must send n_0 to itself. Then we continue recursively, replacing $\mathrm{Aut}(H)$ with the set A of automorphisms which send n_0 to itself. For example, see Fig. 1.

Because FindSubgraphInstances starts with a particular node, we can consider that node already fixed. (Note that the version of FindSubgraphInstances which uses symmetry-breaking only iterates over a set of equivalence class representatives, and not over all nodes of H.) Thus for each representative

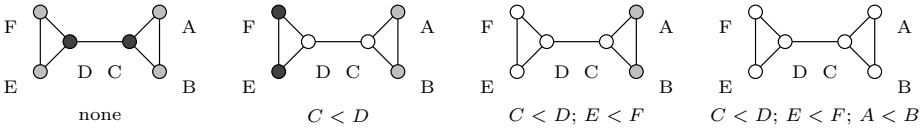


Fig. 1. Finding conditions that will break all the symmetries of a 6-node graph. White nodes are fixed by any automorphism preserving the indicated conditions, and other nodes are shaded according to their equivalence class under the automorphisms which preserve the indicated conditions.

used by FindSubgraphInstances, SymmetryConditions must generate a series of symmetry-breaking conditions which start by fixing that node.

To find the automorphisms of H, we use IsomorphicExtensions *without* symmetry-breaking, which returns an exhaustive list of all isomorphisms from H to itself. To find the automorphisms which fix a node or a set of nodes, the algorithm filters this list in a single pass.

Finding the automorphisms of a graph is thought to be computationally expensive[1], but in practice we have found this is far from the bottleneck in motif-finding algorithms. We were able to *exhaustively* find the automorphisms of all 11,117 8-node undirected graphs in under 30 seconds on a standard laptop, and McKay's tools [14] can find all the automorphisms of very large graphs very rapidly (e.g. some graphs with thousands of nodes and millions of automorphisms, in less than one second on a standard laptop).

SymmetryConditions:
Finds symmetry-breaking conditions for H given H_E, $\mathbf{Aut}(H)$

Let M be an empty map from equivalence representatives to sets of conditions.
For each $n \in H_E$
 Let C be an empty set of conditions.
 $n' \leftarrow n$, and $A \leftarrow \mathrm{Aut}(H)$.
 Do until $|A| = 1$:
  Add "$\textsc{label}(n') < \textsc{min}\{\textsc{label}(m) | m \sim_A n' \text{ and } m \neq n'\}$" to C.
  $A \leftarrow \{f \in A | f(n') = n'\}$.
  Find the largest A-equivalence class E.
  Pick $n' \in E$ arbitrarily.
 Let $M(n) = C$.
Return M.

2.3 Subgraph Evaluation and Network Motif Discovery

To find network motifs we enumerate candidate subgraphs H (exhaustively or by sampling), and evaluate each candidate based on its instances.

[1] Finding graph automorphisms is at least as hard as determining if two graphs are isomorphic. Like the graph isomorphism problem, the graph automorphism problem is not known to be either in P or to be NP-complete.

Evaluating candidate subgraphs. We find all instances of a query graph H in the network G, as well as in a random ensemble of networks with the same degree distribution and same distribution of 3-node subgraphs as G.[2] We evaluate the overrepresentation of the query graph H based on the z-score of its abundance in G against the distribution of its abundance in the random ensemble, as in [18,21].

Exhaustive subgraph enumeration. Our method can be used to find all instances of subgraphs of a given size, similar to previous exhaustive methods. To do this, we generate all non-isomorphic graphs of a particular size using McKay's `geng` and `directg` tools [15]. Then for each graph, we evaluate its significance as above.

Subgraph sampling. Our method can also be used in combination with subgraph sampling. We sample connected subgraphs (usually relatively large, compared to previous network motifs: 10, 15, or 20 nodes) by picking a node at random, and taking a random walk until we have as many nodes as desired [16]. Then we assess the significance of this subgraph as above.

Sampling subgraphs to find anti-motifs. Some studies have also considered *anti*-motifs: subgraphs which are significantly *under*represented compared to randomized versions of the network. To use a sampling method to find anti-motifs, it might be more fruitful to sample initial subgraphs from the random ensemble rather than the network being studied. Anti-motifs will be more prevalent in the ensemble than in the target network, and thus are more likely to be discovered by sampling from the ensemble.

3 Results and Evaluation

We applied our algorithm to the PPI network (1379 nodes, 2493 edges) [4] and transcription network (685 nodes, 1052 edges) [2] of *S. cerevisiae* and compared its performance to previous methods of motif disccovery.

Comparison with previous methods: time. We compare the time requirements of our method to those of Milo *et al.* [18] (Fig. 2). We make this comparison on the undirected PPI network of *S. cerevisiae* [4], by exhaustively counting subgraphs up to 7 nodes.

We implemented both our algorithm and two versions of the Milo *et al.* algorithm [18]: both as originally presented [18], and also by additionally hashing subgraphs by their degree sequence (Sec. 1.3). Fig. 2 shows that our algorithm provides an *exponential* improvement in time, even compared to the modified version of the previous algorithm [18].

[2] Although Shen-Orr *et al.*[21] use a model in which the distribution of $(n-1)$-node subgraphs is preserved when looking for n-node motifs, they only applied this to the case $n = 4$, and we have found it computationally infeasible to preserve this distribution for $n > 4$. Nonetheless, we have found it fruitful to preserve the distribution of 3-node subgraphs, regardless of n.

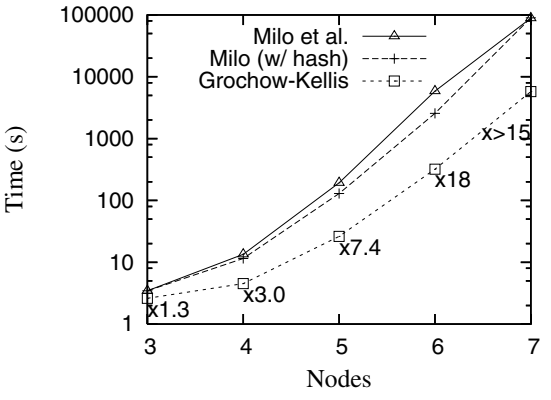


Fig. 2. The runtimes of the original algorithm of Milo *et al.* [18], an improved version of their algorithm, and our new algorithm, as applied to the undirected PPI network of *S. cerevisiae* [4]. The speed-up from the original algorithm of Milo *et al.* [18] to our algorithm is indicated. (Note: the values for 7 nodes for the two variants of Milo *et al.*'s algorithm are underestimates: the program ran out of memory before finishing.)

Comparison with previous methods: space. Our method gains an exponential memory advantage over previous exhaustive methods by not keeping a list of previously visited subgraphs. In the previous exact counting method [18], a list of the subgraphs encountered at each node is necessary in order to avoid duplication, even when the instances of the subgraphs are not desired as output. Thus the space required by the previous method is proportional to the number of subgraphs of a given size going through a given node, which can be exponential in the size of the subgraphs. Because our method does not need to keep such a list, its asymptotic memory requirements are determined by the maximum depth of recurion of IsomorphicExtensions, which is linear in the size of the query graph. Our method thus uses exponentially less space than previous exhaustive methods.

Disk usage. Furthermore, our algorithm uses no more memory to find a list of all instances than to simply count the instances. Since each instance is encountered exactly once, it can be written to disk and removed from active memory as soon as it is encountered, using effectively no additional memory.

Improvement due to symmetry-breaking. The main reason for these improvements is our novel symmetry-breaking technique. Symmetry-breaking ensures that each instance is discovered exactly once, so our method does not have to check a list of the subgraphs previously encountered at a node in order to avoid duplicate counting, while the previous method of exact counting does. Table 1 quantifies this contribution explicitly, showing that the average number of automorphisms of graphs weighted by their occurences in the PPI network and regulatory network of yeast – i.e. the savings gained by symmetry-breaking – appears to grow exponentially.

Table 1. The number of subgraphs encountered by our algorithm with and without symmetry-breaking (including multiple encounters for the version without symmetry-breaking). The improvement factor is exactly the average number of automorphisms of subgraphs of the associated size.

Nodes	Undirected PPI Network			Directed Regulatory Network		
	Total Subgraphs Searched	With Symmetry-Breaking	Improvement	Total Subgraphs Searched	With Symmetry-Breaking	Improvement
3	3.7×10^4	1.1×10^4	$\times3.13$	2.6×10^4	1.3×10^4	$\times2.02$
4	4.0×10^5	7.0×10^4	$\times5.77$	9.7×10^5	1.8×10^5	$\times5.41$
5	4.4×10^6	4.1×10^5	$\times10.9$	4.4×10^7	2.5×10^6	$\times18.0$
6	5.1×10^7	2.3×10^6	$\times22.2$	2.3×10^9	3.2×10^7	$\times73.3$
7	5.7×10^8	1.2×10^7	$\times46.3$	1.3×10^{11}	4.0×10^8	$\times334$
8	6.4×10^9	6.6×10^7	$\times96.2$	—	—	—

4 Discovered Motifs and Their Biological Significance

Discovered motifs. We exhaustively evaluated all candidate motifs and anti-motifs up to 7 nodes in the PPI network of *S. cerevisiae*[4] (1379 nodes, 2493 edges). We used a random ensemble of networks with the same degree distribution and the same distribution of 3-node subgraphs as the PPI network.[3] The most significant subgraphs tend to be motifs rather than anti-motifs: using a z-score cutoff of 4.0, only 3 of the 54 significant subgraphs of size at most 7 were anti-motifs. Two of the motifs were trees, and the most dense motif had 18 edges. Most of the significant graphs were of moderate density: the mean number of edges for 7-node motifs and anti-motifs is 11.49 ± 2.89.

Large discovered motifs. We also discovered larger motifs by first sampling connected subgraphs from the PPI network of *S. cerevisiae*, and then assessing their significance using our method. We sampled approximately 100 connected subgraphs of 15 and 20 nodes, and found several motifs. One such 15-node motif (Fig. 3) represents a common connectivity pattern found within the transcriptional machinery of *S. cerevisiae* (see discussion below).

Clustering of discovered motifs and larger network structures. We noted that almost all of the larger subgraphs we evaluated have large numbers of overlapping instances, which become apparent since our method reports all network instances of a discovered motif. To quantify this property, we developed a subgraph clustering score, based on the number of subgraph instances overlapping a given node, averaged over all nodes in any subgraph instance. We applied this score to evaluate the clustering properties of all discovered motifs, and we found that indeed some of the most abundant motifs show striking clustering properties.

[3] See footnote 2.

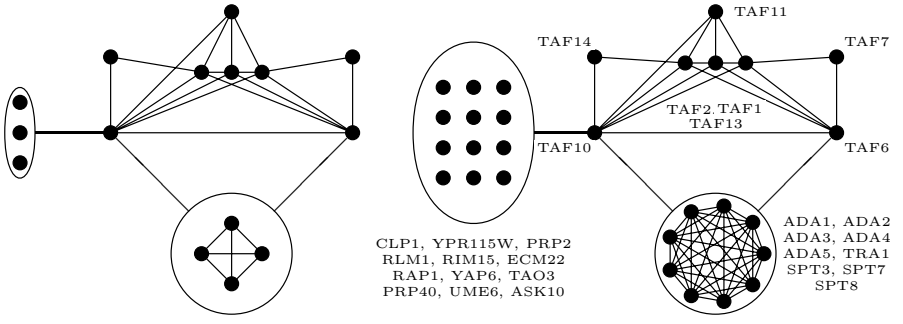


Fig. 3. A motif of 15 nodes and 34 edges (left). An edge from a group of nodes to a node n indicates that each node in the group is connected to n. This motif appears 27,720 times in the PPI network of *S. cerevisiae*[4], and does not appear at all in random ensembles which preserve the degree distribution and the distribution of 3-node subgraphs. All 27,720 instances are clustered into a total of 29 nodes (right), corresponding to the cellular transcription machinery.

The clustered instances frequently reveal important larger network structures. For example, the 15-node motif of Fig. 3 occurs 27,720 times in a single subnetwork of 29 nodes, part of the core transcription machinery of *S. cerevisiae*. This includes a complete 11-node graph (including the two central hubs) corresponding to the SAGA complex, and consisting almost entirely of chromatin modification and histone acetylation factors an 8-node core (shared by all instances of the 15-node motif) corresponding to the TFIID complex, and 12 attachments, which are known activators and suppressors of these two complexes [11]. Similarly, the subgraph of 20 nodes shown in Fig. 4 occurs 5,020 times in a total of 31 nodes, enriched in cell-cycle regulation.

The role of combinatorial effects. The extreme clustering properties of the most abundant motifs appear to result from combinatorial connectivity patterns prevalent in larger network structures. For example, all 27,720 instances of the 15-node motif in Fig. 3 result by choosing 3 attachments from the left and 4 attachments from the the bottom of Fig. 3 ($\binom{12}{3}\binom{9}{4} = 27,720$), and similarly for the 5,020 instances of the 20-node subgraph in Fig. 4. Additionally, in the random ensemble, these combinatorially appearing structures occur either thousands of times, or not at all – they almost never occur just a few or a few hundred times.

Although motif clustering has previously been observed [3] and demonstrated analytically [24], previous motifs studied do not have enough nodes to exhibit the extreme combinatorial clustering we observed for large subgraphs (at least 15 nodes). The magnitude of this combinatorial clustering effect brings into question the current definition of network motif, when applied to larger structures. We propose that additional statistics, either alone or in combination, might be well-suited to identify larger meaningful network structures: our subgraph clustering score, the total number of nodes covered by all instances of the query graph, the

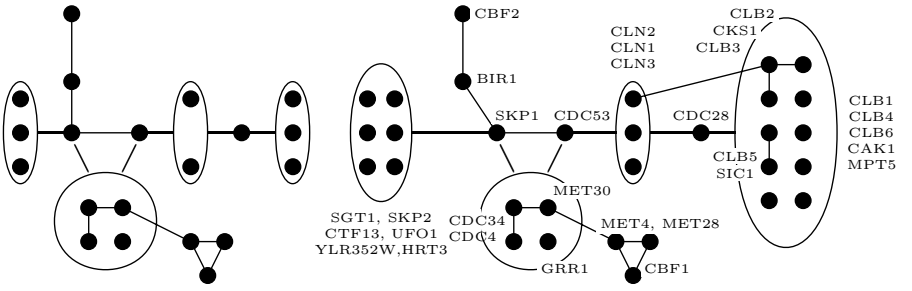


Fig. 4. A subgraph of 20 nodes and 27 edges (left). An edge from a group of nodes to a node n indicates that each node in the group is connected to n. This subgraph appears 5,020 times in the PPI network of *S. cerevisiae* [4]. All 5,020 instances are clustered into a total of 31 nodes (right), enriched in cell-cycle regulation.

total number of edges, and the weighting of the number of nodes/edges based on the number of overlapping instances. All of these statistics can be easily calculated using our algorithm, since it finds and stores all motif instances, and these will be the subject of future studies.

5 Discussion

We presented a novel approach to the discovery of network motifs, based on a solution to the subgraph isomorphism problem that uses a new symmetry-breaking technique, an improved isomorphism test, and hashing based on degree sequences. Several of the techniques presented in our algorithm can also be used in previous algorithms, and lead to significant improvements.

We implemented our algorithm and used it to find significant structures of 15 and 20 nodes in the PPI network and the regulatory network of *S. cerevisiae*, where previous methods had been limited to motifs of 6 and 8 nodes. Using our approach to motif-finding, we re-discovered the cellular transcription machinery – as a 29-node cluster of 15-node motifs – based solely on the structure of the protein interaction network.

Previous methods of motif discovery were network-centric, and could therefore not take advantage of subgraph symmetries. By using a motif-centric algorithm instead, we are able to use symmetry-breaking to get an exponential improvement.

5.1 Applications and Advantages of the New Method

(1) Finding larger motifs. Our improvements have enabled the exhaustive discovery of motifs up to 7 nodes. To find even larger motifs, we sample a connected subgraph as in [16], and then find *all* its instances and assess its significance using our method. This technique has enabled us to find motifs up to 15 nodes and examine subgraphs up to 31 nodes.

(2) Querying a particular subgraph. Our method can be used to query whether a particular subgraph is significant, whereas previous methods can only do this

by examining all subgraphs of the same size, which quickly becomes prohibitive for even moderate sizes. This technique could be used to explore *in silico* the prevalence of a subgraph of interest, identified experimentally (e.g. known pathways), computationally (e.g. motif generalizations [10]), or genetically.

(3) Exploring motif clustering. Because our algorithm finds all instances of a given subgraph, it can be used to explore how these instances cluster together to form larger structures. For example, after finding a 15-node motif, we were able to determine that all of its 27,720 instances clustered in 29 nodes (Fig. 3).

(4) Time and space. Our method, applied to all subgraphs of a given size, takes exponentially less time than previous methods, even when we implement the previous method with our hashing scheme (Sec. 3). Additionally, there are essentially no space limitations on our method: since each instance is found exactly once due to our symmetry-breaking technique, it can be written to disk and removed from active memory as soon as it is found.

(5) Parallelization. Our method is more easily parallelizable than previous motif-finding methods, since each subgraph can be counted on a separate processor. We have found this attribute to be very useful, and we believe other researchers will as well, as cluster computing becomes commonplace in the computational biology community.

5.2 Clustering Properties of Large Subgraphs and Motifs

We revealed that larger subgraphs tend to cluster together *combinatorially* – that is, all instances share a significant core of nodes, and each instance represents a choice of attachments to these core nodes. This combinatorial clustering brings into question the relevance of the standard definition of network motif for large subgraphs of 15 nodes or more. We proposed several statistics which may be more appropriate in this domain.

Finally, we mention that the statistics of Ziv *et al.* [26] may not suffer from these combinatorial effects. The main drawback of these statistics is their lack of one-to-one correspondence with subgraphs. In combination with our algorithm, however, the large subgraphs encompassed by these statistics could be further explored, allowing for a clearer interpretation of the most significant statistics.

Moving forward, we expect the network motifs and methodology presented here will open a window into the large structures and global organization of biological and other networks.

Acknowledgements. The authors would like to thank Pouya Kheradpour, Mike Lin, Matt Rasmussen, Alex Stark, and Radek Szklarczyk (all at the M.I.T. Computer Science and AI Laboratory) for many useful and interesting discussions.

All algorithms were implemented in Java using the Java Universal Networks and Graphs (JUNG) framework [22]. Our software is available on request. McKay's `geng` and `directg` tools [15] were used to enumerate all graphs of a given size. Much of the computational work was carried out on the compute cluster at the Broad Institute of M.I.T. and Harvard.

This work was supported in part by startup funds from Professor Kellis.

References

1. K. Baskerville and M. Paczuski. Subgraph ensembles and motif discovery using a new heuristic for graph isomorphism, 2006. arxiv.org:q-bio/0606023.
2. M. C. Costanzo, M. E. Crawford, J. E. Hirschman, J. E. Kranz, P. Olsen, L. S. Robertson, M. S. Skrzypek, B. R. Braun, K. L. Hopkins, P. Kondu, C. Lengieza, J. E. Lew-Smith, M. Tillberg, and J. I. Garrels. Ypd(tm), pombepd(tm), and wormpd(tm): model organism volumes of the bioknowledge(tm) library, an integrated resource for protein information. *Nucleic Acids Res.*, 29:75–79, 2001.
3. R. Dobrin, Q. K. Beg, A.-L. Barabási, and Z. N. Oltvai. Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network. *BMC Bioinformatics*, 5:10, Jan 2004.
4. J.-D. J. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. M. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995):88–93, Jul 2004.
5. A. Jaimovich, G. Elidan, H. Margalit, and N. Friedman. Towards an integrated protein-protein interaction network: a relational markov network approach. *J. Comp. Bio.*, 13:145–164, 2006.
6. H. Jeong, S. Mason, A.-L. Barabási, and Z. N. Oltvai. Centrality and lethality of protein networks. *Nature*, 411, 2001.
7. H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407, 2000.
8. S. Kalir, J. McClure, K. Pabbaraju, C. Southward, M. Ronen, S. Leibler, M. G. Surette, and U. Alon. Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. *Science*, 292(5524):2080–2083, Jun 2001.
9. N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20(11):1746–1758, Jul 2004. Evaluation Studies.
10. N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Topological generalizations of network motifs. *Phys. Rev. E*, 70:031909, 2004.
11. T. I. Lee and R. A. Young. Transcription of eukaryotic protein-coding genes. *Annu. Rev. Genet.*, 34:77–137, 2000.
12. S. Mangan and U. Alon. Structure and function of the feed-forward loop network motif. *PNAS*, 100(21):11980–11985, Oct 2003.
13. S. Mangan, A. Zaslaver, and U. Alon. The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *J. Mol. Biol.*, 334(2):197–204, Nov 2003.
14. B. D. McKay. Practical graph isomorphism. In *Proceedings of the Tenth Manitoba Conference on Numerical Mathematics and Computing, Vol. I (Winnipeg, Man., 1980)*, volume 30, pages 45–87, 1981. http://cs.anu.edu.au/∼bdm/nauty/.
15. B. D. McKay. Isomorph-free exhaustive generation. *J. Algorithms*, 26:306–324, 1998.
16. M. Middendorf, E. Ziv, and Chris H. Wiggins. Inferring network mechanisms: the Drosophila melanogaster protein interaction network. *PNAS*, 102(9):3192–3197, Mar 2005.
17. R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, Mar 2004.

18. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, Oct 2002.
19. T. M. Przytycka. An important connection between network motifs and parsimony models. In *RECOMB 2006*, pages 321–335, 2006.
20. M. Ronen, R. Rosenberg, B. I. Shraiman, and U. Alon. Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc Natl Acad Sci U S A*, 99(16):10555–10560, Aug 2002.
21. S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics*, 31(1):64–68, May 2002.
22. JUNG Framework Development Team. Jung: The java universal network/graph framework, 2005.
23. J. R. Ullman. An algorithm for subgraph isomorphism. *J. Assoc. Comp. Mach.*, 23(1):31–42, Jan 1976.
24. A. Vazquez, R. Dobrin, D. Sergi, J.-P. Eckmann, Z. N. Oltvai, and A.-L. Barabasi. The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *PNAS*, 101(52):17940–17945, Dec 2004.
25. A. Zaslaver, A. E. Mayo, R. Rosenberg, P. Bashkin, H. Sberro, M. Tsalyuk, M. G. Surette, and U. Alon. Just-in-time transcription program in metabolic pathways. *Nature Genetics*, 36(5):486–491, May 2004.
26. E. Ziv, R. Koytcheff, M. Middendorf, and C. Wiggins. Systematic identification of statistically significant network measures. *Phys. Rev. E*, 71:016110, 2005.

Downloaded from www.genesdev.org on March 13, 2007

Whole-genome ChIP–chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the *Drosophila* embryo

Julia Zeitlinger,[1] Robert P. Zinzen,[2] Alexander Stark,[3,4] Manolis Kellis,[3,4] Hailan Zhang,[5] Richard A. Young,[1,6] and Michael Levine[2,7]

[1]Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA; [2]Department of Molecular Cell Biology, Center for Integrative Genomics, University of California, Berkeley, California 94720, USA; [3]Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts 02141, USA; [4]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; [5]Department of Medicine, Mount Sinai School of Medicine, New York, New York 10029, USA; [6]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Genetic studies have identified numerous sequence-specific transcription factors that control development, yet little is known about their in vivo distribution across animal genomes. We determined the genome-wide occupancy of the dorsoventral (DV) determinants Dorsal, Twist, and Snail in the *Drosophila* embryo using chromatin immunoprecipitation coupled with microarray analysis (ChIP–chip). The in vivo binding of these proteins correlate tightly with the limits of known enhancers. Our analysis predicts substantially more target genes than previous estimates, and includes Dpp signaling components and anteroposterior (AP) segmentation determinants. Thus, the ChIP–chip data uncover a much larger than expected regulatory network, which integrates diverse patterning processes during development.

Supplemental material is available at http://www.genesdev.org.

Received November 7, 2006; revised version accepted January 11, 2007.

Systematic genetic studies of embryonic development have revealed a key role for sequence-specific transcription factors in axis formation and cell differentiation (St Johnston and Nusslein-Volhard 1992). However, it has been difficult to unravel the detailed regulatory interactions through which these transcription factors regulate development. One of the best-studied gene regulatory networks controls the dorsoventral (DV) patterning of the *Drosophila* embryo (see Stathopoulos et al. 2002). This network is initiated by a nuclear concentration gradient of the NF-κB transcription factor Dorsal, which is established by differential activation of the Toll receptor in ventral and ventrolateral regions of precellular embryos (for review, see Moussian and Roth 2005). The ensuing Dorsal gradient leads to the differential regulation of at least 50 target genes across the DV axis. Approximately 40 of these genes encode either sequence-specific transcription factors or components of the Notch, EGF, FGF, and Dpp signaling pathways, which further specify diverse cell fates across the DV axis (e.g., see Stathopoulos et al. 2002). More than 100 functional interconnections among these regulatory factors drive the initial phases of DV patterning (Levine and Davidson 2005).

At the core of the DV patterning network are Dorsal and the products encoded by two of its earliest target genes, Twist and Snail. Twist functions as a basic helix–loop–helix (bHLH) activator, while Snail is a zinc-finger repressor. At least half of the tissue-specific enhancers that are regulated by different levels of Dorsal also contain binding sites for Twist and/or Snail (Stathopoulos et al. 2002; Markstein et al. 2004). Our current understanding of the DV patterning network is based on decades of genetic screens (e.g., see Nusslein-Volhard et al. 1980), investigations of genetic interactions (e.g., see Simpson 1983; Kosman et al. 1991; Leptin 1991), *lacZ* reporter assays (e.g., see Jiang et al. 1991), in vitro DNA-binding assays (e.g., see Huang et al. 1993), and more recently, bioinformatics methods (Markstein et al. 2002, 2004; Papatsenko and Levine 2005). This network has served as a model for understanding axis formation, tissue specification, and cell shape changes during gastrulation (Ip and Gridley 2002). Nonetheless, the genome-wide physical contacts through which the relevant transcription factors interact with the genome and regulate gene expression in vivo are essentially unknown.

Here we determine the genome-wide occupancy of Dorsal, Twist, and Snail in the *Drosophila* embryo using chromatin immunoprecipitation coupled with microarray analysis (ChIP–chip). These assays accurately identify known DV enhancers and further identify several hundred additional potential target enhancers, a dramatic increase over previous estimates. We present evidence that a significant fraction of these enhancers is functional, and suggest that extensive integration occurs between diverse patterning networks, including those controlling DV patterning and segmentation.

Results and Discussion

Known enhancers are bound by Dorsal, Twist, and Snail

ChIP–chip assays were performed with antibodies directed against Dorsal, Twist, or Snail on *Toll*^{10b} mutant embryos, aged 2–4 h. These embryos contain a constitutively activated form of the Toll receptor, which results in high levels of nuclear Dorsal protein and uniform expression of Twist and Snail throughout the embryo (e.g., see Stathopoulos et al. 2002). The high levels of Dorsal, Twist, and Snail cause all cells to form derivatives of the

[*Keywords*: Dorsoventral patterning; segmentation; gene regulatory networks; ChIP–chip]
[7]**Corresponding author.**
E-MAIL mlevine@berkeley.edu; FAX (510) 643-5785.
Article is online at http://www.genesdev.org/cgi/doi/10.1101/gad.1509607.

Downloaded from www.genesdev.org on March 13, 2007

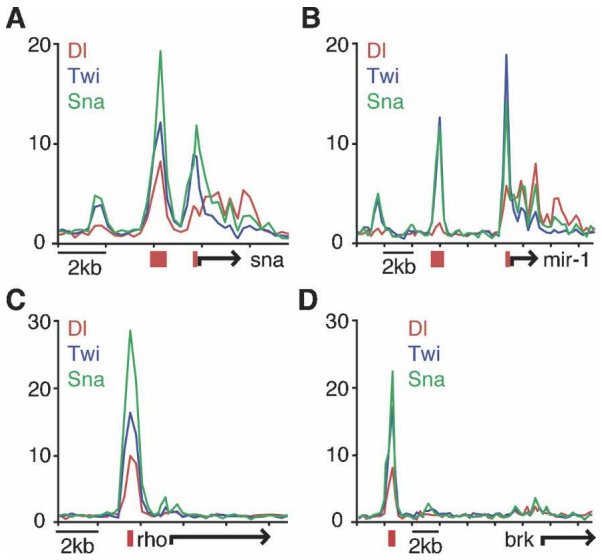


Figure 1. *Drosophila* ChIP–chip identifies known enhancers of Dorsal, Twist, and Snail. *Drosophila* embryos, aged 2–4 h, from Toll^{10b} mothers were used to perform ChIP using antibodies against Dorsal, Twist, and Snail. Most enhancers known to be regulated by these factors were successfully identified, including type 1 target genes *sna* (*A*), *mir-1* (*B*), and type 2 target genes *rho* (*C*) and *brk* (*D*). The graphs show unprocessed ChIP enrichment ratios (*Y*-axis) for Dorsal (red), Twist (blue), and Snail (green), across chromosomal regions (*X*-axis). Gene model (black arrow) and the known enhancer (red) are shown *below*.

mesoderm at the expense of neurogenic and dorsal ectoderm. Thus, these embryos represent a uniform cell type with respect to DV fate.

The whole-genome ChIP–chip experiments reveal several hundred strong binding clusters of Dorsal, Twist, and Snail with up to 40-fold ChIP enrichment, most of which span regions of ~1 kb in length. To identify the binding patterns of bona fide target enhancers of the Dorsal regulatory network, we first analyzed known enhancers. The 22 known enhancers fall into three classes—type 1, type 2, and type 3—based on which levels of nuclear Dorsal regulate their expression (Supplementary Table S1; see Biemar et al. 2006).

The 10 type 1 enhancers (associated with *twi*, *sna*, *miR-1*, *htl*, *hbr*, *mes3*, *CG12177*, *ady43A*, *tin*, and *Phm*) are activated by peak levels of Dorsal in the presumptive mesoderm, and are all constitutively activated in *Toll^{10B}* mutant embryos (e.g., see Stathopoulos et al. 2002; Biemar et al. 2006). The ChIP–chip experiments identify strong binding peaks (greater than fivefold enrichment) of Dorsal, Twist, and Snail (DTS) within five of the 10 enhancers (*twi*, *sna*, *miR-1*, *CG12177* and *Phm*) (Fig. 1A,B). Another three enhancers, those associated with *htl*, *tin*, and *ady43A*, show significant but lower (less than fivefold) binding peaks restricted to Twist and Snail (TS) binding. This observation is consistent with earlier studies indicating that these enhancers might be primarily activated by Twist (Papatsenko and Levine 2005). Hence, eight of the 10 known type 1 enhancers exhibit significant in vivo occupancy by Twist and Snail.

An even greater correspondence between known enhancers and in vivo occupancy is seen for the type 2 [*sim*, *E(spl)*, *vn*, *rho*, *vnd* and *brk*] and type 3 enhancers (*ths*, *sog*, *ind*, *dpp*, *zen* and *tld*), which are regulated by intermediate and low levels of the Dorsal gradient, respectively (Fig. 1C,D). All 12 enhancers are silenced in *Toll^{10B}* mutant embryos due to constitutive expression of the Snail repressor (e.g., see Biemar et al. 2006). Remarkably, every enhancer exhibits strong DTS or TS peaks with greater than fivefold enrichment in the ChIP–chip assays. Thus, ChIP–chip assays correctly identified 20 of the 22 known Dorsal target enhancers.

Genome-wide identification of putative target enhancers

Most known DV enhancers are associated with overlapping binding clusters of Dorsal, Twist, and Snail regardless of whether they mediate activation or repression. Moreover, 17 of the 20 binding clusters at known enhancers display greater than fivefold enrichment of Twist and/or Snail. Using these binding criteria, we identified 428 high-confidence DTS regions and 433 high-confidence TS regions across the genome (see also Supplemental Material).

To confirm these regions through independent evidence, we performed sequence analysis on these regions using the known consensus binding motifs of Dorsal, Twist, and Snail. As expected, the identified regions are highly enriched in all three binding motifs (Fig. 2). Moreover, a large fraction of the motifs is conserved across the 12 sequenced *Drosophila* species providing evidence that the discovered regions are functionally important (Fig. 2). Finally, when motifs that are enriched in these regions were identified de novo, the known binding motifs can be rediscovered (Supplemental Material). Hence, the regions we identified represent putative target gene enhancers of the DV network.

Identification of new DV enhancers

To show that newly identified regions indeed function as enhancers in vivo, we selected putative enhancers of pri-

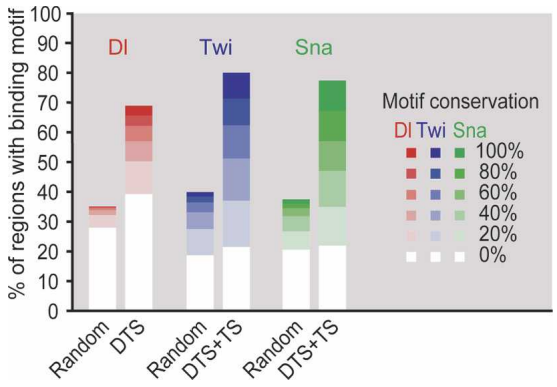


Figure 2. Enrichment of Dorsal-, Twist-, and Snail-binding motifs in bound regions. Sequences from DTS and TS regions with greater than fivefold ChIP enrichment were searched for the presence of the known Dorsal (red), Twist (blue), or Snail (red) motifs, and the evolutionary conservation of these motifs across the 12 sequenced *Drosophila* genomes was determined (as percent of branch length within the phylogenetic tree). The graph shows the fraction of regions at each conservation cutoff (0%, 20%, 40%, 60%, 80%, and 100%). As a control, the same analysis was performed with regions of identical length that were randomly distributed among intronic and intergenic regions. Using a χ^2 test, the motif enrichment and evolutionary conservation was highly significant for Dorsal ($p < 10^{-148}$), Twist ($p < 10^{-228}$), and Snail ($p < 10^{-306}$).

Downloaded from www.genesdev.org on March 13, 2007

mary DV genes; i.e., those genes that are expressed as localized stripes across the DV axis (Stathopoulos et al. 2002; Biemar et al. 2006). In addition to the 22 known DV enhancers, we identified 47 new putative enhancers, some of which appear to regulate the same gene (Supplementary Table S1). By attaching the genomic sequence to a *lacZ* reporter and expressing the construct in transgenic embryos, we show that seven of these enhancers are bona fide DV enhancers and that regulation by multiple enhancers occurs (Fig. 3; Supplemental Material).

The *wntD* gene is expressed in portions of the presumptive mesoderm where it mediates feedback inhibition of Toll signaling (Ganguly et al. 2005; Gordon et al. 2005). A cluster of DTS-binding peaks was identified in the 5′-flanking region, and the corresponding genomic DNA fragment mediates *lacZ* expression in the same region of the mesoderm as the endogenous gene (Fig. 3A). Similar results were obtained with the DTS-binding cluster located in the 5′-flanking region of *mes5/mdr49* (Supplemental Material).

The *vnd* locus contains a well-documented intronic enhancer that mediates expression in the neurogenic ectoderm and recapitulates the spatial and temporal expression pattern of the endogenous gene (Stathopoulos et al. 2002; Markstein et al. 2004). The ChIP–chip analysis detected this enhancer but also revealed two novel clusters further upstream (Fig. 3B). When tested for *lacZ* reporter activity, these novel genomic sequences directed *lacZ* expression in a pattern resembling that of the endogenous gene over different time periods: One directs early *vnd* expression in the presumptive ventral neurogenic ectoderm (vNE) (Fig. 3C) while the other directs later expression in the medial column (mc) (Fig. 3D) of the developing nervous system. All three enhancers contain evolutionarily conserved binding sites for Dorsal, Twist, and Snail, suggesting that the enhancers are not redundant but may function to fine-tune the *vnd* expression pattern. Overlapping enhancer activity was also observed for multiple miR-1 enhancers (Supplemental Material). Overall, as many as a third of all DV genes have multiple binding clusters, and thus might be subject to similar regulatory control.

Several of the occupied regions are associated with Dpp target genes expressed in the dorsal ectoderm. When the *tup* and *pnr* intronic sequences are tested in transgenic embryos, both fragments function as authentic enhancers and direct localized expression in the dorsal ectoderm, comparable to the endogenous *tup* and *pnr* expression patterns (Fig. 3E; Supplemental Material). These results suggest that the Dorsal patterning network directly regulates the expression of Dpp target genes (see below).

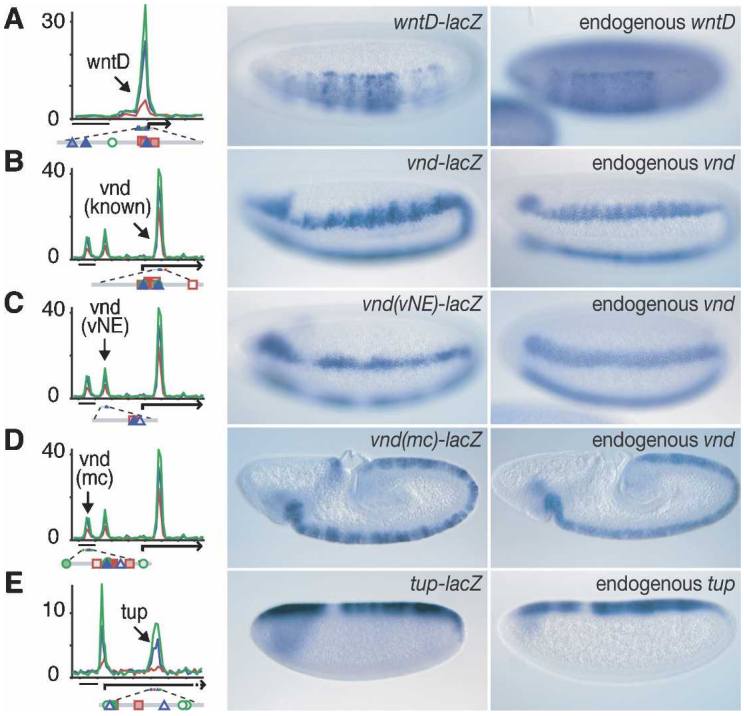


Figure 3. Identification of novel DV enhancers. ChIP–chip data identified enhancers for the DV genes *wntD* (*A*), *vnd* (*B–D*), and *tup* (*E*). The *left* column shows the ChIP enrichment ratios of Dorsal (red), Twist (blue), and Snail (green). Bar, 2 kb. The tested enhancer (arrow) is shown with transcription factor motifs *below* (Dorsal site, red square; Twist site, blue triangle; Snail site, green circle; shading corresponds to their conservation across the 12 sequenced *Drosophila* genomes). The pattern of enhancer-driven *lacZ* expression in transgenic embryos (*middle* column) resembles that of the respective endogenous genes (*right* column). Unexpectedly, in addition to the previously known *vnd* enhancer (*B*), two regulatory regions (*C*,*D*) were identified that drive *lacZ* expression in subsets of the endogenous *vnd* expression pattern: The region in *C* drives early expression in the vNE, while the region in *D* drives later expression in the medial column (mc). Embryos are oriented anterior to the left, dorsal is up.

Connections between the DV and the anterioposterior (AP) patterning networks

We noticed that many of the new DTS/TS clusters are associated with AP genes involved in segmentation (Supplementary Table S2). Although classical genetic studies argue that AP and DV patterning of the early embryo are controlled by separate maternal genetic programs (for review, see St Johnston and Nusslein-Volhard 1992), it is conceivable that the expression of AP target genes is modulated by the DV network. Indeed, DV modulation of segmentation gene expression has been observed previously (e.g., see Carroll et al. 1987; Liaw and Lengyel 1993; Rothe et al. 1994; Gao and Finkelstein 1998).

The gap gene *orthodenticle* (*otd*) is expressed in two stripes across the AP axis in the early embryo. The anterior stripe shows diminished expression on the ventral side (Fig. 4A). Previous studies identified a 5′ enhancer that recapitulates the normal expression pattern, including Dorsal-dependent suppression in ventral regions (Gao and Finkelstein 1998). ChIP–chip identified a strong DTS cluster within the limits of this enhancer (Fig. 4A). A similar DV bias in the expression pattern was found for the gap gene *tailless* (*tll*) (Fig. 4B) and the pair-rule genes *runt* and *hairy* (Fig. 4C,D). In each case, the regions identified by ChIP–chip overlap or map close to known regulatory regions (Pankratz et al. 1990; Butler et al. 1992; Liaw and Lengyel 1993) and contain several Dorsal-binding motifs (Fig. 4B–D).

At the gap gene *knirps,* a DTS-binding cluster was found in a region distinct from the known Bicoid-dependent enhancer (Fig. 4E; Pelegri and Lehmann 1994). This newly identified genomic

Downloaded from www.genesdev.org on March 13, 2007

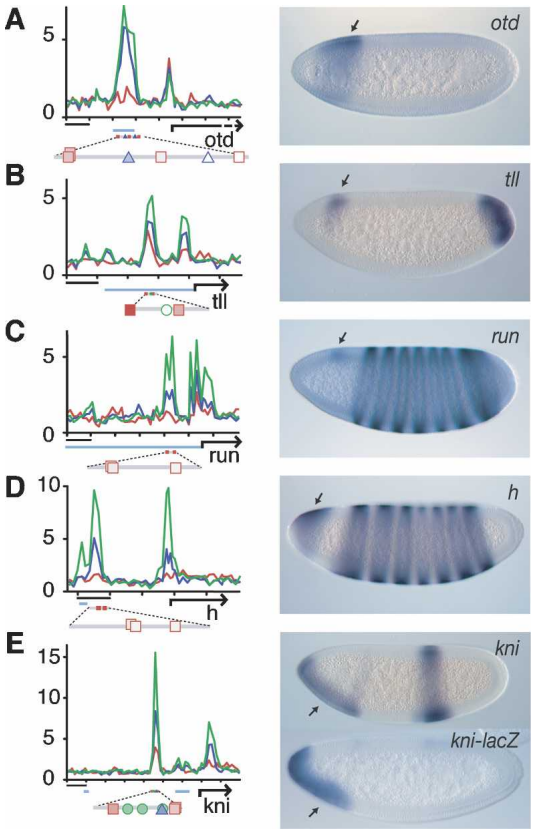


Figure 4. Regulation of AP genes by Dorsal, Twist, and Snail. ChIP–chip binding data of Dorsal (red), Twist (blue), and Snail (green) at the loci of AP-regulated genes are shown in the *left* column (fold enrichment is indicated on the *Y*-axis): *otd* (*A*), *tll* (*B*), *run* (*C*), *h* (*D*), and *kni* (*E*). Bar, 2 kb. The presence of binding motifs and their conservation across the 12 sequenced *Drosophila* genomes at the occupied regions are indicated *below* the graphs (Dorsal site, red square; Twist site, blue triangle; Snail site, green circle). Many identified regions overlap with previously identified regulatory regions (light blue). Association with, and hence regulation by, Dorsal, Twist, and Snail may be responsible for the DV modulation exhibited by these AP genes, as shown in the *right* column (arrows). Embryos are oriented anterior to the left, dorsal is up.

region functions as a bona fide enhancer directing expression in the anteroventral domain like endogenous *knirps* (Fig. 4E). Thus, the ChIP–chip analysis identified novel AP regulatory regions modulated by DV activity.

In summary, many segmentation genes contain DTS/TS-binding clusters, and at least some of these regions modulate gene expression across the DV axis, particularly in anterior regions of the embryo. We conclude that the Dorsal gradient does not only regulate primary DV target genes, but rather appears to fine-tune a large number of genes that do not contribute to DV axis formation themselves, at least based on their known genetic function.

Broad integration of patterning networks

Many DTS/TS-binding clusters are also found at genes encoding signal transduction components (Supplementary Table S3). Analysis of the network formed by these pathways suggests that the Dorsal gradient controls the expression of many target genes by multiple regulatory pathways (Fig. 5).

Dorsal directly represses Dpp expression in the mesoderm and neuroectoderm, leading to localized Dpp signaling in the dorsal ectoderm (Huang et al. 1993). Dpp activates a variety of genes, including *tup* and *pnr*. Our accurate identification of intronic *tup* and *pnr* enhancers suggests that these genes are directly regulated by the Snail repressor (Fig. 3E), in addition to indirect regulation by the Dorsal gradient via Dpp signaling (Fig. 5A). *zen* is another well-known target gene of Dorsal in the dorsal ectoderm, and its product, a homeodomain transcription factor, functions synergistically with Dpp signaling (e.g., see Xu et al. 2005). Target genes of Zen also appear to be subject to additional regulation by the Dorsal gradient (Fig. 5B). In the dorsal ectoderm, Dorsal may regulate gene expression by two mechanisms: direct repression (Kirov et al. 1993), and indirect repression via Snail (Fig. 5C).

Similar network configurations regulate gene expression in the neuroectoderm. High levels of Dorsal repress the expression of *rho* via Snail in the mesoderm (e.g., see Zinzen et al. 2006), thereby blocking EGF signaling in $Toll^{10b}$ mutant embryos. Our ChIP–chip data suggest that the Dorsal network regulates additional genes encoding EGF signaling components as well as EGF target genes such as *pnt*, *aop/yan*, and *argos* (Fig. 5D). In the case of Notch signaling, it is known that the Dorsal network represses Notch target genes such as *sim* in $Toll^{10B}$ mutant embryos through Snail (e.g., see Markstein et al. 2004). The Dorsal network may also regulate Notch sig-

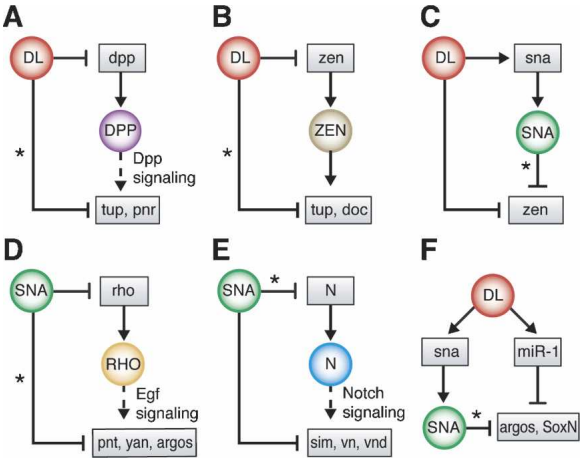


Figure 5. Integration of new putative target genes into the Dorsal network. Many newly identified putative target genes of Dorsal, Twist, and Snail are genes that were previously thought to be induced further downstream in the DV patterning network. Thus, Dorsal appears to regulate many target genes both directly and indirectly (a configuration also known as a feed-forward motif) in the suppression of dorsal ectodermal fate (*A–C*) and neurectodermal fate (*D–F*). Newly identified connections are marked by a star. Examples of such configurations are shown for target genes of Dpp signaling (*A*), which are also directly regulated by Dorsal; target genes of the transcription factor Zen (*B*), which are also directly regulated by Dorsal; target gene suppression by Dorsal (*C*), which can also be mediated by Snail; targets of EGF signaling (*D*), which are regulated by Snail at multiple levels; targets of Notch signaling (*E*), which are also regulated by Snail at multiple levels; and suppression of target genes by Snail or through induction of microRNAs (*F*). The regulatory relationship (activation, arrow; repression, orthogonal bar) of new connections was inferred based on the tissue-specific expression patterns of the target genes (Stathopoulos et al. 2002; Biemar et al. 2006).

Downloaded from www.genesdev.org on March 13, 2007

naling more directly, by suppressing genes encoding components of the signaling pathway including Notch itself (Fig. 5E).

Although repression of neuroectodermal target genes is likely to occur predominantly through Snail, Dorsal also induces the expression of a number of microRNAs in *Toll*^{10b} mutant embryos, including *miR-1* (Biemar et al. 2005; Sokol and Ambros 2005). We found that some of the neuroectodermal genes repressed by Snail are also predicted targets of these microRNAs (Fig. 5F). Hence, there may be multiple tiers of repression in the DV system, similar to the activities of the gap repressors in the AP system (e.g., see Clyde et al. 2003).

In summary, the present ChIP–chip study revealed an unexpectedly broad distribution of binding peaks for Dorsal, Twist, and Snail in the genome, and suggests extensive integration of the Dorsal regulatory network with additional patterning processes, such as Dpp signaling in the dorsal ectoderm and segmentation across the AP axis. In addition to the observed tight correlation between binding peaks and known enhancers, two lines of evidence suggest that a significant fraction of the newly identified regions is functional: First, the bound regions are highly enriched in evolutionarily conserved Dorsal, Twist, and Snail sequence motifs; and, second, several of the identified enhancers were experimentally confirmed by *lacZ* reporter gene expression in transgenic embryos. Thus, while genetic studies identified core sets of regulators for each developmental process in *Drosophila*, we found that gene regulation integrates information more widely from several different systems. It is likely that integration of diverse patterning processes will also apply to mammalian development, including stem cell differentiation.

Materials and methods

ChIP–chip experiments

The ChIP–chip protocol was developed based on protocols from the Young and Maschat laboratories (Chanas et al. 2004; Lee et al. 2006; see Supplemental Material). Cross-linking of the embryos was performed for 5 min in a 5% formaldehyde/hexane solution (Toth and Biggin 2000). The Cy5-labeled amplified immunoprecipated DNA and the Cy3-labeled amplified input DNA were hybridized to whole-genome *Drosophila* arrays using the CGH protocol and CGH reagents provided by Agilent Technologies.

Drosophila *whole-genome array design*

A set of 11 arrays was designed to cover the entire non-repeat-masked euchromatic regions of *Drosophila melanogaster*. This was achieved by 446,931 features (60mers) that were spaced on average ~280 base pairs (bp) from each other. Eight-hundred-sixty-six control spots (oligos from desert regions, the middle of long exons, *Arabidopsis thaliana*, and repetitive sequences) were added to each array. The array was printed by Agilent Technologies.

Drosophila *stocks*

ChIP assays were performed on 2–4-h *Toll*^{10b} embryos (Stathopoulos et al. 2002). P-element transformations and in situ hybridizations were performed with the *yw*^{67} wild-type strain as described (e.g., see Stathopoulos et al. 2002).

Antibodies and beads

Polyclonal antibodies were raised against the HIS-purified (1) N-terminal 50 amino acids of Dorsal in rabbits, (2) full-length Twist in rats, and (3) full-length Snail in guinea pigs (Zinzen et al. 2006). Antibodies were preabsorbed overnight against ~12–24-h embryos (Dorsal), or against overnight collections of dorsalized embryos derived from the *gd*7/FM3 strain (Stathopoulos et al. 2002). The antibodies were then incubated with magnetic Dynabeads coated with protein A (Dorsal, Snail) or protein G (Twist) overnight.

Data processing

Processing of the raw data through an error model and identification of bound regions were performed as previously described for yeast (Pokholok et al. 2005). To identify DTS and TS regions, the union of Twist and Snail regions was determined and regions with an at least fivefold enrichment of either Twist or Snail were selected. Regions with at least twofold enrichment of Dorsal qualified as DTS regions and the remaining regions were called TS regions.

Gene assignment

First, the bound regions located within 2 kb from known gene models were assigned to that gene. The remaining binding regions were then assigned to the gene with the closest transcription start site.

Motif analysis

The binding motifs for Dorsal, Twist, and Snail used for the analysis were obtained from Markstein et al. (2002, 2004), Papatsenko and Levine (2005), Zinzen et al. (2006), Jaspar (MA0022, MA0023, MA0086), and Transfac (M00043, M00120, M00044, M00060). The motifs were matched to the genome of *D. melanogaster* and the orthologous regions of 11 additional sequenced *drosophilid* genomes. Conservation was assessed by the relative total branch length (substitutions per neutrally evolving site) of the phylogenetic tree connecting *D. melanogaster* with the other species.

Cloning, injection, and visualization of enhancers

DNA fragments were PCR-amplified from genomic DNA (for primers, see Supplemental Material), T/A-cloned into the Promega pGEM T-Easy vector, and subcloned via NotI into the gypsy-insulated pCaSpeR vector E2G (Markstein et al. 2004). Each construct was introduced into the germline by microinjection and three to eight independent transgenic lines were tested. Embryos were collected and fixed according to standard methods, hybridized with digoxygenin-labeled antisense RNA probes against *lacZ* or the endogenous gene, and visualized colorimetrically (Jiang et al. 1991).

Acknowledgments

We thank J. Love for microarray experimental support, F. Lewitter and B. Yuan for help in the *Drosophila* whole-genome microarray design, E. Herbolsheimer for bioinformatics support, X. Li for technical advice, and M. Markstein, F. Biemar, and Young laboratory members for discussions. The work was supported by NIH grants HG002668 and GM069676 (to R.A.Y.), GM46638 (to M.L.), and a grant by the Moore Foundation and a postdoctoral fellowship by Schering (to A.S.). R.A.Y. consults for Agilent Technologies.

References

Biemar, F., Zinzen, R., Ronshaugen, M., Sementchenko, V., Manak, J.R., and Levine, M.S. 2005. Spatial regulation of microRNA gene expression in the *Drosophila* embryo. *Proc. Natl. Acad. Sci.* **102:** 15907–15911.

Biemar, F., Nix, D.A., Piel, J., Peterson, B., Ronshaugen, M., Sementchenko, V., Bell, I., Manak, J.R., and Levine, M.S. 2006. Comprehensive identification of *Drosophila* dorsal–ventral patterning genes using a whole-genome tiling array. *Proc. Natl. Acad. Sci.* **103:** 12763–12768.

Butler, B.A., Soong, J., and Gergen, J.P. 1992. The *Drosophila* segmentation gene runt has an extended *cis*-regulatory region that is required for vital expression at other stages of development. *Mech. Dev.* **39:** 17–28.

Carroll, S.B., Winslow, G.M., Twombly, V.J., and Scott, M.P. 1987. Genes that control dorsoventral polarity affect gene expression along the anteroposterior axis of the *Drosophila* embryo. *Development* **99:** 327–332.

Chanas, G., Lavrov, S., Iral, F., Cavalli, G., and Maschat, F. 2004. Engrailed and polyhomeotic maintain posterior cell identity through

Downloaded from www.genesdev.org on March 13, 2007

cubitus-interruptus regulation. *Dev. Biol.* **272:** 522–535.
Clyde, D.E., Corado, M.S., Wu, X., Pare, A., Papatsenko, D., and Small, S. 2003. A self-organizing system of repressor gradients establishes segmental complexity in *Drosophila*. *Nature* **426:** 849–853.
Ganguly, A., Jiang, J., and Ip, Y.T. 2005. *Drosophila* WntD is a target and an inhibitor of the Dorsal/Twist/Snail network in the gastrulating embryo. *Development* **132:** 3419–3429.
Gao, Q. and Finkelstein, R. 1998. Targeting gene expression to the head: The *Drosophila* orthodenticle gene is a direct target of the Bicoid morphogen. *Development* **125:** 4185–4193.
Gordon, M.D., Dionne, M.S., Schneider, D.S., and Nusse, R. 2005. WntD is a feedback inhibitor of Dorsal/NF-κB in *Drosophila* development and immunity. *Nature* **437:** 746–749.
Huang, J.D., Schwyter, D.H., Shirokawa, J.M., and Courey, A.J. 1993. The interplay between multiple enhancer and silencer elements defines the pattern of decapentaplegic expression. *Genes & Dev.* **7:** 694–704.
Ip, Y.T. and Gridley, T. 2002. Cell movements during gastrulation: Snail dependent and independent pathways. *Curr. Opin. Genet. Dev.* **12:** 423–429.
Jiang, J., Kosman, D., Ip, Y.T., and Levine, M. 1991. The dorsal morphogen gradient regulates the mesoderm determinant twist in early *Drosophila* embryos. *Genes & Dev.* **5:** 1881–1891.
Kirov, N., Zhelnin, L., Shah, J., and Rushlow, C. 1993. Conversion of a silencer into an enhancer: Evidence for a co-repressor in dorsal-mediated repression in *Drosophila*. *EMBO J.* **12:** 3193–3199.
Kosman, D., Ip, Y.T., Levine, M., and Arora, K. 1991. Establishment of the mesoderm-neuroectoderm boundary in the *Drosophila* embryo. *Science* **254:** 118–122.
Lee, T.I., Johnstone, S.E., and Young, R.A. 2006. Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nat. Protoc.* **1:** 729–748.
Leptin, M. 1991. twist and snail as positive and negative regulators during *Drosophila* mesoderm development. *Genes & Dev.* **5:** 1568–1576.
Levine, M. and Davidson, E.H. 2005. Gene regulatory networks for development. *Proc. Natl. Acad. Sci.* **102:** 4936–4942.
Liaw, G.J. and Lengyel, J.A. 1993. Control of tailless expression by bicoid, dorsal and synergistically interacting terminal system regulatory elements. *Mech. Dev.* **40:** 47–61.
Markstein, M., Markstein, P., Markstein, V., and Levine, M.S. 2002. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Proc. Natl. Acad. Sci.* **99:** 763–768.
Markstein, M., Zinzen, R., Markstein, P., Yee, K.P., Erives, A., Stathopoulos, A., and Levine, M. 2004. A regulatory code for neurogenic gene expression in the *Drosophila* embryo. *Development* **131:** 2387–2394.
Moussian, B. and Roth, S. 2005. Dorsoventral axis formation in the *Drosophila* embryo—Shaping and transducing a morphogen gradient. *Curr. Biol.* **15:** R887–R899.
Nusslein-Volhard, C., Lohs-Schardin, M., Sander, K., and Cremer, C. 1980. A dorso–ventral shift of embryonic primordia in a new maternal-effect mutant of *Drosophila*. *Nature* **283:** 474–476.
Pankratz, M.J., Seifert, E., Gerwin, N., Billi, B., Nauber, U., and Jackle, H. 1990. Gradients of Kruppel and knirps gene products direct pair-rule gene stripe patterning in the posterior region of the *Drosophila* embryo. *Cell* **61:** 309–317.
Papatsenko, D. and Levine, M. 2005. Quantitative analysis of binding motifs mediating diverse spatial readouts of the Dorsal gradient in the *Drosophila* embryo. *Proc. Natl. Acad. Sci.* **102:** 4966–4971.
Pelegri, F. and Lehmann, R. 1994. A role of polycomb group genes in the regulation of gap gene expression in *Drosophila*. *Genetics* **136:** 1341–1353.
Pokholok, D.K., Harbison, C.T., Levine, S., Cole, M., Hannett, N.M., Lee, T.I., Bell, G.W., Walker, K., Rolfe, P.A., Herbolsheimer, E., et al. 2005. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* **122:** 517–527.
Rothe, M., Wimmer, E.A., Pankratz, M.J., Gonzalez-Gaitan, M., and Jackle, H. 1994. Identical transacting factor requirement for knirps and knirps-related gene expression in the anterior but not in the posterior region of the *Drosophila* embryo. *Mech. Dev.* **46:** 169–181.
Simpson, P. 1983. Maternal–zygotic gene interactions during formation of the dorsoventral pattern in *Drosophila* embryos. *Genetics* **105:** 615–632.
Sokol, N.S. and Ambros, V. 2005. Mesodermally expressed *Drosophila* microRNA-1 is regulated by Twist and is required in muscles during larval growth. *Genes & Dev.* **19:** 2343–2354.
St Johnston, D. and Nusslein-Volhard, C. 1992. The origin of pattern and polarity in the *Drosophila* embryo. *Cell* **68:** 201–219.
Stathopoulos, A., Van Drenth, M., Erives, A., Markstein, M., and Levine, M. 2002. Whole-genome analysis of dorsal–ventral patterning in the *Drosophila* embryo. *Cell* **111:** 687–701.
Toth, J. and Biggin, M.D. 2000. The specificity of protein–DNA crosslinking by formaldehyde: In vitro and in drosophila embryos. *Nucleic Acids Res.* **28:** e4.
Xu, M., Kirov, N., and Rushlow, C. 2005. Peak levels of BMP in the *Drosophila* embryo control target genes by a feed-forward mechanism. *Development* **132:** 1637–1647.
Zinzen, R.P., Senger, K., Levine, M., and Papatsenko, D. 2006. Computational models for neurogenic gene expression in the *Drosophila* embryo. *Curr. Biol.* **16:** 1358–1365.

Cell

Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in *Drosophila*

Julius Brennecke,[1] Alexei A. Aravin,[1,3] Alexander Stark,[2,3] Monica Dus,[1] Manolis Kellis,[2] Ravi Sachidanandam,[1] and Gregory J. Hannon[1,*]

[1] Cold Spring Harbor Laboratory, Watson School of Biological Sciences and Howard Hughes Medical Institute, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA

[2] Broad Institute, MIT Center for Genome Research, 320 Charles Street, Cambridge, MA 02141, USA

[3] These authors contributed equally to this work.

*Correspondence: hannon@cshl.edu

DOI 10.1016/j.cell.2007.01.043

SUMMARY

Drosophila Piwi-family proteins have been implicated in transposon control. Here, we examine piwi-interacting RNAs (piRNAs) associated with each *Drosophila* Piwi protein and find that Piwi and Aubergine bind RNAs that are predominantly antisense to transposons, whereas Ago3 complexes contain predominantly sense piRNAs. As in mammals, the majority of *Drosophila* piRNAs are derived from discrete genomic loci. These loci comprise mainly defective transposon sequences, and some have previously been identified as master regulators of transposon activity. Our data suggest that heterochromatic piRNA loci interact with potentially active, euchromatic transposons to form an adaptive system for transposon control. Complementary relationships between sense and antisense piRNA populations suggest an amplification loop wherein each piRNA-directed cleavage event generates the 5′ end of a new piRNA. Thus, sense piRNAs, formed following cleavage of transposon mRNAs may enhance production of antisense piRNAs, complementary to active elements, by directing cleavage of transcripts from master control loci.

INTRODUCTION

Mobile genetic elements, or their remnants, populate the genomes of nearly every living organism. Potential negative effects of mobile elements on the fitness of their hosts necessitate the development of strategies for transposon control. This is critical in the germline, where transposon activity can create a substantial mutational burden that would accumulate with each passing generation.

Hybrid dysgenesis exemplifies the deleterious effects of colonization of a host by an uncontrolled mobile element. The progeny of intercrosses between certain *Drosophila* strains reproducibly show high germline mutation rates with elevated frequencies of chromosomal abnormalities and partial or complete sterility (Bucheton, 1990; Castro and Carareto, 2004; Kidwell et al., 1977). Studies of the molecular basis of this phenomenon linked the phenotype to transposon mobilization (Pelisson, 1981; Rubin et al., 1982).

Hybrid dysgenesis occurs when a transposon, carried by a male that has established control over that element, is introduced into a naïve female that does not carry the element. The transposon becomes active in the progeny of the naïve female, causing a variety of abnormalities in reproductive tissues that ultimately result in sterility (Engels and Preston, 1979). Since the dysgenic phenotype is often not completely penetrant, a fraction of the progeny from affected females may survive to adulthood. Such animals can develop resistance to the mobilized element, although in many cases, several generations are required for resistance to become fully established (Pelisson and Bregliano, 1987). Immunity to transposons can only be passed through the female germline, indicating that there are both cytoplasmic and genetic components to inherited resistance (Bregliano et al., 1980).

Studies of hybrid dysgenesis have served a critical role in revealing mechanisms of transposon control. In general, two seemingly contradictory models have emerged. The first model correlates resistance with an increasing copy number of the mobile element. A second model suggests that discrete genomic loci encode transposon resistance.

The first model is supported by studies of the I element. Crossing a male carrying full-length copies of the I element to a naïve female leads to I mobilization and hybrid dysgenesis (Bregliano et al., 1980; Bucheton et al., 1984). The number of I copies builds during subsequent crosses of surviving female progeny until it reaches an average of 10–15 per genome (Pelisson and Bregliano, 1987). At this point, I mobility is suppressed, as the initially naïve strain

gains control over this element. Thus, a gradual increase in I element copy number over multiple generations was implicated in the development of transposon resistance.

The second model, which attributes transposon resistance to specific genetic loci, is illustrated by studies of *gypsy* transposon control (Bucheton, 1995). Genetic mapping of *gypsy* resistance determinants led to a discrete locus in the pericentric β-heterochromatin of the X chromosome that was named *flamenco* (Pelisson et al., 1994). Females carrying a permissive *flamenco* allele (one that allows *gypsy* activity) showed a dysgenic phenotype when crossed to males carrying functional *gypsy* elements. Permissive *flamenco* alleles exist in natural *Drosophila* populations but can also be produced by insertional mutagenesis of animals carrying a restrictive *flamenco* allele (Robert et al., 2001). Despite extensive deletion mapping over the *flamenco* locus, no transposon repressor from *flamenco* has been identified. For P elements, a repressor of transposition has been identified as a 66 kDa version of the P element transposase. Expression of the repressor was proposed to correlate with increasing P element copy number, leading to a self-imposed limitation on P element mobility (Misra and Rio, 1990). However, studies of resistance determinants indicated that control over P elements could also be established by insertion of P elements into specific genomic loci, arguing for an alternative, copy number-independent control pathway (Biemont et al., 1990). Studies of inbred lines or of wild isolates with natural P element resistance indicated that P insertions near the telomere of X (cytological position 1A) were sufficient to confer resistance if maternally inherited (Biemont et al., 1990; Ronsseray et al., 1991). Additionally, several groups isolated insertions of incomplete P elements in this same cytological location that acted as dominant transposition suppressors (Marin et al., 2000; Stuart et al., 2002). Importantly, these defective P elements lacked sequences encoding the repressor fragment of transposase.

Both models of transposon resistance, those determined by specific genomic loci and those caused by copy number-dependent responses might be linked to small RNA-based regulatory pathways. Copy number-dependent silencing of mobile elements is reminiscent of copy number-dependent transgene silencing in plants (cosuppression) (Smyth, 1997) and *Drosophila* (Pal-Bhadra et al., 1997). In both cases, silencing occurs through an RNAi-like response where high-copy transgenes provoke the generation of small RNAs, presumably through a double-stranded RNA intermediate (Hamilton and Baulcombe, 1999; Pal-Bhadra et al., 2002). Moreover, mutations in RNAi pathway genes impact transposon mobility in flies (Kalmykova et al., 2005; Sarot et al., 2004; Savitsky et al., 2006) and *C.elegans* (Ketting et al., 1999; Tabara et al., 1999). Finally, small RNAs (rasiRNAs) corresponding to transposons and repeats have been isolated from flies and zebrafish (Aravin et al., 2001, 2003; Chen et al., 2005).

At the core of the RNAi machinery are the Argonaute proteins, which directly bind to small RNAs and use these as guides for the identification and cleavage of their targets (Liu et al., 2004). In animals, Argonautes can be divided into two clades (Carmell et al., 2002). One contains the Argonautes, which act with microRNAs and siRNAs to mediate gene silencing. The second contains the Piwi proteins. Genetic studies have implicated Piwi proteins in germline integrity (Cox et al., 1998; Harris and Macdonald, 2001). For example, *piwi* mutations cause sterility and loss of germline stem cells (Cox et al., 1998; Lin and Spradling, 1997). *aubergine* is a spindle-class gene that is required in the germline for the production of functional oocytes (Harris and Macdonald, 2001). The third *Drosophila* Piwi gene, *Ago3*, has yet to be studied. Mutation of Piwi-family genes also affects mobile elements. For example, *piwi* mutations mobilize *gypsy* (Sarot et al., 2004), and *aubergine* mutations impact *TART* (Savitsky et al., 2006) and P elements (Reiss et al., 2004). Finally, both Piwi and Aubergine bind rasiRNAs (Saito et al., 2006; Vagin et al., 2006) targeting a number of mobile and repetitive elements. These complexes are enriched for antisense small RNAs, as might be expected if they were actively involved in silencing transposons by recognition of their RNA products.

Recently, a new class of small RNAs, the piRNAs, was identified through association with Piwi proteins in mammalian testes (Aravin et al., 2006; Girard et al., 2006; Grivna et al., 2006; Lau et al., 2006). These 26–30 nt RNAs are produced from discrete loci, generally spanning 50–100 kb. Interestingly, mammalian piRNAs are relatively depleted of transposon sequences. Despite apparent differences in the content of Piwi-associated RNA populations in mammals and *Drosophila*, Piwi-family proteins share essential roles in gametogenesis, with all three murine family members, Miwi2 (M.A. Carmell et al., submitted), Mili (Kuramochi-Miyagawa et al., 2004), and Miwi (Deng and Lin, 2002), being required for male fertility.

In order to probe mechanisms of transposon control in *Drosophila* and to understand the relationship between Piwi protein function in flies and mammals, we undertook a detailed analysis of small RNAs associated with Piwi proteins in the *Drosophila* female germline. Our studies indicate that *Drosophila* Piwi-family members function in a transposon surveillance pathway that not only preserves a genetic memory of transposon exposure but also has the potential to adapt its response upon contact with active transposons.

RESULTS

Piwi-Family Members Have Distinct Expression Patterns

In *Drosophila*, Piwi-clade consists of three members: Piwi, Aubergine (Aub), and Ago3. As a prerequisite to further studies of this family, we experimentally determined the sequence of *Ago3* and raised antisera specific to each *Drosophila* Piwi-family protein (Figures S1 and S2).

Previous studies have used myc-tagged Piwi and green fluorescent protein (GFP)-tagged Aub transgenes to investigate their spatial and temporal expression patterns

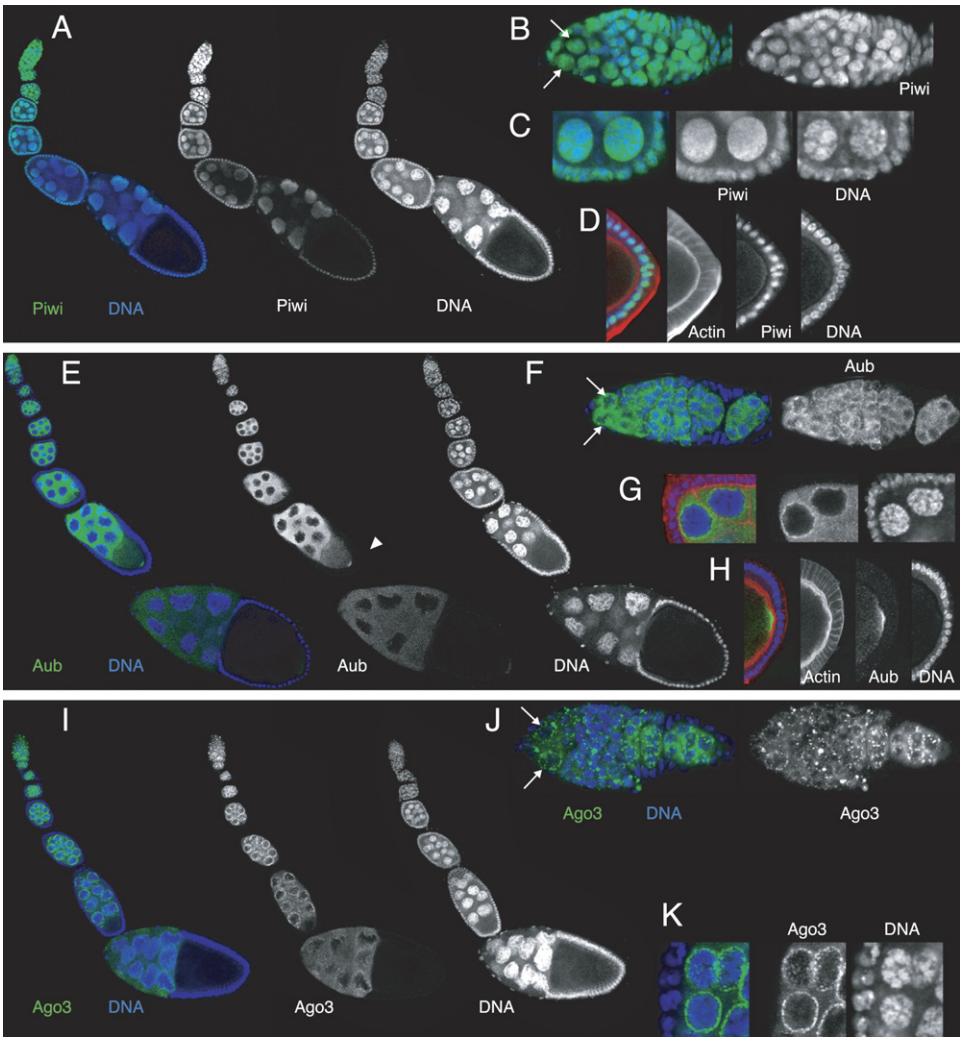


Figure 1. Expression and Localization of Piwi-Family Members in Ovarioles

In all panels, signal from the indicated Piwi-family member is in green; DNA is in blue and actin in red.

(A) Overview of Piwi localization in the ovariole.

(B) Detailed view of the germarium containing the two stem cells (arrows).

(C) Nuclear localization of Piwi in nurse cells and surrounding follicle cells.

(D) Indicates a weak accumulation of maternally deposited Piwi at the posterior pole of stage 10 oocytes.

(E) Overview of Aubergine localization in the ovariole with enrichment of Aub at the posterior oocyte pole in late egg chambers (arrow head).

(F) Detailed view of Aub localization in the germarium containing the two stem cells (arrows).

(G) Enrichment of Aub in the cytoplasm and perinuclear nuage of germline cells; staining is absent from surrounding somatic follicle cells.

(H) Accumulation of Aub at the posterior pole of a stage 10 oocyte.

(I) Overview of Ago3 localization in the ovariole.

(J) Detailed view of Ago3 staining in the germarium showing strong enrichment around the stem cell nuclei (arrows) and in discrete foci.

(K) Detailed view of Ago3 localization to nuage in nurse cells.

during oogenesis (Cox et al., 2000; Harris and Macdonald, 2001). We used our specific antibodies to examine the expression patterns of the endogenous proteins and to extend analyses to the third family member, Ago3.

As previously reported (Cox et al., 2000), Piwi is predominantly nuclear and is present not only in germline cells but also in the somatic cells of the ovary (Figures 1A–1D). Strong Piwi staining is seen in the cap cells that surround the germline stem cells and in the follicle cells that envelop the developing egg chamber. In later stage egg chambers, Piwi is detectable in the cytoplasm of the developing oocyte with a slight enrichment at the posterior where the germline primordia of the embryo will form.

Aubergine is expressed at very low or undetectable levels outside the germline (Figures 1E–1H). Aub is primarily cytoplasmic. As reported previously for GFP-Aub, we detect endogenous protein in germline stem cells, developing cystoblasts, and the nurse cells of developing egg

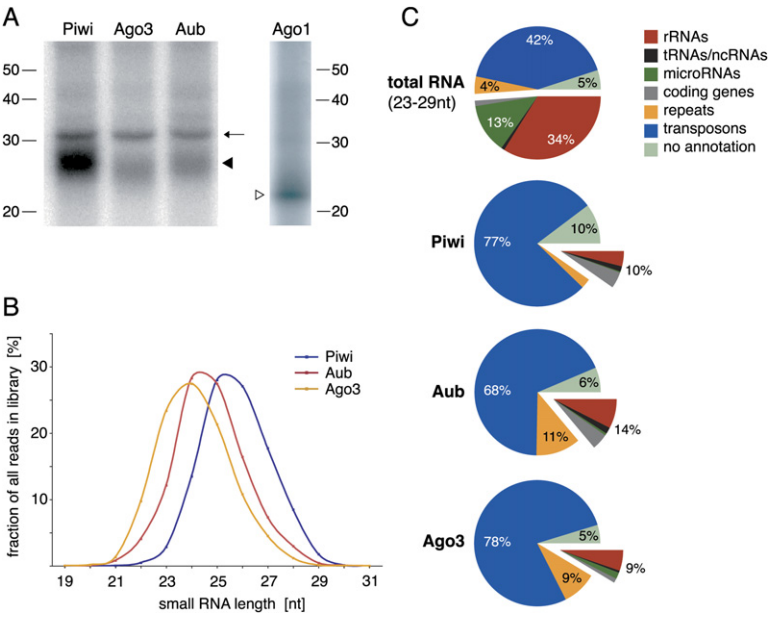


Figure 2. Characteristics of *Drosophila* piRNAs

(A) Radioactively labeled RNA isolated under identical conditions from specific Piwi-family RNPs and Ago1 was analyzed on a denaturing polyacrylamide gel. The positions of RNA size markers, electrophoresed in parallel, are shown to the left. Indicated are piRNAs (solid arrowhead), miRNAs (open arrow head), and 2S rRNA (arrow), which is also present in purifications using control antibodies.

(B) Size distributions of sequenced piRNAs specifically bound by the three Piwi-family members.

(C) Pie chart summarizing the annotation of piRNA populations in total RNA and those bound by Piwi, Aub, and Ago3.

chambers. Aubergine is enriched in nuage, a perinuclear, electron dense structure, displaying a localization pattern very similar to Vasa. Like Vasa, Aubergine is deposited into the developing oocyte from early stage 10 onward and is localized to the pole plasm.

Ago3 protein is also predominantly cytoplasmic (Figure 1I–1K). It is present in the germline but not detectable in somatic cells of the egg chamber, although we do find Ago3 in the somatic cap cells of the germarium. Ago3 shows a more striking accumulation in nuage than does Aub, and it is also found in prominent but discrete foci of unknown character in the germarium. Despite its localization to nuage, Ago3 does not accumulate at the posterior pole of the developing oocyte, and Ago3 is not detected in the pole plasm of early embryos.

Piwi-Family Members Bind Distinct Populations of Small RNAs

To investigate the small RNA populations bound by each *Drosophila* Piwi protein, we purified RNP complexes from ovary lysates. All three proteins associate with small RNAs ranging in length from 23 to 29 nt (Figure 2A). We prepared cDNA libraries from each complex and from 23–29 nt RNAs purified from total ovary RNA. 454 sequencing yielded 60,691 reads (17,709 for Piwi; 23,376 for Ago3; 14,872 for Aub; and 4,734 for ovary total RNA) that match perfectly to Release 5 of the *Drosophila* genome or to nonassembled *Drosophila* sequences from GenBank. These were used for subsequent analysis. It should be noted that we isolated small RNAs from the Oregon R strain, rather than the sequenced strain (y^1; cn^1 bw^1 sp^1). A detailed discussion of potential differences between these genomes and any impacts on the data that we present can be found in the Supplemental Data.

Based both upon gel mobility (Figure 2A) and size distributions (Figure 2B), each Piwi protein bound a specific class of piwi-interacting RNA (piRNA). Piwi-associated RNAs were the largest (25.7 nt), followed by Aub (24.7) and Ago3 bound (24.1) RNAs. As with mammalian piRNAs, Piwi and Aub bound RNAs have a strong preference for a 5′ terminal uridine (83% and 72%, respectively), a trend that is essentially absent from the Ago3 bound population (37% terminal U). *Drosophila* piRNA populations are quite complex, with most RNAs being cloned only once (87% for Piwi, 81% for Aub, and 73% for Ago3).

Despite their size differences, the small RNAs obtained from each complex were remarkably similar in the types of genomic elements to which they correspond. Overall, roughly three quarters of all sequences from each complex could be assigned to annotated transposons or transposon remnants, with nearly all known transposons being represented (Figures 2C and S3). An additional 2%–11% of small RNAs were derived from regions of local repeat structure, such as the subtelomeric TAS repeats or pericentromeric satellite repeats. Thus, nearly 80% of *Drosophila* piRNAs can also be characterized as rasiRNAs. An additional group of sequences (10% for Piwi, 6% for Aub, and 5% for Ago3) map to unannotated regions of the genome, most often to heterochromatic, transposon-rich loci.

Drosophila piRNAs Are Derived from Discrete Genomic Loci

In *Drosophila*, intact and potentially active transposons populate the euchromatin as well as pericentromeric and telomeric heterochromatin. There are also numerous transposon remnants that have mutated sufficiently to negate their potential for transposition. These are strongly enriched in the β-heterochromatin that borders

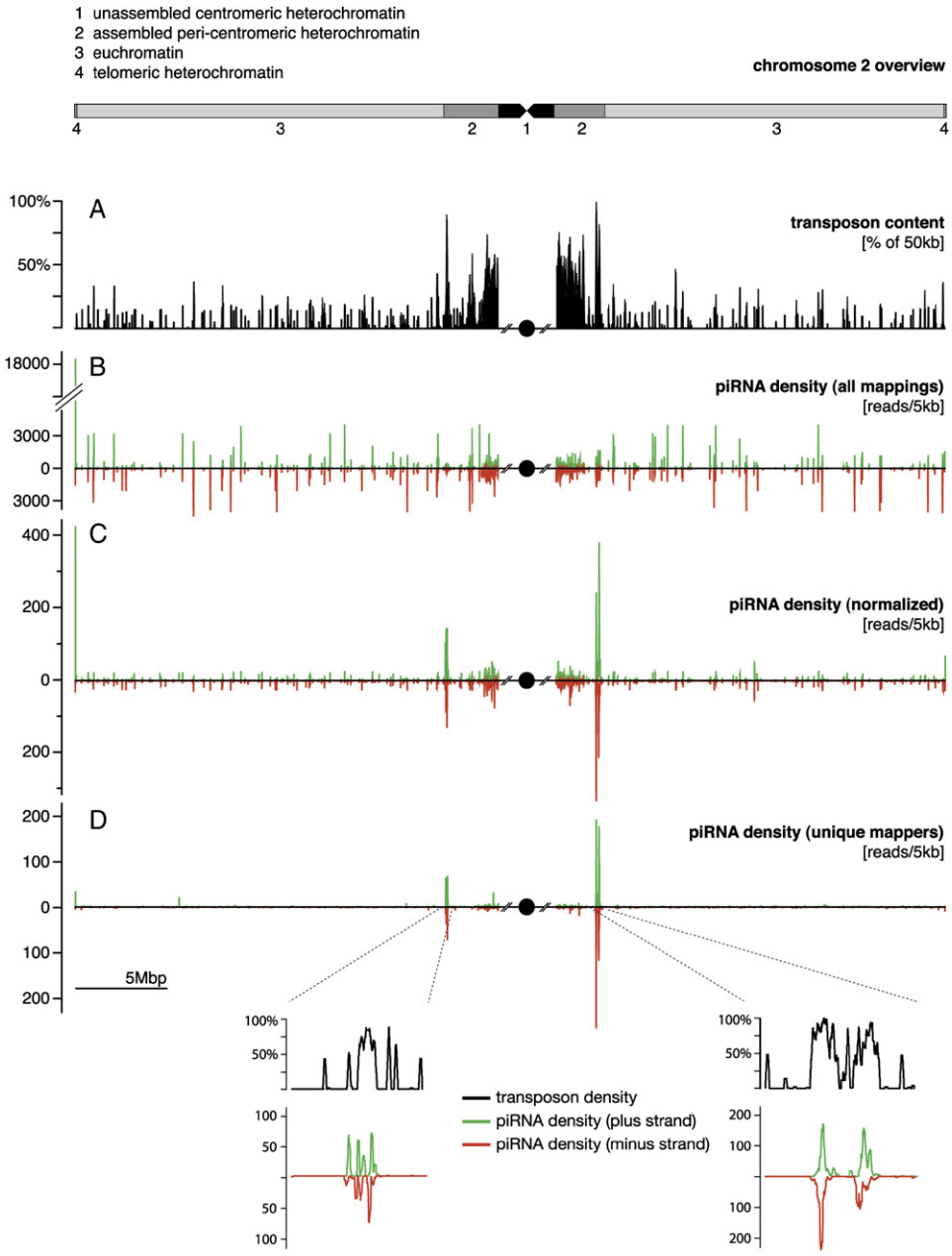


Figure 3. *Drosophila* piRNAs Map to Discrete Genomic Loci

For reference, a sketch of chromosome 2 with the major chromatin domains is provided.

(A) Density of annotated transposons.

(B–D) Density of cloned piRNAs (green: sense; red: antisense) along chromosome 2; the y axis indicates relative units, and the centromere is shown as a circle. In (B) each genomic position corresponding to a cloned piRNA is given equal weight; in (C) each piRNA-genomic match is normalized for the number of times it maps to the genome, resulting in proportionally less signal for piRNAs that map many times. In (D) only those piRNAs that map uniquely to the genome are plotted. The most prominent piRNA clusters (enlarged below and shown together with transposon density in black) are typically found at the border of the pericentromeric heterochromatin.

centromeres and are generally absent from euchromatic chromosome arms (Hoskins et al., 2002).

A depiction of the chromosomal locations matched by *Drosophila* piRNAs closely resembles a plot of transposon density (Figure 3). However, most piRNAs match multiple chromosomal sites. Therefore, to address the genomic origin of piRNAs, it was necessary to restrict our analysis to the ~20% that match the genome at a unique position (Figure S4). A density plot of this small RNA subset shows a striking clustering of piRNAs at discrete loci (e.g.,

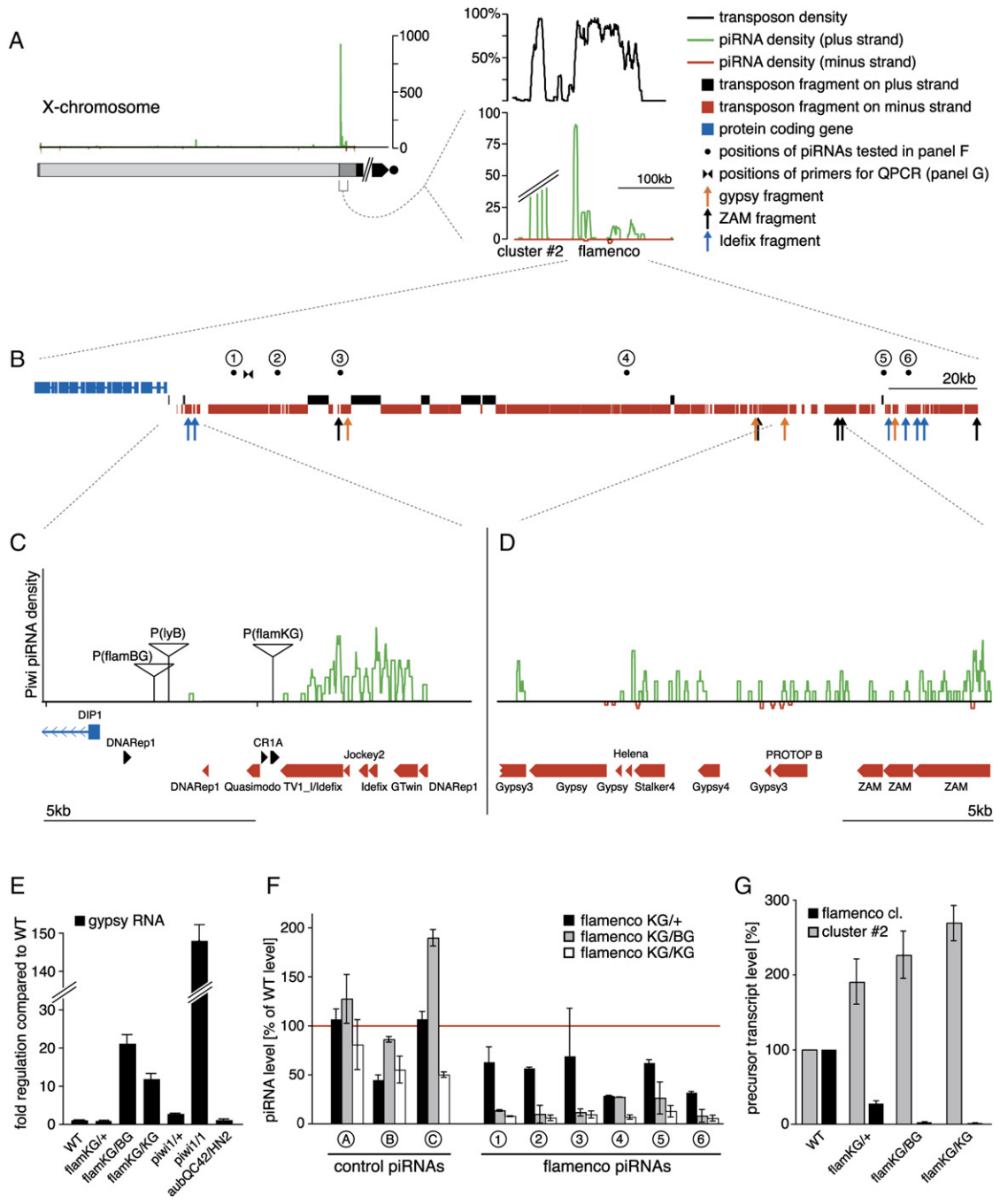


Figure 4. *flamenco* Is a piRNA Cluster

(A) piRNA density on the X chromosome showing two large clusters (numbers as in Table S1) in the pericentromeric heterochromatin (chromatin domains are shaded as in Figure 3). Densities of uniquely mapping sense (green) and antisense (red) piRNAs in two clusters is shown enlarged alongside with corresponding transposon density (black).

(B) Shown is a more detailed map of the *flamenco* locus, indicating protein-coding genes (blue), sense (black), and antisense (red) transposon fragments. The *flamenco* cluster ends ~180 kb proximal to *DIP1* in a gap of unknown size in the genome assembly. Arrows highlight the retroelements *gypsy* (orange), *Idefix* (blue), and *ZAM* (black), which are known to be regulated by the locus.

(C) Shown are the first 10 kb of the *flamenco* locus with the flanking *DIP1* gene (blue), annotated transposon fragments, the P element insertions that result in inactive *flamenco* alleles (triangles), and the density of Piwi-associated piRNAs that map to this region (note that over 99% of the uniquely mapping piRNAs in the *flamenco* cluster derive from the sense strand and that more than 95% associate with Piwi).

(D) Fifteen kilobases of the *flamenco* locus showing Piwi bound piRNAs and a detailed view of transposon fragments from the region, including *gypsy* and *ZAM*.

Figure 3D). A similar plot can be obtained for all piRNAs if each piRNA-genomic match is divided by the number of genomic hits for that sequence (normalized) (Figure 3C). We identified 142 genomic locations as sites of abundant piRNA generation (Tables S1 and S2). These clusters produce 81% of all piRNAs that match the genome at a single site. Although these sites comprise only 3.5% of the assembled genome, more than 92% of all sequenced piRNAs could potentially be derived from these loci.

Only seven piRNA clusters occur in potentially euchromatic regions, with the remainder being present in pericentromeric and telomeric heterochromatin. Telomeric clusters most often consist of satellite sequences and correspond to the subtelomeric terminal associated sequence (TAS) repeats (Karpen and Spradling, 1992). These flank the telomeric arrays of *HetA* and *TART* transposons, for which we also find corresponding piRNAs. Telomeric clusters, supported by the presence of uniquely mapping piRNAs, are found on most chromosome arms (X, 2R, 2L, and 3R). Clusters in the pericentromeric β-heterochromatin display a high content of annotated transoposons (typically from 70% to 90%), with the majority being partial or defective nested elements.

The size of *Drosophila* piRNA clusters varies substantially with the largest being a 240 kB locus in the pericentromeric heterochomatin of chromosome 2R (cytological position 42AB). This cluster produces 20.8% of all uniquely mapping piRNA sequences and could potentially give rise to 30.1% of all the piRNAs, which we identified (Table S1). Overall, the largest 15 clusters account for 57% of the unique piRNAs and up to 70% of the total.

piRNA Clusters Are Master Regulators of Transposon Activity

Numerous genetic studies have pointed to discrete genomic loci that suppress the activity of specific transposons. The best understood is the recessive *flamenco/COM* locus (Prud'homme et al., 1995). *flamenco* was originally identified as a locus controlling the activity of the retroviral *gypsy* element (Pelisson et al., 1994). This locus has subsequently been shown to regulate two additional retroelements, *Idefix* and *ZAM* (Desset et al., 2003).

Through the use of numerous deficiencies, *flamenco* was mapped proximally to the *DIP-1* gene and proposed to span a region of at least 130 kb. This locus corresponds precisely to a piRNA cluster (cluster 8; Table S1, Figures 4A and 4B). Eighty-seven percent of the sequence in the locus consists of nested transposable elements spanning a total length of 179 kb. The locus includes numerous fragments of all three transposable elements that were shown to be controlled by *flamenco/COM* (*gypsy*, *Idefix*, and *ZAM*; Figurse 4B–4D) in addition to many other transposon families.

A second piRNA cluster that has been genetically linked to transposon control corresponds to the subtelomeric TAS repeat on the X chromosome (X-TAS) (Table S1; cluster 4). Numerous studies indicate that insertions of one or two P elements into X-TAS are sufficient to suppress P-M hybrid dysgenesis (Marin et al., 2000; Ronsseray et al., 1991; Stuart et al., 2002). Transposon silencing by these insertions has been linked to the Piwi family, as it is relieved by mutations in *aubergine* (Reiss et al., 2004). The precise insertion sites of three suppressive P elements in X-TAS have been mapped, and they correspond to areas, which give rise to multiple small RNAs (not shown). In accord with a *trans*-acting model for silencing, lacZ-containing P elements inserted into X-TAS can suppress euchromatic lacZ transgenes in the female germline (Roche and Rio, 1998; Ronsseray et al., 1998).

A Functional Pathway Links *flamenco*-Derived piRNAs to *gypsy* Suppression

To probe the relevance of the piRNA cluster mapped to *flamenco*, we made use of mutations that negate the ability of this locus to silence *gypsy*. The only molecularly defined *flamenco* allele corresponds to a P element insertion ~2 kb proximal to *DIP1* and 550 bp upstream of the first piRNA uniquely mapped to this cluster (P(lyB); Figure 4C) (Robert et al., 2001). We obtained two additional lines that harbor P element insertions near P(lyB) and tested their effects on *gypsy* expression. *gypsy* RNA levels increased by ~20-fold in strains carrying homozygous or trans-heterozygous insertions, indicating that P(KG00476) (flamKG) and P(BG02658) (flamBG) (Figure 4C) are indeed *flamenco* mutant alleles (Figure 4E).

We next examined the levels of mature piRNAs from *flamenco* in wild-type animals and *flamenco* mutants. Using quantitative RT-PCR, we found substantial reductions in piRNAs that uniquely map to the *flamenco* piRNA cluster in mutant animals (2, 3, and 5; Figures 4B and 4F). In contrast, piRNAs definitively derived from other clusters were unaffected. We also probed levels of piRNAs that did not map uniquely to *flamenco* (1,4, and 6; Figures 4B and 4F). These were also substantially reduced in *flamenco* mutants, indicating that they arise mainly from this cluster despite not being uniquely assignable based upon sequence information alone.

The *flamenco* piRNA cluster preferentially loads the Piwi protein, with 94% of its uniquely mapping RNAs being Piwi partners. An examination of *gypsy* RNA levels reveals a 150-fold increase in *piwi* mutants (Figure 4E) (Sarot et al., 2004). In contrast, *aubergine* mutations show no

(E) Quantitative RT-PCR analysis on *gypsy* transcript levels in RNA isolated from wild-type, *flamenco* mutant, *piwi* mutant, and *aub* mutant ovaries. Shown are average levels (n = 3), and error bars indicate standard deviation (SD).

(F) Quantitative RT-PCR analysis on several individual piRNAs from small RNA libraries prepared from wild-type or *flamenco* mutant ovaries. Positions of tested piRNAs in *flamenco* are indicated in (B). Error bars indicate SD.

(G) Quantitative RT-PCR analysis on precursor transcripts from two different piRNA clusters in ovaries from wild-type and *flamenco* mutant flies. Position of primers used for the *flamenco* primary transcript are indicated in (B). Shown are average levels (n = 3), and error bars indicate SD.

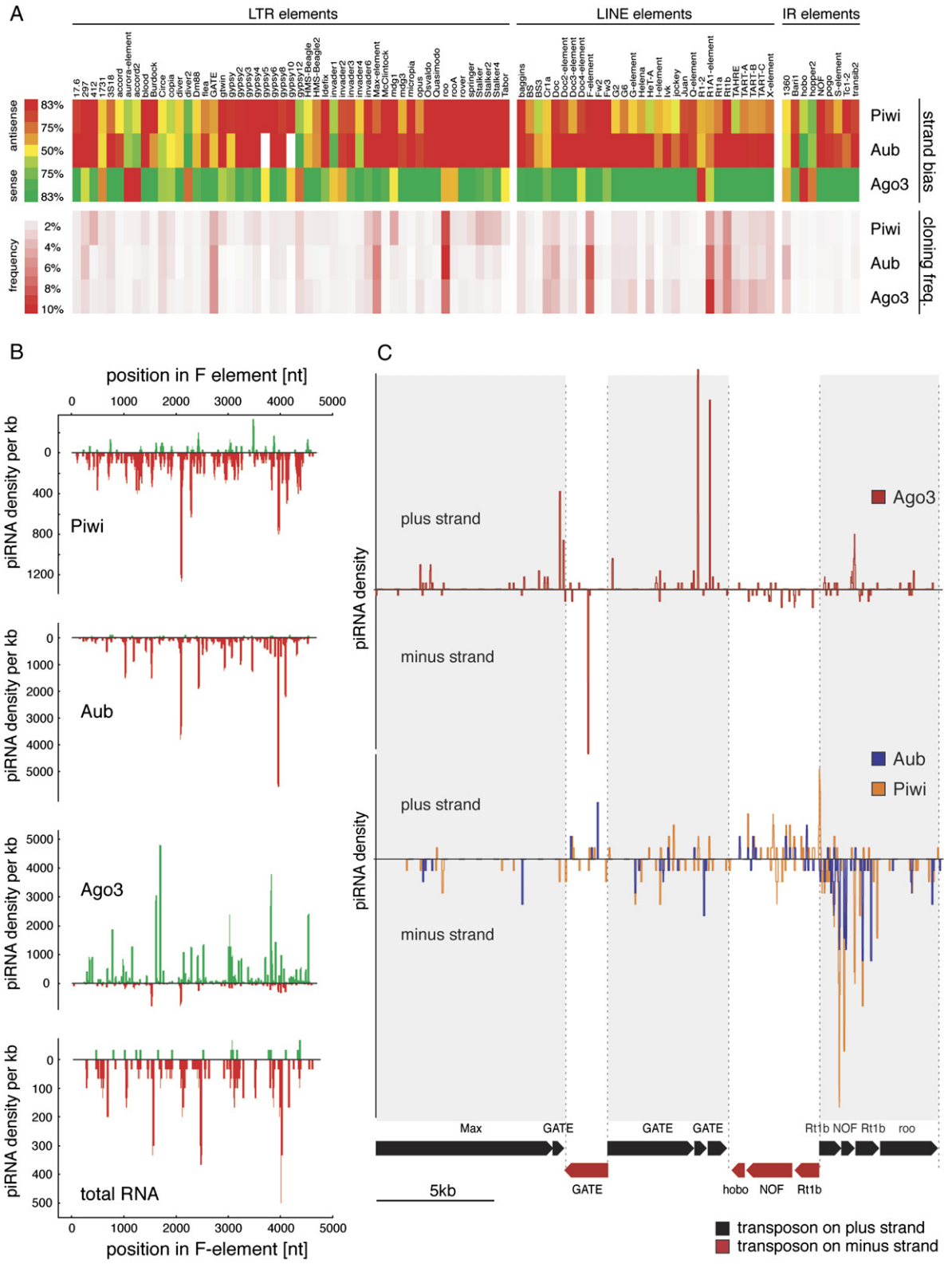


Figure 5. Piwi Complexes Show Profound Strand Biases

(A) Shown in the upper panel is a heat map indicating the strand bias of cloned piRNAs with respect to canonical transposon sequences (indicated at the top). Transposons are grouped into long terminal repeat (LTR) elements, long interspersed nuclear elements (LINE) elements, and inverted repeat

elevation of *gypsy* levels, consistent with a minority of *flamenco* piRNAs entering this complex and with a lack of Aub expression in follicle cells where *gypsy* is expressed. The greater effect of *piwi* compared to *flamenco* mutations is consistent with *flamenco* locus contributing a substantial fraction of but not all *gypsy* repressive piRNAs. In this regard, the *flamenco* cluster has the potential to produce 79% of all piRNAs that target *ZAM*, 30% of those matching *Idefix*, and 33% of piRNAs complementary to *gypsy*.

The P element insertions that we analyzed strongly reduced the abundance of piRNAs generated from sequences up to 168 kb away (Figures 4B and 4F). Considering sequences that map uniquely, *flamenco* produces piRNAs with a marked strand asymmetry that correlates with a strongly biased orientation of transposon fragments in the locus. These observations can be accommodated by a model in which piRNAs are produced from long, unidirectional, precursor transcripts that traverse *flamenco*. Indeed transcripts containing multiple transposons, and several kb of the *flamenco* locus can be easily detected by RT-PCR, and these are lost in *flamenco* mutants (Figure 4G and not shown).

Argonaute3 Binds Sense-Strand piRNAs

Drosophila rasiRNAs show a strong bias for sequences that are antisense to transposable elements (Vagin et al., 2006). We asked whether this observation held for our sequenced piRNAs by examining the strand biases of those derived from Piwi, Aub, and Ago3 complexes. We aligned all piRNA sequences to a comprehensive set of consensus sequences for *D. melanogaster* transposons (canonical set v9.41, Flybase). Piwi and Aub preferentially incorporate piRNAs matching the antisense strand of transposons (76% and 83%, respectively). In contrast, Ago3 complexes contain piRNAs that are strongly biased for sense transposon strands (75%). As piRNAs derived from total RNA retain an antisense bias, Ago3 complexes must be less abundant overall.

The pattern of asymmetry among the three RNPs was preserved when we evaluated each transposable element separately (Figure 5A). As an example, a plot of piRNAs along the F element reveals numerous antisense piRNAs that are loaded into Piwi and Aub and numerous sense piRNAs that enter Ago3 (Figure 5B). There are a few notable exceptions where asymmetry remains marked but is reversed for Piwi/Aub and Ago3 complexes (see for example *accord2*, *gypsy12*, *diver2*, and *hopper2*; Figure 5A).

Unlike *flamenco*, transposons within most piRNA clusters lack an orientation bias. For example, the largest piRNA cluster, at 42AB, contains a high density of randomly oriented, nested transposons and produces uniquely mapping piRNAs from both strands (Figure 5C). Even within this cluster, the strand asymmetry of Piwi complexes is preserved. An interesting illustration is two adjacent *GATE* fragments that are in opposite orientations. Uniquely mapping RNAs in the Ago3 complex correspond to the sense strand of each copy, while Aub, and to some extent Piwi, show the opposite trend (Figure 5C).

Mechanisms of piRNA Biogenesis

To investigate mechanisms of piRNA biogenesis, we examined the relationship between sense and antisense piRNAs. A processing mechanism resembling siRNAs or miRNAs would predict the detection of sense-antisense piRNA pairs that reflect the 2 nt 3′ overhangs produced by RNase III enzymes. According to this scenario, complementary sense and antisense piRNAs should have 5′ ends separated by 23 nt (assuming an average piRNA size of 25 nt) and correspondingly show 23 nt of complementary sequence.

We plotted the distance between each piRNA 5′ end and the 5′ end of its neighbors on the opposite strand. Instead of the expected peak at 23 nucleotides, we found that 5′ ends of complementary piRNAs are most frequently separated by exactly 10 nucleotides (Figure 6A). On the whole, 20% of all piRNAs have a partner whose 5′ end can be mapped ten nucleotides away on the complementary strand. We found the strongest complementary relationships between piRNAs in Ago3 and Aub complexes (Figure 6B). Even though our sequencing efforts are not saturating, more than 48% of small RNAs in the Ago3 library had complementary partners in the Aub library. Statistically significant, although less pronounced, interactions are indicated between Piwi and Ago3, and self complementarity is seen in Aub and Ago3 comparisons.

The 10 nt overlap between sense and antisense piRNAs provoked the hypothesis that the Piwi proteins have a role in piRNA biogenesis. In such a model, an antisense piRNA, derived from a piRNA locus and complexed with Aub or Piwi, would recognize and cleave a sense transposon transcript. This cleavage event would occur opposite nucleotides 10 and 11 of the antisense piRNA, generating a 5′ end 10 nt distant, and on the opposite strand, from the end of the original piRNA (see Figures 6A and 7). The cleaved product would be loaded into a second Piwi family protein, likely Ago3 based upon observed strand biases, ultimately becoming new piRNA after processing at the 3′ end by an unknown mechanism (see Figure 7). piRNAs generated by Piwi-mediated cleavage events are designated as secondary piRNAs.

elements (IR). Color intensities indicate the degree of the strand bias (green: sense; red: antisense; yellow: unbiased). In the lower panel, the cloning frequency for individual transposons in all three complexes is indicated as a heat map. The key for relative cloning frequency is shown at the left.
(B) Shown is the density (sense in green, antisense in red) of all cloned piRNAs assigned to the canonical F element sequence with up to three mismatches. Frequencies in each Piwi-family RNP and in total ovarian RNA are shown individually, as indicated. The relative nucleotide position within the consensus sequence is indicated along the x axes.
(C) Shown is a fragment of the largest piRNA cluster (position 42AB) with only those piRNAs that map uniquely to this region (Ago3 piRNAs in red, Piwi piRNAs in orange, and Aub piRNAs in blue). Top strand piRNAs are shown above the x axes, while bottom strand piRNAs are shown below. Annotated transposon fragments are indicated at the bottom of the panel. Shaded areas mark the boundaries where transposon orientation reverses.

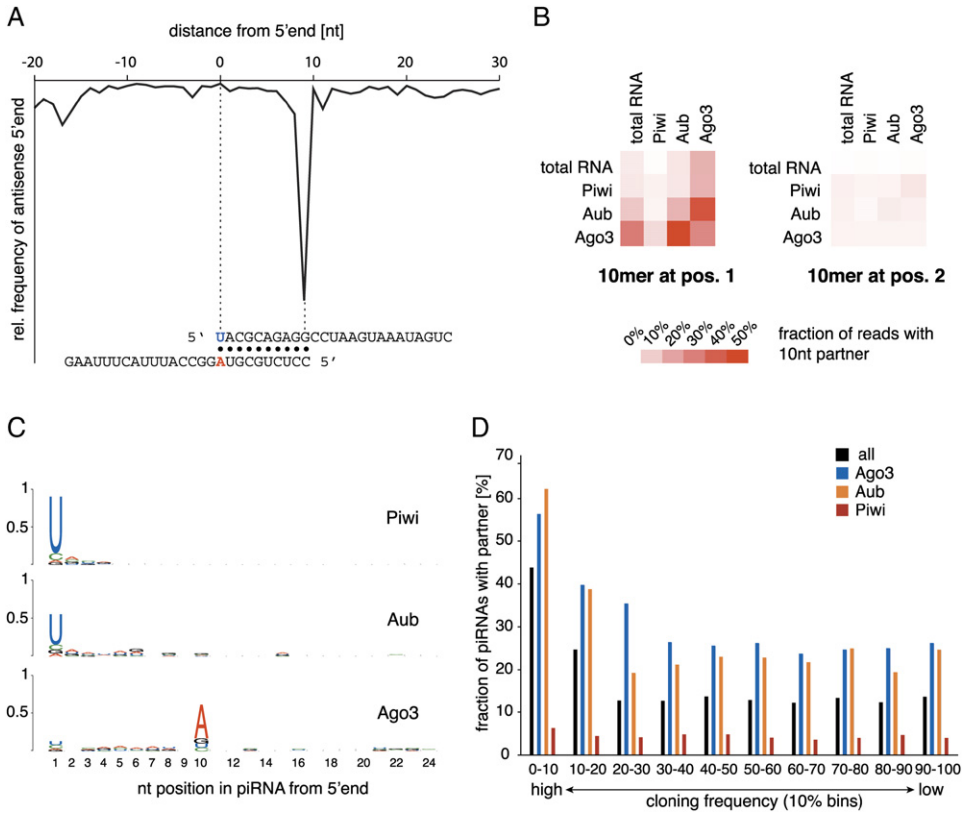


Figure 6. A Slicer-Mediated Amplification Loop for piRNAs

(A) Shown is a frequency map of the separation of piRNAs mapping to opposite genomic strands. The spike at position 9 (the graph starts at 0) indicates the position of maximal probability of finding the 5′ end of a complementary piRNA. One example of this 10 nt offset is shown below.

(B) Shown are heat maps that indicate the degree to which complementary 5′ 10-mers are found in pairwise library comparisons, with a key to the intensity of the signal shown below. The panel to the right represents a control analysis performed with the 10-mer from position 2–11.

(C) The relative nucleotide bias of each position in all piRNAs obtained from Piwi, Aub, and Ago3 complexes, as indicated.

(D) Ten bins were constructed for each Piwi complex (as indicated) and for all sequences combined (all) by sorting piRNAs according to their cloning frequency (e.g., the bin labeled 0–10 contains the 10% of sequences that were most frequently cloned). The fraction of piRNAs within each bin that has a complementary partner was graphed on the y axis.

This model is consistent with the known biochemical properties of Piwi and Argonaute proteins (Lau et al., 2006; Liu et al., 2004; Saito et al., 2006). Moreover, it explains the observed lack of a U bias at the 5′ end of sense-strand piRNAs in Ago3 complexes. If 5′ U-biased antisense piRNAs produce sense piRNAs following cleavage, these should have an unbiased 5′ end. However, sense piRNAs should show an enrichment for A, the complement of the 5′ U, at position 10 (Figure 7). Strikingly, 73% of all Ago3 bound RNAs conform to this prediction, suggesting that the majority of these arise from piRNA-directed cleavage events (Figure 6C).

DISCUSSION

An Amplification Loop that Reinforces Transposon Silencing

In *C. elegans*, effective RNAi depends upon an amplification mechanism (Sijen et al., 2001). Small RNAs from the primary dsRNA trigger are largely dedicated to promoting the use of complementary targets as templates for RNA-dependent RNA polymerases (RdRPs) in the generation of secondary siRNAs. In *Drosophila*, no RdRPs have been identified. However, the ability of Piwi-mediated cleavage to prompt the production of new piRNAs could create an amplification cycle that serves the same purpose as the RdRP-driven secondary siRNA generation systems in worms (Figure 7).

The cycle is initiated by generating primary piRNAs, sampled from the piRNA clusters that we have identified. As these are composed mainly of defective transposon copies, they serve as a genetic memory of transposons to which the population has been exposed. piRNAs that are antisense to expressed, dispersed transposons would identify and cleave their targets, resulting in the genesis of a new, sense piRNA in an Ago3 complex. The Ago3 bound sense piRNA would then seek a target, probably a precursor transcript from a master control locus that contains

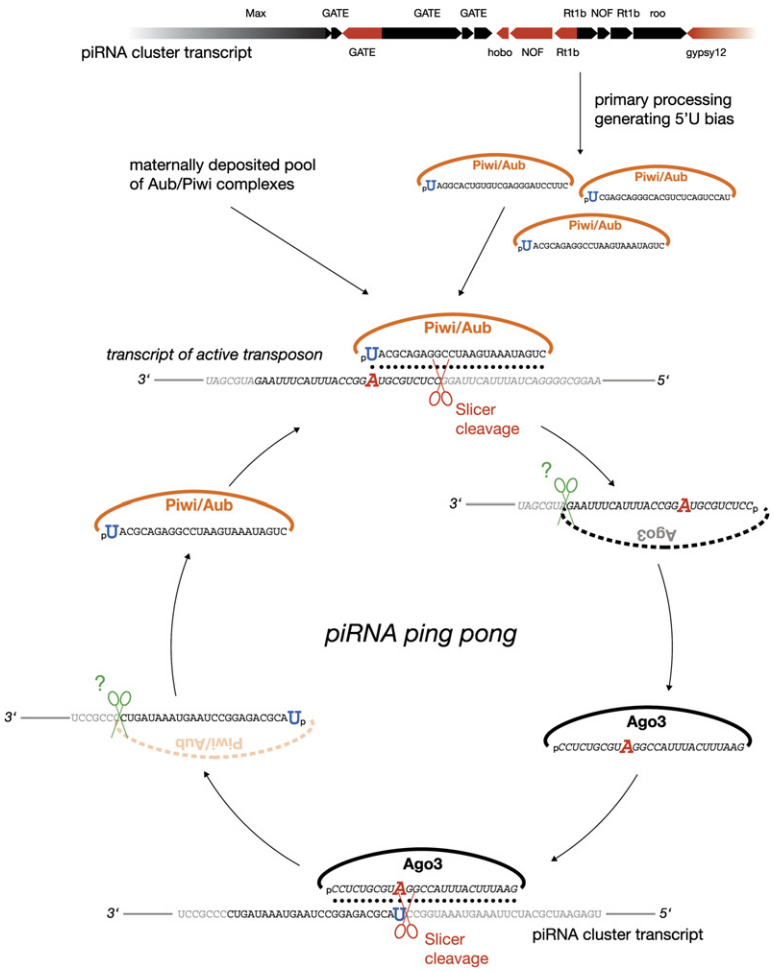


Figure 7. The piRNA Ping-Pong Model

Illustrated is the amplification loop consisting of Piwi/Aub complexes, Ago3 complexes, piRNA cluster transcripts, and transcripts of active transposons. Nucleotide cleavage events are shown as scissors. Potential sources of primary piRNAs are piRNA cluster transcripts and maternally inherited piRNA complexes.

antisense transposon sequences. Ago3-directed cleavage would then generate additional antisense piRNAs capable both of actively silencing their target element and reinforcing the cycle through the creation of additional sense piRNAs.

The existence of such an amplification cycle essentially permits sense and antisense piRNAs act in concert to increase production of silencing-competent RNAs in response to the activity of individual transposons. Since Argonautes act catalytically, a significant amplification of the response could be achieved by even a relatively low level of sense piRNAs in Ago3 complexes. This model predicts that piRNAs participating in this process, namely those with complementary partners, should be more abundant than piRNAs without detectable partners. In accord with this hypothesis, the most frequently cloned Aub and Ago3-associated piRNAs show an increased probability of having partners within the data set (Figure 6D). Interestingly, Piwi-associated RNAs do not follow this pattern. Since the amplification cycle consumes target transposon transcripts as part of its mechanism, post-transcriptional gene silencing may be sufficient to explain transposon repression. However, we cannot rule out the possibility that transcriptional silencing may also be triggered by Piwi-family RNPs.

The amplification cycle may not be absolutely essential for silencing of all elements, as loci such as *flamenco* may operate in somatic follicle cells where the absence of Aub and Ago3 forces it to act in a stoichiometric fashion. In this regard, *flamenco* is unusual in that the vast majority of transposon fragments within this locus exist in a common orientation, which can lead to the production of antisense primary piRNAs given a long, unidirectional, precursor transcript.

Origins of Strand Bias in piRNA Populations

In contrast to *flamenco*, most piRNA loci appear to be both bidirectionally transcribed and contain transposon sequences in random orientation. Nevertheless, the marked asymmetry of Piwi/Aub and Ago3 complexes is conserved in piRNAs that uniquely map to clusters (e.g., Figure 5C). Among piRNAs that match transposons, 77% and 79% of unambiguously cluster-derived Piwi- and Aub-associated RNAs are antisense, while 73% of those in Ago3 are sense. These observations strongly suggest that piRNA clusters themselves participate in

an amplification cycle in a manner that informs strand choice. According to our model, the remarkable strand asymmetry in piRNA populations hinges on informative interactions between master control loci and active transposons, which by their nature produce sense RNAs. Our observations identify Ago3 as the principal recipient of piRNAs derived from transposon mRNAs. Thus, as long as there is an input to the system from active transposon transcripts via Ago3 and a preferential relationship between Ago3 and Aub for generating secondary piRNAs in their reciprocal complexes, a strand bias can be maintained even if primary and secondary piRNAs can both be derived from master control loci.

Biogenesis of Primary piRNAs

The amplification cycle must be initiated by primary piRNAs. Presently, the biogenesis pathway that generates primary piRNAs from piRNA clusters remains obscure. Our data suggest that the piRNA precursor is a long, single-stranded transcript that is cleaved, preferentially at U residues. We detect transcripts from piRNA loci by RT-PCR that encompass multiple transposon fragments (not shown) and find numerous small RNAs that cross junctions between adjacent transposons. In the case of *flamenco*, P element insertions near the proximal end of the locus have a polar effect both on these long RNA transcripts and on *flamenco* piRNAs.

Equally mysterious is the generation of piRNA 3′ ends. Mature piRNAs could arise by two cleavage events and subsequent loading into Piwi complexes. Alternatively, piRNAs could be created following 5′ end formation and incorporation of a long RNA into Piwi by resection of their 3′ ends. The latter model is attractive, since it could provide an explanation for observed size differences between RNAs bound to individual Piwi proteins, as piRNA size would simply reflect the footprint of each Piwi protein.

Although de novo biogenesis mechanisms must exist, maternally inherited piRNA complexes could also serve to initiate the amplification cycle. All three Piwi proteins are loaded into the developing oocyte (Figure S2) (Harris and Macdonald, 2001; Megosh et al., 2006), and Piwi and Aub are concentrated in the pole plasm, which will give rise to the germline of the next generation. Inherited piRNAs could enhance resistance to transposons that are an ongoing challenge to the organism, augmenting zygotic production of primary piRNAs. Indeed, maternally loaded rasiRNAs were detected in early embryos (Aravin et al., 2003), and their presence was correlated with suppression of hybrid dysgenesis in *D. virilis* (Blumenstiel and Hartl, 2005).

A Model for Transposon Silencing in *Drosophila*

Our data point to a comprehensive strategy for transposon repression in *Drosophila* that incorporates both a long-term genetic memory and an acute response to the presence of potentially active elements in the genome. The model that emerges from our studies shows many parallels to adaptive immune systems. The piRNA loci themselves encode a diversity of small RNA fragments that have the potential to recognize invading parasitic genetic elements. Throughout the evolution of *Drosophila* species, a record of transposon exposure may have been preserved by selection for transposition events into these master control loci, as this is one key mechanism through which control over a specific element can be achieved. Evidence has already emerged that X-TAS can act as a transposition hotspot for P elements (Karpen and Spradling, 1992), raising the possibility the piRNA clusters in general may attract transposons. Once an element enters a piRNA locus, it can act, in *trans*, to silencing remaining elements in the genome, either directly through primary piRNAs or by engaging in the amplification model described above. A comparison of *D. melanogaster* piRNAs to transposons present in related *Drosophilids* shows a lack of complementarity when comparisons are made at high stringency. However, when even a few mismatches are permitted, it is clear that piRNA loci might have some limited potential to protect against horizontal transmission of these heterologous elements (Figure S3). The existence of a feed-forward amplification loop can be compared to clonal expansion of immune cells with the appropriate specificity following antigen stimulation, leading to a robust and adaptable response.

EXPERIMENTAL PROCEDURES

Antibodies

Rabbit polyclonal antisera directed against the N-terminal 14–16 AA of Piwi, Aub, and Ago3 (see Figure S1) were raised and affinity purified as described (Denli et al., 2004). Primary antibody dilutions were 1:2000 for western analysis and 1:500 for immunohistochemistry. Actin staining was with Rhodamine-coupled Phalloidin (1:100). Ovaries were dissected on ice (phosphate-buffered saline) and fixed for 20 min in 4% Formaldehyde/PBS/0.1%Triton X-100.

Immunoprecipitation of Piwi RNP Complexes

Ovary extract was prepared in Lysis buffer (20 mM HEPES-NaOH at pH 7.0, 150 mM NaCl, 2.5 mM MgCl2, 250 mM Sucrose, 0.05% NP40, 0.5% Triton X-100, 1x Roche-Complete). Cleared extracts were incubated with primary antibodies (1:50) for 4 hr at 4°C. Antibody complexes were isolated and analyzed for RNA content as described (Aravin et al., 2006).

Small RNA Cloning and Sequencing

Small RNA cloning was performed as described in Pfeffer et al. (2005), and a detailed protocol is available upon request. Sequencing of small RNA libraries was at 454 Life Sciences.

Quantitative RT-PCR

See Supplemental Data for primer sequences and for the amplification of primary cluster transcripts, *gypsy* transcript, and piRNAs.

Bioinformatic Analysis of piRNAs

Sequence extraction and genomic mapping was as described in Girard et al. (2006). We used the Release 5 assembly of the *Drosophila melanogaster* genome and GenBank to identify all piRNAs matching 100% to genomic sequences. The only GenBank entry (L03284) that recovered hits not present in Release 5 corresponds to the tip of the X chromosome, which differs significantly between the sequenced strain and Oregon R, the strain used for our analysis (Abad et al., 2004). Annotation of small RNAs was done using the following

databases: Repbase (http://www.girinst.org/) on the Release 5 assembly, Transposable element canonical sequences (http://www.fruitfly.org/p_disrupt/TE.html), Flybase annotations for coding and noncoding genes (extracted from http://genome.ucsc.edu), and microRNA annotations from Rfam (http://microrna.sanger.ac.uk/sequences). Density analysis of transposons along Release 5 chromosome arms was by counting all nucleotides in a 50 kb window (10 kb offset) annotated as transposons.

piRNA Cluster Analysis

All piRNAs except the 10% of reads corresponding to miRNAs, rRNAs, tRNAs, other ncRNAs, and the sense strand of annotated genes were mapped to Release 5 and the telomeric X-TAS repeat L03284. Nucleotides corresponding to the 5′ end of each piRNA were weighted according to N/M with N = cloning frequency and M = number of genomic mappings. We used a 5 kb sliding window to identify all regions with densities greater than 1 piRNA/kb. Windows within 20 kb of each other were collapsed into clusters. Clusters with at least 5 piRNAs that uniquely matched to the cluster were retained.

Analysis of piRNAs Mapping to Transposable Elements

All piRNAs were matched to the canonical sequences of *Drosophila* transposable elements with high (0 mismatches), medium (3 mm), or low (5 mm) stringency, and the strand relative to the transposon sense strand was determined. We calculated the ratio of all piRNAs per library that match exclusively to the plus or minus strand and excluded those that matched to both (e.g., in inverted repeat [IR] elements). The relative density of piRNAs for each transposon is the number of piRNAs matching a specific element as fraction of all transposon-matching piRNAs.

10 nt Offset Analysis and Nucleotide Bias

The relative relationship of each piRNA to its nearest neighbor was calculated as described in the text, except that all genomic matches were weighted according to the normalized model to avoid skewing by piRNAs that match the genome multiple times. For all subsequent analyses, partners were defined as piRNAs, whose 5′ 10 nt are reverse complements. For Figure 6D we sorted all piRNAs in each library according to their cloning frequency and determined the fraction of piRNAs with partners in bins, each of which contained 10% of all sequences. Position-dependent nt biases for each library were determined by their log-odds score relative to library-specific nt compositions.

Supplemental Data

Supplemental Data include four figures, two table, Supplemental Experimental Procedures, and Supplemental References and can be found with this article online at http://www.cell.com/cgi/content/full/128/6/1089/DC1/.

ACKNOWLEDGMENTS

We thank members of the Hannon laboratory for helpful discussions and support. Ahmet Denli (CSHL) provided antibodies, and Stuart Shuman (MSKCC) provided the expression vector for Rnl2. We thank Andrew Olson and Ted Roeder for bioinformatic support and Stephen Hearn (CSHL) for assistance with confocal microscopy. J.B. and A.S. are supported by fellowships from the Ernst Schering Foundation. A.A. is supported by a fellowship from the CSHL Association. M.D. is an Engelhorn Scholar of the Watson School of Biological Sciences. G.J.H. is an investigator of the Howard Hughes Medical Institute. This work was supported by grants from the N.I.H. (G.J.H.).

Received: October 23, 2006
Revised: December 8, 2006
Accepted: January 19, 2007
Published online: March 8, 2007

REFERENCES

Abad, J.P., De Pablos, B., Osoegawa, K., De Jong, P.J., Martin-Gallardo, A., and Villasante, A. (2004). Genomic analysis of Drosophila melanogaster telomeres: full-length copies of HeT-A and TART elements at telomeres. Mol. Biol. Evol. *21*, 1613–1619.

Aravin, A.A., Naumova, N.M., Tulin, A.V., Vagin, V.V., Rozovsky, Y.M., and Gvozdev, V.A. (2001). Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline. Curr. Biol. *11*, 1017–1027.

Aravin, A.A., Lagos-Quintana, M., Yalcin, A., Zavolan, M., Marks, D., Snyder, B., Gaasterland, T., Meyer, J., and Tuschl, T. (2003). The small RNA profile during Drosophila melanogaster development. Dev. Cell *5*, 337–350.

Aravin, A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., Morris, P., Brownstein, M.J., Kuramochi-Miyagawa, S., Nakano, T., et al. (2006). A novel class of small RNAs bind to MILI protein in mouse testes. Nature *442*, 203–207.

Biemont, C., Ronsseray, S., Anxolabehere, D., Izaabel, H., and Gautier, C. (1990). Localization of P elements, copy number regulation, and cytotype determination in Drosophila melanogaster. Genet. Res. *56*, 3–14.

Blumenstiel, J.P., and Hartl, D.L. (2005). Evidence for maternally transmitted small interfering RNA in the repression of transposition in Drosophila virilis. Proc. Natl. Acad. Sci. USA *102*, 15965–15970.

Bregliano, J.C., Picard, G., Bucheton, A., Pelisson, A., Lavige, J.M., and L'Heritier, P. (1980). Hybrid dysgenesis in Drosophila melanogaster. Science *207*, 606–611.

Bucheton, A. (1990). I transposable elements and I-R hybrid dysgenesis in Drosophila. Trends Genet. *6*, 16–21.

Bucheton, A. (1995). The relationship between the flamenco gene and gypsy in Drosophila: how to tame a retrovirus. Trends Genet. *11*, 349–353.

Bucheton, A., Paro, R., Sang, H.M., Pelisson, A., and Finnegan, D.J. (1984). The molecular basis of I-R hybrid dysgenesis in Drosophila melanogaster: identification, cloning, and properties of the I factor. Cell *38*, 153–163.

Carmell, M.A., Xuan, Z., Zhang, M.Q., and Hannon, G.J. (2002). The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. Genes Dev. *16*, 2733–2742.

Castro, J.P., and Carareto, C.M. (2004). Drosophila melanogaster P transposable elements: mechanisms of transposition and regulation. Genetica *121*, 107–118.

Chen, P.Y., Manninga, H., Slanchev, K., Chien, M., Russo, J.J., Ju, J., Sheridan, R., John, B., Marks, D.S., Gaidatzis, D., et al. (2005). The developmental miRNA profiles of zebrafish as determined by small RNA cloning. Genes Dev. *19*, 1288–1293.

Cox, D.N., Chao, A., Baker, J., Chang, L., Qiao, D., and Lin, H. (1998). A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal. Genes Dev. *12*, 3715–3727.

Cox, D.N., Chao, A., and Lin, H. (2000). piwi encodes a nucleoplasmic factor whose activity modulates the number and division rate of germline stem cells. Development *127*, 503–514.

Deng, W., and Lin, H. (2002). miwi, a murine homolog of piwi, encodes a cytoplasmic protein essential for spermatogenesis. Dev. Cell *2*, 819–830.

Denli, A.M., Tops, B.B., Plasterk, R.H., Ketting, R.F., and Hannon, G.J. (2004). Processing of primary microRNAs by the Microprocessor complex. Nature *432*, 231–235.

Desset, S., Meignin, C., Dastugue, B., and Vaury, C. (2003). COM, a heterochromatic locus governing the control of independent endogenous retroviruses from Drosophila melanogaster. Genetics *164*, 501–509.

Engels, W.R., and Preston, C.R. (1979). Hybrid dysgenesis in Drosophila melanogaster: the biology of female and male sterility. Genetics *92*, 161–174.

Girard, A., Sachidanandam, R., Hannon, G.J., and Carmell, M.A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature *442*, 199–202.

Grivna, S.T., Pyhtila, B., and Lin, H. (2006). MIWI associates with translational machinery and PIWI-interacting RNAs (piRNAs) in regulating spermatogenesis. Proc. Natl. Acad. Sci. USA *103*, 13415–13420.

Hamilton, A.J., and Baulcombe, D.C. (1999). A species of small antisense RNA in posttranscriptional gene silencing in plants. Science *286*, 950–952.

Harris, A.N., and Macdonald, P.M. (2001). Aubergine encodes a Drosophila polar granule component required for pole cell formation and related to eIF2C. Development *128*, 2823–2832.

Hoskins, R.A., Smith, C.D., Carlson, J.W., Carvalho, A.B., Halpern, A., Kaminker, J.S., Kennedy, C., Mungall, C.J., Sullivan, B.A., Sutton, G.G., et al. (2002). Heterochromatic sequences in a Drosophila whole-genome shotgun assembly. Genome Biol. *3*, RESEARCH0085.

Kalmykova, A.I., Klenov, M.S., and Gvozdev, V.A. (2005). Argonaute protein PIWI controls mobilization of retrotransposons in the Drosophila male germline. Nucleic Acids Res. *33*, 2052–2059.

Karpen, G., and Spradling, A. (1992). Analysis of subtelomeric heterochromatin in the Drosophila minichromosome Dp1187 by single P element insertional mutagenesis. Genetics *132*, 737–753.

Ketting, R.F., Haverkamp, T.H., van Luenen, H.G., and Plasterk, R.H. (1999). Mut-7 of C. elegans, required for transposon silencing and RNA interference, is a homolog of Werner syndrome helicase and RNaseD. Cell *99*, 133–141.

Kidwell, M.G., Kidwell, J.F., and Sved, J.A. (1977). Hybrid dysgenesis in drosophila melanogaster: a syndrome of aberrant traits including mutation, sterility and male recombination. Genetics *86*, 813–833.

Kuramochi-Miyagawa, S., Kimura, T., Ijiri, T.W., Isobe, T., Asada, N., Fujita, Y., Ikawa, M., Iwai, N., Okabe, M., Deng, W., et al. (2004). Mili, a mammalian member of piwi family gene, is essential for spermatogenesis. Development *131*, 839–849.

Lau, N.C., Seto, A.G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D.P., and Kingston, R.E. (2006). Characterization of the piRNA complex from rat testes. Science *313*, 363–367.

Lin, H., and Spradling, A.C. (1997). A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the Drosophila ovary. Development *124*, 2463–2476.

Liu, J., Carmell, M.A., Rivas, F.V., Marsden, C.G., Thomson, J.M., Song, J.J., Hammond, S.M., Joshua-Tor, L., and Hannon, G.J. (2004). Argonaute2 is the catalytic engine of mammalian RNAi. Science *305*, 1437–1441.

Marin, L., Lehmann, M., Nouaud, D., Izaabel, H., Anxolabehere, D., and Ronsseray, S. (2000). P-Element repression in Drosophila melanogaster by a naturally occurring defective telomeric P copy. Genetics *155*, 1841–1854.

Megosh, H.B., Cox, D.N., Campbell, C., and Lin, H. (2006). The Role of PIWI and the miRNA Machinery in Drosophila Germline Determination. Curr. Biol. *16*, 1884–1894.

Misra, S., and Rio, D.C. (1990). Cytotype control of Drosophila P element transposition: the 66 kd protein is a repressor of transposase activity. Cell *62*, 269–284.

Pal-Bhadra, M., Bhadra, U., and Birchler, J.A. (1997). Cosuppression in Drosophila: gene silencing of Alcohol dehydrogenase by white-Adh transgenes is Polycomb dependent. Cell *90*, 479–490.

Pal-Bhadra, M., Bhadra, U., and Birchler, J.A. (2002). RNAi related mechanisms affect both transcriptional and posttranscriptional transgene silencing in Drosophila. Mol. Cell *9*, 315–327.

Pelisson, A. (1981). The I–R system of hybrid dysgenesis in Drosophila melanogaster: are I factor insertions responsible for the mutator effect of the I–R interaction? Mol. Gen. Genet. *183*, 123–129.

Pelisson, A., and Bregliano, J.C. (1987). Evidence for rapid limitation of the I element copy number in a genome submitted to several generations of I-R hybrid dysgenesis in Drosophila melanogaster. Mol. Gen. Genet. *207*, 306–313.

Pelisson, A., Song, S.U., Prud'homme, N., Smith, P.A., Bucheton, A., and Corces, V.G. (1994). Gypsy transposition correlates with the production of a retroviral envelope-like protein under the tissue-specific control of the Drosophila flamenco gene. EMBO J. *13*, 4401–4411.

Pfeffer, S., Sewer, A., Lagos-Quintana, M., Sheridan, R., Sander, C., Grasser, F.A., van Dyk, L.F., Ho, C.K., Shuman, S., Chien, M., et al. (2005). Identification of microRNAs of the herpesvirus family. Nat. Methods *2*, 269–276.

Prud'homme, N., Gans, M., Masson, M., Terzian, C., and Bucheton, A. (1995). Flamenco, a gene controlling the gypsy retrovirus of Drosophila melanogaster. Genetics *139*, 697–711.

Reiss, D., Josse, T., Anxolabehere, D., and Ronsseray, S. (2004). aubergine mutations in Drosophila melanogaster impair P cytotype determination by telomeric P elements inserted in heterochromatin. Mol. Genet. Genomics *272*, 336–343.

Robert, V., Prud'homme, N., Kim, A., Bucheton, A., and Pelisson, A. (2001). Characterization of the flamenco region of the Drosophila melanogaster genome. Genetics *158*, 701–713.

Roche, S.E., and Rio, D.C. (1998). Trans-silencing by P elements inserted in subtelomeric heterochromatin involves the Drosophila Polycomb group gene, Enhancer of zeste. Genetics *149*, 1839–1855.

Ronsseray, S., Lehmann, M., and Anxolabehere, D. (1991). The maternally inherited regulation of P elements in Drosophila melanogaster can be elicited by two P copies at cytological site 1A on the X chromosome. Genetics *129*, 501–512.

Ronsseray, S., Marin, L., Lehmann, M., and Anxolabehere, D. (1998). Repression of hybrid dysgenesis in Drosophila melanogaster by combinations of telomeric P-element reporters and naturally occurring P elements. Genetics *149*, 1857–1866.

Rubin, G.M., Kidwell, M.G., and Bingham, P.M. (1982). The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations. Cell *29*, 987–994.

Saito, K., Nishida, K.M., Mori, T., Kawamura, Y., Miyoshi, K., Nagami, T., Siomi, H., and Siomi, M.C. (2006). Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the Drosophila genome. Genes Dev. *20*, 2214–2222.

Sarot, E., Payen-Groschene, G., Bucheton, A., and Pelisson, A. (2004). Evidence for a piwi-dependent RNA silencing of the gypsy endogenous retrovirus by the Drosophila melanogaster flamenco gene. Genetics *166*, 1313–1321.

Savitsky, M., Kwon, D., Georgiev, P., Kalmykova, A., and Gvozdev, V. (2006). Telomere elongation is under the control of the RNAi-based mechanism in the Drosophila germline. Genes Dev. *20*, 345–354.

Sijen, T., Fleenor, J., Simmer, F., Thijssen, K.L., Parrish, S., Timmons, L., Plasterk, R.H., and Fire, A. (2001). On the role of RNA amplification in dsRNA-triggered gene silencing. Cell *107*, 465–476.

Smyth, D.R. (1997). Gene silencing: cosuppression at a distance. Curr. Biol. *7*, R793–R795.

Stuart, J.R., Haley, K.J., Swedzinski, D., Lockner, S., Kocian, P.E., Merriman, P.J., and Simmons, M.J. (2002). Telomeric P elements associated with cytotype regulation of the P transposon family in Drosophila melanogaster. Genetics *162*, 1641–1654.

Tabara, H., Sarkissian, M., Kelly, W.G., Fleenor, J., Grishok, A., Timmons, L., Fire, A., and Mello, C.C. (1999). The rde-1 gene, RNA interference, and transposon silencing in C. elegans. Cell *99*, 123–132.

Vagin, V.V., Sigova, A., Li, C., Seitz, H., Gvozdev, V., and Zamore, P.D. (2006). A distinct small RNA pathway silences selfish genetic elements in the germline. Science *313*, 320–324.

Accession Numbers

The GenBank accession number for Ago3 in this paper is EF211827. The GEO accession numbers for all piRNAs from this study are GSE6734, GSM154618, GSM154620, GSM154621, and GSM154622.

Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites

Xiaohui Xie[†], Tarjei S. Mikkelsen[†‡], Andreas Gnirke[†], Kerstin Lindblad-Toh[†], Manolis Kellis[†§], and Eric S. Lander[†¶‖††]

[†]Broad Institute of MIT and Harvard, Massachusetts Institute of Technology and Harvard Medical School, Cambridge, MA 02142; [‡]Division of Health Sciences and Technology, [§]Computer Science and Artificial Intelligence Laboratory, and [‖]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139; and [¶]Whitehead Institute for Biomedical Research, Cambridge, MA 02142

Contributed by Eric S. Lander, March 3, 2007 (sent for review January 26, 2007)

Conserved noncoding elements (CNEs) constitute the majority of sequences under purifying selection in the human genome, yet their function remains largely unknown. Experimental evidence suggests that many of these elements play regulatory roles, but little is known about regulatory motifs contained within them. Here we describe a systematic approach to discover and characterize regulatory motifs within mammalian CNEs by searching for long motifs (12–22 nt) with significant enrichment in CNEs and studying their biochemical and genomic properties. Our analysis identifies 233 long motifs (LMs), matching a total of ≈60,000 conserved instances across the human genome. These motifs include 16 previously known regulatory elements, such as the histone 3′-UTR motif and the neuron-restrictive silencer element, as well as striking examples of novel functional elements. The most highly enriched motif (LM1) corresponds to the X-box motif known from yeast and nematode. We show that it is bound by the RFX1 protein and identify thousands of conserved motif instances, suggesting a broad role for the RFX family in gene regulation. A second group of motifs (LM2*) does not match any previously known motif. We demonstrate by biochemical and computational methods that it defines a binding site for the CTCF protein, which is involved in insulator function to limit the spread of gene activation. We identify nearly 15,000 conserved sites that likely serve as insulators, and we show that nearby genes separated by predicted CTCF sites show markedly reduced correlation in gene expression. These sites may thus partition the human genome into domains of expression.

comparative genomics | conserved noncoding element

Comparative analysis of the human and several other mammalian genomes has revealed that 5% of the human genome is under purifying selection, with less than one-third of the sequences under selection encoding proteins. The vast majority lies in hundreds of thousands of conserved noncoding elements (CNEs). The functional significance of these CNEs is largely unknown. It seems likely that many are involved in gene regulation, and transgenic experiments have identified some CNEs that are capable of driving highly specific spatiotemporal gene expression patterns (1–4). However, little is known about regulatory motifs contained within CNEs or proteins that recognize these elements.

We and others have previously undertaken large-scale efforts to discover conserved motifs in limited subsets of the human genome (5–8), specifically, gene promoters and 3′-UTRs. The approach has been to search for motifs that are preferentially conserved in these regions by using syntenic alignments of human, mouse, rat, and dog sequences (5). Using this approach we have discovered 174 motifs in promoter regions (within 2 kb of the transcriptional start), most of which are involved in transcriptional regulation and in tissue-specific gene expression control, and 105 motifs in 3′-UTRs, implicated in posttranscriptional regulation with half related to microRNA targeting. These studies were limited in scope because gene promoters and 3′-UTRs contain only a small fraction (≈6%) of the CNEs in the genome. In addition, they were limited in power because they involved comparison with only three non-human mammals.

Here we use the recent availability of sequences of 12 mammalian genomes to extend our motif discovery efforts to the entire human genome. We focus specifically on long regulatory motifs, between 12 and 22 nt, which provide a strong signal for motif discovery. We searched for motifs that are enriched in CNE regions relative to the rest of the genome.

We discovered >200 motifs showing striking enrichment in CNE regions. The analysis automatically rediscovered a dozen previously known regulatory elements. More importantly, most of the discovered motifs are new and show properties distinct from typical promoter elements. In particular, one of the novel motifs defines ≈15,000 potential insulator elements in the human genome, highlighting the diverse role of the CNEs in gene regulation.

Results

Creating a Motif Catalog. We began by compiling a data set of 829,730 CNEs in the human genome (totaling 62 Mb or ≈2% of the euchromatic genome), consisting of sequences showing strong conservation in syntenic regions in comparisons involving 12 mammalian genomes [see supporting information (SI) *Text*]. The vast majority of these elements are located at a considerable distance from the transcriptional start sites (TSS) of protein-coding genes (SI Fig. 4). Approximately 95% are located >2 kb away from the TSS of any gene, and half are >100 kb from a TSS. This suggests that only a small portion of the CNEs serve functions specific to core or proximal promoters.

We sought to create a catalog of sequence motifs enriched in the CNEs (SI Fig. 5). We began by identifying k-mers (for $k \geq 12$) that occur at a significantly higher frequency in the CNE sequences than in the remainder of the genome. We focused only on relatively long k-mers, because the expected number N of random occurrences in the entire CNE database is small (for example, $n < 8$ for $k = 12$; SI Fig. 6). We identified a total of 69,810 enriched k-mers. An example is 5′-GTTGCCATGGAAAC-3′, which appears 698 times in the CNE data set, whereas only 27 sites are expected based on its genome-wide frequency (26-fold enrichment). We noticed that many of enriched k-mers were closely related; therefore, we clustered them based on sequence similarity. The 69,810 enriched k-mers collapsed into 233 distinct groups, denoted LM1, LM2, etc.

Author contributions: X.X. and E.S.L. designed research; X.X. and E.S.L. performed research; X.X., T.S.M., A.G., K.L.-T., M.K., and E.S.L. contributed new reagents/analytic tools; X.X., T.S.M., A.G., K.L.-T., M.K., and E.S.L. analyzed data; and X.X., T.S.M., A.G., K.L.-T., M.K., and E.S.L. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Abbreviations: CNE, conserved noncoding element; TSS, transcriptional start site; PWM, positional weight matrix; NRSE, neuron-restrictive silencer element.

[††]To whom correspondence should be addressed. E-mail: lander@broad.mit.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0701811104/DC1.

© 2007 by The National Academy of Sciences of the USA

Table 1. List of top 50 most highly enriched motifs

ID	*k*-mer sequence	No. of allowed mismatches	No. of initial sites in CNEs	Fold enrichment	*Z*-score	Known motif
LM1	GTTGCCATGGAAAC	1	698	25.9	130.0	X-box
LM2	ACCACTAGATGGCA	1	305	22.9	80.4	
LM3	GTTGCTAGGCAACC	1	204	30.7	76.9	
LM4	GCCTGCTGGGAGTTGTAGTT	3	143	26.3	59.2	
LM5	AACTCCCATTAGCGTTAATGG	3	43	68.1	53.5	
LM6	AAAGGCCCTTTTAAGGGCCAC	3	48	46.2	46.3	Histone 3′-UTR
LM7	CAGCAGATGGCGCTGTT	2	97	22.1	44.4	
LM8	ATGAATTATTCATG	1	280	8.8	44.3	
LM9	TCAGCACCACGGACAG	1	82	25.6	44.2	NRSE
LM10	CTGTTTCCTTGGAAACCAG	3	165	9.3	35.2	
LM11	GAAATGCTGACAGACCCTTAA	3	41	30.7	34.5	
LM12	TGGCCTGAAAGAGTTAATGCA	3	51	22.8	32.7	
LM13	TGCTAATTAGCA	0	82	13.1	30.4	CHX10
LM14	ATCCAGATGTTTGGCA	1	33	27.0	28.9	RP58
LM15	CATTTGCATGCAAATGA	2	124	8.5	28.8	
LM16	TTGAGATCCTTAGATGAAAG	3	64	14.6	28.6	
LM17	CATCTGGTTTGCAT	1	117	8.8	28.5	
LM18	CATTTGCATCTGATTTGCAT	3	80	11.8	28.2	
LM19	TGCTAATTAGCAGC	1	88	10.8	28.1	
LM20	TGACAGCTGTCAAA	1	118	8.5	28.0	
LM21	ATTTGCATCTCATTTGC	2	123	8.2	27.9	
LM22	CAGCTGTTAAACAGCTG	2	80	11.4	27.6	
LM23	AGCACCACCTGGTGGTA	2	65	13.4	27.5	
LM24	AGAACAGATGGC	0	70	12.1	26.8	TAL1BETAITF2
LM25	AAAAGCAATTTCCT	1	202	5.3	26.7	
LM26	TAAACACAGCTG	0	83	10.2	26.3	
LM27	CATTTGCATCTCATTAGCA	3	110	8.0	26.1	
LM28	AGAACATCTGTTTC	1	144	6.3	25.5	
LM29	GCTAATTGCAAATG	1	98	8.4	25.3	
LM30	CTTTGAAATGTCAA	1	182	5.3	25.3	
LM31	CTTTTCATCTTCAAAGCACTT	3	57	13.0	25.2	
LM32	CTGACATTTCCAAA	1	174	5.4	25.0	
LM33	GTAATTGGAAACAGCTG	2	69	10.7	24.8	
LM34	GATTTGCATTGCAAATG	2	84	8.8	24.1	
LM35	ACTTCAAAGGGAGC	1	87	8.5	24.1	
LM36	GAAATGCAATTTGC	1	125	6.4	24.1	
LM37	ATGCAAATGAGCCC	1	85	8.5	23.9	
LM38	GCAAATTAGCAGCT	1	82	8.5	23.4	
LM39	GTCTCCTAGGAAAC	1	84	8.4	23.4	
LM40	TCCCATTGACTTCAATGGGA	3	44	14.2	23.4	
LM41	TTTGAAATGCTAATG	1	80	8.6	23.2	
LM42	AAGCCTAATTAGCA	1	69	9.6	23.1	
LM43	CAGGAAATGAAA	0	141	5.6	23.1	
LM44	GTGTAATTGGAAACAGCTG	3	75	8.9	23.0	
LM45	GCTAATTGGATTTG	1	76	8.7	22.9	
LM46	AACAGCTGTTGAAA	1	128	5.9	22.9	
LM47	AGAGTGCCACCTACTGAAT	3	65	9.8	22.7	
LM48	TAATGAGCTCATTA	1	108	6.5	22.6	
LM49	GTAATTAGCAGCTG	1	68	9.3	22.5	
LM50	TGGGTAATTACATTCTG	2	65	9.6	22.5	

for "long motif." For each of these motifs we derived a positional weight matrix (PWM) representation reflecting the distribution of 4 nt at each position. The enrichment of each motif in the CNE data set was expressed in an enrichment score (see *Methods*). The top 50 motifs are shown in Table 1, and a full list of 233 motifs is given in SI Table 3. The motifs range in size from 12 to 22 bases.

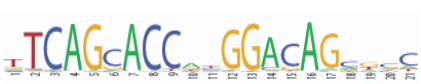
For each of 233 discovered motifs, we searched the entire human genome to identify conserved instances; that is, we identified all human sites matching the PWM and then found those sites that show clear cross-species conservation (see *SI Text*). We found a total of 60,019 conserved instances, with roughly half residing within the CNE data set and roughly half in the remainder of the genome. Importantly, the approach of focusing on motifs enriched in the CNE data set identified many motif instances elsewhere in the genome.

To assess the significance of these results, the procedure was repeated with matched control motifs. For each of the 233 motifs we created a control motif by permuting the columns of the PWM while preserving the occurrence of CpG dinucleotides. These control motifs have only 3,081 conserved instances, which is 20-fold lower than for the discovered motifs. These results indicate that only a small fraction of the 60,019 instances of the discovered motifs are likely to have occurred purely by chance.

The number of conserved instances is highly uneven across the motifs (range 37–7,549, with mean of 266 and median of 61). Most motifs (67%) have <100 conserved instances (SI Table 3). But, remarkably, the two motifs with the highest enrichment scores, LM1 and LM2, both have >5,000 conserved instances in the human genome (Table 2), suggesting a widespread functional role for these elements.

Characterizing the Discovered Motifs. ***Known regulatory elements.*** Among the 233 discovered motifs, 16 match known regulatory elements (Table 1). For example, the LM9 motif is nearly identical

Table 2. Properties of the top 10 discovered motifs

ID	Motif profile	No. of conserved instances	False positive rate*	Conservation rate,† %	Fold increase in conservation rate‡	Correlation between cross-species conservation and motif profile	Positional bias around TSS§
LM1		5,332	0.050	29.3	9.5	0.92	
LM2		7,549	0.048	29.4	14.0	0.91	
LM3		844	0.048	40.1	14.3	0.94	
LM4		1,877	0.046	20.3	13.5	0.89	20.3
LM5		224	0.042	19.4	16.3	0.87	
LM6		79	0.026	20.1	10.1	0.81	25.5
LM7		6,302	0.048	21.6	10.3	0.72	
LM8		608	0.047	17.2	9.6	0.68	
LM9		1,443	0.039	11.8	8.4	0.90	6.1
LM10		5,914	0.050	14.5	6.6	0.77	

*The proportion of conserved instances expected to have occurred by chance.
†The proportion of instances detected in human that are also conserved in orthologous regions of other mammals.
‡Compared to the conservation rates of control motifs.
§Fold enrichment on the number of motif sites located within 1 kb of TSS over those for control motifs. Only motifs with fold enrichment above 4 are shown.

to the consensus sequence of the neuron-restrictive silencer element (NRSE). The NRSE is recognized by the transcription factor REST (RE1 silencing transcription factor), which plays a pivotal role in repressing the expression of neuronal genes in nonneuronal tissues (9–11). Between 800 and 1,900 NRSE sites have been estimated to exist in the human genome (12, 13), which is consistent with our count of the number of conserved instances (1,443). It is reassuring to note that our procedure identified the LM9 motif without any prior knowledge, recovering the correct size of NRSE and showing nearly perfect similarity to NRSE along all of its 21 positions (SI Fig. 7).

Another example is LM6, which is a well studied RNA motif that is present exclusively in the 3′-UTRs of genes encoding histone proteins. In histone mRNAs this sequence is known to fold into a stem-loop structure involved in posttranscriptional regulation, playing a role similar to poly(A) tails on typical mRNAs (14, 15).

Conservation properties. The discovered motifs have two notable conservation properties. First, they show a much higher conservation rate than the control motifs, even outside the CNEs where they were discovered. The conservation rate was defined as the ratio of conserved instances to total instances in the human genome. All of the discovered motifs have a conservation rate that is 2-fold higher than for their matched controls, and 65% have a rate that is 5-fold higher (SI Table 4). If the conservation rate is computed based only on motif instances outside the CNEs, 96% of the discovered motifs have a conservation rate that is 2-fold higher than for their controls, and 63% have a rate that is 5-fold higher.

Second, the motifs show similar patterns of cross-species conservation and within-species conservation (Fig. 1 *a* and *b*). For each motif we asked whether the most conserved positions across the various motif instances within human (and thus those most likely to be involved in motif recognition) are also the positions within individual instances that show the highest conservation across species (and thus are most constrained in their evolution). To measure the within-species conservation of a motif we used the information content (I_k) of its PWM at the position k. To quantify its cross-species conservation we identified its instances located within the CNEs and calculated the proportion (M_k) of the instances with bases not mutated in the orthologous regions of the mouse or dog genomes at position k of the motif. The correlation coefficient between I and M for each of the discovered motifs is shown in Fig. 1*c*. We found that nearly all motifs (95%) show a positive correlation, and 53% have correlation coefficient >0.5. This suggests that the discovered motifs are indeed functional. The results also suggest that these motifs retain similar recognition properties across species.

GENETICS

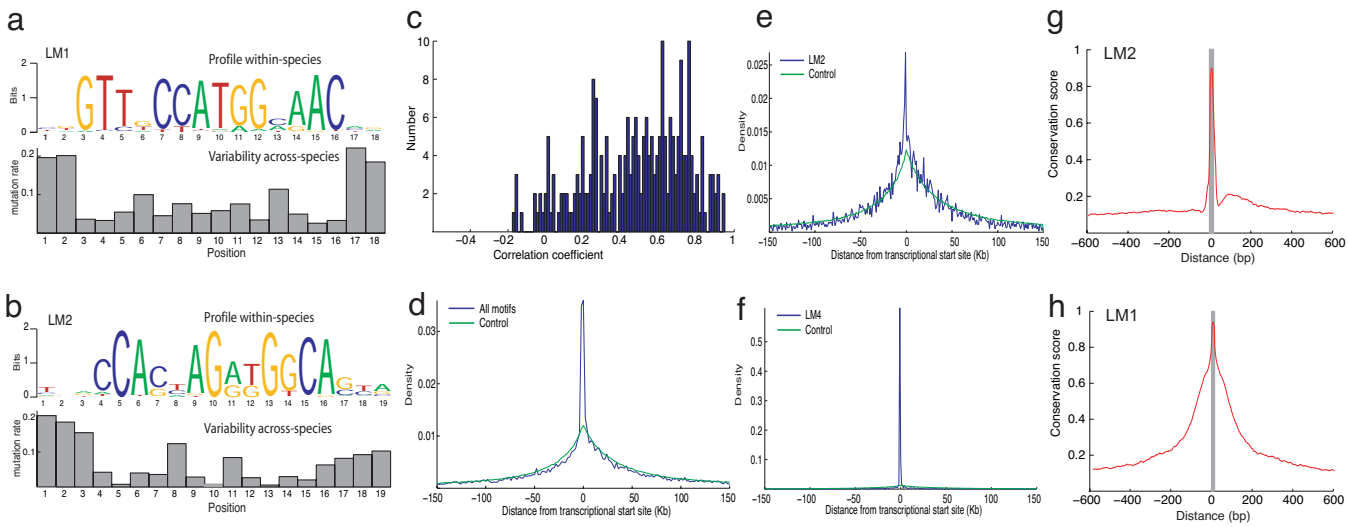


Fig. 1. A summary on properties of the discovered motifs. (*a* and *b*) Motif profile within species and variability across species for LM1 (*a*) and LM2 (*b*). Positions with high information content are less variable than those with low information content in cross-species comparison. (*c*) The correlation coefficients between motif profile and across-species conservation pattern for all discovered motifs. (*d–f*) The location of motif sites relative to TSS for all motifs (*d*), LM2 (*e*), and LM4 (*f*). They demonstrate that most of the discovered motifs, in particular LM2, are broadly distributed relative to TSS, not much different from the distribution of control sites randomly drawn from the genome (green lines). (*h*) Sequences surrounding LM1 sites are also conserved, in contrast to those surrounding LM2 sites (*g*). Gray bars show locations of the motifs. Conservation scores are phastCons scores (28) averaged over motif sites.

Palindromes. A significant proportion (17%) of the 233 motifs are palindromes, forming perfect or nearly perfect matches to their reverse complement over nearly their entire length. For example, LM3 consists of GTTGCY juxtaposed with its reverse complement, RGCAAC, with a central W, itself a self-palindrome (W = A/T). The proportion of palindromes is much higher than for random control sequences (0.13%) (see *SI Text*) and is similar to the proportion seen for the 16 known motifs (18%). The enrichment is especially pronounced among the 20 top-scoring motifs, with 45% being palindromic. Notably, the palindromic motifs are also symmetric in the information content of each base, and weakly specified positions are symmetrically placed with weakly specified positions on the two motif halves. Palindromicity can be indicative of DNA sequences that bind by a protein homodimer. Alternatively, palindromicity can sometimes reflect RNA sequences that form stem-loop structures, as illustrated by the LM6 motif in the 3′-UTRs of histone genes.

Distance from transcriptional starts. Most of the discovered motifs show little or no enrichment near genes. More than 93% have 80% of their conserved instances located >10 kb away from the TSS of any gene (Fig. 1*d*). A typical example is the LM2 motif (Fig. 1*e*). Most of these motifs are likely not to be related to core and proximal promoter functions, but may instead encode distal regulators, insulators, or other functions.

There are five cases, however, with a strong preference for being located near gene starts. A striking example is the LM4 motif, for which ≈60% of the conserved instances lie within 1 kb of the TSS (26-fold enriched over random expectation) and the modal distance is 75 bases upstream of the TSS (Fig. 1*f*). Another example is LM100, a palindromic sequence for which 45% of conserved instances lie within 2 kb of a TSS. These motifs are likely to be related to core and proximal promoter functions.

Local conservation context. We studied the conservation context of the discovered motifs. Because CNE sequences used to discover the motifs tend to occur in large blocks (N50 length = 110 bases, where N50 length is the length x such that 50% of all CNE bases lie in CNEs with the size $\geq x$), conserved motif occurrences lying within CNEs would be expected to be embedded within blocks of conserved sequence. This is indeed the case. For each motif M we examined the block of conserved sequence surrounding the each conserved occurrence in a CNE and defined $d_1(M)$ to be the N50 length of the block. The median value of $d_1(M)$ is 112 bases, with an interquartile range of 88–140 bases.

More revealingly, we examined the corresponding value $d_2(M)$ defined for conserved motif instances that lie outside the CNEs data set. The median value of $d_2(M)$ is 96 bases (interquartile range of 61–133 bases), which is similar to $d_1(M)$. This indicates that the discovered motifs typically function as part of regulatory modules containing many other regulatory elements. These results suggest that CNE motifs here may provide a useful initial entry point for studying the function of diverse large CNEs, including ultra-conserved elements that have been shown to have enhancer function.

Although most motifs appear to function in concert with others, we found eight striking examples among the 233 motifs that appear to act in isolation. This is true both for conserved occurrences within and outside the CNE data set. These motifs are LM9 (NRSE), LM6 (the histone 3′-UTR element), LM4 (a promoter-proximal motif), and four unknown motifs: LM2, LM7, LM23, and LM194. (We show below that LM2, LM7, and LM23 correspond to CTCF binding sites.) The median lengths of surrounding conserved sequences for these motifs are all less than five flanking bases on each side. For example, LM2 has only a median two flanking conserved bases on each side (Fig. 1*g*), whereas LM1 (Fig. 1*h*) has a median of 31 flanking conserved bases on each side.

LM1 Defines RFX Binding Sites. The most highly enriched motif LM1 is similar to the X-box motif, which has been extensively studied in yeast and nematodes (16–18). In yeast, more than three dozen X-box sites have been identified, and these sites have been shown to be bound by the Crt1 protein, an effector of the DNA damage checkpoint pathway (19). In *Caenorhabditis elegans*, >700 X-box sites have been computationally predicted, and several dozen of these sites have been demonstrated to be recognized by the DAF-19 protein, which is known to regulate genes involved in the development of sensory cilia (16, 18).

Individual instances of the X-box motif in vertebrates have been reported, but no systematic survey of X-box motifs in the human genome has been conducted. Approximately three dozen such sites have been reported to be bound by RFX family proteins, which are

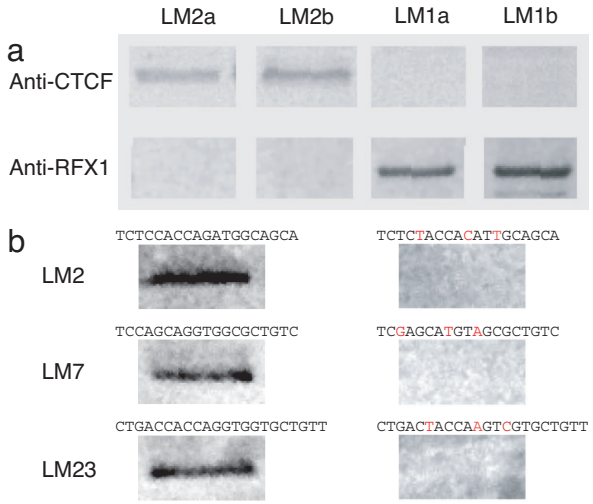


Fig. 2. Confirmation of CTCF and RFX1 binding by *in vitro* affinity capture. (*a*) CTCF was specifically captured by probes LM2a and LM2b constructed for the LM2 motif, whereas RFX1 was specifically captured by probes LM1a and LM1b constructed for the LM1 motif. (*b*) The binding of CTCF to LM2, LM7, and LM23 (*Left*), but not to their corresponding mutant motifs with three core bases altered (*Right*). See *Methods* for probes used in the experiments.

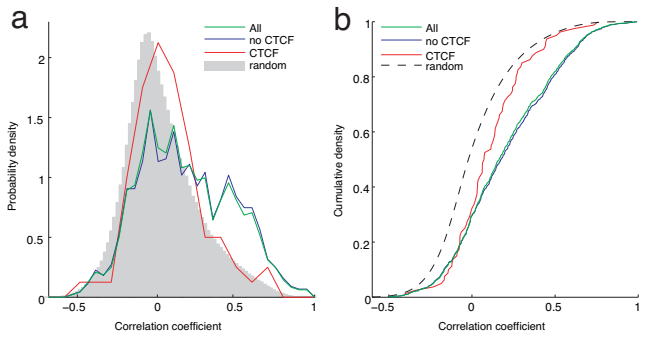


Fig. 3. Genes separated by predicted CTCF sites are less correlated in gene expression. Correlation coefficient between neighboring gene pairs is shown in terms of probability density (*a*) and cumulative distribution (*b*). Green line, correlation between all neighboring genes; red line, correlation between genes separated by at least one CTCF site; gray shading, correlation between randomly chosen gene pairs.

homologous to both Crt1 and DAF-19 and contain a highly conserved winged helix DNA binding domain. The biochemically characterized consensus sequence for RFX binding shows similarity to the LM1 motif (20), although it contains less information.

To test whether LM1 binds RFX proteins, we performed an affinity-capture experiment (see *SI Text*). A biotinylated double-stranded DNA probe containing multiple copies of the LM1 motif was incubated with HeLa cell nuclear extract and then captured with streptavidin. The bound protein was electrophoresed, blotted, and probed with an antibody against RFX1, a prototypical member of the RFX family, revealing that the protein indeed specifically binds LM1 (Fig. 2*a*).

LM2 Defines a Common Insulator Site Across the Human Genome. The most interesting case among the 233 discovered motifs is LM2. It has the largest number of conserved instances (7,549) in the genome, with the vast majority being located far from TSSs (Fig. 1*e*). The LM2 motif is 19 bases in length and does not match the reported consensus sequence of any known motif.

We obtained a hint regarding the possible function of the LM2 motif by using proteomic experiments in which HeLa cell nuclear extract was subjected to affinity capture with a biotinylated double-stranded DNA probe containing multiple copies of the LM2 motif, and the resulting material was analyzed by protease digestion and mass spectrometry. These affinity-capture experiments suggested that the CTCF protein binds the LM2 motif (unpublished data).

CTCF, a protein containing 11 zinc-finger domains, is a major factor implicated in vertebrate insulator activities (21–23). An insulator is a DNA sequence element that prevents a regulatory protein binding to the control region of one gene from influencing the transcription of neighboring genes. When placed between an enhancer and a promoter, an insulator can block the interaction between the two. Several dozen insulator sites have been characterized, and almost all have been shown to contain CTCF binding sites. In some cases, the CTCF site has been directly shown to be both necessary and sufficient for enhancer blocking activities in heterologous settings. The known CTCF sites show considerable sequence variation, and no clear consensus sequence has been derived (22). The well studied CTCF sites in the IGF2/H19 locus show similarity to the LM2 motif (24), although the similarity score is below our threshold used for detecting LM2 sites.

To test directly whether CTCF binds the LM2 motif we analyzed the material obtained by affinity capture with a biotinylated double-stranded DNA probe containing multiple copies of the LM2 motif by immunoblotting with an antibody against the human CTCF protein (see *SI Text*). The results confirmed that CTCF does indeed bind the LM2 motif (Fig. 2). By contrast, mutation of the three core positions with the highest information content (positions 5, 10, and 13 of LM2) (Fig. 1*b*) completely abolished the binding of the CTCF protein.

Given the sequence diversity among reported CTCF sites, we searched for additional motifs in our catalog that show substantial similarity to LM2. The motifs LM7 and LM23 are nearly identical in their first 14 positions, diverging only in the last four or five bases (SI Fig. 8). The two additional motifs also have an unusually large number of conserved instances (6302 for LM7 and 3758 for LM23). Affinity-capture experiments using probes containing copies of the LM7 and LM23 motifs demonstrated that both motifs bind CTCF, whereas mutation of the three core positions with the highest information content completely abolish binding (Fig. 2*b*). The three motifs, LM2, LM7, and LM23, will be referred to as a "supermotif," LM2*.

Altogether the LM2* motif has 14,987 conserved instances in the human genome (which is 20-fold higher than for the corresponding control motifs). Strikingly, this comprises approximately one-fourth of the 60,019 sites for the complete catalog of 233 motifs. We propose that the vast majority of these sites are CTCF-binding sites and function as insulators.

Although the predicted CTCF sites tend to be located far from gene starts, they are not randomly distributed across the genome. Instead, their distribution closely follows the distribution of genes, with a correlation coefficient of 0.6 (SI Fig. 9). This is consistent with the notion that the sites are related to gene regulation, rather than, for example, chromosomal structure.

We sought to test whether the predicted CTCF sites actually serve as functional insulators. Although it is possible to perform insulator assays on individual instances in a heterologous context, we were interested to assess the function of many CTCF sites in their natural context. If the predicted CTCF sites actually function as insulators, we reasoned that the presence of a CTCF site between two genes might "decouple" their gene expression.

It is known that divergent gene pairs, transcribed in opposite directions with transcription start sites close to each other, tend to show correlated gene expression patterns (25, 26). We therefore assembled a data set of 963 divergent gene pairs with intergene distance <20 kb and with expression values measured across 75 human tissues (27). As expected, the divergent gene pairs are more closely correlated in gene expression than randomly chosen gene

GENETICS

pairs (Fig. 3). When the cases are divided into gene pairs separated by a CTCF site (CTCF pairs, 80 cases) and those not separated by a CTCF site (non-CTCF pairs, 883 cases), the former show correlations that are essentially equivalent to the random background. Overall, 37% of non-CTCF pairs are strongly correlated (correlation coefficient $\rho > 0.3$). This proportion is 2-fold higher than the proportion of random genes pairs (12%) showing similarly strong correlation. By contrast, the proportion of CTCF pairs with similarly strong correlation is 16%, which is close to that seen for random gene pairs. This difference persists after correcting for small difference in the lengths of CTCF-containing and CTCF-non-containing intergenic regions (SI Fig. 10). This provides strong evidence that the majority of the predicted CTCF sites do indeed function as insulators.

Finally, we examined the frequency of the CTCF motif LM2* across various vertebrate genomes. The three motifs all occurred frequently in all eutherian mammals, opossum, chicken, and the pufferfish *Tetraodon*. The motif shows a similar total number of instances across all vertebrate species despite a 5-fold variation in genome size (SI Fig. 11). This is consistent with the LM2* motif being related to gene number (which is fairly constant across these species) rather than genome size.

Discussion

Our analysis provided an initial systematic catalog of regulatory motifs in the conserved regions of the entire human genome. The 233 discovered motifs are highly enriched in the CNE sequence, with all being at least 5-fold enriched relative to the rest of the genome. These motifs match 60,019 conserved instances in the human genome, with a typical motif having ≈100 conserved instances. Among the 233 discovered motifs, only 16 could be recognized as previously known regulatory elements, indicating that much more still remains to be learned about the function of CNE.

The most interesting unknown motif is LM2, which has ≈7,500 conserved instances in the genome, more abundant than any other discovered motif. We used affinity-capture assays to demonstrate that LM2, as well as two other closely related motifs, LM7 and LM23, are specifically bound by the CTCF protein, which is involved in insulator function. Together, the three motifs match nearly 15,000 conserved instances in the human genome, corresponding to approximately one-fourth of all matching instances for the entire set of discovered motifs. Although we cannot rule out that CTCF protein can also bind to other, highly dissimilar sites, our findings suggest that a few dominant CTCF motifs are extremely enriched throughout the human genome.

The results here are, of course, only a step toward comprehensive catalog of regulatory motifs across the human genome. In particular, our analysis used stringent threshold to identify only the most highly enriched motifs in the CNEs and therefore have omitted short motifs (e.g., 6–8 nt). Additionally, the current study focused primarily on motifs present in most mammals, and therefore many lineage-specific motifs, such as those unique to primates, still remain to be discovered. The power of motif discovery can be boosted not just by considering the enrichment of sequences in the human CNEs, but by exploiting their detailed conservation pattern across different species. With the availability of genome sequences from an increasing number of related mammals, it should be possible to create a complete dictionary of human motifs in the years ahead.

Methods

We started by enumerating a list of candidate k-mers with $12 \leq k \leq 22$ and counting the number (C) of matching instances of each k-mer present in the CNE data set (SI Fig. 5). A sequence was declared a match to a k-mer if the number of mismatches between the sequence and the k-mer was less than a threshold M (where $M = 0$ for $k \leq 13$; $M = 1$ for $k = 14$, 15, or 16; $M = 2$ for $k = 17$ or 18; and $M = 3$ for $k \geq 19$). For k-mers with $C \geq 30$ we identified all matching instances in the entire human genome and assessed their enrichment in the CNEs using two scores: (*i*) fold enrichment: SNR $= C/\mu$; and (*ii*) Z-score $= (C - \mu)/sqrt(\mu)$, quantifying the significance of the enrichment. Here, μ is the expected number of matching instances within the CNE data set based on the observed frequency of matches in the overall genome. Finally, we collected all k-mers with SNR ≥ 5 and Z-score ≥ 10, resulting in a total of 69,810 k-mers significantly enriched in the CNEs. The probability of a k-mer having a Z-score ≥ 10 at random is $< 2 \times 10^{-12}$, and, after Bonferroni correction, the expected number of such k-mers is $< 10^{-4}$. These k-mers were further clustered and grouped into 233 distinct motifs according to the procedure described in ref. 5. For each motif we selected the k-mer with the highest Z-score to represent the motif.

Please see *SI Text* for additional methods.

We thank David Jaffe, Manuel Garber, and Sarah Calvo for insightful comments and suggestions. We gratefully acknowledge Xiaolan Zhang, Li Wang, Jacob Jaffe, and Steve Carr for assistance with affinity-capture experiments. This work was supported in part by grants from the National Human Genome Research Institute (to E.S.L.) and from the Broad Institute.

1. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, *et al.* (2006) *Nature* 444:499–502.
2. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, James Kent W, Haussler D (2006) *Nature* 441:87–90.
3. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, *et al.* (2005) *PLoS Biol* 3:e7.
4. Dermitzakis ET, Reymond A, Antonarakis SE (2005) *Nat Rev Genet* 6:151–157.
5. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M (2005) *Nature* 434:338–345.
6. Elemento O, Tavazoie S (2005) *Genome Biol* 6:R18.
7. Ettwiller L, Paten B, Souren M, Loosli F, Wittbrodt J, Birney E (2005) *Genome Biol* 6:R104.
8. Jones NC, Pevzner PA (2006) *Bioinformatics* 22:e236–242.
9. Ballas N, Mandel G (2005) *Curr Opin Neurobiol* 15:500–506.
10. Chong JA, Tapia-Ramirez J, Kim S, Toledo-Aral JJ, Zheng Y, Boutros MC, Altshuller YM, Frohman MA, Kraner SD, Mandel G (1995) *Cell* 80:949–957.
11. Schoenherr CJ, Anderson DJ (1995) *Science* 267:1360–1363.
12. Mortazavi A, Thompson EC, Garcia ST, Myers RM, Wold B (2006) *Genome Res* 16:1208–1221.
13. Bruce AW, Donaldson IJ, Wood IC, Yerbury SA, Sadowski MI, Chapman M, Gottgens B, Buckley NJ (2004) *Proc Natl Acad Sci USA* 101:10458–10463.
14. Williams AS, Marzluff WF (1995) *Nucleic Acids Res* 23:654–662.
15. Pandey NB, Marzluff WF (1987) *Mol Cell Biol* 7:4557–4559.
16. Blacque OE, Perens EA, Boroevich KA, Inglis PN, Li C, Warner A, Khattra J, Holt RA, Ou G, Mah AK, *et al.* (2005) *Curr Biol* 15:935–941.
17. Zaim J, Speina E, Kierzek AM (2005) *J Biol Chem* 280:28–37.
18. Efimenko E, Bubb K, Mak HY, Holzman T, Leroux MR, Ruvkun G, Thomas JH, Swoboda P (2005) *Development (Cambridge, UK)* 132:1923–1934.
19. Huang M, Zhou Z, Elledge SJ (1998) *Cell* 94:595–605.
20. Emery P, Strubin M, Hofmann K, Bucher P, Mach B, Reith W (1996) *Mol Cell Biol* 16:4486–4494.
21. Bell AC, West AG, Felsenfeld G (1999) *Cell* 98:387–396.
22. Ohlsson R, Renkawitz R, Lobanenkov V (2001) *Trends Genet* 17:520–527.
23. Gaszner M, Felsenfeld G (2006) *Nat Rev Genet* 7:703–713.
24. Bell AC, Felsenfeld G (2000) *Nature* 405:482–485.
25. Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM (2004) *Genome Res* 14:62–66.
26. Li YY, Yu H, Guo ZM, Guo TQ, Tu K, Li YX (2006) *PLoS Comput Biol* 2:e74.
27. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, *et al.* (2004) *Proc Natl Acad Sci USA* 101:6062–6067.
28. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, *et al.* (2005) *Genome Res* 15:1034–1050.

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes

Michael F. Lin,[1] Joseph W. Carlson,[2] Madeline A. Crosby,[3] Beverley B. Matthews,[3] Charles Yu,[2] Soo Park,[2] Kenneth H. Wan,[2] Andrew J. Schroeder,[3] L. Sian Gramates,[3] Susan E. St. Pierre,[3] Margaret Roark,[3] Kenneth L. Wiley Jr.,[4] Rob J. Kulathinal,[3] Peili Zhang,[3] Kyl V. Myrick,[4] Jerry V. Antone,[4] Susan E. Celniker,[2] William M. Gelbart,[3,4] and Manolis Kellis[1,5,6]

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02139, USA; [2]Berkeley Drosophila Genome Project, Department of Genome Biology, Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA; [3]FlyBase, The Biological Laboratories, Harvard University, Cambridge, Massachusetts 02138, USA; [4]Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA; [5]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts 02139, USA

The availability of sequenced genomes from 12 *Drosophila* species has enabled the use of comparative genomics for the systematic discovery of functional elements conserved within this genus. We have developed quantitative metrics for the evolutionary signatures specific to protein-coding regions and applied them genome-wide, resulting in 1193 candidate new protein-coding exons in the *D. melanogaster* genome. We have reviewed these predictions by manual curation and validated a subset by directed cDNA screening and sequencing, revealing both new genes and new alternative splice forms of known genes. We also used these evolutionary signatures to evaluate existing gene annotations, resulting in the validation of 87% of genes lacking descriptive names and identifying 414 poorly conserved genes that are likely to be spurious predictions, noncoding, or species-specific genes. Furthermore, our methods suggest a variety of refinements to hundreds of existing gene models, such as modifications to translation start codons and exon splice boundaries. Finally, we performed directed genome-wide searches for unusual protein-coding structures, discovering 149 possible examples of stop codon readthrough, 125 new candidate ORFs of polycistronic mRNAs, and several candidate translational frameshifts. These results affect >10% of annotated fly genes and demonstrate the power of comparative genomics to enhance our understanding of genome organization, even in a model organism as intensively studied as *Drosophila melanogaster*.

[Supplemental material is available online at www.genome.org. Additional supplemental materials are available online at http://compbio.mit.edu/fly/genes/. Full-length cDNA sequence data from this study have been submitted to GenBank under accession nos. BT029554–BT029635, BT029637–BT029727, BT029940–BT029957, BT030133–BT030144, BT030416–BT030421, and BT030448–BT030452. RT-PCR amplicon and primer sequence data have been submitted to GenBank under accession nos. ES439769–ES439782.]

The compilation of a complete and accurate catalog of all protein-coding genes is a critical step in fully understanding the functional elements in any genome. In *Drosophila melanogaster*, a century of classical genetics, large-scale EST and cDNA sequencing (Rubin et al. 2000; Stapleton et al. 2002b; http://www.fruitfly.org/EST), and manual curation (Adams et al. 2000; Misra et al. 2002) have led to a gene catalog of very high quality, containing 13,733 euchromatic protein-coding genes (as of FlyBase annotation Release 4.3, the benchmark release for the initial comparative analysis of the 12 sequenced species; see Methods). While all FlyBase genes are assigned a unique numerical identifier (CGid), their level of supporting evidence varies widely. We distinguish the following classes: 4711 genes have a phenotype or molecular function reported in the literature and have been assigned a descriptive name ("named genes"); of these, 893 have at least 50 literature citations ("well-studied genes"). The remaining 9022 genes lack a descriptive name ("CGid-only genes"); of these, 4373 have been assigned a putative molecular function on the basis of homology with known protein domains or genes in other species ("GO-annotated genes"), while the remaining 4649 gene annotations are essentially uncharacterized ("uncharacterized genes"). Most of the CGid-only genes are supported by cDNA sequence data or protein sequence similarity, but a small number are based primarily on de novo predictions.

It is unclear how close to completion the current gene set may be, or what fraction of the current annotations may be inaccurate. On one hand, numerous genes and alternative splice forms may be still missing from the current annotation, and indeed a pilot study suggests an additional 700 genes may lie amidst 10,000 existing de novo and microarray-based predictions (Yandell et al. 2005). On the other hand, existing gene models might be incomplete or contain inaccuracies, and some, especially those based solely on de novo predictions, may be completely spurious. Even some genes supported by cDNA and EST

[6]Corresponding author.
E-mail manoli@mit.edu; fax (617) 262-6121.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6679507. Freely available online through the *Genome Research* Open Access option.

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

evidence or mutation phenotypes could in fact represent RNA-coding genes without any protein-coding function.

Comparative genomic analysis is a powerful approach to the discovery of protein-coding genes. Comparative data have been used to significantly revise the established annotations of the yeast *Saccharomyces cerevisiae* genome (Cliften et al. 2003; Kellis et al. 2003), but the greater complexity of gene structures and other genomic features in large eukaryotic genomes presents many additional challenges. Initial efforts in vertebrates (Thomas et al. 2003) as well as flies (Bergman et al. 2002; Richards et al. 2005) suggest that comparative genomics can similarly lead to substantial improvements in the gene annotations of these species, and the incorporation of comparative data into de novo gene prediction systems has led to great improvements in their accuracy (Brent 2005). However, current de novo gene predictors still cannot be solely relied upon for complete annotation of complex eukaryotic genomes. Moreover, they are of limited use for revisiting existing annotations, since disagreements between predicted gene structures and gene annotations can be due to errors in the predictions at least as often as errors in the annotations. Thus, new methods are necessary in order to discover new genes and exons with high predictive value and to revisit existing annotations using comparative data in complex eukaryotes.

In this study, we use whole-genome alignments of 12 *Drosophila* genomes to systematically review the protein-coding gene annotations of *D. melanogaster*. By studying the conservation properties of known genes, we identify recurrent patterns of evolutionary change that are hallmarks of purifying selection operating upon protein-coding sequences. We use these evolutionary signatures to examine the entire genome and identify conserved protein-coding regions with high accuracy. These signatures confirm the protein-coding function of the vast majority of hypothetical genes and identify more than a thousand new exons. In contrast, these signatures strongly reject several hundred genes, most of which are likely to be spurious predictions or noncoding genes. We also used these signatures to refine the annotation and boundaries of existing genes, including translation initiation sites, splice sites, and functional reading frame of translation. Finally, our methods identify candidates for a variety of exceptional gene structures such as translational readthrough, dicistronic genes, and conserved reading frameshifts in the middle of protein-coding exons. We evaluated many of these proposed changes through manual curation and directed sequencing efforts. Overall, we used comparative data to propose revisions for >10% of *D. melanogaster* protein-coding gene models. While many extensions and future directions remain, this work is a substantial step toward achieving the best possible gene annotations for *D. melanogaster*. It also serves as a model for similar efforts to improve the annotation of other important target genomes, including the human.

Evolutionary signatures for protein-coding gene identification

Protein-coding DNA sequences evolve under distinctive evolutionary constraints since selective forces at the nucleotide level reflect constraints operating on the encoded protein. Thus, mutations to the DNA that preserve properties of the amino acid translation (e.g., synonymous substitutions) tend to be tolerated, while mutations that disrupt the translation (e.g., frame-shifting insertions or deletions or nonsense mutations) tend to be excluded by natural selection. In DNA sequence alignments of closely related species, these constraints manifest themselves as "evolutionary signatures", recurrent patterns of evolutionary change that we can use to uniquely identify protein-coding sequences (Fig. 1).

We applied two independent quantitative metrics that use evidence from multiple informant sequences to distinguish regions under protein-coding selection. The first metric observes reading frame conservation (RFC) and quantifies the strong tendency of insertions or deletions (indels) within coding regions to preserve the reading frame of translation. We have previously applied RFC in yeast species (Kellis et al. 2003). The second metric observes codon substitution frequencies (CSF) and identifies the distinctive biases in the frequency of codon substitutions in protein-coding regions, constrained by the selective preference for synonymous substitutions and amino acid substitutions preserving biochemical properties (Fig. 1). The CSF metric is similar in theme to the well-known K_a/K_s ratio and d_N/d_S rate (Yang and Bielawski 2000; Nekrutenko et al. 2002), but it is more suitable for genome-wide gene identification strategies with many informant genomes.

In contrast to methodologies that focus primarily on high sequence conservation to identify candidate genes, the RFC and CSF metrics focus on distinctive patterns of divergence in protein-coding genes, specific to their unique selective pressures. Therefore, functional RNA-level or DNA-level elements (such as RNA genes and structures, developmental enhancers, or other regulatory regions), which often exhibit high nucleotide conservation (Fig. 2), are very unlikely to show high RFC or CSF scores, enabling these metrics to distinguish coding and noncoding regions with higher accuracy. For example, when used to discriminate between exons of well-studied genes and random noncoding regions with the same length distribution, the CSF metric alone accepts 94% of coding exons while rejecting >99% of the control regions (Supplemental Table 1). This discriminatory power allowed us to systematically review the *D. melanogaster* genome annotation for protein-coding genes. We present detailed benchmarks of these and several other metrics elsewhere (M.F. Lin, A. Deoras, M. Rasmussen, and M. Kellis, in prep.).

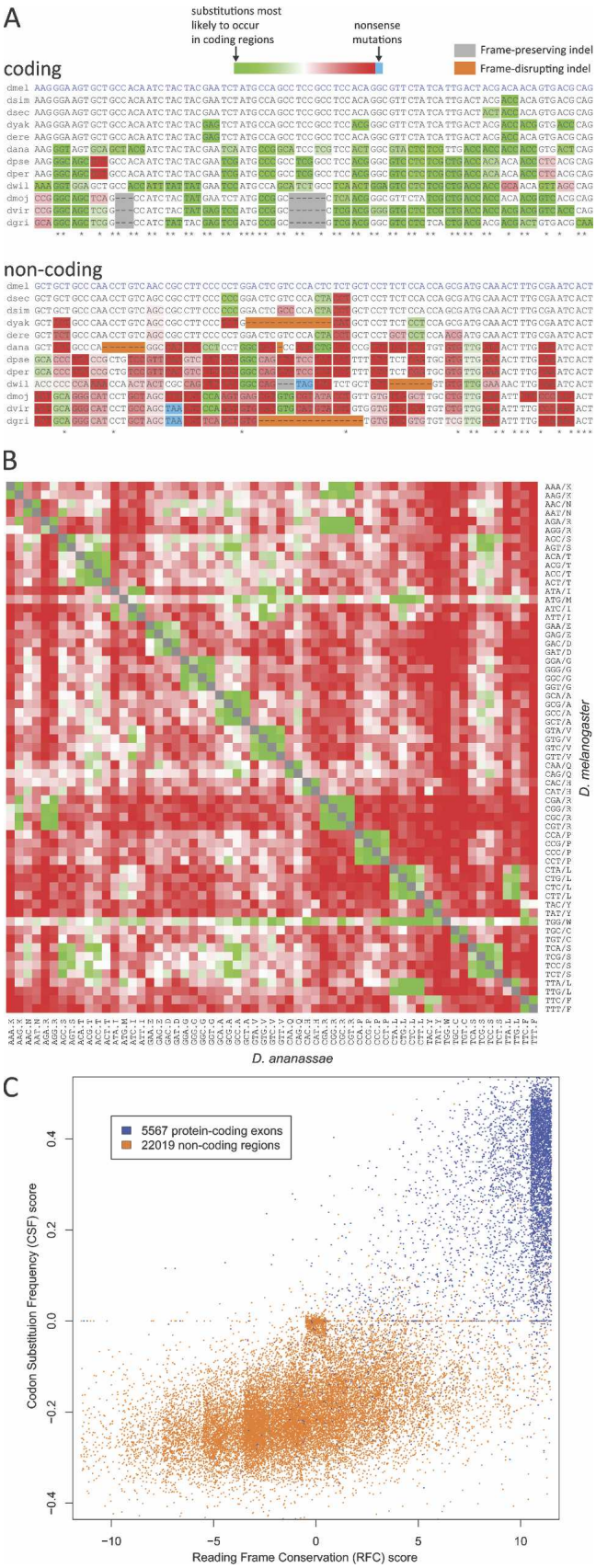
Results

Benchmarking the RFC and CSF evolutionary signatures

Our first goal was to evaluate how well our approach worked on test data sets of well-annotated genes. For this purpose, we used the classes of "named" and "well-studied" genes defined earlier. We scored every gene model covered by whole-genome sequence alignments according to the RFC and CSF metrics. By studying the score distributions for known genes and noncoding control regions, we chose RFC and CSF cutoffs above which a given gene annotation is nearly certain to represent protein-coding sequence, and used these as a test to determine whether the comparative evidence confirms that a candidate gene is indeed protein-coding (although this test does not verify that the annotated gene structure is correct in every detail).

We first scored the 893 well-studied genes. Our test accepts 882 (99%) of these gene models. Only 11 of these genes did not pass our thresholds. Two of these (*y* and *bw*) are well-conserved genes that failed due to previously known strain-specific disrupting mutations in the sequenced strain of *D. melanogaster*. The remainder may represent fast-evolving genes or genes recently evolved from previously noncoding regions. We also applied the same test to the remaining 3818 named genes with <50 citations

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press



and found that it accepts 97% (3684). Overall, the comparative evidence confirms that 4566 of 4711 "named" genes (97%) show the evolutionary signatures of protein-coding genes. We also evaluated 15,564 noncoding regions ≥300 nt in length, randomly chosen throughout the genome, and found that virtually none passed the same thresholds (Table 1; Supplemental Methods). Together, these results illustrate the high sensitivity and specificity of our approach.

Evolutionary confirmation of uncharacterized genes

We then turned our attention to the 9022 CGid-only genes in the Release 4.3 annotation set, which lack a descriptive gene name (including 4373 GO-annotated genes and 4649 uncharacterized genes). The evidence for these gene models varies widely and may include de novo gene model prediction, long open reading frames (ORFs), cDNA sequences, mRNA expression evidence, or homology with genes in other species. Since our evolutionary signatures are specific to protein-coding function, they can provide a powerful additional line of evidence indicating that these genes encode proteins, based on their alignments across *Drosophila* genomes.

Our test accepts 7879 of the 9022 CGid-only genes (87%), confirming that the vast majority of these annotations show the evolutionary signatures of protein-coding genes, and are therefore very likely to encode proteins. (Again, passing our test does not imply that all details of these gene structures are correctly annotated; we also note that it is possible that ancestral genes that have been very recently deactivated in *D. melanogaster*, have not yet acquired many disrupting mutations, and are still annotated as genes may pass our test.) The fraction of accepted CGid-only genes was only slightly higher for the "GO-annotated" subset than for uncharacterized (89% vs. 86%). It is not surprising that the proportion of accepted models for CGid-only genes (87%) is lower than for the named genes (97%): Some uncharacterized genes may be erroneous or spurious annotations (we consider this possibility further below), while others are likely to be under less stringent selective pressure than most named genes, many of which are conserved across very large evolutionary distances (Bergman et al. 2002).

New genes and exons

We next used evolutionary signatures to identify conserved protein-coding sequences missing from the current annotation. This requires not only a way to evaluate the protein-coding potential of a given region, but also a method to discover new coding

Figure 1. Evolutionary signatures for protein-coding gene identification. (*A*) Within coding regions, triplet substitutions are biased toward conservative codon substitutions (Codon Substitution Frequencies, CSF). Additionally, indels in coding regions are strongly biased to be a multiple of three in length (reading frame conservation; RFC). (*B*) The color of each codon substitution between the *D. melanogaster* sequence and an informant sequence corresponds to a log-odds score of observing that substitution in a coding region versus a noncoding region. (*C*) Quantitative metrics of RFC and CSF distinguish coding and noncoding regions. Shown in blue are 5567 coding exons of well-studied genes and in orange are 22,019 regions chosen uniformly at random from the noncoding part of the genome, with the same length distribution as the exons. The CSF score is length-normalized and the discrete RFC score is dithered by adding random noise uniformly from (−0.5,0.5) for the purposes of visualization.

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

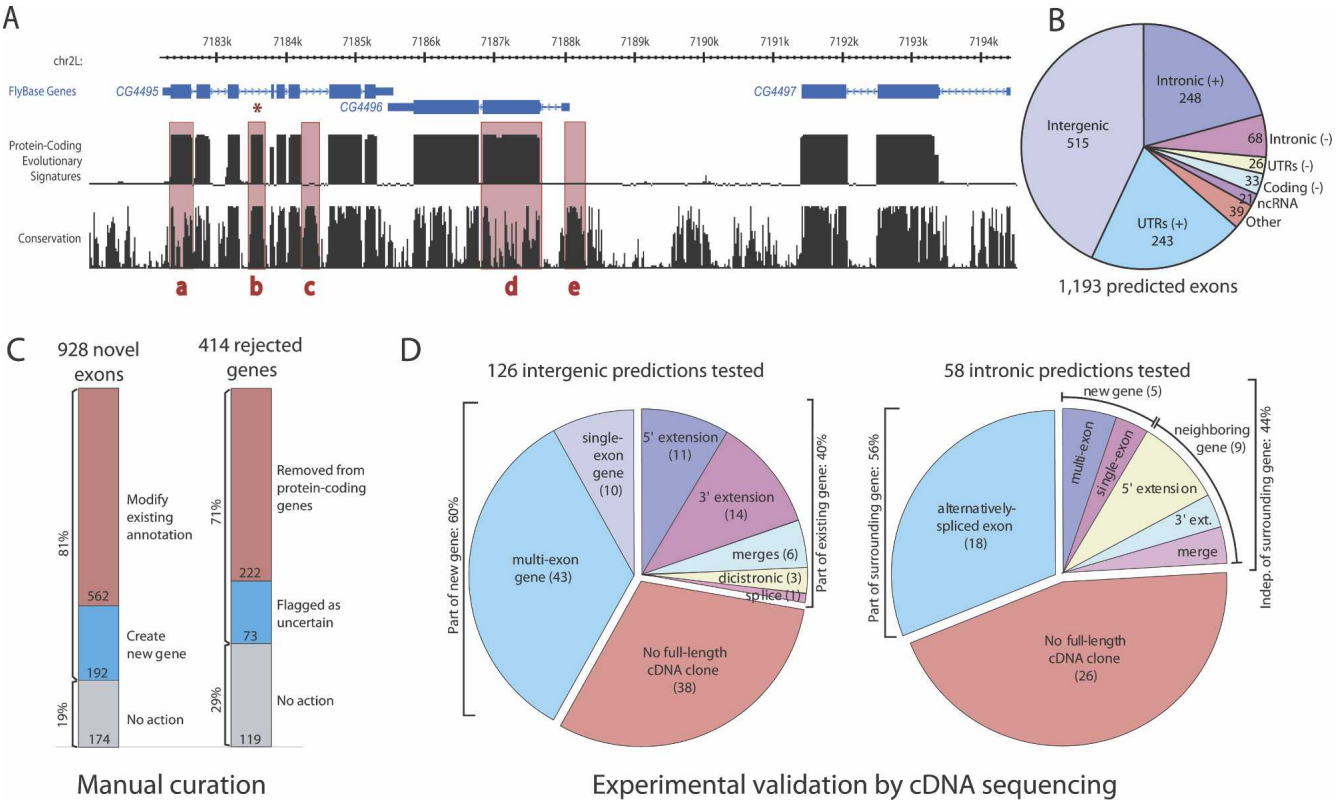


Figure 2. New protein-coding exons predicted by evolutionary signatures, examined by manual curation, and validated by cDNA sequencing. (*A*) The "Evolutionary Signatures" track shows the posterior probability of a protein-coding state in a probabilistic model integrating the RFC and CSF metrics. The "Conservation" track shows the analogous quantity from a model measuring nucleotide conservation only (Siepel et al. 2005). Note the high protein-coding scores of known exons despite lower nucleotide conservation (a,d), the low protein-coding scores of conserved noncoding regions (c,e), and the prediction of a novel exon within an intron of *CG4495* (b), subsequently validated (see Fig. 3). Rendered by the UCSC Genome Browser (Kent et al. 2002). (*B*) Distribution of 1193 new exon predictions throughout the genome. (*C*) Newly predicted exons were examined by manual curation, 81% leading to new and modified FlyBase gene annotations. Additionally, curation of genes rejected by evolutionary signatures led to the recognition of hundreds of spurious annotations. (*D*) A sample of predicted new exons was tested by cDNA sequencing with inverse PCR. Surprisingly, 44% of the validated predictions in "intronic" regions revealed a transcript independent of the surrounding gene, and 40% of the validated predictions in "intergenic" regions were part of existing genes. See Fig. 3 for examples.

intervals and to define their precise boundaries, in the absence of any existing annotation. To this end, we integrated our metrics of protein-coding evolutionary signatures into a probabilistic algorithm that determines an optimal segmentation of the genome into protein-coding and noncoding regions, based on syntenically anchored genome-wide sequence alignments of the 12 *Drosophila* species (see Methods). Our algorithm predicted 1193 new protein-coding exons not overlapping any coding exons in FlyBase Release 4.3. The large majority of these (68%) are in euchromatic intergenic and intronic regions: 515 (43%) are outside any annotated transcripts, and 316 (26%) are within an intron of an existing gene (248 transcribed from the same strand and 68 from the complementary strand). An additional 269 predicted exons (23%) overlap annotated untranslated regions of existing transcripts (243 from the same strand as the overlapping transcript and 26 from the complementary strand). The remaining predicted exons include 21 that overlap existing noncoding annotations, 33 that overlap protein-coding exons on the opposite strand, and 39 that overlap multiple Release 4.3 genes or are located in heterochromatin and cannot be easily categorized as intronic or intergenic. We manually examined most of these predictions, and also validated a subset through directed cDNA sequencing.

Manual curation incorporates most predicted exons into gene annotations

Of the 1193 predicted new exons, 928 were manually reviewed by FlyBase annotators and assessed relative to existing annotations, other gene predictions, cDNA/EST data, and protein sequence similarity evidence according to FlyBase Gene Model Annotation Guidelines (see Supplemental Methods). We excluded from manual review 265 predictions overlapping existing untranslated regions (UTRs), existing noncoding genes, or annotations independently created by FlyBase subsequent to Release 4.3 (our benchmark for this study).

Of the 928 assessed exons, 562 (61%) were incorporated into existing genes, leading to the revision of 438 gene models. The new exons most often led to the creation of alternative transcripts and, less frequently, to the modification of the intron/exon structure of an existing transcript isoform. Many of these changes (58%) were supported by additional evidence such as previously unincorporated BDGP cDNA sequences and/or sequence similarity to known proteins. Some revisions were complex, including 65 merges of two or more Release 4.3 gene models, 10 splits of Release 4.3 gene models, and four new dicistronic transcript models.

An additional 192 (21%) curated exons were incorporated in

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

Table 1. Categorization of existing gene annotations according to comparative evidence

	Total	Confirmed	Unclear	Rejected[a]
Named genes	4711	4566 (96.9%)	105 (2.2%)	40 (0.8%)
Well-studied genes	893	882 (98.8%)	8 (0.9%)	3 (0.3%)
Other named genes	3,818	3684 (96.5%)	97 (2.5%)	37 (1.0%)
CGid-only genes	9022	7879 (87.3%)	729 (8.1%)	414 (4.6%)
GO-annotated	4373	3897 (89.1%)	278 (6.4%)	198 (4.5%)
Uncharacterized	4649	3982 (85.7%)	451 (9.7%)	216 (4.6%)
All genes	13,733	12,445 (90.6%)	834 (6.1%)	454 (3.4%)
Noncoding regions	15,564	3 (0.0%)	131 (0.8%)	15,430 (99.1%)

Each annotated gene in FlyBase Release 4.3 is categorized as "confirmed" if it shows the evolutionary signatures of protein-coding genes, "unclear" if the gene is not alignable or the comparative evidence is otherwise ambiguous, and "rejected" if the gene is alignable to putatively orthologous sequence but appears unlikely to represent a genuine protein-coding gene. "Well-studied" genes are referenced by at least 50 publications in the FlyBase-indexed literature. "Named" genes have been assigned a descriptive symbol by investigators. All remaining genes are "CGid-only." "Noncoding regions" are ≥300 nt regions chosen randomly from the portion of the genome not annotated as protein-coding (see Supplemental Methods).
[a]A minority of rejected genes are falsely rejected; see text for explanation.

142 newly created gene models. Of these, 39% were supported by EST/cDNA and/or protein sequence similarity. Twenty-four of the new gene models (12%) lie within an intron of another gene on the same strand.

The remaining 174 curated exons (19%) were not incorporated into any gene models. Most of these are either small exon predictions, with a median length of 21 amino acids, or encode low-complexity sequence. Typically, these were unsupported by experimental data that would indicate inclusion in a gene model. These 174 exon predictions should be viewed as unresolved with regard to their validity, since future data may provide such experimental support.

Directed cDNA sequencing confirms predicted exons, reveals new genes and splice forms

In parallel to our manual curation efforts, we tested a subset of predictions by directed cDNA sequencing. To identify the most appropriate candidates for sequencing, we filtered the 1193 novel protein-coding exon predictions. We eliminated exon predictions of several types: those that map to certain known genes with incomplete annotations in FlyBase Release 4.3 (including heterochromatic genes and *Dscam*), to BDGP cDNA clones not yet represented in current FlyBase annotations, or to 5′ or 3′ UTRs. Additionally, we excluded any predicted exons that were deemed to be experimentally problematic because of small size or genomically repeated sequences (see Methods). Of the 434 remaining candidates for experimental validation, a sampling of 184, uniformly distributed throughout the genome, was selected. These included 126 within intergenic regions and 58 within introns of existing genes. We tested each of these 184 predictions by attempting to isolate and sequence a full-length cDNA transcript clone using self-ligation of inverse PCR products (Hoskins et al. 2005; Wan et al. 2006).

Of the 126 tested predictions within intergenic regions, we obtained a full-length cDNA for 88 exons (70%). The resulting cDNAs provide evidence for 50 new genes, including 10 single-exon genes and 40 multi-exon genes (which incorporate 43 predicted exons, and additional flanking exons that were not predicted by our algorithm). In addition, these cDNAs provided evidence for the modification of 39 existing Release 4.3 annotations: 11 new 5′ extensions or splice variants, 13 new 3′ extensions or splice variants (14 exons), two dicistronic transcripts (three exons), six transcripts merging multiple Release 4.3 gene models, and one internal splice variant.

Of the 58 tested predictions within introns of existing annotations, we obtained a full-length cDNA for 32 (55%). Only 18 of these represent new internal splice variants of the surrounding gene while the remaining 14 appeared independent of the surrounding gene. These 14 include eight alternative splice forms of previously annotated genes (five 5′ exons and two 3′ exons), two new single-exon genes, two new multi-exon genes, and two gene merges. Most surprising were data supporting an apparent example of overlapping coding sequence on opposite strands (Fig. 3D).

Overall, the cDNA data validated 120 of the 184 targeted predictions (65%). The recovered cDNA sequences also indirectly validated 42 predicted new exons that were not purposely targeted, as they were contained within the transcripts recovered from the 120 targeted predictions, leading to a total of 162 cDNA-validated predictions. The recovered cDNAs also captured additional translated and untranslated exons that were not predicted by our algorithm (see examples in Fig. 3). Finally, we note that the remaining 64 targeted predictions for which we did not obtain a high-quality, full-length cDNA sequence are not necessarily false predictions, since we only screened libraries derived from certain tissues and developmental stages (Hoskins et al. 2005).

Using TBLASTX, we searched other genomes for homologs of the new genes we recovered through cDNA sequencing. We found that many appear to be specific to the *Drosophila* or insect lineages (Supplemental Table 2). For example, 37% have a significant hit in the mosquito (*Anopheles gambiae*) or honeybee (*Apis mellifera*) genome assemblies, compared to 50% of randomly selected genes of comparable length; similarly, only 12% have significant hits to worm, yeast, or vertebrates, compared to 32% of random genes. Because gene annotation often relies on homology with known genes in other species, this might explain in part why these genes have not previously been identified.

An alternative strategy identifies relatively few additional exons

The completeness of our exon predictions is constrained by the coverage and quality of whole-genome alignments, the discriminatory power of our evolutionary metrics applied to the 12 genomes, and limitations of the probabilistic algorithm we used to integrate them. In fact, our exon prediction algorithm failed to identify 24% of exons in named genes (of which 37% were not well aligned; see Supplemental Table 3 for details). In order to

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

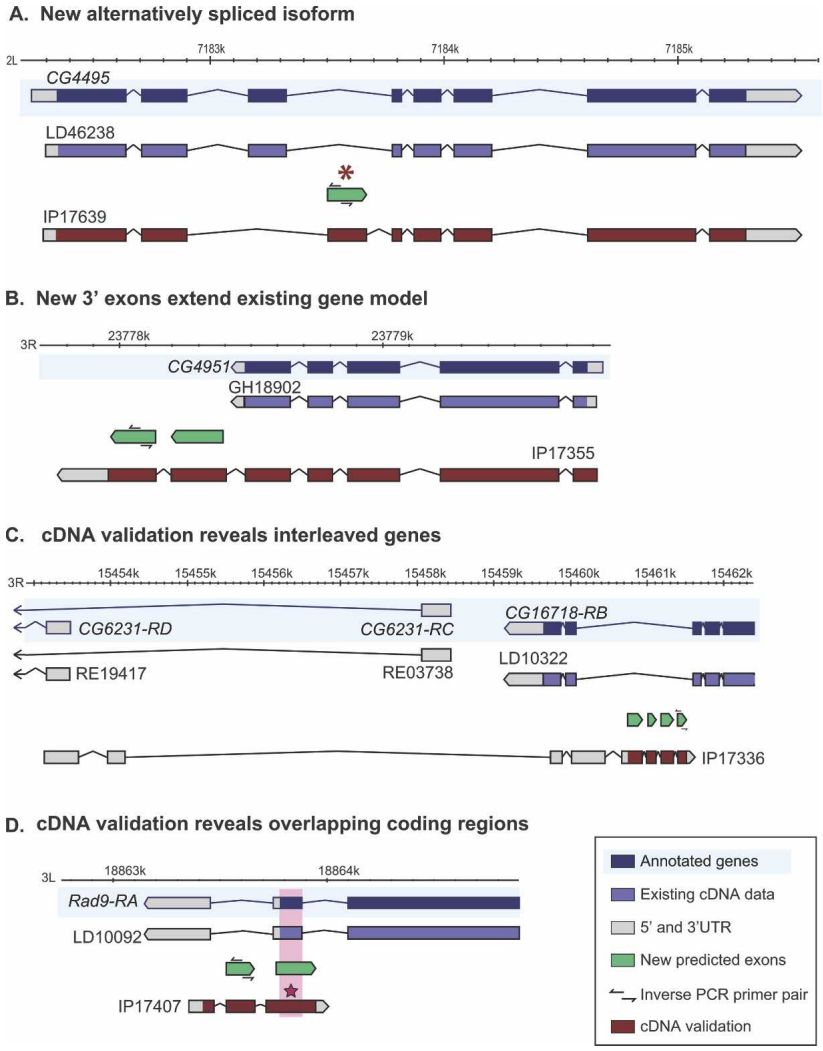


Figure 3. Full-length cDNA sequences recovered from exon predictions through inverse PCR. (*A*) Alternatively spliced transcripts—Exon Shuffling. The clone, IP17639, validates prediction congochr2L7183503 and provides evidence for an alternative transcript of the gene *CG4495*. Analysis of the embryonic microarray data (Manak et al. 2006) shows this exon is not used in embryogenesis, suggesting stage-specific splicing. Interestingly, the two alternative exons encode 20 identical amino acids at the N-terminal side of the exon. (*B*) 3′ CDS extension. The clone, IP17355, validates two predicted exons, congochr3R23777966 and congochr3R23778197, and provides evidence for an alternative transcript encoding an additional 126 aa at the C-terminal end of the gene, *CG4951*. In addition, the clone contains 185 bp of 3′ UTR. (*C*) New spliced interleaved gene. The clone, IP17336, validates four predicted exons, congochr3R15461397, congochr3R15461180, congochr3R15461031, and congochr3R15460742, and provides evidence for four additional exons. (*D*) Novel spliced overlapping gene. The clone, IP17407, validates prediction congochr3L18835687 and extends the CDS by 22 aa at the N terminus and 79 aa at the C terminus. The third coding exon overlaps the coding sequence of the gene on the opposite strand, *Rad9*, such that 45 aa on each strand are encoded in the region of overlap.

evaluate how much additional new protein-coding sequence may have escaped our notice, we undertook an alternative strategy that uses predictions from a variety of gene identification systems, representing several basic algorithmic approaches, and including both de novo and evidence-based strategies. These included AUGUSTUS (Stanke and Waack 2003), CONTRAST (S. Gross, C. Do, M. Sirota, and S. Batzoglou, Stanford University, http://contra.stanford.edu/contrast/), GENSCAN (Burge and Karlin 1997), NCBI GNOMON, geneid (Parra et al. 2000), Genie (Reese et al. 2000), and SNAP (Korf 2004).

We selected 193 "consensus" exons that are predicted by at least five of these algorithms, do not overlap annotated exons, transposable elements, or our predictions, and are at least 100 nt in length. After manual curation, 98 (51%) were incorporated into a gene model: 15 were incorporated into gene models that included exons identified by our algorithm, 63 were incorporated into existing gene models, and 20 were annotated as new or reinstated gene models. To test the validity of this approach, eight of the affected gene models were selected for evaluation by RT-PCR. Seven of the eight newly annotated "consensus" exons were validated. In several cases, additional newly annotated exons based on evolutionary signatures were also validated. Overall, 852 new exons were annotated by manual curation using both analyses, of which 88% were predicted by our algorithm based on evolutionary signatures.

Conclusion: New exons and genes

In summary, we integrated our metrics of protein-coding evolutionary signatures into a probabilistic algorithm that predicted 1193 new exons. Of these, 948 were subjected to manual curation or targeted experimentation, and 787 (83%) were supported by sufficient data to incorporate them into new or revised gene models, resulting in 150 new gene models, 70 gene merges, 10 gene splits of existing annotations, and four pairs of new dicistronic gene models. Some of the 161 predictions that were not supported following manual curation and targeted cDNA sequencing are likely to be validated in the future, e.g., as distant 5′ exons of annotated genes (Manak et al. 2006), when additional data become available. The 245 remaining predictions that were not assessed by either manual curation or experiments, most of which overlap annotated UTRs or noncoding gene models, await analysis.

Although the subsets of the predicted exons that we subjected to curation and sequencing were not selected entirely at random, neither were they selected in a way that would strongly bias them toward the highest-quality predictions. We conclude that our approach was able to identify new exons with very high predictive value, even when all existing gene annotations were excluded. Moreover, the results of an alternative strategy based on a variety of de novo and evidence-based predictions suggest that relatively few protein-coding exons remain unidentified in the euchromatin—at least that can be found at a reasonable false discovery rate using existing computational methods.

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

Many poorly conserved gene annotations are dubious

We next asked whether a subset of the CGid-only genes which failed to be confirmed as protein-coding is in fact spurious. Our previous analysis confirmed 97% of "named" genes, but only 88% of CGid-only genes; we reasoned that the remaining 1119 (12%) may be fast-evolving, recently gained, improperly aligned, or simply spurious. Here, we revisit this set to identify potential spurious annotations that do not correspond to protein-coding genes.

While our previous analysis evaluated each candidate gene over its entire length, here, we searched for any evidence of protein-coding selection. We allowed for fast-evolving domains or partially incorrect annotations by evaluating overlapping windows of 30 amino acids for evidence of protein-coding evolution. We also allowed for lineage-specific genes by searching for evolutionary evidence in groups of species at three different phylogenetic distances from *D. melanogaster*. Moreover, we tested three different genome alignment sets, to allow for potential misalignments (see Methods). Finally, we note that, if a gene is recently gained and its orthologous region is simply absent in the informant genomes, our methods make no statement about its veracity. Instead, we only evaluated regions that do align to putatively orthologous sequences in other species.

We found that 414 CGid-only genes (4.6% of 9022) are rejected even by these very lenient criteria. By comparison, only three of 893 well-studied genes (0.3%) are rejected and only 40 of all 4711 named genes (0.8%). If all rejected well-studied genes are false rejections, we would expect <30 of the 414 rejected CGid-only genes to be false rejections (95% confidence, binomial distribution). Based on named genes, we would expect that <91 of the 414 rejections are false rejections, and that at least 323 of the 414 rejected genes (78%) are indeed spurious. On one hand, this may be an overestimate, as the named and well-studied genes may be biased toward deeply conserved functions with vertebrate orthologs (Bergman et al. 2002). On the other hand, this may be an underestimate if not all rejected named genes are false rejections; some could in fact be incorrect annotations. In particular, we note that a gene can be named on the basis of a mutant phenotype, which does not necessarily imply that it is protein-coding.

Several statistics suggest that most of the genes rejected by our test are likely to be spurious predictions. As a group, they closely resemble random noncoding regions (Supplemental Fig. 1). The majority consist of relatively short, single-exon ORFs, many of which are likely to occur by chance across the whole genome. Their median coding sequence length is 381 nt, considerably shorter than the median length of all genes (1179 nt), and 63% are single-exon.

We manually examined each of the 414 CGid-only genes that were rejected by our test and all evidence supporting them, and we concluded that 222 (54%) can be immediately deleted from the annotations or recategorized as nonprotein-coding genes. These include 55 genes previously annotated as supported by cDNA sequences, which in fact turned out to be due to genomically primed clones. An additional 73 of the rejected genes (18%) had unclear or conflicting evidence and have been flagged as being of uncertain quality in the annotation comments, although they were not immediately deleted. Finally, the remaining 119 (29%) are adequately supported by existing evidence and were kept unchanged in the current database. A subset of these is likely to be rapidly evolving genes, while others may prove to be RNA-coding genes with no protein function.

We also manually examined the 40 named genes that were rejected by our test, and found that six of these should also be deleted or changed to nonprotein-coding annotations. The remaining 34 contain several genes known to be rapidly evolving, including seven male accessory gland peptides or other male-specific genes.

Last, we found that transcript evidence for at least some of the rejected genes may be explained by nonprotein-coding function. In particular, there is strong evidence that the transcripts for *CG33311* and *CG31044* are in fact precursor RNAs of microRNA genes rather than protein-coding mRNAs (Stark et al. 2007). In both cases, newly discovered microRNA genes lie within these transcripts and cluster with neighboring miRNAs of the same family. More generally, we note that some forms of the evidence supporting CGid-only genes, such as transcript cDNA sequence or genomic sequence conservation, do not directly imply translation to protein and could result from noncoding genes.

We conclude that most of the genes rejected by our test in fact do not represent genuine protein-coding genes, and the existence of many of these annotations is due to genomically primed cDNAs, erroneous de novo gene predictions, and sometimes functional RNA genes. A minority is likely to represent fast-evolving or species-specific genes that are not under purifying selection over the evolutionary distances we examined. Overall, our tests based on evolutionary signatures confirmed 7879 of 9022 CGid-only genes (87%) as clearly under protein-coding selection and rejected 414 (4.6%), most of which are likely to be spurious annotations (Table 1). We abstained from making a decision based on comparative evidence for the remaining 729 CGid-only genes (8.1%), which either could not be aligned or were supported by evolutionary signatures weakly or only over a fraction of their length. These results can help guide directed experimentation to resolve the function of all genes and transcripts, and also help focus curation efforts on a relatively small number of problem cases.

Refining existing gene annotations

The deep comparative evidence available within alignments of the 12 *Drosophila* genomes enables more fine-grained analysis than the evaluation of complete genes. We also used our metrics of protein-coding evolutionary signatures to propose a variety of detailed adjustments to existing gene annotations, affecting translation initiation sites, splice boundaries, and reading frame of translation, and to reveal likely species- or strain-specific disruptive mutations.

Translation start sites

Systematic annotation of fly genes has typically designated the longest ORF of each transcript model as the inferred protein translation, starting at its earliest in-frame ATG. However, translation may actually start at a downstream ATG. While the current understanding of the sequence and structural signals that direct translation initiation is still incomplete, the evolutionary signatures of protein-coding selection can often clearly distinguish the preferential site of translation initiation. Our analysis revealed 413 transcripts of 359 genes for which the translation start sites appear to be downstream from the presently annotated AUG, and allowed us to propose corresponding refinements to the annotations. In each case, the previously annotated start AUG is

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

not well-conserved, while the newly proposed AUG is conserved, and the intervening sequence in other species shows an abundance of nonconservative codon substitutions and frequent in-frame stop codons and frame-shifting indels. In many cases, the contrast between the conservation of the regions immediately upstream and downstream of our proposed translation start sites is striking (Fig. 4A). While we cannot rule out that some of these cases could represent species- or lineage-specific N-terminal protein extensions, a majority of our proposed downstream translation start sites are also supported by an independent analysis of the information content in their sequence contexts, to the exclusion of the annotated upstream site (M. Weir and M. Rice, in prep.).

Reading frame of translation

In addition to locating protein-coding regions, the comparative information reveals the reading frame of translation under purifying selection, since the signature of codon substitution frequencies is specific to the reading frame. This has allowed us to distinguish between overlapping ORFs, and reveal the one under selection when multiple ORFs of comparable length are all open (Fig. 4B). Such overlapping ORFs are sometimes found in short single-exon genes, where the systematic annotation has typically selected the longest, while it may in fact be a shorter ORF that is translated. We found five cases (*CG15281*, *CG13244*, *CG7738*, *CG18358*, and *CG12656*) where a shorter ORF is clearly under selection, to the exclusion of the annotated ORF. While this is a small number of cases, we note that this change leads to a completely different protein translation.

Adjustments to existing exons

We searched for potentially erroneous splice sites by identifying gene models in which a splice junction appears to coincide with a shift of the reading frame under selection (Supplemental Fig. 3). We found such events in 210 transcripts of 174 genes. While alternative splicing can use exons in different reading frames, we can at least say in these cases that selection appears to strongly favor one translation of the exon over the alternatives. We conclude that the alternatives should, therefore, be considered suspect, at least in the absence of transcript sequence data clearly indicating their use.

We also identified many existing exons that appear to be incompletely annotated, as the evolutionary signatures of protein-coding selection extend beyond their present splice bound-

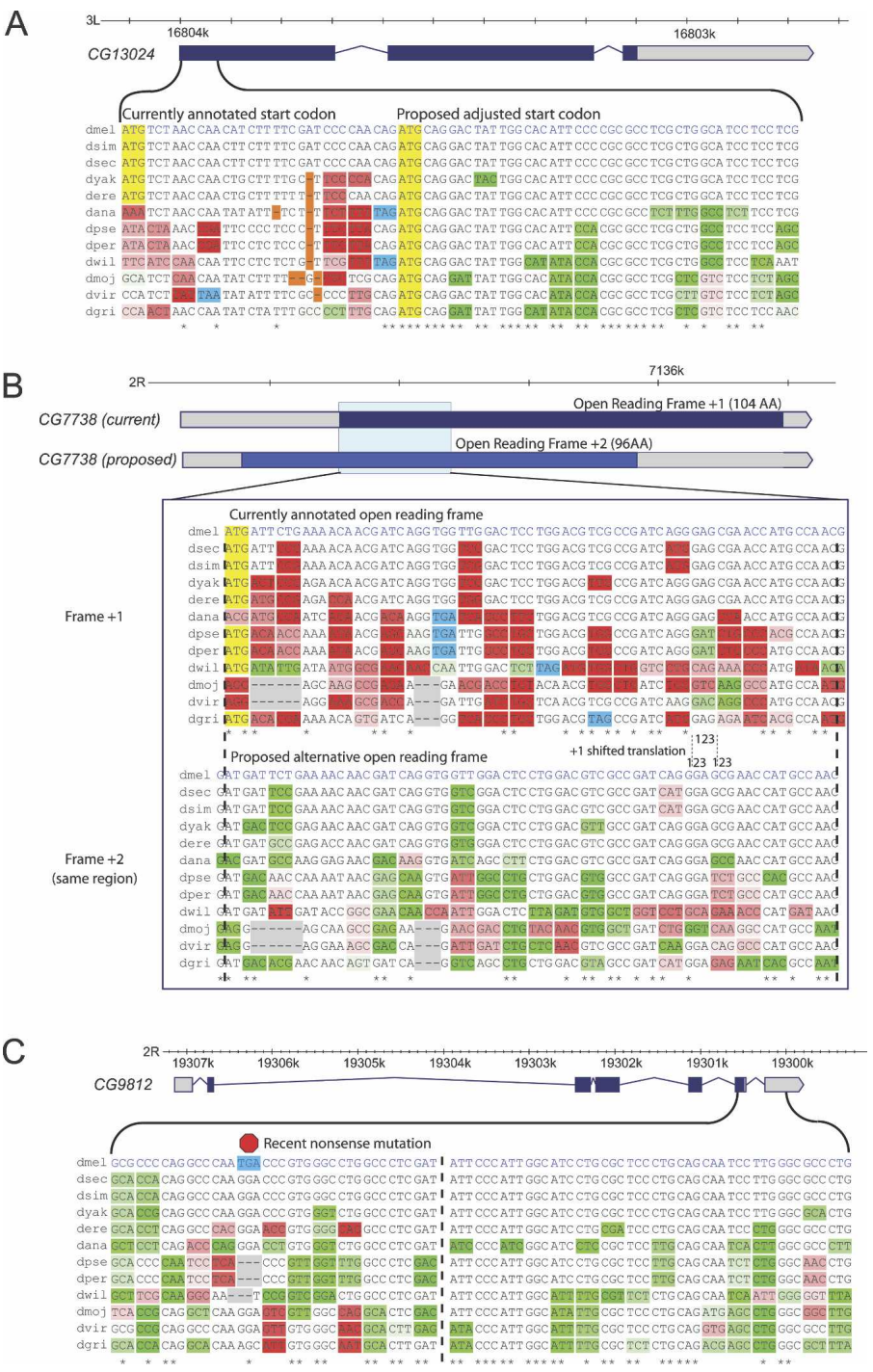


Figure 4. Examples of adjustments to existing annotations based on evolutionary signatures. (*A*) Translation start adjustment. The annotated coding sequence begins at the indicated ATG, but the informant species show frameshifts, nonsense mutations, and nonconservative substitutions in the immediately downstream region. Strikingly, however, coding signatures begin at a slightly downstream ATG. (*B*) Incorrect reading frame annotated. The transcript model contains two overlapping reading frames, the slightly longer of which is annotated as the coding sequence; but the evolutionary signatures clearly show that the other is the frame under selection. (*C*) Nonsense mutation in (the sequenced strain of) *D. melanogaster*.

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

aries, including 912 by at least 30 nt and 600 by at least 45 nt (see Supplemental materials). This may indicate either an alternative splice site or a simple mistaken annotation. When we considered the position of the likely corrected (or alternative) splice site, we found that the "extensions" of at least 30 nt are enriched for lengths divisible by three ($P < 2.2 \times 10^{-16}$, χ^2 test), suggesting that many may be alternatively spliced.

Recent nonsense and frameshift mutations

Finally, we used comparative information to identify several recent disrupting mutations in the sequenced strain of *D. melanogaster* (Celniker et al. 2002), which may have accumulated such characters during many years in laboratory culture. We identified two genes (*CG9812* and *CG33282*) in which an in-frame stop codon is aligned to a sense codon and followed by additional well-conserved protein-coding sequence in all aligned informant species. These appear to be recent nonsense mutations (Fig. 4C). An additional case, *CG14638*, may be a pseudogene.

We also identified locations in the *D. melanogaster* genome where protein-coding evolutionary selection abruptly shifts from one reading frame to another. In five cases, these coincide with a short frame-shifting indel, specific to the sequence of *D. melanogaster*, and absent from all of the other genomes. One of these (within *sdk*) was due to a previously known erroneous genomic sequence on chromosome arm 3L in *D. melanogaster*, while another (within *CG33294*, currently known as *CR33294*) may be a pseudogene. The remaining three cases (within *Ugt86Dd*, *Dscam*, and *CG34143*) are apparently recent frameshift mutations.

Identifying unusual protein-coding structures

The power of evolutionary signatures to distinguish regions under protein-coding selection has allowed us to recognize a variety of unusual phenomena that have not been amenable to systematic discovery, including stop codon readthrough, polycistronic transcripts, and translational frameshifts. We present here the results of this computational analysis, reflecting the best inference from the comparative data available to us. However, the underlying biological mechanisms remain unclear in most cases, and follow-up investigation will be required to explain these observed phenomena.

Stop codon readthrough

Just as evolutionary signatures can often distinguish the preferential site of translation initiation, they can also accurately identify the true site of translation termination. For the vast majority of genes, the comparative data show that protein-coding selection degrades exactly at the stop codon or shortly upstream. For 149 genes, however, evolutionary signatures strongly suggest that translation continues well past a deeply conserved, in-frame stop codon (Fig. 5A), indicating that these "extensions" of the corresponding proteins, which range in length from 15 to hundreds of amino acids, are under selection for their protein-coding function.

Translational readthrough of stop codons can occur through several mechanisms, among which our approach does not distinguish. However, it does not appear that many of these genes represent new selenoproteins, because many (37%) of the putatively readthrough stop codons are not UGA and we were unable to identify convincing examples of the related SECIS elements according to previously established criteria (Kryukov et al. 1999; Castellano et al. 2001). We found the set of putative readthrough genes to be enriched for nervous system expression patterns, according to in situ hybridization data (Tomancak et al. 2002; hypergeometric $P < 4.2 \times 10^{-5}$). For this reason, we speculate a possible role for A → I RNA editing by ADAR, which is most active in the nervous system (Bass 2002) and is known to mediate stop codon readthrough in a viral gene (Luo et al. 1990; Casey and Gerin 1995) and in a *D. melanogaster* neuropeptide receptor (Fig. 1 of Stapleton et al. 2002a). Still, other mechanisms may be responsible, and it is also possible that precisely positioned alternative splicing could lead to the observed signatures without direct readthrough. Overall, our results suggest that translational readthrough is not a rare phenomenon in *Drosophila* and provide candidate genes for further investigation.

Polycistronic messenger RNAs

Polycistronic messenger RNAs are single processed transcripts containing several nonoverlapping ORFs, each of which is individually translated (Andrews et al. 1996; Brogna and Ashburner 1997; Misra et al. 2002). We searched for complete (start-to-stop) ORFs that show clear signs of protein-coding selection and are fully contained within the untranslated region of an existing transcript model (Fig. 5B). This strategy rediscovered 85 of 115 annotated euchromatic dicistronic transcripts (73%) and predicts an additional 135 putative ORFs in 123 genes. We note that many of the ORFs of the previously annotated dicistronic transcripts are also found in single ORF mRNAs. This may also be the case for the genes we have identified. Our results provide a much richer set of candidate genes for further investigation, potentially doubling the number of genes with an annotated dicistronic transcript in the *D. melanogaster* genome.

"Programmed" translational frameshifts

Programmed translational frameshifts are common in viral genomes (Farabaugh 1996), and there is one known example in *D. melanogaster* (Ivanov et al. 1998). We found four locations in fly transcripts where protein-coding selection abruptly shifts from one reading frame to another, that are not readily explained as an incorrect transcript model or a species- or lineage-specific mutation. In these cases, the comparative evidence appears to indicate that a conserved translational frameshift occurs (Fig. 5C). One such case has a striking association with a highly conserved RNA structure (Supplemental Fig. 2), which we speculate might be involved in regulating its usage (Giedroc et al. 2000). We cannot, however, rule out mechanisms other than translational frameshifting, including alternative splicing, and further experimental study is required.

Discussion

A revised fly gene catalog

The availability of whole-genome alignments of the 12 *Drosophila* genomes allowed us to measure evolutionary signatures unique to protein-coding regions. In conjunction with manual curation and large-scale sequencing experimentation, these signatures enabled us to systematically revisit the fly genome annotation, with proposed changes affecting >10% of all genes. (1) We identified 1193 new exons with high predictive value, most of which were integrated into FlyBase gene annotations and many of which were validated by cDNA sequencing experiments, revealing many surprising new gene models and alternative

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

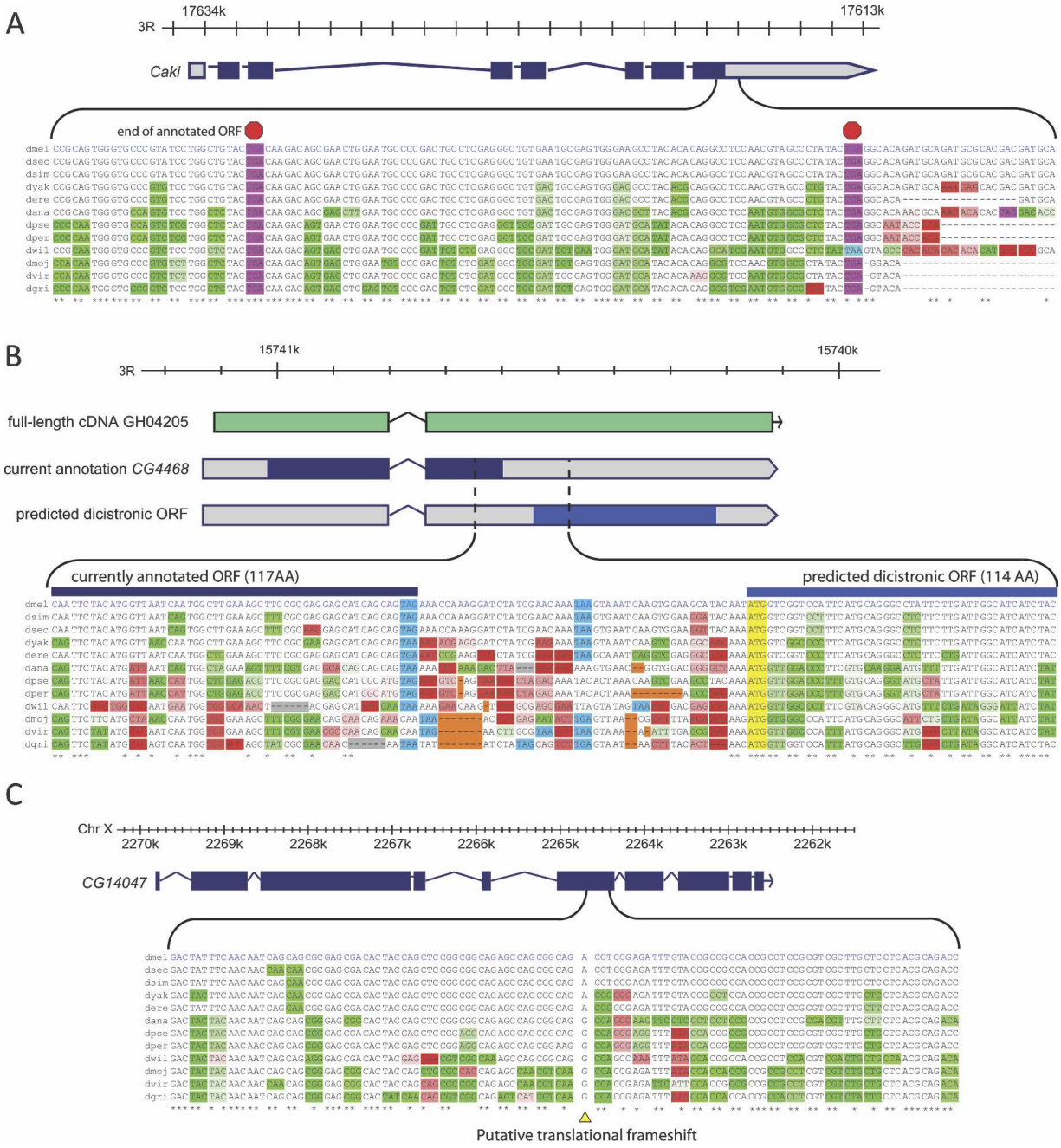


Figure 5. Unusual protein-coding structures identified by evolutionary signatures. (*A*) A well-conserved 30-aa ORF immediately following the stop codon in the gene *Caki* suggests translational readthrough. Note the perfect conservation of the putative readthrough stop codon, the "wobble" of the downstream stop codon, and the precipitous loss of conservation following the downstream stop codon, typical of a true translation stop. (*B*) A well-conserved ORF within the annotated 3′ UTR of *CG4468* suggests a dicistronic transcript structure. Note the region of poor conservation that extends precisely from the upstream stop codon to the downstream start codon, suggesting separate translation of the two ORFs. (*C*) An abrupt change in the reading frame upon which selection appears to act within an exon of *CG14047* is suggestive of a "programmed" translational frameshift (see also Supplemental Fig. 2).

splice forms. (2) In addition to discovering new genes, we used evolutionary signatures to revisit existing gene annotations. This led to confirmation that 87% of CGid-named annotations show evolutionary signatures of protein-coding genes and, conversely, to the identification of 3%–4% of CGid-only annotations that are likely to be spurious predictions or noncoding genes. (3) At a finer-grain level, evolutionary signatures allowed us to propose detailed refinements to hundreds of existing annotations, adjusting the translation start codon, correcting splice boundaries, resolving the functional reading frame in short single-exon transcripts, and identifying strain-specific disrupting mutations. (4) Lastly, the power of evolutionary signatures enabled us to recognize unusual gene structures, which challenge the current assumptions of gene annotation efforts: We found abundant evi-

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

dence of stop codon readthrough, polycistronic transcripts, and several candidates for conserved translational frameshifts.

Challenges for computational prediction of complete gene models

The comparative metrics we used in this study allowed us to distinguish individual protein-coding regions with high predictive value. To tie these exons into complete gene models, we relied on manual curation and large-scale cDNA sequencing experiments directed by our predictions. This allowed us to avoid simplifying assumptions about gene structures typically imposed by de novo gene structure predictors (Brent 2005).

Our results revealed important insights relevant to full gene model prediction. We obtained full-length cDNA clones for 162 of our predicted new exons, many of which fell into surprising gene models, reinforcing the difficulty of de novo gene model prediction. For example, when new exons were discovered within introns of existing genes on the same strand, the simplest expectation would be that they form alternatively spliced transcripts of the surrounding gene. In contrast to this expectation, however, only 56% were alternative transcripts, and the remaining 44% linked to other genes or formed independent transcription units. Such nested and interdigitated genes, as well as mutually exclusive exons within single genes, are refractory to most de novo gene structure predictors.

A further challenge to computational gene structure prediction is presented by exceptional biological phenomena, such as stop codon readthrough, polycistronic transcripts, and translational frameshifts. These are generally assumed to be rare and eukaryotic gene predictors are not built to recognize them. However, 115 dicistronic genes are currently annotated in FlyBase, and our results suggest that the true number may be substantially larger. Similarly, while only one functional translational frameshift has been described in *Drosophila* (Ivanov et al. 1998), our results revealed several new candidates. Most intriguing are the 149 genes we identified as potential targets of stop codon readthrough, which suggest that this phenomenon might be dramatically more common than currently understood (with only three known selenoproteins [Martin-Romero et al. 2001] and a few other cases [Xue and Cooley 1993; Bergstrom et al. 1995; Steneberg et al. 1998]). Although these phenomena may still be considered rare among ~14,000 genes, they represent some of the most intriguing examples of biological versatility, and a complete catalog of protein-coding genes cannot ignore them.

The next major advances in de novo gene prediction methods are likely to come from continued advances in our understanding of the sequence signals governing transcription, splicing, and translation regulation, as well as the advent of more flexible algorithmic frameworks that are well-suited to take advantage of such unconventional signals (Lafferty et al. 2001; Gross et al. 2006; Bernal et al. 2007). Still, the complex and noncanonical gene structures described above present challenges that appear difficult to overcome even for this next generation of eukaryotic gene structure predictors.

Applying the evolutionary signature approach to other target genomes

We believe that the work described in this report clearly demonstrates the power and practicality of our approach to improve the gene annotations of important genomes by complementing existing methodologies, including de novo gene structure prediction, large-scale cDNA sequencing, and manual curation. Our methods are directly applicable to other genome annotation projects that have this infrastructure in place, including the human (Harrow et al. 2006).

More generally, the preexisting, high-quality annotations for *D. melanogaster* allowed us to demonstrate the high sensitivity and specificity of the RFC and CSF tests based on evolutionary signatures. Since these signatures are universal consequences of natural selection and the genetic code, our results suggest that they can provide a strong foundation for the identification of protein-coding genes within any group of closely related species, even when cDNA library sequences are not immediately available or when no genomes with high-quality annotations exist in closely related taxa. Furthermore, it may also be possible to define specific evolutionary signatures—beyond mere sequence conservation—for other classes of functional elements, which suggests a general approach for the identification of functional elements in any genome. The derivation of reliable gene models for protein-coding genes remains a challenge, especially given the abundance of complex gene structures in metazoan genomes. It is also inherently difficult for comparative genomic methods to identify very fast-evolving, species-specific genes, which are centrally important to the study of evolution, speciation, and immunity. Thus, the complete genome annotation of any species will continue to be most effectively pursued through the concerted efforts of computational predictions, manual curation, and large-scale cDNA sequencing.

Methods

Genome alignments

We used several different sets of multiple sequence alignments of the 12 *Drosophila* genomes in this study. Two were based on a synteny map generated by Mercator (C. Dewey [University of Wisconsin, Madison] and L. Pachter [University of California at Berkeley]), with sequence alignments generated by MAVID (Bray and Pachter 2004) and Pecan (B. Paten and E. Birney, European Bioinformatics Institute, Cambridge, UK). Additionally, we used a MULTIZ (Blanchette et al. 2004) alignment of the 12 *Drosophila* genomes and three other insect genomes, excluding the non-*Drosophila* species. We used the Mercator/MAVID synteny-anchored alignments for predicting new exons and all three alignment sets for evaluating existing gene models (taking the highest-scoring version of the gene from the three alignments in order to have some robustness against alignment errors; see Supplemental Methods for details).

Reading frame conservation (RFC)

The RFC score was computed as we have previously described (Kellis et al. 2004). Briefly, given a region of the genome, a pairwise score between *D. melanogaster* in each informant is computed as the percentage of *D. melanogaster* nucleotides in the same reading frame in the informant (taking the largest such percentage out of the three possible reading frames). Each informant then votes +1, −1, or 0 based on an informant-specific cutoff on the pairwise RFC score: +1 if the score is above, −1 if the score is below, or 0 if there was no sequence aligned. These votes are then summed to obtain an overall score for the region.

Codon substitution frequencies (CSF)

CSF assigns a score to each pairwise codon substitution between *D. melanogaster* and an informant equal to the log-likelihood ra-

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

tio of observing that substitution in coding sequence versus noncoding sequence, conditioned on observing a substitution of the *D. melanogaster* codon. These log-likelihood ratio scores, shown in Figure 1B, were computed from codon distance matrices estimated by counting the frequencies of codon substitutions in alignments of annotated genes and noncoding regions for the appropriate pair of species, similar to the BLOSUM amino acid distance matrices estimated from protein alignments (Henikoff and Henikoff 1992). To obtain a score for a given genomic region, the scores of all codon substitutions in its alignment were summed; no score was assigned to gapped or perfectly conserved codons. With multiple informant sequences, the median of the scores of all codon substitutions in each codon column was used as the score of that column, and the score of each column was summed to score the region (see Supplemental Methods for complete details about CSF).

Thorough benchmarks of the RFC and CSF metrics, as well as various other discriminative metrics for protein-coding gene identification, with different alignments and different combinations of informant species, are presented elsewhere (M.F. Lin, A. Deoras, M. Rasmussen, and M. Kellis, in prep.).

"Confirming" genes

We obtained alignments for every transcript model in FlyBase annotation Release 4.3 by extracting them from the genome alignments (see Supplemental Methods). We scored each transcript by the CSF and RFC metrics and normalized the scores by length. Additionally, we scored thousands of disjoint intervals of at least 300 nt, selected uniformly at random from the noncoding part of the euchromatic genome. To define the test for "confirmation," we chose simple cutoffs on the metrics that exclude >99.9% of the control regions (see Table 1 and Supplemental Methods for specific cutoffs used).

"Rejecting" genes

Our test for identifying "rejected" genes was performed by computing the CSF score over every overlapping 30-aa window in every transcript model for each gene. Additionally, we computed these scores using all three genome alignment sets and using three different subsets of the informant species, representing all 12 *Drosophila* genomes, the subgenus *Sophophora*, and the *melanogaster* group. We took the highest scoring 30-amino-acid window in each gene, out of all its transcripts, all of the alignments, and all of the phylogenetic clades, as the score for that gene. We observed the distribution of this score to be bimodal, chose a cutoff to isolate the lower distribution, and found it to closely resemble our random controls (Supplemental Fig. 1).

Predicting new exons

We integrated our evolutionary metrics as features into a semi-Markov conditional random field (SMCRF), a probabilistic model similar to a generalized hidden Markov model but with more flexibility to directly incorporate discriminative metrics such as RFC and CSF (Lafferty et al. 2001; Sarawagi and Cohen 2005). The SMCRF uses the evolutionary metrics to predict only individual exons, not complete gene structures, and therefore may be considered more similar to interval segmentation algorithms that define the boundaries of high-scoring regions than to full gene predictors. The other features used by the SMCRF include sequence-based splice site discriminators (Yeo and Burge 2004), start/stop codon indicator functions, and a length distribution feature; however, it did not contain any explicit coding sequence composition features (e.g., high-order Markov models), nor did it use any information about transcript sequence evidence or homology with known proteins. The SMCRF had seven segment labels or "states": one for each codon position (reading frame) on each strand and one noncoding. The model was trained by maximum conditional likelihood using a training set of known genes, and the Viterbi algorithm was used to generate exon predictions for the whole genome in the Mercator/MAVID alignments (see Supplemental Methods for further details).

Selection of exon candidates for cDNA isolation

We used self-ligation of inverse PCR products (Hoskins et al. 2005; Wan et al. 2006) to screen four cDNA libraries to obtain clones that contained the predicted conserved exons using a modified primer design strategy. Primers were designed for optimal PCR conditions eliminating the requirement for 5′ bias in placement; 172 predicted exons failed the primer design step of our cDNA screening strategy because they were either too small or not unique in the genome. Of the 434 remaining candidates for validation (after exclusion of predictions with existing EST evidence and other filters; see main text), we selected 184 for validation by maximizing the genomic separation between tested predictions. After cloning of the PCR product four sequencing reads were produced: one from each cDNA end and one from each PCR primer. The composite sequence was used to evaluate whether the clone matched the targeted exon. Clones that matched were selected for complete sequencing.

RT-PCR

We extracted total RNA from the *D. melanogaster* sequenced strain at four time points (0–12-h embryos, 12–24-h embryos, first instar larvae, and adults), using the Micro-to-Midi Total RNA Purification System (Invitrogen). Processed mRNA was isolated using Oligotex mini mRNA Kit (Qiagen) and RT-PCR was performed using the OneStep RT-PCR kit (Qiagen) or Invitrogen Superscript II as reverse transcriptase. Gene-specific primers were designed using primer3 (http://primer3.sourceforge.net/). Primers were 20–24 bp in length and were designed to cross intron/exon boundaries. PCR products were directly sequenced and aligned to the genome using est2genome (http://emboss.sourceforge.net). Primer and amplicon sequences were deposited into GenBank under accession nos. ES439769–ES439782.

Refinements to existing annotations and unusual gene structures

We performed directed computational searches for these phenomena using RFC and CSF, and used the resulting lists to guide manual inspection and/or downstream computational analyses, leading to the choice of final cutoffs and data sets. For example, to identify likely recent nonsense mutations, we identified high-scoring regions in FlyBase transcripts downstream from *D. melanogaster* stop codons that align to sense codons in the other species; to identify possible stop codon readthrough genes, we identified similar cases where the stop codon is conserved across the informant species. See Supplemental Methods for further details and the cutoffs used.

Acknowledgments

We are indebted to the community effort for sequencing, assembly, and alignment of the 12 *Drosophila* genome sequences without which this project would not have been possible, and for the early release and collaborative data sharing. We thank Andy Clark, Tim Sackton, and Tony Greenberg for helpful discussions on lineage-specific genes; Gene Yeo and Jade Vinson for sharing code for a splice site discriminator; and Alex Stark, Pouya Kher-

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

apdour, Matt Rasmussen, Ameya Deoras, Josh Grochow, Erez Lieberman, and Aviva Presser for invaluable discussions.

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Andrews, J., Smith, M., Merakovsky, J., Coulson, M., Hannan, F., and Kelly, L.E. 1996. The stoned locus of *Drosophila melanogaster* produces a dicistronic transcript and encodes two distinct polypeptides. *Genetics* **143:** 1699–1711.

Bass, B.L. 2002. RNA editing by adenosine deaminases that act on RNA. *Annu. Rev. Biochem.* **71:** 817–846.

Bergman, C.M., Pfeiffer, B.D., Rincon-Limas, D.E., Hoskins, R.A., Gnirke, A., Mungall, C.J., Wang, A.M., Kronmiller, B., Pacleb, J., Park, S., et al. 2002. Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. *Genome Biol.* **3**. doi: 10.1186/gb-2002-3-12-research0086.

Bergstrom, D.E., Merli, C.A., Cygan, J.A., Shelby, R., and Blackman, R.K. 1995. Regulatory autonomy and molecular characterization of the *Drosophila* out at first gene. *Genetics* **139:** 1331–1346.

Bernal, A.E., Crammer, K., Hatzigeorgiou, A., and Pereira, F.C.N. 2007. Global discriminative learning for higher-accuracy computational gene prediction. *PLoS Comput. Biol.* 3: e54. doi: 10.1371/journal.pcbi.0030054.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14:** 708–715.

Bray, N. and Pachter, L. 2004. MAVID: Constrained ancestral alignment of multiple sequences. *Genome Res.* **14:** 693–699.

Brent, M.R. 2005. Genome annotation past, present, and future: How to define an ORF at each locus. *Genome Res.* **15:** 1777–1786.

Brogna, S. and Ashburner, M. 1997. The Adh-related gene of *Drosophila melanogaster* is expressed as a functional dicistronic messenger RNA: Multigenic transcription in higher organisms. *EMBO J.* **16:** 2023–2031.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Casey, J.L. and Gerin, J.L. 1995. Hepatitis D virus RNA editing: Specific modification of adenosine in the antigenomic RNA. *J. Virol.* **69:** 7593–7600.

Castellano, S., Morozova, N., Morey, M., Berry, M.J., Serras, F., Corominas, M., and Guigo, R. 2001. In silico identification of novel selenoproteins in the *Drosophila melanogaster* genome. *EMBO Rep.* **2:** 697–702.

Celniker, S.E., Wheeler, D.A., Kronmiller, B., Carlson, J.W., Halpern, A., Patel, S., Adams, M., Champe, M., Dugan, S.P., Frise, E., et al. 2002. Finishing a whole-genome shotgun: Release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biol.* **3**. doi: 10.1186/gb-2002-3-12-research0079.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A., and Johnston, M. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301:** 71–76.

Farabaugh, P.J. 1996. Programmed translational frameshifting. *Microbiol. Rev.* **60:** 103–134.

Giedroc, D.P., Theimer, C.A., and Nixon, P.L. 2000. Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting. *J. Mol. Biol.* **298:** 167–185.

Gross, S.S., Russakovsky, O., Do, C.B., and Batzoglou, S. 2006. Training conditional random fields for maximum labelwise accuracy. In *Advances in Neural Information Processing Systems 19 (NIPS 2006)*. http://books.nips.cc/papers/files/nips19/NIPS2006_0891.pdf.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.K., Chrast, J., Lagarde, J., Gilbert, J.G., Storey, R., Swarbreck, D., et al. 2006. GENCODE: Producing a reference annotation for ENCODE. *Genome Biol.* **7:** S4. doi: 10.1186/gb-2006-7-s1-s4.

Henikoff, S. and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89:** 10915–10919.

Hoskins, R.A., Stapleton, M., George, R.A., Yu, C., Wan, K.H., Carlson, J.W., and Celniker, S.E. 2005. Rapid and efficient cDNA library screening by self-ligation of inverse PCR products (SLIP). *Nucleic Acids Res.* **33:** e185. doi: 10.1093/nar/gni184.

Ivanov, I.P., Simin, K., Letsou, A., Atkins, J.F., and Gesteland, R.F. 1998. The *Drosophila* gene for antizyme requires ribosomal frameshifting for expression and contains an intronic gene for snRNP Sm D3 on the opposite strand. *Mol. Cell. Biol.* **18:** 1553–1561.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241–254.

Kellis, M., Patterson, N., Birren, B., Berger, B., and Lander, E.S. 2004. Methods in comparative genomics: Genome correspondence, gene identification and regulatory motif discovery. *J. Comput. Biol.* **11:** 319–355.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res.* **12:** 996–1006.

Korf, I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5:** 59. doi: 10.1186/1471-2105-5-59.

Kryukov, G.V., Kryukov, V.M., and Gladyshev, V.N. 1999. New mammalian selenocysteine-containing proteins identified with an algorithm that searches for selenocysteine insertion sequence elements. *J. Biol. Chem.* **274:** 33888–33897.

Lafferty, J., McCallum, A., and Pereira, F. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282–289. Morgan Kaufmann, San Francisco.

Luo, G.X., Chao, M., Hsieh, S.Y., Sureau, C., Nishikura, K., and Taylor, J. 1990. A specific base transition occurs on replicating hepatitis delta virus RNA. *J. Virol.* **64:** 1021–1027.

Manak, J.R., Dike, S., Sementchenko, V., Kapranov, P., Biemar, F., Long, J., Cheng, J., Bell, I., Ghosh, S., Piccolboni, A., et al. 2006. Biological function of unannotated transcription during the early development of *Drosophila melanogaster*. *Nat. Genet.* **38:** 1151–1158.

Martin-Romero, F.J., Kryukov, G.V., Lobanov, A.V., Carlson, B.A., Lee, B.J., Gladyshev, V.N., and Hatfield, D.L. 2001. Selenium metabolism in *Drosophila*: Selenoproteins, selenoproteins mRNA expression, fertility, and mortality. *J. Biol. Chem.* **276:** 29798–29804.

Misra, S., Crosby, M.A., Mungall, C.J., Matthews, B.B., Campbell, K.S., Hradecky, P., Huang, Y., Kaminker, J.S., Millburn, G.H., Prochnik, S.E., et al. 2002. Annotation of the *Drosophila melanogaster* euchromatic genome: A systematic review. *Genome Biol.* **3**. doi: 10.1186/gb-2002-3-12-research0083.

Nekrutenko, A., Makova, K.D., and Li, W.H. 2002. The K_A/K_S ratio test for assessing the protein-coding potential of genomic regions: An empirical and simulation study. *Genome Res.* **12:** 198–202.

Parra, G., Blanco, E., and Guigo, R. 2000. GeneID in *Drosophila*. *Genome Res.* **10:** 511–515.

Reese, M.G., Kulp, D., Tammana, H., and Haussler, D. 2000. Genie—Gene finding in *Drosophila melanogaster*. *Genome Res.* **10:** 529–538.

Richards, S., Liu, Y., Bettencourt, B.R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M.J., Chen, R., Meisel, R.P., et al. 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene, and *cis*-element evolution. *Genome Res.* **15:** 1–18.

Rubin, G.M., Hong, L., Brokstein, P., Evans-Holm, M., Frise, E., Stapleton, M., and Harvey, D.A. 2000. A *Drosophila* complementary DNA resource. *Science* **287:** 2222–2224.

Sarawagi, S. and Cohen, W. 2005. Semi-Markov conditional random fields for information extraction. *Adv. Neural Inf. Process. Syst.* **17:** 1185–1192.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15:** 1034–1050.

Stanke, M. and Waack, S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19:** II215–II225.

Stapleton, M., Carlson, J., Brokstein, P., Yu, C., Champe, M., George, R., Guarin, H., Kronmiller, B., Pacleb, J., Park, S., et al. 2002a. A *Drosophila* full-length cDNA resource. *Genome Biol.* **3**. doi: 10.1186/gb-2002-3-12-research0080.

Stapleton, M., Liao, G., Brokstein, P., Hong, L., Carninci, P., Shiraki, T., Hayashizaki, Y., Champe, M., Pacleb, J., Wan, K., et al. 2002b. The *Drosophila* gene collection: Identification of putative full-length cDNAs for 70% of *D. melanogaster* genes. *Genome Res.* **12:** 1294–1300.

Stark, A., Kheradpour, P., Parts, L., Brennecke, J., Hodges, E., Hannon, G.J., and Kellis, M. 2007. Systematic discovery and characterization of fly microRNAs using 12 Drosophila genomes. *Genome Res.* (this issue). doi: 10.1101/gr.6593807.

Steneberg, P., Englund, C., Kronhamn, J., Weaver, T.A., and Samakovlis, C. 1998. Translational readthrough in the hdc mRNA generates a novel branching inhibitor in the *Drosophila* trachea. *Genes & Dev.* **12:** 956–967.

Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel,

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

A.C., Thomas, P.J., McDowell, J.C., et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424:** 788–793.

Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., Shu, S., Lewis, S.E., Richards, S., Ashburner, M., Hartenstein, V., Celniker, S.E., et al. 2002. Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol.* **3**. doi: 10.1186/gb-2002-3-12-research0088.

Wan, K.H., Yu, C., George, R.A., Carlson, J.W., Hoskins, R.A., Svirskas, R., Stapleton, M., and Celniker, S.E. 2006. High-throughput plasmid cDNA library screening. *Nat. Protoc.* **1:** 624–632.

Xue, F. and Cooley, L. 1993. kelch encodes a component of intercellular bridges in *Drosophila* egg chambers. *Cell* **72:** 681–693.

Yandell, M., Bailey, A.M., Misra, S., Shu, S., Wiel, C., Evans-Holm, M., Celniker, S.E., and Rubin, G.M. 2005. A computational and experimental approach to validating annotations and gene predictions in the *Drosophila melanogaster* genome. *Proc. Natl. Acad. Sci.* **102:** 1566–1571.

Yang, Z. and Bielawski, J.P. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15:** 496–503.

Yeo, G. and Burge, C.B. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11:** 377–394.

Received May 7, 2007; accepted in revised form September 21, 2007.

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

Evolution, biogenesis, expression, and target predictions of a substantially expanded set of *Drosophila* microRNAs

J. Graham Ruby,[1,2] Alexander Stark,[3,4] Wendy K. Johnston,[1,2] Manolis Kellis,[3,4] David P. Bartel,[1,2,6] and Eric C. Lai[5]

[1]Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA; [2]Howard Hughes Medical Institute and Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; [3]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; [4]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02141, USA; [5]Department of Developmental Biology, Memorial Sloan-Kettering Cancer Center, New York, New York 10021, USA

MicroRNA (miRNA) genes give rise to small regulatory RNAs in a wide variety of organisms. We used computational methods to predict miRNAs conserved among *Drosophila* species and large-scale sequencing of small RNAs from *Drosophila melanogaster* to experimentally confirm and complement these predictions. In addition to validating 20 of our top 45 predictions for novel miRNA loci, the large-scale sequencing identified many miRNAs that had not been predicted. In total, 59 novel genes were identified, increasing our tally of confirmed fly miRNAs to 148. The large-scale sequencing also refined the identities of previously known miRNAs and provided insights into their biogenesis and expression. Many miRNAs were expressed in particular developmental contexts, with a large cohort of miRNAs expressed primarily in imaginal discs. Conserved miRNAs typically were expressed more broadly and robustly than were nonconserved miRNAs, and those conserved miRNAs with more restricted expression tended to have fewer predicted targets than those expressed more broadly. Predicted targets for the expanded set of microRNAs substantially increased and revised the miRNA-target relationships that appear conserved among the fly species. Insights were also provided into miRNA gene evolution, including evidence for emergent regulatory function deriving from the opposite arm of the miRNA hairpin, exemplified by *mir-10*, and even the opposite strand of the DNA, exemplified by *mir-iab-4*.

[Supplemental material is available online at www.genome.org. The small RNA sequence data from this study have been submitted to GEO under accession nos. GPL5061 and GSE7448. Computational tools for miRNA prediction (MiRscan3) are available for anonymous download at http://web.wi.mit.edu/bartel/pub/.]

MicroRNAs (miRNAs) are ~23-nt RNA species that direct the post-transcriptional repression of messenger RNAs (mRNAs) (Bartel 2004). They are generated from primary transcripts (pri-miRNAs) that can fold into characteristic hairpin secondary structures. In animals, those hairpins are typically first cleaved away from the rest of the primary transcript by the nuclear RNase III enzyme Drosha to generate miRNA precursors (pre-miRNA), and are then cleaved near their loops by the cytoplasmic RNase III enzyme Dicer to generate a heteroduplex of two ~23-nt RNAs (Lee et al. 2003). The mature miRNA is preferentially packaged into the RNA-induced silencing complex (RISC), while the other species, known as the miRNA star (miRNA*), is discarded (Lau et al. 2001; Lim et al. 2003b). The decision as to which species is incorporated into the silencing complex is influenced by the difference in pairing stabilities between the two ends of the miRNA:miRNA* duplex, with preferential incorporation of the strand whose 5′ end is less stably paired (Khvorova et al. 2003; Schwarz et al. 2003).

Once incorporated into the silencing complex, metazoan miRNAs pair to the messages of their mRNA targets, primarily in 3′ untranslated regions (3′ UTRs). Complementarity between the message and a segment in the 5′ region of the miRNA known as the "seed" (miRNA nucleotides 2–7) appears to be the most crucial requirement of target recognition. Conserved pairing to the seed region is a feature of most genetically identified miRNA–target interactions (Lee et al. 1993; Wightman et al. 1993; Lai 2002). Indeed, the requirement of conserved pairing to the miRNA seed enables miRNA targets to be predicted in excess of the noise of false-positive predictions (Lewis et al. 2003; Brennecke et al. 2005; Krek et al. 2005; Lewis et al. 2005). Short, 7- to 8-nt sites matching the seed region of the miRNA are not only important but sometimes can suffice for repression in reporter assays (Doench and Sharp 2004; Brennecke et al. 2005; Lai et al. 2005). Consistent with the in vivo sufficiency of 7-mer seed-matching sites in mediating repression, many messages preferentially coexpressed with a highly expressed miRNA are depleted in 7-mer sites matching that miRNA, presumably because of selective avoidance of miRNA-mediated repression during evolution (Farh et al. 2005; Stark et al. 2005). Moreover, miRNAs that share the same seed sequence but are diverse throughout the remainder of their sequences can be functionally redundant (Abbott et al. 2005; Lim et al. 2005), which justifies their grouping into members of the same miRNA "family" (Lewis et al. 2003).

[6]Corresponding author.
E-mail dbartel@wi.mit.edu; fax (617) 258-6768.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6597907. Freely available online through the *Genome Research* Open Access option.

17:1850–1864 ©2007 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/07; www.genome.org

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

The seeds that define families are often conserved throughout diverse species even as the individual miRNA genes within the family vary (Ruby et al. 2006). The arms of the hairpin precursors are less conserved than the seeds, but are more conserved than either the surrounding genomic sequence or the intervening loop sequence (Lai et al. 2003; Lim et al. 2003b).

Most known miRNAs were discovered through the cloning and sequencing of small-RNA cDNAs (Griffiths-Jones 2004). However, this method can miss miRNAs expressed at low levels or in only specific cell types or conditions. One approach for identifying low-abundance miRNAs that has previously been applied in *Drosophila* is to identify candidate miRNA hairpins computationally and then validate their expressions using more directed, and thereby potentially more sensitive, experimental methods (Lai et al. 2003). Because identification of plausible candidates is aided by comparative genomics, this approach gains efficacy as the genome sequences of additional related species become available. A second approach for identifying rare miRNAs is simply to increase the scale of small-RNA sequencing well beyond the reach of prior efforts. This high-throughput sequencing approach has not been used previously in insects, but in other lineages it has revealed miRNAs and miRNA candidates that escaped earlier detection because they are rare or not well conserved in related genomes (Berezikov et al. 2006; Lu et al. 2006; Rajagopalan et al. 2006; Ruby et al. 2006; Fahlgren et al. 2007).

Here, we use the two complementary approaches described above—computational prediction and high-throughput sequencing—to identify nearly 60 additional fly miRNAs and to refine the descriptions of about half of those that had been previously annotated. These results provided insights into miRNA evolution, biogenesis, and expression in insects. When combined with improved target prediction, which used information from all 12 sequenced *Drosophila* genomes (*Drosophila* 12 Genomes Consortium 2007; Stark et al. 2007c) to increase prediction accuracy, these new and revised miRNAs substantially expanded and improved the view of miRNA-directed regulation in flies.

Results

Computational prediction of fly miRNAs

Novel miRNA genes were sought computationally as hairpins with secondary structure and conservation patterns resembling those of previously annotated miRNAs, using an approach with similarities to that described for MiRscan, which had previously been applied to nematodes and vertebrates (Lim et al. 2003a,b). For each of six *Drosophila* genomes (*Drosophila melanogaster*, *Drosophila ananassae*, *Drosophila pseudoobscura*, *Drosophila mojavensis*, *Drosophila virilis*, and *Drosophila grimshawi*) (Adams et al. 2000; Richards et al. 2005; *Drosophila* 12 Genomes Consortium 2007; Stark et al. 2007c), RNAfold (Hofacker et al. 1994) was used to identify candidate hairpins from across the entire genome. Candidate hairpins from each genome were first scored based on the relative frequencies of structural characteristics in the background candidate set versus a foreground training set of annotated miRNA hairpins. This training set comprised 37 miRNA hairpins from *D. melanogaster* that were selected randomly from the 78 previously annotated in miRBase v8.1 (Griffiths-Jones 2004); the remaining 41 miRNAs were reserved as a test set, the performance of which was not evaluated until after the completion of the prediction process. Candidates with scores far below that of the lowest foreground hairpin were removed from the background set, altering its aggregate properties. Unlike the previous method, candidates were then re-scored based on the properties of the minimized background, and the worst candidates were again eliminated. Following several rounds of eliminating candidates from each individual genome, candidates from different genomes (nodes) were paired as putative orthologs (edges) using BLAST (Altschul et al. 1990) and put through the same process of iterative scoring and elimination, now simultaneously evaluating conservation.

The surviving ortholog pairs included 565 hairpin candidates from *D. melanogaster* that could form complete networks between all six genomes, including all 15 possible edges. These successful candidates were ranked by the sum of the 15 pairwise scores from their first round of pairwise scoring and elimination (Supplemental Table S1). Of the 37 members of the training set, 26 survived and fell mostly within the higher scoring tail of the distribution (Fig. 1A). Examination of the test set revealed that 35 of 41 members of the test set survived and that these 35 hairpins had similar score distributions as the training set, indicating that our prediction method did not overfit to the training-set data (Fig. 1A).

Our concept of a candidate miRNA hairpin specified the genomic strand from which the hairpin was derived. However, the secondary structure and conservation properties of a genomic

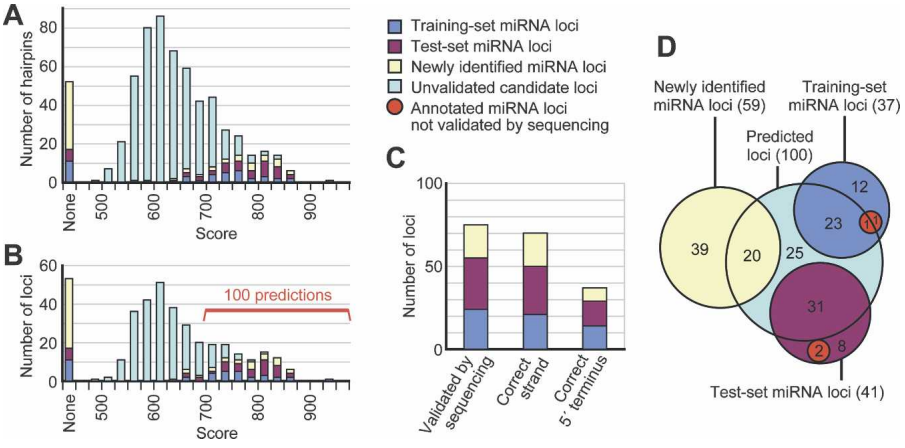


Figure 1. Performance of miRNA gene prediction. (*A*) The summed pairwise scores across all 15 two-species comparisons for each miRNA hairpin candidate. Those candidates overlapping the training, test, newly identified, and unvalidated sets of miRNA hairpins are colored as indicated in the key (*right*) and listed (Supplemental Table S1). (*B*) The candidate loci, following strand collapse and exon filtering, depicted as in *A*. The top 100 candidates, which had scores >698, were carried forward as the set of computational gene predictions (Supplemental Table S1). Of the remaining candidates, only a few were likely to be authentic miRNAs. (*C*) Specificity of the 100 predictions. Plotted are the number of predicted loci that were validated, the number that correctly identified the strand of the miRNA gene, and the number that correctly identified the miRNA 5′ end (Supplemental Table S1), colored as in *A*. (*D*) The overlap of the 100 predicted miRNA loci with the training set, test set, and newly identified miRNA loci. Two loci from the training set and two from the test set were not validated by sequencing (red).

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

sequence frequently match those of its reverse complement. As a result, 174 of our 565 candidate hairpins were paired with a candidate locus from the opposing genomic strand and thus represented only 87 unique genomic loci. We therefore collapsed our predictions into 478 strand-independent genomic loci, which each included a prediction of which strand would give rise to the mature miRNA based on the higher score. Eliminating 151 candidates that overlapped the annotated exons of protein-coding genes (Supplemental Table S1) left 327 candidate loci (Fig. 1B). The top 100 candidate loci were carried forward as our predictions. These included 55 of the previously annotated genes (24 of the 26 surviving training-set genes), as well as 45 novel predictions.

Recent results from plants and worms demonstrate that for the validation of miRNA expression, large-scale sequence data sets are more reliable and sensitive compared to RNA blotting, and more reliable than and roughly equally sensitive as PCR assays (Rajagopalan et al. 2006; Ruby et al. 2006). We therefore used large-scale small RNA sequence data to evaluate the quality of our predictions.

High-throughput sequencing of small RNAs

To survey the miRNAs of flies, we performed high-throughput pyrosequencing (Margulies et al. 2005; Ruby et al. 2006) on libraries of small RNAs isolated from the following 10 *D. melanogaster* tissues or stages: very early embryo (0–1 h), early embryo (2–6 h), mid-embryo (6–10 h), late embryo (12–24 h), larvae (first and third instars), imaginal discs, pupae (0–4 d), adult heads, adult bodies, and tissue-culture cells (S2). Pyrosequencing yielded a total of 1.14 million small RNA reads (55,761–174,031 reads per library) that perfectly matched the *D. melanogaster* genome.

Refinement of prior miRNA annotations

Of the 54 *D. melanogaster* miRNAs (corresponding to 60 hairpins) that had been previously cloned and sequenced (Lagos-Quintana et al. 2001; Aravin et al. 2003), all 54 were represented in our data set of 1.14 million small RNA reads, as exemplified by miR-7 and miR-iab-4 (Fig. 2), and detailed for all the miRNAs (Supplemental Table S2). For the 60 hairpins of these previously cloned miRNAs, read frequencies ranged from 60 (miR-303) to 20,049 (miR-14), with a median of 2415, and for each of these hairpins the miRNA* species was also recovered. Additional *Drosophila* miRNA genes are annotated in miRBase v.8.1 based on homology with other miRNAs or computational predictions supported by RNA blots (Aravin et al. 2003; Lai et al. 2003). Of these 18 genes for which small RNAs had not been previously cloned, 14 were represented in our data set (Fig. 2A). The four that were missing (*mir-280*, *mir-287*, *mir-288*, and *mir-289*) had been predicted computationally and experimentally supported by RNA blots of samples from early embryos, larvae and pupae, and adult males (Lai et al. 2003). Their absence in our libraries from these same developmental stages called their authenticity into question.

In half of the cases (37 of the 74 confirmed genes), the distribution of reads across the hairpin suggested that the mature miRNA differs from the one that had been previously annotated (Supplemental Table S2). Usually, the discrepancy was only at the 3′ terminus of the mature miRNA, as exemplified by miR-7 (Fig. 2B). Although proper 3′ annotation is needed for some miRNA expression profiling methods (Wang et al. 2007), reannotation of the miRNA 3′ terminus was of little consequence because 3′ heterogeneity is a hallmark of mature miRNAs (Lau et al. 2001; Basyuk et al. 2003; Lim et al. 2003b; Ruby et al. 2006). However, in 12 cases, there was discrepancy at the miRNA 5′ terminus (Supplemental Table S2). The reannotation of a miRNA 5′ terminus is far more consequential because of its role in defining the miRNA seed sequence, which, in turn, defines the set of targets. For example, shifting the 5′ terminus by a single nucleotide changes the identity of one or both of the two 7-mers used for target prediction (Lewis et al. 2005), thereby dramatically altering the set of predicted targets, and shifting it by two or more nucleotides would have an even greater effect. Seven of these 12 cases were corrections of annotations that have been based on computational or molecular evidence not expected to identify the 5′ termini with confidence (Supplemental Text). The other five cases were more interesting because they illustrated how a single miRNA hairpin or paralogous hairpins could spawn new miRNA function.

For miR-210, there were 917 reads with the originally annotated 5′ end and 1031 reads with an extra 5′ nucleotide, all of which mapped uniquely to the genome. Combined, the abundance and equivalence of reads indicated that miR-210 was a rare example of a single hairpin generating mature miRNAs with multiple abundant 5′ ends. As was done for miR-248 in *Caenorhabditis elegans* (Ruby et al. 2006), we annotated the species with an extended 5′ end as miR-210.1 and the species with the originally annotated 5′ end as miR-210.2, with the idea that both probably direct repression in the fly (Supplemental Table S2).

In the case of miR-10, the dominant read was from the arm of the hairpin precursor opposite the annotated miRNA and was sevenfold more abundant (Fig. 3), a result expanding on the observation that species from both arms are easily detected (Schwarz et al. 2003). Because conservation criteria supported the function of RNAs from both arms of the hairpin, and in conjunction with a parallel study (Stark 2007), we annotated the two major products of the *mir-10* hairpin as miR-10-5p and miR-10-3p. The seed of the original miR-10 (miR-10-5p) was conserved throughout all annotated *mir-10* genes, including those of vertebrates (Fig. 3A). The seed of the species more abundantly represented in our data set (miR-10-3p) was not conserved in all annotated *mir-10* genes but was nonetheless conserved in at least one *mir-10* gene of each species examined (typically the *mir-10a* paralogs of vertebrates; Fig. 3A).

The *mir-281* and *mir-2* paralogs illustrated how highly related miRNA genes could have divergent functions. For both sets of paralogs, miRNAs deriving from the miRNA arms of the hairpins could be mapped to multiple related hairpins. Nonetheless, the miRNA* species, which mapped uniquely to their hairpins, revealed the likely processing of each hairpin and indicated that one of the two *mir-281* hairpins and two of the five *mir-2* hairpins gave rise to miRNA species that differed from those previously annotated (Supplemental Text).

Novel miRNAs

Having revised many of the previous miRNA annotations of *Drosophila*, we next examined the overlap between our predicted miRNA loci and the small RNA reads. Of the 100 predictions, 45 had not been previously annotated as miRNA genes. Of those, 20 were supported by the reads. In all 20 cases, our prediction method identified the correct strand of the miRNA gene (as well

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

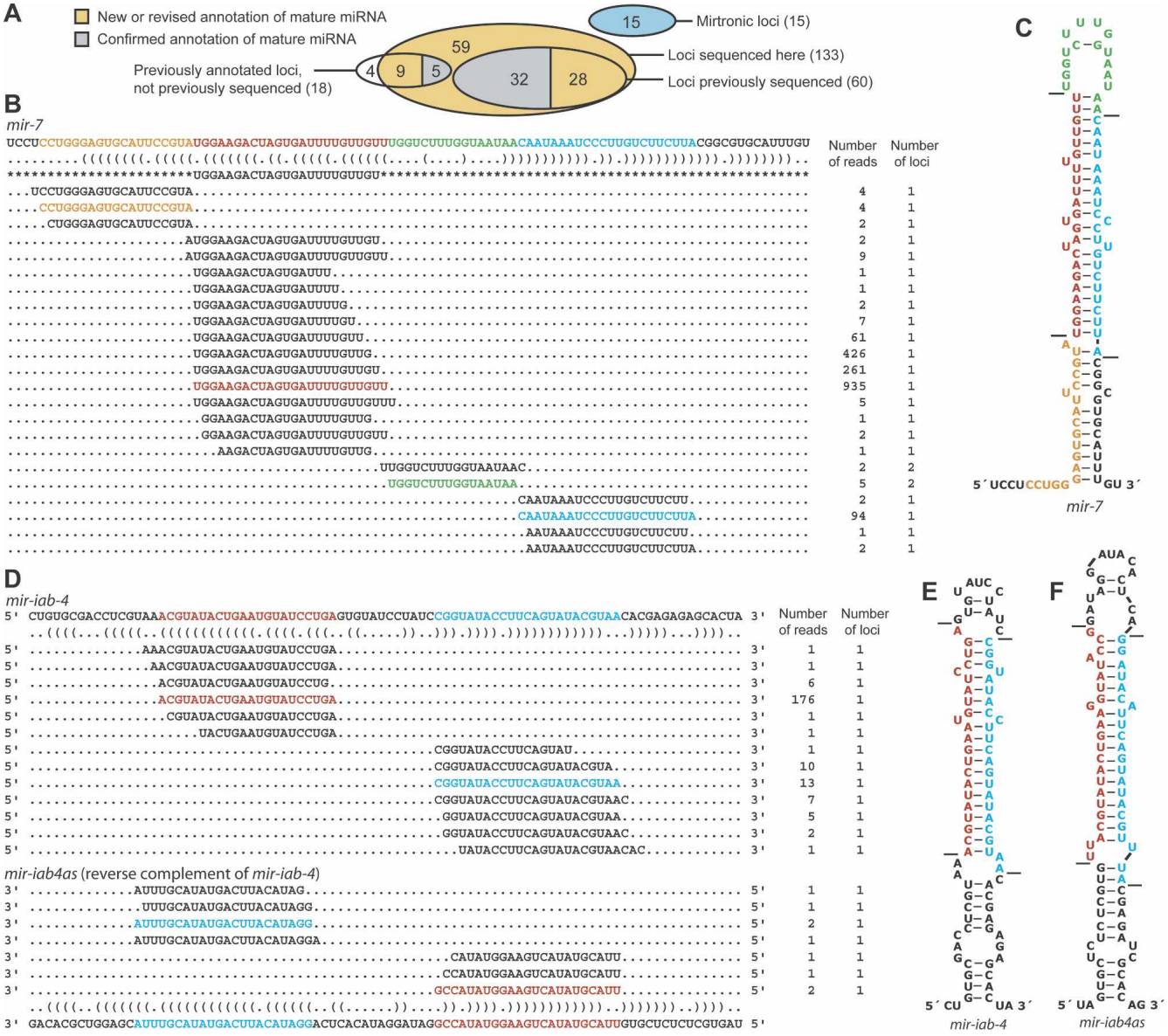


Figure 2. Correspondence between previously annotated miRNA hairpins and sequenced miRNAs. (*A*) Overlap between previously annotated miRNA hairpins and the total set of 133 hairpins of canonical miRNAs supported by our high-throughput sequencing (Supplemental Table S2). Mirtronic loci are described elsewhere (Ruby et al. 2007). (*B*) Small RNAs derived from the *mir-7* hairpin. A portion of the *mir-7* transcript is shown *above* its bracket-notation secondary structure, mature miRNA annotation from miRBase v8.1 (Griffiths-Jones 2004) flanked by asterisks, and sequences from the present study. For each sequence, the number of reads giving rise to that sequence and the number of loci to which the sequence maps in the *D. melanogaster* genome are shown on the *right*. Highlighted are the most abundant sequences corresponding to the miRNA (red), miRNA* (blue), intervening loop (green), and fragment flanking the 5′ Drosha cleavage site (orange) (Supplemental Text). Analogous data for all previously annotated *D. melanogaster* miRNAs are provided (Supplemental Table S2). (*C*) The predicted hairpin structure of the *mir-7* hairpin, colored as in *B*. Lines indicate inferred Drosha and Dicer cleavage sites. (*D*) Small RNAs derived from the *mir-iab-4* and *mir-iab4as* hairpins, displayed as in *B*. (*E*) The predicted secondary structure of the sense *mir-iab-4* hairpin precursor, formatted as in *C*. (*F*) The predicted secondary structure of the *mir-iab-4* reverse complement, *mir-iab4as*, formatted as in *C*.

as 21 of 24 cases from the training set and 29 of 31 cases from the test set). Given correct identification of the miRNA strand, the miRNA 5′ terminus was correctly predicted in eight of 20 cases (plus 14 of 21 cases from the training set and 15 of 29 cases from the test set; Fig. 1C). Some of the remaining 25 predictions that lacked experimental confirmation also might be authentic miRNAs. However, because most might be false positives and because their 5′ termini were not predicted with high confidence, we did not consider any of these 25 suitable for annotation or target-prediction studies.

Looking more broadly at the small RNA reads to identify the miRNAs that were more difficult to recognize computationally increased our count of newly identified miRNA loci from 20 to 59. To confidently identify these as miRNA genes, we considered the following criteria: (1) the pairing characteristics of the hairpin; (2) the miRNA expression, as measured by the abundance of

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

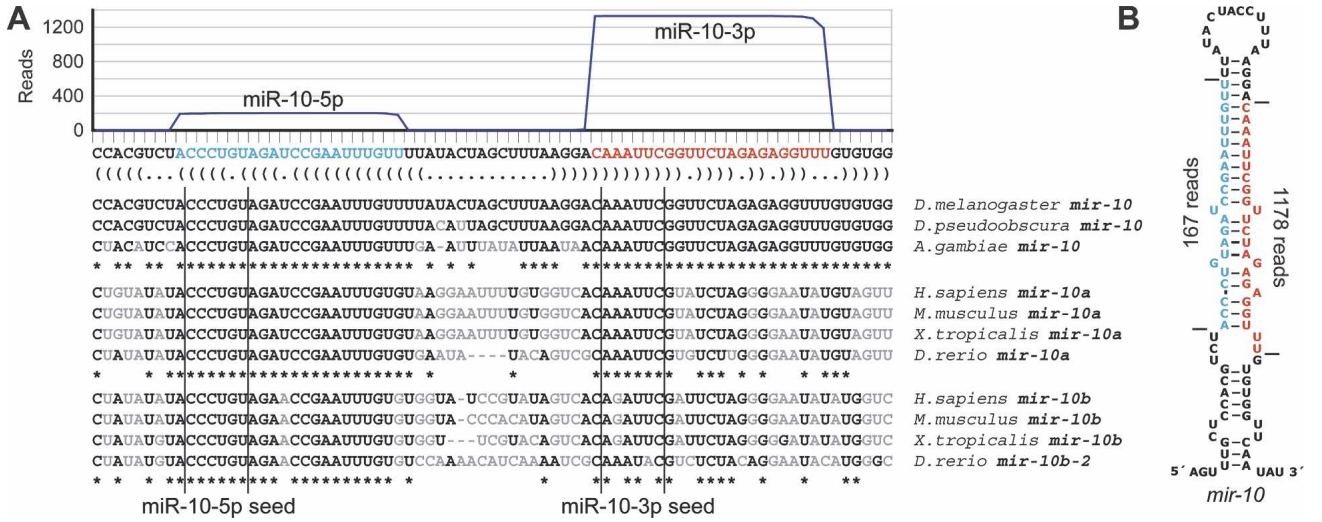


Figure 3. Expression and conservation of *mir-10*. (*A*) The sequence and bracket-notation secondary structure of the *mir-10* hairpin, highlighting the mature miR-10-5p (blue) and the mature miR-10-3p (red), with read abundance along the length of the sequence plotted *above* and orthologous hairpins aligned *below.* Nucleotides differing from the *D. melanogaster* identities are in gray. Vertical lines indicate the edges of the 6-nt seed of each mature RNA. (*B*) The *mir-10* hairpin predicted secondary structure, colored as in *A*. Horizontal lines indicate the inferred Drosha and Dicer cleavage sites.

sequence reads sharing the same 5′ terminus; (3) evolutionary conservation, as evaluated by the apparent conservation of the hairpin in other fly species and grouping of the miRNA candidate into a family based on its seed sequence; (4) the absence of annotation suggesting non-miRNA biogenesis; and (5) the presence of reads corresponding to the predicted miRNA* species.

The observation of both a candidate miRNA and a candidate miRNA* in a set of reads provides particularly compelling evidence for Dicer-like processing from an RNA hairpin (Rajagopalan et al. 2006; Ruby et al. 2006; Fahlgren et al. 2007). As illustrated for *mir-988* (Fig. 4A), 40 newly identified genes satisfied all five of our criteria, and 19 others satisfied a subset of the criteria deemed sufficient for confident annotation as miRNAs (Table 1). Nine additional candidates fell within predicted miRNA-like hairpins and were sequenced more than once (Supplemental Table S2). However, they were considered unlikely to be miRNAs because they did not satisfy the other criteria sufficiently and they mostly mapped to either protein-coding transcripts (candidates 1–5) or heterochromatic DNA (candidates 6–8). Ten of the newly identified miRNAs derived from loci that were among the top 200 predicted to form miRNA precursor hairpins in a previous effort (Lai et al. 2003). Nine of those predictions correctly identified the genomic strand from which the miRNA was derived, but prediction of the mature miRNA had not been attempted.

Two-thirds of the novel miRNAs appeared to be broadly conserved in the *Drosophila* genus (Table 1). Orthologs were sought in six species spanning both the *Sophophora* subgenus (*Drosophila simulans*, *Drosophila yakuba*, *D. ananassae*, and *D. pseudoobscura*) and the *Drosophila* subgenus (*D. mojavensis* and *D. virilis*). Putative orthologs were found in all six species for 28 of the miRNAs and in five of six species for another nine. In 12 of the remaining cases, orthologs were found in two or fewer of the *Drosophila* species.

One newly identified locus, *mir-996*, resided 1.5 kb downstream from a related miRNA (*mir-279*) and within the transcript of *CG31044,* which is annotated as encoding a 140-amino-acid protein (Crosby et al. 2007). We suggest that miR-996, not the putative protein, is the functional product of this gene. Consistent with this proposal, the observed miRNA was perfectly conserved across a wide scope of fly species, whereas in the open reading frame (ORF), sequence polymorphisms in the closely related species *D. simulans*, *D. yakuba*, *D. ananassae*, and *D. pseudoobscura* introduced nonsense mutations at codons 73, 56, 13, and 40, respectively.

Like many known miRNA genes (Griffiths-Jones 2004), 26 of the 59 newly identified loci were clustered with other miRNA loci (Table 1; Fig. 5A), and 13 fell within annotated introns (and from the same genomic strand as the intron; Table 1). Thus, more than half (69 of 133) of the canonical *Drosophila* miRNAs were clustered (Fig. 5A), and more than a quarter (36 of 133) were intronic (Supplemental Table S1).

Although most of the novel miRNA genes closely resembled those previously annotated, three of the hairpin precursors were much larger than those observed previously in animals. For the vast majority of previously annotated fly genes, fewer than 30 nucleotides separated the miRNA and miRNA*, and all had fewer than 60 intervening nucleotides. The distribution of intervening sequence lengths was generally similar among the newly identified miRNAs. However, *mir-956*, *mir-989*, and *mir-997* had abnormally large intervening sequences of 82 nt, 99 nt, and 112 nt, respectively (Fig. 4C,D; Supplemental Table S2). Each of these hairpins gave rise to miRNA* reads, and in each case, the dominant ends of the miRNA versus the miRNA* exhibited 1- or 2-nt 3′ overhangs. In no case was there EST evidence of an intron helping to bridge the distance between the miRNA and miRNA* loci (Karolchik et al. 2003; Crosby et al. 2007). The lack of constraint on the length of the intervening sequence was illustrated by *mir-989*, whose mature miRNA sequence was perfectly conserved across all seven of the *Drosophila* species examined but whose intervening sequence length varied widely, dipping as low as 52 nt in *D. pseudoobscura* (Fig. 4D).

MicroRNA biogenesis in flies

As in nematodes (Ruby et al. 2006), examining the multitude of reads arising from the previously annotated miRNA hairpins pro-

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

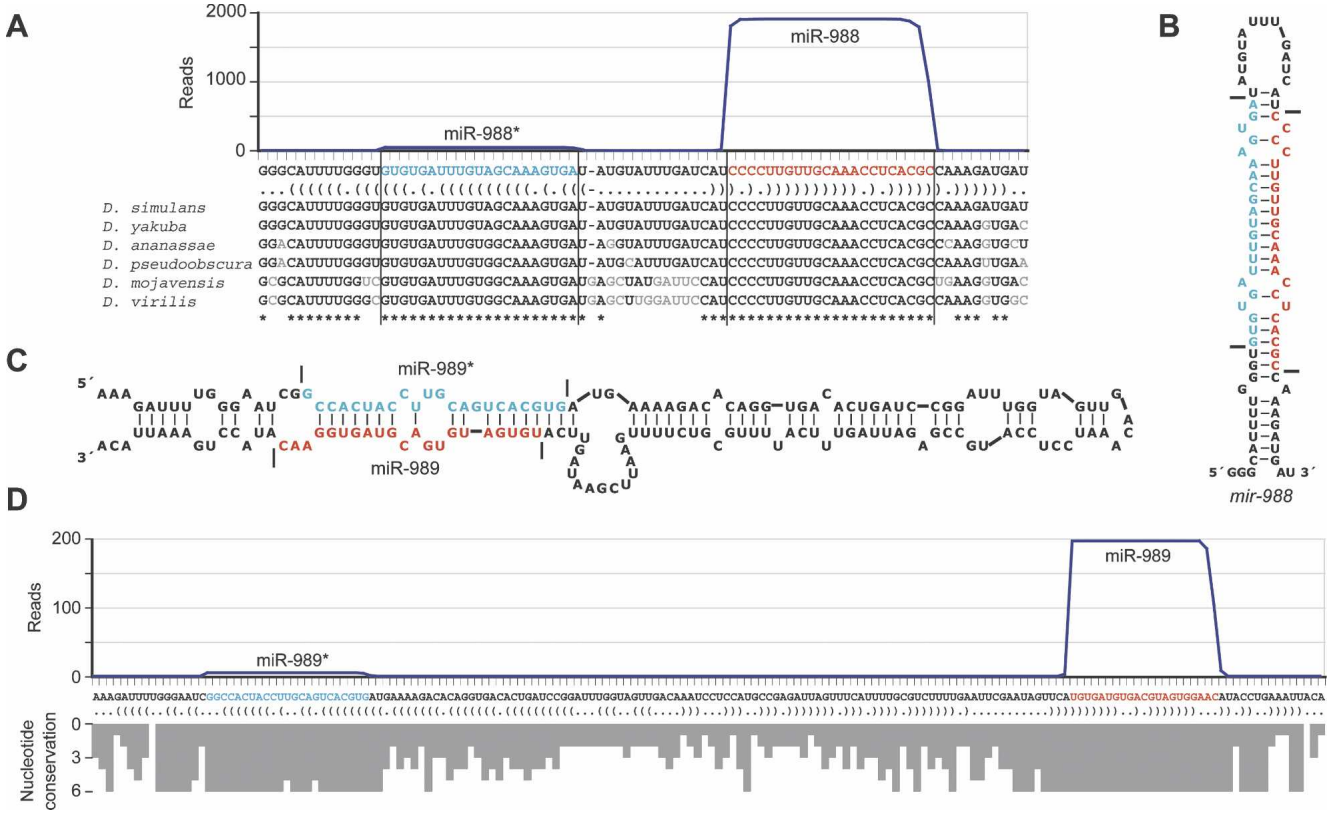


Figure 4. Newly identified miRNAs. (*A*) The sequence and bracket-notation secondary structure of the *mir-988* hairpin, highlighting the miRNA (red) and the miRNA* (blue), with read abundance along the length of the sequence plotted *above* and orthologous hairpins aligned *below;* nonconserved nucleotides in gray (*Drosophila* Comparative Genome Sequencing and Analysis Consortium 2007a,b). Vertical lines indicate the inferred Drosha and Dicer cleavage sites. Analogous data for all newly identified *D. melanogaster* miRNAs are provided (Supplemental Table S2). (*B*) The predicted secondary structure of the *mir-988* hairpin, colored as in *A*. Horizontal lines indicate the inferred Drosha and Dicer cleavage sites. (*C*) The unusually large hairpin of *mir-989*, colored as in *A*. (*D*) The sequence and bracket-notation secondary structure of the *mir-989* hairpin, with coloring and read-abundance display as in *A*. Conservation across the length of the hairpin is shown *below* as a histogram, with bar depth indicating for each nucleotide the number of orthologs from the organisms shown in *A* with that nucleotide conserved.

vided insights into the specificity and precision of Drosha and Dicer processing (Supplemental Table S2). These RNase III enzymes preferentially leave 2-nt 3′ overhangs when cleaving perfect RNA duplexes (Basyuk et al. 2003; Lee et al. 2003), and this configuration between the ends of miRNA and miRNA* species was observed in our data set. The 5′ ends of both the miRNA and miRNA* species were more homogeneous than the 3′ ends, and the miRNA 5′ ends were more consistent than the miRNA* 5′ ends, and the miRNA* 3′ ends were more consistent than the miRNA 3′ ends, regardless of which was generated by Drosha and which by Dicer (Supplemental Table S2; Supplemental Text). The heightened precision of either enzyme when it defined the miRNA seed implied that Dicer does not simply measure from the site of the Drosha cleavage and suggested that additional determinants must be used when needed to more accurately define Dicer cleavage. Similar conclusions arise from the sequencing data in nematodes (Ruby et al. 2006).

As reported in other species (Li et al. 2005; Ruby et al. 2006), untemplated nucleotide addition also contributed to a minor fraction of 3′ heterogeneity. In *Drosophila*, we observed evidence for processivity of the terminal-transferase activity, with a preference for adding adenosines (Supplemental Text).

Reads that were antisense to either the miRNA or miRNA* appeared for four previously annotated hairpins (*mir-iab-4*, *mir-124*, *mir-305*, *mir-307*). The *mir-iab-4* hairpin gave rise to nine antisense reads (Fig. 2D); each of the remaining hairpins gave rise to one read. Antisense transcription from *mir-iab-4* has been noted previously (Cumberledge et al. 1990). Our reads from either side of the antisense hairpin paired to each other with 2-nt 3′ overhangs, indicating that the antisense transcript was processed by the miRNA biogenesis machinery and likely produced a miRNA (miR-iab4as) that enters the silencing complex.

MicroRNA expression patterns

The collection of reads from a variety of developmental stages and anatomical contexts permitted analyses of miRNA expression profiles and overall miRNA expression (Fig. 6A). MicroRNAs were clustered based on their relative expression across all 10 libraries, with the expression values for a particular miRNA in a particular library set to the number of reads corresponding to that miRNA in the given library divided by the total number of reads matching miRNA hairpins in that library. For previously annotated miRNAs, this normalization scheme generated expression profiles similar to those observed previously using 2S rRNA-normalized Northern blots (Fig. 6B) (Aravin et al. 2003), whereas other schemes, such as normalizing to the total reads from each library or to the number of sequenced ribosomal RNA fragments,

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

Table 1. Newly identified miRNAs in *D. melanogaster*

		Reads				Conserved in						Other family members		
miRNA	Sequence	miRNA	miRNA*	Clustered	Intronic	*dsi*	*dya*	*dan*	*dps*	*dmo*	*dvi*	*dme*	*cel*	vert.
miR-137	UAUUGCUUGAGAAUACACGUAG	48	7			Y	Y	Y	Y	Y	Y			miR-137
miR-190	AGAUAUGUUUGAUAUUCUUGGUUG	513	25		Y	Y	Y	Y	Y	Y	Y		miR-50	miR-190
miR-193	UACUGGCCUACUAAGUCCCAAC	755	44			Y	Y	Y	Y	Y	Y		miR-240	miR-193
miR-252	CUAAGUACUAGUGCCGCAGGAG	7271	145		Y	Y	Y	Y	Y	Y	Y	miR-1002	miR-252	
miR-375	UUUGUUCGUUUGGCUUAAGUUA	339	20			Y	Y	Y	Y	Y	Y			miR-375
miR-927	UUUAGAAUUCCUACGCUUUACC	389	14			Y	Y	Y	Y	Y	Y			
miR-929	CUCCCUAACGGAGUCAGAUUG	119	14		Y	Y	Y	Y	Y	Y	Y			
miR-932	UCAAUUCCGUAGUGCAUUGCAG	616	13		Y	Y	Y	Y	Y	Y	Y			
miR-954	UCUGGGUGUUGCGUUGUGUGU	31	10											
miR-955	CAUCGUGCAGAGGUUUGAGUGUC	14				Y	Y	Y		Y	Y			
miR-956	UUUCGAGACCACUCUAAUCCAUU	109	1			Y	Y	Y	Y	Y	Y			
miR-957	UGAAACCGUCCAAAACUGAGGC	137				Y	Y	Y	Y	Y	Y			
miR-958	UGAGAUUCUUCUAUUCUACUUU	1721	110			Y	Y	Y	Y	Y	Y			
miR-959	UUGUCAUCGGGGGUAUUAUGAA	61	18	Y		Y	Y	Y	Y					
miR-960	UGAGUAUUCCAGAUUGCAUAGC	54	14	Y		Y	Y	Y	Y	Y		miR-12		
miR-961	UUUGAUCACCAGUAACUGAGAU	5	4	Y		Y	Y	Y	Y					
miR-962	AUAAGGUAGAGAAAUUGAUGCUGUC	50	9	Y		Y	Y	Y	Y	Y	Y			
miR-963	ACAAGGUAAAUAUCAGGUUGUUUC	92	2	Y		Y	Y	Y		Y	Y			
miR-964	UUAGAAUAGGGGAGCUUAACUU	87	1	Y		Y	Y	Y		Y	Y			
miR-965	UAAGCGUAUAGCUUUUCCCCUU	137	63		Y	Y	Y	Y	Y	Y	Y			
miR-966	UGUGGGUUGUGGGCUGUGUGG	10	2		Y									
miR-967	AGAGAUACCUCUGGAGAAGCG	5	1		Y	Y						miR-977		
miR-968	UAAGUAGUAUCCAUUAAAGGGUUG	84	63	Y		Y	Y	Y	Y		Y			miR-562
miR-969	GAGUUCCACUAAGCAAGUUUU	10		Y		Y	Y	Y	Y	Y	Y			
miR-970	UCAUAAGACACACGCGGCUAU	487	22		Y	Y	Y	Y	Y	Y	Y			
miR-971	UUGGUGUUACUUCUUACAGUGA	52	1				Y	Y	Y	Y	Y			miR-333
miR-972	UGUACAAUACGAAUAUUUAGGC	11		Y		Y								
miR-973	UGGUUGGUGGUUGAACUUCGAUUUU	21	3	Y		Y								
miR-974	AAGCGAGCAAAGAAGUAGUAUU	4	1	Y		Y		Y			Y			
miR-975	UAAACACUUCCUACAUCCUGUAU	66	2	Y		Y	Y	Y						
miR-976	UUGGAUUAGUUAUCAUCAAUGC	31	1	Y		Y	Y	Y		Y	Y			
miR-977	UGAGAUAUUCACGUUGUCUAA	251	8	Y		Y	Y	Y		Y	Y	miR-967		
miR-978	UGUCCAGUGCCGUAAAUUGCAG	51	6	Y		Y	Y	Y						miR-198
miR-979	UUCUUCCCGAACUCAGGCUAA	1	1	Y										
miR-980	UAGCUGCCUUGUGAAGGGCUUA	197	13			Y	Y	Y	Y	Y	Y			miR-22
miR-981	UUCGUUGUCGACGAAACCUGCA	1744	3			Y	Y	Y	Y	Y	Y		miR-76	
miR-982	UCCUGGACAAAUAUGAAGUAAAU	29	3	Y										
miR-983-1	AUAAUACGUUUCGAACUAAUGA	29	39	Y										miR-655
miR-983-2	AUAAUACGUUUCGAACUAAUGA	29	39	Y										miR-655
miR-984	UGAGGUAAAUACGGUUGGAAUUU	173	6	Y								let-7	let-7	let-7
miR-985	CAAAUGUUCCAAUGGUCGGGCA	14	3			Y								
miR-986	UCUCGAAUAGCGUUGUGACUGA	41	1		Y	Y	Y	Y	Y	Y	Y			
miR-987	UAAAGUAAAUAGUCUGGAUUGAUG	875	4			Y	Y	Y	Y	Y	Y			miR-559
miR-988	CCCCUUGUUGCAAACCUCACGC	1908	46		Y	Y	Y	Y	Y	Y	Y			
miR-989	UGUGAUGUGACGUAGUGGAAC	196	5			Y	Y	Y	Y	Y	Y			
miR-990	AUUCACCGUUCUGAGUUGGCC	13	1		Y	Y	Y							
miR-991	UUAAAGUUGUAGUUUGGAAAGU	28		Y		Y								
miR-992	AGUACACGUUUCUGGUACUAAG	148	2	Y			Y							
miR-993	GAAGCUCGUCUCUACAGGUAUCU	287	7			Y	Y	Y	Y	Y	Y		miR-231	
miR-994	CUAAGGAAAUAGUAGCCGUGAU	233	14	Y		Y	Y	Y	Y	Y	Y			
miR-995	UAGCACCACAUGAUUCGGCUU	1326	88		Y	Y	Y	Y	Y	Y	Y	miR-285	miR-49	miR-29
miR-996	UGACUAGAUUUCAUGCUCGUCU	4509	322	Y		Y	Y	Y	Y	Y	Y	miR-279	miR-44	
miR-997	CCCAAACUCGAAGGAGUUUCA	10	2			Y								
miR-998	UAGCACCAUGAGAUUCAGCUC	519	190	Y	Y	Y	Y	Y	Y	Y	Y	miR-285	miR-49	miR-29
miR-999	UGUUAACUGUAAGACUGUGUCU	420	16		Y	Y	Y	Y	Y	Y	Y			
miR-1000	AUAUUGUCCUGUCACAGCAGU	331	31			Y	Y	Y	Y	Y	Y			
miR-1001	UGGGUAAACUCCCAAGGAUCA	35	2		Y	Y								miR-555
miR-1002	UUAAGUAGUGGAUACAAAGGGCGA	73	55	Y		Y	Y	Y	Y		Y	miR-252	miR-252	

did not generate profiles matching the published Northern results (data not shown).

Most miRNAs were observed across several libraries. However, several large sets of miRNAs exhibited strong preference for expression in a single context. The 33 miRNAs that exhibited the narrowest ranges of expression (>70% of their library-normalized reads deriving from a single library) were prevalent in the imaginal discs, adult heads, and to a lesser extent, adult bodies and late embryos (61%, 24%, 12%, and 3% of narrowly expressed miRNAs, respectively), and most were first sequenced in this study (88% of narrowly expressed miRNAs; Fig. 6A).

The normalization of read counts across libraries also permitted an approximate but informative assessment of relative overall expression (Fig. 6A). As reported in vertebrates, worms, and plants (Bartel 2004; Rajagopalan et al. 2006; Ruby et al. 2006), miRNA abundance correlated strongly with the extent of

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

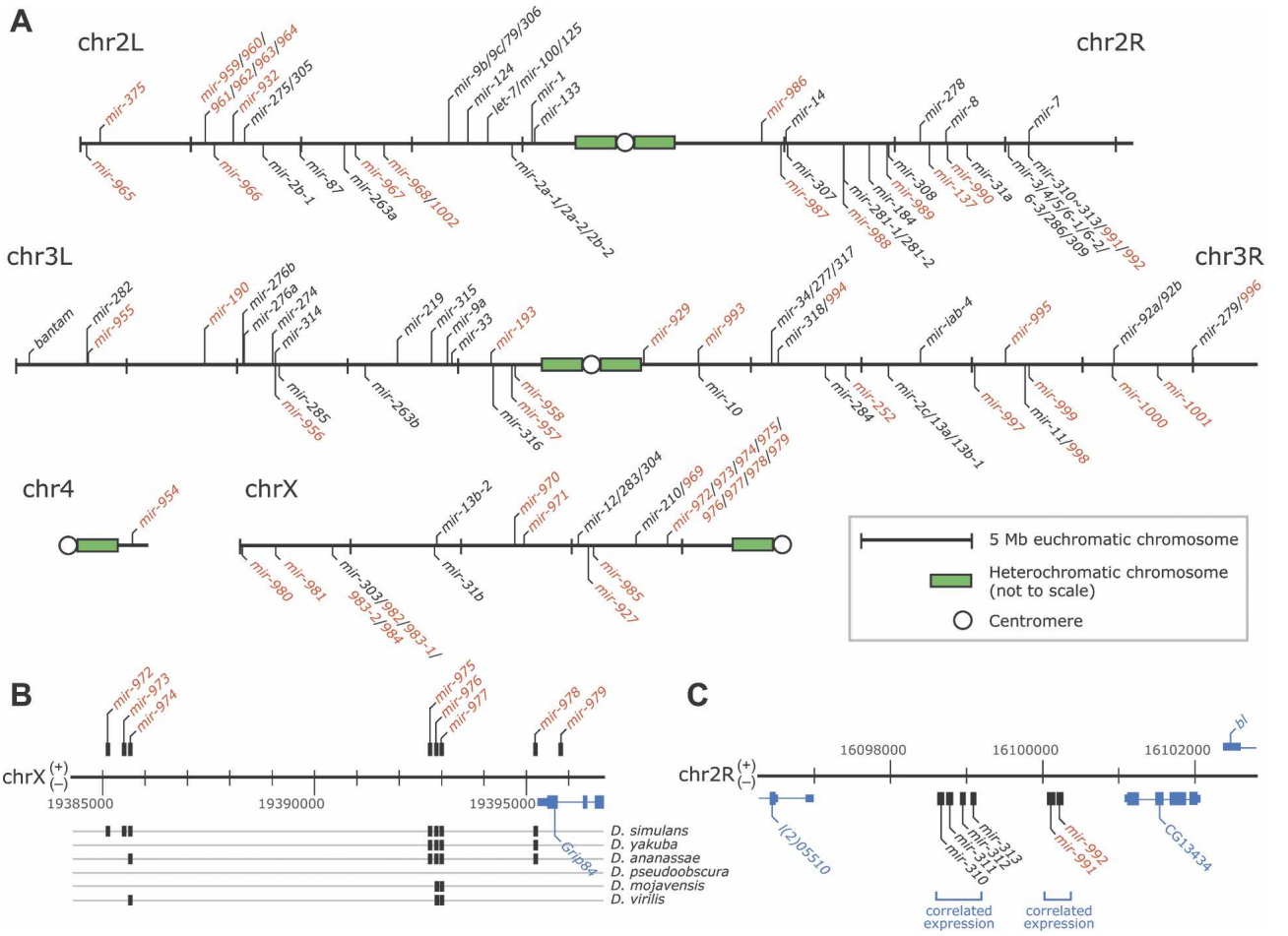


Figure 5. Genomic landscape of miRNA genes. (*A*) The distribution of miRNA genes and clusters across the *D. melanogaster* genome, with newly identified miRNAs indicated (red). Euchromatic portions of the genome are drawn to scale, with (+) strand annotations marked *above* each chromosome and (−) strand annotations marked *below*. MicroRNA gene clusters, listed together (with gene numbers separated by slashes), were each defined as series of miRNA loci on the same strand of a given chromosome with no intervening gaps >10 kb. (*B*) Genomic arrangement and conservation of members of the *mir-972~979* cluster. Detection of an ortholog in the specified species is indicated (black box). (*C*) Genomic arrangement of the *mir-310* cluster. Expression profiles among the constituent miRNAs of each labeled subcluster indicate that the two subclusters were expressed independently (Fig. 6E).

conservation, with those miRNAs found only within the subgenus *Sophophora* expressed significantly less than those conserved beyond that clade (Fig. 6C, Wilcoxon rank-sum test, $P < 2.7 \times 10^{-9}$). Notably, the more highly conserved miRNAs were also observed more evenly across the 10 libraries examined (Wilcoxon rank-sum test, $P < 8.5 \times 10^{-7}$; Fig. 6D).

As observed previously in worms and mammals (Lau et al. 2001; Sempere et al. 2004; Baskerville and Bartel 2005), miRNAs that were clustered in the *Drosophila* genome usually had similar expression profiles (Fig. 6E). The correlation of miRNA expression patterns diminished as the distance separating miRNAs surpassed 10,000 nt. Proximally located miRNAs are thought to derive generally from common primary transcripts (Lagos-Quintana et al. 2001; Lau et al. 2001). The *mir-991/992* and *mir-310~313* clusters, separated from each other by only 1.0 kb, provided a counterexample (Fig. 5C). Although these two clusters each exhibited internally consistent expression patterns, there was little correlation of expression between the two clusters (Fig. 6E), implying that the *mir-991/992* and *mir-310~313* clusters derived from independent transcripts. A more intriguing example was provided by *mir-283*, *mir-12*, and *mir-304*, all three of which map within a single intron. The expression patterns of *mir-12* and *mir-304* correlated very closely with each other (Pearson correlation coefficient = 0.94), but neither correlated well with that of *mir-283* (correlation coefficients of 0.16 and −0.05, respectively), which resided only 1.0 kb upstream of *mir-304* and 1.5 kb upstream of *mir-12*.

MicroRNA targets

In order to gain insight into the functional consequences of the known *D. melanogaster* miRNAs, including those whose annotations were established or modified here, we predicted their targets using comparative genomics of the sequenced genomes of the *Drosophila* genus. As done previously, sites were identified in annotated *D. melanogaster* 3′ UTRs that matched the seed region of each miRNA. Two types of 7-mer sites were sought: the perfect Watson-Crick match to miRNA nucleotides 2–8 (Lewis et al. 2003; Brennecke et al. 2005; Krek et al. 2005; Lewis et al. 2005) and the perfect Watson-Crick match to miRNA nucleotides 2–7,

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

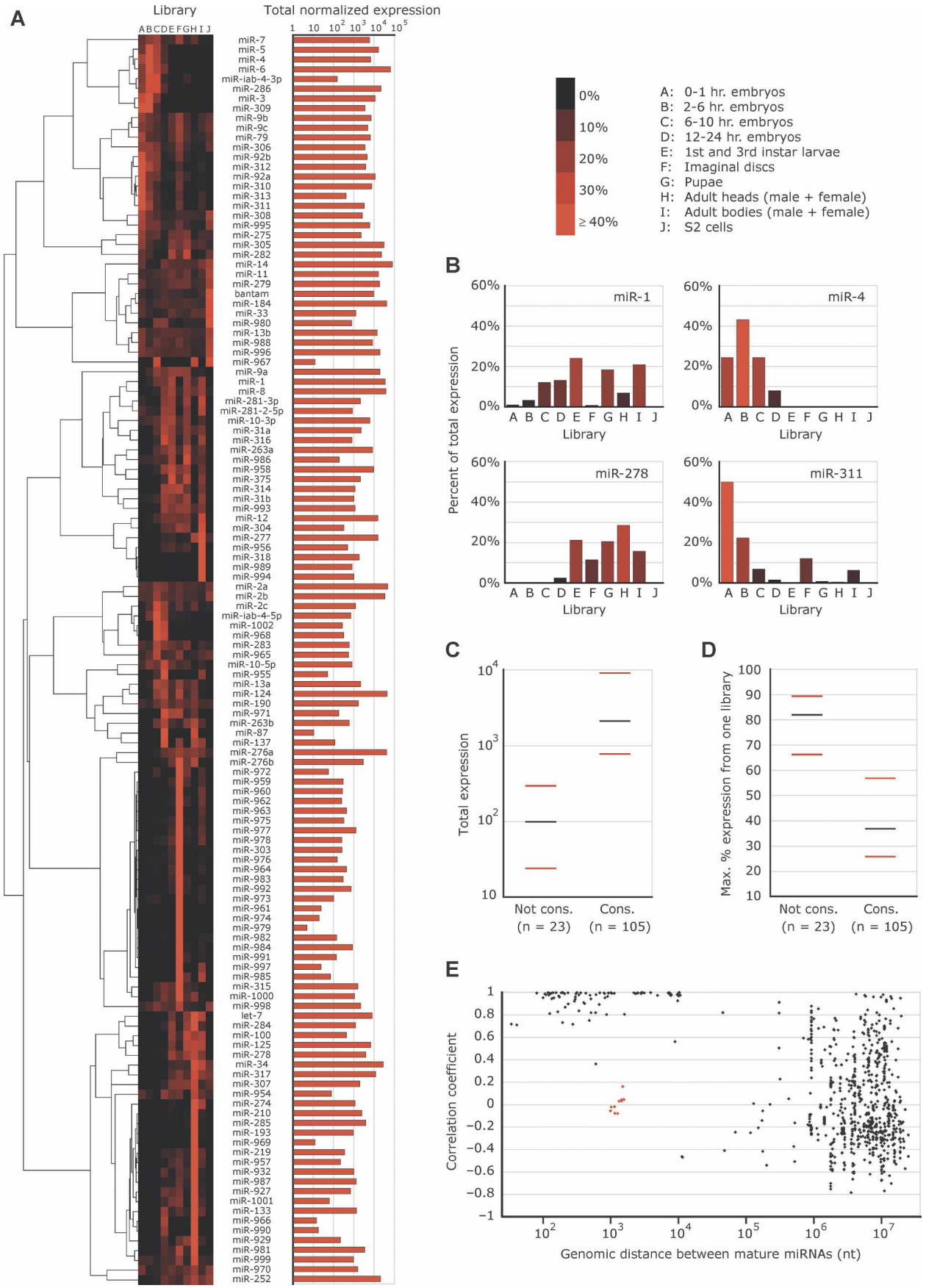


Figure 6. (Legend on next page)

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

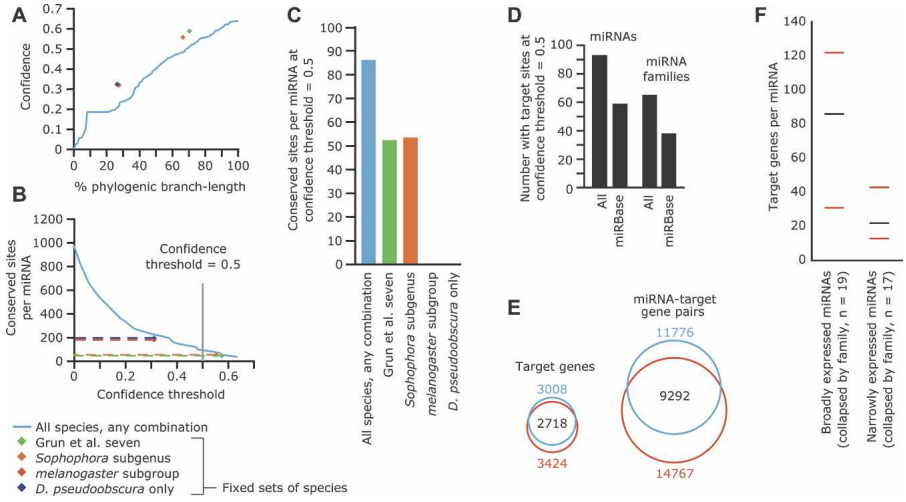


Figure 7. MicroRNA target predictions. (*A*) Confidence of miRNA target prediction versus phylogenic branch length over which sites were conserved in the *Drosophila* genus. Confidence increased with branch length within 12 *Drosophila* species (blue line). Confidence versus branch length values for the following fixed sets of species, strictly requiring conservation in every species, are shown as dots of the indicated colors. (Green) Seven species used by Grun et al. (2005) (*D. melanogaster*, *D. erecta*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura*, *D. mojavensis*, *D. virilis*); (orange) members of the *Sophophora* subgenus (*D. melanogaster*, *D. sechellia*, *D. simulans*, *D. erecta*, *D. yakuba*, *D. ananassae*, *D. persimilis*, *D. pseudoobscura*, *D. willistoni*); (red) members of the *melanogaster* subgroup (*D. melanogaster*, *D. sechellia*, *D. simulans*, *D. erecta*, *D. yakuba*, *D. ananassae*); (purple) *D. melanogaster* and *D. pseudoobscura* only (Enright et al. 2003; Stark et al. 2003). (*B*) Sensitivity of target prediction, shown as the average number of sites per conserved miRNA, versus confidence threshold; colored as in *A*. Note that strict conservation requirements cannot accommodate reduced confidence thresholds, as illustrated by dashed lines. (*C*) Average number of retained target sites per miRNA for each analysis depicted in *A* and *B* at a confidence threshold of 0.5, colored as in *A*. (*D*) The number of miRNAs and miRNA families with targets above a confidence threshold of 0.5. Numbers for miRNAs from miRBase v8.1 (Griffiths-Jones 2004) are compared to those for our expanded/corrected set of miRNA annotations. (*E*) Change to the scope of the predicted miRNA–target network (*left*) and set of genes predicted to be targeted by miRNAs (*right*) as a result of miRNA annotation additions and changes. Target-miRNA pairs and target genes identified based on miRBase v8.1 annotations (Griffiths-Jones 2004) are in blue; those based on the expanded/corrected set of miRNA annotations provided by the present study are in red. (*F*) Specifically expressed miRNAs had fewer predicted targets than did broadly expressed miRNAs. Sets of the most broadly and narrowly expressed miRNAs were collapsed into families based on 6-nt seeds, including only miRNAs conserved beyond the *Sophophora* subgenus. The number of predicted targets for each family was set to the maximum number of predicted targets of any family member. The median (black bars) and 25th and 75th percentiles (red bars) of the number of targets per miRNA family are indicated for each set.

supplemented with an adenosine opposite miRNA position 1 (Lewis et al. 2005).

Conservation of 7-mer sites was assessed using a multigenome alignment of 12 *Drosophila* species (Adams et al. 2000; Richards et al. 2005; *Drosophila* 12 Genomes Consortium 2007; Stark et al. 2007c). The phylogenetic distribution of each seed-match motif was used to calculate the total branch length score (BLS), a measure of evolutionary distance across which the motif was conserved (Kheradpour et al. 2007). Requiring perfect conservation across all of the available species maximized confidence in predicted targets, defined as the fraction of sites that were conserved above chance expectation, but also substantially reduced sensitivity (Fig. 7A–C). This trend extended to arbitrary subsets of the currently available species, including subsets that have been used for other prediction efforts in flies (Enright et al. 2003; Stark et al. 2003; Brennecke et al. 2005; Grun et al. 2005). Such loss of sensitivity is partly attributed to artifacts in sequencing coverage, assembly, or alignment, whose impacts on predictions increase with the number of genomes considered (Grun et al. 2005). Discarding the traditional requirement for perfect conservation within a species set and replacing it with a BLS cutoff enabled confident predictions to be reported in spite of the absence of the motif in particular genomes (Kheradpour et al. 2007). The confidence of miRNA target predictions increased with the total branch length and approached a maximum, averaged over all conserved miRNAs, of 0.64 (Fig. 7A), corresponding to a signal-to-noise ratio of 2.7:1. These improved predictions for the expanded and revised set of miRNAs are available at http://targetscan.org.

For comparison, we used the same procedure to predict targets for the *D. melanogaster* miRNAs as annotated in miRBase v8.1 (Griffiths-Jones 2004). By both increasing the number of annotated conserved miRNAs and correcting the identities of previously annotated miRNAs, our study increased the numbers of both miRNAs and miRNA families with significantly conserved targets (confidence ≥ 0.5) by 1.6-fold (Fig. 7D). While 9292 miRNA–target gene pairs were unaffected by the miRNA annotation additions and changes, 2484

Figure 6. Expression of *D. melanogaster* miRNAs. (*A*) The expression profiles of the *D. melanogaster* miRNAs across the 10 libraries (*left*) and total level of expression (*right*). For each library, miRNA reads are normalized to the total reads deriving from miRNA hairpins in that library. Increasing red color intensity indicates an increasing percentage of normalized reads deriving from that library. Read counts and normalized counts for each miRNA in each library are provided (Supplemental Tables S3 and S4). The summed normalized expressions across all 10 libraries are shown on the *right*; units are the number of miRNA reads per 100,000 total miRNA hairpin reads per library. The tree and image on the *left* were generated using the publicly available software packages Cluster (Eisen et al. 1998) and MapleTree (L. Simirenko, UC Berkeley). (*B*) The expression profiles following normalization of four miRNAs whose profiles can be compared to those determined by stage-specific Northern blot (Aravin et al. 2003). (*C*) The relationship between miRNA conservation and magnitude of total expression. MicroRNAs were separated into two groups based on whether they were conserved (Cons.) or not conserved (Not cons.) beyond the subgenus *Sophophora*. (Black bars) The median expression for each category; (red bars) the 25th and 75th percentiles. Total expression is defined as in *A*. (*D*) The relationship between conservation and breadth of expression, portrayed as in *C*. The *Y*-axis indicates the maximum percentage of expression for a given miRNA derived from a single library. (*E*) The relationship between the genomic distances separating miRNAs and the correlation of their expressions. Each point represents a pair of miRNAs from *A*, including all pairs from the same strand of the same chromosome, but excluding those that can be attributed to multiple genomic loci. The *X*-axis indicates the distance between the mature miRNAs in nucleotides. The *Y*-axis indicates the Pearson correlation coefficient between the normalized expression patterns of the two miRNAs, as displayed in *A*. The red dots represent miR-991 or miR-992 paired with members of the miR-310~313 cluster, and miR-283 paired with miR-12/304. Despite their proximity, these subclusters appeared to be expressed independently.

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

were removed and 5475 were added, thereby changing the predicted network of miRNAs and targets in *D. melanogaster* by 68% [(2484 + 5475)/(9292 + 2484)]. Of the 3424 unique genes predicted to be conserved targets of miRNAs, 706 had conserved sites for only novel miRNAs. Conversely, 290 genes were erroneously predicted to be conserved targets due to miRNA annotations that were adjusted based on our sequencing data (Fig. 7E).

The scope of miRNA targeting varied between those miRNAs broadly expressed across many libraries compared to those expressed more narrowly, independent of the relationship between breadth of expression and conservation discussed above. Of those miRNAs conserved beyond the scope of the *Sophophora* subgenus, the narrowly expressed miRNAs tended to have fewer predicted target genes (Fig. 7F, Wilcoxon rank-sum test, $P < 0.0015$).

Discussion

Hairpin characteristics

The sets of miRNAs initially identified in nematodes, flies, and mammals derive from short hairpins, whereas many of those identified in plants derive from longer precursors (Bartel 2004). Three somewhat longer exceptions have been noted (*Drosophila mir-31b* [Aravin et al. 2003], *C. elegans mir-229* [Ambros et al. 2003; Lim et al. 2003b], and *Caenorhabditis briggsae mir-72* [Ohler et al. 2004]), but the prevalence of short hairpin precursors has seemed to justify limiting the length of the sequenced folded during the initial steps of many prediction protocols (Grad et al. 2003; Lai et al. 2003; Lim et al. 2003b), including the approach described here. Several protocols even explicitly evaluate the distance between the predicted miRNA and miRNA* as a characteristic feature of miRNA hairpins (Bentwich et al. 2005), again including the approach described here. Although imposing these constraints likely boosts the specificity of miRNA prediction, our sequencing results indicated that this comes at the cost of missing some miRNAs with unusually long hairpins, particularly in *Drosophila*, where we found three hairpins (*mir-956*, *mir-989*, and *mir-997*) with at least 80 nt separating the miRNA and miRNA* strands (Fig. 4C,D; Supplemental Table S2).

Our observation that metazoan miRNA precursors can be much longer than previously recognized confirmed that minimal sequence or structural requirements are imposed on the loops of miRNA hairpins (Lai et al. 2003; Berezikov et al. 2005; Han et al. 2006) but raised the question of why long miRNA hairpins are not more frequent in animals. A large, open loop can lead to Drosha processing on the incorrect end of the hairpin by mimicking the single-stranded RNA normally present at the base (Han et al. 2006). This opportunity for dead-end side reactions implies selective pressure for the maintenance of a tight loop. Consistent with this idea, *mir-956*, *mir-989*, and *mir-997* each exhibited extensive secondary structure in the segment connecting the miRNA and miRNA* (Supplemental Table S2). Deletions can tighten a loop even if they disrupt secondary structure, making them more tolerable than insertions, which must be accompanied by compensatory changes in order to maintain the tightness of the loop. Thus, miRNA hairpins might be expected to shorten rather than lengthen over evolutionary time. Another possibility is that shorter pre-miRNAs might be more suitable substrates for downstream events such as nuclear export, and longer pre-miRNAs might only rarely bypass these constraints.

The evolutionary origins of novel miRNA genes

High-throughput sequencing of miRNAs in *D. melanogaster* provided insight into the origins of novel miRNA genes and how their origins might differ from those of protein-coding genes. Generally, the first step in the emergence of a new gene is the duplication of all or part of an ancestral gene (Ohno 1970). A redundant copy of a gene eventually faces one of three fates: the accumulation of mutations that render the copy functionless (nonfunctionalization), the accumulation of mutations that confer a novel and independently selectable function (neofunctionalization), or, in cases in which the ancestral gene had multiple functions, the accumulation of complementary degenerative mutations in both gene copies that specialize each to perform one of the parental functions (subfunctionalization) (Force et al. 1999). Protein-coding genes provide some examples consistent with subfunctionalization and others consistent with neofunctionalization. We observed examples of miRNA genes that were consistent with each of these models and also examples that appeared to be the products of de novo emergence.

The process of subfunctionalization first requires that an ancestral gene acquire multiple functions. Mechanisms that could impart multiple functions on a miRNA locus include imprecise processing that generates alternative miRNA 5′ termini, like that observed for *mir-210*, and transcription from both orientations with subsequent processing of both pri-miRNAs, as observed for *mir-iab-4*. But perhaps the most available mechanism for acquiring new functions is bringing the miRNA* into service. MicroRNA* species are initially generated at an obligate 1:1 stoichiometric ratio compared to mature miRNAs and to varying degrees are incorporated into RISC just like their complementary counterparts, albeit at a generally lower frequency. They thereby represent an easily accessible substrate for the evolution of novel functionality (Lai et al. 2004) (Fig. 8A). Examples of genes in

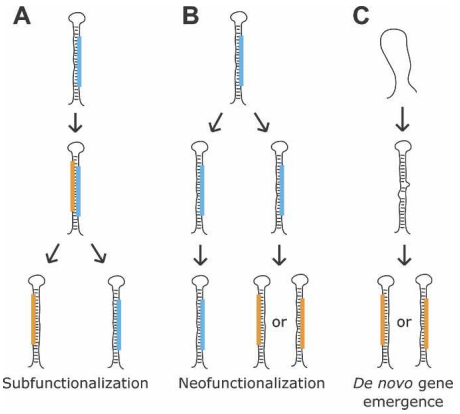


Figure 8. Three models for the genesis of miRNA genes. (Blue bars) Ancestral miRNAs; (orange bars) novel miRNAs. (*A*) An example of subfunctionalization: a miRNA* acquires function; following gene duplication, one daughter copy maintains the function of the original miRNA, while the other maintains the function of the former miRNA*. Another example of subfunctionalization begins with heterologous 5′ processing. (*B*) Neofunctionalization: a miRNA gene duplicates; one daughter copy maintains the function of the original miRNA, while the other accumulates mutations that confer novel functionality to either the former miRNA or miRNA*. (*C*) De novo gene emergence: an unselected portion of a pre-existing transcript, such as an intron or part of a pri-miRNA, acquires the capacity to fold into a hairpin that can be processed into a mature miRNA. That product is selectively maintained because of the fortuitous benefit of gene silencing guided by its seed.

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

which conservation data, read abundance, or experimental data suggested that both strands could be functional included *mir-10*, *mir-iab-4*, and *mir-313*. The miR-313* seed was only conserved within the *melanogaster* complex, which diverged from the *yakuba* and *erecta* complexes of the *melanogaster* subgroup 6–15 million yr ago (Lachaise et al. 1988). Although the *melanogaster* complex species were insufficiently diverged to conclude selective maintenance of the seed, the high abundance of the miRNA* implied the capacity to affect the expressions of target messages.

If a locus with multiple miRNA products, such as one of those listed above, were to duplicate, selective pressure would diminish for each daughter copy to continue producing all miRNA species, and eventually each daughter copy might retain the ability to produce just one of the functional miRNA products. This process might be in progress for the vertebrate *mir-10* paralogs; miR-10-3p is maintained in all of the *mir-10a*, but not the *mir-10b*, hairpins (Fig. 3).

Neofunctionalization requires that gene copies find novel selectable functions after duplication and prior to loss of expressional competence. Because 5′ heterogeneity was very rare among the known miRNAs, the divergent processing of both the *mir-2* paralogs and the tandemly duplicated *mir-281* paralogs likely emerged after duplication and thus represent attractive candidates for neofunctional origins. In the case of *mir-281*, divergent processing not only shifted the seed of the ancestral miRNA, thereby potentially altering its target specificity, but also changed the miRNA:miRNA* pairing asymmetry, which significantly enhanced expression of the presumptive ancestral miRNA* species (Supplemental Text). We speculate that future drift of the two loci, with one increasingly specialized to produce a mature miRNA from the former star strand, would result in two genes with common ancestry yet no recognizable sequence identity.

Many pairs of apparently unrelated modern miRNA hairpins might have arisen from common ancestors through the processes of subfunctionalization and neofunctionalization. However, common descent is not always easy to identify, and the two mechanisms can be difficult to distinguish from each other even when descent from a common ancestor appears evident. For example, the *mir-4*, the three *mir-9*, and the *mir-79* loci appeared to have all derived from a common ancestor, as did the *mir-5* and the three *mir-6* loci (Lai et al. 2004). However, in each of these cases, the structure of the gene family tree was ambiguous.

Novel protein-coding genes derive from duplication and divergence of an ancestral gene, and the active sites of their products generally evolve within the context of the ancestral tertiary structures. For protein-coding genes, the requirements of transcription, translation, protein folding, and protein function impose a myriad of informational constraints, making the completely de novo evolution of novel protein-coding genes highly improbable and therefore exceedingly rare. Canonical miRNAs, in contrast, have much more limited requirements. They must be transcribed, and the subsequent transcript must be capable of folding into a secondary structure that is competent for Drosha/Dicer processing (Han et al. 2006). The secondary structure requirements imposed on miRNA hairpin precursors are not excessively stringent, with a wide variety of bulge distributions, hairpin lengths, and loop sizes tolerated among the miRNAs of any given organism. The minimal informational requirements for miRNA-target interactions make it likely that any expressed miRNA will have a physiological consequence, enabling a young miRNA to find a selectively advantageous physiological role. Perhaps the most difficult obstacle for the emergence of new functional miRNA genes would be the plethora of coexpressed messages with fortuitous sites in their 3′ UTRs whose expression would be dampened as the expression of the emergent miRNA became consequential.

The genomic contexts of many miRNAs, including many of the youngest (most narrowly conserved) miRNAs described here, suggested that the pliancy of the miRNA processing machinery facilitates the emergence of new miRNA genes. Most derived from introns or miRNA clusters. In either of those contexts, a miRNA gene can emerge from otherwise unconstrained portions of pre-existing transcripts with little or no effect on the other products of those transcripts, thereby circumventing the otherwise required de novo acquisition of transcriptional competence. The varying extents of conservation observed within the *mir-972~979* cluster, which was preferentially expressed in the imaginal discs, reflected a variety of ages for the miRNA genes of that transcript (Fig. 5B). The oldest hairpins, *mir-974/976/977*, spanned the *Drosophila* genus, indicating that they are >30 million yr old (Beverley and Wilson 1984). In contrast, the other hairpins of the same cluster appeared to have emerged after the *D. melanogaster/simulans* split, indicating that they are <2.5 million yr old (Lachaise et al. 1988). The presence of hairpins with intermediate scopes of conservation, limited to the *melanogaster* species group (*mir-975/978*) or complex (*mir-972/973*), implied a model of functional miRNA genes emerging and presumably disappearing with some temporal regularity from the context of this transcript.

Two other miRNA genes that appeared to have emerged after the *D. melanogaster/simulans* split deserved special mention. The first, *mir-984*, expressed a miRNA whose 6-nt seed matched that of the *let-7* miRNA and thus could repress many of the same target mRNAs (Lewis et al. 2005). Despite their seed identity, *mir-984* and *let-7* shared little sequence identity and had clearly distinct expression profiles (Fig. 6A; Supplemental Table S2), suggesting that *mir-984* emerged de novo rather than as a paralog of *let-7*. The second gene was *mir-954*, notable for being the first miRNA gene to be identified on the dot chromosome of *D. melanogaster*, Chromosome 4 (Fig. 5A). Portions of the euchromatic chromosome 4 exhibit some heterochromatin-like properties such as variegated expression of inserted reporter constructs, and two such sites of variegated expression flank the *mir-954* locus (Riddle and Elgin 2006).

The scope of miRNA genes and targets in flies

Our current tally of confidently identified miRNA genes in *D. melanogaster* stands at 148. These include 74 of the previously annotated genes, 59 novel genes reported in this study, and another 15 novel genes (*mir-1003~1017*) whose transcripts bypass Drosha processing (Ruby et al. 2007). Forty-five of our top 47 computational predictions and 75 of our 100 predictions were either previously known or newly validated miRNAs (Fig. 1; Supplemental Table S1). Independent predictions from a parallel effort had even greater specificity (Stark 2007), which might be attributed to the use of different training sets; the set used here was smaller and included miRNAs annotated in miRBase but whose authenticity is now in doubt (miR-280, and miR-289), as well as other miRNAs whose 5′ termini appear to have been incorrectly annotated (miR-2a-2, miR-33, miR-274, miR-284, and miR-303). The high specificity of both approaches implied that very few highly conserved miRNAs remain to be discovered in

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

flies. However, most of the miRNAs identified by our sequencing were missed by both prediction methods because these miRNAs were insufficiently conserved. Because the less broadly conserved fly miRNAs tended to be expressed at lower levels, it was impossible to use the cloning results to estimate a lower limit on the overall specificity of the computational gene predictions and thereby derive a meaningful upper limit on the number of miRNAs remaining to be identified in flies. Reliable upper limits on miRNA gene numbers face similar constraints in mammals, worms, and plants (Bartel 2004; Rajagopalan et al. 2006; Ruby et al. 2006).

The implication that there are many more miRNAs to be discovered in flies but almost none of them will be widely conserved, relates to the observed correlation between miRNA conservation and breadth of expression (Fig. 6D), which was likely understated here. All of the libraries from which small RNAs were sampled, with the exception of the S2 library, comprised a conglomerate of cell types, and many of the librariessurveyed thick slices of developmental time. The stronger direct correlation between conservation and total magnitude of expression that was observed here and in other systems may imply that the scarce miRNAs were actually expressed in even narrower contexts that contributed only a small fraction to their encompassing libraries. Thus, most of the remaining undiscovered miRNAs will inhabit niches of increasingly restricted physiological and evolutionary scopes.

Following that conclusion, another observation becomes relevant: given a consistent scope of conservation, the number of predicted targets decreased with more narrow breadth of miRNA expression (Fig. 7F). The regulatory reach of miRNAs, as indicated by the abundance of genes with conserved miRNA target sites, is likely quite vast. However, the as-yet-undiscovered miRNAs appear to have remained hidden thanks to the narrow scopes of their expression. Consequentially, the set of consequential miRNA targets will likely grow at a diminishing rate relative to the catalog of fly miRNAs, and our overall picture of the biological reach of miRNAs will likely not change substantially. This being said, the biology of some of the as-yet-undiscovered miRNAs is still likely to be quite interesting. As illustrated by *lsy-6* in *C. elegans* (Johnston and Hobert 2003), a single miRNA expressed in only a few cells and acting on a limited set of targets can make quite a difference to the animal.

Methods

MicroRNA gene prediction

MicroRNA gene prediction is described in the Supplemental Text.

Library construction and sequencing

Total RNA was extracted from Canton S *D. melanogaster* and from S2 cells using TRIzol. Embryos were collected using a population cage whose food had been changed regularly to minimize egg withholding. Staged collections of 0–1 h, 2–6 h, 6–10 h, and 12–24 h embryos were obtained by culturing at 25°C. First-instar larvae were obtained by aging a 0–12-h embryo collection on a plate for 24 h. Wandering third-instar larvae were collected from vial cultures and rinsed several times in PBS to remove excess food. Total imaginal discs, brains, and salivary glands were isolated from wandering instar larvae to make a pooled "disc" preparation. Separate collections of 0–1 d, 1–2 d, and 2–4 d pupae were prepared and pooled to make a pupal library. Equal numbers of 1- to 5-d-old adult female and male flies were frozen at −80°C, vortexed, and sieved onto dry ice blocks to obtain adult head and body fractions. S2 cells were grown in Schneider's medium and rinsed several times in PBS prior to extraction. A cDNA library was generated from each RNA sample as described (Lau et al. 2001) and was prepared for high-throughput pyrophosphate sequencing (Margulies et al. 2005) as described for run 4 of Ruby et al. (2006). Each library underwent a single sequencing run except for the 2–6-h embryo library, which underwent two sequencing runs. A total of 2,514,465 reads were generated. The processing of sequencing data is described in the Supplemental Text.

Expression analysis

Expression analysis is described in the Supplemental Text.

Target prediction and analysis

For each miRNA, we defined two 7-mer motifs that corresponded to the two types of 7-mers matching the seed region (the Watson-Crick match to miRNA nucleotides 2–8 and the Watson-Crick match to miRNA nucleotides 2–7 followed by an A). All occurrences of the two motifs were identified within annotated *D. melanogaster* 3′ UTRs from FlyBase Release 4.3 (Crosby et al. 2007), and the conservation of each of these sites was assessed using whole-genome alignments of *D. melanogaster* and 11 additional *Drosophila* species (Schwartz and Pachter 2007). To allow for alignment errors or gaps, sites were scored as conserved if they fell within 50 nt of the aligned site in each informant species. For each site, evolutionary conservation was evaluated as the total branch length corresponding to its species distribution (BLS) as described (Kheradpour et al. 2007). A site was considered conserved if its BLS met the specified cutoff. To prevent double-counting of 8-mer sites that contained both of the two 7-mers, target-prediction results reported nonoverlapping sites, obtained by first removing sites that did not meet the specified conservation cutoff and then removing overlapping sites, such that the maximum possible number of nonoverlapping sites was retained.

To estimate the conservation expected by chance, we repeated the target-prediction analyses using control motifs and compared the conservation frequencies of their sites with the conservation frequencies of sites obtained for the authentic miRNA. For each miRNA, nine controls were generated for each of the two motifs. For the motif matching miRNA nucleotides 2–8, each control had equal nucleotide composition and a similar number of matches in *D. melanogaster* 3′ UTRs (deviation <15%) as the authentic motif. The last six nucleotides of these controls were each extended by an A to obtain the nine controls for the other motif. Signal-to-noise ratios were calculated for each individual miRNA by dividing the frequency of conservation for the authentic sites by the average frequency of the control sites. Signal-to-noise ratios for all miRNAs combined were calculated in the same manner, aggregating all sites for all miRNAs under consideration and their controls. In each case, the number of conserved sites expected by chance was determined by multiplying the total number of sites by the control conservation frequency. Confidence was defined as the number of conserved sites above those expected by chance (i.e., above noise) divided by the total number of conserved sites. Confidence reflected the likelihood of a single conserved site being under selection.

For analysis of expression breadth versus number of predicted targets, miRNAs whose conservation did not extend beyond the *Sophophora* subgenus were not considered. A set of narrowly expressed miRNAs was defined as those with >70% of their library-normalized reads deriving from a single library, and a set

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

of broadly expressed miRNAs as those with no more than 25% of their library-normalized reads deriving from a single library. Each set was collapsed into families based on their 6-nt seeds, resulting in 17 narrowly expressed families and 19 broadly expressed families. The number of predicted targets was determined for each family in each set by requiring targets to be conserved across 70% of the available branch length. In cases in which the number of predicted target genes differed among family members because of differences at microRNA nucleotide 8 (which changes one of the two 7-mer sites), the largest number of predicted target genes for the family was used.

Acknowledgments

We thank Ann Hammonds, Michael Axtell, and Gerald Rubin for assistance and support during library construction; Joseph Rodriguez, Robin Ge, Katherine Gurdziel, George Bell, and Fran Lewitter for constructing the TargetScanFly database of predicted targets (http://targetscan.org); and Ramya Rajagopalan for comments on the manuscript. This work was supported by a grant from the NIH (to D.P.B.). D.P.B is an HHMI Investigator. E.C.L. was supported by grants from the Burroughs Wellcome Foundation, the Leukemia and Lymphoma Society, and the V Foundation for Cancer Research. A.S. was supported in part by the Schering AG/Ernst Schering Foundation and in part by the Human Frontier Science Program Organization (HFSPO).

References

Abbott, A.L., Alvarez-Saavedra, E., Miska, E.A., Lau, N.C., Bartel, D.P., Horvitz, H.R., and Ambros, V. 2005. The *let-7* microRNA family members *mir-48*, *mir-84*, and *mir-241* function together to regulate developmental timing in *Caenorhabditis elegans*. *Dev. Cell* **9:** 403–414.

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Ambros, V., Lee, R.C., Lavanway, A., Williams, P.T., and Jewell, D. 2003. MicroRNAs and other tiny endogenous RNAs in *C. elegans*. *Curr. Biol.* **13:** 807–818.

Aravin, A.A., Lagos-Quintana, M., Yalcin, A., Zavolan, M., Marks, D., Snyder, B., Gaasterland, T., Meyer, J., and Tuschl, T. 2003. The small RNA profile during *Drosophila melanogaster* development. *Dev. Cell* **5:** 337–350.

Bartel, D.P. 2004. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* **116:** 281–297.

Baskerville, S. and Bartel, D.P. 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* **11:** 241–247.

Basyuk, E., Suavet, F., Doglio, A., Bordonne, R., and Bertrand, E. 2003. Human *let-7* stem–loop precursors harbor features of RNase III cleavage products. *Nucleic Acids Res.* **31:** 6593–6597.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., et al. 2005. Identification of hundreds of conserved and nonconserved human microRNAs. *Nat. Genet.* **37:** 766–770.

Berezikov, E., Guryev, V., van de Belt, J., Wienholds, E., Plasterk, R.H., and Cuppen, E. 2005. Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120:** 21–24.

Berezikov, E., Thuemmler, F., van Laake, L.W., Kondova, I., Bontrop, R., Cuppen, E., and Plasterk, R.H. 2006. Diversity of microRNAs in human and chimpanzee brain. *Nat. Genet.* **38:** 1375–1377.

Beverley, S.M. and Wilson, A.C. 1984. Molecular evolution in *Drosophila* and the higher Diptera II. A time scale for fly evolution. *J. Mol. Evol.* **21:** 1–13.

Brennecke, J., Stark, A., Russell, R.B., and Cohen, S.M. 2005. Principles of microRNA–target recognition. *PLoS Biol.* **3:** e85. doi: 10.1371/journal.pbio.0030085.

Crosby, M.A., Goodman, J.L., Strelets, V.B., Zhang, P., and Gelbart, W.M. 2007. FlyBase: Genomes by the dozen. *Nucleic Acids Res.* **35:** D486–D491. doi: 10.1093/nar/gkl827.

Cumberledge, S., Zaratzian, A., and Sakonju, S. 1990. Characterization of two RNAs transcribed from the *cis*-regulatory region of the *abd-A* domain within the *Drosophila* bithorax complex. *Proc. Natl. Acad. Sci.* **87:** 3259–3263.

Doench, J.G. and Sharp, P.A. 2004. Specificity of microRNA target selection in translational repression. *Genes & Dev.* **18:** 504–511.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* (in press) doi: 10.1038/nature06341.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* **95:** 14863–14868.

Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D.S. 2003. MicroRNA targets in *Drosophila*. *Genome Biol.* **5:** R1. doi: 10.1186/gb-2003-5-1-r1.

Fahlgren, N., Howell, M.D., Kasschau, K.D., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A., Law, T.F., Grant, S.R., Dangl, J.L., et al. 2007. High-throughput sequencing of *Arabidopsis* microRNAs: Evidence for frequent birth and death of miRNA genes. *PLoS ONE* **2:** e219. doi: 10.1371/journal.pone.0000219.

Farh, K.K., Grimson, A., Jan, C., Lewis, B.P., Johnston, W.K., Lim, L.P., Burge, C.B., and Bartel, D.P. 2005. The widespread impact of mammalian microRNAs on mRNA repression and evolution. *Science* **310:** 1817–1821.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., and Postlethwait, J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151:** 1531–1545.

Grad, Y., Aach, J., Hayes, G.D., Reinhart, B.J., Church, G.M., Ruvkun, G., and Kim, J. 2003. Computational and experimental identification of *C. elegans* microRNAs. *Mol. Cell* **11:** 1253–1263.

Griffiths-Jones, S. 2004. The microRNA Registry. *Nucleic Acids Res.* **32:** D109–D111. doi: 10.1093/nar/gkh023.

Grun, D., Wang, Y.L., Langenberger, D., Gunsalus, K.C., and Rajewsky, N. 2005. microRNA target predictions across seven *Drosophila* species and comparison to mammalian targets. *PLoS Comput. Biol.* **1:** e13. doi: 10.1371/journal.pcbi.0010013.

Han, J., Lee, Y., Yeom, K.H., Nam, J.W., Heo, I., Rhee, J.K., Sohn, S.Y., Cho, Y., Zhang, B.T., and Kim, V.N. 2006. Molecular basis for the recognition of primary microRNAs by the Drosha–DGCR8 complex. *Cell* **125:** 887–901.

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M., and Schuster, P. 1994. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* **125:** 167–188.

Johnston, R.J. and Hobert, O. 2003. A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. *Nature* **426:** 845–849.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al. 2003. The UCSC Genome Browser database. *Nucleic Acids Res.* **31:** 51–54.

Kheradpour, P., Stark, A., Roy, S., and Kellis, M. 2007. Reliable prediction of regulator targets using 12 *Drosophila* genomes. *Genome Res.* (this issue).

Khvorova, A., Reynolds, A., and Jayasena, S.D. 2003. Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115:** 209–216.

Krek, A., Grun, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M., et al. 2005. Combinatorial microRNA target predictions. *Nat. Genet.* **37:** 495–500.

Lachaise, D., Cariou, M.L., David, J.R., Lemeunier, F., Tsacas, L., and Ashburner, M. 1988. Historical biogeography of the *Drosophila melanogaster* species subgroup. *Evol. Biol.* **22:** 159–225.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. 2001. Identification of novel genes coding for small expressed RNAs. *Science* **294:** 853–858.

Lai, E.C. 2002. Micro RNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat. Genet.* **30:** 363–364.

Lai, E.C., Tomancak, P., Williams, R.W., and Rubin, G.M. 2003. Computational identification of *Drosophila* microRNA genes. *Genome Biol.* **4:** R42. doi: 10.1186/gb-2003-4-7-r42.

Lai, E.C., Wiel, C., and Rubin, G.M. 2004. Complementary miRNA pairs suggest a regulatory role for miRNA:miRNA duplexes. *RNA* **10:** 171–175.

Lai, E.C., Tam, B., and Rubin, G.M. 2005. Pervasive regulation of *Drosophila* Notch target genes by GY-box-, Brd-box-, and K-box-class microRNAs. *Genes & Dev.* **19:** 1067–1080.

Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294:** 858–862.

Lee, R.C., Feinbaum, R.L., and Ambros, V. 1993. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

complementarity to *lin-14*. *Cell* **75:** 843–854.
Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S., et al. 2003. The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425:** 415–419.
Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P., and Burge, C.B. 2003. Prediction of mammalian microRNA targets. *Cell* **115:** 787–798.
Lewis, B.P., Burge, C.B., and Bartel, D.P. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120:** 15–20.
Li, J., Yang, Z., Yu, B., Liu, J., and Chen, X. 2005. Methylation protects miRNAs and siRNAs from a 3′-end uridylation activity in *Arabidopsis*. *Curr. Biol.* **15:** 1501–1507.
Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B., and Bartel, D.P. 2003a. Vertebrate microRNA genes. *Science* **299:** 1540.
Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B., and Bartel, D.P. 2003b. The microRNAs of *Caenorhabditis elegans*. *Genes & Dev.* **17:** 991–1008.
Lim, L.P., Lau, N.C., Garrett-Engele, P., Grimson, A., Schelter, J.M., Castle, J., Bartel, D.P., Linsley, P.S., and Johnson, J.M. 2005. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433:** 769–773.
Lu, C., Kulkarni, K., Souret, F.F., Muthu Valliappan, R., Tej, S.S., Poethig, R.S., Henderson, I.R., Jacobsen, S.E., Wang, W., Green, P.J., et al. 2006. MicroRNAs and other small RNAs enriched in the *Arabidopsis* RNA-dependent RNA polymerase-2 mutant. *Genome Res.* **16:** 1276–1288.
Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376–380.
Ohler, U., Yekta, S., Lim, L.P., Bartel, D.P., and Burge, C.B. 2004. Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *RNA* **10:** 1309–1322.
Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, Berlin, New York.
Rajagopalan, R., Vaucheret, H., Trejo, J., and Bartel, D.P. 2006. A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. *Genes & Dev.* **20:** 3407–3425.
Richards, S., Liu, Y., Bettencourt, B.R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M.J., Chen, R., Meisel, R.P., et al. 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene, and *cis*-element evolution. *Genome Res.* **15:** 1–18.
Riddle, N.C. and Elgin, S.C. 2006. The dot chromosome of *Drosophila*: Insights into chromatin states and their change over evolutionary time. *Chromosome Res.* **14:** 405–416.
Ruby, J.G., Jan, C., Player, C., Axtell, M.J., Lee, W., Nusbaum, C., Ge, H., and Bartel, D.P. 2006. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* **127:** 1193–1207.
Ruby, J.G., Jan, C.H., and Bartel, D.P. 2007. Intronic microRNA precursors that bypass Drosha processing. *Nature* **448:** 83–86.
Schwartz, A.S. and Pachter, L. 2007. Multiple alignment by sequence annealing. *Bioinformatics* **23:** e24–e29. doi: 10.1093/bioinformatics/btl311.
Schwarz, D.S., Hutvagner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P.D. 2003. Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115:** 199–208.
Sempere, L.F., Freemantle, S., Pitha-Rowe, I., Moss, E., Dmitrovsky, E., and Ambros, V. 2004. Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation. *Genome Biol.* **5:** R13. http://genomebiology.com/2004/5/3/R13.
Stark, A., Brennecke, J., Russell, R.B., and Cohen, S.M. 2003. Identification of *Drosophila* microRNA targets. *PLoS Biol.* **1:** e60. doi: 10.1371/journal.pbio.0000060.
Stark, A., Brennecke, J., Bushati, N., Russell, R.B., and Cohen, S.M. 2005. Animal microRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. *Cell* **123:** 1133–1146.
Stark, A., Kheradpour, P., Parts, L., Brennecke, J., Hodges, E., Hannon, G.J., and Kellis, M. 2007. Systematic discovery and characterization of fly microRNAs using 12 Drosophila genomes. *Genome Res.* (this issue). doi: 10.1101/gr.6593807.
Stark, A., Lin, M.F., Kheradpour, P., Pedersen, J.S., Parts, L., Carlson, J.W., Crosby, M.A., Rasmussen, M.D., Roy, S., Deoras, A.N., et al. 2007c. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* (in press) doi: 10.1038/nature06340.
Wang, H., Ach, R.A., and Curry, B. 2007. Direct and sensitive miRNA profiling from low-input total RNA. *RNA* **13:** 151–159.
Wightman, B., Ha, I., and Ruvkun, G. 1993. Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell* **75:** 855–862.

Received April 11, 2007; accepted in revised form July 30, 2007.

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

Systematic discovery and characterization of fly microRNAs using 12 *Drosophila* genomes

Alexander Stark,[1,2,4,6] Pouya Kheradpour,[2,4] Leopold Parts,[2,5] Julius Brennecke,[3] Emily Hodges,[3] Gregory J. Hannon,[3] and Manolis Kellis[1,2,6]

[1]*Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02141, USA;* [2]*Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA;* [3]*Cold Spring Harbor Laboratory, Watson School of Biological Sciences and Howard Hughes Medical Institute, Cold Spring Harbor, New York 11724, USA*

MicroRNAs (miRNAs) are short regulatory RNAs that inhibit target genes by complementary binding in 3′ untranslated regions (3′ UTRs). They are one of the most abundant classes of regulators, targeting a large fraction of all genes, making their comprehensive study a requirement for understanding regulation and development. Here we use 12 *Drosophila* genomes to define structural and evolutionary signatures of miRNA hairpins, which we use for their de novo discovery. We predict >41 novel miRNA genes, which encompass many unique families, and 28 of which are validated experimentally. We also define signals for the precise start position of mature miRNAs, which suggest corrections of previously known miRNAs, often leading to drastic changes in their predicted target spectrum. We show that miRNA discovery power scales with the number and divergence of species compared, suggesting that such approaches can be successful in human as dozens of mammalian genomes become available. Interestingly, for some miRNAs sense and anti-sense hairpins score highly and mature miRNAs from both strands can indeed be found in vivo. Similarly, miRNAs with weak 5′ end predictions show increased in vivo processing of multiple alternate 5′ ends and have fewer predicted targets. Lastly, we show that several miRNA star sequences score highly and are likely functional. For *mir–10* in particular, both arms show abundant processing, and both show highly conserved target sites in Hox genes, suggesting a possible cooperation of the two arms, and their role as a master Hox regulator.

[Supplemental material is available online at www.genome.org. All data and predictions are available at http://compbio.mit.edu/fly/mirnas/.]

Regulation of gene expression in response to developmental and environmental stimuli is central to animal development. In addition to regulation at the transcriptional level, it is increasingly recognized that an important fraction of regulation occurs post-transcriptionally, and to a large extent by microRNAs (miRNAs) (Lai 2003; Ambros 2004; Bartel 2004; Alvarez-Garcia and Miska 2005; Plasterk 2006; Chen and Rajewsky 2007). These are short RNA genes that direct the inhibition of target messenger-RNA expression via complementary binding sites in the 3′ untranslated region (3′ UTR) (for reviews, see Zamore and Haley 2005; Valencia-Sanchez et al. 2006). miRNAs play an integral part of animal gene regulatory networks. First, they are one of the most abundant classes of regulators, currently estimated to comprise 1%–5% of animal genes (Bartel 2004; Bentwich et al. 2005a; Berezikov et al. 2005). Second, they target a large fraction of all genes, a typical miRNA regulating hundreds of target genes (Brennecke et al. 2005; Krek et al. 2005; Lewis et al. 2005; Xie et al. 2005). Third, as many genes contain target sites for several miRNAs, combinatorial control similar to that known for transcription factors is likely to increase miRNA regulatory versatility (Hobert 2004). Finally, several miRNAs are deeply conserved in the animal kingdom (e.g., *let-7*; Pasquinelli et al. 2000), suggesting an ancient role similar to some transcription factors in the core of developmental regulatory kernels (Davidson and Erwin 2006).

Thus, a comprehensive understanding of all miRNAs and their targets in an animal genome presents a necessary milestone in our understanding of gene regulation in animal genomes. This is especially true, as knowledge of the miRNA sequence alone can allow the identification of the physiologically relevant target genes (e.g., *mir-iab-4* [Stark et al. 2003; Ronshaugen et al. 2005]; *bantam* [Brennecke et al. 2003]; *mir-9a* [Li et al. 2006]). Although the first miRNA-target pair was identified genetically more than a decade ago (Lee et al. 1993; Wightman et al. 1993), most miRNAs have since been discovered by small RNA cloning, starting in 2001 (Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros 2001). Recent advances in massive parallel sequencing technologies have greatly increased the sensitivity of such approaches, leading to the discovery of many novel miRNAs and suggesting that miRNA discovery in well-studied organisms might be reaching saturation (Berezikov et al. 2006b; Ruby et al. 2006). Nevertheless, miRNAs that are expressed in low levels, only in a few cells, or under highly specific conditions remain difficult to detect experimentally (e.g., *lsy-6* in *Caenorhabditis elegans*; Johnston and Hobert 2003; Ruby et al. 2006). Ideally, computational approaches should be able to complement these studies and discover miRNAs directly from their sequence features in complete genomes. In contrast to experimental approaches, which are inherently limited by the developmental stages and

[4]These authors contributed equally to this work.
[5]Present address: Institute of Computer Science, University of Tartu, Estonia and Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK.
[6]Corresponding authors.
E-mail manoli@mit.edu; fax (617) 253-6652.
E-mail alex.stark@mit.edu; fax (617) 253-6652.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6593807. Freely available online through the *Genome Research* Open Access option.

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

tissues surveyed, computational studies should provide a global view of miRNAs regardless of their expression. Several structural features of miRNA hairpins have been defined and used to predict novel miRNA genes (for review, see Berezikov et al. 2006a). However, in the absence of sufficiently specific miRNA hairpin characteristics (Ng Kwang Loong and Mishra 2007; Ritchie et al. 2007), these are insufficient to recognize the small number of true miRNA genes amidst the enormous number of nonbiological miRNA-like hairpins in large animal genomes (see below and Berezikov et al. 2006a).

Comparative genomics provides an opportunity to discover functional miRNAs systematically, making use of their conservation across multiple species. Comparative approaches have been applied to discover a wide range of functional elements, including protein-coding genes, RNA genes, various classes of regulatory elements or motifs (e.g., for review, see Miller et al. 2004), and have also been used for the prediction of miRNA target genes (for review, see Lai 2004; Rajewsky 2006). For miRNA identification, comparative approaches using a small number of species have led to the discovery of novel miRNAs in flies, worms, and mammals (Lai et al. 2003; Lim et al. 2003a,b; Bentwich et al. 2005b; for review, see Berezikov et al. 2006a). It is still unclear, however, to what extent a purely de novo approach can be used to identify novel miRNAs with high specificity and whether it can define their precise boundaries. It is also unclear how discovery power scales with the number of sequenced species, and whether comparative genomics can lead to functional insights on miRNA function beyond their discovery. The recent availability of 12 fully sequenced *Drosophila* genomes spanning over 40 million years of evolution (*Drosophila* 12 Genomes Consortium 2007; Stark et al. 2007) provides a unique opportunity to address these questions, providing a rich comparative genomics data set, at a range of evolutionary distances.

Drosophila melanogaster presents perhaps the most important model organism for understanding the basic principles and molecular mechanisms of animal development. Similarly, *Drosophila* genetics has played an important role in understanding the functional roles of animal miRNAs. Several miRNA loci were first discovered in the 90s by means of gain-of-functions screens, and only later recognized to be miRNAs (e.g., *mir-7*, Li and Carthew 2005; *bantam*, Brennecke et al. 2003; *mir-278*, Teleman and Cohen 2006). Gain-of-function alleles also led to the discovery of regulatory motifs in the 3′ UTRs of several genes that were later identified as miRNA target sites (Lai and Posakony 1997; Lai et al. 1998; Lai 2002). The sequencing of *Drosophila pseudoobscura* (Richards et al. 2005) enabled the first comparative genomics approaches in *Drosophila*, which proved successful in the identification of novel miRNAs (Lai et al. 2003), the prediction of miRNA target genes (Enright et al. 2003; Stark et al. 2003; Grun et al. 2005), the refinement of rules for miRNA-target recognition (Brennecke et al. 2005), and the description of global patterns of miRNA regulatory networks (Stark et al. 2005). Nevertheless, our picture on *Drosophila* miRNAs is far from complete. In fact, with 60 cloned miRNAs in the Rfam miRNA registry (Griffiths-Jones et al. 2006), only half of the conservative upper estimate of 120 loci are known (Aravin et al. 2003; Lai et al. 2003).

In this paper, we use 12 *Drosophila* genomes (*Drosophila* 12 Genomes Consortium 2007; Stark et al. 2007) to undertake a systematic de novo discovery of miRNA genes and gain insights into miRNA biology. We define distinguishing properties of known miRNA genes, both structural and evolutionary. We use these properties for the de novo discovery of miRNA hairpins and the prediction of the precise start position of mature miRNAs. Our methods predict >41 novel miRNA genes, of which 28 are validated experimentally, and 19 more extend known miRNA families or cluster with known miRNAs. In eight cases, our prediction and/or validation correct the current Rfam annotation of the mature miRNAs by shifts of one or more bases, leading to drastic differences in the set of predicted target genes. The novel predicted miRNAs lead to 37 novel miRNA families with unique 5′ seed sequences, doubling the number of miRNA families and, thus, the diversity of miRNA targeting in *Drosophila*. Our results also lead to several new insights into miRNA biology. We find that, in some cases, both sense and anti-sense sequences have miRNA-like characteristics and can be processed into mature miRNAs. We also show that a single hairpin can give rise to multiple mature miRNAs, whose abundance correlates with the strength of our computational signals. We also show that miRNA star sequences can be functional (miRNA* denotes the small RNA processed from the hairpin arm opposite of the mature miRNA). For example, for *mir-10*, a miRNA in the Hox cluster, both miR-10 and miR-10* appear functional, with highly conserved target sites in multiple Hox genes. In fact, miR-10* shows stronger evolutionary signals and is more abundant, suggesting it may be the primary product. Lastly, we show that the newly discovered miRNAs target an overall similar gene set to known miRNAs, albeit in different combinations stemming from distinct seeds. This implies a much denser miRNA regulatory network than previously thought, with increased potential for combinatorial control.

Results

Structural and evolutionary properties of known miRNA hairpins

We studied the structural and evolutionary properties of the known *Drosophila* miRNAs to derive discriminating features, which distinguish them from nonfunctional miRNA-like hairpins that occur frequently throughout the genome (Fig. 1). We compared the cloned miRNAs in Rfam release 9.0 (Griffiths-Jones et al. 2006) with hairpins in randomly chosen genomic sequences, and with a subset of the random hairpins, selected to resemble known miRNAs in their length and pairing characteristics (see Methods).

We found several distinct structural characteristics of miRNA hairpins (Fig. 1A,B), partly also reported and used in previous prediction efforts (e.g., Lai et al. 2003; Lim et al. 2003b; Bentwich et al. 2005b; for review, see Berezikov et al. 2006a). First, the length of miRNA hairpins is much more precisely defined than that of random hairpins: 90% of all miRNA hairpins are between 73 and 102 nucleotides (nt) long, whereas 90% of random hairpins have lengths between 47 and 117 (2.4× larger range). Second, the lengths of the arms and the hairpin loop are more restricted for miRNAs and show a clear trend to longer arms and shorter loops on average: 90% of all miRNAs have arms lengths between 31 and 47 and loops between 4 and 26 nt, while random hairpins are much less defined. Third, miRNA precursor sequences fold into significantly more stable secondary structures than random hairpins, between 3 and 6 standard deviations above what is expected given their lengths and sequence composition. Fourth, miRNA hairpins have more symmetric loops but fewer asymmetric or bulged loops than random hairpins of similar length. Such loop structures in the arms may direct hairpin cleavage, facilitate asymmetric strand separation, or allow for

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

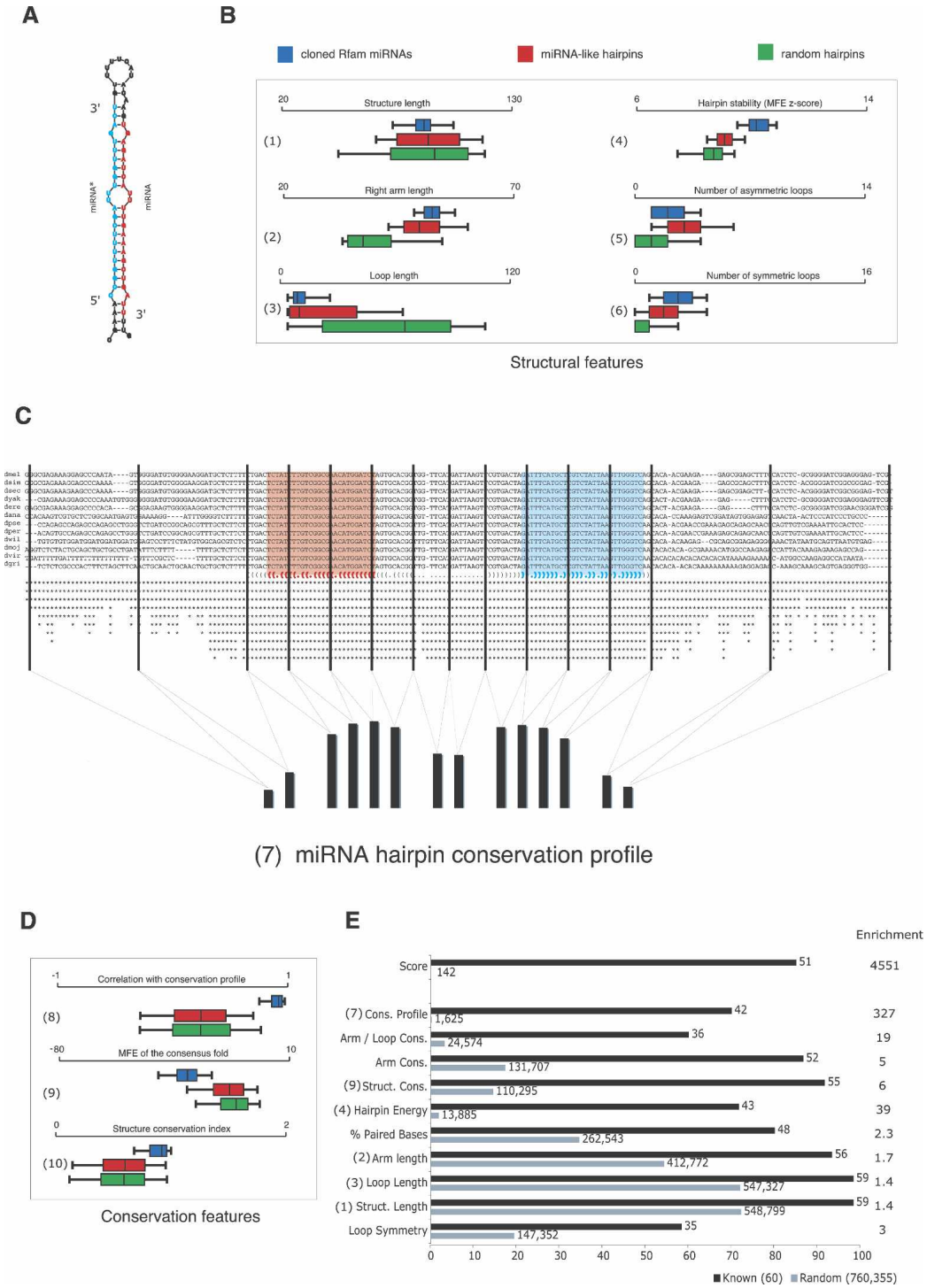


Figure 1. Evolutionary and structural properties of miRNA hairpins. (*A*) Typical miRNA hairpin showing mature miRNA (red) and miRNA* (blue). (*B*) Structural properties of miRNA hairpins for cloned miRNAs (blue), random hairpins of similar lengths and arm pairing (red), and random genomic hairpins (green) (boxes denote 25 percentiles and the whiskers 95 percentiles). miRNAs show more defined hairpin lengths (1) with longer arms (2) and shorter loops (3) than random hairpins. They are more stable than random hairpins (*Z*-scores denote standard deviations above minimum free folding energies of random hairpins with identical lengths and composition) (4), contain fewer asymmetric (5), but more symmetric loops than random hairpins (6). (*C*) Alignment of a typical miRNA hairpin across 12 *Drosophila* genomes shows high conservation in the arms, whereas the loop region and the surrounding sequence are poorly conserved (red, mature miRNA; blue, star sequence). "." and "()" denote unpaired and paired nucleotides according to Hofacker et al. (1994). We measure the conservation in 14 regions distributed over the arms (four each), loop (two), and flanking regions (two each; indicated by vertical bars). The average conservation in these regions across a nonredundant set of cloned *Drosophila* miRNAs shows a characteristic profile (*bottom*) that closely follows the hairpin structure. (*D*) Conservation properties of miRNAs (box-and-whisker plots as in *A*). Cloned miRNAs show a significantly higher correlation to the average conservation profile than the random sets (8). The structures of miRNAs hairpins are generally more highly conserved than random hairpins as assessed by the minimum free energy of the consensus fold (9) (Hofacker et al. 2002) or the structure conservation index (10) (Washietl et al. 2005). (*E*) Discriminatory power of individual features and combined score (features are numbered as in *A*–*C*). The enrichment factor (*right*) indicates how much more miRNAs (black bars) than random hairpins (gray bars) are selected at the cutoff that maximizes the information gain (Mitchell 1997). The numbers next to the bars indicate the total number of miRNAs and random genomic hairpins that are predicted, suggesting that individual features predict many random hairpins in an entire genome.

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

efficient loading of mature miRNAs into effector complexes, processes which are not yet fully understood (Khvorova et al. 2003; Schwarz et al. 2003; Han et al. 2006; Seitz and Zamore 2006).

We next assessed the evolutionary constraints of miRNAs in 12 *Drosophila* genomes (*Drosophila* 12 Genomes Consortium 2007; Stark et al. 2007). We aligned each of the 60 cloned miRNAs and their flanking regions across all 12 species (we determined the corresponding sequence in each of the 12 genomes by BLAST, as existing whole-genome alignments were not found to be reliable). These alignments show a characteristic conservation profile, which closely follows the hairpin structure (Fig. 1C), as previously reported in flies and mammals (Lai et al. 2003; Berezikov et al. 2005). The loop and the flanking regions show abundant mutations, insertions, and deletions, while the arms are very highly conserved. In particular, while compensatory mutations are frequently observed in other RNA genes, they are not found in miRNA hairpins, especially within the mature miRNA. In fact, for all but eight cloned miRNAs, the mature sequence shows a 100% conservation in all species where the corresponding hairpin sequence can be found (Supplemental Table 1). For four of these eight (*mir-277*, *mir-2b-1*, *mir-305*, *mir-309*), the only sequence difference is found in the closely related *Drosophila simulans* or *persimilis*, which have been sequenced at lower coverage, suggesting the differences may represent sequencing errors rather than evolutionary divergence. Perfect conservation of the entire mature miRNAs might reflect additional functional constraints on the arm sequences beyond simply stabilizing an RNA secondary structure (e.g., compensatory pairing of the miRNA 3′ end; Brennecke et al. 2005) and has been found to extend across species as distant as worms and vertebrates in some cases (e.g., *let-7*; Pasquinelli et al. 2000). In contrast, the high sequence divergence of the loop suggests that it may simply be a linker with no additional functions.

Taken together, these structural and evolutionary features set real miRNA hairpins apart from random miRNA-like hairpins. We next use combinations of these features for the genome-wide identification of *Drosophila* miRNAs.

Distinguishing *Drosophila* miRNA hairpins from random genomic hairpins

Discovering novel miRNA genes computationally is an enormous challenge: Cloned miRNAs represent only 60 of 760,355 hairpins in the fly genome with miRNA-like properties (total length between ~60 and 120 nt, arm length >20 nt, with at least 70% paired bases). Therefore, in addition to selecting for miRNA-like hairpins, several additional features must be defined in order to achieve the needed discriminatory power of >99.99% specificity. For example, a 0.5% false-positive rate would result in 3500 spurious predictions.

We tested how well each of the structural and evolutionary properties above discriminates between the known miRNA hairpins and all miRNA-like hairpins in the genome (Fig. 1E). For example, scoring the miRNA conservation profile (Fig. 1C) alone shows >300-fold discrimination between miRNAs (70% pass) and other hairpins (only 0.2% pass) across 12 genomes. Our conservation profile score is related to the metric of arm vs. loop conservation ratio, used for miRNA prediction in two *Drosophila* genomes (Lai et al. 2003), albeit much more precise—for comparison, the previous metric shows an enrichment of only 19-fold across 12 genomes.

Amongst structural features, the free energy score of the hairpin, corrected for hairpin length and sequence composition, is most discriminative with a 39-fold discrimination. The folding energy of the consensus sequence, which measures structural conservation (Washietl et al. 2005), shows a sixfold enrichment, followed by the overall loop-symmetry (threefold) and the extent of base-pairing in the hairpin arms (2.3-fold). Remaining features such as the length of the entire hairpin, the arms, or the loop show only moderate contribution (less than twofold). We conclude that, after selection for hairpin length and stem pairing, the remaining contribution of structural features is only moderate.

Although several individual properties provide a significant discrimination, the specificity of even the best of them is not sufficient to reliably predict miRNAs. Given the large number of genomic hairpins, many would be selected that score well for that property but lack other miRNA characteristics and are likely false. For example, most of the 1625 miRNA hairpins selected by the conservation profile above (99.8% rejection rate for random hairpins) are likely false: 40% alone fall into exons or repeats or have accumulated mutations in their mature sequence. Therefore, several features need to be combined to reach a sufficient discrimination power. To identify optimal combinations and weightings of these features, we used a machine learning approach, using tallied votes of 500 independently constructed decision trees (see Methods; Breiman 2001). The combined approach achieves >4,500-fold discrimination after cross-validation, recovering 85% of known miRNAs in <0.02% of genomic hairpins.

Recovering known miRNA genes

To discover *Drosophila* miRNAs in the genome, we ranked all 760,355 miRNA-like hairpins in the entire genome according to this combined score. The top ranking hairpins are strongly enriched in cloned Rfam miRNAs, whose discovery rate plateaus at a score of 0.95 (Fig. 2B). At this cutoff, our method results in 101 hairpins in the fly genome, including 51 of the 60 cloned miRNAs (85%) and novel miRNA candidates with structural and conservation scores similar to, and sometimes higher than, known miRNAs (Fig. 2; Table 1). Upon inspection, of the remaining nine cloned miRNAs which are not recovered, *mir-9b* ranks only slightly below the cutoff (0.938; rank 111), *mir-278* ranks 278th with significantly lower conservation and structural scores, four appear conserved only in very close species (*mir-303*, *mir-309*, *mir-311*, *mir-312*), and two contain very long insertions or deletions in at least one species (*mir-6-2* and *mir-31b*); finally, *mir-314* shows a notable *D. melanogaster*-specific sequence difference, substituting a highly conserved U with a C (Supplemental Fig. S1).

Our top predictions also contain nine of the 18 previously predicted but not yet cloned Rfam miRNAs (Lai et al. 2003). Of the nine Rfam miRNAs which we do not recover, two are poorly conserved and found only in close species (*mir-310*, *mir-313*), two have a highly conserved hairpin loop, which is very unusual for *Drosophila* miRNAs (*mir-100* and mir-*125*), and one has an unusually large hairpin loop (*mir-2c*). The remaining four appear to be spurious (*mir-280*, *mir-287*, *mir-288*, *mir-289*) and, in contrast to all other Rfam miRNAs (see below), none of them could be verified in a large scale sequencing effort for *Drosophila* miRNAs (Ruby et al. 2007, this issue).

Strong support for our method and score cutoff stems from the genomic positions of the candidate hairpins, a feature not used for scoring. We plotted the fraction of all hairpins residing

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

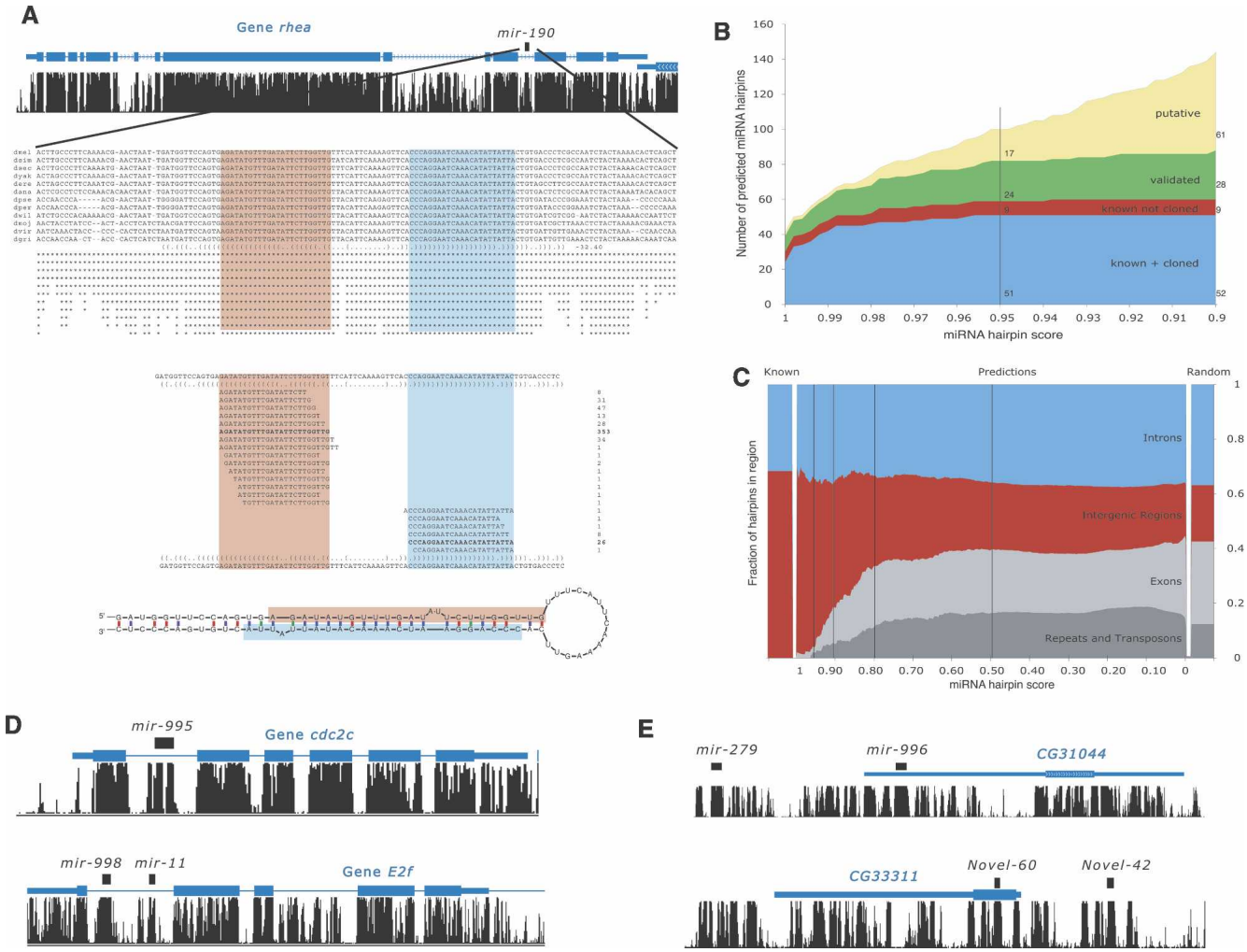


Figure 2. Novel *Drosophila* miRNAs. (*A*) Prediction and validation of miRNA *mir-190*. *mir-190* (black) is predicted in the intron of the cytoskeleton anchor protein rhea (blue, UCSC browser screen-shot) in the direction of transcription, sequence alignment of *mir-190* across 12 *Drosophila* genomes and conservation profile highlighting the mature miRNA (red) and the star sequence (blue). "." and "()" denote unpaired and paired nucleotides according to Hofacker et al. (1994); experimental validation of *mir-190* (total read counts from our and Ruby et al.'s [Ruby et al. 2007] sequencing shown to the *right*). Matching sequence reads show a characteristic pattern of processing with the total reads obtained for the miRNA and the star sequence, indicative of true *Drosophila* miRNAs. (*B*) Recovery of known novel miRNAs. Count of predicted miRNAs (*Y*-axis) at different score cutoffs (*X*-axis), for cloned (training set, blue), previously annotated but not cloned (red), novel and validated (green), and additional novel (yellow). At a conservative cutoff of 0.95, we recover 51 (85%) of cloned and nine previously annotated miRNAs among a total of 101 predictions. Of the 41 novel miRNAs, 24 (59%) are experimentally validated. A more lenient cutoff of 0.9 predicts 49 additional miRNAs, including one additional known miRNA and four novel miRNAs that are validated experimentally. (*C*) High-scoring hairpins are specific to introns and intergenic regions and exclude exons, repeats, and transposons. Shown are percentages for each region (*Y*-axis) for hairpin scores from 1.0 (best) to 0.0 (worst; *X*-axis). This is compared to the ratios obtained for random hairpins (Random) and known miRNAs (Known). For scores <0.8 the distribution of hairpins is indistinguishable from random, arguing that no further conserved hairpins can be expected at a reasonable frequency. (*D*) Examples of novel intronic 1 (four total) and clustered (six total) miRNAs. *mir-995*, that is in an intron of *cdc2c* and *mir-998* is ~500 nt from *mir-11* in the intron of *Ef2*. (*E*) Novel miRNAs explain transcript of erroneously annotated genes. *CG31044* and *CG33311* are likely the precursor transcripts of *mir-996*, which is ~2.5 kb from *mir-279* (a member of the same family) and *Novel-60*, which is ~1300 nt from *Novel-42*.

in exons, introns, repeats, and intergenic regions as a function of the hairpin score (Fig. 2C). Strikingly, while the occurrence of miRNA-like hairpins in the genome is indistinguishable from a random distribution according to the overall region lengths, the top scoring hairpins lie almost exclusively in introns and intergenic regions. Indeed, all known miRNA genes in animals lie in introns of protein coding genes or as separate, intergenic transcription units. For scores <0.95, the fraction of exonic and repeat-derived hairpins increases, and, for scores <0.80, it becomes indistinguishable from random, suggesting that any remaining miRNAs in this score range represent only an extreme minority amidst mostly spurious candidates. We conclude that our approach successfully discriminates real miRNA hairpins from abundant spurious hairpins based on patterns of evolutionary conservation and structural features.

Experimental validation of hairpin predictions

The top 101 hairpins include 41 novel miRNA candidates, for which several measures suggest that they represent functional, novel miRNA genes, nearly doubling the number of miRNA genes in the fly. To experimentally validate our predictions, we

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

Table 1. Predicted miRNAs (score ≥ 0.95)

Score	Validation	Locus	Name	Host Gene	Species	Mature	Solexa	454	Targets
1.000	H	2R 4824837 4824929 −	miR-987		12+	UAAAGUAAAUAGUCUGGAUUGAUG	6	793	173
1.000	H	2R 7674782 7674866 −	miR-988	*prp8*	11+	CCCCUUGUUGCAAACCUCACGC	12060	1796	15
1.000	H	3L 21507860 21507959+	miR-193		12+	UACUGGCCUACUAAGUCCCAAC	14	535	20
1.000	HM	3L 8545756 8545843+	miR-190	*rhea*	12+	AGAUAUGUUUGAUAUUCUUGGUUG	12	506	116
1.000	H	3R 121093 121174+	miR-929	*cpx*	12+	CUCCCUAACGGAGUCAGAUUG	3	112	20
1.000	H	3R 21414588 21414680 −	miR-1000	*msi*	12+	AUAUUGUCCUGUCACAGCAGU	1	325	147
1.000	HM	3R 25042907 25043002+	miR-996		12+	UGACUAGAUUUCAUGCUCGUCU	1309	4478	119
1.000		3R 27091338 27091410 −	Novel-8		12+	CAAAUUAACUGCGACAUGGC			479
1.000	H	3R 6233855 6233952+	miR-994		12+	CUAAGGAAAUAGUAGCCGUGAU	33992	154	65
1.000	H	X 15799886 15799976 −	miR-927		12+	UUUAGAAUUCCUACGCUUUACC		351	134
1.000	H	X 17958301 17958382+	miR-969		12	GAGUUCCACUAAGCAAGUUUU		10	34
0.994	HM	2L 243037 243141 −	miR-965	*kis*	12+	UAAGCGUAUAGCUUUUCCCCUU	18	133	73
0.994	HM	3L 22452052 22452119 −	miR-957		12+	UGAAACCGUCCAAAACUGAGG		66	45
0.994	HM	X 1645021 1645108 −	miR-981		12+	UUCGUUGUCGACGAAACCUGCA	16	1736	147
0.990	HM	2L 857543 857629+	miR-375		12+	UUUGUUCGUUUGGCUUAAGUUA	1438	336	196
0.988	HM	X 12896014 12896100+	miR-971		11+	UUGGUGUUACUUCUUACAGUGA	38	52	121
0.986		3L 4989729 4989829+	Novel-17		10+	AAAAUUAUGCGGAAACGGAAGC			519
0.986	H	X 12470276 12470364+	miR-970	*tomosyn*	12+	UCAUAAGACACACGCGGCUAU	217	484	52
0.982	L	2R 11580121 11580206 −	miR-137		12+	UAUUGCUUGAGAAUACACGUAG	36	48	256
0.982	HM	3R 17623951 17624051 −	miR-999	*Caki*	12+	UGUUAACUGUAAGACUGUGUCU	1114	414	115
0.980		3L 10936322 10936415+	Novel-21		12+	UCGUCGCAUGCGCGUGAUCAAC			26
0.980		3R 23797295 23797385 −	Novel-22	β*Tub97EF*	9	UUUAUUGCGGCCUGGCCUGACA			521
0.978		2R 10136644 10136747+	Novel-23		12+	GAAAGAAUAAGAACGGCCAACU	1		105
0.976		3R 22570390 22570454+	Novel-24		12+	UCAAUCAAAUCACAUGACUGCU			98
0.976		3R 24822601 24822689 −	Novel-25	*CG1443*	12+	UGCAUUUAAGCCAAUUAGCAUA			292
0.974	H	2L 5642112 5642202+	miR-964	*CG31646*	6+	UUAGAAUAGGGGAGCUUAACUU	6393	57	146
0.974		3L 7188122 7188227 −	Novel-27		5	UGAGUCCUUUCACUGGCCACUC			23
0.968		2L 11749802 1174899 −	Novel-28		8+	UGCUUUGAGGUUUAUUAGCUGC			109
0.968	H	3R 17447616 17447730 −	miR-998	*E2f*	12+	UAGCACCAUGAGAUUCAGCUC	515	519	39
0.966	HM	3R 16561652 16561725+	miR-995	*cdc2c*	10+	UAGCACCACAUGAUUCGGCUU	1422	1322	44
0.964		3R 16281787 16281884+	Novel-31		12+	AAUGUCAUUAAAUUCUCAUACA			68
0.962		X 15110474 15110559+	Novel-32		11+	UUUUAUUUGUGUCACUGAGUGG			1399
0.960		3R 21923504 21923596+	Novel-33		12+	UUUGUUCGAGUUGACGUUUGGA			178
0.960		X 18015331 18015398 −	Novel-34		10	UACAUAAUGUCUCUGUAGGCC			320
0.958		2L 3858079 3858150+	Novel-35		12+	AAUUUAAUGUGUCGGCGUGUUU			482
0.958		2R 13077473 13077566+	Novel-36	*Klp54D*	8+	UGUUCUCUCCCAUUUCUGACUC			59
0.954	H	2L 13747747 13747840 −	miR-968		6+	UAAGUAGUAUCCAUUAAAGGGUUG		84	85
0.954	H	3R 9289945 9290028 −	miR-252	*CG17025*	12+	CUAAGUACUAGUGCCGCAGGAG	1173	7130	190
0.952		2L 15654271 15654348 −	Novel-39		11+	UAAUUGCCUGUAAACAUAAAGG			146
0.952	HM	2L 6902071 6902141+	miR-932	*neuroligin*	8+	UCAAUUCCGUAGUGCAUUGCAG	52	615	90
0.952		3L 7528550 7528614 −	Novel-41		12+	UACUUUUACUUUCAUUUAUCAA			193

Sequencing validated the genomic locus only (L), the hairpin with the correct strand (H), and/or the predicted mature 5′ end (M). "Species" summarizes the number of species the hairpin is found in, and "+" indicates if the mature miRNA was perfectly conserved across all those species (see Methods). "Solexa" and "454" summarize the reads supporting the annotated mature 5′ end from ovaries/testes libraries and from Ruby et al. (2007), respectively.

obtained 763,111 Solexa sequencing reads corresponding to 1524 distinct sequences from *Drosophila* ovaries and testes that match our predictions (see Methods and Supplemental data). We required multiple independent reads of the mature miRNA and manually inspected the reads for miRNA-like processing, providing a very stringent validation, essentially free of false positives (Ruby et al. 2006). Sixty-nine of our 101 predictions with scores ≥0.95 were supported by ≥10 reads each, including 17 of the 41 novel predictions. In addition, when we required three or more reads per hairpin position, our data supported 19 novel and 73 hairpins in total. In contrast, none of 500 random hairpins with scores <0.1 were supported by three or more reads. Indeed, only two matched to spurious reads that were each cloned only once. As additional predicted miRNAs might be valid yet not expressed in ovaries or testes, we also intersected our predictions with curated reads from 454 sequencing of small RNA libraries from several developmental stages and tissues (Table 1; Fig. 2A,B; intersection kindly performed by Graham Ruby, Eric Lai, and David Bartel [Ruby et al. 2007]). In total, we validate 84 of our 101 predictions (83%) with scores ≥0.95, including 24 of the 41 novel predictions (59%). At a less stringent cutoff of 0.90, an additional 49 novel miRNA candidates are predicted, of which a total of four are validated (Fig. 2B). This increases the number of cloned *Drosophila* miRNAs by 28 (47%) and suggests that our top predictions include the vast majority of conserved miRNAs.

Additional evidence supports novel miRNAs are functional

We next evaluated the 61 candidate miRNAs, that did not intersect with sufficient sequencing reads (17 with scores ≥0.95 and an additional 44 with scores ≥0.90; Table 1). We evaluated these using additional properties of known miRNA genes, which were not used as features during the prediction of these hairpins. We found that, indeed, numerous candidates show several striking properties. For example, 17 of the 61 miRNAs are family members of known *Drosophila* miRNA families, four have orthologs in mosquito, six in worm, and five in human. An additional five candidates do not belong to known families but have predicted 5′ ends (see below) with 7-mers that scored equally high or higher than most miRNAs (MCS ≥ 40; see below). We found that

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

three of the 61 hairpins are clustered in the genome, a property of many known miRNAs. Overall, 19 of the 61 hairpins are found in introns, common among miRNAs and suggestive of transcription. Indeed, five of these and an additional three intergenic hairpins are expressed during embryogenesis, as revealed by a recent genome-wide expression study using tiling arrays (Manak et al. 2006). Even though this is not significantly more than expected given the overall high prevalence of transcription (~1/6 intergenic probes), it demonstrates the presence of hairpin transcripts necessary for miRNA processing.

We conclude that several of these candidate miRNAs may be functional. These are likely to be of low abundance (similar to worm *lsy-6*; Johnston and Hobert 2003; Ruby et al. 2006), or expressed in tissues or under conditions that were not surveyed. We did not experimentally validate hairpin candidates with scores <0.9, but several metrics suggest they no longer include meaningful miRNAs at a reasonable rate. Amidst 386 hairpins scoring between 0.8 and 0.9, a considerable fraction (25%) falls within protein-coding exons, UTRs, and repeat elements, which are unlikely to contain functional miRNA genes. Below 0.8, the fraction of annotated exons and repeat elements is indistinguishable from random (Fig. 2C). We conclude that few conserved miRNAs are missed by our analysis.

Overall, our analysis resulted in 28 novel miRNA genes with strong evolutionary conservation and experimental support, and additional candidates. Given the recovery rate of 85% of cloned miRNAs, we estimate a total of ~120 conserved miRNAs, which is similar to previous estimates based on comparative information (Lai et al. 2003) and the number of known miRNAs in worm (Ruby et al. 2006). Note, however, that the observation that less deeply conserved miRNAs are expressed at lower levels (see below and Ruby et al. 2006) does not allow an accurate estimation of the number of nonconserved miRNAs.

High accuracy stems from feature set and number of species

Overall, our method recovers 85% of known miRNAs, and at least 83% of all predictions are correct. This high accuracy stems from the features we define, the increased number of species in our comparison, and the large evolutionary distances they span.

To study the effect of evolutionary distance on our performance, we applied our methods to different subsets of species and assessed the number of true miRNAs that were found among the top 100 predictions (Supplemental Fig. S3). For pairwise comparisons, the discovery rate increases approximately linearly with the evolutionary distance of the two species: Close species in the *melanogaster* subgroup (*D. simulans*, *sechellia*, *yakuba*, and *erecta*) allow the discovery of up to 49 valid miRNAs, those in the *D. sophophora* clade (*D. persimilis*, *pseudoobscura*, and *willistonii*) up to 73, and the most remote species (*D. mojavensis*, *virilis*, and *grimshawii*) up to 76 among the top 100 predictions. For multispecies comparisons, inclusion of all species at the same evolutionary distance cutoffs led to 49, 78, and 84 recovered miRNAs, respectively (Supplemental Fig. S3).

To study the effect of our feature set and methodology, we compared the number of recovered miRNAs to those predicted by Lai et al. (2003), by using *D. pseudoobscura* as our only informant species. Lai and colleagues reported 208 predicted miRNA loci, including 60 Rfam miRNA loci, and 13 novel miRNA loci reported by Ruby et al. (2007). By comparison, using the same rank cutoff, we recover a similar number of previously known miRNAs ($n = 61$), but nearly twice as many newly cloned ones ($n = 23$). The difference is likely due to the new features we define, the new methods for combining them, and also the new Rfam miRNAs available for training.

Finally, we asked whether clade-specific miRNAs could be discovered in clade-specific searches, evaluated using a set of 28 miRNAs that are not conserved outside the *melanogaster* species (i.e., beyond *D. ananassae*) and a set of 32 miRNAs that are not conserved outside the *sophophora* species (i.e., beyond *D. willistoni*). We tried 21 species combinations, but we found only one case of a discovered clade-specific miRNA that was not found using all 12 species, suggesting that clade-specific miRNAs are generally missed by this approach. Two factors contribute to this: First, the currently sequenced species do not provide sufficient discovery power for clade-specific miRNAs, due to insufficient genome sequencing at short branch lengths; second, the conservation properties of clade-specific miRNAs are likely to differ from the conservation properties of Rfam miRNAs, which are generally much more deeply conserved. Although we retrained our features for each species combination, we only used Rfam miRNAs as our training set, likely biasing against clade-specific structural and conservation patterns. It is likely that an improved understanding of the structural requirements for precursor processing, a larger training set including clade-specific miRNAs, and additional sequencing of closely related species will enable de novo discovery of additional miRNAs. However, it is currently unclear if structural properties in the absence of evolutionary signatures can be sufficient to reliably detect truly species-specific miRNAs that are diverged even between very close species (for such miRNAs in primates, see Bentwich et al. 2005a; Zhang et al. 2007).

In summary, we found that discovery power continued to increase with additional species and additional evolutionary distance, without reaching saturation. Indeed, using 12 species always performed best, whether we considered all miRNAs, or the clade-specific sets. We thus expect that additional species will lead to continued increase in discovery power, especially in the ability to discover clade-specific genes. Overall, for conserved miRNA genes, the 12 *Drosophila* species provided an ideal substrate for miRNA discovery, resulting in highly accurate de novo predictions.

Genomic clustering of novel miRNA genes and relationship with protein-coding genes

Many of the novel candidates show genomic clustering, a feature frequently observed among animal miRNA genes. Among the 28 validated novel miRNAs, six are clustered with known or novel miRNA genes (Fig. 2D,E). These include three novel miRNA genes within 4 kb of *mir-318*, *mir-279*, and *mir-210*, respectively, one novel miRNA gene 500 nt from *mir-11* in the *E2F* intron, and two novel miRNA genes within 1 kb of each other in consecutive introns of *CG31646*. Among the 61 nonvalidated candidates with scores ≥0.90, three additional hairpins are within 3 kb of each other. Given the abundance of directed knockout experiments in *Drosophila*, recognizing clustered miRNAs is essential, as neighboring miRNAs will undoubtedly influence knockout phenotypes, and sometimes may be primarily responsible for them.

We find that 14 of the 28 novel verified miRNAs are within introns of protein-coding genes (Fig. 2D; Table 1), also a common feature of animal miRNAs. For 10 of these, the miRNA is in the same strand as the host gene, suggesting common transcriptional regulation (Baskerville and Bartel 2005). We particularly point

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

out *mir-1000*, which lies within the intron of *mushashi* (*msi*), an RNA-binding protein that has been reported to down-regulate proteins translationally (Okabe et al. 2001). This novel miRNA might be involved in, or responsible for, the previously reported *mushashi* functions. Similarly, *mir-998* and *mir-995* reside in the introns of the cell-cycle regulators *E2F* and *cdc2c*, respectively (Fig. 2D), and are likely involved in the regulation of cell division. Notably, their vertebrate homolog *mir-29* has been implicated in cancer (Pekarsky et al. 2006).

In the remaining four cases, intronic miRNAs lie on the opposite strand of annotated genes, suggesting expression as independent transcription units (Aboobaker et al. 2005). For example, *mir-964* and *mir-959*, both validated, lie on the opposite strand of *CG31646*. Indeed, these are supported by reverse-strand cDNAs and ESTs, which are correlated to each other in expression, but uncorrelated with *CG31646*.

Finally, two novel miRNAs overlap exons of previously annotated protein coding genes. Candidate *Novel-60* overlapped the annotated coding region of *CG33311* and validated *mir-996* resides in the annotated 5′ UTR of *CG31044* (Fig. 2E). In both cases, an independent analysis of protein-coding evolutionary constraint had in fact rejected these genes as unlikely to encode for proteins (Lin et al. 2007): 12-species alignments for both genes are littered with stop codons, frame-shifting insertions and deletions, and nonconservative codon substitutions, suggesting they are not under protein-coding selection. Consequently, the newly discovered miRNA genes provide an explanation for the previously observed transcripts, which we conclude are not encoding proteins. This finding highlights the importance of systematic high-quality annotation of both protein-coding and non-protein-coding genes, based on their specific evolutionary constraints. It also illustrates the power of our unbiased genome-wide prediction of miRNAs that, unlike previous methods, did not explicitly exclude exons from the regions searched.

Both strands can be transcribed and processed

Intuitively, one may expect that the reverse complementary sequence of a miRNA hairpin would also fold into an equivalent hairpin. However, due to GU base pairs that translate into incompatible AC base pairs, and sequence-specific energy terms, this is generally not the case. In fact, we found large differences between sense and anti-sense hairpins for most cloned miRNAs: while 51 sense hairpins scored ≥0.95, only 21 of the complementary reverse-strand hairpins reached that score, and only four of 51 reverse complements scored more highly than the correct strand (Supplemental Fig. S2). While the contribution of hairpin conservation is independent of the strand, high scores, when considering a wide variety of structural features including loop lengths, structure, and symmetry, suggest that these anti-sense hairpins would constitute bona fide miRNAs if transcribed.

Expression data confirmed that both strands of a miRNA gene can be correctly processed into mature miRNAs and that they may be independently regulated. In four cases (*mir-iab-4*, *mir-307*, *mir-124*, and *mir-305*), sequencing reads were found that corresponded to a correctly processed mature miRNA product for the reverse strand (Ruby et al. 2007), albeit at a much lower level than the forward-strand product. In all four cases, both forward and reverse strand showed scores >0.97 based on our evolutionary and structural metrics. For one case in particular (*mir-iab-4*), both sense and anti-sense transcripts have been detected by in situ analyses in fly embryos (Bae et al. 2002), indicating robust transcription of both strands. Interestingly, the two strands of *mir-iab-4* are expressed in distinct and mutually exclusive embryonic segments, suggesting independent regulation of two distinct miRNAs, likely with distinct functions.

We conclude that anti-sense transcription, whether fortuitous or regulated, can lead to functional processing of anti-sense pre-miRNAs that resemble canonical miRNA hairpins into distinct mature miRNAs. The reverse-strand miRNAs are likely to have distinct promoters and enhancers, and also a distinct target spectrum, since their 5′ ends differ from the forward strand, leading to increased versatility for miRNA regulation.

Accurate prediction of mature miRNAs

In addition to the discovery of miRNA genes themselves, comparative genomics can help pinpoint the exact position of cleavage, allowing accurate prediction of the mature miRNA sequence. Knowledge about the precise 5′ end is particularly important because it dictates the core of the target recognition sequence (seed) (Lai 2002; Lewis et al. 2003; Stark et al. 2003; Doench and Sharp 2004; Kiriakidou et al. 2004; Kloosterman et al. 2004; Brennecke et al. 2005). As a likely consequence, miRNA 5′ ends are under strong selection and exhibit higher processing accuracy than miRNA star 5′ ends or the two respective 3′ ends (Ruby et al. 2006). Several distinctive evolutionary and structural features can be used to identify the precise 5′ end of mature miRNAs (Fig. 3A). First, almost all mature miRNAs are perfectly conserved (see above), such that the completely conserved sequence following the miRNA 5′ end is much longer than for most other positions in the hairpin. Second, as observed previously (Lau et al. 2001), >78% of *Drosophila* miRNA genes start with a 5′ uridine (perhaps reflecting binding affinities of Argonaute proteins or preferred cleavage sites for RNaseIII enzymes), while the overall frequency of uridine in miRNA hairpins is only ~30%. Third, at the structural level, we found that the number of paired bases in a 7-nt window centered at the mature start is highly constrained, requiring at least two to three paired bases, but typically excluding perfect pairing (100% of miRNAs had at least two base pairs in that window; 95% had three or more; but only 25% had perfect pairing). This suggests that miRNA processing cannot occur within large loops, and also that a certain thermodynamic instability is required between the miRNA 5′ ends and the star sequence, which is likely linked to the asymmetric strand loading of miRNAs into RISC (Khvorova et al. 2003; Schwarz et al. 2003).

In addition to these direct signals, computational prediction of mature miRNAs can use indirect signals, stemming from the relationship of miRNAs with their target genes (Fig. 3A). For example, the 7-mers complementary to miRNA 5′ ends (seed matches) are abundant in 3′ UTRs and highly preserved throughout evolution (Lewis et al. 2003, 2005; Stark et al. 2003; Xie et al. 2005). In contrast, 7-mers complementary to other parts of mature miRNAs are generally under no selective constraint. Moreover, these 7-mer seed matches are generally avoided in 3′ UTRs of ubiquitously expressed genes, thus preventing their targeting by miRNAs, while 7-mers starting at other miRNA positions are generally not avoided (Farh et al. 2005; Stark et al. 2005). Because overlapping 7-mers starting at adjacent nucleotide positions share a large fraction of their occurrences, these conservation and avoidance properties showed a larger characteristic profile, peaking at the true 5′ end and extending with moderate scores over multiple positions surrounding it. Interestingly the overall pro-

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

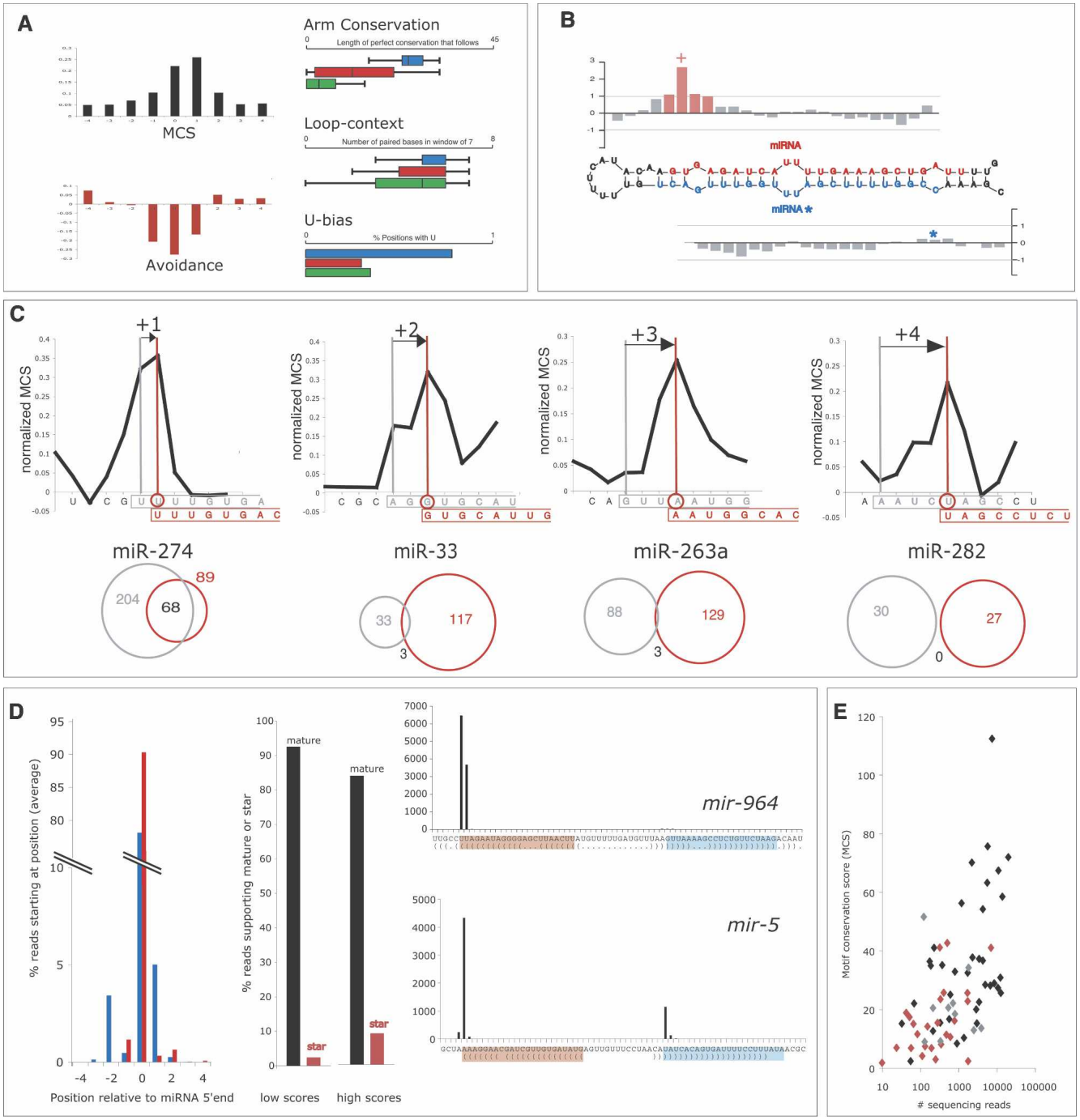


Figure 3. Properties of mature miRNAs. (*A*) Properties mature miRNA 5′ ends. 7-mers complementary to the start of mature miRNAs show a characteristic profile of 3′ UTR motif conservation scores (MCS) and avoidance in 3′ UTR of anti-target genes (normalized *Z*-scores averaged over nonredundant cloned miRNAs). miRNA 5′ ends are followed by a long stretch of perfect conservation that corresponds to the mature miRNA, and are in regions with constraint base-pairing; 78% of *Drosophila* miRNAs start with a uridine, whereas only ~30% of all hairpin positions are uridine (boxes and whiskers denote 25 and 95 percentiles, for miRNA 5′ ends [blue], random arm positions [red], random hairpin positions [green]). (*B*) Combined properties accurately highlight the start of mature miRNAs. For an average miRNA hairpin, SVM *Z*-scores that combine above properties strongly highlight the start of the mature sequence (red, scores > 1), whereas they are low or negative (gray) throughout the remainder of the hairpin. Shown is the average *Z*-score over all nonredundant cloned *Drosophila* miRNAs along a representative hairpin (averages are calculated for each position after aligning all miRNAs and miRNA*s separately at their starts). (*C*) Correction of *Drosophila* miRNA annotation. Predicted and validated changes of mature miR-263a, miR-274, miR-282, and miR-33 (black lines denote the MCS profile for the new start sites; gray and red denote old and new sequences, respectively). These changes (between 1 and 4 nt) shifted the target spectrum of the miRNA drastically (Venn diagrams), such that the target genes of the annotated (gray circles) and corrected sequences (red circles) had generally little or no overlap. (*D*) miRNAs show alternate mature forms. miRNAs for which we correctly predict the 5′ end show more precise processing on average (90% of all reads supporting the mature 5′ end at $x = 0$; red). In contrast, the remaining miRNAs show a significant drop of the central peak with more reads coming from other positions (78%, $P = 6 \times 10^{-3}$; blue; *left*). The fraction of reads supporting the star sequence (red; shown is the median percentage across all cloned miRNAs) increases for high-scoring stars, while the reads supporting the mature miRNA (black) drops (*middle*). Number of reads from Ruby et al. (2007) supporting different 5′ end positions (mature and star sequences are shaded in red and blue, respectively; *right*). Novel miRNA *mir-964* has two mature miRNAs at a ratio of 2:1: The major 5′ end is supported by 6457 (6393 from ovaries/testes) reads, but we predict position +1, which is supported by 3670 (3641 from ovaries/testes) reads. miR-5 and miR-5* are processed in a ratio of 4:1. (*E*) miRNA abundance correlates strongly with 3′ UTR motif conservation. Shown is the number of 454 sequencing reads (Ruby et al. 2007) for the dominant small RNA per hairpin and miRNA family (*X*-axis) versus the motif conservation score (MCS) of the corresponding 3′ UTR motif. A strong correlation between both measures is found with a Pearson correlation coefficient of 0.72. Colors denote Rfam miRNAs (previously cloned in black, others in gray) and novel validated miRNAs.

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

file was more discriminative than the measures at individual positions (Fig. 3A).

Using these direct and indirect features of miRNA 5′ ends, we developed a computational framework to predict mature miRNAs (Fig. 3B). We combined each of these properties and adjusted their relative contributions using a support vector machine trained on a nonredundant set of cloned *Drosophila* miRNAs using cross-validation. We found that the resulting score is much more accurate than either of the scores alone and can pinpoint precise 5′ ends for known and novel miRNA genes: When evaluated based on previously cloned miRNAs (see Methods), we found that our method pinpointed the exact start position in 47/60 cases (78%) and was within 1 bp in 51/60 (85%).

Refined annotation of known miRNA genes leads to refined target spectrum

Among the 14 annotated but not previously cloned Rfam miRNAs, our predictions disagreed with the previous 5′ annotation in nine cases, often by several nucleotides (Fig. 3C). Comparison with our sequencing reads and curated miRNAs from an accompanying paper (Ruby et al. 2007) revealed that, in fact, for six of nine cases our predictions provided a more accurate view of the 5′ end than the previous annotation, pinpointing their exact position in four cases and falling within one nucleotide for the other two (in the five cases where our predictions agreed with the previous annotation, sequencing confirmed four were exactly correct, and one was shifted by 1 nt).

The revised 5′ end annotation of Rfam miRNAs leads to dramatic changes in the inferred target spectrum, according to our published method, which combines seed matches, extended pairing information, and evolutionary conservation (Brennecke et al. 2005; Stark et al. 2005). Strikingly, we found that the overlap in predicted targets between the old and new annotation is as low as 40% for an adjustment of +1 nt, and it drops sharply to 1%–5% target overlap when the discrepancy increases to +2, +3, or more nucleotides (Fig. 3C).

The revised 5′ end annotations also result in several miRNAs now being dissimilar in sequence, family membership, and targeting properties from their currently annotated family members. In particular, miR-2c is no longer a K-box miRNA, and miR-263a no longer matches miR-263b.

Our results emphasize the importance of high-confidence annotations of miRNA 5′ ends to understand the target spectrum, biological function, and family membership of a miRNA. They also illustrate the power of comparative methods to reveal such information, solely based on the genome sequence, when experimental data are missing.

Novel miRNAs show fewer conserved targets and lower miRNA expression

Our prediction accuracy for mature miRNA gene boundaries relies on both structural features, and 7-mer-based scores derived from target 3′ UTRs. Thus, for miRNAs with few or poorly conserved targets, our accuracy may diminish. Indeed, for the 28 novel miRNAs, only 43% of our mature 5′ end predictions were confirmed and a total of 61% were within 1 nt. These show significantly fewer predicted targets on average than previously known miRNAs (100 vs. 154, $P < 10^{-2}$, counting 5′ dominant and 3′ compensatory sites; Brennecke et al. 2005) and in fact show much lower 7-mer-based scores of motif conservation (MCS of 15 vs. 36; $P < 10^{-5}$).

We reasoned that these miRNAs may be of overall lower abundance, perhaps explaining why they have been missed by earlier cloning attempts. In fact, when examining the sequencing data of Ruby et al. (2007), these 28 novel miRNAs showed significantly fewer reads on average than previously cloned miRNAs (700 vs. 4337; $P < 10^{-4}$). Overall, we found a strong correlation (Pearson coefficient: 0.72) between the number of reads supporting a mature miRNA, and the genome-wide motif conservation score of the corresponding 3′ UTR motifs (Fig. 3E).

This strong correlation suggests that the impact of miRNAs on average target 3′ UTRs increases with their abundance, stemming from either their overall expression levels, or range of expression across tissues or developmental stages. miRNAs that are expressed at low levels in restricted domains are likely only able to interact with few genes such that the number of targets but also the number of anti-targets are small. This scenario has, for example, been proposed as a likely evolutionary mechanism for the emergence of novel miRNAs and their targets (Chen and Rajewsky 2007).

Novel miRNA families lead to increased potential for combinatorial regulation

To assess the impact of the new miRNA genes on the miRNA regulatory network, we determined targets of all known and novel miRNAs, applying our previously published algorithm (Brennecke et al. 2005; Stark et al. 2005) and using the validated 5′ end for each miRNA. We find that, despite the relatively smaller number of targets, the novel miRNAs have a significant impact on the overall miRNA regulatory network.

First, they introduce many new miRNA families, as defined by shared 7-mer seeds in positions 2–8 of the mature miRNA. The novel miRNAs were discovered by their structural and evolutionary signatures, rather than by their (sequence) similarity to existing miRNAs, and this is reflected in the many unique seed sequences they introduce. Most novel validated miRNAs ($n = 28$) have unique 5′ ends, an additional six form new families, and only three extend existing families. Thus, although the novel miRNAs introduce 50% more miRNA genes (from 61 to 94), they result in an even >70% increase in the number of families (from 43 to 73).

Second, although the novel miRNAs introduce 3300 new target sites, these are heavily biased toward genes already targeted by miRNAs. On one hand, the new miRNAs target known target categories, such as developmental genes, and exclude anti-target categories, such as ribosomal genes (Stark et al. 2005). On the other hand, they show great overlap with individual genes already targeted by miRNAs, leading to a denser regulatory network. It is already known that the existing miRNAs network is dense, with most target genes targeted by multiple miRNAs (Enright et al. 2003; Grun et al. 2005; Krek et al. 2005; Lewis et al. 2005; Stark et al. 2005). The novel miRNAs reinforce this dense network, leading to >3000 new sites, but only a small number of novel target genes. Overall, the number of miRNA-targeted genes increases only by a relatively modest 16%, from 3000 genes (21% of the genome) to 3500 genes (25% of the genome; note that the accompanying paper by Ruby et al. [2007] extends this analysis with an enlarged set of novel *Drosophila* miRNAs).

This effect is not additive, but it is combinatorial. Since the new miRNA genes introduce many new miRNA families, they introduce new sets of target genes, which can be differentially regulated in response to developmental and environmental

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

stimuli; these are distinct from previous sets of targets and cutting across them. The result is a significantly higher potential for combinatorial regulation, giving opportunity for the precise regulation of individual subsets of genes using different combinations of miRNAs, resulting in a much more versatile and precise network of miRNA regulation.

A single miRNA gene can lead to multiple functional mature products

Our ability to score structural and evolutionary properties of the novel miRNAs, and correlate them with sequencing reads, led to new insights into miRNA biogenesis and function. Interestingly, when our algorithm did not predict a precise miRNA 5′ end, the sequencing reads also showed several alternate mature products, with different 5′ ends (Fig. 3D). For example, the predicted alternate 5′ end of miR-964, which is shifted by 1 nt, is supported by more than half of the reads supporting the annotated 5′ end (3461 vs. 6393 in ovaries/testes and 29 vs. 57 otherwise). Overall, when our algorithm predicted the correct 5′ end, 90% of reads supported that start; when our algorithm did not predict the correct start, only 78% of reads supported it ($P = 6 \times 10^{-3}$), and a significant number of hairpins were in fact processed into alternate mature products. Interestingly, more highly expressed miRNAs showed overall more accurate processing (Pearson correlation: 0.39), suggesting that inaccuracies may be tolerated for less abundant miRNAs (or young miRNAs, which are not yet evolutionarily fixed; Chen and Rajewsky 2007), while highly expressed miRNAs are under strong selection for accurate processing. Overall, our results suggest that, when the evolutionary signal is mixed, the signals for processing are also less accurate. It is currently unclear if the alternate forms are regulated and functional, or arise solely due to processing inaccuracies.

Similarly, in several cases, our algorithm yielded high scores for the miRNA* arm, suggesting these may also have functional targets. While in most cases, the miRNA* arm showed no significant signal (Fig. 3B), we found that, for 10 miRNAs, the 5′ end of the star sequence scored as highly as some miRNAs. Since our score is based on 7-mer conservation and avoidance of target sites, these high scores suggest that the miRNA* may be functional. In fact, in four cases the miRNA* scored even more highly than the primary arm: Two of these star products showed convincing similarity to known miRNAs (miR-5 and miR-4; Lai et al. 2004), likely to contribute to the high scores. However, seven of the 10 have unique 5′ ends such that high 7-mer scores cannot be attributed to a known miRNA, suggesting a distinct role of the star arm in targeting 3′ UTRs.

The sequencing reads strongly support these findings. The 10 miRNAs with high-scoring star sequences also showed abundant reads from the star arm. For example, miR-5* is supported by 1142 reads, far exceeding the abundance of many miRNAs and accounting for greater than one-fourth of the mature reads obtained in the *mir-5* locus (Fig. 3D). On average, the number of reads increases from 19 to 92 and the fraction of star reads among all reads of the hairpin increases from 2% to 9%, suggesting a specifically increased preference of the star arm (Fig. 3D). The abundance of star sequences alone could be explained by RISC incorporation of star arms, due to the thermodynamic stability of the miRNA–miRNA* duplex (Khvorova et al. 2003; Schwarz et al. 2003), and thus star sequences have not received much attention in miRNA target prediction (Lai 2004; Rajewsky 2006). However, the observation that abundant miRNA star sequences also show strong signals stemming from their interaction with target gene 3′ UTRs suggests they may in fact have physiologically relevant targets.

mir–10 as a master Hox regulator

The miRNA *mir-10*, which lies in the Hox cluster, presents a particularly striking case of miRNA* functionality (Fig. 4). On one hand, the miRNA* sequence for *mir-10* shows an even higher score than the mature miRNA itself, in terms of structure, conservation, and also 7-mer motif conservation and avoidance (Fig. 4A). Moreover, we found 34 times as many products of miR-10* than miR-10 in our reads from ovaries and testes (306 vs. 9; Fig. 4B) and this trend also holds for other tissues (1319 vs. 189; Ruby et al. 2007). Further, an independent analysis of miRNA–miRNA* duplex energy showed that both miR-10 and miR-10* can be detected and suggested that both are likely incorporated into RISC (Schwarz et al. 2003). We thus reason that miR-10* may be functional and is in fact likely to represent the primary product.

To infer the potential functional role of miR-10*, we studied its target spectrum, revealing striking insights. First, miR-10* has nearly 10 times more targets than the annotated miR-10, suggesting it plays a major role in gene regulation (Fig. 4C). Moreover, these targets include several Hox genes, in particular *Abdominal-B* (*Abd-B*) and *Ultrabithorax* (*Ubx*), which show highly conserved canonical target sites (Fig. 4D). This provides another example of regulatory relationships between Hox miRNAs and Hox genes, highly reminiscent of two additional Hox miRNAs, *mir-iab-4* in *Drosophila*, and *mir-196* in vertebrates (Stark et al. 2003; Yekta et al. 2004; Ronshaugen et al. 2005).

Unique to *mir-10*, however, both arms appear to be functional and appear to be targeting Hox genes: The originally annotated miR-10 has a highly conserved compensatory target site in the 3′ UTR of *Sex-combs-reduced* (*Scr*; Brennecke et al. 2005; Fig. 4D). To our knowledge, this is the first demonstration of functional relevance of both miRNA arms, and their potential cooperative action, targeting similar genes.

Overall, our results suggest that *mir-10* may in fact play the role of a master Hox regulator, which was previously missed likely due to its incomplete annotation. The observation that both arms of a miRNA are expressed and have highly conserved targets in similar processes suggests that miRNA-star arms might frequently be functional, with major implications for the identification of functional miRNA targets and overall role and versatility of miRNA regulation.

Discussion

A complete knowledge of all miRNAs is extremely important, especially in *Drosophila*, where the study of mutants from genetic screens heavily relies on accurate genome annotation and where powerful reverse genetics tools allow the systematic analysis of miRNA functions. The recent availability of 12 closely related *Drosophila* genomes enables the use of comparative genomics for recognizing miRNAs systematically, based on their structural and evolutionary signatures. We report >41 novel predicted miRNAs and validate 28 experimentally using sequencing reads from small RNA library sequencing (see Supplemental material; Ruby et al. 2007). In addition, the predictions show several miRNA-like properties, which were not used in their discovery, including genomic clustering, family membership, intronic occurrences, and transcription.

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

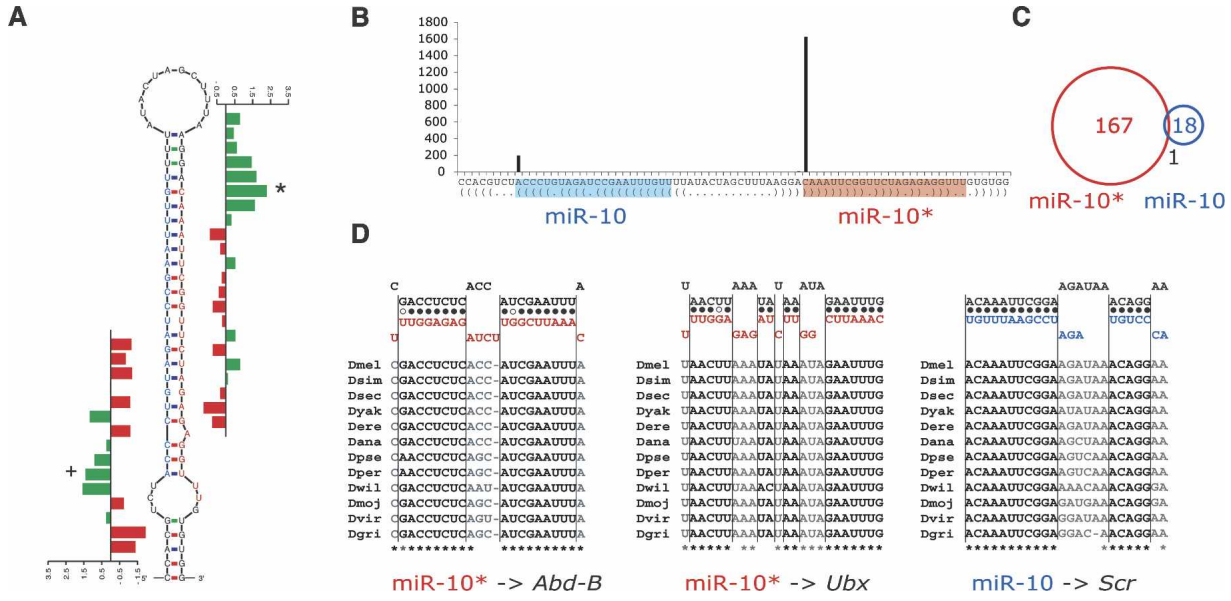


Figure 4. miR-10 and miR-10* target Hox genes. (*A*) SVM *Z*-scores indicate that miR-10 and miR-10* are both functional and that miR-10* is likely the major miRNA (green and red are positive and negative scores, respectively). (*B*) Cloning confirms that both sequences are expressed but that miR-10* is more abundant (306/9 in ovaries/testes and 1319/189 otherwise), consistent with the *Z*-scores. Shown is the number of total Solexa and 454 (Ruby et al. 2007) reads supporting each 5′ end position along the miRNA precursor hairpin (the more abundant miR-10* is shaded in red and miR-10 in blue). (*C*) Target genes of miR-10 and miR-10* are largely different. miR-10* targets 167 genes as compared to the 18 predicted for miR-10, with only one overlapping gene (Venn diagrams). (*D*) miR-10 and miR-10* both target Hox genes. miR-10* has highly conserved canonical sites in the Hox genes *Abdominal-B* (*Abd-B*) and *Ultrabithorax* (*Ubx*). miR-10 has a highly conserved compensatory target site in the Hox gene *Sex combs reduced* (*Scr*; Brennecke et al. 2005). In the absence of sequence-similar miRNAs, these sites argue that both sequences are functioning by regulating Hox genes. This relationship is similar to that of other Hox miRNAs in fly and mammals (Stark et al. 2003; Yekta et al. 2004).

The newly discovered miRNAs significantly increase the number of miRNA genes known in the fly, from 60 prior to this work to 89 experimentally confirmed genes (a 50% increase) and 17 additional candidates. This increase is even more drastic when considering miRNA families: The new miRNAs introduce 30 new miRNA families (a 70% increase), resulting in increased versatility by forming new groups of potentially coregulated target genes. Finally, we found that the genes targeted by these new miRNAs heavily overlap with existing miRNA targets, resulting in a denser miRNA network.

In addition to miRNA annotation, we find that comparative genomics can lead to a deeper understanding of the functions and biogenesis of miRNAs. We show that a single miRNA gene can give rise to several mature sequences: First, both strands can be expressed and correctly processed; second, multiple mature products can be produced at small offsets from the primary miRNA product; third, both arms of the miRNA hairpin can lead to mature miRNAs, with many potential functional targets. These alternate mature forms are sometimes of surprisingly similar abundance, share similarly strong evolutionary and structural signatures, and show similar relationships with target 3′ UTRs. As these alternate forms often have drastically different target gene spectra, they could be employed to increase the number of targets for each miRNA gene and may constitute a powerful evolutionary mechanism for the emergence of new miRNAs.

Particularly striking was the relationship between miRNAs and miRNA stars. For 13 miRNAs, both hairpin arms scored highly, reflecting their interaction with miRNA targets and antitargets and indicating that star arms might have previously overlooked functional roles. In the case of *mir-10*, located in the Hox cluster, we find highly conserved target genes for both arms, targeting multiple distinct genes in the Hox cluster, suggesting *mir-10* is a master regulator of Hox genes, while, previously, Hox miRNAs were assumed to only have a single dominant Hox target gene each (Stark et al. 2003; Yekta et al. 2004; Ronshaugen et al. 2005). As miRNA and miRNA* are coexpressed, the potential of sharing a functional target spectrum is intriguing and has major implications for miRNA targeting and biology.

Conversely, mature products from opposite DNA strands have the potential of distinct regulatory domains. For example, for the *mir-iab-4* locus, we predict high scores for both strands, and both are validated by sequencing reads as expressed and correctly processed. These two strands are known to be transcribed in adjacent and nonoverlapping domains in fly embryos (Bae et al. 2002), suggesting distinct developmental functions for two miRNAs from a single locus. In our study, we find four such examples of opposite-strand high-scoring miRNAs that are confirmed by sequencing reads. The extent of this mechanism is still unclear, but reverse-strand expression may represent a novel biologically relevant regulatory principle for miRNA genes and RNAs more generally.

The methods presented here are general and applicable for the systematic annotation of any species. In particular, this study can serve as a powerful model for the analysis of the human genome: The total evolutionary distance across the 12 flies is comparable to the total evolutionary distance of the mammalian and vertebrate genomes (Stark et al. 2007) and, given the much larger number of 474 annotated human miRNAs (Griffiths-Jones et al. 2006), comparative approaches may reveal many new insights into the biology, regulation, and targeting of human miRNAs. By measuring specific evolutionary constraints, comparative studies can complement ongoing sequencing efforts by rec-

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

ognizing miRNA genes with low abundance or tissue-specific expression, and help indicate if sequencing reads matching hairpin structures constitute noise or functional miRNAs (Berezikov et al. 2006a). More importantly, though, they can reveal biological insights that are not directly accessible from sequencing reads alone. More generally outside the realm of miRNAs, the ability to define precise evolutionary and structural signatures for specific classes of RNA genes represents a powerful approach for the systematic discovery of functional RNA genes and structures in any species.

Methods

miRNA training sets

From miRBase release 9.0 (Griffiths-Jones et al. 2006), we selected all 60 *D. melanogaster* miRNAs that have been cloned. For both hairpin and mature prediction, miRNAs that may bias the score due to overfitting are excluded when scoring the known miRNAs (see the appropriate section below for details).

Collecting all melanogaster hairpins

To identify miRNA-like hairpins, we ran RNAfold from the Vienna package (Hofacker et al. 1994) on 120-nt windows (overlap of 90 nt) in the *D. melanogaster* genome (rel 4). We considered all hairpins in each window (including branching hairpins) and trimmed them to the end of the stem. We use these folds to infer the arms and loop of each hairpin. As a lenient prescreening, we removed all hairpins <63 nt, with an arm of <20 nt or with <70% arm base-pairing. We were left with all the known hairpins and an additional list of 760,000 potentially overlapping putative miRNA hairpins.

Hairpin sequence alignments

For each *melanogaster* hairpin sequence, we selected the best BLAST (Altschul et al. 1997) match with *E*-value $\leq 1 \times 10^{-5}$ in each of the 11 other genomes (CAF1 assemblies). We performed a multiple alignment of the corresponding sequences plus 50 nt flanking sequence on each side using ClustalW (Thompson et al. 1994).

miRNA hairpin discovery

For each hairpin we derived several structural and conservation features. The most important of these features are summarized in Figure 1E and the complete list of features is available in Supplemental Table 2. We scored the list of 760,355 putative miRNAs with a method similar to Random Forests (Breiman 2001). Using the combined conservation and structural feature set, 500 decision trees were trained on the positive training set of 60 miRNAs and a different randomly selected negative set of 250 of the remaining putative miRNAs. The final score for a hairpin was derived through cross validation (the score of each hairpin is evaluated only by trees that exclude it, all redundant sequence similar miRNAs and overlapping miRNAs). From all hairpins that overlap on the same strand, only the hairpin with the highest score is kept. This can lead to known miRNAs having a slightly revised hairpin selected.

miRNA mature 5′ identification

For each position in the hairpin, we computed several features indicative of the start (5′ end) of mature miRNAs. 7-mer scores are determined for the sequence complementary 7-mers for each position in the hairpin. 7-mer conservation scores are motif-conservation scores (MCS) of 7-mers calculated in all annotated 3′ UTRs (FlyBase release 4.3) as described in Kellis et al. (2003, 2004) and Xie et al. (2005). Additionally, we assessed the avoidance of the 7-mers in 3′ UTRs of global anti-target genes (Stark et al. 2005) by computing the deviation relative to all genes by *Z*-scores. In addition, we considered the nucleotide to account for the U-bias often observed in mature miRNAs, the number of paired bases in a window of 7 around the position, and others (see Supplemental Table 2). We excluded potential start positions for which the corresponding miRNA would fall outside the hairpin or span the hairpin loop region (for positions in the left arm, we required at least 15 nt before the start of the loop; for positions in the right arm, we allowed no more than a 3-nt overlap with the loop and required at least 18 nt before the end of the hairpin). Within each hairpin, we linearly normalized each feature to be from 0 to 1 and marked each known mature site as a positive and all remaining sites as negative. We augmented the features for each position with the features of the position of the left and the right. We used the SVMlight (Joachims 1999) package to train an SVM with default parameters (linear kernel and positive gain 1) on all the permissible locations from all the known hairpins. The SVM scores for each hairpin are linearly normalized so that the scores of the permissible regions have mean 0 and standard deviation 1. We predict the mature location by taking the permissible location in each hairpin with the highest SVM score. Each hairpin is only scored by models trained on cloned Rfam hairpins (Rfam 9.0), excluding itself and all family members. For evaluation (not training), we use the partly corrected 5′ end annotation of Ruby et al. (2007). To test if we predicted the star sequences, we determined the star sequence based on the fold-back structure as a 2-nt 5′ overhang of the mature miRNA sequence.

Validation of novel miRNAs

To validate our predictions experimentally, we obtained 763,111 Solexa reads corresponding to 1524 distinct sequences that matched to our predictions. These were cloned from adult *Drosophila* ovaries and testes as described previously (Brennecke et al. 2007). We excluded short reads (<15 nt) and those that matched the genome more than three times and aligned the remaining reads to the predicted hairpins. For validation, we required that at least one position in the hairpin was supported by at least 10 reads and manually inspected the alignments for miRNA-like processing patterns (e.g., dominant sequence, presence of star sequence, no sign of degradation products). In addition, we intersected the predicted hairpins with curated sequencing reads of several *Drosophila* libraries kindly provided by Graham Ruby, David Bartel, and Eric Lai. To validate mature miRNA 5′ end predictions, we used the curated mature miRNAs reported by Ruby et al. (2007) and refer to them by their newly assigned Rfam names (Griffiths-Jones et al. 2006). In Figures 2–4, we report the sum of reads from Solexa and 454 sequencing when showing individual miRNAs. However, given the large number of Solexa reads stemming from only ovaries and testes, we report only the 454 read-count when comparing several miRNAs.

miRNA recovery using different species sets

We investigated the dependency of genome-wide miRNA discovery on the number and evolutionary distance of the contributing species. For this, we obtained all novel miRNAs defined by Ruby et al. (2007) and tested how many we recovered with our protocol when using selected subsets of species. In each case, we allowed for the optimal re-weighting of feature contributions (e.g., to down-weight conservation features when comparing only close species, if appropriate).

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

Acknowledgments

We thank Graham Ruby and David Bartel (Whitehead Institute/HHMI, MA) and Eric Lai (Sloan-Kettering Institute, NY) for providing access to their sequencing data prior to publication and for helpful comments on the final manuscript. We thank Matt Rasmussen, Mike Lin (CSAIL, Broad Institute, MA), and other members of the Kellis lab for helpful discussions and for sharing unpublished data. A.S. and J.B. thank the Schering AG/Ernst Schering Foundation for postdoctoral fellowships. P.K. was supported in part by a National Science Foundation Graduate Research Fellowship. L.P. thanks Jaak Vilo (University of Tartu, Estonia) for support and helpful discussions.

References

Aboobaker, A.A., Tomancak, P., Patel, N., Rubin, G.M., and Lai, E.C. 2005. *Drosophila* microRNAs exhibit diverse spatial expression patterns during embryonic development. *Proc. Natl. Acad. Sci.* **102:** 18017–18022.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Alvarez-Garcia, I. and Miska, E.A. 2005. MicroRNA functions in animal development and human disease. *Development* **132:** 4653–4662.

Ambros, V. 2004. The functions of animal microRNAs. *Nature* **431:** 350–355.

Aravin, A.A., Lagos-Quintana, M., Yalcin, A., Zavolan, M., Marks, D., Snyder, B., Gaasterland, T., Meyer, J., and Tuschl, T. 2003. The small RNA profile during *Drosophila melanogaster* development. *Dev. Cell* **5:** 337–350.

Bae, E., Calhoun, V.C., Levine, M., Lewis, E.B., and Drewell, R.A. 2002. Characterization of the intergenic RNA profile at *abdominal-A* and *Abdominal-B* in the *Drosophila* bithorax complex. *Proc. Natl. Acad. Sci.* **99:** 16847–16852.

Bartel, D.P. 2004. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* **116:** 281–297.

Baskerville, S. and Bartel, D.P. 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* **11:** 241–247.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., et al. 2005a. Identification of hundreds of conserved and nonconserved human microRNAs. *Nat. Genet.* **37:** 766–770.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., et al. 2005b. Identification of hundreds of conserved and nonconserved human microRNAs. *Nat. Genet.* **37:** 766–770.

Berezikov, E., Guryev, V., van de Belt, J., Wienholds, E., Plasterk, R.H., and Cuppen, E. 2005. Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120:** 21–24.

Berezikov, E., Cuppen, E., and Plasterk, R.H. 2006a. Approaches to microRNA discovery. *Nat. Genet.* **38:** S2–S7.

Berezikov, E., Thuemmler, F., van Laake, L.W., Kondova, I., Bontrop, R., Cuppen, E., and Plasterk, R.H. 2006b. Diversity of microRNAs in human and chimpanzee brain. *Nat. Genet.* **38:** 1375–1377.

Breiman, L. 2001. Random forests. *Mach. Learn.* **45:** 5–32.

Brennecke, J., Hipfner, D.R., Stark, A., Russell, R.B., and Cohen, S.M. 2003. *bantam* encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in *Drosophila*. *Cell* **113:** 25–36.

Brennecke, J., Stark, A., Russell, R.B., and Cohen, S.M. 2005. Principles of microRNA-target recognition. *PLoS Biol.* **3:** e85. doi: 10.1371/journal.pbio.0030085.

Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G.J. 2007. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **128:** 1089–1103.

Chen, K. and Rajewsky, N. 2007. The evolution of gene regulation by transcription factors and microRNAs. *Nat. Rev. Genet.* **8:** 93–103.

Davidson, E.H. and Erwin, D.H. 2006. Gene regulatory networks and the evolution of animal body plans. *Science* **311:** 796–800.

Doench, J.G. and Sharp, P.A. 2004. Specificity of microRNA target selection in translational repression. *Genes & Dev.* **18:** 504–511.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* (in press) doi: 10.1038/nature06341.

Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D.S. 2003. MicroRNA targets in *Drosophila*. *Genome Biol.* **5:** R1. doi: 10.1186/gb-2003-5-1-r1.

Farh, K.K., Grimson, A., Jan, C., Lewis, B.P., Johnston, W.K., Lim, L.P., Burge, C.B., and Bartel, D.P. 2005. The widespread impact of mammalian microRNAs on mRNA repression and evolution. *Science* **310:** 1817–1821.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., and Enright, A.J. 2006. miRBase: MicroRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **34:** D140–D144. doi: 10.1093/nar/gkj112.

Grun, D., Wang, Y.L., Langenberger, D., Gunsalus, K.C., and Rajewsky, N. 2005. microRNA target predictions across seven *Drosophila* species and comparison to mammalian targets. *PLoS Comput. Biol.* **1:** e13. doi: 10.1371/journal.pcbi.0010013.

Han, J., Lee, Y., Yeom, K.H., Nam, J.W., Heo, I., Rhee, J.K., Sohn, S.Y., Cho, Y., Zhang, B.T., and Kim, V.N. 2006. Molecular basis for the recognition of primary microRNAs by the Drosha–DGCR8 complex. *Cell* **125:** 887–901.

Hobert, O. 2004. Common logic of transcription factor and microRNA action. *Trends Biochem. Sci.* **29:** 462–468.

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M., and Schuster, P. 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie / Chemical Monthly* **125:** 167–188.

Hofacker, I.L., Fekete, M., and Stadler, P.F. 2002. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* **319:** 1059–1066.

Joachims, T. 1999. Making large-scale SVM learning practical. In *Advances in kernel methods—Support vector learning* (eds. B. Schölkopf et al.), pp. 41–56. MIT Press, Cambridge, MA.

Johnston, R.J. and Hobert, O. 2003. A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. *Nature* **426:** 845–849.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241–254.

Kellis, M., Patterson, N., Birren, B., Berger, B., and Lander, E.S. 2004. Methods in comparative genomics: Genome correspondence, gene identification and regulatory motif discovery. *J. Comput. Biol.* **11:** 319–355.

Khvorova, A., Reynolds, A., and Jayasena, S.D. 2003. Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115:** 209–216.

Kiriakidou, M., Nelson, P.T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z., and Hatzigeorgiou, A. 2004. A combined computational-experimental approach predicts human microRNA targets. *Genes & Dev.* **18:** 1165–1178.

Kloosterman, W.P., Wienholds, E., Ketting, R.F., and Plasterk, R.H. 2004. Substrate requirements for *let-7* function in the developing zebrafish embryo. *Nucleic Acids Res.* **32:** 6284–6291. doi: 10.1093/nar/gkh968.

Krek, A., Grun, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., Macmenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M., et al. 2005. Combinatorial microRNA target predictions. *Nat. Genet.* **37:** 495–500.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. 2001. Identification of novel genes coding for small expressed RNAs. *Science* **294:** 853–858.

Lai, E.C. 2002. Micro RNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat. Genet.* **30:** 363–364.

Lai, E.C. 2003. microRNAs: Runts of the genome assert themselves. *Curr. Biol.* **13:** R925–R936.

Lai, E.C. 2004. Predicting and validating microRNA targets. *Genome Biol.* **5:** 115. doi: 10.1186/gb-2004-5-9-115.

Lai, E.C. and Posakony, J.W. 1997. The Bearded box, a novel 3′ UTR sequence motif, mediates negative post-transcriptional regulation of Bearded and Enhancer of split Complex gene expression. *Development* **124:** 4847–4856.

Lai, E.C., Burks, C., and Posakony, J.W. 1998. The K box, a conserved 3′ UTR sequence motif, negatively regulates accumulation of enhancer of split complex transcripts. *Development* **125:** 4077–4088.

Lai, E.C., Tomancak, P., Williams, R.W., and Rubin, G.M. 2003. Computational identification of *Drosophila* microRNA genes. *Genome Biol.* **4:** R42. doi: 10.1186/gb-2003-4-7-r42.

Lai, E.C., Wiel, C., and Rubin, G.M. 2004. Complementary miRNA pairs suggest a regulatory role for miRNA:miRNA duplexes. *RNA* **10:** 171–175.

Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294:** 858–862.

Lee, R.C. and Ambros, V. 2001. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* **294:** 862–864.

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

Lee, R.C., Feinbaum, R.L., and Ambros, V. 1993. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **75:** 843–854.
Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P., and Burge, C.B. 2003. Prediction of mammalian microRNA targets. *Cell* **115:** 787–798.
Lewis, B.P., Burge, C.B., and Bartel, D.P. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120:** 15–20.
Li, X. and Carthew, R.W. 2005. A microRNA mediates EGF receptor signaling and promotes photoreceptor differentiation in the *Drosophila* eye. *Cell* **123:** 1267–1277.
Li, Y., Wang, F., Lee, J.A., and Gao, F.B. 2006. MicroRNA-9a ensures the precise specification of sensory organ precursors in *Drosophila*. *Genes & Dev.* **20:** 2793–2805.
Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B., and Bartel, D.P. 2003a. Vertebrate microRNA genes. *Science* **299:** 1540.
Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B., and Bartel, D.P. 2003b. The microRNAs of *Caenorhabditis elegans*. *Genes & Dev.* **17:** 991–1008.
Lin, M.F., Carlson, J.W., Crosby, M.A., Matthews, B.B., Yu, C., Park, S., Wan, K.H., Schroeder, A.J., Gramates, L.S., St. Pierre, S.E., et al. 2007. Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes. *Genome Res.* (this issue) 10.1101/gr.6679507.
Manak, J.R., Dike, S., Sementchenko, V., Kapranov, P., Biemar, F., Long, J., Cheng, J., Bell, I., Ghosh, S., Piccolboni, A., et al. 2006. Biological function of unannotated transcription during the early development of *Drosophila melanogaster*. *Nat. Genet.* **38:** 1151–1158.
Miller, W., Makova, K.D., Nekrutenko, A., and Hardison, R.C. 2004. Comparative genomics. *Annu. Rev. Genomics Hum. Genet.* **5:** 15–56.
Mitchell, T.M. 1997. *Machine learning*. McGraw-Hill, New York.
Ng Kwang Loong, S. and Mishra, S.K. 2007. Unique folding of precursor microRNAs: Quantitative evidence and implications for de novo identification. *RNA* **13:** 170–187.
Okabe, M., Imai, T., Kurusu, M., Hiromi, Y., and Okano, H. 2001. Translational repression determines a neuronal potential in *Drosophila* asymmetric cell division. *Nature* **411:** 94–98.
Pasquinelli, A.E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M.I., Maller, B., Hayward, D.C., Ball, E.E., Degnan, B., Muller, P., et al. 2000. Conservation of the sequence and temporal expression of *let-7* heterochronic regulatory RNA. *Nature* **408:** 86–89.
Pekarsky, Y., Santanam, U., Cimmino, A., Palamarchuk, A., Efanov, A., Maximov, V., Volinia, S., Alder, H., Liu, C.G., Rassenti, L., et al. 2006. Tcl1 expression in chronic lymphocytic leukemia is regulated by *miR-29* and *miR-181*. *Cancer Res.* **66:** 11590–11593.
Plasterk, R.H. 2006. Micro RNAs in animal development. *Cell* **124:** 877–881.
Rajewsky, N. 2006. microRNA target predictions in animals. *Nat. Genet.* **38:** S8–S13.
Richards, S., Liu, Y., Bettencourt, B.R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M.J., Chen, R., Meisel, R.P., et al. 2005. Comparative genome sequencing of *Drosophila* pseudoobscura: Chromosomal, gene, and *cis*-element evolution. *Genome Res.* **15:** 1–18.
Ritchie, W., Legendre, M., and Gautheret, D. 2007. RNA stem-loops: To be or not to be cleaved by RNAse III. *RNA* **13:** 457–462.
Ronshaugen, M., Biemar, F., Piel, J., Levine, M., and Lai, E.C. 2005. The *Drosophila* microRNA iab-4 causes a dominant homeotic transformation of halteres to wings. *Genes & Dev.* **19:** 2947–2952.
Ruby, J.G., Jan, C., Player, C., Axtell, M.J., Lee, W., Nusbaum, C., Ge, H., and Bartel, D.P. 2006. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* **127:** 1193–1207.
Ruby, J.G., Stark, A., Johnston, W.K., Kellis, M., Bartel, D.P., and Lai, E.C. 2007. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of *Drosophila* microRNAs. *Genome Res.* (this issue). doi: 10.1101/gr.6597907.
Schwarz, D.S., Hutvagner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P.D. 2003. Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115:** 199–208.
Seitz, H. and Zamore, P.D. 2006. Rethinking the microprocessor. *Cell* **125:** 827–829.
Stark, A., Brennecke, J., Russell, R.B., and Cohen, S.M. 2003. Identification of *Drosophila* microRNA targets. *PLoS Biol.* **1:** e60. doi: 10.1371/journal.pbio.0000060.
Stark, A., Brennecke, J., Bushati, N., Russell, R.B., and Cohen, S.M. 2005. Animal microRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. *Cell* **123:** 1133–1146.
Stark, A., Lin, M.F., Kheradpour, P., Pedersen, J.S., Parts, L., Carlson, J.W., Crosby, M.A., Rasmussen, M.D., Roy, S., Deoras, A.N., et al. 2007. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* (in press) doi: 10.1038/nature06340.
Teleman, A.A. and Cohen, S.M. 2006. *Drosophila* lacking microRNA miR-278 are defective in energy homeostasis. *Genes & Dev.* **20:** 417–422.
Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680. doi: 10.1093/nar/22.22.4673.
Valencia-Sanchez, M.A., Liu, J., Hannon, G.J., and Parker, R. 2006. Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes & Dev.* **20:** 515–524.
Washietl, S., Hofacker, I.L., and Stadler, P.F. 2005. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci.* **102:** 2454–2459.
Wightman, B., Ha, I., and Ruvkun, G. 1993. Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell* **75:** 855–862.
Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. 2005. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434:** 338–345.
Yekta, S., Shih, I.H., and Bartel, D.P. 2004. MicroRNA-directed cleavage of HOXB8 mRNA. *Science* **304:** 594–596.
Zamore, P.D. and Haley, B. 2005. Ribo-gnome: The big world of small RNAs. *Science* **309:** 1519–1524.
Zhang, R., Peng, Y., Wang, W., and Su, B. 2007. Rapid evolution of an X-linked microRNA cluster in primates. *Genome Res.* **17:** 612–617.

Received April 10, 2007; accepted in revised form June 18, 2007.

Reliable prediction of regulator targets using 12 *Drosophila* genomes

Pouya Kheradpour,[1,4] Alexander Stark,[1,2,4,5] Sushmita Roy,[3] and Manolis Kellis[1,2,5]

[1]*Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA;* [2]*Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02141, USA;* [3]*Department of Computer Science, University of New Mexico, Albuquerque, New Mexico 87131, USA*

Gene expression is regulated pre- and post-transcriptionally via *cis*-regulatory DNA and RNA motifs. Identification of individual functional instances of such motifs in genome sequences is a major goal for inferring regulatory networks yet has been hampered due to the motifs' short lengths that lead to many chance matches and poor signal-to-noise ratios. In this paper, we develop a general methodology for the comparative identification of functional motif instances across many related species, using a phylogenetic framework that accounts for the evolutionary relationships between species, allows for motif movements, and is robust against missing data due to artifacts in sequencing, assembly, or alignment. We also provide a robust statistical framework for evaluating motif confidence, which enables us to translate evolutionary conservation into a confidence measure for each motif instance, correcting for varying motif length, composition, and background conservation of the target regions. We predict targets of fly transcription factors and miRNAs in alignments of 12 recently sequenced *Drosophila* species. When compared to extensive genome-wide experimental data, predicted targets are of high quality, matching and surpassing ChIP-chip microarrays and recovering miRNA targets with high sensitivity. The resulting regulatory network suggests significant redundancy between pre- and post-transcriptional regulation of gene expression.

[Supplemental material is available online at www.genome.org. All data and predicted transcription factor and miRNA targets are freely available at http://compbio.mit.edu/fly/motif-instances/.]

Understanding gene expression and its regulation in response to developmental and environmental stimuli is one of the greatest challenges of modern biology. Regulatory control of gene expression occurs at many levels, both pre- and post-transcriptionally, generally based on short DNA and RNA signals known as regulatory motifs. These are recognized in a sequence-specific way by diverse protein and RNA regulators to direct transcription initiation, mRNA export, stability, and translation, ultimately leading to diverse gene-regulatory programs in organogenesis and development, and in response to environmental stimuli.

The sequence-based nature of regulatory control should in principle enable computational identification of regulator targets, by recognizing individual motif instances that constitute functional binding sites. However, due to their short lengths, motifs match very frequently to the genome or in fact any (random) nucleotide sequence by chance alone, and the majority of genome-wide motif occurrences do not lead to functional regulator binding, being either occluded by chromatin structure, separated from necessary cofactor motifs, or otherwise nonconsequential to transcriptional regulation (Wasserman and Sandelin 2004). To address the large signal-to-noise problem and predict functional regulatory elements, previous computational approaches have sought regions of motif clustering across several cooperating motifs, which are often associated with enhancer function (Berman et al. 2002; Markstein et al. 2004; Schroeder et al. 2004; Philippakis et al. 2006). Although these approaches have been successful in identifying novel enhancers, which are functional when tested in vivo, they only identify a small subset of all functional targets of each regulator and are only applicable when the specific combinations of factors are already known. In particular, they are unable to identify individual motif instances when these act in isolation or with diverse sets of cofactors.

Comparative genomics provides a general methodology for distinguishing functional regulatory motif instances, as biologically meaningful elements are typically under negative selection during evolution, with the type and extent of evolutionary conservation generally reflecting the specific requirements of the selected function (Ureta-Vidal et al. 2003; Miller et al. 2004). As closely related species often share substantial parts of their morphology and developmental programs, the expression of important genes, their regulatory connections, and the underlying regulatory elements are also likely conserved. In fact, some gene-regulatory network kernels involved in organogenesis, such as heart specification, are conserved in species as distant as flies and vertebrates (Davidson and Erwin 2006). Thus, although some processes are subject to more rapid divergence or positive selection (e.g., body color and pigmentation [Prud'homme et al. 2006]), this suggests that comparative genomics at a range of evolutionary distances should allow for the identification of many regulatory components of gene expression programs.

Indeed, previous comparative genomics studies have used the conservation of regulatory elements for the de novo discovery of regulatory motifs across related species (Cliften et al. 2003; Kellis et al. 2003; Chan et al. 2005; Ettwiller et al. 2005; Xie et al. 2005). These studies have relied on the average conservation of thousands of motif instances for each regulator, leading to a high genome-wide signal for motif discovery. However, it has re-

[4]These authors contributed equally to this work.
[5]Corresponding authors.
E-mail manoli@mit.edu; fax (617) 253-7512.
E-mail alex.stark@mit.edu; fax (617) 253-7512.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.7090407. Freely available online through the *Genome Research* Open Access option.

mained unclear what fraction of conserved motif instances were functional and what fraction of functional instances were conserved, namely, whether in fact comparative genomics is applicable for high-specificity and high-sensitivity identification of individual motif instances. Moreover, the available genomes have been either too few for sufficient neutral divergence or too distantly related for motif instances to be conserved (e.g., Cooper et al. 2005; Ettwiller et al. 2005). Accurate motif instance identification would thus require many closely related species, which also present novel conceptual and methodological challenges, with respect to sequence coverage, alignment accuracy, and motif movement, gain, and loss (Boffelli et al. 2003; Margulies et al. 2003, 2007; Thomas et al. 2003; Cooper et al. 2005; Eddy 2005).

Methods such as phylogenetic footprinting, evolutionary rate profiling, and phylogenetic hidden Markov models (HMMs) have been successfully used to identify genomic regions under evolutionary selection (Wasserman et al. 2000; Margulies et al. 2003, 2007; Cooper et al. 2005; Siepel et al. 2005), but they cannot determine the regions' functions that are selected for. Similar to more complex models of motif evolution (e.g., Moses et al. 2004; Zhou and Wong 2004), such methods are often restricted to regions that are well aligned and can be sensitive to motif movements or errors in sequencing, assembly, or alignment (Moses et al. 2004; Margulies et al. 2007). Further, methods to predict genomic regions with regulatory potential generally do not allow identification of regulatory targets for individual factors or miRNAs (Elnitski et al. 2003; Taylor et al. 2006). Lastly, the comparative prediction of miRNA binding sites in 3′ UTRs proved successful (for reviews, see Lai 2004; Rajewsky 2006) but has relied on site presence in defined sets of informant species, and a severe loss of sensitivity has been observed when the number of informant species was increased (Lewis et al. 2003; Grun et al. 2005; Stark et al. 2005).

In this paper, we develop a general methodology for identifying functional motif instances based on their evolutionary conservation across many related species and provide a robust statistical framework for evaluating motif confidence, enabling us to achieve both high sensitivity and high specificity. Our approach uses a phylogenetic framework, which allows for motif movements and local alignment inaccuracies and is robust against missing data due to artifacts in sequencing, assembly, or alignment. Our statistical framework enables us to translate evolutionary conservation into a confidence measure for each motif instance, correcting for varying motif length, composition, and background conservation of the target regions.

We apply our framework to whole-genome alignments of 12 recently sequenced *Drosophila* species (Drosophila 12 Genomes Consortium 2007; Stark et al. 2007) and predict targets of 83 transcription factors (TFs) and 78 miRNAs (57 distinct families), leading to 46,525 regulatory connections. We use genome-wide ChIP-chip experiments and direct tests of TF or miRNA targeting (independently published by us [Stark et al. 2005; Zeitlinger et al. 2007] and others [Abrams and Andrew 2005; Sandmann et al. 2006, 2007; Sethupathy et al. 2006]) to show that computationally predicted regulator targets are of very high quality, matching and surpassing ChIP-chip sensitivity and specificity, and can identify seemingly functional instances even when these are not bound in the conditions experimentally surveyed. Lastly, we study properties of the resulting network, which suggest significant redundancy between pre- and post-transcriptional regulation.

Assessing motif-instance conservation across many genomes

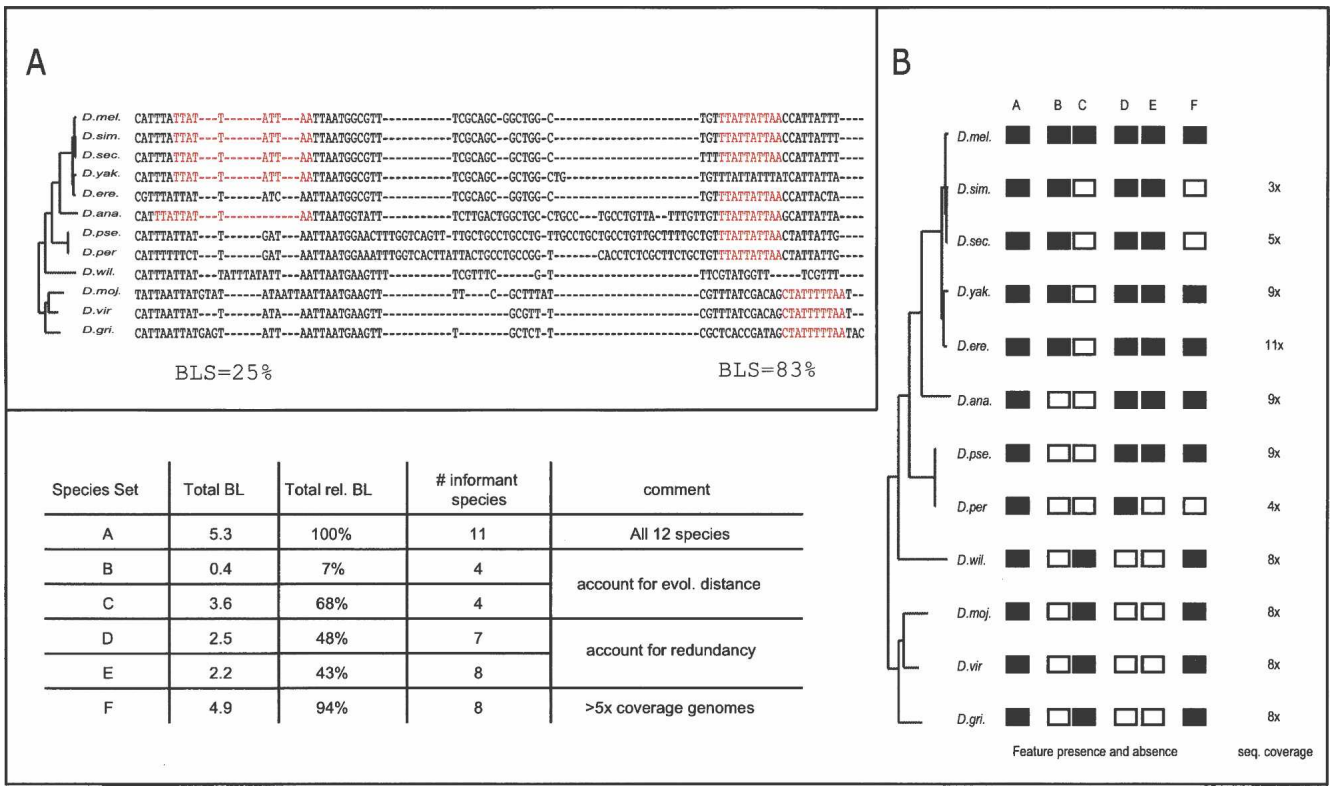
Unlike protein-coding and RNA genes, which are typically well aligned in the multiple sequence alignments of related species, many regulatory motifs are too short to guide alignment algorithms and thus may not appear at orthologous positions in multiple sequence alignments (Wray et al. 2003; Wasserman and Sandelin 2004). As motifs can act at a wide range of distances, individual motif instances may move, either by insertions and deletions, or by "birth" of new motifs and loss of old motifs via compensatory mutational changes (Ludwig et al. 2000). In addition, individual instances of regulatory motifs may actually diverge across different species, and may experience duplication, gain, and loss across the evolutionary tree (Ludwig et al. 2005; Prud'homme et al. 2006; McGregor et al. 2007). Lastly, comparison of many species introduces artifacts due to sequencing, assembly, and alignment, which may affect the alignment of equivalent regulatory motif instances (see Supplemental Fig. S1; Margulies et al. 2007).

To account for these unique evolutionary and alignment properties of regulatory motifs, we developed a phylogenetic framework for motif instance identification which tolerates motif movement and loss, while recognizing their clear selective pressure across the phylogenetic tree. Briefly, we search for motif instances in each of the aligned genomes and, given the set of species that contain motif instances within tolerable distances of the *D. melanogaster* instance, we evaluate the total evolutionary branch length over which the motif appears conserved. The overall score of a motif instance becomes this total branch length of the phylogenetic tree over which the motif is conserved, which we call the Branch Length Score, or BLS (Fig. 1). We thus implicitly assume that all motif instances in *D. melanogaster* are potentially ancestral and count instances in the informant species as evidence when they are conserved. We do not interpret presence/absence patterns of motif instances as evolutionary gain- and loss events, as they could arise from artifacts in sequencing or alignment. The BLS value of a given motif instance ranges from BLS = 0.0 (nonconserved) to BLS = 1.0 (fully conserved), representing the fraction of the total phylogenetic tree covered by the species containing the motif.

This BLS conservation measure has many attractive properties, which enable us to define the conservation level of motif instances across a complete genome, to select conservation thresholds for defining all genome-wide instances of a regulatory motif, and to assign confidence values to the observed conservation, as we describe below. Moreover, because missing instances in the aligned species are not interpreted as evolutionary loss events and are not explicitly penalized, the BLS measure is robust against missing sequence due to low-coverage sequencing, assembly errors, or alignment artifacts. Lastly, BLS provides a direct estimate of the expected neutral divergence of the species compared (Felsenstein 2004), accounting for different divergence times between species and correcting for redundant contributions of individual species in a complex tree and their different rates of divergence (Fig. 1).

Establishing confidence levels for BLS conservation scores

To translate this BLS conservation score to a robust statistic that can be used across different motifs and different types of genomic



Species Set	Total BL	Total rel. BL	# informant species	comment
A	5.3	100%	11	All 12 species
B	0.4	7%	4	account for evol. distance
C	3.6	68%	4	
D	2.5	48%	7	account for redundancy
E	2.2	43%	8	
F	4.9	94%	8	>5x coverage genomes

Figure 1. BLS measure (Branch Length Score) for assessing motif conservation in many genomes. (*A*) Conservation level and corresponding BLS scores for two Mef-2 motif instances. The BLS measure scores the total branch length of the subtree connecting the species with motif instances, as a fraction of the total branch length of all twelve species. As shown in these examples (Mef-2 motif: YTAWWWWTAR), BLS accounts for local alignment inaccuracies, gaps, motif movement, and motif loss. Species abbreviations as follows: *Drosophila melanogaster* (*D. mel.*), *D. simulans* (*D. sim.*), *D. sechellia* (*D. sec.*), *D. yakuba* (*D. yak.*), *D. erecta* (*D. ere.*), *D. ananassae* (*D. ana.*), *D. pseudoobscura* (*D. pse.*), *D. persimilis* (*D. per.*), *D. willistonii* (*D. will.*), *D. mojavensis* (*D. moj.*), *D. virilis* (*D. vir.*), and *D. grimshawii* (*D. gri.*). (*B*) BLS scores for different instance conservation scenarios. Given the pattern of presence (black) and absence (white) within a phylogenetic tree, BLS evaluates the total branch length of the subtree connecting the species that contain the motif: When all species are present, BLS is 100% (column *A*); different sets of species lead to different BLS scores based on their evolutionary distances: distantly related species lead to higher scores as they span larger evolutionary distances (columns *B,C*); species that are very closely related to each other lead to only small incremental contributions, due to their phylogenetic redundancy (columns *D,E*); sequencing, assembly, and alignment artifacts are not penalized, such as those stemming from lower-coverage genomes, as redundancy of branches between close species complements BLS (column *F*). Information about sequence coverage is from Drosophila 12 Genomes Consortium (2007) and Richards et al. (2005).

regions (e.g., promoters, introns, 5′ or 3′ UTRs, etc.), we mapped each BLS score to a confidence value between 0% and 100%, representing the probability that a given motif instance is functional. This probability reflects the increased conservation of motif instances compared to overall sequence similarity and is estimated using control motifs, similar to the signal-to-noise ratio for miRNA target predictions (Lewis et al. 2003). Evaluated in a motif- and region-specific way, it corrects for differences in motif length and composition and for different average conservation levels and nucleotide composition of different genomic regions. Intuitively, longer and highly specific motifs are very unlikely to be conserved by chance and thus result in high confidence levels, even for modest BLS thresholds. Further, regions of overall high conservation (such as protein-coding exons) are likely to contain many conserved motif instances by chance alone and thus require more stringent BLS thresholds to achieve a desired confidence level. Lastly, AT-rich motifs are likely to have many conserved occurrences in AT-rich regions due to chance alone (and GC-rich motifs in GC-rich regions), and thus require higher BLS thresholds if the corresponding control motifs show similarly high conservation.

We found that the number of random motif instances generally decreased rapidly for increasing BLS values, while the number of instances for known motifs remained high (Fig. 2A). For example, at BLS $\geq$ 0.50, the motif for Snail (CAGGTG), has 229 occurrences in promoter regions, compared to 54 motif instances on average for a pool of 10 control motifs. Therefore, we would expect that of these 229 Snail instances, 54 are likely due to chance while 175 of them (76%) are nonrandom, leading to a confidence of $C_{0.5}$ = 76% for each of these motif instances, at BLS $\geq$ 0.5. At a more stringent conservation threshold of BLS $\geq$ 0.70, Snail shows 152 instances while the control motifs show 24 instances on average, leading to a confidence of $C_{0.7}$ = 128/152 = 84%. Similarly, the miRNA K-Box motif (Lai et al. 1998) (CTGTGAT; 5′ seed motif of *Drosophila* miRNAs 2, 6, 11, 13, and 308) reaches confidence values >75% at BLS = 0.4 and >90% at BLS = 0.76 (Fig. 2A). We note that the confidence measure is conservative by nature: 76% confidence for example means that 76% of conserved instances are conserved above background and are thus likely functional. The remaining 24% of conserved instances might contain functional instances that cannot be discerned from noise, suggesting a maximum false-positive rate of up to 24%.

We found that with increasing confidence levels motifs

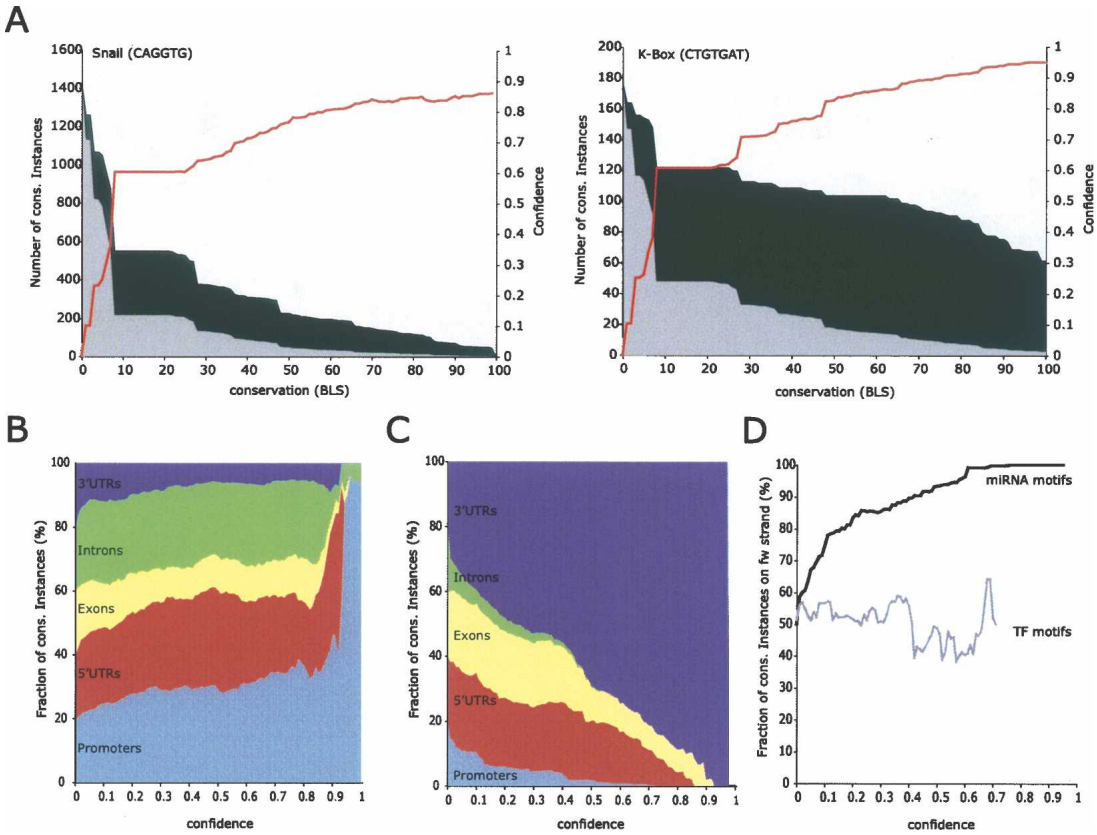


Figure 2. High-confidence recovery of individual motif instances. (*A*) Mapping BLS scores to confidence values. Recovery of conserved motif instances the transcriptional repressor Snail (CAGGTG) in promoter regions (2-kb regions upstream of transcription start sites), and the K-box miRNA (CTGTGAT) in 3′ UTRs, at different BLS cutoffs (*X-axis*). Instances of shuffled control motifs (gray area) decrease much more rapidly than instances of real motifs (height of black curve), leading to a large fraction of motif instances conserved above background (black area). The motif-confidence score (red line) is calculated as the fraction of conserved instances above background. Random motifs are selected to have equal frequency as real motifs at BLS = 0. (*B,C*) Increasing confidence values select functional motif instances. With increasing confidence cutoffs (*X*-axis), transcription factor (TF) motif instances fall increasingly in promoter regions (light blue), 5′ UTRs (red), and introns (green), at the exclusion of 3′ UTRs (dark blue) and coding regions (yellow). In contrast, miRNA motif instances fall increasingly into 3′ UTRs to the exclusion of promoters and other regions. Relative size of regions is normalized at BLS = 0. (*D*) miRNA motif instances at increasing confidence cutoffs are increasingly on the transcribed strand of 3′ UTRs (black curve), while no such trend is seen for TF motifs (gray). Curves are truncated when <10 instances reach the respective confidence.

were predominantly found in regions in which they are known to function. For example, with increasing confidence, the normalized fraction of TF motif instances within promoter regions rises from 20% to 90%, and that of miRNA motif instances within 3′ UTRs from 20% to 100% (Fig. 2B,C). In addition, the percentage of miRNA motif instances on the transcribed strand of 3′ UTRs rises from essentially random (uniform 50%) to exclusively on the transcribed strand (100%), while promoter motifs do not show any strand preference (Fig. 2D). These results illustrate the effectiveness of region-specific confidence values (which require more stringent BLS thresholds for more highly conserved regions), as high-confidence motif instances were not simply biased toward regions with overall high conservation but specifically selected in regions they are known to act.

Effect of allowing motif movements on instance identification

Using confidence cutoffs also allowed us to assess the influence of tolerating motif movements on the recovery of functional motif instances. Allowing for motif movement permits capturing functionally equivalent instances across genomes, independent of their relative positions in the alignment. However, while this approach will always increase the number of conserved instances recovered for real motifs, it also increases the number of spurious motif instances that appear conserved due to increased background conservation for large tolerated movements.

The number of motif instances recovered at a given confidence value presents a robust measure of overall discovery power, as it evaluates sensitivity at a fixed specificity. If the window of tolerated motif movement is too small, many true motif instances will be missed. Conversely, if the window of tolerated motif movement is too large, we would expect both real and control motifs to show increased conservation, thus reducing the confidence and leading to fewer confidently identified instances. Between these two extremes, we would expect the number of high-confidence motif instances to peak for an optimal window of tolerated motif movement, and decrease for lower or higher values.

Indeed, we found that allowing for motif movements of 10–500 nucleotides relative to the *D. melanogaster* instance often increased the number of confident motif instances, while allowing for large movements generally decreased this effect (Fig. 3A). Different window sizes were optimal for different motifs: Longer

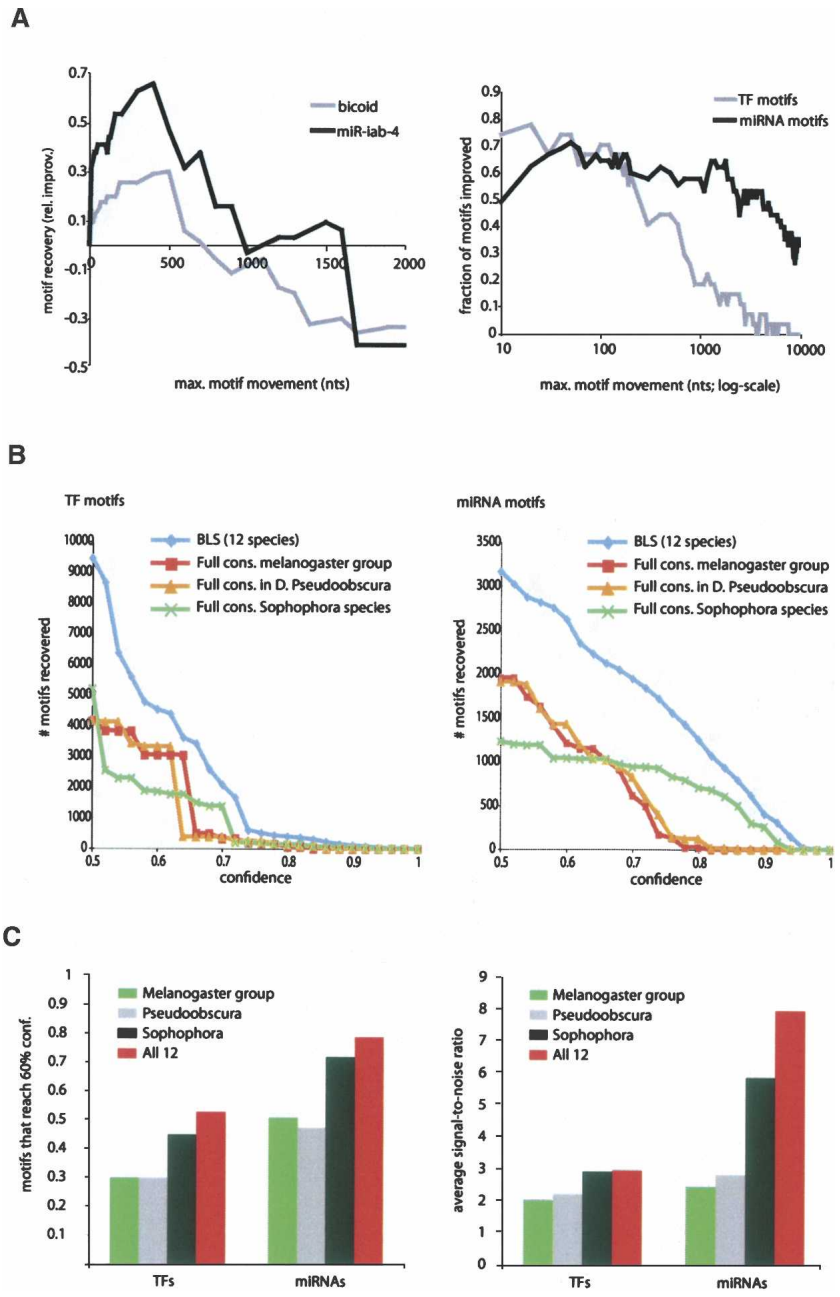


Figure 3. Discovery power for motif instance prediction. (*A*) Effect of tolerated motif movement. Number of recovered motif instances at 60% confidence for TF and miRNA motifs. (*Left* panel) For both TF motifs (gray: *bicoid* motif, VVVBTAATCC) and miRNA motifs (black: miR-iab-4 motif, GTATACG), instance recovery increases until an optimal window size (500 and 400 nucleotides, respectively) and then decreases for larger movements, suggesting that tolerating motif movements increases overall discovery power. (*Right* panel) Performance across all TF motifs (black) and all miRNA motifs (gray) shows improved recovery until windows of 300–500 nucleotides (for 60%–80% of motifs) but reduced performance for larger window sizes. Performance for individual examples (*left* panel) shows a sharper peak than the overall performance across all motifs (*right* panel), as different window sizes are optimal for different motifs. (*B*) BLS measure leads to increased sensitivity. Number of motif instances recovered (*Y*-axis) at each confidence value (*X*-axis) for transcription factor (TF) motifs (*left* panel) and miRNA motifs (*right* panel). The BLS measure applied to the 12 fly genomes (blue) recovers more motif instances at each confidence, as compared to approaches requiring motif presence in all compared species ("full" conservation), applied to the five melanogaster species (red), the pairwise comparison of *D. melanogaster* and *D. pseudoobscura* (yellow), or the nine *Sophophora* species (green). (*C*) Additional species lead to increased specificity. Two measures of discovery power for the BLS measure applied to the five *melanogaster* group species (green), a pairwise comparison of *D. melanogaster* and *D. pseudoobscura* (gray), the nine *Sophophora* species (black), and all 12 *Drosophila* species (red). (*Left* panel) More TF and miRNA motifs reach 60% confidence for increasing number of genomes at larger evolutionary distances. (*Right* panel) Increasing numbers of genomes at larger evolutionary distances also lead to increased signal-to-noise ratio, measured as the conservation level of real motifs vs. control motifs at the most stringent BLS cutoff.

TF motifs with higher information content peaked for longer windows (Pearson correlation 0.40 for information content, 0.33 for length), while motifs with many matches in *D. melanogaster* showed shorter optimal windows (correlation −0.27 for TF, −0.26 for miRNA motifs). We also found a correlation with GC content for miRNA motifs (0.28), as expected since 3′UTRs are AT-rich, but there was little correlation for TF motifs (−0.15). These results illustrate the more rapidly increasing noise levels for motifs with low information content, while motifs with higher information content are less likely to appear by chance within the length of the tolerated window.

Overall, the single best window improved the recovery of 56% of TF motifs (20 nucleotides), and of 71% of miRNA motifs (50 nucleotides; both at 60% confidence). For 71% TF motifs, some window between 10 and 500 nucleotides improved sensitivity, and improvement was substantial for 11% (at 60% confidence; $P \leq 0.05$ after Bonferroni correction to account for testing multiple windows). Similarly, 93% of miRNA motifs showed improved sensitivity, which was substantial for 13%. Improvements were observed over a wide range of confidence cutoffs, showing that tolerating motif movement is important at any desired confidence level for motif instance identification. These results confirm our intuition that indeed, many motif instances are offset considerably in the 12-species alignments, whether due to alignment artifacts or evolutionary plasticity of regulatory motifs.

BLS measure enables increased sensitivity

The confidence measure also enabled us to gauge the sensitivity of the BLS measure, measured as the number of instances recovered at a fixed specificity, compared to different methodological choices. In particular, we asked whether requiring perfect conservation across fewer species (the nine *Sophophora* subgroup species, the four *melanogaster* subgroup species, and *D. pseudoobscura* as the only informant) would lead to higher sensitivity/specificity levels, perhaps due to many lineage-specific motifs.

We found that the BLS measure across all 12 species recovered most instances for all TF and miRNA motifs, at

all confidence levels (Fig. 3B). For TF motifs, our approach recovers more than 1.4-fold more instances than the second most sensitive of the other approaches at 60%, 1.5-fold more at 70%, and threefold more at 80% confidence (for miRNAs motifs, 1.8-fold more at 60%, twofold more at 70%, and 1.8-fold more at 80% confidence). When comparing the three other approaches for confidence thresholds <65%, we found that perfect conservation was indeed more sensitive across the four closely related species in the melanogaster subgroup and in *D. pseudoobscura* compared to perfect conservation across nine *Sophophora* species. However, very few motifs reached higher confidence levels, due to the high overall sequence similarity between these species, resulting in an apparent drop in motif recovery. We also found that the discovery power in *D. pseudoobscura* was comparable to the four melanogaster species, likely due to its position in the phylogenetic tree.

Lastly, the BLS and confidence measures allow us to gauge the effect of additional species. We found that evaluating motif conservation across all 12 species allowed more motifs to reach confidence levels of 60% than was possible with the other species combination and led to higher average signal-to-noise ratios than any other species combination for TFs and miRNAs (Fig. 3C).

These results show that the discovery power for target gene identification continues to increase even with more distantly related species. The usefulness of distant species only becomes effective by the use of the BLS measure, while inclusion of distantly related species resulted in lower performance when perfect conservation was required. Overall, the combination of additional species and a phylogenetic framework for evaluating motif conservation allowed high sensitivity and high specificity in motif-instance identification.

Conserved motif instances identify functional in vivo targets

We then compared our computationally determined conserved motif instances with experimentally determined in vivo targets of known regulators. To define in vivo targets, we used several large-scale experimental datasets: a set of high-confidence direct CrebA targets confirmed with a variety of reporter assays (Abrams and Andrew 2005), three genome-wide chromatin IP (ChIP) experiments for developmental TFs with known motifs (Snail, Mef-2, and Twist) (Sandmann et al. 2006, 2007; Zeitlinger et al. 2007), and a set of experimentally confirmed targets for different miRNAs (Stark et al. 2005; Sethupathy et al. 2006). We note that the experimentally validated miRNA sites were initially predicted based on conservation to *D. pseudoobscura* and thus are biased toward higher conservation (already showing BLS > 0.26). However, the CrebA and the three ChIP data sets were determined independently of any comparative information and thus provide an entirely independent evaluation of our methodology, allowing us to estimate both sensitivity and specificity of our predictions.

For each regulator, we compared motif instances at different confidence cutoffs with the experimentally derived in vivo targets. We found that motif instances at increasing confidence thresholds were strongly enriched for experimentally derived in vivo targets (Fig. 4A). In absence of any comparative information, Mef-2 motif instances in *D. melanogaster* showed no enrichment for experimentally derived targets, while conserved instances showed up to fivefold enrichment (at 60% confidence). Similarly, enrichment rose from threefold to sevenfold for Snail at increasing confidence levels, from fourfold to ninefold for Twist, and from 4.5-fold to 12-fold for CrebA ($P = 4 \times 10^{-11}$, 3×10^{-10}, 2×10^{-6}, and 1×10^{-7} at the highest confidence for the four factors). This illustrates the ability of evolutionary information to select for functional motif occurrences, experimentally shown to be bound and/or functional in vivo. In fact, the enrichment was most pronounced for CrebA (12-fold enrichment; $P = 1.4 \times 10^{-7}$), for which the targets had been shown to be direct transcriptional targets, while some of the ChIP-derived targets may reflect indirect binding or binding that is nonconsequential for transcription.

We also found that even stringent confidence thresholds recovered a large fraction of motif instances in experimentally derived in vivo targets, illustrating the high sensitivity of our approach (Fig. 4B). When ChIP-bound motifs overlapped experimentally defined enhancer elements (Sandmann et al. 2006, 2007; Zeitlinger et al. 2007), 65% Mef-2, 65% Snail, and 25% Twist motif instances were recovered at our 60% confidence cutoff. The lower rate for Twist was possibly due to an overly specific Twist motif used (Markstein et al. 2004). Recovery was again highest for CrebA, for which 76% of motif instances were conserved, illustrating the high sensitivity of comparative genomics methods for validated direct targets (Fig. 4C).

Recovery was much lower when all ChIP-bound regions were considered, regardless of enhancer information, suggesting that some of the ChIP-derived targets may be due to noise and that conservation is able to pinpoint functional enhancers within ChIP-bound regions. Lastly, we recovered 90% of miRNA motif instances in experimentally confirmed targets at 80% confidence (Stark et al. 2005; Sethupathy et al. 2006) (Fig. 4D), showing that despite the added branch length (from BLS > 0.26 for *D. pseudoobscura* to BLS > 0.60 at 80% confidence across the 12 genomes on average), our methods maintain high sensitivity.

In contrast to evaluating conservation by the BLS methodology, requiring perfect conservation across all 12 *Drosophila* species or across the nine *Sophophora* species recovered significantly fewer experimentally validated motif instances for TF and miRNA motifs (see above and Supplementary Fig. S2).

Nonconserved binding events show decreased functional enrichment

Although the overlap between conservation derived motif instances and in vivo binding was highly significant and we recovered a substantial fraction of instances in ChIP-bound enhancers, CrebA targets, and miRNA targets, we noted that numerous motif instances in ChIP-bound regions were not conserved above 60% confidence, especially for regions that had not previously been shown to be enhancers (Fig. 4B). Nonconserved sites might be functional but missed due to unusually large motif movements or sequencing and alignment errors. Alternatively, they may play roles with only lineage-specific selection (and thus not meeting our 60% confidence threshold) or represent largely nonconsequential binding, without a specific biological role subject to evolutionary selection. To distinguish the two possibilities, we studied the enrichment of conserved and nonconserved motif instances of the mesodermal factors Mef-2, Twist, and Snail in muscle genes.

We found that ChIP-bound motif instances that were evo-

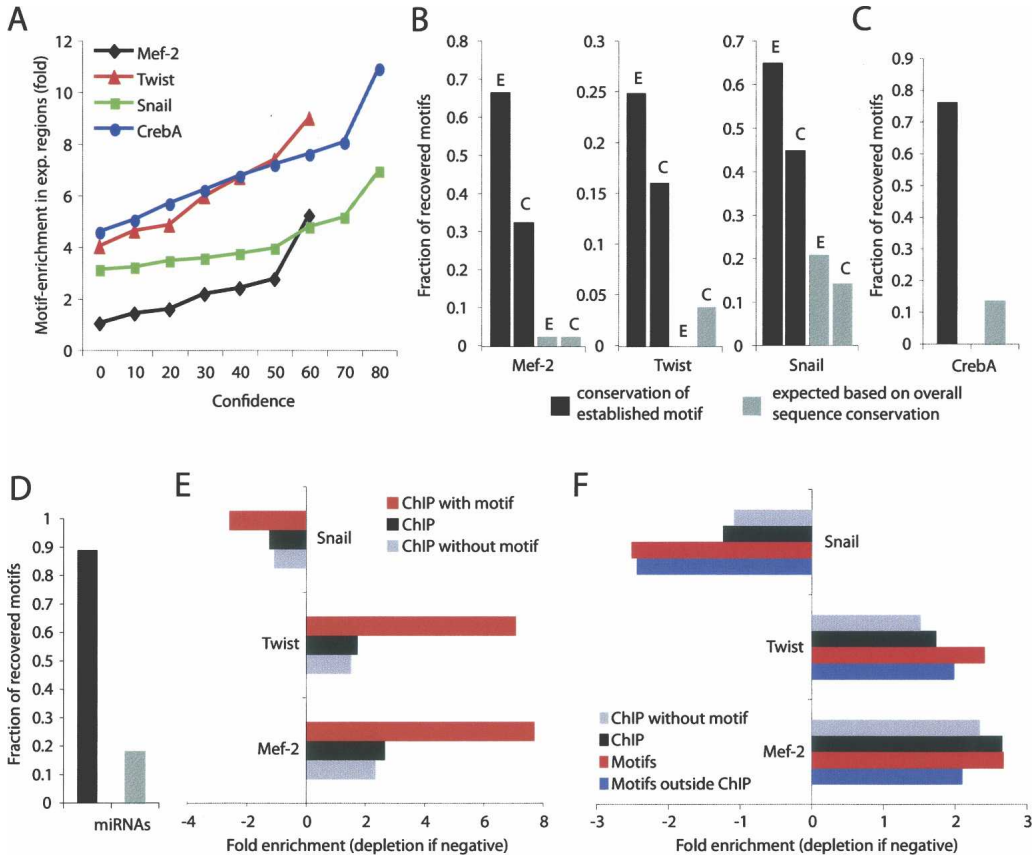


Figure 4. Conserved motif instances identify functional in vivo targets. Functional in vivo targets were determined for Mef-2, Twist, and Snail using ChIP-chip (Sandmann et al. 2006, 2007; Zeitlinger et al. 2007), and direct transcriptional targets were determined for CrebA using various assays (Abrams and Andrew 2005). (*A*) Increasing confidence values show increased enrichment for in vivo sites. Fold enrichment in functional in vivo sites (*Y*-axis) for conserved motif instances at varying confidence values (*X*-axis). Hypergeometric *P*-values for max fold enrichments are 4×10^{-11} for Mef-2, 2×10^{-6} for Twist, 3×10^{-10} for Snail, and 1×10^{-7} for CrebA. Increasing confidence levels selected functional in vivo sites with increased enrichment for all four regulators, showing that high conservation selects for functional motif instances ($X = 0\%$ shows the enrichment in the absence of comparative information, i.e., without requiring conservation). Curves are truncated when motifs do not reach the respective confidence levels. (*B–D*) High-sensitivity recovery of in vivo targets for TF and miRNA regulators. Fraction of motifs in bound regions recovered at 60% confidence (black bars), compared to the fraction expected given the overall conservation of the respective regions, as assessed by control motifs using the same BLS cutoff (gray; suggesting preferential conservation of the corresponding TF motif instances). (*B*) Recovery of ChIP-bound motifs, across all ChIP-bound regions (labeled "C"), and only those instances overlapping known enhancers (labeled "E"). Recovery rates show high sensitivity for TF motif instances, especially when these overlap enhancer regions. (*C*) Recovery of experimentally validated direct CrebA targets shows even higher sensitivity, likely due to the multiple lines of experimental evidence establishing them as direct targets. (*D*) miRNA recovery at 80% confidence is very high. (*E*) Nonconserved ChIP sites show reduced functional enrichments. Enrichment in promoter regions of muscle genes for motif instances of activators Twist and Mef-2, and depletion for motif instances of repressor Snail are reduced for ChIP-bound regions for which motif instances are not conserved, suggesting they may contain a higher fraction of nonfunctional sites. The enrichment/depletion is even weaker for ChIP-bound regions without motif instances (all enrichments are significant with *P*-values between 1.1×10^{-4} and 5.1×10^{-13} except those for Snail). (*F*) Conservation-inferred targets and ChIP-inferred targets show comparable functional enrichments. Conservation-inferred motif targets at 60% confidence (red; all $P < 10^{-4}$) show higher muscle-gene enrichment/depletion than ChIP-inferred targets (black). Even outside ChIP-bound regions, conserved motifs show comparable enrichment and depletion (blue; all $P < 5 \times 10^{-3}$).

lutionarily conserved showed enrichment or depletion in promoters of muscle genes for all three factors: The transcriptional activators Mef-2 and Twist showed eightfold and sevenfold enrichment, respectively, and Snail, a mesodermal repressor, showed threefold depletion in muscle genes. In contrast, ChIP-bound motif instances that were not conserved showed only one- to twofold enrichment or depletion for all three factors (Fig. 4E). This suggests that potential lineage-specific roles corresponding to nonconserved ChIP-bound sites may lie outside the regulators' conserved functions in core development processes (e.g., mesoderm/muscle development). Alternatively, these sites may be of decreased biological significance, perhaps representing nonconsequential binding sites with no role in gene-expression regulation, which are known to be recovered in ChIP experiments (Boyer et al. 2005; Lee et al. 2006). In either case, our results show that nonconserved sites are not simply due to low sensitivity of comparative methods but are functionally distinct from conserved sites.

ChIP-derived and conservation-derived targets show comparable functional significance

Interestingly, evolutionary conservation identified many high-confidence motif instances outside ChIP-bound regions. These may be functional sites reflecting higher coverage for conserva-

tion-derived targets or spurious sites reflecting noise in the methodology. To distinguish the two possibilities, we used the correlation of these additional motif instances with muscle genes, providing an independent assessment of the overall quality of our predictions.

We found that conservation-derived targets outside ChIP regions were enriched in the same categories in which the factors are known to act. In fact, even outside ChIP regions, conserved sites showed comparable or higher enrichment or depletion in muscle genes than those identified by the ChIP methodology (Fig. 4F), suggesting they may be of similar overall quality. For Twist, enrichment was 1.3-fold higher; for Snail, depletion was 2.5-fold higher; and for Mef-2, enrichment was slightly lower (0.9-fold). Overall, when assessing ChIP- and conservation-derived targets independently (i.e., considering all ChIP targets and all conservation-derived targets), our approach showed a consistently higher enrichment or depletion in muscle genes than ChIP-chip (1.4-fold for Twist, twofold for Snail, and 1.01-fold for Mef-2; Fig. 4F).

Our results suggest that the additional sites outside ChIP-bound regions are likely functional and reflect the higher coverage of conservation-derived targets as compared to experimentally derived targets. Indeed, while ChIP-derived targets are constrained by the developmental stages or cell types surveyed, comparative approaches capture all conserved gene targets regardless of their spatial or temporal constraints. Moreover, comparative approaches are not constrained by the abundance of TFs at bound sites, but only by the strength of evolutionary selection; they can thus identify important sites even when these are bound more rarely (or in few cell types). Lastly, comparative genomics enables us to capture additional functional targets that may be missed due to experimental limitations of ChIP technology, for which reported false-negative rates are up to 30% (Boyer et al. 2005; Lee et al. 2006).

Regulatory network of *D. melanogaster* at 60% confidence

We conclude that comparative genomics provides a powerful methodology for identifying functional targets showing high sensitivity and high specificity. For factors with experimentally determined in vivo binding sites, we showed that evolutionary conservation provides comparable discover power as ChIP and importantly reveals additional functional sites that potentially function at stages or tissues not surveyed. More generally, even when ChIP studies are not available, comparative genomics can provide a first overview of the regulatory connections across a complete genome.

We used our comparative approach to present an initial regulatory network of *D. melanogaster* at 60% confidence for both pre- and post-transcriptional regulators (Fig. 5). Overall, 49 of 57 miRNA motifs (86%) and 67 of 83 TF motifs (81%) had instances with confidence values of 60% or higher and were considered (Supplemental Tables S1, S2). The remaining motifs may have too few physiologically relevant and conserved target sites to discern them reliably from background, or they may not accurately reflect the factors' binding properties, potentially being overly specific or degenerate.

We find a total of 46,525 regulatory connections for TF motifs and 3662 for miRNA motifs, targeting 8287 genes and 2003 genes, respectively. The distribution of targets is highly asymmetric: While we find on average 123 targets per TF motif and 41 targets per miRNA motif, some TF motifs have up to 4129 targets (homeobox factors), and some miRNA motifs more than 150 targets (miR-4, miR-92, and miR-1). We note, that some motifs (e.g., the homeobox TF motif or the K-box miRNA motif) correspond to multiple TFs or miRNAs, and thus the numbers likely represent combined targets for all individual factors. The distribution of target sites per gene (indegree) is also highly imbalanced: While a typical gene is regulated by six different TF motifs and two different miRNA motifs on average, some genes are targeted by up to 33 different TF and up to 14 different miRNA motifs. Genes with high indegree were enriched in morphogenesis, organogenesis, neurogenesis, and a variety of tissues, while genes with small indegree were enriched in ubiquitously expressed or maternal genes with functions in DNA, RNA, or protein metabolism for both TF and miRNA motifs (Supplemental Table S3). Many genes with high indegree were TFs ($P < 10^{-9}$ for TF and miRNA motifs), and transcriptional regulators were indeed more densely targeted than other genes, by both TF (10.1 vs. 5.5, $P < 10^{-20}$) and miRNA motifs (2.3 vs. 1.8, $P < 5 \times 10^{-5}$). The similarity between the TF and miRNA motif network was further illustrated by mutual enrichment: Genes with high TF indegree are enriched in genes with high miRNA indegree ($P = 8 \times 10^{-5}$), as are genes with low indegree for both types of regulators ($P = 2 \times 10^{-7}$).

This initial network contained many connections with independent support in the literature (Fig. 5; Supplemental Table S4). For example, we identified the direct regulation of *achaete* by Hairy (Van Doren et al. 1994), several direct targets of Suppressor of Hairless Su(H) in the *Enhancer* of split E(spl) complex (Bailey and Posakony 1995), and direct regulation of the gap gene *giant* by Bicoid (Kraut and Levine 1991). In addition, the network proposed many novel connections supported by experimental evidence, including direct regulation of *bagpipe* by Tinman, which both cooperate in mesoderm induction and heart specification (Yin and Frasch 1998). More generally, when tissue-specific expression data was available, we found that on average 46% of all targets were co-expressed with their factor in at least one tissue (Fig. 5), which is significantly higher than expected by chance ($P = 2 \times 10^{-3}$).

Discussion

We showed that comparative analysis of many related genomes allows us to identify functional motif instances with very high confidence. Overall, 86% miRNA motifs and 81% TF motifs had instances with confidence values of ≥60. The remaining factors may have too few physiologically relevant and conserved target sites to discern them reliably from background, or may contain inaccuracies in their binding site motifs might be artificially specific or degenerate.

We found that the availability of many genomes allowed for very high signal-to-noise levels for many motifs at the most stringent settings. However, more importantly, we showed that the BLS measure allowed us to use the increased number of species to strongly increase sensitivity at any given specificity compared to requiring perfect motif conservation in arbitrary subsets of species. While requiring perfect conservation across many genomes is of limited use, the increased power enables approaches that account for artifacts in sequencing, assembly and alignment, and tolerate diverged, missing, or moved motif instances. Our BLS measure is more generally applicable to PWMs (Stormo 2000), to

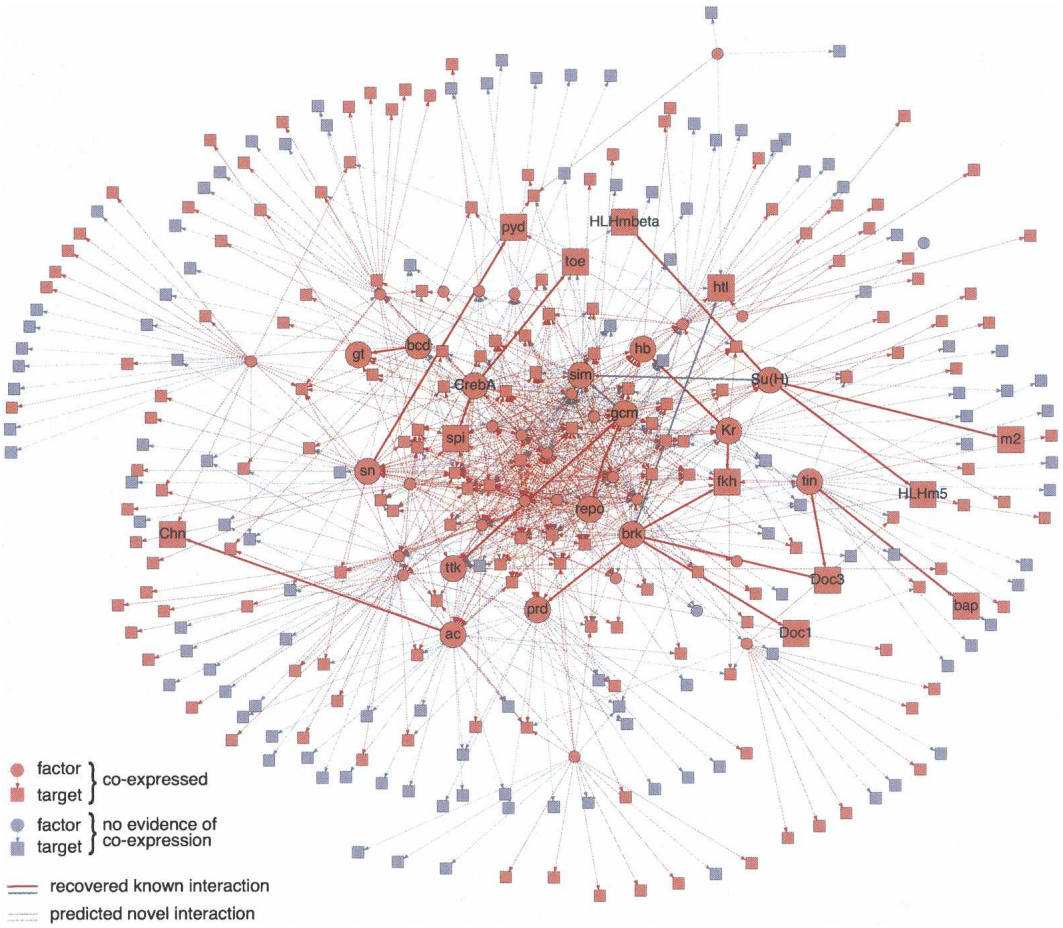


Figure 5. An initial regulatory network in *Drosophila*. Regulatory network with 46,525 connections between 83 TF and 57 miRNA motifs (circles) and their target genes (squares) at 60% confidence. If the regulator and its target are co-expressed in at least one tissue according to ImaGO (Tomancak et al. 2002), the corresponding edges (lines) and nodes (circles or squares) are colored red; otherwise they are gray. The high fraction of red edges (46%, $P = 2 \times 10^{-3}$) highlights the quality of the network. Nodes with gene names and connected by bold edges indicate examples of regulatory connections with evidence in the literature (see Supplemental Table S4).

more complex models of regulatory motifs that account for dependencies between individual motif positions (Yada et al. 1998; Naughton et al. 2006), and to more advanced rules for miRNA-target recognition that for example score the contribution of the 3′pairing energy (Stark et al. 2003; Brennecke et al. 2005).

We found that comparative genomics and ChIP-chip showed similar power for functional target identification. The two approaches are complementary, each with unique advantages: Conservation helps pinpoint evolutionarily selected functional targets across all conditions, while ChIP-chip reveals stage- and tissue-specific binding in vivo, as well as species-specific sites which may play important evolutionary roles in the emergence of new functions. As motifs of additional regulators are derived by experimental (e.g., by SELEX, in Tuerk and Gold [1990] or protein-binding microarrays [Mukherjee et al. 2004]) or computational approaches (e.g., by motif-overrepresentation [Tompa et al. 2005] or genome-wide motif-instance conservation [Kellis et al. 2003; Xie et al. 2005]), and tissue-specific binding becomes available for dozens of factors (e.g., through the ENCODE and modENCODE projects), comparative studies can help establish and refine their genome-wide targets. Indeed, we found that motif instances identified by both approaches had the highest functional enrichments, suggesting that combined approaches may prove useful in the future. Although the regulatory network we present likely lacks many true regulatory relationships that could not be reliably recovered, our comparison with ChIP-chip data and other validated targets showed that the network is of high overall quality. We anticipate that the network and the predicted regulatory connections prove to be a useful resource for the fly community working on the biology of TFs or miRNAs and their target genes and their roles in development. The methodology to assess motif conservation across many genomes and predict functional motif instances with high sensitivity is more generally applicable for the study of any genome.

Methods

Regulatory motifs

We obtained TF motifs from Transfac (Matys et al. 2003), Jaspar (Sandelin et al. 2004), FlyReg (Bergman et al. 2005), and the literature. To remove redundancy for global statements about motif targets, we clustered TF motifs using centroid-linkage hierarchical clustering with a Pearson correlation coefficient cutoff of 0.8 (calculated on the columns of the equivalent PWM) at the

best alignment offset (Pietrokovski 1996; Schones et al. 2005; Xie et al. 2005; Gupta et al. 2007). To avoid the creation of artificial motifs by averaging, we chose the original motif from each cluster that is closest to the cluster average as the cluster representative. We defined miRNA motifs as the nonredundant set of 7mers reverse complementary to miRNA 5′ end positions 2–8 (seeds after Lewis et al. 2003) for all *Drosophila* miRNAs in Rfam release 9.2 (Griffiths-Jones et al. 2006). We represent all motifs as consensus sequences over an alphabet of 15 characters (IUPAC code, http://www.chem.qmul.ac.uk/iupac/) consisting of the four nucleotides A,C,G,T, the six twofold degenerate characters S = (CG), W = (AT), Y = (CT), R = (AG), M = (AC), and K = (GT), the four threefold degenerate characters H = (ACT), B = (GCT), V = (G,A,C), and D = (G,A,T), and the fourfold degenerate character N = (ACGT). A motif instance (or motif occurrence) is a sequence that matches the motif at each position, i.e., containing one of the allowed characters at that position.

We translate consensus sequences to PWMs given the definition of the degenerate characters. We translate PWMs to consensus sequences by choosing the character with the highest sum of the PWM column entries corresponding to that character minus a correction for character degeneracy (1/2 for ACGT, 2/3 for SYRMK, 5/6 for HBVD, and 1 for N).

Genome alignments and annotation

For all analyses, we used whole genome MULTIZ alignments of 12 *Drosophila* genomes (Stark et al. 2007), available from UCSC (Kent et al. 2002). We used the *D. melanogaster* genome-annotations from FlyBase (Release 4.3), and excluded simple repeats, repeat masked regions obtained from UCSC, and noncoding exons according to FlyBase 4.3.

Motif matching and BLS measure

We searched all motif instances in the *D. melanogaster* genome and evaluated their conservation in the 12 species using the whole-genome alignments. For each motif instance in *D. melanogaster*, we recorded all instances in the other genomes that were aligned, allowing for motif movements (see below). We prevented double counting of motif instances by assigning each instance in an informant species to the closest instance in *D. melanogaster*. We evaluated the conservation of all motif instances by summing the branch-lengths of the subtree of the species with conserved motif instances (BLS). This procedure implicitly assumes that all instances are potentially ancestral, such that an instance conserved in a remote informant species would score more highly than instances in closely related informants. One disadvantage of this approach is therefore that chance occurrences or gains in distant species may contribute false positives. The phylogenetic tree branch lengths were obtained from a whole-genome alignment of all 12 species (Dewey et al. 2006; Stark et al. 2007).

P-values

All *P*-values are calculated based on the hypergeometric distribution, and correction for multiple-testing was done with the Bonferroni correction.

Allowing for motif movements

When assessing motif conservation, we allowed motif instances in the informant species to be offset relative to the alignment position of the *D. melanogaster* instances within a given window (counted as distance in either direction in characters excluding gaps). We did not use a prior for a cutoff on maximal tolerable motif movement, as we are not aware of a systematic experimental study that assessed typical movements of functionally equivalent motifs in related species nor systematically assessed of the maximum movement tolerable while maintaining function. We consequently used the window that maximized signal over noise.

While it is clear that increasing tolerated windows may capture additional equivalent instances across genomes, thereby increasing sensitivity, they also increase the number of spurious motif instances that are recovered by chance. We account for the increased background conservation by the use of control motifs (see above), and determine the optimal allowable motif movement window (the one that recovered most motif instances) out of 32 windows between 0 and 500 nucleotides (0, 5, 10, 20, 30, . . . , 90, 100, 120, 140, . . . , 480, 500). For Figure 4B and for analyzing the correlation of optimal window size with different motif properties, we assessed 119 windows between 0 and 10,000 nucleotides (0, 10, 20, . . . , 190, 200, 300, . . . , 9900, 10,000). Similarly, we allow for strand reversals of TF motif instances in informant species, when they help instance recovery in the respective windows. The significance of sensitivity improvement for individual windows and for allowing windows in general was assessed by hypergeometric *P*-values compared to motif instances identified with a window of 0 nucleotides, i.e., perfect alignment of instances.

Estimation of confidence levels of motif instances

For each motif and type of genomic region (promoter, 5′ UTR, 3′ UTR, intron, etc.), we created 100 shuffled control motifs and selected those that had a similar number of matches to the region in the *D. melanogaster* genome (±20%). By requiring the control motifs to have occurrence rates similar to real motifs in the respective genomic regions in *D. melanogaster* (i.e., without conservation), we corrected for biases in di- or trinucleotide frequencies (see discussion in Lewis et al. 2003). To remove possible redundancy, we clustered the control motifs (cutoff 0.8) and selected only one representative per cluster, limiting to 10 motifs total that were least similar to known motifs. For each real motif and its controls, we computed the conservation rate (the number of conserved instances at a given BLS cutoff divided by the total number of instances in the *D. melanogaster* genome) in each region and at each BLS cutoff. We determined the confidence at each BLS as the fraction of conserved motif instances above background conservation, where the latter was estimated using the conservation ratio of the control motifs. This provided a BLS-to-confidence mapping for each motif and region. The variation between the control motifs lead to an average standard-error of 5% for TF motifs, and 4% for miRNA motifs at 60% confidence, indicating an accurate assessment of background conservation.

Comparison with experimental data sets

We obtained all experimentally validated miRNA target gene pairs from TarBase (Sethupathy et al. 2006) and our previous study (Stark et al. 2005). We obtained ChIP-chip regions and the subset that overlapped known enhancers from (Sandmann et al. 2006, 2007; Zeitlinger et al. 2007) and CrebA target genes from Abrams and Andrew (2005). We calculated the enrichment of sites at different confidence cutoffs between 3′UTRs of validated miRNA/target pairs and all 3′UTRs, and between ChIP regions within 2 kb upstream regions and the union of all 2 kb upstream regions. As CrebA targets were originally defined through mostly 5′ UTR instances (Abrams and Andrew 2005) and Mef-2 showed considerable overlap with 5′ UTR regions, we included the 5′ UTR and restricted the upstream region to 500 bp instead. We

assessed the recovery of motif instances as the fraction of motif instances in the functional regions (with the same restrictions) that reached the indicated confidence. To assess the fraction of these that are expected by putatively increased overall conservation in these regions, we assess the recovery of control motifs at the same BLS (not confidence, as the control motifs, by definition, would not reach high confidence levels).

Evaluation of experimental and motif instances by correlation with muscle genes

We used correlation with expression patterns to independently evaluate ChIP-regions and predicted motif instances. Muscle genes were 616 genes annotated as "muscle system (13-16)" by the manually curated BDGP in situ database (ImaGO) (Tomancak et al. 2002). To obtain a unique assignment of regions to genes, we restricted our analysis to the 5′ UTR and 500 bases upstream of each gene. We calculated functional enrichments as the fraction of nucleotides covered by motif instances (at 60% confidence) or ChIP regions in muscle genes divided by the corresponding number in all genes present in ImaGO. Hypergeometric *P*-values were computed for motif instances using control motifs at the same BLS and window and for ChIP regions using the fraction of muscle genes matched versus the fraction of all genes matched (note that individual nucleotides are correlated, such that nucleotide *P*-values would overestimate the significance).

Assessing the indegree distribution

We assessed the nonrandomness of the indegree distribution against a control Erdos–Renyi random network (Bollobás 2001) with the same number of edges. To construct this network, we added edges by selecting a source and target node with probability $1/m$ and $1/n$, where m and n were the number of source and target nodes in the true network, respectively. We assessed the difference of indegree distributions between the true and control network with a Wilcoxon rank-sum test. We also assessed the difference in indegree distribution between all transcription factors (as defined by Adryan and Teichmann 2006) and all other genes also with a Wilcoxon rank-sum test.

Functional/ImaGO enrichment of high and low indegree genes

We considered all genes with a GO (Ashburner et al. 2000) and ImaGO (Tomancak et al. 2002) functional annotation (n = 7495 and 5996, respectively) and computed the indegree (number of incoming edges) for each gene in the transcription factor (TF) and miRNA networks. For both networks we defined high-indegree nodes as the 1% with the highest indegree ($\geq$20 for the TF network and $\geq$4 for the miRNA network) and low-indegree nodes as miRNA antitargets (indegree = 0) and the same fraction of nodes with lowest indegree in the TF network (80%; $\leq$7 edges). For each GO/ImaGO category, we assessed over-representation and depletion with a hypergeometric *P*-value.

Mutual enrichment between high indegree transcriptional and miRNA targets

We considered all genes that were either a target or a regulator in the TF and microRNA networks resulting in a total of 8760 nodes and defined high- and low-indegree sets as above. We then evaluated if nodes in the miRNA network with high indegree were enriched high-indegree nodes of the transcriptional network (or vice versa) using a hypergeometric *P*-value.

Tissue co-expression

For each TF with available expression information (n = 42; ImaGO; see Tomancak et al. 2002), we counted the number of targets that were co-expressed with the TF in any of the annotated tissues and the number of targets that were not annotated to be co-expressed. The statistical significance of co-expression of a TF with its target was estimated using the hypergeometric distribution given the number of co-expressed targets, and the total number of targets of the TF with known tissue expression, and the corresponding counts for all genes.

Network figure

The network figure was drawn in Cytoscape (Shannon et al. 2003) to display genes (nodes) and regulatory connections (edges) of the 60% confidence network. We colored edges and nodes if genes were expressed in the same tissue according to ImaGO (Tomancak et al. 2002). For clarity, we only show 20 randomly picked targets per transcription factor, i.e., without influencing the fraction of colored edges.

Acknowledgments

We thank Matt Rasmussen, Mike Lin (CSAIL, Broad), and other members of the Kellis laboratory for helpful discussions and for sharing unpublished data. A.S. thanks the Human Frontier Science Program Organization (HFSPO) for a postdoctoral fellowship (LT00495/2006-L). P.K. was supported in part by a National Science Foundation Graduate Research Fellowship. S.R. thanks Terran Lane and Maggie Werner-Washburne (University of New Mexico) for their support.

References

Abrams, E.W. and Andrew, D.J. 2005. CrebA regulates secretory activity in the *Drosophila* salivary gland and epidermis. *Development* **132:** 2743–2758.

Adryan, B. and Teichmann, S.A. 2006. FlyTF: A systematic review of site-specific transcription factors in the fruit fly *Drosophila melanogaster*. *Bioinformatics* **22:** 1532–1533.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25:** 25–29.

Bailey, A.M. and Posakony, J.W. 1995. Suppressor of hairless directly activates transcription of enhancer of split complex genes in response to Notch receptor activity. *Genes & Dev.* **9:** 2609–2622.

Bergman, C.M., Carlson, J.W., and Celniker, S.E. 2005. *Drosophila* DNase I footprint database: A systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster*. *Bioinformatics* **21:** 1747–1749.

Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., and Eisen, M.B. 2002. Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl. Acad. Sci.* **99:** 757–762.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L., and Rubin, E.M. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299:** 1391–1394.

Bollobás, B. 2001. *Random graphs*. Cambridge University Press, Cambridge, UK.

Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., et al. 2005. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122:** 947–956.

Brennecke, J., Stark, A., Russell, R.B., and Cohen, S.M. 2005. Principles of microRNA-target recognition. *PLoS Biol.* **3:** e85. doi: 10.1371/journal.pbio.0030085.

Chan, C.S., Elemento, O., and Tavazoie, S. 2005. Revealing posttranscriptional regulatory elements through network-level

conservation. *PLoS Comput. Biol.* **1:** e69. doi: 10.1371/journal.pcbi.0010069.
Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A., and Johnston, M. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301:** 71–76.
Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., and Sidow, A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15:** 901–913.
Davidson, E.H. and Erwin, D.H. 2006. Gene regulatory networks and the evolution of animal body plans. *Science* **311:** 796–800.
Dewey, C.N., Huggins, P.M., Woods, K., Sturmfels, B., and Pachter, L. 2006. Parametric alignment of *Drosophila* genomes. *PLoS Comput. Biol.* **2:** e73. doi: 10.1371/journal.pcbi.0020073.
Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* (in press) doi: 10.1038/nature06341.
Eddy, S.R. 2005. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.* **3:** e10. doi: 10.1371/journal.pbio.0030010.
Elnitski, L., Hardison, R.C., Li, J., Yang, S., Kolbe, D., Eswara, P., O'Connor, M.J., Schwartz, S., Miller, W., and Chiaromonte, F. 2003. Distinguishing regulatory DNA from neutral sites. *Genome Res.* **13:** 64–72.
Ettwiller, L., Paten, B., Souren, M., Loosli, F., Wittbrodt, J., and Birney, E. 2005. The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates. *Genome Biol.* **6:** R104. doi: 10.1186/gb-2005-6-12-r104.
Felsenstein, J. 2004. *Inferring phylogenies*. Sinauer Associates, Sunderland, MA.
Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., and Enright, A.J. 2006. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **34:** D140–D144. doi: 10.1093/nar/gkj112.
Grun, D., Wang, Y.L., Langenberger, D., Gunsalus, K.C., and Rajewsky, N. 2005. microRNA target predictions across seven *Drosophila* species and comparison to mammalian targets. *PLoS Comput. Biol.* **1:** e13. doi: 10.1371/journal.pcbi.0010013.
Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. 2007. Quantifying similarity between motifs. *Genome Biol.* **8:** R24. doi: 10.1186/gb-2007-8-2-r24.
Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241–254.
Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res.* **12:** 996–1006.
Kraut, R. and Levine, M. 1991. Spatial regulation of the gap gene giant during *Drosophila* development. *Development* **111:** 601–609.
Lai, E.C. 2004. Predicting and validating microRNA targets. *Genome Biol.* **5:** 115. doi: 10.1186/gb-2004-5-9-115.
Lai, E.C., Burks, C., and Posakony, J.W. 1998. The K box, a conserved 3′ UTR sequence motif, negatively regulates accumulation of enhancer of split complex transcripts. *Development* **125:** 4077–4088.
Lee, T.I., Jenner, R.G., Boyer, L.A., Guenther, M.G., Levine, S.S., Kumar, R.M., Chevalier, B., Johnstone, S.E., Cole, M.F., Isono, K., et al. 2006. Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125:** 301–313.
Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P., and Burge, C.B. 2003. Prediction of mammalian microRNA targets. *Cell* **115:** 787–798.
Ludwig, M.Z., Bergman, C., Patel, N.H., and Kreitman, M. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403:** 564–567.
Ludwig, M.Z., Palsson, A., Alekseeva, E., Bergman, C.M., Nathan, J., and Kreitman, M. 2005. Functional evolution of a *cis*-regulatory module. *PLoS Biol.* **3:** e93. doi: 10.1371/journal.pbio.0030093.
Margulies, E.H., Blanchette, M., Haussler, D., and Green, E.D. 2003. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13:** 2507–2518.
Margulies, E.H., Cooper, G.M., Asimenos, G., Thomas, D.J., Dewey, C.N., Siepel, A., Birney, E., Keefe, D., Schwartz, A.S., Hou, M., et al. 2007. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res.* **17:** 760–774.
Markstein, M., Zinzen, R., Markstein, P., Yee, K.P., Erives, A., Stathopoulos, A., and Levine, M. 2004. A regulatory code for neurogenic gene expression in the *Drosophila* embryo. *Development* **131:** 2387–2394.
Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., et al. 2003. TRANSFAC: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31:** 374–378. doi: 10.1093/nar/gkg108.
McGregor, A.P., Orgogozo, V., Delon, I., Zanet, J., Srinivasan, D.G., Payre, F., and Stern, D.L. 2007. Morphological evolution through multiple *cis*-regulatory mutations at a single gene. *Nature* **448:** 587–590.
Miller, W., Makova, K.D., Nekrutenko, A., and Hardison, R.C. 2004. Comparative genomics. *Annu. Rev. Genomics Hum. Genet.* **5:** 15–56.
Moses, A.M., Chiang, D.Y., Pollard, D.A., Iyer, V.N., and Eisen, M.B. 2004. MONKEY: Identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol.* **5:** R98. doi: 10.1186/gb-2004-5-12-r98.
Mukherjee, S., Berger, M.F., Jona, G., Wang, X.S., Muzzey, D., Snyder, M., Young, R.A., and Bulyk, M.L. 2004. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.* **36:** 1331–1339.
Naughton, B.T., Fratkin, E., Batzoglou, S., and Brutlag, D.L. 2006. A graph-based motif detection algorithm models complex nucleotide dependencies in transcription factor binding sites. *Nucleic Acids Res.* **34:** 5730–5739. doi: 10.1093/nar/gkl585.
Philippakis, A.A., Busser, B.W., Gisselbrecht, S.S., He, F.S., Estrada, B., Michelson, A.M., and Bulyk, M.L. 2006. Expression-guided in silico evaluation of candidate *cis* regulatory codes for *Drosophila* muscle founder cells. *PLoS Comput. Biol.* **2:** e53. doi: 10.1371/journal.pcbi.0020053.
Pietrokovski, S. 1996. Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.* **24:** 3836–3846.
Prud'homme, B., Gompel, N., Rokas, A., Kassner, V.A., Williams, T.M., Yeh, S.D., True, J.R., and Carroll, S.B. 2006. Repeated morphological evolution through *cis*-regulatory changes in a pleiotropic gene. *Nature* **440:** 1050–1053.
Rajewsky, N. 2006. MicroRNA target predictions in animals. *Nat. Genet.* **38:** S8–S13.
Richards, S., Liu, Y., Bettencourt, B.R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M.J., Chen, R., Meisel, R.P., et al. 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene, and *cis*-element evolution. *Genome Res.* **15:** 1–18.
Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., and Lenhard, B. 2004. JASPAR: An open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32:** D91–D94. doi: 10.1093/nar/gkh012.
Sandmann, T., Girardot, C., Brehme, M., Tongprasit, W., Stolc, V., and Furlong, E.E. 2007. A core transcriptional network for early mesoderm development in *Drosophila melanogaster*. *Genes & Dev.* **21:** 436–449.
Sandmann, T., Jensen, L.J., Jakobsen, J.S., Karzynski, M.M., Eichenlaub, M.P., Bork, P., and Furlong, E.E. 2006. A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. *Dev. Cell* **10:** 797–807.
Schones, D.E., Sumazin, P., and Zhang, M.Q. 2005. Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics* **21:** 307–313.
Schroeder, M.D., Pearce, M., Fak, J., Fan, H., Unnerstall, U., Emberly, E., Rajewsky, N., Siggia, E.D., and Gaul, U. 2004. Transcriptional control in the segmentation gene network of *Drosophila*. *PLoS Biol.* **2:** e271. doi: 10.1371/journal.pbio.0020271.
Sethupathy, P., Corda, B., and Hatzigeorgiou, A.G. 2006. TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA* **12:** 192–197.
Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13:** 2498–2504.
Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15:** 1034–1050.
Stark, A., Brennecke, J., Russell, R.B., and Cohen, S.M. 2003. Identification of *Drosophila* microRNA targets. *PLoS Biol.* **1:** e60. doi: 10.1371/journal.pbio.0000060.
Stark, A., Brennecke, J., Bushati, N., Russell, R.B., and Cohen, S.M. 2005. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. *Cell* **123:** 1133–1146.
Stark, A., Lin, M.F., Kheradpour, P., Pedersen, J.S., Parts, L., Carlson, J.W., Crosby, M.A., Rasmussen, M.D., Roy, S., Deoras, A.N., et al. 2007. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* (in press). doi: 10.1038/nature06340.

Stormo, G.D. 2000. DNA binding sites: Representation and discovery. *Bioinformatics* **16:** 16–23.
Taylor, J., Tyekucheva, S., King, D.C., Hardison, R.C., Miller, W., and Chiaromonte, F. 2006. ESPERR: Learning strong and weak signals in genomic sequence alignments to identify functional elements. *Genome Res.* **16:** 1596–1604.
Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., McDowell, J.C., et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424:** 788–793.
Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., Shu, S., Lewis, S.E., Richards, S., Ashburner, M., Hartenstein, V., Celniker, S.E., et al. 2002. Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol* **3:** RESEARCH0088. doi: 10.1186/gb-2002-3-12-research0088.
Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., et al. 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* **23:** 137–144.
Tuerk, C. and Gold, L. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249:** 505–510.
Ureta-Vidal, A., Ettwiller, L., and Birney, E. 2003. Comparative genomics: Genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet.* **4:** 251–262.
Van Doren, M., Bailey, A.M., Esnayra, J., Ede, K., and Posakony, J.W. 1994. Negative regulation of proneural gene activity: hairy is a direct transcriptional repressor of achaete. *Genes & Dev.* **8:** 2729–2742.
Wasserman, W.W. and Sandelin, A. 2004. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* **5:** 276–287.
Wasserman, W.W., Palumbo, M., Thompson, W., Fickett, J.W., and Lawrence, C.E. 2000. Human–mouse genome comparisons to locate regulatory sites. *Nat. Genet.* **26:** 225–228.
Wray, G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V., and Romano, L.A. 2003. The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* **20:** 1377–1419.
Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. 2005. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434:** 338–345.
Yada, T., Totoki, Y., Ishikawa, M., Asai, K., and Nakai, K. 1998. Automatic extraction of motifs represented in the hidden Markov model from a number of DNA sequences. *Bioinformatics* **14:** 317–325.
Yin, Z. and Frasch, M. 1998. Regulation and function of tinman during dorsal mesoderm induction and heart specification in *Drosophila*. *Dev. Genet.* **22:** 187–200.
Zeitlinger, J., Zinzen, R.P., Stark, A., Kellis, M., Zhang, H., Young, R.A., and Levine, M. 2007. Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the *Drosophila* embryo. *Genes & Dev.* **21:** 385–390.
Zhou, Q. and Wong, W.H. 2004. CisModule: De novo discovery of *cis*-regulatory modules by hierarchical mixture modeling. *Proc. Natl. Acad. Sci.* **101:** 12114–12119.

Received August 29, 2007; accepted in revised form October 10, 2007.

Downloaded from www.genome.org on April 28, 2008 - Published by Cold Spring Harbor Laboratory Press

Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes

Matthew D. Rasmussen[1] and Manolis Kellis[1,2,3]

[1]*MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts 02139, USA;* [2]*The Broad Institute, Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02140, USA*

Comparative genomics provides a general methodology for discovering functional DNA elements and understanding their evolution. The availability of many related genomes enables more powerful analyses, but requires rigorous phylogenetic methods to resolve orthologous genes and regions. Here, we use 12 recently sequenced *Drosophila* genomes and nine fungal genomes to address the problem of accurate gene-tree reconstruction across many complete genomes. We show that existing phylogenetic methods that treat each gene tree in isolation show large-scale inaccuracies, largely due to insufficient phylogenetic information in individual genes. However, we find that gene trees exhibit common properties that can be exploited for evolutionary studies and accurate phylogenetic reconstruction. Evolutionary rates can be decoupled into gene-specific and species-specific components, which can be learned across complete genomes. We develop a phylogenetic reconstruction methodology that exploits these properties and achieves significantly higher accuracy, addressing the species-level heterotachy and enabling studies of gene evolution in the context of species evolution.

[Supplemental material is available online at www.genome.org and http://compbio.mit.edu/spidir/.]

Comparative genomics of multiple related species has emerged as a powerful approach for the systematic discovery of evolutionarily conserved functional elements (Mouse Genome Sequencing Consortium 2002; Kellis et al. 2003; Ureta-Vidal et al. 2003; Miller et al. 2004; Richards et al. 2005), and for the identification of duplicated and rapidly evolving genes involved in the emergence of new functions (Jaillon et al. 2004; Kellis et al. 2004; Dehal and Boore 2005). Both types of analysis rely on an accurate mapping of orthologous and paralogous genes and regions across the species compared, accounting for all duplication and loss events (Fitch 1970).

Phylogenetic trees provide a rigorous framework for genome comparison (Woese et al. 1990; Baldauf et al. 2000; Murphy et al. 2001), naturally capturing gene duplication and loss, and allowing varying rates of sequence divergence across evolutionary time (Goodman et al. 1979; Page 1994; Eisen 1998). Phylogenies of orthologous genes across species can be used to study species evolution, each internal node representing a speciation event (Fig. 1A). Similarly, phylogenies of paralogous genes within a species can be used to study gene-family expansions, each internal node representing a gene-duplication event (Fig. 1B). Phylogenetics in the context of multiple complete genomes, known as phylogenomics (Eisen 1998), combines multiple orthologs and paralogs across many species in a general gene tree (Fig. 1C) and enables a much richer set of questions than ortholog trees or paralog trees alone (Ma et al. 2000; Zmasek and Eddy 2002; Storm and Sonnhammer 2003; Arvestad et al. 2004; Dufayard et al. 2005; Durand et al. 2006; Li et al. 2006; Huerta-Cepas et al. 2007). Its internal nodes thus represent both speciation and duplication events, and their ordering dictates the evolutionary history of a gene family across the species compared (Goodman et al. 1979; Page 1994). Ortholog and paralog relationships can be readily inferred by mapping general gene trees to the known phylogeny relating the species (Fig. 1D), in a process known as reconciliation. Reconciliation assumes that the species tree is known, and that the gene tree is correct. However, these assumptions have been found to be frequently violated (Rokas et al. 2003; Li et al. 2006), and erroneous gene trees can lead to incorrect ortholog and paralog assignments and many extraneous duplications and losses (Fig. 1E), thus distorting inferred patterns of gene-family expansion and contraction (Hahn 2007).

In this work, we use the 12 recently sequenced *Drosophila* genomes (*Drosophila* 12 Genomes Consortium 2007) and nine publicly available fungal genomes (Wolfe and Shields 1997) to study the properties and reconstruction of gene family evolution in the context of complete genomes. Our work has three key contributions:

- We show that many gene-tree incongruences in both flies and fungi are likely due to inaccuracies in phylogenetic reconstruction stemming from the lack of informative sites in the short alignments of individual genes. Indeed, we find that incongruences are most pronounced for short alignments and slow-evolving genes, and lead to the same alternate topologies as found in simulation due to reconstruction inaccuracies, suggesting that they are primarily methodological rather than biological.
- We show that the substitution rate of any gene can be expressed as the product of a gene-specific rate, dictated by the selective constraints on the gene's function (Dickerson 1971; Bromham and Penny 2003), and a species-specific rate, dictated by the reproductive and population dynamics of each lineage (Ohta and Kimura 1971), and we provide specific dis-

[3]Corresponding author.
E-mail manoli@mit.edu; fax (617) 253-7512.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.7105007.

Downloaded from www.genome.org on April 28, 2008 - Published by Cold Spring Harbor Laboratory Press

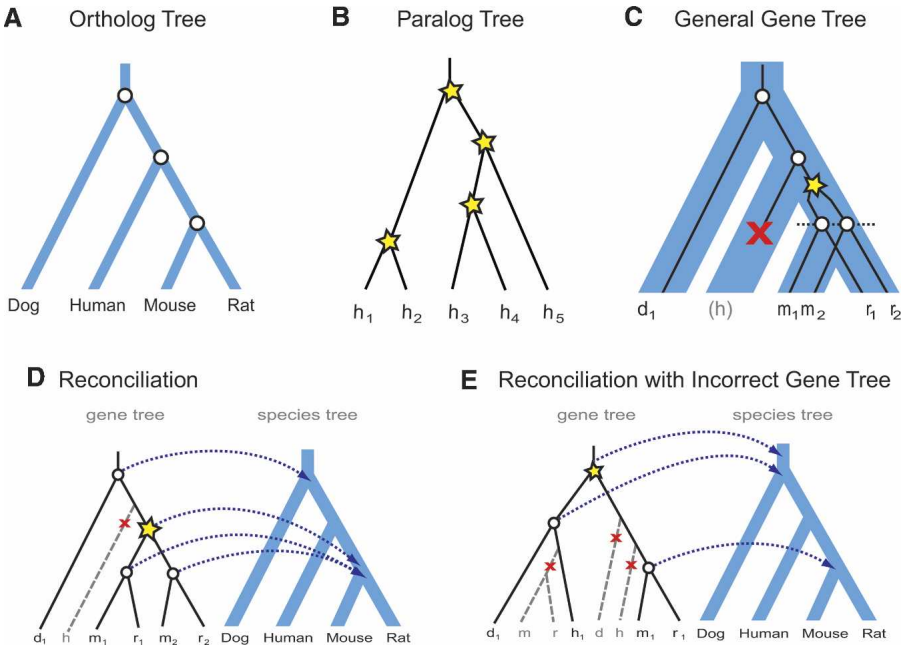


Figure 1. Relationship between gene trees and species trees. (*A*) Ortholog trees used to study species evolution. Each internal node represents a speciation event (circle). (*B*) Paralog trees used to study gene family expansions within a single species. Each internal node represents a duplication event (star). (*C*) General gene trees combine both orthologs and paralogs across multiple species to infer gene duplication (star), gene loss (×), and speciation (circle) events. Each gene is named with the first letter of the corresponding species. The gene tree (black lines) can be viewed as evolving inside the species tree (blue area), implying coordinated speciation events at branching points in the species tree (dotted line). (*D*) Gene duplication and loss events are inferred by reconciling a gene tree to a species tree, mapping each gene-tree node to its closest species-tree common ancestor node (arrows). (*E*) When the gene tree is incorrect, many spurious events will be inferred. In this example, a common misplacement of rodents due to long-branch-attraction leads to four spurious events (one duplication and at least three losses).

tributions for the two. This decomposition provides a surprisingly good fit to actual phylogenies in both flies and fungi, and can be used for accurate gene-tree reconstruction.

- We present a probabilistic framework for distance-based gene-tree reconstruction in complete genomes, based on rate-distribution parameters learned from alignments of unambiguous orthologs and implemented in a publicly available tool, SPIDIR. We used SPIDIR to infer gene trees in both flies and fungi, and show that it leads to significantly higher reconstruction accuracies. In particular, we find that our strategy can address the long-branch attraction problem for species-level heterotachy (Bergsten 2005; Philippe et al. 2005), by learning to expect longer branches for faster-evolving lineages.

Results

Incongruences and inaccuracies of gene trees for syntenic orthologs

Numerous studies have addressed the accuracy of phylogeny reconstruction methods using mainly simulated alignments (Saitou and Imanishi 1989; Kuhner and Felsenstein 1994; Tateno et al. 1994; Philippe et al. 2005) and, in some cases, microevolution observed experimentally (Hillis et al. 1994; Bull et al. 1997; Woods et al. 2006). With multiple complete genomes, regions of conserved gene order (synteny) provide a natural test for phylogenetic methods (Rokas et al. 2003; Ciccarelli et al. 2006), since all genes within these regions are typically coinherited from a single gene in the common ancestor of the species (Fig. 2A). Therefore, in the absence of horizontal gene transfer, gene conversions, and incomplete lineage sorting (Avise et al. 1983; Koonin et al. 2001), their phylogenies should be perfectly congruent to the species phylogeny. However, recent studies have shown that phylogenetic trees obtained for different orthologs frequently disagree with the species phylogeny, resulting in large-scale incongruences (Rokas et al. 2003; Li et al. 2006; Hahn 2007; Huerta-Cepas et al. 2007).

Indeed, using 5154 syntenic one-to-one orthologs across 12 *Drosophila* genomes and 739 syntenic one-to-one orthologs across nine fungal genomes (see Methods), we found that existing phylogenetic methods recovered the known species topology (*Drosophila* 12 Genomes Consortium 2007; Stark et al. 2007), denoted T1, for only a small minority of gene trees (Fig. 2B), between 24% and 42% for flies and between 22% and 31% for fungi. This was true across all methods tested, PHYML (Guindon and Gascuel 2003), DNAML (Felsenstein 2005), MrBayes (Ronquist and Huelsenbeck 2003), BIONJ (Gascuel 1997), Parsimony (Felsenstein 2005), for both protein-coding and nucleotide alignments, and using various substitution models (HKY) (Hasegawa et al. 1985), JTT (Jones et al. 1992), and synonymous-substitution dS (Yang 1997) (see Supplemental information). Moreover, no alternate topology was systematically favored: The next most frequent topologies for PHYML, denoted T2–T5, covered between 4% and 11% of fly trees, and an additional 305 topologies accounted for the remaining 31% of trees (Fig. 2B).

Biological mechanisms proposed for gene-tree incongruence, such as incomplete lineage sorting of pre-speciation alleles (Pollard et al. 2006; Wong et al. 2007), may be contributing to the observed incongruences, but are unlikely to explain all incongruent gene trees. Instead, we found multiple lines of evidence suggesting that algorithmic inaccuracies, rather than biological reasons, are likely responsible for a large fraction of the incongruent gene trees.

First, we found a clear, monotonic increase in recovery of congruent gene trees with the length of the corresponding genes (Fig. 2C), as expected for algorithmic accuracy based on simulation studies (Huelsenbeck 1995). For the length of a typical gene alignment (940 ungapped nucleotides), all methods showed accuracies around 40%. These were as low as 25% for shorter genes (<800 nt) and rose up to 60% for the longest genes (>2300 nt, corresponding to <10% of genes). The observed recovery vs length correlation continued with increasing alignment lengths (90% for 20,000 nt, obtained by concatenating 20 randomly chosen genes) in agreement with lengths typically recommended to produce accurate species trees (Rokas et al. 2003; Ciccarelli et al.

Downloaded from www.genome.org on April 28, 2008 - Published by Cold Spring Harbor Laboratory Press

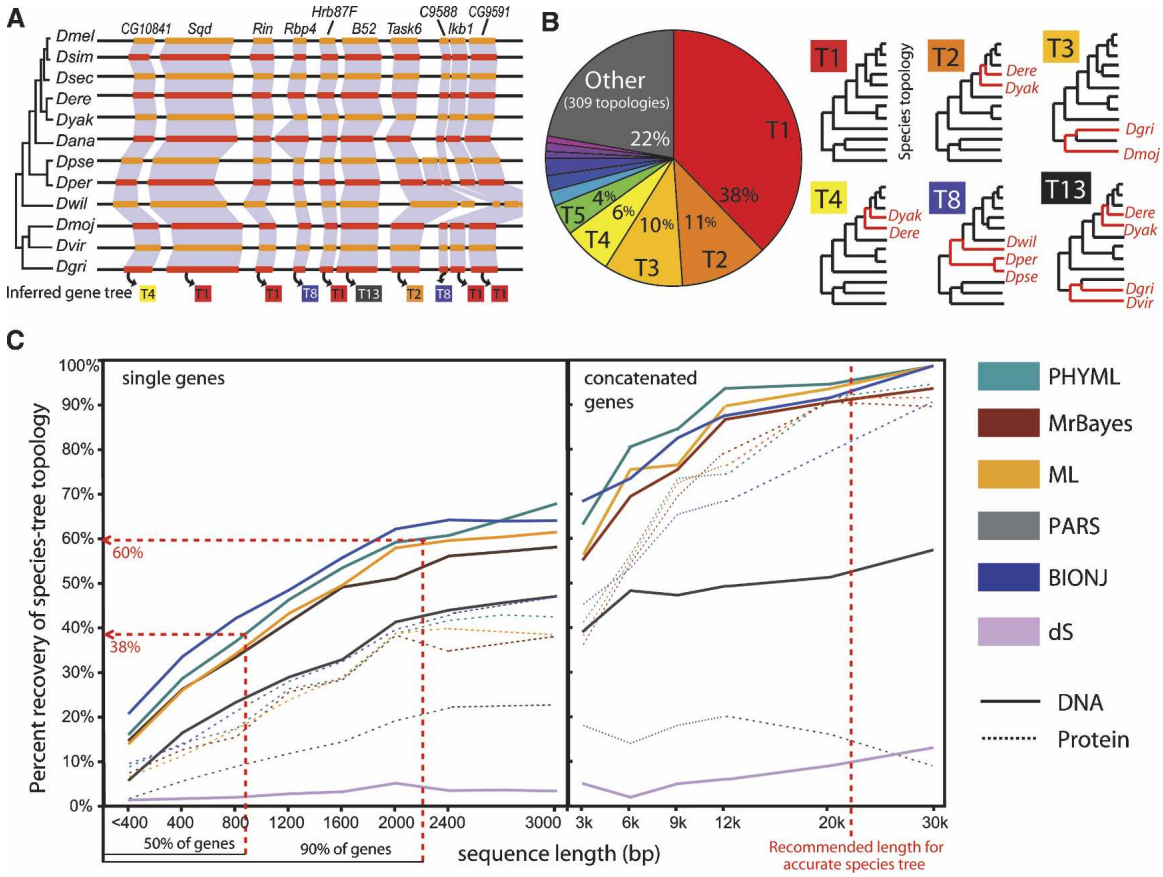


Figure 2. Large-scale gene-tree incongruence correlates with gene length. (*A*) Unambiguously orthologous genes in syntenic regions show diverse PHYML trees, even for consecutive genes (topologies numbered according to their genome-wide frequency). (*B*) Frequency and topology of most abundant ML gene trees for syntenic orthologs across 12 fly genomes. Discrepancies from the species topology (red branches) correspond to rotations of short internal branches. (*C*) Percentage of gene trees congruent to the species phylogeny correlates with gene length, regardless of the method used. DNA-based reconstruction, which uses threefold as many aligned positions, consistently outperformed protein-based reconstruction across all methods.

2006). Of course, such lengths are unrealistic for individual genes, and concatenation is not an option for accurate gene-tree reconstruction.

Second, we found that genes with moderate divergence rates showed the highest performance, while most errors were found in very slow and very fast evolving genes. Reconstruction accuracy peaked for genes with 40%–50% sequence identity (reaching 48% accuracy), but was significantly reduced for slower evolving genes (25% accuracy for 70% identity) or faster evolving genes (35% accuracy for 20% identity; Supplemental Fig. S6). This can also be attributed to a lack of phylogenetically informative sites in slow-evolving genes (lacking sufficient events to resolve phylogenetic divergence order) and also in fast-evolving genes (as sites with many independent substitutions do not distinguish between different topologies). In contrast, incomplete lineage sorting is not expected to show such correlations.

Third, simulated phylogenies with the known species topology and similar branch lengths resulted in the same alternate topologies T1–T5 at comparable frequencies (e.g., 4%–11% vs. 3%–5% for PHYML in flies, Supplemental Fig. S3C), suggesting that even the most frequent incongruent topologies may result from reconstruction errors. In fact, the frequency of T2 + T4 corresponding to previously reported incomplete lineage sorting (Pollard et al. 2006; Wong et al. 2007) only differed by 8% between simulation and real data (9% vs. 17%), providing an estimate of the extent of incomplete lineage sorting. The correct phylogeny was recovered for 72% of simulated gene trees on average (an ~30% increase over real data), potentially reflecting reduced discrepancies from model assumptions in simulated alignments (since the same model of evolution was used for reconstruction and simulation, while real alignments may violate this model), and potentially attributable to incomplete lineage sorting in true phylogenies. However, even if the increase is entirely due to incomplete lineage sorting in true phylogenies, it would only explain incongruences in, at most, 30% of trees, while 62% of fly trees and 76% of fungal trees were found to be incongruent. Thus, a significant portion of incongruences are likely due to reconstruction inaccuracies.

Lastly, if alternate topologies were due to biological reasons rather than methodological inaccuracies, we would expect them to be recovered with multiple methods, show high bootstrap support, and have significantly higher likelihood, neither of which was the case. In fact, the frequencies of T2–T5 were reduced from 4%–11% to 1%–5% when all methods were required to agree (Supplemental Fig. S3D), and the phylogenetic trees that disagreed with the species topology showed significantly lower bootstrap support values (Supplemental Fig. S17). In fact, amongst the 3102 gene trees where an alternative topology was

Downloaded from www.genome.org on April 28, 2008 - Published by Cold Spring Harbor Laboratory Press

selected by PHYML, only 5.7% of these had a significantly higher likelihood than the topology congruent to the species tree (SH test $P < 0.01$) (PAUP) (Shimodaira and Hasegawa 1999), suggesting that many of these alternative topologies have insufficient support.

We conclude that a significant fraction of observed phylogenetic incongruences are due to inaccuracies in phylogenetic reconstruction (attributable to a lack of informative sites in the typical gene), and that additional information is necessary to increase the accuracy of gene-tree reconstruction. When gene trees are studied in isolation, it is likely that such information may not exist. However, in the phylogenomic setting, where thousands of gene trees involve only a relatively small number of species, there is an opportunity to learn common features shared among different gene trees, which can be used to guide gene-tree reconstruction. In the following section, we study fly and fungal gene trees and propose a model capturing their common properties. We then develop a novel inference algorithm that can use this information for accurate gene-tree reconstruction in the phylogenomic setting.

Gene- and species-specific substitution rates in phylogenomics

To take advantage of the phylogenomic setting, we sought to capture the fact that thousands of gene trees all evolve within the same species tree, and explicitly model their common properties. We expressed the substitution rate b_i of each gene in each lineage as the product of two independent rates (Fig. 3A): a gene-specific substitution rate g dictated by the selective constraints imposed on the function of the gene (Bromham and Penny 2003), and a species-specific substitution rate s_i dictated by the time interval and evolutionary dynamics of each lineage i (e.g., population size, generation time, mating behavior, overall mutation rate). Our gene-specific rate is similar to site-specific scaling factors used in previous studies (Yang 1994; Felsenstein and Churchill 1996; Siepel et al. 2005; Kim and Pritchard 2007), and the independence of the two rates agrees with recently reported correlations in mammals and hominids (Cooper et al. 2003; Chimpanzee Sequencing and Analysis Consortium 2005).

To derive the properties and specific distributions for gene- and species-specific substitution rates, we revisited our 5154 syntenic fly orthologs, this time requiring each gene-tree topology to be congruent to the species-tree topology, and inferring branch lengths from pairwise distances by maximum likelihood (see Methods). From our model definition, we expect the gene rate to be proportional to the total branch length for any tree. Thus, for each resulting gene tree, we can estimate the gene rate g as the sum of all "absolute" branch lengths b_i, (representing the overall substitution rate

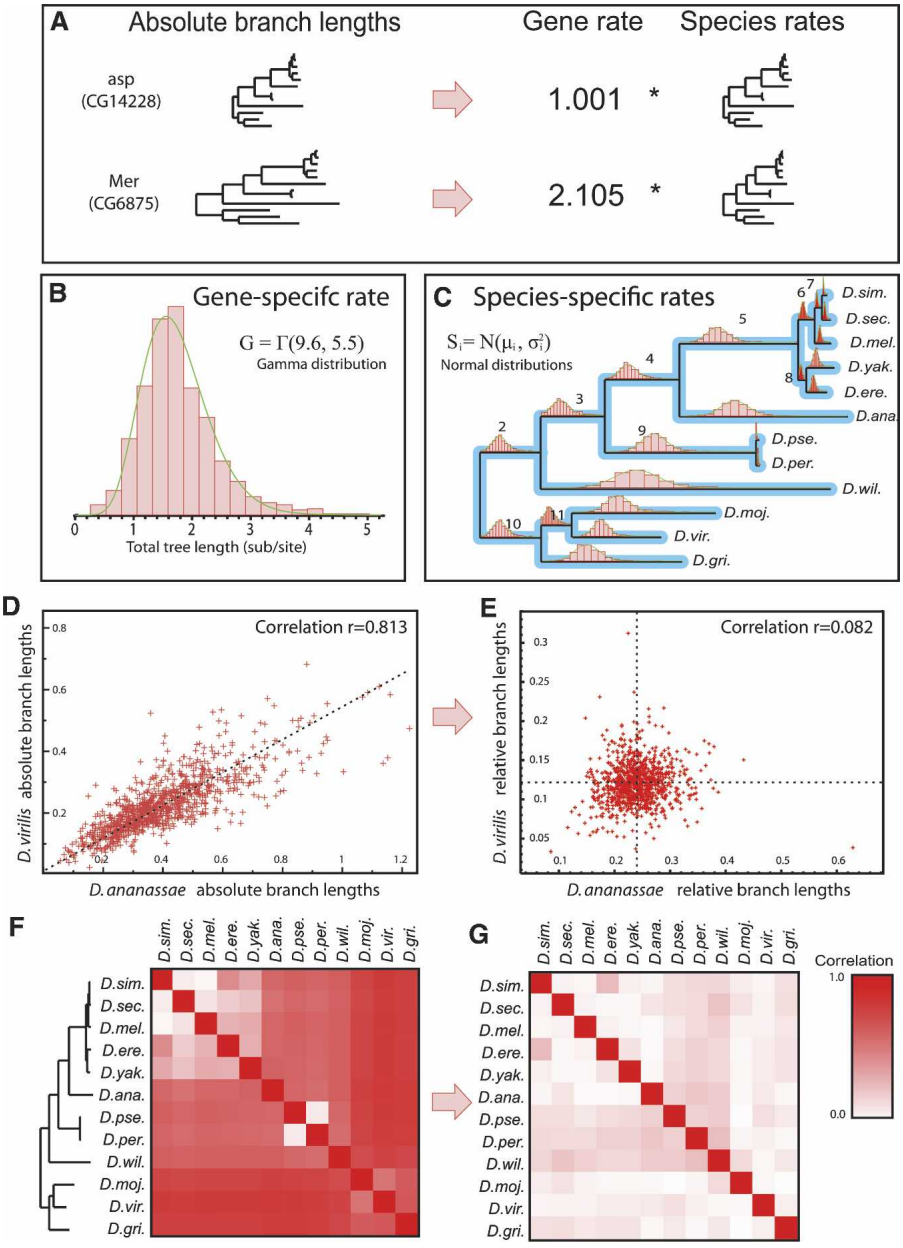


Figure 3. Evolutionary rates decoupled into gene-specific and species-specific components. (*A*) Syntenic ortholog trees appear as scaled versions of a common species tree, and can be expressed as the product of a gene-specific rate and species-specific rates. (*B*) Gene-specific rates of 5154 fly orthologs follow a gamma distribution. (*C*) Species-specific rates for each lineage follow normal distributions. Means and standard deviations shown in Supplemental Figure S7. (*D*) Unnormalized (absolute) branch lengths are highly correlated. Lengths for *D. virilis* and *D. ananassae* since their last common ancestor across the 5154 orthologs show correlation r = 0.813. (*E*) Relative branch lengths become independent after normalization by the gene-specific rate (r = 0.082). (*F*) Correlations are high for all species pairs before normalization, except for very closely related species. (*G*) Relative lengths are uncorrelated for all species pairs, showing that gene-specific rate accounts for their initial dependencies.

Downloaded from www.genome.org on April 28, 2008 - Published by Cold Spring Harbor Laboratory Press

across the entire tree), and the species-specific rate s_i for each branch as its "relative" length b_i/g after normalization by the gene rate (representing the fraction of substitutions attributable to that lineage). We found that the gene rate g was distributed as a gamma distribution (Fig. 3B), as expected for a rate (Uzzell and Corbin 1971). We also found that each s_i was distributed as a normal distribution (Fig. 3C), reflecting small fluctuations around the expected mean rate μ_i for each branch, given the stochastic nature of nucleotide substitution. Relative branch lengths b_i/g showed tighter distributions than absolute branch lengths b_i (typically with standard deviations between 1/3 and 1/4 of the mean; Supplemental Figs. S7–S12).

Our model makes the assumption that species-specific rates are independent of each other and of the gene rate. This assumption implies specific properties of gene trees, which we found to hold in the fungal and fly genomes. First, we would expect gene trees to be uniformly longer or shorter across different orthologs, appearing as scaled versions of an average gene tree, due to the common species rates s_i across gene trees. Indeed, pairs of gene trees showed strong correlations to each other: For example, *Merlin* and *abnormal spindle* showed correlation r = 0.96 (Fig. 3A), and 93% of genes showed correlations above r = 0.8 to the average gene tree (tree with average branch lengths across all genes) (Supplemental Fig. S14). Second, we would expect strong correlations between the absolute branch lengths from any pair of species, stemming from the common gene rate g. Indeed, the average pairwise correlation of absolute branch lengths was 0.61 across all pairs of species (Fig. 3F). Lastly, we would expect relative branch lengths, representing species-specific rates, to be independent of each other if their correlation was truly due primarily to the common gene-specific rate g, and indeed, we found that the average pairwise correlation of relative branch lengths dropped to 0.09 after normalization by the gene rate (Fig. 3G). For example, the correlation between *Drosophila ananassae* and *Drosophila virilis* was r = 0.81 for absolute branch lengths and 0.082 for relative branch lengths after normalization (Fig. 3D,E). All of these relationships, reported here for flies, also held for mammals and fungi (see Supplemental material).

SPIDIR: A machine-learning framework for phylogenomic gene-tree reconstruction

Our results suggest that substitution rates can be decoupled into gene-specific and species-specific rates g and s_i, and that these are well approximated by independent gamma and normal distributions, respectively. Based on these properties, we develop a generative model for gene-tree evolution across multiple complete genomes: A gene tree is generated as the product of a gene-specific rate g, sampled from a gamma distribution $g \sim G = \Gamma(\alpha, \beta)$, and a species-specific rate s_i for each lineage, sampled from a normal distrtibution $s_i \sim S_i = N(\mu_i, \sigma_i^2)$, each distributed independently of each other.

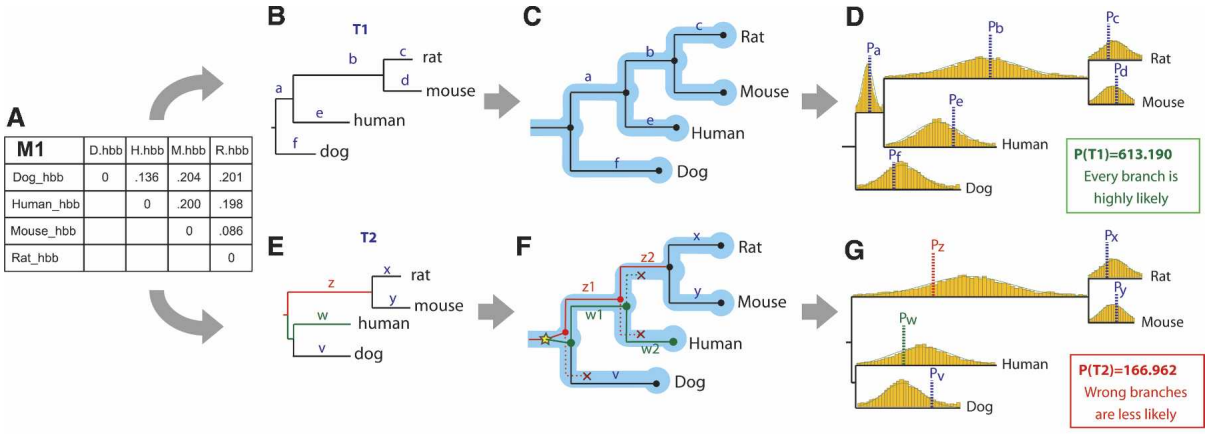
We used this generative model to develop a novel phylogenetic reconstruction method, called SPIDIR, for SPecies-Informed DIstance-based Reconstruction. Similarly to other likelihood-based methods, we search through a large number of gene tree topologies (Huelsenbeck and Ronquist 2001; Guindon and Gascuel 2003), evaluate the likelihood of each, and guide the search toward a maximum-likelihood tree. In contrast to existing methods, both phylogenetic (Gascuel 1997; Guindon and Gascuel 2003; Ronquist and Huelsenbeck 2003; Felsenstein 2005) and phylogenomic (Ma et al. 2000; Zmasek and Eddy 2002; Storm and Sonnhammer 2003; Arvestad et al. 2004; Dufayard et al. 2005; Durand et al. 2006), our algorithm works in two stages, first learning a model of gene and species evolution based on unambiguous orthologs, and then using this model for gene-tree reconstruction.

In the first stage (learning), we estimate the parameters of gene- and species-rate distributions based on alignments of unambiguous one-to-one orthologs across the species compared. As we focus on gene-tree reconstruction, we assume that the species tree is known, or can be reliably inferred using genome-scale information (e.g., using multigene analyses) (Rokas et al. 2003; Gadagkar et al. 2005; Ané et al. 2007; Edwards et al. 2007), or based on unique transposon insertion events (Kriegs et al. 2006). We also assume that a training set of genes with clear one-to-one orthology can be established, with phylogenies that are most likely congruent to the species tree (e.g., by using syntenic one-to-one orthologs, which, in the 12 fly species, include about one-third of all genes). Using the known species tree and multiple alignments of these unambiguous orthologs, we construct gene trees that are congruent to the species topology and estimate their absolute branch lengths b_i using least-square error (see Methods). As each gene tree has exactly one gene from each species, we use the total tree length $\Sigma_i(b_i)$ as an estimate of the gene rate g, and the relative branch lengths b_i/g as estimates of individual s_i (see Methods). This results in thousands of g and s_i estimates, to which we fit gamma and normal distributions, respectively, to infer $(\alpha, \beta, \mu_i, \sigma_i)$ parameters.

In the second stage (inference), we use our model to reconstruct phylogenies of the remaining genes, which may contain duplication and loss events (Fig. 4). In this work, we use our model for distance-based reconstruction, and thus, the input for the inference stage is a pairwise distance matrix M (Fig. 4A) inferred from multiple sequence alignments of the genes in question (extensions directly incorporating sequence characters are possible, but will be the subject of future work). We then search across many proposed topologies to find the maximum-likelihood gene tree. For each proposed gene-tree topology T, branch lengths b are estimated from the distance matrix M using least-square error (Bryant and Waddell 1998) (Fig. 4B,E), and the likelihoods of these branch lengths b_i are calculated according to our model, based on the learned parameters $(\alpha, \beta, \mu_i, \sigma_i)$. When the proposed gene tree is congruent to the species tree (Fig. 4B–D), the rate estimation and probability calculation are straightforward: The probability of the observed branch lengths is simply the product of probabilities of the overall gene rate and the observed relative branch lengths: $P(b|G,S) = P(g|G) \prod_i [P(b_i/g|s_i)]$, each gene-tree branch b_i uniquely mapping to a species-tree branch. When the proposed gene tree contains duplications and losses (Fig. 4E–G), our gene-rate estimate accounts for the missing data (see Methods), and the probability of relative branch lengths is estimated according to derived rate distributions, possibly spanning multiple species branches (see Methods). The nature of gene and species rate distributions within our model enables efficient likelihood computations for any gene-tree to species-tree reconciliation (see Methods).

As each tree is evaluated according to the distributions learned across the genomes, this framework allows us to distinguish between gene trees with unlikely branch lengths and gene trees whose observed branch lengths fit the learned distributions. For example, the correct gene-tree topology T1 (Kriegs et al. 2006) for orthologous mammalian hemoglobin-beta genes showed more than a 3.5-fold higher likelihood than an alterna-

Downloaded from www.genome.org on April 28, 2008 - Published by Cold Spring Harbor Laboratory Press



H

Method	Topology chosen	Ratio	Result	T1	T2
SPIDIR likelihood	T1	3.7	Correct	**613.190**	166.962
Max likelihood	T2	119.9	Wrong	-1689.5	**-1687.4**
Neighbor Joining	T2	8.0	Wrong	128	**828**
Parsimony	T2	1.2	Wrong	307	**300**

I

Species branch	Expected Length	Expected Stdev	Topology 1 Branch	Topology 1 Length	Topology 1 Likelihood	Topology 2 Branch	Topology 2 Length	Topology 2 Likelihood	Likelihood comparison
mouse	0.079	0.065	**d**	0.083	6.133	**y**	0.074	6.132	1.00
rat	0.084	0.068	**c**	0.060	5.510	**x**	0.053	5.299	1.04
human	0.172	0.083	**e**	0.241	3.403				1.04
hmr+human	0.279	0.099				**w**	0.215	3.261	
dog	0.107	0.054				**v**	0.179	3.025	1.22
dog+hmr	0.215	0.107	**a+f**	0.199	3.684				
mr	0.314	0.108	**b**	0.325	3.676				
hmr+mr	0.422	0.121				**z**	0.290	1.830	2.01
Family rate	alpha=1.31	b=4.95	rate=0.602		0.394	rate=0.675		0.285	1.38
Total					**613.190**			166.962	3.67

Figure 4. Evaluating gene-tree likelihood using learned rate distributions. (*A*) Observed distance matrix for mammalian orthologs of hemoglobin-β estimated from an HKY model based on multiple alignments of the four genes. (*B–D*) Likelihood evaluation for proposed topology T1. (*B*) Distance matrix M1 is mapped onto the proposed topology T1, resulting in branch lengths a–f. (*C*) Gene-tree branches are mapped to species-tree branches by reconciliation. Since the gene-tree topology is congruent to the species tree, each branch is mapped to exactly one lineage. (*D*) The probability of each branch length is evaluated based on species-specific rate distributions. T1 results in overall high-likelihood density, since the resulting relative branch lengths a–f fall near the average rate for the corresponding species-specific distribution (dotted lines). (*E–G*) Likelihood evaluation for proposed topology T2. (*E*) Distance matrix M1 is mapped onto the proposed topology T2, resulting in branch lengths v–z. (*F*) Reconciliation results in one gene duplication and three gene losses; gene-tree branches w and z now span two species-tree branches each and are evaluated based on accordingly longer species-tree rate distributions obtained by summing two normals. (*G*) The resulting branch lengths *z, w,* and *v* show large discrepancies from the average species-rate distributions, resulting in a 3.7-fold lower likelihood for branch lengths corresponding to the incorrect topology T2. (*H*) All other methods select the incorrect topology T2 due to long-branch attraction, even though the hemoglobin-β genes are unambiguous one-to-one orthologs and should follow the known mammalian phylogeny T1. (*I*) Branch-level comparison of likelihood scores shows consistently higher scores for T1, the correct topology. Notice that the gene-rate likelihood for T1 is different from that for T2, as the two topologies imply different gene family rates.

tive topology T2 (Fig. 4), each branch providing a much closer fit to the expected rate distributions, and thus resulting in consistently higher likelihood values (Fig. 4I): The ancestral rodent branch alone showed a twofold increase in likelihood for the correct topology, as the observed length *b* is much closer to the mean of the corresponding distribution, while the corresponding branch length *z* in the alternate topology is significantly shorter than would be expected if the gene-tree topology was truly T2. In contrast, all traditional methods systematically selected the incorrect topology T2 for the hemoglobin-beta genes because of long-branch attraction due to the faster-evolving rodent branch (tree rooted by fish ortholog): Neighbor-joining showed eightfold higher bootstrap support for T2, parsimony showed a slight preference for T2, and traditional maximum likelihood showed 100-fold higher likelihood for T2 (Fig. 4H; Supplemental Fig. S18). This effect has also been observed for other genes (Cannarozzi et al. 2006). Our method was able to resolve the correct topology T1 because it expected a longer branch for the rodent lineage, as they have an overall longer species-rate distribution S_i.

When paralogs were compared, and the correct gene-tree topology differed from the species-tree topology, our method again led to the correct answer (Supplemental Fig. S16): Comparing rodent hemoglobin-beta to the paralogous human and dog hemoglobin-alpha correctly resulted in T2, since hemoglobin-alpha and -beta are paralogs resulting from an ancestral duplication well before the mammalian speciation. In this case, the likelihood of topology T1 dropped 50-fold, while the likelihood of topology T2 increased, leading to 14-fold higher likelihood for T2. Thus, our method was not biased to always select the species topology when the correct gene-tree topology differed, and was able to resolve paralogous gene trees even in the presence of gene duplication and loss.

Learning species-specific rates leads to increased accuracy

We implemented and tested two versions of SPIDIR (available online at http://compbio.mit.edu/spidir), one using solely rate-based information to guide the reconstruction without penalizing duplication and loss, and one with an explicit penalty for duplication, similar in nature to those described elsewhere (Goodman et al. 1979; Page and Charleston 1997; Durand et al. 2006) and derived from previously reported gene duplication rates of 0.0023 and 0.0013 dup/gene/myr (Lynch and Conery 2000; Hahn et al. 2007) (see Methods).

We tested both versions extensively using the 12 fly and nine fungal genomes. We trained our evolutionary models using 500 randomly selected fly trees and 200 fungal trees, and used the remaining 4654 fly trees and 539 fungal trees to test our performance compared with existing algorithms. As the vast majority of these gene trees is likely congruent to the species tree, we

Downloaded from www.genome.org on April 28, 2008 - Published by Cold Spring Harbor Laboratory Press

evaluated accuracy as the ability to recover the expected gene tree in each case. We evaluated accuracy separately for the nine fly genomes with >7× sequence coverage from a single strain, and for the complete set of 12 fly species, which includes two species sequenced at 3× coverage and one mosaic genome assembled using seven different strains (*Drosophila* 12 Genomes Consortium 2007) (see Methods).

We found significantly increased performance over existing methods for both flies and fungi, and for both single-copy and duplicated genes (Fig. 5A). For the nine fly genomes, SPIDIR recovered the correct gene tree for 62% of genes, significantly higher than the leading existing reconstruction methods (BIONJ at 48% and PHYML at 40%); this increased to 74% and 86% with inclusion of an explicit parameter for a gene-duplication probability (0.5 and 0.1, respectively). For the full set of 12 genomes, SPIDIR also showed a clear improvement (5% over BIONJ and 9% over PHYML), although the low-coverage lineages showed increased reconstruction errors (Supplemental Fig. S3B), which may be due to sequencing errors affecting our rate estimates for the very short branches of low-coverage species. For the nine fungi, SPIDIR recovered the correct gene tree for 42% of orthologs, a 10% increase over MrBayes, and 18% increase over PHYML, the leading existing methods; again, this increased to 62% and 78% with use of an explicit duplication parameter of 0.5 and 0.1. Lastly, we used doubly syntenic orthologs arising from whole-genome duplication ("ohnologs") (Wolfe and Shields 1997; Kellis et al. 2004), to test SPIDIR's ability to capture gene duplication and loss, inferring model parameters from 739 single-copy syntenic genes and testing performance on 138 duplication-containing gene families; again, we found a 10% improvement on the correct placement of each duplicate pair, which increased by an additional 10% with inclusion of an explicit duplication parameter. Each of these performance improvements was also seen for partial correctness of the gene tree, measured using Robinson-Foulds error (Robinson and Foulds 1981) (Supplemental Fig. S5).

Moreover, SPIDIR accuracy correlated with the number of informative sites, suggesting that it uses available information fully: Performance monotonically increased with increasing gene lengths (Fig. 5B), and peaked for genes with moderate sequence

A

	Syntenic ortholog accuracy			Yeast whole genome duplication accuracy				
program	9 Flies	12 Flies	9 Fungi	pre-dup	s.stricto	S.cer.	S.cas.	C.gla.
SPIDIR	62.2%	47.3%	41.5%	56.2%	67.4%	56.2%	43.7%	60.7%
SPIDIR (D=.5)	74.0%	58.0%	61.8%	76.5%	71.9%	63.6%	48.4%	65.9%
SPIDIR (D=.1)	85.6%	69.5%	77.8%	83.7%	78.5%	68.1%	63.0%	75.6%
PHYML	40.7%	37.9%	23.5%	47.4%	60.7%	45.1%	30.3%	46.6%
BIONJ	47.4%	41.9%	29.6%	49.6%	54.0%	42.2%	28.8%	48.1%
MrBayes	40.1%	34.3%	31.4%	54.1%	57.8%	43.0%	28.9%	45.9%
DNAPARS	34.9%	23.5%	22.2%	40.7%	31.1%	25.2%	28.9%	28.9%

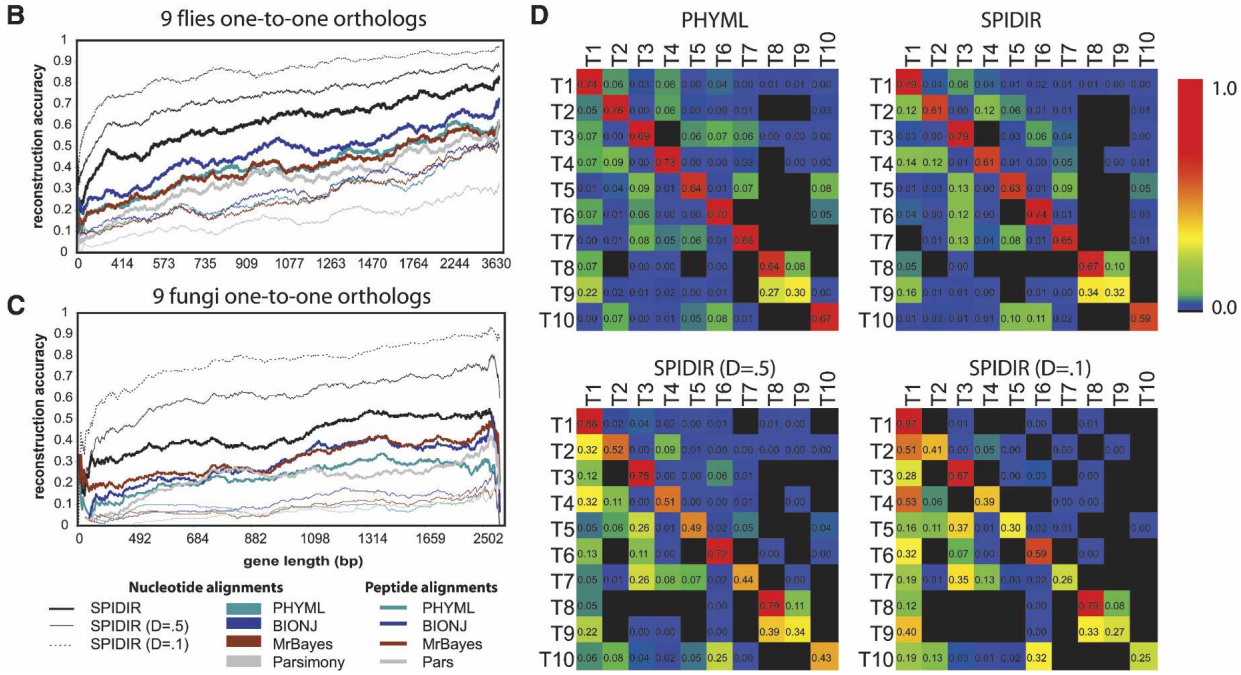


Figure 5. SPIDIR learning methodology leads to significantly higher accuracy. (*A*) Comparison of SPIDIR and several popular phylogenetic methods for syntenic orthologs in nine flies, 12 flies, and nine fungi, and for duplicate genes arising from whole-genome duplication. "pre-dup" gives the accuracy of reconstructing the topology of the three preduplicated species, "*s. stricto*" is the topology accuracy of only the four *sensu strict* species, and "*S. cer*", "*S. cas*", and "*C. gla*" give the accuracy of placing annotated ohnologs of each species on opposing sides of the whole-genome duplication node. (*B*) Reconstruction accuracy for SPIDIR correlates with gene length, similarly to other methods, and is consistently higher. (*C*) Reconstruction accuracy for SPIDIR and PHYML for ROSE simulated fly alignments according to 10 most frequent topologies T1–T10 (*left*). Both methods are unbiased and recover alternate topologies at similar rates (*top*), although SPIDIR was trained on T1. With increasing duplication cost, T1 becomes favored by SPIDIR (D = 0.5 and 0.1).

Downloaded from www.genome.org on April 28, 2008 - Published by Cold Spring Harbor Laboratory Press

divergence (Supplemental Fig. S6), surpassing existing methods for all length and divergence intervals. In addition, we found that reconstruction accuracy was consistently high, regardless of the gene function: Of 3700 GO terms, SPIDIR had higher reconstruction accuracy than PHYML for 3469 (93%), and of the remaining 231 GO terms (7%), none showed significant enrichment for alternate topologies ($P > 0.185$ hypergeometric). This suggests that the evolutionary parameters learned in our training set held across all genes tested, regardless of their specific function.

Finally, we found that our method showed no systematic biases toward the species topology. We simulated evolution according to the 10 most frequent ML topologies T1–T10, and asked which topology was inferred by the different prediction algorithms, summarizing the results in a "confusion matrix" (Fig. 5C; Methods). We found that T1 constituted only 6.2% of the SPIDIR-inferred trees for simulated topologies T2–T10, which is similar to PHYML (also 6.2%), confirming that our method is not biased. With duplication and loss penalties, the percentage of T1 increased to 15% for D = 0.5 and 30% for D = 0.1, as expected, but this may be desirable when, in fact, the species tree is known. In addition, both SPIDIR and PHYML identified 62%–65% of all alternate topologies T2–T10 correctly, although SPIDIR was only trained on T1. These trends should be a necessary test for phylogenetic methods that use species-level information to ensure lack of systematic biases.

Discussion

We showed that gene trees are subject to two complementary forces of evolution: a gene-specific component, summarizing the selective pressures on individual gene functions, and a species-specific component, reflecting the divergence times and evolutionary dynamics of the species compared. We found gene- and species-specific substitution rates are independent and can be described by simple distributions that provide a very good fit to actual phylogenies. For both fly and fungal species, we found that a single gene rate was sufficient to model gene trees across the entire clades studied, although larger evolutionary distances and more diverse species groups may require modeling lineage-specific variations in this rate. More generally, we expect that the study of diverse groups of multiple complete genomes will reveal additional properties of gene and species phylogenies, enabling further increases in accuracy and potentially revealing new insights into gene evolution.

We used the decoupling of gene and species rates to introduce a novel approach for phylogenetic reconstruction that is specifically tailored for application in complete genomes. In contrast to existing methods, which treat each gene-tree reconstruction problem in isolation, our approach enables learning across hundreds of phylogenies to improve the accuracy of reconstructing any gene tree involving these species. We tested our method extensively and showed consistent improvements over existing methods for both flies and fungi, lack of bias with respect to the species topology, and increased performance across all lengths, functional categories, and in the presence of gene duplication and loss. Although we have applied our model solely for distance-based reconstruction, it is also applicable to character-based reconstruction. Specifically, our model can be viewed as specifying a prior probability on gene-tree branch lengths, which could replace the uniform branch length prior that is commonly used in maximum likelihood and Bayesian approaches, providing a promising direction for future development.

Several other models have been developed for modeling gene and species evolution simultaneously. One class of models has primarily addressed the inference of a species tree from many gene trees (Ané et al. 2007; Edwards et al. 2007), typically by considering only orthologous genes and assuming that every incongruent node is due to deep coalescence and incomplete lineage sorting. The second class of models has addressed gene-tree reconstruction, typically by assuming that the prevalent reasons for incongruences are gene duplication and loss (Ma et al. 2000; Zmasek and Eddy 2002; Storm and Sonnhammer 2003; Arvestad et al. 2004; Dufayard et al. 2005; Durand et al. 2006). Our model fits within the second class, demonstrating the effectiveness of learning branch-length distributions for gene-tree reconstruction in a generative model for gene tree evolution. A potential future direction for both types of work may be a joint modeling of deep coalescence and gene duplication/loss.

The methodology introduced here, although general, allowed us to address the problem of long-branch attraction at the species level (Bergsten 2005; Philippe et al. 2005). It is known that when fast-evolving lineages are intermixed with slowly evolving lineages, the longer branches tend to cluster together and join further back in evolutionary time, due to increased rates of homoplasy in rapidly evolving lineages. However, addressing long-branch attraction is still a major challenge in phylogenetics. Our decoupling of evolutionary rates allows us to capture heterotachy at the species level, since fast-evolving lineages are uniformly faster across the entire genome. As illustrated in our mammalian example, our model can learn to expect longer branches for faster-evolving lineages, thus recovering the true topology even when all existing methods suffer from long-branch attraction.

Although in this study we have focused our attention on phylogenetic reconstruction accuracy, decoupling gene-specific and species-specific rates can also be used to identify unusual cases of evolutionary change. In particular, it enables us to distinguish whether a long branch is due to simply an overall faster gene rate, a fast-evolving species, or specific acceleration for a particular gene in a given lineage. This is applicable at the level of individual genes, or for sets of genes within a functional category, to recognize evolutionary adaptation of individual genes or pathways. Such studies of acceleration or deceleration can be coupled with studies of positive selection (e.g., K_a/K_s), to detect lineage-specific changes in selective pressures.

The *Saccharomycete* and *Drosophila* groups are only the first two in an increasingly long series of groups of related species scheduled for dense sequencing, including 32 mammals, five worms, dozens of fungi, hundreds of bacteria, and thousands of viruses. The increasing number of species in comparative studies should lead to increased power, both for biological signal discovery and for evolutionary studies, but these will require increasingly rigorous methods for genome comparison, which can scale to many species. Single-gene phylogenetic methods are unlikely to scale reliably to dozens of species, while phylogenomic methods should benefit from the abundance of information in complete genomes. The methodologies presented here are general and likely to significantly contribute in the comparison and understanding of many complete genomes.

Methods

Genomic sequences

We selected the two largest groups of fully sequenced closely related species with long-range synteny across the entire group.

Downloaded from www.genome.org on April 28, 2008 - Published by Cold Spring Harbor Laboratory Press

The *Drosophila* genus includes *D. melanogaster* (Adams 2000), *D. pseudoobscura* (Richards 2005), and the 10 recently-sequenced species *D. sechellia, D. simulans, D. yakuba, D. erecta, D. ananassae, D. persimilis , D. willistoni, D. mojavensis, D. virilis,* and *D. grimshawi* (*Drosophila* 12 Genomes Consortium 2007). Our fungal clade includes nine species: (*Saccharomyces cerevisiae, S. paradoxus, S. mikatae, S. bayanus, S. castellii* (Cliften 2003; Kellis 2003), *Candida glabrata, Kluyveromyces lactis* (Dujon 2004), *Ashbya gossypii* (Dietrich 2004), and *Kluyveromyces waltii* (Kellis 2004).

Identifying syntenic orthologous genes

We constructed syntenic regions for the 12 flies by defining a syntenic block to be at least three genes within 200 kb of each other with no other blocks in between. For the fungal data set, we required at least three genes per block with a maximum uninterrupted gene separation of 20 kb. Syntenic blocks were filtered to keep only those containing exactly one gene from each species, thus removing potential segmental duplications. For our whole-genome duplication data set, we used *S. cerevisiae*, *S. castellii*, and *C. glabrata* ohnologs from the Yeast Gene Order Browser (YGOB) (Byrne and Wolfe 2005). Ohnologs were clustered by best reciprocal BLAST hits. Clusters were filtered such that exactly one ohnolog pair from each species is present. Clusters were extended to include genes from the remaining species using our syntenic alignments. Alignments were manually curated to remove possible gene conversion events.

Species phylogeny

The currently accepted fly species phylogeny (*Drosophila* 12 Genomes Consortium 2007) is shown in Supplemental Figure S1. The major features of the fungal phylogeny are also widely accepted (Rokas et al. 2003; Hittinger et al. 2004; Byrne and Wolfe 2005); however, there is less agreement on the branch orders of the preduplication species *K. waltii, K. lactis,* and *A. gossypii*. The branching order we used (Supplemental Fig. S1) is the most frequent topology for all methods (ML, MAP, MP) on nucleotide alignments.

Alignments

We study phylogenies of 5154 unambiguous fly orthologs and 739 unambiguous fungal orthologs. These genes are selected from regions of synteny that are filtered to be free of tandem duplications. For each of these ortholog sets, we produced multiple alignments of their protein sequences using MUSCLE (Edgar 2004). To attain nucleotide alignments, we map the nucleotide sequence on to peptide alignments, substituting every amino acid by the corresponding codon and every gap by a triplet of gaps.

Model parameter learning

For each ortholog alignment, a rooted tree congruent to the species was constructed and fitted with our implementation of least-square error on distances estimated by PUZZLE-TREE (Schmidt et al. 2002) using an HKY model. To be consistent, the root is placed at the midpoint of the rooting branch. The branches of each tree were normalized by the total tree length. To estimate the parameters of our model, the mean and variance of relative branches were calculated for each species ($4n-2$ parameters for n species), and the alpha- and beta-parameters for the gene-specific rate were calculated with maximum-likelihood estimates from the total absolute branch lengths.

Generative model of gene-tree evolution

To define a generative model for gene-tree evolution with an arbitrary number of duplications and losses, we use a more general definition of reconciliation than is commonly used (Goodman et al. 1979; Page 1994). We define a *reconciliation R* to be a mapping from gene nodes b_l to a species node i and duplication point k_l: $R(b_l) = (i, k_l)$.

If gene node b_l is a duplication, k_l defines the fraction along the species branch at which the gene duplication occurred: $k_l = \varepsilon$ if the duplication occurs immediately after speciation of species $parent(R(b_l))$, and $k_l = 1\text{-}\varepsilon$ if the duplication occurs immediately before species $R(b_l)$. If b_l represents a speciation, we define k_l to be 1. We define k_l to be distributed uniformly over (0,1), unless an ancestor b_{l2} of b_l reconciles to the same species with duplication point k_{l2}, in which case $k \sim Uniform(k_{l2}, 1)$.

For our model, we also define a reconciliation R_b that maps gene branches to species branches. One complication is that a gene branch may map to a path of species branches and may use only a portion of the starting and ending species branch. Thus, we define: $R_b(b_l) = ((s_1, s_2, \ldots, s_m), (p_1, p_2, \ldots, p_m))$, where the vector $s_1, \ldots, s_m$ defines the path of branches in the species tree and $p_1, \ldots, p_m$ defines the portion of each species branch used by $R_b(b_l)$. Notice that the internal branch portions $p_2, \ldots, p_{m-1}$, if they exist, are always 1. Defining duplication points k_l immediately imply the values of p_j, and vice versa (see Supplemental Methods).

Above, we have presented the generative model for a gene branch that reconciles to exactly one species branch $((s_i), (p_1 = 1))$, namely, $b_l \sim G\ Si = \Gamma(\alpha, \beta)\ N(\mu_i, \sigma_i^2)$, where G and S_i are the gene- and species-specific rates. Here, we specify how to generate a branch length b_l that reconciles across multiple species branches. We model such a b_l to be a product of a gene rate g and a relative rate x_l, that itself is the sum of m independent random variables y_j, each with the distribution: $y_j \sim N(p_j \mu_j, p_j \sigma_j^2)$. Thus, each branch b_l in a gene tree is distributed as: $b_l \sim G \sum_j y_j = \Gamma(\alpha, \beta)\ \sum_j N(p_j \mu_j, p_j \sigma_j^2) = \Gamma(\alpha, \beta)\ N(\sum_j p_j \mu_j, \sum_j p_j \sigma_j^2)$.

In the Supplemental Methods, we present an algorithm that uses this generative model to search for the gene-tree topology with maximum likelihood given its branch lengths, based on a heuristic search over gene-tree topologies.

Performance comparison

We compared our algorithm against a variety of the most popular and successful phylogeny programs. For a maximum-likelihood method, we used PHYML v2.4.4 (Guindon and Gascuel 2003). Nucleotides substitutions were modeled with the HKY model and peptide substitutions were modeled with JTT. For parsimony methods, we used PHYLIP's DNAPARS and PROTPARS programs. MrBayes v3.1.1, a Bayesian-based method, was used to find the maximum a posteriori phylogenetic tree (Ronquist and Huelsenbeck 2003). We used four chains, an automatic stop rule, a 25% burn-in, sampled every 10 generations from a total of 10,000 generations, a fixed BLOSUM model for peptides, a 4by4 model for nucleotides, and we ensured MrBayes reported the most likely binary tree. Lastly, we used the Neighbor-Joining program BIONJ (Gascuel 1997) with a variety of substitution models, including HKY built with PUZZLE-TREE (Schmidt et al. 2002), JTT built with PROTDIST (Felsenstein 2005), and dS built with the YN00 program from PAML v3.15 (Yang 1997). For all programs, unless stated, default options were used. Parsimony methods perform consistently worse than other methods (parsimony is not statistically consistent) (Felsenstein 1978), as well as the synonymous-substitution metric dS, which unfortunately saturates at the evolutionary distances studied.

Downloaded from www.genome.org on April 28, 2008 - Published by Cold Spring Harbor Laboratory Press

Duplication probability

Lynch and Conery (2000) and Hahn et al. (2007) estimated the rate of gene duplication to be 0.0023 dup/gene/myr and 0.0013 dup/gene/myr, respectively. Given that the fly tree has a depth of roughly 40 myr and a total length that is approximately six times that, we expect that a gene duplication occurs in any gene family with a probability of 0.5 (0.0023 * 40 * 6) and 0.312 (0.0013 * 40 * 6).

Simulation

We used simulated sequence evolution to evaluate the accuracy of all of the phylogenetic methods we tested. Unlike other phylogenetic methods, SPIDIR uses a generative model of gene-tree evolution to calculate the likelihood of a phylogeny. Therefore, we must simulate sequences such that they behave like real gene families. To do this, we combined an existing sequence simulation program, ROSE (Stoye et al. 1998), with our generative model of gene-tree evolution. If we use a model trained on fly one-to-one orthologs, we can create simulated fly gene trees that have the same gene-specific and species-specific substitution rates as real fly gene families.

For each simulation, we fix the desired topology (T1–T10). We then use our generative model to choose branch lengths for the topology as described in the generative model. In addition, if a negative branch length is generated, it is discarded and a new one is drawn from the distribution.

Once we have a tree with branch lengths, ROSE is used to simulate sequence evolution down each branch using the HKY model. Base frequencies (A = 0.258, C = 0.267, G = 0.266, T = 0.209), transition bias (3.18), gene length (861 bp), total tree length (1.82 sub/site), and alignment percent identity (0.368) are matched to that of real data.

Acknowledgments

We thank Marcia Lara, Antonis Rokas, and Bruce Birren at the Broad Institute for useful comments and feedback. We thank Mike Lin, Alex Stark, Joshua Grochow, and Radek Sklarczyk for help, advice, and comments. We are indebted to the fly community for early release of data and annotations for use in our benchmarks. We also thank the NIAID and our collaborators at the Broad Institute, and especially Bruce Birren and Christina Cuomo for discussions. This work was funded by the Ruth L. Kirschstein National Research Service Award (NRSA).

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Ané, C., Larget, B., Baum, D.A., Smith, S.D., and Rokas, A. 2007. Bayesian estimation of concordance among gene trees. *Mol. Biol. Evol.* **24:** 412–426.

Arvestad, L., Berglund, A., Lagergren, J., and Sennblad, B. 2004. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. In *Proceedings of the Eighth Annual International Conference on Computational Molecular Biology*, pp. 326–335.

Avise, J.C., Shapira, J., Daniel, S.W., Aquadro, C.F., and Lansman, R.A. 1983. Mitochondrial DNA differentiation during the speciation process in Peromyscus. *Mol. Biol. Evol.* **1:** 38–56.

Baldauf, S.L., Roger, A.J., Wenk-Siefert, I., and Doolittle, W.F. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290:** 972–977.

Bergsten, J. 2005. A review of long-branch attraction. *Cladistics* **21:** 163–193.

Bromham, L. and Penny, D. 2003. The modern molecular clock. *Nat. Rev. Genet.* **4:** 216–224.

Bryant, D. and Waddell, P. 1998. Rapid evaluation of least-squares and minimum-evolution criteria on phylogenetic trees. *Mol. Biol. Evol.* **15:** 1346–1359.

Bull, J.J., Badgett, M.R., Wichman, H.A., Huelsenbeck, J.P., Hillis, D.M., Gulati, A., Ho, C., and Molineux, I.J. 1997. Exceptional convergent evolution in a virus. *Genetics* **147:** 1497–1507.

Byrne, K.P. and Wolfe, K.H. 2005. The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.* **15:** 1456–1461.

Cannarozzi, G.M., Schneider, A., and Gonnet, G. 2006. A phylogenomic study of human, dog and mouse. *PLoS Computat. Biol.* e2.eor. doi: http://dx.doi.org/10.1371/journal.pcbi.0030002.eor.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437:** 69–87.

Ciccarelli, F.D., Doerks, T., v. Mering, C., Creevey, C.J., Snel, B., and Bork, P. 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science* **311:** 1283–1287.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A., and Johnston, M. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301:** 71–76.

Cooper, G.M., Brudno, M., Program, N.I.S.C.C.S., Green, E.D., Batzoglou, S., and Sidow, A. 2003. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res.* **13:** 813–820.

Dehal, P. and Boore, J.L. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* **3:** e314. doi: 10.1371/journal.pbio.0030314.

Dickerson, R.E. 1971. The structures of cytochrome c and the rates of molecular evolution. *J. Mol. Evol.* **1:** 26–45.

Dietrich, F.S., Voegeli, S., Brachat, S., Lerch, A., Gates, K., Steiner, S., Mohr, C., Pöhlmann, R., Luedi, P., Choi, S., et al. 2004. Comparison of the sequence of laboratory yeast to that of a distant relative, a small filamentous fungus, reveals that yeast evolved via genome duplication and separate evolution of homologous genes. *Science* **9:** 304–307.

Drosophila 12 Genomes Consortium. 2007. Genomics on phylogeny: Evolution of genes and genomes in the genus *Drosophila*. *Nature* (in press). doi: 10.1038/nature06341.

Dufayard, J.-F., Duret, L., Penel, S., Gouy, M., Rechenmann, F., and Perriere, G. 2005. Tree pattern matching in phylogenetic trees: Automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* **21:** 2596–2603.

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., de Montigny, J., Marck, C., Neuvéglise, C., Talla, E., et al. 2004. Genome evolution in yeasts. *Nature* 35–44.

Durand, D., Halldorsson, B.V., and Vernot, B. 2006. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J. Comput. Biol.* **13:** 320–335.

Edgar, R.C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32:** 1792–1797.

Edwards, S.V., Liu, L., and Pearl, D.K. 2007. High-resolution species trees without concatenation. *Proc. Natl. Acad. Sci.* **104:** 5936–5941.

Eisen, J.A. 1998. Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* **8:** 163–167.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27:** 401–410.

Felsenstein, J. 2005. *PHYLIP (Phylogeny Inference Package) version 3.6*. Distributed by the author. Department of Genome Sciences, University of Washington,

Felsenstein, J. and Churchill, G.A. 1996. A hidden Markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* **13:** 93–104.

Fitch, W.M. 1970. Distinguishing homologous from analogous proteins. *Syst. Zool.* **19:** 99–113.

Gadagkar, S.R., Rosenberg, M.S., and Kumar, S. 2005. Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. *J. Exp. Zoolog. B Mol. Dev. Evol.* **304:** 64–74.

Gascuel, O. 1997. BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* **14:** 685–695.

Goodman, M., Czelusniak, J., Moore, G., Romero-Herrera, A., and Matsuda, G. 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool.* **28:** 132–163.

Guindon, S. and Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52:** 696–704.

Downloaded from www.genome.org on April 28, 2008 - Published by Cold Spring Harbor Laboratory Press

Hahn, M. 2007. Bias in phylogenetic tree reconciliation methods: Implications for vertebrate genome evolution. *Genome Biol.* **8:** R141. doi: 10.1186/gb-2007-8-7-r141.
Hahn, M.W., Han, M.V., and Han, S.-G. 2007. Gene family evolution across 12 *Drosophila* genomes. *PLoS Genet.* **3:** e197.
Hasegawa, M., Kishino, H., and Yano, T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22:** 160–174.
Hillis, D.M., Huelsenbeck, J.P., and Cunningham, C.W. 1994. Application and accuracy of molecular phylogenies. *Science* **264:** 671–677.
Hittinger, C.T., Rokas, A., and Carroll, S.B. 2004. Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts. *Proc. Natl. Acad. Sci.* **101:** 14144–14149.
Huelsenbeck, J. 1995. The robustness of two phylogenetic methods: Four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. *Mol. Biol. Evol.* **12:** 843–849.
Huelsenbeck, J.P. and Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17:** 754–755.
Huerta-Cepas, J., Dopazo, H., Dopazo, J., and Gabaldon, T. 2007. The human phylome. *Genome Biol.* **8:** R109. doi: 10.1186/gb-2007-8-6-r109.
Jaillon, O., Aury, J.M., Brunet, F., Petit, J.L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., et al. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* **431:** 946–957.
Jones, D.T., Taylor, W.R., and Thornton, J.M. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8:** 275–282.
Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241–254.
Kellis, M., Birren, B.W., and Lander, E.S. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428:** 617–624.
Kim, S.Y. and Pritchard, J.K. 2007. Adaptive evolution of conserved noncoding elements in mammals. *PLoS Genet.* **3:** e147. doi: 10.1371/journal.pgen.0030147.
Koonin, E.V., Makarova, K.S., and Aravind, L. 2001. Horizontal gene transfer in prokaryotes: Quantification and classification. *Annu. Rev. Microbiol.* **55:** 709–742.
Kriegs, J.O., Churakov, G., Kiefmann, M., Jordan, U., Brosius, J., and Schmitz, J. 2006. Retroposed elements as archives for the evolutionary history of placental mammals. *PLoS Biol.* **4:** e91. doi: 10.1371/journal.pbio.0040091.
Kuhner, M.K. and Felsenstein, J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* **11:** 459–468.
Li, H., Coghlan, A., Ruan, J., Coin, L.J., Heriche, J.-K., Osmotherly, L., Li, R., Liu, T., Zhang, Z., Bolund, L., et al. 2006. TreeFam: A curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.* **34:** D572–D580. doi: 10.1093/nar/gkj118.
Lynch, M. and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290:** 1151–1155.
Ma, B., Kowloon, H., Li, M., Waterloo, O., Zhang, L., and Center, B. 2000. From gene trees to species trees. *SIAM J. Comput.* **30:** 729–752.
Miller, W., Makova, K.D., Nekrutenko, A., and Hardison, R.C. 2004. Comparative genomics. *Annu. Rev. Genomics Hum. Genet.* **5:** 15–56.
Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Murphy, W.J., Eizirik, E., Johnson, W.E., Zhang, Y.P., Ryder, O.A., and O'Brien, S.J. 2001. Molecular phylogenetics and the origins of placental mammals. *Nature* **409:** 614–618.
Ohta, T. and Kimura, M. 1971. On the constancy of the evolutionary rate of cistrons. *J. Mol. Evol.* **1:** 18–25.
Page, R. 1994. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst. Biol.* **43:** 58–77.
Page, R.D. and Charleston, M.A. 1997. From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem. *Mol. Phylogenet. Evol.* **7:** 231–240.
Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., and Delsuc, F. 2005. Heterotachy and long-branch attraction in phylogenetics. *BMC Evol. Biol.* **5:** 50. doi: 10.1186/1471-2148-5-50.
Pollard, D.A., Iyer, V.N., Moses, A.M., and Eisen, M.B. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: Evidence for incomplete lineage sorting. *PLoS Genet.* **2:** e173. doi: 10.1371/journal.pgen.0020173.
Richards, S., Liu, Y., Bettencourt, B.R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M.J., Chen, R., Meisel, R.P., et al. 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene, and *cis*-element evolution. *Genome Res.* **15:** 1–18.
Robinson, D.F. and Foulds, L.R. 1981. Comparison of phylogenetic trees. *Math. Biosci.* **53:** 131–147.
Rokas, A., Williams, B.L., King, N., and Carroll, S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425:** 798–804.
Ronquist, F. and Huelsenbeck, J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19:** 1572–1574.
Saitou, N. and Imanishi, T. 1989. Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. *Mol. Biol. Evol.* **6:** 514–525.
Schmidt, H.A., Strimmer, K., Vingron, M., and v. Haeseler, A. 2002. TREE-PUZZLE: Maximum-likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18:** 502–504.
Shimodaira, H. and Hasegawa, M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16:** 1114–1116.
Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15:** 1034–1050.
Stark, A., Lin, M.F., Kheradpour, P., Pedersen, J.S., Parts, L., Carlson, J.W., Crosby, M.A., Rasmussen, M.D., Roy, S., Deoras, A.N., et al. 2007. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* (in press).
Storm, C.E.V. and Sonnhammer, E.L.L. 2003. Comprehensive analysis of orthologous protein domains using the HOPS database. *Genome Res.* **13:** 2353–2362.
Stoye, J., Evers, D., and Meyer, F. 1998. Rose: Generating sequence families. *Bioinformatics* **14:** 157–163.
Tateno, Y., Takezaki, N., and Nei, M. 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Mol. Biol. Evol.* **11:** 261–277.
Ureta-Vidal, A., Ettwiller, L., and Birney, E. 2003. Comparative genomics: Genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet.* **4:** 251–262.
Uzzell, T. and Corbin, K.W. 1971. Fitting discrete probability distributions to evolutionary events. *Science* **172:** 1089–1096.
Woese, C.R., Kandler, O., and Wheelis, M.L. 1990. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci.* **87:** 4576–4579.
Wolfe, K.H. and Shields, D.C. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387:** 708–713.
Wong, A., Jensen, J.D., Pool, J.E., and Aquadro, C.F. 2007. Phylogenetic incongruence in the *Drosophila melanogaster* species group. *Mol. Phylogenet. Evol.* **43:** 1138–1150.
Woods, R., Schneider, D., Winkworth, C.L., Riley, M.A., and Lenski, R.E. 2006. Tests of parallel molecular evolution in a long-term experiment with *Escherichia coli*. *Proc. Natl. Acad. Sci.* **103:** 9107–9112.
Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.* **39:** 306–314.
Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13:** 555–556.
Zmasek, C.M. and Eddy, S.R. 2002. RIO: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics* **3:** 14. doi: 10.1186/1471-2105-3-14.

Received September 1, 2007; accepted in revised form October 16, 2007.

ARTICLES

Evolution of genes and genomes on the *Drosophila* phylogeny

Drosophila 12 Genomes Consortium*

Comparative analysis of multiple genomes in a phylogenetic framework dramatically improves the precision and sensitivity of evolutionary inference, producing more robust results than single-genome analyses can provide. The genomes of 12 *Drosophila* species, ten of which are presented here for the first time (*sechellia*, *simulans*, *yakuba*, *erecta*, *ananassae*, *persimilis*, *willistoni*, *mojavensis*, *virilis* and *grimshawi*), illustrate how rates and patterns of sequence divergence across taxa can illuminate evolutionary processes on a genomic scale. These genome sequences augment the formidable genetic tools that have made *Drosophila melanogaster* a pre-eminent model for animal genetics, and will further catalyse fundamental research on mechanisms of development, cell biology, genetics, disease, neurobiology, behaviour, physiology and evolution. Despite remarkable similarities among these *Drosophila* species, we identified many putatively non-neutral changes in protein-coding genes, non-coding RNA genes, and *cis*-regulatory regions. These may prove to underlie differences in the ecology and behaviour of these diverse species.

As one might expect from a genus with species living in deserts, in the tropics, on chains of volcanic islands and, often, commensally with humans, *Drosophila* species vary considerably in their morphology, ecology and behaviour[1]. Species in this genus span a wide range of global distributions: the 12 sequenced species originate from Africa, Asia, the Americas and the Pacific Islands, and also include cosmopolitan species that have colonized the planet (*D. melanogaster* and *D. simulans*) as well as closely related species that live on single islands (*D. sechellia*)[2]. A variety of behavioural strategies is also encompassed by the sequenced species, ranging in feeding habit from generalist, such as *D. ananassae*, to specialist, such as *D. sechellia*, which feeds on the fruit of a single plant species.

Despite this wealth of phenotypic diversity, *Drosophila* species share a distinctive body plan and life cycle. Although only *D. melanogaster* has been extensively characterized, it seems that the most important aspects of the cellular, molecular and developmental biology of these species are well conserved. Thus, in addition to providing an extensive resource for the study of the relationship between sequence and phenotypic diversity, the genomes of these species provide an excellent model for studying how conserved functions are maintained in the face of sequence divergence. These genome sequences provide an unprecedented dataset to contrast genome structure, genome content, and evolutionary dynamics across the well-defined phylogeny of the sequenced species (Fig. 1).

Genome assembly, annotation and alignment

Genome sequencing and assembly. We used the previously published sequence and updated assemblies for two *Drosophila* species, *D. melanogaster*[3,4] (release 4) and *D. pseudoobscura*[5] (release 2), and generated DNA sequence data for 10 additional *Drosophila* genomes by whole-genome shotgun sequencing[6,7]. These species were chosen to span a wide variety of evolutionary distances, from closely related pairs such as *D. sechellia*/*D. simulans* and *D. persimilis*/*D. pseudoobscura* to the distantly related species of the *Drosophila* and *Sophophora* subgenera. Whereas the time to the most recent common ancestor of the sequenced species may seem small on an evolutionary timescale, the evolutionary divergence spanned by the genus *Drosophila* exceeds that of the entire mammalian radiation when generation time is taken into account, as discussed further in ref. 8. We sequenced seven of the new species (*D. yakuba*, *D. erecta*, *D. ananassae*, *D. willistoni*, *D. virilis*, *D. mojavensis* and *D. grimshawi*) to deep coverage (8.4× to 11.0×) to produce high quality draft sequences. We sequenced two species, *D. sechellia* and *D. persimilis*, to intermediate coverage (4.9× and 4.1×, respectively) under the assumption that the availability of a sister species sequenced to high coverage would obviate the need for deep sequencing without sacrificing draft genome quality. Finally, seven inbred strains of *D. simulans* were sequenced to low coverage (2.9× coverage from w^{501} and ~1× coverage of six other strains) to provide population variation data[9]. Further details of the sequencing strategy can be found in Table 1, Supplementary Table 1 and section 1 in Supplementary Information.

We generated an initial draft assembly for each species using one of three different whole-genome shotgun assembly programs (Table 1). For *D. ananassae*, *D. erecta*, *D. grimshawi*, *D. mojavensis*, *D. virilis* and *D. willistoni*, we also generated secondary assemblies; reconciliation of these with the primary assemblies resulted in a 7–30% decrease in the estimated number of misassembled regions and a 12–23% increase in the N50 contig size[10] (Supplementary Table 2). For *D. yakuba*, we generated 52,000 targeted reads across low-quality regions and gaps to improve the assembly. This doubled the mean contig and scaffold sizes and increased the total fraction of high quality bases (quality score (Q) > 40) from 96.5% to 98.5%. We improved the initial 2.9× *D. simulans* w^{501} whole-genome shotgun assembly by filling assembly gaps with contigs and unplaced reads from the ~1× assemblies of the six other *D. simulans* strains, generating a 'mosaic' assembly (Supplementary Table 3). This integration markedly improved the *D. simulans* assembly: the N50 contig size of the mosaic assembly, for instance, is more than twice that of the initial w^{501} assembly (17 kb versus 7 kb).

Finally, one advantage of sequencing genomes of multiple closely related species is that these evolutionary relationships can be exploited to dramatically improve assemblies. *D. yakuba* and *D. simulans* contigs and scaffolds were ordered and oriented using pairwise alignment to the well-validated *D. melanogaster* genome

*A list of participants and affiliations appears at the end of the paper.

©2007 Nature Publishing Group

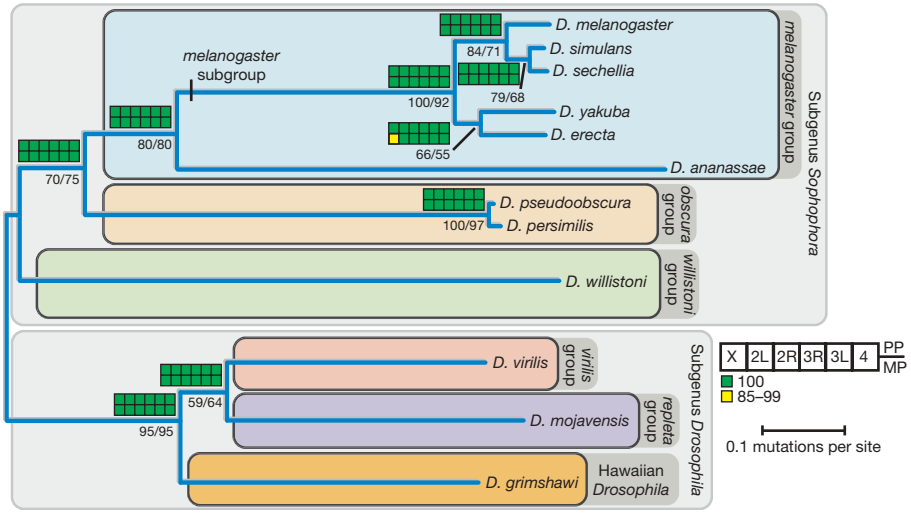


Figure 1 | Phylogram of the 12 sequenced species of *Drosophila*. Phylogram derived using pairwise genomic mutation distances and the neighbour-joining method[152,153]. Numbers below nodes indicate the per cent of genes supporting a given relationship, based on evolutionary distances estimated from fourfold-degenerate sites (left of solidus) and second codon positions (right of solidus). Coloured blocks indicate support from bayesian (posterior probability (PP), upper blocks) and maximum parsimony (MP; bootstrap values, lower blocks) analyses of data partitioned by chromosome arm. Branch lengths indicate the number of mutations per site (at fourfold-degenerate sites) using the ordinary least squares method. See ref. 154 for a discussion of the uncertainties in the *D. yakuba*/*D. erecta* clade.

sequence (Supplementary Information section 2). Likewise, the 4–5× *D. persimilis* and *D. sechellia* assemblies were improved by assisted assembly using the sister species (*D. pseudoobscura* and *D. simulans*, respectively) to validate both alignments between reads and linkage information. For the remaining species, comparative syntenic information, and in some cases linkage information, were also used to pinpoint locations of probable genome misassembly, to assign assembly scaffolds to chromosome arms and to infer their order and orientation along euchromatic chromosome arms, supplementing experimental analysis based on known markers (A. Bhutkar, S. Russo, S. Schaeffer, T. F. Smith and W. M. Gelbart, personal communication) (Supplementary Information section 2).

The mitochondrial (mt)DNA of *D. melanogaster*, *D. sechellia*, *D. simulans* (siI), *D. mauritiana* (maII) and *D. yakuba* have been previously sequenced[11,12]. For the remaining species (except *D. pseudoobscura*, the DNA from which was prepared from embryonic nuclei), we were able to assemble full mitochondrial genomes, excluding the A+T-rich control region (Supplementary Information section 2)[13]. In addition, the genome sequences of three *Wolbachia* endosymbionts (*Wolbachia wSim, Wolbachia wAna* and *Wolbachia wWil)* were assembled from trace archives, in *D. simulans*, *D. ananassae* and *D. willistoni*, respectively[14]. All of the genome sequences described here are available in FlyBase (www.flybase.org) and GenBank (www.ncbi.nlm.nih.gov) (Supplementary Tables 4 and 5).

Repeat and transposable element annotation. Repetitive DNA sequences such as transposable elements pose challenges for whole-genome shotgun assembly and annotation. Because the best approach to transposable element discovery and identification is still an active and unresolved research question, we used several repeat libraries and computational strategies to estimate the transposable element/repeat content of the 12 *Drosophila* genome assemblies (Supplementary Information section 3). Previously curated transposable element libraries in *D. melanogaster* provided the starting point for our analysis; to limit the effects of ascertainment bias, we also developed *de novo* repeat libraries using PILER-DF[15,16] and ReAS[17]. We used four transposable element/repeat detection methods (RepeatMasker, BLASTER-TX, RepeatRunner and CompTE) in conjunction with these transposable element libraries to identify repetitive elements in non-*melanogaster* species. We assessed the accuracy of each method by calibration with the estimated 5.5% transposable element content in the *D. melanogaster* genome, which is based on a high-resolution transposable element annotation[18] (Supplementary Fig. 1). On the basis of our results, we suggest a hybrid strategy for new genome sequences, employing translated BLAST with general transposable element libraries and RepeatMasker with species-specific ReAS libraries to estimate the upper and lower bound on transposable element content.

Protein-coding gene annotation. We annotated protein-coding sequences in the 11 non-*melanogaster* genomes, using four different *de novo* gene predictors (GeneID[19], SNAP[20], N-SCAN[21] and CONTRAST[22]); three homology-based predictors that transfer annotations from *D. melanogaster* (GeneWise[23], Exonerate[24], GeneMapper[25]); and one predictor that combined *de novo* and homology-based evidence (Gnomon[26]). These gene prediction sets

Table 1 | A summary of sequencing and assembly properties of each new genome

Final assembly	Genome centre	Q20 coverage (×)	Assembly size (Mb)	No. of contigs ≥2 kb	N50 contig ≥2 kb (kb)	Per cent of base pairs with quality >Q40
D. simulans	WUGSC*	2.9	137.8	10,843	17	90.3
D. sechellia	Broad†	4.9	166.6	9,713	43	90.6
D. yakuba	WUGSC*	9.1	165.7	6,344	125	98.5
D. erecta	Agencourt†	10.6	152.7	3,283	458	99.2
D. ananassae	Agencourt†	8.9	231.0	8,155	113	98.5
D. persimilis	Broad†	4.1	188.4	14,547	20	93.3
D. willistoni	JCVI‡	8.4	235.5	6,652	197	97.4
D. virilis	Agencourt†	8.0	206.0	5,327	136	98.7
D. mojavensis	Agencourt†	8.2	193.8	5,734	132	98.6
D. grimshawi	Agencourt†	7.9	200.5	9,632	114	97.1

Contigs, contiguous sequences not interrupted by gaps; N50, the largest length *L* such that 50% of all nucleotides are contained in contigs of size ≥*L*. The Q20 coverage of contigs is based on the number of assembled reads, average Q20 readlength and the assembled size excluding gaps. Assemblers used: *PCAP6, †ARACHNE4.5 and ‡Celera Assembler 7.

©2007 Nature Publishing Group

Table 2 | A summary of annotated features across all 12 genomes

	Protein-coding gene annotations		tRNA (pseudo)	Non-coding RNA annotations				Repeat coverage (%)*	Genome size (Mb; assembly†/flow cytometry‡)
	Total no. of protein- coding genes (per cent with *D. melanogaster* homologue)	Coding sequence/ intron (Mb)		snoRNA	miRNA	rRNA (5.8S + 5S)	snRNA		
D. melanogaster	13,733 (100%)	38.9/21.8	297 (4)	250	78	101	28	5.35	118/200
D. simulans	15,983 (80.0%)	45.8/19.6	268 (2)	246	70	72	32	2.73	111/162
D. sechellia	16,884 (81.2%)	47.9/21.9	312 (13)	242	78	133	30	3.67	115/171
D. yakuba	16,423 (82.5%)	50.8/22.9	380 (52)	255	80	55	37	12.04	127/190
D. erecta	15,324 (86.4%)	49.1/22.0	286 (2)	252	81	101	38	6.97	134/135
D. ananassae	15,276 (83.0%)	57.3/22.3	472 (165)	194	76	134	29	24.93	176/217
D. pseudoobscura	16,363 (78.2%)	49.7/24.0	295 (1)	203	73	55	31	2.76	127/193
D. persimilis	17,325 (72.6%)	54.0/21.9	306 (1)	199	75	80	31	8.47	138/193
D. willistoni	15,816 (78.8%)	65.4/23.5	484 (164)	216	77	76	37	15.57	187/222
D. virilis	14,680 (82.7%)	57.9/21.7	279 (2)	165	74	294	31	13.96	172/364
D. mojavensis	14,849 (80.8%)	57.8/21.9	267 (3)	139	71	74	30	8.92	161/130
D. grimshawi	15,270 (81.3%)	54.9/22.5	261 (1)	154	82	70	32	2.84	138/231

* Repeat coverage calculated as the fraction of scaffolds >200 kb covered by repeats, estimated as the midpoint between BLASTER-tx + PILER and RepeatMasker + ReAS (Supplementary Information section 3). †Total genome size estimated as the sum of base pairs in genomic scaffold >200,000 bp. ‡Genome size estimates based on flow cytometry[38].

were combined using GLEAN, a gene model combiner that chooses the most probable combination of start, stop, donor and acceptor sites from the input predictions[27,28]. All analyses reported here, unless otherwise noted, relied on a reconciled consensus set of predicted gene models—the GLEAN-R set (Table 2, and Supplementary Information section 4.1).

Quality of gene models. As the first step in assessing the quality of the GLEAN-R gene models, we used expression data from microarray experiments on adult flies, with arrays custom-designed for *D. simulans*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura*, *D. virilis* and *D. mojavensis*[29] (GEO series GSE6640; Supplementary Information section 4.2). We detected expression significantly above negative controls (false-discovery-rate-corrected Mann–Whitney U (MWU) $P < 0.001$) for 77–93% of assayed GLEAN-R models, representing 50–68% of the total GLEAN-R predictions in each species (Supplementary Table 6). Evolutionarily conserved gene models are much more likely to be expressed than lineage-specific ones (Fig. 2). Although these data cannot confirm the detailed structure of gene models, they do suggest that the majority of GLEAN-R models contain sequence that is part of a poly-adenylated transcript. Approximately 20% of transcription in *D. melanogaster* seems to be unassociated with protein-coding genes[30], and our microarray experiments fail to detect conditionally expressed genes. Thus, transcript abundance cannot conclusively establish the presence or absence of a protein-coding gene. Nonetheless, we believe these expression data increase our confidence in the reliability of the GLEAN-R models, particularly those supported by homology evidence (Fig. 2).

Because the GLEAN-R gene models were built using assemblies that were not repeat masked, it is likely that some proportion of gene models are false positives corresponding to coding sequences of transposable elements. We used RepeatMasker with *de novo* ReAS libraries and PFAM structural annotations of the GLEAN-R gene set to flag potentially transposable element-contaminated gene models (Supplementary Information section 4.2). These procedures suggest that 5.6–32.3% of gene models in non-*melanogaster* species correspond to protein-coding content derived from transposable elements (Supplementary Table 7); these transposable element-contaminated gene models are almost exclusively confined to gene predictions without strong homology support (Fig. 2). Transposable element-contaminated gene models are excluded from the final gene prediction set used for subsequent analysis, unless otherwise noted.

Homology assignment. Two independent approaches were used to assign orthology and paralogy relationships among euchromatic *D. melanogaster* gene models and GLEAN-R predictions. The first approach was a fuzzy reciprocal BLAST (FRB) algorithm, which is an

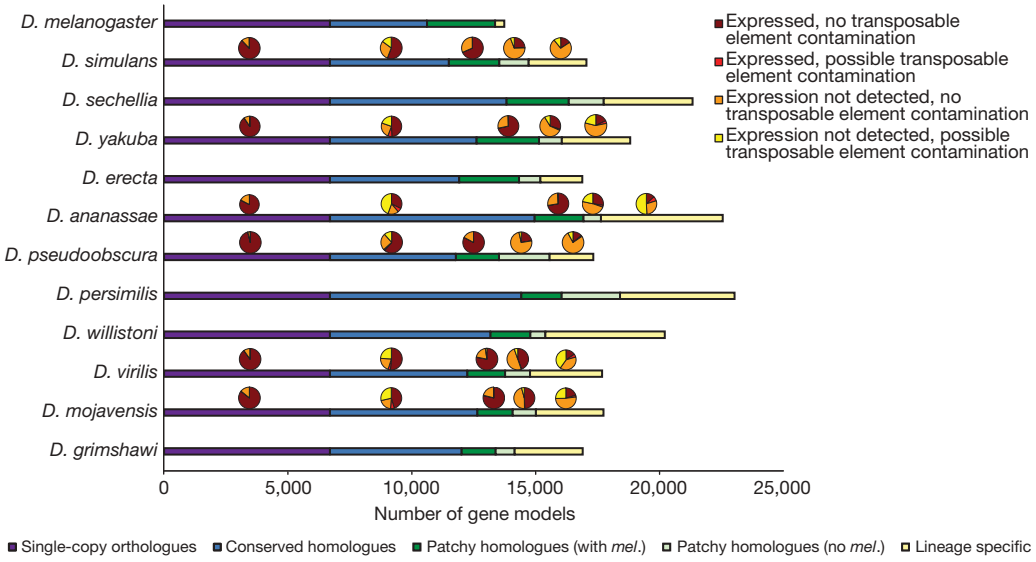


Figure 2 | Gene models in 12 *Drosophila* genomes. Number of gene models that fall into one of five homology classes: single-copy orthologues in all species (single-copy orthologues), conserved in all species as orthologues or paralogues (conserved homologues), a *D. melanogaster* homologue, but not found in all species (patchy homologues with *mel.*), conserved in at least two species but without a *D. melanogaster* homologue (patchy homologues, no *mel.*), and found only in a single lineage (lineage specific). For those species with expression data[29], pie charts indicate the fraction of genes in each homology class that fall into one of four evidence classes (see text for details).

©2007 Nature Publishing Group

extension of the reciprocal BLAST method[31] applicable to multiple species simultaneously (Supplementary Information section 5.1). Because the FRB algorithm does not integrate syntenic information, we also used a second approach based on Synpipe (Supplementary Information section 5.2), a tool for synteny-aided orthology assignment[32]. To generate a reconciled set of homology calls, pairwise Synpipe calls (between each species and *D. melanogaster*) were mapped to GLEAN-R models, filtered to retain only 1:1 relationships, and added to the FRB calls when they did not conflict and were non-redundant. This reconciled FRB + Synpipe set of homology calls forms the basis of our subsequent analyses. There were 8,563 genes with single-copy orthologues in the *melanogaster* group and 6,698 genes with single-copy orthologues in all 12 species; similar numbers of genes were also obtained with an independent approach[33]. Most single-copy orthologues are expressed and are free from potential transposable element contamination, suggesting that the reconciled orthologue set contains robust and high-quality gene models (Fig. 2).

Validation of homology calls. Because both the FRB algorithm and Synpipe rely on BLAST-based methods to infer similarities, rapidly evolving genes may be overlooked. Moreover, assembly gaps and poor-quality sequence may lead to erroneous inferences of gene loss. To validate putative gene absences, we used a synteny-based GeneWise pipeline to find potentially missed homologues of *D. melanogaster* proteins (Supplementary Information section 5.4). Of the 21,928 cases in which a *D. melanogaster* gene was absent from another species in the initial homology call set, we identified plausible homologues for 13,265 (60.5%), confirmed 4,546 (20.7%) as genuine absences, and were unable to resolve 4,117 (18.8%). Because this approach is conservative and only confirms strongly supported absences, we are probably underestimating the number of genuine absences.

Coding gene alignment and filtering. Investigating the molecular evolution of orthologous and paralogous genes requires accurate multi-species alignments. Initial amino acid alignments were generated using TCOFFEE[34] and converted to nucleotide alignments (Supplementary Table 8). To reduce biases in downstream analyses, a simple computational screen was developed to identify and mask problematic regions of each alignment (Supplementary Information section 6). Overall, 2.8% of bases were masked in the *melanogaster* group alignments, and 3.0% of bases were masked in the full 12 species alignments, representing 8.5% and 13.8% of alignment columns, respectively. The vast majority of masked bases are masked in no more than one species (Supplementary Fig. 3), suggesting that the masking procedure is not simply eliminating rapidly evolving regions of the genome. We find an appreciably higher frequency of masked bases in lower-quality *D. simulans* and *D. sechellia* assemblies, compared to the more divergent (from *D. melanogaster*) but higher-quality *D. erecta* and *D. yakuba* assemblies, suggesting a higher error rate in accurately predicting and aligning gene models in lower-quality assemblies (Supplementary Information section 6 and Supplementary Fig. 3). We used masked versions of the alignments, including only the longest *D. melanogaster* transcripts for all subsequent analysis unless otherwise noted.

Annotation of non-coding (nc)RNA genes. Using *de novo* and homology-based approaches we annotated over 9,000 ncRNA genes from recognized ncRNA classes (Table 2, and Supplementary Information section 7). In contrast to the large number of predictions observed for many ncRNA families in vertebrates (due in part to large numbers of ncRNA pseudogenes[35,36]), the number of ncRNA genes per family predicted by RFAM and tRNAscan in *Drosophila* is relatively low (Table 2). This suggests that ncRNA pseudogenes are largely absent from *Drosophila* genomes, which is consistent with the low number of protein-coding pseudogenes in *Drosophila*[37]. The relatively low numbers of some classes of ncRNA genes (for example, small nucleolar (sno)RNAs) in the *Drosophila* subgenus are likely to be an artefact of rapid rates of evolution in these types of genes and the limitation of the homology-based methods used to annotate distantly related species.

Evolution of genome structure

Coarse-level similarities among Drosophilids. At a coarse level, genome structure is well conserved across the 12 sequenced species. Total genome size estimated by flow cytometry varies less than three-fold across the phylogeny, ranging from 130 Mb (*D. mojavensis)* to 364 Mb (*D. virilis*)[38] (Table 2), in contrast to the order of magnitude difference between *Drosophila* and mammals. Total protein-coding sequence ranges from 38.9 Mb in *D. melanogaster* to 65.4 Mb in *D. willistoni*. Intronic DNA content is also largely conserved, ranging from 19.6 Mb in *D. simulans* to 24.0 Mb in *D. pseudoobscura* (Table 2). This contrasts dramatically with transposable element-derived genomic DNA content, which varies considerably across genomes (Table 2) and correlates significantly with euchromatic genome size (estimated as the summed length of contigs > 200 kb) (Kendall's $\tau = 0.70$, $P = 0.0016$).

To investigate overall conservation of genome architecture at an intermediate scale, we analysed synteny relationships across species using Synpipe[32] (Supplementary Information section 9.1). Synteny block size and average number of genes per block varies across the phylogeny as expected, with the number of blocks increasing and the average size of blocks decreasing with increasing evolutionary distance from *D. melanogaster* (A. Bhutkar, S. Russo, T. F. Smith and W. M. Gelbart, personal communication) (Supplementary Fig. 4). We inferred 112 syntenic blocks between *D. melanogaster* and *D. sechellia* (with an average of 122 genes per block), compared to 1,406 syntenic blocks between *D. melanogaster* and *D. grimshawi* (with an average of 8 genes per block). On average, 66% of each genome assembly was covered by syntenic blocks, ranging from 68% in *D. sechellia* to 58% in *D. grimshawi*.

Similarity across genomes is largely recapitulated at the level of individual genes, with roughly comparable numbers of predicted protein-coding genes across the 12 species (Table 2). The majority of predicted genes in each species have homologues in *D. melanogaster* (Table 2, Supplementary Table 9). Moreover, most of the 13,733 protein-coding genes in *D. melanogaster* are conserved across the entire phylogeny: 77% have identifiable homologues in all 12 genomes, 62% can be identified as single-copy orthologues in the six genomes of the *melanogaster* group and 49% can be identified as single-copy orthologues in all 12 genomes. The number of functional non-coding RNA genes predicted in each *Drosophila* genome is also largely conserved, ranging from 584 in *D. mojavensis* to 908 in *D. ananassae* (Table 2).

There are several possible explanations for the observed interspecific variation in gene content. First, approximately 700 *D. melanogaster* gene models have been newly annotated since the FlyBase Release 4.3 annotations used in the current study, reducing the discrepancy between *D. melanogaster* and the other sequenced genomes in this study. Second, because low-coverage genomes tend to have more predicted gene models, we suspect that artefactual duplication of genomic segments due to assembly errors inflates the number of predicted genes in some species. Finally, the non-*melanogaster* species have many more predicted lineage-specific genes than *D. melanogaster*, and it is possible that some of these are artefactual. In the absence of experimental evidence, it is difficult to distinguish genuine lineage-specific genes from putative artefacts. Future experimental work will be required to fully disentangle the causes of interspecific variation in gene number.

Abundant genome rearrangements during *Drosophila* evolution. To study the structural relationships among genomes on a finer scale, we analysed gene-level synteny between species pairs. These synteny maps allowed us to infer the history and locations of fixed genomic rearrangements between species. Although *Drosophila* species vary in their number of chromosomes, there are six fundamental chromosome arms common to all species. For ease of denoting

©2007 Nature Publishing Group

chromosomal homology, these six arms are referred to as 'Muller elements' after Hermann J. Muller, and are denoted A–F. Although most pairs of orthologous genes are found on the same Muller element, there is extensive gene shuffling within Muller elements between even moderately diverged genomes (Fig. 3, and Supplementary Information section 9.1).

Previous analysis has revealed heterogeneity in rearrangement rates among close relatives: careful inspection of 29 inversions that differentiate the chromosomes of *D. melanogaster* and *D. yakuba* revealed that 28 were fixed in the lineage leading to *D. yakuba*, and only one was fixed on the lineage leading to *D. melanogaster*[39]. Rearrangement rates are also heterogeneous across the genome among the 12 species: simulations reject a random-breakage model, which assumes that all sites are free to break in inversion events, but fail to reject a model of coldspots and hotspots for breakpoints (S. Schaeffer, personal communication). Furthermore, inversions seem to have played important roles in the process of speciation in at least some of these taxa[40].

One particularly striking example of the dynamic nature of genome micro-structure in *Drosophila* is the homeotic *homeobox (Hox)* gene cluster(s)[41]. *Hox* genes typically occur in genomic clusters, and this clustering is conserved across many vertebrate and invertebrate taxa, suggesting a functional role for the precise and collinear arrangement of these genes. However, several cluster splits have been previously identified in *Drosophila*[42,43], and the 12 *Drosophila* genome sequences provide additional evidence against the functional importance of *Hox* gene clustering in *Drosophila*. There are seven different gene arrangements found across 13 *Drosophila* species (the 12 sequenced genomes and *D. buzzatii*), with no species retaining the inferred ancestral gene order[44]. It thus seems that, in *Drosophila, Hox* genes do not require clustering to maintain proper function, and are a powerful illustration of the dynamism of genome structure across the sequenced genomes.

Transposable element evolution. Mobile, repetitive transposable element sequences are a particularly dynamic component of eukaryotic genomes. Transposable element/repeat content (in scaffolds >200 kb) varies by over an order of magnitude across the genus, ranging from ~2.7% in *D. simulans* and *D. grimshawi* to ~25% in *D. ananassae* (Table 2, and Supplementary Fig. 1). These data support the lower euchromatic transposable element content in *D. simulans* relative to *D. melanogaster*[45], and reveal that euchromatic transposable element/repeat content is generally similar within the *melanogaster* subgroup. Within the *Drosophila* subgenus, *D. grimshawi* has the lowest transposable element/repeat content, possibly relating to its ecological status as an island endemic, which may minimize the chance for horizontal transfer of transposable element families. Finally, the highest levels of transposable element/repeat content are found in *D. ananassae* and *D. willistoni*. These species also have the highest numbers of pseudo-transfer (t)RNA genes (Table 2), indicating a potential relationship between pseudo-tRNA genesis and repetitive DNA, as has been established in the mouse genome[36].

Different classes of transposable elements can vary in abundance owing to a variety of host factors, motivating an analysis of the intragenomic ecology of transposable elements in the 12 genomes. In *D. melanogaster*, long terminal repeat (LTR) retrotransposons have the highest abundance, followed by LINE (long interspersed nuclear element)-like retrotransposons and terminal inverted repeat (TIR) DNA-based transposons[18]. An unbiased, conservative approach (Supplementary Information section 3) for estimating the rank order abundance of major transposable element classes suggests that these abundance trends are conserved across the entire genus (Supplementary Fig. 5). Two exceptions are an increased abundance of TIR elements in *D. erecta* and a decreased abundance of LTR elements in *D. pseudoobscura*; the latter observation may represent an assembly artefact because the sister species *D. persimilis* shows typical LTR abundance. Given that individual instances of transposable element repeats and transposable element families themselves are not conserved across the genus, the stability of abundance trends for different classes of transposable elements is striking and suggests common mechanisms for host–transposable element co-evolution in *Drosophila*.

Although comprehensive analysis of the structural and evolutionary relationships among families of transposable elements in the 12 genomes remains a major challenge for *Drosophila* genomics, some initial insights can be gleaned from analysis of particularly well-characterized transposable element families. Previous analysis has shown variable dynamics for the most abundant transposable element family *(DINE-1)*[46] in the *D. melanogaster* genome[18,47]: although inactive in *D. melanogaster*[48], *DINE-1* has experienced a recent transpositional burst in *D. yakuba*[49]. Our analysis confirms that this element is highly abundant in all of the other sequenced genomes of *Drosophila*, but is not found outside of Diptera[50,51]. Moreover, the inferred phylogenetic relationship of *DINE-1* paralogues from several *Drosophila* species suggests vertical transmission as the major mechanism for *DINE-1* propagation. Likewise, analysis of the *Galileo*

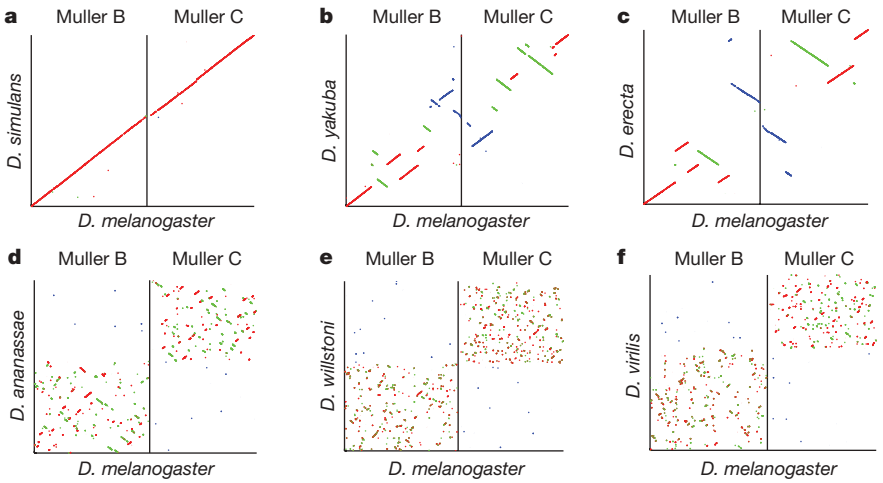


Figure 3 | Synteny plots for Muller elements B and C with respect to *D. melanogaster* gene order. The horizontal axis shows *D. melanogaster* gene order for Muller elements B and C, and the vertical axis maps homologous locations[32,155] in individual species (**a**–**f** in increasing evolutionary distance from *D. melanogaster)*. Left to right on the *x* axis is from telomere to centromere for Muller element B, followed by Muller element C from centromere to telomere. Red and green lines represent syntenic segments in the same or reverse orientation along the chromosome relative to *D. melanogaster*, respectively. Blue segments show gene transposition of genes from one element to the other.

©2007 Nature Publishing Group

and *1360* transposons reveals a widespread but discontinuous phylogenetic distribution for both families, notably with both families absent in the geographically isolated Hawaiian species, *D. grimshawi*[52]. These results are consistent with an ancient origin of the *Galileo* and *1360* families in the genus and subsequent horizontal transfer and/or loss in some lineages.

The use of these 12 genomes also facilitated the discovery of transposable element lineages not yet documented in *Drosophila*, specifically the P instability factor (*PIF*) superfamily of DNA transposons. Our analysis indicates that there are four distinct lineages of this transposon in *Drosophila*, and that this element has indeed colonized many of the sequenced genomes[53]. This superfamily is particularly intriguing given that *PIF*-transposase-like genes have been implicated in the origin of at least seven different genes during the *Drosophila* radiation[53], suggesting that not only do transposable elements affect the evolution of genome structure, but that their domestication can play a part in the emergence of novel genes.

D. melanogaster maintains its telomeres by occasional targeted transposition of three telomere-specific non-LTR retrotransposons (*HeT-A*, *TART* and *TAHRE*) to chromosome ends[54,55] and not by the more common mechanism of telomerase-generated G-rich repeats[56]. Multiple telomeric retrotransposons have originated within the genus, where they now maintain telomeres, and recurrent loss of most of the ORF2 from telomeric retrotransposons (for example, *TAHRE*) has given rise to half-telomeric-retrotransposons (for example, *HeT-A*) during *Drosophila* evolution[57]. The phylogenetic relationship among these telomeric elements is congruent with the species phylogeny, suggesting that they have been vertically transmitted from a common ancestor[57].

ncRNA gene family evolution. Using ncRNA gene annotations across the 12-species phylogeny, we inferred patterns of gene copy number evolution in several ncRNA families. Transfer RNA genes are the most abundant family of ncRNA genes in all 12 genomes, with 297 tRNAs in *D. melanogaster* and 261–484 tRNA genes in the other species (Table 2). Each genome encodes a single selenocysteine tRNA, with the exception of *D. willistoni*, which seems to lack this gene (R. Guigo, personal communication). Elevated tRNA gene counts in *D. ananassae* and *D. willistoni* are explained almost entirely by pseudo-tRNA gene predictions. We infer from the lack of pseudo-tRNAs in most *Drosophila* species, and from similar numbers of tRNAs obtained from an analysis of the chicken genome ($n = 280$)[58], that the minimal metazoan tRNA set is encoded by ~300 genes, in contrast to previous estimates of 497 in human and 659 in *Caenorhabditis elegans*[59,60]. Similar numbers of snoRNAs are predicted in the *D. melanogaster* subgroup ($n = 242$–255), in which sequence similarity is high enough for annotation by homology, with fewer snoRNAs ($n = 194$–216) annotated in more distant members of the *Sophophora* subgenus, and even fewer snoRNAs ($n = 139$–165) predicted in the *Drosophila* subgenus, in which annotation by homology becomes much more difficult.

Of 78 previously reported micro (mi)RNA genes, 71 (91%) are highly conserved across the entire genus, with the remaining seven genes (*mir-2b-1*, *-289*, *-303*, *-310*, *-311*, *-312* and *-313*) restricted to the subgenus *Sophophora* (Supplementary Information section 7.2). All the species contain similar numbers of spliceosomal snRNA genes (Table 2), including at least one copy each of the four U12-dependent (minor) spliceosomal RNAs, despite evidence for birth and death of these genes and the absence of stable subtypes[61]. The unusual, lineage-specific expansion in size of U11 snRNA, previously described in *Drosophila*[61,62], is even more extreme in *D. willistoni*. We annotated 99 copies of the 5S ribosomal (r)RNA gene in a cluster in *D. melanogaster*, and between 13 and 73 partial 5S rRNA genes in clusters in the other genomes. Finally, we identified members of several other classes of ncRNA genes, including the RNA components of the RNase P (1 per genome) and the signal recognition particle (SRP) RNA complexes (1–3 per genome), suggesting that these functional RNAs are involved in similar biological processes throughout the genus. We were only able to locate the *roX* (RNA on X)[63,64] genes involved in dosage compensation using nucleotide homology in the *melanogaster* subgroup, although analyses incorporating structural information have identified *roX* genes in other members of the genus[65].

We investigated the evolution of rRNA genes in the 12 sequenced genomes, using trace archives to locate sequence variants within the transcribed portions of these genes. This analysis revealed moderate levels of variation that are not distributed evenly across the rRNA genes, with fewest variants in conserved core coding regions, more variants in coding expansion regions, and higher still variant abundances in non-coding regions. The level and distribution of sequence variation in rRNA genes are suggestive of concerted evolution, in which recombination events uniformly distribute variants throughout the rDNA loci, and selection dictates the frequency to which variants can expand[66].

Protein-coding gene family evolution. For a general perspective on how the protein-coding composition of these 12 genomes has changed, we examined gene family expansions and contractions in the 11,434 gene families (including those of size one in each species) predicted to be present in the most recent common ancestor of the two subgenera. We applied a maximum likelihood model of gene gain and loss[67] to estimate rates of gene turnover. This analysis suggests that gene families expand or contract at a rate of 0.0012 gains and losses per gene per million years, or roughly one fixed gene gain/loss across the genome every 60,000 yr[68]. Many gene families (4,692 or 41.0%) changed in size in at least one species, and 342 families showed significantly elevated ($P < 0.0001$) rates of gene gain and loss compared to the genomic average, indicating that non-neutral processes may play a part in gene family evolution. Twenty-two families exhibit rapid copy number evolution along the branch leading to *D. melanogaster* (eighteen contractions and four expansions; Supplementary Table 10). The most common Gene Ontology (GO) terms among families with elevated rates of gain/loss include 'defence response', 'protein binding', 'zinc ion binding', 'proteolysis', and 'trypsin activity'. Interestingly, genes involved in 'defence response' and 'proteolysis' also show high rates of protein evolution (see below). We also found heterogeneity in overall rates of gene gain and loss across lineages, although much of this variation could result from interspecific differences in assembly quality[68].

Lineage-specific genes. The vast majority of *D. melanogaster* proteins that can be unambiguously assigned a homology pattern (Supplementary Information section 5) are inferred to be ancestrally present at the genus root (11,348/11,644, or 97.5%). Of the 296 non-ancestrally present genes, 252 are either *Sophophora*-specific, or have a complicated pattern of homology requiring more than one gain and/or loss on the phylogeny, and are not discussed further. The remaining 44 proteins include 14 present in the *melanogaster* group, 23 present only in the *melanogaster* subgroup, 3 unique to the *melanogaster* species complex, and 4 found in *D. melanogaster* only. Because we restricted this analysis to unambiguous homologues of high-confidence protein-coding genes in *D. melanogaster*[8], we are probably undercounting the number of genes that have arisen *de novo* in any particular lineage. However, ancestrally heterochromatic genes that are currently euchromatic in *D. melanogaster* may spuriously seem to be lineage-specific.

The 44 lineage-specific genes (Supplementary Table 11) differ from ancestrally present genes in several ways. They have a shorter median predicted protein length (lineage-specific median 177 amino acids, other median 421 amino acids, MWU, $P = 3.6 \times 10^{-13}$), are more likely to be intronless (Fisher's exact test (FET), $P = 6.2 \times 10^{-6}$), and are more likely to be located in the intron of another gene on the opposite strand (FET, $P = 3.5 \times 10^{-4}$). In addition, 18 of these 44 genes are testis- or accessory-gland-specific in *D. melanogaster*, a significantly greater fraction than is found in the ancestral set (FET, $P = 1.25 \times 10^{-4}$). This is consistent with previous observations that novel genes are often testis-specific in *Drosophila*[69–73] and

©2007 Nature Publishing Group

expression studies on seven of the species show that species-restricted genes are more likely to exhibit male-biased expression[29]. Further, these genes are significantly more tissue-specific in expression (as measured by τ; ref. 74) (MWU, $P = 9.6 \times 10^{-6}$), and this pattern is not solely driven by genes with testis-specific expression patterns.

Protein-coding gene evolution

Positive selection and selective constraints in *Drosophila* genomes. To study the molecular evolution of protein-coding genes, we estimated rates of synonymous and non-synonymous substitution in 8,510 single-copy orthologues within the six *melanogaster* group species using PAML[75] (Supplementary Information section 11.1); synonymous site saturation prevents analysis of more divergent comparisons. We investigate only single-copy orthologues because when paralogues are included, alignments become increasingly problematic. Rates of amino acid divergence for single-copy orthologues in all 12 species were also calculated; these results are largely consistent with the analysis of non-synonymous divergence in the *melanogaster* group, and are not discussed further.

To understand global patterns of divergence and constraint across functional classes of genes, we examined the distributions of ω ($=d_N/d_S$, the ratio of non-synonymous to synonymous divergence) across Gene Ontology categories (GO)[76], excluding GO annotations based solely on electronic support (Supplementary Information section 11.2). Most functional categories of genes are strongly constrained, with median estimates of ω much less than one. In general, functionally similar genes are similarly constrained: 31.8% of GO categories have significantly lower variance in ω than expected (q-value true-positive test[77]). Only 11% of GO categories had statistically significantly elevated ω (relative to the median of all genes with GO annotations) at a 5% false-discovery rate (FDR), suggesting either positive selection or a reduction in selective constraint. The GO categories with elevated ω include the biological process terms 'defence response', 'proteolysis', 'DNA metabolic process' and 'response to biotic stimulus'; the molecular function terms 'transcription factor activity', 'peptidase activity', 'receptor binding', 'odorant binding', 'DNA binding', 'receptor activity' and 'G-protein-coupled receptor activity'; and the cellular location term 'extracellular' (Fig. 4, and Supplementary Table 12). Similar results are obtained when d_N is compared across GO categories, suggesting that in most cases differences in ω among GO categories is driven by amino acid rather than synonymous site substitutions. The two exceptions are the molecular function terms 'transcription factor activity' and 'DNA binding activity', for which we observe significantly decelerated d_S (FDR $= 7.2 \times 10^{-4}$ for both; Supplementary Information section 11.2) and no significant differences in d_N.

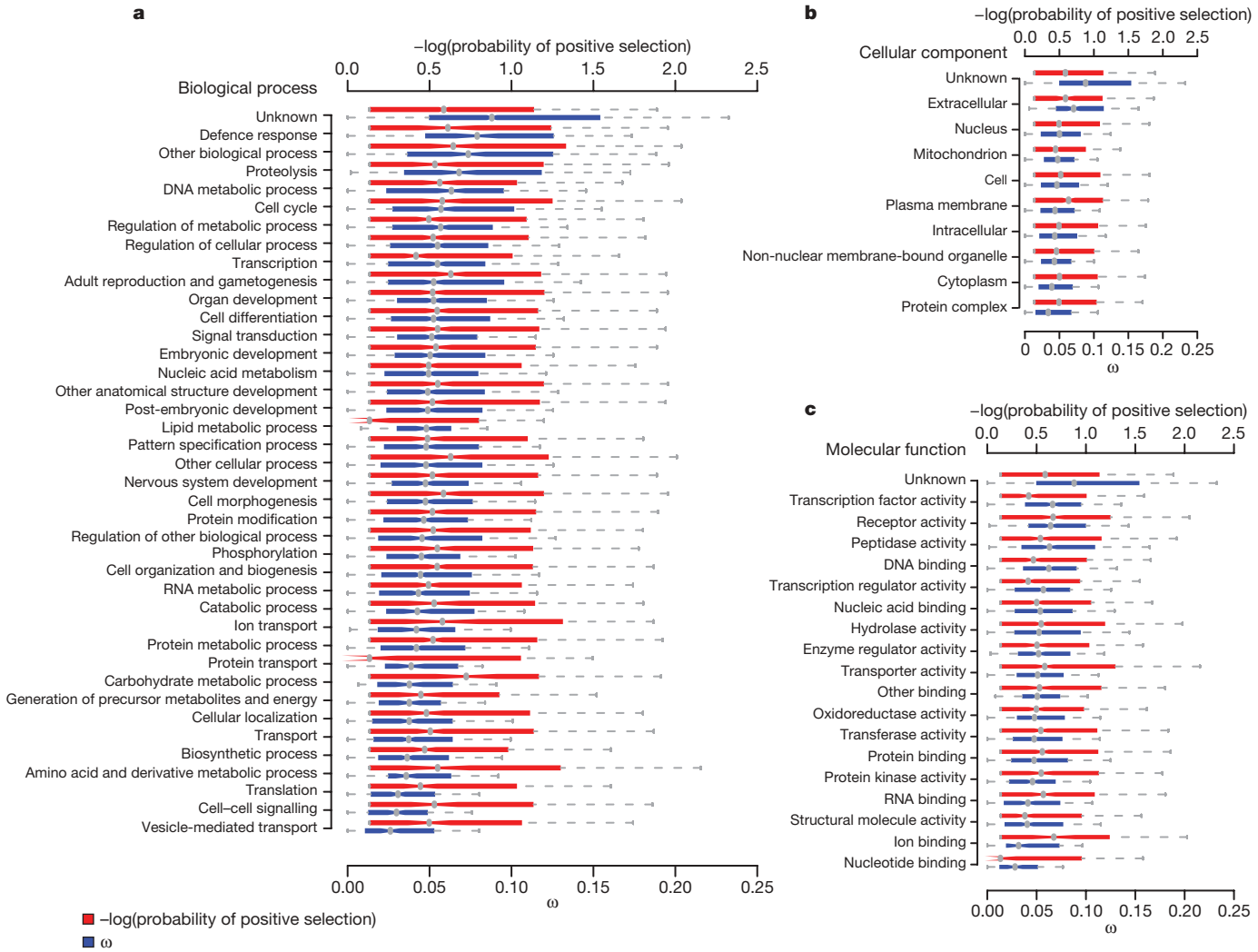


Figure 4 | Patterns of constraint and positive selection among GO terms. Distribution of average ω per gene and the negative $\log_{10}$ of the probability of positive selection (Supplementary Information section 11.2) for genes annotated with: **a**, biological process GO terms; **b**, cellular component GO terms; and **c**, molecular function GO terms. Only GO terms with 200 or more genes annotated are plotted. See Supplementary Table 12 for median values and significance. Note that most genes evolve under evolutionary constraint at most of their sites, leading to low values of ω; even genes that experience positive selection do not typically have an average ω across all codons that exceeds one.

©2007 Nature Publishing Group

To distinguish possible positive selection from relaxed constraint, we tested explicitly for genes that have a subset of codons with signatures of positive selection, using codon-based likelihood models of molecular evolution, implemented in PAML[78,79] (Supplementary Information section 11.1). Although this test is typically regarded as a conservative test for positive selection, it may be confounded by selection at synonymous sites. However, selection at synonymous sites (that is, codon bias, see below) is quite weak. Moreover, variability in ω presented here tends to reflect variability in d_N. We therefore believe that it is appropriate to treat synonymous sites as nearly neutral and sites with $\omega > 1$ as consistent with positive selection. Despite a number of functional categories with evidence for elevated ω, 'helicase activity' is the only functional category significantly more likely to be positively selected (permutation test, $P = 2 \times 10^{-4}$, FDR = 0.007; Supplementary Table 12); the biological significance of this finding merits further investigation. Furthermore, within each GO class, there is greater dispersion among genes in their probability of positive selection than in their estimate of ω (MWU one-tailed, $P = 0.011$; Supplementary Information section 11.1), suggesting that although functionally similar genes share patterns of constraint, they do not necessarily show similar patterns of positive selection (Fig. 4).

Interestingly, protein-coding genes with no annotated ('unknown') function in the GO database seem to be less constrained (permutation test, $P < 1 \times 10^{-4}$, FDR = 0.006)[80] and to have on average lower *P*-values for the test of positive selection than genes with annotated functions (permutation test, $P = 0.001$, FDR = 0.058). It is unlikely that this observation results entirely from an over-representation of mis-annotated or non-protein-coding genes in the 'unknown' functional class, because this finding is robust to the removal of all *D. melanogaster* genes predicted to be non-protein-coding in ref. 8. The bias in the way biological function is ascribed to genes (to laboratory-induced, easily scorable functions) leaves open the possibility that unannotated biological functions may have an important role in evolution. Indeed, genes with characterized mutant alleles in FlyBase evolve significantly more slowly than other genes (median $\omega_{\text{with alleles}} = 0.0525$ and $\omega_{\text{without alleles}} = 0.0701$; MWU, $P < 1 \times 10^{-16}$).

Previous work has suggested that a substantial fraction of non-synonymous substitutions in *Drosophila* were fixed through positive selection[81–85]. We estimate that 33.1% of single-copy orthologues in the *melanogaster* group have experienced positive selection on at least a subset of codons (*q*-value true-positive tests[77]) (Supplementary Information section 11.1). This may be an underestimate, because we have only examined single-copy orthologues, owing to difficulties in producing accurate alignments of paralogues by automated methods. On the basis of the 878 genes inferred to have experienced positive selection with high confidence (FDR < 10%), we estimated that an average of 2% of codons in positively selected genes have $\omega > 1$. Thus, several lines of evidence, based on different methodologies, suggest that patterns of amino acid fixation in *Drosophila* genomes have been shaped extensively by positive selection.

The presence of functional domains within a protein may lead to heterogeneity in patterns of constraint and adaptation along its length. Among genes inferred to be evolving by positive selection at a 10% FDR, 63.7% (*q*-value true-positive tests[77]) show evidence for spatial clustering of positively selected codons (Supplementary Information section 11.2). Spatial heterogeneity in constraint is further supported by contrasting ω for codons inside versus outside defined InterPro domains (genes lacking InterPro domains are treated as 'outside' a defined InterPro domain). Codons within InterPro domains were significantly more conserved than codons outside InterPro domains (median ω: 0.062 InterPro domains, 0.084 outside InterPro domains; MWU, $P < 2.2 \times 10^{-16}$; Supplementary Information section 11.2). Similarly, there were significantly more positively selected codons outside of InterPro domains than inside domains (FET $P < 2.2 \times 10^{-16}$), suggesting that in addition to being more constrained, codons in protein domains are less likely to be targets of positive selection (Supplementary Fig. 6).

Factors affecting the rate of protein evolution in *Drosophila*. The sequenced genomes of the *melanogaster* group provide unprecedented statistical power to identify factors affecting rates of protein evolution. Previous analyses have suggested that although the level of gene expression consistently seems to be a major determinant of variation in rates of evolution among proteins[86,87], other factors probably play a significant, if perhaps minor, part[88–91]. In *Drosophila*, although highly expressed genes do evolve more slowly, breadth of expression across tissues, gene essentiality and intron number all also independently correlate with rates of protein evolution, suggesting that the additional complexities of multicellular organisms are important factors in modulating rates of protein evolution[78]. The presence of repetitive amino acid sequences has a role as well: non-repeat regions in proteins containing repeats evolve faster and show more evidence for positive selection than genes lacking repeats[92].

These data also provide a unique opportunity to examine the impact of chromosomal location on evolutionary rates. Population genetic theory predicts that for new recessive mutations, both purifying and positive selection will be more efficient on the X chromosome given its hemizygosity in males[93]. In contrast, the lack of recombination on the small, mainly heterochromatic dot chromosome[94,95] is expected to reduce the efficacy of selection[96]. Because codon bias, or the unequal usage of synonymous codons in protein-coding sequences, reflects weak but pervasive selection, it is a sensitive metric for evaluating the efficacy of purifying selection. Consistent with expectation, in all 12 species, we find significantly elevated levels of codon bias on the X chromosome and significantly reduced levels of codon bias on the dot chromosome[97]. Furthermore, X-chromosome-linked genes are marginally over-represented within the set of positively selected genes in the *melanogaster* group (FET, $P = 0.055$), which is consistent with increased rates of adaptive substitution on this chromosome. This analysis suggests that chromosomal context also serves to modulate rates of molecular evolution in protein-coding genes.

To examine further the impact of genomic location on protein evolution, we examined the subset of genes that have moved within or between chromosome arms[32,98]. Genes inferred to have moved between Muller elements have a significantly higher rate of protein evolution than genes inferred to have moved within a Muller element (MWU, $P = 1.32 \times 10^{-14}$) and genes that have maintained their genomic position (MWU, $P = 0.008$) (Supplementary Fig. 7). Interestingly, genes that move within Muller elements have a significantly lower rate of protein evolution than those for which genomic locations have been maintained (MWU, $P = 3.85 \times 10^{-14}$). It remains unclear whether these differences reflect underlying biases in the types of genes that move inter- versus intra-chromosomally, or whether they are due to *in situ* patterns of evolution in novel genomic contexts.

Codon bias. Codon bias is thought to enhance the efficiency and/or accuracy of translation[99–101] and seems to be maintained by mutation–selection–drift balance[101–104]. Across the 12 *Drosophila* genomes, there is more codon bias in the *Sophophora* subgenus than in the *Drosophila* subgenus, and a previously noted[105–109] striking reduction in codon bias in *D. willistoni*[110,111] (Fig. 5). However, with only minor exceptions, codon preferences for each amino acid seem to be conserved across 11 of the 12 species. The striking exception is *D. willistoni*, in which codon usage for 6 of 18 redundant amino acids has diverged (Fig. 5). Mutation alone is not sufficient to explain codon-usage bias in *D. willistoni*, which is suggestive of a lineage-specific shift in codon preferences[111,112]. We found evidence for a lineage-specific genomic reduction in codon bias in *D. melanogaster* (Fig. 5), as has been suggested previously[113–119]. In addition, maximum-likelihood estimation of the strength of selection on synonymous sites in 8,510 *melanogaster* group single-copy orthologues revealed a marked reduction in the number of genes under selection

©2007 Nature Publishing Group

for increased codon bias in *D. melanogaster* relative to its sister species *D. sechellia*[120].

Evolution of genes associated with ecology and reproduction. Given the ecological and environmental diversity encompassed by the 12 *Drosophila* species, we examined the evolution of genes and gene families associated with ecology and reproduction. Specifically, we selected genes with roles in chemoreception, detoxification/metabolism, immunity/defence, and sex/reproduction for more detailed study.

Chemoreception. *Drosophila* species have complex olfactory and gustatory systems used to identify food sources, hazards and mates, which depend on odorant-binding proteins, and olfactory/odorant and gustatory receptors (*Or*s and *Gr*s). The *D. melanogaster* genome has approximately 60 *Or*s, 60 *Gr*s and 50 odorant-binding protein genes. Despite overall conservation of gene number across the 12 species and widespread evidence for purifying selection within the *melanogaster* group, there is evidence that a subset of *Or* and *Gr* genes experiences positive selection[121–123]. Furthermore, clear lineage-specific differences are detectable between generalist and specialist species within the *melanogaster* subgroup. First, the two independently evolved specialists (*D. sechellia* and *D. erecta*) are losing *Gr* genes approximately five times more rapidly than the generalist species[121,124]. We believe this result is robust to sequence quality, because all pseudogenes and deletions were verified by direct re-sequencing and synteny-based orthologue searches, respectively. Generalists are expected to encounter the most diverse set of tastants and seem to have maintained the greatest diversity of gustatory receptors. Second, *Or* and *Gr* genes that remain intact in *D. sechellia* and *D. erecta* evolve significantly more rapidly along these two lineages ($\omega = 0.1556$ for *Or*s and 0.1874 for *Gr*s) than along the generalist lineages ($\omega = 0.1049$ for *Or*s and 0.1658 for *Gr*s; paired Wilcoxon, $P = 0.0003$ and 0.003, respectively[124]). There is some evidence that odorant-binding protein genes also evolve significantly faster in specialists compared to generalists[122]. This elevated ω reflects a trend observed throughout the genomes of the two specialists and is likely to result, at least in part, from demographic phenomena. However, the difference between specialist and generalist ω for *Or*/*Gr* genes (0.0292) is significantly greater than the difference for genes across the genome (0.0091; MWU, $P = 0.0052$)[121], suggesting a change in selective regime. Moreover, the observation that elevated ω as well as accelerated gene loss disproportionately affect groups of *Or* and *Gr* genes that respond to specific chemical ligands and/or are expressed during specific life stages suggests that rapid evolution at *Or*/*Gr* loci in specialists is related to the ecological shifts these species have sustained[121].

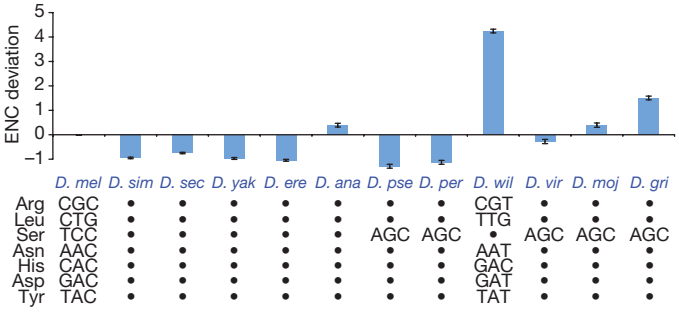


Figure 5 | Deviations in codon bias from *D. melanogaster* in 11 *Drosophila* species. The upper panel depicts differences in ENC (effective number of codons) between *D. melanogaster* and the 11 non-*melanogaster* species, calculated on a gene-by-gene basis. Note that increasing levels of ENC indicates a decrease in codon bias. The *Sophophora* subgenus in general has higher levels of codon bias than the *Drosophila* subgenus with the exception of *D. willistoni,* which shows a dramatic reduction in codon bias. The lower panel shows the 7 codons for which preference changes across the 12 *Drosophila* species. A dot indicates identical codon preference to *D. melanogaster*; otherwise the preferred codon is indicated.

Detoxification/metabolism. The larval food sources for many *Drosophila* species contain a cocktail of toxic compounds, and consequently *Drosophila* genomes encode a wide variety of detoxification proteins. These include members of the cytochrome P450 (P450), carboxyl/choline-esterase (CCE) and glutathione *S*-transferase (GST) multigene families, all of which also have critical roles in resistance to insecticides[125–127]. Among the P450s, the five enzymes associated with insecticide resistance are highly dynamic across the phylogeny, with 24 duplication events and 4 loss events since the last common ancestor of the genus, which is in striking contrast to genes with known developmental roles, eight of which are present as a single copy in all 12 species (C. Robin, personal communication). As with chemoreceptors, specialists seem to lose detoxification genes at a faster rate than generalists. For instance, *D. sechellia* has lost the most P450 genes; these 14 losses comprise almost one-third of all P450 loss events (Supplementary Table 13) (C. Robin, personal communication). Positive selection has been implicated in detoxification-gene evolution as well, because a search for positive selection among GSTs identified the parallel evolution of a radical glycine to lysine amino acid change in GSTD1, an enzyme known to degrade DDT[128]. Finally, although metabolic enzymes in general are highly constrained (median $\omega = 0.045$ for enzymes, 0.066 for non-enzymes; MWU, $P = 5.7 \times 10^{-24}$), enzymes involved in xenobiotic metabolism evolve significantly faster than other enzymes (median $\omega = 0.05$ for the xenobiotic group versus $\omega = 0.045$ overall, two-tailed permutation test, $P = 0.0110$; A. J. Greenberg, personal communication).

Metazoans deal with excess selenium in the diet by sequestration in selenoproteins, which incorporate the rare amino acid selenocysteine (Sec) at sites specified by the TGA codon. The recoding of the normally terminating signal TGA as a Sec codon is mediated by the selenocystein insertion sequence (SECIS), a secondary structure in the 3′ UTR of selenoprotein messenger RNAs. All animals examined so far have selenoproteins; three have been identified in *D. melanogaster* (SELG, SELM and SPS2[129,130]). Interestingly, although the three known *melanogaster* selenoproteins are all present in the genomes of the other *Drosophila* species, in *D. willistoni* the TGA Sec codons have been substituted by cysteine codons (TGT/TGC). Consistent with this finding, analysis of the seven genes implicated to date in selenoprotein synthesis including the Sec-specific tRNA suggests that most of these genes are absent in *D. willistoni* (R. Guigo, personal communication). *D. willistoni* thus seems to be the first animal known to lack selenoproteins. If correct, this observation is all the more remarkable given the ubiquity of selenoproteins and the selenoprotein biosynthesis machinery in metazoans, the toxicity of excess selenium, and the protection from oxidative stress mediated by selenoproteins. However, it remains possible that this species encodes selenoproteins in a different way, and this represents an exciting avenue of future research.

Immunity/defence. *Drosophila*, like all insects, possesses an innate immune system with many components analogous to the innate immune pathways of mammals, although it lacks an antibody-mediated adaptive immune system[131]. Immune system genes often evolve rapidly and adaptively, driven by selection pressures from pathogens and parasites[132–134]. The genus *Drosophila* is no exception: immune system genes evolve more rapidly than non-immune genes, showing both high total divergence rates and specific signs of positive selection[135]. In particular, 29% of receptor genes involved in phagocytosis seem to evolve under positive selection, suggesting that molecular co-evolution between *Drosophila* pattern recognition receptors and pathogen antigens is driving adaptation in the immune system[135]. Somewhat surprisingly, genes encoding effector proteins such as antimicrobial peptides are far less likely to exhibit adaptive sequence evolution. Only 5% of effector genes (and no antimicrobial peptides) show evidence of adaptive evolution, compared to 10% of genes genome-wide. Instead, effector genes seem to evolve by rapid duplication and deletion. Whereas 49% of genes genome-wide, 63%

©2007 Nature Publishing Group

of genes involved in pathogen recognition and 81% of genes implicated in immune-related signal transduction can be found as single-copy orthologues in all 12 species, only 40% of effector genes exist as single-copy orthologues across the genus ($\chi 2 = 41.13$, $P = 2.53 \times 10^{-8}$), suggesting rapid radiation of effector protein classes along particular lineages[135]. Thus, much of the *Drosophila* immune system seems to evolve rapidly, although the mode of evolution varies across immune-gene functional classes.

Sex/reproduction. Genes encoding sex- and reproduction-related proteins are subject to a wide array of selective forces, including sexual conflict, sperm competition and cryptic female choice, and to the extent that these selective forces are of evolutionary consequence, this should lead to rapid evolution in these genes[136] (for an overview see refs 137, 138). The analysis of 2,505 sex- and reproduction-related genes within the *melanogaster* group indicated that male sex- and reproduction-related genes evolve more rapidly at the protein level than genes not involved in sex or reproduction or than female sex- and reproduction-related genes (Supplementary Fig. 8). Positive selection seems to be at least partially responsible for these patterns, because genes involved in spermatogenesis have significantly stronger evidence for positive selection than do non-spermatogenesis genes (permutation test, $P = 0.0053$). Similarly, genes that encode components of seminal fluid have significantly stronger evidence for positive selection than 'non-sex' genes[139]. Moreover, protein-coding genes involved in male reproduction, especially seminal fluid and testis genes, are particularly likely to be lost or gained across *Drosophila* species[29,139].

Evolutionary forces in the mitochondrial genome. Functional elements in mtDNA are strongly conserved, as expected: tRNAs are relatively more conserved than the mtDNA overall (average pairwise nucleotide distance = 0.055 substitutions per site for tRNAs versus 0.125 substitutions per site overall). We observe a deficit of substitutions occurring in the stem regions of the stem-loop structure in tRNAs, consistent with strong selective pressure to maintain RNA secondary structure, and there is a strong signature of purifying selection in protein-coding genes[13]. However, despite their shared role in aerobic respiration, there is marked heterogeneity in the rates of amino acid divergence between the oxidative phosphorylation enzyme complexes across the 12 species (NADH dehydrogenase, 0.059 > ATPase, 0.042 > CytB, 0.037 > cytochrome oxidase, 0.020; mean pairwise d_N), which contrasts with the relative homogeneity in synonymous substitution rates. A model with distinct substitution rates for each enzyme complex rather than a single rate provides a significantly better fit to the data ($P < 0.0001$), suggesting complex-specific selective effects of mitochondrial mutations[13].

Non-coding sequence evolution

ncRNA sequence evolution. The availability of complete sequence from 12 *Drosophila* genomes, combined with the tractability of RNA structure predictions, offers the exciting opportunity to connect patterns of sequence evolution directly with structural and functional constraints at the molecular level. We tested models of RNA evolution focusing on specific ncRNA gene classes in addition to inferring patterns of sequence evolution using more general datasets that are based on predicted intronic RNA structures.

The exquisite simplicity of miRNAs and their shared stem-loop structure makes these ncRNAs particularly amenable to evolutionary analysis. Most miRNAs are highly conserved within the *Drosophila* genus: for the 71 previously described miRNA genes inferred to be present in the common ancestor of these 12 species, mature miRNA sequences are nearly invariant. However, we do find a small number of substitutions and a single deletion in mature miRNA sequences (Supplementary Table 14), which may have functional consequences for miRNA–target interactions and may ultimately help identify targets through sequence covariation. Pre-miRNA sequences are also highly conserved, evolving at about 10% of the rate of synonymous sites.

To link patterns of evolution with structural constraints, we inferred ancestral pre-miRNA sequences and deduced secondary structures at each ancestral node on the phylogeny (Supplementary Information section 12.1). Although conserved miRNA genes show little structural change (little change in free energy), the five *melanogaster* group-specific miRNA genes (*miR-303* and the *mir-310/311/312/313* cluster) have undergone numerous changes across the entire pre-miRNA sequence, including the ordinarily invariant mature miRNA. Patterns of polymorphism and divergence in these lineage-specific miRNA genes, including a high frequency of derived mutations, are suggestive of positive selection[140]. Although lineage-specific miRNAs may evolve under less constraint because they have fewer target transcripts in the genome, it is also possible that recent integration into regulatory networks causes accelerated rates of miRNA evolution.

We further investigated patterns of sequence evolution for the subset of 38 conserved pre-miRNAs with mature miRNA sequences at their 3′ end by calculating evolutionary rates in distinct site classes (Fig. 6, and Supplementary Information section 12.2). Outside the mature miRNA and its complementary sequence, loops had the highest rate of evolution, followed by unpaired sites, with paired sites having the lowest rate of evolution. Inside the mature miRNA, unpaired sites evolve more slowly than paired sites, whereas the opposite is true for the sequence complementary to the mature miRNA. Surprisingly, a large fraction of unpaired bulges or internal loops in the mature miRNA seem to be conserved—a pattern which may have implications for models of miRNA biogenesis and the degree of mismatch allowed in miRNA–target prediction methods. Overall these results support the qualitative model proposed in ref. 141 for the canonical progression of miRNA evolution, and show that functional constraints on the miRNA itself supersede structural constraints imposed by maintenance of the hairpin-loop.

To assess constraint on stem regions of RNA structures more generally, we compared substitution rates in stems (*S*) to those in nominally unconstrained loop regions (*L*) in a wide variety of ncRNAs (Supplementary Information section 12.3). We estimated substitution rates using a maximum likelihood framework, and compared the observed *L*/*S* ratio with the average *L*/*S* ratio estimated from published secondary structures in RFAM, which we normalized to 1.0. *L*/*S* ratios for *Drosophila* ncRNA families range from a highly constrained 2.57 for the nuclear RNase P family to 0.56 for the 5S ribosomal RNA (Supplementary Table 15).

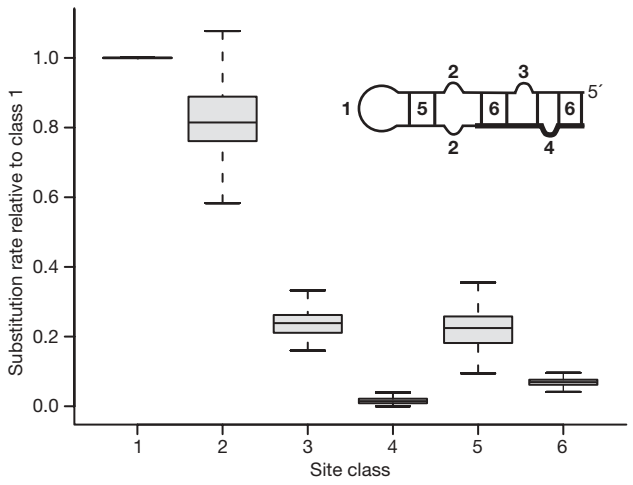


Figure 6 | Substitution rate of site classes within miRNAs. Bootstrap distributions of miRNA substitution rates. Structural alignments of miRNA precursor hairpins were partitioned into six site-classes (inset): (1) hairpin loops; unpaired sites (2) outside, (3) in the complementary region of, and (4) inside the miRNA; and base pairs (5) adjacent to and (6) involving the miRNA. Whiskers show approximate 95% confidence intervals for median differences, boxes show interquartile range.

©2007 Nature Publishing Group

Finally, we predicted a set of conserved intronic RNA structures and analysed patterns of compensatory nucleotide substitution in *D. melanogaster*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura*, *D. virilis* and *D. mojavensis* (Supplementary Information section 13). Signatures of compensatory evolution in RNA helices are detected as covarying nucleotide sites or 'covariations' (that is, two Watson–Crick bases that interact in species A replaced by a different Watson–Crick pair in species B). The number of covariations (per base pair of a helix) depends on the physical distance between the interacting nucleotides (Supplementary Fig. 9), as has been observed for the RNA helices in the *Drosophila bicoid* 3′ UTR region[142]. Short-range pairings exhibit a higher average number of covariations with a larger variance among helices than longer-range pairings. The decrease in rate of covariation with increasing distance may be explained by physical properties of a helix, which may impose selective constraints on the evolution of covarying nucleotides within a helix. Alternatively, if individual mutations at each locus are deleterious but compensated by mutations at a second locus, given sufficiently strong selection against the first deleterious mutation these epistatic fitness interactions could generate the observed distance effect[143].

Evolution of *cis*-regulatory DNAs. Comparative analyses of *cis*-regulatory sequences may provide insights into the evolutionary forces acting on regulatory components of genes, shed light on the constraints of the *cis*-regulatory code and aid in annotation of new regulatory sequences. Here we rely on two recently compiled databases, and present results comparing *cis*-regulatory modules[144] and transcription factor binding sites (derived from DNase I footprints)[145] between *D. melanogaster* and *D. simulans* (Supplementary Information section 8). We estimated mean selective constraint (*C*, the fraction of mutations removed by natural selection) relative to the 'fastest evolving intron' sites at the 5′ end of short introns, which represent putatively unconstrained neutral standards (Supplementary Information section 8.2)[146]. Note that this approach ignores the contribution of positively selected sites, potentially underestimating the fraction of functionally relevant sites[147].

Consistent with previous findings, *Drosophila cis*-regulatory sequences are highly constrained[148,149]. Mean constraint within *cis*-regulatory modules is 0.643 (95% bootstrap confidence interval = 0.621–0.662) and within footprints is 0.692 (0.655–0.723), both of which are significantly higher than mean constraint in non-coding DNA overall (0.555 (0.546–0.563)) and significantly lower than constraint at non-degenerate coding sites (0.862 (0.856–0.868)) and ncRNA genes (0.864 (0.846–0.880)) (Supplementary Fig. 10). The high level of constraint in *cis*-regulatory sequences also extends into flanking sequences, only declining to constraint levels typical of non-coding DNA 40 bp away. This is consistent with previous findings that transcription factor binding sites tend to be found in larger blocks of constraint that cluster to form *cis*-regulatory modules[150]. To understand selective constraints on nucleotides within *cis*-regulatory sequences that have direct contact with transcription factors, we estimated the selective constraint for the best match to position weight matrices within each footprint[151]; core motifs in transcription-factor-binding sites have a mean constraint of 0.773 (0.729–0.814), significantly greater than the mean for the footprints as a whole, and approaching the level of constraint found at non-degenerate coding sites and in ncRNA genes (Supplementary Fig. 10).

We next examined the variation in selective constraint across *cis*-regulatory sequences. Surprisingly, we find no evidence that selective constraint is correlated with predicted transcription-factor-binding strength (estimated as the position weight matrix score *P*-value) (Spearman's $r = 0.0681$, $P = 0.0609$). We observe significant variation in constraint both among target genes (Kruskal–Wallis tests, footprints, $P < 0.0001$; and position weight matrix matches within footprints, $P = 0.0023$) and among chromosomes (*cis*-regulatory modules, $P = 0.0186$; footprints, $P = 0.0388$; and position weight matrix matches within footprints, $P = 0.0108$; Supplementary Table 16).

Discussion and conclusion

Each new genome sequence affords novel opportunities for comparative genomic inference. What makes the analysis of these 12 *Drosophila* genomes special is the ability to place every one of these genomic comparisons on a phylogeny with a taxon separation that is ideal for asking a wealth of questions about evolutionary patterns and processes. It is without question that this phylogenomic approach places additional burdens on bioinformatics efforts, multiplying the amount of data many-fold, requiring extra care in generating multi-species alignments, and accommodating the reality that not all genome sequences have the same degree of sequencing or assembly accuracy. These difficulties notwithstanding, phylogenomics has extraordinary advantages not only for the analyses that are possible, but also for the ability to produce high-quality assemblies and accurate annotations of functional features in a genome by using closely related genomes as guides. The use of multi-species orthology provides especially convincing evidence in support of particular gene models, not only for protein-coding genes, but also for miRNA and other ncRNA genes.

Many attributes of the genomes of *Drosophila* are remarkably conserved across species. Overall genome size, number of genes, distribution of transposable element classes, and patterns of codon usage are all very similar across these 12 genomes, although *D. willistoni* is an exceptional outlier by several criteria, including its unusually skewed codon usage, increased transposable element content and potential lack of selenoproteins. At a finer scale, the number of structural changes and rearrangements is much larger; for example, there are several different rearrangements of genes in the *Hox* cluster found in these *Drosophila* species.

The vast majority of multigene families are found in all 12 genomes, although gene family size seems to be highly dynamic: almost half of all gene families change in size on at least one lineage, and a noticeable fraction shows rapid and lineage-specific expansions and contractions. Particularly notable are cases consistent with adaptive hypotheses, such as the loss of *Gr* genes in ecological specialists and the lineage-specific expansions of antimicrobial peptides and other immune effectors. All species were found to have novel genes not seen in other species. Although lineage-specific genes are challenging to verify computationally, we can confirm at least 44 protein-coding genes unique to the *melanogaster* group, and these proteins have very different properties from ancestral proteins. Similarly, although the relative abundance of transposable element subclasses across these genomes does not differ dramatically, total genomic transposable element content varies substantially among species, and several instances of lineage-specific transposable elements were discovered.

There is considerable variation among protein-coding genes in rates of evolution and patterns of positive selection. Functionally similar proteins tend to evolve at similar rates, although variation in genomic features such as gene expression level, as well as chromosomal location, are also associated with variation in evolutionary rate among proteins. Whereas broad functional classes do not seem to share patterns of positive selection, and although very few GO categories show excesses of positive selection, a number of genes involved in interactions with the environment and in sex and reproduction do show signatures of adaptive evolution. It thus seems likely that adaptation to changing environments, as well as sexual selection, shape the evolution of protein-coding genes.

Annotation of ncRNA genes across all 12 species allows comprehensive analysis of the evolutionary divergence of these genes. MicroRNA genes in particular are more conserved than protein-coding genes with respect to their primary DNA sequence, and the substitutions that do occur often have compensatory changes such that the average estimated free energy of the folding structures remains remarkably constant across the phylogeny. Surprisingly,

©2007 Nature Publishing Group

mismatches in miRNAs seem to be highly conserved, which may impact models of miRNA biogenesis and target recognition. Lineage-restricted miRNAs, however, have considerably elevated rates of change, suggesting either reduced constraint due to novel miRNAs having fewer targets, or adaptive evolution of evolutionarily young miRNAs.

Virtually any question about the function of genome features in *Drosophila* is now empowered by being embedded in the context of this 12 species phylogeny, allowing an analysis of the ways by which evolution has tuned myriad biological processes across the hundreds of millions of years spanned in total by this phylogeny. The analyses presented herein have generated more questions than they have answered, and these results represent a small fraction of that which is possible. Because much of this rich and extraordinary comparative genomic dataset remains to be explored, we believe that these 12 *Drosophila* genome sequences will serve as a powerful tool for gleaning further insight into genetic, developmental, regulatory and evolutionary processes.

METHODS

The full methods for this paper are described in Supplementary Information. Here, we describe the datasets generated by this project and their availability.

Genomic sequence. Scaffolds and assemblies for all genomic sequence generated by this project are available from GenBank (Supplementary Tables 4 and 5), and FlyBase (ftp://ftp.flybase.net/12_species_analysis/). Genome browsers are available from UCSC (http://genome.ucsc.edu/cgi-bin/hgGateway?hgsid = 98180333&clade = insect&org = 0&db = 0) and Flybase (http://flybase.org/cgi-bin/gbrowse/dmel/). BLAST search of these genomes is available at FlyBase (http://flybase.org/blast).

Predicted gene models. Consensus gene predictions for the 11 non-*melanogaster* species, produced by combining several different GLEAN runs that weight homology evidence more or less strongly, are available from FlyBase as GFF files for each species (ftp://ftp.flybase.net/12_species_analysis/). These gene models can also be accessed from the Genome Browser in FlyBase (Gbrowse; http://flybase.org/cgi-bin/gbrowse/dmel/). Predictions of non-protein-coding genes are also available in GFF format for each species, from FlyBase (ftp://ftp.flybase.net/12_species_analysis/).

Homology. Multiway homology assignments are available from FlyBase (ftp://ftp.flybase.net/12_species_analysis/), and also in the Genome Browser (Gbrowse).

Alignments. All alignment sets produced are available in FASTA format from FlyBase (ftp://ftp.flybase.net/12_species_analysis/).

PAML parameters. Output from PAML models for the alignments of single copy orthologues in the *melanogaster* group, including the *q*-value for the test for positive selection, are available from FlyBase (ftp://ftp.flybase.net/12_species_analysis/).

Received 19 July; accepted 5 October 2007.

1. Markow, T. A. & O'Grady, P. M. Drosophila biology in the genomic age. *Genetics* doi:10.1534/genetics.107.074112 (in the press).
2. Powell, J. R. *Progress and Prospects in Evolutionary Biology: The* Drosophila *Model* (Oxford Univ. Press, Oxford, 1997).
3. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287,** 2185–2195 (2000).
4. Celniker, S. E. *et al.* Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biol.* **3,** research0079.1–0079.14 (2002).
5. Richards, S. *et al.* Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and *cis*-element evolution. *Genome Res.* **15,** 1–18 (2005).
6. Myers, E. W. *et al.* A whole-genome assembly of *Drosophila*. *Science* **287,** 2196–2204 (2000).
7. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269,** 496–512 (1995).
8. Stark et al.. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* doi:10.1038/nature06340 (this issue).
9. Begun, D. J. *et al.* Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.* **5,** e310, doi:10.1371/journal.pbio.0050310 (2007).
10. Zimin, A. V., Smith, D. R., Sutton, G. & Yorke, J. A. Assembly reconciliation. *Bioinformatics* (in the press).
11. Clary, D. O. & Wolstenholme, D. R. The mitochondrial DNA molecule of *Drosophila yakuba*: nucleotide sequence, gene organization, and genetic code. *J. Mol. Evol.* **22,** 252–271 (1985).
12. Ballard, J. W. When one is not enough: introgression of mitochondrial DNA in *Drosophila*. *Mol. Biol. Evol.* **17,** 1126–1130 (2000).
13. Montooth, K. L., Abt, D. N., Hoffman, J. & Rand, D. M. Evolution of the mitochondrial DNA across twelve species of *Drosophila*. *Mol. Biol. Evol.* (submitted).
14. Salzberg, S. *et al.* Serendipitous discovery of *Wolbachia* genomes in multiple *Drosophila* species. *Genome Biol.* **6,** R23 (2005).
15. Edgar, R. C. & Myers, E. W. PILER: identification and classification of genomic repeats. *Bioinformatics* **21,** i152–i158 (2005).
16. Smith, C. D. *et al.* Improved repeat identification and masking in Dipterans. *Gene* **389,** 1–9 (2007).
17. Li, Q. *et al.* ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole shotgun. *PloS Comput. Biol.* **1,** e43 (2005).
18. Bergman, C. M., Quesneville, H., Anxolabehere, D. & Ashburner, M. Recurrent insertion and duplication generate networks of transposable element sequences in the *Drosophila melanogaster* genome. *Genome Biol.* **7,** R112 (2006).
19. Guigo, R., Knudsen, S., Drake, N. & Smith, T. Prediction of gene structure. *J. Mol. Biol.* **226,** 141–157 (1992).
20. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5,** 59 (2004).
21. Gross, S. S. & Brent, M. R. Using multiple alignments to improve gene prediction. *J. Comput. Biol.* **13,** 379–393 (2006).
22. Gross, S. S., Do, C. B. & Batzoglou, S. in *BCATS 2005 Symposium Proc.* 82 (2005).
23. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14,** 988–995 (2004).
24. Slater, G. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6,** 31 (2005).
25. Chatterji, S. & Pachter, L. Reference based annotation with GeneMapper. *Genome Biol.* **7,** R29 (2006).
26. Souvorov, A. *et al.* in *NCBI News Fall/Winter, NIH Publication No. 04-3272* (eds Benson, D & Wheeler, D) (2006).
27. Honeybee Genome Sequencing Consortium. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* **443,** 931–949 (2006).
28. Elsik, C. G. *et al.* Creating a honey bee consensus gene set. *Genome Biol.* **8,** R13 (2007).
29. Zhang, Y., Sturgill, D., Parisi, M., Kumar, S. & Oliver, B. Constraint and turnover in sex-biased gene expression in the genus *Drosophila*. *Nature* doi:10.1038/nature06323 (this issue).
30. Manak, J. R. *et al.* Biological function of unannotated transcription during the early development of *Drosophila melanogaster*. *Nature Genet.* **38,** 1151–1158 (2006).
31. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278,** 631–637 (1997).
32. Bhutkar, A., Russo, S., Smith, T. F. & Gelbart, W. M. Techniques for multi-genome synteny analysis to overcome assembly limitations. *Genome Informatics* **17,** 152–161 (2006).
33. Heger, A. & Ponting, C. Evolutionary rate analyses of orthologues and paralogues from twelve *Drosophila* genomes. doi:10.1101/gr6249707 *Genome Res.* (in the press).
34. Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302,** 205–217 (2000).
35. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–521 (2004).
36. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
37. Harrison, P. M., Milburn, D., Zhang, Z., Bertone, P. & Gerstein, M. Identification of pseudogenes in the *Drosophila melanogaster* genome. *Nucleic Acids Res.* **31,** 1033–1037 (2003).
38. Bosco, G., Campbell, P., Leiva-Neto, J. & Markow, T. Analysis of *Drosophila* species genome size and satellite DNA content reveals significant differences among strains as well as between species. *Genetics* doi:10.1534/Genetics107.075069 (in the press).
39. Ranz, J. *et al.* Principles of genome evolution in the *Drosophila melanogaster* species group. *PLoS Biol.* **5,** e152, doi:10.1371/journal.pbio.0050152 (2007).
40. Noor, M. A. F., Garfield, D. A., Schaeffer, S. W. & Machado, C. A. Divergence between the *Drosophila pseudoobscura* and *D. persimilis* genome sequences in relation to chromosomal inversions. *Genetics* doi:10.1534/genetics.107.070672 (in the press).
41. Lewis, E. B. A gene complex controlling segmentation in *Drosophila*. *Nature* **276,** 565–570 (1978).
42. Negre, B., Ranz, J. M., Casals, F., Caceres, M. & Ruiz, A. A new split of the *Hox* gene complex in *Drosophila*: relocation and evolution of the gene labial. *Mol. Biol. Evol.* **20,** 2042–2054 (2003).
43. Von Allmen, G. *et al.* Splits in fruitfly *Hox* gene complexes. *Nature* **380,** 116 (1996).
44. Negre, B. & Ruiz, A. HOM-C evolution in *Drosophila*: is there a need for *Hox* gene clustering? *Trends Genet.* **23,** 55–59 (2007).
45. Dowsett, A. P. & Young, M. W. Differing levels of dispersed repetitive DNA among closely related species of *Drosophila*. *Proc. Natl Acad. Sci.* **79,** 4570–4574 (1982).
46. Kapitonov, V. V. & Jurka, J. DNAREP1_DM. (Repbase Update Release 3.4, 1999).
47. Kapitonov, V. V. & Jurka, J. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. *Proc. Natl Acad. Sci. USA* **100,** 6569–6574 (2003).

©2007 Nature Publishing Group

48. Singh, N. D., Arndt, P. F. & Petrov, D. A. Genomic heterogeneity of background substitutional patterns in *Drosophila melanogaster*. *Genetics* **169,** 709–722 (2004).
49. Yang, H.-P., Hung, T.-L., You, T.-L. & Yang, T.-H. Genomewide comparative analysis of the highly abundant transposable element *DINE-1* suggests a recent transpositional burst in *Drosophila yakuba*. *Genetics* **173,** 189–196 (2006).
50. Yang, H.-P. & Barbash, D. Abundant and species-specific miniature inverted-repeat transposable elements in 12 *Drosophila* genomes. *Genome Biol.* (submitted).
51. Wilder, J. & Hollocher, H. Mobile elements and the genesis of microsatellites in dipterans. *Mol. Biol. Evol.* **18,** 384–392 (2001).
52. Marzo, M., Puig, M. & Ruiz, A. The foldback-like element *Galileo* belongs to the P superfamily of DNA transposons and is widespread within the genus *Drosophila*. *Proc. Natl Acad. Sci. USA* (submitted).
53. Casola, C., Lawing, A., Betran, E. & Feschotte, C. PIF-like transposons are common in *Drosophila* and have been repeatedly domesticated to generate new host genes. *Mol. Biol. Evol.* **24,** 1872–1888 (2007).
54. Abad, J. P. *et al.* Genomic analysis of *Drosophila melanogaster* telomeres: full-length copies of *HeT-A* and *TART* elements at telomeres. *Mol. Biol. Evol.* **21,** 1613–1619 (2004).
55. Abad, J. P. *et al. TAHRE*, a novel telomeric retrotransposon from *Drosophila melanogaster*, reveals the origin of *Drosophila* telomeres. *Mol. Biol. Evol.* **21,** 1620–1624 (2004).
56. Blackburn, E. H. Telomerases. *Annu. Rev. Biochem.* **61,** 113–129 (1992).
57. Villasante, A. *et al. Drosophila* telomeric retrotransposons derived from an ancestral element that as recruited to replace telomerase. *Genome Res.* (in the press).
58. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432,** 695–716 (2004).
59. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
60. *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282,** 2012–2018 (1998).
61. Mount, S. M., Gotea, V., Lin, C. F., Hernandez, K. & Makalowski, W. Spliceosomal small nuclear RNA genes in 11 insect genomes. *RNA* **13,** 5–14 (2007).
62. Schneider, C., Will, C. L., Brosius, J., Frilander, M. J. & Luhrmann, R. Identification of an evolutionarily divergent U11 small nuclear ribonucleoprotein particle in *Drosophila*. *Proc. Natl Acad. Sci. USA* **101,** 9584–9589 (2004).
63. Deng, X. & Meller, V. H. Non-coding RNA in fly dosage compensation. *Trends Biochem. Sci.* **31,** 526–532 (2006).
64. Amrien, H. & Axel, R. Genes expressed in neurons of adult male *Drosophila*. *Cell* **88,** 459–469 (1997).
65. Park, S.-W. *et al.* An evolutionarily conserved domain of roX2 RNA is sufficient for induction of H4-Lys16 acetylation on the *Drosophila* X chromosome. *Genetics* (in the press).
66. Stage, D. E. & Eickbush, T. H. Sequence variation within the rRNA gene loci of 12 *Drosophila* species. *Genome Res.* (in the press).
67. Hahn, M. W., De Bie, T., Stajich, J. E., Nguyen, C. & Cristianini, N. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res.* **15,** 1153–1160 (2005).
68. Hahn, M. W., Han, M. V. & Han, S.-G. Gene family evolution across 12 *Drosophila* genomes. *PLoS Biol.* **3,** e197 (2007).
69. Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A. & Begun, D. J. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc. Natl Acad. Sci. USA* **103,** 9935–9939 (2006).
70. Ponce, R. & Hartl, D. L. The evolution of the novel *Sdic* gene cluster in *Drosophila melanogaster*. *Gene* **376,** 174–183 (2006).
71. Arguello, J. R., Chen, Y., Tang, S., Wang, W. & Long, M. Originiation of an X-linked testes chimeric gene by illegitimate recombination in *Drosophila*. *PLoS Genet.* **2,** e77 (2006).
72. Begun, D. J., Lindfore, H. A., Thompson, M. E. & Holloway, A. K. Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags. *Genetics* **172,** 1675–1681 (2006).
73. Betran, E., Thornton, K. & Long, M. Retroposed new genes out of the X in *Drosophila*. *Genome Res.* **12,** 1854–1859 (2002).
74. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21,** 650–659 (2005).
75. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13,** 555–556 (1997).
76. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
77. Storey, J. D. A direct approach to false discovery rates. *J. R. Stat. Soc. B* **64,** 479–498 (2002).
78. Larracuente, A. M. *et al.* Evolution of protein-coding genes in *Drosophila*. *Trends Genet.* (submitted).
79. Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A. M. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155,** 431–449 (2000).
80. Bergman, C. M. *et al.* Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. *Genome Biol.* **3,** research0086.1–0086.20 (2002).
81. Bierne, N. & Eyre Walker, A. C. The genomic rate of adaptive amino acid substitution in *Drosophila*. *Mol. Biol. Evol.* **21,** 1350–1360 (2004).
82. Sawyer, S. A., Kulathinal, R. J., Bustamante, C. D. & Hartl, D. L. Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. *J. Mol. Evol.* **57** (suppl. 1), S154–S164 (2003).
83. Sawyer, S. A., Parsch, J., Zhang, Z. & Hartl, D. L. Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. *Proc. Natl Acad. Sci. USA* **104,** 6504–6510 (2007).
84. Smith, N. G. & Eyre-Walker, A. Adaptive protein evolution in *Drosophila*. *Nature* **415,** 1022–1024 (2002).
85. Welch, J. J. Estimating the genomewide rate of adaptive protein evolution in *Drosophila*. *Genetics* **173,** 821–837 (2006).
86. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. Why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. USA* **102,** 14338–14343 (2005).
87. Drummond, D. A., Raval, A. & Wilke, C. O. A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.* **23,** 327–337 (2006).
88. Pal, C., Papp, B. & Hurst, L. D. Highly expressed genes in yeast evolve slowly. *Genetics* **158,** 927–931 (2001).
89. Pal, C., Papp, B. & Lercher, M. J. An integrated view of protein evolution. *Nature Rev. Genet.* **7,** 337–348 (2006).
90. Wall, D. P. *et al.* Functional genomic analysis of the rates of protein evolution. *Proc. Natl Acad. Sci. USA* **102,** 5483–5488 (2005).
91. Rocha, E. P. The quest for the universals of protein evolution. *Trends Genet.* **22,** 412–416 (2006).
92. Huntley, M. A. & Clark, A. G. Evolutionary analysis of amino acid repeats across the genomes of 12 *Drosophila* species. *Mol. Biol. Evol.* (in the press).
93. Charlesworth, B., Coyne, J. A. & Barton, N. H. The relative rates of evolution of sex chromosomes and autosomes. *Am. Nat.* **130,** 113–146 (1987).
94. Larsson, J. & Meller, V. H. Dosage compensation, the origin and the afterlife of sex chromosomes. *Chromosome Res.* **14,** 417–431 (2006).
95. Riddle, N. C. & Elgin, S. C. The dot chromosome of *Drosophila*: insights into chromatin states and their change over evolutionary time. *Chromosome Res.* **14,** 405–416 (2006).
96. Gordo, I. & Charlesworth, B. Genetic linkage and molecular evolution. *Curr. Biol.* **11,** R684–R686 (2001).
97. Singh, N. D., Larracuente, A. M. & Clark, A. G. Contrasting the efficacy of selection on the X and autosomes in *Drosophila*. *Mol. Biol. Evol.* (submitted).
98. Bhutkar, A., Russo, S. M., Smith, T. F. & Gelbart, W. M. Genome scale analysis of positionally relocated genes. *Genome Res.* (in the press).
99. Akashi, H. & Eyre-Walker, A. Translational selection and molecular evolution. *Curr. Opin. Genet. Dev.* **8,** 688–693 (1998).
100. Akashi, H., Kliman, R. M. & Eyre-Walker, A. Mutation pressure, natural selection, and the evolution of base composition in *Drosophila*. *Genetica (Dordrecht)* **102–103,** 49–60 (1998).
101. Bulmer, M. The selection–mutation–drift theory of synonymous codon usage. *Genetics* **129,** 897–908 (1991).
102. McVean, G. A. T. & Charlesworth, B. A population genetic model for the evolution of synonymous codon usage: Patterns and predictions. *Genet. Res.* **74,** 145–158 (1999).
103. Sharp, P. M. & Li, W. H. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* **24,** 28–38 (1986).
104. Akashi, H. & Schaeffer, S. W. Natural selection and the frequency distributions of "silent" DNA polymorphism in *Drosophila*. *Genetics* **146,** 295–307 (1997).
105. Powell, J. R., Sezzi, E., Moriyama, E. N., Gleason, J. M. & Caccone, A. Analysis of a shift in codon usage in *Drosophila*. *J. Mol. Evol.* **57,** S214–S225 (2003).
106. Anderson, C. L., Carew, E. A. & Powell, J. R. Evolution of the *Adh* locus in the *Drosophila willistoni* group: The loss of an intron, and shift in codon usage. *Mol. Biol. Evol.* **10,** 605–618 (1993).
107. Rodriguez-Trelles, F., Tarrio, R. & Ayala, F. J. Switch in codon bias and increased rates of amino acid substitution in the *Drosophila saltans* species group. *Genetics* **153,** 339–350 (1999).
108. Rodriguez-Trelles, F., Tarrio, R. & Ayala, F. J. Evidence for a high ancestral GC content in *Drosophila*. *Mol. Biol. Evol.* **17,** 1710–1717 (2000).
109. Rodriguez-Trelles, F., Tarrio, R. & Ayala, F. J. Fluctuating mutation bias and the evolution of base composition in *Drosophila*. *J. Mol. Evol.* **50,** 1–10 (2000).
110. Heger, A. & Ponting, C. Variable strength of translational selection among twelve *Drosophila* species. *Genetics* (in the press).
111. Vicario, S., Moriyama, E. N. & Powell, J. R. Codon Usage in Twelve Species of *Drosophila*. *BMC Evol. Biol.* (submitted).
112. Singh, N. D., Arndt, P. F. & Petrov, D. A. Minor shift in background substitutional patterns in the *Drosophila saltans* and *willistoni* lineages is insufficient to explain GC content of coding sequences. *BMC Biol.* **4,** 10.1186/1741–7007–4-37 (2006).
113. Akashi, H. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. *Genetics* **139,** 1067–1076 (1995).
114. Akashi, H. Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. *Genetics* **144,** 1297–1307 (1996).

©2007 Nature Publishing Group

115. Akashi, H. *et al.* Molecular evolution in the *Drosophila melanogaster* species subgroup: Frequent parameter fluctuations on the timescale of molecular divergence. *Genetics* **172,** 1711–1726 (2006).
116. Bauer DuMont, V., Fay, J. C., Calabrese, P. P. & Aquadro, C. F. DNA variability and divergence at the *Notch* locus in *Drosophila melanogaster* and *D. simulans*: a case of accelerated synonymous site divergence. *Genetics* **167,** 171–185 (2004).
117. McVean, G. A. & Vieira, J. The evolution of codon preferences in *Drosophila*: a maximum-likelihood approach to parameter estimation and hypothesis testing. *J. Mol. Evol.* **49,** 63–75 (1999).
118. Nielsen, R., Bauer DuMont, V., Hubisz, M. J. & Aquadro, C. F. Maximum likelihood estimation of ancestral codon usage bias parameters in *Drosophila*. *Mol. Biol. Evol.* **24,** 228–235 (2007).
119. Begun, D. J. The frequency distribution of nucleotide variation in *Drosophila* simulans. *Mol. Biol. Evol.* **18,** 1343–1352 (2001).
120. Singh, N. S., Bauer DuMont, V. L., Hubisz, M. J., Nielsen, R. & Aquadro, C. F. Patterns of mutation and selection at synonymous sites in *Drosophila*. *Mol. Biol. Evol.* doi:10.1093/mobev.msm196 (in the press).
121. McBride, C. S. & Arguello, J. R. Five *Drosophila* genomes reveal non-neutral evolution and the signature of host specialization in the chemoreceptor superfamily. *Genetics* (in the press).
122. Vieira, F. G., Sanchez-Gracia, A. & Rozas, J. Comparative genomic analysis of the odorant-binding protein family in 12 *Drosophila* genomes: Purifying selection and birth-and-death evolution. *Genome Biol.* **8,** 235 (2007).
123. Gardiner, A., Barker, D., Butilne, R. K., Jordan, W. C. & Ritchie, M. G. *Drosophila* chemoreceptor evolution: Selection, specialisation and genome size. *Genome Biol.* (submitted).
124. McBride, C. S. Rapid evolution of smell and taste receptor genes during host specialization in *Drosophila sechellia*. *Proc. Natl Acad. Sci. USA* **104,** 4996–5001 (2007).
125. Ranson, H. *et al.* Evolution of supergene families associated with insecticide resistance. *Science* **298,** 179–181 (2002).
126. Tijet, N., Helvig, C. & Feyereisen, R. The cytochrome P450 gene superfamily in *Drosophila melanogaster*. *Gene* **262,** 189–198 (2001).
127. Claudianos, C. *et al.* A deficit of detoxification enzymes: pesticide sensitivity and environmental response in the honeybee. *Insect Mol. Biol.* **15,** 615–636 (2006).
128. Low, W. L. *et al.* Molecular evolution of glutathione *S*-transferases in the genus *Drosophila*. *Genetics* (in the press).
129. Castellano, S. *et al. In silico* identification of novel selenoproteins in the *Drosophila melanogaster* genome. *EMBO Rep.* **2,** 697–702 (2001).
130. Martin-Romero, F. J. *et al.* Selenium metabolism in *Drosophila*: selenoproteins, selenoprotein mRNA expression, fertility, and mortality. *J. Biol. Chem.* **276,** 29798–29804 (2001).
131. Lemaitre, B. & Hoffmann, J. The host defense of *Drosophila melanogaster*. *Annu. Rev. Immunol.* **25,** 697–743 (2007).
132. Hughes, A. L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335,** 167–170 (1988).
133. Murphy, P. M. Molecular mimicry and the generation of host defense protein diversity. *Cell* **72,** 823–826 (1993).
134. Schlenke, T. A. & Begun, D. J. Natural selection drives *Drosophila* immune system evolution. *Genetics* **164,** 1471–1480 (2003).
135. Sackton, T. B. *et al.* The evolution of the innate immune system across *Drosophila*. *Nature Genet.* (submitted).
136. Civetta, A. & Singh, R. S. High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in *Drosophila melanogaster* and *Drosophila virilis* group species. *J. Mol. Evol.* **41,** 1085–1095 (1995).
137. Civetta, A. Shall we dance or shall we fight? Using DNA sequence data to untangel controversies surrounding sexual selection. *Genome* **46,** 925–929 (2003).
138. Clark, N. L., Aagard, J. E. & Swanson, W. J. Evolution of reproductive proteins from animals and plants. *Reproduction* **131,** 11–22 (2006).
139. Haerty, W. *et al.* Evolution in the fast lane: rapidly evolving sex- and reproduction-related genes in *Drosophila* species. *Genetics* (in the press).
140. Lu, J. *et al.* Adaptive evolution of newly-emerged microRNA genes in *Drosophila*. *Mol. Biol. Evol.* (submitted).
141. Lai, E. C., Tomancak, P., Williams, R. W. & Rubin, G. M. Computational identification of *Drosophila* microRNA genes. *Genome Biol.* **4,** R42 (2003).
142. Parsch, J., Braverman, J. M. & Stephan, W. Comparative sequence analysis and patterns of covariation in RNA secondary structures. *Genetics* **154,** 909–921 (2000).
143. Stephan, W. The rate of compensatory evolution. *Genetics* **144,** 419–426 (1996).
144. Gallo, S. M., Li, L., Hu, Z. & Halfon, M. S. REDfly: a Regulatory Element Database for *Drosophila*. *Bioinformatics* **22,** 381–383 (2006).
145. Bergman, C. M., Carlson, J. W. & Celniker, S. E. *Drosophila* DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster*. *Bioinformatics* **21,** 1747–1749 (2005).
146. Halligan, D. L. & Keightley, P. D. Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Res.* **16,** 875–884 (2006).
147. Andolfatto, P. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* **437,** 1149–1152 (2005).
148. Bird, C. P., Stranger, B. E. & Dermitzakis, E. T. Functional variation and evolution of non-coding DNA. *Curr. Opin. Genet. Dev.* **16,** 559–564 (2006).
149. Wittkopp, P. J. Evolution of *cis*-regulatory sequence and function in Diptera. *Heredity* **97,** 139–147 (2006).
150. Ludwig, M. Z., Patel, N. H. & Kreitman, M. Functional analysis of *eve stripe 2* enhancer evolution in *Drosophila*. *Development* **125,** 949–958 (1998).
151. Down, A. T. A., Bergman, C. M., Su, J. & Hubbard, T. J. P. Large scale discovery of promoter motifs in *Drosophila melanogaster*. *PloS Comput. Biol.* **3,** e7 (2007).
152. Tamura, K., Subramanian, S. & Kumar, S. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Mol. Biol. Evol.* **21,** 36–44 (2004).
153. Kumar, S., Tamura, K. & Nei, M. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinform.* **5,** 150–163 (2004).
154. Pollard, D. A., Iyer, V. N., Moses, A. M. & Eisen, M. B. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS Genet.* **2,** e173 (2006).
155. Bhutkar, A., Gelbart, W. M. & Smith, T. F. Inferring genome-scale rearrangement phylogeny and ancestral gene order: A *Drosophila* case study. *Genome Biol.* (in the press).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Agencourt Bioscience Corporation, The Broad Institute of MIT and Harvard and the Washington University Genome Sequencing Center were supported by grants and contracts from the National Human Genome Research Insititute (NHGRI). T.C. Kaufman acknowledges support from the Indian Genomics Initiative.

Author Contributions The laboratory groups of A. G. Clark (including A. M. Larracuente, T. B. Sackton, and N. D. Singh) and Michael B. Eisen (including V. N. Iyer and D. A. Pollard) played the part of coordinating the primary writing and editing of the manuscript with the considerable help of D. R. Smith, C. M. Bergman, W. M. Gelbart, B. Oliver, T. A. Markow, T. C. Kaufman and M. Kellis. D. R. Smith served as primary coordinator for the assemblies. The remaining authors contributed either through their efforts in sequence production, assembly and annotation, or in the analysis of specific topics that served as the focus of more than 40 companion papers.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.G.C. (ac347@cornell.edu), M.B.E. (mbeisen@lbl.gov), D.R.S. (douglas.smith@agencourt.com), C.M.B. (casey.bergman@manchester.ac.uk), W.G. (gelbart@morgan.harvard.edu), B.O. (oliver@helix.nih.gov), T.A.M. (tmarkow@public.arl.arizona.edu), T.C.K. (kaufman@indiana.edu), M.K. (manoli@mit.edu), V.N.I. (venky@berkeley.edu), T.B.S. (tbs7@cornell.edu), A.M.L. (aml69@cornell.edu), D.A.P. (danielapollard@alum.bowdoin.edu), N.D.S. (nds25@cornell.edu), or collectively to 12flies@morgan.harvard.edu.

***Drosophila* 12 Genomes Consortium**

Project Leaders Andrew G. Clark[1], Michael B. Eisen[2,3], Douglas R. Smith[4], Casey M. Bergman[5], Brian Oliver[6], Therese A. Markow[7], Thomas C. Kaufman[8], Manolis Kellis[9,10] & William Gelbart[11,12]

Annotation Coordination Venky N. Iyer[13] & Daniel A. Pollard[14]

Analysis/Writing Coordination Timothy B. Sackton[1,15], Amanda M. Larracuente[1] & Nadia D. Singh[1]

Sequencing, Assembly, Annotation and Analysis Contributors Jose P. Abad[16], Dawn N. Abt[17], Boris Adryan[18], Montserrat Aguade[19], Hiroshi Akashi[20], Wyatt W. Anderson[21], Charles F. Aquadro[1], David H. Ardell[22], Roman Arguello[23], Carlo G. Artieri[24], Daniel A. Barbash[1], Daniel Barker[25], Paolo Barsanti[26], Phil Batterham[27], Serafim Batzoglou[28], Dave Begun[29], Arjun Bhutkar[11,30], Enrico Blanco[31], Stephanie A. Bosak[4], Robert K. Bradley[32], Adrianne D. Brand[4], Michael R. Brent[33], Angela N. Brooks[13], Randall H. Brown[33], Roger K. Butlin[34], Corrado Caggese[26], Brian R. Calvi[35], A. Bernardo de Carvalho[36], Anat Caspi[32], Sergio Castrezana[37], Susan E. Celniker[2], Jean L. Chang[10], Charles Chapple[31], Sourav Chatterji[38,39], Asif Chinwalla[40], Alberto Civetta[41], Sandra W. Clifton[40], Josep M. Comeron[42], James C. Costello[43], Jerry A. Coyne[23], Jennifer Daub[44], Robert G. David[4], Arthur L. Delcher[45], Kim Delehaunty[40], Chuong B. Do[28], Heather Ebling[4], Kevin Edwards[46], Thomas Eickbush[47], Jay D. Evans[48], Alan Filipski[49], Sven Findeiß[49,50], Eva Freyhult[22], Lucinda Fulton[40], Robert Fulton[40], Ana C. L. Garcia[51], Anastasia Gardiner[25], David A. Garfield[52], Barry E. Garvin[4], Greg Gibson[53], Don Gilbert[8], Sante Gnerre[10], Jennifer Godfrey[40], Robert Good[27], Valer Gotea[20], Brenton Gravely[54], Anthony J. Greenberg[1], Sam Griffiths-Jones[5,44], Samuel Gross[28], Roderic Guigo[31,55], Erik A. Gustafson[4], Wilfried Haerty[24], Matthew W. Hahn[8,43], Daniel L. Halligan[56], Aaron L. Halpern[57], Gillian M. Halter[20], Mira V. Han[43], Andreas Heger[58,59], LaDeana Hillier[40], Angie S. Hinrichs[60], Ian Holmes[32], Roger A. Hoskins[2], Melissa J. Hubisz[61], Dan Hultmark[62], Melanie A. Huntley[1], David B. Jaffe[10], Santosh Jagadeeshan[24], William R. Jeck[63], Justin Johnson[57], Corbin D. Jones[63], William C. Jordan[64], Gary H. Karpen[13,65], Eiko Kataoka[66], Peter D. Keightley[56], Pouya Kheradpour[9], Ewen F. Kirkness[57], Leonardo B. Koerich[36], Karsten Kristiansen[67], Dave

©2007 Nature Publishing Group

Kudrna[68], Rob J. Kulathinal[69], Sudhir Kumar[49,70], Roberta Kwok[8], Eric Lander[10], Charles H. Langley[29], Richard Lapoint[71], Brian P. Lazzaro[72], So-Jeong Lee[68], Lisa Levesque[41], Ruiqiang Li[67,73], Chiao-Feng Lin[20], Michael F. Lin[9,10], Kerstin Lindblad-Toh[10], Ana Llopart[42], Manyuan Long[23], Lloyd Low[27], Elena Lozovsky[69], Jian Lu[23], Meizhong Luo[68], Carlos A. Machado[7], Wojciech Makalowski[20], Mar Marzo[74], Muneo Matsuda[66], Luciano Matzkin[7], Bryant McAllister[42], Carolyn S. McBride[29], Brendan McKernan[4], Kevin McKernan[4], Maria Mendez-Lago[16], Patrick Minx[40], Michael U. Mollenhauer[20], Kristi Montooth[17], Stephen M. Mount[45,75], Xu Mu[20], Eugene Myers[76], Barbara Negre[77], Stuart Newfeld[70], Rasmus Nielsen[78], Mohamed A. F. Noor[52], Patrick O'Grady[71], Lior Pachter[38], Montserrat Papaceit[19], Matthew J. Parisi[4], Michael Parisi[6], Leopold Parts[9], Jakob S. Pedersen[60,79], Graziano Pesole[80], Adam M. Phillippy[45], Chris P. Ponting[58,59], Mihai Pop[45], Damiano Porcelli[26], Jeffrey R. Powell[81], Sonja Prohaska[49,82], Kim Pruitt[83], Marta Puig[74], Hadi Quesneville[84], Kristipati Ravi Ram[1], David Rand[17], Matthew D. Rasmussen[9], Laura K. Reed[53], Robert Reenan[85], Amy Reily[40], Karin A. Remington[57], Tania T. Rieger[86], Michael G. Ritchie[25], Charles Robin[27], Yu-Hui Rogers[57], Claudia Rohde[87], Julio Rozas[19], Marc J. Rubenfield[4], Alfredo Ruiz[74], Susan Russo[11,12], Steven L. Salzberg[45], Alejandro Sanchez-Gracia[19,88], David J. Saranga[4], Hajime Sato[66], Stephen W. Schaeffer[20], Michael C. Schatz[45], Todd Schlenke[89], Russell Schwartz[90], Carmen Segarra[19], Rama S. Singh[24], Laura Sirot[1], Marina Sirota[91], Nicholas B. Sisneros[68], Chris D. Smith[65,92], Temple F. Smith[30], John Spieth[40], Deborah E. Stage[47], Alexander Stark[9,10], Wolfgang Stephan[93], Robert L. Strausberg[57], Sebastian Strempel[93], David Sturgill[6], Granger Sutton[57], Granger G. Sutton[57], Wei Tao[4], Sarah Teichmann[18], Yoshiko N. Tobari[94], Yoshihiko Tomimura[95], Jason M. Tsolas[4], Vera L. S. Valente[51], Eli Venter[57], J. Craig Venter[57], Saverio Vicario[81], Filipe G. Vieira[19], Albert J. Vilella[19,96], Alfredo Villasante[16], Brian Walenz[57], Jun Wang[67,73], Marvin Wasserman[97], Thomas Watts[7], Derek Wilson[18], Richard K. Wilson[40], Rod A. Wing[68], Mariana F. Wolfner[1], Alex Wong[1], Gane Ka-Shu Wong[73,98], Chung-I Wu[23], Gabriel Wu[32], Daisuke Yamamoto[99], Hsiao-Pei Yang[1], Shiaw-Pyng Yang[40], James A. Yorke[100], Kiyohito Yoshida[101], Evgeny Zdobnov[102], Peili Zhang[11,12], Yu Zhang[6], Aleksey V. Zimin[100], Broad Institute Genome Sequencing Platform* & Broad Institute Whole Genome Assembly Team*

[1]Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853, USA. [2]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [3]Center for Integrative Genomics, Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, California 94720, USA. [4]Agencourt Bioscience Corporation, Beverly, Massachusetts 01915, USA. [5]Faculty of Life Sciences, University of Manchester, Manchester M13 9PT, UK. [6]Laboratory of Cellular and Developmental Biology, National Institutes of Health, Bethesda, Maryland 20892, USA. [7]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA. [8]Department of Biology, Indiana University, Bloomington, Indiana 47405, USA. [9]Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts 02139, USA. [10]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [11]Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [12]FlyBase, The Biological Laboratories, Harvard University, Cambridge, Massachusetts 02138, USA. [13]Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, California 94720, USA. [14]Biophysics Graduate Group, University of California at Berkeley, Berkeley, California 94720, USA. [15]Field of Ecology and Evolutionary Biology, Cornell University, Ithaca, New York 14853, USA. [16]Centro de Biología Molecular Severo Ochoa, Universidad Autónoma de Madrid, Madrid 28049, Spain. [17]Department of Ecology and Evolutionary Biology, Brown University, Providence, Rhode Island 02912, USA. [18]Structural Studies Division, MRC Laboratory of Molecular Biology, Cambridge CB2 2QH, UK. [19]Departament de Genètica, Universitat de Barcelona, Barcelona 08071, Spain. [20]Department of Biology, Pennsylvania State University, University Park, Pennsylvania 16802, USA. [21]Department of Genetics, University of Georgia, Athens, Georgia 30602, USA. [22]Linnaeus Centre for Bioinformatics, Uppsala Universitet, Uppsala, SE-75124, Sweden. [23]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA. [24]Department of Biology, McMaster University, Hamilton, Ontario, L8S 4K1, Canada. [25]School of Biology, University of St. Andrews, Fife KY16 9TH, UK. [26]Dipartimento di Genetica e Microbiologia dell'Università di Bari, Bari, 70126, Italy. [27]Department of Genetics, University of Melbourne, Melbourne 3010, Australia. [28]Computer Science Department, Stanford University, Stanford, California 94305, USA. [29]Section of Evolution and Ecology and Center for Population Biology, University of California at Davis, Davis, California 95616, USA. [30]BioMolecular Engineering Research Center, Boston University, Boston, Massachusetts 02215, USA. [31]Research Group in Biomedical Informatics, Institut Municipal d'Investigacio Medica, Universitat Pompeu Fabra, Barcelona 08003, Catalonia, Spain. [32]Department of Bioengineering, University of California at Berkeley, Berkeley, California 94720, USA. [33]Laboratory for Computational Genomics, Washington University, St Louis, Missouri 63108, USA. [34]Animal and Plant Sciences, The University of Sheffield, Sheffield S10 2TN, UK. [35]Department of Biology, Syracuse University, Syracuse, New York 13244, USA. [36]Departamento de Genética, Universidade Federal do Rio de Janeiro, Rio de Janeiro 21944-970, Brazil. [37]Tucson Stock Center, Tucson, Arizona 85721, USA. [38]Department of Mathematics, University of California at Berkeley, Berkeley, California 94720, USA. [39]Genome Center, University of California at Davis, Davis, California 95616, USA. [40]Genome Sequencing Center, Washington University School of Medicine, St Louis, Missouri 63108, USA. [41]Department of Biology, University of Winnipeg, Winnipeg, Manitoba R3B 2E9, Canada. [42]Department of Biological Sciences, University of Iowa, Iowa City, Iowa 52242, USA. [43]School of Informatics, Indiana University, Bloomington, Indiana 47405, USA. [44]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [45]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland 20742, USA. [46]Department of Biological Sciences, Illinois State University, Normal, Illinois 61790, USA. [47]Department of Biology, University of Rochester, Rochester, New York 14627, USA. [48]Bee Research Lab, USDA-ARS, Beltsville, Maryland 20705, USA. [49]Center for Evolutionary Functional Genomics, Biodesign Institute, Arizona State University, Tempe, Arizona 85287, USA. [50]Department of Computer Science, University of Leipzig, Leipzig 04107, Germany. [51]Departamento de Genética, Universidade Federal do Rio Grande do Sul, Porto Alegre/RS 68011, Brazil. [52]Department of Biology, Duke University, Durham, New Carolina 27708, USA. [53]Department of Genetics, North Carolina State University, Raleigh, North Caroline 27695, USA. [54]Health Center, University of Connecticut, Farmington, Connecticut 06030, USA. [55]Center of Genomic Regulation, Barcelona 8003, Catalonia, Spain. [56]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, UK. [57]J. Craig Venter Institute, Rockville, Maryland 20850, USA. [58]MRC Functional Genetics Unit, University of Oxford, Oxford OX1 3QX, UK. [59]Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford OX1 3QX, UK. [60]Center for Biomolecular Science and Engineering, University of California at Santa Cruz, Santa Cruz, California 95064, USA. [61]Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA. [62]Umeå Center for Molecular Pathogenesis, Umeå University, Umeå SE-90187, Sweden. [63]Department of Biology and Carolina Center for Genome Sciences, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [64]Institute of Zoology, Regent's Park, London NW1 4RY, UK. [65]Drosophila Heterochromatin Genome Project, Department of Genome and Computational Biology, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [66]Kyorin University, School of Medicine, Mitaka, Tokyo 181-8611, Japan. [67]Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M DK-5230, Denmark [68]Arizona Genomics Institute, Department of Plant Sciences and BIO5, University of Arizona, Tucson, Arizona 85721, USA. [69]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [70]School of Life Sciences, Arizona State University, Tempe, Arizona 85287, USA. [71]Department of Environmental Science, Policy and Management, University of California at Berkeley, Berkeley, California 94720, USA. [72]Department of Entomology, Cornell University, Ithaca, New York 14853, USA. [73]Beijing Genomics Institute at ShenZhen, ShenZhen 518083, China. [74]Departament Genètica i Microbiologia, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain. [75]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, Maryland 20742, USA. [76]Janelia Farm Research Campus, Howard Hughes Medical Institue, Ashburn, Viginia 20147-2408, USA. [77]Department of Zoology, University of Cambridge, Cambridge, CB2 3EJ, UK. [78]Institue of Biology, University of Copenhagen, DK-2100 Copenhagen Ø, Denmark. [79]Bioinformatics Centre, Department of Molecular Biology, University of Copenhagen, DK-2200 Copenhagen N, Denmark. [80]Dipartimento di Biochimica e Biologia Molecolare, Università di Bari and Istituto Tecnologie Biomediche del Consiglio Nazionale delle Ricerche, Bari 70126, Italy. [81]Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut 06520, USA. [82]Department of Biomedical Informatics, Arizona State University, Tempe, Arizona 85287, USA. [83]National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland 20894, USA. [84]Bioinformatics and Genomics Laboratory, Institut Jacques Monod, Paris, 75251, France. [85]Department of Molecular Biology, Cell Biology and Biochemistry, Brown University, Providence, Rhode Island 02912, USA. [86]Departamento de Genética, Centro de Ciências Biológicas, Universidade Federal de Pernambuco, Recife/PE 68011, Brazil. [87]Centro Acadêmico de Vitória, Universidade Federal de Pernambuco, Vitória de Santo Antão/PE, Brazil. [88]Cajal Institute, CSIC, Madrid 28002, Spain. [89]Department of Biology, Emory University, Atlanta, Georgia 30322, USA. [90]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA. [91]Biomedical Informatics, Stanford University, Stanford, California 94305, USA. [92]Department of Biology, San Francisco State University, San Francisco, California 94132, USA. [93]Department of Biology, University of Munich, 82152 Planegg-Martinsried, Germany. [94]Institute of Evolutionary Biology, Setagaya-ku, Tokyo 158-0098, Japan. [95]Shiba Gakuen, Minato-ku, Tokyo 105-0011, Japan. [96]European Bioinformatics Institute, Hinxton, CB10 1SD, UK. [97]Department of Biology, City University of New York at Queens, Flushing, New York 11367, USA. [98]Department of Biological Sciences and Department of Medicine, University of Alberta, Edmonton, Alberta T6G 2E9, Canada. [99]Department of Developmental Biology and Neurosciences, Tohoku University, Sendai 980-8578, Japan. [100]Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742, USA. [101]Hokkaido University, EESBIO, Sapporo, Hokkaido 060-0810, Japan. [102]Faculty of Medecine, Universite de Geneve, Geneva CH-1211, Switzerland.

*__Broad Institute Genome Sequencing Platform__ Jennifer Baldwin[10], Amr Abdouelleil[10], Jamal Abdulkadir[10], Adal Abebe[10], Brikti Abera[10], Justin Abreu[10], St Christophe Acer[10], Lynne Aftuck[10], Allen Alexander[10], Peter An[10], Erica Anderson[10], Scott Anderson[10], Harindra Arachi[10], Marc Azer[10], Pasang Bachantsang[10], Andrew Barry[10], Tashi Bayul[10], Aaron Berlin[10], Daniel Bessette[10], Toby Bloom[10], Jason Blye[10], Leonid Boguslavskiy[10], Claude Bonnet[10], Boris Boukhgalter[10], Imane Bourzgui[10], Adam Brown[10], Patrick Cahill[10], Sheridon Channer[10], Yama Cheshatsang[10], Lisa Chuda[10], Mieke Citroen[10], Alville Collymore[10], Patrick Cooke[10], Maura Costello[10], Katie D'Aco[10], Riza Daza[10], Georgius De Haan[10], Stuart DeGray[10], Christina DeMaso[10], Norbu Dhargay[10], Kimberly Dooley[10], Erin Dooley[10], Missole Doricent[10], Passang Dorje[10], Kunsang Dorjee[10], Alan Dupes[10], Richard Elong[10], Jill Falk[10], Abderrahim Farina[10], Susan Faro[10], Diallo Ferguson[10], Sheila Fisher[10], Chelsea D. Foley[10], Alicia Franke[10], Dennis Friedrich[10], Loryn Gadbois[10], Gary Gearin[10], Christina R. Gearin[10], Georgia Giannoukos[10], Tina Goode[10], Joseph Graham[10], Edward Grandbois[10], Sharleen Grewal[10], Kunsang Gyaltsen[10], Nabil Hafez[10], Birhane Hagos[10], Jennifer Hall[10], Charlotte Henson[10], Andrew Hollinger[10], Tracey Honan[10], Monika D. Huard[10], Leanne Hughes[10], Brian Hurhula[10], M Erii Husby[10], Asha Kamat[10], Ben Kanga[10],

©2007 Nature Publishing Group

Seva Kashin[10], Dmitry Khazanovich[10], Peter Kisner[10], Krista Lance[10], Marcia Lara[10], William Lee[10], Niall Lennon[10], Frances Letendre[10], Rosie LeVine[10], Alex Lipovsky[10], Xiaohong Liu[10], Jinlei Liu[10], Shangtao Liu[10], Tashi Lokyitsang[10], Yeshi Lokyitsang[10], Rakela Lubonja[10], Annie Lui[10], Pen MacDonald[10], Vasilia Magnisalis[10], Kebede Maru[10], Charles Matthews[10], William McCusker[10], Susan McDonough[10], Teena Mehta[10], James Meldrim[10], Louis Meneus[10], Oana Mihai[10], Atanas Mihalev[10], Tanya Mihova[10], Rachel Mittelman[10], Valentine Mlenga[10], Anna Montmayeur[10], Leonidas Mulrain[10], Adam Navidi[10], Jerome Naylor[10], Tamrat Negash[10], Thu Nguyen[10], Nga Nguyen[10], Robert Nicol[10], Choe Norbu[10], Nyima Norbu[10], Nathaniel Novod[10], Barry O'Neill[10], Sahal Osman[10], Eva Markiewicz[10], Otero L. Oyono[10], Christopher Patti[10], Pema Phunkhang[10], Fritz Pierre[10], Margaret Priest[10], Sujaa Raghuraman[10], Filip Rege[10], Rebecca Reyes[10], Cecil Rise[10], Peter Rogov[10], Keenan Ross[10], Elizabeth Ryan[10], Sampath Settipalli[10], Terry Shea[10], Ngawang Sherpa[10], Lu Shi[10], Diana Shih[10], Todd Sparrow[10], Jessica Spaulding[10], John Stalker[10], Nicole Stange-Thomann[10], Sharon Stavropoulos[10], Catherine Stone[10], Christopher Strader[10], Senait Tesfaye[10], Talene Thomson[10], Yama Thoulutsang[10], Dawa Thoulutsang[10], Kerri Topham[10], Ira Topping[10], Tsamla Tsamla[10], Helen Vassiliev[10], Andy Vo[10], Tsering Wangchuk[10], Tsering Wangdi[10], Michael Weiand[10], Jane Wilkinson[10], Adam Wilson[10], Shailendra Yadav[10], Geneva Young[10], Qing Yu[10], Lisa Zembek[10], Danni Zhong[10], Andrew Zimmer[10] & Zac Zwirko[10] **Broad Institute Whole Genome Assembly Team** David B. Jaffe[10], Pablo Alvarez[10], Will Brockman[10], Jonathan Butler[10], CheeWhye Chin[10], Sante Gnerre[10], Manfred Grabherr[10], Michael Kleber[10], Evan Mauceli[10] & Iain MacCallum[10]

©2007 Nature Publishing Group

ARTICLES

Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures

Alexander Stark[1,2]*, Michael F. Lin[1,2]*, Pouya Kheradpour[2]*, Jakob S. Pedersen[3,4]*, Leopold Parts[5,6], Joseph W. Carlson[7], Madeline A. Crosby[8], Matthew D. Rasmussen[2], Sushmita Roy[9], Ameya N. Deoras[2], J. Graham Ruby[10,11], Julius Brennecke[12], Harvard FlyBase curators†, Berkeley Drosophila Genome Project†, Emily Hodges[12], Angie S. Hinrichs[4], Anat Caspi[13], Benedict Paten[4,5,14], Seung-Won Park[15], Mira V. Han[16], Morgan L. Maeder[17], Benjamin J. Polansky[17], Bryanne E. Robson[17], Stein Aerts[18,19], Jacques van Helden[20], Bassem Hassan[18,19], Donald G. Gilbert[21], Deborah A. Eastman[17], Michael Rice[22], Michael Weir[23], Matthew W. Hahn[16], Yongkyu Park[15], Colin N. Dewey[24], Lior Pachter[25,26], W. James Kent[4], David Haussler[4], Eric C. Lai[27], David P. Bartel[10,11], Gregory J. Hannon[12], Thomas C. Kaufman[21], Michael B. Eisen[28,29], Andrew G. Clark[30], Douglas Smith[31], Susan E. Celniker[7], William M. Gelbart[8,32] & Manolis Kellis[1,2]

Sequencing of multiple related species followed by comparative genomics analysis constitutes a powerful approach for the systematic understanding of any genome. Here, we use the genomes of 12 *Drosophila* species for the *de novo* discovery of functional elements in the fly. Each type of functional element shows characteristic patterns of change, or 'evolutionary signatures', dictated by its precise selective constraints. Such signatures enable recognition of new protein-coding genes and exons, spurious and incorrect gene annotations, and numerous unusual gene structures, including abundant stop-codon readthrough. Similarly, we predict non-protein-coding RNA genes and structures, and new microRNA (miRNA) genes. We provide evidence of miRNA processing and functionality from both hairpin arms and both DNA strands. We identify several classes of pre- and post-transcriptional regulatory motifs, and predict individual motif instances with high confidence. We also study how discovery power scales with the divergence and number of species compared, and we provide general guidelines for comparative studies.

The sequencing of the human genome and the genomes of dozens of other metazoan species has intensified the need for systematic methods to extract biological information directly from DNA sequence. Comparative genomics has emerged as a powerful methodology for this endeavour[1,2]. Comparison of few (two–four) closely related genomes has proven successful for the discovery of protein-coding genes[3–5], RNA genes[6,7], miRNA genes[8–11] and catalogues of regulatory elements[3,4,12–14]. The resolution and discovery power of these studies should increase with the number of genomes[15–20], in principle enabling the systematic discovery of all conserved functional elements.

The fruitfly *Drosophila melanogaster* is an ideal system for developing and evaluating comparative genomics methodologies. Over the past century, *Drosophila* has been a pioneering model in which many of the basic principles governing animal development and population biology were established[21]. In the past decade, the genome sequence of *D. melanogaster* provided one of the first systematic views

[1]The Broad Institute, Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02140, USA. [2]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, Massachusetts 02139, USA. [3]The Bioinformatics Centre, Department of Molecular Biology, University of Copenhagen, Ole Maaloes Vej 5, 2200 Copenhagen N, Denmark. [4]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064, USA. [5]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [6]Institute of Computer Science, University of Tartu, Estonia. [7]BDGP, LBNL, 1 Cyclotron Road MS 64-0119, Berkeley, California 94720, USA. [8]FlyBase, The Biological Laboratories, Harvard University, 16 Divinity Avenue, Cambridge, Massachusetts 02138, USA. [9]Department of Computer Science, University of New Mexico, Albuquerque, New Mexico 87131, USA. [10]Department of Biology, MIT, Cambridge, Massachusetts 02139, USA. [11]Whitehead Institute, Cambridge, Massachusetts 02142, USA. [12]Cold Spring Harbor Laboratory, Watson School of Biological Sciences, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA. [13]University of California, San Francisco/University of California, Berkeley Joint Graduate Group in Bioengineering, Berkeley, California 97210, USA. [14]EMBL Nucleotide Sequence Submissions, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [15]Department of Cell Biology and Molecular Medicine, G-629, MSB, 185 South Orange Avenue, UMDNJ-New Jersey Medical School, Newark, New Jersey 07103, USA. [16]Department of Biology and School of Informatics, Indiana University, Indiana 47405, USA. [17]Department of Biology, Connecticut College, New London, Connecticut 06320, USA. [18]Laboratory of Neurogenetics, Department of Molecular and Developmental Genetics, VIB, 3000 Leuven, Belgium. [19]Department of Human Genetics, K. U. Leuven School of Medicine, 3000 Leuven, Belgium. [20]Department de Biologie Moleculaire, Universite Libre de Bruxelles, 1050 Brussels, Belgium. [21]Department of Biology, Indiana University, Bloomington, Indiana 47405, USA. [22]Department of Mathematics and Computer Science, Wesleyan University, Middletown, Connecticut 06459, USA. [23]Biology Department, Wesleyan University Middletown, Connecticut 06459, USA. [24]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA. [25]Department of Mathematics, University of California at Berkeley, Berkeley, California 94720, USA. [26]Department of Computer Science, University of California at Berkeley, Berkeley, California 94720, USA. [27]Department of Developmental Biology, Memorial Sloan-Kettering Cancer Center, New York, New York 10021, USA. [28]Graduate Group in Biophysics, Department of Molecular and Cell Biology, and Center for Integrative Genomics, University of California, Berkeley, California 94720, USA. [29]Lawrence Berkeley National Laboratory, Life Sciences Division, Berkeley, California 94720, USA. [30]Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853, USA. [31]Agencourt Bioscience Corporation, 500 Cummings Center, Suite 2450, Beverly, Massachusetts 01915, USA. [32]The Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA.
*These authors contributed equally to this work.
†Lists of participants and affiliations appear at the end of the paper.

©2007 Nature Publishing Group

of a metazoan genome[22], and the ongoing effort by the FlyBase and Berkeley Drosophila Genome Project (BDGP) groups established a systematic high-quality genome annotation[23–25]. Moreover, the fruitfly benefits from extensive experimental resources[26–28], which enable novel functional elements to be systematically tested and used in the evaluation of genetic screens[29,30].

The fly research community has sequenced, assembled and annotated the genomes of 12 *Drosophila* species[22,31,32] at a range of evolutionary distances from *D. melanogaster* (Fig. 1a, b). The analysis of these genomes was organized around two complementary aims. The first, described in an accompanying paper[32], was to understand the evolution of genes and chromosomes on the *Drosophila* phylogeny, and how it relates to speciation and adaptation. The second goal, described here, was to develop general comparative methodologies to discover and refine functional elements in *D. melanogaster* using the 12 genomes, and to investigate the scaling of discovery power and its implications for studies in vertebrates (Fig. 1c).

Here, we report genome-wide alignments of the 12 species (Supplementary Information 1), and the systematic discovery of euchromatic functional elements in the *D. melanogaster* genome. We predict and refine thousands of protein-coding exons, RNA genes and structures, miRNAs, pre- and post-transcriptional regulatory motifs and regulatory targets. We validate many of these elements using complementary DNA (cDNA) sequencing, human curation, small RNA sequencing, and correlation with experimentally supported transcription factor and miRNA targets. In addition, our analysis leads to several specific biological findings, listed below.

- We predict 123 novel polycistronic transcripts, 149 genes with apparent stop-codon readthrough and several candidate programmed frameshifts, with potential roles in regulation, localization and function of the corresponding protein products.
- We make available the first systematic prediction of general RNA genes and structures (non-coding RNAs (ncRNAs)) in *Drosophila*, including several structures probably involved in translational regulation and adenosine-to-inosine RNA editing (A-to-I editing).
- We present comparative and experimental evidence that some miRNA loci yield multiple functional products, from both hairpin arms or from both DNA strands, thereby increasing the versatility and complexity of miRNA-mediated regulation.
- We provide further comparative evidence for miRNA targeting in protein-coding exons.
- We report an initial network of pre- and post-transcriptional regulatory targets in *Drosophila* on the basis of individual high-confidence motif occurrences.

Comparative genomics and evolutionary signatures. Although multiple closely related genomes provide sufficient neutral divergence for recognition of functional regions in stretches of highly conserved nucleotides[16,17,33], measures of nucleotide conservation alone do not distinguish between different types of functional elements. Moreover, functional elements that tolerate abundant 'silent' mutations, such as protein-coding exons and many regulatory motifs, might not be detected when searching on the basis of strong nucleotide conservation.

Across many genomes spanning larger evolutionary distances, the information in the patterns of sequence change reveals evolutionary signatures (Fig. 2) that can be used for systematic genome annotation. Protein-coding regions show highly constrained codon substitution frequencies[34] and insertions and deletions that are heavily

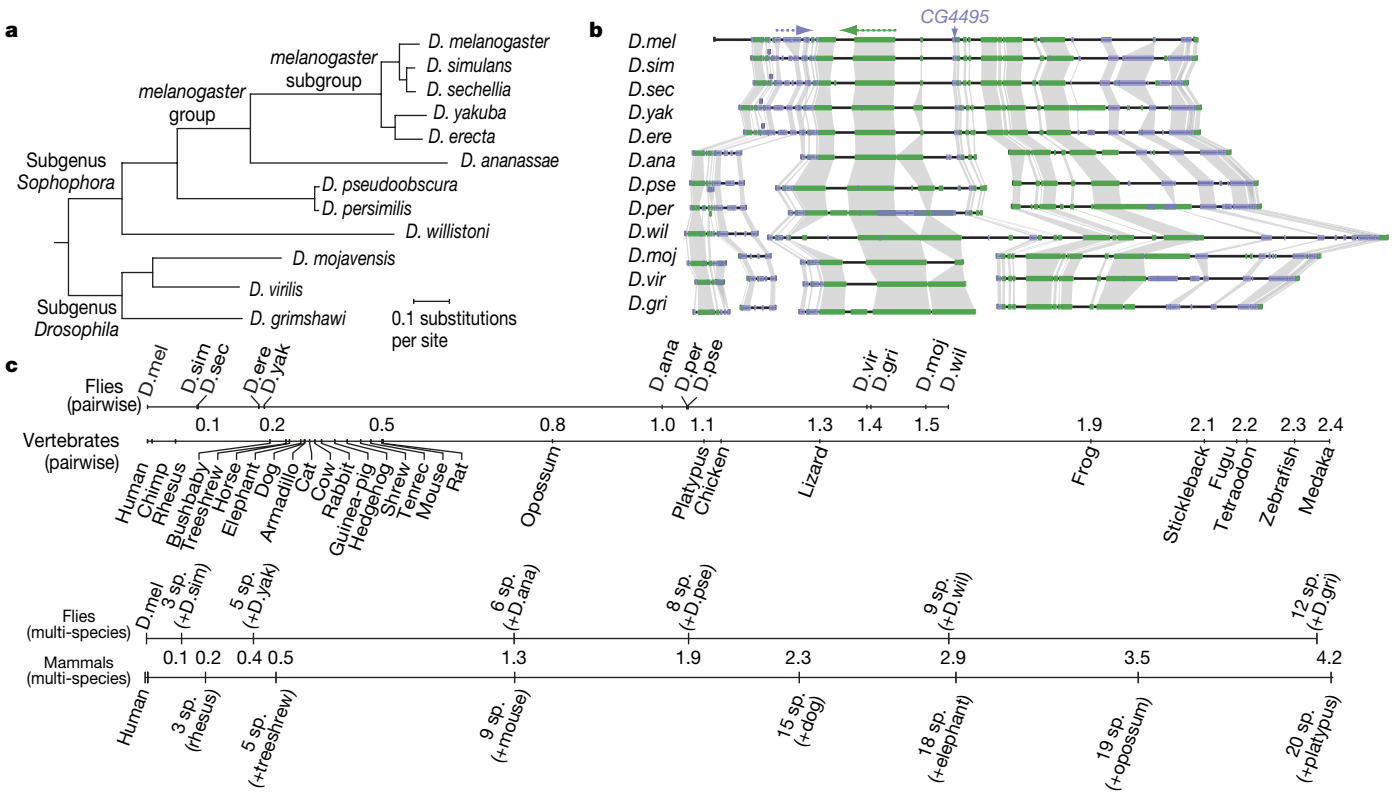


Figure 1 | Phylogeny and alignment of 12 *Drosophila* species. **a**, Phylogenetic tree relating the 12 *Drosophila* species, estimated from fourfold degenerate sites (Supplementary Methods 1). The 12 species span a total branch length of 4.13 substitutions per neutral site. **b**, Gene order conservation for a 0.45-Mb region of chromosome 2L centred on *CG4495*, for which we predict a new exon (Fig. 3a), and spanning 35 genes. Colour represents the direction of transcription. Boxes represent full gene models. Individual exons and introns are not shown. **c**, Comparison of evolutionary distances spanned by fly and vertebrate trees. Pairwise and multi-species distances (in substitutions per fourfold degenerate site) are shown from *D. melanogaster* and from human as reference genomes. Note that species with longer branches (for example, mouse) show higher pairwise distances, not always reflecting the order of divergence. Multi-species distances include all species within a phylogenetic clade.

©2007 Nature Publishing Group

biased to be multiples of three[3] (Fig. 2a). RNA genes and structures tolerate substitutions that preserve base pairing[35,36] (Fig. 2b). MicroRNA hairpins show a characteristic conservation profile with high conservation in the stem and mutations in loop regions[10,11] (Fig. 2c). Finally, regulatory motifs are marked by high levels of genome-wide conservation[3,4,12–14], and post-transcriptional motifs show strand-biased conservation[12] (Fig. 2d, e).

We find that these signatures can be much more precise for genome annotation than the overall level of nucleotide conservation (for example, Fig. 3a).

Revisiting the protein-coding gene catalogue

The annotation of protein-coding genes remains difficult in metazoan genomes owing to short exons and complex gene structures with abundant alternative splicing. Comparative information has improved computational gene predictors[5], but their accuracy still falls far short of well-studied gene catalogues such as the FlyBase annotation, which combines computational gene prediction[37], high-throughput experimental data[38–42] and extensive manual curation[23]. Recognizing this, we set out not only to produce an independent computational annotation of protein-coding genes in the fly genome, but also to assess and refine its already high-quality annotations[43].

Our analyses of *D. melanogaster* coding genes are based on two independent evolutionary signatures unique to protein-coding regions (Fig. 2a): (1) reading frame conservation (RFC)[3], which observes the tendency of nucleotide insertions and deletions to preserve the codon reading frame; and (2) codon substitution frequencies

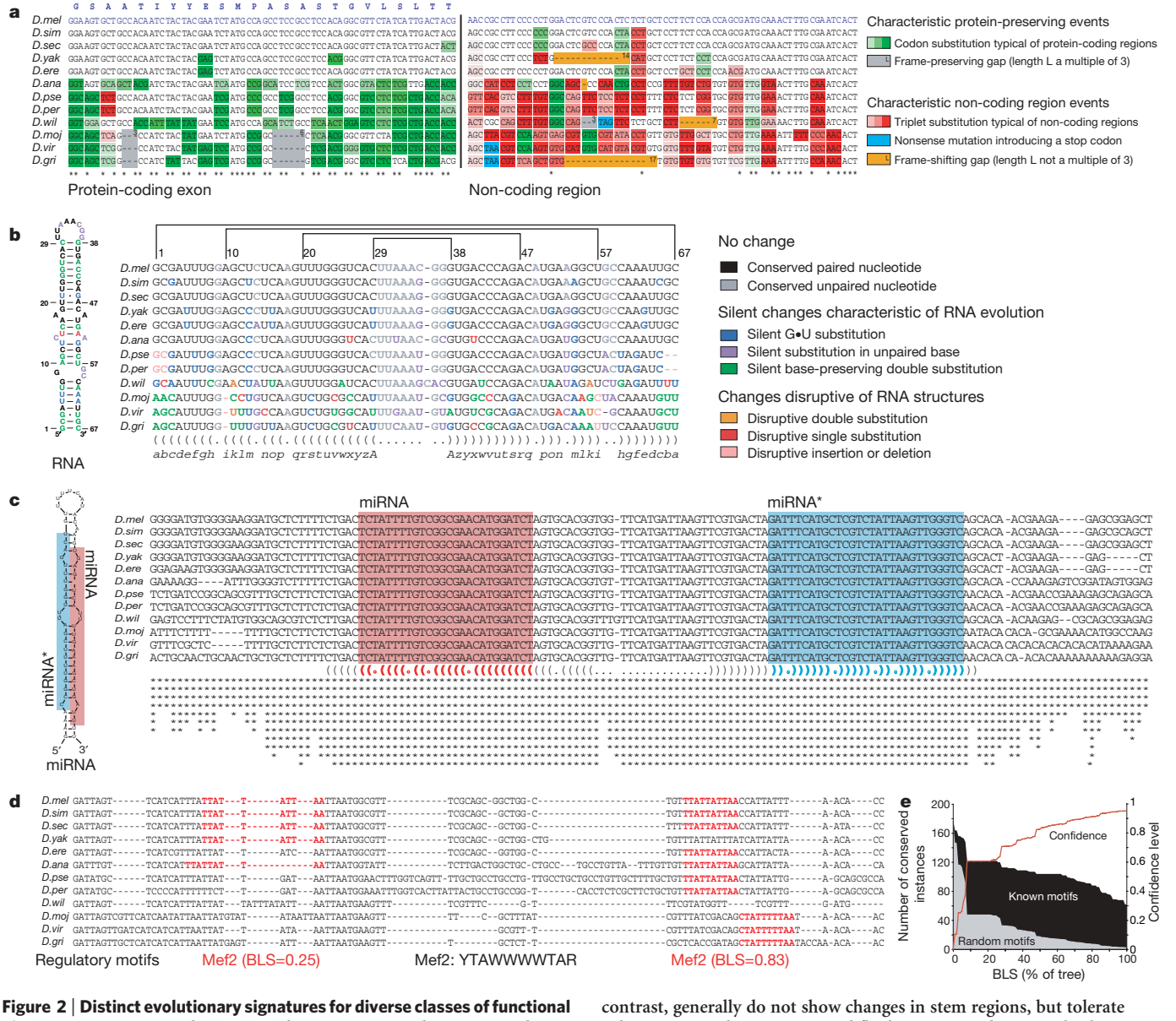


Figure 2 | Distinct evolutionary signatures for diverse classes of functional elements. **a**, Protein-coding genes tolerate mutations that preserve the amino-acid translation, leading to abundant conservative codon substitutions (green). Insertions and deletions are largely constrained to be a multiple of three (grey). In contrast, non-coding regions show abundant non-conservative triplet substitutions (red), nonsense mutations (blue) and frame-shifting insertions and deletions (orange). **b**, RNA genes tolerate mutations that preserve the secondary structure (for example, single substitutions involving G•U base pairs and compensatory changes) and exclude structure-disrupting mutations. Matching parentheses and matching letters of the alphabet indicate paired bases. **c**, MicroRNA genes, in contrast, generally do not show changes in stem regions, but tolerate substitutions in loop regions and flanking unpaired regions, leading to a distinctive conservation profile. Asterisks denote the number of informant species matching the *melanogaster* sequence at each position. **d**, Regulatory motifs tolerate local movement and nucleotide substitutions consistent with their degeneracy patterns, and show increased conservation across the phylogenetic tree, measured as the branch length score (BLS; Supplementary Methods 5a). **e**, Increasing BLS thresholds select for instances of known motifs (black) at increasing confidence (red), as the number of conserved instances of control motifs (grey) drops significantly faster.

©2007 Nature Publishing Group

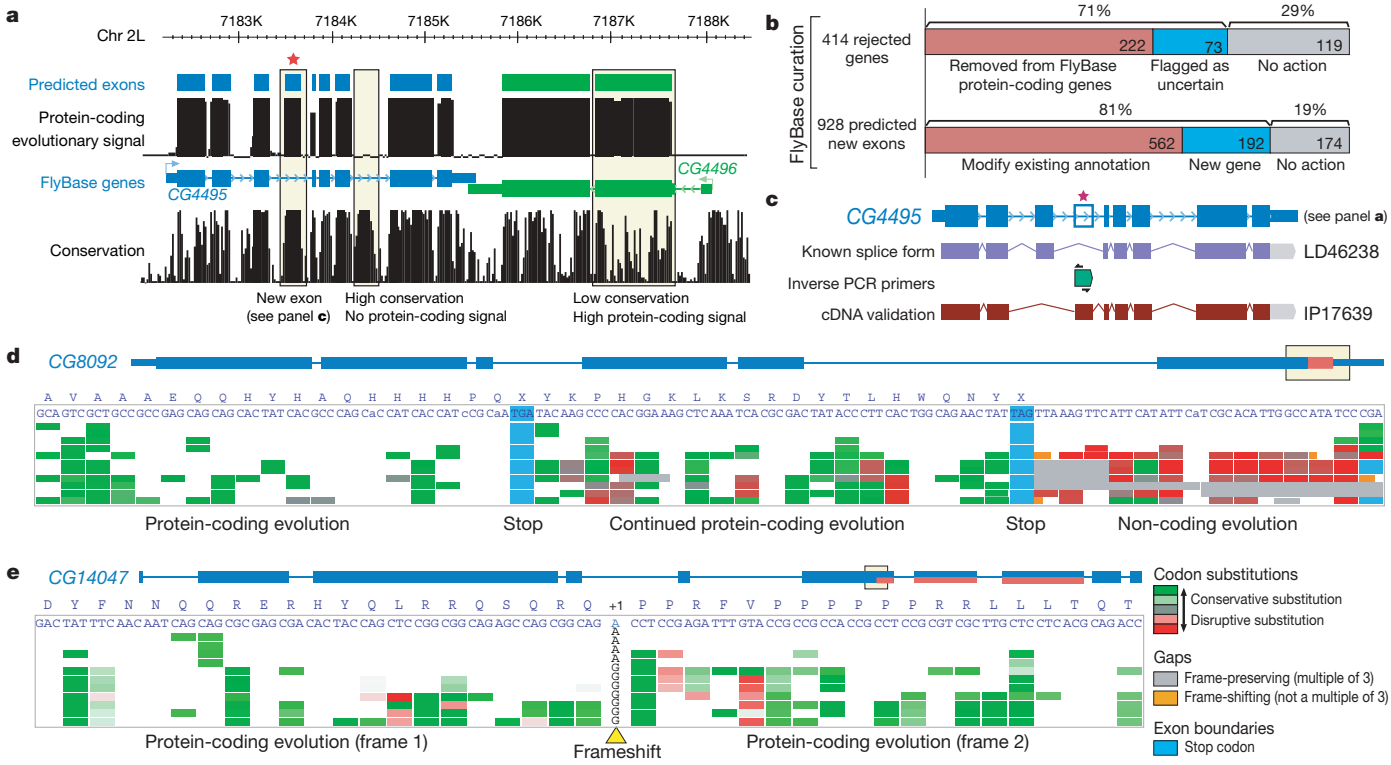


Figure 3 | Revisiting the protein-coding gene catalogue and revealing unusual gene structures. **a**, Protein-coding evolutionary signatures correlate with annotated protein-coding exons more precisely than the overall conservation level (phastCons track[33]), for example excluding highly conserved yet non-coding elements. Asterisk denotes new predicted exon, which we validate with cDNA sequencing (see panel **c**). The height of the black tracks indicates protein-coding potential according to evolutionary signatures (top) and overall sequence conservation (bottom). Blue and green boxes indicate predicted coding exons (top) and the current FlyBase annotation (bottom). The region shown represents the central 6 kb of Fig. 1b, rendered by the UCSC genome browser[126]. **b**, Results of FlyBase curation of 414 genes rejected by evolutionary signatures (Table 1), and 928 predicted new exons. **c**, Experimental validation of predicted new exon from panel **a**. Inverse PCR with primers in the predicted exon (green) results in a full-length cDNA clone, confirming the predicted exon and revealing a new alternative splice form for *CG4495*. **d**, Protein-coding evolution continues downstream of a conserved stop codon in 149 genes, suggesting translational readthrough. **e**, Codon-based evolutionary signatures (CSF score) abruptly shift from one reading frame to another within a protein-coding exon, suggesting a conserved, 'programmed' frameshift.

(CSF, see Supplementary Methods 2a), which observes mutational biases towards synonymous codon substitutions and conservative amino acid changes, similar to the non-synonymous/synonymous substitution ratio K_A/K_S[34] and other methods[44–46].

Assessing and refining existing gene annotations. We first assessed the 13,733 euchromatic genes in FlyBase[47] release 4.3. Using the above measures, we defined tests that 'confirmed' genes supported by the evolutionary evidence, 'rejected' genes inconsistent with protein-coding selection, or 'abstained' for genes that were not aligned or with ambiguous comparative evidence (Supplementary Methods 2a). Of the 4,711 genes with descriptive names, we confirmed 97%, rejected 1% and abstained for 2%, whereas the same criteria applied to 15,000 random non-coding regions ≥300 nucleotides rejected 99% of candidates and confirmed virtually none (Table 1). Together, these results illustrate the high sensitivity and specificity of our criteria.

Applying the same criteria to the 9,022 genes lacking a descriptive name (genes designated only by a CG identifier, referred to hereafter as CGid-only genes), our tests accepted 87%, rejected 5% (414 genes) and abstained for 8%. This provides strong evidence that most CGid-only genes encode proteins, but also suggests that they may be less constrained[20,32] and/or may include incorrect annotations. Indeed, on manual review, 222 (54%) of the 414 rejected CGid-only genes were re-categorized as non-protein-coding or deleted (of which 55 were due to genomically primed clones), 73 (18%) were flagged as being of uncertain quality, and the remaining 119 (29%) were kept unchanged (Fig. 3b). Some of these are probably rapidly evolving protein-coding genes, but others may also prove to be non-protein-coding genes or spurious; in fact, none of these had any functional gene ontology (GO) annotation[48].

In addition, we proposed specific corrections and adjustments to hundreds of existing transcript models, including translation start site adjustments (Supplementary Fig. 2b), alternative splice boundaries (Supplementary Fig. 2b), recent nonsense mutations (Supplementary Fig. 2c) and alternative translational reading frames[43].

Identifying new genes and exons. To predict new protein-coding exons, we integrated our metrics into a probabilistic algorithm that determines an optimal segmentation of the genome into protein-coding and non-coding regions (Fig. 3a) on the basis of whole-genome sequence alignments of the 12 fly species (Supplementary

Table 1 | Assessment of FlyBase euchromatic protein-coding gene annotations

Regions evaluated	Total	Confirm	Abstain	Reject*
Named genes	4,711	4,566 (96.9%)	105 (2.2%)	40 (0.8%)
CGid-only genes	9,022	7,879 (87.3%)	729 (8.1%)	414 (4.6%)
Non-coding regions†	15,564	3 (0.0%)	131 (0.8%)	15,430 (99.1%)

* A minority of rejected genes are false rejections; see Fig. 3b and text for details.
† Regions ≥300 nucleotides in length randomly chosen from the non-coding part of the genome (see Supplementary Methods 2a).

©2007 Nature Publishing Group

Methods 2a). Our genome-wide search predicted 1,193 new protein-coding exons, mostly in euchromatic regions annotated as intergenic (43%), intronic (26%), or 5′/3′ untranslated region (UTR; 23%) in FlyBase annotation release 4.3.

We manually reviewed 928 of these predictions according to FlyBase standards[23] (Supplementary Methods 2a), leading to 142 new gene models (incorporating 192 predictions) and 438 revised gene models (incorporating 562 predictions) (Fig. 3b). In parallel, we tested 184 predictions (126 intergenic, 58 intronic) by directed cDNA sequencing using inverse polymerase chain reaction (inverse PCR) of circularized full-length clones[49–51] (Fig. 3c), which validated 120 targeted predictions (65%) and an additional 42 predictions not directly targeted but contained within the recovered transcripts. Predictions in intergenic regions yielded 88 full-length cDNAs, providing evidence for 50 new genes and modification of 39 gene models. Predictions within introns of existing annotations yielded 32 full-length cDNAs, of which only 18 (56%) represent new splice variants of the surrounding gene, whereas the remaining 14 revealed nested or interleaved gene structures. This provides additional evidence that such complex gene structures are not rare in *Drosophila*[23].

Overall, 83% of the 948 predicted exons that we assessed by manual curation or cDNA sequencing were incorporated into FlyBase, resulting in 150 new genes and modifications to hundreds of existing gene models. Finally, the 245 predictions that we did not assess were in non-coding regions of existing transcript models, or were already included in FlyBase independent of our study. In an independent analysis[52], we predicted 98 new genes on the basis of inferred homology to predicted genes in the informant species[32], of which 63% matched the above predictions.

Discovering unusual features of protein-coding genes. Our analysis also predicted an abundance of unusual protein-coding genes that call for follow-up experimental investigation. First, we found open reading frames with clear protein-coding signatures and conserved start and stop sites on the transcribed strand of annotated UTRs, indicative of polycistronic transcripts[23,53,54]. These include 73% of 115 annotated dicistronic transcripts and 135 new candidate cistrons of 123 genes (Supplementary Fig. 2b).

Second, we predicted that 149 genes undergo stop codon readthrough, with protein-coding selection continuing past a deeply conserved stop codon (Fig. 3d), in some cases for hundreds of amino acids. It is unlikely that these genes are selenoproteins, as they appear to lack SECIS elements that direct selenocysteine recoding[55–58]. Other mechanisms may instead be at work, such as regulation of ribosomal release factors[59], A-to-I editing[39,60,61], alternative splicing, or other less-characterized mechanisms[62]. In fact, these genes are significantly enriched in neuronal proteins ($P = 10^{-4}$), which frequently undergo A-to-I editing[63].

Third, we found four genes in which CSF signatures abruptly shift from one reading frame to another in the absence of nearby intron–exon boundaries or insertions and deletions (Fig. 3e). These are suggestive of conserved 'programmed' frameshifts[64], which are thought to be rare in eukaryotes.

Overall, our results affected over 10% of protein-coding genes, and will be available in future releases of FlyBase. They also suggest that several types of unusual protein-coding gene structure may be more prevalent in the fly than previously appreciated.

RNA genes and structures

Several comparative approaches to RNA gene identification have been developed[6,7,65] that recognize their characteristic properties: compensatory double substitutions of paired nucleotides (for example, A•U↔C•G), structure-preserving single-nucleotide mutations involving G•U base pairs (G•U↔G•C and G•U↔A•U), and few nucleotide substitutions disrupting functional base pairs (Fig. 2b). To predict new structures, we applied EvoFold[7] in highly conserved segments of the 12 *Drosophila* species and focused on high-stringency candidates with strong support by compensatory changes (Supplementary Methods 4).

Our search led to 394 predictions, recovering 68 known RNA structures (primarily transfer RNA genes) in 0.02% of the genome (570-fold enrichment). The novel candidates consisted of 177 structures in intergenic regions (54%), 103 in introns (32%), 36 in 3′ UTRs (11%) and 10 in 5′ UTRs (3%). In addition, we predicted 200 structures in protein-coding regions (Supplementary Methods 3). Notably, 75% of 3′ UTR structures and 80% of 5′ UTR structures were predicted on the transcribed strand, suggesting that they are frequently part of the messenger RNA. In contrast, only 47% of intronic structures are on the transcribed strand, suggesting that they are largely independent of the surrounding genes.

Known and novel types of RNA genes. Of the 177 predicted intergenic structures, 30 were detected in a tiling-array expression study[42]. This fraction (17%) is significantly above that for all conserved intergenic regions (12%, $P = 0.007$), but lower than that of known

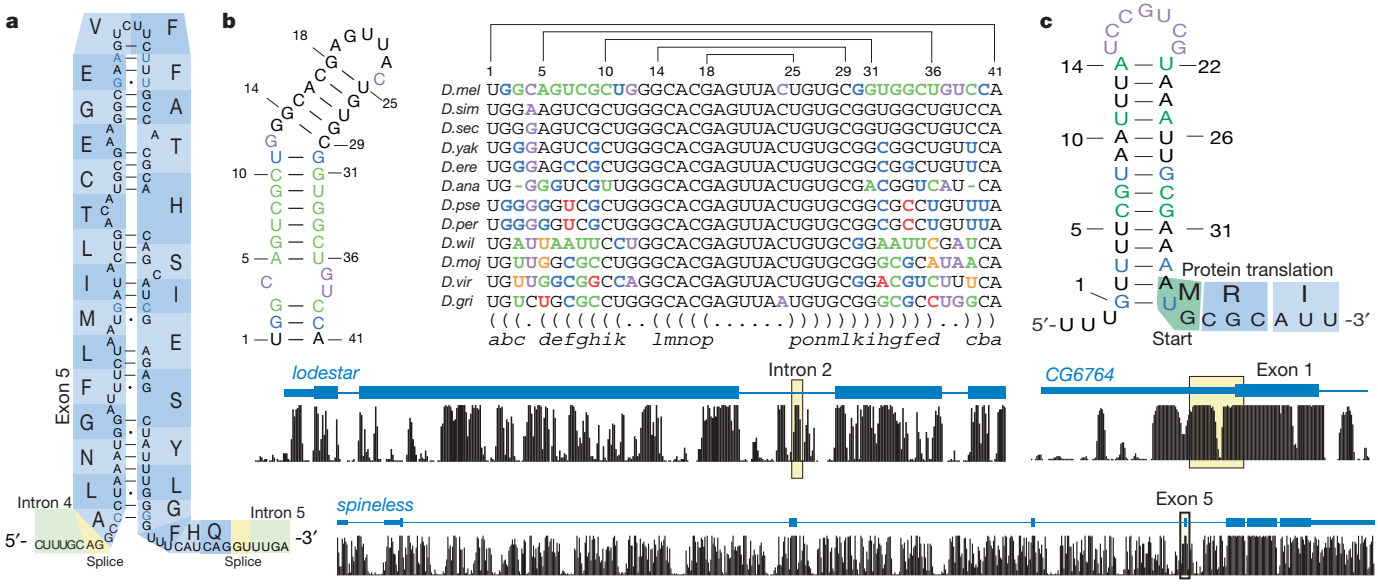


Figure 4 | Novel RNA structures. a, New exonic RNA structure spanning 78 of 90 nucleotides of *spineless* exon 5. **b**, New intronic RNA structure in *lodestar* shows 11 compensatory substitutions and 10 silent G•U substitutions, providing strong evidence of structural selection (colours as in Fig. 2b). **c**, New 5′ UTR structure that overlaps the translation start site of *CG6764*, the fly orthologue of yeast ribosomal protein *RPL24*, suggesting a potential role in translational regulation. **a–c**, Structure shown corresponds to shaded region in the gene model.

©2007 Nature Publishing Group

intergenic ncRNAs (21%), suggesting that these candidates may be of lower abundance, temporally or spatially constrained, or might include false positives. Two predictions were expressed throughout development, one extending the annotation of a previously reported but uncharacterized ncRNA[66] and the other probably representing a novel type of ncRNA. The predictions also included nine novel H/ACA-box small nucleolar RNA candidates in introns of ribosomal genes, known to frequently contain small nucleolar RNAs that guide post-transcriptional base modifications of ncRNAs[67].

Likely A-to-I editing structures. Many of the 48 intronic candidates on the transcribed strand and many of the 200 hairpins in coding sequence are probably involved in A-to-I editing or post-transcriptional regulation (Fig. 4a). Hairpins in coding sequence were associated with 11 of the 157 known editing sites (120-fold enrichment) and both intronic and coding-sequence hairpins showed a strong enrichment for ion-channel genes (6%, $P = 0.007$ and 10%, $P = 2 \times 10^{-12}$, respectively), known to be frequent editing targets. Editing is known to occur at multiple sites in the same gene[63], and we find an additional 10 hairpins in known editing targets, as well as 40 additional hairpins clustered in 18 genes not previously known to be edited (for example *huntingtin*[68], which harbours four predicted hairpins, more than any other gene). Intronic predictions also showed the highest abundance of compensatory substitutions: for example, *Resistant to dieldrin* (Fig. 2b) contained a 26-base-pair (bp) intronic hairpin flanked by exons known to be edited[69] with a striking 16 compensatory changes, *lodestar* showed one hairpin with 11 compensatory changes, and *Inverted repeat-binding protein* showed one hairpin with 10 compensatory substitutions (Fig. 4b).

Likely regulatory UTR structures. We predicted 38 structures in 3′ UTRs, a density twofold higher than the genomic average, whereas fewer than 10 such examples are currently known[70]. A considerable fraction of these lies in regulatory genes (14 out of 38; $P = 10^{-4}$), including several transcriptional regulators (for example, *cas*, *spen* and *Alh*), the tyrosine phosphatase *PTP-ER* and the translation initiation factor *eIF3-S8*. This suggests that many regulatory genes may themselves be regulated post-transcriptionally through these structures.

3′ UTR structures were also enriched for genes involved in mRNA localization (3 out of 38, $P = 2.7 \times 10^{-4}$), including *oo18 RNA-binding protein* (*orb*) and *staufen* (*stau*), both of which contain double-stranded RNA-binding domains, are involved in axis specification during oogenesis, and interact with the mRNA of maternal effect protein *oskar*. The hairpin in *orb* is known to be important for mRNA transport and localization[71], whereas the highly similar *stau* hairpin has not been previously described to our knowledge.

The ten structures found in 5′ UTRs probably contain binding sites for factors that regulate translation. For example, the fly homologue of yeast ribosomal protein *RPL24* contains a hairpin structure overlapping its start codon (Fig. 4c). This is interesting in light of high conservation upstream of the start codon in yeast ribosomal proteins[3,4], and findings that ribosomal proteins bind to their mRNAs and control translation in prokaryotes[72,73].

Conserved RNA structures in *roX2* recruit MSL. In an independent study[74], we searched for conserved regions in the non-coding *roX1* and *roX2* (*RNA on the X*) genes to gain insights into their function. Both RNAs are components of the MSL (Male-specific lethal) complex and are crucial for dosage compensation in male flies, inducing lysine 16 acetylation of histone H4, leading to upregulation of hundreds of genes on the X chromosome[75]. We identified several stem-loop structures with repeated sequence motifs (for example, GUUNUACG), and found that tandem repeats of one of these were sufficient to recruit MSL complexes to the X chromosome and to induce acetylation of lysine 16 of histone H4. Although this structure could not fully rescue roX-deficient males, our results suggest that it mediates MSL recruitment during *roX2*-dependent chromatin modification and dosage compensation, illustrating the power of evolutionary evidence for directing experimental studies.

Prediction and characterization of miRNA genes

Focusing on specific classes of RNA genes markedly increases the accuracy of RNA gene prediction, reviewed in refs 35, 76 and illustrated here for *Drosophila* miRNA genes. The common biogenesis and function of miRNAs[77] lead to evolutionary and structural signatures (Fig. 2c) that can be used for their systematic *de novo* discovery[8–11]. Using such signatures in the 12 fly genomes (Supplementary Methods 4a, b), we predicted 101 miRNAs[78] (Supplementary Table 4d), which include 60 of the 74 verified Rfam miRNAs (81%), while spanning less than 0.006% of the fly genome (13,500-fold nucleotide enrichment).

Comparison of our predictions with high-throughput sequencing data of short RNA libraries from different stages and tissues of *D. melanogaster*[78,79] revealed that 84 of the 101 predictions (83%), including 24 of the 41 novel predictions (59%), were authentic miRNA genes (Fig. 5a and Supplementary Table 4d). An independent computational method[79] had 20 of its 45 novel predictions validated when used across six *Drosophila* species. Additional candidates may represent genuine miRNAs whose temporal or spatial expression pattern does not overlap with the surveyed libraries.

Several of the validated miRNAs were on the transcribed strand of introns or clustered with other miRNAs. For example, *mir-11* and *mir-998* (the vertebrate homologue of which, *mir-29*, has been implicated in cancer[80]) were both found in the last intron of *E2f*, and might be involved in cell-cycle regulation (Fig. 5b). Notably, two predictions overlapped exons of previously annotated protein-coding genes that were independently rejected above (Fig. 5c), providing an explanation for the previously observed transcripts of these annotations and highlighting the importance of specific signatures for genome annotation.

High-throughput sequencing data discovered an additional 50 miRNAs not found computationally[79,81], thereby illustrating the limitations of purely computational approaches. Some of these had precursor structures not seen previously for animal miRNAs, including unusually long hairpins[79] and hairpins corresponding to short introns (mirtrons)[81,82]. The remaining were often less broadly conserved or showed unusual conservation properties.

Signatures for mature miRNA annotation. The exact position of 5′ cleavage of mature miRNAs is important, because it dictates the core of the target recognition sequence[83–85]. This leads to unique structural and evolutionary signatures, including direct signals, present at the 5′ cleavage site, and indirect signals, stemming from the relationship of miRNAs with their target genes (Supplementary Methods 4a, c). Combined into a computational framework[78], these signatures predicted the exact start position in 47 of the 60 cloned Rfam miRNAs (78%), and were within 1 bp in 51 cases (85%). The method disagreed with the previous annotation in 9 of the 14 Rfam miRNAs that were not previously cloned, of which 6 were confirmed by sequencing reads[78,79], leading to marked changes in the inferred target spectrum (Fig. 5d). Prediction accuracy was significantly lower (41% exact, 61% within 1 nucleotide) for novel miRNAs, which, however, also showed less accurate processing *in vivo*[78,79].

New insights into miRNA function and biogenesis. We predicted targets for all conserved miRNAs identified by high-throughput sequencing[79] searching for conserved matches to the seed region (similar to ref. 86) evaluated using the branch length score (Supplementary Methods 5a), a new scoring scheme described below. Whereas the resulting miRNA targeting network changed substantially[79], we found that the novel and revised miRNAs shared many of their predicted targets with previously known miRNAs, resulting in a denser network with increased potential for combinatorial regulation[78,79].

For ten miRNA hairpins, the mature miRNA and the corresponding miRNA star sequence (miRNA*, the small RNA from the opposite arm of the hairpin) both appeared to be functional: both reached high computational scores and were frequently sequenced[78,79], often exceeding the abundance of many mature miRNAs (Supplementary Table 4e). The Hox miRNA *mir-10* showed a particularly striking example of a functional star sequence (Fig. 5e): both arms showed

©2007 Nature Publishing Group

abundant reads, high scores and highly conserved Hox gene targets[78,79], suggesting a key role in Hox regulation.

In addition, for 20 miRNA loci, the anti-sense strand also folded into a high-scoring hairpin suggestive of a functional miRNA[78] (Supplementary Table 4f). Indeed, sequencing reads confirmed that four of these anti-sense hairpins are processed into small RNAs *in vivo*[79]. Thus, a single genomic miRNA locus may produce up to four miRNAs, each with distinct targets.

Regulatory motif discovery and characterization

Regulatory motifs recognized by proteins and RNAs to control gene expression have been difficult to identify due to their short length, their many weakly specified positions, and the varying distances at which they can act[87,88]. Recent studies have shown that comparative genomics of a small number of species can be used for motif discovery[3,4,12–14], on the basis of hundreds of conserved instances across the genome (Fig. 2d). Many related genomes should lead to increased discovery power, but also pose new challenges, arising from sequencing, assembly, or alignment artefacts, and from movement or loss of motif instances in individual species.

To account for the unique properties of regulatory motifs, we developed a phylogenetic framework to assess the conservation of each motif instance across many genomes[89]. Briefly, we searched for motif instances in each of the aligned genomes, and based on the set

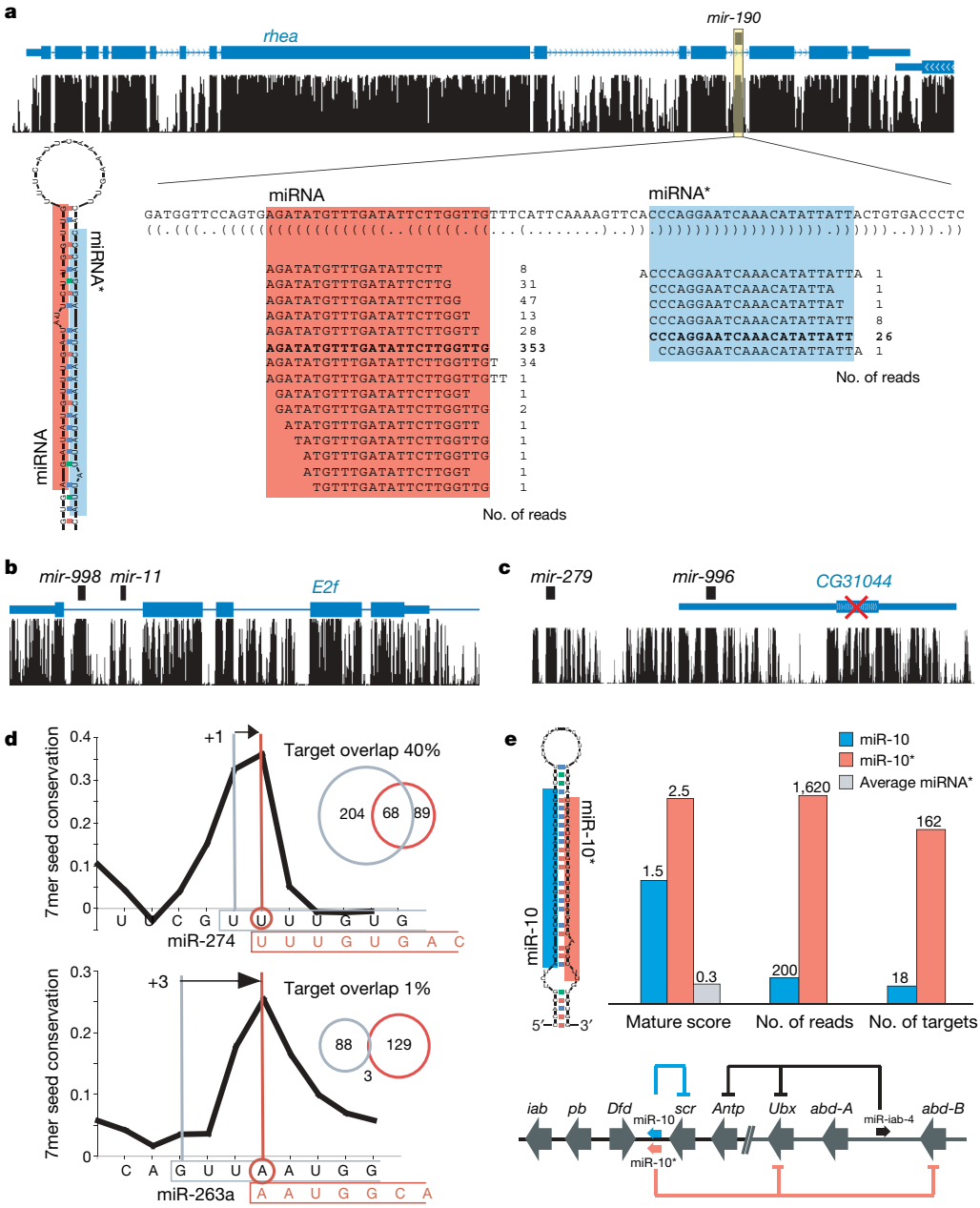


Figure 5 | MicroRNA gene identification and functional implications. **a**, New predicted miRNA (*mir-190*) and its validation by sequencing reads. Total read counts for mature miRNA (red) and miRNA* (blue) show a characteristic pattern of processing indicative of miRNAs. Highlighted regions indicate most abundant processing products. **b**, Example of clustered known (*mir-11*) and new (*mir-998*) miRNAs in the intron of cell-cycle regulator *E2f*. **c**, Example of a new miRNA (*mir-996*) in the transcript of a spurious gene. *CG31044* was rejected by our protein-coding analysis, its transcript probably representing the precursor of *mir-996*, with no protein-coding function. **d**, Revisions to the 5′ end of miR-274 and miR-263a are proposed on the basis of evolutionary evidence (for example, 7mer seed conservation; black curve) and confirmed by sequencing reads. Changes at the 5′ end of more than one nucleotide results in marked changes to the predicted target spectra (venn diagrams). **e**, Evidence from evolutionary signals (mature score), sequencing reads and target predictions suggests that both miR-10 and miR-10* are functional, each targeting distinct Hox genes.

©2007 Nature Publishing Group

of species that contained them, we evaluated the total branch length over which the *D. melanogaster* motif instance appears to be conserved (Supplementary Methods 5a, b), which we call the branch length score (BLS). We used BLS for the discovery of novel motifs (this section) and for the prediction of individual functional motif instances (next section).

Predicted motifs recover known regulators. To discover motifs, we estimated the conservation level of candidate sequence patterns with a motif excess conservation (MEC) score compared to overall conservation levels in promoters, UTRs, introns, protein-coding exons and intergenic regions (Supplementary Methods 5a).

Our search in regions with roles in pre-transcriptional regulation resulted in 145 distinct motifs (Table 2), obtained by collapsing variants across 83 motifs discovered in promoters, 35 in enhancers, 20 in 5′ UTRs, 35 in core promoters, 30 in introns and 84 in the remaining intergenic regions. Motifs discovered in each region showed similar properties and large overlap: 66 (46%) were discovered independently in at least two regions and 40 (28%) in at least three, consistent with shared regulatory elements in these regions[90].

The 145 discovered motifs match 40 (46%) of the 87 known transcription factors in *Drosophila* (Supplementary Table 5c) compared to 8% expected at random ($P = 1 \times 10^{-20}$). Several of the non-discovered known motifs are involved in early anterior–posterior segmentation of the embryo, consistent with reports that they are largely non-conserved[91]; indeed, 74% of these did not exceed the conservation expected by chance in promoter regions. Other non-discovered motifs often lacked characteristics expected for transcription factor motifs, suggesting that some may be spurious: 49% were unusually long (>10 nucleotides) compared to 23% of recovered ones, and showed only one or a few total instances genome-wide, suggestive of individual regulatory sites rather than motifs.

Tissue-specific and functional enrichment of novel motifs. The discovered motifs showed strong signals with respect to embryonic expression patterns (Fig. 6a). Overall, 75 (52%) were either enriched or depleted in genes expressed in at least one tissue, compared to 59% of known motifs and 3% of random controls. Motif depletion may represent either specific repressors for individual tissues, or activators excluded from these tissues. Motif depletion was found more generally in ubiquitously expressed genes (30% of discovered and 34% of known motifs compared with 1% expected at random), similar to findings for *in vivo* binding sites[92], and probably reflecting less complex regulation. We also found significant motif enrichment in groups of genetically interacting genes (collected by FlyBase) that often function in common developmental contexts or signalling pathways, genes of metabolic pathways (Kyoto Encyclopedia of Genes and Genomes, KEGG[93]), and genes with shared functions (GO).

In total, 68% of discovered and 70% of known motifs were enriched or depleted in one of the functional categories (14% random). Noteworthy examples include motif ME93 (GCAACA), which was more highly enriched in neuroblasts ($P = 4 \times 10^{-12}$) than either of the two well-known regulators of neuroblast development, *prospero* and *asense* ($P = 4 \times 10^{-5}$ and 2×10^{-7}, respectively). Similarly,

Table 2 | Pre-transcriptional motifs

Name	Motif consensus	MEC	MCS	Region*	Known transcription factor†	Multiplicity score‡	ImaGO enrichment§	ImaGO score§
ME1	GTCACGTD	0.448	45.41	PIG	–	–	–	–
ME2	AWNTGGGTCA	0.393	26.97	PIG	Hr46	–	Oesophagus (13–16)	4.52
ME3	BCATAAATYA	0.369	36.02	PCEIG	Caudal	–	Ubiquitous (13–16)	−6.22
ME4	HAATTAYGCRH	0.365	32.71	PCE5IG	Engrailed	–	–	–
ME5	STATAWAWR	0.358	24.31	C	TATA	–	Ventral nerve cord (13–16)	−5.1
ME6	VATTWGCAT	0.356	44.06	PE5IG	–	3.73	Ubiquitous (11–12)	−7.15
ME7	BYAATTARH	0.338	15.45	PCE5IG	Engrailed	7.08	Ubiquitous (11–12)	−10.26
ME8	HRTCAATCA	0.338	42.32	PIG	–	–	Dorsal pharyngeal muscle PR (11–12)	−4.15
ME9	TGACANNNNNNTGACA	0.336	9	G	–	–	–	–
ME10	RCGTGNNNNGCAT	0.329	15.94	PIG	–	–	–	–
ME11	MATTAAWNATGCR	0.324	12.43	PIG	acj6	–	Tracheal PR (11–12)	4.11
ME12	TTAATGATG	0.32	20.31	PG	–	–	–	–
ME13	WTGACANBT	0.318	63.45	PE5IG	–	4.14	Ubiquitous (13–16)	−3.97
ME14	YGACMTTGA	0.313	27.06	PIG	–	–	Midgut (13–16)	4.32
ME15	AATTRNNNNCAATT	0.309	21.17	PG	–	–	–	–
ME16	TGACGTCAT	0.304	12.24	PC5IG	CrebA	–	–	–
ME17	MAATTNAATT	0.304	51.57	PE5IG	–	–	Ubiquitous (11–12)	−6.66
ME18	MRYTTCCGYY	0.304	39.04	PEIG	Dorsal	–	Ubiquitous (11–12)	−4.4
ME19	MATTRRCACNY	0.303	25.24	PIG	–	–	–	–
ME20	YTAATGAVS	0.298	44.5	PEIG	–	–	Foregut PR (11–12)	4.19
ME21	TAATTRANNTTNATG	0.294	8.67	G	–	–	–	–
ME22	WAATGCGCNT	0.291	18.17	G	–	–	–	–
ME23	MATTWRTCA	0.288	46.25	PEIG	–	–	Dorsal epidermis PR (11–12)	4.4
ME24	YAATTWNRYGC	0.287	30.91	PG	–	4.27	Ubiquitous (11–12)	−4.79
ME25	TTAYGTAA	0.283	13.06	5	Giant	–	Midgut (13–16)	5.32
ME26	YGCGTHAATTR	0.283	13.61	PEG	–	–	–	–
ME27	AATTRYGWCA	0.28	22.85	PEIG	–	–	Pericardial cell (13–16)	4.1
ME28	GCGCATGH	0.28	30.17	PCEG	–	–	Ventral nerve cord PR (11–12)	5.75
ME29	WAATCARCGC	0.275	13.82	G	–	–	–	–
ME30	AATTAANNNNNCATNA	0.271	16.44	G	Antennapedia	–	–	–
ME31	GCGTSAAA	0.271	29.95	PG	–	–	–	–
ME32	YGCGYRTCAWT	0.269	12.87	G	–	–	–	–
ME33	GCGTTGAYA	0.269	15.1	PG	–	–	–	–
ME34	AAATKKCATTA	0.266	14.04	PG	–	–	–	–
ME35	RACASCTGY	0.266	28.38	PCEG	Scute	–	Ventral sensory complex SA (11–12)	4.08
ME36	TGTCAATTG	0.265	12.65	PG	–	–	Tracheal system (13–16)	4.56
ME37	WAATKNNNNNCRCGY	0.261	23.34	PEG	–	–	–	–
ME38	CASGTAR	0.261	9.24	PEG	Single-minded	4.58	Ventral epidermis PR (11–12)	7.41
ME39	WCACGTGC	0.26	10.54	PCE5IG	Enhancer of split	–	–	–
ME40	CATTANNNWAATT	0.259	19.02	G	–	–	–	–

The top 40 of 145 are shown. MEC, motif excess conservation; MCS, motif conservation score. See Supplementary Table 5c for the full table.
* Region where the motif was found: P, promoter, C, core promoter; E, enhancers; 5, 5′ UTR; I, intron; G, intergenic genome.
† The known transcription factor motif matching the consensus sequence.
‡ A multiplicity score is reported for motifs with many repeated occurrences.
§ Tissue where motif is most strongly enriched or depleted, and corresponding score (positive, enrichment; negative, depletion). PR, primordium; SA, specific anlage.

©2007 Nature Publishing Group

motifs ME89 (CACRCAC), ME11 (MATTAAWNATGCR) and ME117 (MAAMNNCAA) were highly enriched in malpighian tubule ($P = 4 \times 10^{-7}$), trachea ($P = 4 \times 10^{-5}$) and surface glia (6×10^{-7}), respectively, in each case ranking above motifs for factors known to be important in these tissues (Supplementary Table 5c). These presumably correspond to as-yet-unknown regulators for these tissues.

Exclusion, clustering and positional constraints. A large number of motifs were depleted in coding sequence (57% of discovered versus 57% of known and 10% of random motifs, $P = 3 \times 10^{-18}$) and in 3′ UTRs (30% versus 22% and 0%, $P = 4 \times 10^{-11}$), suggesting specific exclusion similar to *in vivo* binding[92].

Many of the intergenic or intronic instances occurred in clusters, a property of motifs that has been used to identify enhancer elements[91,94–96]. We assessed increased conservation of motifs when found near other instances of the same motif (whether conserved or not, to correct for regional conservation biases), and found significant multiplicity for 19% of the discovered motifs (compared to 24% of known and 4% of random motifs).

In addition, 15 of the discovered motifs (10%) were significantly enriched near transcription start sites (compared to 14% of known and 1% of random motifs). Several were enriched at precise positions and preferred orientations (Fig. 6b), including close matches to several known core promoter motifs involved in transcription initiation[97]. For example, ME5 (STATAWAWR), which matches the TATA-box motif, displayed a sharp peak on the transcribed strand, 27 nucleotides upstream of the transcription start site. Similarly, ME120 (TCAGTT), corresponding to the known initiator motif (Inr) strongly peaked directly on the transcription start site, and ME54 (RCGYRCGY), which matches a known downstream promoter element (DPE), peaked 30 nucleotides downstream of the transcription start site.

Regulatory motifs involved in post-transcriptional regulation. We also used BLS/MEC to discover motifs involved in post-transcriptional regulation, and developed methods to distinguish motifs acting at the DNA level, motifs acting at the RNA level and motifs stemming from protein-coding codon biases (Supplementary Methods 5a). Motifs acting post-transcriptionally at the RNA level generally showed highly asymmetric conservation[12], as functional instances can only occur on the transcribed strand. Indeed, 71 of 90 motifs (79%) discovered in 3′ UTRs showed strand-specific conservation (compared with only 3% of 5′ UTR motifs and 5% of intron motifs, suggesting that these act primarily in pre-transcriptional regulation).

Overall, 33 motifs discovered in 3′ UTRs were complementary to the 5′ end of Rfam miRNAs, recovering 72% of known miRNAs (68% of 5′ unique miRNA families). An additional 21 motifs matched to 5′ ends of novel miRNAs predicted above, of which 12

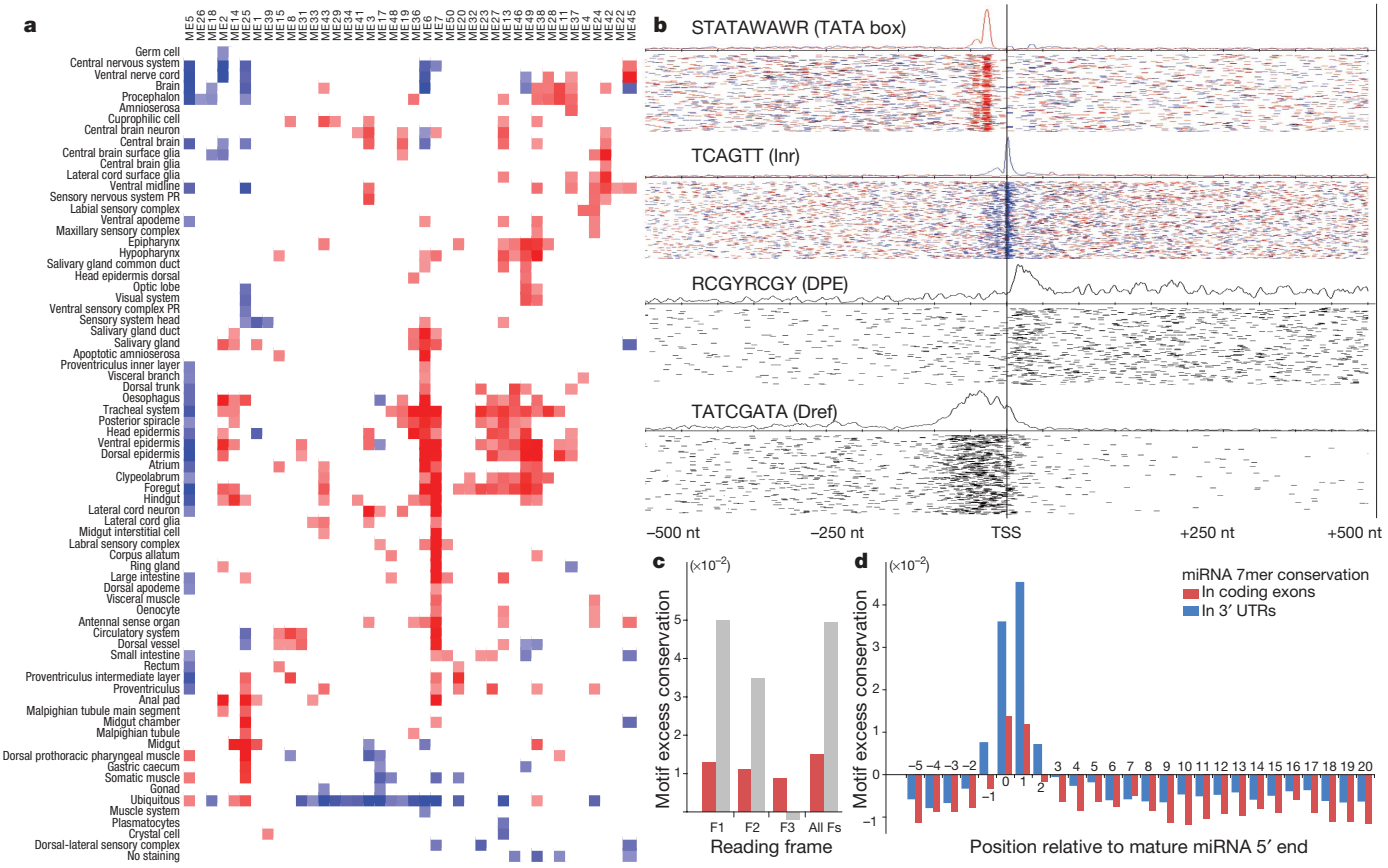


Figure 6 | Regulatory motif discovery. **a**, Discovered motifs show enrichment (red) or depletion (blue) in genes expressed in a given tissue (log colour range from $P = 10^{-5}$ enrichment to $P = 10^{-5}$ depletion). Bi-clustering reveals groups of motifs with similar tissue enrichment and groups of tissues with similar motif content. Full matrix and randomized control is shown in Supplementary Fig. 6d. **b**, Positional bias of discovered motifs relative to transcription start sites (TSS). Peaks with highly specific distances from the transcription start site (for example, first three plots) are characteristic of core promoter elements, and broad peaks (for example, fourth plot) are characteristic of transcription factors. For non-palindromic motifs, colours indicate forward-strand (red) and reverse-strand (blue) instances. Curves denote the density of all instances and individual segments denote individual motif instances, summed across groups of 50 genes (each line). **c**, Coding regions show reading-frame-invariant conservation for miRNA motifs (red) and reading-frame-biased conservation for protein motifs (grey). MEC scores are evaluated for each of the three reading frame offsets (F1–F3) and also without frame correction (all Fs). Plots show average MEC for all miRNA motifs and 500 top-scoring protein-coding motifs (based on MEC without frame correction). **d**, Motif excess conservation (MEC) of 7mer complements at different offsets with respect to miRNA 5′ end, averaged across all Rfam miRNAs. MEC scores evaluated in protein-coding regions and 3′ UTRs show a highly similar profile (correlation coefficient 0.96), suggesting similar evolutionary constraints.

©2007 Nature Publishing Group

were validated experimentally[78,79], and 3 motifs matched uniquely to miRNA star sequences, all of which were abundantly expressed *in vivo* (Supplementary Table 4e).

We found 33 additional motifs in 3′ UTRs that were apparently not associated with miRNAs. MO40 (TGTANWTW) closely matches the Puf-family Pumilio motif[98]. MO32 (AATAAA) corresponds to the polyadenylation signal and displays both very strong conservation and a sharply defined distance preference with respect to the end of the annotated 3′ UTR ($P = 10^{-69}$). Finally, several motifs (for example, MO24 = TAATTTAT; MO94 = TTATTTT) are variants of known AU-rich elements, which are known to mediate mRNA instability and degradation[99].

MicroRNA targeting in protein-coding regions. Protein-coding regions can also harbour functional regulatory motifs, such as exonic splicing regulatory elements[100]. However, motif conservation is difficult to assess within protein-coding regions because of the overlapping selective pressures. Indeed, the most highly conserved nucleotide sequence patterns of length seven (7mers) in coding sequence showed strong reading-frame-biased conservation, suggesting that they reflect protein-coding constraints rather than regulatory roles at the DNA or RNA level (Fig. 6c).

MicroRNA motifs, which function at the RNA level, instead showed high conservation in all three reading frames, suggesting that they are specifically selected within coding regions for their RNA-level function. Indeed, previous studies have shown that miRNA motifs in coding regions are preferentially conserved in vertebrates[86], that they can lead to repression in experimental assays[101,102], and that they are avoided in genes co-expressed with the miRNA[103]. Frame-invariant conservation allows us to demonstrate the coding-region targeting of individual miRNAs, and also enables the *de novo* discovery of miRNA motifs in coding regions. Using frame-invariant conservation, we recovered 11 miRNA motifs within the top 20 coding-region motifs (Supplementary Table 5g), whereas using overall conservation required several hundred candidates to recover 11 miRNA motifs.

Moreover, 7mers complementary to different positions in the mature miRNA show a distinctive conservation pattern indicative of functional targeting in coding regions (Fig. 6d) and similar to that found in 3′ UTRs[12,83] (correlation coefficient 0.96). Finally, 6mers complementary to miRNA 5′ ends were depleted in coding exons of anti-target genes (Supplementary Fig. 5f), similar to findings for these genes' 3′ UTRs[103,104]. Overall, these results, together with findings in vertebrates[86,101–103], suggest that important miRNA targets have been overlooked by many target prediction methods[105] that have traditionally focused exclusively on 3′ UTR sequences.

Prediction of individual regulator binding sites

Previous methods for regulatory motif discovery[3,4,12–14] integrated conservation information over hundreds of motif instances across the genome, leading to an exceedingly clear signal for motif discovery even if many of these instances are only marginally conserved. In contrast, the reliable identification of individual motif instances has been hampered by lack of neutral divergence and would require many related genomes[15–19]. In the absence of such data, previous studies have relied on motif clustering[91,94–96] or other sequence characteristics[106] to predict regulatory targets or regions.

With the availability of the 12 fly genomes, we inferred high-confidence instances of regulatory motifs by mapping the BLS of each motif instance to a confidence value (Supplementary Methods 5a). This value represents the probability that a motif instance is functional, on the basis of the conservation level of appropriate control motifs evaluated in the same type of region (promoter, 3′ UTR, coding, and so on). Because the number of conserved instances decreases much more rapidly for control motifs than for real motifs, the many genomes allowed us to reach high confidence values for many transcription factors and miRNAs, even at relatively modest BLS thresholds (Fig. 2e).

Conserved motif instances identify functional *in vivo* targets. We found that increasing confidence levels selected for functional instances for both transcription factor and miRNA motifs: the normalized fraction of transcription factor motif instances within promoter regions rose from 20% to 90%; that of miRNA motif instances within 3′ UTRs rose from 20% to 90%; and the fraction of miRNA motif instances on the transcribed strand of 3′ UTRs rose from 50% (uniform) to 100% (Fig. 7a); in each case selecting the regions and strands where the motifs are known to be functional.

We further assessed how predicted motif instances compared with *in vivo* targets in promoter regions, defined experimentally (without comparative information). We used a set of high-confidence direct CrebA targets[107] and three genome-wide chromatin

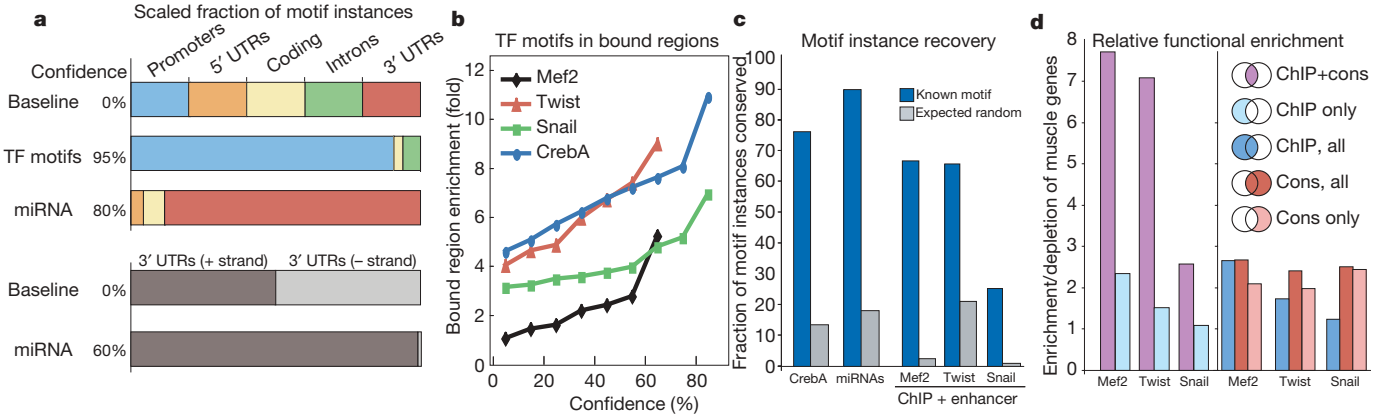


Figure 7 | Identification of individual motif instances. **a**, Increasing confidence levels select for motif instances in regions they are known to be functional: conserved transcription factor (TF) motifs enrich for promoters; miRNA motifs for 3′UTRs, and specifically the transcribed strand. Regions are normalized for their overall length, measured by the number of motif instances without conservation (0% confidence baseline). **b**, Increasing confidence levels select for transcription factor motif instances with experimental support for each factor tested. **c**, The high fraction of experimentally supported motif instances that are recovered at 60% confidence for transcription factors and 80% confidence for miRNAs illustrates the high sensitivity of the BLS approach. **d**, Comparison of chromatin immunoprecipitation (ChIP) and conservation in their ability to identify functional motif instances. Motif instances that are both ChIP-bound and conserved (purple) show the strongest functional enrichment in muscle genes for Mef2 and Twist (depletion for Snail), whereas motif instances derived by ChIP alone (light blue) show substantially reduced enrichment levels. Comparing the enrichment of all instances recovered by ChIP (blue) and all instances recovered by conservation (red) suggests that the two approaches perform comparably. Even the sites recovered by conservation alone outside bound regions (pink) show enrichment levels comparable to ChIP, suggesting that they are also functional.

©2007 Nature Publishing Group

immunoprecipitation (ChIP) data sets for Snail, Mef2 and Twist[92,108,109], and in each case found that the enrichment between conserved motif instances and known *in vivo* regions increased sharply for increasing confidence values (Fig. 7b).

We also found that a large fraction of motif instances in experimentally determined target regions was conserved (Fig. 7c): 76% of motif instances in direct CrebA targets and 90% of motif instances in experimentally supported miRNA targets[104,110] were recovered at 60% confidence. Although many of the miRNA targets stem from comparative predictions and are expected to be well conserved, their high recovery rate illustrates the increased sensitivity of the BLS measure compared to perfect conservation (Supplementary Fig. 7d). Similar results were found for motifs in known enhancers that were determined to be bound by ChIP ('ChIP-bound'): 65% of Mef2 motifs, 65% of Snail motifs and 25% of Twist motifs were conserved (Fig. 7c).

ChIP-determined and conservation-determined targets show similar enrichment. To determine whether ChIP-bound motifs that lack conservation are biologically meaningful, we studied their enrichment in muscle gene promoters. We found that motifs that were both bound and evolutionarily conserved showed very strong correlation with muscle genes for all three factors: Mef2 showed eightfold enrichment, Twist showed sevenfold enrichment and Snail, a mesodermal repressor, showed threefold depletion for muscle genes. However, when only non-conserved sites were considered, the correlation dropped significantly to 1–2-fold for all three factors, suggesting that non-conserved ChIP-bound sites may be of decreased biological significance (Fig. 7d).

We also used the correlation with muscle genes to compare ChIP-on-chip and evolutionary conservation as two complementary methods for target identification (Fig. 7d). We found that the enrichment of conservation-inferred targets was consistently higher than the enrichment of ChIP-inferred targets for each of the three factors. Finally, we assessed the functional significance of motif instances that were only found by the conservation approach, specifically excluding those in ChIP-bound regions, and found that these were also enriched in the same functional categories as ChIP-bound sites with comparable or higher functional correlations (Fig. 7d). This suggests that the additional conserved instances are indeed functional, probably reflecting the higher coverage of conservation-based approaches, which are not restricted to the experimental conditions surveyed, or that they may be bound *in vivo* yet missed by ChIP-on-chip technology[111,112].

In an independent study[113] we compared several strategies for the prediction of motif instances and *cis*-regulatory modules and found that using the 12 fly genomes led to substantial improvements. In another study, we reported the recovery of conserved motifs for several known regulators, including *Suppressor of Hairless*, in genes of the *Enhancer of split* complex[114].

A regulatory network of *D. melanogaster* at 60% confidence. Having established the accuracy of conserved motif instances, we present an initial regulatory network for *D. melanogaster* at 60% confidence (Supplementary Fig. 5i), containing 46,525 regulatory connections between 67 transcription factors and 8,287 genes, and 3,662 connections between 81 cloned miRNAs (clustered in 49 families with unique seed sequences) and 2,003 genes.

The distribution of predicted sites per target gene is highly non-uniform and indicative of varying levels of regulatory control. Genes with the highest number of sites appeared to be enriched in morphogenesis, organogenesis, neurogenesis and a variety of tissues, whereas ubiquitously expressed genes and maternal genes with housekeeping functions had the fewest sites[104]. Interestingly, transcription factors appeared to be more heavily targeted than other genes, both by transcription factors (10 sites versus 5.5 on average, $P = 10^{-15}$) and by miRNAs (2.3 versus 1.8 miRNAs, $P = 5 \times 10^{-5}$). Moreover, genes with many transcription factor sites also had many miRNA sites, and conversely, genes with few transcription factor sites also had few miRNA sites ($P = 10^{-4}$ and $P = 7 \times 10^{-3}$, respectively).

Several of the predicted regulatory connections have independent experimental support (Supplementary Table 5h), including direct regulation of *achaete* by Hairy[115], of *giant* by Bicoid[116], of *Enhancer of split* complex genes by Suppressor of Hairless[117], and of *bagpipe* by Tinman (known to cooperate in mesoderm induction and heart specification[118]). More generally, when tissue-specific expression data were available, we found that on average 46% of all targets were co-expressed with their factor in at least one tissue (Supplementary Fig. 5i), which is significantly higher than expected by chance ($P = 2 \times 10^{-3}$).

Scaling of comparative genomics power

Theoretical considerations and pilot studies on selected genomic regions showed that the discovery power of comparative methods scales with the number and phylogenetic distance of the species compared[16–20,46,119,120]. We extended these analyses by investigating the scaling of genome-wide discovery power using evolutionary signatures for each class of functional elements (Fig. 8), on the basis of the recovery of known elements using different subsets of informant species (at a fixed stringency).

We found that recovery consistently increased with the total number of informant species, and that multi-species comparisons outperformed pairwise comparisons within the same phylogenetic clade. When we examined subsets of informants with similar total branch length (for example, several close species versus one distant species),

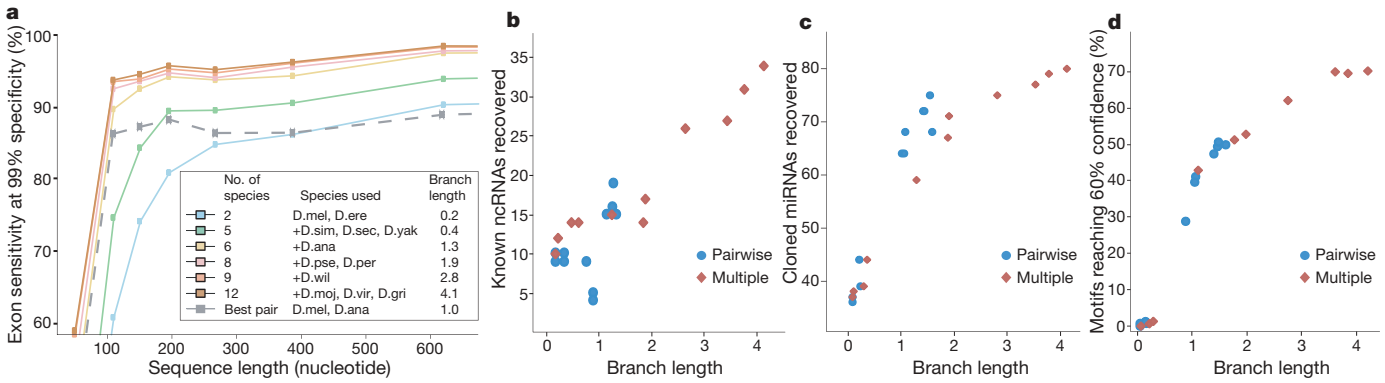


Figure 8 | Scaling of discovery power with the number and distance of informant species. a, Discriminatory power of CSF protein-coding evolutionary metric for varying exon lengths and using different numbers of informant species. Sensitivity is shown for known exons at a fixed false-positive rate based on random non-coding regions. Mean length is shown for each exon length quantile. Multi-species comparisons increase discovery power, especially among short exons. **b**, Recovery of known ncRNAs (among the top 100 predictions) for pairwise (blue) and multi-species (red) comparisons. **c**, Recovery of cloned miRNAs (among the top 100 predictions). **d**, Recovery of transcription factor and miRNA motifs with instances at 60% confidence.

©2007 Nature Publishing Group

multi-species comparisons sometimes performed better (protein-coding exons, ncRNAs), comparably (motifs), or worse (miRNAs) than pairwise comparisons. This complex relationship between total branch length and actual discovery power probably reflects imperfect genome assemblies/alignments, characteristics of each class of functional elements, and the specific methods we used. For example, ncRNA discovery probably benefits from observing more compensatory changes across more genomes, whereas miRNA discovery may be more sensitive to artefacts in low-coverage genomes, given the expected high conservation of miRNA arms.

As expected, longer elements were easier to discover than shorter elements. Long protein-coding exons (>300 nucleotides) were recovered at very high rates even with few species at close distances (leaving little room for improvement with additional species). In contrast, more informant species and larger distances were crucial for recovering short exons, miRNAs and regulatory motifs.

Notably, the optimal evolutionary distance for pairwise comparisons to *D. melanogaster* also seemed to depend on element length: for long protein-coding exons, the best pairwise informant was the closely related *D. erecta*, for exons of intermediate lengths *D. ananassae*, and for the shortest exons the distant *D. willistoni* (Supplementary Table 7a). Distant species were also optimal for other classes of short elements (ncRNAs, miRNAs and motifs, Fig. 8b–d). This suggests that a small number of species at close evolutionary distances may generally allow the discovery of long elements, possibly including clade-specific elements, whereas short clade-specific elements may not be reliably detectable without many genomes at close distances.

Finally, we investigated the effect of alignment choice on our results (Supplementary Fig. 8). We found high similarity between different alignment strategies for longer elements (>93% agreement for exons), whereas shorter elements showed larger discrepancies between alignments (81% and 59% agreement for miRNA and motif instances, respectively).

Although factors such as genome size, repeat density, pseudogene abundance and physiological differences might confound a simple analogy to the vertebrate phylogeny based on neutral branch length (Fig. 1c), our results suggest that comparisons spanning marsupials, birds and reptiles may prove surprisingly useful for biological signal discovery in the human genome.

Discussion

Our results demonstrate the potential of comparative genomics for the systematic characterization of functional elements in a complete genome. Even in a species as intensely studied as *D. melanogaster*, our methods predicted several thousand new functional elements, including protein-coding genes and exons, novel RNA genes and structures, miRNA genes, regulatory motifs, and regulator targets. Our novel predictions have overwhelming statistical support, often surpassing that of known functional elements, and are additionally supported by experimental evidence in hundreds of cases. The common underlying methodology in this study has been the recognition of specific evolutionary signatures associated with each class of functional elements, which can be much more informative for genome annotation than overall measures of nucleotide conservation. These signatures are general and are immediately relevant to the analysis of the human genome and more generally of any species.

In addition to the many new elements, we gained specific biological insights and formulated hypotheses that we hope will guide follow-up experiments. We found 149 genes with potential translational readthrough, showing protein-like evolution downstream of a highly conserved stop codon, and possibly encoding additional protein domains or peptides specific to certain developmental contexts. We also found several candidate programmed frameshifts, which might be part of regulatory circuits (as for *ODC*/*Oda* [64]) or help expand the diversity of protein products generated from one mRNA, similar to their role in prokaryotes[121]. We also presented evidence of miRNA processing from both arms of a miRNA hairpin and from both DNA strands of a miRNA locus in some cases, potentially leading to as many as four functional miRNAs per locus. As miRNA/miRNA* pairs are expressed from a single precursor and thus co-regulated, whereas sense/anti-sense pairs are expressed from distinct promoters, the use of both arms or both strands provides compelling general building blocks for higher-level miRNA-mediated regulation.

The newly discovered elements did not dramatically increase the total number of annotated nucleotides. Known and predicted elements explain 42% of nucleotides in phastCons elements[33], compared to 35.5% for previous annotations (Supplementary Fig. 6), an 18% increase (mostly owing to conserved motif instances). The remaining phastCons elements and independent estimates based on transcriptional activity[42] would suggest that a much higher fraction of the genome may be functional (Supplementary Fig. 6). Although it is possible that these estimates are artificially high and that we are in fact converging on a complete annotation of the fly genome, they might instead indicate that much remains to be discovered, which may require the recognition of as-yet-unknown classes of functional elements with distinct evolutionary signatures.

Our results also allowed us to compare and contrast evolutionary and experimental methods for the recovery of functional elements, particularly for the identification of regulator targets. We found that comparative genomics resulted in many functionally meaningful sites for transcription factors Mef2, Twist and Snail outside ChIP-bound regions, probably representing targets from diverse conditions not surveyed experimentally. Similarly, ChIP resulted in many additional sites outside those recovered by comparative genomics: some of these may have been replaced by functionally equivalent non-orthologous sequence, rendering them apparently non-conserved in sequence alignments[122–124]; others may have species- or lineage-specific roles, thus lacking sufficient signal for their comparative detection; finally, some bound sites may be biochemically active yet selectively neutral[125]. It is worth noting, however, that ChIP-bound motifs that were not conserved showed decreased enrichment in muscle/mesoderm development where the factors are known to act, suggesting that potential lineage-specific roles may lie outside the regulators' conserved functions. To resolve these questions, comparative genomics studies would benefit greatly from experimental studies in several related species in parallel.

Overall, comparative genomics and species-specific experimental studies provide complementary approaches to biological signal discovery. Comparative studies help pinpoint evolutionarily selected functional elements across diverse conditions, whereas experimental studies reveal stage- and tissue-specific information, as well as species-specific sites. Ultimately, their integration is a necessary step towards a comprehensive understanding of animal genomes.

METHODS SUMMARY

The Methods are described in Supplementary Information, with more details found in the cited companion papers for each section. The sections of the Supplementary Methods are arranged in the same order as the manuscript to facilitate cross-referencing, with an index on the first page to aid navigation.

Received 21 July; accepted 4 October 2007.

1. Miller, W., Makova, K. D., Nekrutenko, A. & Hardison, R. C. Comparative genomics. *Annu. Rev. Genomics Hum. Genet.* **5,** 15–56 (2004).
2. Ureta-Vidal, A., Ettwiller, L. & Birney, E. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nature Rev. Genet.* **4,** 251–262 (2003).
3. Kellis, M. *et al.* Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423,** 241–254 (2003).
4. Cliften, P. *et al.* Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301,** 71–76 (2003).
5. Brent, M. R. Genome annotation past, present, and future: how to define an ORF at each locus. *Genome Res.* **15,** 1777–1786 (2005).
6. Washietl, S., Hofacker, I. L. & Stadler, P. F. Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA* **102,** 2454–2459 (2005).
7. Pedersen, J. S. *et al.* Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.* **2,** e33 (2006).

©2007 Nature Publishing Group

8. Lim, L. P. *et al.* The microRNAs of *Caenorhabditis elegans*. *Genes Dev.* **17,** 991–1008 (2003).
9. Lim, L. P. *et al.* Vertebrate microRNA genes. *Science* **299,** 1540 (2003).
10. Lai, E. C., Tomancak, P., Williams, R. W. & Rubin, G. M. Computational identification of *Drosophila* microRNA genes. *Genome Biol.* **4,** R42 (2003).
11. Berezikov, E. *et al.* Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120,** 21–24 (2005).
12. Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434,** 338–345 (2005).
13. Ettwiller, L. *et al.* The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates. *Genome Biol.* **6,** R104 (2005).
14. Chan, C. S., Elemento, O. & Tavazoie, S. Revealing posttranscriptional regulatory elements through network-level conservation. *PLoS Comput. Biol.* **1,** e69 (2005).
15. Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299,** 1391–1394 (2003).
16. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15,** 901–913 (2005).
17. Margulies, E. H., Blanchette, M., Haussler, D. & Green, E. D. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13,** 2507–2518 (2003).
18. Thomas, J. W. *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424,** 788–793 (2003).
19. Eddy, S. R. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.* **3,** e10 (2005).
20. Bergman, C. M. *et al.* Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome. *Genome Biol.* **3,** RESEARCH0086 (2002).
21. Rubin, G. M. & Lewis, E. B. A brief history of *Drosophila*'s contributions to genome research. *Science* **287,** 2216–2218 (2000).
22. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287,** 2185–2195 (2000).
23. Misra, S. *et al.* Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. *Genome Biol.* **3,** RESEARCH0083 (2002).
24. Celniker, S. E. & Rubin, G. M. The *Drosophila melanogaster* genome. *Annu. Rev. Genomics Hum. Genet.* **4,** 89–117 (2003).
25. Ashburner, M. & Bergman, C. M. *Drosophila melanogaster*: a case study of a model genomic sequence and its consequences. *Genome Res.* **15,** 1661–1667 (2005).
26. Matthews, K. A., Kaufman, T. C. & Gelbart, W. M. Research resources for *Drosophila*: the expanding universe. *Nature Rev. Genet.* **6,** 179–193 (2005).
27. Venken, K. J., He, Y., Hoskins, R. A. & Bellen, H. J. P[acman]: a BAC transgenic platform for targeted insertion of large DNA fragments in *D. melanogaster*. *Science* **314,** 1747–1751 (2006).
28. Dietzl, G. *et al.* A genome-wide transgenic RNAi library for conditional gene inactivation in *Drosophila*. *Nature* **448,** 151–156 (2007).
29. Spradling, A. C. *et al.* The Berkeley *Drosophila* Genome Project gene disruption project: Single P-element insertions mutating 25% of vital *Drosophila* genes. *Genetics* **153,** 135–177 (1999).
30. St Johnston, D. The art and design of genetic screens: *Drosophila melanogaster*. *Nature Rev. Genet.* **3,** 176–188 (2002).
31. Richards, S. *et al.* Comparative genome sequencing of *Drosophila* pseudoobscura: chromosomal, gene, and cis-element evolution. *Genome Res.* **15,** 1–18 (2005).
32. Drosophila 12 Genomes Consortium.. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* doi:10.1038/nature06341 (this issue) (2007).
33. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).
34. Nekrutenko, A., Makova, K. D. & Li, W. H. The K_A/K_S ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res.* **12,** 198–202 (2002).
35. Eddy, S. R. Computational genomics of noncoding RNA genes. *Cell* **109,** 137–140 (2002).
36. Bompfuenewerer, A. F. *et al.* Evolutionary patterns of non-coding RNAs. *Theor. Biosci.* **123,** 301–369 (2004).
37. Reese, M. G. *et al.* Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.* **10,** 483–501 (2000).
38. Rubin, G. M. *et al.* A *Drosophila* complementary DNA resource. *Science* **287,** 2222–2224 (2000).
39. Stapleton, M. *et al.* A *Drosophila* full-length cDNA resource. *Genome Biol.* 3, RESEARCH0080 (2002).
40. Hild, M. *et al.* An integrated gene annotation and transcriptional profiling approach towards the full gene content of the *Drosophila* genome. *Genome Biol.* **5,** R3 (2003).
41. Yandell, M. *et al.* A computational and experimental approach to validating annotations and gene predictions in the *Drosophila melanogaster* genome. *Proc. Natl Acad. Sci. USA* **102,** 1566–1571 (2005).
42. Manak, J. R. *et al.* Biological function of unannotated transcription during the early development of *Drosophila melanogaster*. *Nature Genet.* **38,** 1151–1158 (2006).
43. Lin, M. F. *et al.* Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using twelve fly genomes. *Genome Res.* doi:10.1101/gr.6679507 (in the press).
44. Yang, Z. & Bielawski, J. P. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15,** 496–503 (2000).
45. Mignone, F., Grillo, G., Liuni, S. & Pesole, G. Computational identification of protein coding potential of conserved sequence tags through cross-species evolutionary analysis. *Nucleic Acids Res.* **31,** 4639–4645 (2003).
46. Zhang, L., Pavlovic, V., Cantor, C. R. & Kasif, S. Human-mouse gene identification by comparative evidence integration and evolutionary analysis. *Genome Res.* **13** (6A), 1190–1202 (2003).
47. Crosby, M. A. *et al.* FlyBase: genomes by the dozen. *Nucleic Acids Res.* 35 (Database issue) D486–D491 (2007).
48. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25,** 25–29 (2000).
49. Ochman, H., Ajioka, J. W., Garza, D. & Hartl, D. L. Inverse polymerase chain reaction. *Bio/Technology* **8,** 759–760 (1990).
50. Hoskins, R. A. *et al.* Rapid and efficient cDNA library screening by self-ligation of inverse PCR products (SLIP). *Nucleic Acids Res.* **33,** e185 (2005).
51. Wan, K. H. *et al.* High-throughput plasmid cDNA library screening. *Nature Protocols* **1,** 624–632 (2006).
52. Hahn, M. W., Han, M. V. & Han, S.-G. Gene family evolution across 12 *Drosophila* genomes. *PLoS Genet* **3,** e197 (2007).
53. Andrews, J. *et al.* The stoned locus of *Drosophila melanogaster* produces a dicistronic transcript and encodes two distinct polypeptides. *Genetics* **143,** 1699–1711 (1996).
54. Brogna, S. & Ashburner, M. The Adh-related gene of *Drosophila melanogaster* is expressed as a functional dicistronic messenger RNA: multigenic transcription in higher organisms. *EMBO J.* **16,** 2023–2031 (1997).
55. Hatfield, D. L. & Gladyshev, V. N. How selenium has altered our understanding of the genetic code. *Mol. Cell. Biol.* **22,** 3565–3576 (2002).
56. Kryukov, G. V. *et al.* Characterization of mammalian selenoproteomes. *Science* **300,** 1439–1443 (2003).
57. Copeland, P. R. Regulation of gene expression by stop codon recoding: selenocysteine. *Gene* **312,** 17–25 (2003).
58. Castellano, S. *et al. In silico* identification of novel selenoproteins in the *Drosophila melanogaster* genome. *EMBO Rep.* **2,** 697–702 (2001).
59. von der Haar, T. & Tuite, M. F. Regulated translational bypass of stop codons in yeast. *Trends Microbiol.* **15,** 78–86 (2007).
60. Luo, G. X. *et al.* A specific base transition occurs on replicating hepatitis delta virus RNA. *J. Virol.* **64,** 1021–1027 (1990).
61. Casey, J. L. & Gerin, J. L. Hepatitis D virus RNA editing: specific modification of adenosine in the antigenomic RNA. *J. Virol.* **69,** 7593–7600 (1995).
62. Steneberg, P. *et al.* Translational readthrough in the hdc mRNA generates a novel branching inhibitor in the *Drosophila* trachea. *Genes Dev.* **12,** 956–967 (1998).
63. Bass, B. L. RNA editing by adenosine deaminases that act on RNA. *Annu. Rev. Biochem.* **71,** 817–846 (2002).
64. Ivanov, I. P. *et al.* The *Drosophila* gene for antizyme requires ribosomal frameshifting for expression and contains an intronic gene for snRNP Sm D3 on the opposite strand. *Mol. Cell. Biol.* **18,** 1553–1561 (1998).
65. Eddy, S. R. Non-coding RNA genes and the modern RNA world. *Nature Rev. Genet.* **2,** 919–929 (2001).
66. Yuan, G. *et al.* RNomics in *Drosophila melanogaster*: identification of 66 candidates for novel non-messenger RNAs. *Nucleic Acids Res.* **31,** 2495–2507 (2003).
67. Lestrade, L. & Weber, M. J. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.* **34** (Database issue), D158–D162 (2006).
68. Bier, E. *Drosophila*, the golden bug, emerges as a tool for human genetics. *Nature Rev. Genet.* **6,** 9–23 (2005).
69. Hoopengardner, B., Bhalla, T., Staber, C. & Reenan, R. Nervous system targets of RNA editing identified by comparative genomics. *Science* **301,** 832–836 (2003).
70. Mignone, F. *et al.* UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res.* **33** (Database issue), D141–D146 (2005).
71. Cohen, R. S., Zhang, S. & Dollar, G. L. The positional, structural, and sequence requirements of the *Drosophila* TLS RNA localization element. *RNA* **11,** 1017–1029 (2005).
72. Allemand, F. *et al. Escherichia coli* ribosomal protein L20 binds as a single monomer to its own mRNA bearing two potential binding sites. *Nucleic Acids Res.* **35,** 3016–3031 (2007).
73. Okumura, T., Matsumoto, A., Tanimura, T. & Murakami, R. An endoderm-specific GATA factor gene, dGATAe, is required for the terminal differentiation of the *Drosophila* endoderm. *Dev. Biol.* **278,** 576–586 (2005).
74. Park, S. W. *et al.* An evolutionarily conserved domain of roX2 RNA is sufficient for induction of H4-Lys16 acetylation on the *Drosophila* X chromosome. *Genetics* (in the press).
75. Park, Y. & Kuroda, M. I. Epigenetic aspects of X-chromosome dosage compensation. *Science* **293,** 1083–1085 (2001).
76. Berezikov, E., Cuppen, E. & Plasterk, R. H. Approaches to microRNA discovery. *Nature Genet.* **38** (Suppl 1), S2–S7 (2006).
77. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116,** 281–297 (2004).
78. Stark, A. *et al.* Systematic discovery and characterization of fly microRNAs using 12 *Drosophila* genomes. *Genome Res.* doi:10.1101/gr.6593807 (in the press).
79. Ruby, J. G. *et al.* Evolution, biogenesis, expression, and target predictions of a substantially expanded set of *Drosophila* microRNAs. *Genome Res.* doi:10.1101/gr.6597907 (in the press).
80. Pekarsky, Y. *et al.* Tcl1 expression in chronic lymphocytic leukemia is regulated by miR-29 and miR-181. *Cancer Res.* **66,** 11590–11593 (2006).
81. Ruby, J. G., Jan, C. H. & Bartel, D. P. Intronic microRNA precursors that bypass Drosha processing. *Nature* **448,** 83–86 (2007).

©2007 Nature Publishing Group

82. Okamura, K. *et al.* The mirtron pathway generates microRNA-class regulatory RNAs in *Drosophila*. *Cell* **130,** 89–100 (2007).
83. Lewis, B. P. *et al.* Prediction of mammalian microRNA targets. *Cell* **115,** 787–798 (2003).
84. Stark, A., Brennecke, J., Russell, R. B. & Cohen, S. M. Identification of *Drosophila* microRNA targets. *PLoS Biol.* **1,** E60 (2003).
85. Lai, E. C. Micro RNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation. *Nature Genet.* **30,** 363–364 (2002).
86. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120,** 15–20 (2005).
87. Tompa, M. Identifying functional elements by comparative DNA sequence analysis. *Genome Res.* **11,** 1143–1144 (2001).
88. Stormo, G. D. DNA binding sites: representation and discovery. *Bioinformatics* **16,** 16–23 (2000).
89. Kheradpour, P., Stark, A., Roy, S. & Kellis, M. Reliable prediction of regulator targets using 12 *Drosophila* genomes. *Genome Res.* doi:10.1101/gr.7090407 (in the press).
90. Stathopoulos, A. & Levine, M. Genomic regulatory networks and animal development. *Dev. Cell* **9,** 449–462 (2005).
91. Schroeder, M. D. *et al.* Transcriptional control in the segmentation gene network of *Drosophila*. *PLoS Biol.* **2,** e271 (2004).
92. Zeitlinger, J. *et al.* Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the *Drosophila* embryo. *Genes Dev.* **21,** 385–390 (2007).
93. Kanehisa, M. *et al.* The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32 (Database issue) D277–D280 (2004).
94. Berman, B. P. *et al.* Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl Acad. Sci. USA* **99,** 757–762 (2002).
95. Markstein, M. *et al.* A regulatory code for neurogenic gene expression in the *Drosophila* embryo. *Development* **131,** 2387–2394 (2004).
96. Philippakis, A. A. *et al.* Expression-guided *in silico* evaluation of candidate cis regulatory codes for *Drosophila* muscle founder cells. *PLoS Comput. Biol.* **2,** e53 (2006).
97. Smale, S. T. & Kadonaga, J. T. The RNA polymerase II core promoter. *Annu. Rev. Biochem.* **72,** 449–479 (2003).
98. Gerber, A. P. *et al.* Genome-wide identification of mRNAs associated with the translational regulator PUMILIO in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **103,** 4487–4492 (2006).
99. Zubiaga, A. M., Belasco, J. G. & Greenberg, M. E. The nonamer UUAUUUAUU is the key AU-rich sequence motif that mediates mRNA degradation. *Mol. Cell. Biol.* **15,** 2219–2230 (1995).
100. Fairbrother, W. G., Yeh, R. F., Sharp, P. A. & Burge, C. B. Predictive identification of exonic splicing enhancers in human genes. *Science* **297,** 1007–1013 (2002).
101. Kloosterman, W. P., Wienholds, E., Ketting, R. F. & Plasterk, R. H. Substrate requirements for let-7 function in the developing zebrafish embryo. *Nucleic Acids Res.* **32,** 6284–6291 (2004).
102. Grimson, A. *et al.* MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell* **27,** 91–105 (2007).
103. Farh, K. K. *et al.* The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science* **310,** 1817–1821 (2005).
104. Stark, A. *et al.* Animal microRNAs confer robustness to gene expression and have a significant impact on 3′ UTR evolution. *Cell* **123,** 1133–1146 (2005).
105. Rajewsky, N. microRNA target predictions in animals. *Nature Genet.* **38** (suppl. 1), S8–S13 (2006).
106. Elnitski, L. *et al.* Distinguishing regulatory DNA from neutral sites. *Genome Res.* **13,** 64–72 (2003).
107. Abrams, E. W. & Andrew, D. J. CrebA regulates secretory activity in the *Drosophila* salivary gland and epidermis. *Development* **132,** 2743–2758 (2005).
108. Sandmann, T. *et al.* A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. *Dev. Cell* **10,** 797–807 (2006).
109. Sandmann, T. *et al.* A core transcriptional network for early mesoderm development in *Drosophila melanogaster*. *Genes Dev.* **21,** 436–449 (2007).
110. Sethupathy, P., Corda, B. & Hatzigeorgiou, A. G. TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA* **12,** 192–197 (2006).
111. Lee, T. I. *et al.* Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125,** 301–313 (2006).
112. Boyer, L. A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122,** 947–956 (2005).
113. Aerts, S., van Helden, J., Sand, O. & Hassan, B. Fine-tuning enhancer models to predict transcriptional targets across multiple genomes. *PLoS ONE* **2** (11), e1115 (2007).
114. Maeder, M., Polansky, B., Robson, B. & Eastman, D. Phylogenetic footprinting analysis in the upstream regulatory regions of the *Drosophila* Enhancer of split genes. *Genetics* (in the press).
115. Van Doren, M. *et al.* Negative regulation of proneural gene activity: hairy is a direct transcriptional repressor of achaete. *Genes Dev.* **8,** 2729–2742 (1994).
116. Kraut, R. & Levine, M. Spatial regulation of the gap gene giant during *Drosophila* development. *Development* **111,** 601–609 (1991).
117. Bailey, A. M. & Posakony, J. W. Suppressor of hairless directly activates transcription of enhancer of split complex genes in response to Notch receptor activity. *Genes Dev.* **9,** 2609–2622 (1995).
118. Yin, Z. & Frasch, M. Regulation and function of tinman during dorsal mesoderm induction and heart specification in *Drosophila*. *Dev. Genet.* **22,** 187–200 (1998).
119. Margulies, E. H. *et al.* An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl Acad. Sci. USA* **102,** 4795–4800 (2005).
120. Margulies, E. H., Chen, C. W. & Green, E. D. Differences between pair-wise and multi-sequence alignment methods affect vertebrate genome comparisons. *Trends Genet.* **22,** 187–193 (2006).
121. Farabaugh, P. J. Programmed translational frameshifting. *Annu. Rev. Genet.* **30,** 507–528 (1996).
122. Odom, D. T. *et al.* Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature Genet.* **39,** 730–732 (2007).
123. Ludwig, M. Z. & Kreitman, M. Evolutionary dynamics of the enhancer region of even-skipped in *Drosophila*. *Mol. Biol. Evol.* **12,** 1002–1011 (1995).
124. Ludwig, M. Z. *et al.* Functional evolution of a cis-regulatory module. *PLoS Biol.* **3,** e93 (2005).
125. The ENCODE Project Consortium.. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
126. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the National Human Genome Research Institute (NHGRI) for continued support. A.S. was supported in part by the Schering AG/Ernst Schering Foundation and in part by the Human Frontier Science Program Organization (HFSPO). P.K. was supported in part by a National Science Foundation Graduate Research Fellowship. J.S.P. thanks B. Raney and R. Baertsch, and the Danish Medical Research Council and the National Cancer Institute for support. J.B. thanks the Schering AG/Ernst Schering Foundation for a postdoctoral fellowship. L.Parts thanks J. Vilo. S.R. was supported by a HHMI-NIH/NIBIB Interfaces Training Grant and thanks T. Lane and M. Werner-Washburne. D.H., D.P.B., G.J.H. and T.C.K. are Investigators of the Howard Hughes Medical Institute, and B.P., J.G.R., E.H. and J.B. are affiliated with these investigators. J.W.C. and S.E.C. were supported by the NHGRI. M.K. was supported by start-up funds from the MIT Electrical Engineering and Computer Science Laboratory, the Broad Institute of MIT and Harvard, and the MIT Computer Science and Artificial Intelligence Laboratory, and by the Distinguished Alumnus (1964) Career Development Professorship.

Author Contributions Organizing committee: Manolis Kellis, William Gelbart, Doug Smith, Andrew G. Clark, Michael E. Eisen, Thomas C. Kaufman; protein-coding gene prediction: Michael F. Lin, Ameya N. Deoras, Mira V. Han, Matthew W. Hahn, Donald G. Gilbert, Michael Weir, Michael Rice, Manolis Kellis; manual curation of protein-coding genes: Madeline A. Crosby, Harvard FlyBase curators, William M. Gelbart; validation of protein-coding genes: Joseph W. Carlson, Berkeley Drosophila Genome Project, Susan E. Celniker; non-coding RNA gene prediction: Jakob S. Pedersen, David Haussler, Yongkyu Park, Seung-Won Park, Manolis Kellis; microRNA gene prediction: Alexander Stark, Pouya Kheradpour, Leopold Parts, Manolis Kellis; microRNA cloning and sequencing: Julius Brennecke, Emily Hodges, Gregory J. Hannon; microRNA target prediction: Alexander Stark, J. Graham Ruby, Manolis Kellis, Eric C. Lai, David P. Bartel; motif identification: Alexander Stark, Pouya Kheradpour, Manolis Kellis; motif instance prediction: Alexander Stark, Pouya Kheradpour, Sushmita Roy, Morgan L. Maeder, Benjamin J. Polansky, Bryanne E. Robson, Deborah A. Eastman, Stein Aerts, Bassem Hassan, Jacques van Helden, Manolis Kellis; genome alignments: Angie S. Hinrichs, W. James Kent, Anat Caspi, Lior Pachter, Colin N. Dewey, Benedict Paten; phylogeny and branch length estimation: Matthew D. Rasmussen, Manolis Kellis; final manuscript preparation: Alexander Stark, Michael F. Lin, Pouya Kheradpour, Jakob Pedersen, Manolis Kellis.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.K. (manoli@mit.edu).

Harvard FlyBase curators Madeline A. Crosby[1], Beverley B. Matthews[1], Andrew J. Schroeder[1], L. Sian Gramates[1], Susan E. St Pierre[1], Margaret Roark[1], Kenneth L. Wiley Jr[1], Rob J. Kulathinal[1], Peili Zhang[1], Kyl V. Myrick[1], Jerry V. Antone[1] & William M. Gelbart[1]

Berkeley Drosophila Genome Project Joseph W. Carlson[2], Charles Yu[2], Soo Park[2], Kenneth H. Wan[2] & Susan E. Celniker[2]

[1]FlyBase, The Biological Laboratories, Harvard University, 16 Divinity Avenue, Cambridge, Massachusetts 02138, USA. [2]BDGP, LBNL, 1 Cyclotron Road MS 64-0119, Berkeley, California 94720, USA.

©2007 Nature Publishing Group

nature genetics

© 2007 Nature Publishing Group http://www.nature.com/naturegenetics

RNA polymerase stalling at developmental control genes in the *Drosophila melanogaster* embryo

Julia Zeitlinger[1,7], Alexander Stark[2,3], Manolis Kellis[2,3], Joung-Woo Hong[4], Sergei Nechaev[5], Karen Adelman[5], Michael Levine[4] & Richard A Young[1,6]

It is widely assumed that the key rate-limiting step in gene activation is the recruitment of RNA polymerase II (Pol II) to the core promoter[1]. Although there are well-documented examples in which Pol II is recruited to a gene but stalls[2–12], a general role for Pol II stalling in development has not been established. We have carried out comprehensive Pol II chromatin immunoprecipitation microarray (ChIP-chip) assays in *Drosophila* embryos and identified three distinct Pol II binding behaviors: active (uniform binding across the entire transcription unit), no binding, and stalled (binding at the transcription start site). The notable feature of the ~10% genes that are stalled is that they are highly enriched for developmental control genes, which are either repressed or poised for activation during later stages of embryogenesis. We propose that Pol II stalling facilitates rapid temporal and spatial changes in gene activity during development.

Pol II stalling is probably best studied at heat shock genes in *Drosophila*, where Pol II engages in transcription but pauses immediately downstream of the transcriptional start site[2,3,13]. Upon activation by heat shock, Pol II is able to rapidly transcribe these genes. Regulation of Pol II activity after recruitment has also been described in bacteria[14], yeast[11] and mammalian cell lines[2,5–10], and it includes instances where Pol II is found in an inactive preinitiation complex[15,16]. We will collectively refer to inactive Pol II near the transcription start site as stalled Pol II.

To determine at which genes Pol II stalling occurs during development, we analyzed global Pol II occupancy in whole *Drosophila* embryos. Although this is one of the few systems in which genomics approaches can easily be applied to developmental questions, interpretation is complicated by the occurrence of multiple tissues. To reduce the complexity, we used *Toll*^{10b} embryos (2–4 h after fertilization), a well-characterized mutant that contains a homogeneous population of mesodermal precursor cells at the expense of neuronal and ectodermal cells[17–23]. In *Toll*^{10b} mutants, mesodermal genes are uniformly activated, whereas genes required for the development of ectodermal and neural tissues are repressed throughout the embryo[17–19]. Previous whole-genome microarray experiments have identified the transcript levels of all genes in these mutants[21,22]. To distinguish between stalled and active Pol II, we used a mixture of antibodies that recognizes both the initiating and elongating forms of Pol II (see Methods) and carried out whole-genome ChIP-chip assays as previously described[23].

The results show that many genes known to be repressed in *Toll*^{10b} embryos show notably high Pol II signal near the transcription start site (**Fig. 1a–d**). In some cases, the prominent Pol II peak was tightly restricted to the promoter region (for example, at the tail-up (*tup*) gene; **Fig. 1a**), whereas at other genes Pol II was also found at low abundance throughout the transcription unit (for example, the *sog* and *brk* genes; **Fig. 1c,d**). This is consistent with previous evidence that some genes, such as *sog*, are transiently activated but then repressed at later stages[24], whereas others, such as *tup*, are never activated in *Toll*^{10b} mutants[21,22].

The Pol II profiles of repressed genes are clearly distinct from those of active genes (**Fig. 1e,f**). For example, the *stumps* (also known as *Hbr*) gene, which encodes a fibroblast growth factor (FGF) receptor specifically expressed in mesodermal precursors (**Fig. 1e**), and ribosomal genes such as *RpL3* (**Fig. 1f**) show uniformly high levels of Pol II throughout the transcription unit. Furthermore, genes that are silent in the early embryo simply lack Pol II binding altogether (**Fig. 1g,h**). Thus, there appear to be three distinct classes of genes: those with Pol II distributed throughout the transcription unit, those with preferential enrichment of Pol II at the transcription site and those that lack Pol II binding altogether.

To further characterize these three groups, we developed a principled method that classifies genes on the basis of their Pol II enrichment profiles (**Fig. 2** and **Supplementary Note** online). In an

[1]Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA. [2]Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts 02141, USA. [3]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [4]Department of Molecular Cell Biology, Center for Integrative Genomics, University of California, Berkeley, California 94720, USA. [5]Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, North Carolina 27709, USA. [6]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [7]Present address: Stowers Institute for Medical Research, 1000 East 50th St., Kansas City, Missouri 64110, USA. Correspondence should be addressed to M.L. (mlevine@berkeley.edu) or R.A.Y. (young@wi.mit.edu).

Received 10 May; accepted 7 September; published online 11 November 2007; doi:10.1038/ng.2007.26

© 2007 Nature Publishing Group http://www.nature.com/naturegenetics

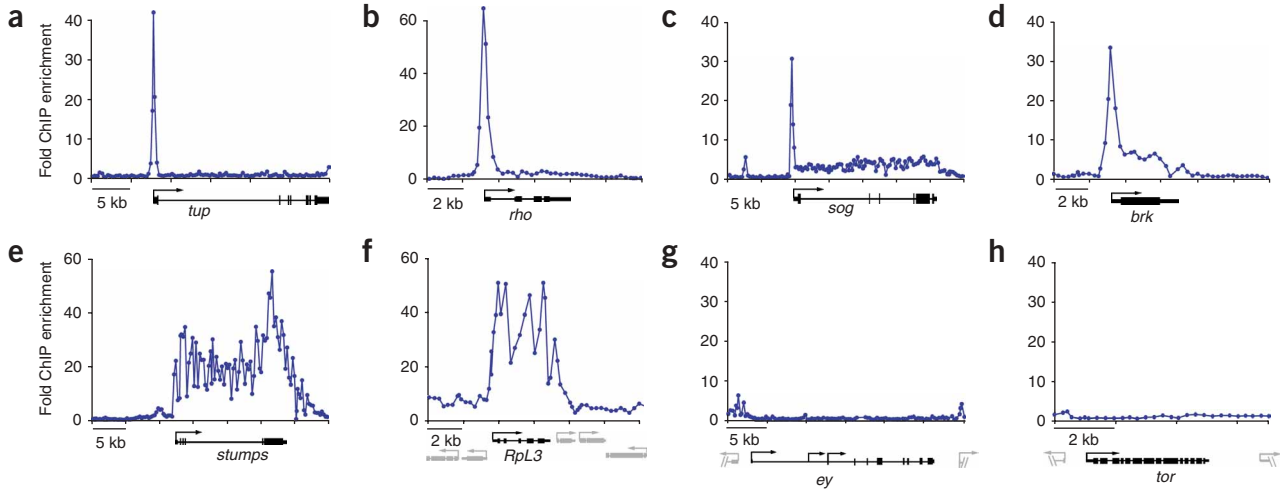


Figure 1 Different classes of Pol II binding profiles. ChIP-chip assays were carried out with 2–4 h *Toll*^{10b} embryos using antibodies that recognize both the initiating and the elongating forms of Pol II. The enrichment ratios of Pol II are shown on the *y* axis. (**a**–**d**) Binding patterns across genes that are repressed in *Toll*^{10b} embryos. All four genes show high Pol II signals near the transcription start sites. At some genes, such as *tup* (**a**), Pol II is tightly restricted to this region, whereas at other genes, including *sog* (**c**) and *brk* (**d**), Pol II is also detected at lower signals throughout the transcription unit. (**e**,**f**) Pol II is uniformly distributed across the transcription units of genes that are actively transcribed. The *stumps* (also called *Hbr* or *Dof*—downstream of FGF) gene (**e**) is specifically activated in mesodermal precursor cells, whereas *RpL3* (**f**) is a highly expressed ribosomal gene. (**g**,**h**) No Pol II binding is found at many genes that are inactive during embryogenesis. The eyeless (*ey*) gene (**g**) is expressed during eye development at larval stages but not in the early embryo. Likewise, the torso (*tor*) gene (**h**) is active only during oogenesis but not in the early embryo.

analysis similar to one carried out in *Escherichia coli*[14], we calculated the ratio between Pol II enrichment at the transcription start site versus internal regions of the transcription unit (**Fig. 2**). We were able to assign 76% of the protein coding genes (10,220 of 13,448 genes) to one of the three classes (**Fig. 2b**). At least 27% of all genes had an active Pol II profile in which Pol II was detected uniformly throughout the transcription unit. At least 12% of all genes (1,614 of 13,448) showed disproportionate accumulation of Pol II near the transcription start site. Among this group, Pol II was tightly restricted to the transcription start site at 62% of genes. At the remaining 38% of these genes, Pol II was also detected within the transcription unit, presumably because these genes—such as *sog*—are expressed at low levels in at least a subset of cells during the time frame of the analysis (2–4 h after fertilization). Finally, 37% of all genes lacked Pol II binding altogether.

Several lines of evidence confirm that the ~1,600 genes with disproportionate enrichment of Pol II at the transcription start site have a form of stalled Pol II (**Fig. 3**). First, all heat shock genes, which provide the classical example of Pol II stalling[3,13], fall into this class (**Fig. 3a**). Second, the Pol II peaks map an average of ~50 bp downstream of the transcription start site, consistent with the location of stalled Pol II at heat shock genes[3,4,13] (**Fig. 3b**). Because this is an average profile, it is possible that a fraction of Pol II occupancy comes from inactive preinitiation complexes. However, the majority of detected Pol II signal seems to come from Pol II that is stalled downstream of the transcription start site. Third, Pol II stalling at these genes is consistent with comprehensive expression analysis using whole-genome tiling arrays[22]. Genes with Pol II tightly restricted to the transcription start site are either silent or only weakly expressed in *Toll*^{10b} mutants (**Fig. 3c**). In contrast, genes with similar levels of Pol II binding but uniform distribution throughout the transcription unit are expressed at substantial levels in these mutants (**Fig. 3c**). Finally, we used permanganate footprint assays as an independent method to confirm stalled Pol II at selected genes[4,12]. For example, the *rho* gene showed clear permanganate sensitivity downstream of the transcription start site (+37 bp), consistent with the Pol II stalling profile seen in *Toll*^{10b} mutants (**Fig. 3d** and **Fig. 1b**).

There are considerable differences in the expression and functions of genes in the active, stalled or no Pol II classes based on *in situ* expression patterns (ImaGO database)[25] and functional annotations (Gene Ontology database)[26] (**Fig. 4**). The set of genes with stalled Pol II is highly enriched for developmentally regulated genes, particularly those expressed in ectodermal and neuronal precursor cells (**Fig. 4a**).

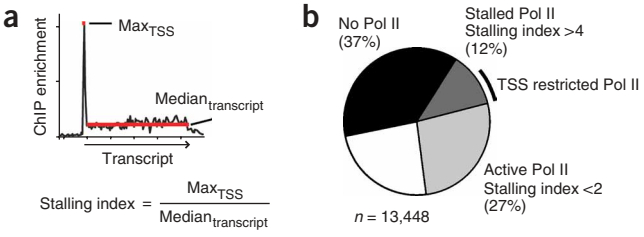


Figure 2 Whole-genome analysis of Pol II binding. (**a**) Genes were assigned to one of three classes—stalled Pol II, active Pol II or no Pol II—on the basis of their stalling index. The stalling index is the ratio between the maximum enrichment near the transcription start site (Max_{TSS}; ±300 bp) and the median enrichment of the probes distributed across the transcription unit ($Median_{transcript}$; excluding the first 600 bp). Stalling index values of >4 qualified as 'stalled Pol II', whereas stalling index values <2 qualified as 'uniform (active) Pol II'. If no probe within the TSS region was significantly bound, the gene was assigned to the 'no Pol II' category. (**b**) Over 76% of all protein-coding genes could be assigned to one of three categories on the basis of stalling index values: 12% have stalled Pol II, 27% show the active form of Pol II and 37% have no Pol II. Among the genes with stalled Pol II, 62% have Pol II tightly restricted to the transcription start site.

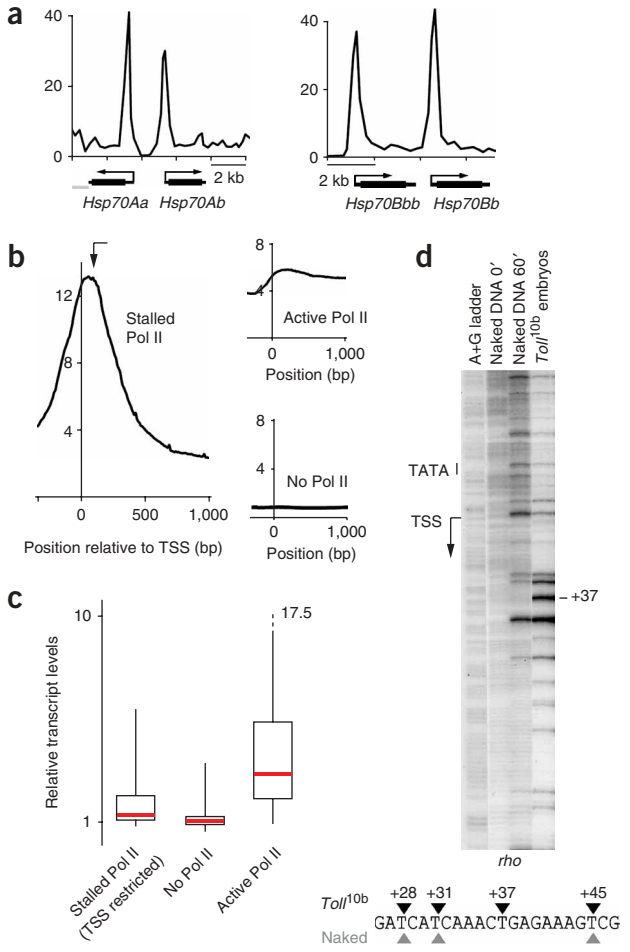


Figure 3 Confirmation of the class of genes with Pol II stalling. (**a**) The group of stalled genes contains all known heat shock genes where Pol II stalling has been well documented. (**b**) Metagene analysis shows that the average peak of 'stalled Pol' II is ~50 bp downstream of transcription (arrow). The profile for 'no Pol II' and 'active Pol II' is shown as a comparison. (**c**) Analysis of the transcript levels confirms that genes with the stalled Pol II profile (tightly restricted to the transcription site) are either silent or expressed at low levels. Genes that show Pol II enrichment at comparable levels throughout the transcription unit (active Pol II) are expressed at significantly higher levels. Genes with no Pol II are not expressed. The transcript levels are represented as box and whiskers plot of the fold ratios (measured by whole-genome tiling arrays[22]). The box represents the 25th and 75th percentiles, with the median as a red bar. The whiskers refer to the 1st and 99th percentiles. The scale on the *y* axis is a log scale. (**d**) A permanganate footprint assay of the *rho* gene confirms stalled Pol II downstream of the transcription start site. Genomic sequences of A+G are shown as marker (lane 1). In comparison with purified genomic DNA, which was either not treated (lane 2) or treated (lane 3) with $KMnO_4$, a prominent hypersensitive T residue is detected in *Toll*^{10b} mutant embryos (lane 4), implying the existence of a transcription bubble at the region around +37 *in vivo*. The bottom panel shows actual sequences from +26 to +47 of the *rho* locus (relative to TSS as +1 bp).

Consistent with these results, genes with stalled Pol II are highly enriched for functions in development, including neurogenesis, ectoderm development and muscle differentiation (**Fig. 4b**). Many of these genes encode sequence-specific transcription factors (Hox, T-box, bHLH, zinc fingers and HMG) and components of cell signaling pathways (FGF, Wnt, Notch, EGF, TGFβ, JNK and TNF) (**Supplementary Note**).

In contrast, the set of genes with uniform Pol II binding is highly enriched for ubiquitously expressed genes (**Fig. 4a**), which function mostly in metabolism and cell proliferation (**Fig. 4b**). The set of genes that lacks Pol II binding is highly enriched in genes that show no staining in whole-embryo *in situ* hybridizations, confirming that they are not expressed during early embryogenesis (**Fig. 4a**). Many of these genes encode proteins that have functions in adult cells, such as cuticle proteins or proteins required for vision (**Fig. 4b**).

Pol II stalling could reflect two nonexclusive developmental functions. It could be indicative of active transcriptional repression, or it could prepare genes for activation at later stages of embryogenesis. The second model is particularly attractive, because Pol II stalling has already been shown to prepare heat shock genes for rapid induction[3,13]. We found evidence for both models.

Pol II stalling is particularly prevalent among genes expressed in the neuroectoderm and dorsal ectoderm, which are repressed in *Toll*^{10b} embryos. To test whether Pol II stalling is specific for repressed genes, we examined the Pol II profile of these genes in mutant embryos

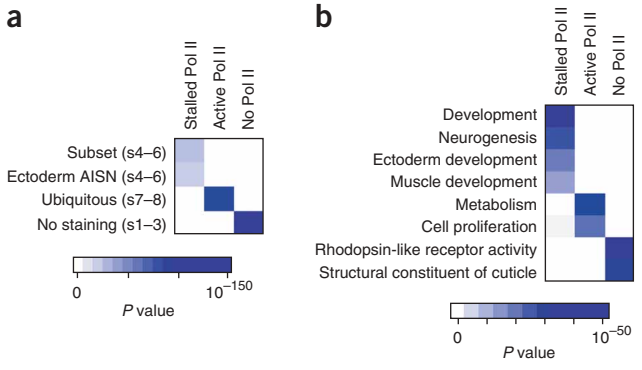


Figure 4 Functional analysis of the three classes of genes. (**a**,**b**) Representative categories of enrichment among genes with stalled Pol II, active Pol II or no Pol II for gene sets are shown for the ImaGO database (**a**), which contains the *in situ* expression patterns of a substantial fraction of all protein coding genes in the *Drosophila* genome, and the gene sets from the biological process categories in the Gene Ontology (GO) database (**b**). The scale bars below indicate the significance of each test (hypergeometric distribution). Genes containing stalled Pol II are significantly enriched for genes expressed in a subset of cells and in those of developing ectoderm (AISN = *anlage in statu nascendi*) at the time of the analysis (120–240 min). The stage of the enriched categories is indicated in parenthesis (s1–3 = ~0–100 min, s4–6 = ~100–200 min, s7–8 = ~200–250 min). The genes expressed as subset (s4–6) are largely identical to those of the ectoderm AISN (s4–6) category, because mesoderm, neurectoderm and dorsal ectoderm are specified at that stage. Functional analysis confirms that genes with stalled Pol II are enriched for genes involved in development, in particular those required for neurogenesis, ectoderm differentiation and muscle development. Genes showing active Pol II are enriched for genes that are ubiquitously expressed in developing embryos. They are enriched for functions that mediate cell proliferation and metabolic functions such as protein and nucleotide metabolism. Genes lacking Pol II tend to be inactive during embryogenesis, and they are deployed at later stages of the life cycle, such as cuticle function and vision.

© 2007 Nature Publishing Group http://www.nature.com/naturegenetics

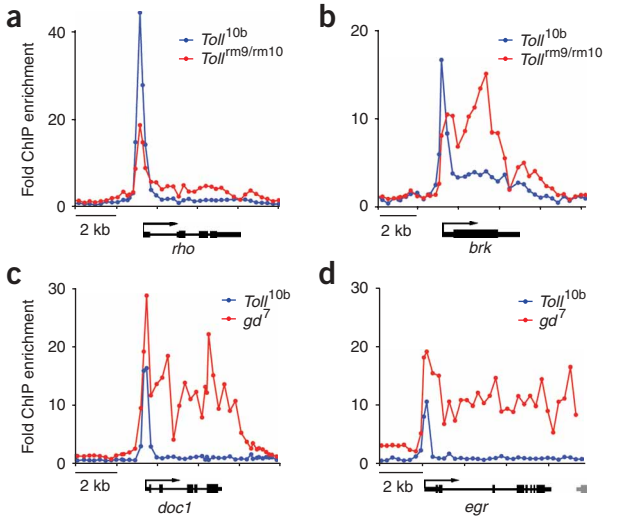


Figure 5 Pol II profile comparison of genes in the repressed versus active state. (**a**–**d**) Pol II ChIP enrichment ratios are shown for genes that are repressed in *Toll*^{10b} embryos but are active in either *Toll*$^{rm9/rm10}$ embryos (**a**,**b**) or *gd*^{7} embryos (**c**,**d**) at 2–4 h after fertilization. The enrichment ratios of Pol II are shown on the *y* axis: active state (red) and repressed state (blue). The results show that the degree of Pol II stalling is dependent on the gene's activity, with some genes showing a complete switch between the stalled and active forms of Pol II.

in which they are active. For this, we used two well-defined mutants, *Toll*$^{rm9/rm10}$ and *gd*^{7} (2–4 h), in which cells adopt neurecto-dermal and dorsal ectodermal fates, respectively[17,21,22,27]. Indeed, at these genes, Pol II is redistributed into the transcription unit in these mutants (**Fig. 5**), and some genes now show the active Pol II profile (**Fig. 5b–d**). These results indicate that Pol II stalling is associated with cell-type specific repression and is subject to dynamic changes during development.

Previous studies have shown that the repression of a large set of genes in *Toll*^{10b} embryos depends on snail, a well-studied repressor that is constitutively expressed in *Toll*^{10b} embryos but not in *Toll*$^{rm9/rm10}$ and *gd*^{7} embryos[18,19,21–23,28]. We found a statistically significant association between repression by snail and Pol II stalling. For example, among the 139 genes that are occupied by snail[23] and show reduced expression in the *Toll*^{10b} mutant[21], 54% have stalled Pol II, whereas only 19% of all genes with reduced expression show Pol II stalling ($P < 10^{-23}$, **Supplementary Note**). This suggests that Pol II stalling in *Toll*^{10b} embryos may be regulated by snail. A role of developmental repressors in regulating Pol II stalling is also consistent with a recent study of *Drosophila* segmentation[12].

Multiple lines of evidence suggest that Pol II stalling also occurs at genes that are poised for activation in older embryos. Genes with stalled Pol II are highly over-represented among genes that are rapidly induced within 12 h after the time frame of our analysis ($P < 10^{-27}$, **Supplementary Note**). Moreover, genes with stalled Pol II are enriched for genes expressed in the derivatives of the mesoderm precursors present in *Toll*^{10b} mutants, such as the developing heart and muscle cells ($P < 10^{-15}$, **Supplementary Note**). These genes, such as Drop (*Dr*) and *bap*, are not yet activated at the time frame of the analysis[29,30], but they nonetheless show high levels of Pol II near the transcription start site (**Supplementary Note**).

To confirm that muscle genes indeed show stalled Pol II before activation, we carried out permanganate assays on wild-type *Drosophila* embryos at 2–4 h after fertilization (**Fig. 6**). *Dr* and *lbe* showed a clear permanganate footprint downstream of transcription. These footprints were specific to the early embryo stage, as S2 cells, a cell line derived from older embryos, did not show a permanganate footprint under the same conditions (**Fig. 6**). These results confirm that Pol II stalling is dynamically regulated and suggest that one of its functions is to prepare genes for activation.

Our genome-wide analysis showed that genes in *Drosophila* embryos are found in three distinct dynamic states: active, stalled or no Pol II. Stalled Pol II is particularly associated with developmental genes that are repressed and poised for activation. We propose that Pol II stalling prepares genes for rapid response to developmental signals during embryogenesis and thus may represent a key regulatory step for gene transcription in development.

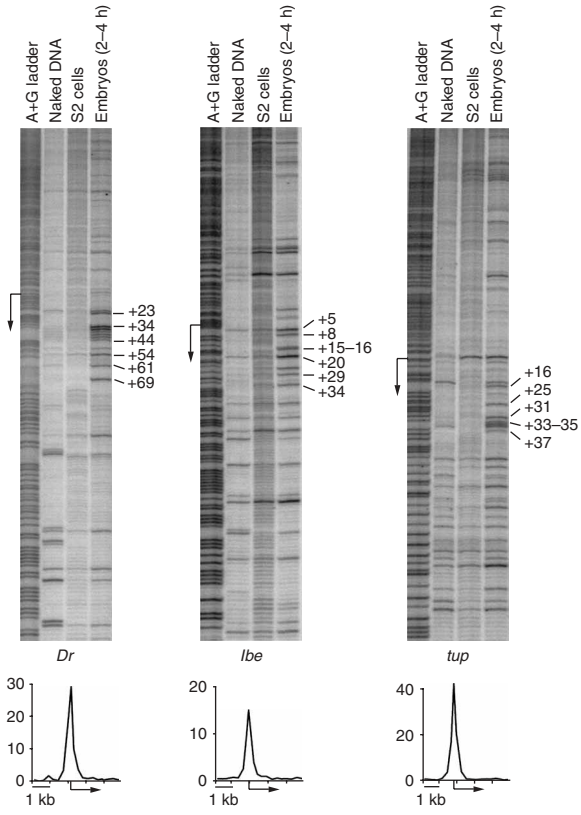


Figure 6 Pol II stalling at genes before activation. *In vivo* permanganate footprints were carried out on the genes *Dr* and ladybird *(lbe)*, which show a strong 'stalled Pol II' profile in ChIP assays (lower panel) and are activated at later stages. Results for *tup,* a gene known to be repressed at that stage, are shown for comparison. The permanganate footprints were carried out on early wild-type embryos (2–4 h) and S2 cells (107 cells), a cell line derived from older embryos. Positions of A+G (lane 1) and the transcription start site (TSS, arrow) are shown for orientation. T residues sensitive to $KMnO_4$ treatment are shown for naked DNA as control (lane 2), S2 cells (lane 3) and early embryos (lane 4). Hypersensitive T residues in the early embryo sample (selected positions marked on the right) indicate an open transcription bubble. The footprints found at the muscle regulatory genes *Dr* and *lbe* are similar to that seen for *tup*. Thus, Pol II is found in a stalled form at the *Dr* and *lbe* genes before activation during embryonic development.

© 2007 Nature Publishing Group http://www.nature.com/naturegenetics

METHODS

Drosophila stock. *Toll*^{10b} is a dominant gain-of-function mutation of the maternal gene *Toll* (ref. 17). Embryos were collected from *Toll*$^{10b/+}$ females obtained directly from the balanced stock (*Toll*^{10b}*/TM3 Sb Ser* and *Toll*^{10b}*/OR60*). *Toll*^{rm9} and *Toll*^{rm10} are recessive *Toll* mutations[17], and embryos were collected from *Toll*$^{rm9/rm10}$ females. *gd*^{7} embryos were collected from *gd*$^{7/7}$ females[27]. Wild-type embryos were *white*^{1118}.

ChIP-chip assays. The chromatin immunoprecipitation (ChIP) experiments coupled to microarrays (chip) were carried out as described[23].

Antibodies. The antibodies against Pol II were 8WG16 and H14 (see **Supplementary Note**).

Arrays. We used *Drosophila* whole-genome tiling arrays (Agilent) as described[23]. Probes of 60-mers span the entire eukaryotic portion of the *Drosophila melanogaster* genome. Although the spacing of these probes is ~280 bp on average, an additional probe is present between the two probes that flank each known TSS. Thus, the resolution around transcriptional start sites is ~140 bp. The data are available from ArrayExpress and our website, http://web.wi.mit.edu/young/pol2/.

Data analysis. We used the Rosetta error model to control for noise at probes; thus a probe required a P value <0.001. We did not use our previous algorithms for detecting bound probes and then assigning genes. Rather, we calculated parameters indicating Pol II enrichment directly for each gene (see **Supplementary Note**). A combination of Pol II enrichment at the start site and median enrichment across the gene were used to classify Pol II as either absent, stalled or active.

Statistics. For the identification of gene sets that are over-represented in the three classes of genes, we used the hypergeometric distribution test (see **Supplementary Note**).

Permanganate footprint assays. We carried out transcription bubble assays with $KMnO_4$ as described previously[4,12]. Embryos were collected 2–4 h after egg deposition (AED), dechorionated and partially homogenized before treatment with $KMnO_4$. Embryos were treated with 20 mM or 40 mM $KMnO_4$ for 60 s on ice. The transcription start sites of the examined genes were identified and confirmed using ESTs in FlyBase and previous expression analysis using tiling arrays[22]. The linker primers and gene-specific primers used for ligation-mediated PCR are listed in the **Supplementary Note**.

Accession numbers. ArrayExpress: Data have been deposited with accession code E-TABM-322.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We would like to thank R. Zinzen for collecting the *Toll*^{10b}, *Toll*$^{rm9/rm10}$ and *gd*^{7} embryos, T. Volkert and J. Love for microarray experimental support and members of the Young laboratory for critical reading of the manuscript. This research was supported in part by the Intramural Research Program of the US National Institutes of Health, National Institute of Environmental Health Sciences (K.A.), US National Institutes of Health grants HG002668 and GM069676 to R.A.Y. and GM34431 to M.L., a grant by the Moore Foundation and a postdoctoral fellowship by the Human Frontier Science Program Organization (HFSPO, A.S.).

AUTHOR CONTRIBUTIONS

J.Z. and M.L. designed the experiment. J.Z. designed the arrays and carried out the experiments and analysis. A.S., M.K. and J.Z. analyzed expression data and functional categories. J.-W.H., S.N. and K.A. carried out the permanganate footprint assays. J.Z., M.L. and R.A.Y. prepared the manuscript.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturegenetics/.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Ptashne, M. & Gann, A. Transcriptional activation by recruitment. *Nature* **386**, 569–577 (1997).
2. Conaway, J.W., Shilatifard, A., Dvir, A. & Conaway, R.C. Control of elongation by RNA polymerase II. *Trends Biochem. Sci.* **25**, 375–380 (2000).
3. Saunders, A., Core, L.J. & Lis, J.T. Breaking barriers to transcription elongation. *Nat. Rev. Mol. Cell Biol.* **7**, 557–567 (2006).
4. Gilmour, D.S. & Lis, J.T. RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in *Drosophila melanogaster* cells. *Mol. Cell. Biol.* **6**, 3984–3989 (1986).
5. Bender, T.P., Thompson, C.B. & Kuehl, W.M. Differential expression of c-myb mRNA in murine B lymphomas by a block to transcription elongation. *Science* **237**, 1473–1476 (1987).
6. Strobl, L.J. & Eick, D. Hold back of RNA polymerase II at the transcription start site mediates down-regulation of c-myc in vivo. *EMBO J.* **11**, 3307–3314 (1992).
7. Krumm, A., Meulia, T., Brunvand, M. & Groudine, M. The block to transcriptional elongation within the human c-myc gene is determined in the promoter-proximal region. *Genes Dev.* **6**, 2201–2213 (1992).
8. Laspia, M.F., Wendel, P. & Mathews, M.B. HIV-1 Tat overcomes inefficient transcriptional elongation in vitro. *J. Mol. Biol.* **232**, 732–746 (1993).
9. Cheng, C. & Sharp, P.A. RNA polymerase II accumulation in the promoter-proximal region of the dihydrofolate reductase and gamma-actin genes. *Mol. Cell. Biol.* **23**, 1961–1967 (2003).
10. Aida, M. *et al.* Transcriptional pausing caused by NELF plays a dual role in regulating immediate-early expression of the junB gene. *Mol. Cell. Biol.* **26**, 6094–6104 (2006).
11. Radonjic, M. *et al.* Genome-wide analyses reveal RNA polymerase II located upstream of genes poised for rapid response upon *S. cerevisiae* stationary phase exit. *Mol. Cell* **18**, 171–183 (2005).
12. Wang, X., Lee, C., Gilmour, D.S. & Gergen, J.P. Transcription elongation controls cell fate specification in the *Drosophila* embryo. *Genes Dev.* **21**, 1031–1036 (2007).
13. Lis, J. & Wu, C. Protein traffic on the heat shock promoter: parking, stalling, and trucking along. *Cell* **74**, 1–4 (1993).
14. Reppas, N.B., Wade, J.T., Church, G.M. & Struhl, K. The transition between transcriptional initiation and elongation in *E. coli* is highly variable and often rate limiting. *Mol. Cell* **24**, 747–757 (2006).
15. Agalioti, T. *et al.* Ordered recruitment of chromatin modifying and general transcription factors to the IFN-beta promoter. *Cell* **103**, 667–678 (2000).
16. Soutoglou, E. & Talianidis, I. Coordination of PIC assembly and chromatin remodeling during differentiation-induced gene activation. *Science* **295**, 1901–1904 (2002).
17. Schneider, D.S., Hudson, K.L., Lin, T.Y. & Anderson, K.V. Dominant and recessive mutations define functional domains of Toll, a transmembrane protein required for dorsal-ventral polarity in the *Drosophila* embryo. *Genes Dev.* **5**, 797–807 (1991).
18. Leptin, M. twist and snail as positive and negative regulators during *Drosophila* mesoderm development. *Genes Dev.* **5**, 1568–1576 (1991).
19. Kosman, D., Ip, Y.T., Levine, M. & Arora, K. Establishment of the mesoderm-neuroectoderm boundary in the *Drosophila* embryo. *Science* **254**, 118–122 (1991).
20. Furlong, E.E., Andersen, E.C., Null, B., White, K.P. & Scott, M.P. Patterns of gene expression during *Drosophila* mesoderm development. *Science* **293**, 1629–1633 (2001).
21. Stathopoulos, A., Van Drenth, M., Erives, A., Markstein, M. & Levine, M. Whole-genome analysis of dorsal-ventral patterning in the *Drosophila* embryo. *Cell* **111**, 687–701 (2002).
22. Biemar, F. *et al.* Comprehensive identification of *Drosophila* dorsal-ventral patterning genes using a whole-genome tiling array. *Proc. Natl. Acad. Sci. USA* **103**, 12763–12768 (2006).
23. Zeitlinger, J. *et al.* Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the *Drosophila* embryo. *Genes Dev.* **21**, 385–390 (2007).
24. Markstein, M., Markstein, P., Markstein, V. & Levine, M.S. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Proc. Natl. Acad. Sci. USA* **99**, 763–768 (2002).
25. Tomancak, P. *et al.* Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biol.* **3**, research0088.1–research0088.14 (2002).
26. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
27. Huang, A.M., Rusch, J. & Levine, M. An anteroposterior Dorsal gradient in the *Drosophila* embryo. *Genes Dev.* **11**, 1963–1973 (1997).
28. Boulay, J.L., Dennefeld, C. & Alberga, A. The *Drosophila* developmental gene snail encodes a protein with nucleic acid binding fingers. *Nature* **330**, 395–398 (1987).
29. Sandmann, T. *et al.* A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. *Dev. Cell* **10**, 797–807 (2006).
30. Furlong, E.E. Integrating transcriptional and signalling networks during muscle development. *Curr. Opin. Genet. Dev.* **14**, 343–350 (2004).

© 2007 Nature Publishing Group http://www.nature.com/naturegenetics

Distinguishing protein-coding and noncoding genes in the human genome

Michele Clamp*†, Ben Fry*, Mike Kamal*, Xiaohui Xie*, James Cuff*, Michael F. Lin‡, Manolis Kellis*‡, Kerstin Lindblad-Toh*, and Eric S. Lander*†§¶‖

*Broad Institute of Massachusetts Institute of Technology and Harvard, 7 Cambridge Center, Cambridge, MA 02142; ¶Department of Biology and ‡Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139; §Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142; and ‖Department of Systems Biology, Harvard Medical School, Boston, MA 02115

Contributed by Eric S. Lander, October 3, 2007 (sent for review August 1, 2007)

Although the Human Genome Project was completed 4 years ago, the catalog of human protein-coding genes remains a matter of controversy. Current catalogs list a total of ≈24,500 putative protein-coding genes. It is broadly suspected that a large fraction of these entries are functionally meaningless ORFs present by chance in RNA transcripts, because they show no evidence of evolutionary conservation with mouse or dog. However, there is currently no scientific justification for excluding ORFs simply because they fail to show evolutionary conservation: the alternative hypothesis is that most of these ORFs are actually valid human genes that reflect gene innovation in the primate lineage or gene loss in the other lineages. Here, we reject this hypothesis by carefully analyzing the nonconserved ORFs—specifically, their properties in other primates. We show that the vast majority of these ORFs are random occurrences. The analysis yields, as a by-product, a major revision of the current human catalogs, cutting the number of protein-coding genes to ≈20,500. Specifically, it suggests that nonconserved ORFs should be added to the human gene catalog only if there is clear evidence of an encoded protein. It also provides a principled methodology for evaluating future proposed additions to the human gene catalog. Finally, the results indicate that there has been relatively little true innovation in mammalian protein-coding genes.

comparative genomics

An accurate catalog of the protein-coding genes encoded in the human genome is fundamental to the study of human biology and medicine. Yet, despite its importance, the human gene catalog has remained an elusive target. The twofold challenge is to ensure that the catalog includes all valid protein-coding genes and excludes putative entries that are not valid protein-coding genes. The latter issue has proven surprisingly difficult. It is the focus of this article.

Putative protein-coding genes are identified based on computational analysis of genomic data—typically, by the presence of an open-reading frame (ORF) exceeding ≈300 bp in a cDNA sequence. The underlying premise, however, is shaky. Recent studies have made clear that the human genome encodes an abundance of non-protein-coding transcripts (1–3). Simply by chance, noncoding transcripts may contain long ORFs. This is particularly so because noncoding transcripts are often GC-rich, whereas stop codons are AT-rich. Indeed, a random GC-rich sequence (50% GC) of 2 kb has a ≈50% chance of harboring an ORF ≈400 bases long [supporting information (SI) Fig. 4].

Once a putative protein-coding gene has been entered into the human gene catalogs, there has been no principled way to remove it. Experimental evidence is of no utility in this regard. Although one can demonstrate the validity of protein-coding gene by direct mass-spectrometric evidence of the encoded protein, one cannot prove the invalidity of a putative protein-coding gene by failing to detect the putative protein (which might be expressed at low abundance or in different tissues or at different developmental stages).

The lack of a reliable way to recognize valid protein-coding transcripts has created a serious problem, which is only growing as large-scale cDNA sequencing projects yield ever-larger numbers of transcripts (2). The three most widely used human gene catalogs [Ensembl (4), RefSeq (5), and Vega (6)] together contain a total of ≈24,500 protein-coding genes. It is broadly suspected that a large fraction of these entries is simply spurious ORFs, because they show no evidence of evolutionary conservation. [Recent studies indicate that only ≈20,000 show evolutionary conservation with dog (7).] However, there is currently no scientific justification for excluding ORFs simply because they fail to show evolutionary conservation; the alternative hypothesis is that these ORFs are valid human genes that reflect gene innovation in the primate lineage or gene loss in other lineages. As a result, the human gene catalog has remained in considerable doubt. The resulting uncertainty hampers biomedical projects, such as systematic sequencing of all human genes to discover those involved in disease.

The situation also complicates studies of comparative genomics and evolution. Current catalogs of protein-coding genes vary widely among mammals, with a recent analysis of the dog genome (8) reporting ≈19,000 genes and a recent article on the mouse genome (2) reporting at least 33,000 genes. The difference is attributable to nonconserved ORFs identified in cDNA sequencing projects. It is currently unclear whether it reflects meaningful evolutionary differences among species or simply varying numbers of spurious ORFs in species with more cDNAs in current databases. In addition, the confusion about protein-coding genes clearly complicates efforts to create accurate catalogs of non-protein-coding transcripts.

The purpose of this article is to test whether the nonconserved human ORFs represent bona fide human protein-coding genes or whether they are simply spurious occurrences in cDNAs. Although it is broadly accepted that ORFs with strong cross-species conservation to mouse or dog are valid protein-coding genes (7), no work has addressed the crucial issue of whether nonconserved human ORFs are invalid. Specifically, one must reject the alternative hypothesis that the nonconserved ORFs represent (*i*) ancestral genes that are present in our common mammalian ancestor but were lost in mouse and dog or (*ii*) novel genes that arose in the human lineage after divergence from mouse and dog.

Here, we provide strong evidence to show that the vast majority of the nonconserved ORFs are spurious. The analysis begins with a thorough reevaluation of a current gene catalog to identify conserved protein-coding genes and eliminate many putative genes resulting from clear artifacts. We then study the remaining set of nonconserved ORFs. By studying their properties in primates, we

Author contributions: M.C. and E.S.L. designed research; M.C., B.F., M. Kamal, X.X., J.C., M.F.L., M. Kellis, K.L.-T., and E.S.L. performed research; M.C., B.F., M. Kamal, X.X., J.C., M.F.L., M. Kellis, K.L.-T., and E.S.L. analyzed data; and M.C. and E.S.L. wrote the paper.

The authors declare no conflict of interest.

†To whom correspondence may be addressed. E-mail: mclamp@broad.mit.edu or lander@broad.mit.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0709013104/DC1.

© 2007 by The National Academy of Sciences of the USA

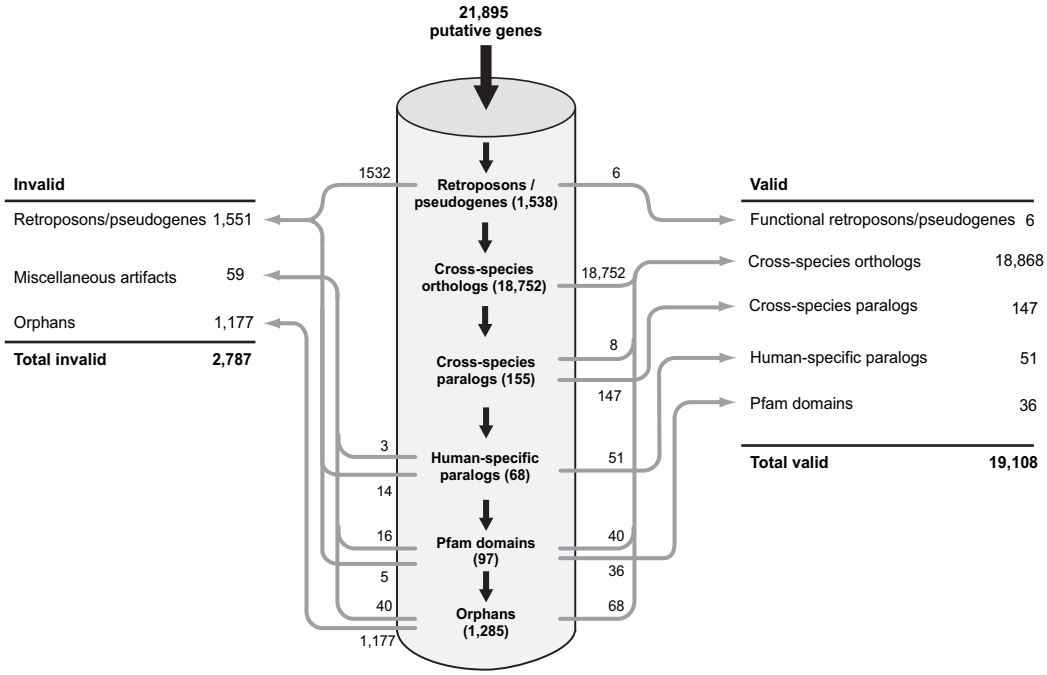


Fig. 1. Flowchart of the analysis. The central pipeline illustrates the computational analysis of 21,895 putative genes in the Ensembl catalog (v35). We then performed manual inspection of 1,178 cases to obtain the tables of likely valid and invalid genes. See text for details.

show that the vast majority are neither (*i*) ancestral genes lost in mouse and dog nor (*ii*) novel genes that arose after divergence from mouse or dog.

The results have three important consequences. First, the analysis yields as a by-product a major revision to the human gene catalog, cutting the number of genes from ≈24,500 to ≈20,500. The revision eliminates few valid protein-coding genes while dramatically increasing specificity. Second, the analysis provides a scientifically valid methodology for evaluating future proposed additions to the human gene catalog. Third, the analysis implies that the mammalian protein-coding genes have been largely stable, with relatively little invention of truly novel genes.

Results

Identifying Orphans. Our analysis requires studying the properties of human ORFs that lack cross-species counterparts, which we term "orphans." Such study requires carefully filtering the human gene catalogs, to identify genes with counterparts and to eliminate a wide range of artifacts that would interfere with analysis of the orphans. For this reason, we undertook a thorough reanalysis of the human gene catalogs.

We focused on the Ensembl catalog (version 35), which lists 22,218 protein-coding genes with a total of 239,250 exons. Our analysis considered only the 21,895 genes on the human genome reference sequence of chromosomes 1–22 and X. (We thus omitted the mitochondrial chromosome, chromosome Y, and "unplaced contigs," which involve special considerations; see below.)

We developed a computational protocol by which the putative genes are classified based on comparison with the human, mouse, and dog genomes (Fig. 1; see *Materials and Methods*). The mouse and dog genomes were used, because high-quality genomic sequence is available (7, 8), and the extent of sequence divergence is well suited for gene identification. The nucleotide substitution rate relative to human is ≈0.50 per base for mouse and ≈0.35 for dog, with insertion and deletion (indel) events occurring at a frequency that is ≈10-fold lower (8, 9). These rates are low enough to allow reliable sequence alignment but high enough to reveal the differential mutation patterns expected in coding and noncoding regions.

After the computational pipeline, we undertook visual inspection of ≈1,200 cases to detect instances misclassifications due to limitations of the algorithms or apparent errors in reported human gene annotations; this process revised the classification of 417 cases. We briefly summarize the results.

Class 0: Transposons, pseudogenes, and other artifacts. Some of the putative genes consist of transposable elements or processed pseudogenes that slipped through the process used to construct the Ensembl catalog. Using a more stringent filter, we identified 1,538 such cases. These were 487 cases consisting of transposon-derived sequence, 483 processed pseudogenes derived from a multiexon parent gene (recognizable because the introns had been eliminated by splicing), and 568 processed pseudogenes derived from a single-exon parent gene (recognizable because the pseudogene sequence almost precisely interrupts the aligned orthologous sequence of human with mouse or dog).

Class 1: Genes with cross-species orthologs. We next identified putative genes with a corresponding gene in the syntenic region of mouse or dog. We examined the orthologous DNA sequence in each species, checking whether an orthologous gene was already annotated in current gene catalogs for mouse or dog and, if not, whether we could identify an orthologous gene. Such cases are referred to as "simple orthology" (or 1:1 orthology). We then expanded the search to a surrounding region of 1 Mb in mouse and dog to allow for cases of local gene family expansion. Such cases are referred to as "complex orthology" (or "coorthology"). In both circumstances, the orthologous gene was required to have an ORF that aligns to a substantial portion (≥80%) of the human gene and have substantial peptide identity (≥50% for mouse, ≥60% for dog). Orthologous genes were identified for 18,752 of the putative human genes, with 16,210 involving simple orthology and 2,542 involving coorthology.

Class 2: Genes with cross-species paralogs. The pipeline then identified 155 cases of putative human genes that have a paralog within the human genome, that, in turn, has an ortholog in mouse or dog. These genes largely represent nonlocal duplications in the human lineage (three-quarters lie in segmental duplications) or possibly gene losses in the other lineages. Among these genes, close inspec-

GENETICS

tion revealed eight cases in which a small change to the human annotation allowed the identification of a clear human ortholog.

Class 3: Genes with human-only paralogs. The pipeline identified 68 cases of putative human genes that have one or more paralogs within the human genome, but with none of these paralogs having orthologs in mouse or dog. Close inspection eliminated 17 cases as additional retroposons or other artifacts (see *SI Appendix*). The remaining 51 cases appear to be valid genes, with 15 belonging to three known families of primate-specific genes (DUF1220, NPIP, and CDRT15 families) and the others occurring in smaller paralogous groups (two to eight members) that may also represent primate-specific families.

Class 4: Genes with Pfam domains. The pipeline identified 97 cases of putative genes with homology to a known protein domain in the Pfam collection (10). Close inspection eliminated 21 cases as additional retroposons or other artifacts (see *SI Appendix*) and 40 cases in which a small change to the human annotation allowed the identification of a clear human ortholog. The remaining 36 genes appear to be valid genes, with 10 containing known primate-specific domains and 26 containing domains common to many species.

Class 5: Orphans. A total of 1,285 putative genes remained after the above procedure. Close inspection identified 40 cases that were clear artifacts (long tandem repeats that happen to lack a stop codon) and 68 cases in which a cross-species ortholog could be assigned after a small change correction to the human gene annotation. The remaining 1,177 cases were declared to be orphans, because they lack orthology, paralogy, or homology to known genes and are not obvious artifacts. We note that the careful review of the genes was essential to obtaining a "clean" set of orphans for subsequent analysis.

Characterizing the Orphans. We characterized the properties of the orphans to see whether they resemble those seen for protein-coding genes or expected for randoms ORFs arising in noncoding transcripts.

ORF lengths. The orphans have a GC content of 55%, which is much higher than the average for the human genome (39%) and similar to that seen in protein-coding genes with cross-species counterparts (53%). The high-GC content reflects the orphans' tendency to occur in gene-rich regions.

We examined the ORF lengths of the orphans, relative to their GC-content. The orphans have relatively small ORFs (median = 393 bp), and the distribution of ORF lengths closely resembles the mathematical expectation for the longest ORF that would arise by chance in a transcript-derived form human genomic DNA with the observed GC-content (SI Fig. 4).

Conservation properties. We then focused on cross-species conservation properties. To assess the sensitivity of various measures, we examined a set of 5,985 "well studied" genes defined by the criterion that they are discussed in more than five published articles. For each well studied gene, we selected a matched random control sequence from the human genome, having a similar number of "exons" with similar lengths, a similar proportion of repeat sequence and a similar proportion of cross-species alignment, but not overlapping with any putative genes.

The well studied genes and matched random controls differ with respect to all conservation properties studied (SI Fig. 5 and SI Table 1). The nucleotide identity and Ka/Ks ratio clearly differ, but the distributions are wide and have substantial overlap. The indel density has a tighter distribution: 97.3% of well studied genes, but only 2.8% of random controls, have an indel density of <10 per kb. The sharpest distinctions, however, were found for two measures that reflect the distinctive evolution of protein-coding genes: the reading frame conservation (RFC) score and the codon substitution frequency (CSF) score.

Reading frame conservation. The RFC score reflects the percentage of nucleotides (ranging from 0% to 100%) whose reading frame is conserved across species (SI Fig. 6). The RFC score is determined by aligning the human sequence to its cross-species ortholog and calculating the maximum percentage of nucleotides with conserved reading frame, across the three possible reading frames for the ortholog. The results are averaged across sliding windows of 100 bases to limit propagation of local effects due to errors in sequence alignment and gene boundary annotation. We calculated separate RFC scores relative to both the mouse and dog genomes and focused on a joint RFC score, defined as the larger of two scores. The RFC score was originally described in our work on yeast, but has been adapted to accommodate the frequent presence of introns in human sequence (see *SI Appendix*).

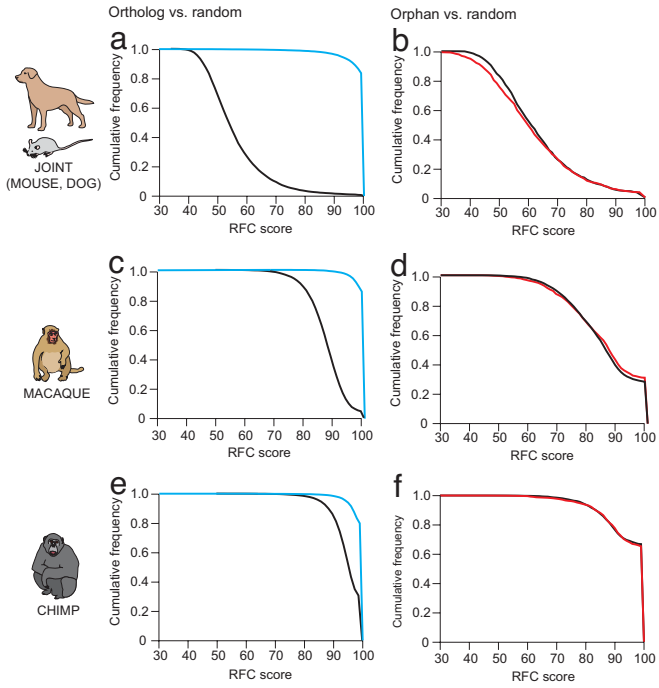


Fig. 2. Cumulative distributions of RFC score. (*Left*) Human genes with cross-species orthologs (blue) versus matched random controls (black). (*Right*) Human orphans (red) versus matched random controls (black). RFC scores are calculated relative to mouse and dog together (*Top*), macaque (*Middle*) and chimpanzee (*Bottom*). In all cases, the orthologs are strikingly different from their matched random controls, whereas the orphans are essentially indistinguishable from their matched random controls.

The RFC score shows virtually no overlap between the well studied genes and the random controls (SI Fig. 5). Only 1% of the random controls exceed the threshold of RFC >90, whereas 98.2% of the well studied genes exceed this threshold. The situation is similar for the full set of 18,752 genes with cross-species counterparts, with 97% exceeding the threshold (Fig. 2*a*). The RFC score is slightly lower for more rapidly evolving genes, but the RFC distribution for even the top 1% of rapidly evolving genes is sharply separated from the random controls (SI Fig. 5).

By contrast, the orphans show a completely different picture. They are essentially indistinguishable from matched random controls (Fig. 2*b*) and do not resemble even the most rapidly evolving subset of the 18,572 genes with cross-species counterparts. In short, the set of orphans shows no tendency whatsoever to conserve reading frame.

Codon substitution frequency. The CSF score provides a complementary test of for the evolutionary pattern of protein-coding genes. Whereas the RFC score is based on indels, the CSF score is based on the different patterns of nucleotide substitution seen in protein-coding vs. random DNA. Recently developed for comparative genomic analysis of *Drosophila* species (11), the method calculates a codon substitution frequency (CSF) score based on alignments

across many species. We applied the CSF approach to alignments of human to nine mammalian species, consisting of high-coverage sequence (≈7×) from mouse, dog, rat, cow, and opossum and low-coverage sequence (≈2×) from rabbit, armadillo, elephant, and tenrec.

The results again showed strong differentiation between genes with cross-species counterparts and orphans. Among 16,210 genes with simple orthology, 99.2% yielded CSF scores consistent with the expected evolution of protein-coding genes. By contrast, the 1,177 orphans include only two cases whose codon evolution pattern indicated a valid gene. Upon inspection, these two cases were clear errors in the human gene annotation; by translating the sequence in a different frame, a clear cross-species orthologs can be identified.

Orphans Do Not Represent Protein-Coding Genes. The results above are consistent with the orphans being simply random ORFs, rather than valid human protein-coding genes. However, consistency does not constitute proof. Rather, we must rigorously reject the alternative hypothesis.

Suppose the orphans represent valid human protein-coding genes that lack corresponding ORFs in mouse and dog. The orphans would fall into two classes: (*i*) some may predate the divergence from mouse and dog—that is, they are ancestral genes that were lost in both mouse and dog, and (*ii*) some may postdate the divergence—that is, they are novel genes that arose in the lineage leading to the human. How can we exclude these possibilities? Our solution was to study two primate relatives: macaque and chimpanzee. We consider the alternatives in turn.

1. Suppose that the orphans are ancestral mammalian genes that were lost in dog and mouse but are retained in the lineage leading to human. If so, they would still be present and functional in macaque and chimpanzee, except in the unlikely event that they also underwent independent loss events in both macaque and chimpanzee lineages.
2. Suppose that the orphans are novel genes that arose in the lineage leading to the human, after the divergence from dog and mouse [≈75 million years ago (Mya)]. Assuming that the generation of new genes is a steady process, the birthdates should be distributed across this period. If so, most of the birthdates will predate the divergence from macaque (≈30 Mya) and nearly all will predate the divergence from chimpanzee (≈6 Mya) (12).

Under either of the above scenarios, the vast majority of the orphans must correspond to functional protein-coding genes in macaque or chimpanzee.

We therefore tested whether the orphans show any evidence of protein-coding conservation relative to either macaque or chimpanzee, using the RFC score. Strikingly, the distribution of RFC scores for the orphans is essentially identical to that for the random controls (Fig. 2 *d* and *f*). The distribution for the orphans does not resemble that seen even for the top 1% of most rapidly evolving genes with cross-species counterparts (SI Figs. 7–9).

The set of orphans thus shows no evidence whatsoever of reading-frame conservation even in our closest primate relatives. (It is of course possible that the orphans include a few valid protein-coding genes, but the proportion must be small enough that it has no discernable effect on the overall RFC distribution.) We conclude that the vast majority of orphans do not correspond to functional protein-coding genes in macaque and chimpanzee, and thus are neither ancestral nor newly arising genes.

If the orphans represent valid human protein-coding genes, we would have to conclude that the vast majority of the orphans were born after the divergence from chimpanzee. Such a model would require a prodigious rate of gene birth in mammalian lineages and a ferocious rate of gene death erasing the huge number of genes born before the divergence from chimpanzee. We reject such a model as wholly implausible. We thus conclude that the vast majority of orphans are simply randomly occurring ORFs that do not represent protein-coding genes.

Finally, we note that the careful filtering of the human gene catalog above was essential to the analysis above, because it eliminated pseudogenes and artifacts that would have prevented accurate analysis of the properties of the orphans.

Experimental Evidence of Encoded Proteins. As an independent check on our conclusion, we reviewed the scientific literature for published articles mentioning the orphans to determine whether there was experimental evidence for encoded proteins. Whereas the vast majority of the well studied genes have been directly shown to encode a protein, we found articles reporting experimental evidence of an encoded protein *in vivo* for only 12 of 1,177 orphans, and some of these reports are equivocal (SI Table 2). The experimental evidence is thus consistent with our conclusion that the vast majority of nonconserved ORFs are not protein-coding. In the handful of cases where experimental evidence exists or is found in the future, the genes can be restored to the catalog on a case-by-case basis.

Revising the Human Gene Catalogs. With strong evidence that the vast majority of orphans are not protein-coding genes, it is possible to revise the human gene catalogs in a principled manner.

Ensembl catalog. Our analysis of the Ensembl (v35) catalog indicates that it contains 19,108 valid protein-coding genes on chromosomes 1–22 and X within the current genome assembly. The remaining 15% of the entries are eliminated as retroposons, artifacts or orphans. Together with the mitochrondrial chromosome [well known to contain 13 protein-coding genes (13)] and chromosome Y [for which careful analysis indicates 78 protein-coding genes (14)], the total reaches 19,199.

We extended the analysis to the Ensembl (v38) catalog, in which 2,212 putative genes were added and many previous entries were revised or deleted. Our computational pipeline found 598 additional valid protein-coding genes based on cross-species counterparts, 1,135 retroposons, and 479 orphans. The RFC curves for the orphans again closely matched the expectation for random DNA.

Other catalogs. We applied the same approach to the Vega (v34) and RefSeq (March 2007) catalog. Both catalogs contain a substantial proportion of entries that appear not to be valid protein-coding genes (16% and 10%, respectively), based on the lack of a cross-species counterpart (see SI Fig. 10 and *SI Appendix*). If we restrict the RefSeq entries to those with the highest confidence (with the caveat that this set contains many fewer genes), only 1% appear invalid. Together, these two catalogs add an additional 673 protein-coding genes.

Combined analysis. Combining the analysis of the three major gene catalogs, we find that only 20,470 of the 24,551 entries appear to be valid protein-coding genes.

Limitations on the Analysis. Our analysis of the current gene catalogs has certain limitations that should be noted.

First, we eliminated all pseudogenes and orphans. We found six reported cases in which a processed pseudogene or transposon underwent exaptation to produce a functional gene (SI Tables 1 and 3) and 12 reported cases of orphans with experimental evidence for an encoded protein. These 18 cases can be readily restored to the catalog (raising the count to 20,488). There are additional cases of potentially functional retroposons that are not present in the current gene catalogs (15). If any are found to produce protein, they should also be included.

Second, we have not considered the 197 putative genes that lie in the "unmapped contigs." These regions are sequences that were omitted from the finished assembly of the human genome. They largely consist of segmental duplications, and most of the genes are highly similar to others in the assembly. Many of the sequence may

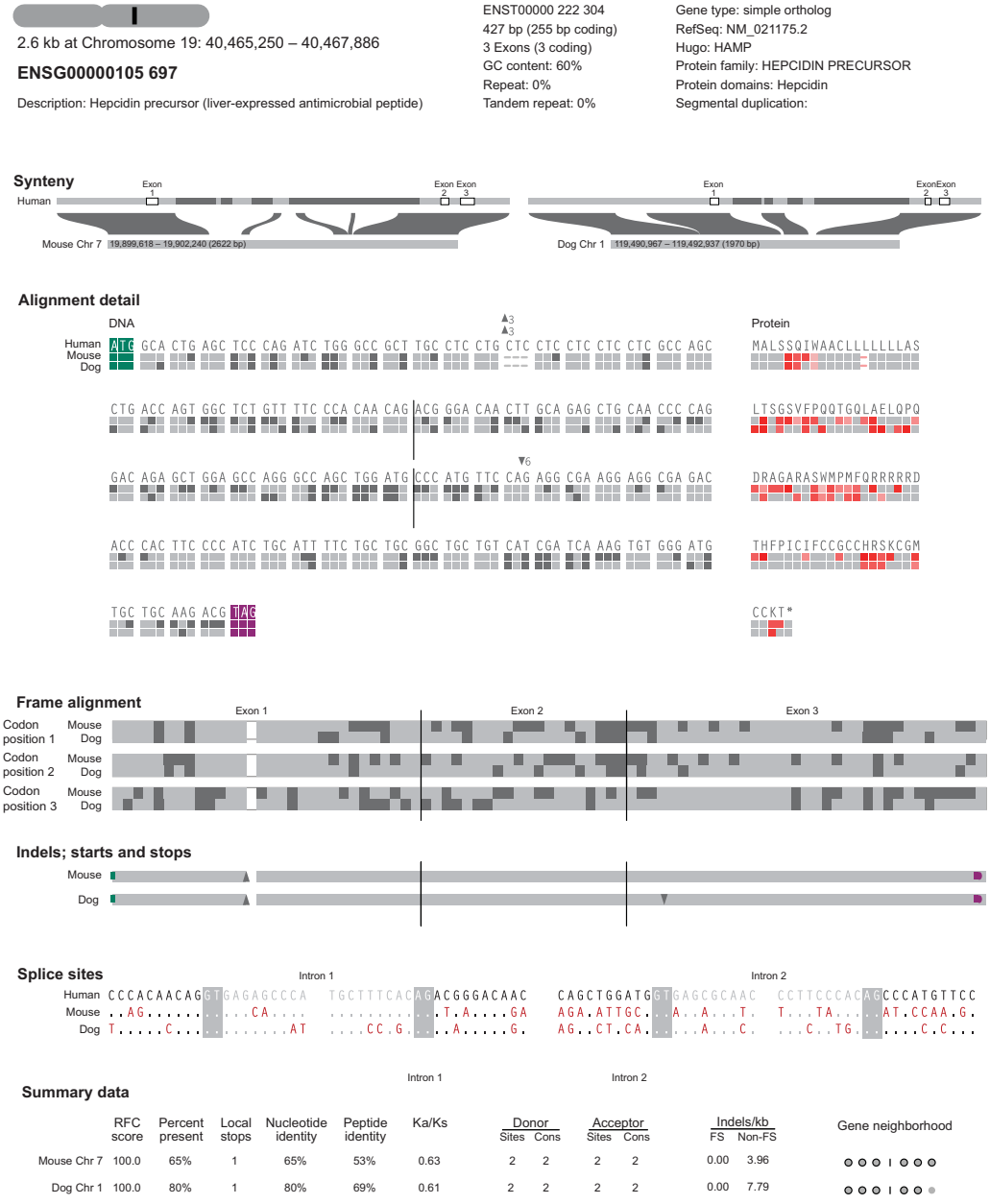


Fig. 3. An example gene report card for a small gene, HAMP, on chromosome 19. Report cards for all 22,218 putative genes in Ensembl v35 are available at www.broad.mit.edu/mammals/alpheus. The report cards provide a visual framework for studying cross-species conservation and for spotting possible problems in the human gene annotation. Information at the top shows chromosomal location; alternative identifiers; and summary information, such as length, number of exons, and repeat content. Various panels below provide graphical views of the alignment of the human gene to the mouse and dog genomes. ''Synteny'' shows the large-scale alignment of genomic sequence, indicating both aligned and unaligned segments. The human sequence is annotated with the exons in white and repetitive sequence in dark gray. ''Alignment detail'' shows the complete DNA sequence alignment and protein alignment. In the DNA alignment, the human sequence is given at the top, bases in the other species are marked as matching (light gray) or nonmatching (dark gray), exon boundaries are marked by vertical lines, indels are marked by small triangles above the sequence (vertex down for insertions, vertex up for deletions, number indicating length in bases), the annotated start codon is in green, and the annotated stop codon is in purple. In the protein alignment, the human amino acid sequence is given at the top, and the sequences in the other species are marked as matching (light gray), similar (pink), or nonmatching (red). ''Frame alignment'' shows the distribution of nucleotide mismatches found in each codon position, with excess mutations expected in the third position. Matching are shown in light gray, and mismatches are shown in dark gray. ''Indels, starts and stops'' provides an overview of key events. Indels are indicated by triangles (vertex down for insertions, vertex up for deletions) and marked as frameshifting (red) or frame-preserving (gray). Start codons are marked in green and stop codons in purple. ''Splice sites'' shows sequence conservation around splice sites, with two-base donor and acceptor sites highlighted in gray and mismatching bases indicated in red. ''Summary data'' lists various conservation statistics relative to mouse and dog, including RFC score, nucleotide identity, number of conserved splice sites, frameshifting and nonframeshifting indel density/kb, and gene neighborhood. The gene neighborhood shows a dot for the three upstream and downstream genes, which is colored gray if synteny is preserved and red otherwise.

represent alternative alleles or misassemblies of the genome. However, regions of segmental duplication are known to be nurseries of evolutionary innovation (16) and may contain some valid genes. They deserve focused attention.

Third and most importantly, the nonconserved ORFs studied here were typically included in current gene catalogs because they have the potential to encode at least 100 amino acids. We thus do not know whether our conclusions would apply to much shorter ORFs. In principle, there exist many additional protein-coding genes that encode short proteins, such as peptide hormones, which are usually translated from much larger precursors and may evolve rapidly. It should be possible to investigate the properties of smaller ORFs by using additional mammalian species beyond mouse and dog.

Improving Gene Annotations. In the course of our work, we generated detailed graphical "report cards" for each of the 22,218 putative genes in Ensembl (v35). The report cards show the gene structure, sequence alignments, measures of evolutionary conservation, and our final classification (Fig. 3).

The report cards are valuable for studying gene evolution and for refining gene annotation. By examining local anomalies by cross-species comparison, we have identified 23 clear errors in gene annotation (including cases in which changing the reading frame or coding strand reveals unambiguous cross-species orthologs) and

332 cases in which cross-species conservation suggests altering the start or stop codon, eliminating an internal exon, or moving a splice site. Of these latter cases, most are likely to be errors in the human gene annotation, although some may represent true cross-species differences. The report cards, together with search tools and summary tables, are available at www.broad.mit.edu/mammals/alpheus.

Discussion

The analysis here addresses an important challenge in genomics—determining whether an ORF truly encodes a protein. We show that the vast majority of ORFs without cross-species counterparts are simply random occurrences. The exceptions appear to represent a sufficiently small fraction that the best course is would be consider such ORFs as noncoding in the absence of direct experimental evidence.

We propose that it is time to undertake a thorough revision of the human gene catalogs by applying this principle to filter the entries. Specifically, we propose that nonconserved ORFs should be included in the human gene catalog if there is clear experimental evidence of an encoded protein. We report here an initial attempt to apply this principle, resulting in a catalog with 20,488 genes.

Our focus has been on excluding putative genes from the human catalogs. We have not explored whether there are additional protein-coding genes that have not yet been included, although it is clear that cross-species analysis can be helpful in identifying such genes. Preliminary analysis from our own group and others suggests that there may be a few hundred additional protein-coding genes to be found but that the final total is likely to remain under ≈21,000. The largest open question concerns very short peptides, which may still be seriously underestimated.

One important biological implication of our results is that truly novel protein-coding genes (encoding at least 100 amino acids) arise only rarely in mammalian lineages. With the current gene catalogs, there are only 168 "human-specific" genes (<1% of the total; only 11 are manually reviewed entries in RefSeq; see SI Table 4). These genes lack clear orthologs or paralogs in mouse and dog, but are recognizable because they belong to small paralogous families within the human genome (2 to 9 members) or contain Pfam domains homologous to other proteins. These paralogous families shows a range of nucleotide identities, consistent with their having arisen over the course of ≈75 million years since the divergence from the mouse lineage. In fact, many of these 168 genes are not entirely novel inventions: One-third show strong similarity to mouse or dog genes across at least 50% of their length; although this falls short of our threshold for declaring orthologs or paralogs (80%), it is nonetheless substantial. Among the orphans, there are only 12 cases with reported experimental evidence of an encoded protein. These cases, which comprise ≈0.06% of the gene catalog, have similar RFC and nucleotide identity scores to neutral sequence and have no similarity with any mouse or dog genes, suggesting these are truly novel inventions. We conclude that mammals thus share largely the same repertoire of protein-coding genes, modified primarily by gene family expansions and contractions.

Finally, the creation of more rigorous catalogs of protein-coding genes for human, mouse, and dog will also aid in the creation of catalogs of noncoding transcripts. This should help propel understanding of these fascinating and potentially important RNAs.

Materials and Methods

All annotations were based on the NCBI35 (hg17) assembly and all genome alignments were taken from the pairwise BLASTZ alignment to mouse assembly NCBI36 (mm4) and dog Broad, Version 1.0 (canFam1; available from http://genome.ucsc.edu). We identified retroposons, using the Ensembl annotation (www.ensembl.org). We then eliminated pseudogenes by identifying transcripts with either retained introns or through interrupted synteny at the boundaries of the transcript. The set of well studied genes were found by using those transcripts whose RefSeq entry contained references to more than five articles. Orthologous genes were identified by using synteny (across >80% of the gene) and peptide identity (>50% for mouse and >60% for dog). The combined RFC score was the highest independent score (taking into account the length of the transcript) for alignments to both mouse and dog. For more details, see *SI Appendix*.

We thank colleagues at the University of California, Santa Cruz, genome browser and the Ensembl genome browser for providing data (BLASTZ alignments, synteny nets, genes, and annotations); L. Gaffney for assistance in preparing the manuscript and figures; S. Fryc and N. Anderson for resequencing data; and a large collection of colleagues around the world for many helpful discussions over the past 3 years that have helped shape and improve this work. This work was supported by the National Institutes of Health National Human Genome Research Institute.

1. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, *et al*. (2005) *Science* 308:1149–1154.
2. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, *et al*. (2005) *Science* 309:1559–1563.
3. ENCODE Project Consortium (2007) *Nature* 447:799–816.
4. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, *et al*. (2007) *Nucleic Acids Res* 35:D610–D617.
5. Pruitt KD, Tatusova T, Maglott DR (2007) *Nucleic Acids Res* 35:D61–D65.
6. Ashurst JL, Chen CK, Gilbert JG, Jekosch K, Keenan S, Meidl P, Searle SM, Stalker J, Storey R, Trevanion S, *et al*. (2005) *Nucleic Acids Res* 33:D459–D465.
7. Goodstadt L, Ponting CP (2006) *PLoS Comput Biol* 2:e133:1134–1150.
8. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ, III, Zody MC, *et al*. (2005) *Nature* 438:803–819.
9. Mouse Genome Sequencing Consortium (2002) *Nature* 420:520–562.
10. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, *et al*. (2006) *Nucleic Acids Res* 34:D247–D251.
11. Lin MF, Carlson JW, Crosby MA, Matthews BB, Yu C, Park S, Wan KH, Schroeder AJ, Gramates LS, St. Pierre SE, *et al*. (2007) *Genome Res*, 10.1101/gr.6679507.
12. Pilbeam D, Young N (2004) *C R Palevol* 3:305–321.
13. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, *et al*. (1981) *Nature* 290:457–465.
14. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, *et al*. (2003) *Nature* 423:825–837.
15. Vinckenbosch N, Dupanloup I, Kaessmann H (2006) *Proc Natl Acad Sci USA* 103:3220–3225.
16. Eichler EE (2001) *Trends Genet* 17:661–669.

Downloaded from www.genesdev.org on January 7, 2008 - Published by Cold Spring Harbor Laboratory Press

RESEARCH COMMUNICATION

A single Hox locus in *Drosophila* produces functional microRNAs from opposite DNA strands

Alexander Stark,[1,2,6,8] Natascha Bushati,[3,6] Calvin H. Jan,[4] Pouya Kheradpour,[1,2] Emily Hodges,[5] Julius Brennecke,[5] David P. Bartel,[4] Stephen M. Cohen,[3,7] and Manolis Kellis[1,9]

[1]Broad Institute of Massachussetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02141, USA; [2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; [3]European Molecular Biology Laboratory, 69117 Heidelberg, Germany; [4]Department of Biology, Howard Hughes Medical Institute and Whitehead Institute for Biomedical Research, Massachusetts Institute of Technology Cambridge, Massachusetts 02139, USA; [5]Watson School of Biological Sciences and Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

MicroRNAs (miRNAs) are ~22-nucleotide RNAs that are processed from characteristic precursor hairpins and pair to sites in messages of protein-coding genes to direct post-transcriptional repression. Here, we report that the miRNA *iab-4* locus in the *Drosophila* Hox cluster is transcribed convergently from both DNA strands, giving rise to two distinct functional miRNAs. Both sense and antisense miRNA products target neighboring Hox genes via highly conserved sites, leading to homeotic transformations when ectopically expressed. We also report sense/antisense miRNAs in mouse and find antisense transcripts close to many miRNAs in both flies and mammals, suggesting that additional sense/antisense pairs exist.

Supplemental material is available at http://www.genesdev.org.

Received September 6, 2007; revised version accepted November 2, 2007.

Hox genes are highly conserved homeobox-containing transcription factors crucial for development in animals (Lewis 1978; for reviews, see McGinnis and Krumlauf 1992; Pearson et al. 2005). Genetic analyses have identified them as determinants of segmental identity that specify morphological diversity along the anteroposterior body axis. A striking conserved feature of Hox complexes is the spatial colinearity between Hox gene transcription in the embryo and the order of the genes along the chromosome (Duboule 1998). Hox clusters also give rise to a variety of noncoding transcripts, including microRNAs (miRNAs) *mir-10* and *mir-iab-4/mir-196*, which derive from analogous positions in Hox clusters in flies and vertebrates (Yekta et al. 2004). miRNAs are ~22-nucleotide (nt) RNAs that regulate gene expression post-transcriptionally (Bartel 2004). They are transcribed as longer precursors and processed from characteristic pre-miRNA hairpins. In particular, Hox miRNAs have been shown to regulate Hox protein-coding genes by mRNA cleavage and inhibition of translation, thereby contributing to the extensive regulatory connections within Hox clusters (Mansfield et al. 2004; Yekta et al. 2004; Hornstein et al. 2005; Ronshaugen et al. 2005). Several Hox transcripts overlap on opposite strands, providing evidence of extensive antisense transcription, including antisense transcripts for *mir-iab-4* in flies (Bae et al. 2002) and its mammalian equivalent *mir-196* (Mainguy et al. 2007). However, the function of these transcripts has been elusive. Here we show that the *iab4* locus in *Drosophila* produces miRNAs from opposite DNA strands that can regulate neighboring Hox genes via highly conserved sites. We provide evidence that such sense/antisense miRNA pairs are likely employed in other contexts and a wide range of species.

Results and Discussion

Our examination of the antisense transcript that overlaps *Drosophila mir-iab-4* revealed that the reverse complement of the *mir-iab-4* hairpin folds into a hairpin reminiscent of miRNA precursors (Fig. 1A). Moreover, 17 sequencing reads from small RNA libraries of *Drosophila* testes and ovaries mapped uniquely to one arm of the *iab-4* antisense hairpin (Fig. 1B). All reads were aligned at their 5′ end, suggesting that the *mir-iab-4* antisense hairpin is processed into a single mature miRNA in vivo, which we refer to as miR-iab-4AS. For comparison, we found six reads consistent with the known miR-iab-4-5p (or miR-iab-4 for short) and one read for its star sequence (miR-iab-4-3p). Interestingly, the relative abundance of mature miRNAs and star sequences for *mir-iab-4AS* (17:0) and *mir-iab-4* (6:1) reflects the thermodynamic asymmetry of the predicted miRNA/miRNA* duplexes (Khvorova et al. 2003; Schwarz et al. 2003). Because they derived from complementary near palindromes, miR-iab-4 and miR-iab-4AS had high sequence similarity, only differing in four positions at the 3′ region (Fig. 1B). However, they differed in their 5′ ends, which largely determine miRNA target spectra (Brennecke et al. 2005; Lewis et al. 2005): miR-iab-4AS was shifted by 2 nt, suggesting targeting properties distinct from those of miR-iab-4 and other known *Drosophila* miRNAs.

We confirmed robust transcription of *mir-iab-4* sense and antisense precursors by in situ hybridization to *Drosophila* embryos (Fig. 1C). Both transcripts were detected in abdominal segments in the posterior part of the embryo, but intriguingly in nonoverlapping domains. As described previously (Bae et al. 2002; Ronshaugen et al. 2005), *mir-iab-4* sense was expressed highly in abdominal segments A5–A7, showing modulation in levels within the segments: *abdominal-A* (*abd-A*)-expressing cells (Fig. 1D; Karch et al. 1990; Macias et al. 1990) ap-

[*Keywords*: *Drosophila*; miR-iab-4; Hox; antisense miRNAs]
[6]**This authors contributed equally to this work.**
[7]**Present address: Temasek Life Sciences Laboratory, The National University of Singapore, Singapore 117604.**
Corresponding authors.
[8]**E-MAIL alex.stark@mit.edu; FAX (617) 253-7512.**
[9]**E-MAIL manoli@mit.edu; FAX (617) 253-7512.**
Article is online at http://www.genesdev.org/cgi/doi/10.1101/gad.1613108.

Downloaded from www.genesdev.org on January 7, 2008 - Published by Cold Spring Harbor Laboratory Press

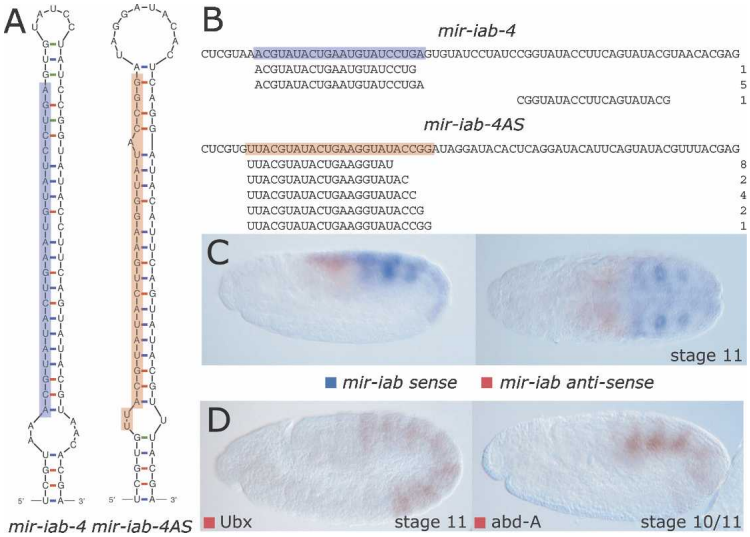


Figure 1. *Drosophila iab-4* contains sense and antisense miRNAs. (*A*) *mir-iab-4* sense and antisense sequences can adopt fold-back stem–loop structures characteristic for miRNA precursors (structure predictions by Mfold [Zuker 2003]; mature miRNAs shaded in blue [miR-iab-4] and red [miR-iab-4AS]). (*B*) Solexa sequencing reads that uniquely align to the *mir-iab-4* hairpin sequence (*top*) or its reverse complement (*bottom*; numbers on the *right* indicate the cloning frequency for each sequence). The mature miRNAs have very similar sequences that are shifted by 2 nt and are different in only four additional positions. (*C*) Expression of primary transcripts for *mir-iab-4* (blue) and *mir-iab-4AS* (red) in nonoverlapping abdominal segments determined by in situ hybridization (lateral [*left* panel] and dorsal [*right* panel] view of embryonic stage 11, anterior is to the left). (*D*) Lateral views of stage 10/11 embryos in which *Ubx* and *abd-A* proteins are visualized (anterior is to the left, and dorsal is upwards).

peared to have more *mir-iab-4*, whereas *Ultrabithorax* (*Ubx*)-positive cells appeared to have little or none (Fig. 1D; Ronshaugen et al. 2005). In contrast, *mir-iab-4AS* transcription was detected in the segments A8 and A9, where *Abdominal-B* (*Abd-B*) is known to be expressed (Fig. 1C; Yoder and Carroll 2006). Primary transcripts for *mir-iab-4* and *mir-iab-4AS* were also detected by strand-specific RT–PCR in larvae, pupae, and male and female adult flies (Supplemental Fig. S1), suggesting that both miRNAs are expressed throughout fly development.

To assess the possible biological roles of the two *iab-4* miRNAs, we examined fly genes for potential target sites by searching for conserved matches to the seed region of the miRNAs (Lewis et al. 2005). We found highly conserved target sites for miR-iab-4AS in the 3′ untranslated regions (UTRs) of several Hox genes that are proximal to the *iab-4* locus and are expressed in the neighboring more anterior embryonic segments: *abd-A*, *Ubx*, and *Antennapedia* (*Antp*) have four, five, and two seed sites, respectively, most of which are conserved across 12 *Drosophila* species that diverged 40 million years ago (Fig. 2A; Supplemental Fig. S2; *Drosophila* 12 Genomes Consortium 2007; Stark et al. 2007a). More than two highly conserved sites for one miRNA is exceptional for fly 3′ UTRs, placing these messages among the most confidently predicted miRNA targets and suggesting that they might be particularly responsive to the presence of the miRNA. The strong predicted targeting of proximal Hox genes was reminiscent of previously characterized miR-iab-4 targeting of *Ubx* in flies and miR-196 targeting of *HoxB8* in vertebrates (Mansfield et al. 2004; Yekta et al. 2004; Hornstein et al. 2005; Ronshaugen et al. 2005).

To test whether miR-iab4AS is functional and can directly target *abd-A* and *Ubx*, we constructed Luciferase reporters carrying the corresponding wild-type 3′ UTRs and control 3′ UTRs in which each seed site was disrupted by point substitutions. *mir-iab-4AS* potently repressed reporter activity for *abd-A* and *Ubx* (Fig. 2B). This repression was specific to the miR-iab-4AS seed sites, as expression of the control reporters with mutated sites was not affected. We also tested whether *mir-iab-4AS* reduced expression of a Luciferase reporter with the *Abd-B* 3′ UTR, which has no seed sites. As expected, *mir-iab-4AS* expression did not affect reporter activity, consistent with a model where miRNAs do not target genes that are coexpressed at high levels (Farh et al. 2005; Stark et al. 2005). In addition to demonstrating specific repression dependent on the predicted target sites, these assays confirmed the processing of the *mir-iab-4AS* hairpin into a functional mature miRNA.

If miR-iab-4AS were able to potently down-regulate

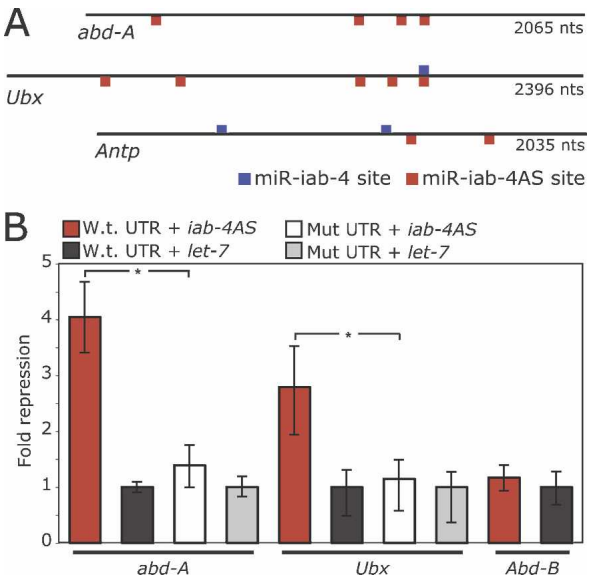


Figure 2. miR-iab-4AS targets neighboring Hox genes. (*A*) miR-iab-4AS has five 3′ UTR seed sites (red) in *Ubx*, four in *abd-A*, and two in *Antp* of which three, four, and one are conserved across 12 *Drosophila* species, respectively (Supplemental Fig. S2). miR-iab-4 has one 3′ UTR seed site (blue) in *Ubx* and two in Antp, while *abd-A* has no such sites. (*B*) miR-iab-4AS mediates repression of luciferase reporters through complementary seed sites in 3′ UTRs from *abd-A* and *Ubx*, but not *Abd-B* (*Antp* was not tested). Luciferase activity in S2 cells cotransfected with plasmid expressing the indicated miRNA with either wild-type luciferase reporters or mutant reporters bearing a single point mutation in the seed. Bars represent geometric means from 16 replicates, normalized to the transfection control and noncognate miRNA control (*let-7*; see Materials and Methods). Error bars represent the fourth largest and smallest values from 16 replicates ([*] $P < 0.0001$, Wilcoxon rank-sum test).

Downloaded from www.genesdev.org on January 7, 2008 - Published by Cold Spring Harbor Laboratory Press

Ubx in the fly, its misexpression should result in a *Ubx* loss-of-function phenotype, a line of reasoning that has often been used to study the functions and regulatory relationships of Hox genes. *Ubx* is expressed throughout the haltere imaginal disc, where it represses wing-specific genes and specifies haltere identity (Weatherbee et al. 1998). When we expressed *mir-iab-4AS* in the haltere imaginal disc under *bx-Gal4* control, a clear homeotic transformation of halteres to wings was observed (Fig. 3). The halteres developed sense organs characteristic of the wing margin and their size increased severalfold, features typical of transformation to wing (Weatherbee et al. 1998). Consistent with the increased number of miR-iab4AS target sites, the transformation was stronger than that reported for expression of *iab-4* (Ronshaugen et al. 2005), for which we confirmed changes in morphology but did not find wing-like growth (Fig. 3D).

We conclude that both strands of the *iab-4* locus are expressed in nonoverlapping embryonic domains and that each transcript produces a functional miRNA in vivo. In particular, the novel *mir-iab-4AS* is able to strongly down-regulate neighboring Hox genes. Interestingly, vertebrate *mir-196*, which lies at an analogous position in the vertebrate Hox clusters, is transcribed in the same direction as *mir-iab-4AS* and most other Hox genes, and targets homologs of both *abd-A* and *Ubx* (Mansfield et al. 2004; Yekta et al. 2004; Hornstein et al. 2005). With its shared transcriptional orientation and homologous targets, *mir-iab-4AS* appears to be the functional equivalent of *mir-196*.

The expression patterns and regulatory connections between Hox genes and the two *iab-4* miRNAs show an intriguing pattern in which the miRNAs appear to reinforce Hox gene-mediated transcriptional regulation (Fig. 4A). In particular, miR-iab-4AS would reinforce the posterior expression boundary of *abd-A*, *Ubx*, and *Antp*, supporting their transcriptional repression by *Abd-B*. *mir-iab-4* appears to support *abd-A*- and *Abd-B*-mediated repression of *Ubx*, reinforcing the *abd-A*/*Ubx* expression domains and the posterior boundary of *Ubx* expression. Furthermore, both *iab-4* miRNAs have conserved target sites in *Antp*, which is also repressed by *Abd-B*, *abd-A*, and *Ubx*. The *iab-4* miRNAs thus appear to support the established regulatory hierarchy among Hox transcription factors, which exhibits "posterior prevalence," in that more posterior Hox genes repress more anterior ones and are dominant in specifying segment identity (for reviews, see McGinnis and Krumlauf 1992; Pearson et al. 2005). Interestingly, *Abd-B* and *mir-iab-4AS* are expressed in the same segments, and the majority of *cis*-regulatory elements controlling *Abd-B* expression are located 3′ of *Abd-B* (Boulet et al. 1991). This places them near the inferred transcription start of *mir-iab-4AS*, where they potentially direct the coexpression of these genes. Similarly, *abd-A* and *mir-iab-4* may be coregulated as both are transcribed divergently, potentially under the control of shared upstream elements.

Our data demonstrate the transcription and processing of sense and antisense *mir-iab-4* into functional miRNAs with highly conserved functional target sites in neighboring Hox genes. In an accompanying study (Bender 2008), genetic and molecular analyses in *mir-iab-4* mutant *Drosophila* revealed that the proposed regulation of *Ubx* by both sense and antisense miRNAs occurs under physiological conditions and, in particular, the regulation by miR-iab-4AS is required for normal development. These lines of evidence establish miR-iab-4AS as a novel Hox gene, being expressed from within the Hox cluster and regulating Hox genes during development.

The genomic arrangement of two miRNAs that are expressed from the same locus but on different strands

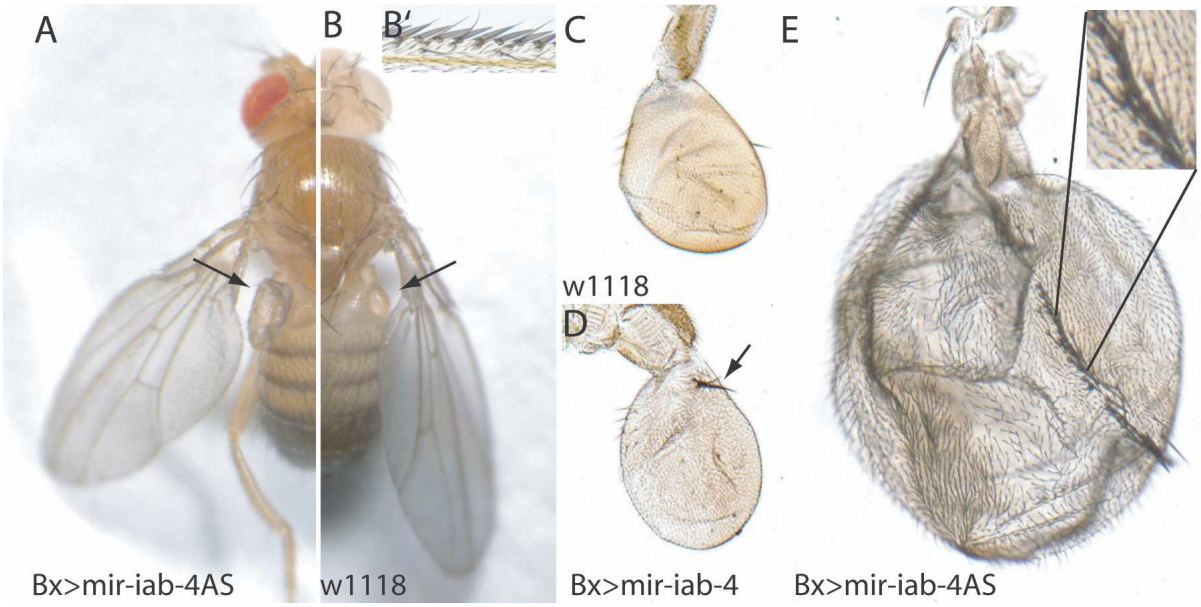


Figure 3. Misexpression of miR-iab-4AS transforms halteres to wings. (*A*,*B*) Overview of an adult wild-type *Drosophila* (*B*) and an adult expressing *mir-iab-4AS* using *bx*-Gal4 (*A*). The halteres, balancing organs of the third thoracic segment, are indicated by arrows. (*C*) Wild-type haltere. (*D*) Expression of *mir-iab-4* using *bx-Gal4* induces a mild haltere-to-wing transformation. Sensory bristles characteristic of wild-type wing margins (shown in *B*′) are indicated by an arrow. (*E*) Expression of *mir-iab-4AS* using *bx-Gal4* induces a strong haltere-to-wing transformation, displaying the triple row of sensory bristles (*inset*) normally seen in wild-type wings (shown in *B*′). Note that *C*–*E* are at the same magnification.

Downloaded from www.genesdev.org on January 7, 2008 - Published by Cold Spring Harbor Laboratory Press

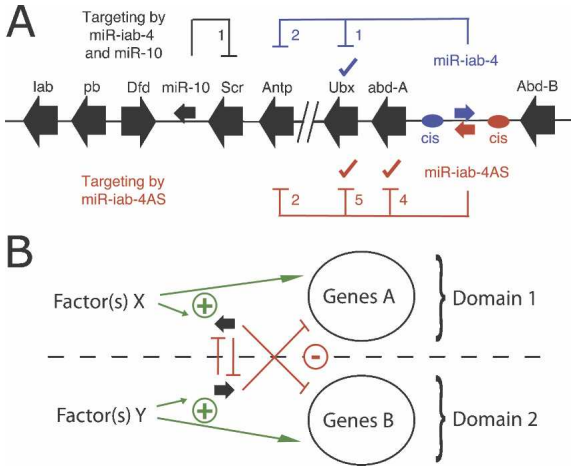


Figure 4. Regulation of gene expression by antisense miRNAs. (*A*) miRNA-mediated control in the *Drosophila* Hox cluster. Schematic representation of the *Drosophila* Hox cluster (Antennapedia and Bithorax complex) with miRNA target interactions (check marks represent experimentally validated targets). miR-iab-4 (blue) and miR-iab-4AS (red) target anterior neighboring Hox genes and miR-10 (black) targets posterior *Sex-combs-reduced* (*Scr*) (Brennecke et al. 2005). *abd-A* and *mir-iab-4* and *Abd-B* and *mir-iab-4AS* might be coregulated from shared control elements (*cis*). Note that *mir-iab-4AS* is expressed in the same direction as most other Hox genes and its mammalian equivalent, *mir-196*. (*B*) General model for defining different expression domains with pairs of antisense miRNAs (black). Different transcription factor(s) activate the transcription of miRNAs and genes in each of the two domains separately (green lines). Both miRNAs might inhibit each other by transcriptional interference or post-transcriptionally (vertical red lines), leading to essentially nonoverlapping expression and activity of both miRNAs. Further, both miRNAs likely target distinct sets of genes (diagonal red lines), potentially re-enforcing the difference between the two expression domains.

might provide a simple and efficient means to create nonoverlapping miRNA expression domains (Fig. 4B). Such sense/antisense miRNAs could restrict each other's transcription, either by direct transcriptional interference, as shown for overlapping convergently transcribed genes (Shearwin et al. 2005; Hongay et al. 2006), or post-transcriptionally, possibly via RNA–RNA duplexes formed by the complementary transcripts. Sense/antisense miRNAs would usually differ at their 5′ ends and thereby target distinct sets of genes, which might help define and establish sharp boundaries between expression domains. Coupled with feedback loops or coregulation of miRNAs and genes in *cis* or *trans*, this arrangement could provide a powerful regulatory switch. The *iab-4* miRNAs might be a special case of tight regulatory integration in which miRNAs and proximal genes appear coregulated transcriptionally in *cis* and repress each other both transcriptionally and post-transcriptionally.

It is perhaps surprising that no antisense miRNA had been found previously, even though, for example, the intriguing expression pattern of the *iab-4* transcripts had been reported nearly two decades ago (Cumberledge et al. 1990; Bae et al. 2002), and *iab-4* lies in one of the most extensively studied regions of the *Drosophila* genome. The frequent occurrence of antisense transcripts (Yelin et al. 2003; Katayama et al. 2005) suggests that more antisense miRNAs might exist. Indeed, up to 13% of known *Drosophila*, 20% of mouse, and 31% of human miRNAs are located in introns of host genes transcribed on the opposite strand or are within 50 nt of antisense ESTs or cDNAs (Supplemental Table S1). These include an antisense transcript overlapping human *mir-196* (see also Mainguy et al. 2007). However, because of the contribution of noncanonical base pairs, particularly G:U pairs that become less favorable A:C in the antisense strand, many miRNA antisense transcripts will not fold into hairpin structures suitable for miRNA biogenesis, which explains the propensity of miRNA gene predictions to identify the correct strand (Lim et al. 2003). Nonetheless, in a recent prediction effort, 22 sequences reverse-complementary to known *Drosophila* miRNAs showed scores seemingly compatible with miRNA processing (Stark et al. 2007b). Deep sequencing of small RNA libraries from *Drosophila* confirmed the processing of small RNAs from four of these high-scoring antisense candidates (Ruby et al. 2007), and the ovary/testes libraries used here showed antisense reads for an additional *Drosophila* miRNA (*mir-312*) (see Supplemental Tables S2, S3). In addition, using high-throughput sequencing of small RNA libraries from mice, we found sequencing reads that uniquely matched the mouse genome in loci antisense to 10 annotated mouse miRNAs. Eight of the inferred antisense miRNAs were supported by multiple independent reads, and two of them had reads from both the mature miRNA and the star sequence (Supplemental Table S2). These results suggest that sense/antisense miRNAs could be more generally employed in diverse contexts and in species as divergent as flies and mammals.

Materials and methods

Plasmids

3′ UTRs were amplified from *Drosophila melanogaster* genomic DNA and cloned in pCR2.1 for site-directed mutagenesis. The following primer pairs were used to amplify the indicated 3′ UTR: abd-A (tctagaGCGGTCAGCAAAGTCAACTC; gtcgacATGGATGGGTTCTCGTTGCAG), Ubx (tctagaATCCTTAGATCCTTAGATCCTTAG; ctcgagATGGTTTGAATTTCCACTGA), and Abd-B (tctagaGCCACCACCTGAACCTTAG; aactcgagCGGAGTAATGCGAAGTAATTG). QuickChange multisite-directed mutagenesis was used to mutate all miR-iab-4AS seed sites from ATACGT to ATAGGT, per the manufacturer's directions (Stratagene). Wild-type and mutated 3′ UTRs were subcloned into pCJ40 between SacI and NotI sites to make *Renilla* luciferase reporters. Plasmid pCJ71 contains the abd-A wild-type 3′ UTR, pCJ72 contains the Ubx wild-type 3′ UTR, pCJ74 contains the Abd-B wild-type 3′ UTR, pCJ75 contains the abd-A mutated 3′ UTR, and pCJ76 contains the Ubx mutated 3′ UTR fused to *Renilla* luciferase. The control *let-7* expression vector was obtained by amplifying *let-7* from genomic DNA with primers 474 base pairs (bp) upstream of and 310 bp downstream from the let-7 hairpin and cloning it into pMT-puro. To express miR-iab-4 and miR-iab-4AS, a 430-bp genomic fragment containing the miR-iab-4 hairpin was cloned, in either direction, downstream from the tubulin promoter as described in Stark et al. (2005). For the UAS-miR-iab-4 and UAS-miR-iab-4AS constructs, the same 430-bp genomic fragment containing the miR-iab-4 hairpin was cloned downstream from pUAST-DSred2 (Stark et al. 2003) in either direction.

Reporter assays

For the luciferase assays, 2 ng of p2129 (firefly luciferase), 4 ng of *Renilla* reporter, 48 ng of miRNA expression plasmid, and 48 ng of p2032 (GFP) were cotransfected with 0.3 µL Fugene HD per well of a 96-well plate. Twenty-four hours after transfection, expression of *Renilla* luciferase was induced by addition of 500 µM $CuSO_4$ to the culture media. Twenty-four hours after induction, reporter activity was measured with the Dual-Glo luciferase kit (Promega), per the manufacturer's instructions on a Tecan Safire II plate reader.

Downloaded from www.genesdev.org on January 7, 2008 - Published by Cold Spring Harbor Laboratory Press

The ratio of *Renilla*:firefly luciferase activity was measured for each well. To calculate fold repression, the ratio of *Renilla*:firefly for reporters cotransfected with *let-7* was set to 1. The Wilcoxon rank-sum test was used to assess the significance of changes in fold repression of wild-type reporters compared with mutant reporters. Geometric means from 16 transfections representing four replicates of four independent transfections are shown. Error bars represent the fourth highest and lowest values of each set.

Drosophila *strains*

UAS-miR-iab-4 and *UAS-miR-iab-4AS* flies were generated by injection of the corresponding plasmids into w^{1118} embryos. bx^{MS1096}*-GAL4* flies were obtained from the Bloomington Stock Center.

In situ hybridization and protein stainings

Double in situ hybridization for the miRNA primary transcripts was performed as described in Stark et al. (2005). Probes were generated using PCR on genomic DNA with primers TCAGAGCATGCAGAGACAT AAAG, TTGTAGATTGAAATCGGACACG for iab-4 sense and ATTT TACTGGGTGTCTGGGAAAG, TAGAAACTGAGACGGAGAAGCAG for iab-4 antisense. Protein stainings were performed as described in Patel (1994). Antibodies used were mouse anti-Ubx (1:30), mouse anti-abd-A (1:5), and HRP-conjugated goat anti-mouse (Dianova, 1:3000).

RT–PCRs

Total RNA was isolated using Trizol (Invitrogen), treated with RQI DNase (Promega), and used for strand-specific cDNA synthesis with SuperScript III (Invitrogen). Primers for cDNA synthesis were CATATAA CAAAGTGCTACGTG (iab-4 sense) and CTTTATCTGCATTTG GATCCG (iab-4 antisense). Both primers were used for subsequent amplification.

Small library sequencing

Drosophila small RNAs were cloned from adult ovaries and testes as described previously (Brennecke et al. 2007) and sequenced using Solexa sequencing. A total of 657,251 sequencing reads uniquely matched known *Drosophila* miRNAs (Rfam release 9.2), and the 69 miRNAs with unique matches had 1011 matches on average (Stark et al. 2007b). Two miRNAs had unique matches to the antisense hairpin (Supplemental Tables S2, S3). Mouse small RNAs were cloned from wild-type and c-kit mutant ovaries (Supplemental Table S4; G. Hannon, pers. comm.) and from Comma-Dbgeo cells, a murine mammary epithelial cell line (Ibarra et al. 2007), and were sequenced using Solexa sequencing. A total of 4,217,883 reads uniquely matched known mouse miRNAs (Rfam release 9.2), and the 286 miRNAs with unique reads showed 256 reads on average. Sequencing reads matching to the plus and minus strand of known mouse miRNAs with antisense reads are listed in Supplemental Table S3.

Multiple sequence alignments and target site prediction

The multiple sequence alignments for the indicated Hox 3′ UTRs were obtained from the University of California at Santa Cruz (UCSC) genome browser (Kent et al. 2002) and were slightly manually adjusted. We predicted target sites according to Lewis et al. (2005) by searching for 3′ UTR seed sites (reverse-complementary to miRNA positions 2–8 or matching to "A" + reverse complement of miRNA positions 2–7).

Antisense transcripts near known miRNAs

To assess the fraction of *Drosophila*, human, and mouse miRNAs that are also putatively transcribed on both strands and might give rise to antisense miRNAs, we determined the number of miRNAs that are near known transcripts on the opposite strand. We obtained the coordinates of all introns of protein-coding genes and all mapped ESTs or cDNAs for the three species from the UCSC genome browser (Kent et al. 2002). We intersected them with the miRNA coordinates from Rfam (release 9.2; Griffiths-Jones et al. 2006), requiring miRNAs and transcripts to be on opposite strands and at a distance of at most 50 nt. For each miRNA, we recorded the number of antisense transcripts and their identifiers. Note that some of the transcripts might have been mapped to more than one place in the genome, such that the intersection represents an upper estimate based on the currently known transcripts.

Acknowledgments

We thank Greg Hannon for providing Solexa sequencing data and support, Juerg Mueller for the anti-Ubx antibody, Thomas Sandmann for *Drosophila* embryos, and Sandra Mueller for preparing transgenic flies. We thank the *Drosophila* genome sequencing centers and the UCSC genome browser for access to the 12 *Drosophila* multiple sequence alignments prior to publication, and Welcome Bender for sharing data prior to publication. A.S. was partly supported by a post-doctoral fellowship from the Schering AG and partly by a post-doctoral fellowship from the Human Frontier Science Program Organization (HFSPO). C.H.J. is an NSF graduate fellow. J.B. thanks the Schering AG for a post-doctoral fellowship. This work was also partially supported by a grant from the NIH.

References

Bae, E., Calhoun, V.C., Levine, M., Lewis, E.B., and Drewell, R.A. 2002. Characterization of the intergenic RNA profile at abdominal-A and Abdominal-B in the *Drosophila* bithorax complex. *Proc. Natl. Acad. Sci.* **99:** 16847–16852.

Bartel, D.P. 2004. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* **116:** 281–297.

Bender, W. 2008. MicroRNAs in the *Drosophila* bithorax complex. *Genes & Dev.* (this issue), doi: 10.1101/gad.1614208.

Boulet, A.M., Lloyd, A., and Sakonju, S. 1991. Molecular definition of the morphogenetic and regulatory functions and the *cis*-regulatory elements of the *Drosophila* Abd-B homeotic gene. *Development* **111:** 393–405.

Brennecke, J., Stark, A., Russell, R.B., and Cohen, S.M. 2005. Principles of microRNA-target recognition. *PLoS Biol.* **3:** e85. doi: 10.1371/journal.pbio.0030085.

Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G.J. 2007. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **128:** 1089–1103.

Cumberledge, S., Zaratzian, A., and Sakonju, S. 1990. Characterization of two RNAs transcribed from the *cis*-regulatory region of the abd-A domain within the *Drosophila* bithorax complex. *Proc. Natl. Acad. Sci.* **87:** 3259–3263.

Drosophila 12 Genomes Consortium 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450:** 203–218.

Duboule, D. 1998. Vertebrate hox gene regulation: Clustering and/or colinearity? *Curr. Opin. Genet. Dev.* **8:** 514–518.

Farh, K.K., Grimson, A., Jan, C., Lewis, B.P., Johnston, W.K., Lim, L.P., Burge, C.B., and Bartel, D.P. 2005. The widespread impact of mammalian microRNAs on mRNA repression and evolution. *Science* **310:** 1817–1821.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., and Enright, A.J. 2006. miRBase: MicroRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **34 (Database issue):** D140–D144. doi: 10.1093/nar/gkj112.

Hongay, C.F., Grisafi, P.L., Galitski, T., and Fink, G.R. 2006. Antisense transcription controls cell fate in *Saccharomyces cerevisiae*. *Cell* **127:** 735–745.

Hornstein, E., Mansfield, J.H., Yekta, S., Hu, J.K., Harfe, B.D., McManus, M.T., Baskerville, S., Bartel, D.P., and Tabin, C.J. 2005. The microRNA miR-196 acts upstream of Hoxb8 and Shh in limb development. *Nature* **438:** 671–674.

Ibarra, I., Erlich, Y., Muthuswamy, S.K., Sachidanandam, R., and Hannon, G.J. 2007. A microRNA fingerprint of mammary epithelial stem cells. *Genes & Dev.* **21:** 3238–3243.

Karch, F., Bender, W., and Weiffenbach, B. 1990. abdA expression in *Drosophila* embryos. *Genes & Dev.* **4:** 1573–1587.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C.C., Suzuki, M., Kawai, J., et al. 2005. Antisense transcription in the mammalian transcriptome. *Science* **309:** 1564–1566.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res.* **12:** 996–1006.

Khvorova, A., Reynolds, A., and Jayasena, S.D. 2003. Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115:** 209–216.

Lewis, E.B. 1978. A gene complex controlling segmentation in *Dro-*

Downloaded from www.genesdev.org on January 7, 2008 - Published by Cold Spring Harbor Laboratory Press

sophila. *Nature* **276:** 565–570.
Lewis, B.P., Burge, C.B., and Bartel, D.P. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120:** 15–20.
Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B., and Bartel, D.P. 2003. The microRNAs of *Caenorhabditis elegans*. *Genes & Dev.* **17:** 991–1008.
Macias, A., Casanova, J., and Morata, G. 1990. Expression and regulation of the abd-A gene of *Drosophila*. *Development* **110:** 1197–1207.
Mainguy, G., Koster, J., Woltering, J., Jansen, H., and Durston, A. 2007. Extensive polycistronism and antisense transcription in the mammalian hox clusters. *PLoS ONE* **2:** e356. doi: 10.1371/journal.pone.0000356.
Mansfield, J.H., Harfe, B.D., Nissen, R., Obenauer, J., Srineel, J., Chaudhuri, A., Farzan-Kashani, R., Zuker, M., Pasquinelli, A.E., Ruvkun, G., et al. 2004. MicroRNA-responsive 'sensor' transgenes uncover Hox-like and other developmentally regulated patterns of vertebrate microRNA expression. *Nat. Genet.* **36:** 1079–1083.
McGinnis, W. and Krumlauf, R. 1992. Homeobox genes and axial patterning. *Cell* **68:** 283–302.
Patel, N.H. 1994. Imaging neuronal subsets and other cell types in whole-mount *Drosophila* embryos and larvae using antibody probes. *Methods Cell Biol.* **44:** 445–487.
Pearson, J.C., Lemons, D., and McGinnis, W. 2005. Modulating Hox gene functions during animal body patterning. *Nat. Rev. Genet.* **6:** 893–904.
Ronshaugen, M., Biemar, F., Piel, J., Levine, M., and Lai, E.C. 2005. The *Drosophila* microRNA iab-4 causes a dominant homeotic transformation of halteres to wings. *Genes & Dev.* **19:** 2947–2952.
Ruby, J.G., Stark, A., Johnston, W.K., Kellis, M., Bartel, D.P., and Lai, E.C. 2007. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of *Drosophila* microRNAs. *Genome Res.* doi: 10.1101/gr.6597907.
Schwarz, D.S., Hutvagner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P.D. 2003. Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115:** 199–208.
Shearwin, K.E., Callen, B.P., and Egan, J.B. 2005. Transcriptional interference—A crash course. *Trends Genet.* **21:** 339–345.
Stark, A., Brennecke, J., Russell, R.B., and Cohen, S.M. 2003. Identification of *Drosophila* microRNA targets. *PLoS Biol.* **1:** E60. doi: 10.1371/journal.pbio.0000060.
Stark, A., Brennecke, J., Bushati, N., Russell, R.B., and Cohen, S.M. 2005. Animal microRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. *Cell* **123:** 1133–1146.
Stark, A., Lin, M.F., Kheradpour, P., Pedersen, J.S., Parts, L., Carlson, J.W., Crosby, M.A., Rasmussen, M.D., Roy, S., Deoras, A.N., et al. 2007a. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* **450:** 219–232.
Stark, A., Kheradpour, P., Parts, L., Brennecke, J., Hodges, E., Hannon, G.J., and Kellis, M. 2007b. Systematic discovery and characterization of fly microRNAs using 12 *Drosophila* genomes. *Genome Res.* doi: 10.1101/gr.6593807.
Weatherbee, S.D., Halder, G., Kim, J., Hudson, A., and Carroll, S. 1998. Ultrabithorax regulates genes at several levels of the wing-patterning hierarchy to shape the development of the *Drosophila* haltere. *Genes & Dev.* **12:** 1474–1482.
Yekta, S., Shih, I.H., and Bartel, D.P. 2004. MicroRNA-directed cleavage of HOXB8 mRNA. *Science* **304:** 594–596.
Yelin, R., Dahary, D., Sorek, R., Levanon, E.Y., Goldstein, O., Shoshan, A., Diber, A., Biton, S., Tamir, Y., Khosravi, R., et al. 2003. Widespread occurrence of antisense transcription in the human genome. *Nat. Biotechnol.* **21:** 379–386.
Yoder, J.H. and Carroll, S.B. 2006. The evolution of abdominal reduction and the recent origin of distinct Abdominal-B transcript classes in *Diptera*. *Evol. Dev.* **8:** 241–251.
Zuker, M. 2003. Mfold Web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31:** 3406–3415.

The evolutionary dynamics of the *Saccharomyces cerevisiae* protein interaction network after duplication

Aviva Presser*†, Michael B. Elowitz‡, Manolis Kellis†§, and Roy Kishony*¶‖

*School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138; †Broad Institute, Cambridge, MA 02142; ‡Division of Biology and Department of Applied Physics, California Institute of Technology, Pasadena, CA 91125; §Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139; and ¶Department of Systems Biology, Harvard Medical School, Boston, MA 02115

Edited by Leonid Kruglyak, Princeton University, Princeton, NJ, and accepted by the Editorial Board November 20, 2007 (received for review August 2, 2007)

Gene duplication is an important mechanism in the evolution of protein interaction networks. Duplications are followed by the gain and loss of interactions, rewiring the network at some unknown rate. Because rewiring is likely to change the distribution of network motifs within the duplicated interaction set, it should be possible to study network rewiring by tracking the evolution of these motifs. We have developed a mathematical framework that, together with duplication data from comparative genomic and proteomic studies, allows us to infer the connectivity of the preduplication network and the changes in connectivity over time. We focused on the whole-genome duplication (WGD) event in *Saccharomyces cerevisiae*. The model allowed us to predict the frequency of intergene interaction before WGD and the post-duplication probabilities of interaction gain and loss. We find that the predicted frequency of self-interactions in the preduplication network is significantly higher than that observed in today's network. This could suggest a structural difference between the modern and ancestral networks, preferential addition or retention of interactions between ohnologs, or selective pressure to preserve duplicates of self-interacting proteins.

gene duplication | network motifs | self-interacting proteins | whole-genome duplication

Complex biological networks result from the evolutionary growth of simpler networks with fewer components. Gene duplication is thought to be a key mechanism by which networks evolve and new components are added (1–6, 43). These duplication events can act on a single gene, a chromosomal segment, or even a whole genome (1, 7–11). After duplication, the duplicate genes may assume one of several fates, including differentiation of sequence and function, or loss of one of the duplicates (12–17, 44). These outcomes are thought to be affected by genetic factors including redundancy, modularization, and expression dosage (9, 12, 15, 18–22, 45).

Little is known about the rules that govern the modification of gene interactions after a duplication event or the effects of gene interaction on the fate of duplicate genes. Here, we report a mathematical framework for inferring the preduplication connectivity properties of a network and for describing its postduplication dynamics. Our method decomposes a protein interaction network into a vector of network motifs and tracks the evolution of this vector over time. We apply our methodology to the protein interaction network of *Saccharomyces cerevisiae* (23–29), which has undergone a whole-genome duplication (WGD) event, resulting in hundreds of coordinately duplicated gene pairs (ohnologs) (8, 9, 11).

Results and Discussion

Network motifs are small subgraphs, or interaction patterns, that occur in networks more frequently than would be expected by chance (30). Motifs have been a valuable tool in identifying functional structure in many biological networks including in transcriptional, neural, and developmental networks (30, 31). We applied the concept of network motifs to WGD genes in *S. cerevisiae* and analyzed network motifs composed of pairs of ohnologs (namely, motifs of interactions within four proteins, Fig. 1*A*). There are six possible interactions between any four proteins, hence 64 possible motifs (2^6). This number is reduced

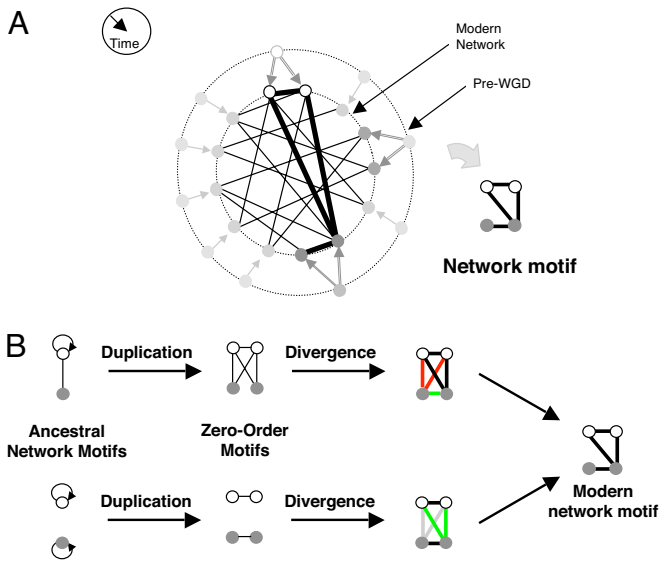


Fig. 1. Whole-genome duplication (WGD) produces network motifs between ohnolog pairs. (*A*) The paths genes take through time after a WGD. In most cases only one of the duplicated genes is retained (light gray). Surviving gene duplicate pairs are present as ohnologs in the modern network (white, dark gray). Interactions between any two pairs of ohnologs form a four-node subgraph (network motif) in the proteome. (*B*) Modern ohnolog motifs are formed through a process of duplication and divergence. Preduplication self-interacting proteins lead to a postduplication interaction between ohnologs. If two ancestral genes interacted, 4 interactions are formed between their pairs of descendants. The duplication step thus yields an initial ohnolog motif (zero-order motifs), which is subsequently modified over time. During the divergence step, interactions might be gained (green) and others are lost (red). Not everything changes: some interactions are retained (black) and other interactions remain absent (gray).

Author contributions: A.P., M.B.E., and R.K. designed research; A.P. performed research; A.P., M.K., and R.K. analyzed data; and A.P., M.B.E., M.K., and R.K. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. L.K. is a guest editor invited by the Editorial Board.

‖To whom correspondence should be addressed. E-mail: roy_kishony@hms.harvard.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0707293105/DC1.

© 2008 by The National Academy of Sciences of the USA

Table 1. Motif distribution in the modern protein interaction network

Motif class no.	Motif class	No. of motifs present in today's yeast proteome	Modern motif frequency (m_{modern})
1		81,983	8.15×10^{-1}
2		17,748	1.76×10^{-1}
3		215	2.13×10^{-3}
4		925	9.16×10^{-2}
5		14	1.39×10^{-4}
6		2	1.98×10^{-5}
7		93	9.21×10^{-4}
8		15	1.48×10^{-4}
9		6	5.94×10^{-5}
10		0	0
11		16	1.58×10^{-4}
12		0	0
13		1	9.90×10^{-6}
14		1	9.90×10^{-6}
15		0	0
16		4	3.96×10^{-5}
17		0	0
18		1	9.90×10^{-6}
19		1	9.90×10^{-6}

to 19 different motif classes after accounting for the symmetry between the motif's ohnolog pairs and the symmetry of the genes within each ohnolog pair [supporting information (SI) Table 3].

The proteins we considered for our motif analysis are the 450 WGD ohnolog pairs, as listed in Kellis *et al.* (8). Interactions between these proteins are listed in the Database of Interacting Proteins (DIP) (23–29). From these data we determined the modern distribution (m_{modern}) of our 19 motif classes (Table 1). We observe a rich variability in motif prevalences. Even for motifs with the same number of interactions, we observed that frequencies vary across several orders of magnitude, indicating that motif frequencies reflect evolutionary processes rather than stochastic effects. We then asked how much of the motif distribution observed today could be explained by a neutral model accounting for the evolutionary dynamics of gene duplication after the WGD event.

We developed a model describing protein connectivity within the subnetwork of surviving ohnologs (Fig. 1*A*) (5, 36). The model consists of two steps: duplication and divergence (Fig. 1*B*). The duplication step assumes that each protein is duplicated along with all its interactions. Because the two daughter proteins are initially identical to each other, the resulting interaction sets are identical. Accordingly, if a protein was self-interacting, each of its duplicates will be self-interacting, and an interaction will

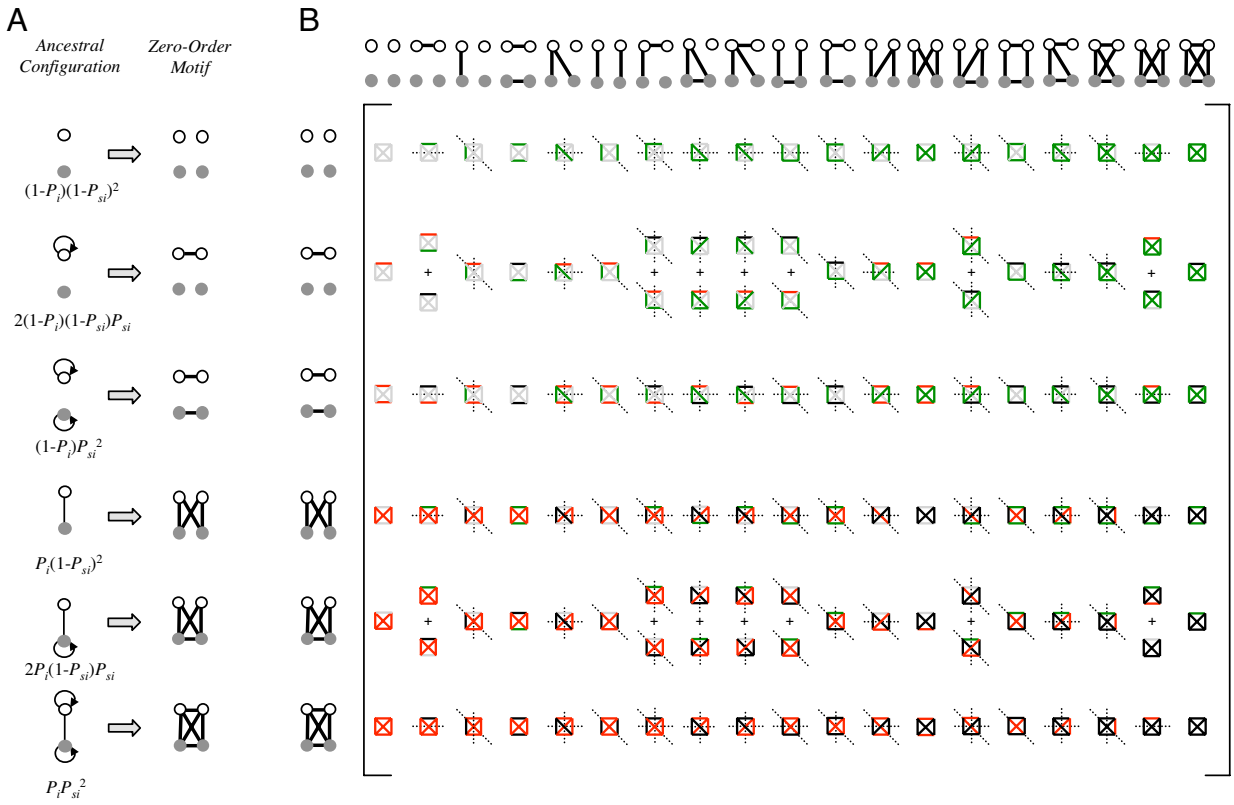


Fig. 2. Ohnolog motif frequencies provide a method for estimating ancestral connectivity and rewiring parameters. (*A*) Immediately after duplication, ohnolog motifs can be one of six zero-order motifs with probability vector m_0 (row vector shown as its transpose). The probabilities of observing each ancestral configuration, and hence each zero-order motif, are listed as functions of the ancestral interaction (P_i) and self-interaction (P_{si}) probabilities. Thirteen of the 19 motifs cannot arise in this fashion, enforcing a strong constraint on the initial conditions of the system. (*B*) The six zero-order motifs can evolve into any one of the 19 possible motifs. The transition probabilities are given by a matrix T, whose entries T_{ij} represent the probability of a member of the motif class in row *i* becoming a member of the motif class in column *j*. This matrix is represented iconographically, with each entry showing the interaction changes necessary to go from one motif to another. Edges are colored as in Fig. 1*B* and symmetry axes are shown by dotted lines. Horizontal and vertical symmetry axes indicate reflections that yield alternative icon-procedures for getting from class *i* to class *j*. A diagonal symmetry axis indicates that exchanging the positions of the vertices in either ohnolog pair yields an alternative icon-procedure for getting from class *i* to class *j* (SI Table 3). The value of each entry is given by $T_{i,j} = 2^S P_+^{n_G} P_-^{n_L} (1 - P_-)^{n_R} (1 - P_+)^{n_A}$, where n_G, n_L, n_R, and n_A represent the number of edges that are gained (green), lost (red), retained (black), or remain absent (gray). The values of the icons in each row sums to 1. As an illustration, the probability of a motif in class becoming a motif of class graphically is that equals $2^2 \cdot P_+^2 \cdot P_-^1 \cdot (1-P_+)^3 \cdot (1-P_-)^0 = 4\, P_+^2\, P_-\, (1-P_+)^3$.

exist between the duplicates. This duplication process can generate only 6 different motifs of the possible 19 (Fig. 2*A*). We term these initial patterns "zero-order motifs," and represent their distribution by a vector, m_0. The frequencies of these zero-order motifs are governed by P_{si} and P_i, defined as the probabilities of protein self-interaction and of interaction between two different proteins in the preduplication network, respectively (Fig. 2*A*).

The second step in the model encompasses the evolutionary dynamics after duplication (1). Mutations leading to the addition or deletion of an interaction are assumed to occur with probabilities P_+ and P_-, respectively. We define these probabilities as describing the overall period from the WGD event until today, accounting for the possibility of multiple rounds of addition and deletion.** We assume that rewiring events are independent, so that the probability of adding or removing multiple interactions is described by the product of the individual probabilities. This rewiring dynamic is described mathematically by a transition matrix (T, Fig. 2*B*) whose elements are the probabilities of evolution from the initial, six-element condition vector, m_0, to an observed, 19-element vector, m_0T. For example, the probability of a motif in class becoming a motif of class is $P_-(1 - P_+)^5$—the probability of losing the one interaction multiplied by the probability of not gaining an interaction at any of the five open positions. The final outcome of duplication and divergence should yield the motif distribution observed today, m_{modern}. We obtain a system of 19 equations, one for each motif class, with four variables: P_i, P_{si}, P_+, and P_-:

$$m_0(P_i,P_{si}) \cdot T(P_+,P_-) = m_{\text{modern}}. \qquad [1]$$

The transition matrix elements are functions of P_+ and P_-,and the initial condition zero-order motif vector m_0 is a function of the preduplication parameters P_i and P_{si}. Because these four parameters are overdetermined by the 19 equations of Eq. **1**, the existence of a solution is not mathematically guaranteed. We solved the equations for the best-fit values of P_i, P_{si}, P_+, and P_- (*Methods* and Table 2). Fig. 3*A* shows that the observed number of motifs is in good agreement with the predictions of the model given the best-fit parameters obtained. This indicates that our simplified model is able to capture much of the complexity of the

**Explicitly, we allow one edge transition per site. This would not include cases where we have multiple transitions at a single site (e.g., is equivalent in our method to). In practice, multiple transitions are improbable, but we define our transitions to include these higher-order transitions for completeness.

Table 2. Best-fit values of preduplication network connectivity and postduplication dynamics inferred from the proteomic network motif distribution of *Saccharomyces cerevisiae*

Parameter	Parameter value ± SD
P_i	0.0023 ± 0.0003
P_{si}	0.25 ± 0.04
P_+	0.0007 ± 0.0001
P_-	0.61 ± 0.03

preduplication network and its rewiring dynamics. Our model is less predictive for some of the motifs, in particular some low-frequency ones (see *SI Text* for further discussion on potential reasons for these outliers). As shown in Table 2, postduplication rewiring of the network involved a high probability of interaction loss, whereas the likelihood of gaining an interaction was small. This result is consistent with previous work (5, 38).

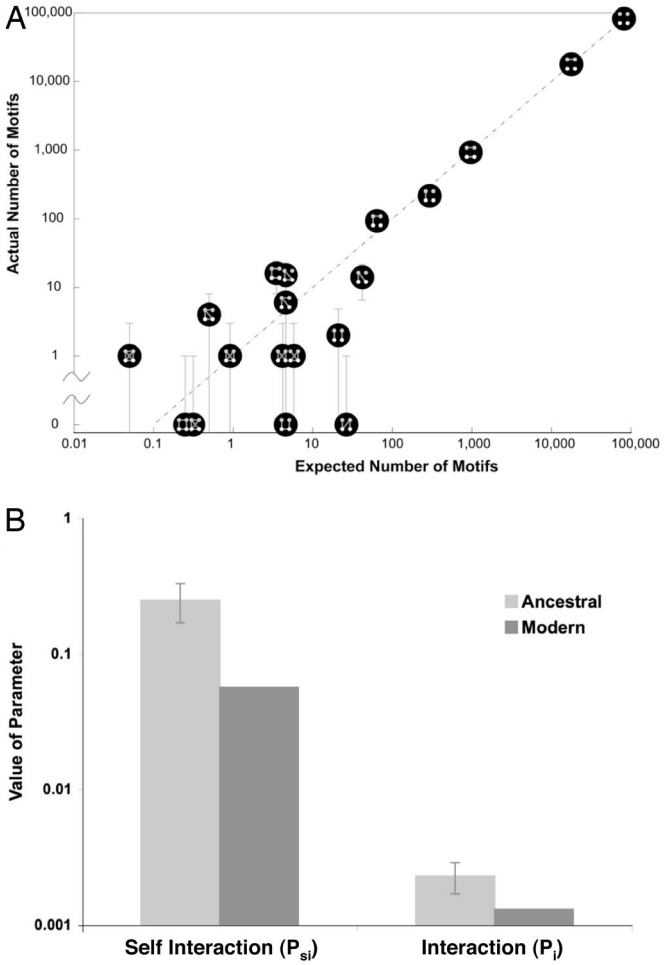


Fig. 3. The modern motif distribution closely resembles the expected distribution. (*A*) We solved our system of 19 equations in 4 unknowns to compute the best-fit network. The expected number of motifs given the best-fit parameters P_i, P_{si}, P_+, and P_- (*x* axis) is plotted against actual motif data from today's *S. cerevisiae* proteome (*y* axis). The error bars in the modern motif distribution are estimated as $\sqrt{p(1-p)}\mathrm{N}$. Best-fit parameter values are listed in Table 2. (*B*) Observed values for P_i and P_{si} in the modern network are compared with the inferred P_i and P_{si} parameters for the ancestral preduplication network. Although the intergene connectivity (P_i) is very similar, the inferred self-interaction frequency (P_{si}) of that network differs by a factor of five from the equivalent modern value. Error estimation is described in *SI Text*. Similar analyses with the database of Batada *et al.* (47) yield consistent results (SI Fig. 4).

We also observe an enrichment of interactions between the ohnologs themselves. Based on the modern frequency of protein interactions (0.13%), we would expect <1 ohnolog pair to interact. We observe 44 interactions of this type (binomial $P \ll 10^{-10}$)—nearly 10% of our ohnologs (see, for example, refs. 18, 32, 33, and 37). This phenomenon translates itself in the context of our model to a high probability of self-interaction in the preduplication network ($P_{si} = 0.25$). This frequency of self-interaction is nearly fivefold higher than observed in the modern value (0.056,[††] Fig. 3*B*).

A simple explanation for this phenomenon is that the ancestral network contained more self-interacting proteins than exist in the modern network and that the ohnolog interactions are descendents of the frequent ancestral self-interactions. This would suggest a structural difference between the ancient and modern proteome. Because a network's structure can reflect its functional capabilities, such a difference might imply unique functional capabilities of the ancestral proteome or potentially proteomic subfunctionalization between the pre- and postduplication organisms (36–38). Alternatively, these ohnologous interactions might be *de novo*. Because overall P_+ is small, this would suggest an evolutionary preference for adding or retaining ohnolog interactions (i.e., $P_{+,ohnolog} > P_{+,nonohnolog}$, or $P_{-,ohnolog} < P_{-,nonohnolog}$) (36).

Another intriguing explanation is that the high estimate for P_{si} results from selective retention of duplicates descended from ancestrally self-interacting proteins. Assuming that self-interactions were not more common in the ancestral network, our data may suggest that these pairs were under selective pressure to be maintained (46). Because they would be retained over long periods of time, they are more likely to have evolved a novel function (22, 38, 49). We suggest a simple dose-dependent model (described in *SI Text*) consistent with the idea that duplicated self-interacting proteins are selectively preserved (39). This could be an important contributor to the evolution of protein complexes (38, 45, 49).

Our model explains the current prevalence of the 19 ohnolog motifs and provides an estimate for pre- and postduplication parameters of the interaction network. The estimated frequency of self-interaction in the ancestral network is significantly higher than in today's network. This could indicate preferential retention of self-interacting protein duplicates, structural differences between the networks, or an inherent asymmetry between ohnologous and nonohnologous protein interaction dynamics. Our results are based on DIP and should be taken with caution because of possible bias and inherent noise associated with the high-throughput data that make up a significant portion of the DIP (23–29, 48). It will be interesting to see whether similar observations appear in other sources of interaction data for *S. cerevisiae* and other species (1, 21, 40, 41).

Methods

Databases. We used the protein interactions listed in the DIP database (23, 26–29). Data can be downloaded at http://dip.doe-mbi.ucla.edu/. The whole-genome duplicates are listed in the supplemental material of Kellis *et al.* (8).

Minimization Algorithm. We solve Eq. **1** for the parameters that best fit the data by minimizing the error associated with the fit. The right hand side, m_{modern}, is directly derived from the data (Table 1). The left hand side, $m_0(P_i,P_{si})\cdot T(P_+,P_-)$ yields a vector m_{expected} that depends on the four parameters P_i, P_{si}, P_+, and P_-. For a motif *i*, the goodness of fit is given by the square of the difference

[††]According to DIP, the dataset on which we base our analysis. In other datasets, this parameter ranges in value, with the largest being 0.138 [large literature-curated dataset (35)].

between the observed abundance $m_{\text{modern},i}$ and the expected abundance $m_{\text{expected},i}$, scaled by the expected number of motifs:

$$E = \sum_i \frac{(m_{\text{modern},i} - m_{\text{expected},i})}{m_{\text{expected},i}}.$$

We then minimize E using the simplex search method (42) implemented by the *fminsearch* function in Matlab, obtaining best-fit values of P_i, P_{si}, P_+, and P_- (see Table 2). The algorithm to estimate the error in the parameters is described in *SI Text*. We tested the model on simulated networks (*SI Text* and SI Table 4) before running on the actual yeast proteome.

ACKNOWLEDGMENTS. We acknowledge N. Barkai, M. Brenner, A. DeLuna, E. Lieberman, I. Nachman, I. Wapinski, and K. Wolfe for their advice and helpful discussions and E. Lieberman and R. Milo for critical readings of the manuscript. This work was supported in part by National Institutes of Health Grants GM068763 (to M.B.E.) and R01GM081617 (to R.K.). A.P. was supported by a National Science Foundation Graduate Fellowship and a National Defense Science and Engineering Graduate Fellowship.

1. Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Segurens B, Daubin V, Anthouard V, Aiach N, *et al.* (2006) *Nature* 444:171–178.
2. Barabasi AL, Albert R (1999) *Science* 286:509–512.
3. Dehal P, Boore JL (2005) *PLoS Biol* 3:e314.
4. Ispolatov I, Krapivksy PL, Mazo I, Yuryev A (2005) *New J Phys* 7:145.
5. Pastor-Satorras R, Smith E, Sole RV (2003) *J Theor Biol* 222:199–210.
6. Hughes AL (1994) *Proc R Soc London Ser B* 256:119–123.
7. Wolfe K (2004) *Curr Biol* 14:R392–R394.
8. Kellis M, Birren BW, Lander ES (2004) *Nature* 428:617–624.
9. Langkjaier RB, Cliften PF, Johnston M, Piskur J (2003) *Nature* 421:848–852.
10. Ohno S (1970) *Evolution by Gene Duplication* (Allen and Unwin, London).
11. Wolfe KH, Shields DC (1997) *Nature* 387:708–713.
12. Conant GC, Wolfe KH (2006) *PLoS Biol* 4:545–554.
13. Ihmels J, Collins SR, Schuldiner M, Krogan NJ, Weissman JS (2007) *Mol Syst Biol* 3.
14. Kafri R, Bar-Even A, Pilpel Y (2005) *Nat Genet* 37:295–299.
15. Lynch M, Force A (2000) *Genetics* 154.
16. Tirosh I, Barkai N (2007) *Genome Biol* 8:R50.
17. Wagner A (2002) *Mol Biol Evol* 19:1760–1768.
18. Papp B, Pal C, Hurst LD (2003) *Nature* 424:194–197.
19. Cliften PF, Fulton RS, Wilson RK, Johnston M (2006) *Genetics* 172:863–872.
20. Mintseris J, Weng Z (2005) *Proc Natl Acad Sci USA* 102:10930–10935.
21. Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe K (2006) *Nature* 440:341–345.
22. Wapinski I, Pfeffer A, Friedman N, Regev A (2007) *Nature* 449:54–61.
23. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, *et al.* (2002) *Nature* 415:180–183.
24. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) *Proc Natl Acad Sci USA* 98:4569–4574.
25. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) *Nucleic Acids Res* Database Issue 32:D449–D451.
26. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, *et al.* (2000) *Nature* 403:623–627.
27. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D (2000) *Nucleic Acids Res* 28:289–291.
28. Xenarios I, Fernandez E, Salwinski L, Duan XJ, Thompson MJ, Marcotte EM, Eisenberg D (2001) *Nucleic Acids Res* 29:239–241.
29. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim S, Eisenberg D (2002) *Nucleic Acids Res* 30:303–305.
30. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chkovskii D, Alon U (2002) *Science* 298:824–827.
31. Shen-Orr S, Milo R, Mangan S, Alon U (2003) *Nat Genet* 32:64–68.
32. DeLuna A, Avendaño A, Riego L, González A (2001) *J Biol Chem* 276:43775–43783.
33. Gibson TJ, Spring J (1999) *TiG* 14:46–49.
34. Guldner U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A, Mewes HW, Stumpflen V (2006) *Nucleic Acids Res* 34:D436–D441.
35. Reguly T, Breitkreutz A, Boucher L, Breitkreutz B-J, Hon GC, Myers CL, Parsons A, Friesen H, Oughtred R, Tong A, *et al.* (2006) *J Biol* 5:11.11–11.28.
36. Wagner A (2003) *Proc R Soc London Ser B* 270:457–466.
37. Ispolatov I, Yuryev A, Mazo I, Maslov S (2005) *Nucleic Acids Res* 33:3629–3635.
38. Pereira-Leal JB, Levy ED, Kamp C, Teichmann SA (2007) *Genome Biol* 8:R51.51–R51.12.
39. Hughes T, Ekman D, Ardawatia H, Elofsson A, Liberles DA (2007) *Genome Biol* 8:8:213.211–218:213.214.
40. Britten RJ (2006) *Proc Natl Acad Sci USA* 103:19027–19032.
41. Jaillon O, Aury J-M, Brunet F, Petit J-L, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, *et al.* (2004) *Nature* 431:946–957.
42. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) *Numerical Recipes in C* (Cambridge Univ Press, Cambridge, UK).
43. Prince VE, Pickett FB (2002) *Not Rev Genet* 3:827–837.
44. Wagner A (2001) *Mol Biol Evol* 18:1283–1292.
45. Pereira-Leal JB, Teichmann SA (2005) *Genome Res* 15:552–559.
46. Marianayagam NJ, Sunde M, Mathews JM (2004) *Trends Biochem Sci* 29:618–625.
47. Batada NN, Reguly T, Breitkreutz A, Boucher L, Breitkreutz B-J, Hurst LD, Tyers M (2007) *PLoS Biol* 5:e154.
48. Yu H, Paccanaro A, Trifonov V, Gerstein M (2006) *Bioinformatics* 22:823–829.
49. Musso G, Zhang Z, Emili A (2007) Retention of protein–protein interactions by ancient duplicated gene products in budding yeast. *Trends Genet* 23:266–269.

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Performance and Scalability of Discriminative Metrics for Comparative Gene Identification in 12 *Drosophila* Genomes

Michael F. Lin[1], Ameya N. Deoras[2], Matthew D. Rasmussen[2], Manolis Kellis[1,2]*

1 Broad Institute of MIT and Harvard University, Cambridge, Massachusetts, United States of America, **2** Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America

Abstract

Comparative genomics of multiple related species is a powerful methodology for the discovery of functional genomic elements, and its power should increase with the number of species compared. Here, we use 12 *Drosophila* genomes to study the power of comparative genomics metrics to distinguish between protein-coding and non-coding regions. First, we study the relative power of different comparative metrics and their relationship to single-species metrics. We find that even relatively simple multi-species metrics robustly outperform advanced single-species metrics, especially for shorter exons ($\leq$240 nt), which are common in animal genomes. Moreover, the two capture largely independent features of protein-coding genes, with different sensitivity/specificity trade-offs, such that their combinations lead to even greater discriminatory power. In addition, we study how discovery power scales with the number and phylogenetic distance of the genomes compared. We find that species at a broad range of distances are comparably effective informants for pairwise comparative gene identification, but that these are surpassed by multi-species comparisons at similar evolutionary divergence. In particular, while pairwise discovery power plateaued at larger distances and never outperformed the most advanced single-species metrics, multi-species comparisons continued to benefit even from the most distant species with no apparent saturation. Last, we find that genes in functional categories typically considered fast-evolving can nonetheless be recovered at very high rates using comparative methods. Our results have implications for comparative genomics analyses in any species, including the human.

Citation: Lin MF, Deoras AN, Rasmussen MD, Kellis M (2008) Performance and Scalability of Discriminative Metrics for Comparative Gene Identification in 12 *Drosophila* Genomes. PLoS Comput Biol 4(4): e1000067. doi:10.1371/journal.pcbi.1000067

Editor: Roderic Guigó, Centre de Regulació Genòmica (CRG), Spain

Received October 22, 2007; **Accepted** March 20, 2008; **Published** April 18, 2008

Copyright: © 2008 Lin et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The study's authors were supported by the MIT Department of Electrical Engineering and Computer Science, the MIT Computer Science and Artifical Intelligence Laboratory, and the Broad Institute of MIT and Harvard. No sponsors or funders influenced the study.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: manoli@mit.edu

Introduction

The recent availability of complete genome sequences from many closely related species has enabled the use of comparative genomics for systematic gene identification. In practice, the discovery power of comparative genomics is intrinsically linked to specific methods for extracting information from from multi-species alignments. Numerous such methods have been developed for gene identification, capturing diverse signals that distinguish protein-coding genes from non-coding regions. These signals are found in the primary sequence of the target genome (e.g. nucleotide frequencies and codon usage biases) and also in the distinctive evolutionary signatures of protein-coding regions (e.g. favoring synonymous vs. non-synonymous substitutions) that only become apparent when informant species are used for comparison.

In this paper, we study the discovery power of diverse discriminative metrics that capture comparative genomics as well as single-species evidence. Given a region of the genome and, when available, its alignment across multiple species, discriminative metrics produce a score that indicates how likely the region is to be protein-coding. Similar to previous studies of the performance of single-sequence metrics [1–3], we measure discovery power in a binary classification framework, based on each metric's ability to discriminate between known protein-coding exons and random non-coding regions.

The goals of our study are twofold. First, we seek to determine the relative power of different metrics, their independence, and the power obtained by combining them. Such metrics can be applied to assess and correct existing gene annotations [4,5], and to decide whether experimentally derived cDNA sequences represent protein-coding mRNAs or non-coding transcripts [6,7]. In addition, our study is immediately applicable to the design of discriminative features for comparative gene structure predictors that can incorporate artibrary metrics to determine precise exon boundaries, such as systems based on semi-Markov conditional random fields (SMCRFs). While initial studies on such discriminative gene prediction systems have successfully focused on their training algorithms [8–10] and advantages over their generative predecessors [11,12], here we focus on the discriminative features they can use, which ultimately enable their increased power.

Second, we seek to understand how discovery power scales with the phylogenetic distance and number of species compared. On one hand, increasing either distance or number of species should, in principle, provide more signal and therefore increased discovery

Author Summary

Comparing the genomes of related species is a powerful approach to the discovery of functional elements such as protein-coding genes. Theoretically, using more species should lead to more discovery power. Many questions remain, however, surrounding the optimal choice of species to compare and how to best use multi-species alignments. It is even possible that practical limitations in the sequencing, assembly, and alignment of genomes could effectively negate the benefit of using more species. Here, we used 12 complete fly genomes to study a variety of metrics used to identify protein-coding genes, including methods that analyze only the genome of interest and comparative methods that examine evolutionary signatures in genome alignments. We found that species over a surprisingly broad range of phylogenetic distances were effective in comparative analyses, and that discovery power continued to scale with each additional species without apparent saturation. We also examined whether comparative methods systematically miss genes considered fast-evolving, and studied how performance is influenced by genome alignment strategies. Our results can help guide species selection for future comparative studies and provide methodological guidance for a variety of gene identification tasks, including the design of future *de novo* gene predictors and the search for unusual gene structures.

power [13], as shown in several pilot studies in selected genomic regions [14–18]. On the other hand, greater phylogenetic distance and more informant species can also lead to conflicting evidence arising from elements that have undergone evolutionary divergence. Moreover, additional species may in practice result in increased noise and systematic errors in the sequencing, assembly, and alignment of complete genomes. In fact, initial studies using *de novo* gene structure predictors with multiple informants led to mixed results [19,20]. Thus, empirical studies of the scalability of gene identification power in multiple complete genomes are needed, to help address several remaining questions surrounding comparative gene identification that are still unresolved: is there an optimal pairwise distance for gene identification, does multi-species discovery power saturate after a small number of compared species, are some classes of genes systematically missed by comparative methods, are synteny-anchored alignments necessary for achieving high specificity?

To address these two goals, we have assembled a large benchmark dataset consisting of tens of thousands of coding and non-coding sequences aligned across twelve recently sequenced *Drosophila* genomes [21,22]. We measure the discriminatory power of diverse metrics and how it varies with sequence length, phylogenetic distance, total number of informant sequences, and the genome alignment strategy. We also study the redundancy and independence of different metrics, and the discovery power of metric combinations. Finally, we discuss the overall strategic implications of our results for comparative approaches to gene identification.

Discriminative Metrics for Gene Identification

We evaluate both well-known methods for gene identification as well as several metrics that we have developed. These metrics are briefly summarized here and in Table 1, while we provide full implementation details in the Methods section.

Pairwise comparative metrics

Most initial efforts at comparative gene identification used a single informant genome to support the annotation of a target genome [15,23–29]. We selected several metrics that capture the essential properties of coding sequence evolution that they observe: the K_A/K_S ratio [30,31] and the Codon Substitution Frequencies

Table 1. Discriminative metrics for gene identification.

	Metric	Description	References
Pairwise comparative	K_A/K_S	Ratio of non-synonymous to synonymous substitutions per site	[30,31]
	Codon Substitution Frequencies (CSF)	Log-likelihood ratio of coding vs. non-coding based on empirical frequencies of all codon substitutions	[5]
	Reading Frame Conservation (RFC)	Percent of nucleotides in same reading frame offset based on indel pattern	[4,32]
	TBLASTX	Significance of protein sequence similarity (bit score), independent of genome alignments	[33]
	Seq. conservation	(baseline) Percent identity	-
Multi-species comparative	*dN/dS* test	Pr($dN/dS < 1$), probability that synonymous substitution rate exceeds non-synonymous substitution rate, based on maximum likelihood phylogenetic models	[34,35,36]
	Codon Substitution Frequencies (CSF)	Pairwise CSF log-likelihood ratios combined by median in each column	[5]
	Reading Frame Conservation (RFC)	Pairwise RFC scores for each informant combined by voting scheme	[4,32]
	Seq. conservation	(baseline) Averaged identity in each column	-
Single sequence	Fourier transform	Three-base periodicity in genetic code	[37]
	Codon bias	Unequal usage of synonymous codons	[38]
	Interpolated context models (ICMs)	Generative probabilistic models measuring *k*-mer frequency biases	[39]
	Z curve	Linear discriminant analysis on *k*-mer frequencies	[2]

Additional details are provided in Methods.
doi:10.1371/journal.pcbi.1000067.t001

(CSF) score [5] observe biases towards synonymous and other conservative codon substitutions; the Reading Frame Conservation (RFC) score observes the strong bias of indels within coding regions to be multiples of three in length [4,32]; TBLASTX measures the genome-wide significance of protein sequence similarity [33]; finally, a baseline sequence conservation metric simply measures the percent nucleotide identity between the target and informant sequences.

Multi-species comparative metrics

We also selected several metrics that use multi-species alignments: the *dN/dS* test observes biases towards synonymous codon substitution using a statistical test based on maximum likelihood phylogenetic algorithms [34–36]; the multi-species CSF and RFC scores use *ad hoc* strategies to efficiently combine their respective pairwise scores; lastly, a baseline multi-species sequence conservation metric measures the largest fraction of species having the same nucleotide in each column (plurality), averaged across the alignment.

Single-sequence metrics

We also included several single-sequence metrics in our benchmarks to compare them to the comparative methods. Since previous studies have benchmarked many single-sequence metrics extensively [1–3], we chose only a representative set here: the Fourier transform measures the strength of the three-base periodicity in coding sequences [37]; codon bias observes the unequal usage of synonymous codons, resulting in part from how different synonymous codons affect translation efficiency [38]; interpolated context models (ICMs) are generative probabilistic models that observe reading frame-dependent biases in the frequencies of *k*-mers in coding sequences, simultaneously for several different *k*-mer sizes [39]; lastly, Z curve observes reading frame-dependent biases in *k*-mer frequencies using a discriminative approach based on Fisher linear discriminant analysis [2].

Benchmarks for Gene Identification Metrics in 12 Fly Genomes

To benchmark the discriminatory power of each of these metrics, we assembled a test set consisting of 10,722 known protein-coding exons (from 2,734 genes) in the fruit fly *Drosophila melanogaster*, and 39,181 random intergenic regions with the same length and strand distribution (see Methods). These provide an ideal setting in which to evaluate genome-wide comparative genomics methods given the high quality of the FlyBase gene annotations [5] and the recent sequencing of ten *Drosophila* genomes [21,22], in addition to *D. melanogaster* [40] and *D. pseudoobscura* [41]. We extracted each of these regions from two different sets of whole-genome sequence alignments of the twelve fly genomes [22], one generated by MULTIZ [42], which uses local alignments of high-similarity regions, and the second generated by the Mercator orthology mapper (C. Dewey and L. Pachter) and MAVID sequence aligner [43], based on the identification of orthologous segments in each genome by conserved gene order (synteny).

For each metric, we scored all the 49,903 regions in our test set (10,722 exons and 39,181 non-coding regions) and then measured its ability to correctly classify them as coding or non-coding. We used four-fold cross-validation to train and apply the metrics that require training data. We evaluated the performance of each metric by examining receiver-operator characteristic (ROC) curves showing its sensitivity and specificity at different score cutoffs. (Here and throughout this paper, we use the term *specificity* as it is defined in binary classification problems: the fraction of true negatives that are correctly classified as negative. This differs from the common usage of the term in the gene prediction field to refer to the fraction of the examples classified as positive that are true positives. Additionally, we use the term *false positive rate* to mean 1-Specificity, or the fraction of true negatives incorrectly classified as positive.)

Based on the ROC curve for each metric, we also computed two different summary error measures, to facilitate comparing the performance of different metrics and methodological choices:

- The *minimum average error* (MAE) is the average of the false negative rate and the false positive rate at the cutoff where this average is minimized; intuitively, this is the "elbow" of the ROC curve. This represents the fraction of examples that are incorrectly classified (if the positive and negative classes are the same size), at a single point on the ROC curve.
- The *area above the curve* (AAC) is the area lying above the ROC curve in the unit square. Although it lacks a simple interpretation, the AAC summarizes more information about classification performance over all sensitivity/specificity regimes, providing a measure complementary to MAE.

Results

Performance, Independence, and Combinations of the Metrics

We first compared the overall performance of the metrics (Figure 1). All of the metrics we evaluated demonstrated high classification performance, but some general trends were apparent. The comparative metrics (using the MULTIZ alignments of all twelve fly genomes) generally outperformed the single-sequence metrics (except for the baseline sequence conservation metric). For example, the best comparative metric resulted in 24% lower error than the best single-sequence metric (0.050 MAE for the *dN/dS* test vs. 0.065 for Z curve). Different metrics were preferable at different sensitivity/specificity tradeoffs. For example, the CSF and *dN/dS* metrics achieved the highest specificity (99.9% for CSF) even at fairly high sensitivities (85.2%). RFC tended towards higher sensitivity and lower specificity than CSF and *dN/dS*.

We also compared the pairwise metrics, using the best pairwise informant (*D. ananassae*; we investigate different pairwise informants below), and found similar trends (Figure S1). For example, CSF and K_A/K_S performed comparably, showing the highest specificity, while RFC tended towards higher sensitivity and lower specificity. TBLASTX performed substantially worse than K_A/K_S, CSF, and RFC, but it was still better than our baseline conservation metric. Notably, none of the pairwise comparative metrics outperformed the best single-sequence metric (Z curve) according to MAE and AAC error, and they exhibited generally lower sensitivity. CSF and K_A/K_S were, however, able to achieve higher specificity at a moderate sensitivity tradeoff. For example, at 80% sensitivity, CSF had a nearly ten-fold lower false positive rate than Z curve (0.15% and 1.39%); the specificity of CSF exceeded Z curve at less than 85% sensitivity, compared to 93% sensitivity at Z curve's MAE point.

Comparative methods are strongly preferred for short exons

We next assessed each metric's discriminatory power for different sequence length categories (Figure 1C). All of the metrics performed better on longer sequences than shorter sequences. Single-sequence metrics performed comparably or slightly better

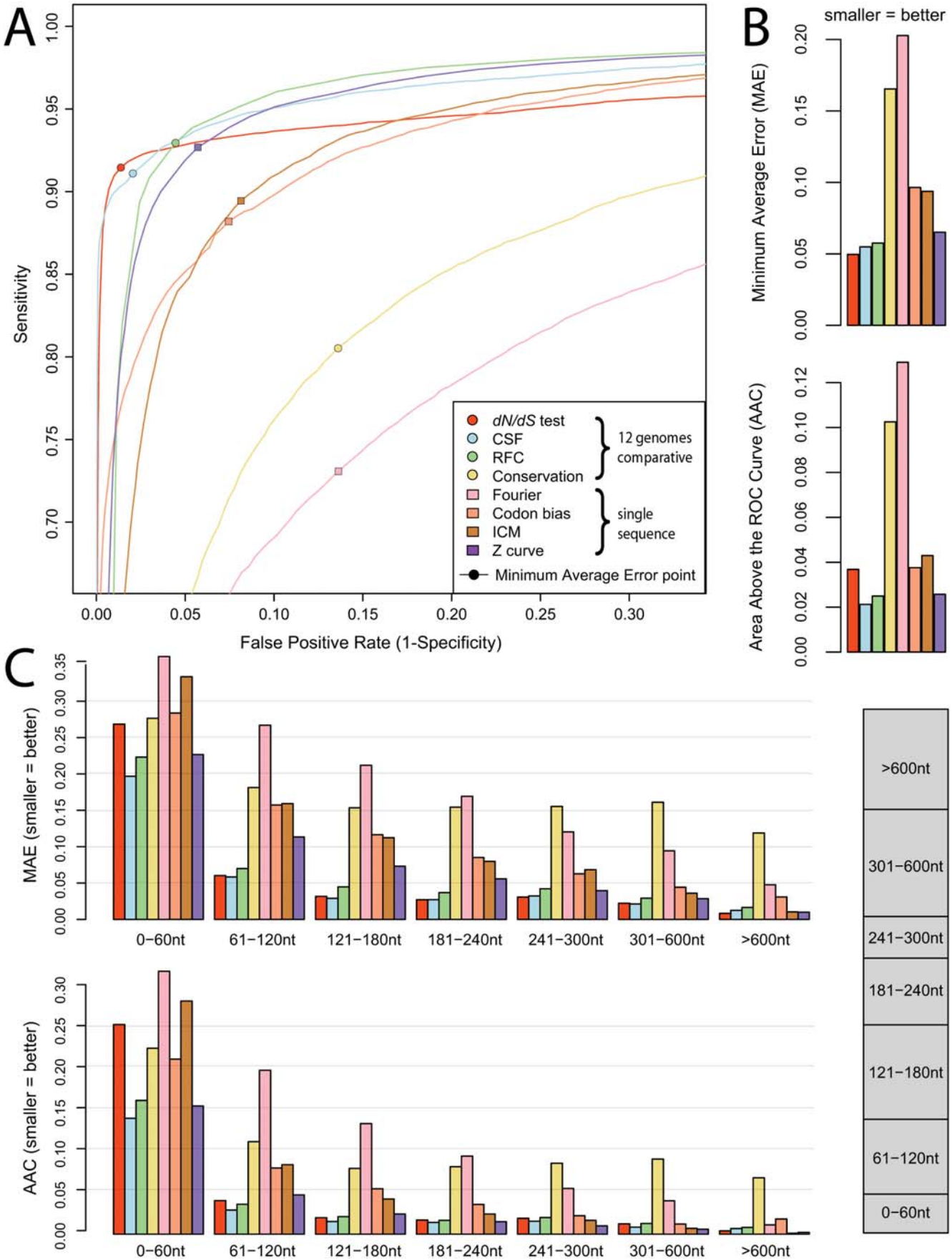

A
Sensitivity
False Positive Rate (1-Specificity)
dN/dS test
CSF
RFC
Conservation
12 genomes comparative
Fourier
Codon bias
ICM
Z curve
single sequence
Minimum Average Error point
B
smaller = better
Minimum Average Error (MAE)
Area Above the ROC Curve (AAC)
C
MAE (smaller = better)
AAC (smaller = better)
0–60nt
61–120nt
121–180nt
181–240nt
241–300nt
301–600nt
>600nt

Figure 1. Overall discovery power of discriminative metrics using 12 genomes. (A) ROC curves showing sensitivity and specificity of each metric on classifying 10,722 known exons and 39,181 random non-coding regions. Comparative methods tended to outperform single-sequence metrics, with the exception of a baseline sequence conservation metric. CSF and the *dN/dS* test achieved near-perfect specificity, while RFC achieved high sensitivity. (B) Summary error statistics for each metric computed from the ROC curves. Minimum Average Error (MAE) is the minimum average of the false negative rate and false positive rate. Area Above the Curve (AAC) is the area above the ROC curve in the unit square. (C) MAE and AAC error statistics for each metric when the dataset is partitioned into several sequence length categories. All metrics tended to perform better on longer sequences than on shorter sequences. Comparative methods strongly outperformed single-sequence metrics on short sequences (60–240 nt). Inset: relative size of each sequence length category.
doi:10.1371/journal.pcbi.1000067.g001

than comparative methods for long sequences (>240 nt), but comparative methods strongly outperformed single-sequence metrics on shorter sequences. For example, in the length range of 181–240 nt (which includes the median exon length) the best comparative metric resulted in 51% lower error than the best single-sequence metric (0.027 MAE for the *dN/dS* test and 0.056 MAE for Z curve). In the shorter length range of 121–180 nt, the best comparative metric resulted in 60% lower error than the best single-sequence metric (0.029 MAE for CSF and 0.073 MAE for Z curve). Different comparative methods were also preferred at different lengths. For example, CSF strongly outperformed the *dN/dS* test on the shortest sequences (≤60 nt), while they performed comparably on longer sequences.

Independence of the metrics

While each of the metrics we studied exhibited unique performance characteristics, some measure similar fundamental lines of evidence, and thus may tend to err on the same examples. We investigated the independence of the metrics, indicated by how differently they rank the exons in our test set, using a dimensionality reduction technique called multidimensional scaling (MDS; see Methods). This analysis led to a two-dimensional visualization shown in Figure 2A, in which each point represents one of the metrics and the distance between the points approximately represents their dissimilarity.

We found that the *dN/dS* test and CSF behaved very similarly, while RFC was clearly distinct. The sequence conservation metric was separate from each of these, while TBLASTX clustered with CSF and *dN/dS*. The four single-sequence metrics formed two additional clusters distinct from the comparative metrics. These findings agree with intuition: CSF and the *dN/dS* test both observe the distinctive biases in codon substitutions in protein-coding sequences, while RFC observes patterns of insertions and deletions that are essentially orthogonal to codon substitutions, and the single-sequence metrics observe compositional biases and periodicities that are ignored by the comparative metrics.

Combining metrics

The relative independence of several of the metrics suggests that combining them could lead to higher performance. We selected five metrics representing each of the MDS clusters (CSF, RFC, sequence conservation, Z curve, and codon bias) and combined them using cross-validated linear discriminant analysis (LDA). As expected, the hybrid metric outperformed any of its inputs: by MAE error, the LDA hybrid resulted in 27% lower error than its best input metric (0.040 MAE for LDA vs. 0.055 for CSF). The hybrid metric demonstrated much higher sensitivity than any of its input metrics (Figure 2B), and higher specificity than all of the input metrics except CSF. We obtained almost identical results using a second hybrid metric based on a linear support vector machine instead of LDA. Thus, although CSF and the *dN/dS* test remain the methods of choice for the highest specificity, the hybrid metrics achieved higher overall performance.

Dependence of Comparative Methods on Genome Alignments

We next investigated how strongly the performance of the comparative methods depends on genome sequence alignments. We compared the above results, based on MULTIZ local similarity-based alignments, with the corresponding results based on the synteny-anchored Mercator/MAVID alignments. Overall, the two alignments led to highly concordant results, with similar trends in the performance of the metrics relative to each other and across different sequence lengths. There were, however, some notable differences in their absolute levels of performance.

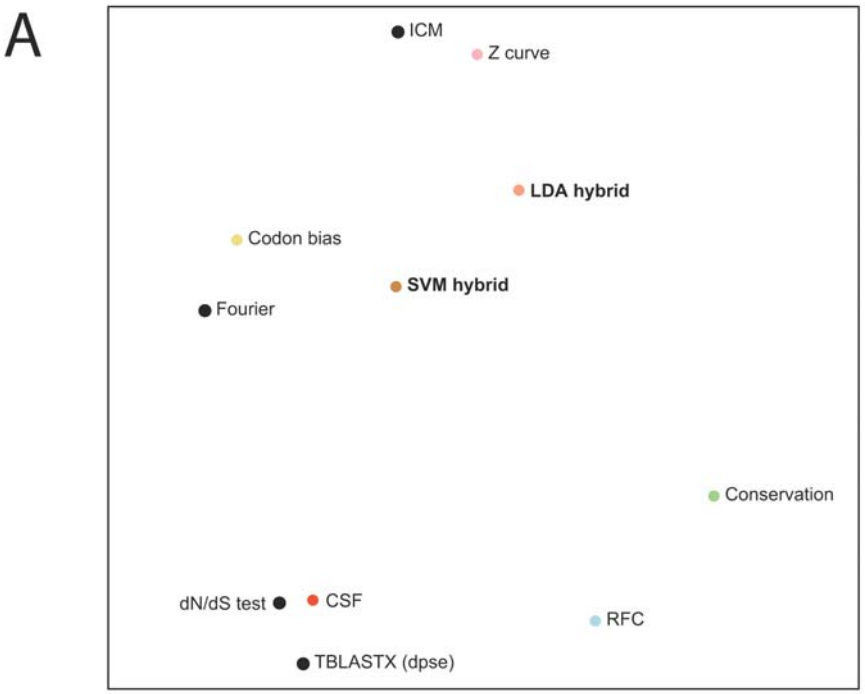
We expected the local alignment approach to give higher sensitivity than the synteny-anchored alignments, since it should be better able to align exons that have undergone rearrangements [45]. Indeed, we found that MULTIZ tended to align more species for each region (Figure S2) and led to higher sensitivity than the Mercator/MAVID alignments (e.g. 90% vs. 87% for CSF at 99% specificity, with 85% of exons detected in both alignments; Figure S3). Conversely, we expected the synteny-anchoring approach used by Mercator/MAVID to give higher specificity than the local alignment approach of MULTIZ, since it may generate fewer spurious non-orthologous alignments [45]. However, we found that while the Mercator/MAVID alignment could lead to slightly higher specificity, it did so only at disproportionate sensitivity tradeoffs. For example, with the baseline sequence conservation metric, specificity using the Mercator/MAVID alignments exceeded that of the MULTIZ alignments only at lower than 58% sensitivity (compared to 80% sensitivity at the MULTIZ-based MAE point). Similarly, with RFC, specificity resulting from the Mercator/MAVID alignments was greater only at lower than 63% sensitivity (compared to 92% MAE sensitivity).

Overall, the Mercator/MAVID alignments led to somewhat lower sensitivity without a clear specificity advantage, and this was reflected in worse MAE and AAC error statistics (Figure S3). We therefore focused on the MULTIZ alignments for the remainder of our analysis. We note, however, that the Mercator/MAVID alignments did allow detection of some exons not detected in the MULTIZ alignments (~2% of all exons). More generally, these empirical observations could be highly dependent on parameter settings of the genome alignment programs, and further investigation of these strategies is required.

A Wide Range of Phylogenetic Distances Is Effective in Pairwise Analysis

To investigate which species are the most and least effective informants for gene identification, we evaluated each pairwise comparative metric using informant genomes at increasing evolutionary distance from *D. melanogaster*. We applied each metric to pairwise alignments of *D. melanogaster* with *D. erecta*, *D. ananassae*, *D. pseudoobscura*, *D. willistoni*, and *D. grimshawi*, each representing various clades within the genus *Drosophila* (Figure 3).

We found that *D. ananassae* was overall the most effective informant, outperforming other species on most metrics. However,



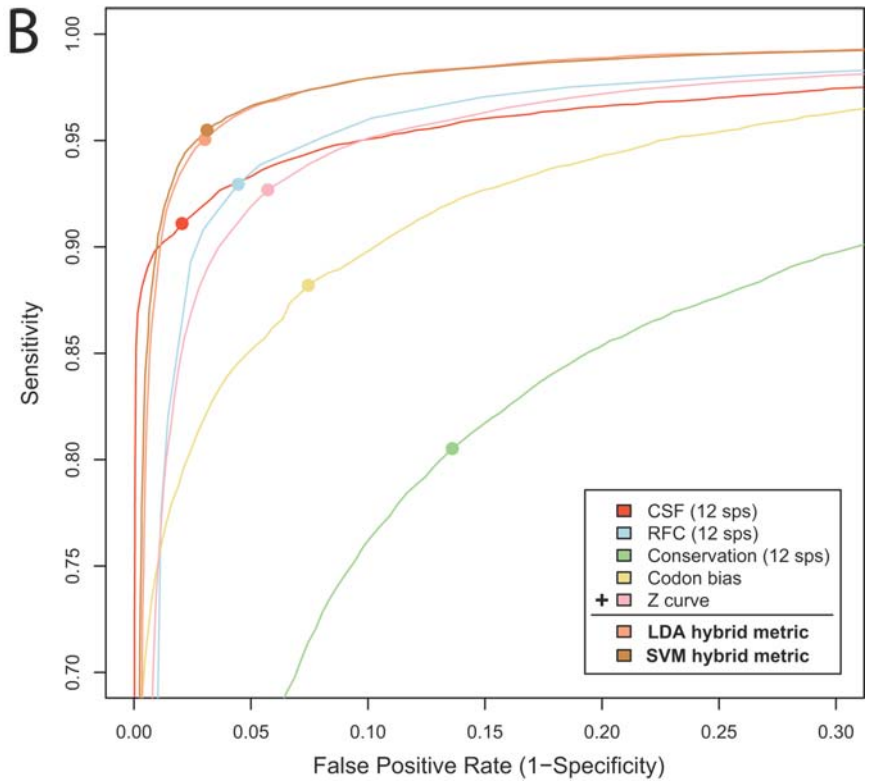


Figure 2. Independence of metrics and discovery power of metric combinations. (A) Multidimensional scaling (MDS) visualization in which each point represents a metric and the distance between any two points approximately represents their dissimilarity, measured as 1-(rank correlation of the scores of the known exons). Hybrid metrics appear closer to the center, suggesting that they successfully combine distinct information from the individual metrics. (B) ROC curves showing the performance of two hybrid metrics created by combining five comparative and single-sequence metrics using Linear Discriminant Analysis (LDA) or a Support Vector Machine (SVM). The hybrid metrics outperformed all of their input metrics.
doi:10.1371/journal.pcbi.1000067.g002

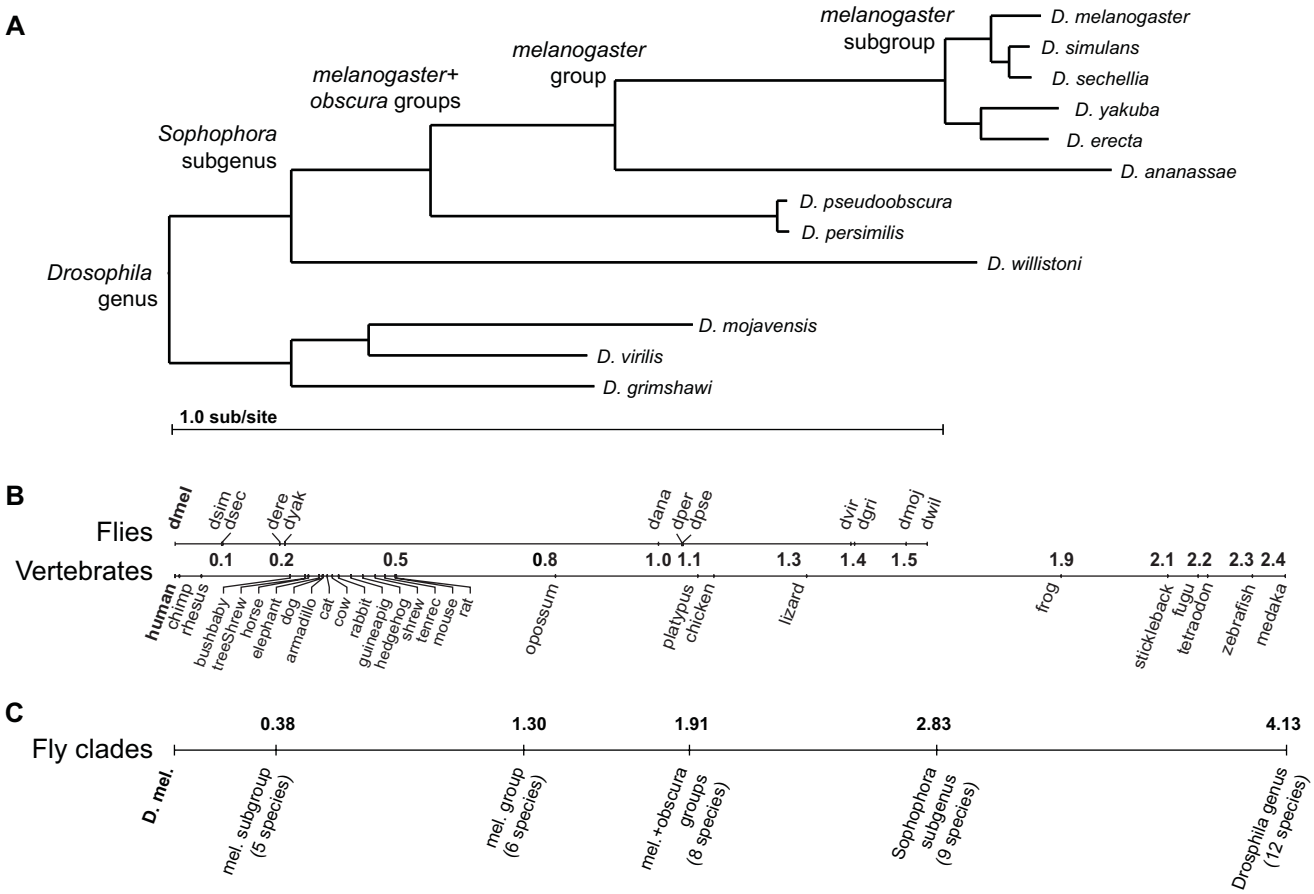


Figure 3. Evolutionary distances relating 12 *Drosophila* species. (A) Phylogenetic tree and estimated neutral branch lengths for the species. Tree topology follows the accepted phylogeny of these species [21,22]. Neutral substitution rates estimated from 12,861 4-fold degenerate sites in syntenic one-to-one orthologs (see Methods). (B) Pairwise distance of each of the 11 other *Drosophila* species from *D. melanogaster*, as compared to similarly estimated distances for vertebrates. (C) Total independent branch length provided by several subsets of the *Drosophila* species used to benchmark multi-species methods.
doi:10.1371/journal.pcbi.1000067.g003

inspection of the corresponding ROC curves often revealed a more complex situation, with multiple species showing similar performance, and sometimes higher for certain sensitivity/specificity tradeoffs. For example, with K_A/K_S, *D. ananassae* and *D. willistoni* performed comparably, with *D. ananassae* leading to slightly higher sensitivity and *D. willistoni* leading to slightly higher specificity (Figure 4A). Similarly, with RFC, closely related species led to slightly higher sensitivities, and more distant species led to slightly higher specificities (Figure S4). Hence, while *D. ananassae* was overall the most effective informant, it did not *robustly* outperform the other pairwise informants we studied. The only exception was *D. erecta*, the most closely related to *D. melanogaster* of the species we studied. *D. erecta* was consistently less informative than the others, leading to the lowest overall classification performance on most of the pairwise metrics.

To investigate more distant species for which we lacked whole-genome alignments, we also applied TBLASTX to the genomes of the mosquito [46] and honeybee [47]. We found that these species led to much worse performance than the *Drosophila* species as informants for *D. melanogaster* (Figure 4B).

We conclude that a broad range of species within the genus *Drosophila* (outside of the *melanogaster* subgroup) make effective pairwise informants for gene identification in *D. melanogaster*, while the mosquito and honeybee, the next most closely related species with fully sequenced genomes, are likely to be too distant for this application. These findings are consistent with a previous smaller-scale study of comparative gene identification power in flies [14], and previous theoretical and simulation studies suggesting that, while some mathematically optimal distance may exist, species at a broad range of phylogenetic distances should be comparably effective informants for identifying exons and other conserved elements [13,15].

Multi-Species Comparisons Lead to Higher Performance

We next investigated the effectiveness of increasing numbers of informant species on the metrics that can use multiple informants. We evaluated each metric using subsets of the available species corresponding to increasingly broad clades within the genus *Drosophila* (see phylogeny in Figure 3): the *melanogaster* subgroup (5 species including *D. melanogaster*), the *melanogaster* group (6 species), the *melanogaster* and *obscura* groups (8 species), the subgenus *Sophophora* (9 species), and finally all 12 species of the genus *Drosophila*.

We found that for each of the metrics we benchmarked in this way, discriminatory power tended to increase as additional informant species were used (Figure 5A). In contrast to our

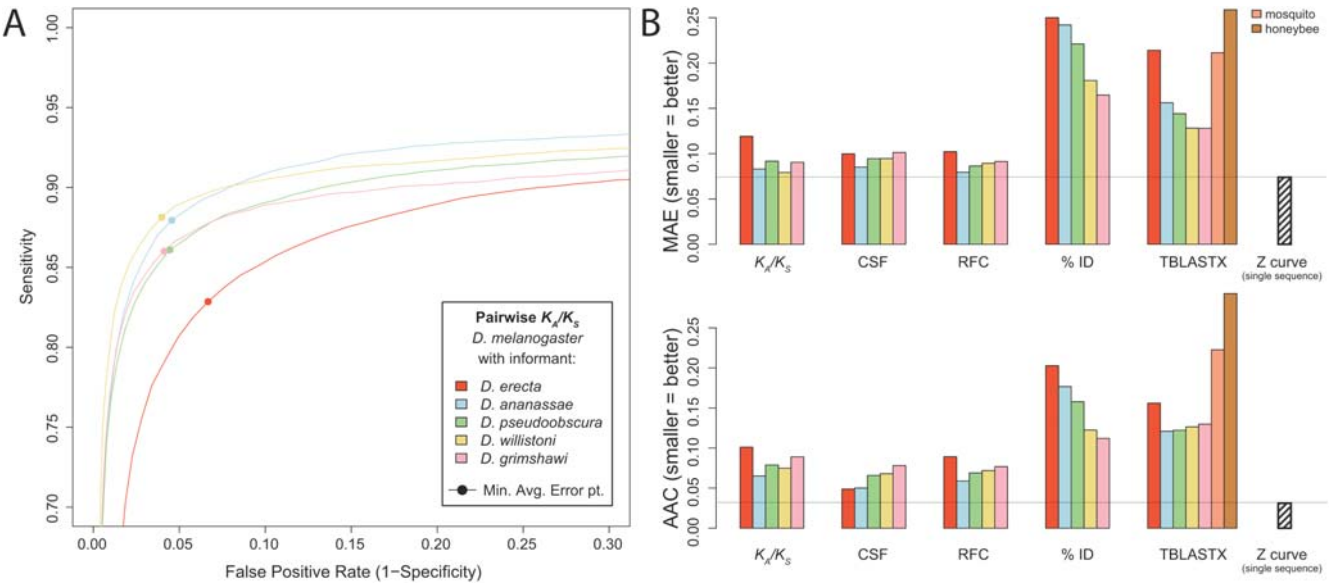


Figure 4. Pairwise discovery power using different informant species. (A) ROC curves for K_A/K_S using *D. melanogaster* with each of five different informant species. Species at a wide range of evolutionary distances performed comparably, except for *D. erecta*, the most closely related to *D. melanogaster*, which clearly underperformed the others. (B) MAE and AAC error statistics for each pairwise comparative metrics applied to the same five informants. *D. ananassae* (blue) is overall the preferred informant, but not uniformly so. For TBLASTX, the performance is also shown using mosquito (*Anopheles gambiae*) and honeybee (*Apis mellifera*), which led to worse performance than the *Drosophila* species. No pairwise comparison outperformed the best single-sequence metric (Z curve).
doi:10.1371/journal.pcbi.1000067.g004

previous pairwise analysis, in which the most distant *Drosophila* informants led to similar or slightly worse performance than closer species, *adding* informants at increasing distances led to a clear trend in higher classification performance. The *dN/dS* test, RFC, and the sequence conservation metric each showed a smooth progression of increasing performance with each successively larger group of informant species. For example, starting from the four informants within the *melanogaster* subgroup, the *dN/dS* test achieved an MAE of 0.103. With the addition of each successive group of informants, the MAE was reduced relatively by 35%, 43%, 48%, and finally by 52%. CSF showed a similar trend through the subgenus *Sophophora*, but did not clearly benefit from the subsequent addition of the final three informants of subgenus *Drosophila*. In all cases, the improvement with multiple species was most pronounced for short exons (Figure 5B).

With a sufficient number of informants, the multi-species metrics surpassed single-sequence metrics according to MAE (Figure 5C). This also stands in contrast to our pairwise analysis, in which no informant enabled any comparative metric to outperform the best single-sequence metric (Z curve). CSF exceeded the performance of Z curve once we used at least six species ($\geq$1.3 sub/site), *dN/dS* with at least eight species ($\geq$1.9 sub/site), and RFC, using its simplistic vote-tallying scheme, with all twelve species (4.1 sub/site). The baseline sequence conservation metric never outperformed Z curve, although its performance also increased with additional species. (We note that while these results show that a certain number of informants is *sufficient*, they do not imply that they are all *necessary* to achieve some level of performance; removing informants that contribute very little independent branch length might not substantially reduce performance.)

In most cases, the four informants of the *melanogaster* subgroup together yielded worse performance than pairwise analysis with the best pairwise informant, *D. ananassae*. In contrast, all of the informant clades that combined *D. ananassae* with more distant species led to better performance than any pairwise analysis. This affirms our earlier conclusion, based on a pairwise analysis with *D. erecta*, that the species within the *melanogaster* subgroup are suboptimal informants for the metrics we studied, presumably because they are too closely related to *D. melanogaster*. Indeed, the neutral distance of *D. ananassae* from *D. melanogaster* is 1.0 substitutions per neutral site, while the *total* independent branch length provided by the four *melanogaster* subgroup informants is only 0.4 sub/site.

Characterizing Genes that Comparative Methods Fail to Detect

It is well-known that genes in certain categories of biological function tend to be faster-evolving [41,46–48]. We lastly investigated whether comparative metrics therefore systematically fail to distinguish such genes from non-coding regions. We obtained Gene Ontology (GO) annotations [49,50] for each of the 2,734 genes comprising our test set. For each of the 192 GO terms represented by at least thirty genes in our test set, we determined the fraction of those genes with at least one exon scoring above a stringent cutoff ("detected genes").

We found that all of the functional categories we investigated had very high detection rates (Table S1). For example, with a CSF cutoff corresponding to 85% exon sensitivity and 99.9% specificity using all twelve fly genomes, the overall fraction of detected genes was 92%, and the detection rates surpassed 90% for all but two functional categories: serine-type endopeptidase activity (89% detected genes) and its superset, serine-type peptidase activity (86%). Serine proteases play key roles in insect innate immunity, and some likely evolve under positive selection [46,51,52]. Several other categories that intuition suggests might relate to more rapidly evolving genes, however, were not problematic, including immune response (94%), gametogenesis (95%) and G-protein coupled receptor activity (100%).

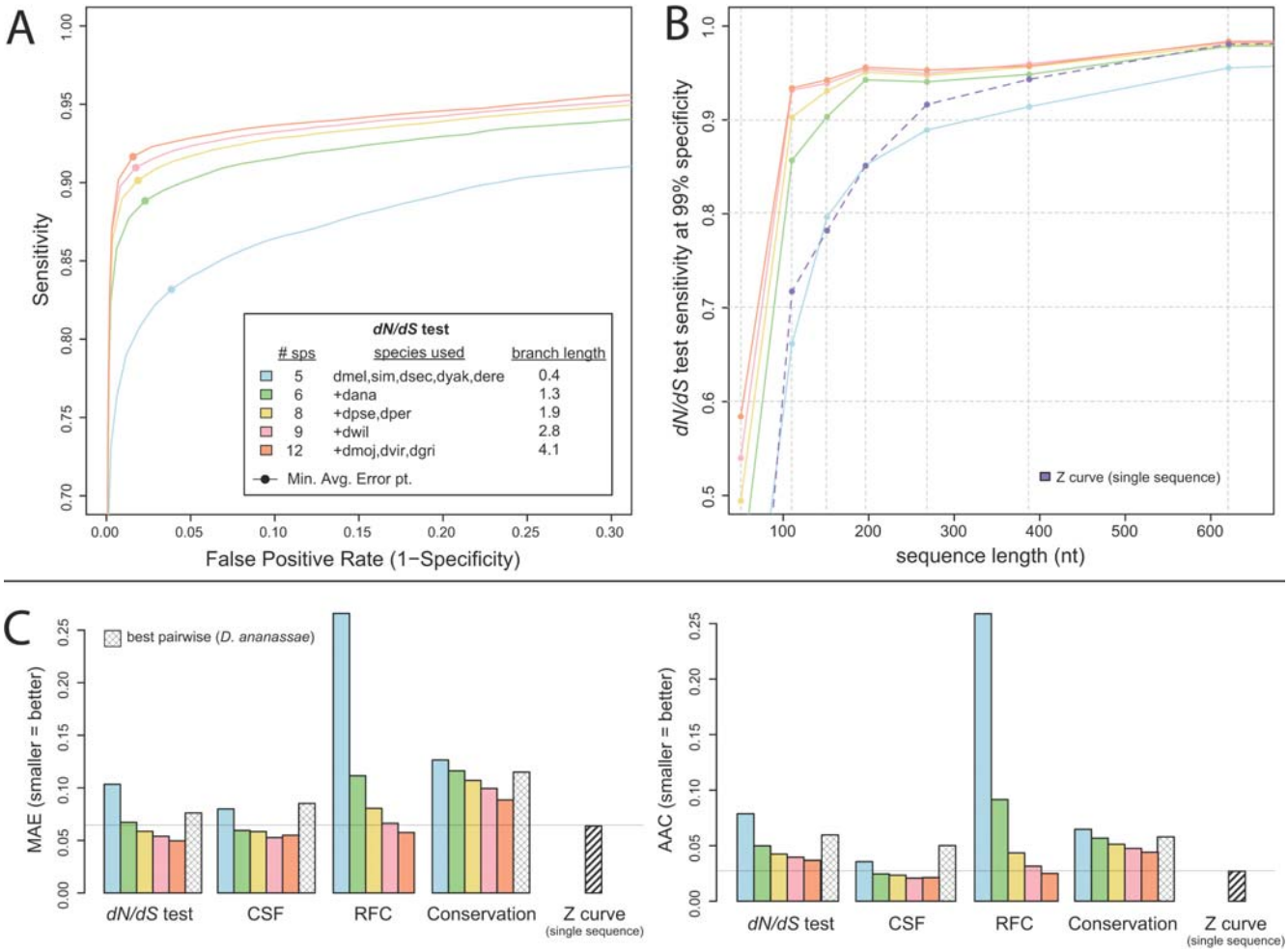


Figure 5. Multi-species discovery power using increasing numbers of informant species. (A) ROC curves for the *dN/dS* test using subsets of *Drosophila* species corresponding to increasingly broad phylogenetic clades from *D. melanogaster* (see Figure 1). Discriminatory power steadily increased as more informants were used, leading to strictly better sensitivity and specificity. (B) Effect of additional species was most pronounced for short exon lengths. (*x*-axis) mean length within a quantile of the sequence length distribution (*y*-axis) sensitivity of the *dN/dS* test within each quantile at fixed specificity (99%). (C) MAE and AAC error statistics for each multi-species comparative metric using the same subsets of informants. Also shown for comparison are the best pairwise analysis and the best single-sequence metric, both of which are outperformed by multi-species methods with sufficient informants.
doi:10.1371/journal.pcbi.1000067.g005

Instead, comparative metrics had the most difficulty detecting genes of unknown function. Three GO terms indicating unknown function (unknown cellular component, molecular function, and biological process) had only 67%, 61%, and 60% detected genes. In fact, of the genes that were not detected at this cutoff, 85% were of unknown function or lacked any GO term, compared to 49% of all the genes in our dataset. These trends held for all of the comparative metrics and cutoffs we investigated (Table S1).

Overall, these results indicate that comparative methods using the twelve fly genomes were able to detect the vast majority of genes in all of the functional categories we investigated (which were represented by at least 30 genes in our dataset; a larger sample might reveal more specific functional categories that are, in fact, very difficult for comparative methods to detect). They had much greater difficulty detecting genes of unknown function, which may be under less selective constraint overall [14,21] but could also include a higher proportion of incorrect or spurious annotations [5]. Interestingly, Z curve, a single-sequence metric, also showed much lower sensitivity to genes of unknown function (Table S1), suggesting that these genes, if they are correctly annotated, tend to be unusual in several ways.

Discussion

In this paper, we investigated discriminative metrics for distinguishing protein-coding sequences from non-coding sequences. We found that multi-species comparative methods outperform single-sequence metrics, particularly on short sequences ($\leq$240 nt). On the other hand, the pairwise comparative methods we studied achieved higher specificity, but did not outperform advanced single-sequence metrics overall. We showed that several comparative and single-sequence metrics can be combined into a more powerful hybrid metric. We found that a broad range of species within the genus *Drosophila* are comparably effective pairwise informants for *D. melanogaster*, in agreement with theoretical predictions. We showed that adding more species to comparative analysis progressively increased genome-wide discovery power, for a variety of different methods. Contrary to expectation, we found no evidence that synteny-anchored alignments lead to appreciably

higher specificity, and no evidence that comparative methods systematically fail to detect genes in functional categories typically considered fast-evolving.

Among the three multi-species comparative metrics we studied (CSF, the *dN/dS* test, and RFC; excluding the baseline sequence conservation metric), none strictly outperformed the others. RFC tended towards lower specificity but higher sensitivity than CSF and the *dN/dS* test. CSF was more effective than the *dN/dS* test on the shortest exons, but they performed comparably overall, and both achieved near-perfect specificity at moderate sensitivity tradeoffs. We developed CSF as a simpler alternative to the computationally expensive phylogenetic algorithms upon which the *dN/dS* test is based, and we consider it successful in this respect, considering its comparable results and its much faster total compute time (on our dataset, completed in several minutes for CSF vs. a few weeks for the *dN/dS* test using PAML).

On the other hand, our tests with different numbers of informant species suggest that the CSF method may benefit from future improvements to take advantage of ever-larger numbers of informants. Both CSF and RFC are discriminative methods that use heuristic approaches to combine multi-species evidence, making them less theoretically appealing than generative phylogenetic models such as those used in the *dN/dS* test. It is likely that such principled statistical frameworks can lead to further improvements for both CSF and RFC. Presently, however, the fact that both of these relatively simple methods outperformed advanced single-sequence metrics, and even competed with a maximum-likelihood phylogenetic algorithm, speaks to the power of the underlying comparative data. Lastly, we note that simple methods such as RFC and K_A/K_S might be preferable in certain ways when working with species for which high-accuracy training data is not available. In our setting, the best performing metrics tended to be highly parameterized approaches that require reliable training data, and thus probably benefited from the excellent FlyBase/BDGP annotation of the *D. melanogaster* genome.

Selection of Informants for Comparative Gene Identification

Using a variety of different methods, we found that species ranging from 1.0–1.4 substitutions per neutral site from *D. melanogaster* are comparably effective informants for pairwise gene identification, with slight preference given to the closer end of this range. This ''optimal'' range might extend both towards closer species (between *D. erecta* and *D. ananassae*) and towards more distant species (between *D. grimshawi* and *A. gambiae*), but these distances were not explored in the currently sequenced genomes. This range is comparable to the distance from human of the opossum (0.8 sub/site), chicken (1.1 sub/site), and lizard (1.3 sub/site), suggesting that species more distant than the eutherian mammals (the farthest of which are less than 0.5 sub/site; Figure 3) may prove to be excellent informants for human gene identification.

Moreover, our study showed that comparative genomics power did not saturate with the number of species compared, as the multi-species metrics tended to show continued improvement from each progressively larger group of informants studied (Figure 5). The overall improvement did become more incremental as the number of informants grew, which could be interpreted either as diminishing returns from additional genomes, or simply as the expected asymptotic increase in performance towards an achievable optimum. Importantly, the improvement from more informants was far more pronounced among short exons than long exons (Figure 5B); this suggests that, while long exons are easy to discover even with few species, still more informants may significantly improve the discovery of short coding exons, and perhaps other classes of small elements. Thus, especially for small elements, we apparently have not yet reached a saturation point with twelve metazoan species spanning a total of 4.13 substitutions per neutral site.

We chose to express discovery power as a function of the neutral substitution rate estimated for the species compared (Figure 3). While this rate provides a compelling measure of expected discovery power [13], it is important to note that genetic distance between species (whether measured by neutral substitution rate or other metrics [21,53]) is far from the only consideration that should guide comparative informant selection. For example, population dynamics affect the strength of selection relative to neutral drift, and thus may skew the relationship between neutral divergence and the significance of observed conservation in some lineages [54,55]. Additionally, the genome size and the density and type of repetitive elements in an informant genome may affect the ability to sequence, assemble, and align it to a target genome, especially if low-coverage [18] or short-read [56,57] sequencing strategies are used. Accurate alignment is further complicated by variation in the rates of chromosomal rearrangement and segmental duplication and loss, which are likely to affect the proportion of the genome that can be accurately recognized as orthologous, even for species that show similar nucleotide divergence.

Much more fundamentally, distant species share less in common biologically; indeed, the 12 *Drosophila* species were selected in part to represent the diverse ecological niches they occupy [58] and the neutral distance they span (approximately corresponding to the distance between human and reptiles). Thus, while our results suggest that such distant species may nonetheless be highly informative given high-quality sequences and alignments, future empirical studies should compare them to the use of many species at closer distances, such as those represented by the eutherian mammals, for gene identification.

Implications for Gene Prediction Strategies

One application of the metrics we have studied will be their integration into *de novo* gene structure predictors based on semi-Markov conditional random fields, which can combine multiple discriminative metrics in a manner not unlike our LDA hybrid. Our results suggest that these systems should be able to use multiple informant species and multiple metrics to identify protein-coding sequences with higher accuracy, especially on short exons. Still, it is not obvious that these trends in the metrics' performance necessarily imply higher-accuracy prediction of complete gene structures, since the latter also strongly depends on the detection of splice sites and other sequence signals [12,59]. Additionally, like the more advanced metrics we studied, such systems tend to be highly parameterized and thus dependent on high-quality training data, which may not be available in less well-studied species. More fundamentally, the probabilistic models used in gene predictors make simplifying assumptions about gene structures that lead to many incorrect predictions, and that cannot be relaxed just by using more powerful metrics. For example, they currently cannot predict nested and interleaved genes, which are fairly common in metazoan genomes [5,50,60–62], since these structures violate Markov independence assumptions. A similar challenge is presented by alternative splice isoforms with mutually exclusive exons that do not splice to each other in-frame.

The methods we have studied also have other important applications, such as assessing and refining existing annotations, and searching the genome for coding regions that are systematically missed or erroneously modeled by other methods. In

particular, the effectiveness of comparative methods for detecting short coding regions may prove crucial in identifying short proteins, which are known to serve important biological roles but have probably been systematically under-represented in genome annotations [63–66]. They also provide a promising way to search for gene structures that violate traditional assumptions entirely, such as stop codon readthrough, translational frameshifts and polycistronic transcripts, which also might be more common in animal genomes than currently appreciated [5].

Methods

Genomes, Alignments, Annotations, and Phylogeny

We used "Comparative Analysis Freeze 1" assemblies of the twelve *Drosophila* genomes [21] available from the following web site: http://rana.lbl.gov/drosophila/assemblies.html. We used two different genome alignment sets [22]. One was derived from a synteny map generated by Mercator (C. Dewey, http://www.biostat.wisc.edu/~cdewey/mercator/) and sequence alignments generated by MAVID [43]. The other genome alignments were generated by MULTIZ [42]. These alignments are available from the following web site: http://rana.lbl.gov/drosophila/wiki/index.php/Alignment.

We obtained FlyBase release 4.3 annotations from the following web site:

ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_tr4.3_20060303/gff.

We estimated branch lengths in the phylogenetic tree for the flies (shown in Figure 3) based on four-fold degenerate sites in alignments of orthologous protein-coding genes. We identified one-to-one orthologs based on FlyBase annotation release 4.3 for *D. melanogaster* and community annotations for the 11 other species [21], yielding 12,861 four-fold sites. Then, to estimate branch lengths, we ran PHYML v2.4.4 [67] with an HKY model of sequence evolution, a fixed tree topology (Figure 3A), and remaining parameters at default values. For comparison with vertebrates, we estimated the branch lengths for 28 vertebrates using 10,340 four-fold sites, based on alignments of genes with one-to-one orthologs in human, dog, and mouse [68]. We obtained the MULTIZ vertebrate alignments from the UCSC Genome Browser [69].

Dataset Preparation

We randomly sampled 2,734 of the 13,733 euchromatic genes in FlyBase annotation release 4.3, and then selected all 10,722 non-overlapping exons of all transcripts of those genes. We chose this strategy of randomly sampling genes and selecting all exons of those genes, rather than directly sampling exons, to facilitate studying how the power of each metric varies across different functional categories of genes. Although not by design, the length distribution of sequences in our test set (median = 224 nt, mean = 404 nt, sd = 570 nt) is very similar to the length distribution of exons in the genome (median = 220 nt, mean = 408 nt, sd = 568 nt). Each known exon was evaluated in its annotated reading frame.

For each known exon in our dataset, we selected four non-coding regions of the same length and strand. We selected each of these regions by randomly choosing a start coordinate in the BDGP Release 4 assembly of the *D. melanogaster* euchromatic chromosome arms, and ensuring that the resulting region did not overlap an annotated coding exon. We also chose only regions consisting of at least 50% nucleotide characters (as opposed to Ns). The codon reading frame for the non-coding regions was always set arbitrarily to 0 (that is, they were always considered to begin with a complete codon). We removed in-frame stop codons in *D. melanogaster* from the non-coding regions (the length of each control region matched the corresponding exon *after* removing stop codons). All the regions in the dataset were selected without regard to how well they were aligned in either genome alignment set we used.

The coordinates, sequences, and alignments of our dataset are available for download (Text S1).

Metric Training and Evaluation

CSF and the single-sequence metrics (except for Fourier transform) require training to estimate parameters. To avoid overfitting, we trained and applied them using four-fold cross validation: we randomly partitioned the dataset into four subsets, and then generated scores for each subset by training on the other three subsets. We then combined the scores for the subsets to obtain scores for the entire dataset. We applied the other metrics directly to each sequence.

We computed ROC curves for each metric by choosing 250 cutoffs representing quantiles of the score distribution over the entire dataset, and at each cutoff, evaluating sensitivity and specificity when sequences scoring above the cutoff are considered positively classified, and sequences scoring less than or equal to the cutoff are negatively classified. Some metrics failed to produce a score for some sequences; for example, comparative metrics produced no score for sequences in which no alignment was present. These sequences were regarded as negatively classified at all cutoffs, reflecting a non-coding default hypothesis. Our ROC curves may therefore underestimate the sensitivity or overestimate the specificity that each comparative method would exhibit if given perfect alignments of all orthologous elements.

We computed the MAE as the highest average sensitivity and specificity among the 250 points on the ROC curve, and the AAC by trapezoidal integration over these points.

Metric Implementation Details

K_A/K_S. To estimate K_A/K_S, we used the method of Nei and Gojobori [30], which is simple and widely used although it is known to have certain inherent biases [31]. We considered only codons with ungapped alignments between *D. melanogaster* and the informant.

TBLASTX. We used the blastall program in NCBI BLAST 2.2.15 [33] with the parameters -p tblastx -m 9 against the repeat-masked genome assembly of the informant species. For each sequence, we used the best "bit score" among the resulting hits as the score for that sequence. We applied TBLASTX to the mosquito and honeybee in addition to the *Drosophila* species. We obtained these genome assemblies [46,47] from the UCSC Genome Browser [69], assembly versions anoGam1 and apiMel2.

dN/dS test. We carried out the *dN/dS* test by using PAML 3.14 [34] to compute likelihoods of each sequence alignment under the assumption of either $dN/dS = 1$ or dN/dS estimated by maximum likelihood. Each multiple sequence alignment was pre-processed to make it acceptable to PAML as follows: gaps in the *D. melanogaster* sequence were removed, ends were trimmed so that the sequence only contains complete codons, and in-frame stop codons were changed to gaps in the informant sequences. Additionally, rows (informant species) with more than 50% gapped positions were removed, to reduce the computational cost of marginalizing over such heavily gapped rows.

PAML was then run twice on each alignment, once with fix_omega = 1 and once with fix_omega = 0. The other paramaters, common to both runs, were runmode = 0, seqtype = 1, CodonFreq = 2, model = 0. The tree was specified as shown in Figure 3.

The log-likelihood values computed by the two runs were subtracted to obtain a log likelihood ratio used as the score for the region.

For practical reasons, PAML was not allowed to run for more than one hour on any individual alignment. Cases in which PAML exceeded this time limit, where no informant sequences remained after preprocessing, or otherwise failed were regarded as negatively classified at all cutoffs. This occurred in only 70 of 49,903 cases with 12 flies and 242 of 49,903 cases with the *melanogaster* subgroup informants.

CSF. The CSF metric is based on estimates of the frequencies at which all pairs of codons are substituted between genes in the target species and the informants [5]. First, let us consider computing the score for a pairwise alignment only. Consider the alignment of a putative ORF/exon as two sequences of codons A and B, where A_k is the target codon that aligns to the informant codon B_k at position k in the target codon sequence (position $3k$ in the in-frame target nucleotide sequence). CSF assigns a score to each codon position k where: (1) A_k and B_k are both un-gapped triplets, (2) A_k is not a stop codon, and (3) $A_k \neq B_k$. CSF then sums these scores to obtain an overall score for the sequence.

The score assigned to a codon substitution (a,b) is a log-likelihood ratio indicating how much more frequently that substitution occurs in coding regions than in non-coding regions. Each likelihood compared in this ratio is derived from a Codon Substitution Matrix (CSM), where

$$CSM_{a,b} = \mathrm{P}(\text{informant codon } b \mid \text{target codon } a,\ a \neq b)$$

The entries of the CSM are estimated for each target and informant by counting aligned codon pairs in training data, and then normalizing the rows to obtain the desired conditional probabilities. We train two CSMs, one for which the training data is alignments of known protein-coding genes (CSM^C) and one for which the training data is alignments of random non-coding regions (CSM^N). The score that CSF assigns a codon substitution (a,b) is then

$$\log \frac{CSM^C_{a,b}}{CSM^N_{a,b}}.$$

With multiple informants, CSF uses an *ad hoc* strategy to combine evidence from the informants without double-counting multiple apparent substitutions among extant species that result from fewer evolutionary events in their ancestors. For each target codon position k, CSF assigns a score to codon substitutions between the target and each informant exactly as in the pairwise case, using the appropriate CSMs for each informant. CSF then takes the median of these scores to obtain a composite score for position k, and sums these composite scores to obtain an overall score for the sequence. Note that the median is usually taken on fewer than n pairwise scores, since the pairwise scores are only assigned to ungapped informant codons that differ from the target codon.

RFC. We applied the RFC metric exactly as previously described [4,32]. Briefly, given an alignment of a region of the target genome (*D. melanogaster*), a pairwise score between the target and each informant wass computed as the percentage of target nucleotides that aligned in the same reading frame in the informant (taking the largest such percentage out of the three possible reading frame offsets). With multiple informant species, each species votes +1, −1, or 0 based on a species-specific cutoff on the pairwise RFC score: +1 if the score is above, −1 if the score is below, or 0 if there was no sequence aligned. These votes are then summed to obtain an overall score for the region. The cutoff for each species is chosen by examining the typically bimodal distribution of the score between known coding and non-coding regions, and usually ranges between 70% and 80%.

Sequence conservation metrics. The pairwise sequence conservation metric is simply the percent identity between the target and informant sequences (as a fraction of the target sequence length). For multiple alignments, we assigned a score to each target nucleotide column corresponding to the largest fraction of species having the same nucleotide in that column (plurality), and averaged these scores across the columns of the alignment.

Fourier transform. The Fourier transform metric is an aggregate measure of the three-base periodicities of each nucleotide character in coding sequences [37,70]. First, the DNA sequence is converted into four binary indicator sequences, one for each nucleotide, e.g.

$$u_A(n) = \begin{cases} 1 & \text{if nucleotide } A \text{ occurs at position } n \\ 0 & \text{otherwise} \end{cases}$$

For each nucleotide, a three-base periodicity is then calculated by computing the magnitude of the discrete Fourier transform (DFT) of its indicator sequence at 1/3 frequency, e.g.

$$U_A(1/3) = \left| \sum_{n=1}^{N} u_A(n) e^{-\frac{2\pi i (n-1)}{3}} \right|^2$$

The overall score of the sequence is then computed by summing the contribution of each nucleotide periodicity normalized by the length of the sequence,

$$S = \frac{1}{N}\left(U_A(1/3) + U_G(1/3) + U_C(1/3) + U_T(1/3)\right)$$

We found that the discriminative performance of this metric is identical to that obtained by computing the signal-to-noise ratio of the 1/3 frequency component of the DFT [71]. We chose the former because it has fewer free parameters.

Codon bias. Let a_i be the amino acid translation of codon i. The metric utilizes codon usage vectors C and N for coding and non-coding sequences, where C_i is the likelihood of codon i conditional on amino acid a_i in coding regions, and N_i is the corresponding likelihood for non-coding regions. C_i is estimated from training data by determining the ratio of the number of times codon i occurs in-frame to the total number of times amino acid a_i occurs in-frame; N_i is estimated similarly with an arbitrary frame. To evaluate a given sequence, a total log-likelihood ratio *LLR* is computed by summing $\log \frac{C_i}{N_i}$ for each putative in-frame codon i in the sequence. *LLR* is positive if the codon bias in the given sequence is more similar to the coding regions in the training set than to the non-coding regions, and negative otherwise.

ICMs. We used Glimmer 3.02 [39] to build and evaluate the ICMs. In the training step, we used the build-icm program to estimate parameters for coding and non-coding ICMs. For both models, we used the default *depth* = 6. We found a choice of *width* = 6 improved discrimination over the default setting. The coding ICM was trained with the default *period* = 3 while the non-coding model was constrained to *period* = 1. In the testing step, the coding and non-coding ICMs were used to score the sequences using the glimmer3 program with the *linear* and *multifasta* options.

The ICM metric score was computed as the log-ratio of the coding and non-coding likelihoods.

Z curve. The Z curve score for a sequence of DNA is a linear combination of 189 frame-specific mono-, di-, and tri-nucleotide occurrence frequencies [2,72]. The weights assigned to these frequencies are trained by Fisher linear discriminant analysis on the frequency vectors computed from the coding and non-coding sequences in the training set, which we carried out using MATLAB with default settings.

Hybrid Metrics

We created hybrid metrics by combining the pre-computed scores of the input metrics using linear discriminant analysis (LDA) and a support vector machine (SVM). In both cases, prior to combination, the scores of each input metric were normalized to have zero mean and unit variance across the entire dataset. The normalized scores from each input metric were then used as feature vectors representing each sequence in the dataset.

We trained and applied the hybrid metrics using four-fold cross-validation. We applied LDA with default settings in MATLAB. For SVM, we used SVM^{light} 4.00 [73] with a linear kernel and default cost parameters. We used the prediction confidence computed by the svm_classify program as the SVM hybrid metric score for each sequence.

Multidimensional Scaling

Multidimensional scaling (MDS) takes a high-dimensional matrix of pairwise similarities between items (in our case, metrics), and assigns each item to a point in a low-dimensional space (in our case, two dimensions for visualization), such that the distance between any two points approximately represents the dissimilarity of the corresponding items. We applied MDS to generate the visualization in Figure 2A using the R function cmdscale with default parameters. We defined the similarity between two metrics as $S(i,j) = \mathrm{cor}(R_i, R_j)$, where R_i is the vector of ranks of the known exons according to the scores computed by metric i. For example, if the known exons are ordered in some way E_1, E_2, E_3, and metric i assigns them scores $M_i([E_1, E_2, E_3]) = [0.2, 1.0, -0.5]$, then $R_i = [3,1,2]$.

Supporting Information

Figure S1 Comparison of pairwise comparative metrics with *D. ananassae* as the informant species. Pairwise comparisons using the metrics we studied did not in general outperform the best single sequence metric (Z curve), although CSF and KA/KS achieve higher specificity.
Found at: doi:10.1371/journal.pcbi.1000067.s001 (0.17 MB PDF)

Figure S2 Comparison of alignment depth provided by MULTIZ and Mercator/MAVID alignments. Shown on each plot is the cumulative proportion of regions in our dataset that have a certain number of species aligned (top) and the total branch length of those species (bottom), in the MULTIZ (red) or Mercator/MAVID (blue) alignments. For each region, an informant species was considered to align if at least 50% of the *D. melanogaster* nucleotides were aligned to an informant nucleotide (as opposed to gaps). The total branch lengths for the species aligning to each region were computed by taking the corresponding subtree of the neutral tree shown in Figure 3. In all cases, the MULTIZ alignments tend to align more species than the Mercator/MAVID alignments, consistent with their somewhat higher overall sensitivity (see Figure S3). (These results were generated from static genome alignment sets, and may not be representative of what is possible with the two approaches under different parameter settings.)
Found at: doi:10.1371/journal.pcbi.1000067.s002 (0.23 MB PDF)

Figure S3 Comparison of discovery power provided by MULTIZ and Mercator/MAVID alignments. (Top) The MULTIZ alignments lead to higher sensitivity than the Mercator/MAVID alignments. The Mercator/MAVID alignments can lead to slightly higher specificity, but only at low sensitivities ($<60\%$). (Bottom) The two alignments overall lead to concordant sets of detected exons, with $>93\%$ of exons detected in either alignment detected in both alignments. Although the MULTIZ alignments have higher overall sensitivity, the Mercator/MAVID alignments do uniquely allow the detection of $\sim 1.5\%$ of exons. (These results were generated from static genome alignment sets, and may not be representative of what is possible with the two approaches under different parameter settings.)
Found at: doi:10.1371/journal.pcbi.1000067.s003 (0.31 MB PDF)

Figure S4 Pairwise discovery power for RFC with different informants. More closely related species tend to yield higher sensitivity, while more distant species yield higher specificity.
Found at: doi:10.1371/journal.pcbi.1000067.s004 (0.15 MB PDF)

Table S1 Gene detection rates within Gene Ontology (GO) categories. Each entry shows the percentage of genes with at least one exon detected at a fixed exon sensitivity cutoff for each metric.
Found at: doi:10.1371/journal.pcbi.1000067.s005 (0.08 MB XLS)

Text S1 Information about access to test dataset coordinates, sequences, and alignments, and metric score data.
Found at: doi:10.1371/journal.pcbi.1000067.s006 (0.03 MB DOC)

Acknowledgments

We thank Huy L. Nguyen for informatics assistance; David DeCaprio, Jade Vinson, and James Galagan for helpful discussions regarding SMCRFs; and Alexander Stark and Pouya Kheradpour for additional comments and discussions.

Author Contributions

Conceived and designed the experiments: ML AD MK. Performed the experiments: ML AD. Analyzed the data: ML AD. Contributed reagents/materials/analysis tools: ML AD MR. Wrote the paper: ML MK.

References

1. Fickett JW, Tung CS (1992) Assessment of protein coding measures. Nucleic Acids Res 20: 6441–6450.
2. Gao F, Zhang CT (2004) Comparison of various algorithms for recognizing short coding sequences of human genes. Bioinformatics 20: 673–681.
3. Saeys Y, Rouze P, Van de Peer Y (2007) In search of the small ones: improved prediction of short exons in vertebrates, plants, fungi and protists. Bioinformatics 23: 414.
4. Kellis M, Patterson N, Endrizzi M, Birren B, Lander E (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423: 241–254.
5. Lin MF, Carlson JW, Crosby MA, Matthews BB, Yu C, et al. (2007) Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. Genome Res 17: 1823–1836.
6. Frith M, Bailey T, Kasukawa T, Mignone F, Kummerfeld S, et al. (2006) Discrimination of non-protein-coding transcripts from protein-coding mRNA. RNA Biol 3: 40–48.
7. Liu J, Gough J, Rost B (2006) Distinguishing protein-coding from non-coding RNAs through support vector machines. PLoS Genet 2: e29.
8. Gross S, Russakovsky O, Do C, Batzoglou S (2007) Training conditional random fields for maximum parse accuracy. In: Schölkopf B, Platt J, Hoffman T, eds (2007) Advances in Neural Information Processing Systems 19. Cambridge, MA: MIT Press. pp 529–536.
9. Bernal A, Crammer K, Hatzigeorgiou A, Pereira F (2007) Global discriminative learning for higher-accuracy computational gene prediction. PLoS Comput Biol 3: e54. 10.1371/journal.pcbi.0030054.

10. Vinson JP, DeCaprio D, Pearson MD, Luoma S, Galagan JE (2007) Comparative gene prediction using conditional random fields. In: Schölkopf B, Platt J, Hoffman T, eds (2007) Advances in Neural Information Processing Systems 19. Cambridge, MA: MIT Press. pp 1441–1448.
11. Decaprio D, Vinson JP, Pearson MD, Montgomery P, Doherty M, et al. (2007) Conrad: Gene prediction using conditional random fields. Genome Res 17: 1389–1398.
12. Gross S, Do C, Sirota M, Batzoglou S (2007) Contrast: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction. Genome Biol 8: R269.
13. Eddy S (2005) A model of the statistical power of comparative genome sequence analysis. PLoS Biol 3: e10. doi:10.1371/journal.pbio.0030010.
14. Bergman C, Pfeiffer B, Rincon-Limas D, Hoskins R, Gnirke A, et al. (2002) Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome. Genome Biology 3: research0086.1–research0086.20.
15. Zhang L, Pavlovic V, Cantor CR, Kasif S (2003) Human-mouse gene identification by comparative evidence integration and evolutionary analysis. Genome Res 13: 1190–1202.
16. Margulies EH, Blanchette M, Program NISCCS, Haussler D, Green ED (2003) Identification and characterization of multi-species conserved sequences. Genome Res 13: 2507–2518.
17. Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, et al. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424: 788–793.
18. Margulies EH, Vinson JP, Program NISCCS, Miller W, Jaffe DB, et al. (2005) An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. Proc Natl Acad Sci U S A 102: 4795–4800.
19. Gross S, Brent M (2006) Using multiple alignments to improve gene prediction. J Comput Biol 13: 379–393.
20. Brent M (2005) Genome annotation past, present, and future: How to define an ORF at each locus. Genome Research 15: 1777–1786.
21. Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the Drosophila phylogeny. Nature 450: 203–218.
22. Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, et al. (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature 450: 219–232.
23. Badger JH, Olsen GJ (1999) Critica: coding region identification tool invoking comparative analysis. Mol Biol Evol 16: 512–524.
24. Batzoglou S, Pachter L, Mesirov JP, Berger B, Lander ES (2000) Human and mouse gene structure: Comparative analysis and application to exon prediction. Genome Res 10: 950–958.
25. Korf I, Flicek P, Duan D, Brent M (2001) Integrating genomic homology into gene structure prediction. Bioinformatics 17: S140–S148.
26. Meyer I, Durbin R (2002) Comparative ab initio prediction of gene structures using pair HMMs. Bioinformatics 18: 1309–1318.
27. Alexandersson M, Cawley S, Pachter L (2003) Slam: Cross-species gene finding and alignment with a generalized pair hidden Markov model. Genome Res 13: 496–502.
28. Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW, et al. (2003) Comparative gene prediction in human and mouse. Genome Res 13: 108–117.
29. Mignone F, Grillo G, Liuni S, Pesole G (2003) Computational identification of protein coding potential of conserved sequence tags through cross-species evolutionary analysis. Nucleic Acids Res 31: 4639–4645.
30. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3: 418–426.
31. Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Molecular Biology and Evolution 17: 32–43.
32. Kellis M, Patterson N, Birren B, Berger B, Lander E (2004) Methods in comparative genomics: genome correspondence, gene identification and motif discovery. J Comput Biol 11: 319–355.
33. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389–3402.
34. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555–556.
35. Nekrutenko A, Makova KD, Li WH (2002) The KA/KS ratio test for assessing the protein-coding potential of genomic regions: An empirical and simulation study. Genome Res 12: 198–202.
36. Yang Z, Bielawski J (2000) Statistical methods for detecting molecular adaptation. Trends in Ecology and Evolution 15: 496–503.
37. Anastassiou D (2001) Genomic signal processing. IEEE Signal Processing Magazine 18: 8–20.
38. Akashi H (2001) Gene expression and molecular evolution. Curr Opin Genet Dev 11: 660–666.
39. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27: 4636–4641.
40. Adams M, Celniker S, Holt R, Evans C, Gocayne J, et al. (2000) The genome sequence of Drosophila melanogaster. Science 287: 2185–95.
41. Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, et al. (2005) Comparative genome sequencing of Drosophila pseudoobscura: Chromosomal, gene, and cis-element evolution. Genome Res 15: 1–18.
42. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, et al. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 14: 708–715.
43. Bray N, Pachter L (2004) Mavid: Constrained ancestral alignment of multiple sequences. Genome Res 14: 693–699.
44. Cox T, Cox M (2001) Multidimensional Scaling. Chapman & Hall/CRC.
45. Blanchette M (2007) Computation and analysis of genomic multi-sequence alignments. Annual Review of Genomics and Human Genetics 8: 193–213.
46. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, et al. (2002) The genome sequence of the malaria mosquito Anopheles gambiae. Science 298: 129–149.
47. Honeybee Genome Sequencing Consortium (2006) Insights into social insects from the genome of the honeybee Apis mellifera. Nature 443: 931–949.
48. Zdobnov EM, von Mering C, Letunic I, Torrents D, Suyama M, et al. (2002) Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster. Science 298: 149–159.
49. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. Nat Genet 25: 25–29.
50. Misra S, Crosby M, Mungall C, Matthews B, Campbell K, et al. (2002) Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. Genome Biol 3: 1–0083.
51. Reichhart JM (2005) Tip of another iceberg: Drosophila serpins. Trends Cell Biol 15: 659–665.
52. Jiggins FM, Kim KW (2007) A screen for immunity genes evolving under positive selection in Drosophila. J Evol Biol 20: 965–970.
53. Rasmussen MD, Kellis M (2007) Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. Genome Res 17: 1932–1942.
54. Cherry JL (1998) Should we expect substitution rate to depend on population size? Genetics 150: 911–919.
55. Gillespie JH (1999) The role of population size in molecular evolution. Theor Popul Biol 55: 145–156.
56. Whiteford N, Haslam N, Weber G, Prügel-Bennett A, Essex JW, et al. (2005) An analysis of the feasibility of short read sequencing. Nucleic Acids Res 33: e171.
57. Sundquist A, Ronaghi M, Tang H, Pevzner P, Batzoglou S (2007) Whole-genome sequencing and assembly with high-throughput, short-read technologies. PLoS ONE 2: e484.
58. Markow TA, O'Grady PM (2007) Drosophila biology in the genomic age. Genetics 177: 1269–1276.
59. Zhang MQ (2002) Computational prediction of eukaryotic protein-coding genes. Nat Rev Genet 3: 698–709.
60. Karlin S, Chen C, Gentles AJ, Cleary M (2002) Associations between human disease genes and overlapping gene groups and multiple amino acid runs. Proc Natl Acad Sci U S A 99: 17008–17013.
61. Yu P, Ma D, Xu M (2005) Nested genes in the human genome. Genomics 86: 414–422.
62. ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799–816.
63. Oyama M, Itagaki C, Hata H, Suzuki Y, Izumi T, et al. (2004) Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. Genome Res 14: 2048–2052.
64. Frith MC, Forrest AR, Nourbakhsh E, Pang KC, Kai C, et al. (2006) The abundance of short proteins in the mammalian proteome. PLoS Genet 2: e52.
65. Galindo MI, Pueyo JI, Fouix S, Bishop SA, Couso JP (2007) Peptides encoded by short ORFs control development and define a new eukaryotic gene family. PLoS Biol 5: e106. 10.1371/journal.pbio.0050106.
66. Kondo T, Hashimoto Y, Kato K, Inagaki S, Hayashi S, et al. (2007) Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. Nat Cell Biol 9: 660–665.
67. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696–704.
68. Clamp M, Fry B, Kamal M, Xie X, Cuff J, et al. (2007) Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci U S A 104: 19428–19433.
69. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome Res 12: 996–1006.
70. Voss RF (1992) Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Phys Rev Lett 68: 3805–3808. 10.1103/PhysRevLett.68.3805.
71. Tiwari S, Ramachandran S, Bhattacharya A, Bhattacharya S, Ramaswamy R (1997) Prediction of probable genes by Fourier analysis of genomic sequences. Comput Appl Biosci 13: 263–270.
72. Zhang C, Wang J (2000) Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve. Nucleic Acids Research 28: 2804–2814.
73. Joachims T (1999) Advances in Kernel Methods - Support Vector Learning, MIT Press, chapter Making Large-Scale SVM Learining Practical.

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

Conservation of small RNA pathways in platypus

Elizabeth P. Murchison,[1] Pouya Kheradpour,[2] Ravi Sachidanandam,[1] Carly Smith,[1] Emily Hodges,[1] Zhenyu Xuan,[1] Manolis Kellis,[2,3] Frank Grützner,[4] Alexander Stark,[2,3] and Gregory J. Hannon[1,5]

[1]*Watson School of Biological Sciences, Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA;* [2]*Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA;* [3]*Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02141, USA;* [4]*Discipline of Genetics, School of Molecular & Biomedical Science, The University of Adelaide, Adelaide, 5005 SA, Australia*

Small RNA pathways play evolutionarily conserved roles in gene regulation and defense from parasitic nucleic acids. The character and expression patterns of small RNAs show conservation throughout animal lineages, but specific animal clades also show variations on these recurring themes, including species-specific small RNAs. The monotremes, with only platypus and four species of echidna as extant members, represent the basal branch of the mammalian lineage. Here, we examine the small RNA pathways of monotremes by deep sequencing of six platypus and echidna tissues. We find that highly conserved microRNA species display their signature tissue-specific expression patterns. In addition, we find a large rapidly evolving cluster of microRNAs on platypus chromosome XI, which is unique to monotremes. Platypus and echidna testes contain a robust Piwi-interacting (piRNA) system, which appears to be participating in ongoing transposon defense.

[Supplemental material is available online at www.genome.org. The microRNA sequence data from this study have been submitted to miRBase under accession nos. MI0006658–MI0007078 (http://microrna.sanger.ac.uk/sequences/).

Initially regarded as a scientific hoax by European biologists, the platypus *(Ornithorhynchus anatinus)*, along with four species of echidna (family Tachyglossidae), represent the only remaining members of the basal mammalian clade of prototheria. These animals diverged from remaining mammalian lineages around 166 million years ago (Bininda-Emonds et al. 2007). Morphologically, monotremes display a mixture of mammalian, reptilian, and unique features. As mammals, monotremes have fur, but they share with reptiles a cloaca, the single external opening for both the gastrointestinal and urogenital systems. Monotremes lay eggs that are nourished in part through a placental structure before deposition, but they provide milk to their hatched young through mammary glands. Monotreme embryos show meroblastic rather than holoblastic cleavage, which is more typical of reptiles and fish than of mammals.

The platypus is a specialized semiaquatic carnivore that displays several unique features. Its leathery bill is equipped with specialized neurons, which serve as electrosensors, and male platypuses have spikes on their hind legs, which are used in an elaborate venom delivery system. The platypus karyotype is also very unusual, with small chromosomes reminiscent of reptilian microchromosomes and multiple sex chromosomes (5X, 5Y) that form a chain during meiosis (Grutzner et al. 2004; Rens et al. 2004). Remarkably, the unique mix of mammalian, reptilian, and specialized features of monotremes is also revealed by its genomic sequence. This together with its phylogenetic position makes it a uniquely placed comparative tool for studying mammalian evolution (Warren et al. 2007).

RNAi pathways are deeply conserved and can be traced at least as far back as the divergence of prokaryotes and eukaryotes, ~2.7 billion years ago. Individual eukaryotic lineages have adapted these pathways in different ways, although most use RNAi generally as both a genome defense and a gene regulatory mechanism. The RNAi machinery uses small RNAs of between 18 and 33 nucleotides (nts) in length as specificity determinants to regulate RNA stability and gene expression, through both post-transcriptional and transcriptional modes.

Two endogenous RNAi pathways have been described in mammals, the microRNA (miRNA) pathway and the Piwi-interacting (piRNA) pathway. The miRNA pathway uses small RNAs, around 18–24 nt in length, to regulate genes post-transcriptionally (Bartel 2004). miRNA duplexes are liberated from genome-encoded hairpins by successive processing by RNase III enzymes RNASEN (also known as Drosha) and DICER1 (Bartel 2004). Incorporation into complexes containing members of the EIF2C (also known as Argonaute) family of RNaseH-like enzymes coincides with the release of the second, or star, strand, leaving the mature strand to direct sequence-specific silencing events. miRNAs act in diverse biological networks, with many having roles in reinforcing cell-type and tissue identity (Plasterk 2006). Some families of miRNA genes have deeply conserved expression patterns, extending throughout animal lineages, whereas others are more rapidly evolving and lineage specific (Plasterk 2006). piRNAs are a second class of small RNAs, slightly larger in length than miRNAs (~25–33 nt), which interact with the Piwi clade of Argonaute-family proteins. piRNAs show restricted expression patterns, being prominent in gonad, with mammalian piRNAs thus far being restricted to the germline compartment (Aravin et al. 2007a). Genetic and sequence analyses of piRNAs in *Drosophila* have revealed a role for piRNAs in transposon control via a pathway that uses small RNA generating loci as transposon sensors (Brennecke et al. 2007; Gunawardane

[5]Corresponding author.
E-mail hannon@cshl.edu; fax (516) 367-8874.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.073056.107.

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

et al. 2007). In post-pubescent mammalian testis, piRNAs arise from dense strand-biased clusters whose biogenesis and functions remain largely unknown, although roles in transposon defense are supported for piRNAs that are expressed in germ cell precursors earlier in development (Aravin et al. 2007a).

Using the platypus genome sequence, we have performed a comparative analysis of small RNA pathways in platypus, echidna, chicken, and eutherian mammals and have found deep conservation of many small RNA species. We also note miRNAs with mammalian or chicken/monotreme lineage-specific conservation as well as a surprisingly extensive set of monotreme-specific and platypus-specific small RNAs. Strikingly, our analysis implicates small RNA functions in monotreme reproduction, as we find large clusters of germline-expressed fast-evolving miRNA loci on one of the X chromosomes in platypus and evidence for ongoing transposon defense by piRNAs in the adult platypus testis.

Results

Comparative analysis of RNAi components in platypus

The completion of the platypus genome sequence offers an opportunity to examine the conservation of small RNA pathways in prototherian mammals (Warren et al. 2007). Indeed, homologs of key RNAi components *DICER1*, *RNASEN*, and *EIF2C/Argonaute* family members 1 through 4 could be readily identified in platypus (Fig. 1A; Supplemental Table 1). Moreover, synteny of the *EIF2C/Argonaute* gene family was conserved between platypus and eutherian mammals, with platypus *EIF2C1, EIF2C3*, and *EIF2C4* closely linked on Ultracontig 472, and *EIF2C2* on chromosome 4. Platypus also had obvious orthologs for PIWI family members *PIWIL1, PIWIL2*, and *PIWIL4* but appeared to be missing *PIWIL3* (Fig. 1B; Supplemental Tables 1 and 2). As *PIWIL3* is only found in a subset of eutherian mammals and could not be found in a marsupial genome (*Monodelphis domestica*), it is likely that this gene emerged after the divergence of the eutherian and prototherian lineages. Interestingly, the DDH motif, necessary for the endoribonuclease activity of EIF2C/Argonaute enzymes, was conserved in platypus EIF2C2 and all three PIWI proteins, as well as in EIF2C3, even though mouse and human orthologs of EIF2C3 appear to be noncatalytic (Liu et al. 2004; Meister et al. 2004). The conservation of this motif in the EIF2C/Argonaute clade throughout mammalian evolution is mysterious, as the biological significance of Argonaute-mediated RNA cleavage in mammals is unclear (Yekta et al. 2004; Davis et al. 2005; Yu et al. 2005; O'Carroll et al. 2007). Overall, our findings suggest that the platypus maintains a fully functional set of RNAi enzymes, the evolution of which reconstructs the phylogeny of the vertebrate lineage.

Identification of conserved and novel miRNAs in monotremes

To determine the extent of conservation of miRNA pathways in monotremes, we cloned and deep-sequenced the 18- to 24-nt small RNA content of six tissues (brain, kidney, heart, lung, liver, and testis) from adult male platypus *(Ornithorhynchus anatinus)* and adult male short-beaked echidna (*Tachyglossus aculeatus)*. A total of 2,982,604 unique small RNAs were identified from these twelve libraries that mapped to the platypus genome: 1,409,428 small RNAs from platypus and 1,532,352 small RNAs from echidna. We devised a miRNA discovery pipeline combining a computational heuristic with manual curation of miRNAs (see Methods). This approach yielded 332 miRNA candidates, 149 orthologs of known miRNAs, and 183 novel miRNAs (Supplemental Table 3). Three of the novel miRNAs (miR-1329, miR-1397, and miR-1416) were also identified in chicken (*Gallus gallus*), whereas 180 miRNAs were apparently unique to monotremes.

The platypus genome sequence has revealed commonalities with both therian mammals and nonmammalian vertebrates (Warren et al. 2007). Monotremes share at least 10 miRNA genes with eutherian mammals (mouse/human) but not chicken, 4 with chicken but not mouse or human, and one with fish (*Dania rerio, Fugu rubripes*, and *Tetraodon nigroviridis*) but not mouse, human, or chicken (Warren et al. 2007) (Fig. 2; Supplemental Table 3). Although it is possible that some miRNAs that have apparently been lost during vertebrate evolution are missing because of incomplete miRNA identification or genome sequence, it is

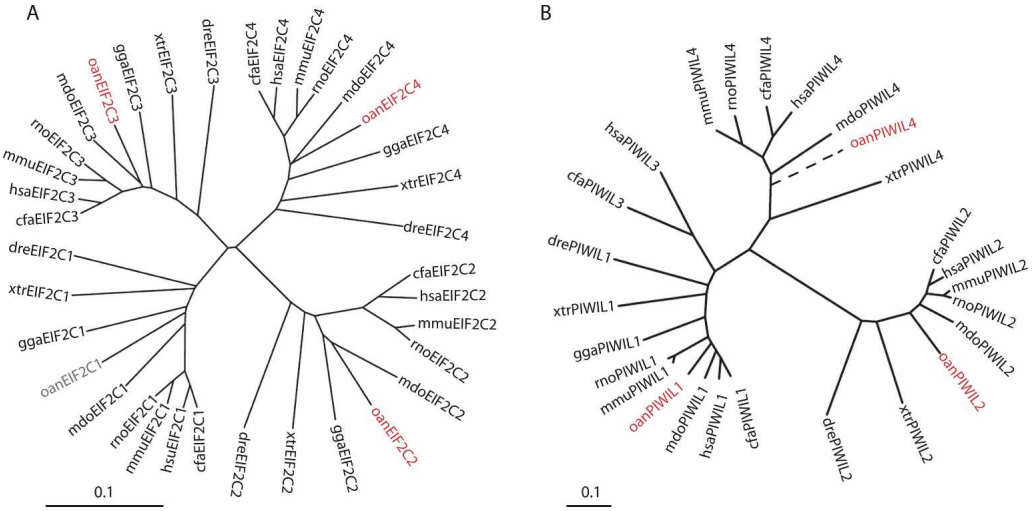


Figure 1. RNA interference genes in platypus. Phylogenetic trees of *EIF2C* (*A*) and *PIWI* (*B*) families in multiple vertebrate species. The platypus *PIWIL4* (*oanPIWIL4*) sequence was incomplete, and a partial coding sequence was used to place this gene on the tree (broken line) (Supplemental Table 2). Platypus genes are highlighted in red. Scale: 0.1 = 0.1 base substitutions expected. (*oan*) *Ornithorhynchus anatinus;* (*cfa*) *Canis familiaris;* (*hsa*) *Homo sapiens;* (*mmu*) *Mus musculus;* (*rno*) *Rattus norvegicus;* (*mdo*) *Monodelphis domesctica;* (*gga*) *Gallus gallus;* (*xtr*) *Xenopus tropicalis;* (*dre*) *Danio rerio.*

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

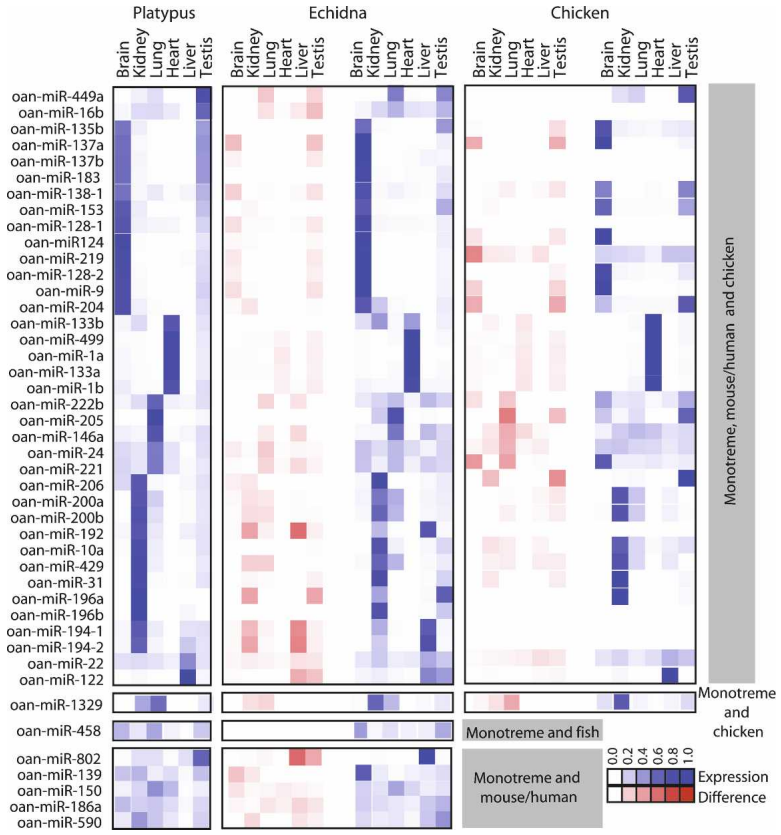


Figure 2. Conservation of microRNA expression patterns in monotremes. Comparison of platypus/echidna and platypus/chicken miRNA tissue expression profiles. miRNA orthology was determined by perfect match (platypus and echidna) or at least 16-nt identity in the mature miRNA (mouse, human, chicken, and fish). The expression score represented by blue squares is a ratio of the number of normalized reads in that tissue versus all tissues in that organism, and the difference score represented by red squares is the difference between the confidence intervals of the expression ratios for each platypus/echidna or platypus/chicken pair of miRNAs. A subset of miRNAs shared between monotremes, mouse/human, and chicken; monotreme and chicken; monotreme and fish; and monotreme and mouse/human are shown. miRNAs with the most distinct tissue-specific expression patterns are shown here clustered by tissue expression; the complete data sets can be found in Supplemental Figs. 1–4.

likely that some miRNA gain and loss reflects genuine evolutionary adaptation.

miRNA expression patterns have been extensively characterized in several vertebrates (Chen et al. 2005; Wienholds et al. 2005; Ason et al. 2006; Darnell et al. 2006; Kloosterman et al. 2006a,b; Takada et al. 2006; Xu et al. 2006; Landgraf et al. 2007). Using miRNA cloning frequency data, we determined miRNA signatures for platypus and echidna brain, heart, lung, kidney, liver, and testis (Fig. 2; Supplemental Fig. 1). Although expression patterns of miRNAs tend to be conserved across vertebrates, variation has been observed in some cases (Ason et al. 2006). We found strong conservation of expression domains and expression levels of orthologous miRNAs among monotremes (platypus and echidna), among mammals (platypus and mouse/human), and between platypus and chicken (Fig. 2; Supplemental Figs. 1–4). Interestingly, the tissue with the greatest differences in miRNA expression level between the aquatic platypus and the terrestrial, predominantly ant-eating echidna was liver, perhaps reflecting the unique dietary adaptations of these two monotremes (Fig. 2; Supplemental Fig. 1).

The platypus and echidna lineages are thought to have diverged ~16.5 million years ago, although recent re-examination of fossil evidence suggest the possibility of an earlier divergence (Warren et al. 2007; Rowe et al. 2008). Of the 180 novel monotreme miRNA candidates identified through our screening approach, 95 were found only in platypus, 4 only in echidna (although their mature sequence was perfectly conserved in the platypus genome), and 81 from both platypus and echidna (Fig. 3A; Supplemental Table 3). Although we used stringent criteria for determining platypus and echidna orthology (perfect identity) and likely missed a number of echidna orthologs, it is striking that such a large number of miRNAs were found only in platypus. This suggests either that platypus and echidna indeed diverged earlier than previously estimated (Rowe et al. 2008) or that a subset of platypus miRNAs have evolved rapidly.

Identification of a novel X-linked miRNA cluster in monotremes

We investigated tissue expression patterns of a subset of the novel monotreme miRNAs by comparing total normalized cloning frequencies across tissues (Fig. 3B). Strikingly, a large number of miRNAs in both the monotreme- and platypus-specific sets was almost exclusively expressed in testis (Fig. 3B; Supplemental Table 3). Indeed, closer inspection revealed that 92 of the monotreme-specific miRNAs (51%), which constituted the majority of the highly expressed species (Warren et al. 2007), fell into seven testis-expressed miRNA groups (miR-1347, miR-1376, miR-1414, miR-1419, miR-1420, miR-1421, and miR-1422) with related precursor hairpins (Fig. 3C). miRNAs were named based on the classification of their precursors, that is, miRNAs processed from a miR-1421-related hairpin were called miR-1421a through miR-1421am.

The testis-expressed miRNAs arose from nine large miRNA clusters on chromosome X1 and contigs 1754, 7160, 7359, 8388, 11,344, 22,847, 198,872, and 191,065. Physical mapping of six of these contigs placed five on the long arm of chromosome X1, raising the possibility that the testis miRNAs are derived from a single large cluster on chromosome X1 (Fig. 3D). The miRNA clusters are also related at the sequence level and appear to have arisen from a complex series of tandem duplications.

One miRNA cluster on chromosome X1 is ~40 kb long and encodes at least 27 mature miRNAs (Fig. 3E; Supplemental Table 3). Although the miRNAs arising from this cluster are from highly related precursors, they are cloned at a range of frequencies (Fig. 3E; Supplemental Table 3), suggesting that there is differential regulation of miRNA expression, processing, or stability. Interestingly, although most miRNAs in this cluster are predominantly cloned from platypus testis, some are also detectable in

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

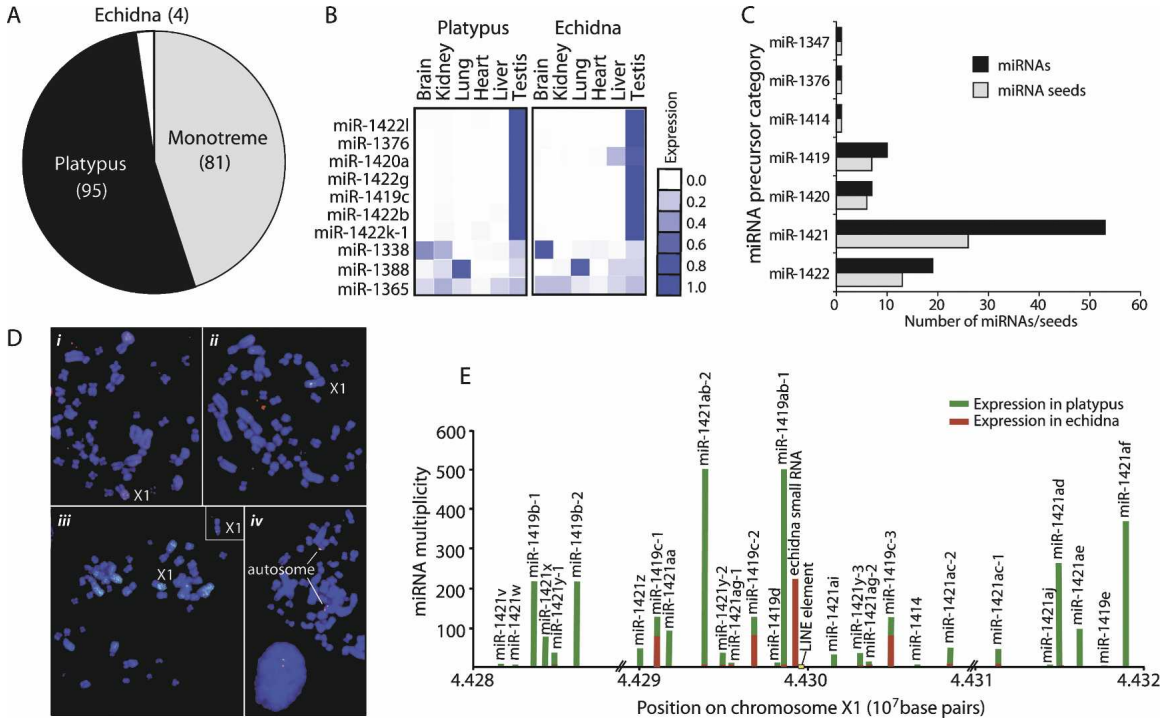


Figure 3. (Continued on next page)

echidna testis, and one small RNA mapping to this cluster is almost entirely restricted to echidna (Fig. 3E). This echidna small RNA is highly related to miR-1421 miRNAs but was not predicted as a miRNA because of the absence of a complete hairpin precursor in the platypus genome. This small RNA may be a miRNA that has been lost in platypus because of the insertion of a LINE element into the miRNA precursor since the divergence of platypus and echidna (Fig. 3E).

miRNAs are often classified into families on the basis of identity at the "seed" heptamer region at nucleotide positions 2–8 of the mature miRNA. The seed sequence participates in miRNA target recognition and binding, and 5′ end miRNA processing must be very precise to ensure the production of the correct seed. Interestingly, although the 92 testis-expressed monotreme miRNAs share only seven related precursor sequences, they generate a great diversity of seed sequences (55 unique seeds) (Fig. 3C; Supplemental Table 3). Indeed, the 51 miRNAs processed from miR-1421-related precursors give rise to 26 unique seed sequences (Fig. 3E).

To understand whether the unexpectedly high diversity of seed sequences in testis miRNAs arose from sequence divergence in the seed region or irregular processing, we mapped both the mature and star sequences for each of 53 miR-1421-related miRNAs on a multiple sequence alignment (Fig. 3F). Interestingly, the sites of processing are slightly offset for some miR-1421 miRNAs (Fig. 3F). In addition, although the mature miRNA is usually processed from the 5′ arm and the less frequently cloned star from the 3′ arm of the miR-1421 precursor, in 6 cases this orientation was switched, thus generating seed diversity (Fig. 3F).

The seed is usually the most highly conserved region of a miRNA hairpin. To examine sequence conservation in miR-1421 hairpins, we calculated percent identity to a consensus for each nucleotide. Overall, the miR-1421 hairpins were highly related and shared on average 82% identity across the entire hairpin (Fig. 3F). However, the most highly conserved positions were in the 5′ stem of the hairpin and 3′ end of the 5′ arm mature miRNA. Interestingly, the seed region of the 5′ arm mature miRNA is considerably less conserved, with an average of 76% identity with the consensus, than its foldback region at the 3′ end of the star sequence, which has an average of 88% identity (Fig. 3F). This is probably at least partly due to the predominance of A and G nucleotides in the seed region, both of which can base pair with uracil residues, and suggests that these miRNAs are exploring sequence space within the constraints of hairpin architecture (Fig. 3F).

Overall, the lack of conservation in processing site, mature strand choice, and seed sequence, as well as the striking divergence in miRNA repertoire between platypus and echidna suggest that testis miRNAs are fast evolving and may regulate a diverse set of targets.

piRNAs in platypus

piRNAs are a class of germline-expressed small RNAs that interact with Piwi proteins (Aravin et al. 2007a). Many piRNAs function as the specificity determinants of an RNA-based innate immune system that protects the germline from transposable elements and other nucleic acid parasites (Aravin et al. 2007a). In eutherian mammals, piRNAs are ~26–30 nt long, tend to start with a uracil, and are abundantly expressed from discrete genomic intervals in testis (Aravin et al. 2007a). Radioactive end labeling of total RNA from platypus, echidna, and mouse testes revealed an abundant class of ~30-nt piRNAs in monotreme testis (Fig. 4A).

To investigate the relationships between the Piwi/piRNA pathways of monotremes and eutherians, we cloned and deeply sequenced RNAs sized between 25 and 33 nt from platypus testis.

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

F

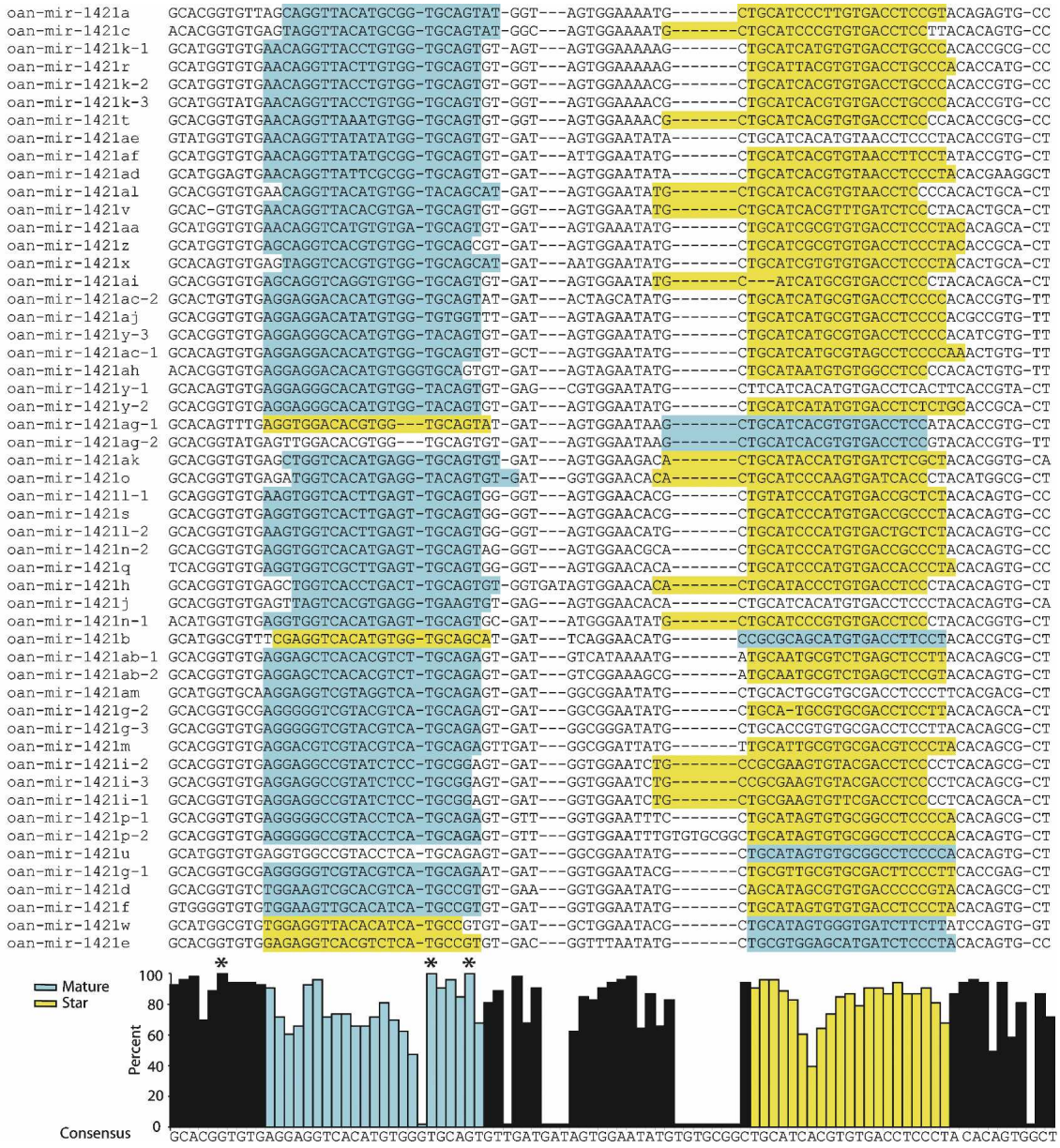


Figure 3. Novel microRNAs in monotremes. (*A*) Pie graph of monotreme species distribution of 180 novel miRNAs. Eighty-one miRNAs were sequenced in both platypus and echidna small (gray); 95 were only in platypus (black); and four were cloned only in echidna (white), although their sequences were present in the platypus genome. (*B*) Novel miRNA tissue expression profiles. The expression score represented by blue squares is a ratio of the number of normalized reads in that tissue versus all tissues in that organism. Expression data for 10 of the most highly expressed novel miRNAs is shown. (*C*) Testis miRNA classification. Ninety-two novel miRNAs were highly expressed in testis and were derived from one of seven related hairpin precursors. The number of miRNAs and miRNA seed sequences derived from each precursor category is shown. (*D*) Five out of nine contigs containing testis-expressed novel miRNAs map to X1q. Fluorescence in situ hybridization with BAC and Fosmid clones containing miRNA clusters on male platypus metaphase chromosomes. (*i*) BAC 175h10 from contig 8388 (green) and BAC 273L19 from contig 7359 (red) on the long arm of X1. (*ii*) Colocalization of BAC 175h10 from contig 8388 (green) and BAC 272K3 from contig 22,847 (red) on X1q. (*iii*) Localization of Fosmid 0879F15 from contig 7160 (green) on X1q. (*Insert*) Localization of Fosmid 1002L21 (contig 11,344) on X1q. (*iv*) BAC clones 686C10 (green) and 802P20 (red) from contig 1754 map to a metacentric autosome. (*E*) Novel testis-expressed chromosome X1 miRNA cluster. Only small RNAs that map to the genome three times or less are shown. The cloning frequency divided by the number of genome matches for each miRNA is represented by a bar. Platypus miRNAs are plotted in green, echidna in red; platypus and echidna bars are independent and not additive. Gaps in the genome assembly are shown by //. A LINE element insertion that disrupted an echidna small RNA precursor is represented in yellow. (*F*) Orientation and sequence conservation of 53 miRNAs derived from the miR-1421 group of precursors. Multiple sequence alignment of 53 miR-1421 precursors with the sequence of the mature miRNA highlighted in blue and the miRNA star in yellow. The percent of nucleotides in the alignment that match the consensus for each position is graphed below. Perfectly conserved nucleotides are marked with an asterisk (*).

Seventy-two percent of uniquely mapping RNAs in this size range could be assigned to one of 50 large intervals in the platypus genome (Supplemental Table 4). These appeared similar to piRNA clusters observed in eutherian mammals, being an average of ~30 kb long and exhibiting a strong strand bias for small RNA production. Based on these characteristics and their 5′ U bias, we

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

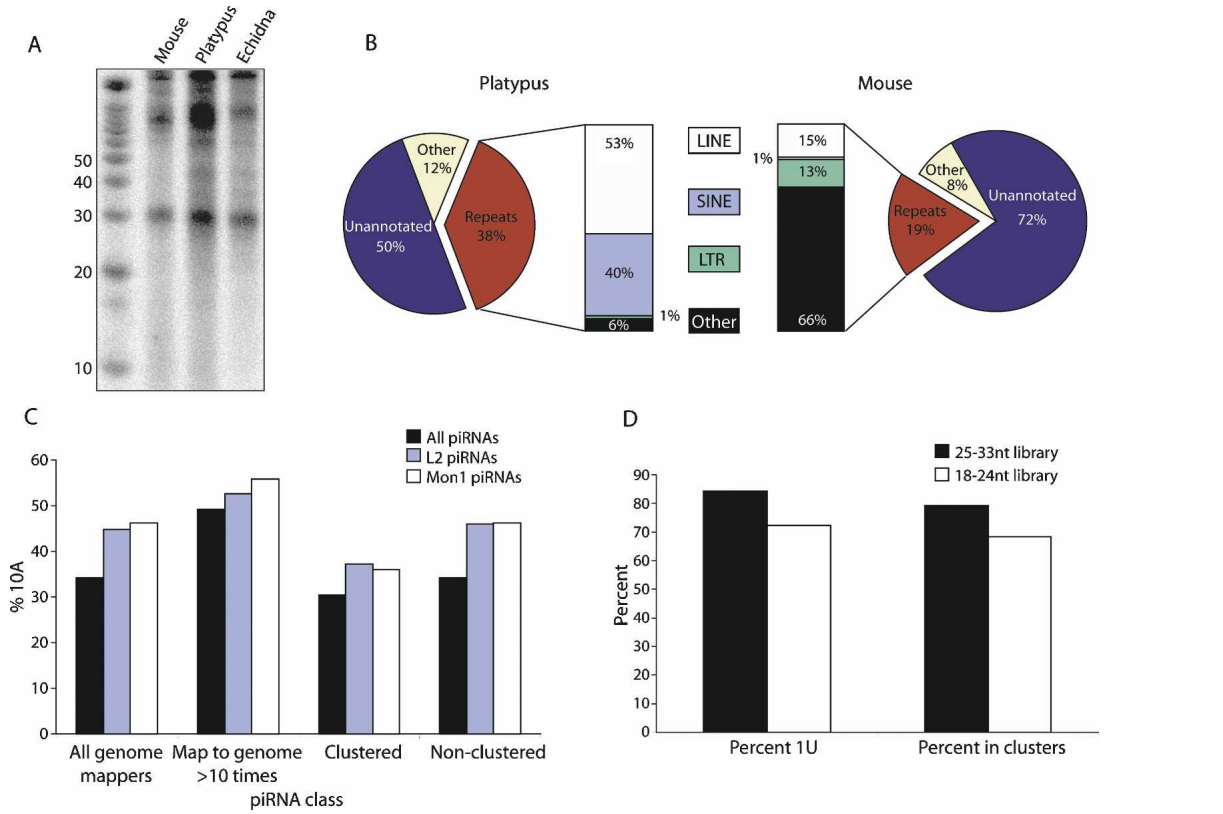


Figure 4. piRNAs in platypus. (*A*) Total RNA labeling from mouse, platypus, and echidna testis reveals a distinct ~30-nt piRNA species in monotremes. RNA was dephosphorylated, end labeled with ^{32}P, and run on a denaturing 15% polyacrylamide gel. Sizes of markers (*left*) are in nucleotides. (*B*) Comparison of platypus and mouse piRNA libraries. A platypus piRNA library was generated by cloning small RNAs in a 25- to 33-nt size window from total platypus testis RNA. Sequences from this library and from an equivalent mouse piRNA library (Girard et al. 2006) were classed as repeat (red), unannotated (blue), or other (beige). The repeat piRNAs were broken down into classes of repeat (bars). (*C*) 10A profiles for platypus piRNAs. An A at piRNA position 10 is a signature of an active piRNA amplification system. "Clustered" piRNAs are defined as those that map to the genome once and fall into one of 50 piRNA clusters (Supplemental Table 4), while "non-clustered" piRNAs are those that map to the genome once but do not fall into a cluster. 10A percentages for all piRNAs are in black, those that are derived from LINE2 elements in blue, and those from Mon1 SINE elements in white. (*D*) Characteristics of "small" piRNAs. "Percent 1U" includes all small RNAs that match the genome; "percent in clusters" includes only small RNAs that map only once to the genome.

conclude that these species are likely platypus piRNAs. As is the case in eutherian mammals, no piRNA clusters were observed on X chromosomes in platypus; however, as the platypus genome is incomplete and many of the platypus piRNA clusters are located on unmapped contigs, we cannot rule out the possibility that platypus has piRNA clusters on sex chromosomes (Supplemental Table 4).

Platypus piRNAs were annotated based on genome mapping (Supplemental Table 5). In mouse, piRNAs expressed in prepubertal testis are repeat-rich and have a function in transposon control, but piRNAs expressed in adult testis are repeat poor and have no known function (Aravin et al. 2007a,b). Interestingly, the platypus piRNA library contained a much larger proportion of repeat-associated species than an equivalent piRNA library from adult mouse testis (Girard et al. 2006) (Fig. 4B). Furthermore, the repeat composition of adult platypus and mouse repeat-associated piRNAs was quite distinct. Whereas in platypus the majority of repeat-associated piRNAs were from LINE and SINE elements, these repeats only made up 16% of the repeat-derived adult mouse piRNAs (Fig. 4B). As LINE and SINE elements, specifically L2 and Mon1, are thought to be active in platypus (Warren et al. 2007), we speculated that perhaps a difference in transposon expression may underlie the differential contribution of these repeat elements to the piRNA populations of adult testis in monotremes and eutherian mammals.

Enrichment for A residues at position 10 of piRNAs is a signature of their participation in ongoing transposon defense (Aravin et al. 2007a; Brennecke et al. 2007). In other systems, piRNA populations with a 10A bias are secondary piRNAs and largely arise from transcripts from active transposons (Brennecke et al. 2007). If LINE and SINE elements are active in platypus, a higher proportion of 10A piRNAs might be expected from these elements than from the bulk piRNA population. In accord with this prediction, the percentage of 10A was significantly greater in L2 and Mon1 piRNAs (Fig. 4C; $P < 10^{-125}$). We also reasoned that the more times a repeat-derived piRNA matched the genome, the greater the likelihood that it is derived from an active transposon, and thus greater the likelihood of 10A. To test whether we could see this effect in the platypus piRNA population, we calculated the 10A bias for piRNAs that mapped to the genome at least ten times. Strikingly, 10A was at 49% in this population and increased to 53% and 56% in L2 and Mon1 piRNAs, respectively (Fig. 4C; $P < 10^{-4}$). Moreover, piRNAs that mapped once to the genome but lay outside of the 50 piRNA clusters (Supplemental Table 4), and thus presumably represented individual active transposons recognizable through their divergent sequences, had

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

a higher proportion of 10A, particularly if they were derived from L2 or Mon1 elements (Fig. 4C). Overall, this analysis provides evidence that piRNA pathways in platypus are active in transposon defense against L2 and Mon1 elements, and this may underlie the high proportion of LINE and SINE piRNAs in platypus.

One characteristic feature of piRNAs is that they are generally larger than miRNAs and siRNAs. However, a surprisingly large proportion of the platypus testis 18- to 24-nt library was composed of small RNAs with piRNA-specific properties. Indeed, 68% of small RNAs in this library that mapped to the genome once were derived from one of the 50 piRNA clusters (Fig. 4D). Furthermore, 79% of small RNAs in this library that mapped to the genome at least once had a U at position 1 (Fig. 4D). Further studies will be required to determine whether this small piRNA class has unique biogenesis or Piwi binding characteristics as compared with the abundant, larger piRNA species.

Discussion

Overall, these studies indicate that the RNAi machinery and its functions in genome defense and gene regulation are conserved in monotremes. Cloning and deep sequencing of miRNAs and piRNAs in monotremes have revealed small RNAs and regulatory systems similar to those found in other mammals, as well as monotreme-specific components.

In general, we found conservation of both sequence and expression patterns of miRNAs in monotremes, chicken, and eutherian mammals. Some platypus miRNAs had unique conservation patterns. For example, orthologs of platypus miR-458 were found only in fish, but not in chicken or mammals. We also identified four miRNAs that are shared between platypus and chicken but that are not shared with other mammals (Fig. 2; Supplemental Table 3). As these miRNAs appear to have been lost from some lineages, they are presumably involved in biological functions that have either become redundant or obsolete in some branches of mammalian evolution.

We have identified 180 novel miRNA candidates unique to platypus and echidna, including at least one large miRNA cluster on platypus chromosome X1. These clustered miRNAs were almost exclusively expressed in testis among the six tissues we analyzed. There are several examples of such expanded fast-evolving testis-expressed miRNA clusters on X chromosomes in other mammals. For instance, the *MIRN506* to *MIRN514* cluster on the X chromosome of humans and other primates contains 10 rapidly evolving miRNAs whose testis-specific expression and developmental regulation suggest a role in spermatogenesis (Bentwich et al. 2005; Zhang et al. 2007). Likewise, the *Mirn743* to *Mirn465* cluster is a testis-expressed cluster of 19 miRNAs on the rodent X chromosome (Yu et al. 2005; Ro et al. 2007). A large miRNA cluster has also been described on the X chromosome of the marsupial opossum *Monodelphis domestica* (Devor and Samollow 2007). Male reproduction is generally fast-evolving in mammals, and genes on sex chromosomes, present as single copy in males, are particularly susceptible to positive selection for male reproductive success (Wyckoff et al. 2000; Delbridge and Graves 2007). Identification of miRNA target genes and analysis of miRNAs in additional vertebrate groups will illuminate the significance of such fast-evolving miRNA clusters in spermatogenesis.

The testis-expressed miRNAs in monotreme testis generate at least 92 mature miRNAs with 55 unique seed sequences. Although most of these seeds are novel, seven of them are shared with miRNAs in mouse, human, or chicken (Supplemental Table 3). Interestingly, six of the shared seeds are found in uniquely mammalian clustered miRNAs, which tend to be expressed in embryonic, placental, and reproductive tissues. These include a subset of miRNAs in the mouse *Mirn290* to *Mirn295* cluster, *Mirn467a*, *Mirn764*, and members of the human *MIRN512* to *MIRN373* cluster (Houbaviy et al. 2003, 2005; Bentwich et al. 2005). It is unclear whether these seed similarities imply functional or evolutionary relatedness.

Monotremes are unique among mammals for their multiple sex chromosomes, which form a chain during meiosis (Grutzner et al. 2004; Rens et al. 2004). It is unclear whether the ten paired X and Y chromosomes of monotremes undergo meiotic sex chromosome inactivation and sex body formation as is observed in eutherian mammals and marsupials (Handel 2004; Namekawa et al. 2007). Abundant testis expression of at least one X1-linked miRNA cluster in monotremes suggests that either this cluster is expressed in premeiotic or somatic cells of the testis or that sex chromosomes are not completely silenced in monotremes.

The testis-clustered miRNAs identified in platypus and echidna are processed irregularly. In some cases, the mature strand and star strands are switched, and the RNase III processing sites are slightly offset between individual hairpins (Fig. 3F). These differences may reflect subtle differences in hairpin structure or sequence. Alternatively, monotremes may possess unique components of the miRNA processing machinery that are able to process miRNA hairpins in an unusual manner.

The platypus genome encodes three proteins of the PIWI family involved in piRNA biogenesis and function, PIWIL1, PIWIL2, and PIWIL4 (Fig. 1B). The conservation pattern and longer branch length of *PIWIL2* suggest that it might encode the ancestral PIWI protein in vertebrates and that subsequent PIWI family members may have arisen by gene duplication events. This theory is consistent with observations that PIWIL2 binds a smaller size range of piRNAs, more akin to the piRNAs found in invertebrates, and that it is the only PIWI protein in mammals that is expressed in both the embryonic and adult gonad (Kuramochi-Miyagawa et al. 2001, 2004; Aravin et al. 2006).

Although platypus piRNAs are organized in large strand-biased clusters similar to those found in other mammals, the piRNA system appears to be actively engaged in combating transposon activity in mature platypus testis. This is in contrast to piRNAs in adult testis of other mammals, which are depleted of repeat sequences and have no known function (Aravin et al. 2006; Girard et al. 2006; Grivna et al. 2006; Lau et al. 2006; Watanabe et al. 2006). The piRNAs of platypus adult testis have features of both fish and eutherian piRNA systems (Houwing et al. 2007) and suggest that mammalian piRNA clusters that are expressed late in germ cell development may have progressively lost repeat content as they acquired novel functions (Aravin et al. 2007b; Carmell et al. 2007).

The small RNA complement of platypus has highlighted not only unique and conserved roles for miRNAs and piRNAs in monotremes but also provides insight into the evolution of small RNA pathways in mammals.

Methods

Tree building

The phylogenetic trees of *EIF2C* and *PIWI* families in multiple species were constructed using the Phylip package (version 3.67) (Felsenstein 1997). The full-length coding region of each gene

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

was used to calculate the distance, and a neighbor-joining method was used to obtain the topology of the tree.

Small RNA cloning, sequencing, and annotation

RNA was isolated by a standard TRizol method from snap-frozen tissue of adult male platypus *(O. anatinus)*, echidna (*T. aculeatus),* and chicken (*G. gallus*) heart, liver, lung, kidney, brain, and testis. Animals (platypus and echidna) were captured at the Upper Barnard River, New South Wales, Australia, during breeding season (AEC permit no. S-49-2006 to F.G.). miRNA (18–24 nt) and piRNA (25–33 nt) fractions of total RNA were used to clone libraries as described (Brennecke et al. 2007). Small RNA libraries were sequenced on the Illumina 1G sequencer. In the case of miRNA libraries, 3′ linkers were clipped using a dynamic alignment algorithm, which required perfect matches at positions 25–30 but allowed mismatches toward the end of the read. The piRNA library was uniformly clipped to 26 nt in length. Platypus and echidna small RNAs with perfect matches to the platypus genome were annotated using Ensembl and RepeatMasker platypus genome annotations. piRNA clusters were defined manually and may be incomplete.

miRNA prediction

Clipped small RNAs of 18 nt or longer from platypus/echidna or chicken 18- to 24-nt libraries were mapped to the platypus or chicken genomes, respectively, requiring perfect sequence identity. Echidna small RNAs were mapped to the platypus genome because of the lack of availability of an echidna genome. The potential for each position of the genome to be a 5′ end of a mature miRNA sequence was evaluated by counting the total number of sequence counts starting at that position. Two windows of 150 base pairs were folded around the potential mature miRNA, one starting 30 bases upstream (for a mature in the left arm) and one starting 100 bases upstream (for a mature in the right arm). Folding was done by sampling 500 structures using RNAsubopt from the Vienna RNA Package (Wuchty et al. 1999) and taking the one that had the greatest number of paired arm bases. The longest pair of substrings whose bases only matched to each other and whose starts and ends were paired were defined as the two arms, and each hairpin was trimmed to the ends of the arms. Only hairpins that had at least 20 paired bases in the arms were considered. It was required that at least 18 of 20 bases following the potential mature 5′ end overlapped with the intended arm and that none of the bases overlapped with the opposite arm. We considered only hairpins for which at least 40% of the reads within the hairpin and 50% of reads within 10 bases of the putative 5′ end of the mature belonged to the putative 5′ end of the mature miRNA. To avoid repeat associated sequences, no read within the hairpin could have more than five total matching positions in the genome, and 99% of the reads must not have more than three matching positions. It was required that there be no more than six positions that had more than 1% of the hairpin reads and no more than four positions that had more than 10% of the hairpin reads. Candidate miRNAs that were sequenced only once were discarded. The 2350 hairpins from platypus and echidna that fit these criteria were inspected manually to remove piRNAs, repeats, degradation products, and other types of noncoding RNAs, leaving 332 candidate miRNAs. A total of 999 hairpins were predicted for chicken, which were not manually curated. Chicken orthologs of platypus miRNAs were assigned by BLASTN requiring at least 16-nt identity in the mature miRNA. In 13 cases in which chicken orthologs were cloned but not predicted (e.g., because of missing genome sequence), the cloned chicken sequences are shown in Supplemental Table 3. The most frequently cloned small RNA in the candidate miRNA hairpin was designated as the mature sequence, and the most frequently cloned sequence from the other arm was designated as the miRNA* sequence. Platypus and echidna orthologs of known miRNAs were identified by BLASTN with the Rfam miRNA database, requiring at least 16 nucleotide matches. miRNAs were named according to their ortholog; novel miRNAs were given unique names. Four miRNAs cloned only from echidna (oan-miR-1331, oan-miR-1336, oan-miR-1352, and oan-miR-1384) were given platypus names because of the lack of availability of the echidna genome.

Heatmap construction

Orthologs for candidate mature miRNAs in platypus brain, heart, liver, and testis (for comparison with human) or platypus brain, heart, liver, testis, kidney, and lung (for comparison with mouse and chicken) were found by BLASTN using the human or mouse Rfam database and 999 predicted chicken miRNAs, requiring at least 16 matched bases. miRNA read counts in platypus and those in corresponding tissues in chicken and in a human and mouse miRNA expression atlas (Landgraf et al. 2007) were first normalized to the tissue with the least number of total miRNA reads. miRNAs that did not have at least 10 normalized reads in both platypus and human/mouse/chicken were discarded. Each miRNA in each tissue was assigned an expression score, a ratio of the number of reads in that tissue versus all tissues in that organism. The miRNAs were then clustered by their tissue expression ratios in platypus using the repeated bisection option in the Cluto clustering package (Zhao and Karypis 2005). The distance between the confidence intervals of the expression ratios for each platypus/human, platypus/mouse, or platypus/chicken pair of miRNAs was computed after Wilson correction with $Z = 2$ (Brown et al. 2001) and was displayed as a difference score on the heatmap. The following human and mouse libraries from the miRNA expression atlas (Landgraf et al. 2007) were used for comparison with platypus: brain, hsa_Frontal-cortex-adult (human), average mmu_Brain-WT, and mmu_Cortex (mouse); heart, hsa_Heart (human) and mmu_Heart (mouse); liver, hsa_Liver (human) and mmu_Liver (mouse); testis, hsa_Testis (human) and mmu_Testis (mouse); lung, mmu_Lung (mouse); kidney, mmu_Kidney (mouse). For platypus/echidna expression profiles, the same procedure was used, except read counts for candidate mature miRNAs in platypus for brain, kidney, lung, heart, liver, and testis were compared with those for identical miRNAs found in echidna. Missing orthologs may in some cases reflect a failure to predict the miRNA rather than absence.

Multiple sequence alignment

Fifty-three miR-1421 related precursor sequences were aligned using ClustalW. The most highly sequenced small RNA from the most abundantly cloned arm of the hairpin was designated as the mature miRNA, and the small RNA with the greatest number of reads from the other arm was designated as the miRNA star. The consensus sequence indicates the most frequently occurring nucleotide for each position.

Platypus cell culture

Mitotic metaphase chromosomes were prepared from established platypus and echidna fibroblast cell lines. Primary cultures were set up from toe web from animals captured at the Upper Barnard River, New South Wales, Australia, during breeding season (AEC permit no. S-49-2006 to F.G.) following standard procedures. The cells where maintained in AmnioMAX C100 medium (GIBCO).

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

Physical mapping of BAC clones

A total of 100 ng of DNA from BAC clones was directly labeled with spectrum orange or spectrum green (Vysis, Abbot Molecular) using random primers and Klenow polymerase. Physical mapping by fluorescence in situ hybridization to male platypus metaphase chromosomes was performed under standard conditions. Briefly, the slides were treated with 100 µg/mL RNase A/2× SSC at 37°C for 30 min and with 0.01% pepsin in 10 mM HCl at 37°C for 10 min. After refixing for 10 min in 1× PBS, 50 mM $MgCl_2$, 1% formaldehyde, the preparations were dehydrated in an ethanol series. Slides were denatured for 2.5 min at 75°C in 70% formamide, 2× SSC at pH 7.0 and again dehydrated. For hybridization of one-half slide, 10 µL of probe DNA was coprecipitated with 10–20 µg of boiled genomic platypus DNA as competitor and 50 µg salmon sperm DNA as carrier and dissolved in 50% formamide, 10% dextran sulfate, 2× SSC. The hybridization mixture was denatured for 10 min at 80°C. Preannealing of repetitive DNA sequences was carried out for 30 min at 37°C. The slides were hybridized overnight in a moist chamber at 37°C and thereafter washed three times for 5 min in 50% formamide, 2× SSC at 42°C and once for 5 min in 0.1× SSC. Chromosomes and cell nuclei were counterstained with 1 µg/mL DAPI in 2× SSC for 1 min and mounted (Vectashield). Images were taken with a Zeiss AxioImager Z.1 epifluorescence microscope equipped with a CCD camera and Zeiss Axiovision software.

Acknowledgments

The v5.0.1 (ornAna1) platypus genome sequence and assembly was produced by the Genome Sequencing Center at Washington University, St. Louis. We thank Michelle Rooks, Michael Regulski, Laura Cardone, and Danea Rebolini for assistance with miRNA cloning and sequencing; Rachel O'Neill and Dawn Carone (University of Connecticut), Michael Darre (UCONN Poultry Farm), and Russell Jones (Newcastle University) for help with the tissue collection; John Wallis (Washington University) and Natasha McInnes (University of Adelaide) for assistance with the BAC end database and BAC identification; and Sam Griffiths-Jones (miRBase) for help with miRNA predictions. Facilities were provided by Macquarie Generation and Glenrock Station, New South Wales. Approval to collect animals was granted by the New South Wales National Parks and Wildlife Services, New South Wales Fisheries, and the Animal Ethics Committee, University of Adelaide. E.P.M. is a Sir Keith Murdoch Fellow of the American Australian Association. F.G. is an Australian Research Council Research Fellow. A.S. was supported in part by the Schering AG/Ernst Schering Foundation and in part by the Human Frontier Science Program Organization (HFSPO). P.K. was supported in part by a National Science Foundation Graduate Research Fellowship. This work was funded in part by grants from the NIH (G.J.H.) and by a kind gift from Kathryn W. Davis (G.J.H.).

References

Aravin, A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., Morris, P., Brownstein, M.J., Kuramochi-Miyagawa, S., Nakano, T., et al. 2006. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **442:** 203–207.

Aravin, A.A., Hannon, G.J., and Brennecke, J. 2007a. The Piwi/piRNA pathway: Adaptive defense for the transposon arms race. *Science* **318:** 761–764.

Aravin, A.A., Sachidanandam, R., Girard, A., Fejes-Toth, K., and Hannon, G.J. 2007b. Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* **316:** 744–747.

Ason, B., Darnell, D.K., Wittbrodt, B., Berezikov, E., Kloosterman, W.P., Wittbrodt, J., Antin, P.B., and Plasterk, R.H. 2006. Differences in vertebrate microRNA expression. *Proc. Natl. Acad. Sci.* **103:** 14385–14389.

Bartel, D.P. 2004. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* **116:** 281–297.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., et al. 2005. Identification of hundreds of conserved and nonconserved human microRNAs. *Nat. Genet.* **37:** 766–770.

Bininda-Emonds, O.R., Cardillo, M., Jones, K.E., MacPhee, R.D., Beck, R.M., Grenyer, R., Price, S.A., Vos, R.A., Gittleman, J.L., and Purvis, A. 2007. The delayed rise of present-day mammals. *Nature* **446:** 507–512.

Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G.J. 2007. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **128:** 1089–1103.

Brown, L.D., Cai, T.T., and DasGupta, A. 2001. Interval estimation for a binomial proportion. *Stat. Sci.* **16:** 101–133.

Carmell, M.A., Girard, A., van de Kant, H.J., Bourc'his, D., Bestor, T.H., de Rooij, D.G., and Hannon, G.J. 2007. MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Dev. Cell* **12:** 503–514.

Chen, P.Y., Manninga, H., Slanchev, K., Chien, M., Russo, J.J., Ju, J., Sheridan, R., John, B., Marks, D.S., Gaidatzis, D., et al. 2005. The developmental miRNA profiles of zebrafish as determined by small RNA cloning. *Genes & Dev.* **19:** 1288–1293.

Darnell, D.K., Kaur, S., Stanislaw, S., Konieczka, J.H., Yatskievych, T.A., and Antin, P.B. 2006. MicroRNA expression during chick embryo development. *Dev. Dyn.* **235:** 3156–3165.

Davis, E., Caiment, F., Tordoir, X., Cavaille, J., Ferguson-Smith, A., Cockett, N., Georges, M., and Charlier, C. 2005. RNAi-mediated allelic *trans*-interaction at the imprinted Rtl1/Peg11 locus. *Curr. Biol.* **15:** 743–749.

Delbridge, M.L. and Graves, J.A. 2007. Origin and evolution of spermatogenesis genes on the human sex chromosomes. *Soc. Reprod. Fertil. Suppl.* **65:** 1–17.

Devor, E.J. and Samollow, P.B. 2007. In vitro and in silico annotation of conserved and nonconserved microRNAs in the genome of the marsupial *Monodelphis domestica*. *J. Hered.* **99:** 66–72.

Felsenstein, J. 1997. An alternating least squares approach to inferring phylogenies from pairwise distances. *Syst. Biol.* **46:** 101–111.

Girard, A., Sachidanandam, R., Hannon, G.J., and Carmell, M.A. 2006. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442:** 199–202.

Grivna, S.T., Beyret, E., Wang, Z., and Lin, H. 2006. A novel class of small RNAs in mouse spermatogenic cells. *Genes & Dev.* **20:** 1709–1714.

Grutzner, F., Rens, W., Tsend-Ayush, E., El-Mogharbel, N., O'Brien, P.C., Jones, R.C., Ferguson-Smith, M.A., and Marshall Graves, J.A. 2004. In the platypus a meiotic chain of ten sex chromosomes shares genes with the bird Z and mammal X chromosomes. *Nature* **432:** 913–917.

Gunawardane, L.S., Saito, K., Nishida, K.M., Miyoshi, K., Kawamura, Y., Nagami, T., Siomi, H., and Siomi, M.C. 2007. A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in *Drosophila*. *Science* **315:** 1587–1590.

Handel, M.A. 2004. The XY body: A specialized meiotic chromatin domain. *Exp. Cell Res.* **296:** 57–63.

Houbaviy, H.B., Murray, M.F., and Sharp, P.A. 2003. Embryonic stem cell-specific microRNAs. *Dev. Cell* **5:** 351–358.

Houbaviy, H.B., Dennis, L., Jaenisch, R., and Sharp, P.A. 2005. Characterization of a highly variable eutherian microRNA gene. *RNA* **11:** 1245–1257.

Houwing, S., Kamminga, L.M., Berezikov, E., Cronembold, D., Girard, A., van den Elst, H., Filippov, D.V., Blaser, H., Raz, E., Moens, C.B., et al. 2007. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in zebrafish. *Cell* **129:** 69–82.

Kloosterman, W.P., Steiner, F.A., Berezikov, E., de Bruijn, E., van de Belt, J., Verheul, M., Cuppen, E., and Plasterk, R.H. 2006a. Cloning and expression of new microRNAs from zebrafish. *Nucleic Acids Res.* **34:** 2558–2569.

Kloosterman, W.P., Wienholds, E., de Bruijn, E., Kauppinen, S., and Plasterk, R.H. 2006b. In situ detection of miRNAs in animal embryos using LNA-modified oligonucleotide probes. *Nat. Methods* **3:** 27–29.

Kuramochi-Miyagawa, S., Kimura, T., Yomogida, K., Kuroiwa, A., Tadokoro, Y., Fujita, Y., Sato, M., Matsuda, Y., and Nakano, T. 2001. Two mouse piwi-related genes: miwi and mili. *Mech. Dev.* **108:** 121–133.

Kuramochi-Miyagawa, S., Kimura, T., Ijiri, T.W., Isobe, T., Asada, N., Fujita, Y., Ikawa, M., Iwai, N., Okabe, M., Deng, W., et al. 2004. Mili, a mammalian member of piwi family gene, is essential for

Downloaded from genome.cshlp.org on October 13, 2010 - Published by Cold Spring Harbor Laboratory Press

spermatogenesis. *Development* **131:** 839–849.
Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A.O., Landthaler, M., et al. 2007. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129:** 1401–1414.
Lau, N.C., Seto, A.G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D.P., and Kingston, R.E. 2006. Characterization of the piRNA complex from rat testes. *Science* **313:** 363–367.
Liu, J., Carmell, M.A., Rivas, F.V., Marsden, C.G., Thomson, J.M., Song, J.J., Hammond, S.M., Joshua-Tor, L., and Hannon, G.J. 2004. Argonaute2 is the catalytic engine of mammalian RNAi. *Science* **305:** 1437–1441.
Meister, G., Landthaler, M., Patkaniowska, A., Dorsett, Y., Teng, G., and Tuschl, T. 2004. Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Mol. Cell* **15:** 185–197.
Namekawa, S.H., VandeBerg, J.L., McCarrey, J.R., and Lee, J.T. 2007. Sex chromosome silencing in the marsupial male germ line. *Proc. Natl. Acad. Sci.* **104:** 9730–9735.
O'Carroll, D., Mecklenbrauker, I., Das, P.P., Santana, A., Koenig, U., Enright, A.J., Miska, E.A., and Tarakhovsky, A. 2007. A Slicer-independent role for Argonaute 2 in hematopoiesis and the microRNA pathway. *Genes & Dev.* **21:** 1999–2004.
Plasterk, R.H. 2006. Micro RNAs in animal development. *Cell* **124:** 877–881.
Rens, W., Grutzner, F., O'Brien, P.C., Fairclough, H., Graves, J.A., and Ferguson-Smith, M.A. 2004. Resolution and evolution of the duck-billed platypus karyotype with an X1Y1X2Y2X3Y3X4Y4X5Y5 male sex chromosome constitution. *Proc. Natl. Acad. Sci.* **101:** 16257–16261.
Ro, S., Park, C., Sanders, K.M., McCarrey, J.R., and Yan, W. 2007. Cloning and expression profiling of testis-expressed microRNAs. *Dev. Biol.* **311:** 592–602.
Rowe, T., Rich, T.H., Vickers-Rich, P., Springer, M., and Woodburne, M.O. 2008. The oldest platypus and its bearing on divergence timing of the platypus and echidna clades. *Proc. Natl. Acad. Sci.* **105:** 1238–1242.
Takada, S., Berezikov, E., Yamashita, Y., Lagos-Quintana, M., Kloosterman, W.P., Enomoto, M., Hatanaka, H., Fujiwara, S., Watanabe, H., Soda, M., et al. 2006. Mouse microRNA profiles determined with a new and sensitive cloning method. *Nucleic Acids Res.* **34:** e115.
Warren, W.C., Hillier, L.W., Marshall Graves, J.A., Birney, E., Ponting, C.P., Grutzner, F., Belov, K., Miller, W., Clarke, L., Chinwalla, A.T., et al. 2007. Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453:** 175–183.
Watanabe, T., Takeda, A., Tsukiyama, T., Mise, K., Okuno, T., Sasaki, H., Minami, N., and Imai, H. 2006. Identification and characterization of two novel classes of small RNAs in the mouse germline: Retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. *Genes & Dev.* **20:** 1732–1743.
Wienholds, E., Kloosterman, W.P., Miska, E., Alvarez-Saavedra, E., Berezikov, E., de Bruijn, E., Horvitz, H.R., Kauppinen, S., and Plasterk, R.H. 2005. MicroRNA expression in zebrafish embryonic development. *Science* **309:** 310–311.
Wuchty, S., Fontana, W., Hofacker, I.L., and Schuster, P. 1999. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* **49:** 145–165.
Wyckoff, G.J., Wang, W., and Wu, C.I. 2000. Rapid evolution of male reproductive genes in the descent of man. *Nature* **403:** 304–309.
Xu, H., Wang, X., Du, Z., and Li, N. 2006. Identification of microRNAs from different tissues of chicken embryo and adult chicken. *FEBS Lett.* **580:** 3610–3616.
Yekta, S., Shih, I.H., and Bartel, D.P. 2004. MicroRNA-directed cleavage of HOXB8 mRNA. *Science* **304:** 594–596.
Yu, Z., Raabe, T., and Hecht, N.B. 2005. MicroRNA Mirn122a reduces expression of the posttranscriptionally regulated germ cell transition protein 2 (Tnp2) messenger RNA (mRNA) by mRNA cleavage. *Biol. Reprod.* **73:** 427–433.
Zhang, R., Peng, Y., Wang, W., and Su, B. 2007. Rapid evolution of an X-linked microRNA cluster in primates. *Genome Res.* **17:** 612–617.
Zhao, Y. and Karypis, G. 2005. Data clustering in life sciences. *Mol. Biotechnol.* **31:** 55–80.

Received October 15, 2007; accepted in revised form February 26, 2008.

ARTICLES

Genome analysis of the platypus reveals unique signatures of evolution

A list of authors and their affiliations appears at the end of the paper

We present a draft genome sequence of the platypus, *Ornithorhynchus anatinus*. This monotreme exhibits a fascinating combination of reptilian and mammalian characters. For example, platypuses have a coat of fur adapted to an aquatic lifestyle; platypus females lactate, yet lay eggs; and males are equipped with venom similar to that of reptiles. Analysis of the first monotreme genome aligned these features with genetic innovations. We find that reptile and platypus venom proteins have been co-opted independently from the same gene families; milk protein genes are conserved despite platypuses laying eggs; and immune gene family expansions are directly related to platypus biology. Expansions of protein, non-protein-coding RNA and microRNA families, as well as repeat elements, are identified. Sequencing of this genome now provides a valuable resource for deep mammalian comparative analyses, as well as for monotreme biology and conservation.

The platypus (*Ornithorhynchus anatinus*) has always elicited excitement and controversy in the zoological world[1]. Some initially considered it to be a true mammal despite its duck-bill and webbed feet. The platypus was placed with the echidnas into a new taxon called the Monotremata (meaning 'single hole' because of their common external opening for urogenital and digestive systems). Traditionally, the Monotremata are considered to belong to the mammalian subclass Prototheria, which diverged from the therapsid line that led to the Theria and subsequently split into the marsupials (Marsupialia) and eutherians (Placentalia). The divergence of monotremes and therians falls into the large gap in the amniote phylogeny between the eutherian radiation about 90 million years (Myr) ago and the divergence of mammals from the sauropsid lineage around 315 Myr ago (Fig. 1). Estimates of the monotreme–theria divergence time range between 160 and 210 Myr ago; here we will use 166 Myr ago, recently estimated from fossil and molecular data[2].

The most extraordinary and controversial aspect of platypus biology was initially whether or not they lay eggs like birds and reptiles. In 1884, William Caldwell's concise telegram to the British Association announced ''Monotremes oviparous, ovum meroblastic'', not holoblastic as in the other two mammalian groups[3,4]. The egg is laid in an earthen nesting burrow after about 21 days and hatches 11 days later[5,6]. For about 4 months, when most organ systems differentiate, the young depend on milk sucked directly from the abdominal skin, as females lack nipples. Platypus milk changes in protein composition during lactation (as it does in marsupials, but not in most eutherians[5]). The anatomy of the monotreme reproductive system reflects its reptilian origins, but shows features typical of mammals[7], as well as unique specialized characteristics. Spermatozoa are filiform, like those of birds and reptiles, but, uniquely among amniotes, form bundles of 100 during passage through the epididymis. Chromosomes are arranged in defined order in sperm[8] as they are in therians, but not birds[9]. The testes synthesize testosterone and dihydrotestosterone, as in therians, but there is no scrotum and testes are abdominal[10].

Other special features of the platypus are its gastrointestinal system, neuroanatomy (electro-reception) and a venom delivery system, unique among mammals[11]. Platypus is an obligate aquatic feeder that relies on its thick pelage to maintain its low (31–32 °C) body temperature during feeding in often icy waters. With its eyes, ears and nostrils closed while foraging underwater, it uses an electro-sensory system in the bill to help locate aquatic invertebrates and other prey[12,13]. Interestingly, adult monotremes lack teeth.

The platypus genome, as well as the animal, is an amalgam of ancestral reptilian and derived mammalian characteristics. The platypus karyotype comprises 52 chromosomes in both sexes[14,15], with a few large and many small chromosomes, reminiscent of reptilian macro- and microchromosomes. Platypuses have multiple sex chromosomes with some homology to the bird Z chromosome[16]. Males have five X and five Y chromosomes, which form a chain at meiosis and segregate into 5X and 5Y sperm[17,18]. Sex determination and sex chromosome dosage compensation remain unclear.

Platypuses live in the waterways of eastern and southern Australia, including Tasmania. Its secretive lifestyle hampers understanding of its population dynamics and the social and family structure. Platypuses are still relatively common in the wild, but were recently reclassified as 'vulnerable' because of their reliance on an aquatic environment that is under stress from climate change and degradation by human activities. Water quality, erosion, destruction of habitat and food resources, and disease now threaten populations. Because the platypus has rarely bred in captivity and is the last of a long line of ornithorhynchid monotremes, their continued survival is of great importance. Here we describe the platypus genome sequence and compare it to the genomes of other mammals, and of the chicken.

Sequencing and assembly

All sequencing libraries were prepared from DNA of a single female platypus (Glennie; Glenrock Station, New South Wales, Australia) and were sequenced using established whole-genome shotgun (WGS) methods[19]. A draft assembly was produced from ~6× coverage of whole-genome plasmid, fosmid and bacterial artificial chromosome (BAC) reads (Supplementary Table 1) using the assembly program PCAP[20] (Supplementary Notes 1). A BAC-based physical map was developed in parallel with the sequence assembly and subsequently integrated with the WGS assembly to provide the primary means of scaffolding the assembly into larger ordered and oriented groupings (ultracontigs; Supplementary Notes 2 and 3 and Supplementary Table 2). Because there were no platypus linkage maps available, we used fluorescent *in situ* hybridization (FISH) to localize a subset of the sequence scaffolds to chromosomes following the agreed nomenclature[21]. Of the 1.84 gigabases (Gb) of assembled sequence, 437 megabases (Mb) were ordered and oriented along 20 of

©2008 Nature Publishing Group

the platypus chromosomes. We analysed numerous metrics of assembly quality (Supplementary Notes 4–11) and we conclude that despite the adverse contiguity, the existing platypus assembly, given its structural and nucleotide accuracy, provides a reasonable substrate for the analyses presented here.

Non-protein-coding genes

In general, the platypus genome contains fewer computationally predicted non-protein-coding (nc)RNAs (1,220 cases excluded high repetitive small nucleolar RNA (snoRNA) copies; see below) than do other mammalian species (for example, human with 4,421 Rfam hits), similar to observations in chicken[19] (655 Rfam-based ncRNAs). This is probably because of the extensive retrotransposition of ncRNAs in therian mammals and the apparent lack of L1-mediated retrotransposition in chicken and platypus. The exception to this is the platypus family of snoRNAs, which is markedly expanded (~2,000 matches to the Rfam covariant models) compared to that for therian mammals (~200). snoRNAs are involved in RNA modifications, in particular of ribosomal RNA, and are often located in introns of protein-coding genes[22]. Our investigations revealed a novel short-interspersed-element (SINE)-like, snoRNA-related retrotransposon—which we have labelled snoRTEs—that has duplicated in platypus to ~40,000 full-length or truncated copies. It is retrotransposed by means of retrotransposon-like non-LTR (long terminal repeat) transposable elements (RTE) as opposed to the L1-mediated transposition mechanism in therians[23]. We constructed a complementary DNA library of small, ncRNAs and identified 371 consensus sequences of small RNAs that included 166 snoRNAs[23] (Supplementary Table 3). Ninety-nine of these cloned snoRNAs are found in paralogous families, and 21 of them belong to the snoRTE class. The presence of both the structural requirements known to be important in snoRNA function[24] and evidence of their expression are consistent with these snoRTE elements being functional in the platypus. Similar to other unrelated ncRNAs that have proliferated in therian mammals (for example, 7SL RNA-derived primate *Alu* elements, tRNA-derived rodent identifier (ID) elements), this recent SINE-like expansion is probably due to chance events. However, given the RNA modification activity of snoRNAs, and our increasing awareness of the cellular importance of RNA molecules, it might be that some of the retrotranspositionally duplicated RNAs were exapted into new functions in this species.

Other small RNAs. Overall, we found commonalities with small RNA (sRNA) pathways of other mammals, but also features that are unique to monotremes. Components of the RNA interference machinery are conserved in platypus, including elements of biogenesis pathways (Dicer and Drosha) and RNA-interference effector complexes (argonaute proteins; Supplementary Table 4). Of 20,924,799 platypus and echidna sRNA reads derived from liver, kidney, brain, lung, heart and testis, 67% could be assigned to known microRNA (miRNA) families. Established patterns of miRNA expression were generally recapitulated in monotremes.

To determine the conservation patterns of miRNAs in platypus, we identified platypus miRNAs sharing at least 16-nucleotide identity with miRNAs in eutherian mammals (mouse/human) and chicken. Although most conserved miRNAs were identified across these vertebrate lineages (137 miRNAs), 10 miRNAs were shared only with eutherians (mouse/human) and 4 only with chicken (Fig. 2a). miRNAs can be classified into families based on identity of the functional 'seed' region at position 2–8 of the mature miRNA strand. We identified miRNA families that were shared between platypus and eutherians but not chicken (40 families), or between platypus and chicken but not eutherians (8 families), suggesting that for some miRNAs only

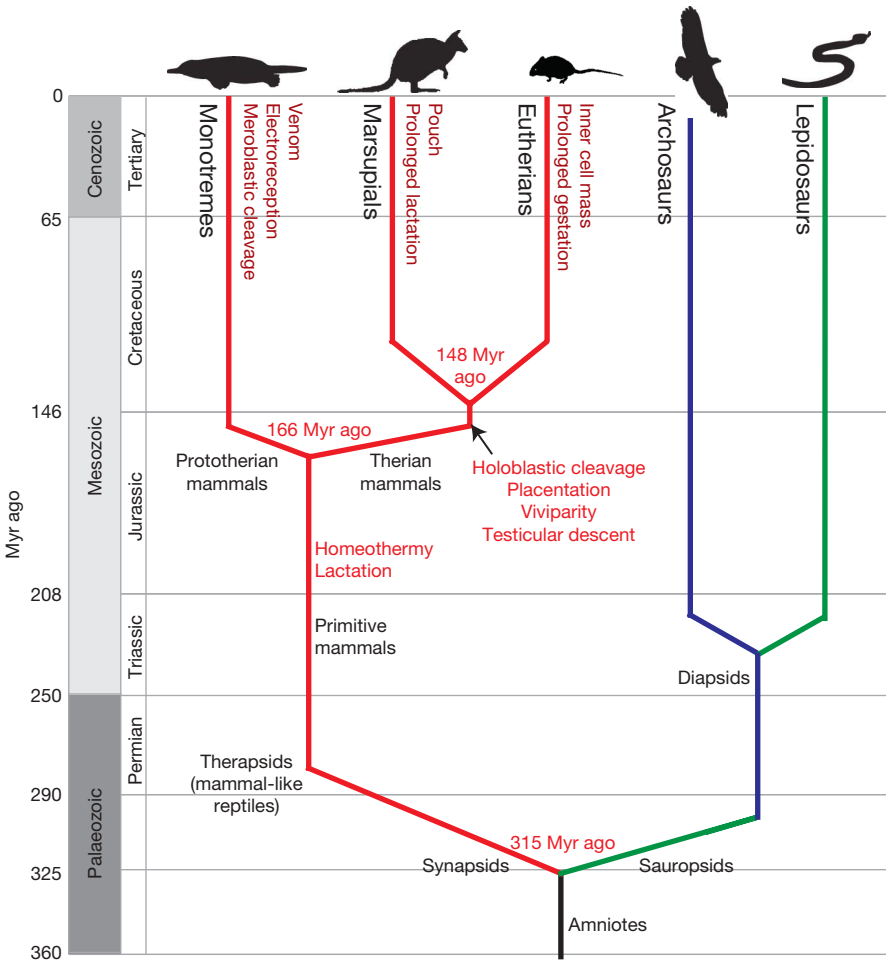


Figure 1 | Emergence of traits along the mammalian lineage. Amniotes split into the sauropsids (leading to birds and reptiles) and synapsids (leading to mammal-like reptiles). These small early mammals developed hair, homeothermy and lactation (red lines). Monotremes diverged from the therian mammal lineage ~166 Myr ago[2] and developed a unique suite of characters (dark-red text). Therian mammals with common characters split into marsupials and eutherians around 148 Myr ago[2] (dark-red text). Geological eras and periods with relative times (Myr ago) are indicated on the left. Mammal lineages are in red; diapsid reptiles, shown as archosaurs (birds, crocodilians and dinosaurs), are in blue; and lepidosaurs (snakes, lizards and relatives) are in green.

©2008 Nature Publishing Group

the seed region may have been selectively conserved (Fig. 2a). Conserved miRNAs tended to be more robustly expressed in the platypus tissues analysed than lineage-restricted miRNAs (Fig. 2b).

To identify miRNAs unique to monotremes we used a heuristic search that identifies miRNA candidates in deep-sequencing data sets[25]. This method predicted 183 novel miRNAs in platypus and echidna (Fig. 2a). Notably, 92 of these lay in 9 large clusters, on platypus chromosome X1 and contigs 1754, 7160, 7359, 8388, 11344, 22847, 198872 and 191065. Physical mapping confirmed that at least five of these contigs are linked to the long arm of chromosome X1 (ref. 25). These abundantly expressed clusters were sequenced almost exclusively from platypus and echidna testis (Fig. 2b). The expansion of this unique miRNA class and its expression domain suggest possible roles in monotreme reproductive biology[25].

Piwi-interacting RNAs (piRNAs) associate with a germline-expressed clade of argonaute proteins, known as Piwis[26], and have a role in transposon silencing and genome methylation[26]. Monotreme piRNAs bear strong structural similarity to those in eutherians. They are ~29 nucleotides in length and arise from large testis-specific genomic clusters with distinct genomic strand asymmetry, often with a typical 'bidirectional' organization. We identified 50 major platypus piRNA clusters as well as numerous smaller clusters[25]. In contrast to piRNAs in mouse, platypus piRNAs are repeat-rich and bear strong signatures of active transposon defence.

Gene evolution

We set out to define the protein-coding gene content of platypus to illuminate both the specific biology of the monotreme clade and for comparisons to eutherians and marsupials, or to chicken, the representative sauropsid. Protein-coding genes were predicted using the established Ensembl pipeline[27] suitably modified for platypus (Supplementary Notes 14), with a greater emphasis placed on similarity matches to mammalian genes. Overall this resulted in 18,527 protein-coding genes being predicted from the current platypus assembly. The number of platypus protein-coding genes thus is similar to estimates (18,600–20,800) for human and opossum[28,29].

We were interested first in identifying platypus genes that contribute most to core biological functions that are conserved across the mammals. These will typically be 'simple' 1:1 orthologues, genes that have remained as single copies without duplication or deletion in platypus, in Eutheria (specifically, in dog, human and mouse) and in opossum, a representative marsupial. Subsequently, we considered genes that have been duplicated or deleted in the monotreme lineage, or that have been lost in eutherian and/or marsupial lineages. Such genes are proposed to contribute most to the lineage-specific biological functions that distinguish individual mammals[30]. These studies required the use of an outgroup species, here chicken, a representative of the sauropsids.

As expected, the majority of platypus genes (82%; 15,312 out of 18,596) have orthologues in these five other amniotes (Supplementary Table 5). The remaining 'orphan' genes are expected to primarily reflect rapidly evolving genes, for which no other homologues are discernible, erroneous predictions, and true lineage-specific genes that have been lost in each of the other five species under consideration. Simple 1:1 orthologues, which have been conserved without duplication, deletion or non-functionalization across the five mammalian species, were greatly enriched in housekeeping functions, such as metabolism, DNA replication and mRNA splicing (Supplementary Table 6).

We then identified evolutionary lineages that experienced the most stringent purifying selection. The mouse terminal lineage exhibited a significantly higher degree of purifying selection (the ratio of amino acid replacement to silent substitution rates, $d_N/d_S = 0.105$, $P < 0.001$) than dog, opossum and chicken terminal branches (values of 0.123–0.128); human and platypus terminal lineages showed significantly reduced purifying selection (both 0.132, $P < 0.03$). These values probably reflect the increased efficiency of purifying selection in populations of larger effective size, such as that of mouse[31]. We find that at least one nucleotide substitution has occurred, on average, in synonymous sites of platypus and human orthologues since their last common ancestor (Supplementary Notes 17 and Supplementary Fig. 1). This means that most neutral sequence cannot be aligned accurately between monotreme and eutherian genomes.

Next, we determined the genetic distance of echidna (*Tachyglossus aculeatus*) from platypus. The median d_S value of 0.125 for the orthologues of echidna and platypus, when compared to the value for the monotreme lineage, predicts that platypus and echidna last shared a common ancestor 21.2 Myr ago. Although similar to previous estimates[32], this value seems to be at odds with fossil evidence, perhaps owing to relatively recent reductions of mutational rates in the monotreme lineage[33].

Monotreme biology

We next investigated whether the ancestral reptilian characters of monotremes are reflected in the set of genes that have been retained in platypus, sauropsids and other vertebrates from outside of the amniote clade (such as frogs and fish), but have been lost from eutherian and marsupial lineages (Fig. 1). These ancestral, sauropsid-like, characters of platypus include oviparity (egg laying) and the outward appearances of its spermatozoa and retina. Simultaneously, we sought genetic evidence within the platypus genome both for characteristics peculiar to monotremes, such as venom production and

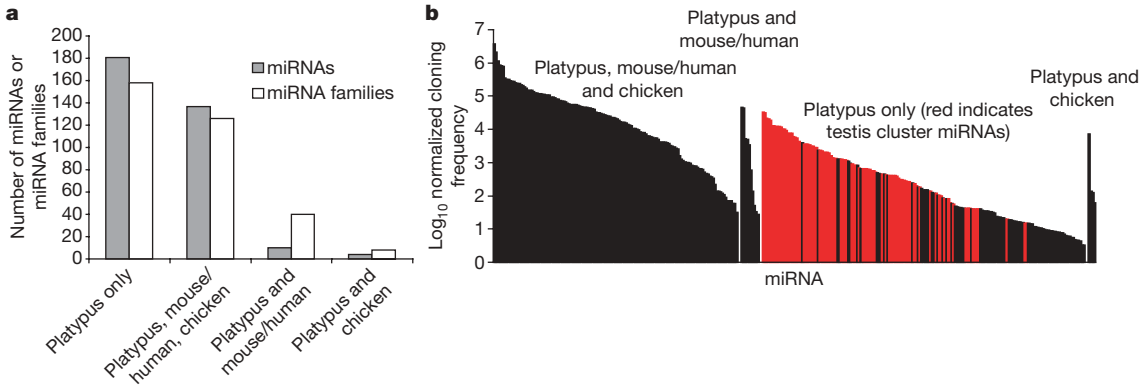


Figure 2 | Platypus miRNAs. a, Platypus has miRNAs shared with eutherians and chickens, and a set that is platypus-specific. miRNAs cloned from six platypus tissues were assigned to families based on seed conservation. Platypus miRNAs and families were divided into classes (indicated) based on their conservation patterns with eutherian mammals (mouse/human) and with chicken. **b**, Expression of platypus miRNAs. The cloning frequency of each platypus mature miRNA sequenced more than once is represented by a vertical bar and clustered by conservation pattern. miRNAs from a set of monotreme-specific miRNA clusters that are expressed in testis are shaded in red.

©2008 Nature Publishing Group

electro-reception, and for characteristics unique to mammals, in particular lactation. By investigating platypus homologues of genes already known to be involved in specific physiological processes (see Methods), we highlight those platypus genes for which evolution exemplifies the ancestral or derived physiological characters of monotremes.

Chemoreception. The semi-aquatic platypus was expected to sense its terrestrial, but not aquatic, environment by detecting airborne odorants using olfactory receptors and vomeronasal receptors (types 1 and 2: V1Rs, V2Rs). Nevertheless large numbers of odorant receptor, V1R and V2R homologues (approximately 700, 950 and 80, respectively) are apparent in the platypus genome assembly, although for each family only a minority lack frame disruptions (approximately 333, 270 and 15, respectively)[34]. Many of these platypus genes and pseudogenes are monophyletic, having arisen by duplication in the 166 Myr since the last common ancestor of monotremes and therians. Although mouse and rat genomes possess greater numbers of odorant receptors and V2Rs than the platypus genome[35,36], the platypus repertoire of V1Rs, showing undisrupted reading frames, is the largest yet seen, 50% more than for mouse (Fig. 3b). This is particularly noteworthy as the *Anolis carolinensis* lizard (sequence data used with the permission of the Broad Institute) and the chicken[19] seem to possess no such receptors. The large expansion of the platypus V1R gene family might reflect sensory adaptations for pheromonal communication or, more generally, for the detection of water-soluble, non-volatile odorants, during underwater foraging.

The platypus odorant receptor gene repertoire is roughly one-half as large as those in other mammals[37]. Nevertheless, platypus odorant receptors fall into class, family and subfamily structures that are well represented from across the mammals, with a few notable exceptions such as family 14 (Fig. 3a). Together with the finding that lizard contains only ~200 odorant receptor genes and pseudogenes, this indicates that the platypus olfactory repertoire is, as expected, more akin to other mammals than it is to sauropsids.

Eggs. Fertilization in the platypus exhibits both sauropsid and therian characteristics. Platypus ova are small (4 mm diameter) relative to comparably sized reptiles and birds, and eggs hatch at an early stage of development so that most growth of the embryo and infant is dependent on lactation, as in marsupials. Like all mammals and many other amniotes, when fertilization occurs the ovum is invested with a zona pellucida. The platypus genome encodes each of the four proteins of the human zona pellucida[38], as well as two ZPAX genes (Table 1) that previously were observed only in birds, amphibians and fish. The aspartyl-protease nothepsin is present in platypus, but has been lost from marsupial and eutherian genomes (Table 1). In zebrafish, this gene is specifically expressed in the liver of females under the action of oestrogens, and accumulates in the ovary[39]. These are the same characteristics as of the vitellogenins, indicating that nothepsin may be involved in processing vitellogenin or other egg-yolk proteins. We find that platypus has retained a single vitellogenin gene and pseudogene, whereas sauropsids such as chicken have three and the viviparous marsupials and eutherians have none.

Spermatozoa. Orthologues of many of the eutherian sperm membrane proteins related to fertilization[40] are present in platypus (and marsupial) genomes. These include the genes for a number of putative zona pellucida receptors and proteins implicated in sperm–oolemma fusion. Testis-specific proteases, which in eutherians participate in degradation of the zona pellucida during fertilization, are all absent from the platypus genome assembly.

Monotreme spermatozoa undergo some post-testicular maturational changes, including the acquisition of progressive motility, loss of cytoplasmic droplets and aggregation of single spermatozoa into bundles during passage through the epididymis[11]. Nevertheless, maturational changes in the sperm surface that are both unique and essential in other mammals for fertilization of the ovum have yet to be identified. Also, the epididymis of monotremes is not highly adapted for sperm storage as in most marsupial and eutherian mammals. Consistent with these findings is the absence of platypus genes for the epididymal-specific proteins that have been implicated in sperm maturation and storage in other mammals. The most abundant secreted protein in the platypus epididymis is a lipocalin, the homologues of which are the most secreted proteins in the reptilian epididymis[41]. Notably, *ADAM7*, a protease that is secreted in the epididymis of eutherians, has an orthologue in the platypus. This is a bona fide protease with a characteristic Zn^{2+}-coordinating sequence HExxH in the platypus, in the opossum and the tree shrew (*Tupaia belangeri*). However, loss of its proteolytic activity is predicted in eutherians[42] owing to a single point mutation within its active site (E to Q).

Lactation and dentition. Lactation is an ancient reproductive trait whose origin predates the origin of mammals. It has been proposed that early lactation evolved as a water source to protect porous parchment-shelled eggs from desiccation during incubation[43] or as a protection against microbial infection. Parchment-shelled egg-laying monotremes also exhibit a more ancestral glandular mammary patch or areola without a nipple that may still possess roles in egg protection. However, in common with all mammals, the milk of monotremes has evolved beyond primitive egg protection into a true milk that is a rich secretion containing sugars, lipids and milk proteins with nutritional, anti-microbial and bioactive functions. In a reflection of this eutherian similarity platypus casein genes are tightly clustered together in the genome, as they are in other mammals, although platypus contains a recently duplicated β-casein gene (Supplementary Fig. 2).

Mammalian casein genes are thought to have originally arisen by duplication of either enamelin or ameloblastin[44], both of which are tooth enamel matrix protein genes that are located adjacent to the casein gene cluster in eutherians and, we find, also in platypus. Adult platypuses, as well as echidnas, lack teeth but the conservation of these enamel protein genes is consistent with the presence of teeth and enamel in the juvenile, as well as the fossil platypuses[45].

Venom. Only a handful of mammals are venomous, but the male platypus is unique among them in delivering its poison not via a bite but from hind-leg spurs. Despite the obvious difficulties in obtaining

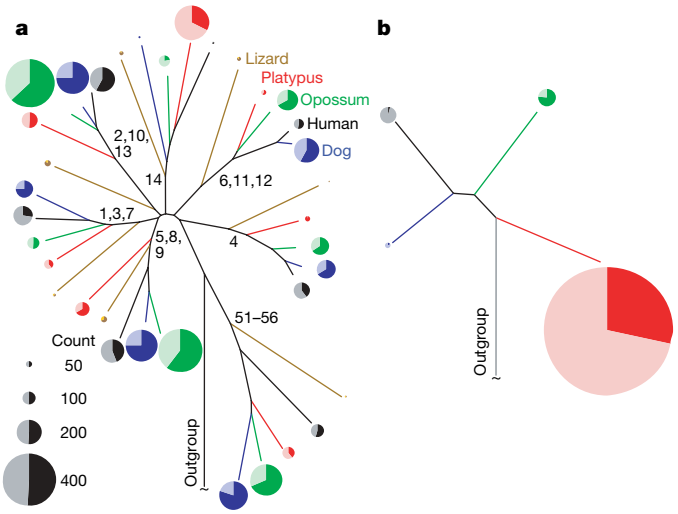


Figure 3 | The platypus chemosensory receptor gene repertoire. a, b, The platypus genome contains only few olfactory receptor genes from olfactory receptor families that are greatly expanded among therians (three other mammals and a reptile shown), but many genes in olfactory receptor family 14 (**a**), and relatively numerous vomeronasal type 1 (V1R) receptors (**b**). These schematic phylogenetic trees show relative family sizes and pseudogene contents of different gene families (enumerated beside internal branches) and the V1R repertoire in platypus. Pie charts illustrate the proportions of intact genes (heavily shaded) versus disrupted pseudogenes (lightly shaded).

©2008 Nature Publishing Group

samples, it is now known that platypus venom is a cocktail of at least 19 different substances[46] including defensin-like peptides (vDLPs), C-type natriuretic peptide (vCNP) and nerve growth factor (vNGF). When analysed phylogenetically and mapped to the platypus genome assembly, these sequences are revealed to have arisen from local duplications of genes possessing very different functions (Fig. 4). Notably, duplications in each of the β-defensin, C-type natriuretic peptide and nerve growth factor gene families have also occurred independently in reptiles during the evolution of their venom[47]. Convergent evolution has thus clearly occurred during the independent evolution of reptilian and monotreme venom[48].

Immunity. Although the major organs of the monotreme immune system are similar to those of other mammals[49], the repertoire of immunity molecules shows some important differences from those of other mammals. In particular, the platypus genome contains at least 214 natural killer receptor genes (Supplementary Notes 18) within the natural killer complex, a far larger number than for human (15 genes[50]), rat (45 genes[50]) or opossum (9 genes[51]).

Both platypus and opossum genomes contain gene expansions in the cathelicidin antimicrobial peptide gene family (Supplementary Fig. 3). Among eutherians, primates and rodents have a single cathelicidin gene[52,53], whereas sheep and cows have numerous genes that have been duplicated only recently[54]. The expanded repertoire of cathelicidin genes in both marsupials and monotremes may arm their immunologically naive young with a diverse arsenal of innate immune responses. In eutherians, with their increases in length of gestation and advances in development *in utero* of their immune systems, the diversity of antimicrobial peptide genes may have become less critical. The platypus genome also contains an expansion in the macrophage differentiation antigen *CD163* gene family (Supplementary Notes 18).

Genome landscape

First, we analyse the phylogenetic position of platypus and confirm that marsupials and eutherians are more closely related than either is to monotremes (Supplementary Notes 19). We then describe platypus chromosomes and observe some properties of platypus interspersed and tandem repeats. We also discuss a potential relationship between interspersed repeats and genomic imprinting and investigate how the extremely high G+C fraction in platypus affects the strong association seen in eutherians between CpG islands and gene promoters.

Platypus chromosomes. Platypus chromosomes provide clues to the relationship between mammal and reptile chromosomes, and to the origins of mammal sex chromosomes and dosage compensation. Our analysis provides further insight with the following findings: the 52 platypus chromosomes show no correlation between the position of orthologous genes on the small platypus chromosomes and chicken microchromosomes; for the unique 5X chromosomes of platypus we reveal considerable sequence alignment similarity to chicken Z and no orthologous gene alignments to human X, implying that the platypus X chromosome evolved directly from a bird-like ancestral reptilian system[55]; and the genes on the five platypus X chromosomes appear to be partially dosage compensated (Supplementary Fig. 5), perhaps parallel to the incomplete dosage compensation recently described in birds[56].

Repeat elements. About one-half of the platypus genome consists of interspersed repeats derived from transposable elements. The most abundant and still active repeats are (severely truncated) copies of the 5-kb long-interspersed-element (LINE2) and its non-autonomous SINE-companion mammalian-wide interspersed repeat (MIR, Mon-1 in monotremes) that became extinct in marsupials and in eutherians 60–100 Myr ago. We estimate that there are 1.9 and 2.75 million copies of LINE2 and MIR/Mon-1, respectively, in the 2.3-Gb platypus genome. DNA transposons and LTR retroelements are quite

Table 1 | Platypus genes that have been lost from the eutherian lineage

Description	Platypus Ensembl gene	Proposed function
Retinal guanylate cyclase activator 1A	ENSOANG00000012043	In zebrafish, expressed in retina
Enoyl-CoA hydratase/isomerase	ENSOANG00000012890	Involved in fatty acid metabolism
Ferric reductase/cytochrome *b*561	ENSOANG00000019725	Absorption of dietary iron
Nothepsin, aspartic proteinase	ENSOANG00000005955	Processes egg-yolk proteins
Glutamine synthetase	ENSOANG00000008089	Role in nitrogen metabolism
Vitellogenin II	Contig 10010	Major egg-yolk protein
Cytochrome P450, CYP2-like	ENSOANG00000004537	Toxin degradation
ATP6AP1 paralogue	ENSOANG00000004825	Retinal pigmentation
Organic solute transporter alpha (2 genes)	Ultracontig 462, Contig 159089	Bile acid transport
Neuropeptide Y7 receptor	ENSOANG00000014966	Regulator of food intake
Melatonin receptor 1C	ENSOANG00000011638	Circadian rhythm regulation
Epidermal differentiation-specific proteins (3 genes)	ENSOANG00000005335, ENSOANG00000003767, ENSOANG00000013512	Neural and epidermal differentiation
TRPV7/TRPV8 transient receptor potential cation channels	ENSOANG00000015080, ENSOANG00000015083	Novel epithelial calcium channels
Shortwave-sensitive-2 (SWS2) opsin gene	Ultracontig 401	Cone visual pigment
Opsin 5 paralogue	ENSOANG00000009478	Light-sensitive receptor
Indigoidine synthase A	Contig 29616	Pigmentation
ZPAX, egg envelope glycoprotein	ENSOANG00000007840, ENSOANG00000002187	Egg envelope protein
Galanin receptor	ENSOANG00000020606	Neuropeptide receptor
Kainate-binding protein	ENSOANG00000007006	Glutamate receptor
Anti-dorsalizing morphogenetic protein	ENSOANG00000002980	Patterning of the body axis during gastrulation
Retinal genes (2 genes)	ENSOANG00000001054, ENSOANG00000004065	Unknown function
Uteroglobin-like secretoglobins (3 genes)	ENSOANG00000020019, ENSOANG00000022350, ENSOANG00000021122	Unknown function
Testis homeobox C14-like proteins (>2 genes)	ENSOANG00000020069, ENSOANG00000022694	Unknown function
Parvalbumin	ENSOANG00000000764	Muscle function
Slc7a2-prov protein	ENSOANG00000009602	Cationic amino acid transporter
Cystine/glutamate transporter	ENSOANG00000005615	Amino acid transporter
SOUL protein	ENSOANG00000013998	Retina and pineal gland haem protein, oxygen sensing
Twin-pore potassium channel Talk-1-like	ENSOANG00000011839	Potassium channel
Alpha-aspartyl dipeptidase	ENSOANG00000009001	Unknown function
Monovalent cation/H^+ antiporter	ENSOANG00000012961	Unknown function; conserved in other metazoa and in yeast

Sequences without Ensembl nomenclature are found in Supplementary Information.

©2008 Nature Publishing Group

rare in platypus, but there are thousands of copies of an ancient gypsy-class LTR element (all LTR elements previously identified in mammals, birds, or reptiles belong to the retrovirus clade). Overall, the frequency of interspersed repeats (over 2 repeats per kb) is higher than in any previously characterized metazoan genome. Population analysis using LINE2/Mon-1 elements distinguished the Tasmanian population from three other mainland clusters (Supplementary Fig. 4a, b), in good agreement with tree-based analysis, physical proximity and previous knowledge of platypus population relationships[57].

Cluster analysis of all LINE2 copies revealed a phylogenetic relationship lacking branches, as if a single-locus, fast-evolving gene has steadily spread an exceptional number of pseudogenes over time (Supplementary Fig. 6). This 'master gene' appearance is, to a lesser degree, also observed for LINE1 in eutherians[58], but not to the same extent for MIR/Mon-1 or other retrotransposons in mammals. The phylogeny of LINE2 and Mon-1 was also supported by a genome-wide transposition-in-transposition (TinT) analysis[59] (Supplementary Tables 7 and 8). LINE2 density is similar on all chromosomes (Supplementary Fig. 7); it does not correlate with chromosome length (and recombination rate) as the CR1 LINE density does in the chicken genome[19], nor is it higher on sex chromosomes than on autosomes, as LINE1 density is in eutherians (which has led to postulations on a function in dosage compensation)[60].

We compared microsatellites in the platypus genome with those of representative vertebrates (Supplementary Notes 22). The mean microsatellite coverage of platypus genomic sequences assembled into chromosomes is 2.67 ± 0.34%; significantly lower than all other mammalian genomes sequenced so far and most similar to that observed in chicken (Supplementary Fig. 8). Microsatellites are on average shorter in platypus than in other genomes (Supplementary Table 9), but microsatellite coverage surpasses chicken owing to very long tri- and tetranucleotide repeats (Supplementary Fig. 9). The platypus has a higher proportion of microsatellites with high A+T content, in comparison to the other vertebrates examined, an abundance distribution that has more in common with reptiles than with mammals (Supplementary Fig. 10).

Genomic imprinting. Genomic imprinting is an epigenetic phenomenon that results in monoallelic gene expression. In the vertebrates, imprinting seems to have evolved recently and has only been confirmed in marsupials and eutherian mammals[61,62]. The autosomal localization of some imprinted orthologues in platypus is known[63]. However, we examined the conservation of synteny and the distribution of retrotransposed elements in all orthologous eutherian-imprinted clustered and non-clustered genes in the platypus genome. A representative cluster is shown in Fig. 5 (see also Supplementary Fig. 12).

Clusters that became imprinted in therians (with the exception of the Prader–Willi–Angelman locus[64]) have not been assembled recently and reside in ancient syntenic mammalian groups, although some regions have expanded by mechanisms such as gene duplication or transposition. There were significantly fewer LTR and DNA elements across all platypus orthologous regions relative to eutherian imprinted genes ($P < 0.04$ and 0.04, respectively), whereas there was a significant increase in the sequences masked by SINEs ($P < 0.03$). The chicken had fewer total repeats and no SINEs or sRNAs. Comparison of all regions in the platypus with the orthologous regions in opossum, mouse, dog and human demonstrates that accumulation of LTR, DNA elements, and simple and low complexity repeats coincides with, and may be a driving force in, the acquisition of imprinting in these regions in therian mammals.

The CpG fraction. The eutherian and chicken genomes generally average around 41% G+C content, although many intervals differ substantially from the average, particularly in humans (Supplementary Notes 23). In contrast, the platypus genome averages 45.5% G+C content and rarely deviates far from the average. The opossum genome averages only 38% G+C content and also has a narrow distribution (Supplementary Fig. 13). The source of the elevated G+C fraction in platypus remains unclear. It is explained only in part by monotreme interspersed repeat elements, as platypus DNA outside of known interspersed repeats is 44.7% G+C. Furthermore, tandem repeats of short DNA motifs (microsatellites) in platypus show an A+T bias, as with other mammals. Recombination-driven biased gene conversion may be a factor, in agreement with what has been shown for eutherians[65] and marsupials[66]. This is suggested by the observation that the six platypus chromosomes where the currently mapped DNA sequence averages over 45% G+C content (that is, 17, 20, 15, 14, 10 and 11 in order of decreasing G+C fraction) are among the 10 shortest (Supplementary Fig. 14), because short chromosomes have a higher recombination rate[67]. However, a direct test

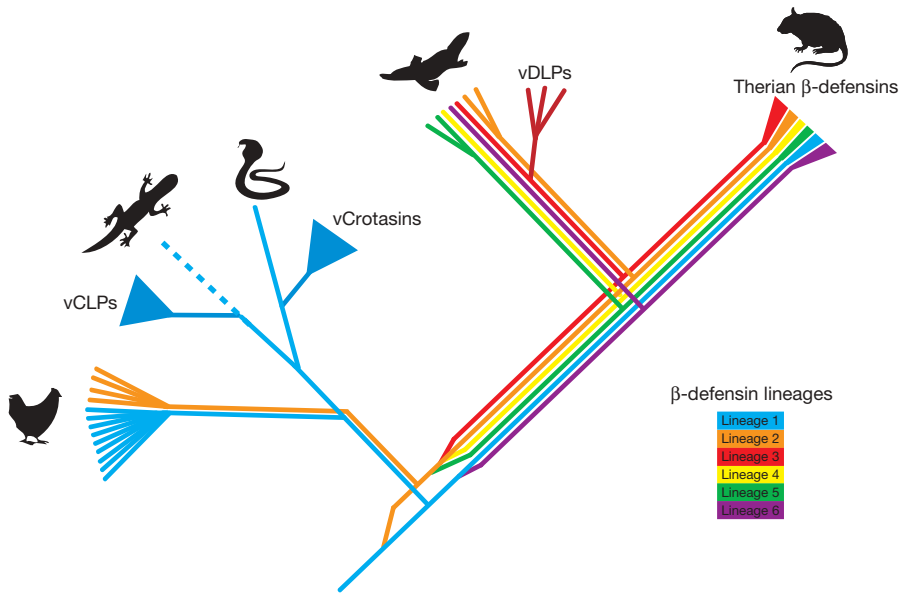


Figure 4 | The evolution of β-defensin peptides in platypus venom gland. The diagram illustrates separate gene duplications in different parts of the phylogeny for platypus venom defensin-like peptides (vDLPs), for lizard venom crotamine-like peptides (vCLPs) and for snake venom crotamines. These venom proteins have thus been co-opted from pre-existing non-toxin homologues independently in platypus and in lizards and snakes[48].

©2008 Nature Publishing Group

is currently lacking because platypus recombination rates have not been measured. A further examination of the CpG fraction, that associated with promoter elements, is found in Supplementary Notes 24 and Supplementary Fig. 15.

Conclusions

The egg-laying platypus is a remarkable species with many biological features unique among mammals. Our sequencing of the platypus genome now enables us to compare its sequence characteristics and organization with those of birds and therian mammals in order to address the questions of platypus biology and to date the emergence of mammalian traits. We report here that sequence characteristics of the platypus genome show features of reptiles as well as mammals.

Platypus contains a largely standard repertoire of non-protein-coding, ncRNAs, except for the snoRNAs, which exhibit a marked expansion associated with at least one retrotransposed subfamily. Some of these retrotransposed snoRNAs are expressed and thus may have functional roles. The platypus has fully elaborated piRNA and miRNA pathways, the latter including many monotreme-specific miRNAs and miRNAs that are shared with either mammals or chickens. Many functional assessments of these novel miRNAs remain to be carried out and will surely add to our knowledge of mammalian miRNA evolution.

The 18,527 protein-coding genes predicted from the platypus assembly fall within the range for therian genomes. Of particular interest are families of genes involved in biology that links monotremes to reptiles, such as egg-laying, vision and envenomation, as well as mammal-specific characters such as lactation, characters shared with marsupials such as antibacterial proteins, and platypus-specific characters such as venom delivery and underwater foraging. For instance, anatomical adaptations for chemoreception during underwater foraging are reflected in an unusually large repertoire of vomeronasal type 1 receptor genes. However, the repertoire of milk protein genes is typically mammalian, and the arrangement of milk protein genes seems to have been preserved since the last common ancestor of monotremes and therian mammals.

Since its initial description, the platypus has stood out as a species with a blend of reptilian and mammalian features, which is a characteristic that penetrates to the level of the genome sequence. The density and distribution of repetitive sequence, for example, reflects this fact. The high frequency of interspersed repeats in the platypus genome, although typical for mammalian genomes, is in contrast with the observed mean microsatellite coverage, which appears more reptilian. Additionally, the correlation of parent-of-origin-specific expression patterns in regions of reduced interspersed repeats in the platypus suggests that the evolution of imprinting in therians is linked to the accumulation of repetitive elements.

We find that the mixture of reptilian, mammalian and unique characteristics of the platypus genome provides many clues to the function and evolution of all mammalian genomes. The wealth of new findings and confirmation of existing knowledge immediately evident from the release of these data promise that the availability of

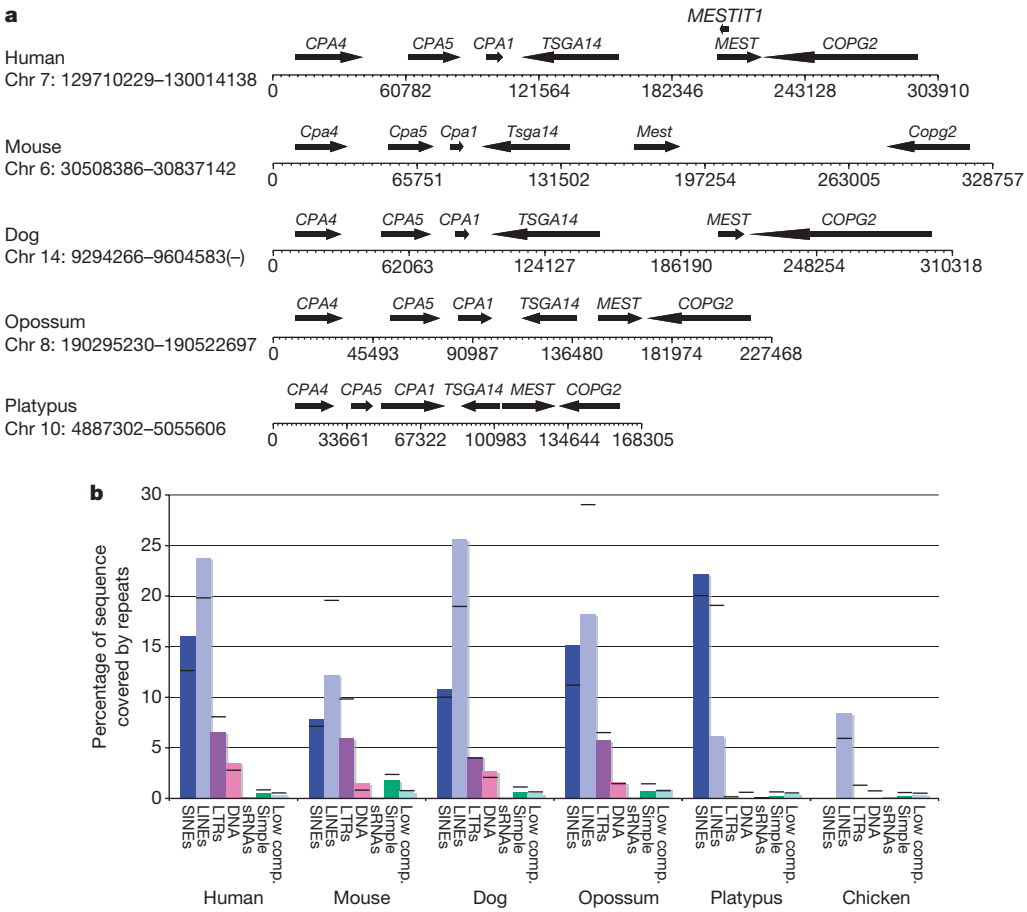


Figure 5 | Comparative mammalian analysis for a representative eutherian imprinted gene cluster (*PEG1/MEST*). **a**, The gene arrangement is conserved between mammals. However, non-coding regions are expanded in therians. Arrows indicate genes and the direction of transcription; the scale shows base pairs. **b**, Summary of repeat distribution for the *PEG1/MEST* cluster. Histograms represent the sequence (%) masked by each repeat element within the *MEST* cluster; black bars represent repeat distribution across the entire genome. With the exception of SINEs, platypus has fewer repeats of LINEs, LTRs, DNA and simple repeats (Simple) than eutherian mammals. Low comp., low complexity; sRNAs, small RNAs.

©2008 Nature Publishing Group

the platypus genome sequence will provide the critically needed background to inspire rapid advances in other investigations of mammalian biology and evolution.

METHODS SUMMARY

Tissue resources. Tissue was obtained from animals captured at the Upper Barnard River, New South Wales, Australia, during breeding season (AEEC permit number R.CG.07.03 to F. Grützner; Environment ACT permit number LI 2002 270 to J. A. M. Graves; NPWS permit number A193 to R. C. Jones; AEC permit number S-49-2006 to F. Grützner).

Sequence assembly. A total of 26.9 million reads was assembled using the PCAP software[20]. Attempts were made to assign the largest contiguous blocks of sequence to chromosomes using standard FISH techniques.

Non-coding RNAs. We used the established Rfam pipeline[68] and *de novo* sequencing to detect non-protein-coding RNAs (ncRNAs). Cloning, sequencing and annotation of sRNAs from platypus, echidna and chicken as well as miRNA sequences are described in ref. 25.

Genes. Protein-coding and non-protein-coding genes were computed using a modified version of the Ensembl pipeline (Supplementary Notes 14). Gene orthology assignment followed a procedure implemented previously[69]. Orthology rate estimation was performed with PAML[70] using the model of ref. 71. In all cases, codon frequencies were estimated from the nucleotide composition at each codon position (F3X4 model).

Genome landscape. Pairwise alignments between human and dog, mouse, opossum, platypus and chicken were projected from whole-genome alignments of 28 species (http://genome.cse.ucsc.edu/). These alignments were the basis for phylogeny, chromosome synteny, interspersed repeats, imprinting and CpG fraction analyses.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 14 September 2007; accepted 25 March 2008.

1. Home, E. A. A description of the anatomy of *Ornithorynchus paradoxus*. *Phil. Trans. R. Soc. Lond. B* **92,** 67–84 (1802).
2. Bininda-Emonds, O. R. P. *et al.* The delayed rise of present-day mammals. *Nature* **446,** 507–512 (2007).
3. Caldwell, H. The embryology of Monotremata and Marsupialia. *Phil. Trans. R. Soc. Lond. B* **178,** 463–486 (1887).
4. Griffiths, M. *Echidnas* (Pergamon, Oxford, 1968).
5. Griffiths, M. *The Biology of the Monotremes* (Academic, New York, 1978).
6. Renfree, M. B. Monotreme and marsupial reproduction. *Reprod. Fertil. Dev.* **7,** 1003–1020 (1995).
7. Renfree, M. B. *Mammal Phylogeny* (Springer, New York, 1993).
8. Watson, J. M., Meyne, J. & Graves, J. A. M. Ordered tandem arrangement of chromosomes in the sperm heads of monotreme mammals. *Proc. Natl Acad. Sci. USA* **93,** 10200–10205 (1996).
9. Greaves, I. K., Rens, W., Ferguson-Smith, M. A., Griffin, D. & Graves, J. A. M. Conservation of chromosome arrangement and position of the X in mammalian sperm suggests functional significance. *Chromosome Res.* **11,** 503–512 (2003).
10. Temple-Smith, P. & Grant, T. Uncertain breeding: a short history of reproduction in monotremes. *Reprod. Fertil. Dev.* **13,** 487–497 (2001).
11. Temple-Smith, P. D. *Seasonal Breeding Biology of the Platypus,* Ornithorynchus anatinus *(Shaw 199), with Special Reference to Male*. PhD thesis, Australian National University (1973).
12. Scheich, H., Langner, G., Tidemann, C., Coles, R. B. & Guppy, A. Electroreception and electrolocation in platypus. *Nature* **319,** 401–402 (1986).
13. Pettigrew, J. D. Electroreception in monotremes. *J. Exp. Biol.* **202,** 1447–1454 (1999).
14. Bick, Y. A. E. & Sharman, G. B. The chromosomes of the platypus (*Ornithorynchus*: Monotremata). *Cytobios* **14,** 17–28 (1975).
15. Wrigley, J. M. & Graves, J. A. M. Two monotreme cell lines, derived from female platypuses (*Ornithorhynchus anatinus;* Monotremata, Mammalia). *In Vitro* **20,** 321–328 (1984).
16. El-Mogharbel, N. *et al. DMRT* gene cluster analysis in the platypus: new insights into genomic organization and regulatory regions. *Genomics* **89,** 10–21 (2007).
17. Rens, W. *et al.* Resolution and evolution of the duck-billed platypus karyotype with an X1Y1X2Y2X3Y3X4Y4X5Y5 male sex chromosome constitution. *Proc. Natl Acad. Sci. USA* **101,** 16257–16261 (2004).
18. Grutzner, F. *et al.* In the platypus a meiotic chain of ten sex chromosomes shares genes with the bird Z and mammal X chromosomes. *Nature* **432,** 913–917 (2004).
19. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432,** 695–716 (2004).
20. Huang, X. *et al.* Application of a superword array in genome assembly. *Nucleic Acids Res.* **34,** 201–205 (2006).
21. McMillan, D. *et al.* Characterizing the chromosomes of the platypus (*Ornithorhynchus anatinus*). *Chromosome Res.* **15,** 961–974 (2007).
22. Maxwell, E. S. & Fournier, M. J. The small nucleolar RNAs. *Annu. Rev. Biochem.* **64,** 897–934 (1995).
23. Schmitz, J. *et al.* Retroposed SNOfall—A mammalian-wide comparison of platypus snoRNAs. *Genome Res.* doi:10.1101/gr.7177908 (in the press).
24. Zemann, A., op de Bekke, A., Kiefmann, M., Brosius, J. & Schmitz, J. Evolution of small nucleolar RNAs in nematodes. *Nucleic Acids Res.* **34,** 2676–2685 (2006).
25. Murchison, E. P. *et al.* Conservation of small RNA pathways in platypus. *Genome Res.* doi:10.1101/gr.73056.107 (in the press).
26. Aravin, A. A., Hannon, G. J. & Brennecke, J. The Piwi/piRNA pathway provides adaptive defense for the transposon arms race. *Science* **318,** 761–764 (2007).
27. Curwen, V. *et al.* The Ensembl automatic gene annotation system. *Genome Res.* **14,** 942–950 (2004).
28. Goodstadt, L. & Ponting, C. P. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput. Biol.* **2,** e133 (2006).
29. Goodstadt, L., Heger, A., Webber, C. & Ponting, C. P. An analysis of the gene complement of a marsupial, *Monodelphis domestica*: evolution of lineage-specific genes and giant chromosomes. *Genome Res.* **17,** 969–981 (2007).
30. Emes, R. D., Goodstadt, L., Winter, E. E. & Ponting, C. P. Comparison of the genomes of human and mouse lays the foundation of genome zoology. *Hum. Mol. Genet.* **12,** 701–709 (2003).
31. Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* **246,** 96–98 (1973).
32. Kirsch, J. A. & Mayer, G. C. The platypus is not a rodent: DNA hybridization, amniote phylogeny and the palimpsest theory. *Phil. Trans. R. Soc. Lond. B* **353,** 1221–1237 (1998).
33. Rowe, T., Rich, T. H., Vickers-Rich, P., Springer, M. & Woodburne, M. O. The oldest platypus and its bearing on divergence timing on platypus and echidna clades. *Proc. Natl Acad. Sci. USA* **105,** 1238–1242 (2008).
34. Grus, W. E., Shi, P. & Zhang, J. Largest vertebrate vomeronasal type 1 receptor (V1R) gene repertoire in the semi-aquatic platypus. *Mol. Biol. Evol.* **24,** 2153–2157 (2007).
35. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
36. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–521 (2004).
37. Aloni, R., Olender, T. & Lancet, D. Ancient genomic architecture for mammalian olfactory clusters. *Genome Biol.* **7,** R88 (2006).
38. Jovine, L., Qi, H., Williams, Z., Litscher, E. S. & Wassarman, P. M. Features that affect secretion and assembly of zona pellucida glycoproteins during mammalian oogenesis. *Soc. Reprod. Fertil.* **63** (suppl.), 187–201 (2007).
39. Riggio, M., Scudiero, R., Filosa, S. & Parisi, E. Sex- and tissue-specific expression of aspartic proteinases in *Danio rerio* (zebrafish). *Gene* **260,** 67–75 (2000).
40. Nixon, B., Aitken, R. J. & McLaughlin, E. A. New insights into the molecular mechanisms of sperm-egg interaction. *Cell. Mol. Life Sci.* **64,** 1805–1823 (2007).
41. Morel, L., Dufaure, J. P. & Depeiges, A. The lipocalin sperm coating lizard epididymal secretory protein family: mRNA structural analysis and sequential expression during the annual cycle of the lizard, *Lacerta vivipara*. *J. Mol. Endocrinol.* **24,** 127–133 (2000).
42. Schlomann, U. *et al.* The metalloprotease disintegrin ADAM8. Processing by autocatalysis is required for proteolytic activity and cell adhesion. *J. Biol. Chem.* **277,** 48210–48219 (2002).
43. Oftedal, O. The origin of lactation as a water source for parchment-shelled eggs. *J. Mammary Gland Biol. Neoplasia* **7,** 253–266 (2002).
44. Kawasaki, K. & Weiss, K. M. Mineralized tissue and vertebrate evolution: the secretory calcium-binding phosphoprotein gene cluster. *Proc. Natl Acad. Sci. USA* **100,** 4060–4065 (2003).
45. Lester, K. S., Boyde, A., Gilkeson, C. & Archer, M. Marsupial and monotreme enamel structure. *Scanning Microsc.* **1,** 401–420 (1987).
46. de Plater, G., Martin, R. L. & Milburn, P. J. A pharmacological and biochemical investigation of the venom from the platypus (*Ornithorhynchus anatinus*). *Toxicon* **33,** 157–169 (1995).
47. Fry, B. G. From genome to "venome": molecular origin and evolution of the snake venom proteome inferred from phylogenetic analysis of toxin sequences and related body proteins. *Genome Res.* **15,** 403–420 (2005).
48. Whittington, C. *et al.* Defensins and the convergent evolution of platypus and reptile venom genes. *Genome Res.* doi:10.1101/gr.7149808 (in the press).
49. Diener, E. & Ealey, E. H. Immune system in a monotreme: studies on the Australian echidna (*Tachyglossus aculeatus*). *Nature* **208,** 950–953 (1965).
50. Kelley, J., Walter, L. & Trowsdale, J. Comparative genomics of natural killer cell receptor gene clusters. *PLoS Genet.* **1,** 129–139 (2005).
51. Belov, K. *et al.* Characterization of the opossum immune genome provides insights into the evolution of the mammalian immune system. *Genome Res.* **17,** 982–991 (2007).
52. Durr, U. H., Sudheendra, U. S. & Ramamoorthy, A. LL-37, the only human member of the cathelicidin family of antimicrobial peptides. *Biochim. Biophys. Acta* **1758,** 1408–1425 (2006).
53. Bals, R. & Wilson, J. M. Cathelicidins—a family of multifunctional antimicrobial peptides. *Cell. Mol. Life Sci.* **60,** 711–720 (2003).

©2008 Nature Publishing Group

54. Zanetti, M. Cathelicidins, multifunctional peptides of the innate immunity. *J. Leukoc. Biol.* **75,** 39–48 (2004).
55. Veyrunes, F. *et al.* Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. *Genome Res.* doi:10.1101/gr.7101908 (in the press).
56. Itoh, Y. *et al.* Dosage compensation is less effective in birds than in mammals. *J. Biol.* **6,** 2 (2007).
57. Gemmell, N. J. *et al.* Determining platypus relationships. *Aust. J. Zool.* **43,** 283–291 (1995).
58. Smit, A. F. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9,** 657–663 (1999).
59. Kriegs, J. O. *et al.* Waves of genomic hitchhikers shed light on the evolution of gamebirds (Aves: Galliformes). *BMC Evol. Biol.* **7,** 190 (2007).
60. Ross, M. T. *et al.* The DNA sequence of the human X chromosome. *Nature* **434,** 325–337 (2005).
61. Killian, J. K. *et al.* M6P/IGF2R imprinting evolution in mammals. *Mol. Cell* **5,** 707–716 (2000).
62. Suzuki, S. *et al.* Retrotransposon silencing by DNA methylation can drive mammalian genomic imprinting. *PLoS Genet.* **3,** e55 (2007).
63. Edwards, C. A. *et al.* The evolution of imprinting: chromosomal mapping of orthologues of mammalian imprinted domains in monotreme and marsupial mammals. *BMC Evol. Biol.* **7,** 157 (2007).
64. Rapkins, R. W. *et al.* The Prader-Willi/Angelman imprinted domain was assembled recently from non-imprinted components. *PLoS Genet.* **2,** 1–10 (2006).
65. Meunier, J. & Duret, L. Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.* **21,** 984–990 (2004).
66. Mikkelsen, T. S. *et al.* Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* **447,** 167–177 (2007).
67. Coop, G. & Przeworski, M. An evolutionary view of human recombination. *Nature Rev. Genet.* **8,** 23–34 (2007).
68. Griffiths-Jones, S. *et al.* Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33,** D121–D124 (2005).
69. Heger, A. & Ponting, C. P. Evolutionary rate analyses of orthologues and paralogues from twelve *Drosophila* genomes. *Genome Res.* **17,** 1837–1849 (2007).
70. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13,** 555–556 (1997).
71. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11,** 725–736 (1994).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The sequencing of platypus was funded by the National Human Genome Research Institute (NHGRI). This research was supported by grant HG002238 from the NHGRI (W.M.), NGFN (0313358A; to J.S. and J.B.), the DFG (SCHM 1469; to J.S. and J.B.), National Science Foundation BCS-0218338 (M.A.B.) and EPS-0346411 (M.A.B.), National Institutes of Health RO1 GM59290 (M.A.B.), National Institutes of Health RO1HG02385 (E.E.E), Australian Research Council (F.G.), UK Medical Research Council (C.P.P. and A.H.), Ministry of Science-Spain (X.S.P. and C.L.-O.) and the State of Louisiana Board of Regents Support Fund (M.A.B.). We thank T. Grant, S. Akiyama, P. Temple-Smith, R. Whittington and the Queensland Museum for platypus sample collection and DNA, and Macquarie Generation and Glenrock station for providing access and facilities during sampling. Approval to collect animals was granted by the New South Wales National Parks and Wildlife Services, New South Wales. Funding support for some platypus samples was provided by Australian Research Council and W.V. Scott Foundation. We thank M. Shelton, I. Elton and the Healesville Sanctuary for platypus pictures. We thank L. Duret for assistance on genome landscape analysis; G. Shaw for use of the silhouettes on Fig. 1; and Z.-X. Luo, M. Archer and R. Beck for advice on the Fig. 1 phylogeny. We acknowledge the approved use of the green anole lizard sequence data provided by the Broad Institute. Resources for exploring the sequence and annotation data are available on browser displays available at UCSC (http://genome.ucsc.edu), Ensembl (http://www.ensembl.org) and the NCBI (http://www.ncbi.nlm.nih.gov).

Author Information The *Ornithorhynchus anatinus* whole-genome shotgun project has been deposited in DDBJ/EMBL/GenBank under the project accession AAPN00000000. The version described in this paper is the first version, AAPN01000000. The SNPs have been deposited in the dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/) with Submitter Method IDs PLATYPUS-ASSEMBLY_SNPS_200801 and PLATYPUS-READS_SNPS_200801.

Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. Correspondence and requests for materials should be addressed to W.C.W. (wwarren@wustl.edu) or R.K.W. (rwilson@wustl.edu).

Wesley C. Warren[1], LaDeana W. Hillier[1], Jennifer A. Marshall Graves[2], Ewan Birney[3], Chris P. Ponting[4], Frank Grützner[5], Katherine Belov[6], Webb Miller[7], Laura Clarke[8], Asif T. Chinwalla[1], Shiaw-Pyng Yang[1], Andreas Heger[4], Devin P. Locke[1], Pat Miethke[2], Paul D. Waters[2], Frédéric Veyrunes[2,9], Lucinda Fulton[1], Bob Fulton[1], Tina Graves[1], John Wallis[1], Xose S. Puente[10], Carlos López-Otín[10], Gonzalo R. Ordóñez[10], Evan E. Eichler[11], Lin Chen[11], Ze Cheng[11], Janine E. Deakin[2], Amber Alsop[2], Katherine Thompson[2], Patrick Kirby[2], Anthony T. Papenfuss[12], Matthew J. Wakefield[12], Tsviya Olender[13], Doron Lancet[13], Gavin A. Huttley[14], Arian F. A. Smit[15], Andrew Pask[16], Peter Temple-Smith[16,17], Mark A. Batzer[18], Jerilyn A. Walker[18], Miriam K. Konkel[18], Robert S. Harris[7], Camilla M. Whittington[6], Emily S. W. Wong[6], Neil J. Gemmell[19], Emmanuel Buschiazzo[19], Iris M. Vargas Jentzsch[19], Angelika Merkel[19], Juergen Schmitz[20], Anja Zemann[20], Gennady Churakov[20], Jan Ole Kriegs[20], Juergen Brosius[20], Elizabeth P. Murchison[21], Ravi Sachidanandam[21], Carly Smith[21], Gregory J. Hannon[21], Enkhjargal Tsend-Ayush[5], Daniel McMillan[2], Rosalind Attenborough[2], Willem Rens[9], Malcolm Ferguson-Smith[9], Christophe M. Lefèvre[22,23], Julie A. Sharp[23], Kevin R. Nicholas[23], David A. Ray[24], Michael Kube[25], Richard Reinhardt[25], Thomas H. Pringle[26], James Taylor[27], Russell C. Jones[28], Brett Nixon[28], Jean-Louis Dacheux[29], Hitoshi Niwa[30], Yoko Sekita[30], Xiaoqiu Huang[31], Alexander Stark[32], Pouya Kheradpour[32], Manolis Kellis[32], Paul Flicek[3], Yuan Chen[3], Caleb Webber[4], Ross Hardison[7], Joanne Nelson[1], Kym Hallsworth-Pepin[1], Kim Delehaunty[1], Chris Markovic[1], Pat Minx[1], Yucheng Feng[1], Colin Kremitzki[1], Makedonka Mitreva[1], Jarret Glasscock[1], Todd Wylie[1], Patricia Wohldmann[1], Prathapan Thiru[1], Michael N. Nhan[1], Craig S. Pohl[1], Scott M. Smith[1], Shunfeng Hou[1], Marilyn B. Renfree[16], Elaine R. Mardis[1] & Richard K. Wilson[1]

[1]Genome Sequencing Center, Washington University School of Medicine, Campus Box 8501, 4444 Forest Park Avenue, St Louis, Missouri 63108, USA. [2]Australian National University, Canberra, Australian Capital Territory 0200, Australia. [3]EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [4]MRC Functional Genetics Unit, University of Oxford, Department of Human Physiology, Anatomy and Genetics, South Parks Road, Oxford OX1 3QX, UK. [5]Discipline of Genetics, School of Molecular & Biomedical Science, The University of Adelaide, 5005 South Australia, Australia. [6]Faculty of Veterinary Science, The University of Sydney, Sydney, New South Wales 2006, Australia. [7]Center for Comparative Genomics and Bioinformatics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA. [8]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [9]Cambridge University, Department of Veterinary Medicine, Madingley Road, Cambridge CB3 0ES, UK. [10]Instituto Universitario de Oncologia, Departamento de Bioquimica y Biologia Molecular, Universidad de Oviedo, 33006-Oviedo, Spain. [11]Department of Genome Sciences, Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA. [12]The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3050, Australia. [13]Crown Human Genome Center, Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel. [14]John Curtin School of Medical Research, The Australian National University, Canberra, Australian Capital Territory 0200, Australia. [15]Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103-8904, USA. [16]Department of Zoology, The University of Melbourne, Victoria 3010, Australia. [17]Monash Institute of Medical Research, 27–31 Wright Street, Clayton, Victoria 3168, Australia. [18]Department of Biological Sciences, Center for Bio-Modular Multi-Scale Systems, Louisiana State University, 202 Life Sciences Building, Baton Rouge, Louisiana 70803, USA. [19]School of Biological Sciences, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand. [20]Institute of Experimental Pathology, University of Muenster, Von-Esmarch-Strasse 56, D-48149 Muenster, Germany. [21]Cold Spring Harbor Laboratory, Howard Hughes Medical Institute, Cold Spring Harbor, New York 11724, USA. [22]Victorian Bioinformatics Consortium, Monash University, Clayton, Victoria 3080, Australia. [23]CRC for Innovative Dairy Products, Department of Zoology, University of Melbourne, Victoria 3010, Australia. [24]Department of Biology, West Virginia University, Morgantown, West Virginia 26505, USA. [25]MPI Molecular Genetics, D-14195 Berlin-Dahlem, Ihnestrasse 73, Germany. [26]Sperling Foundation, Eugene, Oregon 97405, USA. [27]Courant Institute, New York University, New York, New York 10012, USA. [28]Discipline of Biological Sciences, School of Environmental and Life Sciences, University of Newcastle, New South Wales 2308, Australia. [29]UMR INRA-CNRS 6073, Physiologie de la Reproduction et des Comportements, Nouzilly 37380, France. [30]Laboratory for Pluripotent Cell Studies, RIKEN Center for Developmental Biology (CDB), 2-2-3 Minatojima-minamimachi, Chuo-ku, Kobe 6500047, Japan. [31]Department of Computer Science, Iowa State University, 226 Atanasoff Hall, Ames, Iowa 50011, USA. [32]The Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02139, USA.

©2008 Nature Publishing Group

METHODS

Sequence assembly. A total of 26.9 million reads was assembled using the PCAP software[20]. Assembly quality assessment accounted for read depth, chimaeric reads, repeat content, cloning bias, G+C content and heterozygosity (Supplementary Notes 4–11). We identified a total of ~1.2 million single nucleotide polymorphisms (SNPs) within the 1.84-Gb sequenced female platypus genome using two independent analyses, SSAHA2 (SSAHA: a fast search method for large DNA databases[72]) and PCAP output[20] (Supplementary Notes 11).

Non-coding RNAs. snoRNA annotation is as described in ref. 23. miRNAs sharing a heptamer at nucleotide position 2–8 were defined as a family. Homology with mouse/human miRNAs was based on annotated miRNAs in Rfam (http://microrna.sanger.ac.uk/sequences/index.shtml). piRNA sequences have been submitted to GEO (http://www.ncbi.nlm.nih.gov/geo/). miRNA total cloning frequency was normalized across tissue libraries by scaling cloning frequency per library by a factor representing total number of miRNA reads per library.

Genes. Orthologue groups were selected based on whether they contained genes predicted only from the platypus, and not from the chicken, opossum, dog, mouse or human genome assemblies (Supplementary Notes 15–17). Other groups were selected where the number of in-paralogous platypus genes exceeded the numbers of the other (chicken, opossum, dog, mouse and human) terminal lineages. Some of these groups represent erroneous gene predictions where, for example, protein-coding sequence predictions represented instead transposed element or highly repetitive sequence, or overlapped, on the reverse strand, other well-established coding sequence. Such instances were discarded. Lineage-specific gene loss was detected by inspection of BLASTZ alignment chains and nets at the UCSC Genome Browser (http://genome.cse.ucsc.edu/); by the interrogation of all known cDNA, EST and protein sequences held in GenBank using BLAST; and by attempting to predict orthologous genes within genomic intervals flanked by syntenic anchors.

Genome landscape. To establish phylogeny we extended the basic data sampling approach described previously[73] to protein-coding genes, and used established techniques to analyse protein-coding indels[74] and retrotransposon insertions[75] (Supplementary Notes 19).

The population structure of 90 platypuses from different regions in Australia was determined using Structure software v2.1 (ref. 76) using genotypes of 57 polymorphic Mon-1 and LINE2 loci. Five thousand replications were examined (Supplementary Notes 21).

Microsatellites were identified across the platypus genome (ornAna1) combining two programs: Tandem Repeat Finder (TRF)[77] and Sputnik[78] (Supplementary Notes 22).

For the imprinting cluster of *PEG1/MEST*, comparative maps were complied from Vega annotations for the mouse and human, and Ensembl gene builds for other species. Multiple alignments of each region for repeat distribution analyses were constructed using MLAGAN[79] with translated anchoring.

We examined genomic assemblies for human (hg18), mouse (musMus8), dog (canFam2), opossum (monDom4), platypus (ornAna1) and chicken (galGal3), downloaded from the UCSC Genome Browser (http://genome.ucsc.edu), and computed the fraction of G+C nucleotides in each non-overlapping 10,000-bp window free of ambiguous bases. Bases in repeats were not distinguished and were counted along with non-repeat bases. For platypus all assembled sequence was analysed; for the other species only bases assigned to chromosomes were used.

72. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11,** 1725–1729 (2001).
73. Huttley, G. A., Wakefield, M. J. & Easteal, S. Rates of genome evolution and branching order from whole genome analysis. *Mol. Biol. Evol.* **24,** 1722–1730 (2007).
74. Murphy, W. J., Pringle, T. H., Crider, T. A., Springer, M. S. & Miller, W. Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Res.* **17,** 413–421 (2007).
75. Kriegs, J. O. *et al.* Retroposed elements as archives for the evolutionary history of placental mammals. *PLoS Biol.* **4,** e91 (2006).
76. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959 (2000).
77. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27,** 573–580 (1999).
78. La Rota, M., Kantety, R. V., Yu, J. K. & Sorrells, M. E. Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. *BMC Genomics* **6,** 23–29 (2005).
79. Brudno, M. *et al.* LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13,** 721–731 (2003).

©2008 Nature Publishing Group

LETTERS

An endogenous small interfering RNA pathway in *Drosophila*

Benjamin Czech[1]*, Colin D. Malone[1]*, Rui Zhou[2], Alexander Stark[3,4], Catherine Schlingeheyde[1], Monica Dus[1], Norbert Perrimon[2], Manolis Kellis[3], James A. Wohlschlegel[5], Ravi Sachidanandam[1], Gregory J. Hannon[1] & Julius Brennecke[1]

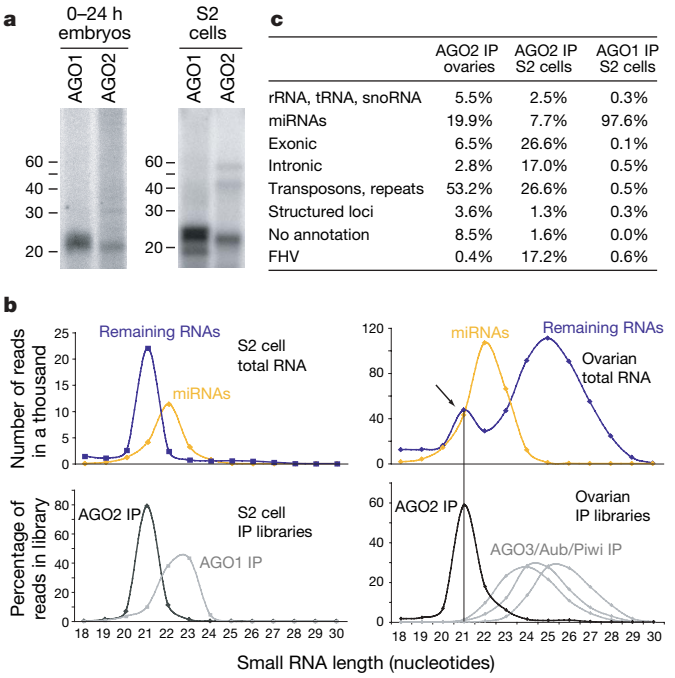
***Drosophila* endogenous small RNAs are categorized according to their mechanisms of biogenesis and the Argonaute protein to which they bind. MicroRNAs are a class of ubiquitously expressed RNAs of ~22 nucleotides in length, which arise from structured precursors through the action of Drosha–Pasha and Dicer-1–Loquacious complexes[1–7]. These join Argonaute-1 to regulate gene expression[8,9]. A second endogenous small RNA class, the Piwi-interacting RNAs, bind Piwi proteins and suppress transposons[10,11]. Piwi-interacting RNAs are restricted to the gonad, and at least a subset of these arises by Piwi-catalysed cleavage of single-stranded RNAs[12,13]. Here we show that *Drosophila* generates a third small RNA class, endogenous small interfering RNAs, in both gonadal and somatic tissues. Production of these RNAs requires Dicer-2, but a subset depends preferentially on Loquacious[1,4,5] rather than the canonical Dicer-2 partner, R2D2 (ref. 14). Endogenous small interfering RNAs arise both from convergent transcription units and from structured genomic loci in a tissue-specific fashion. They predominantly join Argonaute-2 and have the capacity, as a class, to target both protein-coding genes and mobile elements. These observations expand the repertoire of small RNAs in *Drosophila*, adding a class that blurs distinctions based on known biogenesis mechanisms and functional roles.**

Drosophila melanogaster expresses five Argonaute proteins, which segregate into two classes. The Piwi proteins (Piwi, Aubergine and AGO3) are expressed in gonadal tissues and act with Piwi-interacting RNAs (piRNAs) to suppress mobile genetic elements[10,11]. The Argonaute class contains AGO1 and AGO2. AGO1 binds microRNAs (miRNAs) and regulates gene expression[8,9]. The endogenous binding partners of AGO2 have remained enigmatic.

We generated transgenic flies expressing epitope-tagged AGO2 under the control of its endogenous promoter. Tagged AGO2 localized to the cytoplasm of germline and somatic cells of the ovary (Supplementary Fig. 1). Immunoprecipitated AGO2-associated RNAs differed in their mobility from those bound to AGO1 (Fig. 1a). Deep sequencing of small RNAs from AGO1 and AGO2 complexes yielded 2,094,408 AGO1-associated RNAs and 916,834 AGO2-associated RNAs from Schneider (S2) cells, and 455,227 AGO2-associated RNAs from ovaries that matched perfectly to the *Drosophila* genome. We also sequenced three libraries derived from 18–29-nucleotide RNAs (936,833 sequences from wild-type ovaries, 1,042,617 sequences from *Dicer-2* (*Dcr-2*) mutant ovaries, and 1,946,339 sequences from *loquacious* (*loqs*) mutant ovaries) and an 18–24-nucleotide library from wild-type testes (522,848 sequences). Finally, we added to our analysis 92,363 published sequences derived from 19–26-nucleotide RNAs from S2 cells[15].

We noted that among the ~50% of AGO2-associated RNAs from S2 cells that did not match the genome, ~17% matched the flock house virus (FHV), a pathogenic RNA virus and reported target for RNAi in flies[16,17]. These probably arose because of persistent infection of our S2 cultures.

After excluding presumed degradation products of abundant cellular RNAs, we divided each of the total RNA libraries into two categories: annotated miRNAs and the remainder (Fig. 1b). For the S2 cell library, the size distribution of these populations formed two



c

	AGO2 IP ovaries	AGO2 IP S2 cells	AGO1 IP S2 cells
rRNA, tRNA, snoRNA	5.5%	2.5%	0.3%
miRNAs	19.9%	7.7%	97.6%
Exonic	6.5%	26.6%	0.1%
Intronic	2.8%	17.0%	0.5%
Transposons, repeats	53.2%	26.6%	0.5%
Structured loci	3.6%	1.3%	0.3%
No annotation	8.5%	1.6%	0.0%
FHV	0.4%	17.2%	0.6%

Figure 1 | AGO2 binds endogenous small RNAs. a, RNA was isolated from AGO1 and AGO2 immunoprecipitates (IP) from embryos and S2 cells. **b**, Length profiles for small RNAs isolated from S2 cells and ovaries are shown. Species are split into miRNAs and remainder, and are compared to those obtained from individual Argonaute complexes (as indicated). **c**, Annotation of AGO1- and AGO2-associated small RNAs from ovaries and S2 cells. rRNA, ribosomal RNA; snoRNA, small nucleolar RNA; tRNA, transfer RNA.

[1]Watson School of Biological Sciences, Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA. [2]Harvard Medical School, Department of Genetics, Howard Hughes Medical Institute, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA. [3]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02141, USA. [4]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [5]Department of Biological Chemistry, David Geffen School of Medicine, University of California at Los Angeles, Los Angeles, California 90095, USA.
*These authors contributed equally to this work.

©2008 Nature Publishing Group

peaks, with non-miRNAs lying at 21 nucleotides and miRNAs exhibiting a broader peak from 21 to 23 nucleotides. Libraries derived from AGO1 and AGO2 complexes almost precisely mirrored these two size classes. In the ovary library, this approach revealed three size classes. Whereas two reflected those seen in S2 cells, a third class comprised piRNAs. Again, RNA size profiles from AGO2 or Piwi family immunoprecipitates[12] mirrored those within the total ovary library. These data demonstrate that AGO2 is complexed with a previously uncharacterized population of small RNAs.

Whereas known miRNAs comprised more than 97% of AGO1-associated RNAs in S2 cells, they made up only 8% or 20% of the AGO2-bound species in S2 cells or ovaries, respectively. The remaining small RNAs in AGO2 complexes formed a complex mixture of endogenous siRNAs (endo-siRNAs; Fig. 1c). Among these, transposons and satellite repeats contributed substantially to AGO2-associated small RNAs in S2 cells (27%) and ovaries (53%). The nature of the transposons giving rise to abundant siRNAs in ovaries and S2 cells differed substantially (Fig. 2a), probably reflecting differential expression of specific transposons in these tissues. Unlike piRNAs[12,13,18,19], neither somatic nor germline siRNAs exhibited a pronounced enrichment for sense or antisense species (Supplementary Fig. 2a).

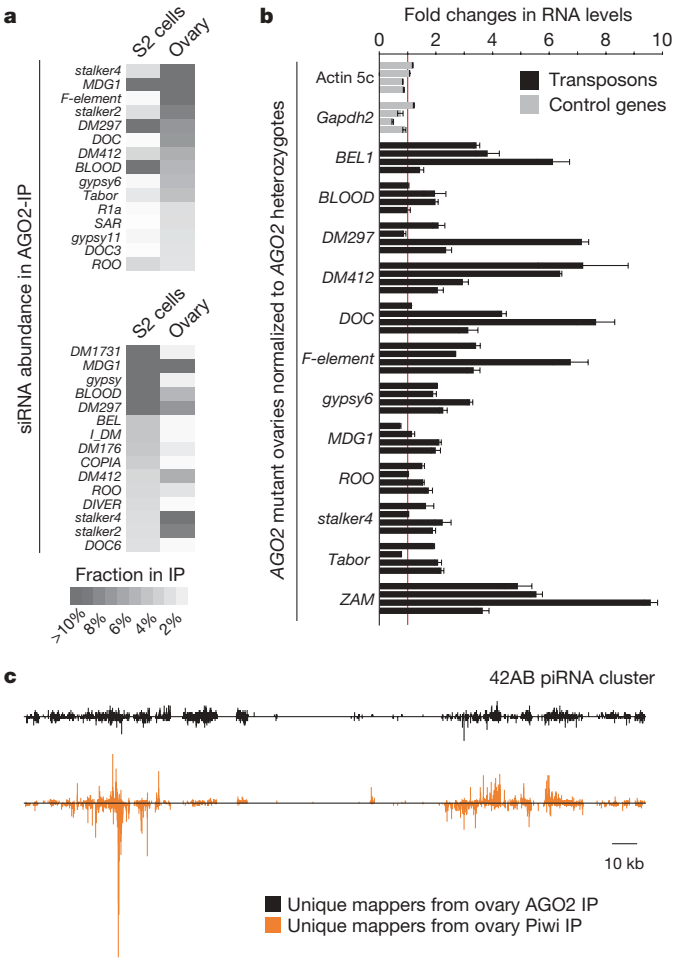


Figure 2 | A subset of endo-siRNAs originates from transposons. **a**, Indicated are the cloning frequencies of AGO2-bound siRNAs in ovaries and S2 cells that match individual transposons. **b**, RNA levels of twelve transposons and two control genes in ovaries mutant for *AGO2* as compared to *AGO2* heterozygotes (four biological replicates; error bars indicate technical variation). **c**, Distributions of AGO2-bound siRNAs (black) and Piwi-bound piRNAs (orange) from ovaries on the piRNA cluster at cytological position 42AB (ref. 12; relative abundances of both populations can be estimated from Supplementary Fig. 2d).

In accord with these findings, knockdown of *AGO2* in S2 cells leads to increased expression of several mobile elements[20]. In the germ line, the Piwi–piRNA system has been reported as the dominant transposon-silencing pathway[19]. Nevertheless, we found that several transposons, with a potential to be targeted by siRNAs, were substantially derepressed in *AGO2* mutant or *Dcr-2* mutant ovaries (Fig. 2b and Supplementary Fig. 2c). Although comparisons of relative abundance were difficult, both piRNAs and siRNAs mapped to piRNA clusters, with the regions that generate uniquely mapping species generally overlapping (Fig. 2c and Supplementary Fig. 2d). Thus, piRNA loci are a possible source for antisense RNAs matching transposons and might serve a dual function in small RNA generation. Considered together, these data suggest that endo-siRNAs repress the expression of mobile elements, in some tissues acting alongside piRNA pathways.

To probe the nature of the remaining endo-siRNAs, we computationally extracted genomic sites, which give rise to multiple uniquely mapping RNAs that do not fall into heterochromatic regions. These generally segregated into two categories, which we term structured loci and convergently transcribed loci.

Transcripts from structured loci can fold to form extensive double-stranded RNA directly. The two major loci, termed *esi-1* and *esi-2* (Fig. 3a and Supplementary Fig. 3), gave rise to half of the 20 most abundant endo-siRNAs in ovaries and also generated siRNAs in embryos, larvae and adults (not shown). *esi-1*, annotated as CG18854, can produce an ~400-base pair (bp) dsRNA through interaction of its 5′ and 3′ untranslated regions (UTRs; Supplementary Fig. 3). *esi-2* overlaps with CG4068 and consists of 20 palindromic ~260-nucleotide repeats (Fig. 3a). All siRNAs derived from these two loci arise from one genomic strand. In some previously characterized instances (for example, *Arabidopsis trans*-acting-siRNAs[21]) Dicer generates 'phased' siRNAs with 5′ ends showing a 21-nucleotide periodicity. In all tissues examined, *esi-1* and *esi-2* produced phased siRNAs, consistent with a defined initiation site for Dicer processing (Fig. 3a and Supplementary Fig. 3). Phasing was not observed for viral or repeat-derived siRNAs. Finally, siRNAs from both loci also joined AGO1 in proportions greater than siRNAs produced from transposons and repeats, perhaps owing to the imperfect nature of the dsRNA that they produce[22,23] (Fig. 1c).

AGO2 regulates gene expression by cleavage of complementary sites rather than by recognition of seed sites typical of AGO1–miRNA-mediated regulation[23]. We searched for possible targets of endo-siRNAs by identifying transcripts with substantial complementarity. A highly abundant siRNA from *esi-2* is highly complementary to the coding sequence of the DNA-damage-response gene *mutagen-sensitive 308* (*mus308*). Using a modified rapid amplification of cDNA ends (RACE) protocol, we detected *mus308* fragments with 5′ ends corresponding precisely to predicted endo-siRNA cleavage sites (Fig. 3b). Moreover, *AGO2* and *Dcr-2* loss consistently increased *mus308* expression in testis and to a lesser extent in ovaries, consistent with the relative abundance of *esi-2* siRNAs in these tissues (Fig. 3b, c). Finally, a reporter gene containing two *mus308* target sites was significantly derepressed in S2 cells on depletion of Dcr-2 or AGO2 but not of Dcr-1 or AGO1 (Fig. 3c). Although extensive complementarity between other endo-siRNAs and messenger RNAs was rare, we found several *esi-1*-derived siRNAs complementary to CG8289 (Supplementary Fig. 3), suggesting a potential regulatory interaction *in vivo*.

A second group of siRNA-generating loci contained regions in which dsRNAs can arise from convergent transcription. If sorted for siRNA density, most of the top 50 ovarian and S2 cell siRNA loci lay in regions where annotated 3′ UTRs or expressed-sequence-tags corresponding to convergently transcribed protein-coding genes overlap (Supplementary Tables 1 and 2). Typically, siRNAs arise on both genomic strands but only from overlapping portions of convergent transcripts (Fig. 3d). Examining all 998 convergently transcribed gene pairs in the *Drosophila* genome with annotated

©2008 Nature Publishing Group

overlapping transcripts, we found the peak abundance of ovarian siRNAs to be at the centre of the overlap, with sharp declines away from this region (Supplementary Fig. 4). In an alternative arrangement, *Pgant35A* produces sense and antisense siRNAs across its entire annotated transcript, consistent with expressed-sequence-tag support for antisense transcription traversing this locus (Supplementary Fig. 5).

Thus, a large number of *Drosophila* genes generate endogenous siRNAs, with most having perfect complementarity to the 3′ UTRs of neighbouring genes. Relative levels of endo-siRNAs generated from each convergent transcription unit were low (not shown), and we found no or little change (up to a ~1.3-fold increase) in the expression of such genes in *AGO2* mutant ovaries. Possibly, the level of small RNAs produced by this genomic arrangement is

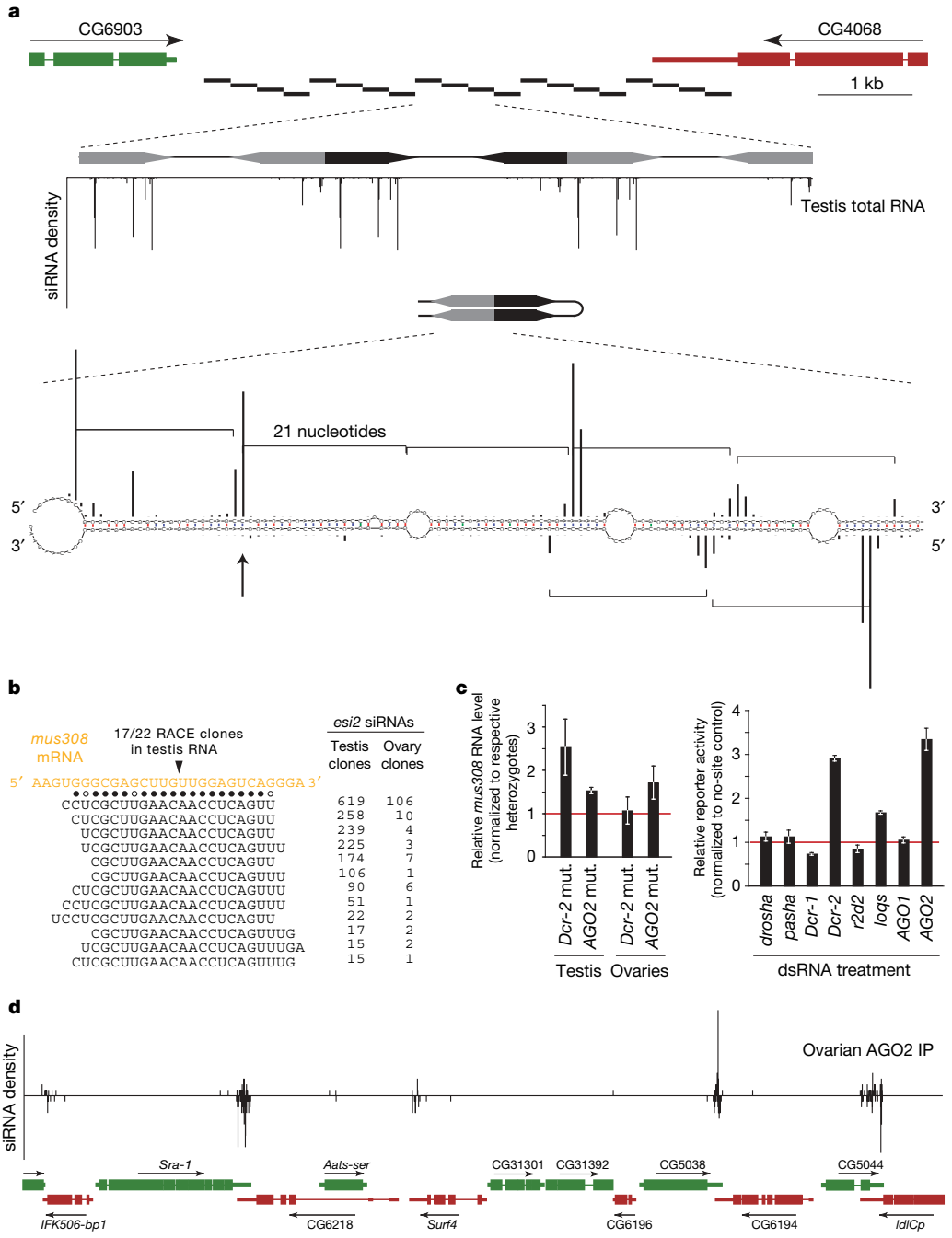


Figure 3 | Two types of genic endo-siRNA loci in *Drosophila*. **a**, FlyBase gene structure at the *esi-2* locus indicating the twenty ~260-bp repeats of *esi-2* (black bars). Below, three repeats are magnified and the small RNA density is depicted. Most siRNAs match up to 20 times and cannot be unambiguously assigned to a chromosomal location. Each repeat is a palindromic sequence, offering a multitude of possible structures. One structure is shown with the 5′ ends of cloned siRNAs as black bars (the height correlates with cloning frequency; phasing is indicated by brackets). **b**, An abundant siRNA from *esi-2* (indicated by an arrow in **a**) is highly complementary to the *mus308* mRNA. Shown is the *mus308* target site and cloned siRNAs (solid dots indicate canonical base pairs; open dots indicate GU base pairs). The only detected cleavage site within this duplex is indicated above. **c**, Shown are *mus308* transcript levels from *AGO2* and *Dcr-2* mutant (mut.) flies compared to their respective heterozygotes (error bars indicate standard deviation). To the right, average reporter levels (error bars indicate standard deviation) of a construct containing two *mus308* target sites in S2 cells depleted of the indicated genes by RNAi are shown. **d**, An example of siRNAs arising from convergent transcription units. A 30-kb region containing multiple instances of convergently transcribed genes is displayed, with the density of AGO2-associated RNAs in ovaries shown above.

©2008 Nature Publishing Group

inconsequential, amounting to noise within silencing pathways. However, there are probably circumstances wherein regulation by such arrangements might substantially impact expression.

In S2 cells, two neighbouring loci encoded nearly 16% of AGO2-associated RNAs (Supplementary Table 2). These reside within a large intron of *klarsicht* (Supplementary Fig. 6) and did not generate siRNAs in any other tissue. A similar locus, corresponding to CG14033, was found within an intron of *thickveins* (Supplementary Fig. 7) and gave rise to testis-specific siRNAs. Although the function of both siRNA clusters is unclear, the *thickveins* cluster shares considerable complementarity to CG9203, and loss of *AGO2* and *Dcr-2* mildly increased CG9203 mRNA levels in testis but not in ovaries (Supplementary Fig. 7).

Dcr-2 has been implicated in the production of siRNAs from viral replication intermediates or exogenously introduced dsRNAs, whereas Dcr-1 has been linked to miRNA biogenesis[6,16,17]. In agreement with these observations, all endo-siRNA classes were lost in *Dcr-2* mutant ovaries (Fig. 4a). To obtain more insight into the genetic requirements for endo-siRNA biogenesis and stability, we depleted components of siRNA and miRNA pathways in S2 cells and analysed levels of abundant siRNAs derived from structured loci (Fig. 4b and Supplementary Fig. 8). Although depletion of Dcr-2 and AGO2 resulted in substantial reductions in siRNA levels, little or no changes were observed on Drosha, Pasha, Dcr-1 or AGO1 depletion. Unexpectedly, we found virtually no requirement for the Dcr-2 partner R2D2 (ref. 14) but a strong requirement for the Dcr-1 partner Loquacious[1,4,5]. Only one analysed siRNA exhibited partial dependence on R2D2, potentially correlating with the extensive dsRNA character of its precursor duplex (Supplementary Fig. 9). Artificial sensors for endo-siRNAs from *esi-1* and *esi-2* in S2 cells gave patterns of de-repression that matched our analysis of endo-siRNA levels (Figs 3c and 4c).

Analysis of the most abundant siRNA from *esi-2* in flies mutant for *Dcr-2*, *AGO2*, *r2d2* or *loqs* extended our findings from cell culture (Supplementary Fig. 10). To examine the unexpected requirement for *loqs* more broadly, we sequenced small RNAs from *loqs*-mutant ovaries and observed a near complete loss of endo-siRNAs from structured loci (Fig. 4a). A much smaller impact of *loqs* was seen on endo-siRNAs derived from repeats and convergent transcription units. However, an involvement of Loqs and not R2D2 in the function of siRNAs derived from perfect dsRNA precursors was supported by analysing the impact of depleting siRNA/miRNA pathway components on the ability to suppress FHV replication in our infected S2 cell cultures (Supplementary Fig. 11).

Our results uncover an unanticipated role for Loqs in siRNA biogenesis and suggest that R2D2 has a lesser impact on at least two types of endogenous siRNAs. It is well established that Loqs partners with Dcr-1 for miRNA processing. To probe a molecular interaction with Dcr-2, we catalogued Loqs binding partners using quantitative proteomics. Dcr-1 and Dcr-2 were both abundant in Loqs immunoprecipitates from cultured cells and flies (Supplementary Fig. 12), supporting a physical interaction between Dcr-2 and Loqs.

Among animals, endo-siRNA pathways have so far been restricted to *Caenorhabditis elegans*[24–27]. Our results extend the prevalence of

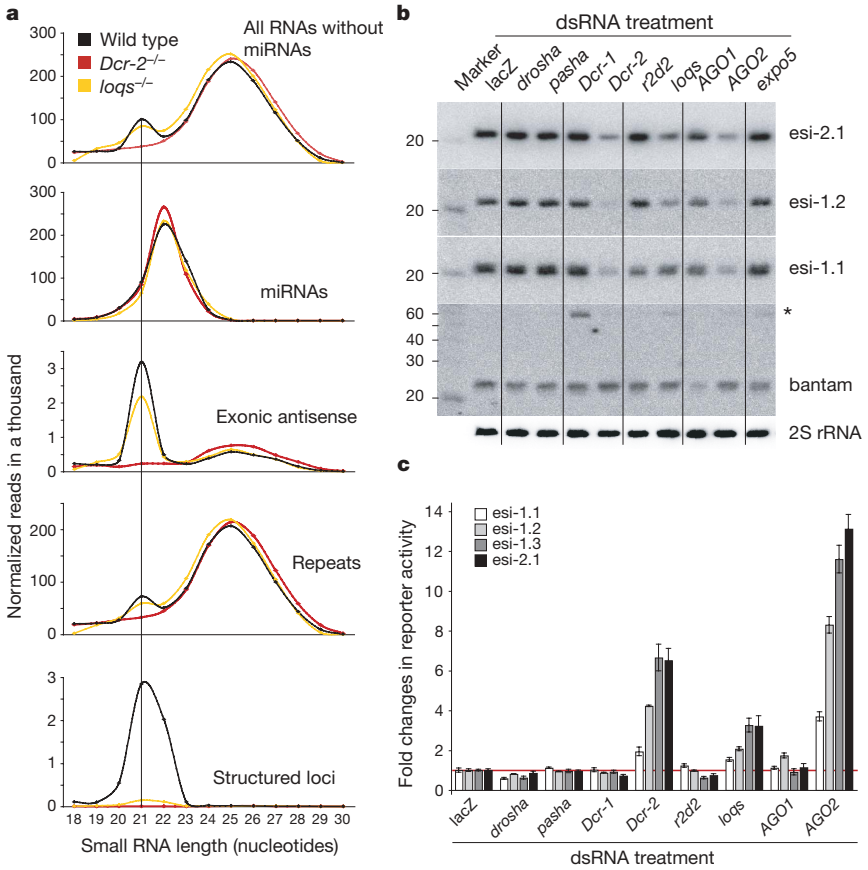


Figure 4 | Genetic requirements for siRNA biogenesis. **a**, Length distributions of small RNAs from total RNA libraries obtained from wild-type (black), *Dcr-2* mutant (red) and *loqs* mutant (yellow) ovaries. Excluding miRNAs and piRNAs (23–29 nucleotides), the population of 21-nucleotide siRNAs mapping to structured loci, genes and repeats is lost from *Dcr-2* mutants and those mapping to structured loci are strongly reduced in *loqs* mutants (all libraries are adjusted to the same total small RNA count). **b**, Northern blots showing levels of three siRNAs encoded from structured loci *esi-1* and *esi-2* (esi-2.1, esi-1.1 and esi-1.2) in S2 cells treated with dsRNA against the genes indicated. As controls, northern blots of *bantam* (pre-miRNA indicated by the asterisk) and 2S rRNA are shown below. **c**, *Renilla* luciferase reporter assays are shown for the siRNAs examined in **b** and an additional esi-1-derived species (esi-1.3) in S2 cells treated with dsRNAs against the indicated genes (error bars indicate standard deviation; $n = 3$).

©2008 Nature Publishing Group

such systems to *Drosophila* and parallel recent discoveries of an endo-siRNA pathway in mouse oocytes[28,29]. These systems have many common features but also key differences. In both, siRNAs collaborate with piRNAs to repress transposons. Also, mouse and *Drosophila* both generate endo-siRNAs from structured loci. In mouse, dsRNAs can form by pairing of sense protein-coding transcripts with antisense transcripts from pseudogenes. Whether or not transcripts from unlinked sites lead to siRNA production in *Drosophila* is unclear. However, transposon sense transcripts may hybridize to antisense sequences transcribed from piRNA clusters to form endo-siRNA precursors. In flies, a much larger number of genic loci enter the pathway as compared to mice because convergent transcription of neighbouring genes frequently creates overlapping transcripts. Overall, annotation of the *Drosophila* genome indicates that a significant proportion is transcribed in both orientations, providing widespread potential for dsRNA formation. This property is shared by many other annotated genomes, raising the possibility that the RNAi pathway has broad impacts on gene regulation. Viewed in combination, our studies suggest an evolutionarily widespread adoption of dsRNAs as regulatory molecules, a property previously ascribed only to miRNAs.

METHODS SUMMARY

The fly stocks used were *Dcr-2*^{L811Fsx} (ref. 6), *AGO2*^{414} (ref. 30), *loqs*^{f00791} (ref. 1) and *r2d2*^{1} (ref. 14). Recombineering was used to insert a Flag–haemagglutinin (HA) tag at the amino terminus of the *AGO2* coding sequence in the context of the genomic *AGO2* locus including flanking regulatory regions (for details, see Methods). Polyclonal anti-AGO1 antibody was obtained from Abcam (lot number 113754). Small RNAs for library production were isolated from ovarian total RNA or from Argonaute immunoprecipitates. Libraries were produced as described[12] and sequenced using the Illumina platform (protocol available on request). A description of the bioinformatics methods can be found online. For quantitative real-time PCR (qRT–PCR) analyses, we used total RNA preparations and random hexamer primers. Details and all primer sequences are given in the Supplementary Information. S2 cell knockdown treatments were for eight days with two sequential dsRNA soakings. For the reporter experiments, inducible expression plasmids for *Renilla* and firefly were transfected into S2 cells together with dsRNA for the desired knockdown target. Renilla constructs contained two target sites for endogenous siRNAs, whereas the firefly construct was used for normalization. For details on plasmids, dsRNAs and target sites, see Supplementary Information.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 27 January; accepted 18 April 2008.
Published online 7 May 2008.

1. Forstemann, K. *et al.* Normal microRNA maturation and germ-line stem cell maintenance requires Loquacious, a double-stranded RNA-binding domain protein. *PLoS Biol.* **3,** e236 (2005).
2. Lee, Y. *et al.* The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425,** 415–419 (2003).
3. Denli, A. M., Tops, B. B., Plasterk, R. H., Ketting, R. F. & Hannon, G. J. Processing of primary microRNAs by the Microprocessor complex. *Nature* **432,** 231–235 (2004).
4. Saito, K., Ishizuka, A., Siomi, H. & Siomi, M. C. Processing of pre-microRNAs by the Dicer-1–Loquacious complex in *Drosophila* cells. *PLoS Biol.* **3,** e235 (2005).
5. Jiang, F. *et al.* Dicer-1 and R3D1-L catalyze microRNA maturation in *Drosophila*. *Genes Dev.* **19,** 1674–1679 (2005).
6. Lee, Y. S. *et al.* Distinct roles for *Drosophila* Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. *Cell* **117,** 69–81 (2004).
7. Bernstein, E., Caudy, A. A., Hammond, S. M. & Hannon, G. J. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409,** 363–366 (2001).
8. Eulalio, A., Huntzinger, E. & Izaurralde, E. Getting to the root of miRNA-mediated gene silencing. *Cell* **132,** 9–14 (2008).
9. Bushati, N. & Cohen, S. M. microRNA functions. *Annu. Rev. Cell Dev. Biol.* **23,** 175–205 (2007).
10. Aravin, A. A., Hannon, G. J. & Brennecke, J. The Piwi–piRNA pathway provides an adaptive defense in the transposon arms race. *Science* **318,** 761–764 (2007).
11. Klattenhoff, C. & Theurkauf, W. Biogenesis and germline functions of piRNAs. *Development* **135,** 3–9 (2008).
12. Brennecke, J. *et al.* Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **128,** 1089–1103 (2007).
13. Gunawardane, L. S. *et al.* A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in *Drosophila*. *Science* **315,** 1587–1590 (2007).
14. Liu, Q. *et al.* R2D2, a bridge between the initiation and effector steps of the *Drosophila* RNAi pathway. *Science* **301,** 1921–1925 (2003).
15. Ruby, J. G. *et al.* Evolution, biogenesis, expression, and target predictions of a substantially expanded set of *Drosophila* microRNAs. *Genome Res.* **17,** 1850–1864 (2007).
16. Galiana-Arnoux, D., Dostert, C., Schneemann, A., Hoffmann, J. A. & Imler, J. L. Essential function *in vivo* for Dicer-2 in host defense against RNA viruses in *Drosophila*. *Nature Immunol.* **7,** 590–597 (2006).
17. Wang, X. H. *et al.* RNA interference directs innate immunity against viruses in adult *Drosophila*. *Science* **312,** 452–454 (2006).
18. Saito, K. *et al.* Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the *Drosophila* genome. *Genes Dev.* **20,** 2214–2222 (2006).
19. Vagin, V. V. *et al.* A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313,** 320–324 (2006).
20. Rehwinkel, J. *et al.* Genome-wide analysis of mRNAs regulated by Drosha and Argonaute proteins in *Drosophila melanogaster*. *Mol. Cell. Biol.* **26,** 2965–2975 (2006).
21. Allen, E., Xie, Z., Gustafson, A. M. & Carrington, J. C. microRNA-directed phasing during *trans*-acting siRNA biogenesis in plants. *Cell* **121,** 207–221 (2005).
22. Tomari, Y., Du, T. & Zamore, P. D. Sorting of *Drosophila* small silencing RNAs. *Cell* **130,** 299–308 (2007).
23. Forstemann, K., Horwich, M. D., Wee, L., Tomari, Y. & Zamore, P. D. *Drosophila* microRNAs are sorted into functionally distinct argonaute complexes after production by Dicer-1. *Cell* **130,** 287–297 (2007).
24. Ruby, J. G. *et al.* Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* **127,** 1193–1207 (2006).
25. Sijen, T., Steiner, F. A., Thijssen, K. L. & Plasterk, R. H. Secondary siRNAs result from unprimed RNA synthesis and form a distinct class. *Science* **315,** 244–247 (2007).
26. Pak, J. & Fire, A. Distinct populations of primary and secondary effectors during RNAi in *C. elegans*. *Science* **315,** 241–244 (2007).
27. Yigit, E. *et al.* Analysis of the *C. elegans* Argonaute family reveals that distinct Argonautes act sequentially during RNAi. *Cell* **127,** 747–757 (2006).
28. Tam, O. H. *et al.* Pseudogene-derived siRNAs regulate gene expression in mouse oocytes. *Nature* advance online publication, doi:10.1038/nature06904 (10 April 2008).
29. Watanabe, T. *et al.* Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* advance online publication, doi:10.1038/nature06908 (10 April 2008).
30. Okamura, K., Ishizuka, A., Siomi, H. & Siomi, M. C. Distinct roles for Argonaute proteins in small RNA-directed RNA cleavage pathways. *Genes Dev.* **18,** 1655–1666 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. Carthew, H. Siomi, P. Zamore and D. Smith for reagents. We are grateful to M. Rooks, E. Hodges and D. McCombie for help with deep sequencing. B.C. was supported by the German Academic Exchange Service. C.D.M. is a Beckman fellow of the Watson School of Biological Sciences and is supported by an NSF Graduate Research Fellowship. R.Z. is a Special Fellow of the Leukemia and Lymphoma Society. M.D. is an Engelhorn fellow of the Watson School of Biological Sciences. J.B. is supported by the Ernst Schering foundation. A.S. is supported by an HFSP fellowship. This work was supported in part from grants from the NIH to G.J.H. and N.P. and a gift from K. W. Davis (G.J.H.).

Author Information Small RNA sequences were deposited in the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/) under accession number GSE11086. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to G.J.H. (hannon@cshl.edu) or J.B. (brenneck@cshl.edu).

©2008 Nature Publishing Group

METHODS

Fly stocks. A 3× Flag–HA tag was inserted at the N terminus of the Flybase RB transcript of *AGO2* into BAC RP98-21A13 by means of bacterial red/ET recombination (Gene Bridges GmbH). A 13.9-kb AvrII/XhoI fragment of the modified BAC encompassing the *AGO2* locus including parts of flanking genes (chromosome 3L coordinates: 15,544,405–15,558,309) was cloned into pCasper4 (XbaI/XhoI). Transgenic flies were generated at Bestgene Inc. Expression of tagged AGO2 in embryos, ovaries and whole flies was verified in multiple lines by western blotting using a monoclonal anti-HA-peroxidase antibody (1:500; catalogue number 12013819001, Roche). *Dcr-2*^{L811Fsx} flies were a gift of R. Carthew[6], *AGO2*^{414} flies were a gift of H. Siomi[30], *loqs*^{f00791} flies were a gift of P. Zamore[1] and *r2d2*^{1} flies were a gift of D. Smith[14]. For wild-type fly stocks, stock number 2057 from Bloomington (Celera sequencing strain) and OregonR flies were used.

Small RNA libraries. Twenty-six 10-cm plates with 50–70% confluent Schneider cells were transfected with pCasper_Flag-HA-AGO2 using calcium phosphate, harvested after 36 h and lysed in buffer A (20 mM HEPES, pH 7.0, 150 mM NaCl, 2.5 mM $MgCl_2$, 0.3% Triton, 30% glycerol) supplemented with 1 mM PMSF and protease inhibitors (Complete, Roche). Cleared extract was split and incubated with rabbit polyclonal AGO1 antibody (1:20; lot number 113754, Abcam) or mouse anti-Flag M2-agarose (1:25, Sigma) for 4 h at 4 °C. AGO1 antibodies were isolated by adding protein G beads (1:10; Roche) for 1 h. Beads were washed six times each for 10 min in buffer B (30 mM HEPES, pH 7.4, 800 mM NaCl, 2 mM $MgCl_2$, 0.1% NP-40), which contained equal supplements as buffer A. The immunoprecipitation was analysed by western blotting using anti-AGO1 (1:2,000; Abcam) and anti-HA-peroxidase (1:500; catalogue number 12013819001, Roche). AGO1- and AGO2-associated RNAs were isolated with phenol/chloroform and ethanol precipitated. For the ovarian AGO2 IP library, ~500 mg ovaries from transgenic Flag-HA-AGO2 flies were dissected and lysed mechanically in buffer A. AGO2 complexes and associated RNAs were purified as above.

AGO1- and AGO2-bound small RNAs as well as small RNAs from total RNA were cloned as described[12] (the detailed protocol is available on request). The following small RNA libraries from total RNA were prepared for this study: 18–28 nucleotides from ovaries of the Celera sequenced strain (Bloomington, number 2057); 18–28 nucleotides from ovaries of *Dcr-2*^{L811Fsx} homozygous flies; 18–28 nucleotides from *loqs*^{f00791} homozygous flies; and 18–24 nucleotides from the testis of OregonR flies.

Libraries were sequenced in house using the Illumina platform. Published libraries used in this study were a 16–26-nucleotide S2 cell total RNA library[15] and libraries from ovarian Piwi/Aub/AGO3 immunoprecipitates[12].

Bioinformatic analysis of small RNA libraries. Small RNA sequences were matched to the *Drosophila* release 5 genome and to genomes of *Drosophila* C virus, FHV and cricket paralysis virus. Only reads matching the fly genome 100% and viral genomes with up to three mismatches were used for further analysis. For annotations, we used Flybase for protein-coding genes, UCSC for non-coding RNAs and transposons/repeats (http://genome.ucsc.edu/) and the most recent miRNA catalogue[15,31].

siRNA clusters were extracted by mapping all 20–22-nucleotide-long RNAs from the AGO2-IP libraries to the genome (only uniquely mapping RNAs were used) and retaining 200-nucleotide windows that contained at least three distinct small RNAs. Windows separated by a maximum of 200 nucleotides were fused and those with more than 40 unique reads were sorted after the density of siRNAs per base pair.

For the transposon analysis, 20–22-nucleotide AGO2-bound RNAs from ovaries and S2 cells were mapped onto the Repbase collection of transposons[32] with up to three mismatches to construct heatmaps indicating cloning frequency and strand bias of siRNAs. For the latter analysis, only siRNAs unambiguously mapped to one strand were considered.

Cleavage site mapping for endo-siRNA targets. Wild-type testes were dissected on ice into 1× PBS. Total RNA was isolated using Trizol (Invitrogen) according to the manufacturer's protocol. Total RNA (5 μg) was used as starting material. Ligation of an RNA adaptor, reverse transcription using the GeneRacer oligo(dT) primer and 5′ RACE–PCR were performed according to the manufacturer's instructions (GeneRacer kit, Invitrogen). 5′ RACE–PCR was carried out using the GeneRacer 5′ primer (5′-CGACTGGAGCACGAGGACACTGA-3′) and a *mus308* gene-specific reverse primer (5′-TGCTTTGCAGAGTCGAA-GCTGATTG-3′), and followed by one round of nested PCR using the GeneRacer 5′ nested primer (5′-GGACACTGACATGGACTGAAGGAGTA-3′) and a nested primer specific to *mus308* (5′-CCGCTAGCTCTACCAAACTGGTGAT-3′). PCR products were gel purified and cloned into pCR4Blunt-TOPO (Invitrogen). Twenty-two clones were sequenced with T7 (5′-GTAATACGA-CTCACTATAGGGC-3′) and T3 (5′-AATTAACCCTCACTAAAGGG-3′) primers, and subjected to further analysis.

dsRNA treatment of Schneider cells. Approximately 3×10^6 S2-NP cells were soaked in 1.5 ml serum-free Schneider's medium containing 10 μg dsRNAs in 6-well plates, and 3 ml serum-containing medium was added 45 min later. After 4 days of initial dsRNA treatment, cells were treated with a second round of dsRNAs using the same procedure, and were harvested another 4 days later. Total RNA was extracted with Trizol (Invitrogen). Sequences of the primers for generating dsRNAs are listed below.

siRNA reporter constructs. A SalI/BglII fragment from pGL3-Basic (Promega) was cloned to pRmHa-3 using SalI/BamHI (pMT-Firefly-long). The coding region of the *Renilla* luciferase gene was amplified by PCR and cloned into pRmHa-3 using BamHI/EcoRI sites (pMT-Renilla). A pair of oligonucleotides containing two perfect binding sites for si1_1, si1_2 or si2 were annealed and cloned into pMT-Renilla (BamHI/SalI) to generate sensor constructs (Supplementary Fig. 13).

Transfection was performed in a 384-well plate format. For each well, ~100 ng plasmid DNA (5 ng pMT-Renilla, 20 ng pMT-Renilla-si1_1, 50 ng pMT-si1_2 or 100 ng pMT-si2, 5 ng pMT-Firefly-long, and corresponding amounts of pRmHa-3 serving as carrier DNA) and ~80 ng dsRNA were mixed with 0.8 μl enhancer in 15 μl EC (Qiagen) and the mixture was incubated at room temperature (23 °C) for 5 min. After this, 0.35 μl of Effectene reagent was added and the mixture was immediately dispensed into each well containing dsRNA. After incubation at room temperature for 10 min, 40 μl S2-NP cells (10^6 cells ml^{-1}) were dispensed into the well. Cells were induced with 200 μM $CuSO_4$ 132 h post transfection, and luciferase assays were performed 36 h later using DualGlo reagents (Promega). For each well, the reporter activity was calculated as the ratio of *Renilla* luciferase to firefly luciferase. Each data point was normalized against the data points where dsRNA against LacZ was transfected. Presented are average results with standard deviation ($n = 3$).

Northern blotting. Total RNA was isolated using Trizol (Invitrogen). RNA (30 μg) was separated on a 15% denaturing polyacrylamide gel and transferred onto a Hybond-N+ membrane (Amersham Biosciences) in 1× TBE (Tris-Borate-EDTA) buffer. The RNA was crosslinked using ultraviolet light (Stratalinker) to the membrane and pre-hybridized in ULTRAhyb buffer (Ambion) for 1 h. DNA probes complementary to the indicated endo-siRNAs, bantam and 2S rRNA were 5′ radiolabelled and added to the hybridization buffer (hybridization overnight at 37 °C). Membranes were washed 4–6 times in 1× SSC with 0.1% SDS at 37 °C and exposed to PhosphoImager screens. Probes were stripped by boiling the membrane twice in 0.2× SSC containing 0.1% SDS in a microwave.

Quantitative real-time PCR. Ovaries and testis from homozygous or heterozygous flies were dissected on ice into 1× PBS. Total RNA of dissected tissues or S2 cells was extracted using Trizol (Invitrogen). RNA was treated with DNase I Amplification Grade (Invitrogen) according to the manufacturer's instructions. Complementary DNA was prepared by reverse transcription using SuperScript III Reverse Transcriptase (Invitrogen) and random hexamer primer. qRT–PCR was carried out using SYBR GREEN PCR Master Mix (Applied Biosystems) and a Chromo4 Real-Time PCR Detector (BioRad). C_t values were calculated within the log-linear phase of the amplification curve using the Opticon Monitor 3.1.32 software (BioRad). Quantification was normalized to the mRNA coding for the endogenous ribosomal protein rp49, and relative expression levels were calculated using the following equation: $A = 1.8^{[Ct(ref) - Ct(ref\text{-}control)] - [Ct(sample) - Ct(sample\text{-}control)]}$. Transposon analysis was carried out with four biological replicates (individually shown; error bars indicate technical replicates); siRNA target analysis was carried out with three biological replicates. Oligonucleotide primers used in this study are listed below.

DNA oligonucleotides. PCR primers for the generation of dsRNAs were as follows: T7-Dicer-1-F-14 TAATACGACTCACTATAGGGTGCGACAACAA-TCTGC; T7-Dicer-1-R-565 TAATACGACTCACTATAGGGTCAGTTGCTG-CAGCTCAC; T7-Dicer-2-F-5 TAATACGACTCACTATAGGGAAGATGTGG-AAATCAAGCC; T7-Dicer-2-R-555 TAATACGACTCACTATAGGGCCACG-TTCGTAATTTC; T7-Drosha-F-3356 TAATACGACTCACTATAGGGTGAA-TCAGGACTGGAACG; T7-Drosha-R-3910 TAATACGACTCACTATAGGG-AGCCATCGCTATCACTGC; T7-Exportin5-F-55 TAATACGACTCACTATA-GGGATCTAGTCATGAACCCG; T7-Exportin5-R-623 TAATACGACTCACT-ATAGGGAACGCAGTCACATGCTGC; T7-AGO1-F1225 TAATACGACTC-ACTATAGGGAACGGACAGACCGTAGAG; T7-AGO1-R1858 TAATACGAC-TCACTATAGGGTGGCGTACTTACAGAAGC; T7-AGO2-F2211 TAATACG-ACTCACTATAGGGAGCCACATCGACGAACG; T7-AGO2-R2855 TAATAC-GACTCACTATAGGGCGAGGATCATCCTTGATC; T7-R2D2-PZ-F TAAT-ACGACTCACTATAGGGCATACACGGCTTGATGAAGGATTC; T7-R2D2-PZ-R TAATACGACTCACTATAGGGTTGCTTGTGCTCGCTACTTGC; T7-Pasha-F452 TAATACGACTCACTATAGGGACTTTGAAGTCCTACCCG; T7-Pasha-R1177 TAATACGACTCACTATAGGGCTCCTTGAACTCATAGG; T7-Loqs-F-1 TAATACGACTCACTATAGGGATGGACCAGGAGAATTTCC; T7-

©2008 Nature Publishing Group

Loqs-R-540 TAATACGACTCACTATAGGGAAGGGCGTATCCTTGTC; T7-LacZ-F TAATACGACTCACTATAGGGCATTATCCGAACCATCC; T7-LacZ-R TAATACGACTCACTATAGGGCAGAACTGGCGATCGTTCG. endo-siRNA sensor oligonucleotides were as follows: BamHI-esi-1_2-S2-2P-F, GATCCCAACAGTTTATTTACTTGGAGGCAACATAATCAAATGAACTGAGGGTTACTTGGAGGCAACATAATCAG; SalI-esi-1_2-S2-2P-R, TCGACTGATTATGTTGCCTCCAAGTAACCCTCAGTTCATTTGATTATGTTGCCTCCAAGTAAATAAACTGTTGG; BamHI-esi-2_1-S2-2P-F, GATCCCAACAGTTTATTGGAGCGAACTTGTTGGAGTCAAAATGAACTGAGGGTGGAGCGAACTTGTTGGAGTCAAG; SalI-esi-2_1-S2-2P-R, TCGACTTGACTCCAACAAGTTCGCTCCACCCTCAGTTCATTTTGACTCCAACAAGTTCGCTCCAATAAACTGTTGG; BamHI-esi-1_3-S2-2P-F, GATCCCAACAGTTTATTCATTTGATCCATAGTTTCCCGAATGAACTGAGGGTCATTTGATCCATAGTTTCCCGG; SalI-esi-1_3-S2-2P-R, TCGACCGGGAAACTATGGATCAAATGACCCTCAGTTCATTCGGGAAACTATGGATCAAATGAATAAACTGTTGG; BamHI-esi-1_1_ -S2-2P-F, GATCCCAACAGTTTATTGCCAAGGTACGTGGTCGACCGAAATGAACTGAGGGTGCCAAGGTACGTGGTCGACCGAG; SalI-esi-1_1_BC36-S2-2P-R, TCGACTCGGTCGACCACGTACCTTGGCACCCTCAGTTCATTTCGGTCGACCACGTACCTTGGCAATAAACTGTTGG; BamHI-mus308_target-S2-2P-F, GATCCCAACAGTTTATTGGGCGAGCTTGTTGGAGTCAGAATGAACTGAGGGTGGGCGAGCTTGTTGGAGTCAGG; SalI-mus308_target-S2-2P-R, TCGACCTGACTCCAACAAGCTCGCCCACCCTCAGTTCATTCTGACTCCAACAAGCTCGCCCAATAAACTGTTGG.

Northern probes were as follows: esi-2.1, GGAGCGAACTTGTTGGAGTCAA; esi-1.1, GCCAAGGTACGTGGTCGACCGA; esi-1.2, CATTTGATCCATAGTTTCCCG; miR-bantam, AATCAGCTTTCAAAATGATCTCA; 2S rRNA, TACAACCCTCAACCATATGTAGTCCAAGCA.

Quantitative real-time PCR. For transposon analysis: rp49_F, ATGACCATCCGCCCAGCATAC; rp49_R, CTGCATGAGCAGGACCTCCAG; GAPDH2_F, TGATGAAATTAAGGCCAAGGTTCAGGA; GAPDH2_R, TCGTTGTCGTACCAAGAGATCAGCTTC; actin5c_F, AAGTTGCTGCTCTGGTTGTCG; actin5c_R, GCCACACGCAGCTCATTGTAG; BEL1_F, ATTATACAAACGCCCAATTGCCAAAA; BEL1_R, TCCGATGAAGCTGCAGACAAATAAGA; BLOOD_F, AGACGTTCATTACAGATCAAGGTACGGA; BLOOD_R, AGTTCGTATGGGCAATAGTCATGGACT; DM412_F, AAAGTACGGTCCAATGAAGACG; DM412_R, GTGGTGATGAGCTGTTGATGTT; F-element_F, TTGTTGAACAGCATACCACTCC; F-element_R, CCAGAGTTGATGAGCCAGTGTA; gypsy6_F, GACAAGGGCATAACCGATACTGTGGA; gypsy6_R, AATGATTCTGTTCCGGACTTCCGTCT; MDG1_F, AACAGAAACGCCAGCAACAGC; MDG1_R, CGTTCCCATGTCCGTTGTGAT; ROO_F, CGTCTGCAATGTACTGGCTCT; ROO_R, CGGCACTCCACTAACTTCTCC; stalker4_F, TTTGGAAGATTACCAAGGCAGTTCGC; stalker4_R, GGATCTAACTTATGACCCGATTCGTTCC; ZAM_F, ACTTGACCTGGATACACTCACAAC; ZAM_R, GAGTATTACGGCGACTAGGGATAC; FHV_F, CCCTGGAGTCGCTTACTTGAGTGCT; FHV_R, ATGGAAGCGTACCTGAAGGAGGACA; DOC_F, TACCTTAAACAAACAAACATGCCCACC; DOC_R, TTTGTATGGGTGGTCAGCTTTTCGT; DM297_F, GCCAGTACACACGAACGAAATA; DM297_R, AATTGAATTTTGGCAATTTTGG; Tabor_F, GAGCAAGAATTATGCTCGAAGAA; Tabor_R, AATTTATGTCCGGTTTCGTTTTT.

For endo-siRNA target analysis: *mus308*_F, AAGGATTAGCGCCAAGCTGGAGGAT; *mus308*_R, ACCACGACCACTGCCACAGAGATTC; CG9203_F, AGCTGGCAGAAAAACCATGACCAGT; CG9203_R, CAATTCTTTTGGCGTAGCTTGAGCA.

For S2 cell knockdown analysis: S2-Dcr-1_F, ACGCCTTCCATCTCCCAGTTTTACC; S2-Dcr-1_R, GCCACCCTGCTTATTCTGACTGCTC; S2-Dcr-2_F, AAACGAGAGATTCGTGCCCAAAACA; S2-Dcr-2_R, CTGTCCTTGCTCTTATCGGCCTTGT; S2-Drosha_F, AGATGCCAGAGAACTTCACCATCCA; S2-Drosha_R, GAAAGAAGTGAAAAGCTGGGCAGGA; S2-Pasha_F, TGTCAAGGACAAGATAACGGGCAACA; S2-Pasha_R, GTTGGGAGATGGCTCCGTCGTCT; S2-AGO1_F, ACTACCACGTTCTGTGGGACGACAA; S2-AGO1_R, GAATCGTGCTCCTTCTCCACCAGAT; S2-AGO2_F, AACCCTCAAAAGTAAATCATGGGAAA; S2-AGO2_R, ATTTTTGCTGTTGGCCTCCTTG; S2-Loqs_F, GTGTGTGCGTCTGGATTTTGCTGTA; S2-Loqs_R, GTTTTCGGGAGGATTCGGTGTGTAT; S2-R2D2_F, GCGAAGACGGAGGGTACGTCTGTAA; S2-R2D2_R, AGTCGAATCCTTCATCAAGCCGTGT.

31. Stark, A. *et al.* Systematic discovery and characterization of fly microRNAs using 12 *Drosophila* genomes. *Genome Res.* **17,** 1865–1879 (2007).
32. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110,** 462–467 (2005).

©2008 Nature Publishing Group

doi:10.1038/nature07672

LETTERS

Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals

Mitchell Guttman[1,2], Ido Amit[1], Manuel Garber[1], Courtney French[1], Michael F. Lin[1], David Feldser[3], Maite Huarte[1,6], Or Zuk[1], Bryce W. Carey[2,8], John P. Cassady[2,8], Moran N. Cabili[7], Rudolf Jaenisch[2,8], Tarjei S. Mikkelsen[1,4], Tyler Jacks[2,3], Nir Hacohen[1,9], Bradley E. Bernstein[1,10,11], Manolis Kellis[1,5], Aviv Regev[1,2], John L. Rinn[1,6,11]* & Eric S. Lander[1,2,7,8]*

There is growing recognition that mammalian cells produce many thousands of large intergenic transcripts[1–4]. However, the functional significance of these transcripts has been particularly controversial. Although there are some well-characterized examples, most (>95%) show little evidence of evolutionary conservation and have been suggested to represent transcriptional noise[5,6]. Here we report a new approach to identifying large non-coding RNAs using chromatin-state maps to discover discrete transcriptional units intervening known protein-coding loci. Our approach identified ~1,600 large multi-exonic RNAs across four mouse cell types. In sharp contrast to previous collections, these large intervening non-coding RNAs (lincRNAs) show strong purifying selection in their genomic loci, exonic sequences and promoter regions, with greater than 95% showing clear evolutionary conservation. We also developed a functional genomics approach that assigns putative functions to each lincRNA, demonstrating a diverse range of roles for lincRNAs in processes from embryonic stem cell pluripotency to cell proliferation. We obtained independent functional validation for the predictions for over 100 lincRNAs, using cell-based assays. In particular, we demonstrate that specific lincRNAs are transcriptionally regulated by key transcription factors in these processes such as p53, NFκB, Sox2, Oct4 (also known as Pou5f1) and Nanog. Together, these results define a unique collection of functional lincRNAs that are highly conserved and implicated in diverse biological processes.

There are at present only about a dozen well-characterized lincRNAs in mammals, with transcript sizes ranging from 2.3 to 17.2 kilobases (kb)[7,8]. These lincRNAs have distinctive biological roles through diverse molecular mechanisms, including functioning in X-chromosome inactivation (Xist, Tsix)[8,9], imprinting (H19, Air)[7,10], *trans*-acting gene regulation (HOTAIR)[11] and regulation of nuclear import (Nron)[12]. Importantly, these well-characterized lincRNAs show clear evolutionary conservation confirming that they are functional.

Genomic projects over the past decade have used shotgun sequencing and microarray hybridization[1–4] to obtain evidence for many thousands of additional non-coding transcripts in mammals. Although the number of transcripts has grown, so too have the doubts as to whether most are biologically functional[5,6,13]. The main concern was raised by the observation that most of the intergenic transcripts show little to no evolutionary conservation[5,13]. Strictly speaking, the absence of evolutionary conservation cannot prove the absence of function. But, the markedly low rate of conservation seen in the current catalogues of large non-coding transcripts (<5% of cases) is unprecedented and would require that each mammalian clade evolves its own distinct repertoire of non-coding transcripts. Instead, the data suggest that the current catalogues may consist largely of transcriptional noise, with a minority of bona fide functional lincRNAs hidden amid this background. Thus, to expand our understanding of functional lincRNAs, we are faced with two important challenges: (1) identifying lincRNAs that are most likely to be functional; and (2) inferring putative functions for these lincRNAs that can be tested in hypothesis-driven experiments.

To address the first challenge, we took an entirely different approach to discovering functional lincRNAs on the basis of exploiting chromatin structure. We recently developed an efficient method[14] to create genome-wide chromatin-state maps, using chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-Seq). We observed that genes actively transcribed by RNA polymerase II (Pol II) are marked by trimethylation of lysine 4 of histone H3 (H3K4me3) at their promoter and trimethylation of lysine 36 of histone H3 (H3K36me3) along the length of the transcribed region[14]. We will refer to this distinctive structure as a 'K4–K36 domain'. We proposed that, by identifying K4–K36 structures that reside outside known protein-coding gene loci, we could systematically discover lincRNAs.

To test this hypothesis, we searched for K4–K36 domains in genome-wide chromatin-state maps of four mouse cell types: mouse embryonic stem cells (ESCs), mouse embryonic fibroblasts (MEF), mouse lung fibroblasts (MLF) and neural precursor cells (NPC). We identified K4–K36 domains of at least 5 kb in size that did not overlap regions containing protein-coding genes as well as known microRNAs[15] and endogenous short interfering RNAs (siRNAs)[16,17]. This analysis revealed 1,675 K4–K36 (1,250 conservatively defined) domains that do not overlap with known annotations; examples are shown in Fig. 1 (Supplementary Table 1).

Having identified K4–K36 loci with no previous annotation, we addressed: (1) whether these gene loci produce large multi-exonic RNA molecules; (2) whether the RNA molecules encode proteins or are non-coding transcripts; and (3) whether the RNA molecules, their promoters and their chromatin structure show conservation across mammals.

To test whether the intergenic K4–K36 domains produce RNA transcripts, we selected a random sample of 350 regions and designed

[1]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [2]Department of Biology, [3]The Koch Institute for Integrative Cancer Research, [4]Division of Health Sciences and Technology, and [5]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [6]Department of Pathology, Beth Israel Deaconess Medical Center, Boston, Massachusetts 02215, USA. [7]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02114, USA. [8]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA. [9]Center for Immunology and Inflammatory Diseases, [10]Molecular Pathology Unit and Center for Cancer Research, Massachusetts General Hospital, Charlestown, Massachusetts 02129, USA. [11]Department of Pathology, Harvard Medical School, Boston, Massachusetts 02115, USA.
*These authors contributed equally to this work.

©2009 Macmillan Publishers Limited. All rights reserved

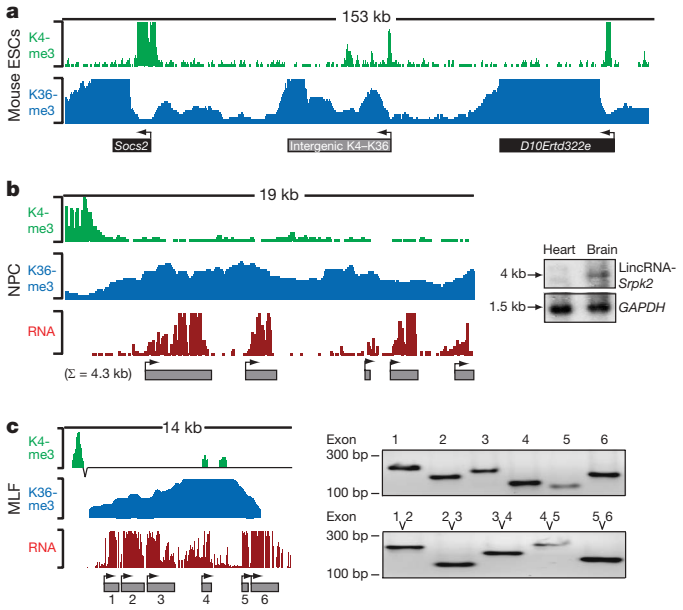


Figure 1 | Intergenic K4–K36 domains produce multi-exonic RNAs. **a**, Example of an intergenic K4–K36 domain and the K4–K36 domains of two flanking protein-coding genes. Each histone modification is plotted as the number of DNA fragments obtained by ChIP-Seq at each position. Black boxes indicate known protein-coding regions and grey boxes are intergenic K4–K36 domains. Arrowheads indicate the orientation of transcription. **b**, Intergenic K4–K36 domains were interrogated for presence of transcription by hybridizing RNA to DNA tiling arrays. The RNA hybridization intensity is plotted in red. RNA peaks were determined and are represented by grey boxes. The presence of a spliced transcript was validated by hybridization to a northern blot (right). **c**, Connectivity between the inferred exons was validated by PCR with reverse transcription (RT–PCR). Right top shows RT–PCR validation of each exon, right bottom shows RT–PCR across each consecutive exon.

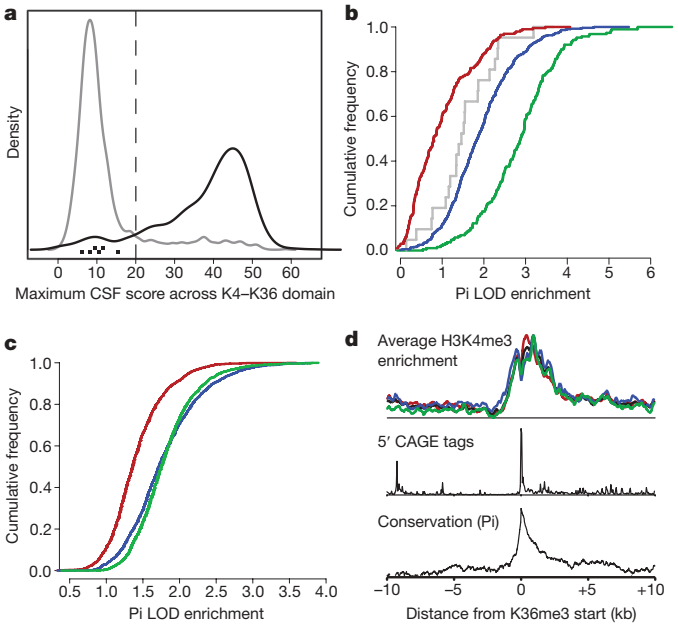


Figure 2 | lincRNA K4–K36 domains do not encode proteins and are conserved in their exons and promoters. **a**, Density plot of the maximum CSF score (Methods) across intergenic K4–K36 domains (grey) and known protein-coding genes (black). The maximum CSF scores for known lincRNAs are indicated as black points at the bottom. **b**, Cumulative distribution of sequence conservation across mammals for lincRNA exons (blue), protein-coding exons (green), introns (red) and known non-coding RNA exons (grey). **c**, Cumulative distribution of sequence conservation for lincRNA promoters (blue), random intergenic regions (red), and protein-coding promoters (green). LOD, logarithm of the odds ratio; Pi is the conservation metric (see Supplementary Methods). **d**, Enrichment of various promoter features plotted as the distance from the start of the K36me3 region averaged across all lincRNAs. Enrichment in each cell type of K4me3 domains across mouse ESCs (red), MEF (black), MLF (blue) and NPC (green) is shown (top panel). Enrichment of 5′ CAGE-tag density representing the 5′ end of RNA molecules (middle panel) and conservation scores in the K4me3 region are shown (bottom panel).

DNA microarrays containing oligonucleotides that tile across the regions (Methods) as well as various control regions. We hybridized poly(A)$^+$-selected RNA from each of the four cell types to the arrays. We developed an algorithm (Methods) to identify regions of significant hybridization and used it to define putative exons of transcripts detected at the loci. For ~70% of the intergenic loci with K4–K36 domains present in a cell type, we found clear evidence of RNA transcription in that cell type (Fig. 1 and Supplementary Tables 2 and 3). The proportion is similar to that for the protein-coding genes: ~72% of K4–K36 domains corresponding to known protein-coding genes show significant hybridization ($P < 0.05$, Wilcoxon test). In addition, we confirmed the presence of 93 out of 107 (87%) randomly selected exons, representing at least one exon from 19 out of 20 K4–K36 domains tested. We also confirmed the connectivity of consecutive exons in 52 out of 67 (78%) cases, including one from each of 16 K4–K36 domains tested (Fig. 1c and Supplementary Table 4). Furthermore, we validated the presence of discrete transcripts by hybridization to RNA northern blots in 15 of 17 tested loci (Fig. 1b, Supplementary Fig. 1, Supplementary Table 5 and Methods).

To determine whether the transcripts encode previously unknown protein-coding genes or non-coding RNAs, we used an established metric (the codon substitution frequency, CSF[18,19]) to assess characteristic evolutionary signatures of protein-coding domains. Analysing both the overall genomic locus (Fig. 2a and Supplementary Table 6) and the exons themselves (Supplementary Fig. 2 and Methods), we found that >90% of the intergenic K4–K36 domains fall well below the threshold of known protein-coding genes and resemble known lincRNAs (Fig. 2a). The result indicates that most of the loci do not encode protein-coding genes. Consistent with this, fewer than 2.5% of the exons show any similarity to known protein-coding genes, using the BLASTX program (Methods).

To assess the extent of nucleotide sequence conservation in the RNA transcripts, we used a method that explicitly models the underlying substitution rate (Methods) across 21 mammalian genomes (M.G. and X. Xie, submitted, Methods). We found that the lincRNA exons show clear sequence conservation when compared to other intergenic regions (Fig. 2b, Supplementary Fig. 3 and Supplementary Tables 2 and 7). Furthermore, the transcribed regions are highly enriched for conserved elements (defined by the PhastCons program[20]) compared to other intergenic regions ($P < 0.0001$, permutation test). The conservation level is similar to that seen for known lincRNAs, although it is lower than that seen for protein-coding exons, probably reflecting a lower degree of constraint on RNA structures than on amino-acid codons. The presence of strong purifying selection provides firm evidence that most K4–K36-defined lincRNAs must be biologically functional in mammals.

We used the same method to assess the conservation of the lincRNAs promoters (marked by the K4 domain). The lincRNA promoter regions show strong conservation, being essentially indistinguishable from known protein-coding genes (Fig. 2c and Supplementary Table 8). Furthermore, the lincRNA promoters show a notable enrichment of 'CAGE tags' (obtained by capturing the 7-methylguanosine cap at the 5′-end of Pol II transcripts) that mark transcriptional start sites[21] (Fig. 2d). Most of the lincRNA promoters regions (85%) contain a significant cluster of CAGE tags, with the density tightly localized around the promoter. In addition, the

©2009 Macmillan Publishers Limited. All rights reserved

lincRNA promoters show strong enrichment for binding of RNA Pol II in mouse ESCs ($P < 2 \times 10^{-16}$; Supplementary Fig. 4).

To investigate whether the K4–K36 chromatin structures observed at the loci are conserved across species, we constructed chromatin-state maps in human lung fibroblasts and MLF. Notably, ~70% of the K4–K36 domains in human also had a K4–K36 domain in the orthologous region of the mouse genome (Supplementary Table 9). The proportion is similar to that seen for protein-coding genes (~80%).

Together, the results show that most of the K4–K36 domains encode multi-exonic, non-protein-coding transcripts and the loci show clear conservation of nucleotide sequence and chromatin structure. Moreover, transcription and processing of these lincRNAs appears to be similar to that for protein-coding genes—including Pol II transcription, 5′-capping and poly-adenylation.

Having identified a large set of conserved lincRNAs, the next important challenge is to develop a method to infer putative functions that can be tested experimentally. To this end, we began by creating an RNA expression compendium of both lincRNAs and protein-coding genes across a wide range of tissues. We hybridized poly-adenylated RNA from 16 mouse samples to a custom lincRNA array. The samples included the original four cell types (mouse ESCs, NPC, MEF and MLF), a time course of embryonic development (whole embryo, hindlimb and forelimb at embryonic days 9.5, 10.5 and 13.5), and four normal adult tissues (brain, lung, ovary and testis) (Supplementary Fig. 5 and Supplementary Table 10).

The expression data contains a wealth of information about the lincRNAs. As an example, we searched for lincRNAs with an expression pattern opposite to the known lincRNA *HOTAIR*. Notably, we found that the most highly anti-correlated lincRNA in the genome lies in the HOXC cluster, in the same euchromatic domain as *HOTAIR*; we call this lincRNA *Frigidair* (Fig. 3c). This suggests that *Frigidair* may repress *HOTAIR* or perhaps activate genes in the HOXD cluster.

To take a more systematic approach, we also analysed RNA expression data for protein-coding genes from published sources[14,22] and generated further data for the embryonic development time course. We clustered the lincRNA and protein-coding genes into sets with correlated expression patterns (Supplementary Fig. 6a). We used Gene Set Enrichment Analysis (GSEA) to construct a matrix of the association of each lincRNA with each of ~1,700 functional gene sets (Fig. 3a and Supplementary Tables 10 and 11)[23]. We next performed biclustering on the gene set matrix to identify sets of lincRNAs that are associated with distinct sets of functional categories[24]. This analysis revealed numerous sets of lincRNAs associated with distinct and diverse biological processes (Fig. 3a). These include cell proliferation, RNA binding complexes, immune surveillance, ESC pluripotency, neuronal processes, morphogenesis, gametogenesis and muscle development (Fig. 3b and Supplementary Fig. 7).

To assess the validity of the inferred functional associations, we examined the gene sets associated with HOTAIR. *HOTAIR* showed negative association with *HOXD* genes (false discovery rate (FDR) < 0.018) and positive association with 'Chang Serum Response' (FDR < 0.001), a known predictor of breast cancer metastasis[25]. Both results are consistent with the known properties of *HOTAIR*, including a role in breast cancer metastasis[11,26].

We then sought to obtain independent experimental validation of the inferred biological functions for many of the lincRNAs. We focused on three large clusters of lincRNAs associated with the p53-mediated DNA damage response in MEF, NFκB signalling in dendritic cells, and ESC pluripotency, on the basis of their expression pattern across tissues.

We exposed $p53^{+/+}$ and $p53^{-/-}$ MEF to a DNA damaging agent and profiled the resulting expression changes on our lincRNA microarray (Methods)[27]. We found 39 lincRNAs that were significantly induced in $p53^{+/+}$ but not in $p53^{-/-}$ cells (Methods, Supplementary Fig. 8 and Supplementary Table 12). Approximately half of these lincRNAs resided in the cluster associated with p53-mediated DNA damage response, confirming the validity of the functional inference ($P < 10^{-7}$). Notably, we found that the promoters of these 39 lincRNAs were significant enriched for the p53 *cis*-regulatory binding element (versus all lincRNA promoters, $P < 0.01$, Wilcoxon test; Supplementary Fig. 8 and Supplementary Table 13). This suggests that p53 directly binds and regulates the expression of at least some of these lincRNA genes.

We stimulated $CD11C^{+}$ bone-marrow-derived dendritic cells with a specific agonist of the Toll-like receptor Tlr4, which signals through NFκB. We found that 20 lincRNAs showed marked upregulation after Tlr4-stimulation (Supplementary Table 14). Consistent with the inferences described earlier, 80% of these induced lincRNAs resided in the cluster associated with NFκB signalling. The greatest change in expression was observed in a lincRNA that is located ~51 kb upstream of the protein-coding gene *COX2* (also known as *Ptgs2*), a critical inflammation mediator that is directly induced by NFκB on Tlr4 stimulation; we refer to this as lincRNA-*COX2*. We found that lincRNA-*COX2* is induced ~1,000-fold over the course of 12 h after Tlr4 stimulation (Fig. 3d). In contrast, stimulation of Tlr3, which signals through IRF3, led to only weak induction of lincRNA-*COX2* (Fig. 3d).

Using published data from mouse ESCs, we identified 118 lincRNAs in which the promoter loci were bound by the core transcription factors Oct4 and Nanog[28] (Supplementary Table 15). Of those represented on our expression array 72% resided in the cluster associated with pluripotency, again supporting the validity of the functional inference. We noticed that one of these lincRNAs, which is only expressed in ESCs, is located ~100 kb from the *Sox2* locus, which encodes another key transcription factor associated with pluripotency (Fig. 3e). We cloned the promoter of this locus (which we will refer to as lincRNA-*Sox2*) upstream of a luciferase reporter gene and transfected the construct into mouse cells transiently expressing *Oct4*, *Sox2*, or both, as well as several controls. We found that *Sox2* and *Oct4* were each sufficient to drive expression of this lincRNA promoter, and the expression of both *Oct4* and *Sox2* together caused synergistic increases in expression (Fig. 3f). To our knowledge, this is the first experimental validation of a lincRNA promoter being directly regulated by key transcription factors such as Sox2 and Oct4.

The ultimate proof-of-function will be to demonstrate that RNA-interference-mediated knockout of each lincRNAs has the predicted phenotypic consequences. Towards this end, we examined a recently published short hairpin RNA screen of (presumed) protein-coding genes to identify genes that regulate cell proliferation rates in mouse ESCs[29]. The screen involved genes and some unidentified transcripts that had been identified as expressed in ESCs and showing rapid decrease in expression after retinoic acid treatment. Of the top ten hits in the screen, one corresponded to a gene of unknown function. We discovered that this gene corresponds to one of our lincRNAs (located ~181 kb from *Enc1*) contained in both the 'cell cycle and cell proliferation' cluster (FDR < 0.001) and the 'ESC' cluster (FDR < 0.001; Supplementary Fig. 9 and Supplementary Table 16). This provides functional confirmation that this lincRNA has a direct role in cell proliferation in ESCs, consistent with the analysis above.

Our results address the two key issues in the study of lincRNAs. We show that chromatin structure can identify sets of lincRNAs that show a high degree of evolutionary conservation, indicating that they are biologically functional. (We do not exclude the possibility that lincRNAs identified by shotgun sequencing that fail to show conservation are nonetheless functional, but other evidence will be required to establish this point.) We also provide a functional genomics pipeline for inferring putative roles for lincRNAs. The approach suggested functional roles for 150 lincRNAs that we studied on microarrays, and the independent experiments provided support for the predicted pathways for ~85 lincRNAs. The pipeline thus provides a useful guide for hypothesis-driven functional studies.

A fundamental issue will now be to determine the biological functions and the mechanisms by which lincRNAs act. One clue may come from the distribution of lincRNAs across the genome. We

©2009 Macmillan Publishers Limited. All rights reserved

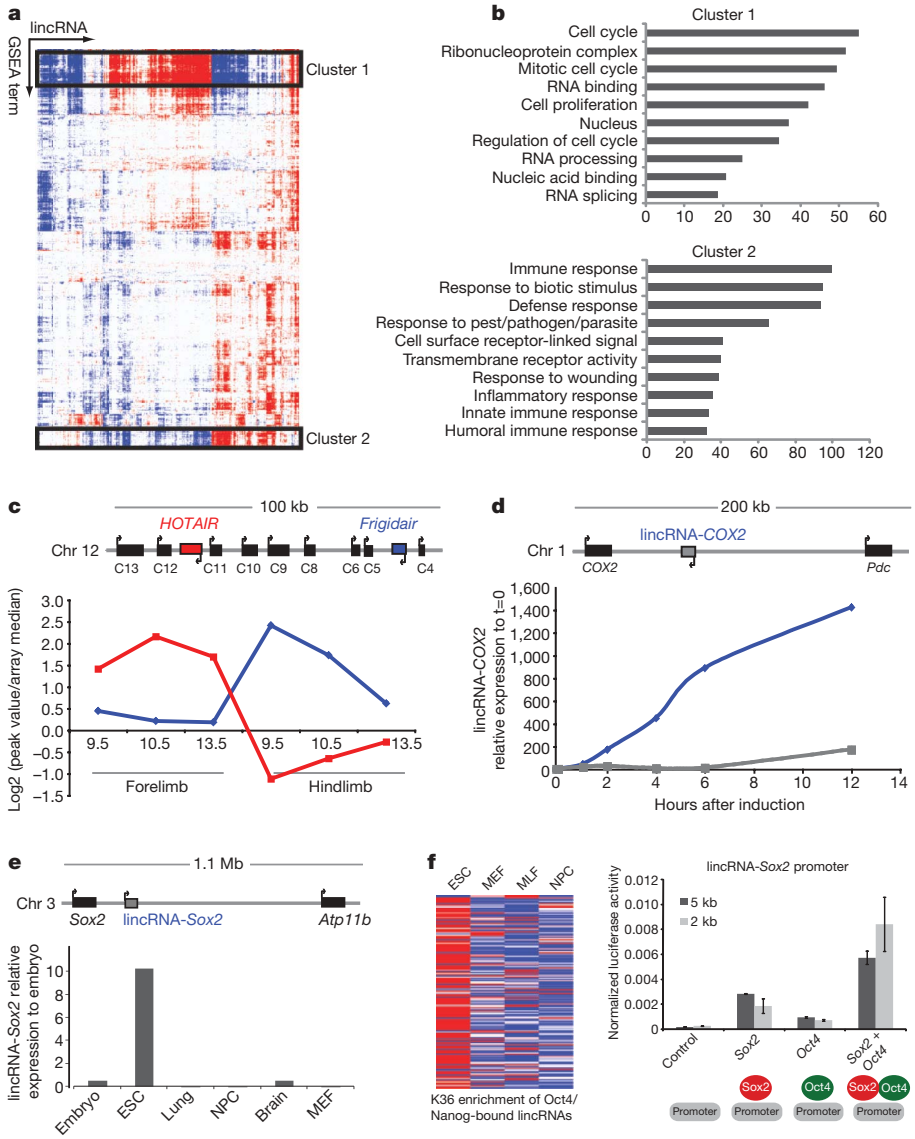


Figure 3 | lincRNAs show strong associations with other lincRNAs and with several biological processes. **a**, Association matrix of lincRNA and functional gene sets. Functional gene sets (columns) and lincRNAs (rows) are shown as positively (red), negatively (blue) or not associated (white) with lincRNA expression profiles. The black boxes highlight two significant biclusters in the matrix. **b**, Gene ontology of the protein-coding genes in these clusters is shown and plotted as the $-\log(P \text{ value})$ for the enrichment of each Gene Ontology term. **c**, Map of mouse genomic locus (*Hoxc*) containing *HOTAIR*. *HOTAIR* (red) and *Frigidair* (blue) show diametrically opposed expression patterns between mouse forelimb (anterior) and mouse hindlimb (posterior). **d**, Map of genomic locus containing *COX2* along with the location of lincRNA-*COX2*. Quantitative RT–PCR shows that lincRNA-*COX2* is upregulated in TLR4-stimulated cells (blue) but not TLR3-stimulated cells (grey). **e**, A map of the genomic locus containing *Sox2* shows a lincRNA ~50 kb upstream that is expressed specifically in ESCs. **f**, K36me3 enrichment across four cell types for lincRNAs bound by Oct4 or Nanog (left). Red indicates high enrichment, blue denotes low enrichment. The lincRNA-*Sox2* promoter was cloned into a luciferase reporter construct and assayed for transcriptional activity with *Sox2* and *Oct4* alone, together and controls (right). The *y*-axis represents the transcriptional activity of this promoter relative to a Renilla construct. Error bars are ± s.d. of three replicate transfections.

noted that several of the lincRNAs were located near genes encoding transcription factors (such as Sox2, Klf4, Myc and Brn1). Analysing the set of lincRNAs, we found that the genes neighbouring lincRNAs were strongly biased towards those encoding transcription factors ($P < 0.001$, permutation test; Supplementary Fig. 10 and Supplementary Table 17) and other proteins factors related to transcription. A second clue may come from our previous observation that HOTAIR[11] represses gene expression and is associated with chromatin remodelling proteins, together with recent similar observations for XIST[30]. On the basis of these observations, we speculate that many lincRNAs may be involved in transcriptional control—perhaps by guiding chromatin remodelling proteins to target loci—and that some transcription factors and lincRNAs may act together, with the transcription factor activating a transcriptional program and the lincRNA repressing a previous transcriptional program. Testing these speculations will require biochemical and genetic studies, including gene knockdown in appropriate settings. Whatever their functions, the highly conserved lincRNAs represent an important new contingent in the growing population of the modern 'RNA world'.

METHODS SUMMARY

Identifying intergenic K4–K36 domains and RNA. Enriched K4–K36 domains were identified using a sliding window approach across the genome and assessing the significance of each window. We filtered the list of K4–K36 enriched domains to eliminate known annotations. DNA tiling arrays (Nimblegen) were designed to tile intergenic K4–K36 domains. Transcribed regions were defined using a sliding window approach.

Conservation and coding potential. To detect sequence constraint we used a method that explicitly modelled the rate of mutation and level of constraint. We

©2009 Macmillan Publishers Limited. All rights reserved

took the maximum 12-base-pair window score for each exonic region. We normalized for the size differences between exons by computing a size matched random genomic score (Supplementary Methods).

We tested the protein-coding potential of K4–K36 domains by determining the maximum CSF score observed across the entire genomic locus. We computed the CSF scores across sliding windows of 90 base pairs and scanned all six possible reading frames in each window.

Protein-coding gene expression profiles. We generated a correlation matrix between lincRNAs and between lincRNAs and protein-coding genes by computing the Pearson correlation for all pairwise combinations. This matrix was clustered and visualized using the Gene Pattern platform for integrative genomics (http://genepattern.broad.mit.edu/). Functional associations were computed using Gene Set Enrichment Analysis (GSEA) (Supplementary Methods). In brief, we used each lincRNA as a profile, computed the Pearson correlation for each protein-coding gene and then ranked the protein-coding genes by their correlation coefficient. Gene sets were filtered by an FDR < 0.05 and an association matrix was generated.

Received 5 August; accepted 25 November 2008.
Published online 1 February 2009.

1. Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306,** 2242–2246 (2004).
2. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309,** 1559–1563 (2005).
3. Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296,** 916–919 (2002).
4. Rinn, J. L. *et al.* The transcriptional activity of human chromosome 22. *Genes Dev.* **17,** 529–540 (2003).
5. Ponjavic, J., Ponting, C. P. & Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.* **17,** 556–565 (2007).
6. Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nature Struct. Mol. Biol.* **14,** 103–105 (2007).
7. Brannan, C. I., Dees, E. C., Ingram, R. S. & Tilghman, S. M. The product of the *H19* gene may function as an RNA. *Mol. Cell. Biol.* **10,** 28–36 (1990).
8. Brown, C. J. *et al.* A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* **349,** 38–44 (1991).
9. Lee, J. T., Davidow, L. S. & Warshawsky, D. *Tsix*, a gene antisense to *Xist* at the X-inactivation centre. *Nature Genet.* **21,** 400–404 (1999).
10. Sotomaru, Y. *et al.* Unregulated expression of the imprinted genes *H19* and *Igf2r* in mouse uniparental fetuses. *J. Biol. Chem.* **277,** 12474–12478 (2002).
11. Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human *HOX* loci by noncoding RNAs. *Cell* **129,** 1311–1323 (2007).
12. Willingham, A. T. *et al.* A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* **309,** 1570–1573 (2005).
13. Wang, J. *et al.* Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs. *Nature* **431,** 1–2, doi: 10.1038/nature03016 (2004).
14. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560 (2007).
15. Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A. & Enright, A. J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **34,** D140–D144 (2006).
16. Tam, O. H. *et al.* Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453,** 534–538 (2008).
17. Watanabe, T. *et al.* Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* **453,** 539–543 (2008).
18. Clamp, M. *et al.* Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl Acad. Sci. USA* **104,** 19428–19433 (2007).
19. Lin, M. F. *et al.* Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes. *Genome Res.* **17,** 1823–1836 (2007).
20. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).
21. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genet.* **38,** 626–635 (2006).
22. Su, A. I. *et al.* Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA* **99,** 4465–4470 (2002).
23. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102,** 15545–15550 (2005).
24. Tanay, A., Sharan, R. & Shamir, R. Discovering statistically significant biclusters in gene expression data. *Bioinformatics* **18** (Suppl 1), S136–S144 (2002).
25. Chang, H. Y. *et al.* Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc. Natl Acad. Sci. USA* **102,** 3738–3743 (2005).
26. Carrio, M., Arderiu, G., Myers, C. & Boudreau, N. J. Homeobox D10 induces phenotypic reversion of breast tumor cells in a three-dimensional culture model. *Cancer Res.* **65,** 7177–7185 (2005).
27. Ventura, A. *et al.* Cre-lox-regulated conditional RNA interference from transgenes. *Proc. Natl Acad. Sci. USA* **101,** 10380–10385 (2004).
28. Loh, Y. H. *et al.* The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nature Genet.* **38,** 431–440 (2006).
29. Ivanova, N. *et al.* Dissecting self-renewal in stem cells with RNA interference. *Nature* **442,** 533–538 (2006).
30. Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J. & Lee, J. T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322,** 750–756 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We would like to thank our colleagues at the Broad Institute, especially J. P. Mesirov for discussions and statistical insights, X. Xie for statistical help with conservation analyses, J. Robinson for visualization help, M. Ku, E. Mendenhall and X. Zhang for help generating ChIP samples, and N. Novershtern and A. Levy for providing transcription factor lists. M. Guttman is a Vertex scholar, I.A. acknowledges the support of the Human Frontier Science Program Organization. This work was funded by Beth Israel Deaconess Medical Center, National Human Genome Research Institute, and the Broad Institute of MIT and Harvard.

Author Contributions J.L.R., E.S.L., A.R. and M. Guttman conceived and designed experiments. The manuscript was written by M. Guttman, A.R., J.L.R. and E.S.L. J.L.R., I.A., C.F., D.F., M.H., B.W.C., J.P.C. and M. Guttman performed molecular biology experiments. All data analyses were performed by M. Guttman in conjunction with M. Garber (conservation analyses), M.F.L. (codon substitution frequency), T.S.M. (ChIP-seq data), O.Z. (motif analysis) and M.N.C. (lincRNA genomic location analysis). Reagents were provided by M. Garber (pre-published conservation analysis tools); T.J. and D.F. (p53 wild-type and knockout MEFs); N.H., A.R. and I.A. (dendritic cell stimulated time course); B.E.B. (ChIP data); R.J., B.W.C. and J.P.C. (luciferase assays); and M.K. and M.F.L. (codon substitution frequency code).

Author Information Microarray data have been deposited in the Gene Expression Omnibus (GEO) under accession number GSE13765. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.L.R. (jrinn@broad.mit.edu).

©2009 Macmillan Publishers Limited. All rights reserved

LETTERS

Histone modifications at human enhancers reflect global cell-type-specific gene expression

Nathaniel D. Heintzman[1,2]*, Gary C. Hon[1,3]*, R. David Hawkins[1]*, Pouya Kheradpour[5], Alexander Stark[5,6], Lindsey F. Harp[1], Zhen Ye[1], Leonard K. Lee[1], Rhona K. Stuart[1], Christina W. Ching[1], Keith A. Ching[1], Jessica E. Antosiewicz-Bourget[7], Hui Liu[8], Xinmin Zhang[8], Roland D. Green[8], Victor V. Lobanenkov[9], Ron Stewart[7], James A. Thomson[7,10], Gregory E. Crawford[11], Manolis Kellis[5,6] & Bing Ren[1,4]

The human body is composed of diverse cell types with distinct functions. Although it is known that lineage specification depends on cell-specific gene expression, which in turn is driven by promoters, enhancers, insulators and other *cis*-regulatory DNA sequences for each gene[1–3], the relative roles of these regulatory elements in this process are not clear. We have previously developed a chromatin-immunoprecipitation-based microarray method (ChIP-chip) to locate promoters, enhancers and insulators in the human genome[4–6]. Here we use the same approach to identify these elements in multiple cell types and investigate their roles in cell-type-specific gene expression. We observed that the chromatin state at promoters and CTCF-binding at insulators is largely invariant across diverse cell types. In contrast, enhancers are marked with highly cell-type-specific histone modification patterns, strongly correlate to cell-type-specific gene expression programs on a global scale, and are functionally active in a cell-type-specific manner. Our results define over 55,000 potential transcriptional enhancers in the human genome, significantly expanding the current catalogue of human enhancers and highlighting the role of these elements in cell-type-specific gene expression.

We performed ChIP-chip analysis as described previously[5] to determine binding of CTCF (insulator-binding protein) and the coactivator p300 (also known as EP300), and patterns of histone modifications in five human cell lines: cervical carcinoma HeLa, immortalized lymphoblast GM06690 (GM), leukaemia K562, embryonic stem cells (ES) and BMP4-induced ES cells (dES). We first investigated 1% of the human genome selected by the ENCODE consortium[7], using DNA microarrays consisting of 385,000 50-base oligonucleotides that tile 30-million base pairs (bp) at 36 bp resolution. We examined mono- and tri-methylation of histone H3 lysine 4 (H3K4me1, H3K4me3) and acetylation of histone H3 lysine 27 (H3K27ac) at well-annotated promoters, reasoning that the state of these histone modifications would vary in a cell-type-specific manner. To our surprise, the chromatin signatures at promoters are remarkably similar across all cell types (Fig. 1a). Quantitative comparison of ChIP-chip enrichment (see Supplementary Information) revealed highly correlated histone modification patterns at promoters across all cell types, with an average Pearson correlation coefficient of 0.71 (Supplementary Fig. 1a). This observation also holds for the larger set of Gencode promoters (Supplementary Fig. 2).

Next, we identified putative insulators in the ENCODE regions for these cell types based on CTCF binding, because mammalian insulators are generally understood to require CTCF to block promoter–enhancer interactions[3]. We observed nearly identical CTCF occupancy (Supplementary Table 1 and Supplementary Fig. 1e) and highly correlated CTCF enrichment patterns across all five cell types (Supplementary Fig. 1b), providing experimental support for the mostly cell-type-invariant function of CTCF as suggested by DNase hypersensitivity mapping results[8].

We then investigated transcriptional enhancers in the ENCODE regions, performing ChIP-chip in HeLa, K562 and GM cells to locate binding sites for the transcriptional coactivator protein p300 (Supplementary Tables 2–4) because p300 is known to localize at enhancers[9]. We observed highly cell-type-specific histone modification patterns at distal p300-binding sites (Supplementary Fig. 1f), in marked contrast to the similarities in histone modifications across cell types at promoters. We then used our chromatin-signature-based prediction method[5] to identify additional enhancers in the ENCODE regions in these cell types (Fig. 1b and Supplementary Tables 5–9). In addition to the characteristic H3K4me1 enrichment, predicted enhancers are frequently marked by acetylation of H3K27, DNaseI hypersensitivity and/or binding of transcription factors and coactivators, and many contain evolutionarily conserved sequences (Supplementary Figs 3 and 4; see Supplementary Information). Unlike promoters and insulators, but similar to p300-binding sites, the histone modification patterns at predicted enhancers are largely cell-type-specific (Fig. 1b and Supplementary Fig. 1d), in agreement with observations that H3K4me1 is distributed in a cell-type-specific manner[10].

These results indicate that enhancers are the most variable class of transcriptional regulatory element between cell types and are probably of primary importance in driving cell-type-specific patterns of gene expression. Knowledge of enhancers is therefore critical for understanding the mechanisms that control cell-type-specific gene expression, yet our incomplete knowledge of enhancers in the human genome has confined previous studies of gene regulatory networks mainly to promoters. To identify enhancers on a genome-wide scale and facilitate global analysis of gene regulatory mechanisms, we performed ChIP-chip throughout the entire human genome as described[6], mapping enrichment patterns of H3K4me1 and H3K4me3 in HeLa cells. Using previously described chromatin signatures for enhancers[5], we

[1]Ludwig Institute for Cancer Research, [2]Biomedical Sciences Graduate Program, [3]Bioinformatics Program, and [4]Department of Cellular and Molecular Medicine, UCSD School of Medicine, 9500 Gilman Drive, La Jolla, California 92093-0653, USA. [5]MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar Street, Cambridge, Massachusetts 02139, USA. [6]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [7]Morgridge Institute for Research, Madison, Wisconsin 53707-7365, USA. [8]Roche NimbleGen, Inc., 500 South Rosa Road, Madison, Wisconsin 53719, USA. [9]National Institutes of Allergy and Infectious Disease, 5640 Fishers Lane, Rockville, Maryland 20852, USA. [10]University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin 53706, USA. [11]Institute for Genome Sciences and Policy, and Department of Pediatrics, Duke University, 101 Science Drive, Durham, North Carolina 27708, USA.
*These authors contributed equally to this work.

©2009 Macmillan Publishers Limited. All rights reserved

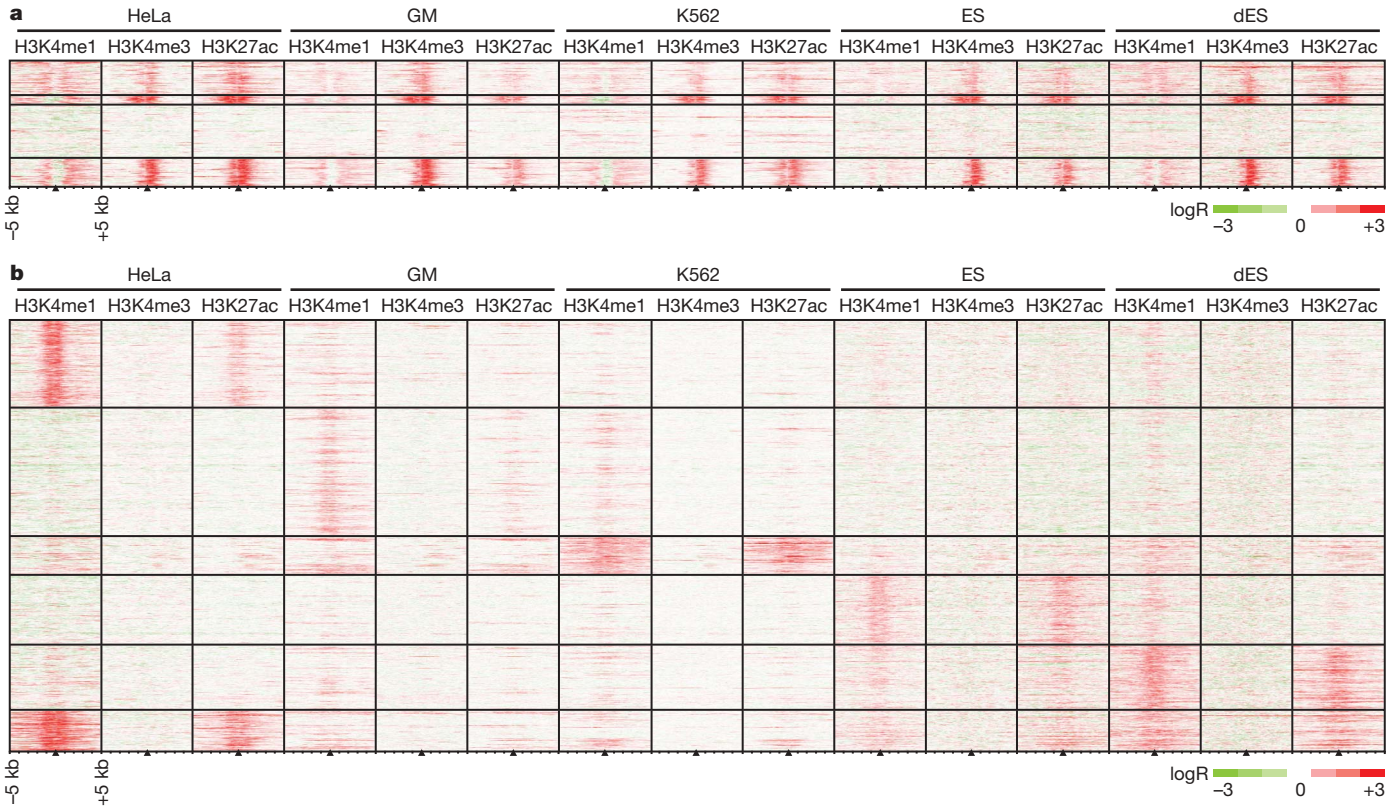


Figure 1 | Chromatin modifications at promoters are generally cell-type-invariant whereas those at enhancers are cell-type-specific. We used ChIP-chip to map histone modifications (H3K4me1, H3K4me3 and H3K27ac) in the ENCODE regions in five cell types (HeLa, GM, K562, ES, dES). **a**, We performed *k*-means clustering on the chromatin modifications found ±5 kb from 414 promoters, and observe them to be generally invariant across cell types. **b**, As in **a**, but clustering on 1,423 non-redundant enhancers predicted on the basis of chromatin signatures, revealing the cell-type-specificity of enhancers. LogR is the log ratio of enrichment of each marker as determined by ChIP-chip. Promoters and predicted enhancers are located at the centre of 10-kb windows as indicated by black triangles.

predicted 36,589 enhancers in the HeLa genome (Fig. 2a and Supplementary Table 10; see Supplementary Information). This method correctly located several previously characterized enhancers, including the β-globin HS2 enhancer[11] and distal enhancers for the PAX6 (ref. 12) and PLAT[13] genes (Fig. 2b). Most predicted enhancers are distal to promoters (Fig. 2c), have strong evolutionary conservation (see Supplementary Information) and are marked by histone acetylation (H3K27ac), binding of coactivator proteins (p300, MED1) or DNaseI hypersensitivity (DHS) (Fig. 2a, d; see Supplementary Information). We verified the functional potential of predicted HeLa enhancers using luciferase reporter assays as described[5] (see Supplementary Methods). Out of nine predicted enhancers that we evaluated, seven (78%) were active in reporter assays (Fig. 2e and Supplementary Table 11), with median activity significantly different from that in random genomic regions ($P = 3.25 \times 10^{-4}$). These results support the suitability of using chromatin signatures to identify genomic regions with enhancer function.

We evaluated the predicted enhancers for conserved motif-like sequences using several-hundred shuffled TRANSFAC motifs across ten mammals in a phylogenetic framework that tolerates motif movement, partial motif loss and sequencing or alignment discrepancies (see Supplementary Methods). Predicted enhancers showed conservation for 4.3% of instances (at branch-length-score >50%, see Supplementary Methods), which is substantially greater than for the remaining intergenic regions (2.9%, $P < 1 \times 10^{-100}$) and even promoter regions (3.9%, $P = 1 \times 10^{-57}$). Additionally, testing a list of 123 unique TRANSFAC motifs as described[14] (see Supplementary Information), we found that 67 (54%) are over-conserved and 39 (32%) are enriched in predicted enhancers (Supplementary Table 12). We also performed *de novo* motif discovery in enhancer regions using multiple alignments of 10 mammalian genomes (see Supplementary Methods), revealing 41 enhancer motifs, 19 of which match known transcription factor motifs whereas 22 are new (Supplementary Table 13). These motifs show conservation rates between 7% and 22% in enhancers (median 9.3%), compared to only 1.1% for control shuffled motifs of identical composition. Furthermore, over 90% of these motifs seem to be unique to enhancers, as only 4 motifs are enriched in promoter regions and 12 are in fact depleted in promoters (Supplementary Table 13), indicating that predicted enhancers contain unique regulatory sequences that may be specific to enhancer function.

To investigate the association of predicted enhancers with HeLa-specific gene expression, we used Shannon entropy[15] to rank genes by the specificity of their expression levels in HeLa compared to that in three other cell lines (K562, GM06990, IMR90) (Supplementary Fig. 5; see Supplementary Information), and then plotted the distribution of enhancers around genes within insulator-delineated domains (as defined by CTCF-binding sites in Supplementary Fig. 6; see Supplementary Information). Predicted enhancers are markedly enriched near HeLa-specific expressed genes (Fig. 3a), particularly within 200 kb of promoters. We observed a 1.83-fold enrichment ($P = 4.71 \times 10^{-279}$) of predicted enhancers around HeLa-specific expressed genes relative to a random distribution (see Supplementary Information) and significant depletion of enhancers around non-specific expressed genes ($P = 5.43 \times 10^{-15}$) and HeLa-specific repressed genes ($P = 4.63 \times 10^{-2}$).

To investigate more directly the relationship between chromatin modification patterns at enhancers and cell-type-specific gene expression, we expanded our global analysis to another cell type. We performed genome-wide ChIP-chip for H3K4me1 and H3K4me3 in K562 cells and identified 24,566 putative enhancers in this cell type using our chromatin-signature-based enhancer-prediction method (Supplementary Table 14; see Supplementary Information).

©2009 Macmillan Publishers Limited. All rights reserved

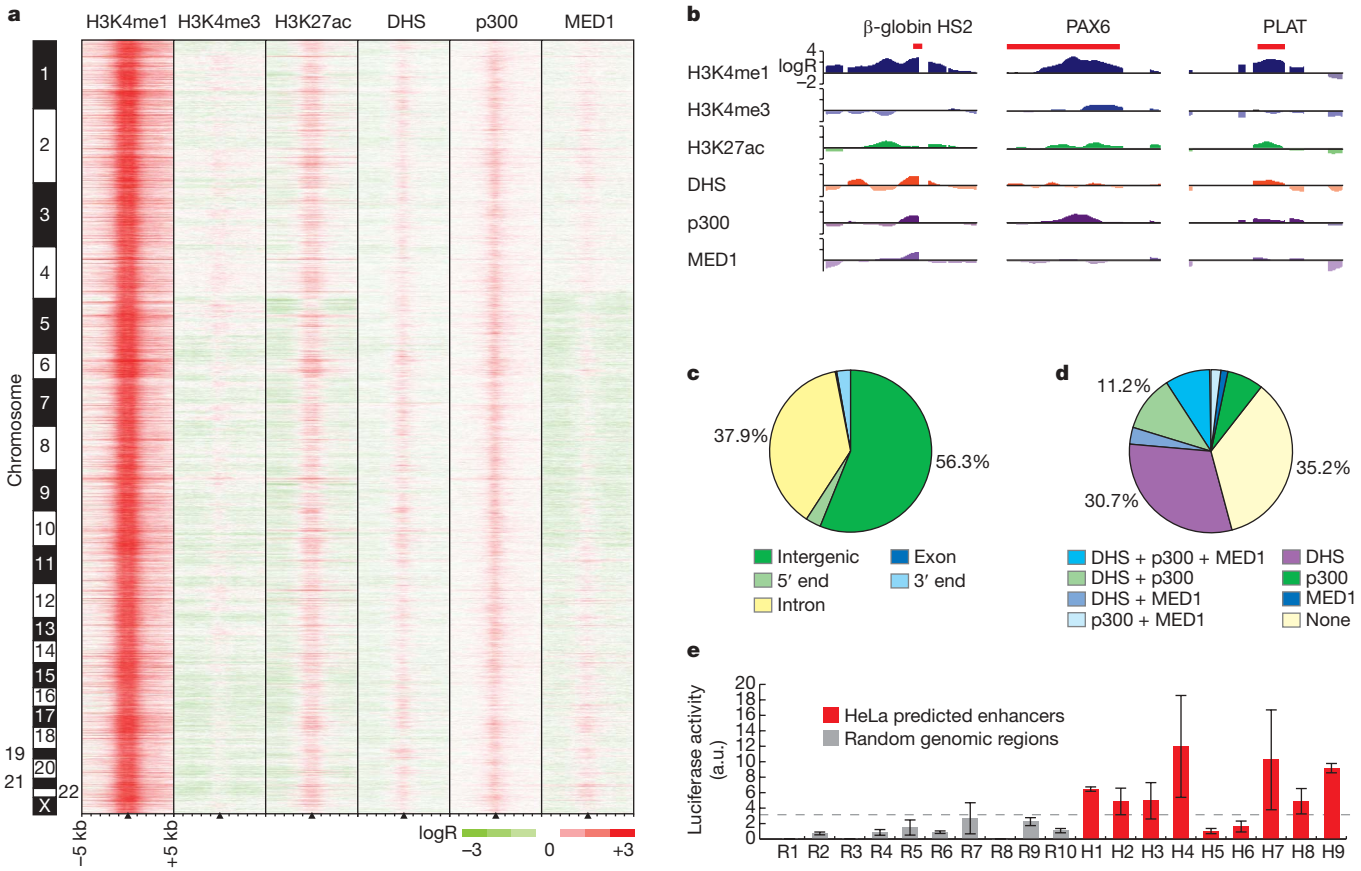


Figure 2 | Genome-wide enhancer predictions in human cells. a, We predict 36,589 enhancers in HeLa cells on the basis of chromatin signatures for H3K4me1 and H3K4me3 as determined by ChIP-chip using genome-wide tiling microarrays and condensed enhancer microarrays (see Supplementary Information). Enhancer predictions are located at the centre of 10-kb windows as indicated by black triangles, and ordered by genomic position. Enrichment data are shown for histone modifications (H3K4me1, H3K4me3 and H3K27ac), DNaseI hypersensitivity (DHS), and binding of p300 and MED1. **b**, ChIP-chip enrichment profiles at several known enhancers (indicated in red) recovered by prediction: β-globin HS2 (chromosome 11: 5,258,371–5,258,665)[11], PAX6 (chromosome 11: 31,630,500–31,635,000)[12], PLAT (chromosome 8: 42,191,500–42,192,400)[13] (5-kb windows centred on enhancer predictions; images generated in part at the UCSC Genome Browser). **c**, Predicted enhancer distribution relative to UCSC Known Genes. Most enhancers have intergenic (56.3%) or intronic (37.9%) localization relative to UCSC Known Gene 5′-ends. **d**, Most enhancers (64.8%) are significantly marked by DNaseI hypersensitivity, binding of p300, binding of MED1, or some combination thereof. **e**, Seven out of nine enhancers predicted in HeLa cells were active in reporter assays (red bars) as compared to none of the random fragments selected as controls (grey), where activity is defined as relative luciferase value greater than 2.33 standard deviations ($P = 0.01$) above the median random activity (grey dashed line). Error bars, standard deviation. Regions of ~1–2 kb in size were randomly selected for validation in reporter assays based on histone modification patterns as in **a**, overlap with features in **d**, and sequence features amenable to cloning by means of polymerase chain reaction (see Supplementary Information). a.u., arbitrary units.

Consistent with results in the ENCODE regions, most enhancers predicted in K562 and HeLa cells are unique to either cell type (Fig. 3b), even though most expressed genes are common between the cell types (Fig. 3c). Chromatin modification profiles at predicted enhancers throughout the genome are highly cell-type-specific (Fig. 3d), with a Pearson correlation coefficient of −0.32. Furthermore, these differences seem to have regulatory implications, because domains with HeLa-specific expressed genes are enriched in HeLa enhancers but depleted in K562 enhancers, and vice versa (Fig. 3e; see Supplementary Information). These observations hold across all five cell types in the ENCODE regions (see Supplementary Information). To assess the cell-type-specificity of enhancer activity, we cloned enhancers predicted specifically in K562 cells (and not in HeLa cells) and subjected them to reporter assays in HeLa cells as described above. Out of nine K562-specific enhancers tested, only two (22%) were active in HeLa cells (Supplementary Fig. 7), and the median activity of the K-562-specific enhancers was not significantly different from random ($P = 0.11$), indicating that the enhancer chromatin signature is a reliable marker of cell-type-specific enhancer function.

Although most enhancers are cell-type-specific, the presence of predicted enhancers shared by HeLa and K562 (Fig. 3b, d) indicates that some enhancers may be active in multiple cell types or conditions. We compared the HeLa enhancer predictions with the results of several genome-wide studies of binding sites for sequence-specific transcription factors in different cell types, namely oestrogen receptor[16] (ER, also known as ESR1), p53 (TP53; ref. 17) and p63 (TP63; ref. 18) in MCF7, HCT116 and ME180 cells, respectively. Interestingly, significant percentages of binding sites for each transcription factor (from 21.4% to 32.6%) overlap with predicted enhancers in HeLa cells (Fig. 4a and Supplementary Table 15), in contrast to a significant depletion of the repressor NRSF (also known as REST)[19] at predicted enhancers and minimal overlap with CTCF-binding sites (see Supplementary Information).

To examine the potential role of enhancers in regulating inducible gene expression, we treated HeLa cells with the cytokine interferon-γ (IFN-γ) and identified binding sites for the transcription factor STAT1 throughout the genome using ChIP-chip. STAT1 generally binds its target DNA sequences only after IFN-γ induction[20], with a small fraction of binding possible before induction[21]. In IFN-γ-treated HeLa cells, we identified 1,969 STAT1-binding sites (Supplementary Table 16), with 85.8% of STAT1-binding sites occurring distal to Known Gene 5′-ends. Comparison of these distal STAT1-binding sites with recent

©2009 Macmillan Publishers Limited. All rights reserved

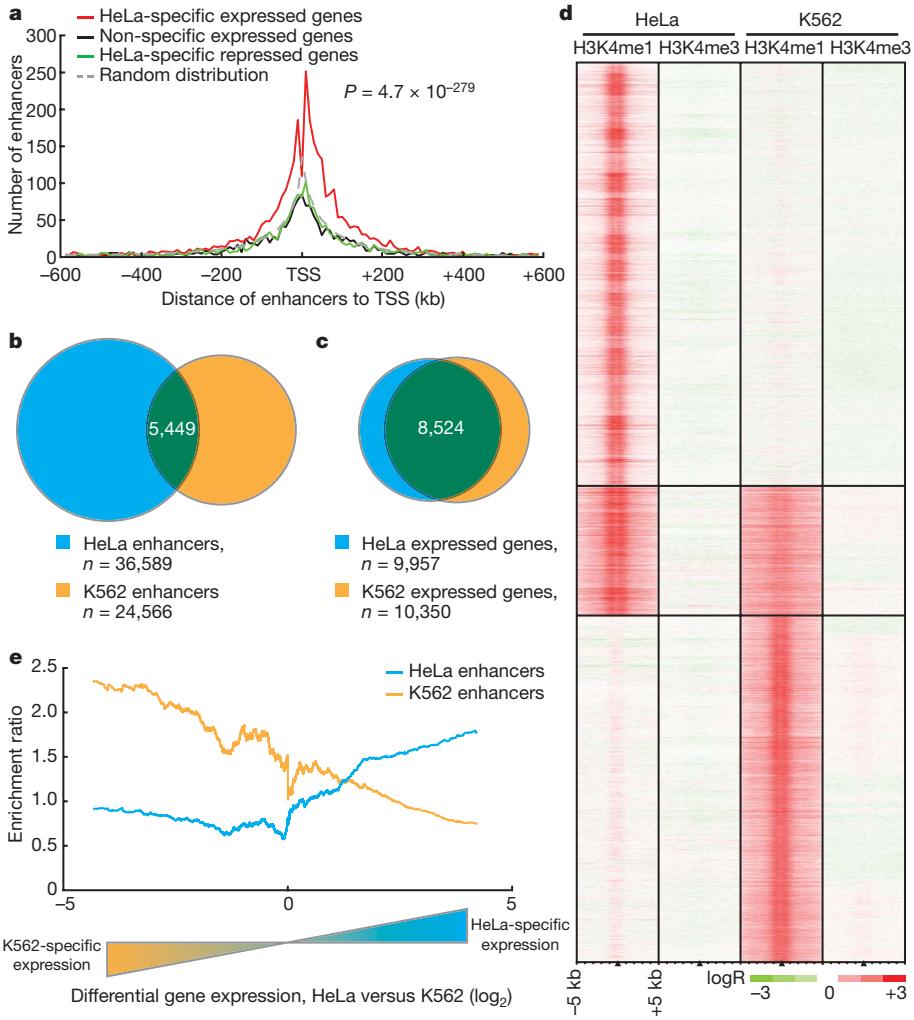


Figure 3 | Chromatin modifications at enhancers are globally related to cell-type-specific gene expression. **a**, Enhancer localization relative to HeLa-specific expressed genes compared to K562, GM06990 and IMR90 cells (red), non-specific expressed (green), HeLa-specific repressed (black), and a random distribution (dashed grey). Predicted enhancers are enriched around HeLa-specific expressed genes within insulator-defined domains and depleted in domains of ubiquitous or non-expressed genes (*P*-value reflects significance of enhancer enrichment in domains of HeLa-specific expressed genes, see Supplementary Information). TSS, transcription start site. **b**, **c**, Most enhancers predicted in HeLa and K562 cells are cell-type-specific (**b**) whereas most genes in HeLa and K562 cells are not specifically expressed (**c**); *n* = integer number of enhancers or genes in each set. **d**, Chromatin modification patterns are cell-type-specific at most of the 55,454 enhancers predicted in HeLa and K562 cells. **e**, Comparison of enhancer enrichment and differential gene expression between HeLa cells and K562 cells revealed that HeLa enhancers are enriched near HeLa-specific expressed genes (blue line) whereas K562 enhancers are enriched near K562-specific expressed genes (orange line).

analysis of STAT1 binding in uninduced HeLa cells[21] shows that only 6.5% of IFN-γ-induced STAT1-binding sites are occupied by STAT1 before induction. We observed that 429 distal STAT1-binding sites overlapped enhancers predicted in HeLa cells before induction (Fig. 4a and Supplementary Table 15). The H3K4me1 enhancer chromatin signature is present before induction at these STAT1-binding sites, which we designated as STAT1 group I, whereas no evidence of this signature was visible at the remaining 1,260 distal STAT1-binding sites, designated STAT1 group II (Fig. 4b). Intriguingly, we observed significant relative induction of expression of genes in the domains of STAT1-group-I-binding sites after just 30 min of IFN-γ induction, whereas induction levels remained relatively unchanged for genes in the domains of other distal STAT1-group-II-binding sites during this time (Fig. 4c). These findings indicate that an enhancer chromatin signature confers increased regulatory responsiveness to a STAT1-binding site, in agreement with our previous discovery of functional enhancers in HeLa cells that were marked by the enhancer chromatin signature but were not active until they were bound by STAT1 (ref. 5).

Our findings offer, to our knowledge, the first genome-wide evaluation of the relationship between chromatin modifications at transcriptional enhancers and global programs of cell-type-specific gene expression. We determined over 55,000 potential enhancers in the human genome and show that the chromatin modifications at the

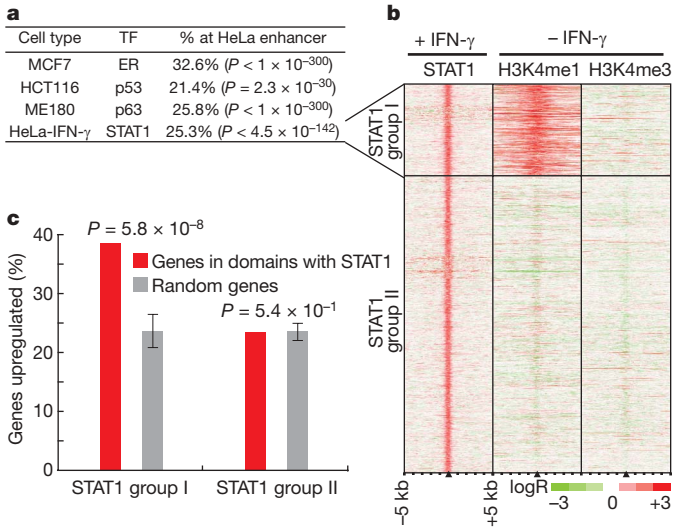


Figure 4 | Chromatin modifications are associated with an increased regulatory response of transcription-factor-binding sites at enhancers. **a**, Predicted enhancers in steady-state HeLa cells overlap with significant fractions of transcription-factor-binding sites (ER, p53, p63) in diverse cell types (MCF7, HCT116, ME180), as well as with STAT1-binding sites in HeLa cells treated with the cytokine interferon-γ (HeLa-IFN-γ) (TF, transcription factor; TFBS, transcription factor binding sites). **b**, Hundreds of STAT1-binding sites after treatment (+IFN-γ) are marked by the enhancer chromatin signature in HeLa cells even before treatment (−IFN-γ). **c**, In HeLa cells treated with IFN-γ (upper panel), gene expression is significantly ($P = 5.8 \times 10^{-8}$) more likely to be induced by STAT1 binding at sites with the enhancer chromatin signature (red, STAT1 group I) than by STAT1 binding at other distal sites (red, STAT1 group II) relative to a random distribution (grey). Error bars, standard deviation.

©2009 Macmillan Publishers Limited. All rights reserved

enhancers correlate with cell-type-specific gene expression and functional enhancer activity. Perhaps the most intriguing observation is the large number of enhancers identified from the investigation of just two cell lines. Because enhancers are mostly cell-type-specific, our data indicate the existence of a vast number of enhancers in the human genome, on the order of 10^5-10^6, that are used to drive specific gene expression programs in the 200 cell types of the human body. Future experiments with diverse cell types and experimental conditions will be necessary to comprehensively identify these regulatory elements and understand their roles in the specific gene expression program of each cell type.

METHODS SUMMARY

HeLa, K562 and IMR90 cells were obtained from ATCC. GM06990 cells were acquired from Coriell. All cells were cultured under recommended conditions. Passage 32 H1 cells were cultured as described[22] with/without 200 ng ml^{-1} BMP4 for 6 days (RND Systems). Chromatin preparation, ChIP, DNA purification and ligation-mediated PCR were performed as described using commercially available and custom-made antibodies, and ChIP samples were hybridized to tiling microarrays and to custom-made condensed enhancer microarrays (NimbleGen Systems, Inc.) as described[5,6]. DNase-chip was performed and the data analysed as described[23]. Cloning and reporter assays were performed as described[5]. Data were normalized as described[5], and ChIP-chip targets for CTCF, p300, MED1 and STAT1 were selected with the Mpeak program. We used MA2C (ref. 24) to normalize and call peaks on Nimblegen HD2 arrays. Enhancers were predicted, and *k*-means clustering, intersection analysis and evolutionary conservation analysis were performed as described[5]. Motif analysis was performed as described[25]. Gene expression was analysed using HGU133 Plus 2.0 microarrays (Affymetrix) as described[5]. Specificity of expression was determined using a function of Shannon entropy[15]. We use the MAS5 algorithm from the Bioconductor R package to generate gene expression present/absent calls. Detailed methods can be found in Supplementary Information. Supplementary data for the microarray experiments have been formatted for viewing in the UCSC genome browser via http://bioinformatics-renlab.ucsd.edu/enhancer.

Received 17 October 2008; accepted 26 January 2009.
Published online 18 March 2009.

1. Heintzman, N. D. & Ren, B. The gateway to transcription: identifying, characterizing and understanding promoters in the eukaryotic genome. *Cell. Mol. Life Sci.* **64,** 386–400 (2007).
2. Nightingale, K. P., O'Neill, L. P. & Turner, B. M. Histone modifications: signalling receptors and potential elements of a heritable epigenetic code. *Curr. Opin. Genet. Dev.* **16,** 125–136 (2006).
3. Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.* **7,** 29–59 (2006).
4. Kim, T. H. *et al.* A high-resolution map of active promoters in the human genome. *Nature* **436,** 876–880 (2005).
5. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genet.* **39,** 311–318 (2007).
6. Kim, T. H. *et al.* Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128,** 1231–1245 (2007).
7. ENCODE Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306,** 636–640 (2004).
8. Xi, H. *et al.* Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genet.* **3,** e136 (2007).
9. Wang, Q., Carroll, J. S. & Brown, M. Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking. *Mol. Cell* **19,** 631–642 (2005).
10. Koch, C. M. *et al.* The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res.* **17,** 691–707 (2007).
11. King, D. C. *et al.* Evaluation of regulatory potential and conservation scores for detecting *cis*-regulatory modules in aligned mammalian genome sequences. *Genome Res.* **15,** 1051–1060 (2005).
12. Kleinjan, D. A. *et al.* Aniridia-associated translocations, DNase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of PAX6. *Hum. Mol. Genet.* **10,** 2049–2059 (2001).
13. Wolf, A. T., Medcalf, R. L. & Jern, C. The t-PA -7351C>T enhancer polymorphism decreases Sp1 and Sp3 protein binding affinity and transcriptional responsiveness to retinoic acid. *Blood* **105,** 1060–1067 (2005).
14. Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434,** 338–345 (2005).
15. Schug, J. *et al.* Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol.* **6,** R33 (2005).
16. Carroll, J. S. *et al.* Genome-wide analysis of estrogen receptor binding sites. *Nature Genet.* **38,** 1289–1297 (2006).
17. Wei, C. L. *et al.* A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124,** 207–219 (2006).
18. Yang, A. *et al.* Relationships between p63 binding, DNA sequence, transcription activity, and biological function in human cells. *Mol. Cell* **24,** 593–602 (2006).
19. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* **316,** 1497–1502 (2007).
20. Brivanlou, A. H. & Darnell, J. E. Jr. Signal transduction and the control of gene expression. *Science* **295,** 813–818 (2002).
21. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods* **4,** 651–657 (2007).
22. Ludwig, T. E. *et al.* Feeder-independent culture of human embryonic stem cells. *Nature Methods* **3,** 637–646 (2006).
23. Crawford, G. E. *et al.* DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nature Methods* **3,** 503–509 (2006).
24. Song, J. S. *et al.* Model-based analysis of two-color arrays (MA2C). *Genome Biol.* **8,** R178 (2007).
25. Kheradpour, P., Stark, A., Roy, S. & Kellis, M. Reliable prediction of regulator targets using 12 *Drosophila* genomes. *Genome Res.* **17,** 1919–1931 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank members of the Ren laboratory for comments. This work was supported by funding from American Cancer Society (R.D.H.), NIAID Intramural Research Program (V.V.L.), LICR (B.R.), NHGRI (B.R.), NCI (B.R.) and CIRM (B.R.).

Author Contributions R.D.H., N.D.H., G.C.H. and B.R. designed the experiments; R.D.H., N.D.H., L.F.H., Z.Y., L.K.L., R.K.S., C.W.C., H.L. and X.Z. conducted the ChIP-chip experiments; G.C.H. and K.A.C. analysed the ChIP-chip data; G.C.H. predicted enhancers; R.D.H. and L.K.L. conducted the reporter assays; J.E.A.-B., R.S. and J.A.T. provided human ES cells and expression data; V.V.L. provided advice and antibodies for CTCF-ChIP experiments; P.K., A.S. and M.K. analysed the transcription factor motifs; G.E.C. performed and analysed the DNaseI-chip experiments; and N.D.H., G.C.H., R.D.H. and B.R. wrote the manuscript.

Author Information Microarray data have been submitted to the GEO repository under accession numbers GSE14083, GSE8098, GSE7872 and GSE7118. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to B.R. (biren@ucsd.edu).

©2009 Macmillan Publishers Limited. All rights reserved

ARTICLES

Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes

Geraldine Butler[1], Matthew D. Rasmussen[2], Michael F. Lin[2,3], Manuel A. S. Santos[4], Sharadha Sakthikumar[3], Carol A. Munro[5], Esther Rheinbay[2,6], Manfred Grabherr[3], Anja Forche[7], Jennifer L. Reedy[8], Ino Agrafioti[9], Martha B. Arnaud[10], Steven Bates[11], Alistair J. P. Brown[5], Sascha Brunke[12], Maria C. Costanzo[10], David A. Fitzpatrick[1], Piet W. J. de Groot[13], David Harris[14], Lois L. Hoyer[15], Bernhard Hube[12], Frans M. Klis[13], Chinnappa Kodira[3]†, Nicola Lennard[14], Mary E. Logue[1], Ronny Martin[12], Aaron M. Neiman[16], Elissavet Nikolaou[5], Michael A. Quail[14], Janet Quinn[17], Maria C. Santos[4], Florian F. Schmitzberger[10], Gavin Sherlock[10], Prachi Shah[10], Kevin A. T. Silverstein[18], Marek S. Skrzypek[10], David Soll[19], Rodney Staggs[18], Ian Stansfield[5], Michael P. H. Stumpf[9], Peter E. Sudbery[20], Thyagarajan Srikantha[19], Qiandong Zeng[3], Judith Berman[7], Matthew Berriman[14], Joseph Heitman[8], Neil A. R. Gow[5], Michael C. Lorenz[21], Bruce W. Birren[3], Manolis Kellis[2,3]* & Christina A. Cuomo[3]*

***Candida* species are the most common cause of opportunistic fungal infection worldwide. Here we report the genome sequences of six *Candida* species and compare these and related pathogens and non-pathogens. There are significant expansions of cell wall, secreted and transporter gene families in pathogenic species, suggesting adaptations associated with virulence. Large genomic tracts are homozygous in three diploid species, possibly resulting from recent recombination events. Surprisingly, key components of the mating and meiosis pathways are missing from several species. These include major differences at the mating-type loci (*MTL*); *Lodderomyces elongisporus* lacks *MTL*, and components of the a1/α2 cell identity determinant were lost in other species, raising questions about how mating and cell types are controlled. Analysis of the CUG leucine-to-serine genetic-code change reveals that 99% of ancestral CUG codons were erased and new ones arose elsewhere. Lastly, we revise the *Candida albicans* gene catalogue, identifying many new genes.**

Four *Candida* species, *C. albicans*, *C. glabrata*, *C. tropicalis* and *C. parapsilosis*, together account for ~95% of identifiable *Candida* infections[1]. Although *C. albicans* is still the most common causative agent, its incidence is declining and the frequency of other species is increasing. Of these, *C. parapsilosis* is a particular problem in neonates, transplant recipients and patients receiving parenteral nutrition; *C. tropicalis* is more commonly associated with neutropenia and malignancy. Other *Candida* species, including *C. krusei*, *C. lusitaniae* and *C. guilliermondii*, account for <5% of invasive candidiasis. Almost all *Candida* species, with the exceptions of *C. glabrata* and *C. krusei*, belong in a single *Candida* clade (Fig. 1) characterized by the unique translation of CUG codons as serine rather than leucine[2]. Within this, haploid and diploid species occupy two separate subclades (Fig. 1).

To determine the genetic features underlying their diversity of biology and pathogenesis, we sequenced six genomes from the *Candida* clade (Fig. 1). These include a second sequenced isolate of *C. albicans* (WO-1) characterized for white–opaque switching, a phenotypic change that correlates with host specificity and mating[3,4]. We also sequenced the major pathogens *C. tropicalis* and *C. parapsilosis*; *L. elongisporus*, a close relative of *C. parapsilosis* recently identified as a cause of bloodstream infection[5]; and two haploid emerging pathogens, *C. guilliermondii* and *C. lusitaniae*. We compared these with the previously sequenced *C. albicans* strain (SC5314)[6–8], *Debaryomyces hansenii*[9], a marine yeast rarely associated with disease, and nine species from the related *Saccharomyces* clade (Fig. 1). These species span a wide evolutionary range and show large phenotypic differences in pathogenicity and mating, allowing us to study the genomic basis for these traits.

Genome sequence and comparative annotation

We found enormous variation in genome size and composition between the *Candida* genomes sequenced (Table 1). Each genome assembly displayed high continuity, ranging from nine to 27 scaffolds

[1]UCD School of Biomolecular and Biomedical Science, Conway Institute, University College Dublin, Belfield, Dublin 4, Ireland. [2]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, Massachusetts 02139, USA. [3]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. [4]Department of Biology and CESAM, University of Aveiro, 3810-193 Aveiro, Portugal. [5]School of Medical Sciences, Institute of Medical Sciences, University of Aberdeen, Foresterhill, Aberdeen AB25 2ZD, UK. [6]Bioinformatics Program, Boston University, Boston, Massachusetts 02215, USA. [7]Department of Genetics, Cell Biology and Development, University of Minnesota, Minneapolis, Minnesota 55455, USA. [8]Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, North Carolina 27710, USA. [9]Centre for Bioinformatics, Imperial College London, Wolfson Building, South Kensington, London SW7 2AY, UK. [10]Department of Genetics, Stanford University Medical School Stanford, California 94305-5120, USA. [11]School of Biosciences, University of Exeter, Exeter EX4 4QD, UK. [12]Department of Microbial Pathogenicity Mechanisms, Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll Institute, D-07745 Jena, Germany. [13]Swammerdam Institute for Life Sciences, University of Amsterdam, 1090 GB Amsterdam, The Netherlands. [14]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK. [15]Department of Pathobiology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61802, USA. [16]Department of Biochemistry and Cell Biology, SUNY Stony Brook, Stony Brook, New York 11794, USA. [17]Institute for Cell and Molecular Biosciences, Newcastle University, Newcastle upon Tyne NE2 4HH, UK. [18]Biostatistics and Bioinformatics Group, Masonic Cancer Center, University of Minnesota, Minneapolis, Minnesota 55455, USA. [19]Department of Biology, The University of Iowa, Iowa City, Iowa 52242, USA. [20]Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield S10 2TN, UK. [21]Department of Microbiology and Molecular Genetics, The University of Texas Health Science Center at Houston, Houston, Texas 77030, USA. †Present address: 454 Life Sciences, 20 Commercial Street, Branford, Connecticut 06405, USA.
*These authors contributed equally to this work.

©2009 Macmillan Publishers Limited. All rights reserved

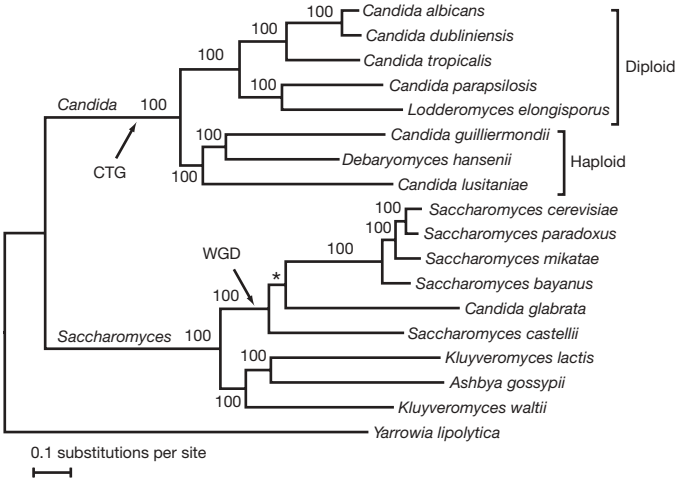


Figure 1 | Phylogeny of sequenced *Candida* and *Saccharomyces* clade species. Tree topology and branch lengths were inferred with MRBAYES (Supplementary Information, section S5). Posterior probabilities are indicated for each branch. The asterisk marks a branch that was constrained on the basis of syntenic conservation[40].

(Supplementary Table 1). Scaffold number and size largely match pulsed-field gel electrophoresis estimates for all genomes, and telomeric repeat arrays are linked to the ends of nearly all large scaffolds (Supplementary Information, section S2). Genome size ranges from 10.6 to 15.5 megabases (Mb), a difference of nearly 50%, with haploid species having smaller genomes. GC content ranges from 33% to 45% (Table 1). Transposable elements and other repetitive sequences vary in number and type between assemblies (Supplementary Information, section S6). Regions similar to the major repeat sequence (MRS) of *C. albicans* were found only in *C. tropicalis*, suggesting that MRS-associated recombination could contribute to the observed karyotypic variation among *C. tropicalis* strains[10].

Despite the genome size and phenotypic variation among the species, the predicted numbers of protein-coding genes are very similar, ranging from 5,733 to 6,318 (Table 1). Even the small differences in gene number are not correlated with genome size; the smallest genome, *C. guilliermondii*, has more genes than the largest genome, *L. elongisporus*. Instead, genome size differences are explained by an approximately threefold variation in intergenic spacing (Table 1). Large syntenic blocks of conserved gene order were detected among the four diploid species and between two of the haploid species, *C. guilliermondii* and *D. hansenii* (Supplementary Fig. 5). Although synteny blocks have been shuffled by local inversions and rearrangements, these have been primarily intrachromosomal as chromosome boundaries have been largely preserved across the diploid genomes (Supplementary Fig. 5).

Given the high conservation of protein-coding genes across the *Candida* clade, we used multiple alignments of the related genomes to revise the annotation of *C. albicans*. We identified 91 new or updated genes, of which 80% are specific to the *Candida* clade (Supplementary Information, section S4). We also corrected existing annotations in *C. albicans*, revealing 222 dubious genes, and also identified 190 probable frame shifts and 36 nonsense sequencing errors in otherwise well-conserved genes (Supplementary Information, section S4). In each case, manual curation confirmed ~80% of these predictions.

Polymorphism in diploid genomes

To gain insights into the recent history of *C. albicans*, we compared the two diploid strains, SC5314 and WO-1, which belong to different population subgroups[11]. Variation in the karyotype of these strains is primarily due to translocations at MRS sequences (ref. 12 and Supplementary Fig. 1). The two assemblies are largely co-linear with 12 inversions of 5–94 kilobases between them, except that in WO-1 some non-homologous chromosomes have recombined at the MRS (Supplementary Information, section S8). We found similar rates of single nucleotide polymorphisms (SNPs) within each strain (one SNP per 330–390 bases), and twice this rate between them, suggesting relatively recent divergence. Polymorphism rates in the other diploids range from one SNP per 222 bases in *L. elongisporus* and one SNP per 576 bases in *C. tropicalis* to a particularly low one SNP per 15,553 bases in *C. parapsilosis*, more than 70-fold lower than in the closely related *L. elongisporus*.

Notable regions of extended homozygosity are found in three of the four diploid genomes, which may reflect break-induced replication or recent passage through a parasexual or sexual cycle. *Candida albicans*, *C. tropicalis* and *L. elongisporus* each shows large chromosomal regions devoid of SNPs, extending up to ~1.2 Mb (Fig. 2 and Supplementary Figs 6–8). In contrast, the few SNPs in *C. parapsilosis* are randomly distributed across the genome (Supplementary Fig. 9a). A total of 4.3 Mb (30%) of the WO-1 assembly is homozygous for SNPs, approximately twice that found in SC5314 (Supplementary Information, section 7, and refs 7, 13, 14). There is at least one homozygous region per chromosome, none of which spans the predicted centromeres and only one of which starts at a MRS (Fig. 2). Whereas nearly all homogeneous regions are present at diploid levels and are therefore homozygous, WO-1 has lost one copy of a >300-kilobase region on chromosome 3 comprising nearly 200 genes (Supplementary Fig. 10). The pressure to maintain this region as homozygous in both strains is apparently high, as it is diploid but homogenous in SC5314.

Usage and evolution of CUG codons

All *Candida* clade species translate CUG codons as serine instead of leucine[15]. This genetic-code change altered the decoding rules of CUN codons in the *Candida* clade: whereas *Saccharomyces cerevisiae* uses two transfer RNAs, each of which translates two codons, *Candida* species use a dedicated $tRNA_{CAG}^{Ser}$ for CUG codons and a single $tRNA_{IAG}^{Leu}$ for CUA, CUC, and CUU codons, as inosine can base-pair with A, C and U (Fig. 3). This alteration in decoding rules forced the reduced usage of CUG—and also CUA, probably as a result of the weaker wobble—in *Candida* genes (Fig. 3a). CUU and CUC codons do not display the same bias for infrequent usage

Table 1 | *Candida* genome features

Species*	Genome size (Mb)	GC content (%)	No. of genes	Ave. gene size (bp)	Intergenic ave. (bp)	Ploidy	Pathogen†
C. albicans WO-1	14.4	33.5	6,159	1,444	921	diploid	++
C. albicans SC5314	14.3	33.5	6,107	1,468	858	diploid	++
C. tropicalis	14.5	33.1	6,258	1,454	902	diploid	++
C. parapsilosis	13.1	38.7	5,733	1,533	752	diploid	++
L. elongisporus	15.4	37.0	5,802	1,530	1,174	diploid	−
C. guilliermondii	10.6	43.8	5,920	1,402	426	haploid	+
C. lusitaniae	12.1	44.5	5,941	1,382	770	haploid	+
D. hansenii	12.2	36.3	6,318	1,382	550	haploid	−

bp, base pair.
* *C. albicans* SC5314 assembly 21 and gene set dated 28 January 2008 downloaded from the *Candida* Genome Database; *D. hansenii* assembly from GenBank[9]. The remaining assemblies are reported as part of this work, and are available in GenBank and at the Broad Institute *Candida* Database website.
† Relative level of pathogen strength: ++, strong pathogen; +, moderate pathogen; −, rare pathogen.

©2009 Macmillan Publishers Limited. All rights reserved

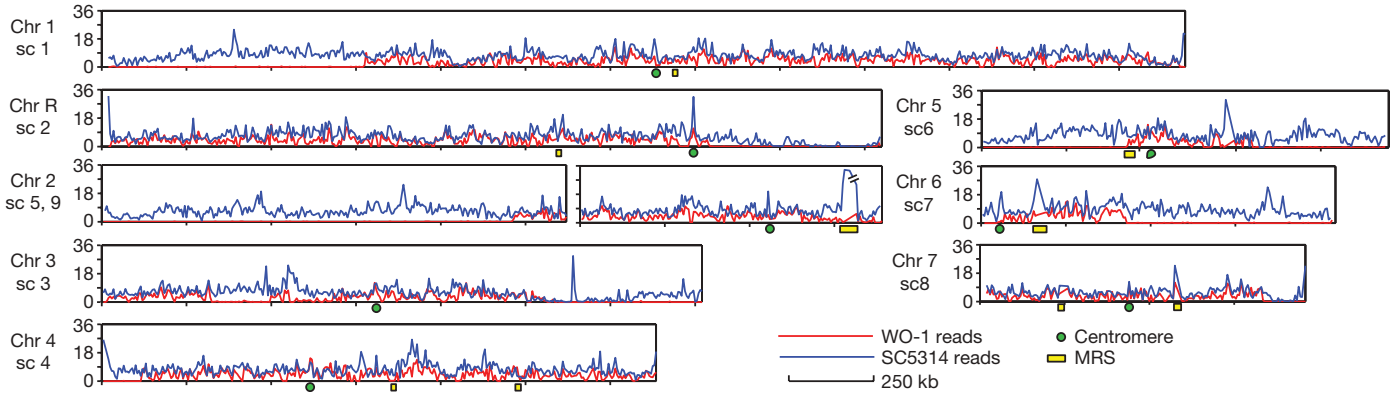


Figure 2 | *C. albicans* WO-1 is highly homozygous. Red lines show SNPs per kilobase, normalized by coverage, within WO-1, and blue lines show SNPs per kilobase between WO-1 and SC5314. Although both copies of chromosome 5 have rearranged at the MRS (yellow box) in WO-1, we show this as a single chromosome to allow a haploid reference for polymorphism. Relative to SC5314, chromosomes 1, 4 and 6 have the opposite orientation (Supplementary Fig. 6). Chr, chromosome; sc, supercontig; kb, kilobase.

(Fig. 3b). An additional pressure influencing codon usage may be the GC content, as usage of leucine codons in *Candida* species is correlated with the percentage GC composition (Supplementary Table 11).

We also examined the evolutionary fate of ancestral CUG codons and the origin of new CUG codons (Supplementary Table 12). CUG codons in *C. albicans* almost never (1%) align opposite CUG codons in *S. cerevisiae*. Instead, CUG serine codons in *C. albicans* align primarily to *Saccharomyces* codons for serine (20%) and other hydrophilic residues (49%). CUG leucine codons in *S. cerevisiae* align primarily to leucine codons in *Candida* (50%) and to other hydrophobic-residue codons (30%). This suggests a complete functional replacement of CUG codons in *Candida*.

Gene family evolution

To identify gene families likely to be associated with *Candida* pathogenicity and virulence, we used a phylogenomic approach across seven *Candida* and nine *Saccharomyces* genomes (Supplementary Information, section S10). Among 9,209 gene families, we identified 21 that are significantly enriched in the more common pathogens (Table 2). These include families encoding lipases, oligopeptide transporters and adhesins, which are all known to be associated with pathogenicity[8], as well as poorly characterized families not previously associated with pathogenesis.

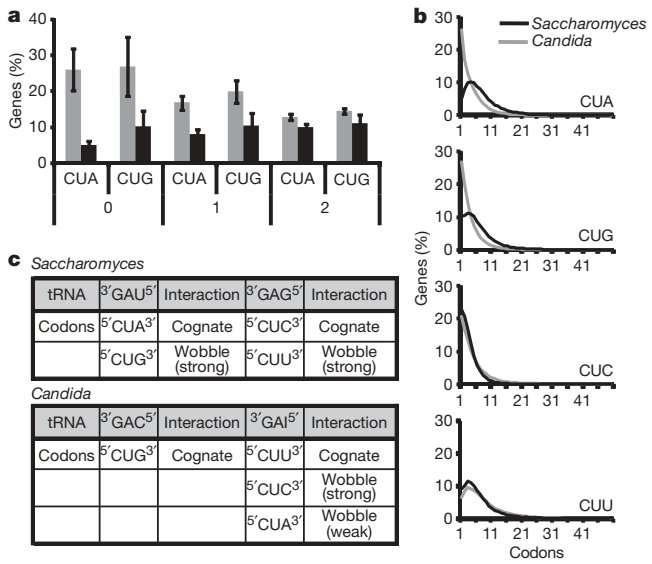


Figure 3 | Evolutionary effects of CUG coding. **a**, Average percentage of genes with zero, one or two CUG and CUA codons in *Candida* (grey bars) and *Saccharomyces* (black bars). Error bars, s.d. All differences are significant with $P \leq 0.0004$ (*t*-test, $N = 8$ *Candida* spp. and 6 *Saccharomyces* spp.) except for those in genes with two CUA codons ($P = 0.02$). '*Candida*' and '*Saccharomyces*' here refer to the CTG and WGD clades in Fig. 1, but including *Pichia stipitis* and excluding *C. dubliniensis*. **b**, CUN codon usage for all codon counts. **c**, Decoding rules for CUN codons in *Saccharomyces* and *Candida*.

Three cell wall families are enriched in the pathogens: those encoding Hyr/Iff proteins (ref. 16), Als adhesins[17] and Pga30-like proteins (Table 2 and Supplementary text S11). The Als family (family 17 in Supplementary Tables 22 and 23) in *C. albicans* is associated with virulence, and in particular with adhesion to host surfaces[18], invasion of host cells[19] and iron acquisition[20]. All these families are absent from the *Saccharomyces* clade species and are particularly enriched in the more pathogenic species (Supplementary Table 18). All three families are highly enriched for gene duplications (Supplementary Table 19), including tandem clusters of two to six genes, and show high mutation rates (fastest 5% of families) (Supplementary Information, section S10d). This variable repertoire of cell wall proteins is likely to be of profound importance to the niche adaptations and relative virulence of these organisms.

Als[17] and Hyr/Iff genes frequently contain intragenic tandem repeats, which modulate adhesion and biofilm formation in *S. cerevisiae*[21] (Fig. 4 and Supplementary Figs 19 and 20). The sequence of intragenic tandem repeats is conserved at the protein level across species (Supplementary Figs 19 and 20). Two proteins contain both an Als domain and repeats characteristic of the Hyr/Iff family.

Candida clade pathogens show expansions of extracellular enzyme and transmembrane transporter families (Table 2 and Supplementary Table 22). These families are either not found in *Saccharomyces* (including amino-acid permeases, lipases and superoxide dismutases) or are present in *S. cerevisiae* but significantly expanded in pathogens (including phospholipase B, ferric reductases, sphingomyelin phosphodiesterases and GPI-anchored yapsin proteases, which have been linked to virulence in *C. glabrata*[22]). Several groups of cell-surface transporters are also enriched (including oligopeptide transporters, amino-acid permeases and the major facilitator superfamily). Overall, these family expansions illustrate the importance of extracellular activities in virulence and pathogenicity. Genes involved in stress response are also variable between species (Supplementary Information, section S12).

C. albicans also showed species-specific expansion of some families, including two associated with filamentous growth, a leucine-rich repeat family and the Fgr6-1 family (Table 2). *Candida albicans* forms hyphae whereas *C. tropicalis* and *C. parapsilosis* produce only pseudohyphae, so these families may contribute to differences in hyphal growth.

©2009 Macmillan Publishers Limited. All rights reserved

Table 2 | Gene families enriched in pathogenic *Candida* spp.

No.	Annotation	Pathogen genes	Non-pathogen genes	*P* val.	Dup.	Loss	Gene rate	*C. alb.*	*C. tro.*	*C. par.*	*L. elo.*	*C. gui.*	*C. lus.*	*D. han.*	*C. gla.*	Yeast (ave.)
1	GPI family 18 (Hyr/Iff-like)	56	10	1.4×10^{-16}	52	11	16.2	11	18	17	9	3	7	1	0	0.0
2	Leucine-rich repeat (Ifa/Fgr38-like)	34	0	4.2×10^{-16}	32	5	18.3	33	1	0	0	0	0	0	0	0.0
3	Ferric reductase family	45	10	1.9×10^{-12}	30	25	2.5	12	19	7	7	3	4	2	0	0.1
4	Reductase family	43	11	3.2×10^{-11}	31	30	2.3	7	12	9	6	13	2	4	0	0.1
5	GPI family 17 (Als-like adhesins)	31	5	4.4×10^{-10}	29	4	20.5	8	16	5	4	2	0	1	0	0.0
6	GPI family 13 (Pga30-like)	34	7	5.0×10^{-10}	25	5	14.8	12	14	6	6	1	1	1	0	0.0
7	Unclassified	20	0	9.0×10^{-10}	13	3	15.9	9	9	0	0	2	0	0	0	0.0
8	Cell wall mannoprotein biosynthesis	38	18	7.2×10^{-7}	19	34	2.1	8	7	8	8	11	4	9	0	0.1
9	Major facilitator transporters	25	7	9.2×10^{-7}	14	17	2.0	3	3	7	3	10	2	4	0	0.0
10	Oligopeptide transporters	31	13	2.2×10^{-6}	23	11	6.7	6	9	9	4	4	3	1	0	0.9
11	Unclassified	25	9	6.3×10^{-6}	15	6	11.1	7	9	3	5	3	1	4	2	0.2
12	Amino-acid permeases	27	11	7.7×10^{-6}	11	18	1.7	6	6	6	4	6	3	6	0	0.1
13	Sphingomyelin phosphodiesterases	18	5	3.2×10^{-5}	11	9	7.4	4	5	4	2	3	2	1	0	0.2
14	Fgr6 family (filamentous growth)	12	1	3.3×10^{-5}	7	1	14.5	8	1	1	0	1	1	1	0	0.0
15	Secreted lipases	20	7	4.6×10^{-5}	17	8	9.6	10	5	4	4	1	0	3	0	0.0
16	Cytochrome P450 family	34	21	5.5×10^{-5}	23	22	6.0	6	8	10	7	5	4	6	1	1.0
17	Amino-acid permeases	16	4	5.6×10^{-5}	14	10	1.5	2	3	6	3	2	3	1	0	0.0
18	Zinc-finger transcription factors	31	18	6.2×10^{-5}	17	14	12.3	5	8	7	7	7	4	11	0	0.0
19	Unclassified	13	2	6.3×10^{-5}	8	0	8.1	3	1	6	1	2	1	1	0	0.0
20	Predicted transmembrane family	17	5	7.2×10^{-5}	9	2	7.5	4	4	5	3	3	1	2	0	0.0
21	Unclassified secreted family	20	8	1.1×10^{-4}	7	6	9.3	4	4	6	4	4	2	4	0	0.0

Pathogen genes: total genes in family for *C. albicans*, *C. tropicalis*, *C. parapsilosis*, *C. guilliermondii*, *C. lusitaniae* and *C. glabrata*. Non-pathogen genes: total genes in family for *L. elongisporus*, *D. hansenii* and all *Saccharomyces* clade species (Fig. 1) except *C. glabrata*. *P* val., *P* value of the hypergeometric test; all families shown above have a false discovery rate of less than 0.05 (Supplementary Information, section S10c). Dup., duplications; Loss, losses (Supplementary Information, section S10b). Gene rate: average mutation rate for each family (Supplementary Information, section S10d); the average gene rate across all families is 5.8. Yeast (ave.): average count for all *Saccharomyces* species. GPI, glycosyl phosphatidylinositol.

We identified 64 families showing positive selection in the highly pathogenic *Candida* species (Supplementary Table 32). These are highly enriched for cell wall, hyphal, pseudohyphal, filamentous growth and biofilm functions (Supplementary Information, section S13). Six of the families have been previously associated with pathogenesis, including that of *ERG3*, a C-5 sterol desaturase essential for ergosterol biosynthesis, for which mutations can cause drug resistance[23].

Structure of the *MTL* locus

Pathogenic fungi may have limited their sexual cycles to maximize their virulence[24], and the sequenced *Candida* species show tremendous diversity in their apparent abilities to mate. Among the four diploids, *C. albicans* has a parasexual cycle (mating of diploid cells followed by mitosis and chromosome loss instead of meiosis[25]), *L. elongisporus* has been described as sexual and homothallic (self-mating)[26], whereas *C. tropicalis* and *C. parapsilosis* have never been observed to mate. Among the three haploids, *C. guilliermondii* and *C. lusitaniae* are heterothallic (cross-mating only) and have a complete sexual cycle, whereas *D. hansenii* is haploid and homothallic[27] (Supplementary Information, section 14c).

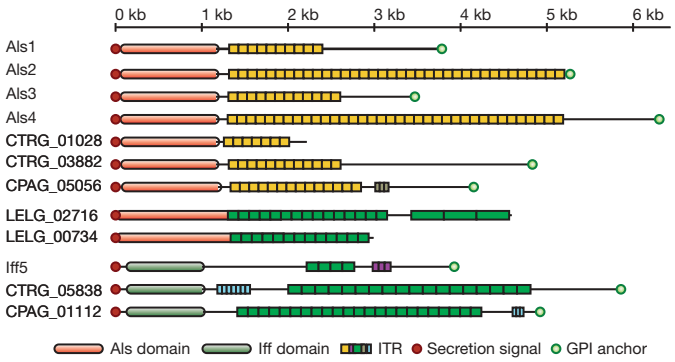


Figure 4 | Conserved domains of Als and Hyr/Iff cell wall families. The amino-terminal Als domain (red) and Hyr/Iff domain (green) are shown as ovals. Intragenic tandem repeats (ITRs; see Supplementary Information, section S11) are shown as rectangles, coloured to represent similar amino-acid sequences.

To understand the genomic basis for this diversity, we studied the *Candida MTL* locus, which determines mating type, similar to the *MAT* locus in *S. cerevisiae*. In both *C. albicans* and *S. cerevisiae*, the mating locus has two idiomorphs, **a** and α, encoding the regulators **a**1, α1 and α2, respectively. *Candida albicans MTL**a*** also encodes **a**2, and both idiomorphs in this species contain alleles of three additional genes without known roles in mating: *PAP*, *OBP* and *PIK*[28]. The *MTL*α and *MTL**a*** genes, alone or in combination, specify one of the three possible cell-type programs (**a** haploid, α haploid, **a**/α diploid). In *C. albicans*, the alpha-domain protein α1 activates α-specific mating genes, the high-mobility group (HMG) factor **a**2 activates **a**-specific mating genes, and the **a**1/α2 homeodomain heterodimer represses mating genes in **a**/α cells[29,30].

Despite extended conservation of the genomic context flanking the mating-type locus, there is great variability in *MTL* gene content (Fig. 5). *MTL**a**1* has become a pseudogene in *C. parapsilosis*[31], and is probably a recent loss because target genes retain predicted **a**1/α2 binding sites (Supplementary Information, section S14). *MTL*α*2* is missing in both *C. guilliermondii* and *C. lusitaniae* (J. L. Reedy, A. Floyd and J. Heitman, submitted). A fused mating-type locus containing both **a** and α genes is found in *D. hansenii* and *Pichia stipitis*[32,33].

Most surprisingly, all four mating-type genes are missing in *L. elongisporus*. It contains a site syntenic to *MTL**a*** in other species, but this contains only 508 base pairs of apparently non-coding DNA,

©2009 Macmillan Publishers Limited. All rights reserved

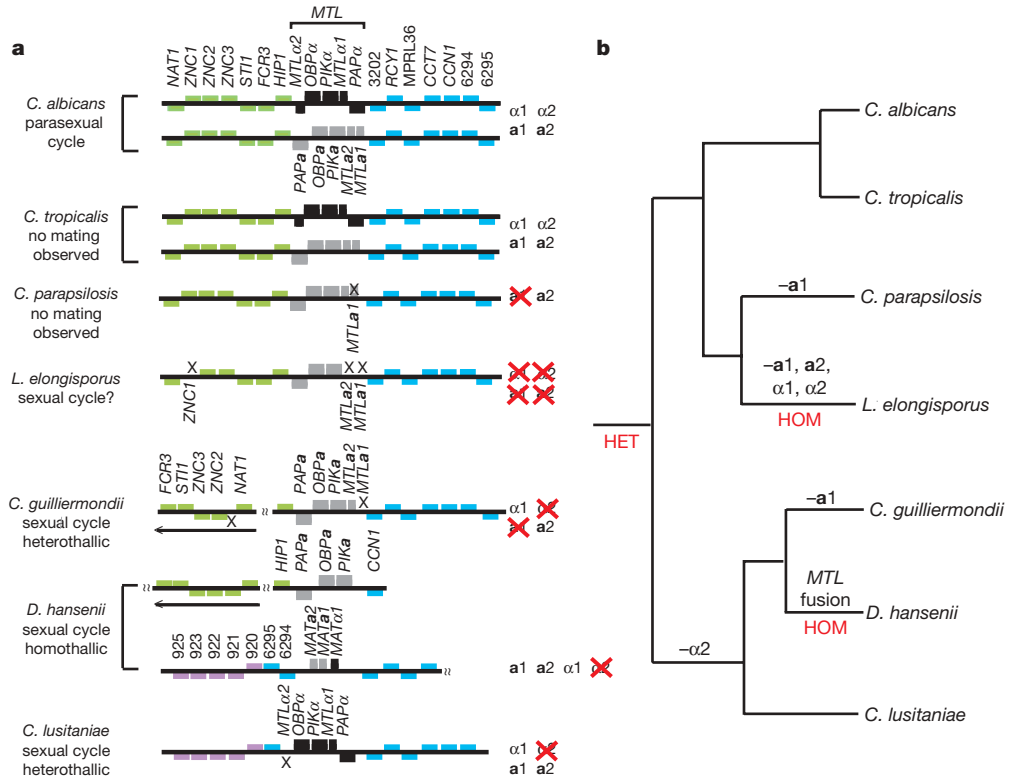


Figure 5 | Organization of *MTL* loci in the *Candida* clade. **a**, *MTLa*-specific genes are shown in grey, *MTLα*-specific genes are shown in black and other orthologues are shown in colour. Two idiomorphs from *C. albicans* and *C. tropicalis* are shown. Arrows indicate inversions relative to *C. albicans*. Crosses show gene losses and the *C. parapsilosis MTLa1* pseudogene. There is no *MTL* locus in *L. elongisporus*; gene order around the *PAPa*, *OBPa* and *PIKa* genes is shown. For *C. guilliermondii* and *C. lusitaniae*, the genome project sequenced one idiomorph; the second was obtained by J. Reedy. In *D. hansenii*, *PAP*, *OBP* and *PIK* are separate from the fused *MTL* locus[32]. **b**, Placement of gene losses on the phylogenetic tree. HOM, homothallic; HET, heterothallic.

a length insufficient to encode **a**1 or **a**2 even if they were extensively divergent in sequence. We confirmed this finding in seven other *L. elongisporus* isolates (not shown). The sexual state of *L. elongisporus* has been assumed to be homothallic because asci are generated from identical cells[5,26], but the absence of an **a**-factor pheromone, receptor and transporter (Supplementary Table 34) as well as *MTL*, suggests that it may not have a sexual cycle. Alternatively, mating may require only one pheromone and receptor (α), and *L. elongisporus* may be the first identified ascomycete that can mate independently of *MTL* or *MAT*.

Our discovery that *MTLα2* and *MTLa1* are frequently absent challenges our understanding of how the mating locus operates. The model derived for mating regulation in *C. albicans*, including the role of **a**1/α2 in white–opaque switching[3], must differ substantially in the other sexual species. This is particularly interesting because the major regulators of the white–opaque switch in *C. albicans* (*WOR1*, *WOR2*, *CZF1* and *EFG1*[34]) are generally conserved in other species, but Wor1 control over white–opaque genes appears to be a recent innovation in *C. albicans* and *Candida dubliniensis*[35]. Our data also suggest that the loss of α2 occurred early in the haploid sexual lineage.

Mating and meiosis

To gain further insight into their diversity in sexual behaviour, we examined whether 227 genes required for meiosis in *S. cerevisiae* and other fungi have orthologues in the *Candida* species (Supplementary Information, section S14). A previous report[36] that some components of meiosis, such as the major regulator *IME1*, are missing from *C. albicans* led us to propose that their loss could be correlated with lack of meiosis. Surprisingly, however, we find that these genes are missing in all *Candida* species, suggesting that sexual *Candida* species undergo meiosis without them. Conversely, even seemingly non-mating species showed highly conserved pheromone response pathways, suggesting that pheromone signalling plays an alternative role such as regulation of biofilm formation[37]. These findings suggest considerable plasticity and innovation of meiotic pathways in *Candida*.

Moreover, we find that sexual *Candida* species have undergone a recent dramatic change in the pathways involved in meiotic recombination, with loss of the Dmc1-dependent pathway in the heterothallic species *C. lusitaniae* and *C. guilliermondii* (Supplementary Information, section 14c). We also found that mechanisms of chromosome pairing and crossover formation have changed recently in these two species, because they (and to a lesser extent *D. hansenii*) have lost several components of the synaptonemal and synapsis initiation complexes (Supplementary Table 35). They have also lost components of the major crossover-formation pathway in *S. cerevisiae* (*MSH4*, *MSH5*), but have retained a minor pathway (*MUS81*, *MMS4*)[38,39]. Overall, if *Candida* species undergo meiosis it is with reduced machinery, or different machinery, suggesting that unrecognized meiotic cycles may exist in many species, and that the model of meiosis developed in *S. cerevisiae* varies significantly, even among yeasts.

The genome sequences reported here provide a resource that will allow current knowledge of *C. albicans* biology, the product of decades of research, to be applied with maximum effect to the other pathogenic species in the *Candida* clade. They also allow many of the unusual features of *C. albicans*—such as cell wall gene family amplifications, and its apparent ability to undergo mating and a parasexual cycle without meiosis—to be understood in an evolutionary context that shows that the genes involved in virulence and mating have highly dynamic rates of turnover and loss.

METHODS SUMMARY

The methods for this paper are described in Supplementary Information. Here we outline the resources generated by this project.

©2009 Macmillan Publishers Limited. All rights reserved

Assemblies, gene sets, and single nucleotide polymorphisms are available in GenBank and at the Broad Institute *Candida* Database website (http://www.broad.mit.edu/annotation/genome/candida_group/MultiHome.html). The Broad Institute website provides search, visualization, BLAST and download of assemblies and gene sets. Gene families can be accessed by searching either for individual genes or with family identifiers (CF#####). The *C. parapsilosis* assembly is also available at the Wellcome Trust Sanger Institute website (http://www.sanger.ac.uk/sequencing/Candida/parapsilosis/). The revised annotation of *C. albicans* (SC5314) is available at the *Candida* Genome Database (www.candidagenome.org).

Received 22 February; accepted 15 April 2009.
Published online 24 May 2009.

1. Pfaller, M. A. & Diekema, D. J. Epidemiology of invasive candidiasis: a persistent public health problem. *Clin. Microbiol. Rev.* **20,** 133–163 (2007).
2. Santos, M. A. & Tuite, M. F. The CUG codon is decoded *in vivo* as serine and not leucine in *Candida albicans*. *Nucleic Acids Res.* **23,** 1481–1486 (1995).
3. Lockhart, S. R. *et al.* In *Candida albicans*, white-opaque switchers are homozygous for mating type. *Genetics* **162,** 737–745 (2002).
4. Slutsky, B., Buffo, J. & Soll, D. R. High-frequency switching of colony morphology in *Candida albicans*. *Science* **230,** 666–669 (1985).
5. Lockhart, S. R., Messer, S. A., Pfaller, M. A. & Diekema, D. J. *Lodderomyces elongisporus* masquerading as *Candida parapsilosis* as a cause of bloodstream infections. *J. Clin. Microbiol.* **46,** 374–376 (2008).
6. Braun, B. R. *et al.* A human-curated annotation of the *Candida albicans* genome. *PLoS Genet.* **1,** e1 (2005).
7. Jones, T. *et al.* The diploid genome sequence of *Candida albicans*. *Proc. Natl Acad. Sci. USA* **101,** 7329–7334 (2004).
8. van het Hoog, M, *et al.* Assembly of the *Candida albicans* genome into sixteen supercontigs aligned on the eight chromosomes. *Genome Biol.* **8,** R52 (2007).
9. Dujon, B. *et al.* Genome evolution in yeasts. *Nature* **430,** 35–44 (2004).
10. Zhang, J., Hollis, R. J. & Pfaller, M. A. Variations in DNA subtype and antifungal susceptibility among clinical isolates of *Candida tropicalis*. *Diagn. Microbiol. Infect. Dis.* **27,** 63–67 (1997).
11. Tavanti, A. *et al.* Population structure and properties of *Candida albicans*, as determined by multilocus sequence typing. *J. Clin. Microbiol.* **43,** 5601–5613 (2005).
12. Chu, W. S., Magee, B. B. & Magee, P. T. Construction of an SfiI macrorestriction map of the *Candida albicans* genome. *J. Bacteriol.* **175,** 6637–6651 (1993).
13. Forche, A., Magee, P. T., Magee, B. B. & May, G. Genome-wide single-nucleotide polymorphism map for *Candida albicans*. *Eukaryot. Cell* **3,** 705–714 (2004).
14. Legrand, M. *et al.* Haplotype mapping of a diploid non-meiotic organism using existing and induced aneuploidies. *PLoS Genet.* **4,** e1 (2008).
15. Massey, S. E. *et al.* Comparative evolutionary genomics unveils the molecular mechanism of reassignment of the CTG codon in *Candida* spp. *Genome Res.* **13,** 544–557 (2003).
16. Bates, S. *et al. Candida albicans* Iff11, a secreted protein required for cell wall structure and virulence. *Infect. Immun.* **75,** 2922–2928 (2007).
17. Hoyer, L. L., Green, C. B., Oh, S. H. & Zhao, X. Discovering the secrets of the *Candida albicans* agglutinin-like sequence (ALS) gene family–a sticky pursuit. *Med. Mycol.* **46,** 1–15 (2008).
18. Yeater, K. M. *et al.* Temporal analysis of *Candida albicans* gene expression during biofilm development. *Microbiology* **153,** 2373–2385 (2007).
19. Phan, Q. T. *et al.* Als3 is a *Candida albicans* invasin that binds to cadherins and induces endocytosis by host cells. *PLoS Biol.* **5,** e64 (2007).
20. Almeida, R. S. *et al.* The hyphal-associated adhesin and invasin Als3 of *Candida albicans* mediates iron acquisition from host ferritin. *PLoS Pathog.* **4,** e1000217 (2008).
21. Verstrepen, K. J., Jansen, A., Lewitter, F. & Fink, G. R. Intragenic tandem repeats generate functional variability. *Nature Genet.* **37,** 986–990 (2005).
22. Kaur, R., Ma, B. & Cormack, B. P. A family of glycosylphosphatidylinositol-linked aspartyl proteases is required for virulence of *Candida glabrata*. *Proc. Natl Acad. Sci. USA* **104,** 7628–7633 (2007).
23. Chau, A. S. *et al.* Inactivation of sterol $\Delta^{5,6}$-desaturase attenuates virulence in *Candida albicans*. *Antimicrob. Agents Chemother.* **49,** 3646–3651 (2005).
24. Nielsen, K. & Heitman, J. Sex and virulence of human pathogenic fungi. *Adv. Genet.* **57,** 143–173 (2007).
25. Noble, S. M. & Johnson, A. D. Genetics of *Candida albicans*, a diploid human fungal pathogen. *Annu. Rev. Genet.* **41,** 193–211 (2007).
26. van der Walt, J. P. *Lodderomyces*, a new genus of the Saccharomycetaceae. *Antonie Van Leeuwenhoek* **32,** 1–5 (1966).
27. van der Walt, J. P., Taylor, M. B. & Liebenberg, N. V. D. W. Ploidy, ascus formation and recombination in *Torulaspora* (*Debaryomyces*) *hansenii*. *Antonie Van Leeuwenhoek* **43,** 205–218 (1977).
28. Hull, C. M., Raisner, R. M. & Johnson, A. D. Evidence for mating of the ''asexual'' yeast *Candida albicans* in a mammalian host. *Science* **289,** 307–310 (2000).
29. Tsong, A. E., Miller, M. G., Raisner, R. M. & Johnson, A. D. Evolution of a combinatorial transcriptional circuit: a case study in yeasts. *Cell* **115,** 389–399 (2003).
30. Tsong, A. E., Tuch, B. B., Li, H. & Johnson, A. D. Evolution of alternative transcriptional circuits with identical logic. *Nature* **443,** 415–420 (2006).
31. Logue, M. E., Wong, S., Wolfe, K. H. & Butler, G. A genome sequence survey shows that the pathogenic yeast *Candida parapsilosis* has a defective MTLa1 allele at its mating type locus. *Eukaryot. Cell* **4,** 1009–1017 (2005).
32. Fabre, E. *et al.* Comparative genomics in hemiascomycete yeasts: evolution of sex, silencing, and subtelomeres. *Mol. Biol. Evol.* **22,** 856–873 (2005).
33. Jeffries, T. W. *et al.* Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*. *Nature Biotechnol.* **25,** 319–326 (2007).
34. Zordan, R. E., Miller, M. G., Galgoczy, D. J., Tuch, B. B. & Johnson, A. D. Interlocking transcriptional feedback loops control white-opaque switching in *Candida albicans*. *PLoS Biol.* **5,** e256 (2007).
35. Tuch, B. B., Galgoczy, D. J., Hernday, A. D., Li, H. & Johnson, A. D. The evolution of combinatorial gene regulation in fungi. *PLoS Biol.* **6,** e38 (2008).
36. Tzung, K. W. *et al.* Genomic evidence for a complete sexual cycle in *Candida albicans*. *Proc. Natl Acad. Sci. USA* **98,** 3249–3253 (2001).
37. Daniels, K. J., Srikantha, T., Lockhart, S. R., Pujol, C. & Soll, D. R. Opaque cells signal white cells to form biofilms in *Candida albicans*. *EMBO J.* **25,** 2240–2252 (2006).
38. Argueso, J. L., Wanat, J., Gemici, Z. & Alani, E. Competing crossover pathways act during meiosis in *Saccharomyces cerevisiae*. *Genetics* **168,** 1805–1816 (2004).
39. de los Santos, T. *et al.* The Mus81/Mms4 endonuclease acts independently of double-Holliday junction resolution to promote a distinct subset of crossovers during meiosis in budding yeast. *Genetics* **164,** 81–94 (2003).
40. Scannell, D. R., Byrne, K. P., Gordon, J. L., Wong, S. & Wolfe, K. H. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440,** 341–345 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the US National Human Genome Research Institute (NHGRI) for support under the Fungal Genome Initiative at the Broad Institute. We thank C. Kurtzman for providing the sequenced strains of *L. elongisporus*, *C. guilliermondii* and *C. lusitaniae*, M. Koehrsen for Broad Institute website support, D. Park for informatics support and K. Wolfe and A. Regev for comments on the manuscript. We acknowledge the contributions of the Broad Institute Sequencing Platform and A. Barron, L. Clark, C. Corton, D. Ormond, D. Saunders, K. Seeger and R. Squares from the Wellcome Trust Sanger Institute for the *C. parapsilosis* sequencing and assembly. N.A.R.G., A.J.P.B., M.B. and co-workers were supported by the Wellcome Trust; M.G., S.S., Q.Z., C.K., B.W.B. and C.A.C. were supported by NHGRI and the National Institute of Allergy and Infectious Disease, US National Institutes of Health (NIH), Department of Health and Human Services; G.B. and co-workers were supported by Science Foundation Ireland; J.B., A.F., J.H., A.M.N. and co-workers were supported by the NIH; and M.K., M.D.R. and M.F.L. by the NIH, the US National Science Foundation and the Sloan Foundation.

Author Information Assemblies reported here have been deposited in GenBank under the following project accession numbers: AAFO00000000 (*C. albicans* WO-1), AAFN00000000 (*C. tropicalis*), AAPO00000000 (*L. elongisporus*), AAFM00000000 (*C. guilliermondii*), AAFT00000000 (*C. lusitaniae*), CABE01000001–CABE01000024 (*C. parapsilosis*). Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. Correspondence and requests for materials should be addressed to C.A.C. (cuomo@broad.mit.edu), M.K. (manoli@mit.edu) or G.B. (geraldine.butler@ucd.ie).

©2009 Macmillan Publishers Limited. All rights reserved

The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes

Kim D. Pruitt,[1,9] Jennifer Harrow,[2] Rachel A. Harte,[3] Craig Wallin,[1] Mark Diekhans,[3] Donna R. Maglott,[1] Steve Searle,[2] Catherine M. Farrell,[1] Jane E. Loveland,[2] Barbara J. Ruef,[4] Elizabeth Hart,[2] Marie-Marthe Suner,[2] Melissa J. Landrum,[1] Bronwen Aken,[2] Sarah Ayling,[5] Robert Baertsch,[3] Julio Fernandez-Banet,[2] Joshua L. Cherry,[1] Val Curwen,[2] Michael DiCuccio,[1] Manolis Kellis,[6,7] Jennifer Lee,[1] Michael F. Lin,[6,7] Michael Schuster,[8] Andrew Shkeda,[1] Clara Amid,[4] Garth Brown,[1] Oksana Dukhanina,[1] Adam Frankish,[2] Jennifer Hart,[1] Bonnie L. Maidak,[1] Jonathan Mudge,[2] Michael R. Murphy,[1] Terence Murphy,[1] Jeena Rajan,[2] Bhanu Rajput,[1] Lillian D. Riddick,[1] Catherine Snow,[2] Charles Steward,[2] David Webb,[1] Janet A. Weber,[1] Laurens Wilming,[2] Wenyu Wu,[1] Ewan Birney,[8] David Haussler,[3] Tim Hubbard,[2] James Ostell,[1] Richard Durbin,[2] and David Lipman[1]

[1]National Center for Biotechnology Information, National Library of Medicine, Bethesda, Maryland 20894, USA; [2]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, United Kingdom; [3]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064, USA; [4]Zebrafish Information Network, University of Oregon, Eugene, Oregon 97403-5291, USA; [5]The University of Manchester, Faculty of Life Sciences, Manchester Interdisciplinary Biocentre, Manchester M1 7DN, United Kingdom; [6]Computer Science and Artificial Intelligence Laboratory, Institute of Technology, Cambridge, Massachusetts 02139, USA; [7]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02141, USA; [8]European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, United Kingdom

Effective use of the human and mouse genomes requires reliable identification of genes and their products. Although multiple public resources provide annotation, different methods are used that can result in similar but not identical representation of genes, transcripts, and proteins. The collaborative consensus coding sequence (CCDS) project tracks identical protein annotations on the reference mouse and human genomes with a stable identifier (CCDS ID), and ensures that they are consistently represented on the NCBI, Ensembl, and UCSC Genome Browsers. Importantly, the project coordinates on manually reviewing inconsistent protein annotations between sites, as well as annotations for which new evidence suggests a revision is needed, to progressively converge on a complete protein-coding set for the human and mouse reference genomes, while maintaining a high standard of reliability and biological accuracy. To date, the project has identified 20,159 human and 17,707 mouse consensus coding regions from 17,052 human and 16,893 mouse genes. Three evaluation methods indicate that the entries in the CCDS set are highly likely to represent real proteins, more so than annotations from contributing groups not included in CCDS. The CCDS database thus centralizes the function of identifying well-supported, identically-annotated, protein-coding regions.

[Supplemental material is available online at www.genome.org. Data sets and documentation are available in the CCDS database at http://www.ncbi.nlm.nih.gov/CCDS.]

One key goal of genome projects is to identify and accurately annotate all protein-coding genes. The resulting annotations add functional context to the sequence data and make it easier to traverse to other rich sources of gene and protein information. Accurately annotating known genes, identifying novel genes, and tracking annotations over time are complex processes that are best achieved through a combination of large-scale computational analyses and expert curation. These methods must (1) process repetitive sequences in multiple categories including retrotransposons, segmental duplications, and paralogs; (2) process variation including copy number variation (CNV) (Feuk et al. 2006) and microsatellites; (3) distinguish functional genes and alleles from pseudogenes; (4) define alternate splice products; and (5) avoid erroneous interpretation based on experimental error.

Genome annotation information is available from many sources including publications on the sequencing and annotation

[9]Corresponding author.
E-mail Pruitt@ncbi.nlm.nih.gov; fax (301) 480-2918.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.080531.108.

19:1316–1323; ISSN 1088-9051/09; www.genome.org

of genes for whole genomes, individual chromosomes, and whole-genome annotation computed by multiple bioinformatics groups. Ensembl and the National Center for Biotechnology Information (NCBI) independently developed computational processes to annotate vertebrate genomes (Kitts 2002; Potter et al. 2004). Both pipelines predict genes, transcripts, and proteins based on interpretations of gene prediction programs, transcript alignments, and protein alignments. In addition, manual annotation is provided by the Havana group at the Wellcome Trust Sanger Institute (WTSI) and the Reference Sequence (RefSeq) group at the National Center for Biotechnology Information (NCBI).

The abundance of different data sources has been problematic for the scientific community since annotated models may change over time as more experimental data accumulate, or may differ among annotation groups owing to differences in methodology or data used. Differences in presentation can also compound the problem. Assigning a unique, tracked accession (CCDS ID) to identical coding region annotations removes some of the uncertainty by explicitly noting where consensus protein annotation has been identified, independent of the website being used. "Consensus" is defined as protein-coding regions that agree at the start codon, stop codon, and splice junctions, and for which the prediction meets quality assurance benchmarks. Thus, a distinguishing feature of the collaborative consensus coding sequence (CCDS) project compared to other protein databases is the integrated tracking of protein sequences in the context of the genome sequence. The current CCDS data sets for human and mouse can be accessed from several public resources including the members of the CCDS collaboration, namely: (1) the Ensembl Genome Browser, which is a joint project between the European Bioinformatics Institute (EBI) and WTSI (Birney et al. 2004); (2) the NCBI Map Viewer (Dombrowski and Maglott 2002); (3) the University of California Santa Cruz (UCSC) Genome Browser (Karolchik et al. 2008; Zweig et al. 2008); and (4) the WTSI Vertebrate Genome Annotation (Vega) Genome Browser (Ashurst et al. 2005). The CCDS collaborators provide access to the same reference genomic sequence and CCDS data set.

The CCDS set is built by consensus; each member of the collaboration contributes annotation, quality assessments, and curation. The collaboration pragmatically defines the initial focus on coding region annotations, rather than the annotated transcripts including untranslated regions (UTRs), because it is critical to identify encoded proteins and because there is more variation in UTR annotation. Protein-coding region annotations that do not satisfy the criteria for assigning a CCDS ID are evaluated between releases so that the annotation continues to improve. The key goal of the CCDS project is to provide a complete set of high-quality annotations of protein-coding genes on the human and mouse genomes.

Results

We have developed the process flow, quality assurance tests, curation infrastructure, and web resources to support identification, tracking, and reporting of identical protein annotations. Table 1 summarizes the growth of the CCDS database since its first release in 2005.

Following a coordinated annotation update of the reference genome annotation, results are compared to identify identical protein-coding regions. Each coding sequence (CDS) annotation must then pass quality assessment tests before being assigned a CCDS ID and version number (see Methods; Supplemental Fig. 4). The CCDS ID is stable, and every effort is made to ensure that all protein-coding regions with existing CCDS IDs are consistently annotated with each whole-genome annotation update. The protein sequence defined for the CCDS ID is the predicted translation of the coding sequence that is annotated on the genomic reference chromosome. Thus it is identical to the sequence reported by Ensembl, UCSC, and Vega. The sequence may differ at individual amino acids from associated RefSeq records (Pruitt et al. 2009) because the latter are often based on translations of independently generated mRNA sequences that may be selected to represent a different allele.

Quality assessment

We assessed the content and quality of the current public CCDS collection using three metrics: (1) evaluation of NCBI HomoloGene clusters to determine the number of homologous pairs of human and mouse CCDS proteins, (2) comparison of CCDS proteins to the curated UniProtKb/SWISS-PROT protein data set (hereto after referred to as "SWISS-PROT"), and (3) evaluation of genome conservation. HomoloGene (http://www.ncbi.nlm.nih.gov/sites/entrez?db=homologene) (Wheeler et al. 2008) reports groups of related proteins annotated on reference chromosomes for select species and includes a consideration of local conserved synteny that is limited to an assessment of flanking genes. HomoloGene is a gene-oriented resource as calculation is based on

Table 1. Growth and current size of the CCDS set

NCBI build[a]	Ensembl release[a]	Existing CCDS	Reinstated[b]	New candidate	QA tests, rejection	QA tests, curation override	Withdrawn, curation	Withdrawn, other[c]	Total no. of CCDS	Total no. of genes
Hs35.1	23	0		16,068	1273	0	0	0	14,795	13,142
Hs36.2	39	14,795		4955	1033	5	243	189	18,290	16,008
Hs36.3	49	18,290	23	2808	692	40	281	29	20,159	17,052
Mm36.1	39	0		15,744	2370	0	0	0	13,374	13,014
Mm37.1	47	13,374		5716	1371	37	6	43	17,707	16,893

[a]NCBI build numbers and Ensembl release numbers (e.g., build 35.1 and release 23, etc.) are displayed in the Map Viewer and Ensembl browser, respectively, and represent distinct whole-genome annotation runs. The values reported here reflect the input annotation data set used to calculate new candidates. (Hs) *Homo sapiens*; (Mm) *Mus musculus*.
[b]If unexpected losses, indicated in the "Withdrawn, Other" column, are found in a later build, the CCDS ID is reinstated with a "public" status. Reinstatement requires that the CDS structure be identical to the version that was previously lost, or, if the CDS structure has changed and is found as identical in both input data sets, then the CCDS version number is incremented. For example, see CCDS2672.
[c]Unexpected loss of consistent CDS annotation includes changed or removed annotation that is not tracked by the CCDS database as curation-based change. The large accidental loss in human build 36.2 resulted in improved tracking of annotation input data by both the NCBI and Ensembl annotation pipelines. Robust CCDS tracking continues to be a goal of annotation pipelines.

a single longest protein per annotated locus and excludes additional annotated proteins that may be available from consideration (Sayers et al. 2009). There are 16,590 HomoloGene clusters that include proteins from both human and mouse; 15,963 of these have at least one protein with a CCDS ID, and 13,329 clusters include proteins with CCDS IDs from both human and mouse. When the latter subset is evaluated for length of protein product, 68% are within 30 amino acids and 25% are identical. Note that because there are fewer CCDS proteins in mouse, assessing the quality of the annotation by counts of HomoloGene clusters with both human and mouse CCDS members may underrepresent the conservation of annotations. Of the human and mouse protein-coding genes with an associated CCDS ID, 96.5% of the human genes and 95.4% of the mouse genes (16,461 and 16,114, respectively) are clustered by HomoloGene with at least one homolog from any other species (see Fig. 1).

A second approach to assess the CCDS data set is to compare the derived protein sequences to another curated protein data set, namely, the SWISS-PROT records available for human and mouse. Of the 35,505 human and mouse SWISS-PROT records available at the time of this analysis (release 55.5), 81% match a CCDS protein at or above 95% identity, of which 66% are identical. Similar numbers are found for the human and mouse CCDS data sets (see Fig. 2). A complete match between the CCDS data set and that of SWISS-PROT is not expected because these two resources are generated using very different data models; SWISS-PROT is a protein-oriented resource that doesn't require consistency with the reference genome sequence, whereas CCDS proteins do have that requirement. Lack of a SWISS-PROT match may indicate differential representation of alternate splice products, differences in the gene type designation (protein-coding vs. non-coding), known gaps in CCDS for genes with limited sequence data, and differences that can be correlated to the reference genome sequence including small sequence differences (mismatches, insertions, or deletions), assembly gaps, copy number variation, or alternate haplotypes. Since SWISS-PROT is centered on proteins, it includes records for which a direct correlation to a CCDS cannot be expected because the categories are not in scope for CCDS. Among those are entries for human endogenous retrovirus (HERV) proteins, mature immunoglobulin proteins resulting from genomic rearrangement events, small physiologically active peptides that may result from enzymatic reactions, and putative uncharacterized proteins that may be alternatively interpreted by another group as a pseudogene or non-protein-coding RNA.

The third assessment uses the reading frame conservation (RFC) methodology (Kellis et al. 2003) to compare the protein-coding evolutionary signature of genes in CCDS with those from the RefSeq and Ensembl collections that do not contain CCDS proteins (see Supplemental Table 2). The RFC score is the percentage of nucleotides whose reading frame is evolutionarily conserved across species. Previous work has shown that 98% of the well-studied human genes have RFC scores >90 (Clamp et al. 2007). Figure 3 shows the distribution of the RFC scores of the CCDS, RefSeq, and Ensembl loci for human and mouse, along with a control set of non-protein-coding DNA for human. In the human set, 95.9% of the CCDS genes have an RFC score above 90.

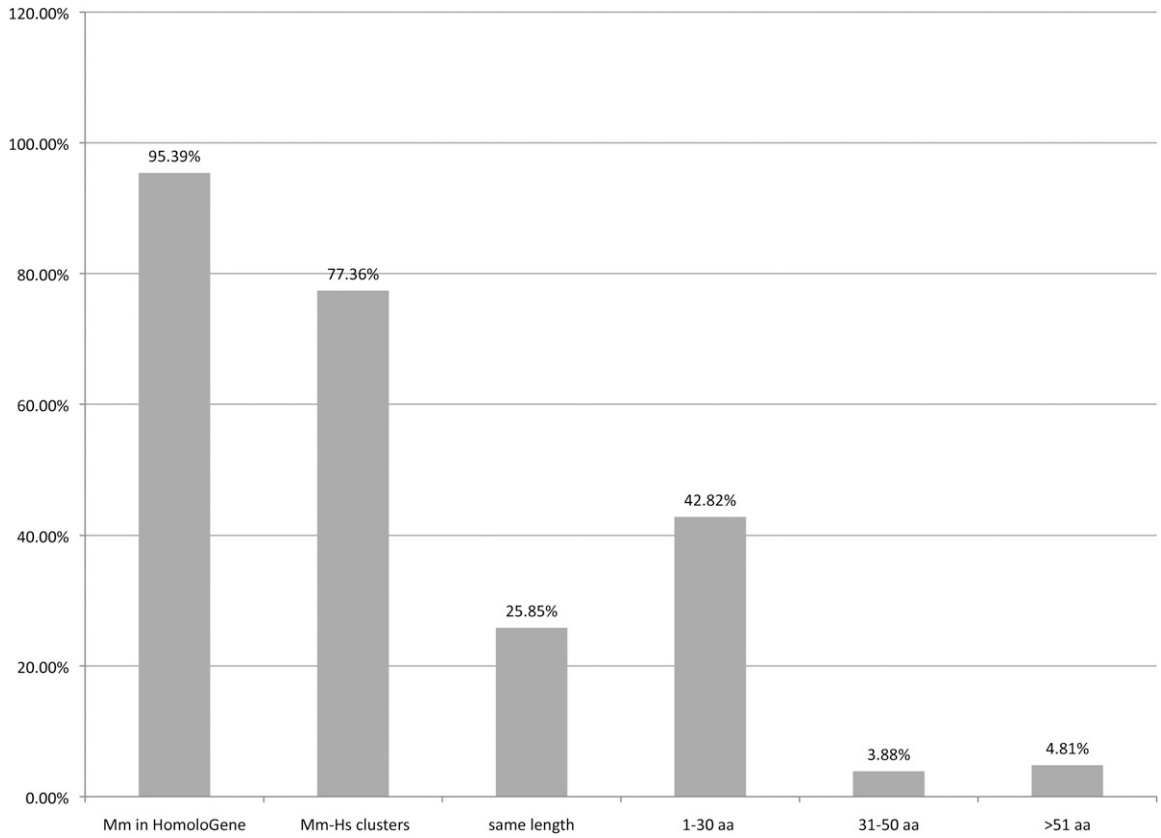


Figure 1. The percentage of mouse CCDS proteins that are found in any HomoloGene cluster versus those in a cluster that also contains a human CCDS protein (first two bars, respectively). For the latter category, results are further categorized based on protein length differences for the human and mouse homologous proteins.

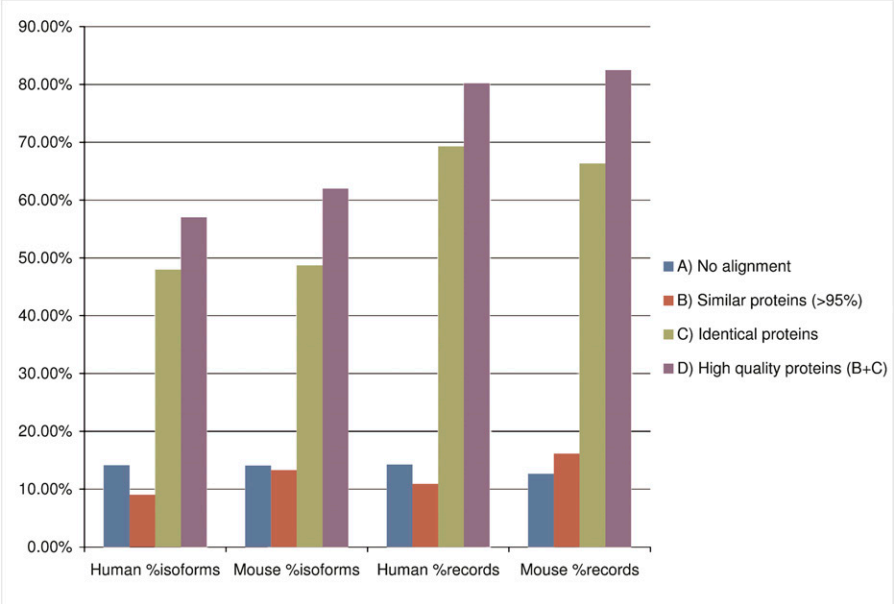


Figure 2. The percentage of human and mouse genes, with associated CCDS IDs for one or more proteins that are identical (*C*), similar (*B*) , or unique (*A*) when compared to SWISS-PROT records and to SWISS-PROT isoforms that were extracted from record annotation (see Methods). (*D*) The total number of high-quality matches.

In contrast, 37.6% and 44.3% of the non-CCDS RefSeq and Ensembl loci, respectively, have scores above RFC 90, and only 1.2% of the control set has RFC scores >90. In mouse, 93.5% of the CCDS genes exceed the RFC 90 threshold, compared to 36.8% of the RefSeq non-CCDS genes and 46.3% of the Ensembl non-CCDS genes scoring above RFC 90. In both human and mouse, the CCDS gene set shows significantly stronger evidence of evolving in a manner consistent with protein-coding genes than the RefSeq and Ensembl loci that do not contain CCDS proteins. The weak-scoring genes tend to be the genes that require careful review by annotators. For instance, 35% of human RefSeq genes with low RFC scores (<90) are single-exon genes, while 6% of those with a high RFC score (≥90) are single-exon genes. Genes with low RFC scores are also enriched for segmental duplications with 31% of low RFC RefSeq genes overlapping regions of segmental duplication (Bailey et al. 2001), compared to only 9% of the high RFC scoring genes being in segmental duplications. Similarly, 41% of the low RFC human RefSeqs are identified as originating by retrotransposition (Baertsch et al. 2008), while only 18% of high RFC ones are categorized as retro-copies.

Access to CCDS data

NCBI hosts a public website for the CCDS project (http://www.ncbi.nlm.nih.gov/CCDS/) that includes information about the collaboration, provides links to reports and FTP download, and provides a query interface to retrieve information about CCDS sequences and locations. The interface supports query by multiple identifiers including official gene symbols, CCDS ID, Entrez Gene ID (Maglott et al. 2007), or sequence ID (RefSeq, Ensembl ID, or Vega ID). Query results are presented in a table format with links provided to access the full report details for each CCDS ID (see Supplemental Fig. 1). Multiple CCDS IDs are reported for a gene if both data sets consistently annotated more than one CDS location. The CCDS ID-specific report page, shown in Figure 4, provides a detailed report for the CCDS ID. Some records also include a Public Note, provided by a curator, summarizing the rationale for an update, withdrawal, or to explain representation choices. Reports may include an update history for associated sequence records when relevant (Supplemental Fig. 2).

All collaborating groups indicate the CCDS ID on gene and/or protein report pages, and Genome Browsers indicate when a CCDS is available for a locus using either display style (coloration), text labels, or by providing as a data track (Supplemental Fig. 3). For those interested in downloading the full CCDS

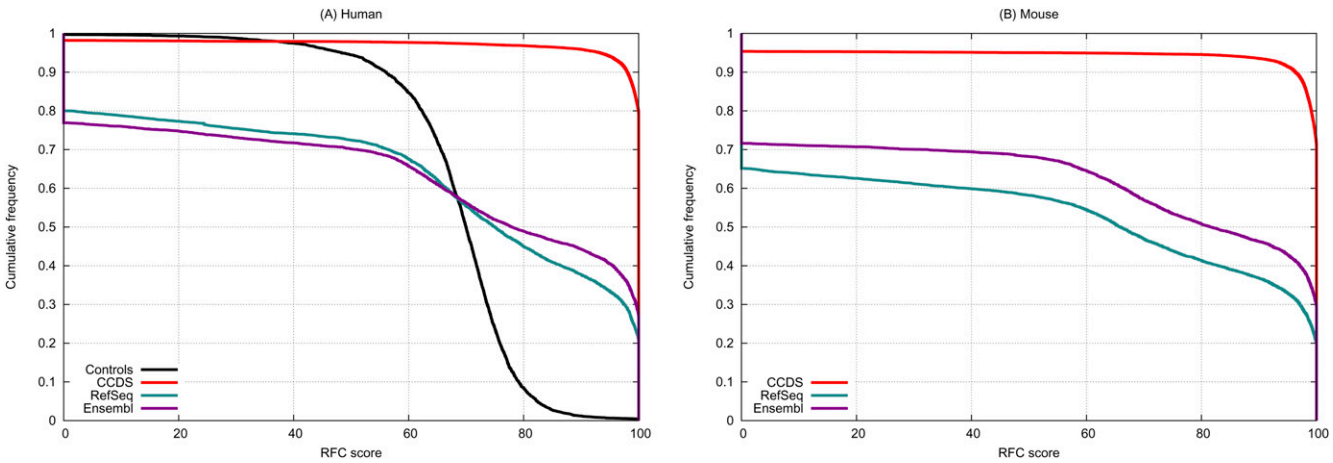


Figure 3. Cumulative distributions of RFC scores for human (*A*) and mouse (*B*). These graphs compare the RFC scores for CCDS loci with those of RefSeq and Ensembl loci that do not contain a CCDS protein, as well as a control data set for human. Since the controls were designed to have a similar alignment coverage to well-known genes, loci in other gene sets with less alignment coverage will score less than the controls.

Report for CCDS ID 7564

CCDS	Status	Species	Chrom.	Gene	NCBI Builds	Links
7564.2	Public	*Homo sapiens*	10	MXI1	35.1 - 36.3	H G G

Public Note

The coding region has been updated to include an alternative exon that is more supported by the available transcript and protein data. The update completes the helix-loop-helix dimerization domain, which was previously truncated.

Sequence IDs included in CCDS 7564.2

Original	Current	Source	Nucleotide ID	Protein ID	Status in CCDS	Seq. Status	Links
✔	✔	EBI,WTSI	ENST00000239007	ENSP00000239007	Accepted	alive	N P N P
✔	✔	EBI,WTSI	OTTHUMT00000050316	OTTHUMP00000020469	Accepted	alive	N P N P
✔	✔	NCBI	NM_005962.4	NP_005953.4	Accepted	alive	N P N P

Chromosomal Locations for CCDS 7564.2

On '+' strand of Chromosome 10 (NC_000010.9)

Genome Browser links: N U E V

Chromosome	Start	Stop	Links
10	111975959	111976031	N N U E V
10	111977937	111978069	N N U E V
10	111994576	111994605	N N U E V
10	112028928	112029042	N N U E V
10	112029662	112029833	N N U E V
10	112034572	112034735	N N U E V

CCDS Sequence Data

Blue highlighting indicates alternate exons.

Red highlighting indicates amino acids encoded across a splice junction.

Mouse over the nucleotide or protein sequence below and click on the highlighted codon or residue to select the pair.

Nucleotide Sequence (687 nt):

ATGGAGCGGGTGAAGATGATCAACGTGCAGCGTCTGCTGGAGGCTGCCGAGTTTTTGGAGCGCCGGGAGC
GAGAGTGTGAACATGGCTACGCCTCTTCATTCCCGTCCATGCCGAGCCCCCGACTGCAGCATTCAAAGCC

Figure 4. Detailed CCDS ID report page for a MXI1 protein. The CCDS report page presents three tables of information followed by nucleotide and protein sequences for the annotated CDS. The first table summarizes the status for the specified CCDS ID. Colored icons provide links to related resources, to a history report (orange H icon), or to re-query the CCDS database with a different type of identifier. (Red G icon re-queries the CCDS database by GeneID to return all CCDS IDs available for a gene.) A Public Note is provided for a subset of curated records to explain the nature of, and/or the reason for, an update or withdrawal. The sequence identifiers table reports sequences tracked as members of the CCDS ID. A checkmark in the first column (Original) identifies sequence identifiers represented on the annotated genome and included in the analyzed input data sets, and a checkmark in the second column (Current) identifies those that are considered current members (see Supplemental Fig. 2) The Chromosome Locations table reports the genomic coordinates of each exon of the CDS with links to view the annotation in different browsers—the violet icons (N) NCBI; (U) UCSC; (E) Ensembl; and (V) Vega. The nucleotide sequence of the CDS, derived from the genome sequence using the reported exon.

data set, the collaboration provides reports of the associated identifiers as well as the sequence data for anonymous FTP (ftp://ftp.ncbi.nih.gov/pub/CCDS/). Please refer to provided README files for descriptions of information provided and file formats.

Manual curation of the CCDS data set

Coordinated manual curation, a critical aspect of the CCDS project, is supported by a restricted-access website and a discussion e-mail list. The collaboration has generated standardized curation guidelines for selection of the initiation codon and interpretation of upstream ORFs and transcripts that are predicted to be candidates for nonsense-mediated decay (NMD) (Lejeune and Maquat 2005). Curation occurs continuously, and any of the collaborating centers can flag a CCDS ID as a potential update or withdrawal. Planned updates that are either under discussion or have achieved consensus agreement are indicated in the public CCDS website by a change in the status of the CCDS ID (Supplemental Table 1). Conflicting opinions are addressed by consulting with scientific experts or other annotation curation groups such as the HUGO Gene Nomenclature Committee (HGNC) (Bruford et al. 2008) and Mouse Genome Informatics (MGI) (Eppig et al. 2005). If a conflict cannot be resolved, then collaborators agree to withdraw the CCDS ID until more information becomes available.

To date, we have reviewed more than 14,000 CCDS proteins and confirmed the existing CDS annotation with no change. Review also resulted in the removal of 530 CCDS IDs and suggested updates to 1014 proteins. If a CCDS protein is updated, the CCDS ID version number is incremented, and often a note is provided explaining the update. For example, review of the protein annotation (CCDS ID 10689) for the human *SRCAP* gene resulted in an N-terminal extension that adds an HSA domain that is found in DNA-binding proteins and is often associated with helicases. The HSA domain is consistent with the other domains found in the protein and with the presumed function of this protein as a component of the SRCAP chromatin remodeling complex (Johnston et al. 1999; Wong et al. 2007). This significant improvement to the protein annotation was immediately available in the RefSeq transcript and protein sequences and in the Vega and UCSC Genome Browsers. This improvement became available in the NCBI and Ensembl Genome Browsers following a recalculation of the annotation for the human genome. Note there is no corresponding CCDS ID for the mouse *Srcap* locus yet, primarily owing to insufficient transcript data but confounded by the observation of differences in the exon definition compared to the human coding sequence.

Discussion

Although the human and mouse genome sequences have been of "finished" quality for several years, refinements to the annotation of protein-coding genes continue to take place. Until the

inception of the CCDS project, it was difficult to identify which protein annotations were represented consistently by the major browsers. The CCDS project solves this problem and establishes a framework to support well-supported, consistent, comprehensive annotation of the protein-coding content of the human and mouse genomes. The CCDS project has already identified 37,866 consistently annotated human and mouse CDSs for 33,945 genes and assigned them stable, versioned identifiers. By developing annotation standards, coordinating review of automated annotation, and documenting annotation decisions, the CCDS group continues to make a major contribution to the usability of the human and mouse genomic sequences.

The three independent methods used to assess the CCDS collection demonstrate that the genes included are highly likely to be protein-coding loci. Comparison to HomoloGene data indicates that at least 77% of the mouse genes with an associated CCDS ID have a homologous gene in human with an associated CCDS ID. Of these, the majority of the homologous CCDS proteins have a comparable length. Review of identified homologs with larger length differences indicates that the majority of them reflect valid differences due to alternate splicing. Comparison to another highly curated data set, SWISS-PROT, showed that 81% of the SWISS-PROT proteins are identical or highly similar to those encoded by CCDS, with similar results for mouse and human. The RFC analysis results indicate that 95% of the CCDS proteins do have an evolutionary signature that is consistent with their protein-coding designation.

The absence of a CCDS ID for a putative protein-coding gene annotation does not necessarily indicate that annotation is of poor quality; it indicates only that annotation is not yet consistent and requires additional review. Causes of annotation differences include resource-specific automatic annotation methods, timing of manual curation updates, conflicts between genomic and cDNA evidence, and incomplete curation guidelines on evaluating whether or not a locus is protein-coding, how much evidence is required to provide annotation, or where splice junctions should be annotated in repetitive regions. Until we have robust data from proteomic analyses, it is indeed a challenge to identify genes that are protein-coding, whether or not they are in the CCDS set. Supplemental Table 3 summarizes one approach to this problem, namely, classifying officially named genes thought to be protein-coding that have not yet been assigned a CCDS ID. Some have been assigned a RefSeq protein accession, many of which became available since the last CCDS analysis and are expected to gain a CCDS ID in the next analysis; some are associated with genome assembly issues that prevent representing the preferred protein from the genomic sequence; some are associated with protein sequence but are not in the RefSeq NM/NP set (often because the protein data appear to be partial); and some loci, perhaps historical, have no associated protein sequence at all. It is important to note that a CCDS ID represents consistency between annotation resources—it does not indicate that the annotation has been manually reviewed. We welcome feedback from the scientific community either regarding current annotations or to provide data and help with annotating new loci (see the CCDS website for contact information).

The benefits of the CCDS project extend beyond the CCDS data set currently available for the human and mouse genomes. The collaboration supporting the CCDS analysis process has resulted in improvements in automated annotation methods, quality assessment, and manual curation that are applied to many genomes. Discussions about evidence for annotation and publications between the RefSeq, Havana, and UCSC curation staff have resulted in re-evaluation of genomic sequence including assembly issues, correction of annotation errors, and identification of loci for which additional experimental validation is needed. Questions about the genomic sequence are reported to the Genome Reference Consortium (http://www.ncbi.nlm.nih.gov/genome/assembly/grc/index.shtml); annotation errors are resolved collaboratively ensuring consistent representation at all sites, and loci in need of experimental validation are reported to the GENCODE project. Experimental validation of transcripts and splice sites will occur as part of the GENCODE scale-up project (http://www.sanger.ac.uk/encode/), which builds on the successful GENCODE pilot project (Harrow et al. 2006). GENCODE is part of the extended human Encyclopedia of DNA Elements (ENCODE) project (The ENCODE Project Consortium 2007). Annotated transcripts highlighted for validation will be confirmed in an array of tissues using RT-PCR or RNAseq. The resulting sequence will be fed back into the CCDS project as supporting evidence.

The CCDS group is thus a key participant in improving the representation of the human and mouse genomes. For example, we are collaborating with the HGNC to match loci they have named to the human genomic sequence. Curation is also focused on human–mouse homologous proteins for which one lacks a CCDS ID, and protein-coding loci with associated SWISS-PROT proteins for which there is no corresponding CCDS ID. An additional long-term goal is to add attributes that indicate where transcript annotation is also identical (including the UTRs) and to indicate splice variants with different UTRs that have the same CCDS ID. It is also anticipated that as more complete and high-quality genome sequence data become available for other organisms, annotations from these organisms may be in scope for CCDS representation.

Methods

Identifying the candidate CCDS groups and tracking updates

Following the release of a re-annotation of the human or mouse reference genomes by both NCBI and Ensembl, we compared the genome annotation data sets provided by NCBI and Ensembl (Supplemental Fig. 4) to identify protein-coding annotations that are identical (start codon, stop codon, splice junctions) and do not include in-frame stop codons or apparent frameshifts. The full-length translation of proteins that include the amino acid selenocysteine (identified as the codon UGA) is provided when collaborators are consistent in annotating an internal stop codon. Identical annotations are subject to quality assessment tests and assessed to identify whether they correspond to existing CCDS IDs or are novel. Existing CCDS IDs are tracked using the combination of sequence identifiers and chromosomal coordinates. The version number is incremented if the annotated exon coordinates and predicted protein product have changes, in which case it is required that they have been identically updated in both annotation sources owing to coordinated curation (see below and Supplemental Table 1). Novel entries are assigned a unique CCDS ID with an initial version of 1. Although mechanisms are in place in the Ensembl and NCBI annotation pipelines to ensure that existing CCDS entries are stably incorporated in the whole genome re-annotation process, it is possible that final annotation of a CCDS protein may not be included or may no longer match, and so the CCDS entry is determined to be "lost" by the comparative process described above, in which case, they are flagged for manual follow-up.

The primary data represented by a CCDS ID are the chromosome coordinates of the annotated protein-coding exons and the nucleotide and conceptually translated protein sequence obtained from those coordinates. Ancillary data associated with the CCDS ID include sequence IDs and gene IDs included in the NCBI and Ensembl data sets. Each CCDS ID includes at least one protein identifier from both NCBI and Ensembl; a CCDS ID can include additional protein identifiers from either data set when they are predicted from transcripts that differ only because of alternate splicing in the untranslated region (i.e., when the protein-coding annotation is identical).

Quality assessment for a CCDS build

We have implemented a series of tests that assess the protein sequence, conservation, and likelihood that annotation erroneously represents a pseudogene as protein-coding. The Ensembl and NCBI protein length and sequence are compared by alignment to identify and discard proteins that are discordant owing to annotation or processing error or insertion/deletion differences between NCBI RefSeq proteins and the protein annotated on the reference genome. Two types of protein comparisons are done: (1) protein sequences provided by each annotation source as FASTA files are compared to the conceptually translated CDS sequence that is extracted de novo from the genome annotation coordinates, and (2) proteins provided as FASTA by each annotation data source are compared to each other. Putative retrotransposed pseudogenes are identified from mRNAs [with poly(A) tails removed] that align to the genome using BLASTZ (Schwartz et al. 2003) at more than one location. Alignments are scored for a series of features to identify putative retrotransposed pseudogenes as previously described (Baertsch et al. 2008). Protein models are also evaluated for genome conservation patterns that may indicate that the gene is not functional in human. Analysis of BLASTZ cross-species alignments to the human genome gene annotation detects potential problems including: nonconserved start and stop codons, nonconserved splice sites, uncompensated insertion- or deletion-associated frameshifts, and in-frame stop codons. Cross-species alignments included chimpanzee, mouse, rat, dog, chicken, and rhesus; only syntenic alignments are used from the assembled genomes. Additional QA tests are applied by NCBI to the RefSeq sequences included in the CCDS database as previously described (Pruitt et al. 2007).

Exchange of curation data

NCBI maintains a relational database that tracks CCDS candidates; locations and identifiers; results of quality assurance tests, curator comments; and CCDS IDs and versions. A database extraction is distributed via a private ftp site to the members of the CCDS collaboration on a daily basis. A restricted-access website supports the collaboration, and a public access website and ftp site disseminate data for CCDS IDs.

Quality assessment of the current human and mouse CCDS data set

HomoloGene

The number of mouse and human CCDSs with homologs in other species was calculated by determining whether the current CCDS genes for human and mouse are members of a HomoloGene group (release 62) that has a gene from at least one other species as a member. Additional filtering identified HomoloGene clusters containing human and mouse genes where the corresponding human and mouse loci are both associated with a CCDS ID.

Comparison to SWISS–PROT

Manually curated SWISS-PROT records (Apweiler et al. 2008) (release 55.5) were obtained via the EBI ftp site (ftp://ftp.ebi.ac.uk/pub/databases/uniprot/). A SWISS-PROT accession number may represent more than one sequence isoform. For example, Q8NCE2 includes annotation indicating that amino acids 479 through 538 are missing in isoform 3. Therefore, the SWISS-PROT data set was processed with VARSPLIC (Kersey et al. 2000) to derive an expanded data set (isoforms) by extracting alternate splice products based on record annotation. Exonerate (Slater and Birney 2005) was used with the affine:local model and a sequence identity threshold of 95% to align UniProt sequences to CCDS proteins. Alignments were analyzed and binned into several different categories, with interpretations based on an evaluation of alignments to the expanded set of alternate protein variants versus alignments to the UniProt record where the record is defined as a CCDS protein alignment with the highest coverage score to any of the set of splice variants extracted from the UniProt record. Results were binned as follows: (1) alignment found or not, (2) overall sequence identity, (3) N terminus identical or not, (4) C terminus identical or not, and (5) the alignment coverage of the UniProt protein.

RFC conservation analysis

Conservation analysis was performed using genomic annotations of transcripts for human assembly NCBI 36 (UCSC hg18) and mouse assembly NCBI 37 (UCSC mm9) for CCDS along with the corresponding RefSeq and Ensembl protein-coding gene sets (Supplemental Table 2). Human transcripts were scored against mouse and dog, with mouse transcripts scored using human and dog. The score for a gene is the maximum RFC for any of its transcripts against either of the aligned genomes. A control data set of randomized human sequences was also scored, as previously described (Clamp et al. 2007). Controls are non-protein-coding regions of the genome that serve as a null model, with similar structure, GC content, alignment coverage, and mutation rates as for well-known protein-coding genes.

Acknowledgments

We thank the programmer, database, and curation staff at Ensembl, NCBI, WTSI, and UCSC for their contribution to the CCDS analysis, maintenance, and continuing curation efforts. We thank the UniProt Consortium, the HGNC, and MGI for many useful discussions that improve protein representation in all data sets. UCSC thanks the UCSC Genome Browser team for their tools, data, and assistance, and Michele Clamp (Broad Institute) for providing controls for conservation analysis. NCBI thanks Zev Hochberg for his contributions toward the initial CCDS database schema and CCDS build analysis. UCSC was funded for this work from subcontract no. 0244-03 from NHGRI grant no. 1U54HG004555-01 to the Wellcome Trust Sanger Institute. Work at the Wellcome Trust Sanger Institute was supported by the Wellcome Trust (grant nos. WT062023, WT077198) and by NHGRI grant no. 1U54HG004555-01. Work at NCBI was supported by the Intramural Research Program of the NIH, National Library of Medicine.

References

Apweiler R, Bougueleret L, Altairac S, Amendolia V, Auchincloss A, Argoud-Puy G, Axelsen K, Baratin D, Blatter MC, Boeckmann B, et al. 2008. The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res* **37:** D169–D174.

Ashurst JL, Chen CK, Gilbert JG, Jekosch K, Keenan S, Meidl P, Searle SM, Stalker J, Storey R, Trevanion S, et al. 2005. The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res* **33:** D459–D465.

Baertsch R, Diekhans M, Kent WJ, Haussler D, Brosius J. 2008. Retrocopy contributions to the evolution of the human genome. *BMC Genomics* **9:** 466. doi: 10.1186/1471-2164-9-466.
Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. 2001. Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Res* **11:** 1005–1017.
Birney E, Andrews T, Bevan P, Caccamo M, Chen Y, Clarke L, Coates G, Cuff J, Curwen V, Cutts T, et al. 2004. An overview of Ensembl. *Genome Res* **14:** 925–928.
Bruford EA, Lush MJ, Wright MW, Sneddon TP, Povey S, Birney E. 2008. The HGNC Database in 2008: A resource for the human genome. *Nucleic Acids Res* **36:** D445–D448.
Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES. 2007. Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci* **104:** 19428–19433.
Dombrowski SM, Maglott D. 2002. Using the Map Viewer to explore genomes. In *The NCBI handbook*. National Library of Medicine, Bethesda, MD. http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.chapter.ch20.
Eppig JT, Bult CJ, Kadin JA, Richardson JE, Blake JA, Anagnostopoulos A, Baldarelli RM, Baya M, Beal JS, Bello SM, et al. 2005. The Mouse Genome Database (MGD): From genes to mice—a community resource for mouse biology. *Nucleic Acids Res* **33:** D471–D475.
The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.
Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. *Nat Rev Genet* **7:** 85–97.
Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, et al. 2006. GENCODE: Producing a reference annotation for ENCODE. *Genome Biol* **7:** S4. doi: 10.1186/gb-2006-7-s1-s4.
Johnston H, Kneer J, Chackalaparampil I, Yaciuk P, Chrivia J. 1999. Identification of a novel SNF2/SWI2 protein family member, SRCAP, which interacts with CREB-binding protein. *J Biol Chem* **274:** 16370–16376.
Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, et al. 2008. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* **36:** D773–D779.
Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241–254.
Kersey P, Hermjakob H, Apweiler R. 2000. VARSPLIC: Alternatively-spliced protein sequences derived from SWISS-PROT and TrEMBL. *Bioinformatics* **16:** 1048–1049.
Kitts P. 2002. Genome assembly and annotation process. In *The NCBI handbook*. National Library of Medicine, Bethesda, MD. http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.chapter.ch14.
Lejeune F, Maquat LE. 2005. Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. *Curr Opin Cell Biol* **17:** 309–315.
Maglott D, Ostell J, Pruitt KD, Tatusova T. 2007. Entrez Gene: Gene-centered information at NCBI. *Nucleic Acids Res* **35:** D26–D31.
Potter SC, Clarke L, Curwen V, Keenan S, Mongin E, Searle SM, Stabenau A, Storey R, Clamp M. 2004. The Ensembl analysis pipeline. *Genome Res* **14:** 934–941.
Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35:** D61–D65.
Pruitt KD, Tatusova T, Klimke W, Maglott DR. 2009. NCBI Reference Sequences: Current status, policy and new initiatives. *Nucleic Acids Res* **37:** D32–D36.
Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, et al. 2009. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **37:** D5–D15.
Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. 2003. Human mouse alignments with BLASTZ. *Genome Res* **13:** 103–107.
Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6:** 31. doi: 10.1186/1471-2105-6-31.
Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, et al. 2008. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **36:** D13–D21.
Wong MM, Cox LK, Chrivia JC. 2007. The chromatin remodeling protein, SRCAP, is critical for deposition of the histone variant H2A.Z at promoters. *J Biol Chem* **282:** 26132–26139.
Zweig AS, Karolchik D, Kuhn RM, Haussler D, Kent WJ. 2008. UCSC Genome Browser tutorial. *Genomics* **92:** 75–84.

Received December 5, 2008; accepted in revised form April 20, 2009.

FEATURE

Unlocking the secrets of the genome

Despite the successes of genomics, little is known about how genetic information produces complex organisms. A look at the crucial functional elements of fly and worm genomes could change that.

Susan E. Celniker, Laura A. L. Dillon, Mark B. Gerstein, Kristin C. Gunsalus, Steven Henikoff, Gary H. Karpen, Manolis Kellis, Eric C. Lai, Jason D. Lieb, David M. MacAlpine, Gos Micklem, Fabio Piano, Michael Snyder, Lincoln Stein, Kevin P. White and Robert H. Waterston, for the modENCODE Consortium

The primary objective of the Human Genome Project was to produce high-quality sequences not just for the human genome but also for those of the chief model organisms: *Escherichia coli*, yeast (*Saccharomyces cerevisiae*), worm (*Caenorhabditis elegans*), fly (*Drosophila melanogaster*) and mouse (*Mus musculus*). Free access to the resultant data has prompted much biological research, including development of a map of common human genetic variants (the International HapMap Project)[1], expression profiling of healthy and diseased cells[2] and in-depth studies of many individual genes. These genome sequences have enabled researchers to carry out genetic and functional genomic studies not previously possible, revealing new biological insights with broad relevance across the animal kingdom[3,4].

Nevertheless, our understanding of how the information encoded in a genome can produce a complex multicellular organism remains far from complete. To interpret the genome accurately requires a complete list of functionally important elements and a description of their dynamic activities over time and across different cell types. As well as genes for proteins and non-coding RNAs, functionally important elements include regulatory sequences that direct essential functions such as gene expression, DNA replication and chromosome inheritance.

Although geneticists have been quick to decode the functional elements in the yeast *S. cerevisiae*, with its small compact genome and powerful experimental tools[5–6], our understanding of the more complex genomes of human, mouse, fly and worm is still rudimentary. Intrinsic signals that define the boundaries of protein-coding genes can only be partly recognized by current algorithms, and signals for other functional elements are even harder to find and interpret. Experimental approaches, notably the sequencing of complementary DNA and expressed sequence tags, have been invaluable, but unfortunately these data sets remain incomplete[7]. Non-coding RNA genes present an even greater challenge[8–10], and many remain to be discovered, particularly those that have not been strongly conserved during evolution. Flies and worms have roughly the same number of known transcription factors as humans[11], but comprehensive molecular studies of gene regulatory networks have yet to be tackled in any of these species.

In an attempt to remedy this situation, the National Human Genome Research Institute (NHGRI) launched the ENCODE (Encyclopedia of DNA Elements) project in 2003, with the goal of defining the functional elements in the human genome. The pilot phase of the project focused on 1% of the human genome and a parallel effort to foster technology development[12]. The initial ENCODE analysis revealed new findings but also made clear just how complex the biology is and how our grasp of it is far from complete[13]. On the basis of this experience, the NHGRI launched two complementary programmes in 2007: an expansion of the human ENCODE project to the whole genome (www.genome.gov/ENCODE) and the model organism ENCODE (modENCODE) project to generate a comprehensive annotation of the functional elements in the *C. elegans* and *D. melanogaster* genomes (www.modencode.org; www.genome.gov/modENCODE).

These two model organisms, with their ease of husbandry and genetic manipulation, are pillars of modern biological research, and a systematic catalogue of their functional genomic elements promises to pave the way to a more complete understanding of the human genome. Studies of these animals have provided key insights into many basic metazoan processes, including developmental patterning, cellular signalling, DNA replication and inheritance, programmed cell death and RNA interference (RNAi). The genomes are small enough to be investigated comprehensively with current technologies and findings can be validated *in vivo*. The research communities that study these two organisms will rapidly make use of the modENCODE results, deploying powerful experimental approaches that are often not possible or practical in mammals, including genetic, genomic, transgenic, biochemical and RNAi assays. modENCODE, with its potential for biological validation, will add value to the human ENCODE effort by illuminating the relationship between molecular and biological events.

The modENCODE project (Table 1) complements other systematic investigations into these highly studied organisms. In both organisms, RNAi collections have been developed and used to uncover novel gene functions[14–18]. Mutants are being recovered through insertional mutagenesis[19] and targeted deletions (http://celeganskoconsortium.omrf.org;

TABLE 1 modENCODE CONSORTIUM

Elements	Worm	Fly	Primary experimental data
Transcripts (mRNAs, non-coding RNAs, transcription start sites, untranslated regions, miRNAs)	Robert Waterston (University of Washington), Fabio Piano (New York University)	Susan Celniker (LBNL), Eric Lai (Sloan-Kettering Institute)	Tiling arrays, RNASeq, RT-PCR/RACE, mass spectrometry, 3′ untranslated region clone library, UAS-miRNA flies, knockdowns of RNA-binding proteins
Transcription-factor-binding sites	Michael Snyder (Yale University)	Kevin White (University of Chicago)	ChIP-chip, ChIP-seq, transcription-factor-tagged strains, anti-transcription factor antibodies
Chromatin marks	Jason Lieb (University of North Carolina), Steven Henikoff (University of Washington)	Gary Karpen (LBNL), Steven Henikoff	ChIP-chip and ChIP-seq of chromosome-associated proteins and nucleosomes
DNA replication		David MacAlpine (Duke University Medical Center)	ChIP-chip and ChIP-seq of essential initiator proteins, origin mapping and DNA copy number in differentiated tissues

© 2009 Macmillan Publishers Limited. All rights reserved

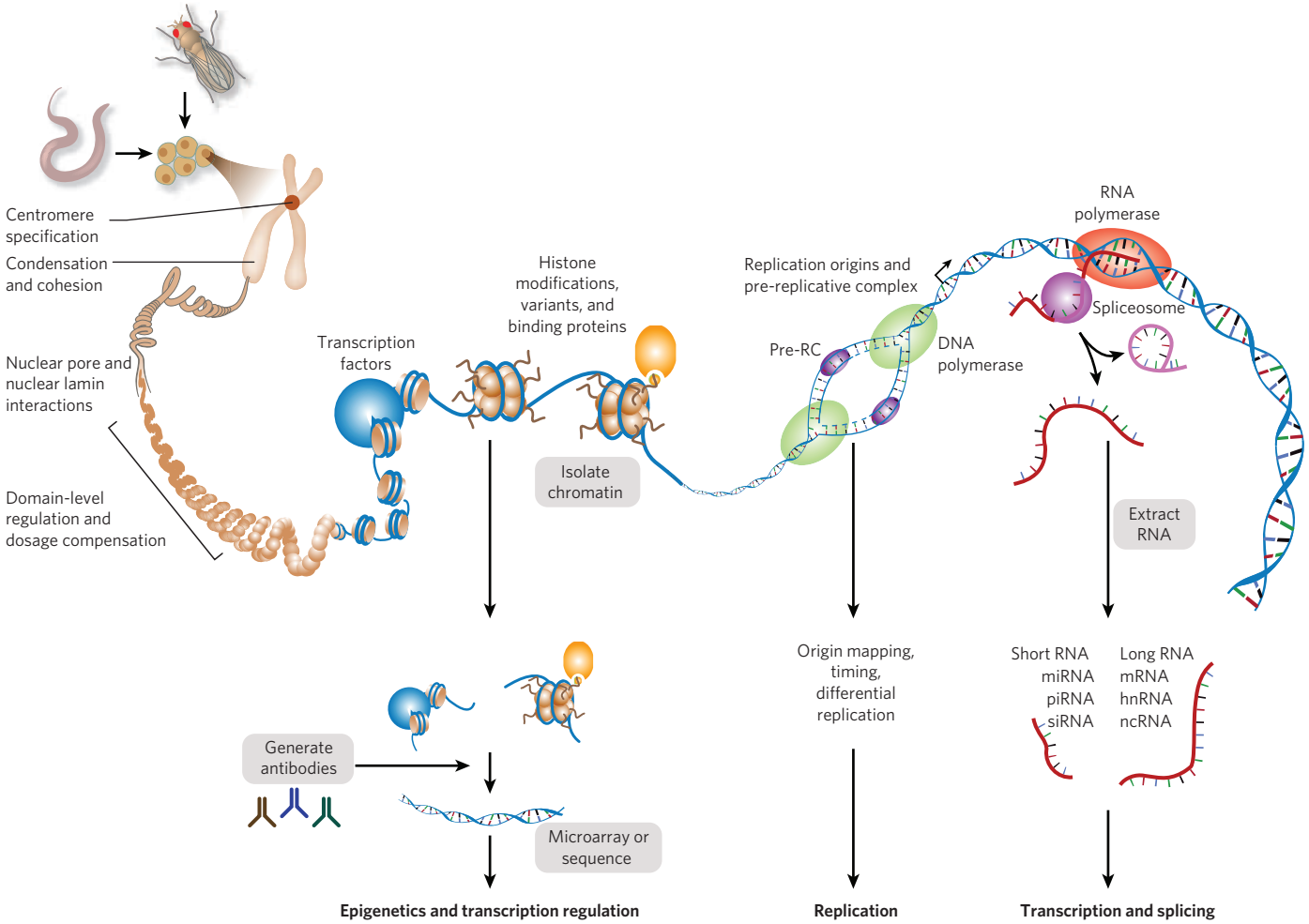


Figure 1 | DNA element functions and identification process.

www.shigen.nig.ac.jp/c.elegans), with the eventual goal of one for every known gene. Genome sequences of related species are now also available for both fly[20,21] and worm[22], and multiple independent wild isolates are being characterized (T. MacKay, personal communication, www.dpgp.org[23]; R.H.W.). First-generation catalogues have been assembled of gene expression patterns during development and in different tissues[24–34].

Research and analysis

The modENCODE project will operate as an open consortium and participants can join on the understanding that they will abide by the set criteria (www.genome.gov/26524644). An important aim of the project is to respond to the needs of the broader *Drosophila* and *C. elegans* scientific communities, and several avenues will be open for suggestions on which experiments to prioritize. For example, researchers can visit www.modencode.org/Vote.shtml now to help prioritize transcription factors for studies using chromatin immunoprecipitation followed by DNA microarray or DNA sequencing (ChIP-chip and ChIP-seq), and can also indicate whether they have useful antibodies. We will seek community input on other issues as the opportunities arise.

The core of the modENCODE project consists of ten groups who use high-throughput methods to identify functional elements (see Table 1). A Data Coordinating Center (DCC) will collect, integrate and display the data. Together, the groups expect to identify the principal classes of functional element for *D. melanogaster* and *C. elegans*. They will work closely together to complete the precise annotation of protein-coding genes, identify small RNAs and non-coding RNA transcripts, map transcription start sites, identify promoter motif elements, elucidate functional elements within 3′ untranslated regions, and identify alternatively spliced transcripts as well as the signals required for splicing. Genomic sites bound by sequence-specific transcription factors will also be comprehensively identified. Charting the chromatin 'landscapes' will include the characterization of key histone modifications and variants, nucleosome phasing, RNA polymerase II isoforms and proteins involved in dosage compensation, centromere function, replication, homologue pairing, recombination and associations of chromosomes with the nuclear envelope.

Integrative analysis of these data across the different types of functional element will be used to reveal fundamental principles of fly and worm genome biology and to begin to uncover the emergent properties of these complex genomes. Some topics the modENCODE groups, along with interested members of the wider community, intend to explore are outlined below, but these are only a beginning. Our intention is to create a resource that will provide the foundation for ongoing analysis by scientists for years to come.

Our two model organisms share many similarities with other metazoans, including humans. They also differ from other organisms in some striking ways, particularly in details of the establishment and maintenance of cellular identity, centromere biology and heterochromatin function. To help understand how the similarities and differences in worm and fly biology are reflected in their genome sequences and how they are specified by genome function at the molecular level, we will carry out comparative analyses of transcription, splicing, *cis*-regulatory and post-transcriptional elements and chromatin function. We will subsequently investigate how our findings apply to the control of gene expression in the human genome.

© 2009 Macmillan Publishers Limited. All rights reserved

We also plan to use genome-wide data on pre- and post-transcriptional functional elements to expand our understanding of gene-regulatory networks. We will study how these two layers of control complement or reinforce each other during development. For example, the availability of full-length transcripts and promoter structures for microRNA (miRNA) genes will enable us to develop models of regulatory circuits that integrate the upstream regulation of miRNA genes with that of other regulatory factors (such as transcription factors) and the effects of miRNAs on their downstream targets. We will search global patterns identified in the regulatory programs for emerging principles of gene regulation within and across species; as part of this endeavour, we will evaluate evidence for the modular structure of regulatory networks.

Because several developmental stages and diverse tissues will be sampled in both animals, we will be able to investigate the global and dynamic activities of functional elements across the entire genome in multiple cell types and stages of differentiation. We aim to define the characteristics and rules that distinguish regulatory programs in different cell types and developmental stages at the DNA, chromatin, and post-transcriptional levels. This will enable us to identify the types of element that function together in various spatio-temporal environments and find new types of functional element, perhaps including those used in restricted developmental contexts.

An important objective is to generate specific biological hypotheses that can be refined and tested experimentally by the broader scientific community. For example, these analyses might identify transcribed regions with novel regulatory roles, structural regions that function in the establishment of chromatin structure or three-dimensional conformation, enhancers far away from the gene they control, and alternative promoter regions. In addition, we will use comparative analyses of the sequenced genomes from different species to clarify the extent of conservation and the functional constraints associated with potential new classes of element and to characterize their evolutionary signatures[21].

Another objective of the modENCODE project is the creation of reference data sets of maximum utility. We have agreed that, whenever possible, a common set of reagents will be used to facilitate comparison of data sets generated by different groups. For example, the fly and worm groups using ChIP-chip and related methods to map the genome-wide distributions of histone modifications will use a common set of validated antibodies. In addition, we will use common fly and worm strains, and in the case of *Drosophila*, the common cell lines Kc167, S2-DRSC, CME W1 Cl.8+ and ML-DmBG3-c2.

The fly and worm genomes are about a thirtieth of the size of their mammalian counterparts, making current methods for high-throughput genomic analysis cost-effective. We will use high-density tiling DNA microarrays to interrogate the genome on a single microarray (*C. elegans*, 26 base pair (bp) median spacing; *D. melanogaster*, 38 bp median spacing) at a resolution sufficient for ChIP-chip experiments. Denser arrays (*D. melanogaster*, 7 bp median spacing), which promise higher resolution, will be used in a move to high-throughput sequencing platforms such as the Illumina Genome Analyzer to generate sufficient sequence coverage for transcript mapping and miRNA and ChIP experiments.

The biological significance of the genomic features identified will be tested in experiments designed to evaluate the accuracy and functionality of subsets of the structural and regulatory annotations. For example, we will carry out ChIP experiments on extracts from whole animals or cells that lack selected regulators (using mutants or RNAi). The tissue-specific DNA-binding patterns of selected regulators will be validated in transgenic animals. Figure 1 summarizes the DNA elements to be interrogated and the methods to be used.

Data management and accessibility

Data generated by the modENCODE Consortium, including those from validation experiments, will be collected, quality checked, integrated and distributed through the modENCODE DCC (www.modencode.org). The DCC will collate detailed metadata for each submitted data set to ensure broad and long-term usability. Where appropriate, the data will also be submitted to public databases, for example, GenBank (www.ncbi.nlm.nih.gov) and the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) or Array Express (www.ebi.ac.uk/microarray-as/aer/entry) and the University of California, Santa Cruz Genome Bioinformatics Site (http://genome.ucsc.edu). The DCC will also work closely with WormBase (www.wormbase.org) and FlyBase (www.flybase.org) to facilitate integration of the modENCODE data with selected data from these databases and with other information about these organisms.

All data will be available for bulk download through an FTP site and through a number of Generic Model Organism Database tools (www.gmod.org): BioMart (www.biomart.org) will provide powerful data-mining capabilities, and InterMine (www.intermine.org) will provide a flexible interface for complex querying of the data, a library of canned queries, and powerful list-based tools and operations (http://intermine.modencode.org). As for the ENCODE pilot project data (www.genome.gov/10005107), new data can be examined alongside existing data using interactive genome browsers[35] for both the fly (www.modencode.org/cgi-bin/gbrowse/fly) and the worm (www.modencode.org/cgi-bin/gbrowse/worm).

The *Drosophila* and *C. elegans* communities have thrived because of their open culture. In keeping with this tradition and with those of the genome sequencing projects, HapMap and the ENCODE pilot project, modENCODE is a 'community resource project' subject to the NHGRI's data-sharing policy. The success of this policy is based on mutual and independent responsibilities for the production and use of the resource. We will release data rapidly (Table 2), before publication, once they have been established to be reproducible (verification; see www.modencode.org/'Publication Policy link' for the criteria), even if the data have not been sampled to determine if there is biological meaning (validation). In turn, users are asked to recognize the source of the data and to respect the legitimate interest of the resource producers to publish an initial report of their work (see www.genome.gov/modencode for more details). Finally, the funding agencies

TABLE 2 GLOBAL ANALYSIS GOALS

Elements and processes	Specific examples
Transcribed regions	Define cell- and tissue-specific transcriptional landscapes. Annotate transcription start sites, exons, untranslated region structures, small regulatory RNAs and short single-exon open reading frames
Gene regulation, transcriptional regulation	Identify transcription-factor binding sites in various cell and tissue types. Correlate chromatin structure marks and transcriptional activities for protein-coding and non-protein-coding genes
Post-transcriptional regulation	Identify tissue-specific binding sites for miRNAs and other small RNAs, RNA secondary structures and alternative splicing regulatory motifs
Chromatin structure and function	Identify sites of association between DNA and chromosomal proteins involved in centromere specification, meiotic recombination, dosage compensation, nuclear envelope and matrix interactions and chromosome condensation. Identify sites of incorporation of histone variants and specifically modified histones. Correlate transcription maps for meta-analysis of developmental chromatin dynamics.
DNA replication	Identify cell- and tissue-specific origins of replication. Correlate with cell- and tissue-specific transcription and chromatin marks

© 2009 Macmillan Publishers Limited. All rights reserved

recognize the need to support the analysis and dissemination of the data.

In addition, a variety of physical resources (for example, DNA constructs and transgenic strains) will be produced that are likely to be of use to the broader community and to which that community will have unrestricted access. We expect to cooperate with data users in the worm and fly communities to set the gold standard for data release and openness.

Conclusion

The Human Genome Project benefited enormously from the technology developed and the experience acquired in sequencing the significantly smaller genomes of model organisms, particularly *C. elegans* and *D. melanogaster*. The modENCODE project is dedicated to the next phase of decoding the information stored in these genomes: the comprehensive identification of sequence-based functional elements. Having laid the foundation for the discovery of many of the genetic programs underlying metazoan development and behaviour, *Drosophila* and *Caenorhabditis* will serve as ideal model systems to identify DNA-based functional elements on a genome-wide basis. In the future, these data will provide a powerful platform for characterizing the functional networks that direct multicellular biology, thereby linking genomic data with the biological programs of higher organisms, including humans. ■

1. Sabeti, P. C. *et al. Nature* **449,** 913–918 (2007).
2. Neve, R. M. *et al. Cancer Cell* **10,** 515–527 (2006).
3. Chintapalli, V. R., Wang, J. & Dow, J. A. *Nature Genet.* **39,** 715–720 (2007).
4. Nichols, C. D. *Pharmacol. Ther.* **112,** 677–700 (2006).
5. Ross-Macdonald, P. *et al. Nature* **402,** 413–418 (1999).
6. Boone, C., Bussey, H. & Andrews, B. J. *Nature Rev. Genet.* **8,** 437–449 (2007).
7. Celniker, S. E. & Rubin, G. M. *Annu. Rev. Genomics Hum. Genet.* **4,** 89–117 (2003).
8. Tupy, J. L. *et al. Proc. Natl Acad. Sci. USA* **102,** 5495–5500 (2005).
9. Ruby, J. G. *et al. Cell* **127,** 1193–1207 (2006).
10. Ruby, J. G. *et al. Genome Res.* **17,** 1850–1864 (2007).
11. Reece-Hoyes, J. S. *et al. Genome Biol.* **6,** R110 (2005).
12. The ENCODE Project Consortium *Science* **306,** 636–640 (2004).
13. Birney, E. *et al. Nature* **447,** 799–816 (2007).
14. Boutros, M. *et al. Science* **303,** 832–835 (2004).
15. Kamath, R. S. *et al. Nature* **421,** 231–237 (2003).
16. Rual, J. F. *et al. Genome Res.* **14,** 2162–2168 (2004).
17. Sonnichsen, B. *et al. Nature* **434,** 462–469 (2005).
18. Dietzl, G. *et al. Nature* **448,** 151–156 (2007).
19. Bellen, H. J. *et al. Genetics* **167,** 761–781 (2004).
20. Clark, A. G. *et al. Nature* **450,** 203–218 (2007).
21. Stark, A. *et al. Nature* **450,** 219–232 (2007).
22. Stein, L. D. *et al. PLoS Biol.* **1,** E45 (2003).
23. Hillier, L. W. *et al. Nature Methods* **5,** 183–188 (2008).
24. Tomancak, P. *et al. Genome Biol.* **3,** research0088.1–0088.14 (2002).
25. Arbeitman, M. N. *et al. Science* **297,** 2270–2275 (2002).
26. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. *Science* **302,** 249–255 (2003).
27. Li, T. R. & White, K. P. *Dev. Cell* **5,** 59–72 (2003).
28. Stolc, V. *et al. Science* **306,** 655–660 (2004).
29. Manak, J. R. *et al. Nature Genet.* **38,** 1151–1158 (2006).
30. Tomancak, P. *et al. Genome Biol.* **8,** R145 (2007).
31. Jiang, M. *et al. Proc. Natl Acad. Sci. USA* **98,** 218–223 (2001).
32. Reinke, V., Gil, I. S., Ward, S. & Kazmer, K. *Development* **131,** 311–323 (2004).
33. Baugh, L. R., Hill, A. A., Slonim, D. K., Brown, E. L. & Hunter, C. P. *Development* **130,** 889–900 (2003).
34. Kim, S. K. *et al. Science* **293,** 2087–2092 (2001).
35. Stein, L. D. *et al. Genome Res.* **12,** 1599–1610 (2002).

Supplementary Information A full list of names and addresses of current consortium participants is linked to the online version of this feature at http://tinyurl.com/modENCODE

Acknowledgements We thank Brenda Andrews and Tim Hughes for discussions on the status of yeast functional genomics.

Author Information Correspondence should be addressed to S.E.C. (celniker@fruitfly.org).

Authors

Susan E. Celniker[1], Laura A. L. Dillon[2], Mark B. Gerstein[3,4], Kristin C. Gunsalus[5], Steven Henikoff[6], Gary H. Karpen[7], Manolis Kellis[8,9], Eric C. Lai[10], Jason D. Lieb[11], David M. MacAlpine[12], Gos Micklem[13], Fabio Piano[5], Michael Snyder[14], Lincoln Stein[15], Kevin P. White[16,17], Robert H. Waterston[18]

[1]Department of Genome Biology, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [2]Division of Extramural Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [3]Program in Computational Biology and Bioinformatics, [4]Department of Computer Science and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA. [5]Center for Genomics and Systems Biology, New York University, New York, New York 10003, USA. [6]Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA. [7]Department of Genome and Computational Biology, Lawrence Berkeley National Laboratory, Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA. [8]Broad Institute, Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02140, USA. [9]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [10]Sloan-Kettering Institute, New York, New York 10065, USA. [11]Department of Biology and Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [12]Department of Pharmacology and Cancer Biology, Duke University Medical Center, Durham, North Carolina 27710, USA. [13]Department of Genetics, University of Cambridge, CB2 3EH, UK, and Cambridge Systems Biology Centre, Tennis Court Road, Cambridge CB2 1QR, UK. [14]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut 06824, USA. [15]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11542 USA. [16]Institute for Genomics & Systems Biology, University of Chicago, Chicago, Illinois 60637, USA. [17]Institute for Genomics & Systems Biology, Argonne National Laboratory, Argonne, Illinois 60439, USA. [18]Department of Genome Sciences and University of Washington School of Medicine, Seattle, Washington 98195, USA.

© 2009 Macmillan Publishers Limited. All rights reserved

Motif Discovery in Physiological Datasets: A Methodology for Inferring Predictive Elements

ZEESHAN SYED
University of Michigan
and
COLLIN STULTZ, MANOLIS KELLIS, PIOTR INDYK, and JOHN GUTTAG
Massachusetts Institute of Technology

In this article, we propose a methodology for identifying predictive physiological patterns in the absence of prior knowledge. We use the principle of conservation to identify activity that consistently precedes an outcome in patients, and describe a two-stage process that allows us to efficiently search for such patterns in large datasets. This involves first transforming continuous physiological signals from patients into symbolic sequences, and then searching for patterns in these reduced representations that are strongly associated with an outcome.

Our strategy of identifying conserved activity that is unlikely to have occurred purely by chance in symbolic data is analogous to the discovery of regulatory motifs in genomic datasets. We build upon existing work in this area, generalizing the notion of a regulatory motif and enhancing current techniques to operate robustly on non-genomic data. We also address two significant considerations associated with motif discovery in general: computational efficiency and robustness in the presence of degeneracy and noise. To deal with these issues, we introduce the concept of active regions and new subset-based techniques such as a two-layer Gibbs sampling algorithm. These extensions allow for a framework for information inference, where precursors are identified as approximately conserved activity of arbitrary complexity preceding multiple occurrences of an event.

We evaluated our solution on a population of patients who experienced sudden cardiac death and attempted to discover electrocardiographic activity that may be associated with the endpoint of death. To assess the predictive patterns discovered, we compared likelihood scores for motifs in the sudden death population against control populations of normal individuals and those with non-fatal supraventricular arrhythmias. Our results suggest that predictive motif discovery may be able to identify clinically relevant information even in the absence of significant prior knowledge.

Categories and Subject Descriptors: I.2.6 [**Artificial Intelligence**]: Learning; I.5.4 [**Pattern Recognition**]: Applications; J.3 [**Life and Medical Sciences**]

Z. Syed was previously affiliated with the Massachusetts Institute of Technology. This work was supported in part by the Center for the Integration of Medicine and Innovative Technology (CIMIT) and the Harvard-MIT Division of Health Sciences and Technology (HST).
Author's address: Z. Syed; email: zhs@umich.edu.
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© 2010 ACM 1556-4681/2010/01-ART2 $10.00
DOI 10.1145/1644873.1644875 http://doi.acm.org/10.1145/1644873.1644875

General Terms: Algorithms, Design, Experimentation, Performance

Additional Key Words and Phrases: Gibbs sampling, knowledge discovery, data mining, motifs, physiological signals, inference

ACM Reference Format:

Syed, Z., Stultz, C., Kellis, M., Indyk, P., and Guttag, J. 2010. Motif discovery in physiological datasets: A methodology for inferring predictive elements. ACM Trans. Knowl. Discov. Data. 4, 1, Article 2 (January 2010), 23 pages.
DOI = 10.1145/1644873.1644875 http://doi.acm.org/10.1145/1644873.1644875

1. INTRODUCTION

The subject of finding predictive elements has been extensively studied in a wide variety of contexts including geodesic, medical, and financial data. In this article, we present a motif discovery methodology for discovering precursors. While we focus mainly on physiological datasets, we present general techniques that may be broadly applicable to a wider group of signals.

We model prediction as the problem of identifying activity that consistently precedes an event of interest. In the absence of any prior knowledge, this activity can be discovered by observing multiple occurrences of the event and detecting statistically significant commonalities in the data preceding it, by searching for conserved elements unlikely to occur purely by chance prior to the event of interest (Figure 1). To handle noise, we further adopt a relaxed view of conservation, whereby precursors may approximately match or be altogether absent on some observations of the event. A further practical consideration is that the search be computationally efficient to handle large amounts of data resulting from multiple observations.

This model of prediction is similar to the search for regulatory motifs in the setting of computational biology. Motif discovery techniques operate on genomic datasets and search for DNA sequences that are conserved across genomes. We generalize this model and describe how the search for precursors to acute clinical events can be carried out in an analogous manner, by first converting continuous physiological signals into an alphabetical representation, and then mining this representation for conserved activity. A variety of randomized greedy algorithms can be used to efficiently carry out the search for such patterns. We use techniques such as TCM and Gibbs sampling as the foundation of our work, and enhance them to operate on data with highly divergent background distributions of symbols, frequent noise and patterns of increased degeneracy relative to genomic data.

The rest of this article describes the proposed unsupervised inference methodology. While the techniques we suggest can be used on a variety of signals and are sufficiently general-purpose, we motivate them in the more concrete setting of searching for predictive activity in physiological signals. We detail the challenges associated with such an approach and describe its benefits and limitations.

Section 2 details the concept and challenges of representing continuous physiological signals as symbolic strings. Section 3 presents a similar discussion of the problem of detecting predictive motifs in string data. Section 4

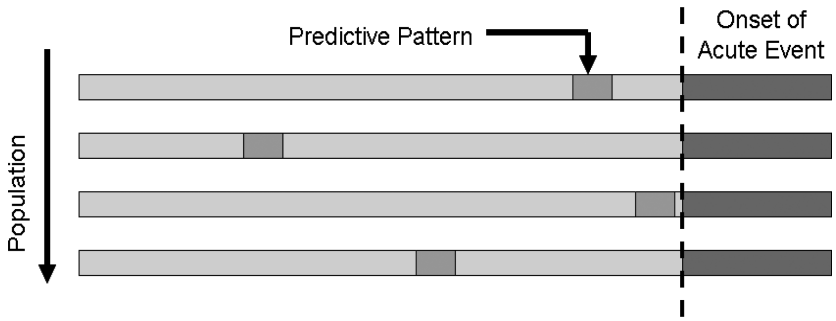


Fig. 1. Prediction through conservation in the context of a population of patients affected by a common acute clinical event.

describes existing computational biology algorithms for motif detection, while Section 5 proposes data transformations and algorithms (including a two-level Gibbs sampling technique) that have been augmented to search for motifs in a computationally-efficient manner in the presence of noise and degeneracy. An application of our work to sudden cardiac death data is discussed in Section 6. Related work is presented in Section 7. Finally, a summary and conclusions appear in Section 8.

2. SYMBOLIZATION

2.1 Symbolic Representation of Physiological Data

The notion of representing physiological signals as symbolic sequences follows from the quasi-periodic nature of many important signals. For example, data from the heart and lungs often comprises units such as heart beats or breaths, which are repetitive. It is often more natural to analyze physiological signals in terms of these units than at the level of raw samples.

We use the property of quasi-periodicity in physiological signals to determine appropriate boundaries for segmentation, and then replace each unit with a single symbol. In doing so, we exploit the underlying repetitive structure and redundancy to obtain a layer of data reduction. The raw physiological data is reexpressed to retain salient differences between units of quasi-periodic activity while abstracting away the common structure. For example, as shown in Figure 2, raw ECG data can be partitioned at the level of heart beats into different equivalence classes, each of which is assigned a unique alphabetic label for identification. This reduces the data rate from around 4000 bits/second (for a beat lasting one second in a signal sampled at 360 Hz with 11 bit quantization) to n bits/second (where n depends upon the number of bits needed to differentiate between symbols, two for this case).

The data reduction introduced by symbolization reduces the search space for the detection of interesting activity and provides a significant computational advantage over working in the original space of the raw signal. A further advantage of using symbolization is that it implicitly abstracts away some of the time-normalization issues that complicate the use of cross-correlation and other techniques that operate on raw time samples.

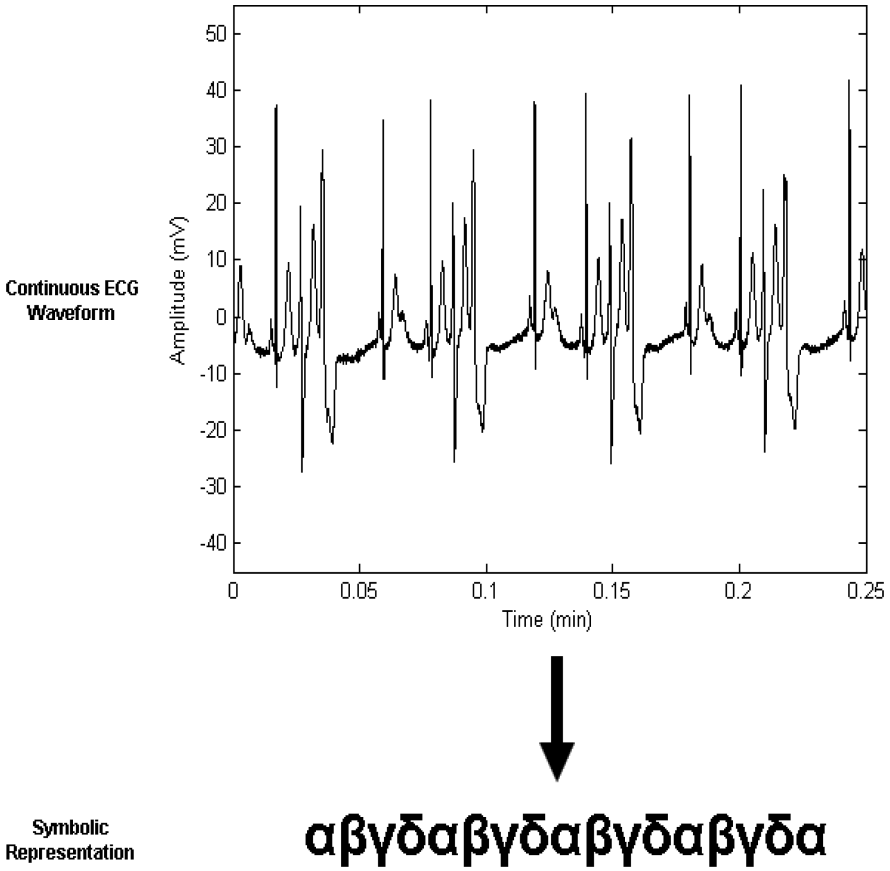


Fig. 2. Example transformation of continuous ECG waveform to a string of symbols. Each of the symbols shown corresponds to a different class of electrophysiological activity.

2.2 Creating Symbolic Representations

To transform continuous waveforms into a string representation that can be mined for patterns more efficiently, we propose segmenting the original signal into intervals and then assigning an alphabetic label to each token. This effectively transforms the original data into a sequence of symbols, and maps the problem into the domain of string algorithms.

The task of assigning labels can be carried out in a number of different ways. One approach is to use clinical information to partition segmented tokens into equivalence classes. This approach provides a set of symbols that have a fixed meaning in a medical context and can be shared across a population. For example, the ECG signal in Figure 3 can be decomposed into RR-intervals as shown (each RR-interval corresponds to the period between two successive contractions of the ventricles of the heart: the period between successive sharp spikes in the raw ECG tracings). Each RR-interval can then be labeled using existing annotations for electrophysiological activity. RR-intervals associated

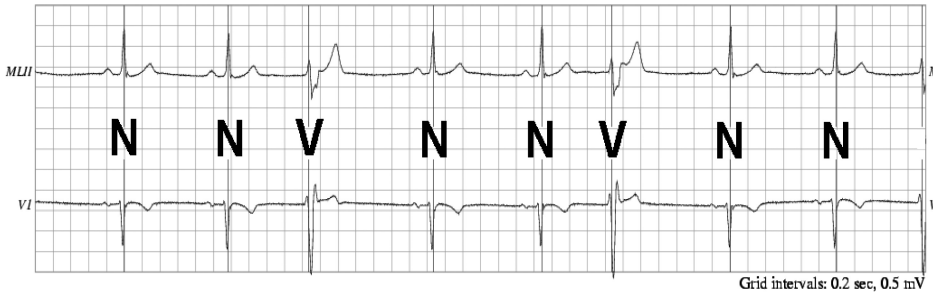


Fig. 3. Example symbolization of continuous ECG waveforms using clinical annotations (N = normal, V = premature ventricular contraction).

with normal heart beats are labeled N, while those associated with abnormal contractions originating from ventricular regions are labeled V.

The approach of using clinical annotations is restricted to detecting predictive activity that expresses itself in terms of known clinical classes. It does not allow for the isolation of changes at the level of variations within particular classes. This is important because the granularity of later analysis is constrained by the granularity of labeling.

From a knowledge discovery goal, it is appealing to derive the alphabet for symbolization directly from the data itself. Techniques such as those in Syed et al. [2007] can be employed to achieve this goal. While the approach of generating a patient-specific symbolic representation is powerful in its ability to capture significant changes across a patient, it poses the problem that the clusters are derived separately for each patient. This restricts comparisons across a population. A possible means for addressing this issue is to use a semisupervised approach where the symbols derived for each patient are related by a human expert. This allows for the symbols to be dynamically derived based on characteristics inherent in the data itself, and for these symbols to be related and compared across a population.

At present, registering patient-specific symbols in a fully automated manner across a population represents an area of continuing work. The discussion that follows therefore focuses on the use of clinical annotations (or of semisupervised symbols related manually across patients) despite the possible benefits of patient-specific symbols.

3. PHYSIOLOGICAL MOTIFS

3.1 Physiological Motifs in Symbolic Data

In the setting of computational biology, regulatory motifs correspond to short DNA sequences that regulate gene expression. This notion of a genetic switch that controls activity further downstream is well-suited to our model for prediction. We generalize this idea and choose to model regulatory motifs as sequential triggers that precede abrupt clinical events and are conserved across a population of patients owing to an association with the event.

A recent strategy for regulatory motif discovery that has gained popularity is to make use of comparative genomics [Kellis et al. 2003]. This allows for the discovery of regulatory elements by exploiting their evolutionary conservation across related species. Under this approach, regulatory motif discovery can be viewed computationally as finding sequences that are recurrent is a group of strings, upstream of specified endpoints.

The problem of regulatory motif discovery can be stated more formally in either a combinatorial or probabilistic framework [Jones and Pevzner 2004]. While the two frameworks both attempt to identify similar preceding subsequences, they may lead to slightly different results and require distinct algorithmic techniques.

—*Combinatorial*. Given a set of sequences $\{s_1, \ldots, s_N\}$, find a subsequence $m_1, \ldots, m_W$ that occurs in all s_i with k or fewer differences.

—*Probabilistic*. Given a set of sequences $\{s_1, \ldots, s_N\}$, find a set of starting positions $\{p_1, \ldots, p_N\}$ in the sequences that lead to the best (as defined in the following) $A \times W$ profile matrix M (where A is the number of different symbols in the data and W is the length of the motif).

For the probabilistic case, the profile matrix is derived from the subsequences of length W immediately following the starting positions $p_1, \ldots, p_N$ in each of $s_1, \ldots, s_N$. These subsequences are lined up and the probability of each of the A unique symbols at every one of the W motif positions is estimated. $M(x, y)$ then gives the probability that the motif has character x at position y. The resulting profile matrix can be scored using different criteria with the implicit goal of seeking a nontrivial profile that is strongly conserved at each position and best explains the data. The scoring function most often used is the log-odds likelihood:

$$score = \sum_{i=1}^{N} \sum_{j=1}^{W} \log \left[\frac{M(s_i(p_i + j - 1), j)}{B(s_i(p_i + j - 1))} \right], \tag{1}$$

where B gives the background distribution of each unique symbol in the data. Effectively, this calculates the log-likelihood of a motif while compensating for trivial occurrences that would be seen in the data merely due to the frequent occurrence of certain symbols.

3.2 Challenges Associated with Motif Detection in Symbolic Signals

The problem of motif discovery gives rise to a number of issues in the physiological setting. This section discusses the major challenges faced when modeling acute clinical events as physiological motifs.

3.2.1 *Symbol Distribution Skews.* A complication arising in the context of physiological signals is that of the sparsity of abnormal activity. Periods with interesting events are typically separated by long, variable-sized runs of normal behavior—the distribution of the symbols is significantly skewed in favor of normal labels. This increases the number of trivial motifs in the data and consequently the running time of the motif discovery algorithms. In

addition, for algorithms such as TCM and Gibbs sampling, discussed in Section 4, a secondary effect resulting from the presence of long stretches of normal behavior is that the starting locations chosen randomly may often correspond to uninteresting regions of the signal, further increasing time to convergence.

3.2.2 *Motif Degeneracy.* The issue of degeneracy is frequently encountered in DNA sequences and assumes a critical role for physiological motifs as well. Predictive patterns may be approximately conserved across some patients in a population, while in others, they may be missing altogether. This results from a variety of factors, including differences in the age, gender, clinical history, medications, and lifestyle of patients, as well as noise obscuring predictive patterns in some recordings.

The goal of detecting imperfectly conserved activity represents a significant challenge to the task of discovering precursors. Since patterns can vary, the process of determining whether a pattern appears in a patient, is required to explore a large search space, spanning all possible variations. Similarly, the fact that some patients may have the predictive activity obscured due to noise requires recognizing these cases and preventing motif discovery algorithms from forcibly incorporating this data in the search process.

4. COMPUTATIONAL BIOLOGY ALGORITHMS FOR MOTIF DISCOVERY

In this section, we review three popular algorithms for finding regulatory motifs using comparative genomics; the Two Component Mixture (TCM) algorithm using expectation-maximization, Gibbs sampling, and Consensus. TCM and Gibbs sampling attempt to solve the probabilistic formulation of motif discovery, while Consensus focuses on the combinatorial problem.

4.1 Two Component Mixture (TCM)

TCM is an enhancement to the basic EM algorithm [Bailey and Elkan 1995], which essentially reduces the search into two smaller, decoupled problems. The first (the M-step) involves constructing the profile for a motif given a set of fuzzy starting positions $p_1, \ldots, p_N$ in the input sequences (the M-step). The second (the E-step) then uses this matrix profile representation to score all possible starting positions in every sequence and then update the initial $p_1, \ldots, p_N$.

The overall TCM algorithm operates in the following manner.

TCM Algorithm

TCM($\{s_1, \ldots, s_N\}$, W):
1. Set random starting positions $p_1, \ldots, p_N$
2. Do
   I. M-step to update profile matrix
   II. E-step to update starting positions
   Until the change in the score of M is less than some threshold ε.

The M-step of TCM estimates the profile matrix using the probability Z_{ij} that the motif starts in sequence i at position j. As a first step, the values $n_{c,k}$,

which indicate how often the character c occurs at position k in the motif, are estimated.

$$n_{c,k} = \begin{cases} \sum_{i} \sum_{j \mid s_{i,j}=c} Z_{ij} & k > 0 \\ n_c - \sum_{j=1}^{W} n_{c,j} & k = 0 \end{cases} \tag{2}$$

$k = 0$ represents the case where character c occurs in the sequence outside the motif, while n_c gives the total number of times c occurs in the data. Using these values, we can obtain a profile matrix M as follows:

$$M_{c,k} = \frac{n_{c,k} + d_{c,k}}{\sum_{a} \left(n_{a,k} + d_{a,k}\right)}, \tag{3}$$

where $d_c,_k$ denotes the pseudocount for character c and helps ensure that the probability of c at position k is not zero while estimating frequencies from finite data [Bailey and Elkan 1995].

In addition to computing the profile matrix during the M-step, TCM also calculates a prior probability that a motif might start arbitrarily at any position in the data. This is denoted by λ and is obtained by taking the average of Z_{ij} across all sequences and positions.

TCM primarily differs from other EM approaches to motif discovery in its E-step. For every sequence s_i in the dataset, TCM assigns a likelihood L_{ij} to the W-mer starting at each position j:

$$L_{ij}(1) = \Pr(s_{ij} | Z_{ij} = 1, M, b) = \prod_{k=j}^{j+W-1} M_{k-j+1,c_k} \tag{4}$$

and

$$L_{ij}(0) = \Pr(s_{ij} | Z_{ij} = 0, M, b) = \prod_{k=j}^{j+W-1} b_{c_k}, \tag{5}$$

where b gives the background probability for each character in the dataset. For iteration t of TCM, the values of Z_{ij} can then be estimated using:

$$Z_{ij}^{(t)} = \frac{L_{ij}^{(t)}(1)\lambda^{(t)}}{L_{ij}^{(t)}(0)\left[1 - \lambda^{(t)}\right] + L_{ij}^{(t)}(1)\lambda^{(t)}}. \tag{6}$$

4.2 Gibbs Sampling

Gibbs sampling [Gert et al. 2002] can be viewed as a stochastic analogue of EM for finding regulatory motifs and is less susceptible to local minima than EM. It is also much faster and uses less memory in practice. This is because unlike EM, the Gibbs sampling approach keeps track only of the starting locations, $p_1, \ldots, p_N$, of the motif in each sequence and does not maintain a distribution over all possible starting positions for the motif (the Z_{ij} in TCM representing fuzzy starting positions, are replaced by hard $p_1, \ldots, p_N$).

The Gibbs sampling algorithm for motif discovery can then be written as follows.

Gibbs Sampling Algorithm

GIBBS($\{s_1, \ldots, s_N\}$, W):
1. Set random initial values for p
2. Do
   i. Select sequence s_i at random
   ii. Estimate M from set $\{s_1, \ldots, s_N\}$-s_i
   iii. Use M to score all starts in s_i
   iv. Pick start p_i with probability proportional to its score
   Until the change in the score of M is less than some threshold ε .

Gibbs sampling is less dependent on the initial parameters than TCM and therefore more versatile. However, it is dependent on all sequences having the motif. This is an inefficiency we address in our work.

4.3 Consensus

Consensus [Stormo and Hartzell 1989] is a greedy motif clustering algorithm that picks out two sequences at random, finds the most conserved pairs among them and then iterates over all the remaining sequences adding the W-mers that best match the results of the previous iteration at every stage.

The Consensus algorithm is as follows.

Consensus Algorithm

CONSENSUS($\{s_1, \ldots, s_N\}$, W):
1. Pick sequences s_i and s_j at random
2. Find most similar W-mers in s_i and s_j
3. For each unprocessed sequence s_k
   i. Expand solution set with W-mers from s_k that match best with previous ones.

5. DATA TRANSFORMATIONS AND SUBSET-BASED TECHNIQUES

5.1 Active Regions

The issue of skewed symbol distributions can be addressed by removing long stretches of activity that are known to be uninteresting. By definition, a predictive motif is associated with an acute clinical event and must be associated with abnormal activity. As a result, trivial motifs comprising normal activity can be trimmed away to reduce the running time associated with the motif-discovery algorithms. For example, given the sequence:

V JV J J N N N N N N N N N N N V N V N B B r,

a possible reduction of this data would be:

V J V J J N + V N + V N + B B r.

This technique is associated with a significant loss of information. Specifically, the search for motifs proceeds in the transformed space, and the N+ regular expression may occur in motifs without a consistent meaning (it may be arbitrarily long in some patients). The more general issue here is that conservation of a pattern in the transformed space does not imply conservation in the original signals.

To avoid this issue, we identify regions of abnormal activity—active regions—by splicing out trivial periods in the signal. Given a motif length W, this involves iterating over the data and removing all normal symbols that would occur only in trivial motifs. This approach preserves the temporal structure of abnormal stretches of the signal, ensuring that the motifs correspond to patterns that are conserved in all of the original signals. For example, using this approach for a motif of length 3, the original example pattern would map to:

V J V J J N N V N V N B B r.

5.2 Gibbs^2 and Seeded Consensus

The Gibbs sampling algorithm in Section 4 assumes that a motif is present in all sequences. To deal with the issue of degeneracy, where noise may obscure the predictive pattern completely for some patients, we propose a new algorithm that provides a layer of robustness while dealing with a population where activity may be altogether absent in some of the observed examples. This is achieved by adding a second layer of Gibbs sampling to the original algorithm, leading to the Gibbs^2 algorithm presented here.

The Gibbs^2 algorithm operates at any time on a working subset $V = \{v_1, \ldots, v_C\}$ of the original sequences $\{s_1, \ldots, s_N\}$. Sequences are dynamically swapped into and out of this set with the goal of replacing poor matches with potentially better options. The underlying goal is to arrive at a cluster of sequences that share a strongly conserved motif.

The initial subset of sequences is chosen at random, and at each iteration, a single sequence, v_i, in the working set is scored at every position p_i, using the profile generated from $V–v_i$, i.e.:

$$score\,(v_i\,(p_i)) = \sum_{j=1}^{W} \log\left[\frac{M(s_i(p_i + j - 1), j)}{B(s_i(p_i + j - 1))}\right]. \tag{7}$$

With some probability, v_i is swapped out and replaced by one of the sequences outside the working set. The probability of being swapped out varies inversely with the maximum score seen for the sequence at any position, the score at the position that corresponds most strongly to the profile matrix:

$$\log[\Pr(swap)] \propto -\max_{pi}[score(v_i(p_i))]. \tag{8}$$

The proportionality factor depends on the length of the motifs being searched for.

The intuition behind the Gibbs^2 algorithm is that if a sequence scores high for a motif, it matches quite well with other sequences used to derive the profile and is retained with a higher probability. Conversely, if a sequence does not

score highly, it matches poorly with the remaining sequences in the working set used to derive the profile.

Ideally, the sequence swapped out should be replaced by one that scores highest on the profile matrix being used. This approach is computationally intensive since all outstanding sequences need to be scored before the optimal one can be chosen. To avoid this, once a sequence is swapped out, it is replaced by any of the sequences outside the working set at random. This avoids the need to score all previously excluded sequences to find the one with the best match. Furthermore, after each swap, further swapping is temporarily disabled to allow the new sequence to be absorbed and contribute to the profile matrix.

The Gibbs^2 algorithm can be written as follows (with C denoting the size of the working set and K representing the number of iterations, swapping is disabled after a sequence is replaced from one outside the working set):

Gibbs^2 Algorithm

GIBBS^2($\{s_1, \ldots, s_N\}$, W, C, K):
1. Choose C sequences at random from $\{s_1, \ldots, s_N\}$
2. Set random initial values for p
3. Do
   i. Select sequence v_i at random
   ii. Estimate M from set $V - v_i$
   iii. Use M to score all starts in v_i
   iv. Swap out v_i with Pr($swap$) and replace it with a random sequence outside the working set
   v. If swap occurs,
      a. Disable swapping for K iterations
   vi. Pick start p_i with probability proportional to its score
   Until the change in the score of M is less than some threshold ε.

The Gibbs^2 approach can be used to iteratively partition the data into a set containing a strongly conserved motif and an outstanding set that can be broken into further subsets sharing a common pattern. This allows for the discovery of multiple predictive motifs occurring in subsets of the population.

We propose choosing the working set size by studying how the log-odds likelihood of motifs changes for different selections of C. The average contribution to the log-odds likelihood by each sequence in the working set can be measured using (1) as:

$$\frac{1}{C}\sum_{i=1}^{C}\sum_{j=1}^{W}\log\left[\frac{M(s_i(p_i+j-1),j)}{B(s_i(p_i+j-1))}\right]. \tag{9}$$

As sequences are added to the working set, the average contribution measured in (9) decreases significantly if the addition of a further sequence prevents the working set from sharing a common motif—if the additional sequence does not allow a strong motif to be identified. The size of the working set for the Gibbs^2 algorithm can therefore be determined by searching for a knee in the curve relating the average contribution to the log-odds likelihood by each

sequence with C. This process may be approximated by a binary search to reduce computation.

The use of Gibbs^2 also allows for the performance of the Consensus algorithm from Section 4 to be improved. Specifically, Consensus can be seeded using a strongly conserved pattern obtained by Gibbs^2. This reduces the likelihood that Consensus will be affected by a poor choice of the initial two strings.

6. EVALUATION

6.1 Testing Methodology

We applied our techniques to the Physionet Sudden Cardiac Death Holter Database (SDDB) [Goldberger et al. 2000]. This database contains several hours of ECG data recorded using Holter monitors from 23 patients who experienced sudden cardiac death. The recordings were obtained in the 1980s in Boston area hospitals and were compiled as part of a later study of ventricular arrhythmias. Owing to the retrospective nature of this collection, there are important limitations. Patient information including drug regimens and dosages is limited, and sometimes completely unavailable. Furthermore, sudden cardiac death may result from a variety of underlying causes and it is likely that among the 23 patients there are multiple groups sharing different regulatory factors. Despite these shortcomings, the SDDB ECG signals represent an interesting dataset since they represent a population sharing a common acute event. In addition, the recordings are sufficiently long (up to 24 hours prior to death in some cases) that it is likely the predictive factors occurred during the recording period. Finally, the signals in SDDB are generally well-annotated, with cardiologists providing labels at the level of each beat, and this yields a source of clinically relevant symbols that can be used to search for motifs.

For the 23 SDDB patients TCM, Gibbs sampling, Gibbs^2, and Consensus were used to discover potentially predictive motifs of lengths 4, 10, and 16. Since TCM, Consensus and the variants of the Gibbs sampling algorithms, are stochastic in nature, a hundred runs were executed with the strongest motifs being automatically returned as the solution. The scoring function used was the log-likelihood score described in Section 3.

In each case, the endpoint used to signify the acute event associated with death was the occurrence of ventricular fibrillation (VF). This was annotated for all patients and only regions preceding VF were searched for conserved motifs.

For visualization purposes, we used WebLogo [Crooks et al. 2004] to display the motifs returned by our algorithms. This uses the profile matrix to represent motifs as sequence logos, which are graphical representations consisting of stacks of symbols. For each position in the motif, the overall height of the stack indicates how strongly the motif is conserved at that position, while the height of symbols within the stack indicates the relative frequency of each symbol at that position. For example, for the length 10 motif in Figure 4, the sequence logo shows that the motif is strongly conserved at positions 8 and 10, where the predictive sequence was found to contain normal beats across patients. The

motif is also conserved at positions 1, 3, and 5, where ventricular activity was seen for most patients, with some occurrences of normal beats (position 1) and supraventricular beats (positions 3 and 5) as well.

For position j in the motif, the height of symbol i at that location is given by:

$$M(i, j)[2 - H(j)], \tag{10}$$

where:

$$H(j) = -\sum_{k} M(k, j)\log_2(M(k, j)). \tag{11}$$

For Consensus, where a profile matrix is not explicitly constructed, the best-matching subsequences were used to derive a profile that could be represented using WebLogo. This allowed for results to be consistently visualized, irrespective of the algorithm used to discover motifs.

More information on WebLogo can be found at their Web site.[1]

6.2 Data Reduction

The transformations discussed in Section 5 can be evaluated in terms of the data compression realized using these approaches. This allows for an appreciation of the extent to which the original data contains long runs of normal activity that can be compacted. The original sequences across the 23 patients contained 1,216,435 symbols in total, each corresponding to a single beat annotated by a skilled cardiologist. Using the notion of active regions and stripping away uninteresting normal motifs reduced the size of the data to 257,479 characters—a reduction of 78.83%.

6.3 TCM, Gibbs Sampling, and Consensus

Figures 4–6 present the results returned by TCM, Gibbs sampling, and Consensus, as sequence logos. Commonly occurring labels are N = normal, V = premature ventricular contraction, and S = supraventricular premature or ectopic beats.

The motifs discovered by all three algorithms were similar and comprised runs of premature ventricular contractions. For each choice of motif length, TCM returned more strongly conserved motifs than both Gibbs sampling and Consensus. This can be explained by the fact that TCM scores all starting positions in every sequence during each iteration, and is stochastic only in the choice of an initial profile matrix. It employs significantly more computation than either Gibbs sampling or Consensus and is able to find more strongly conserved patterns as a result. On the other hand, the Gibbs sampling algorithm depends on both a random set of initial starting positions and probabilistic choices during each iteration to select a string, s_i, and a new starting position within that string. Consensus is similar to TCM in that it is stochastic only in its initial choice of sequences to use as seed, but unlike TCM, where a poor initial

[1]http://weblogo.berkely.edu

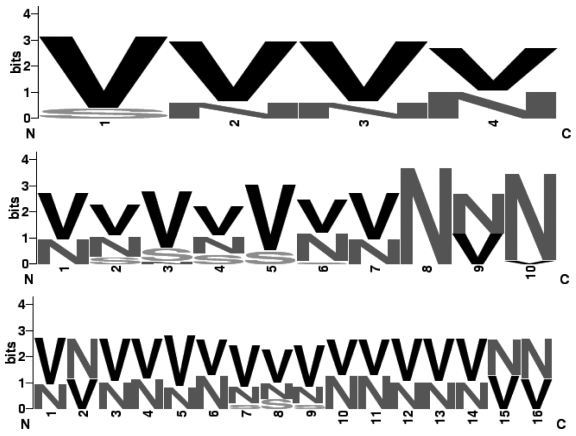


Fig. 4. Motifs of length 4, 10, and 16 found using TCM.

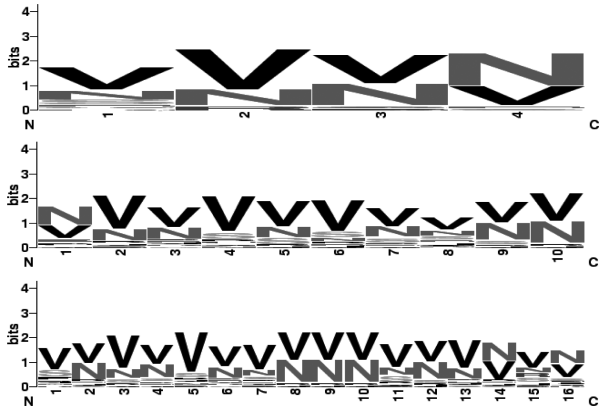


Fig. 5. Motifs of length 4, 10, and 16 found using Gibbs sampling.

choice can be corrected during subsequent iterations, in the case of Consensus, the effects of a poor initial choice propagate all the way through.

Although TCM produced the best results in this case, the process of scoring every starting position in each sequence was considerably more time consuming and took an order of magnitude more time than either Gibbs sampling or Consensus.

6.4 Gibbs^2 and Seeded Consensus

Figure 7 shows the motifs discovered by the Gibbs^2 algorithm with an initial working set of size 12, containing sequences chosen at random. The size of the initial working set was determined from the average contribution of each sequence to the log-odds likelihood of the best scoring motif, as described in Section 5.2. Figure 8 illustrates how the average contribution of the log-odds likelihood changed with increased values of C.

In this case, the predictive motif once again found comprised runs of premature ventricular contractions, but was more strongly conserved than the best

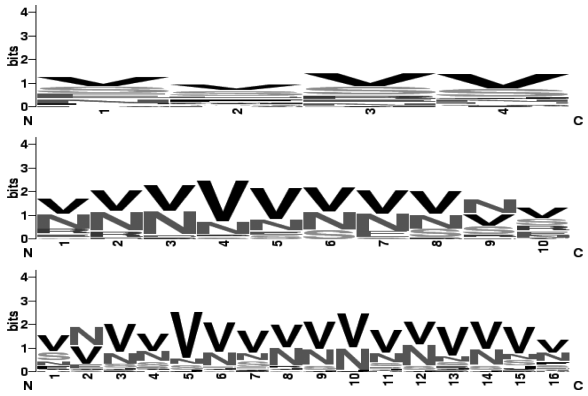


Fig. 6. Motifs of length 4, 10, and 16 found using Consensus.

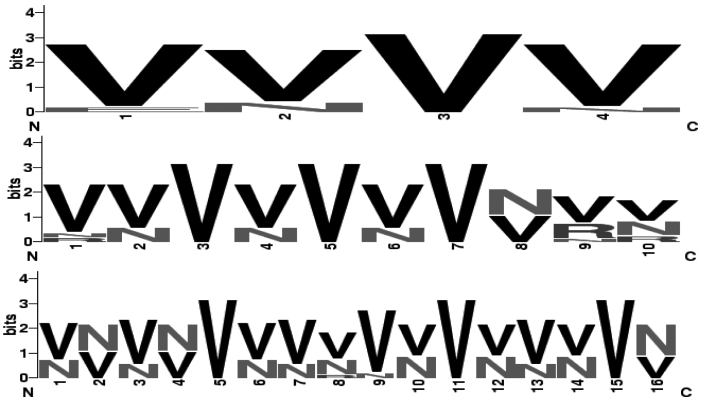


Fig. 7. Motifs of length 4, 10, and 16 found using Gibbs^2.

results produced earlier by TCM, Gibbs sampling, and Consensus. Specifically, comparing Figures 4–7, the stack of symbols in Figure 7 shows the premature ventricular activity figuring more prominently at positions within the motifs.

This effect may be attributed to the ability of Gibbs^2 to select a group of patients who had matching motifs comprising premature ventricular activity, unlike TCM, Gibbs sampling, and Consensus, which were constrained to find a less conserved intermediate that was a best fit for data from all the different patients in the population. For this reason, Gibbs^2 provided an improvement not only over the original Gibbs sampling algorithm but also the more computationally intensive TCM. The Gibbs^2 algorithm has the same basic structure as the original Gibbs sampling technique, but is able to outperform TCM by addressing the issue of subsets of the population exhibiting different regulatory activity. Section 6.5 explores this aspect of subset-based algorithms in more detail, and motivates the idea of searching for different motives in subpopulations iteratively.

Figure 9 presents the result of using Seeded Consensus to detect motifs of length 4 relative to the original Consensus algorithm. In this case, the Gibbs^2 algorithm with a working set of size 5 was used to first find an initial seed for

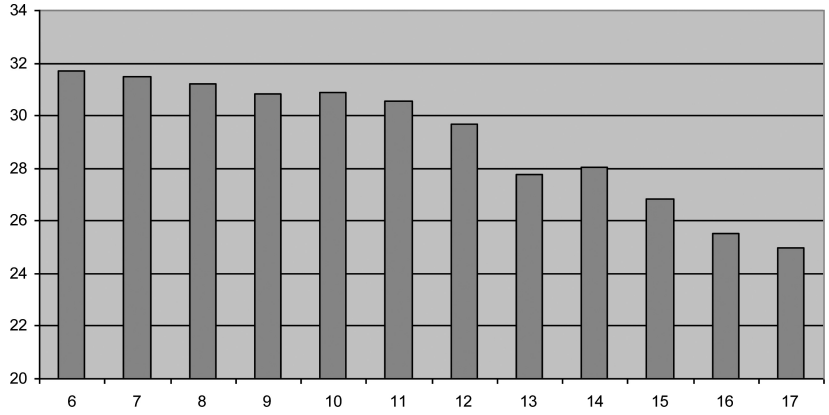


Fig. 8. Relation of the average contribution of each sequence to the log-odds likelihood for the best scoring motif with increasing values of C.

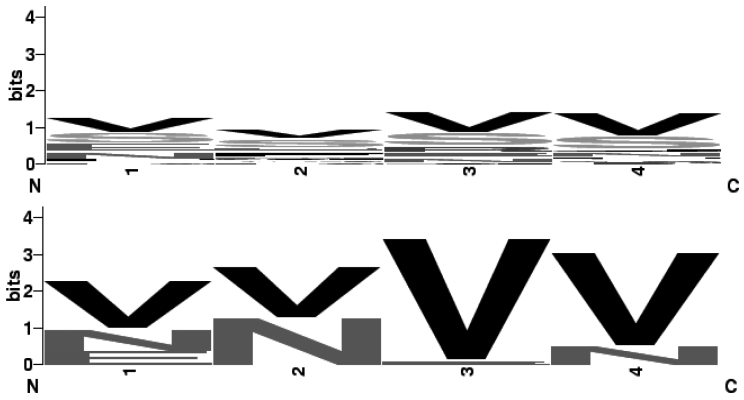


Fig. 9. Motifs of length 4 found using Consensus (top) and Seeded Consensus (bottom).

the Consensus algorithm. As the data shows, Seeded Consensus produced more strongly conserved results than the original Consensus algorithm. This effect followed from reducing the chance that a poor initial choice of sequences would propagate and adversely affect the search for motifs.

The motif found using Seeded Consensus in Figure 9 is not as strongly conserved as the one discovered by Gibbs^2 in Figure 7. This can be explained by the fact that Seeded Consensus uses Gibbs^2 to discover an initial seed but otherwise still operates on all the sequences in the data. The issue of motifs occurring only in a subset of patients does not therefore get addressed, although Seeded Consensus is still able to produce results that are comparable with TCM without the need for intensive computation.

The results of these experiments suggest that subset based techniques using Gibbs^2 either to search for motifs directly, or for the purpose of providing seeds that can be fed into the Consensus algorithm, may allow for more strongly conserved motifs to be discovered than through use of TCM, Gibbs sampling, and the original Consensus algorithm. Moreover, the improvement provided by the

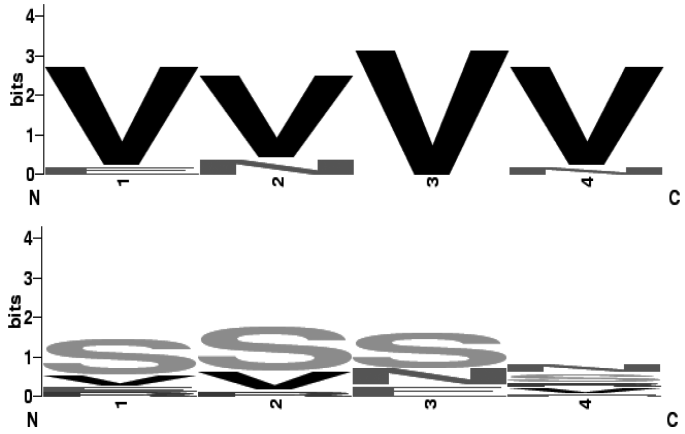


Fig. 10. Two-stage Gibbs^2 motifs of length 4. The top motif comprises a working set of size 12, while the second motif corresponds to those 11 sequences (from a total population of 23) that were not included in the original working set.

Gibbs^2 algorithm proposed in our work is not associated with a significant computational overhead. In addition, the ability to partition the data into groups with homogenous motifs allows for the discovery of more than one predictive pattern, each of which may be associated with the outcome in a different group of patients. We explore this idea in more detail in the next section.

6.5 Two-Stage Gibbs^2

For the motif of length 4, the sequences remaining outside the working set at the termination the of Gibbs^2 algorithm were searched for a second motif common to this group. Figure 10 shows the results of this approach.

In this case, a second motif comprising runs of supraventricular premature or ectopic beats was found among this subgroup of the population. Notably, these patients did not show a motif similar to the ones found earlier, comprising premature ventricular beats, during any of the multiple executions of the motif discovery algorithm. This suggests that the subset of patients left outside the working set by Gibbs^2 did not exhibit regulatory activity similar to the ones for whom a premature ventricular motif was discovered. Including these patients in the search for a predictive motif, as would be the case for non-subset-based techniques, would therefore lead to a less informative motif and would obscure the fact that different groups of patients show varied predictive patterns associated with an endpoint.

6.6 Motif-Event Delay

Using the motif of length 10 shown in Figure 7, for each sequence, the time delay between the starting location of the motif: p_i, and the clinical endpoint (the occurrence of VF in the patients) was calculated for the Gibbs^2 algorithm. For one of the 23 patients in the dataset, the motif occurred less than a minute prior to the event itself. In all other cases, the motif discovered preceded the actual event by at least 20 minutes or more. The median motif-event delay was

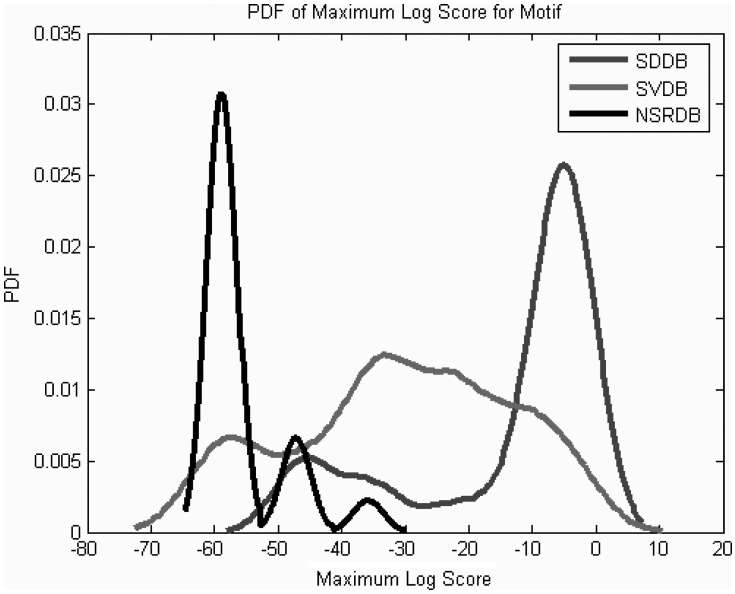


Fig. 11. Motif-matching scores for patients in the Sudden Death Database (SDDB), Supraventricular Arrhythmia Database (SVDB), and Normal Sinus Rhythm Database (NSRDB). The graph shows the probability distributions estimated using kernel density estimation.

60 minutes, while the 25% and 75% quartile times were 42 and 179 minutes respectively. The maximum time separation of the motif and the event was 604 minutes.

These results suggest that the motif occurred sufficiently in advance of the endpoint not to be considered merely an extension of the final event itself. Furthermore, the fact that the motif may occur at a wide range of times prior to the endpoint reinforces the need to carry out the search for predictive patterns in an automated manner, which is able to relate information across a range of positions within each sequence.

6.7 Comparison with Controls

For each patient in the SDDB population, the log-likelihood score was calculated for each starting position in the ECG label sequence. The overall score for the patient was the maximum log-likehood score found. Intuitively, this strategy assigns each patient the risk score associated with the occurrence of the discovered motif of length 10 shown in Figure 7 at any point during the recording: if activity similar to the motif associated with sudden death is seen anywhere, the patient is assumed to be at higher risk for the event.

Figure 11 shows the probability density function that can be estimated from the scores for the SDDB population. A similar strategy was adopted to score patients in two control datasets: the Physionet Normal Sinus Rhythm Database (NSRDB) and the Physionet Supraventricular Arrhythmia Database (SVDB). The decision to use SVDB data in addition to normal individuals was owing to the fact that the SVDB signals contained the same labels as the SDDB data with

a higher background frequency of abnormal symbols. This ensured that a difference in scores across populations did not result from an absence of labels, but more so because activity was organized in different forms. Specifically, 1.45% of the beats in the SDDB data were premature ventricular contractions. By comparison, 5.39% of the beats in the SVDB signals and 0.002% of the NSRDB beats fell into the same category. This suggests that if the motifs seen in the SDDB population were random occurrences, then they would be expected to be seen more frequently in the SVDB dataset. With this in mind, the fact that SVDB patients had a higher percentage of premature ventricular activity but still scored lower on the discovered motifs provides further indication that the motif corresponded to activity that was not purely a random occurrence in the sudden death population.

Using a maximum likelihood separator, we were able to use our motif to correctly identify 70% of the patients who suffered sudden cardiac death during 24 hours of recording while classifying none of the normal individuals, and only 8% of the patients from the supraventricular dataset as being at risk. The small number of patients in the dataset, however, does not allow for us to make statistically significant clinical statements about these findings.

7. RELATED WORK

In this section, we review existing knowledge-discovery work to detect potentially predictive activity. A discussion of aspects of our work extending computational biology techniques appears earlier in Sections 3 and 4.

An extensive literature exists in the areas of data mining and machine learning on the subject of prediction. A common approach is to infer prediction rules from data of the form:

$$\text{IF } cond_1 \text{ AND } \ldots \; cond_i \; \ldots \text{ AND } cond_m \text{ THEN } pred.$$

These rules correspond to a set of conditions associated with a specific outcome. The challenge in this case is to select conditions that are able to distinguish between whether an event occurs or not, but do not overfit available training data. A number of different techniques exist for this purpose, ranging from decision trees [Helmbold and Schapire 1997] to more recent work using evolutionary algorithms [Freitas 2001].

We supplement this work by finding precursors that exist at a lower level of the data. As an alternative to rules based on the outcomes of a series of diagnostic tests or a sophisticated feature set, we attempt to find interesting patterns by analyzing the specific sequences a system moves through. The motivation for such an approach is provided by the ever-increasing amounts of data collected in various fields, for example, medicine, geodesic studies, space, and planetary sciences. In many of these cases, well-formulated predictive attributes do not exist. Unsupervised techniques can, however, be used to decompose signals into stationary or periodic tokens. These can then be assigned labels to reexpress the original data as a sequence of symbols. Our work allows for the discovery of a specific class of regulatory activity (occurring as subsequences) in this representation without assuming higher-level features for classification.

The idea of transforming time-series data into symbols has been proposed in different forms earlier (an excellent review on this subject can be found in Daw et al. [2003] with some additional techniques for symbolization described in Giles and Lawrence [2001]; Lin et al. [2003]; and Syed et al. [2007]). We build upon this work and use symbolization in the context of large physiological datasets to make the search for predictive patterns more efficient.

While symbolization is an integral component of our approach, we note that there has also been much promising work in recent years on the discovery of motifs directly in time-series signals. These approaches search for patterns either in the raw signal without any transformation [Chiu et al. 2003], or by using piecewise aggregate approximation (PAA) [Patel et al. 2002; Lin et al. 2002] to achieve a simple yet efficient symbolization. The focus in both these cases is on the discovery of patterns that occur frequently in the data. These methods do not explicitly address the goal of prediction and are not designed to find patterns that may occur infrequently but have a consistent regulatory effect. The discovery of such patterns, e.g., those associated with an acute endpoint, is a potentially more computationally intensive problem, whereas symbolization plays an important role in making the search process more scalable.

Our pattern discovery algorithms add to an extensive body of existing work to analyze symbolic sequences. A fairly rich literature can be found on techniques to discover local patterns [Mannila et al. 1999; Jin et al. 2002], frequent sequential associations [Harms et al. 2002], and generalized episodes [Mannila et al. 1997]. These methods search for sequences of discrete events or symbols that occur frequently in the data with a partially specified order. This information can help make future predictions about the behavior of symbolic sequences. For example, the observation that symbols A and B frequently occur in the data and are always followed by symbol C allows for the symbol C to be predicted whenever A and B are next encountered. Such techniques address a similar goal to our work and offer the advantage of finding more general predictive patterns than our methods. However, these methods are fundamentally focused on the discovery of frequent patterns in single symbolic sequences and on learning structure between symbols that occur close to each other. This is in contrast to the techniques proposed in our work, which searches for patterns that may not be frequent but are associated with an endpoint in multiple symbolic sequences, and where the symbols constituting the predictive pattern and the acute event can be far apart. In this way, the motifs discovered by our algorithms can be viewed as global patterns, spanning both a subsequence of symbols and an annotated event that may be variably delayed.

Our analysis of sequential signals is also similar to the use of Markov models to study systems [Durbin et al. 1998]. Our work differs from a purely Markovian approach in that we do not attempt to develop a model explaining the data and focus instead on explicitly identifying predictive elements. Furthermore, in many cases, including the sudden death study conducted as part of this project, the regulatory activity may occur well in advance of the event. Developing a Markov model containing sufficient memory and complexity to handle these cases would prove to be challenging in such situations.

A different form of prediction in learning theory is to approach the task in an online manner and consistently refine a hypothesis based on errors and temporal differences [Cesa-Bianchi and Lugosi 2006]. This approach is similar to the inference of prediction rules in that decisions are made on attributes or features and not individual sequences. Our techniques further differ in that they attempt to exploit the availability of batch data and do not address the issue of online learning.

In addition to suggesting methods to discover motifs, we propose subset-based techniques that can isolate subsets of the data that share common predictive motifs. For example, in Section 6.5, we discuss how two-stage Gibbs^2 can find subpopulations sharing different predictive sequences. This is important since the same event may be associated with different causes. We consider the selection of sequences that share regulatory activity as being internal to the problem of motif discovery. Specifically, partitioning the data in advance without information on the specific predictive pattern is difficult—subsets of the sequences sharing a motif can only be isolated once the motif is known. For this reason, we address the issue of degeneracy and heterogeneous predictive patterns as part of motif discovery and tailor our algorithms to automatically recognize and handle these cases.

A key component of our approach to address motif degeneracy is to randomly swap symbolic sequences into and out of a working subset. The notion of randomized swaps has also been proposed earlier, in the context of data mining to assess the significance of mining results on high-dimensional data [Gionis et al. 2006]. The focus in that work is on randomly swapping 0-1 features within a matrix to create a new matrix with preserved row and column margins that can be used for testing. The process of random swapping does not attempt to isolate a subset of features with desired properties but is used to randomly perturb the matrix instead. Finally, our work is similar to unary classification techniques [Scholkopf et al. 2001] in that the algorithms proposed do not require the presence of both positive and negative examples. Instead, they are geared towards selecting subsequences of labels that can be found across a population in a form unlikely to occur purely by chance. The goal is to better understand similarities that can be analyzed for a predictive relationship with the acute event being considered.

8. SUMMARY AND CONCLUSIONS

In this article, we propose a framework for discovering potentially predictive activity preceding acute events. We generalize the notion of regulatory motifs from computational biology and adapt existing algorithms to operate robustly and efficiently on a broad set of data. We develop and evaluate this work in the context of physiological signals, detailing the challenges associated with fitting a motif-detection model to signals other than DNA. We also describe the performance of subset-based techniques to discover activity associated with sudden cardiac death, comparing discovered patterns against control populations comprising normal individuals and those with supraventricular arrhythmias.

Our work represents a fully-automated approach for discovering a specific class of possible precursors: patterns that are sequential in that a given ordering of different classes is associated with an end result. We impose no restrictions on the patterns to be discovered, and our tools are able to identify sequences of arbitrary complexity that occur in a possibly degenerate form across a population sharing an event.

A central requirement for the techniques described in this article is that the data being mined is symbolic. In the context of physiological signals, this requires transforming continuous waveforms into alphabetical sequences. Creating a set of labels that can be applied to the data can be achieved in a number of different ways. In the work described here, we use clinical labels that have a fixed meaning and can be applied across patients. It is possible that potentially predictive activity may occur at a more subtle level, where differences within clinical classes are important. For this reason, an important future direction of this research is to extend approaches to annotate signals in a patient-specific, data-derived manner to achieve symbolization.

Finally, it is important to stress that although our initial results on detecting a predictive pattern associated with sudden cardiac death appear promising, the small number of patients in the dataset and limited patient histories means that further investigation on a larger set of ECG signals is necessary.

ACKNOWLEDGMENTS

We would like to thank Ebad Ahmed for helping implement the software routines used in this work. We also thank the editors and the anonymous reviewers for their feedback in helping improve the presentation of this work.

REFERENCES

Bailey, T. and Eklan, C. 1995. The value of prior knowledge in discovery motifs with MEME. In *Proceedings of the International Conference on Intelligence Systems in Molecular Biology*. 21–29.

Chiu, B., Keogh, E., and Lonardi, S. 2003. Probabilistic discovery of time series motifs. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Cesa-Bianchi, N. and Lugosi, G. 2006. *Prediction, Learning and Games*. Cambridge University Press, Cambridge, UK.

Crooks, G., Hon, G., Chandonia, J., and Brenner, S. 2004. WebLogo: a sequence long generator. *Genome Res. 14*, 1188–1190.

Daw, C., Finney C., and Tracy, E. 2003. A review of symbolic analysis of experimental data. *Rev. Sci. Instr. 74*, 915–930.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. 1998. *Biological Sequence Analysis*. Cambridge University Press, Cambridge,UK.

Freitas, A. 2001. *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Springer-Verlag, Berlin, Germany.

Gert, T., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouze, P., and Moreau Y. 2002. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J. Comput. Biol. 9*, 447–464.

Giles, C. and Lawrence, S. 2001. Noisy time series prediction using a recurrent neural network and grammatical inference. *Mach. Learn. 44*, 161–183.

Gionis, A., Mannila, H., Mielikainen, T., and Tsaparas, P. 2006. Assessing data mining results via swap randomization. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

GOLDBERGER, A., AMARAL, L., GLASS, L., HAUSDORFF, J., IVANOV, P., MARK, R., MIETUS, J., MOODY, G., PENG, C., AND STANLEY, H. 2000. Components of a new research resource for complex physiologic signals. *Circulation 101*, 215–220.

HARMS, S., DEOGUN, J. AND TADESSE, T. 2002. Discovering sequential association rules with constraints and time lags in multiple sequences. In *Proceedings of the 13th International Symposium on Foundation of Intelligent Systems*.

HELMBOLD, D. AND SCHAPIRE, R. 1997. Predicting nearly as well as the best pruning of a decision tree. *Mach. Learn. 27*, 51–68.

JIN, X., WANG, L., LU, Y., AND SHI, C. 2002. Indexing and mining of the local patterns in sequence database. In *Proceedings of the 3rd International Conference on Intelligent Data Engineering and Automated Learning*.

JONES, N. AND PEVZNER, P. 2004. *An Introduction to Bioinformatics Algorithms*. The MIT Press. Cambridge, MA.

KELLIS, M., PATTERSON, N., ENDRIZZI, M., BIRREN, B., AND LANDER, E. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature 423*, 241–254.

LIN, J., KEOGH, E., LONARDI, S., AND PATEL, P. 2002. Finding motifs in time series. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

LIN, J., KEOGH, E., LONARDI, S., AND CHIU, B. 2003. A symbolic representation of time series with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*.

MANNILA, H., TOIVONEN, H., AND VERKAMO, A. 1997. Discovery of frequent episodes in event sequences. *Data Mining and Knowl. Discov. 1*, 259–289.

MANNILA, H., PAVLOV, D., AND SMYTH, P. 1999. Prediction with local patterns using cross-entropy. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

PATEL, P., KEOGH, E., LIN, J., AND LONARDI, S. 2002. Mining motifs in massive time series databases. In *Proceedings of the International Conference on Data Mining*.

SCHOLKOPF, B., PLATT, J., SHAWE-TAYLOR, J., SMOLA, A., AND WILLIAMSON, R. 2001. Estimating the support of a high-dimensional distribution. *Neural Comput. 13*, 1443–1471.

STORMO, G. AND HARTZELL, G. 1989. Identifying protein-binding sites from unaligned DNA fragments. *Proc. Nat. Acad. Sciences 86*, 1183–1187.

SYED, Z., STULTZ, C., AND GUTTAG, J. 2007. Clustering and symbolic analysis of cardiovascular signals: Discovery and visualization of medically relevant patterns in long-term data using limited prior knowledge. *EURASIP J. Advances Sig. Proc. 2007*, Article ID 67938.

Received February 2008; revised September 2008; accepted May 2009

selective transport would involve a coupling between cofactor binding and pore opening. A similar conformational change was reported recently in a carboxysome shell protein, CsoS1D (*23*) (figs. S4 and S5), suggesting that gated transport could be a common mechanism in microcompartments.

Like some other BMC proteins, EutL was found to associate into tightly packed molecular layers within both of the crystal forms characterized here (Fig. 3B). This observation lends further support to models for microcompartment architecture in which the shell is formed by the tight side-by-side packing of hexameric (or pseudohexameric) protein building blocks into a molecular layer or sheet (*9*–*11*). The appearance of the EutL protein layer, composed of subunits in either the open or closed configuration, illustrates how strongly the porosity of the microcompartment would be affected by the conformational change in EutL (Fig. 3B).

The fourth shell protein, EutK, is distinct among the shell proteins in the Eut microcompartment. First, although all other BMC proteins studied to date form hexamers (or pseudohexamers, like EutL), EutK is a monomer in solution (fig. S8). The apparent inability of EutK to assemble into a hexamer by itself suggests that different BMC paralogs might form mixed hexamers during assembly of the shell. Second, EutK has an extra protein domain of about 60 amino acids following the conserved BMC-type domain (Fig. 4A). Numerous instances of BMC-type proteins fused to other uncharacterized protein domains are evident in the protein sequence databases, but structural and functional data are limited (*27*). Based on sequence similarity of marginal statistical certainty (for example, a 40% probability of similarity due to random chance), the extra domain of EutK could be only tentatively assigned to a broad family of proteins bearing a well-known helix-turn-helix motif, which is common among nucleic acid–binding proteins.

Whereas crystals could not be grown using the full-length EutK protein, the crystal structure of the C-terminal domain was elucidated at a resolution of 2.1 Å. The crystal structure of this 59–amino acid fragment, referred to hereafter as EutK-Ctail, demonstrates that it is indeed a helix-turn-helix domain (Fig. 4B). A computational search for similar structures in the protein structure database identified many known helix-turn-helix domains as close matches, with coordinate differences as low as 1.0 Å. About 90% of the helix-turn-helix domain structures retrieved from the search bind nucleic acids, whereas a minority perform other varied cellular functions (SOM text). A comparison of the surface features of EutK-Ctail with previously characterized helix-turn-helix domains, including those that bind nucleic acids and those that do not, suggests strongly that EutK is a nucleic acid–binding protein. A prominent, positively charged surface is conserved in EutK-Ctail and in those domains that bind nucleic acids (Fig. 4C). Interestingly, on the basis of structure-guided alignments, EutK-Ctail shows the highest sequence similarity to the helix-turn-helix domains of the Ic1R family of transcription factors, which regulate, among other things, carbon metabolism in enterobacteria (*28*, *29*) (fig. S9). The specific function of EutK and the identity of its cognate nucleic acid are unknown. Nonetheless, the key prediction that the Eut microcompartment shell binds a nucleic acid points up the possibility of unexpected parallels to viral capsids, which bind to and encapsulate their viral genomes.

Taken together, the structures of the shell proteins from the Eut microcompartment paint a picture of a complex, highly evolved organelle. The structural features and conformational changes observed illustrate how these evolutionarily related proteins have diverged to fulfill specialized architectural and biochemical roles in a shell that participates actively in the function of the microcompartment.

References and Notes

1. G. C. Cannon *et al.*, *Appl. Environ. Microbiol.* **67**, 5351 (2001).
2. T. O. Yeates, C. A. Kerfeld, S. Heinhorst, G. C. Cannon, J. M. Shively, *Nat. Rev. Microbiol.* **6**, 681 (2008).
3. S. Cheng, Y. Liu, C. S. Crowley, T. O. Yeates, T. A. Bobik, *Bioessays* **30**, 1084 (2008).
4. E. Kofoid, C. Rappleye, I. Stojiljkovic, J. Roth, *J. Bacteriol.* **181**, 5317 (1999).
5. P. Chen, D. I. Andersson, J. R. Roth, *J. Bacteriol.* **176**, 5474 (1994).
6. J. M. Shively *et al.*, *Int. Rev. Cytol.* **113**, 35 (1988).
7. J. M. Shively, F. Ball, D. H. Brown, R. E. Saunders, *Science* **182**, 584 (1973).
8. M. R. Badger, G. D. Price, *J. Exp. Bot.* **54**, 609 (2003).
9. C. A. Kerfeld *et al.*, *Science* **309**, 936 (2005).
10. Y. Tsai *et al.*, *PLoS Biol.* **5**, e144 (2007).
11. S. Tanaka *et al.*, *Science* **319**, 1083 (2008).
12. S. Tanaka, M. R. Sawaya, M. Phillips, T. O. Yeates, *Protein Sci.* **18**, 108 (2009).
13. C. V. Iancu *et al.*, *J. Mol. Biol.* **372**, 764 (2007).
14. M. F. Schmid *et al.*, *J. Mol. Biol.* **364**, 526 (2006).
15. G. D. Havemann, E. M. Sampson, T. A. Bobik, *J. Bacteriol.* **184**, 1253 (2002).
16. J. M. Shively *et al.*, *Can. J. Bot.* **76**, 906 (1998).
17. M. R. Rondon, A. R. Horswill, J. C. Escalante-Semerena, *J. Bacteriol.* **177**, 7119 (1995).
18. E. M. Sampson, T. A. Bobik, *J. Bacteriol.* **190**, 2966 (2008).
19. M. R. Rondon, R. Kazmierczak, J. C. Escalante-Semerena, *J. Bacteriol.* **177**, 5434 (1995).
20. J. T. Penrod, J. R. Roth, *J. Bacteriol.* **188**, 2865 (2006).
21. R. S. English, S. C. Lorbach, X. Qin, J. M. Shively, *Mol. Microbiol.* **12**, 647 (1994).
22. C. S. Crowley, M. R. Sawaya, T. A. Bobik, T. O. Yeates, *Structure* **16**, 1324 (2008).
23. M. G. Klein *et al.*, *J. Mol. Biol.* **392**, 319 (2009).
24. M. Sagermann, A. Ohtaki, K. Nikolakakis, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 8883 (2009).
25. D. L. Caspar, A. Klug, *Cold Spring Harb. Symp. Quant. Biol.* **27**, 1 (1962).
26. M. Carrillo-Tripp *et al.*, *Nucleic Acids Res.* **37**, D436 (2009).
27. J. B. Parsons *et al.*, *J. Biol. Chem.* **283**, 14366 (2008).
28. A. J. Molina-Henares, T. Krell, M. Eugenia Guazzaroni, A. Segura, J. L. Ramos, *FEMS Microbiol. Rev.* **30**, 157 (2006).
29. R. G. Zhang *et al.*, *J. Biol. Chem.* **277**, 19183 (2002).

30. We thank T. Bobik, members of the Yeates lab, J. Escalante-Semerena, F. Guo, M. Faller, and Y. Chen for helpful discussions and M. Phillips for equilibrium sedimentation experiments. This work was supported by the Biological and Environmental Research program of the Department of Energy Office of Science and by NSF grant MCB-0843065. Coordinates and structure factors of the EutK-Ctail, EutL open form, EutL closed form, EutM, EutS, and EutS-G39V have been deposited in the Protein Data Bank under accession numbers 3I71, 3I87, 3I82, 3I6P, 3I96, and 3IA0, respectively.

Supporting Online Material

www.sciencemag.org/cgi/content/full/327/5961/81/DC1
Materials and Methods
SOM Text
Figs. S1 to S9
Tables S1 to S3
References

22 July 2009; accepted 22 October 2009
10.1126/science.1179513

The Tasmanian Devil Transcriptome Reveals Schwann Cell Origins of a Clonally Transmissible Cancer

Elizabeth P. Murchison,[1,2]*† Cesar Tovar,[3] Arthur Hsu,[4] Hannah S. Bender,[1,2] Pouya Kheradpour,[5] Clare A. Rebbeck,[1] David Obendorf,[3] Carly Conlan,[1] Melanie Bahlo,[4] Catherine A. Blizzard,[3] Stephen Pyecroft,[6] Alexandre Kreiss,[3] Manolis Kellis,[5,7] Alexander Stark,[5,7]‡ Timothy T. Harkins,[8] Jennifer A. Marshall Graves,[2] Gregory M. Woods,[3] Gregory J. Hannon,[1] Anthony T. Papenfuss[4]

The Tasmanian devil, a marsupial carnivore, is endangered because of the emergence of a transmissible cancer known as devil facial tumor disease (DFTD). This fatal cancer is clonally derived and is an allograft transmitted between devils by biting. We performed a large-scale genetic analysis of DFTD with microsatellite genotyping, a mitochondrial genome analysis, and deep sequencing of the DFTD transcriptome and microRNAs. These studies confirm that DFTD is a monophyletic clonally transmissible tumor and suggest that the disease is of Schwann cell origin. On the basis of these results, we have generated a diagnostic marker for DFTD and identify a suite of genes relevant to DFTD pathology and transmission. We provide a genomic data set for the Tasmanian devil that is applicable to cancer diagnosis, disease evolution, and conservation biology.

Devil facial tumor disease (DFTD) is a transmissible cancer affecting the Tasmanian devil (*Sarcophilus harrisii*), an endemic Tasmanian marsupial carnivore. First observed in 1996 in northeastern Tasmania, DFTD has been implicated in devil population collapse (*1*, *2*). DFTD is a rapidly fatal disease that culminates in large tumors, primarily on the face

Downloaded from www.sciencemag.org on October 12, 2010

and mouth, which frequently metastasize to internal organs (*3*). There are no diagnostic tests, treatments, or vaccines available for DFTD, and models predict that the disease could lead to extinction of wild Tasmanian devils within 25 to 35 years (*4*).

DFTD appears to be a clonal cell line, transmitted (by biting) as an allograft between devils (*5*, *6*) and may be similar in transmission to canine transmissible venereal tumor (CTVT) and a transmissible sarcoma affecting Syrian hamsters (*7*–*9*). The prevalence and biology of such somatic cell parasites is generally unknown (*10*).

Studies of captive Tasmanian devils have suggested that the species is prone to developing tumors, particularly carcinomas (*11*, *12*). However, DFTD does not resemble previously described devil cancers (*3*, *13*), and determining its etiology is critical for developing management strategies for the disease. Cytologically, DFTD appears as a soft tissue neoplasm consisting of undifferentiated round to spindle-shaped cells with few defining ultrastructural features (*3*, *13*). Immunohistochemistry suggests that the tumor is derived from neuroectoderm (*13*).

Clonal transmission of DFTD was proposed on the basis of karyotypic evidence (*5*) and was supported by genetic analysis of DFTD tumors at microsatellite and major histocompatibility complex loci (*6*). We genotyped at 14 microsatellite loci 25 paired tumor and host samples, as well as 10 samples from DFTD-unaffected devils from 16 locations in Tasmania (*14*) (figs. S1 and S2 and table S1). All DFTD tumors shared a similar genotype across all loci, regardless of location, sex, or age of the animal ($P = 0.18$, permutation test) (figs. S1 and S2). Furthermore, the tumor genotype was distinct from that of both the hosts and unaffected devils ($P < 0.001$, permutation test) (figs. S1 and S2). These data were consistent with previous studies (*5*, *6*) and support the supposition that DFTD is a single clonal cell line propagated as a tumor allograft.

To further assess the clonal origin of DFTD, we sequenced a 1180–base pair fragment of the mitochondrial locus control region (LCR) from 14 tumors, 14 hosts, and 9 DFTD-unaffected devils (table S2). We found that all devils and tumors shared a single LCR haplotype in this region, except for one single-nucleotide polymorphism at position 15,711, which supported the idea that the tumors are clonal. Furthermore, this nucleotide variant was observed only in DFTD-free devils from western Tasmania (figs. S1 and S2 and table S2), a genetically distinct population (*15*). The karyotypic and genetic consistency between DFTD tumors (figs. S1 and S2) (*5*, *6*) reinforces epidemiological evidence for a recent origin of DFTD (*1*, *3*).

We cloned and deeply sequenced microRNA (miRNA) from 10 devil tissues and five DFTD tumors, including one DFTD mammary metastasis, to gain insight into the histogenesis of DFTD. The 114 Tasmanian devil miRNAs, identified with strict conservation parameters, showed characteristic tissue-specific expression profiles (Fig. 1 and table S3). DFTD had a relatively consistent miRNA profile that was distinct from the other 10 tissues (Fig. 1 and fig. S3). Differential expression of four miRNAs in DFTD relative to a non-DFTD tissue (testis) was confirmed by quantitative polymerase chain reaction (PCR) (fig. S4). Hierarchical clustering based on Pearson's correlation statistic showed that the DFTD tumors were clustered (Fig. 1 and table S4) and that the non-DFTD miRNA profile most highly correlated with DFTD was that of brain (Fig. 1 and table S4).

In cancer, miRNAs can both promote and suppress tumors, as well as regulate processes including cell proliferation, angiogenesis, and metastasis (*16*). It is noteworthy that the DFTD profile included a number of miRNAs commonly up-regulated in tumors, including *miR-21*, *miR-24*, and *miR-19b* (*16*), plus a miRNA that has been linked to tumor immune evasion (*miR-222*) (*17*) (Fig. 1 and fig. S3). In contrast, DFTD expresses very low levels of *miR-29b* and *miR-126*, two miRNAs suggested to suppress tumors (Fig. 1 and fig. S3) (*16*).

To create a catalog of genes expressed in DFTD, we sequenced the transcriptome of DFTD and testis (chosen because of its expression of a broad range of genes) from an individual Tasmanian devil resulting in 13,665 and 16,438

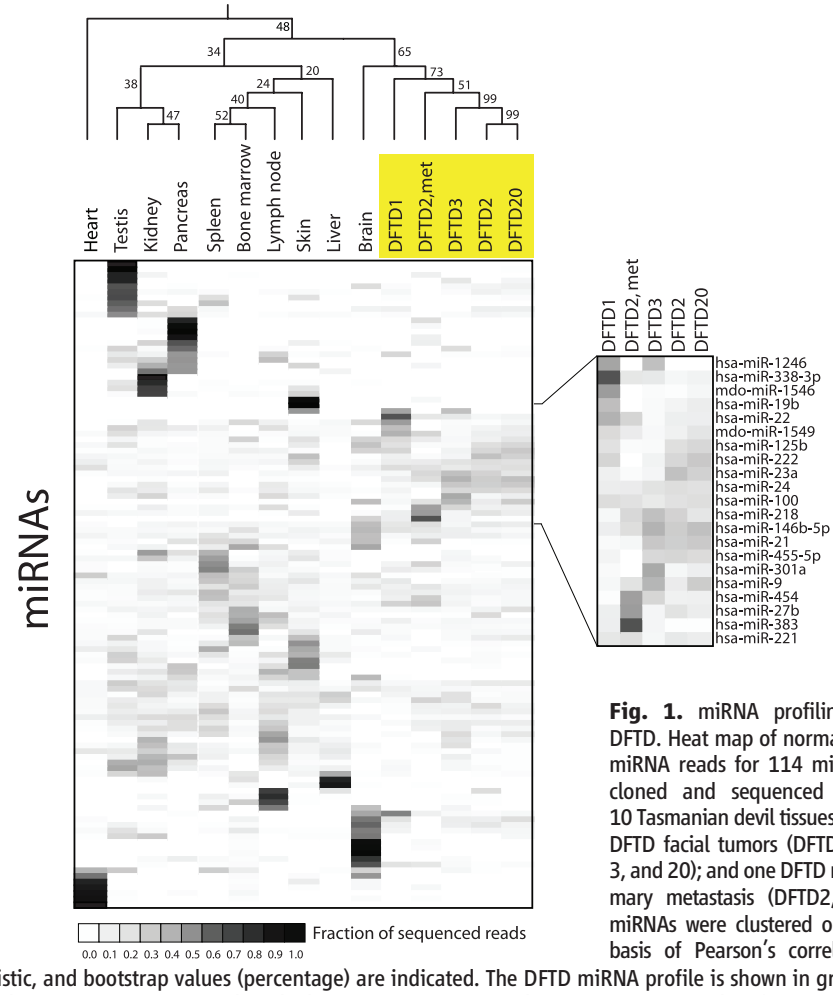


Fig. 1. miRNA profiling of DFTD. Heat map of normalized miRNA reads for 114 miRNAs cloned and sequenced from 10 Tasmanian devil tissues; four DFTD facial tumors (DFTD1, 2, 3, and 20); and one DFTD mammary metastasis (DFTD2,met). miRNAs were clustered on the basis of Pearson's correlation statistic, and bootstrap values (percentage) are indicated. The DFTD miRNA profile is shown in greater detail. miRNAs were annotated on the basis of conservation with comparison species hsa, human; mdo, opossum.

[1]Watson School of Biological Sciences, Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA. [2]Research School of Biology and the ARC Centre of Excellence in Kangaroo Genomics, The Australian National University, Canberra, ACT 0200, Australia. [3]Menzies Research Institute, The University of Tasmania, Hobart, Tasmania 7000, Australia. [4]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Melbourne, Victoria 3052, Australia. [5]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [6]Mt. Pleasant Animal Health Laboratories, Department of Primary Industries and Water, Kings Meadows, Tasmania 7249, Australia. [7]Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA. [8]Roche Applied Sciences, Indianapolis, IN 46250, USA.

*Present address: Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK.

†To whom correspondence should be addressed. E-mail: elizabeth.murchison@sanger.ac.uk

‡Present address: Research Institute of Molecular Pathology (IMP), Dr. Bohr-Gasse 7, A-1030, Vienna, Austria.

Downloaded from www.sciencemag.org on October 12, 2010

unique transcripts identified on the basis of similarity with human proteins and opossum transcripts, respectively (see SOM text). Of the annotated transcripts, 0.4% were differentially represented between tumor and testis libraries (fig. S5). The 31 transcripts with the most significant enrichment in the tumor library (tumor-to-testis ratio of at least 2.5, chi-squared test $P \leq 0.05$) were validated by semiquantitative PCR (Fig. 2A). Of these, 20 transcripts were at least twice as highly expressed in tumor as in testis (Fig. 2A).

The gene with the highest expression in DFTD relative to a housekeeping gene, *GAPDH* (encoding glyceraldehyde-3-phosphate dehydrogenase), was *MBP*, the gene that encodes myelin basic protein (Fig. 2A). Indeed, 9 of the 20 validated tumor genes (45%) were found to be involved in the myelination pathway (table S6). Myelin, an insulating membranous layer that ensheathes nerve axons, is produced by oligodendrocytes in the central nervous system and by Schwann cells in the peripheral nervous system. Components of a molecular network that controls the differentiation of Schwann cells, including transcription factors (*SOX10*, *SOX2*, *POU5F1*, and *JUN*) and structural myelin genes (*MPZ*, *PRX*, *MBP*, and *PMP22*), were apparent in the differentially expressed DFTD transcriptome (Fig. 2A) (*18*).

We also measured the expression of 31 tumor-enriched genes from a panel of 10 Tasmanian devil tissues by semiquantitative PCR. Hierarchical clustering of this data set grouped DFTD together with peripheral nerves, a tissue enriched for Schwann cells (Fig. 2B). To confirm the expression of myelin proteins in DFTD, we stained DFTD tumor tissues with the antibody against Schwann cell–specific myelin protein, periaxin (PRX) (Fig. 3). All tumor cells in DFTD lesions, as well as the myelinated sheaths of peripheral nerve bundles, were strongly and specifically positive for PRX (Fig. 3). We also detected protein expression in DFTD of myelin proteins MBP, MPZ, and PMP22, as well as NES, NGFR, and S100 (figs. S6 and S7, and table S7). These experiments confirm that DFTD expresses genes found in myelinating cells, including Schwann cell–specific markers (PRX, PMP22, and MPZ).

At present, diagnosis of DFTD is based on clinical features and histology. This can pose particular difficulties for the diagnosis of atypical DFTD such as nonfacial DFTD or DFTD metastases. We tested PRX as a potential diagnostic marker for DFTD by staining a panel of both DFTD and non-DFTD Tasmanian devil tumors with PRX. All (n = 20) of the DFTD tumors were positive for PRX, whereas none of the nine non-DFTD tumors tested were positive (table S7). In addition, all (n = 10) of the DFTD metastases collected from a variety of organs were positive for PRX (table S7). Thus, PRX is a strong and specific marker for DFTD and is suitable for diagnostic evaluation.

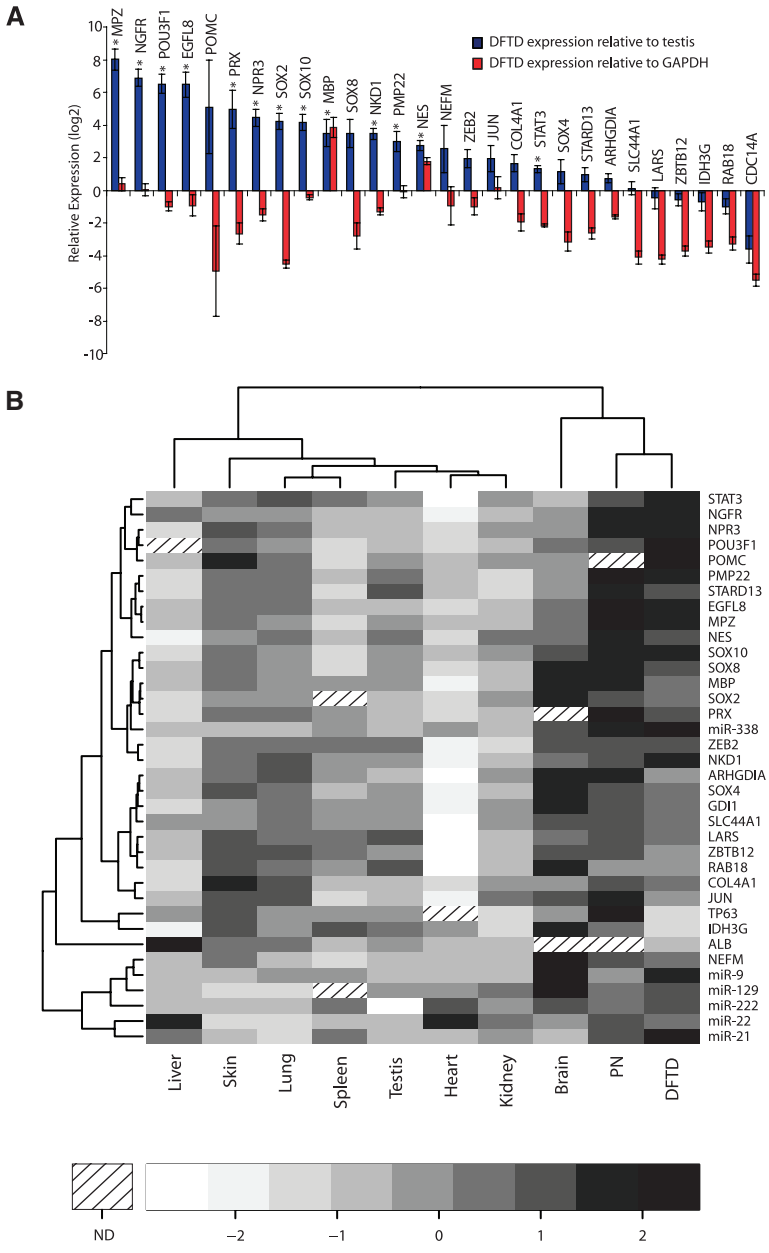


Fig. 2. DFTD transcriptome. (**A**) Semiquantitative RT-PCR expression profiling of 31 genes with enriched expression in tumor relative to testis [454–read count fold change ≥2.5, $P \leq 0.05$, chi-squared test; (green points in fig. S5)]. Log values of the mean expression difference of DFTD genes relative to testis (blue bars) and relative to *GAPDH* (red bars) are shown. Error bars represent standard deviation. Significant differences between tumor and testis expression levels ($P \leq 0.05$) are indicated by an asterisk (two-sample t test, Holm's correction for multiple testing). (**B**) Heat map of semiquantitative PCR expression profiles of 31 genes across a panel of tissues including peripheral nerve (PN), a Schwann cell–enriched tissue. Panel color represents the mean gene expression level, standardized across tissues (z score). Hierarchical clustering based on Pearson's correlation statistic is indicated by dendrograms. For each tissue three biological replicates were performed (n = 3). ND, not determined.

It is striking that DFTD, a cytologically undifferentiated tumor, expresses markers of highly differentiated Schwann cells because human Schwann cell tumors rarely express a complete set of terminally differentiated myelin genes (*19*). Although it is possible that Schwann cell genes have been activated in DFTD cells after carcinogenic transformation, it is more likely that the myelin program observed here reflects the DFTD cell of origin. We therefore propose that DFTD is a peripheral nerve sheath tumor that arose from a Schwann cell or Schwann cell precursor; this is supported by the miRNA profile of DFTD (Figs. 1 and 2). Schwann cells participate in nerve repair after injury and also modulate

Downloaded from www.sciencemag.org on October 12, 2010

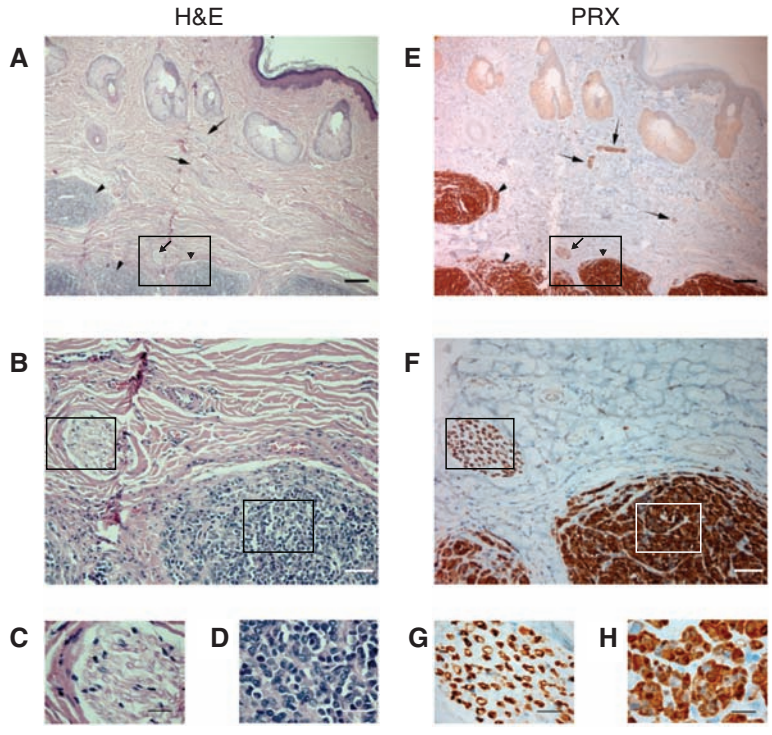


Fig. 3. Identification of a diagnostic marker for DFTD. (**A** to **D**) Hematoxylin and eosin (H&E) and (**E** to **H**) PRX antibody stains used for DFTD tumor histology. Arrowheads, DFTD tissue; arrows, peripheral nerve bundles (containing Schwann cells). Magnification, 4× (A and E; scale bar, 200 μm); 20× (B and F; scale bar, 50 μm); and 100× (C and G) peripheral nerve bundle (scale bar, 20 μm) and (D and H) DFTD tumor (scale bar, 20 μm). Boxed areas indicate approximate locations of areas magnified in lower panels.

local immune reactions in the peripheral nervous system (*20*, *21*). The plasticity and immunocompetence of Schwann cells may be significant in the evolution of DFTD as a transmissible cancer.

Interestingly, *proopiomelanocortin* (*POMC*), a gene encoding a peptide hormone precursor, was expressed in DFTD (Fig. 2). This transcript varies by more than 100-fold between tumors (fig. S8). Technical limitations, due to lack of antibody cross-reactivity with devil, prevented the detection of physiological levels of adrenocorticotropic hormone (ACTH), a labile POMC cleavage product. However, compared with healthy individuals, diseased devils had higher, but not significantly different, levels of serum cortisol (fig. S9, $P = 0.06$, Student's t test). Cortisol is a steroid hormone, released by ACTH activity, that is involved in stress responses and has immunosuppressive activity (*22*). These studies raise the intriguing possibility that DFTD may secrete factors that could alter host physiology and/or behavior, an interesting area for further study.

Our study highlights the activities of two miRNA gene networks that may function in DFTD. *PMP22*, a myelin gene expressed in DFTD, is a target of *miR-29*, a miRNA that is down-regulated in DFTD (Figs. 1 and 2 and fig. S3) (*23*). The expression of transcription factor *ZEB2* is negatively correlated with low expression of the *miR-200* family, which controls tumor invasiveness in a negative-feedback loop (*24*)(Figs. 1 and 2 and fig. S3). Furthermore, this work suggests that commonly mutated Schwann cell cancer genes, such as *NF1*, are interesting candidates for further analysis in DFTD (*25*).

DFTD is believed to be of neuroendocrine origin on the basis of the expression of the vimentin, S100, neuron-specific enolase, chromogranin A, and synaptophysin markers (*13*). As Schwann cells and neuroendocrine cells are both derived from the neural crest and overlap in gene expression, it is possible that DFTD has elements of both tissue types.

A Schwann cell origin for DFTD contrasts with that of the canine clonally transmissible cancer, which has been proposed to be of histiocytic origin (*26*, *27*). It will be of interest to define common and unique features of these two cancers and to determine how the histogenesis of transmissible cancers may influence their occurrence, evolution, and biology. Our catalog of genes provides a framework for this work, as well as for use in the effort to develop a DFTD preclinical test and vaccine.

References and Notes

1. C. E. Hawkins *et al.*, *Biol. Conserv.* **131**, 307 (2006).
2. S. Lachish, M. Jones, H. McCallum, *J. Anim. Ecol.* **76**, 926 (2007).
3. R. Loh *et al.*, *Vet. Pathol.* **43**, 890 (2006).
4. H. McCallum *et al.*, *EcoHealth* **4**, 318 (2007).
5. A. M. Pearse, K. Swift, *Nature* **439**, 549 (2006).
6. H. V. Siddle *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 16221 (2007).
7. H. L. Cooper, C. M. Mackay, W. G. Banfield, *J. Natl. Cancer Inst.* **33**, 691 (1964).
8. C. Murgia, J. K. Pritchard, S. Y. Kim, A. Fassati, R. A. Weiss, *Cell* **126**, 477 (2006).
9. C. A. Rebbeck, R. Thomas, M. Breen, A. M. Leroi, A. Burt, *Evolution* **63**, 2340 (2009).
10. H. McCallum, *Trends Ecol. Evol.* **23**, 631 (2008).
11. P. J. Canfield, A. A. Cunningham, *J. Zoo Wildl. Med.* **24**, 158 (1993).
12. L. A. Griner, *J. Natl. Cancer Inst.* **62**, 589 (1979).
13. R. Loh *et al.*, *Vet. Pathol.* **43**, 896 (2006).
14. Materials and methods are available as supporting material on *Science* Online.
15. M. E. Jones, D. Paetkau, E. Geffen, C. Moritz, *Mol. Ecol.* **13**, 2197 (2004).
16. R. Schickel, B. Boyerinas, S. M. Park, M. E. Peter, *Oncogene* **27**, 5959 (2008).
17. R. Ueda *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 10746 (2009).
18. J. Svaren, D. Meijer, *Glia* **56**, 1541 (2008).
19. H. B. Clark, J. J. Minesky, D. Agrawal, H. C. Agrawal, *Am. J. Pathol.* **121**, 96 (1985).
20. L. B. Dahlin, *Scand. J. Surg.* **97**, 310 (2008).
21. R. Gold, J. J. Archelos, H. P. Hartung, *Brain Pathol.* **9**, 343 (1999).
22. D. Greco, G. H. Stabenfeldt, in *Textbook of Veterinary Physiology*, J. G. Cunningham, Ed. (Saunders, Philadelphia, 1997).
23. J. D. Verrier *et al.*, *Glia* **57**, 1265 (2009).
24. P. A. Gregory, C. P. Bracken, A. G. Bert, G. J. Goodall, *Cell Cycle* **7**, 3112 (2008).
25. S. L. Carroll, N. Ratner, *Glia* **56**, 1590 (2008).
26. E. Mozos, A. Méndez, J. C. Gómez-Villamandos, J. Martín De Las Mulas, J. Pérez, *Vet. Pathol.* **33**, 257 (1996).
27. T. Marchal, L. Chabanne, C. Kaplanski, D. Rigal, J. P. Magnol, *Vet. Immunol. Immunopathol.* **57**, 1 (1997).
28. We thank C. Harmsen and E. Noonan (DPIPWE) and J. Smith (Fort Wayne Children's Zoo) for assistance with sample collection. We thank K. Claudio Campos, T. Dickson, S. Peck, V. Parameswaran, J. Harris, L. Stimmler, E. Hatchwell, R. Sachidanandam, Z. Xuan, M. Rasmussen, and the students and instructors of the 2006 and 2007 Cold Spring Harbor Laboratory deep-sequencing courses. The 454 sequencing was made possible by a grant from Roche Applied Sciences. E.P.M. was supported by a Sir Keith Murdoch Fellowship from the American Australian Association and an Overseas Postdoctoral Biomedical Fellowship from the Australian National Health and Medical Council (NHMRC). A.T.P., M.B., and A.H. were supported by grants from the NHMRC. A.S. was funded by a fellowship from the Human Frontier Science Program Organization. G.J.H. was supported by grants from the National Institutes of Health. H.S.B. was funded by an Australian Research Council Linkage Grant (LP0562190). E.P.M., C.T., A.K., and G.M.W. were supported by Dr. Eric Guiler Tasmanian Devil Research Grants. Sequences generated by this study have been deposited in GEO, www.ncbi.nlm.nih.gov/geo/ accession GSE18352 (miRNA) and the Short Read Archive http://www.ncbi.nlm.nih.gov/sites/entrez?db=sra accession SRA009772 (cDNA).

Supporting Online Material

www.sciencemag.org/cgi/content/full/327/5961/84/DC1
Materials and Methods
Figs. S1 to S9
Tables S1 to S11
References

14 August 2009; accepted 29 October 2009
10.1126/science.1180616

Downloaded from www.sciencemag.org on October 12, 2010

OPEN ACCESS Freely available online

PLoS GENETICS

A Comprehensive Map of Insulator Elements for the *Drosophila* Genome

Nicolas Nègre[1☯], Christopher D. Brown[1☯], Parantu K. Shah[1], Pouya Kheradpour[2], Carolyn A. Morrison[1], Jorja G. Henikoff[3,4], Xin Feng[5], Kami Ahmad[6], Steven Russell[7], Robert A. H. White[8], Lincoln Stein[9,10], Steven Henikoff[3,4], Manolis Kellis[2,11], Kevin P. White[1]*

1 Institute for Genomics and Systems Biology, Department of Human Genetics, and Department of Ecology and Evolution, University of Chicago, Chicago, Illinois, United States of America, **2** Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **3** Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America, **4** Howard Hughes Medical Institute, Seattle, Washington, United States of America, **5** Department of Biomedical Engineering, State University of New York at Stony Brook, Stony Brook, New York, United States of America, **6** Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts, United States of America, **7** Department of Genetics, University of Cambridge, Cambridge, United Kingdom, **8** Department of Physiology, Development, and Neuroscience, University of Cambridge, Cambridge, United Kingdom, **9** Ontario Institute for Cancer Research, Toronto, Canada, **10** Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States of America, **11** Broad Institute, Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts, United States of America

Abstract

Insulators are DNA sequences that control the interactions among genomic regulatory elements and act as chromatin boundaries. A thorough understanding of their location and function is necessary to address the complexities of metazoan gene regulation. We studied by ChIP–chip the genome-wide binding sites of 6 insulator-associated proteins—dCTCF, CP190, BEAF-32, Su(Hw), Mod(mdg4), and GAF—to obtain the first comprehensive map of insulator elements in *Drosophila* embryos. We identify over 14,000 putative insulators, including all classically defined insulators. We find two major classes of insulators defined by dCTCF/CP190/BEAF-32 and Su(Hw), respectively. Distributional analyses of insulators revealed that particular sub-classes of insulator elements are excluded between *cis*-regulatory elements and their target promoters; divide differentially expressed, alternative, and divergent promoters; act as chromatin boundaries; are associated with chromosomal breakpoints among species; and are embedded within active chromatin domains. Together, these results provide a map demarcating the boundaries of gene regulatory units and a framework for understanding insulator function during the development and evolution of *Drosophila*.

Citation: Nègre N, Brown CD, Shah PK, Kheradpour P, Morrison CA, et al. (2010) A Comprehensive Map of Insulator Elements for the *Drosophila* Genome. PLoS Genet 6(1): e1000814. doi:10.1371/journal.pgen.1000814

Editor: Yoshihide Hayashizaki, RIKEN Genomic Sciences Center, Japan

Received July 16, 2009; **Accepted** December 14, 2009; **Published** January 15, 2010

Copyright: © 2010 Nègre et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: CDB is supported by a Lily Life Sciences Research Fellowship. This research is funded by an NIH-NHGRI, modENCODE grant awarded to KPW. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: kpwhite@uchicago.edu

☯ These authors contributed equally to this work.

Introduction

The spatiotemporal regulation of transcription is controlled by the binding of transcription factors to their target *cis*-regulatory modules (CRM) and their resulting interactions with promoters. Such regulatory interactions between CRMs and promoters can occur over short distances when regulatory sequences are immediately proximal to their target promoter or, in many cases, over longer distances involving many thousands of base pairs. Because of the variability in the distances over which CRMs can act, delineating the molecular boundaries of genes can be challenging. Mechanisms by which a CRM targets the appropriate promoter among a collection of adjacent promoters are poorly defined. However, one such mechanism involves the partitioning of the genome into regulatory domains by genome features known as insulators, or boundary DNA elements.

Since their initial characterization twenty years ago [1–4], insulator elements have been thought to create distinct regulatory domains, and thus allow enhancers to find their proper target promoter [5]. Insulators have been identified in *Drosophila* as well as in vertebrate genomes [6] based on their ability to disrupt the communication between an enhancer and a promoter when inserted between them. This enhancer-blocking activity is dependent upon the binding of insulators by several proteins. The CCCTC-binding Factor (CTCF) was first identified in vertebrates [7]; its *Drosophila* homolog, dCTCF, is known to bind several insulators and is necessary for their function [8–11]. CTCF is currently the only vertebrate protein known to be associated with insulator elements. In *Drosophila* however, several other proteins have been identified for their insulator function. Su(Hw) is associated with the *gypsy* retrotransposon insulator and other endogenous binding sites [4,12–14]. The insulator activity of *gypsy* is dependent on the recruitment by Su(Hw) of two other proteins: Modifier of mdg4 [Mod(mdg4)] [15,16] and CP190 [17]. Three additional proteins have been linked to insulator function in *Drosophila*. The binding of Zw5 and BEAF-32 on the scs/scs' elements of the hsp70 locus is required for their enhancer-blocking activity [18,19]. Similarly, the ubiquitous transcription factor GAF

Author Summary

The spatiotemporal specificity of gene expression is controlled by interactions among regulatory proteins, *cis*-regulatory elements, chromatin modifications, and genes. These interactions can occur over large distances, and the mechanisms by which they are controlled are poorly understood. Insulators are DNA sequences that can both block the interaction between regulatory elements and genes, as well as block the spread of regions of modified chromatin. To date, relatively few insulators have been identified in developing *Drosophila* embryos. We here present the genome wide identification of over 14,000 binding sites for 6 insulator-associated proteins. We demonstrate the existence of two broad classes of insulators. Insulators of both classes are enriched at the boundaries of a particular chromatin modification. However, only insulators bound by BEAF-32, CP190, and dCTCF are enriched in regions of open chromatin or demarcate gene boundaries, with a particular enrichment between differentially expressed promoters. Furthermore, insulators of this class are enriched at points of chromosomal rearrangement among the 12 species of sequenced *Drosophila*, suggesting that insulator defined regulatory boundaries are evolutionarily conserved.

(GAGA Associated Factor) is necessary for the enhancer-blocking activity of particular insulators [20–22]. Apart from their enhancer-blocking activity, insulators act as chromatin boundary elements. Such boundaries block the spreading of epigenetic marks or chromatin proteins such as repressive heterochromatin proteins or Polycomb Group-dependent (PcG) silencing [23–27].

While genetic and molecular studies of insulator function suggest that insulators play a major role in the regulatory organization of the genome, functional data have been collected on only a dozen insulator sequences in *Drosophila* and mammals. The identification of new insulators in flies and mammals by genome-wide approaches has only recently been initiated in different biological sources [10,14,27–30]. Here we provide a uniformly collected dataset and comprehensive analysis from developing embryos for six different insulator proteins.

Results

Genome-wide mapping of insulator-associated proteins

We mapped the genome-wide binding sites of 6 insulator-associated proteins: CTCF, CP190, BEAF-32, Su(Hw), Mod(mdg4) and GAF by Chromatin ImmunoPrecipitation coupled with microarrays (ChIP-chip) in *Drosophila* embryos (0–12 h of development). For CTCF and Su(Hw), 2 different antibodies for each factor were used as controls to demonstrate the reproducibility of our experiments. At a 1% False Discovery Rate (FDR), we identified between 2,500 and 6,600 binding sites for each factor (Figure 1 and Table S1), which included all functionally verified *Drosophila melanogaster* insulator sequences (Figure 1, Table S2, and Figure S1). The reproducibility of different ChIP-chip experiments for 2 different antibodies for CTCF and Su(Hw) is very high, with 94% of CTCF and 87% of Su(Hw) binding sites overlapping (Figure S2 and Figure S3). Moreover, we were able to recapitulate the profiles for CTCF and Su(Hw) generated in *Drosophila* embryos

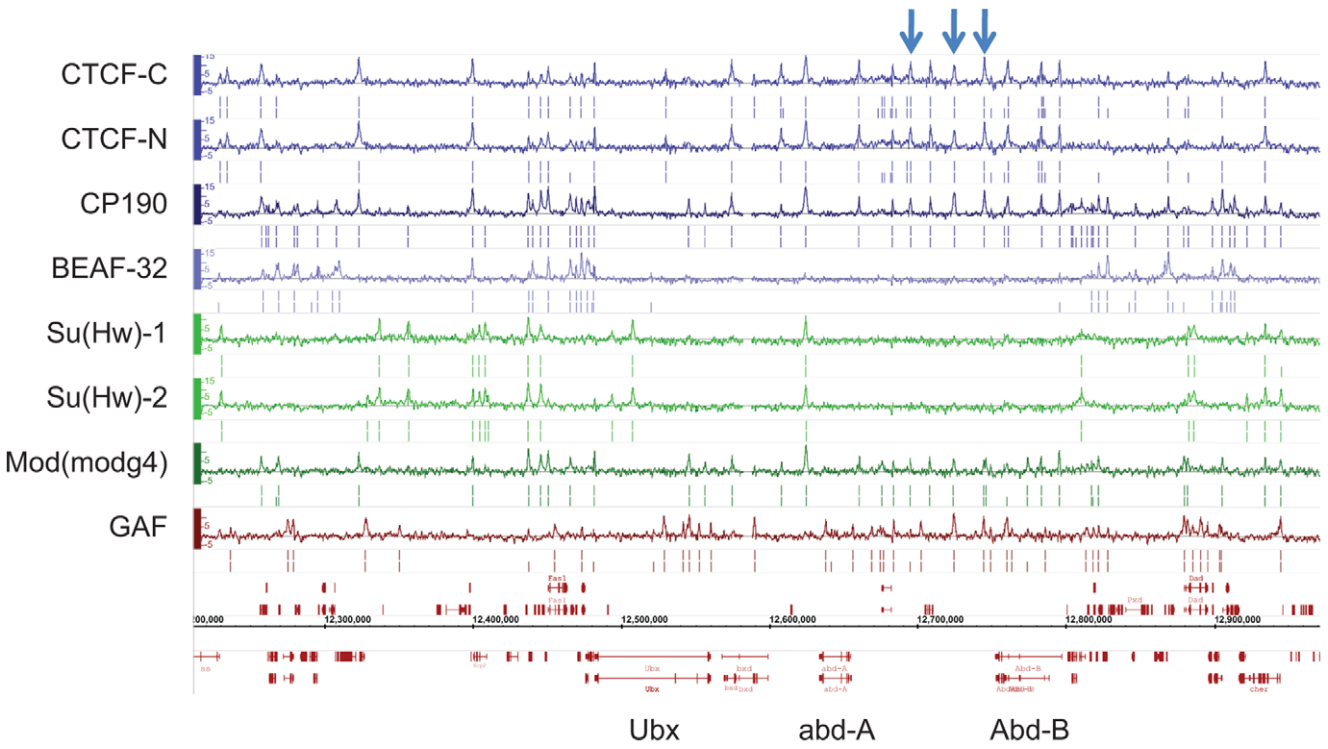


Figure 1. Binding profile of insulator-associated proteins in *Drosophila*. Binding profile of the 6 proteins studied on a large region of the chromosome 3R including the Bithorax-Complex, which contains 3 well-characterized insulators (blue arrows). For each protein, the track depicts the MAT score of each probe plotted on the y-axis versus chromosomal position, plotted along the x-axis. Called binding sites at 1% and 5% FDR are marked below each track. Note CTCF-C/CTCF-N and Su(Hw)-1/Su(Hw)-2 represent data from 2 independent antibodies. Flybase annotated genes are represented in red as the two bottom tracks.
doi:10.1371/journal.pgen.1000814.g001

for the homeotic complexes and 3 Mb of the *Adh* region [10,14] with an overlap of 94% (31/33) for CTCF and 70% (27/41) for Su(Hw) between the published dataset and our mapping in the same genomic region.

To characterize the sequence specificity of each insulator-associated protein, we identified significantly enriched sequence motifs for each set of target sites (Figure S4). The most enriched motif identified for CTCF strongly resembles the CTCF motif identified in vertebrates [31] and *Drosophila* [10]. Likewise, the motif for Su(Hw) that was discovered in this study is similar to a motif previously identified in *Drosophila* from a limited number of Su(Hw) sites [14,32,33]. The discovered motifs are present in 75.6% of CTCF, 86.8% of BEAF-32, 84% of Su(Hw) and 88.6% of GAF binding sites (Table S3). Additionally, the motifs identified for one insulator-factor were often also enriched at the binding sites of other insulator-factors (Figure 2A). This cross-enrichment was not observed however, when only binding sites associated with a single factor were considered (Figure 2B), suggesting that each factor retains unique DNA-level binding specificity but associates with other insulator proteins via clustered binding sites and/or protein-protein interactions.

Cell-type specificity of CTCF binding

Previous analyses have suggested that, in human cells, insulator binding sites are remarkably conserved across cell types [27,31,34]. Given the large overlap between binding sites here identified in whole embryos and data previously produced in *Drosophila* S2 cells for CTCF and CP190 [30], we investigated this trend further. We performed ChIP-chip experiments for CTCF in S2 and Kc cells. Approximately 74–81% of CTCF binding sites identified independently in each cell type overlap (specifically, have a midpoint to midpoint distance less than 250 bases)(Figure S5 and Figure S6A). This observation is consistent with a recently published analysis of CTCF binding sites in S2 and Mbn2 cells [35], in which, by the same criteria, 77–86% of binding sites overlap.

However, given the technical differences in protocols for embryos and suspension cell culture and the loss of information inherent in a comparison of independently thresholded binding site calls, we regard this as a conservative estimate. Qualitative observation of binding profiles suggests that many putatively differential binding sites may result from the threshold applied and normalization issues (Figure S5 and Figure S6B). Indeed, we note that the IP signals at non-overlapping binding sites are, on average, four-fold greater than input background, while overlapping binding sites are six-fold greater. In an attempt to avoid such biases, we used a linear mixed model framework to build a binding site detection model that jointly analyzes the data from multiple cell types (see Text S1). This model identifies 2,784 CTCF binding sites, only 166 of which show significant cell type specificity (Figure S7). In summary, while most insulator sites identified in this study appear to be conserved across cell types, a small fraction appear to function in a regulated fashion.

Binding-site clustering identifies two major categories of insulator sequences

While the six insulator associated proteins mapped in this study often bind independently, we find clusters of overlapping binding sites far more often than would be expected by chance, indicating insulator-associated proteins often bind jointly to the same sequence. Indeed, 45% of the 14,145 binding sites identified in this study are occupied by more than one insulator associated protein. For example, 77% of CTCF binding sites cluster with at least another factor (Figure 2 and Table S1). Analysis of binding site cluster types revealed several notable trends (Figure 2C and Figure S3). CP190 is frequently (5690 out of 6651 total sites) found to bind with additional factors, BEAF-32 being its most common partner (3329/6651). BEAF-32, CTCF, and CP190 cluster together (1378/8872), as do Mod(mdg4) and Su(Hw) (1101/5381), while GAF displays a significant lack of clustering with other insulator proteins (2973 single sites out of 3905 total sites). This binding site clustering and the functional data presented below suggest a previously underappreciated compositional complexity of insulator sequences but also clearly identifies two major classes of insulators: Class I principally representing binding sites for BEAF-32/CP190/CTCF and Class II representing Su(Hw)-associated binding sites.

Positional classification of insulators

The distribution of insulator binding sites relative to different classes of functional genomic elements further supports the existence of several distinct functional classes of insulators. BEAF-32, CP190, CTCF, GAF, and Mod(mdg4) are clearly enriched at promoters (Figure 3A, Figure S8, and Table S4), while Su(Hw) is depleted. BEAF-32, CP190, CTCF and Mod(mdg4) binding sites are also strongly enriched within 5′UTRs as well as in intergenic regions (Figure S8) and at transcription end sites (Figure 3B). In contrast, they are largely excluded from transposable elements and coding exons (Figure S8 and Figure S9), suggesting a role of Class I insulator proteins, but not Class II, in regulating the transcription of genes.

We reasoned that if insulators act as gene boundaries, they should partition genes into distinct regulatory environments. Indeed, we find that four of the six insulator-associated proteins binding sites are significantly enriched between adjacent consecutive promoters (Figure 3C) with a stronger enrichment of BEAF-32, CP190, CTCF and Mod(mdg4) between adjacent divergently oriented promoters (Figure 3D). Additionally, as suggested previously in vertebrates for CTCF [31], Class I and Class II insulator proteins are significantly enriched between alternative promoters, providing a potential mechanism for their independent regulation (Figure 3E).

Insulators demarcate differentially expressed genes

The distribution of insulators relative to a variety of genomic functional element classes suggests a pervasive role in controlling gene regulatory environments. To further address this hypothesis we mapped active promoters in embryos of the same developmental stage that we used for insulator mapping. To identify active promoters, we performed ChIP-chip with antibodies directed against the trimethylated lysine 4 of Histone H3 (H3K4me3), which is a clear mark of activation [36–39], and against the largest subunit of the RNA Polymerase II (PolII). We combined these two mappings with hybridization on tiling arrays of total RNA extracted from the same material. In *Drosophila* embryos, H3K4me3 is associated with gene Transcription Start Sites (TSS) and colocalizes with PolII immediately downstream of the TSS of active genes (Figure S10A and S10B). We extracted from this dataset a set of high confidence actively transcribed promoters, which overlap with H3K4me3 and PolII signals and whose exons overlap significant RNA signal (Figure S10C). We hypothesized that if insulators do indeed demarcate regulatory units, insulators would separate promoters with differing expression status. We repeated the positional analysis of insulator proteins between divergent, adjacent, and alternative promoters while taking into account the transcriptional status of the promoters (Figure 4A–4C). We observed that the enrichment of BEAF-32, CP190,

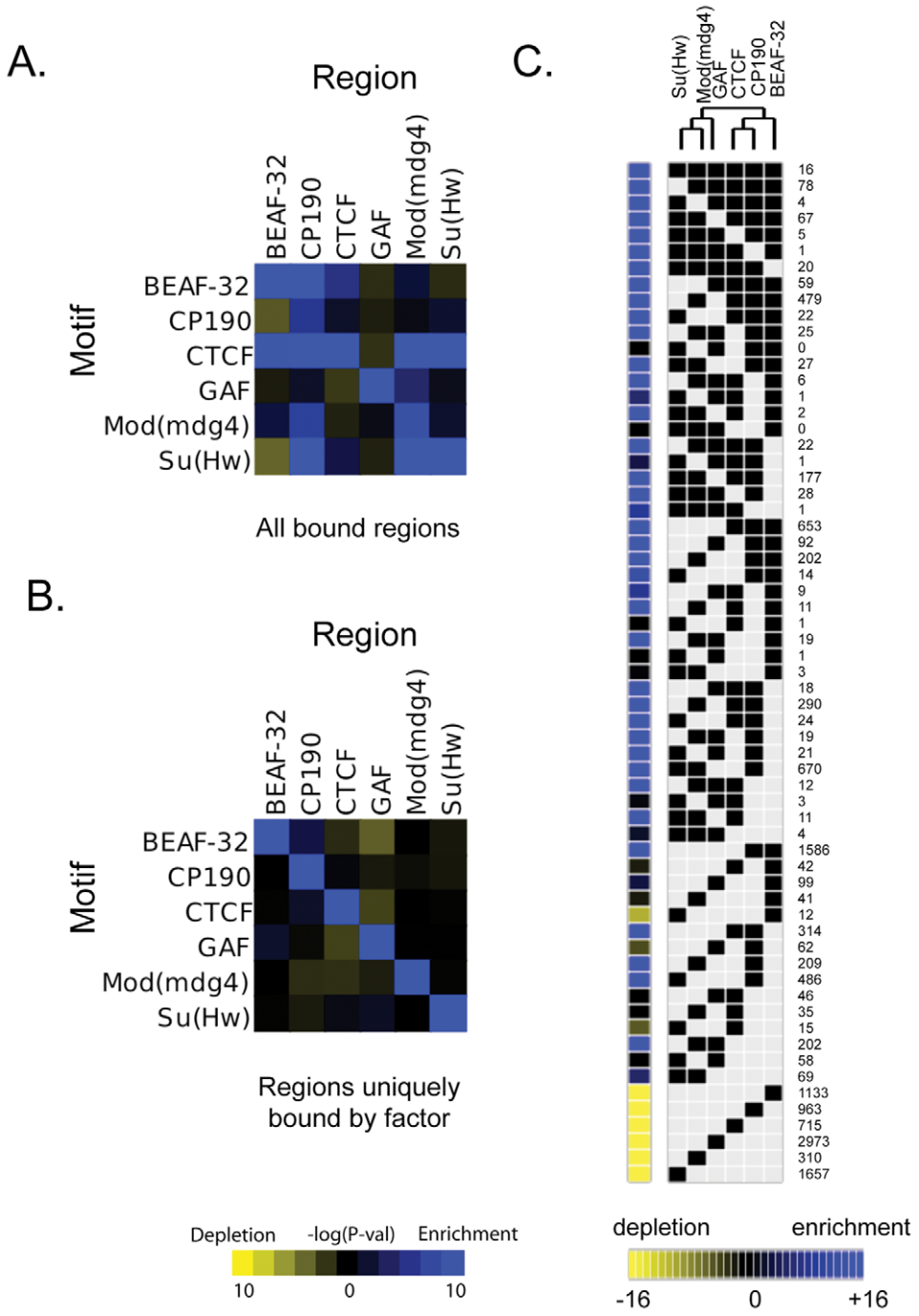


Figure 2. Combinatorial protein binding on insulators. For each insulator, intergenic bound regions were determined using the peaks ±100 bp (as described in Motif Discovery methods). The enrichment or depletion of instances of each motif with respect to these peak regions for each insulator is determined using a hypergeometric p-value as compared to control instances at 0.0 confidence, using (A) all bound regions for a particular protein or (B) peaks whose center is at least 1 kb away from the peak of any other insulator (uniquely bound regions). (C) All binding sites for all 6 insulator proteins have been classified based on their protein composition. Each of the categories is represented in the matrix as a black square for the factors associated with the binding site. The number of occurrences of each particular combination is indicated at the right of the matrix. The yellow to blue squares at the left of the matrix represent the enrichment or depletion p-value for each category when compared to simulated data. This matrix and the significant associations of factors have been use to build the dendrogram at the top of the figure.
doi:10.1371/journal.pgen.1000814.g002

CTCF, GAF, and Mod(mdg4) is greater between promoter pairs when they are differentially expressed (Figure 4, Figure S11, and Figure S12).

It is possible however that this result comes from an averaging of promoter activity across all the cell types present in the embryo at this developmental stage. We then repeated H3K4me3 ChIP-chip as a marker of active promoters in 2 embryonic *Drosophila* cell types: S2 and Kc cells. The overlap of H3K4me3 between embryos and Kc and S2 cell lines is between 71 and 75% respectively, while it is 85% between S2 and Kc cells (Figure

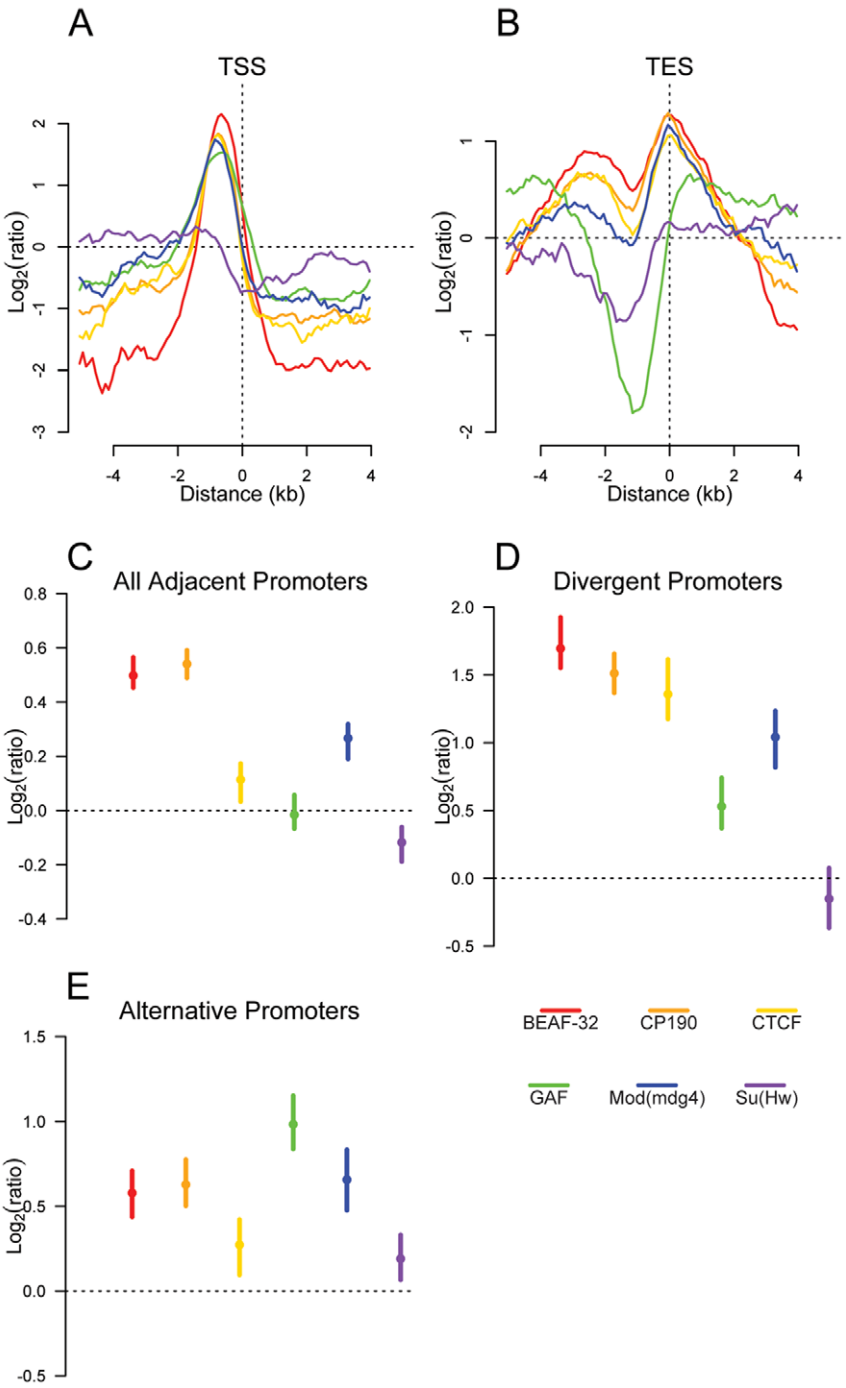


Figure 3. Insulator proteins are enriched between promoters. Plot of $\log_2$ enrichment or depletion of insulator binding sites (y-axis) by binding site base pair position (x-axis), relative to genomic transcript annotations. (A) Transcription start sites; negative and positive values depict upstream and downstream binding, respectively. (B) Transcription end sites; negative and positive values depict upstream and downstream binding. (C–E) Enrichment of insulator binding sites between adjacent promoters (C), divergent promoters (D), and alternative promoters (E). Points and lines depict enrichment estimates and 95% confidence intervals. A dotted line at 0 indicates no enrichment relative to the random expectation.
doi:10.1371/journal.pgen.1000814.g003

S13A). Using H3K4me3 binding sites as a guide, we identified active promoters in each cell type. As in whole embryos, genes flanking CTCF binding sites identified in S2 and Kc cells show a significant enrichment of differentially expressed divergent and alternative promoters (Figure S13B and S13C) further demonstrating that Class I insulators delimit the boundaries of gene regulatory units.

Insulators partition CRMs and promoters

Consistent with the limited previous functional data demonstrating the enhancer-blocking activity of insulators, we find binding sites for BEAF-32, CP190, and Su(Hw) are significantly depleted between annotated CRMs and their target promoters across the entire genome (Figure 4D, Figure S14), while CP190, CTCF, GAF, and Mod(mdg4) are enriched between *cis*-regulatory

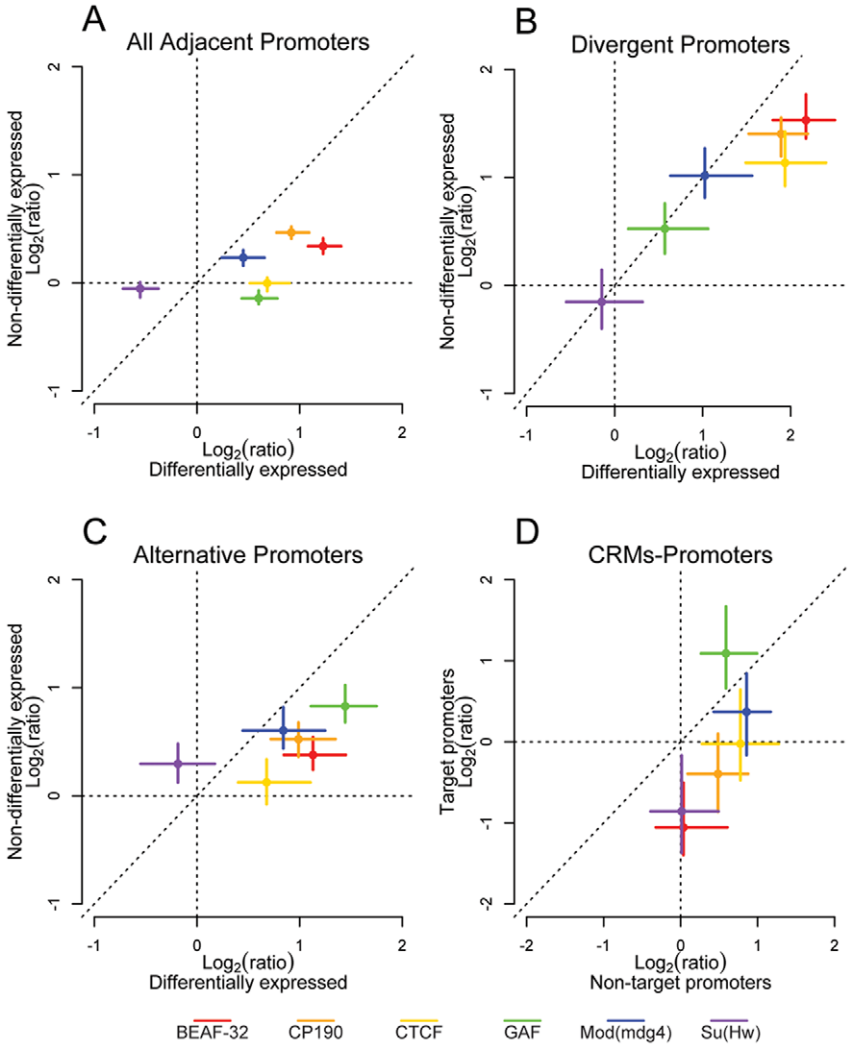


Figure 4. Insulator proteins segregate differentially expressed promoters. Log_2 enrichment or depletion of insulator binding sites between (A) adjacent promoters, (B) divergently transcribed promoters, (C) alternative promoters, and (D) *cis*-regulatory elements and promoters. (A–C) X and Y axes depict enrichment between differentially and non-differentially expressed promoters, respectively. (D) X and Y axes depict enrichment between CRMs and their nearest non-target promoter and their target promoter. Points and lines depict enrichment estimates and 95% confidence intervals.
doi:10.1371/journal.pgen.1000814.g004

elements and their nearest non-target promoter, distributions that strongly support their proposed enhancer blocking function. Interestingly, we note that binding sites for GAF are significantly enriched between CRMs and their target promoters. Similarly, we find that BEAF-32, CP190, and Su(Hw) binding sites are depleted between distinct CRMs of the same gene, while GAF is found more frequently than expected (Figure S14).

We note that the enrichment of insulators within such genomic features may, in part, be driven by the effects of differential promoter density or biases in chromatin accessibility. In order to understand how such factors could affect any interpretation of our data, we reanalyzed binding site data for 36 recently published datasets corresponding to 21 transcription factors, from the Berkeley *Drosophila* Transcription Network Project (BDTNP) [40]. We first observed that none of our insulator binding sites preferentially localize with this transcription factor set (Figure S15). Despite several transcription factors that preferentially bind promoter-proximal sequences (Figure S16), the enrichment of insulators between promoter pairs is greater than for any of the published transcription factors (Figure S17A, S17B, S17C). In contrast to these findings, and as expected, the published BDTNP transcription factors are not as strongly biased towards CRM, non-target promoter separation (Figure S17D).

Insulators mark the boundary of chromatin domains

Previous studies have demonstrated that insulators delimit distinct organizational domains of a genome [27,30]. One such chromatin domain is marked by the trimethylated Lysine 27 of Histone H3 (H3K27me3), a histone modification deposited and recognized by the repressive Polycomb protein complexes [41]. We mapped by ChIP-chip the H3K27me3 mark in *Drosophila* embryos. We observed in whole embryos, as described previously [42,43], that H3K27me3 is distributed throughout the genome in large domains (Figure S18). To better define the boundaries of these large genomic regions, we used a hidden Markov model based segmentation algorithm. We confirm that the genes affected by this silencing mark correspond to the previously described Polycomb target genes [42–45]. We identified 140 regions of

substantial H3K27me3 density and quantified the distribution of each insulator binding site type with respect to the domain boundaries. Interestingly we find that all 6 factors are significantly depleted within and enriched outside these regions (Figure 5A). In addition, CTCF, GAF, and Mod(mdg4) are enriched at the boundaries of regions of high H3K27me3 density, with this enrichment significantly decreasing at increasing distances, further supporting the insulators' role in chromatin domain boundary determination (Figure 5A and Figure S18). It is possible that this result is confounded by the fact that insulators are enriched at TSSs. We performed Pearson's chi-squared contingency table tests to assess if the frequency of insulator-H3K27me3 boundary overlaps are independent of (and greater than) the frequency of TSS- H3K27me3 boundary overlaps. Indeed, CP190 ($p<9.8e-6$), BEAF-32 ($p<1.8e-5$), CTCF ($p<0.00013$), GAF ($p<0.0022$), Mod(mdg4) ($p<0.00035$), and Su(Hw) ($p<0.0088$) are independently associated with H3K27me3 breakpoints.

Insulator binding site sequence constraint and conservation of synteny

Given their apparently pervasive role in the establishment of gene regulatory units, we examined the role insulator sequences have played in shaping the evolution of the *Drosophila* genome. First, insulators show evidence of local sequence constraint. Based on either 15-way insect multiple sequence alignments or pair-wise alignments between the closely related *Drosophila melanogaster* and *Drosophila simulans*, insulator binding sites evolve significantly slower than fast evolving introns, although more swiftly than either coding exons or most transcription factor binding sites [46] (Figure S19). Second, we find that BEAF-32, CP190, CTCF, and Mod(mdg4) are significantly enriched near the 12 *Drosophila* species syntenic breakpoints (Figure 5B) [47]. Chi-squared tests demonstrate that for CP190 ($p<0.0031$), BEAF-32 ($p<0.0086$), GAF ($p<0.027$), and Mod(mdg4) ($p<0.034$), this result is independent of the association of TSSs and syntenic breaks. This finding provides evidence to support the hypothesis [48] that selective pressure has maintained gene regulatory units established by flanking insulators.

Insulators are sites of dynamic chromatin

We find that binding sites for 5 of the 6 insulator-associated proteins (Su(Hw)is the exception) are regions of reduced nucleosome density relative to surrounding regions (Figure 6A). Reduced nucleosome density often corresponds to sites of high histone replacement or displacement [49,50] and classical "active" chromatin as defined by salt solubility properties [51]. We also find that the same 5 of the 6 insulator proteins are preferentially bound in regions characterized by low-salt soluble nucleosomes (Figure 6B and 6C), depleted in the remaining high-salt-soluble fraction (Figure 6D) and highly enriched in the salt-washed insoluble pellet (Figure 6E). Similar analyses of only non-promoter proximal insulators reveal the same trends, indicating that the shared solubility properties of insulators and promoters are indeed independent (Figure S20). Given the correspondence between these results and the regulatory boundary analyses presented above, we hypothesize that this difference in chromatin properties may explain why Su(Hw), defining ClassII insulators, does not act as a gene boundary in the genome.

Discussion

Insulator identification has been the source of much recent interest. Indeed, in the last 6 months CTCF was mapped in S2 cells [52]; BEAF-32 in embryos (6–16 h of development) [53], CTCF and CP190 in S2 cells [54] and more recently CTCF, Su(Hw), CP190 and BEAF-32 in Kc cells and Mbn2 cells [35]. Interestingly, the latter paper describes three subclasses of insulators, with CP190/BEAF association being distinct from CP190/CTCF and CP190/Su(Hw).

We present in this study the embryonic binding profile of six factors previously known to be associated with insulator function in *Drosophila*. Our analysis of insulator binding site distributions

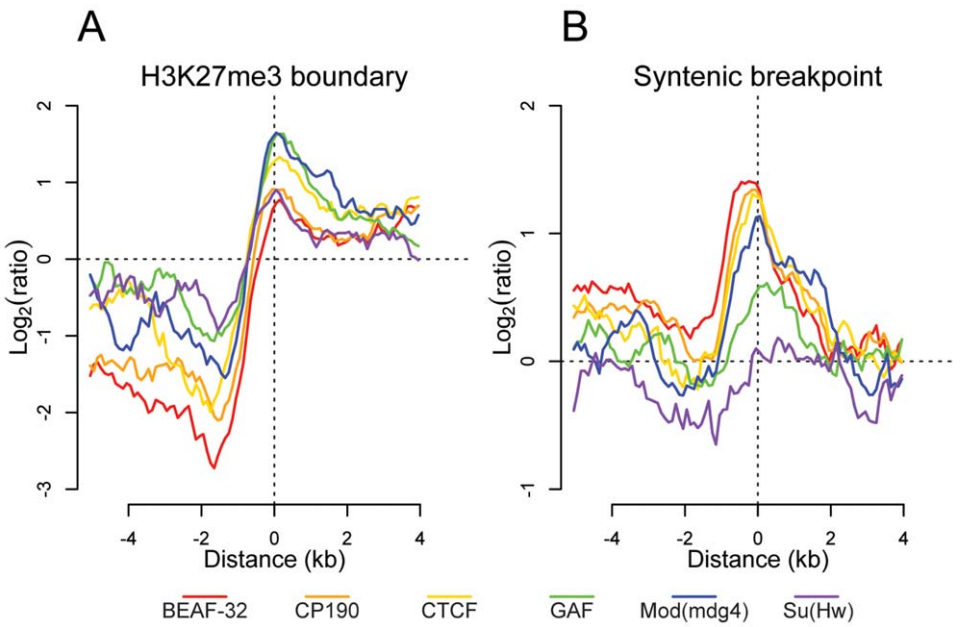


Figure 5. Insulator proteins mark chromatin and syntenic block boundaries. Log_2 enrichment or depletion of insulator binding sites. (A) H3K27me3 boundaries; negative and positive values depict binding within and outside regions of histone modification. (B) Syntenic breakpoints; negative and positive values depict binding within and outside syntenic blocks.
doi:10.1371/journal.pgen.1000814.g005

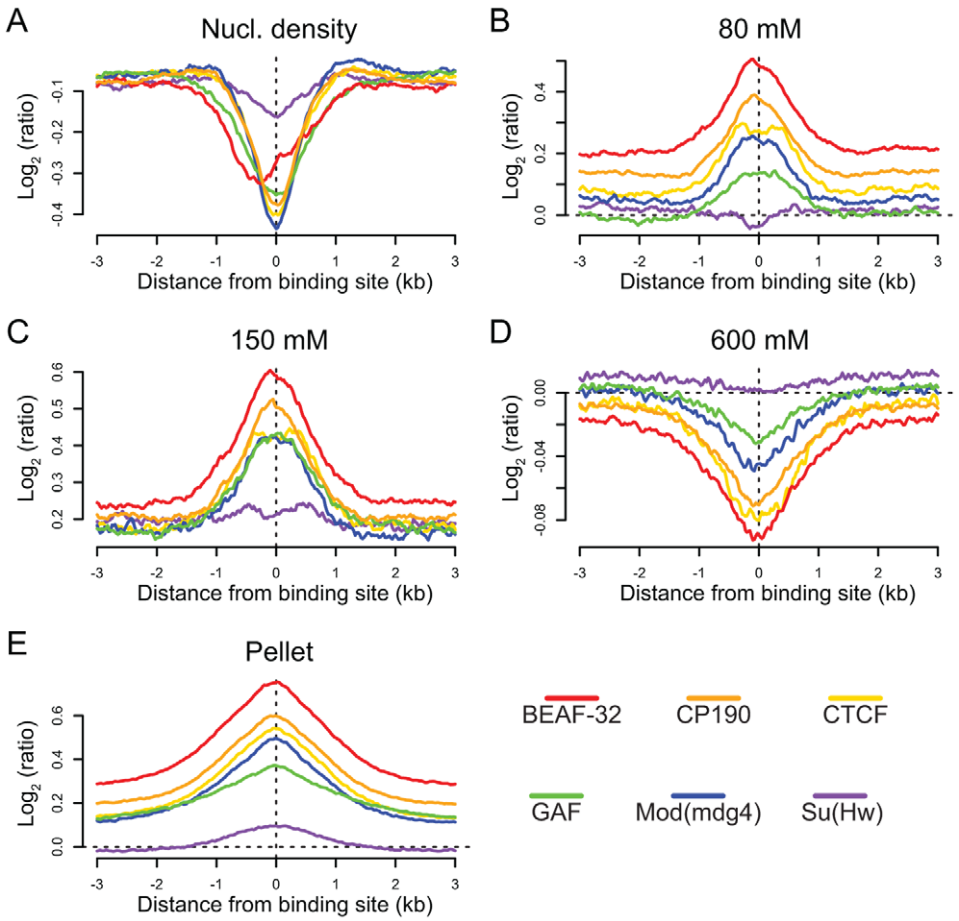


Figure 6. Insulators are sites of dynamic chromatin. Nucleosome density and salt fractionation profiles for *Drosophila* S2 cell chromatin, aligned at sites of insulator protein binding as indicated by color and averaged over a ±3 kb region. (A) nucleosome density, (B) 80 mM salt fraction, (C) 150 mM salt fraction (D) 600 mM salt fraction, (E) salt-washed pellet.
doi:10.1371/journal.pgen.1000814.g006

and protein composition suggest there exist 2 principal categories of insulator elements (Class I and Class II). In particular, we have shown that Class I insulators, identified by the binding of CTCF, CP190 or BEAF-32, segregate differentially expressed genes and delimit the boundaries of chromatin silencing, while they are depleted between known CRMs and their target genes. We do not find evidence supporting a significant distinction between CP190/BEAF and CP190/CTCF or CTCF/BEAF. In contrast, our analyses suggest that BEAF-32, CP190, and CTCF are distributed and function quite similarly, while Su(Hw) appears distinct. The Class II insulators, bound by Su(Hw), are often exceptional in our analyses. We note that the analysis of genome-wide mapping data, expression data, and genome annotation provides an endogenous boundary assay that demonstrates that, while Su(Hw) has been described as an insulator before, it is not systematically associated with the boundaries of the gene units.

By helping to delimit the regulatory boundaries of genes, the Class I insulator map presented here will aid in the identification of transcription factor target genes and the construction of transcriptional regulatory networks. As an example of this concept, we illustrate the distribution of known regulatory elements and insulators across the Antennapedia Complex (ANT-C) of homeotic genes (Figure 7). This region quite strikingly demonstrates the potential utility of insulator binding data for *cis*-regulatory annotation. Across approximately 500 kb, *cis*-regulatory elements and their target promoters are found between insulator pairs. For example, a single insulator separates the *lab* and *Edg84A* genes, with their respective *cis*-regulatory elements narrowly partitioned on either side. The adjacent regulatory elements and promoters of *zen* and *bcd* are similarly insulator segregated.

The presence of an insulator 3′ of *ftz* was previously hypothesized [55] to explain the ability of distal *Scr* regulatory elements to bypass *ftz* by pairing with the proximal SF1 insulator, located between *Scr* and *ftz*. Lastly, at *Antp*, as we observe genome wide, two alternative promoters and their proximal regulatory elements are segregated by a single insulator. We are currently developing analysis methods to systematically partition the entire genome into such regulatory domains.

Consistent with their observed regulatory boundary functions, Class I insulators are embedded within local regions of active chromatin and are frequently associated with syntenic breakpoints between species. Previous work has demonstrated that active promoters in yeast and *Drosophila* are associated with reduced nucleosome occupancy and low-salt soluble and high-salt insoluble chromatin [50,56] (Figure S20). Therefore, surprisingly, dynamic chromatin is a shared feature between promoters and most classes of insulators. It is notable however that some studies have revealed functional similarities between insulators and promoters in transgenic assays [57]. These results have been described as paradoxical, as insulators can negatively affect promoters by

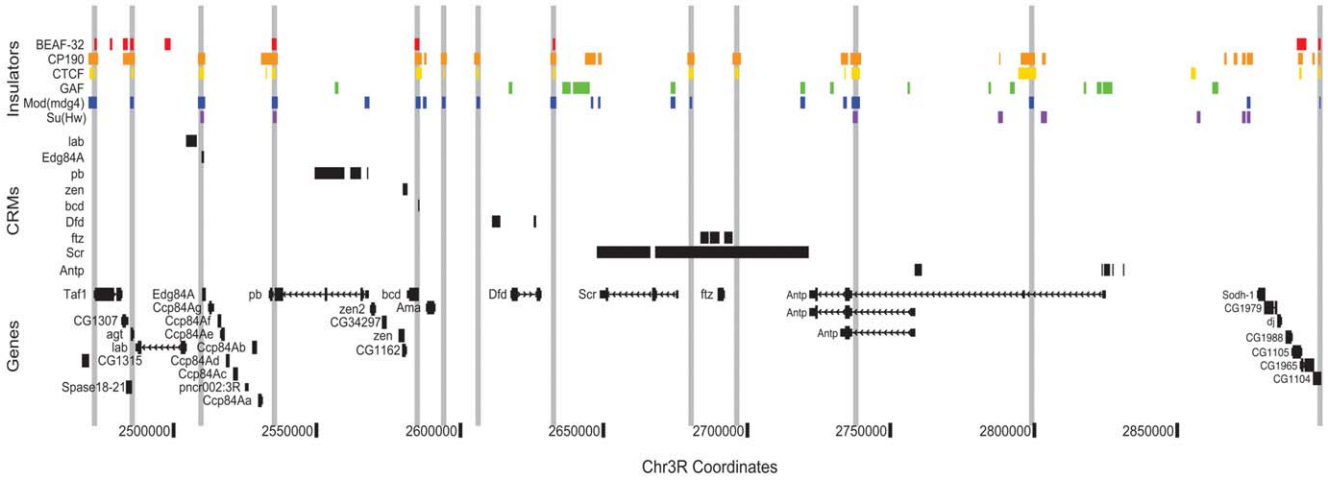


Figure 7. Class I insulators demarcate regulatory boundaries in the Antennapedia Complex (ANT-C) region. Binding sites for insulator proteins are depicted as colored boxes at top. For display purposes, grey vertical lines are drawn through all positions bound by two or more class I insulators. ORegAnno (www.oreganno.org) defined *cis*-regulatory modules for each of 9 genes are plotted as black boxes across the middle. RefSeq (www.ncbi.nlm.nih.gov/RefSeq/) gene models and coordinates across the 500 kb Antp region are displayed at bottom.
doi:10.1371/journal.pgen.1000814.g007

blocking communication between enhancers and promoters. One proposed model for insulator function is that they act as promoter "decoys" by recruiting away factors necessary for transcriptional initiation [57]. Alternatively, insulators and promoters might require common chromatin features to function by mechanisms that are still unknown. One potential interpretation is that the dynamic chromatin at insulators forms a flexible chromatin joint that would affect the probability of productive contact between separated regulatory elements. In this way, the similarity between promoters and insulators would be a consequence of their common requirement for dynamic chromatin, although with very different consequences. This model may explain why promoters are so frequently scored as insulators in the classical insulator assay, when an element is placed between an enhancer and a promoter [1,58].

Materials and Methods

ChIP–chip

Chromatin immunoprecipitations have been performed as described previously [59]. Briefly, the biological material is homogenized in the presence of 1.8% formaldehyde. The cross-linked chromatin is sonicated using a Bioruptor (Diagenode) to an average size of 500 bp. Pre-cleared chromatin extract is incubated overnight at 4C with the specific antibody and immunoprecipitated with protein-A Sepharose beads. After purification of the DNA and amplification of the libraries by linker-mediated PCR, the samples are labeled according to Affymetrix protocols and hybridized in parallel with an input sample onto the Affymetrix *Drosophila* Tiling Array, v2.0 R.

Antibodies

CTCF-C and CTCF-N antibodies are described in [8], CP190 antibody is described in [60], BEAF-32 antibody is described in [18], Mod(mdg4) antibody is directed against the 67.2 isoform and is described in [61], Su(Hw)-1 antibody is described in [62], Su(Hw)-2 is described in [63], GAF antibody is described in [64], H3K27me3 antibody is from Upstate (07-449 lot DAM1387952), H3K4me3 antibodies is from Abcam (ab8580 lot 411277) and PolII antibody is from Covance (8wG16 lot 14861301).

Analysis of arrays

Insulator binding data was processed with Model based Analysis Tiling-arrays (MAT) software [65]. We ran paired MAT analysis with MaxGap of 500, MinProbe of 10, and a Bandwidth of 250. H3K4me3, PolII and RNA data were analyzed with TAS (Tiling Array Software) and a threshold of 5% of the highest pValues was applied to identify the high intensity signals. The same parameters as for the MAT analysis have been applied to then call the peaks with TAS.

We developed a new HMM-based segmentation algorithm to identify H3K27me3 domains, as well as a novel mixed model framework for the joint analysis of ChIP-chip data from more complicated experimental designs, here applied to CTCF binding data from multiple cell types (see details in Text S1).

Motif discovery

Motif discovery was performed separately for each insulator. Peak centers that were at least 1 kb away from the peak center of any other insulator were taken ("uniquely bound peaks") and +/− 100 bp windows were generated excluding coding exons, repeats, transposons, 3′ untranslated regions and non-coding RNAs ("excluded regions"). For each insulator up to 500 of the regions were randomly selected and enriched motifs were identified using MEME [66], AlignACE [67], and MDscan [68]. All programs were run with default parameters except for MEME, which was restricted to a maximum of 3 iterations and a maximum motif width of 25. Instances of each of the motifs at conservation levels from 0.0 to 1.0 confidence (in steps of 0.1) were identified in all Intergenic regions (defined as genomic regions excluding those noted above) using the motif instance pipeline described in [69] with a PWM threshold corresponding to a p-value of 4^{-8} as determined by TFM-Pvalue [70]. The motifs were ranked using the fraction of instances found in the uniquely bound regions divided by the fraction for instances of shuffled control motifs at the same conservation cutoff (Wilson's confidence interval at $Z = 1.5$ was used on the ratios to give a conservative enrichment). This procedure is designed to reduce biases due to composition or conservation level. The motif with the highest enrichment at any confidence level was selected. This procedure was repeated using the MAT peak regions (rather than +/− 100 bp) to produce the

comparison in Figure S4 (otherwise the +/− 100 bp motifs are used throughout).

Genomic distribution analyses

Genomic distribution analyses only used insulators mapped to chromosomes 2L, 2R, 3L, 3R, 4, and X. All gene annotations, including transcription start site locations and alternative promoter presence were defined according to RefSeq annotations. Transposable element locations were based on Flybase annotations. Divergently transcribed genes were identified as all adjacent transcription start sites, on opposite strands, between 500 and 2500 bases apart. Alternative promoters were identified as all RefSeq annotated genes with more than one distinct transcription start site. The 'all adjacent' gene set included all adjacent gene pairs whose transcription start sites were between 1500 and 20000 bases apart, regardless of strand. *Cis*-regulatory elements and their target genes were defined according to the RedFly database [71]. Breakpoints of regions of conserved synteny across the 12 sequenced Drosophilids were identified in [47].

All genomic distributional analyses were first conducted by mapping protein binding sites relative to the genomic feature of interest. This mapping was performed in one of two ways; First, for genomic features that can be faithfully represented as a single base (e.g., a transcription start site), the distance from each insulator to its nearest feature was tabulated, second, for paired genomic features (e.g., divergent promoters), the number of intervening insulators for each feature pair was tabulated. To quantify if the distribution of mapped insulators relative to the genomic feature of interest is significantly different than would be expected by chance (given the number of insulators and the distribution of the particular feature of interest), we performed simulations as follows. First, permuted insulator binding sites were generated by sampling n sites from a random, uniform distribution, the length of each chromosome, where n is the number of observed insulator binding sites, by chromosome. In other words, a simulated insulator is equally likely to be placed at any location across a chromosome. Second, the simulated binding sites were mapped relative to the genomic feature of interest, as with each real dataset. This procedure was repeated 10,000 times for each insulator, target element combination. The median simulated values were used to normalize the real data counts to produce enrichment estimates. The 2.5 and 97.5 percentiles of the simulated distributions were used to produce confidence intervals for display purposes and significance estimates. Empirical p-values were calculated as the fraction of simulations that produced a number of mapped features as extreme as observed in the real data.

Nucleosome enrichment and salt fractionation

The position of binding sites have been compared to data of nucleosome density and salt fractionation of the chromatin extraction as described in [51]. Binding sites are defined by their midpoint and nucleosome density and salt fractionation data from S2 cells are plotted as a log ratio of enrichment in a 3 kb interval around the midpoint of the binding site.

GEO accession number of described datasets

GSE16245

Supporting Information

Figure S1 Example of mapping around some known insulators. The vertical dotted line indicates the location of the known insulators: (A) the 1A2 insulator [16,17] in the *yellow* locus, (B) the scs and scs' elements [18] in the hsp70 locus, (C) the SF1 insulator in the ANT-C region [19].
Found at: doi:10.1371/journal.pgen.1000814.s001 (0.92 MB JPG)

Figure S2 Pair-wise overlap at varying distance thresholds. In this example, the overlap between peaks at 1% FDR for CTCF-N and each of the other factors is plotted. The y axis represents the number of overlapping binding sites, while the x-axis represents the minimal distance between two peaks to call them overlapping. The plateau between CTCF-N and CTCF-C, which correspond to two independent antibodies for CTCF, is reached at a distance of 250 bp, which is the minimal distance we used for further analyses.
Found at: doi:10.1371/journal.pgen.1000814.s002 (0.31 MB JPG)

Figure S3 Overall pair-wise comparison between different factors. The axes in the radar plots indicate the percentage of overlapping binding site for one factor compared to each of the other factors. Data for CTCF and Su(Hw) corresponds to the CTCF_C and Su(Hw)-1 datasets respectively. This representation allows a quick identification of the preference of association between factors. For example, GAF is principally associated with itself and no other factor, while CTCF overlaps to a greater extent with CP190, Mod(mdg4), and BEAF-32, but not with GAF and Su(Hw).
Found at: doi:10.1371/journal.pgen.1000814.s003 (0.53 MB JPG)

Figure S4 *de novo* Identification of DNA motifs. The newly discovered motifs for each factor are represented in color logos, while the previously known motifs are represented in gray scale. We present the motifs corresponding to 2 different discovery regions: the original peak regions as called by MAT (noted Binding Regions; median size ~1,000 bp) and ±100 bp around the center of each peak (see Materials and Methods). The newly discovered motifs for CTCF, Su(Hw) and GAF are in agreement with previously described motifs [8,10,20], while the motif discovered for BEAF only agrees with previous studies [21,22] when discovery is performed using the smaller ±100 bp regions. Interestingly, using the larger MAT regions, high information content motifs are identified for both CP190 and Mod(mdg4) which are not thought to bind DNA directly. The CP190 motif matches a known Vertebrate centromeric sequence [23]. However, the top motifs discovered using the ±100 bp regions are highly degenerate suggesting that while the factors may not bind the DNA directly, co-factors might bind in the more distant vicinity of their peaks.
Found at: doi:10.1371/journal.pgen.1000814.s004 (0.75 MB JPG)

Figure S5 CTCF is a constitutive feature of the *Drosophila* genome. (A,B) In these genome browser views the ChIP-chip profiles for CTCF-C and CTCF-N in embryos are represented as top two tracks. Also represented are the ChIP-chip profiles for CTCF-N in two different cell lines: S2 cells and Kc cells.
Found at: doi:10.1371/journal.pgen.1000814.s005 (0.62 MB JPG)

Figure S6 Decreased signal intensity at cell-type specific CTCF binding sites. (A) A Venn diagram showing the overlap between the binding sites for CTCF in embryos, in S2 cells and Kc cells. (B) The mean and standard deviation of the fold change for each pair-wise comparison between CTCF-C [embryos] and CTCF-N [embryos, S2 cells, Kc cells] is plotted for the peaks that do overlap, and the peaks that don't. The same statistical criteria applied to different datasets might not represent the variation between the different biological samples.
Found at: doi:10.1371/journal.pgen.1000814.s006 (0.39 MB JPG)

Figure S7 A joint-model analysis of the binding sites of CTCF in different tissues. All the raw data from CTCF ChIP-chip in

different tissues have been analysed together with a joint model (see Text S1). A p value corresponding to 1% FDR has been applied to identify the binding sites. The same p value threshold has been applied to estimate the statistical difference of a peak in one condition compared to the others. (A,B) A comparative genome browser view of the results obtained by the joint model and a MAT analysis. In the first example (A) no difference is detected among the 3 profiles, while in (B) a binding site for CTCF upstream of the Fas3 gene is absent in Kc cells.
Found at: doi:10.1371/journal.pgen.1000814.s007 (0.65 MB JPG)

Figure S8 Distribution of the different classes of insulator binding sites compared to genomic features of *Drosophila*. (A) Barchart indicating the number of insulator binding sites of each class mapping to 5′ UTRs (red), exons (blue), introns (green), 3′ UTRs (purple), and intergenic regions (orange). For comparison, this distribution is also plotted for the set of transcription factors from MacArthur et al. [24] and for H3K4me3. (B) Data as in (A) normalized within each class to illustrate the fraction of insulators mapping to each annotation type. Also plotted at the right of the graph is the percentage of each region present in the Dm3 assembly of the *Drosophila* genome.
Found at: doi:10.1371/journal.pgen.1000814.s008 (0.57 MB JPG)

Figure S9 Distribution of the distance of insulator proteins binding sites relative to Transposable Elements. Estimated enrichment of insulator binding sites (black lines), with flanking 95% confidence intervals (gray lines) (Y-axis) are plotted against binding site base pair position (x-axis), relative to transposable element boundaries. Negative positions indicate binding sites within an annotated transposable element, 0 indicates the element boundary, and positive values represent positions outside and flanking element annotations.
Found at: doi:10.1371/journal.pgen.1000814.s009 (0.53 MB JPG)

Figure S10 Expression status of *Drosophila* embryos. (A,B) Enrichment and 95% confidence intervals (Y-axis) plotted against distance to transcription start sites (x-axis) for identified PolII enriched regions (A) or H3K4Me3 enriched regions (B). (C) Venn Diagram representing genes associated with a PolII binding sites at their TSS, an H3K4me3 mark at their TSS and a RNA signal on their exon.
Found at: doi:10.1371/journal.pgen.1000814.s010 (0.34 MB JPG)

Figure S11 Example of position of insulator binding sites at divergent promoters. A genome browser example of signal obtained by ChIP-chip for H3K4me3 (purple), PolII (red), as well as total RNA profiling on tiling microarrays (orange). Insulator binding sites are also represented in this example where we can observe that a Class I insulator, defined by the binding of CTCF, CP190, BEAF-32, and Mod(mdg4), is located between the divergent genes CG6509 and CG4970 which are separated by approximately 350 bp. CG6509 is transcribed as identified by its RNA level and have an active promoter, as identified by the presence of PolII and H3K4me3 at its TSS. CG4970, however, is inactive, thus suggesting that the presence of the insulator allows CG4970 to be activated independently of CG6509.
Found at: doi:10.1371/journal.pgen.1000814.s011 (0.31 MB JPG)

Figure S12 Distribution of insulator binding sites around the TSS of genes dependent of their transcription status. Log enrichment or depletion of insulator binding sites (y-axis) are plotted against binding site base pair position (x-axis), relative to the transcription start sites; negative and positive values depict upstream and downstream binding, respectively. Each panel corresponds to cases where the promoter is either active (On) or inactive (Off), as defined by the presence or absence of H3K4me3 and PolII (Figure S10C) and the transcriptional status of the nearest upstream promoter. (A) The gene TSS is on and the nearest upstream TSS is on. (B) The gene TSS is On and the nearest upstream TSS is Off. (C) The gene TSS is Off and the nearest upstream TSS is Off. (D) The gene TSS is Off and the nearest upstream promoter is On.
Found at: doi:10.1371/journal.pgen.1000814.s012 (0.50 MB JPG)

Figure S13 Enrichment of CTCF binding sites between Adjacent Promoters in different cell lines. (A) Venn diagram between H3K4me3 associated promoters in embryos, S2 cells and Kc cells. Enrichment and 95% confidence intervals of CTCF binding sites in embryos, S2, and Kc cells between (B) divergent and (C) alternative promoters.
Found at: doi:10.1371/journal.pgen.1000814.s013 (0.34 MB JPG)

Figure S14 Enrichment of different classes of binding sites between CRMs and Promoters. For each insulator binding site class, enrichment estimates and flanking confidence intervals (Y-axis) are plotted for genomic intervals with 0, 1, or 2 insulator binding sites (X-axis). Intervals are defined by the region between *cis*-regulatory elements and their target promoters (red), or between *cis*-regulatory elements and their nearest non-target promoters (black), or between adjacent *cis*-regulatory elements that regulate the same gene (gray).
Found at: doi:10.1371/journal.pgen.1000814.s014 (0.36 MB JPG)

Figure S15 Insulator proteins do not associate with transcription factors binding sites. Binding sites from 36 datasets corresponding to 21 transcription factors [24] were downloaded from UCSC and compared to our set of insulator binding sites. A distance matrix was constructed as 1 minus the fraction of sites with midpoint to midpoint distances less than 250 bases (data in Table S4) and hierarchically clustered using the average linkage method. Cell colors range from blue to white to red to depict increasing site overlap.
Found at: doi:10.1371/journal.pgen.1000814.s015 (0.59 MB JPG)

Figure S16 Enrichment at the TSS of insulator proteins and transcription factors. Binding sites from 36 datasets corresponding to 21 transcription factors (BDTNP; [24]) were downloaded from UCSC and compared to our set of insulator binding sites. Log2 enrichment or depletion of insulator binding sites and associated 95% confidence intervals (Y-axis) are plotted against binding site base pair position (X-axis), relative to the transcription start sites; negative and positive values depict upstream and downstream binding, respectively. A point of comparison for a promoter associated factor TFIIB is represented in light blue.
Found at: doi:10.1371/journal.pgen.1000814.s016 (0.59 MB JPG)

Figure S17 Promoter demarcation by insulators and transcription factors. Same Legend as in Figure 4. Now represented in black are the data corresponding to the BDTNP datasets. TFIIB is represented in light blue.
Found at: doi:10.1371/journal.pgen.1000814.s017 (0.53 MB JPG)

Figure S18 CTCF at the boundaries of H3K27me3 domains. Genome browser example showing signals for CTCF and H3K27me3 ChIP-chip experiments performed in embryos and S2 cells. The H3K27me3 data in S2 cells is reanalyzed from [3] (see Text S1). HMM segmentation is used to define the boundaries of H3K27me3 better. This can be visualized on these examples when compared with a MAT analysis performed on embryos. The dashed vertical lines show CTCF binding sites at domain boundaries. (A) Example of CTCF bordering an H3K27me3 domain covering the btd/Sp1 locus. (B) In this example, an entire H3K27me3 domain corresponding to the *Abd-B* gene disappears,

while this chromatin mark is maintained in the rest of the Bithorax Complex region. This depletion of the H3K27me3 mark corresponds to *Abd-B* being expressed in S2 cells while *Ubx* and *abd-A* are repressed. Embryos corresponding to a mixed population of cells, the H3K27me3 signal is coming from its presence in a subpopulation of cells. All CTCF binding sites in this region are conserved between S2 cells and embryos, even inside the H3K27me3 depleted domain (between the dashed vertical lines) and are therefore independent of the transcriptional status of *Abd-B*. Furthermore, the breakpoints of the H3K27me3 depleted domain in S2 cells, compared to embryos, correspond to CTCF binding sites (represented by the two vertical dashed lines).
Found at: doi:10.1371/journal.pgen.1000814.s018 (0.78 MB JPG)

Figure S19 Conservation of insulator binding sites. Phastcons between 15 insect species, including the 12 sequenced *Drosophilae* species, have been calculated for each category of insulator binding sites. The bars correspond to the median (dot) and median absolute deviation (bars) of the scores. The dark red bar (FEI) corresponds to the same scores calculated for fast evolving introns (neutral reference). Also plotted for reference are exons (pink), H3K4me3 (gray), and BDTNP binding sites (black).
Found at: doi:10.1371/journal.pgen.1000814.s019 (0.30 MB JPG)

Figure S20 Dynamic chromatin at insulator binding sites at non-promoter and promoter sites. Each insulator site, defined as the midpoint of the binding site interval, was classified as to whether it fell within the interval defined by a transcriptional start site and 500-bp upstream, using the 12,807 unique 5′ ends annotated in FlyBase r5.13. Based on these criteria, the number of sites in non-promoters and promoters are: BEAF-32 (5546 nonpromoters, 2281 promoters); CP190 (7758 non-promoters, 2698 promoters); CTCF (3286 non-promoters, 1146 promoters); Mod(mdg4) (3154 non-promoters, 821 promoters); GAF (5551 non-promoters, 887 promoters); Su(Hw) 4565 non-promoters, 214 promoters). Displays are for non-promoters (A-E) and promoters (F-J) using the same datasets shown in Figure 5. (A,F) Nucleosome density; (B,G) 80 mM salt fraction; (C,H) 150 mM salt fraction; (D,I) 600 mM salt fraction; (E,J) salt-washed pellet.
Found at: doi:10.1371/journal.pgen.1000814.s020 (0.82 MB JPG)

Table S1 Binding sites of the insulator-associated proteins. The number of binding sites per factor at different confidence interval generated by MAT analysis.
Found at: doi:10.1371/journal.pgen.1000814.s021 (0.02 MB XLS)

Table S2 Known insulators detected by our ChIP-chip analysis. For each published functional insulator element, an X mark indicates if they are associated with a peak for each factor identified in our ChIP-Chip experiments.
Found at: doi:10.1371/journal.pgen.1000814.s022 (0.02 MB XLS)

Table S3 Binding sites containing a discovered consensus motif. Number of regions containing the corresponding discovered motif (see Motif Discovery methods) for each factor at different PWM thresholds (e.g. 6 indicates matching the genome with 4^{-6} p-value and is the most lenient threshold). The 4 panels represent the set of regions studied (all regions versus uniquely bound regions) at different scanning windows ($\pm$100 bp and $\pm$1000 bp around the peak centers). The numbers in each cell indicate the number of intergenic peaks that contain a motif/the total number of intergenic regions. The number after each colon indicates the enrichment of motif instances inside the considered regions (compared to the fraction of the intergenic genome the regions represent). Motifs are in general good predictors for CTCF, Su(Hw), BEAF-32, and GAF (as evidenced by the high enrichment). Within a distance of $\pm$1,000 bp, and at a PWM p-value of 4^{-6}, the discovered motifs are present in 75.6% of CTCF, 86.8% of BEAF-32, 84% of Su(Hw) and 88.6% of GAF binding sites.
Found at: doi:10.1371/journal.pgen.1000814.s023 (0.03 MB XLS)

Table S4 Binding sites overlaps. Fraction of each factor's binding sites with midpoint to midpoint distances less than 250 bases, for each other factor type studied.
Found at: doi:10.1371/journal.pgen.1000814.s024 (0.07 MB XLS)

Text S1 Supplementary methods.
Found at: doi:10.1371/journal.pgen.1000814.s025 (0.07 MB DOC)

Acknowledgments

We would like to thank V. Corces, C. Wu, and P. Geyer for the generous gift of affinity reagents and the Functional Genomics Facility at the University of Chicago for processing the Affymetrix arrays. We also thank Gregory Cooper, Cheryl Smith, and Thomas Stricker for critical reading of the manuscript.

Author Contributions

Conceived and designed the experiments: NN CDB KPW. Performed the experiments: NN CAM KA. Analyzed the data: NN CDB PKS PK JGH XF SH MK. Contributed reagents/materials/analysis tools: SR RAHW LS. Wrote the paper: NN CDB KPW.

References

1. Kellum R, Schedl P (1991) A position-effect assay for boundaries of higher order chromosomal domains. Cell 64: 941–950.
2. Kellum R, Schedl P (1992) A group of scs elements function as domain boundaries in an enhancer-blocking assay. Mol Cell Biol 12: 2424–2431.
3. Holdridge C, Dorsett D (1991) Repression of hsp70 heat shock gene transcription by the suppressor of hairy-wing protein of *Drosophila* melanogaster. Mol Cell Biol 11: 1894–1900.
4. Geyer PK, Corces VG (1992) DNA position-specific repression of transcription by a *Drosophila* zinc finger protein. Genes Dev 6: 1865–1873.
5. Wallace JA, Felsenfeld G (2007) We gather together: insulators and genome organization. Curr Opin Genet Dev 17: 400–407.
6. Chung JH, Whiteley M, Felsenfeld G (1993) A 5′ element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in *Drosophila*. Cell 74: 505–514.
7. Bell AC, West AG, Felsenfeld G (1999) The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell 98: 387–396.
8. Moon H, Filippova G, Loukinov D, Pugacheva E, Chen Q, et al. (2005) CTCF is conserved from *Drosophila* to humans and confers enhancer blocking of the Fab-8 insulator. EMBO Rep 6: 165–170.
9. Ciavatta D, Rogers S, Magnuson T (2007) *Drosophila* CTCF is required for Fab-8 enhancer blocking activity in S2 cells. J Mol Biol 373: 233–239.
10. Holohan EE, Kwong C, Adryan B, Bartkuhn M, Herold M, et al. (2007) CTCF genomic binding sites in *Drosophila* and the organisation of the bithorax complex. PLoS Genet 3: e112. doi:10.1371/journal.pgen.0030112.
11. Mohan M, Bartkuhn M, Herold M, Philippen A, Heinl N, et al. (2007) The *Drosophila* insulator proteins CTCF and CP190 link enhancer blocking to body patterning. EMBO J 26: 4203–4214.
12. Golovnin A, Biryukova I, Romanova O, Silicheva M, Parshikov A, et al. (2003) An endogenous Su(Hw) insulator separates the yellow gene from the Achaete-scute gene complex in *Drosophila*. Development 130: 3249–3258.
13. Parnell TJ, Viering MM, Skjesol A, Helou C, Kuhn EJ, et al. (2003) An endogenous suppressor of hairy-wing insulator separates regulatory domains in *Drosophila*. Proc Natl Acad Sci U S A 100: 13436–13441.
14. Adryan B, Woerfel G, Birch-Machin I, Gao S, Quick M, et al. (2007) Genomic mapping of Suppressor of Hairy-wing binding sites in *Drosophila*. Genome Biol 8: R167.
15. Gause M, Morcillo P, Dorsett D (2001) Insulation of enhancer-promoter communication by a gypsy transposon insert in the *Drosophila* cut gene:

cooperation between suppressor of hairy-wing and modifier of mdg4 proteins. Mol Cell Biol 21: 4807–4817.
16. Ghosh D, Gerasimova TI, Corces VG (2001) Interactions between the Su(Hw) and Mod(mdg4) proteins required for gypsy insulator function. EMBO J 20: 2518–2527.
17. Pai CY, Lei EP, Ghosh D, Corces VG (2004) The centrosomal protein CP190 is a component of the gypsy chromatin insulator. Mol Cell 16: 737–748.
18. Zhao K, Hart CM, Laemmli UK (1995) Visualization of chromosomal domains with boundary element-associated factor BEAF-32. Cell 81: 879–889.
19. Gaszner M, Vazquez J, Schedl P (1999) The Zw5 protein, a component of the scs chromatin domain boundary, is able to block enhancer-promoter interaction. Genes Dev 13: 2098–2107.
20. Ohtsuki S, Levine M (1998) GAGA mediates the enhancer blocking activity of the eve promoter in the *Drosophila* embryo. Genes Dev 12: 3325–3330.
21. Belozerov VE, Majumder P, Shen P, Cai HN (2003) A novel boundary element may facilitate independent gene regulation in the Antennapedia complex of *Drosophila*. EMBO J 22: 3113–3121.
22. Schweinsberg S, Hagstrom K, Gohl D, Schedl P, Kumar RP, et al. (2004) The enhancer-blocking activity of the Fab-7 boundary from the *Drosophila* bithorax complex requires GAGA-factor-binding sites. Genetics 168: 1371–1384.
23. Roseman RR, Pirrotta V, Geyer PK (1993) The su(Hw) protein insulates expression of the *Drosophila* melanogaster white gene from chromosomal position-effects. EMBO J 12: 435–442.
24. Kahn TG, Schwartz YB, Dellino GI, Pirrotta V (2006) Polycomb complexes and the propagation of the methylation mark at the *Drosophila* ubx gene. J Biol Chem 281: 29064–29075.
25. Mallin DR, Myung JS, Patton JS, Geyer PK (1998) Polycomb group repression is blocked by the *Drosophila* suppressor of Hairy-wing [su(Hw)] insulator. Genetics 148: 331–339.
26. Sigrist CJ, Pirrotta V (1997) Chromatin insulator elements block the silencing of a target gene by the *Drosophila* polycomb response element (PRE) but allow trans interactions between PREs on different chromosomes. Genetics 147: 209–221.
27. Cuddapah S, Jothi R, Schones DE, Roh TY, Cui K, et al. (2009) Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Res 19: 24–32.
28. Emberly E, Blattes R, Schuettengruber B, Hennion M, Jiang N, et al. (2008) BEAF regulates cell-cycle genes through the controlled deposition of H3K9 methylation marks into its conserved dual-core binding sites. PLoS Biol 6: e327. doi:10.1371/journal.pbio.0060327.
29. Jothi R, Cuddapah S, Barski A, Cui K, Zhao K (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Res 36: 5221–5231.
30. Bartkuhn M, Straub T, Herold M, Herrmann M, Rathke C, et al. (2009) Active promoters and insulators are marked by the centrosomal protein 190. EMBO J.
31. Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, et al. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell 128: 1231–1245.
32. Spana C, Harrison DA, Corces VG (1988) The *Drosophila* melanogaster suppressor of Hairy-wing protein binds to specific sequences of the gypsy retrotransposon. Genes Dev 2: 1414–1423.
33. Pollard D (2008) Bergman 2004 matrices.
34. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, et al. (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature.
35. Bushey AM, Ramos E, Corces VG (2009) Three subclasses of a *Drosophila* insulator show distinct and cell type-specific genomic distributions. Genes Dev 23: 1338–1350.
36. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125: 315–326.
37. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553–560.
38. Pan G, Tian S, Nie J, Yang C, Ruotti V, et al. (2007) Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. Cell Stem Cell 1: 299–312.
39. Zhao XD, Han X, Chew JL, Liu J, Chiu KP, et al. (2007) Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. Cell Stem Cell 1: 286–298.
40. MacArthur S, Li XY, Li J, Brown JB, Chu HC, et al. (2009) Developmental roles of 21 *Drosophila* transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. Genome Biol 10: R80.
41. Cao R, Zhang Y (2004) The functions of E(Z)/EZH2-mediated methylation of lysine 27 in histone H3. Curr Opin Genet Dev 14: 155–164.
42. Schwartz YB, Kahn TG, Nix DA, Li XY, Bourgon R, et al. (2006) Genome-wide analysis of Polycomb targets in *Drosophila* melanogaster. Nat Genet 38: 700–705.
43. Schuettengruber B, Ganapathi M, Leblanc B, Portoso M, Jaschek R, et al. (2009) Functional anatomy of polycomb and trithorax chromatin landscapes in *Drosophila* embryos. PLoS Biol 7: e13. doi:10.1371/journal.pbio.1000013.
44. Tolhuis B, de Wit E, Muijrers I, Teunissen H, Talhout W, et al. (2006) Genome-wide profiling of PRC1 and PRC2 Polycomb chromatin binding in *Drosophila* melanogaster. Nat Genet 38: 694–699.
45. Negre N, Hennetin J, Sun LV, Lavrov S, Bellis M, et al. (2006) Chromosomal distribution of PcG proteins during *Drosophila* development. PLoS Biol 4: e170. doi:10.1371/journal.pbio.0040170.
46. Halligan DL, Keightley PD (2006) Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. Genome Res 16: 875–884.
47. Bhutkar A, Schaeffer SW, Russo SM, Xu M, Smith TF, et al. (2008) Chromosomal rearrangement inferred from comparisons of 12 *Drosophila* genomes. Genetics 179: 1657–1680.
48. Engstrom PG, Ho Sui SJ, Drivenes O, Becker TS, Lenhard B (2007) Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Res 17: 1898–1908.
49. Dion MF, Kaplan T, Kim M, Buratowski S, Friedman N, et al. (2007) Dynamics of replication-independent histone turnover in budding yeast. Science 315: 1405–1408.
50. Mito Y, Henikoff JG, Henikoff S (2005) Genome-scale profiling of histone H3.3 replacement patterns. Nat Genet 37: 1090–1097.
51. Henikoff S, Henikoff JG, Sakai A, Loeb GB, Ahmad K (2009) Genome-wide profiling of salt fractions maps physical properties of chromatin. Genome Res 19: 460–469.
52. Smith ST, Wickramasinghe P, Olson A, Loukinov D, Lin L, et al. (2009) Genome wide ChIP-chip analyses reveal important roles for CTCF in *Drosophila* genome organization. Dev Biol 328: 518–528.
53. Jiang N, Emberly E, Cuvier O, Hart CM (2009) Genome-wide mapping of boundary element-associated factor (BEAF) binding sites in *Drosophila* melanogaster links BEAF to transcription. Mol Cell Biol 29: 3556–3568.
54. Bartkuhn M, Straub T, Herold M, Herrmann M, Rathke C, et al. (2009) Active promoters and insulators are marked by the centrosomal protein 190. EMBO J 28: 877–888.
55. Maeda RK, Karch F (2003) Ensuring enhancer fidelity. Nat Genet 34: 360–361.
56. Lee CK, Shibata Y, Rao B, Strahl BD, Lieb JD (2004) Evidence for nucleosome depletion at active regulatory regions genome-wide. Nat Genet 36: 900–905.
57. Geyer PK (1997) The role of insulator elements in defining domains of gene expression. Curr Opin Genet Dev 7: 242–248.
58. Chopra VS, Cande J, Hong JW, Levine M (2009) Stalled Hox promoters as chromosomal boundaries. Genes Dev 23: 1505–1509.
59. Negre N, Lavrov S, Hennetin J, Bellis M, Cavalli G (2006) Mapping the distribution of chromatin proteins by ChIP on chip. Methods Enzymol 410: 316–341.
60. Whitfield WG, Chaplin MA, Oegema K, Parry H, Glover DM (1995) The 190 kDa centrosome-associated protein of *Drosophila* melanogaster contains four zinc finger motifs and binds to specific sites on polytene chromosomes. J Cell Sci 108 (Pt 11): 3377–3387.
61. Gerasimova TI, Corces VG (1998) Polycomb and trithorax group proteins mediate the function of a chromatin insulator. Cell 92: 511–521.
62. Gerasimova TI, Gdula DA, Gerasimov DV, Simonova O, Corces VG (1995) A *Drosophila* protein that imparts directionality on a chromatin insulator is an enhancer of position-effect variegation. Cell 82: 587–597.
63. Parnell TJ, Kuhn EJ, Gilmore BL, Helou C, Wold MS, et al. (2006) Identification of genomic sites that bind the *Drosophila* suppressor of Hairy-wing insulator protein. Mol Cell Biol 26: 5983–5993.
64. Tsukiyama T, Wu C (1996) Purification of GAGA factor of *Drosophila* and its role in nucleosome disruption. Methods Enzymol 274: 291–299.
65. Johnson WE, Li W, Meyer CA, Gottardo R, Carroll JS, et al. (2006) Model-based analysis of tiling-arrays for ChIP-chip. Proc Natl Acad Sci U S A 103: 12457–12462.
66. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2: 28–36.
67. Hughes JD, Estep PW, Tavazoie S, Church GM (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J Mol Biol 296: 1205–1214.
68. Liu XS, Brutlag DL, Liu JS (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat Biotechnol 20: 835–839.
69. Kheradpour P, Stark A, Roy S, Kellis M (2007) Reliable prediction of regulator targets using 12 *Drosophila* genomes. Genome Res 17: 1919–1931.
70. Touzet H, Varre JS (2007) Efficient and accurate P-value computation for Position Weight Matrices. Algorithms Mol Biol 2: 15.
71. Halfon MS, Gallo SM, Bergman CM (2008) REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in *Drosophila*. Nucleic Acids Res 36: D594–598.

Genome Biology

MEETING REPORT

Sequences to systems

Manolis Kellis[1,2] and John L Rinn[1,3,4]*

Abstract

A report of the seventh annual meeting on Systems Biology: Global Regulation of Gene Expression, 23-27 March 2010, Cold Spring Harbor, USA.

From a chemical standpoint, a cell can seem as a densely intertwined mess of molecules, but their emergent systems-level properties exhibit structure, coordination and order, leading to the orchestrated complexity we call life. As the seventh annual meeting on Systems Biology: Global Regulation of Gene Expression at the Cold Spring Harbor Laboratory showcased, systems biology aims to bridge the vast gap of knowledge between the pairwise molecular interactions of individual cellular components and the large-scale behavior of cellular circuits that result from their synergism. The talks given at the meeting exemplified a new definition of systems biology, where genomic technologies can revolutionize our understanding of classical problems by enabling molecular interactions to be linked to systems behavior in a mechanistic way.

A large number of definitions have been proposed for systems biology, including 'the study of a single pathway by someone with an attention-deficit disorder' humorously offered by Irv Bernstein and quoted in a keynote address by Steve Henikoff (Fred Hutchinson Cancer Research Center, Seattle, USA). Henikoff went on to demonstrate that Trithorax and Polycomb response elements that trigger cellular differentiation could be recognized by their patterns of histone turnover. He argued that histone variant H2A.Z is mutually exclusive with DNA methylation, drawing a novel connection between histones, their dynamics and the epigenetic inheritance of DNA methylation.

The importance of epigenetic information in mediating the stable inheritance of cellular states was underscored in the context of human disease by the second keynote speaker, Kevin Struhl (Havard Medical School, Boston, USA). By dissecting the transcriptional circuits responding to Src signaling that can yield stably transformed cells, Struhl uncovered a feedback loop mediated by the transcription factor NFκB that is sustainable for many generations in the absence of further signaling. Struhl went further to demonstrate that this pathway is targetable using FDA-approved drugs, opening up novel avenues of cancer therapeutics.

Parsing the genome

The study of any system begins with a comprehensive list of parts and their interactions, and so a major theme of the meeting was the mapping of the building blocks of gene regulation: the control regions, the DNA words within them and their sentence structure.

On one hand, several talks focused on defining sites of transcriptional factor binding, histones and sites of DNA methylation at ever-increasing resolution and complexity. For example, two talks by Barbara Wold (California Institute of Technology) and Stephen Tapscott (Fred Hutchinson Cancer Research Center) presented data revealing thousands of binding sites for the master regulator of muscle differentiation MyoD, concluding that numerous additional functions and or combinatorics result in cell-fate decisions such as muscle differentiation. One of us (MK), presented a new algorithm specifically for revealing biologically meaningful combinations of chromatin marks, and their corresponding chromatin states, that reflect rich information about the underlying biological function of the associated genomic regions. The algorithms have distinguished more than 50 chromatin states, including 11 different classes of promoter regions, associated with different functional categories, and several classes of enhancer regions associated with different expression levels for downstream genes.

On the other hand, several talks focused on the DNA sequences predictive of regulatory activity. Eric Mendenhall (Massachusetts General Hospital, Boston, USA) presented remarkable data that distinct types of DNA sequences can recruit key chromatin factors such as Polycomb. Through multiple transgenic studies Mendenhall showed that he could insert DNA sequence motifs, even those 'that have never seen Polycomb before', into

*Correspondence: jrinn@bidmc.harvard.edu
[1]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142, USA
Full list of author information is available at the end of the article

BioMed Central © 2010 BioMed Central Ltd

the mouse genome that subsequently recruited Polycomb binding, suggesting a model of sequence-based recruitment of this complex. Eileen Furlong (European Molecular Biology Laboratory, Heidelberg, Germany) demonstrated the importance of combinatorial binding events in determining cell fate decisions, using a support vector machine to model transcription factor binding data and reveal the *cis*-regulatory motifs and motif combinations that define mesodermal fate in *Drosophila melanogaster*. A major conclusion was clear evidence that a single expression pattern could result from several different combinatorial codes of transcription factors and their dynamics. Arend Sidow (Stanford University, Stanford, USA) presented interesting insight into what determines nucleosome occupancy during the differentiation of distinct blood cell types both *in vitro* and *in vivo*. He presented clear evidence that *in vitro* nucleosomes associate with sequence-specific motifs, such as GC motifs, and exclude others, such as AT sites, whereas *in vivo* these sequence preferences are largely overridden. Moreover, Sidow demonstrated that nucleosome positioning is globally remodeled across different states of cellular differentiation.

Collectively these studies demonstrate the power of chromatin-level and DNA-level information for predicting regulatory activity, the importance of word combinations and mark combinations, and the interplay of static sequences and dynamic cell-type specific changes in defining the processes of differentiation and, more generally, gene regulation.

Unconventional interactions of the secondary type

The late Ed Lewis famously said 'Progress will still need to be driven by the logic of genetics and by further increases in abstraction'. This sentiment was echoed at this year's meeting, most prominently in the unveiling of numerous unconventional protein-DNA and protein-RNA interactions.

A dramatic example was presented by Heng Zhu (Johns Hopkins Medical Center, Baltimore, USA), who has used human proteome arrays to map global DNA-binding interactions. His analysis identified a striking and atypical interaction between kinases and DNA. Specifically, Zhu identified a carboxy-terminal domain of the kinase ERK2 that binds to DNA with nanomolar affinity and through this interaction serves as a transcriptional repressor. He also presented initial evidence for more than 300 unconventional DNA-binding proteins, including unusual suspects such as RNA-binding proteins and mitochondrial proteins.

Howard Chang (Stanford University, Stanford, USA) presented a comprehensive study demonstrating modular binding of the large intergenic noncoding RNA (lincRNA) HOTAIR to chromatin-modifying complexes such as PRC2, the histone demethylase LSD1, and COREST, a DNA-binding protein. Chang presented a model, based on functional assays, whereby HOTAIR binds all three of these unconventional RNA-binding proteins through a different specific structured domain and in this way can guide them specifically to their target genomic loci. Thus, lincRNAs could serve as adaptors to facilitate the formation of specific RNA-protein interactions. One of us (JR) presented evidence that such mediation is a global phenomenon that extends beyond HOTAIR and includes hundreds of lincRNAs interacting with chromatin-modifying complexes and other DNA-binding proteins involved in cancer. Overall, the genome could be replete with unorthodox RNA-binding interactions serving as RNA adaptors to facilitate the establishment and or maintenance of specific epigenetic states. Tom Gingeras (Cold Spring Harbor Laboratory, New York, USA) expanded the horizons of unorthodox RNA functions, presenting a rich diversity of noncoding RNA molecules that localize to distinct cellular compartments and are derived from unconventional processes, emphasizing the further need for abstraction in our understanding of the mammalian transcriptome and the potential regulatory roles of noncoding RNAs.

Close encounters of the three-dimensional type

At this year's meeting the genome came to life in three dimensions, as a result of 'proximity ligation' techniques that define the three-dimensional neighbors of a given genomic sequence relative to all others in the genome. This reveals potentially important regulatory interactions mediated by DNA looping and compartmentalization.

Job Dekker (University of Massachusetts, Worcester, USA) described work using several variants of chromatin-conformation capture (3C) to reveal these interactions. Dekker and his colleagues have used carbon-copy 3C (5C) to resolve the three-dimensional structure of the beta-globin locus in both the activated and repressed states. He proposed a model by which a large genomic-linker region serves to open up or compact the locus upon activation or repression, respectively. Erez Lieberman-Aiden (Broad Institute, Cambridge, USA) has expanded the technique to take on the entire genome, and reported that megabases of DNA fold into a dense, highly organized, and knot-free structure known as a 'fractal globule', thus providing a glimpse of the three-dimensional genome.

Using a different proximity ligation approach called ChIA-PET, Yijun Ruan (Genome Institute of Singapore, Singapore) has revealed three-dimensional genomic architecture in the context of transcriptional regulation. Ruan specifically demonstrated tertiary DNA interactions mediated by transcription factors and RNA polymerase II. An interesting connection to this finding was presented

by Martin Hemberg (Children's Hospital, Harvard Medical School, Boston, USA) where he described that numerous enhancers, which often result from long-range DNA interactions, are associated with RNA polymerase II and are transcribed at low levels.

Collectively these studies have added a new dimension to systems biology and provide novel insights into the three-dimensional architecture of the human genome.

Systematic evolution

The evolutionary dimension of gene regulation was addressed in regard to the selective pressures and plasticity of regulatory regions and elements. Ollie Rando (University of Massachusetts, Worcester, USA) presented an evolutionary analysis of nucleosome positioning in 12 ascomycete species. Perhaps most notably, the nucleosome positioning corresponded to specific gene sets, such as mitochondrial genes or splicing genes, and differed in physiologically relevant ways in different species.

Continuing the theme of evolution and binding localization, Mike Snyder (Stanford University, Stanford, USA) presented an unprecedented dataset comprising the binding sites for several key transcription factors across ten different human individuals. Snyder showed that genetic variation between individuals had resulted in the loss and gain of binding sites for key transcriptional players such as RNA polymerase II and the transcription factor NFκB. Thus, individual variation results in unique binding patterns of key factors, which should provide potential insights for personalized medicine and disease susceptibility.

Alexander Stark (IMP, Vienna, Austria) presented comparative work on the DNA binding of transcriptional regulators of mesoderm development in six *Drosophila* species at a range of evolutionary distances from each other. He showed that not only were the locations of the peaks representing binding-site occupancy conserved, but also the intensity of the peaks, suggesting that lower-occupancy sites are also functional and are selectively maintained. He presented evidence that these sites were subsequently occupied at later developmental stages, suggesting that lower occupancy may be associated with priming for subsequent activation.

Cell circuits and dynamics

The ultimate goal in systems biology is to define cellular circuits and their resulting cellular outputs. This year's meeting reported unprecedented progress in mapping cell circuits in response to specific stimuli, perturbing them and monitoring their cellular responses. A comprehensive example was presented by Aviv Regev (Broad Institute, Cambridge, USA), who described work with a model mammalian cell system - mouse dendritic cells - which trigger distinct inflammatory or antiviral responses depending on the input stimuli. Regev's strategy was to monitor the complete transcriptional response to each stimulus, identify the key regulators (transcription factors, chromatin-modifying complexes and RNA-binding proteins) in each pathway, and then individually perturb the actions of hundreds of these regulators via stable gene knockdown and monitor the resulting transcriptional changes. This not only identified numerous new regulators across these critical cellular pathways, but also revealed the cooption of cell-cycle circuits in the antiviral response. This approach to systems biology could also be applied to numerous other cell-based models.

Underlying seemingly stable cell circuits are multiple additional layers of modulation, which include both stochastic single-cell variability in transcript levels and regulated processes of post-transcriptional inhibition and splicing. Alexander Van Oudenaarden (Massachusetts Institute of Technology, Cambridge, USA) demonstrated that cell-fate decisions in *Caenorhabditis elegans* intestinal specification arise from an underlying process of variable transcription levels. These stochastic processes resulted in threshold-dependent ON or OFF decisions of master regulators that guide cellular differentiation. David Bartel (Whitehead Institute, Cambridge, USA) described how the dramatic effects of post-transcriptional regulation by microRNAs (miRNAs) can result from even the slightest fine-tuning of transcript levels and translation efficiency. In a potential reversal of the dominant dogma in miRNA regulation, he also presented evidence from mammalian cells that for two miRNAs at least, the major effect of targeting appears to be transcript degradation rather than translational inhibition. Chris Burge (Massachusetts Institute of Technology, Cambridge, USA) presented studies that revealed a remarkable complexity in mRNA transcript splicing and how these isoforms are altered in cancer.

Collectively, this year's meeting revealed new insight into genomic sequences, binding sites and structures that has moved systematic study of gene regulation to new dimensions of both understanding and further abstraction in systems biology.

Author details
[1]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142, USA. [2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [3]Department of Pathology, Beth Israel Deaconess Medical Center, Boston, MA 02215, USA. [4]Department of Pathology, Harvard Medical School, Boston, MA 02115, USA.

Published: 25 May 2010

doi:10.1186/gb-2010-11-5-303
Cite this article as: Kellis M, Rinn JL: Sequences to systems. *Genome Biology* 2010, **11**:303.

MBE	msq189	GM
Journal Name	Art. No.	CE Code

NOT FOR PUBLIC RELEASE

A Bayesian Approach for Fast and Accurate Gene Tree Reconstruction

[AQ1] Matthew D. Rasmussen *,[1] and Manolis Kellis*,[1,2]

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology

[AQ2] [2]Broad Institute of MIT and Harvard

***Corresponding author:** E-mail: manoli@mit.edu; rasmus@mit.edu.

Associate editor: Sudhir Kumar

Abstract

Recent sequencing and computing advances have enabled phylogenetic analyses to expand to both entire genomes and large clades, thus requiring more efficient and accurate methods designed specifically for the phylogenomic context. Here, we present SPIMAP, an efficient Bayesian method for reconstructing gene trees in the presence of a known species tree. We observe many improvements in reconstruction accuracy, achieved by modeling multiple aspects of evolution, including gene duplication and loss (DL) rates, speciation times, and correlated substitution rate variation across both species and loci. We have implemented and applied this method on two clades of fully sequenced species, 12 *Drosophila* and 16 fungal genomes as well as simulated phylogenies and find dramatic improvements in reconstruction accuracy as compared with the most popular existing methods, including those that take the species tree into account. We find that reconstruction inaccuracies of traditional phylogenetic methods overestimate the number of DL events by as much as 2–3-fold, whereas our method achieves significantly higher accuracy. We feel that the results and methods presented here will have many important implications for future investigations of gene evolution.

Key words: phylogenetics, gene tree, species tree, gene duplication and loss, reconciliation, Bayesian.

Research article

Introduction

Phylogenetic analysis has become an increasingly popular and fruitful approach for studying genomes (Hahn et al. 2005; Li et al. 2006; Hobolth et al. 2007; Wapinski et al. 2007; Butler et al. 2009). Methods for reconstructing phylogenies from sequence data have a long history (Felsenstein 1981; Saitou and Nei 1987; Rannala and Yang 1996) and new methods are continually developed to address a wide range of evolutionary questions. The question we approach in this work is the study of gene family evolution, namely how each family of genes has expanded and contracted over evolutionary time in a clade of related species. "Phylogenomics" has been proposed (Eisen 1998) as a systematic approach for studying gene families, where every gene family in several fully sequenced genomes is reconstructed and compared with a common species tree to infer orthologs, paralogs, and all evolutionary events, including gene duplications, losses, and horizontal transfers (Zmasek and Eddy 2002; Li et al. 2006; Huerta-Cepas et al. 2007; Wapinski et al. 2007; Butler et al. 2009; Vilella et al. 2009). However, as with any computational approach, the quality of the conclusions of phylogenomic studies are heavily dependent on the accuracy of the underlying methodologies. Accordingly, there has been much recent work on measuring and improving methods for phylogenetic reconstruction for both species trees and individual gene family trees. Advances have come from increased sequencing data for both additional taxa and loci as well as from new methods for leveraging that data.

For the problem of "species tree reconstruction," many advances have been made by combining data across loci either by concatenating multiple aligned loci into a "supermatrix" (Rokas et al. 2003; Ciccarelli et al. 2006), combining multiple gene trees into a"supertree" (Creevey and McInerney 2005), or by using a model for how such loci are correlated and coordinated in their evolution (Maddison and Knowles 2006; Liu and Pearl 2007). For example, in the BEST model (Liu and Pearl 2007), the correlated evolution of loci is captured by modeling a common species tree that constrains the evolution of each locus while still allowing some topological differences at each locus to occur via a coalescent process (Wakeley 2009). A probabilistic approach such as this allows one to use sequence alignments from multiple loci to estimate the posterior distribution of the species tree.

The problem of "gene tree reconstruction" also needs a similar strategy for exploiting abundant sequence data. Many recent efforts to reconstruct gene families in isolation (i.e., not accounting for their shared species tree or correlated evolution) have met many challenges. For example, the TreeFam project (Li et al. 2006) had found that automatic methods of reconstruction (such as maximum likelihood [ML], Felsenstein 1981; maximum a posteriori [MAP], Rannala and Yang 1996; neighbor joining [NJ], Saitou and Nei 1987; and parsimony, Felsenstein 2005) were not sufficiently accurate for systematic use and thus relied on human curators to adjust trees using additional information from the species tree, syntenic alignments, and the relevant literature. In a study by Hahn (2007), simulations were used to study how errors in gene tree reconstruction propagate into later inferences of gene duplication and loss (DL)

© The Author 2010. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oxfordjournals.org

events. In particular, the study showed that methods such as NJ frequently make reconstruction errors that lead to a biased inference of many erroneous duplications in ancestral lineages followed by numerous compensating losses in recent lineages.

In our own empirical work, we have found that the phylogenetic information available within a single locus is quite limited for most genes (Rasmussen and Kellis 2007). For example, in the recently sequenced 12 *Drosophila* species, we found that for alignments of orthologous genes, the inferred gene trees, regardless of the method used, have only a 38% chance of congruence with the species tree. For the 62% of alignments that supported an incongruent ML gene tree topology, only 5.7% did so with sufficient statistical significance ($P < 0.01$; SH test; Shimodaira and Hasegawa 1999). This along with several other measures of information content indicated that most loci lack enough information to confidently support one gene tree topology over the many other competing alternatives.

As we show below, the phylogenomic setting allows us to overcome the issue of limited information within individual loci by studying many gene families across the genome simultaneously. The additional information ultimately improves our ability to reconstruct gene trees but requires properly integrating information from both across species and genes while building upon several recent advances that we describe next.

Modeling Gene Trees and Species Trees

Our work fits within a growing body of literature addressing the simultaneous modeling of gene and species evolution. In one branch of this field, the primary concern is to model orthologous loci whose phylogeny may become incongruent with the species phylogeny due to incomplete lineage sorting (Maddison and Knowles 2006; Liu and Pearl 2007). In that case, gene trees are often modeled with the coalescent process (Wakeley 2009), which defines how topologies and branch lengths are distributed across loci (Rannala and Yang 2003), and has been used to reconstruct both gene trees (Hobolth et al. 2007; Dutheil et al. 2009) and species trees (Maddison and Knowles 2006; Liu and Pearl 2007), as well as many population related statistics, such as ancestral population sizes and recombination rates.

In another branch of the field, the loci of interest are those whose phylogeny is incongruent because of evolutionary events such gene duplication, loss, and horizontal transfer, and several models have been developed for each of these events. In the specific case of modeling DL, both probabilistic approaches (Arvestad et al. 2004; Gu and Zhang 2004; Hahn et al. 2005) and nonprobabilistic or parsimony-based methods have been developed (Goodman et al. 1979; Page 1994; Chen et al. 2000; Wapinski et al. 2007) to improve the reconstruction of either gene trees (Arvestad et al. 2004; Rasmussen and Kellis 2007; Wapinski et al. 2007) or species trees (Page and Charleston 1997). Our focus will be in this part of the field and specifically on the goal of the probabilistic reconstruction of gene trees in the context of a common and previously determined species tree.

For studying gene trees, Hahn et al. (2005) used the birth–death (BD) process to track changes in the number of paralogs in a gene family across a clade of species. Although it provides a way to look for significantly changing paralog copy counts, the method lacks a way of incorporating information from DNA or peptide sequences.

A method for incorporating such sequences was later developed by Wapinski et al. (2007) and was implemented in their SYNERGY gene tree reconstruction program. The method makes use of peptide sequences by combining a species-aware NJ algorithm along with an optimization for minimizing DLs while maximizing synteny (i.e., conserved gene order) between orthologs. However, this combination is ad hoc and nonprobabilistic, making it difficult to determine the best way to weigh conflicting information (Åkerborg et al. 2009). For example, in the cases where synteny information can be misleading, such as cases of gene conversions, SYNERGY shows significantly reduced reconstruction accuracy, suggesting that the primary sequence information is not sufficiently incorporated into the reconstruction (fig. 6).

A fully Bayesian model was proposed by Arvestad et al. (2004) that combined a model for gene DLs with sequence evolution. This was done by defining a prior for gene tree topologies and branch lengths using a BD process, which when combined with a sequence substitution model (e.g., JC69; Jukes and Cantor 1969) produced a Bayesian method for gene tree reconstruction and reconciliation. One disadvantage of this approach was the assumption of a clock model for substitution (i.e., constant substitution rates).

In 2007, we introduced a distance-based ML method for gene tree reconstruction that incorporates information from the species tree but avoids the clock model assumption (Rasmussen and Kellis 2007). Our model decomposes substitution rates into gene-specific and species-specific components, which was motivated by our observation of substitution rate correlations across the genomes of 12 *Drosophila* and 9 fungal species. By first learning parameters for gene- and species-specific rate distributions from genome-wide information and then using that model to reconstruct gene trees, SPIDIR showed significantly increased reconstruction accuracy compared with several other popular phylogenetic algorithms at the time. However, despite these improvements, the approach was distance based and thus did not fully utilize all the information available in sequence data.

Recently, Arvestad et al. (2004) have introduced PrIME-GSR, an extension of their previous work, which relaxes the clock assumption by using identical independent gamma distributions to model rate variation (Åkerborg et al. 2009), however, no species-specific rate variation is learned or modeled. In our evaluations (see Results), we find that modeling these rates can provide a significant benefit in gene tree reconstruction.

In summary, although much progress has been made in gene tree reconstruction, what remains missing is a principled, fast, and accurate method that incorporates all of

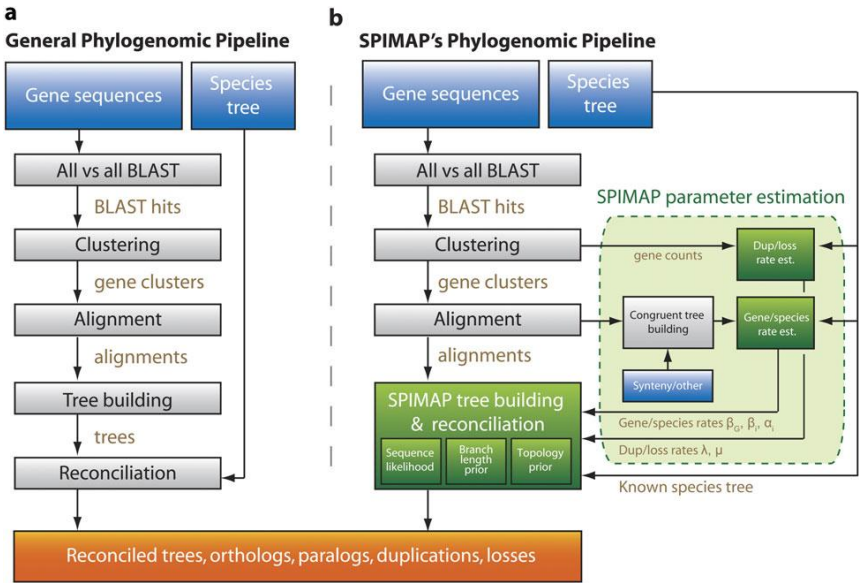


FIG. 1. Overview of the phylogenomic pipeline. (*a*) The typical phylogenomic pipeline consists of several common steps, although particular implementations may vary. The pipeline input is the set of all gene sequences across several species and the known species tree relating the species (blue boxes). Gene sequences are then compared across species and clustered according to their sequence similarity, resulting in a set of homologous gene families. A multiple sequence alignment is then constructed for each gene family, followed by phylogenetic reconstruction of each aligned family to produce gene trees. Each gene tree is then reconciled to the known species tree in order to infer orthologs, paralogs, and gene duplications and loss events, which are the pipeline outputs (orange box). (*b*) Our phylogenomic pipeline follows similar steps, except that SPIMAP includes a model parameter estimation step (dashed light green box) for DL rates (learned from the per-species gene counts in the gene families resulting from the clustering step), and gene- and species-specific substitution rates (learned from a subset of trusted orthologous alignments supported by synteny or other information and congruent to the species trees). These learned evolutionary parameters are then used in a joint tree building and reconciliation step (dark green box), specifically informing our topology prior (duplication/loss model) and our branch length prior (gene/species-specific substitution model). The joint step also enables us to use the known species tree and duplication/loss model to rapidly score topology proposals and speed up tree search in contrast to the traditional pipeline that only uses the known species tree in the reconciliation step.

[AQ3]

these various models. In addition, freely available software is needed to facilitate further analyses in this field.

Here, we present SPIMAP, a Bayesian gene tree reconstruction method that incorporates within a unified framework models for gene DL, gene- and species-specific rate variations, and sequence substitution. We model gene DL using the BD process (Arvestad et al. 2004). Similar to the other methods, we do not attempt to model incomplete lineage sorting or horizontal transfers, although approaches for doing so in the future could be useful. We have implemented a relaxed clock, defined using the rate variation model we have previously developed (Rasmussen and Kellis 2007). A key distinction of our method is that we employ an empirical Bayes approach, where the parameters of the rate model are learned using a novel Expectation–Maximization (EM) training algorithm that incorporates sequence data across many loci. Once these parameters are estimated, we use them along with the species tree to reconstruct gene trees for thousands of sequence alignments from across the genome. Our method also achieves significant speed increases by using a novel tree search strategy derived from our gene tree topology prior. Lastly, we demonstrate the feasibility and increased performance of this method on several real and simulated data sets. The SPIMAP software is written in C++ and is available for download at http://compbio.mit.edu/spimap/.

Methods

Method Overview

The reconstruction of gene trees for every gene family in several genomes typically requires a computational pipeline similar to the one shown in figure 1*a*. Databases that have followed this general outline include TreeFam (Li et al. 2006), Ensembl (Vilella et al. 2009), and many others (Huerta-Cepas et al. 2007; Datta et al. 2009), whereas other methods such as SYNERGY (Wapinski et al. 2007) perform similar tasks but not necessarily as separate consecutive steps. The general pipeline goes as follows: The input (blue boxes in fig. 1*a*) consists of nucleotide or peptide sequences for all genes in all genomes under consideration as well as a species tree estimated prior to the pipeline computation using any method or information desired. Next, the sequences are compared with each other using a method such as an all-vs-all Blast search (Altschul et al. 1990) or HMMER (Eddy 2000). The Blast hits are then clustered using a method such as OrthoMCL (Li et al. 2003) or a method like that of PHIGs (Dehal and Boore 2006) in order to form clusters of highly

[AQ4]

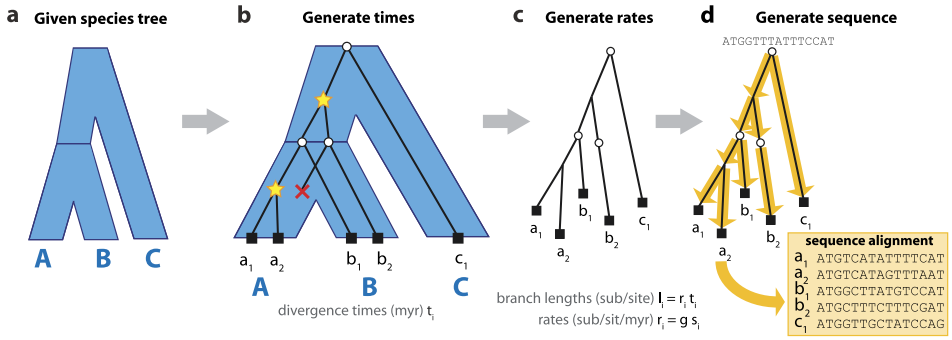


FIG. 2. SPIMAP's generative model. (*a*) First, the process begins with a given species tree S and divergence times. (*b*) Second, a gene tree T (black lines and labels) is evolved inside the species tree according to a DL model. The gene tree bifurcates either at speciation events (white circles located at species tree nodes) or at duplication events (stars located along species tree branches). Gene tree lineages can also terminate within a species tree branch at gene loss events (red "X"). (*c*) Third, substitution rates are generated according to our relaxed clock rates model of species-specific and gene-specific rates. (*d*) Lastly, sequences are evolved down the gene tree according to a continuous-time Markov process to produce a sequence alignment (yellow box) which is emitted from the process.

similar genes that are likely to represent gene families. For each cluster, a multiple sequence alignment is then constructed (e.g., MUSCLE, Edgar 2004) followed by gene tree reconstruction using a phylogenetic algorithm (e.g., PhyML, Guindon and Gascuel 2003; BIONJ, Gascuel 1997; or MrBayes, Ronquist and Huelsenbeck 2003). Lastly, a "reconciliation" algorithm is used to compare each gene tree to the species tree in order to infer all DL events as well as all ortholog and paralog relationships. Reconciliation methods include maximum parsimony reconciliation (MPR) (Page 1994; Zmasek and Eddy 2001), RAP (Dufayard et al. 2005), and Notung (Chen et al. 2000), each of which take different approaches to inferring gene DL events in presence of possibly uncertain gene trees. The duplications, losses, orthology, paralogy, and the gene trees themselves typically constitute the outputs of a phylogenomic pipeline (orange box; fig. 1).

The pipeline we have constructed for SPIMAP follows the same general structure (fig. 1*b*). For clustering, we have implemented our own method (Butler et al. 2009) similar to that of PHIGs. For multiple sequence alignment, we have used the MUSCLE (Edgar 2004) program. In contrast to other methods, however, ours takes an Empirical Bayes approach by including a "training" step (dashed green box; fig. 1*b*) that supplies several species-level evolutionary parameters to SPIMAP's gene tree reconstruction step. In the training step, we estimate the average genome-wide gene DL rates $\theta_t = (\lambda, \mu)$ based on gene counts within each gene family cluster using a method similar to that of Hahn et al. (2005) (see Estimating DL Rate Parameters). We also estimate substitution rate parameters $\theta_b = (\alpha_G, \beta_G, \boldsymbol{\alpha}, \boldsymbol{\beta})$ based on a subset of the alignments using a novel EM method (see Estimating Substitution Rate Parameters). These parameters are then used in a combined gene tree reconstruction and reconciliation step (dark green box; fig. 1*b*) performed simultaneously within a single probabilistic model. From this model, we compute the MAP gene tree using a novel rapid gene tree search that incorporates information from the species tree and from DL rates. In the following sections, we will discuss how we compute the posterior probability of a gene tree and describe the details of our rapid tree search.

Gene Tree and Species Tree Definitions

We define a "gene family" as the set of all genes descended from a single gene in the last common ancestor of all species in consideration. We represent the rooted phylogenetic tree of n genes by a tree with topology $T = (V, E)$, which describes the set of nodes (vertices) $V(T)$ and a set of branches (edges) $E(T)$ of the tree. The leaves $L(T) \subset V(T)$ of a gene tree represent observed genes from extant species, whereas the internal nodes $I(T) = V(T) \setminus L(T)$ represent ancestral genes from ancestral species. We will use several functions to discuss how nodes are related to one another. For example, we use $\text{child}(v)$ to represent the set of children of v, $\text{left}(v)$ and $\text{right}(v)$ to represent the left and right children, and $\textit{parent}(v)$ to represent its parental node. For any node v, we use $b(v)$ to denote the branch $(v, \text{parent}(v))$ and $l(v)$ to be the length of that branch, measured in substitutions per site. Lastly, we use $\boldsymbol{l}$ to denote the vector of all branch lengths of a tree, namely $\boldsymbol{l} = (l(v_1), \ldots, l(v_{2n-2}))$. Thus, a "gene tree" is represented by the tuple $(T, \boldsymbol{l})$.

In addition, we will also consider a phylogeny S relating species, called a "species tree." The branch lengths $\boldsymbol{t}$ of S are expressed in units of time (e.g., millions of years) and are thus typically ultrametric. For a node $u \in V(S)$, we express its length as time $t(u)$. We will assume all trees are rooted and all nodes to have at most two children.

Each gene tree can be viewed as evolving "inside" the species tree (fig. 2*a*). A reconciliation R is a mapping from gene nodes to species nodes that defines the species to which each extant and ancestral gene belongs (Goodman et al. 1979) (fig. 3*a*). In this setting, a gene tree is "congruent" if R is an isomorphic mapping between T and S and "incongruent" otherwise. Also, all internal nodes of a gene tree represent either "gene duplication" or speciation events (represented as stars and white circles, respectively; fig. 2*a*).

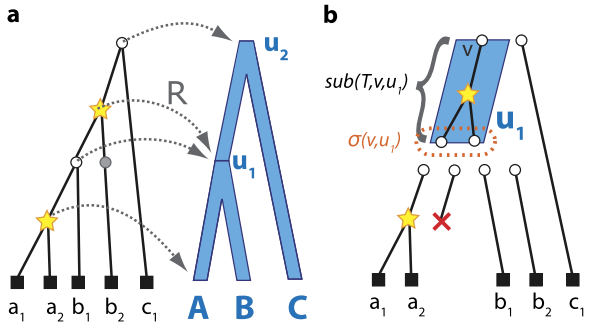


FIG. 3. Reconciliation and duplication subtrees. (*a*) A reconciliation R maps gene nodes to species nodes for both speciation events (white circles) and duplication events (stars). "Implied speciation nodes" (gray circle) are then inferred based on the reconciliation. (*b*) Our algorithm breaks the gene tree T into subtrees $\mathrm{sub}(T, v, u_1)$, where the subtree root v is a speciation and the subtree leaves $\sigma(v, u_1)$ are the next speciation nodes below v that reconcile to species u_1.

Generative Model of Gene Family Evolution

In our generative model, gene trees are generated in three steps: given a species tree with specified topology and speciation times, 1) we first generate a gene tree topology and duplication times by repeated use of a BD process, 2) we then generate substitution rates from gene and species-specific distributions, and 3) lastly, we use these rates to generate molecular sequences according to a continuous-time Markov process (fig. 2).

The parameters of our model are $\theta = (S, \boldsymbol{t}, \theta_t, \theta_b)$, where S and $\boldsymbol{t}$ are the species tree topology and branch lengths, θ_t are the topology parameters λ and μ, and θ_b are the branch length parameters α_G, β_G, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$, the details of which are given below.

Generating Topology and Divergence Times

We use the gene DL model first developed by Arvestad et al. (2003), which is based on a repeated use of the BD process (Feller 1939) to define the topologies and branch lengths (in units of time) of a gene tree evolving inside a species tree (fig. 2*a*).

The BD process is a continuous-time process that generates a binary tree according to a constant rate λ of lineage bifurcation (which will represent gene duplication) and rate μ of lineage termination (representing gene loss). After running a BD process for a time t, all lineages that exist at time t are called "surviving," whereas all others are called "extinct." A node is "doomed" if it has no surviving descendants. The BD process has been used widely in phylogenetics (Arvestad et al. 2004; Gu and Zhang 2004; Hahn et al. 2005), although typically for defining priors for species trees (Rannala and Yang 1996).

The gene DL model is defined by repeatedly using the BD process to generate a gene tree. To initialize, we begin with a single gene node v reconciled to the root of S (i.e., $R(v) = \mathrm{root}(S)$) and mark it as a speciation node. We then recursively apply the following: 1) For each speciation node v at the top of a species branch $b(u)$ of length $t(u)$, we generate a tree according to the BD process for $t(u)$ units of time. 2) For each newly created node w, we record its reconciliation as $R(w) = u$. 3) For each w that survives across that species branch, we mark it as an "extant gene" if u is a leaf species, otherwise mark it as a speciation. 4) We recursively apply steps 1–3 until all speciation nodes have been processed. 5) We mark all nodes in the gene tree not marked as extant genes or speciation as duplications. 6) As a postprocessing step, we prune all doomed lineages, namely lineages with no extant descendants.

Generating Substitution Rates

We use a relaxed clock model where substitution rates are allowed to vary between lineages (fig. 2*b*). Each branch has a length $l(v)$ (measured in substitutions/site) that is the product of a duration of time $t(v)$ and a substitution rate $r(v)$. The times are given by the DL model. The substitution rates indicate the number of substitutions per site per unit time and are described by a rates model. Previously (Rasmussen and Kellis 2007), we developed a rates model that captured the substitution rate $r(v)$ as the production of two components, a gene-specific rate and a species-specific rate. Here, we define these components with the following distributions:

(a) For each gene family j, the "gene-specific rate" g_j scales all rates in a tree. We represent the gene rate as a random variable G_j that is distributed across families as an inverse-gamma distribution with shape and scale parameters, α_G and β_G. Without loss of generality, we constrain G_j to have a mean value of one across all gene families (i.e., $\alpha_G = \beta_G + 1, \alpha_G > 1$). Thus, we have

$$P(G_j = g_j | \beta_G) = \mathrm{InvGamma}(g_j | \alpha_G = \beta_G + 1, \beta_G).$$

(b) For each branch $b(v_k)$, the "species-specific rate" s_k defines a rate specific to that branch in the gene tree. It is represented by a random variable S_k and has a gamma distribution whose scale and shape parameters (α_i, β_i) depend on the species $u_i = R(v_k)$. This allows one to model rate accelerations and decelerations that are specific to a species u_i and exists across all genes of that species. Thus,

$$P(S_k = s_k | \alpha_i, \beta_i) = \mathrm{Gamma}(s_k | \alpha_i, \beta_i), \quad \text{where } u_i = R(v_k). \quad (1)$$

We also assume that each S_k is independent of the others and of the gene rate G. Given these definitions for the substitution rate, we can then express the branch length $l(v_k)$ of a gene tree j as

$$l(v_k) = r(v_k) \times t(v_k) = g_j \times s_k \times t(v_k). \quad (2)$$

In total, our rate model has parameters $\theta_b = (\beta_G, \alpha, \beta)$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)$, and m is the number of species branches $|E(S)|$.

Generating Sequence

After generating a gene tree with a topology, divergence times, and substitution rates, we finally evolve a molecular sequence down the tree using a continuous-time Markov chain to model sequence substitution. Specifically, we have

implemented Hasegawa–Kishino–Yano (HKY; Hasegawa et al. 1985) to generating nucleotide sequences. The HKY process uses the branch lengths $l(v_k) = r(v_k)t(v_k)$ as parameters for sampling derived sequences. Only sequences on the leaves of the tree are emitted, whereas ancestral sequences are hidden (fig. 2c). In our current formulation, sequence insertion and deletion (indels) are not modeled. Instead, gaps in the sequence alignment are treated as missing data.

MAP Reconstruction of Gene Family Evolution

In our current implementation of the algorithm, we compute the MAP gene tree according to our model. Thus, we seek to calculate

$$\hat{\boldsymbol{l}}, \hat{T}, \hat{R} = \underset{\boldsymbol{l},T,R}{\operatorname{argmax}}\ P(\boldsymbol{l}, T, R \mid \boldsymbol{D}, \theta) \tag{3}$$

$$= \underset{\boldsymbol{l},T,R}{\operatorname{argmax}}\ P(\boldsymbol{D} \mid \boldsymbol{l}, T, R, \theta) P(\boldsymbol{l} \mid T, R, \theta) \times P(T, R \mid \theta)/P(\boldsymbol{D} \mid \theta) \tag{4}$$

$$= \underset{\boldsymbol{l},T,R}{\operatorname{argmax}}\ P(\boldsymbol{D} \mid \boldsymbol{l}, T) P(\boldsymbol{l} \mid T, R, \theta) P(T, R \mid \theta). \tag{5}$$

The first term in equation (5) is the likelihood of a gene tree with branch lengths $\boldsymbol{l}$ and topology T given the sequence data $\boldsymbol{D}$. The probability is defined by the sequence evolution model (e.g., HKY) and can be computed efficiently using the pruning algorithm (Felsenstein 1981), which we have implemented for SPIMAP. Because this model only depends on the topology and branch lengths of the gene tree, the likelihood term is conditionally independent of the reconciliation R and parameters θ.

The prior of our model is factored into two terms: the prior of the topology and the prior of the branch lengths. The topology prior $P(T, R \mid \theta)$ is defined based on the DL model and it can be computed efficiently (see Computing the Topology Prior). We have also found that factoring out the topology prior from the branch lengths provides a unique advantage for fast tree search (see Rapid Tree Search).

Lastly, the branch length prior $P(\boldsymbol{l} \mid T, R, \theta)$ represents the probability of observing of gene tree branch lengths $\boldsymbol{l}$. This prior incorporates both divergence times of duplications in the BD process as well as the distribution of substitution rates. We present how to compute this term numerically (see Computing the Branch Length Prior).

Computing the Topology Prior

The topology prior $P(T, R \mid \theta)$ (from eq. 5) helps SPIMAP reconstruct gene trees that have plausible patterns of gene DL. For completeness, we describe how to compute this term.

According to the DL model introduced by Arvestad et al. (2004, 2009), the BD process is repeatedly used to generate the gene tree topology T as it evolves from the root of the species tree S to the leaves. Therefore, T can be viewed as a union of several subtrees, each of which was generated by one BD process. Because these processes are independent of one another, we can view the topology prior $P(T, R \mid \theta)$ of gene tree T as a product of the probabilities of the BD process generating each of the subtrees. Performing this factoring is the key step in computing the topology prior, but, there are two additional caveats to consider: 1) how to account for lineages in the gene tree that are hidden from observation due to extinction and 2) how to account for labeled and unlabeled nodes in the gene tree. By combining these ideas, we can compute the prior of a gene tree topology.

Factoring the Gene Tree

Given a gene tree topology T, we first decompose it into the subtrees that were generated from each individual BD process (fig. 3). We call each of these subtrees "duplication subtrees" because all of their internal nodes consist of duplication nodes. To identify these subtrees, first notice that each speciation node v is the root of two such subtrees. If v has reconciliation $R(v) = w$ and $w \in V(S)$, then the two subtrees perfectly reconcile within the child species branches left(w) and right(w). Also notice that the leaves of each duplication subtree are either speciation nodes or extant genes.

Some speciation nodes (e.g., the gray node in fig. 3a) may be initially hidden in a gene tree due to gene losses. We call such nodes "implcit speciation nodes" and they can be added to a gene tree by identifying gene tree branches that span multiple branches in the species tree (e.g., branch b_2 in fig. 3a). If a given gene tree T lacks implied speciation nodes, we can add them by locating each v and $w = \text{parent}(v)$, where $\text{parent}(R(v)) \neq R(w)$. Next, the edge (v, w) is replaced by a new speciation node x and two new edges (v, x) and (x, w) while setting $R(x) = \text{parent}(R(v))$. This procedure can applied repeatedly until all implied speciation nodes are identified.

When all speciation nodes are explicit, we can identify duplication subtrees by partitioning the gene tree at all speciation nodes $\text{spec}(T)$ (fig. 3). We denote a particular subtree as $\text{sub}(T, v, u)$, where $v \in \text{spec}(T)$ is the root of the subtree and $u \in \text{child}(R(v))$ is the species to which the leaves $L(\text{sub}(T, v, u))$ reconcile. The leaves are defined by the set

$$\sigma(v, u) = \{w : w \in \text{spec}(T) \cup L(T), R(w) = u, w \in V(T_v)\}, \tag{6}$$

where T_v is a subtree of T containing node v and all of its descendants.

For each duplication subtree, we can derive its probability from the BD process (Rannala and Yang 1996). First, for a BD process with a birth rate λ and death rate μ, the probability that one lineage will leave s survivors after time t is

$$p(s, t) = (\lambda/\mu)^s p(1, t)(p(0, t))^{s-1}, \tag{7}$$

where

$$p(0, t) = \frac{\mu(1 - e^{-(\lambda-\mu)t})}{\lambda - \mu e^{-(\lambda-\mu)t}}$$

$$p(1, t) = \frac{(\lambda - \mu)^2 e^{-(\lambda-\mu)t}}{(\lambda - \mu e^{-(\lambda-\mu)t})^2}. \tag{8}$$

Second, for s survivors there are $\xi_s = s!(s-1)!/2^{s-1}$ equally likely "labled histories," which are leaf labeled topologies whose internal nodes are order by their time. Thus, for a topology T with s leaves and $H(T)$ labeled histories, its probability is

$$P(T|t,\lambda,\mu) = \frac{H(T)}{\xi_s} p(s,t), \quad \text{where} \tag{9}$$

$$H(T) = \prod_{v \in I(T)} \binom{|I(T_{\text{right}(v)})| + |I(T_{\text{left}(v)})|}{|I(T_{\text{right}(v)})|}. \tag{10}$$

Doomed Lineages

In addition to factoring the tree, there are two caveats to consider. The first to consider is the possibility of lineages in the gene tree that are hidden from observation because they have gone extinct, that is, they leave no descendants in the leaves of the species tree. We call such lineages "doomed," and this extinction process must be accounted for in our topology prior.

Let $d(u)$ be the probability that a lineage starting at node u in the species tree will be doomed, that is, losses occur such that no descendants exist at the leaves of the species tree. This probability $d(u)$ is the product of the probability of extinction occurring in both the left and the right subtrees beneath node u. For a child branch $b(c)$, where $\text{parent}(c) = u$, we must consider two possibilities. Either the gene lineage goes extinct in $b(c)$ with probability $p(0, t(c))$ (eq. 8) or it survives and leaves i survivors, each of which themselves are doomed with probability $d(c)$. Thus, this probability can be expressed recursively as

$$d(u) = \begin{cases} \prod_{c \in \text{child}(u)} \sum_{i=0}^{\infty} p(i, t(c)) d(c)^i & \text{if } u \in I(S), \\ 0 & \text{if } u \in L(S). \end{cases} \tag{11}$$

The value $d(u)$ can be computed efficiently for each node u in the species tree S by dynamic programing following a postorder traversal of S.

Labeled and Unlabeled Nodes

The second caveat of the topology prior computation is distinguishing between labeled and unlabeled nodes within the gene tree. In equation (9), we give the probability of a BD process generating a labeled topology T. Each duplication subtree $\text{sub}(T, v, u)$ is generated by one BD process, however, only duplication subtrees with extant leaves (i.e., $L(\text{sub}(T, v, u)) \subseteq L(T)$) are labeled topologies. All other duplication subtrees have leaves that are speciation nodes and thus are unlabeled topologies.

To properly account for labeled and unlabeled nodes, we envision the DL model as a three step process. First, a gene tree T' is generated by repeated use of the BD process after which as all extant and speciation nodes are labeled. The probability of this tree is $P(T', R|\theta)$ and it can be computed by factoring T' into duplication subtrees, each of which has a known probability (eq. 9).

Second, a mapping U is applied to T' that removes all labels to produce an unlabeled gene tree T''. The probability $P(T'', R|\theta)$ is thus the sum of the probability of each T' that becomes T'' after removing labels,

$$P(T'', R|\theta) = \sum_{\{T' : T'' = U(T')\}} P(T', R|\theta). \tag{12}$$

We call two trees T_i' and T_j' equivalently labeled if $U(T_i') = U(T_j')$. Because equivalently labeled trees T_i' all have equal probability, the probability $P(T'', R|\theta)$ is simply the probability of T' times the number of equivalent labelings. The number of equivalent labelings is computed as a product of correction terms, one for each duplication subtree. Specifically, for each internal subtree T_2 (i.e., leaves are speciations nodes), we multiply by the term $N_2(T, T_2, R)$ and for each external subtree T_2 (i.e., leaves are extant genes), we multiply by $N_1(T_2, R)$. See supplementary section 2.2, Supplementary Material online for the definition of these terms. [AQ5]

In the third and final step, labels are added back to the leaves of T'' to create our desired leaf labeled gene tree topology T. Because each labeling is equally likely to be generated by this process, the probability $P(T, R|\theta)$ is $P(T'', R|\theta)$ divided by the number of ways to relabel T''. This final correction factor is $1/N_1(T, R)$ and is derived in supplementary section 2.2, Supplementary Material online.

The Full Topology Prior

Combining these ideas, we can compute the probability of a gene tree T being generated by the DL model as

$$P(T, R|S, \boldsymbol{t}, \lambda, \mu) = \frac{1}{N_1(T, R)} \times \prod_{v \in \text{spec}(T)} \prod_{u \in \text{child}(R(v))} g(v, u, \text{sub}(T, v, u)), \tag{13a}$$

$$g(v, u, T_2) = f(T, T_2, R) \sum_{i=0}^{\infty} \binom{|L(T_2)| + i}{i} \times p(|L(T_2)| + i, t(u)) d(u)^i, \tag{13b}$$

$$f(T, T_2, R) = \begin{cases} N_2(T, T_2, R) H(T_2) / \xi_{|L(T_2)|} & \text{if } L(T_2) \subseteq I(S), \\ N_1(T_2, R) H(T_2) / \xi_{|L(T_2)|} & \text{if } L(T_2) \subseteq L(S). \end{cases} \tag{13c}$$

The sum in equation (13) is a sum over how many doomed lineages i might have been present at node u. Within the sum, we find the probability that a BD process generates the survivors $L(T_2)$ that are present plus i hidden doomed lineages. The term $d(u)^i$ is the probability that those i lineages go extinct. The permutation term describes the number of ways to choose i doomed lineages from the total number of survivors $i + |L(T_2)|$.

Although this calculation involves an infinite sum, it can be computed analytically and the total computation of the topology prior takes at most $O(|V(T)||V(S)|)$ run time

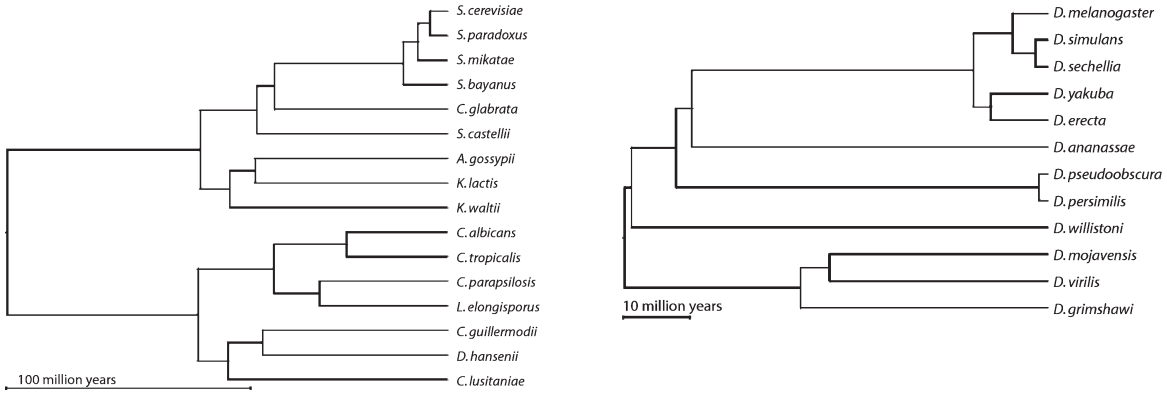


FIG. 4. Species and phylogenies used in evaluation. (*a*) Phylogeny of 16 fungal species used for the reconstruction pipelines of real and simulated evaluation data sets. The phylogeny was estimated in Butler et al. (2009) with divergence times estimated by the r8s program (Sanderson 2003) assuming 180 My (Massey et al. 2003) for the divergence depth. (*b*) The phylogeny of 12 *Drosophila* species used in our simulation evaluation. Phylogeny was estimated by Tamura et al. (2004).

(Arvestad et al. 2009). Currently, we only consider reconciliations R that are maximally parsimonious for DLs. This approximation is likely reasonable, as we find that the true reconciliation is the most parsimonious one in 98% of gene trees simulated using our species tree (fig. 4) and independently estimated DL rates (Hahn et al. 2005), agreeing with results from similar studies (Doyon et al. 2009).

Computing the Branch Length Prior

The final term in our model is the branch length prior $P(\boldsymbol{l}|T,R,\theta)$, which is the prior probability of the branch lengths $\boldsymbol{l}$ given the topology T, reconciliation R, and model parameters θ. This term helps SPIMAP choose gene trees that have branch lengths that are more reasonable given the time span implied by the reconciliation and our prior knowledge of the substitution rates.

We will explain the calculation of this term in a top-down fashion, breaking it into smaller parts until each part is defined. We begin by viewing the branch prior as a marginal over the gene rate g of the family in consideration

$$P(\boldsymbol{l}|T,R,\theta) = \int P(\boldsymbol{l}|g,T,R,\theta)P(g|\alpha_G,\beta_G)\mathrm{d}g. \quad (14)$$

Once conditioned on the gene rate g, many of the branch lengths of T become independent because we know their common scale factor g. However, those branches that surround a duplication node are still nonindependent because their lengths depend on the time of the duplication, which is unknown. However, if we partition T into a set of subtrees $\mathbb{T}$ by segmenting at each speciation node $v \in \mathrm{spec}(T)$ (without adding implied speciation nodes), each subtree $\tau \in \mathbb{T}$ will contain branch lengths that are independent of the other subtrees. In particular, each subtree τ is rooted by a speciation node, its leaves are either extant or speciation nodes, and all other internal nodes are duplication nodes. We refer to branch lengths for each subtree τ as $\boldsymbol{l}^\tau$, its divergence times as $\boldsymbol{t}^\tau$, and its substitution rates as $\boldsymbol{r}^\tau$. Thus, $\boldsymbol{l}^\tau = (l(w_1), l(w_2), \ldots, l(w_k))$ and $\boldsymbol{t}^\tau = (t(w_1), \ldots, t(w_k))$, where $w_1, w_2, \ldots, w_k$ are the nonroot nodes of subtree τ. Using this notation, we can continue to factor,

$$P(\boldsymbol{l}|g,T,R,\theta) = \prod_{\tau\in\mathbb{T}} P(\boldsymbol{l}^\tau|g,T,R,\theta). \quad (15)$$

The branch lengths within $\boldsymbol{l}^\tau$ are nonindependent because they depend on the duplication times. However, if we condition on the branch times $\boldsymbol{t}^\tau$, each branch length l_i^τ becomes a simple function of the branch rate r_i^τ because $l_i^\tau = t_i^\tau r_i^\tau$. Because we model all branch rates as being independent of one another, we can then finally factor the branch prior as a product of the probability of each branch length l_i^τ,

$$P(\boldsymbol{l}^\tau|g,T,R,\theta) = \int P(\boldsymbol{l}^\tau|\boldsymbol{t}^\tau,g,T,R,\theta)P(\boldsymbol{t}^\tau|g,T,R,\theta)\mathrm{d}\boldsymbol{t}^\tau, \quad (16)$$

$$\text{where } P(\boldsymbol{l}^\tau|\boldsymbol{t}^\tau,g,T,R,\theta) = \prod_i P(l_i^\tau|t_i^\tau,g,T,R,\theta), \quad (17)$$

and where $P(\boldsymbol{t}^\tau|g,T,R,\theta)$ describes the distribution of branch times in subtree τ which is defined by the BD process. We have integrated over the branch times $\boldsymbol{t}^\tau$ because they are unknown.

The last term to define is the distribution of a single branch length $l(v_i)$. In the simplest case (see the next section for a caveat), the distribution can be derived as follows:

$$\begin{aligned} l(v_i) &= g \times t(v_i) \times s(v_i) \sim g \times t(v_i) \\ &\quad \times \text{Gamma}\left(\alpha_{R(v_i)}, \beta_{R(v_i)}\right) \\ &= \text{Gamma}\left(\alpha_{R(v_i)}, \frac{\beta_{R(v_i)}}{g \times t(v_i)}\right), \end{aligned} \quad (18)$$

where, $s(v_k)$ is the species-specific rate for branch $b(v_k)$. In our implementation of computing the branch prior, we integrate over gene rates g (eq. 14) by approximating with a summation with equally probable gene rates. Also, the integral over times $\boldsymbol{t}^\tau$ (eq. 17) is performed with Monte Carlo by

sampling from $P(\boldsymbol{t}^{\tau}|g,T,R,\theta)$. The run time of this calculation is implemented to be linear with the size of the gene tree.

Handling Implied Speciation Nodes

One complexity not considered in equation (18) is the effect of implied speciation nodes. In such a case, we can have a branch length $l(v_i)$ that spans multiple species branches. For example, the branch b_2 in figure 3*a* spans the species B and u_1. Also note that the length of branch b_2 is the sum of two smaller branches: one within species branch B and one within species branch u_1. Thus, to complete our description of the branch prior, we must define the probability $P(l(v_i)|t(v_i),g,T,R,\theta)$ for branches that span multiple species.

To handle these cases, we introduce a topology T' that is defined as the topology T with implied speciation nodes added. Also let $\boldsymbol{l}'$ and $\boldsymbol{t}'$ be the length and time vectors of T', and R' be a reconciliation of T' to S. For each branch $b(v_i) = (v_i, w_i)$ in T, where w_i is the parent of v_i in T, there is a path $p = (v_i, \ldots, w_i)$ in T'. Let $p(v_i)$ be the set of all vertices in p excluding the top node w_i. Thus, the branch lengths and times in tree T can be expressed as sums of branch lengths and times in tree T',

$$l(v_i) = \sum_{v'_k \in p(v_i)} l(v'_k) \quad \text{and} \quad t(v_i) = \sum_{v'_k \in p(v_i)} t(v'_k). \tag{19}$$

The distribution of each $l(v'_k)$ is the same as the distribution given in equation (18) using R' as the reconciliation. To define the probability $P(l(v_i)|t(v_i),g,T,R,\theta)$, we note that $l(v_i)$ is simply the sum of independent gamma random variables, and methods exist to compute this probability efficiently (Moschopoulos 1985).

Branches Near the Root

If a gene branch contains the root, then it is still distributed by a sum of gamma distributions and thus can use the same methods developed here. For nodes that reconcile before the species tree root, we still treat them as being generated by a BD process in the basal branch of the species tree. We model the length T_0 of the basal branch as expontentially distributed with mean λ_0 and model the species-specific substitution rate as a gamma-distributioned random variable with mean and variance that is the average of the other species-specific rate distributions.

Rapid Tree Search

To compute the argmax in equation (5), we search over the space of possible gene tree topologies T, branch lengths $\boldsymbol{l}$, and reconciliations R using a hill climbing approach to find the MAP reconciled gene tree $(\hat{T}, \hat{\boldsymbol{l}}, \hat{R})$. We begin our search with an initial tree constructed using the NJ algorithm (Saitou and Nei 1987). We use subtree pruning and regrafting to propose additional topologies T. For each T, branch lengths $\boldsymbol{l}$ are proposed using numerical optimization (Newton–Raphson) of the likelihood term $P(\boldsymbol{D}|\boldsymbol{l},T)$.

One unique feature of our search is that we use the gene tree topology prior $P(T,R|\theta)$, a relatively fast computation compared with computing $P(\boldsymbol{D}|\boldsymbol{l},T)$ by 2–3 orders of magnitude to prescreen topology proposals for those that are likely to have high posterior probability. Given the best topology T thus far, we make $N \in [100, 1,000]$ unique rearrangements T_i and compute their topology prior $k_i = P(T_i, R_i|\theta)$, where R_i is the MPR. As our next proposal, we then choose a topology T_i from $T_1, \ldots, T_N$ with probability $p_i = \frac{c}{N} + \frac{(1-c)k_i}{\sum_i k_i}$, where parameter $c \in (0,1)$ defines a mixing between the weights k_i and the uniform distribution. In practice, we use $c = 0.2$.

We have found that this simple adjustment to our search strategy greatly increases the speed of finding the MAP gene tree (See Results and table 2).

Estimating Substitution Rate Parameters

As discussed previously, our substitution rate model is able to describe rate variation that occurs in both gene- and species-specific ways. In order to achieve this, it requires the estimation of several parameters $\theta_b = (\alpha_G, \beta_G, \boldsymbol{\alpha}, \boldsymbol{\beta})$. One unique approach in our method is that we estimate these parameters prior to reconstruction by analyzing substitution rates from multiple loci with known phylogenetic trees. This constitutes a "training step" in an empirical Bayes approach. Figure 1*b* illustrates how this estimation fits within the larger phylogenomic pipeline.

Currently for our training data set, we use trees of one-to-one orthologous gene alignments (e.g., syntenic orthologs or unambiguous best reciprocal Blast hits) where we can be reasonably confident that the gene tree topology is congruent to the species tree. Fixing the gene tree topology, we estimate the ML branch lengths for N trees with $M = |E(S)|$ branches each in order to construct a matrix $\boldsymbol{L}$ of branch lengths, such that l_{ij} is the length of the jth branch in the ith tree. We then use the $\boldsymbol{L}$ matrix along with a species tree S and its branch lengths $\boldsymbol{t}$ to estimate the parameters θ_b. Because the gene rates $\boldsymbol{g}$ of these trees are not known, we treat them as hidden data and use an EM algorithm to estimate our parameters.

The variables of the substitution rate training model are as follows. A gene tree will have a "gene rate" g, a vector of "species rates" $\boldsymbol{s}$ (measured in substitutions/site/unit time), and a vector of "branch lengths" $\boldsymbol{l}$ (measured in substitutions/site). Thus, for a single gene tree, we have the following variables:

$$g, \boldsymbol{l} = [l_1, \ldots, l_M]^T, \quad \boldsymbol{s} = [s_1, \ldots, s_M]^T,$$
$$\boldsymbol{t} = [t_1, \ldots, t_M]^T, \quad \text{with } l_i = g s_i t_i. \tag{20}$$

For a set of N gene trees indexed by j, we can describe them using the variables

$$\boldsymbol{g} = [g_1, \ldots, g_N]^T, \quad \boldsymbol{L} = [\boldsymbol{l}_1, \ldots, \boldsymbol{l}_N],$$
$$\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N], \quad \text{with } l_{ij} = g_j s_{ij} t_i. \tag{21}$$

We have designed this method to assume that $\boldsymbol{L}$ is directly observed and is given as input along with the divergence times $\boldsymbol{t}$. In contrast, the gene rates $\boldsymbol{g}$ and species rates $\boldsymbol{S}$ are not directly observed and have to be inferred from the model.

As for the distribution of these variables, recall that g_j are independent and identically distributed (i.i.d.) by the inverse gamma InvGamma(α_G, β_G) and that s_{ij} are independently distributed by Gamma(α_i, β_i). Thus, the distribution of the branch length matrix $\boldsymbol{L}$ is

$$P(\boldsymbol{L}|\boldsymbol{t}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \alpha_G, \beta_G) = \prod_j P(\boldsymbol{l}_j|\boldsymbol{t}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \alpha_G, \beta_G) \quad (22)$$

$$= \prod_j \int_0^\infty P(g_j|\alpha_G, \beta_G) P(\boldsymbol{l}_j|g_j, \boldsymbol{t}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \mathrm{d}g_j \quad (23)$$

$$= \prod_j \int_0^\infty \mathrm{InvGamma}(g_j|\alpha_G, \beta_G) \times \prod_i \mathrm{Gamma}\left(l_{ij}|\alpha_i, \frac{\beta_i}{g_j t_i}\right) \mathrm{d}g_j. \quad (24)$$

In our EM algorithm, the branch length matrix $\boldsymbol{L}$ is the observed data and the gene rate vector $\boldsymbol{g}$ is the hidden data. In EM, the goal is to iteratively find better estimates of θ_b by maximizing this function

$$\theta_b{}^{k+1} = \underset{\theta_b}{\mathrm{argmax}} \sum_j \int P(g_j|\boldsymbol{l}_j, \theta_b{}^k) \log P(\boldsymbol{l}_j, g_j|\theta_b) \mathrm{d}g_j. \quad (25)$$

For the derivation of equation (25), see supplementary section 2.3, Supplementary Material online. Conditioning on the hidden data allows us to find the next estimates of the gene- and species-specific rate parameters separately,

$$\beta_G^{k+1} = \underset{\beta_G}{\mathrm{argmax}} \sum_j \left[\int P(g_j|\boldsymbol{l}_j, \theta_b{}^k) \log \mathrm{InvGamma}(g_j|\beta_G) \mathrm{d}g_j \right] \quad (26)$$

$$\alpha_i^{k+1}, \beta_i^{k+1} = \underset{\alpha_i, \beta_i}{\mathrm{argmax}} \sum_j \left[\int P(g_j|\boldsymbol{l}_j, \theta_b{}^k) \log \mathrm{Gamma}\left(l_{ij}|\alpha_i, \frac{\beta_i}{g_j t_i}\right) \mathrm{d}g_j \right]. \quad (27)$$

[AQ6]

These expressions are computed using the Brent root finding algorithm for β_G and BFGS for α_i, β_i as implemented in the GNU scientific library. The gradients of these expressions are given in supplementary section 2.3, Supplementary Material online.

Computing the term $P(g_j|\boldsymbol{l}_j, \theta_b{}^k)$ (i.e., the probability of hidden data) constitutes the E-step. By exploiting conjugate priors (see supplementary section 2.3, Supplementary Material online), we have

$$P(g_j|\boldsymbol{l}_j, \boldsymbol{t}, \theta_b) = \mathrm{InvGamma}\left(g_j|\alpha_G + \sum_i \alpha_i, \beta_G + \sum_i \frac{\beta_i l_{ij}}{t_i}\right). \quad (28)$$

We have currently implemented the EM algorithm such that $P(g_j|\boldsymbol{l}_j, \boldsymbol{t}, \theta_b)$ is discretized. Thus, the integrals in the argmax expressions (26) and (27) are approximated as sums. See supplementary figure S8, Supplementary Material online for an example of parameters learned from data sets of 12 flies and 16 fungi species.

Estimating DL Rate Parameters

We have also implemented a training procedure for estimating the genome-wide average duplication rate λ and loss rate μ. We use the algorithm of Hahn et al. (2005), which uses the gene counts in each gene family cluster (fig. 1*b*) to estimate λ and μ. However, unlike Hahn *et al.*, we do not require λ and μ to be equal. Examples of parameters estimated from data are given in supplementary table S1, Supplementary Material online.

Results

Phylogenomic Data Sets

To evaluate our approach for gene tree reconstruction, we have reconstructed gene trees for both real and simulated data sets. For our real data set, we have used 16 fungi species (fig. 4*a*) whose genomes have been sequenced to either draft or high coverage quality (Goffeau et al. 1996; Cliften et al. 2003; Kellis et al. 2003; Dietrich et al. 2004; Dujon et al. 2004; Jones et al. 2004; Kellis et al. 2004; Butler et al. 2009). For our simulated data sets, we simulated gene alignments that share many properties of real gene trees, by using a model with parameters estimated from real data sets. Thus, we have simulated gene trees that capture the properties of the 16 fungal genomes as well as 12 fully sequenced *Drosophila* genomes (Adams et al. 2000; Richards et al. 2005; Clark et al. 2007) (fig. 4*b*). By using both clades, we can evaluate the performance of phylogenetic methods across a variety of species tree topologies, divergence times, and gene DL rates.

For the species trees, we obtained the topologies and divergence times from several data sources. For the 16 fungi, we used the species phylogeny as constructed in Butler et al. (2009) and estimated time divergence using the r8s program (Sanderson 2003) with an estimate of 180 My (Massey et al. 2003) for the clade depth (fig. 4*a*). For the 12 flies, we used the same topology and divergence times as used in several recent studies (Tamura et al. 2004; Hahn et al. 2007) (fig. 4*b*).

Training SPIMAP's Model Parameters

To run SPIMAP in our evaluations, we applied our training algorithms to estimate the parameters of our gene family model. These parameters were also used to generate the simulated data sets. Here, we describe how we prepared the input data for our training procedure for both the 16 fungi and 12 *Drosophila* data sets. Our training procedure contains two methods: one to estimate our substitution rate parameters $\theta_b = (\beta_G, \boldsymbol{\alpha}, \boldsymbol{\beta})$ and one to estimate our DL rates $\theta_t = (\lambda, \mu)$.

The first method (see Estimating Substitution Rate Parameters) estimates our substitution rate parameters from a

data set of one-to-one orthologous gene trees that are congruent to the species tree. To obtain such trees, we identified families that are highly likely to be one-to-one orthologous (i.e., one gene from each species in the clade). For the 16 fungi, we previously identified 739 confident one-to-one orthologous families (Butler et al. 2009). This was done by identifying synteny blocks containing at least three consecutive genes and spanning across the *Saccharomyces* or *Candidae* clades. Pairs of syntenic clusters with best reciprocal Blast hits spanning across the clades were merged, resulting in 739 families. For the 12 flies clade, we previously identified 5,154 one-to-one families where genes belong to a synteny block spanning all 12 species and contains at least three consecutive genes along each chromosome (Rasmussen and Kellis 2007). Next, for each one-to-one family, we made peptide multiple alignments using MUSCLE (Edgar 2004). Coding sequences were mapped onto the alignments to produce codon-aligned nucleotide alignments, substituting every amino acid with the corresponding codon and every gap with a triplet of gaps. PhyML v2.4.4 (Guindon and Gascuel 2003) was run on each nucleotide alignment using the HKY + Γ + I model and a fixed topology (congruent with the species tree), resulting in estimates for the branch lengths of each gene tree. Lastly, these branch lengths $\boldsymbol{L}$ were used in our EM method to estimate the model parameters (see supplementary figs. S8 and S9, Supplementary Material online).

The second method (see Estimating DL rate parameters) estimates our DL parameters from gene counts present within gene family clusters that contain DLs. For the 16 fungi, we used gene counts from gene families previously clustered (Butler et al. 2009) to estimate the gene DL rates $\lambda = 0.000732$, $\mu = 0.000859$ (events/gene/My). For the 12 *Drosophila* clade, we used the DL rates $\lambda = 0.0012$, $\mu = 0.0012$ that were previously estimated by Hahn et al. (2007) using a similar method as our fungi rate estimation, except that DL rates were assumed to be equal.

Reconstructing Gene Families from 16 Fungi

In our first evaluation, we analyzed the performance of SPIMAP versus several other popular phylogenetic programs on a data set of 16 fungi species. We have included three traditional "sequence-only" methods: PhyML v2.4.4 (Guindon and Gascuel 2003; ML), RAxML 7.0.4 (Stamatakis et al. 2005; ML), BIONJ (Gascuel 1997; NJ), and MrBayes v3.1.1 (Ronquist and Huelsenbeck 2003; Bayesian). We have also evaluated several other methods that use species-related information, which we call "species tree aware." These include our previous method SPIDIR (Rasmussen and Kellis 2007), SYNERGY (Wapinski et al. 2007), and PrIME-GSR (Åkerborg et al. 2009).

For our 16 fungi real data set, we downloaded coding sequences and peptides from the January 2009 update of fungi data set used by the SYNERGY method (Wapinski et al. 2007; Wapinski et al. 2009). By using this data as the input for all the other methods, we can compare against the trees constructed by SYNERGY (also downloaded from the January 2009 update). We focused the analysis on the same 16 species as used in Butler et al. (2009), which is a tree that also agrees with the one used by SYNERGY. We used the same gene clusters as defined by SYNERGY's trees, in effect using SYNERGY as the clustering step for the phylogenomic pipeline (fig 1*a*). Peptide alignments were made using MUSCLE (Edgar 2004), and coding sequences were mapped onto them to produce nucleotide alignments. In addition, from the nucleotide alignments, we also produced RY-encoded alignments, which only indicate whether a base is purine (R) or pyrimidine (Y). No other information from SYNERGY trees was made available to the other methods.

We used the following parameters for each of the methods. For PhyML and BIONJ, we used a HKY + Γ + I model of nucleotide substitution, whereas for RAxML, we used the GTRCAT model. We configured MrBayes with four chains, an automatic stop rule, a 25% burn-in, sampled every ten generations from a total of 10,000 generations, a 4×4 model for nucleotides, and enforced a binary tree. For methods that do not produce reconciled trees (i.e., PhyML, RAxML, MrBayes, and BIONJ), we have used MPR to infer DLs. For SPIDIR, we used DL penalties of 0.001 and an error cost of -600. For PrIME-GSR, we used 50,000 iterations, the Jones, Taylor, and Thorton model, gamma-distributed rates, and our own species tree (fig. 4). The tree search was initialized by an ML tree found by PhyML. We also ran PrIME-GSR with 1,000,000 iterations (as recommended by Åkerborg et al. 2009) but for only 500 trees randomly chosen from the data set in order to limit the computational run time. SPIMAP was executed with two settings: "long" (2,000 iterations with 1,000 prescreening iterations) and "short" (100 iterations with 1,000 prescreening iterations). For all other programs and options, defaults were used.

Although, a ground truth is not known for real data sets, we have used several informative metrics to assess the quality of gene trees, gene duplications, and losses inferred by these methods. Each of these metrics also illustrate different advantages and shortcomings of each method.

Recovering Syntenic Orthologs

The first metric we investigated was the ability to infer syntenic orthologs—pairs of genes that are highly likely to be orthologous given their surrounding conserved gene order. Although not all orthologous pairs are syntenic, synteny information does allow us to identify a conservative set of orthologous genes using a method independent of phylogenetics and thus provides a useful gold standard to test against. See supplementary section 2.1, Supplementary Material online for a description of our synteny determination method. When we construct trees on families that contain such genes, we expect a syntenic gene pair to appear within the reconstructed gene tree such that their most recent common ancestor is a speciation and thus are inferred as orthologs.

SPIMAP recovered syntenic orthologs with 96.5% sensitivity, followed by PrIME-GSR at 88.9% and PhyML at 64.1% (table 1). Because SYNERGY uses synteny as one of its inputs, this test alone cannot assess its accuracy, and indeed 99.2%

Table 1. Evaluation of Several Phylogenetic Programs on Gene Trees from 16 Fungi.

Program	Orthologs[a] (%)	Number of Orthologs[b]	Number of Dup[b]	Number of Loss[b]	Average run time[c]
SPIMAP (quick)[d]	96.2	550,800	5,541	10,884	1.0 min
SPIMAP (long)[d]	96.5	557,981	5,407	10,384	21.9 min
SPIMAP (i.i.d.)[e]	93.9	547,976	6,201	13,428	21.6 min
SPIDIR (quick)[d]	83.3	524,292	10,177	33,550	2.2 min
SYNERGY	99.2	595,289	4,604	8,179	—[f]
PrIME-GSR	88.9	527,153	7,951	21,099	53.1 min
PrIME-GSR (long)[d]	90.7	—	—	—	20.7 h
RAxML	63.8	463,020	21,485	65,392	18.4 s
MrBayes	63.9	460,510	21,307	65,238	43.2 s
PhyML	64.2	464,479	21,264	64,391	45.3 s
BIONJ	60.4	439,193	22,396	71,231	0.5 s

[a]Percentage of syntenic orthologs recovered.
[b]Number of pairwise orthologs, duplications, and losses inferred from trees.
[c]Average run time for reconstructing each gene tree.
[d]Both SPIMAP and PrIME-GSR were run with a few iterations (quick) of 100 and 50,000 and with many iterations (long) 2,000 and 1,000,000.
[e]SPIMAP was also run using a i.i.d. species-specific rate model.
[f]Because SYNERGY trees were downloaded, no run time was estimated.

of syntenic genes are orthologs in SYNERGY's trees. When given more iterations, PrIME-GSR's accuracy increases to 90.7% but computational time increases dramatically, 24-fold from 53 min to 20 h per gene tree. In contrast, SPIMAP achieved its accuracy of 96.5% in 29.1 min on average per tree and can achieve as much as 96.2% accuracy even when limited to an average run time of 1.0 min ("quick" mode). Also, SPIMAP achieves 96.3% ortholog accuracy when assessing the same 500 tree subset as PrIME-GSR's long mode. Note that the species tree aware programs (SPIMAP, SYNERGY, and PrIME-GSR) predict as much as 20% more ortholog pairs than the leading competing sequence-only program (PhyML).

For SPIMAP, performance was greater on RY-encoded alignments (96.5%) versus the full nucleotide alignments (92%, data not shown). This is likely due to that fact that the nucleotide alignments contained a gas chromatography (GC) bias that varies across species (supplementary table S1, Supplementary Material online), thus violating the stationarity assumption made in our implemented sequence evolution model (HKY). Reconstruction accuracy of PhyML and MrBayes was slightly diminished on RY-encoded alignments (63.0% and 61.1%, respectively), most likely due to the their lower information content. We also found that PrIME-GSR performs best on peptide data (88.9%), and that syntenic ortholog recovery decreased to 86.2% on nucleotide alignments using HKY (parameters estimated using PHYML) and 81.2% on RY-encoded alignments using JC69.

One important distinction between SPIMAP and PrIME-GSR is that SPIMAP models species-specific rate variation. To investigate the effect of this difference, we configured SPIMAP to learn an i.i.d. rates model similar to PrIME-GSR. For each branch, our modified training step estimated ($\alpha_i = 2.819, \beta_i = 663.0$) as the parameters for the i.i.d. gamma distributions. Reconstructing gene trees using these parameters, we found fewer syntenic orthologs (93.9%) and greater numbers of DLs.

Counting DL Events

Second, we evaluated the total numbers of DLs inferred across the clade (supplementary fig. S1, Supplementary Material online). Both SPIMAP and SYNERGY inferred at least 32% fewer duplications and 50% fewer losses than PrIME-GSR and the three sequence-only methods. The sequence-only methods, which do not use the species tree, infer many more events on nearly every branch, especially for short interior branches. The distribution of DL events that occur within each gene tree is illustrated in supplementary figure S2, Supplementary Material online. Interestingly, each of the other sequence-only methods inferred over four times as many gene duplication events and six times as many gene loss events as SPIMAP. For the sequence-only methods, duplications are more frequent near the root of the species tree and losses are more frequent near the leaves, a pattern suggesting that these events are erroneous (Hahn 2007).

Duplication Consistency Score

With our third metric, we sought to characterize the plausibility of the inferred duplications using the "duplication consistency score," introduced by Ensembl for evaluating their phylogenomic pipeline (Vilella et al. 2009). The consistency of a duplication node with children l and r is defined as $|A \cap B|/|A \cup B|$, where A and B are the set of species represented in descendants of l and r, respectively (see example in fig. 5*a*). The consistency score is designed to detect duplications that are wrongly inferred due to phylogenetic reconstruction errors because such false duplications are often followed by many compensating losses (Hahn 2007; Vilella et al. 2009) (i.e., low species overlap $|A \cap B|$). Figure 5 depicts the distribution for the duplication consistency score for each program. Both SPIMAP and SYNERGY showed similar consistency distributions that are heavily shifted toward 1 (47.8–49.0% and 4.2–17.2% of duplications with a score of 1 and 0, respectively; fig. 5). The sequence-only methods have many low scoring duplications (<11% and >70% with scores 1 and 0, respectively), an effect seen previously (Vilella et al. 2009). PrIME-GSR's distribution lies in between these extremes with 30.0% and 42.1% for scores 1 and 0, respectively. Lastly, the i.i.d. version

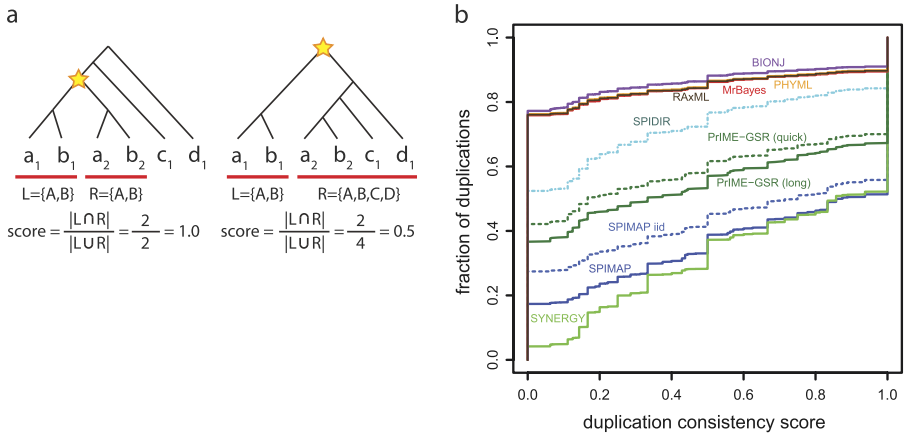


FIG. 5. The duplication consistency score for assessing phylogenetic methods. (*a*) Duplication consistency score computed on two example trees. For each duplication node (star), this score computes the number of species present in both the left and right subtrees divided by the total number of species descendant from the duplication node. Erroneous duplications show an increased rate of compensating losses and thus lower scores. (*b*) Cumulative distribution of duplication consistency scores for all duplications inferred in the 16 fungi data set by each method. SPIMAP (blue) and SYNERGY (green) perform best according to this metric, having the fewest duplications with low consistency scores. SPIMAP trained with an i.i.d. model similar to PrIME-GSR (dashed blue) infers duplications with overall lower consistency scores. These are followed by PrIME-GSR (dark green) and SPIDIR (dashed light blue) that show more moderate performance. Lastly, the four traditional methods implemented in the programs MrBayes, RAxML, PHYML, and BIONJ, all have similar and significantly lower score distributions.

of SPIMAP also scored lower than SPIMAP, inferring nearly twice the number of duplications with a consistency score of zero (fig. 5).

Recovering Gene Conversions

The fourth metric was specifically designed to test the case where species-level information is misleading, effectively testing the ability of species-aware methods to properly weigh species information against conflicting sequence information.

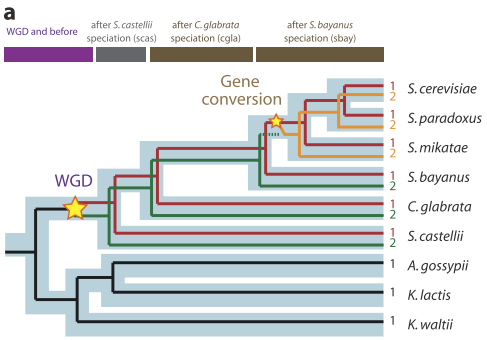
The fungal clade contains a whole-genome duplication (WGD) event, such that every gene simultaneously duplicated followed by many gene losses (Wolfe and Shields 1997; Kellis et al. 2004). Of the paralog pairs that are still present in the *Saccharomyces cerevisiae* genome, 37 of them have a K_s less than the average K_s between the *S. cerevisiae* and *S. bayanus* genomes of 1.05, indicating that these paralogs have undergone recent gene conversions near or after the speciation of the *S. cerevisiae* and *S. bayanus* lineages (Gao and Innan 2004) (see an example in fig. 6*a*). Also indicative of gene conversion (Noonan et al. 2004), these genes have a significantly elevated GC frequency of 42.0% in the third codon position, compared with a frequency of 37.9% for all *S. cerevisiae* genes ($P < 1.7 \times 10^{-09}$; Mann–Whitney *U*). Of these paralogs, SPIMAP infers 15 of them happening after the *S. bayanus* speciation and 31 after the *Candida glabrata* speciation (fig. 6*b*). In comparison, SYNERGY infers none of the paralogs duplicating after the *S. bayanus* speciation and only one after the *C. andidaglabrata* speciation. Instead 34 of the 37 paralogs are inferred as occurring on the branch containing the WGD, thus indicating that synteny information between *S. cerevisiae* and other postduplication species overrides sequence information in the vast majority of cases. For 33 families, the SPIMAP-constructed tree has a higher likelihood than the SYNERGY tree and for 22 families the likelihood is significantly higher ($P < 0.01$; SH test). In contrast, SYNERGY never has significantly higher likelihood.

Together these four metrics applied to real gene trees from 16 fungi suggest that SPIMAP often outperforms both sequence-only and species-aware methods. From these trees, we observe what appears to be an over estimation of DL events by the other methods, an error that has been observed in previous empirical studies (Hahn 2007). To better understand how phylogenetic errors influence the accuracy of event inference, we turn now to simulated data.

Reconstructing Simulated Gene Trees

To test our method on a data set where the correct phylogeny is unambiguously known, we implemented a simulation program based on our model for gene family evolution. Our intent was to make the simulations realistic by capturing the same gene and species-specific rate variation as well as gene DL rates as seen in real gene trees. Thus, the same model parameters and species phylogeny were used as those estimated for both the 12 flies and 16 fungi clades.

For each clade, we simulated 1,000 gene trees and generated the corresponding nucleotide alignments (supplementary figs. S3 and S4, Supplementary Material online). Next, we reconstructed gene trees from these simulated alignments using SPIMAP, PrIME-GSR, and the other sequence-only phylogenetic methods. Note, SYNERGY was excluded from the analysis because synteny, which is one of its inputs, was not simulated. SPIMAP's substitution rate parameters were estimated on a simulated data set with no DLs (supplementary figs. S8 and S9, Supplementary Material online). Its DL parameters were trained from the gene counts of each



b

Program	WGD	scas	cgla	sbay	% conversion recovered
SPIMAP	5	1	16	15	83.8%
SPIDIR	9	4	18	6	64.9%
SYNERGY	34	2	1	0	2.7%
PrIME-GSR	4	0	24	9	89.2%
MrBayes	5	0	21	11	86.5%
RAxML	6	0	22	9	83.8%
PHYML	5	0	19	13	86.5%
BIONJ	5	1	19	12	83.8%

FIG. 6. Inferred duplication times for recent *Saccharomyces cerevisiae* gene conversions. (*a*) Typical gene tree topology for 37 paralogous gene pairs originally arising from WGD and previously reported (Gao and Innan 2004) to have undergone gene conversion events (small star) near or after the speciation of *S. cerevisiae* and *S. bayanus*, such that one gene copy (green) is replaced by the other (red), followed by subsequent nucleotide divergence (orange). The correct inferred duplication of the two *S. cerevisiae* paralogs (red and orange lines, denoted 1 and 2) should occur within the time span indicated by the top brown bars. However, we expect methods that are heavily biased to follow the known species tree to incorrectly infer these events further up the tree. (*b*) We evaluated both traditional and species-aware methods in their ability to recover the correct trees in these cases and report the counts of where different gene conversion events are inferred for each method. We find that both SPIMAP and PrIME-GSR, as well as all traditional methods find the vast majority of these paralogs duplicates near or after *S. bayanus* speciation. However, SYNERGY incorrectly infers a WGD topology, most likely due to strong reliance in synteny information which is misleading in this case.

simulated data set (supplementary table S1, Supplementary Material online).

First, we measured topology accuracy across all the methods. SPIMAP outperforms the sequence-only programs by 7–29% on the simulated 12 flies data set and by 52–81% for the 16 fungi data set (fig. 7*a*). SPIMAP also showed a 3–8% accuracy increase over PrIME-GSR. To test whether lower reconstruction accuracy of the sequence-only methods was

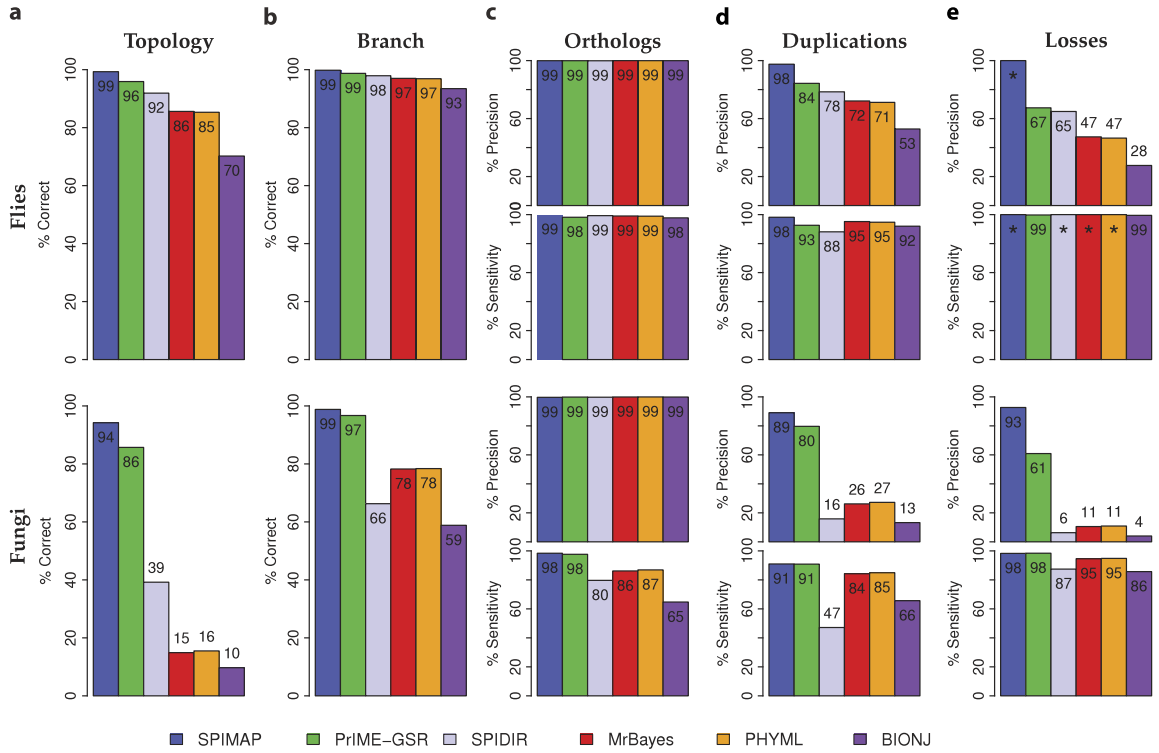


FIG. 7. Metrics of phylogenetic accuracy on simulated data sets for 12 *Drosophila* and 16 fungal species. (*a*) SPIMAP has a higher reconstruction accuracy for correctly inferring the full gene tree topology for both fly and fungal data sets. (*b*) The percent of accurately reconstructed branches is similar across methods for the 12 flies but a larger improvement is seen for SPIMAP on the larger and more diverse 16 fungi clade. (*c*) Despite topological and branch inaccuracies, pairwise ortholog detection is robust in both precision and sensitivity. (*d*, *e*) In contrast, DL inference is very sensitive to phylogenetic errors, especially in terms of precision. Stars indicate 100%.

due to insufficient search, we performed an addition run of PhyML where the tree search was initialized with the correct tree. However, topology accuracy for these runs only increased from 15.5% to 16.7%, indicating errors due to insufficient search play only a minor role. Overall the accuracy improvement for SPIMAP is larger on the fungi data set, which has a more complex and divergent phylogeny.

Second, we assessed partial topology correctness using the percent of branches accurately reconstructed. For the flies, SPIMAP consistently performs better but by only a few percent (fig. 7*b*). However, for the fungi, SPIMAP again shows a larger accuracy improvement at 20–39% over the sequence-only methods.

Third, we looked at the percentage of orthologs inferred correctly, where we noticed a surprising trend. Although topologies and branches had high error rates for many methods, there was also a high percentage of correctly inferred ortholog pairs (fig. 7*c*). Upon closer inspection, we found that often when a branch is misplaced it only disrupts a small fraction of the pairwise orthologs. Thus, it appears that orthology discovery at the pairwise level is quite robust to phylogenetic errors. In addition, we noticed that false positive orthologs calls are rarely made, although false negatives are more frequent, especially on the fungal clade.

Fourth, we looked at the accuracy of inferring gene DLs, which is very important for studies interested in study the rate of such events. As opposed to the ortholog pairwise metric, we find that DLs are very sensitive to phylogenetic errors. Notice that although branch accuracy may be high for some programs and data sets, even a small number of errors can lead to dramatic overestimation DLs (fig. 7*c* and *d* and supplementary fig. S6*b* and S6*a*). In general, all programs are able to recover DL events for the flies and fungi data sets with similar sensitivity (fraction of true positives among all actual positives). Programs differ by less than 6% difference, with SPIDIR and BIONJ as outliers. However, in terms of precision (fraction of true positives among all predicted positives), SPIMAP has a dramatic improvement in event inference over sequence-only methods: 2127% and 4553% for the flies DL, respectively, and 58–69% and 75–80% for fungi DLs (with BIONJ as an outlier in each case). Compared with PrIME-GSR, SPIMAPshows a 9–12% increase in duplication precision and 32–33% increase in loss precision. The 2–3-fold over prediction of events by the sequence-only phylogenetic methods (fig. 7*c* and *d*) is an effect similar to that seen in the real data.

Lastly, we find that these results also hold when simulations are performed with unusually high DL rates at twice ($2\times$) and four times ($4\times$) the estimated true rates ($1\times$). We performed simulations with five different settings $1\times$–$1\times$, $2\times$–$2\times$, $4\times$–$4\times$, $4\times$–$1\times$, $1\times$–$4\times$ for DL rates, respectively. We find that SPIMAP has increased performance for topology, branch, and event accuracy for all these rate settings (supplementary figs. S5*a* and S6*a*, Supplementary Material online). In addition, we found that SPIMAP was robust to errors in the DL rate parameters. When we reconstructed trees from the $1\times$–$1\times$ data set using DL rate parameters that were mis-specified to be four times faster than the true rates, topology accuracy was still 92.0% compared with 94.2% when using properly estimated parameters. Similarly, when reconstructing the $4\times$–$4\times$ data set using DL rate parameters that were one fourth the true rate, we obtained 71.1% topology accuracy compared with 69.2%.

Search Efficiency

In addition to evaluating reconstruction accuracy, we also evaluated reconstruction speed. Our goal with SPIMAP was to develop a method that is feasible enough to include in a phylogenomic pipeline containing thousands of trees and a variety of family sizes.

From the reconstruction of genes from our real data set (table 1), we found that SPIMAP has an average reconstruction time per tree (1.0 min) that is only slightly longer than that of PhyML (43.2 s). To investigate how our search strategy influences reconstruction run time, we generated a simulated data set of 500 gene families using 16 fungi species tree. For this simulation, we used i.i.d. species-specific rates ($\alpha_i = 2.819, \beta_i = 663.0$), no variation occurs in the gene rate, and the Jukes–Cantor model. We also used the same gene DL rates as estimated from real fungi gene families ($\lambda = 0.000732, \mu = 0.000859$). SPIMAP's substitution rate model was trained on a data set with the same parameters but no DLs. The parameters used by SPIMAP during reconstruction are given in supplementary figure S10, Supplementary Material online.

Although we have not implemented many optimizations for SPIMAP, our prescreening search strategy allows SPIMAP to compete with the highly optimized PhyML program (table 2). The RAxML program achieves significantly faster run times, but this occurs with a small decrease in accuracy for this data set. We believe our speed increase is because the gene family model, through the use of the species tree in the prior, produces a posterior distribution that is far more concentrated than the likelihood. Thus, many seemingly equivalent trees from a likelihood perspective are significantly different based on their priors and posteriors. In addition, our prescreening search strategy (see Rapid Tree Search) appears to greatly help in speeding up discovery of the MAP gene tree. For example, with no prescreening, SPIMAP achieves a topology accuracy of 32.4% with an average run time of 7.2 s. By using 100 prescreening iterations, accuracy increases to 84.8%, whereas run time only increases to 8.5 s. For comparison, PhyML achieves 26.0% topology accuracy in about 25.8 s on average.

Whether this prescreening strategy can scale to much larger trees with thousands of sequences (Stamatakis et al. 2005; Price et al. 2010), remains to be seen. However, the duplication subtree factoring within the topology prior may allow reuse of many computations between tree local rearrangements and could be combined with existing strategies for speeding up tree search.

SPIMAP is currently implemented as a MAP method, thus if branch support values are needed, bootstrapping will be required. Given the speed of our search, we can perform 100 bootstraps in about 11.1 minutes to achieve 86.4%

Table 2. Evaluation of Search Time for Several Phylogenetic Methods.

Program	Iterations[a]	Prescreens[b]	Bootstraps	Topology(%)	Branch(%)	Run time
RAxML	—	—	0	20.8	80.3	3.5 s
RAxML	—	—	100	22.8	1.1	39.8 s
PhyML	—	—	0	26.0	83.9	25.8 s
PhyML	—	—	100	26.0	83.9	13.9 m
SPIMAP	50	1	0	32.4	81.4	7.2 s
SPIMAP	100	1	0	50.8	87.1	12.7 s
SPIMAP	500	1	0	83.8	96.0	1.2 mmin
SPIMAP	1,000	1	0	88.6	97.5	2.0 min
SPIMAP	50	100	0	84.8	96.7	8.5 s
SPIMAP	1,000	100	0	90.8	98.1	2.3 min
SPIMAP	50	100	100	86.4	97.1	11.1 min

[a]Number of iterations used for each method.
[b]Number of prescreening iterations used for SPIMAP.

accuracy. This run time is comparable to 100 bootstraps of PhyML at 13.9 min and 26.0% accuracy. Thus, bootstrap analysis is quite feasible for SPIMAP, and the method should be efficient and practical enough for any pipeline that uses phylogenetic programs with run times on the order of PhyML's.

Lastly, we evaluated the influence of run time and family size on reconstruction accuracy. Using the same parameters above, we simulated more gene trees from the 16 fungal species tree and divided them into six classes based on the their number of extant genes: 5–9, 10–19, 20–29, 30–39, 40–49, and 50–59. Each size class was populated with 100 simulated trees and alignments. SPIMAP was run in two modes, one without bootstrapping (1,000 iterations and 100 prescreens) and one with 100 bootstraps (100 iterations and 100 prescreens). For the middle gene size class 20–29, SPIMAP achieved average run times of 5.3 and 50.4 min, respectively. For each data set, PrIME-GSR was also executed, using the same amount of time as SPIMAP, which required 7,300 iterations (quick mode) and 77,000 iterations (long mode) . We find that for smaller trees with 5–29 extant genes that both SPIMAP runs and PrIME-GSR's long mode achieve similar topology accuracy in the range of 80–100% (supplementary fig. S11, Supplementary Material online). However, for larger gene trees with 30–49 extant genes, as accuracy degrades for all methods, both modes of SPIMAP have a 20% increase in topology accuracy over PrIME-GSR. Improvements in inferring DL accuracy is also seen for the larger trees (>10% increase in duplication precision and >30% increase in loss precision).

Discussion

We present a novel probabilistic model and algorithm for gene tree reconstruction. Our approach uses a Bayesian framework to model sequence evolution, gene duplication, loss, and substitution rate variation, thus incorporating many disparate types of information in a principled way. This unified framework presents many advantages.

In contrast to previous gene tree reconstruction methods (Zmasek and Eddy 2002; Dufayard et al. 2005; Durand et al. 2006; Vilella et al. 2009), where a gene tree is reconciled only after full reconstruction by a method such as NJ or ML, our method finds a reconciliation and gene tree simultaneously. In addition, the parameters of our model are interpretable (e.g., substitutions rates and duplication/loss rates), and we have provided training algorithms for each one. This provides an advantage over a method like SYNERGY (Wapinski et al. 2007) that optimizes a parsimony-based cost function for several different events such duplications, loss, and syntenic relationships. Without a probabilistic basis, the weights of these costs and the behavior of their combination are more difficult to determine and analyze. Our study of gene conversions demonstrates more work is needed to understand how synteny information should be weighed against conflicting sources of information.

Our method models rate variation that is correlated across all branches of the tree (gene-specific rate) as well as rates specific to each species lineage (species-specific rates). We have found that when both these effects are modeled, the result is a more informative prior which leads to increased reconstruction accuracy (see the i.i.d. version of SPIMAP in table 1 and fig. 5). In contrast, PrIME-GSR uses identical and independent gamma distributions for rate variation which do not model species-specific rate variation. Thus, species with rate acceleration or decelerated across the genome will have branches that are consistently penalized by an i.i.d. rate prior. One complication for modeling species-specific rates is possibility of overparameterizing the model. We addressed this issue by learning the rate distributions prior to reconstruction from a data set of multiple orthologous gene trees. This learning step provides an advantage even when i.i.d. rates are used (see SPIMAP with i.d.d. and PrIME-GSR in table 1). By combining data across loci, the rate variation prior can be estimated more accurately than if the gene trees were considered in isolation.

The rate prior of our current work builds upon a previously developed method, SPIDIR (Rasmussen and Kellis 2007). We designed SPIDIR to be a distance-based likelihood method that exploits the rate variations we had observed in the 12 fly and 9 fungal genomes. Although the method proved effective, its reliance on pairwise distances did not fully utilize the available character information and it lacked an explicit model for DL rates. Indeed, we find in our latest comparison that SPIMAP has more consistent accuracy improvements than SPIDIR even for larger species trees

(16 fungi) and fast rates of DL (fig. 7 and supplementary figs. S5*a* and S6*a*, Supplementary Material online).

Thus far, we have only implemented a very simple model for sequence evolution within the SPIMAP program. Currently SPIMAP uses the HKY model, although providing additional models as well as modeling rate variation across sites may lead to additional improvements to reconstruction accuracy. However, we note that for the evaluations we present here, that modeling rate variation across sites did not contributed significantly to improved reconstructions. In fact, the PhyML program found very similar recovery rates for syntenic orthologs using both rate variation (64.2% recovery using four categories and an estimated α) or using none (64.9% recovery).

Lastly, we envision this method participating in a larger phylogenomic pipeline. We believe that within most clades of interest, there will be sufficient data for training our model. For example, in the 12 sequenced *Drosophila* species, about one-third of all genes are syntenic across all 12 species (Clark et al. 2007; Rasmussen and Kellis 2007) and thus can serve as a training set for our substitution rates model. Once a model is learned from simple gene families, it can then be applied to reconstruct gene families with more complicated histories of gene DL. Given these advances and many others to come, phylogenetics will likely play an ever increasing role in understanding the evolution and function of genomes.

Supplementary Material

Supplementary tables S1, sections 2.1–2.3, and figures S1–S11 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Michael F. Lin and the rest of the MIT Compbio group for helpful comments, feedback, and discussions. We thank M.R.'s thesis committee, Scott V. Edwards, Eric J. Alm, and Tommi S. Jaakkola for frutiful discussions at various stages of this work. This work was supported by National Science Foundation (NSF) CAREER award NSF 0644282 to M.K.

References

Adams MD, Celniker SE, Holt RA, et al. (195 co-authors). 2000. The genome sequence of Drosophila melanogaster. *Science* 287: 2185–2195.

Åkerborg O, Sennblad B, Arvestad L, Lagergren J. 2009. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc Natl Acad Sci U S A*. 106:5714–5719.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. 215:403–410.

Arvestad L, Berglund AC, Lagergren J, Sennblad B. 2003. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics* 19(Suppl 1):i7–i15.

Arvestad L, Berglund A, Lagergren J, Sennblad B. 2004. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. Proceedings of the Eighth Annual International Conference on Computational Molecular Biology 326–335.

Arvestad L, Lagergren J, Sennblad B. 2009. The gene evolution model and computing its associated probabilities. J. ACM 56:1–44.

Butler G, Rasmussen MD, Lin MF, et al. (51 co-authors). 2009. Evolution of pathogenicity and sexual reproduction in eight candida genomes. *Nature* 459:657–662.

Chen K, Durand D, Farach-Colton M. 2000. Notung: a program for dating gene duplications and optimizing gene family trees. *J Comput Biol*. 7:429–447.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science* 311:1283–1287.

Clark AG, Eisen MB, Smith DR, et al. (417 co-authors). 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* 450:203–218.

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M. 2003. Finding functional features in saccharomyces genomes by phylogenetic footprinting. *Science* 301:71–76.

Creevey CJ, McInerney JO. 2005. Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics* 21: 390–392.

Datta RS, Meacham C, Samad B, Neyer C, Sjlander K. 2009. Berkeley PHOG: phylofacts orthology group prediction web server. *Nucleic Acids Res*. 37:W84–W89.

Dehal PS, Boore JL. 2006. A phylogenomic gene cluster resource: the phylogenetically inferred groups (PHIGs) database. *BMC Bioinformatics* 7:201.

Dietrich FS, Voegeli S, Brachat S, et al. (14 co-authors). 2004. The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. *Science* 304:304–307.

Doyon JP, Chauve C, Hamel S. 2009. Space of gene/species trees reconciliations and parsimonious models. *J Comput Biol*. 16:1399–1418.

Dufayard JF, Duret L, Penel S, Gouy M, Rechenmann F, Perrière G. 2005. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* 21:2596–2603.

Dujon B, Sherman D, Fischer G, et al. (67 co-authors). 2004. Genome evolution in yeasts. *Nature* 430:35–44.

Durand D, Halldorsson BV, Vernot B. 2006. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J Comput Biol*. 13:320–335.

Dutheil JY, Ganapathy G, Hobolth A, Mailund T, Uyenoyama MK, Schierup MH. 2009. Ancestral population genomics: the coalescent hidden Markov model approach. *Genetics* 183:259–274.

Eddy SR. 2000. HMMER: profile hidden Markov models for biological sequence analysis. Availabel from: http://hmmer.org. **[AQ10]**

Edgar RC. 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32:1792–1797.

Eisen JA. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res*. 8:163–167.

Feller W. 1939. Die grundlagen der Volterraschen Theorie des kampfes ums dasein in Wahrscheinlichkeitstheoretischer Behandlung. *Acta Biotheo*. 5:11–40.

Felsenstein J. 1981. Evolutionary trees from dna sequences: a maximum likelihood approach. *J Mol Evol*. 17:368–376.

Felsenstein J. 2005. PHYLIP (phylogeny inference package). Version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington. **[AQ11]**

Gao L-z, Innan H. 2004. Very low gene duplication rate in the yeast genome. *Science* 306:1367–1370.

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol*. 14: 685–695.

Goffeau A, Barrell BG, Bussey H, et al. (16 co-authors). 1996. Life with 6000 genes. *Science* 274:546, 563–546, 567.

Goodman M, Czelusniak J, Moore G, Romero-Herrera A, Matsuda G. 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst Zool*. 28:132–163.

Gu X, Zhang H. 2004. Genome phylogenetic analysis based on extended gene contents. *Mol Biol Evol*. 21:1401–1408.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 52:696–704.

Hahn M. 2007. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol*. 8:R141.

Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianini N. 2005. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res*. 15:1153–1160.

Hahn MW, Han MV, Han S-G. 2007. Gene family evolution across 12 Drosophila genomes. *PLoS Genet*. 3:e197.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*. 22:160–174.

Hobolth A, Christensen OF, Mailund T, Schierup MH. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet*. 3:e7.

Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldon T. 2007. The human phylome. *Genome Biol*. 8:R109.

Jones T, Federspiel NA, Chibana H, et al. (12 co-authors). 2004. The diploid genome sequence of candida albicans. *Proc Natl Acad Sci U S A*. 101:7329–7334.

Jukes T, Cantor C. 1969. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic Press. p. 21–132.

Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature* 428:617–624.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423:241–254.

Li H, Coghlan A, Ruan J, et al. (15 co-authors). 2006. Treefam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res*. 34:D572–D580.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 13:2178–2189.

Liu L, Pearl DK. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst Biol*. 56:504–514.

Maddison WP, Knowles LL. 2006. Inferring phylogeny despite incomplete lineage sorting. *Syst Biol*. 55:21–30.

Massey SE, Moura G, Beltro P, Almeida R, Garey JR, Tuite MF, Santos MAS. 2003. Comparative evolutionary genomics unveils the molecular mechanism of reassignment of the CTG codon in Candida spp. *Genome Res*. 13:544–557.

Moschopoulos P. 1985. The distribution of the sum of independent gamma random variables. *Ann Inst Stat Math*. 37:541–544.

Noonan JP, Grimwood J, Schmutz J, Dickson M, Myers RM. 2004. Gene conversion and the evolution of protocadherin gene cluster diversity. *Genome Res*. 14:354–366.

Page R. 1994. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst Biol*. 43:58–77.

Page RD, Charleston MA. 1997. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol Phylogenet Evol*. 7:231–240.

Price MN, Dehal PS, Arkin AP. 2010. Fasttree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490.

Rannala B, Yang Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J Mol Evol*. 43:304–311.

Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci. *Genetics* 164:1645–1656.

Rasmussen MD, Kellis M. 2007. Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. *Genome Res*. 17:1932–1942.

Richards S, Liu Y, Bettencourt BR, et al. (52 co-authors). 2005. Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. *Genome Res*. 15:1–18.

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425:798–804.

Ronquist F, Huelsenbeck JP. 2003. Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 4:406–425.

Sanderson MJ. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19:301–302.

Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol*. 16:1114–1116.

Stamatakis A, Ludwig T, Meier H. 2005. RAxML: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 21:456–463.

Tamura K, Subramanian S, Kumar S. 2004. Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. *Mol Biol Evol*. 21:36–44.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. Ensemblcompara genetrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*. 19:327–335.

Wakeley J. 2009. Coalescent theory: an introduction. Roberts & Company Publishers.

Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449:54–61.

Wapinski I, Pfeffer A, Friedman N, Regev A. 2009. Synergy dataset January 2009 update. http://www.broadinstitute.org/regev/orthogroups/.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387:708–713.

Zmasek CM, Eddy SR. 2001. A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics* 17:821–828. **[AQ12]**

Zmasek CM, Eddy SR. 2002. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics* 3:14.

Discovery and characterization of chromatin states for systematic annotation of the human genome

Jason Ernst[1,2] & Manolis Kellis[1,2]

A plethora of epigenetic modifications have been described in the human genome and shown to play diverse roles in gene regulation, cellular differentiation and the onset of disease. Although individual modifications have been linked to the activity levels of various genetic functional elements, their combinatorial patterns are still unresolved and their potential for systematic *de novo* genome annotation remains untapped. Here, we use a multivariate Hidden Markov Model to reveal 'chromatin states' in human T cells, based on recurrent and spatially coherent combinations of chromatin marks. We define 51 distinct chromatin states, including promoter-associated, transcription-associated, active intergenic, large-scale repressed and repeat-associated states. Each chromatin state shows specific enrichments in functional annotations, sequence motifs and specific experimentally observed characteristics, suggesting distinct biological roles. This approach provides a complementary functional annotation of the human genome that reveals the genome-wide locations of diverse classes of epigenetic function.

© 2010 Nature America, Inc. All rights reserved.

The primary DNA sequence of the human genome encodes the genetic information of each cell, but numerous epigenetic modifications can modulate the interpretation of the primary sequence. These modifications contribute to the diversity of phenotypes found across different human cell types, play key roles in the establishment and maintenance of cellular identity during development and have been associated with DNA repair, replication and human disease. Post-translational modifications in the tails of histone proteins that package DNA into chromatin constitute perhaps the most versatile type of such epigenetic information. More than a dozen positions of multiple histone proteins can undergo a number of modifications, such as acetylation and mono-, di- or tri-methylation[1,2].

More than 100 distinct histone modifications have been described, leading to the 'histone code hypothesis' that specific combinations of chromatin modifications would encode distinct biological functions[3]. Others, however, have instead proposed that individual epigenetic marks act in additive ways and the multitude of modifications simply contributes to stability and robustness[4]. The specific combinations of epigenetic modifications that are biologically meaningful, and their corresponding functional roles, are still largely unknown.

To directly address these questions, we introduce an approach for the *de novo* discovery of 'chromatin states' (**Fig. 1**, **Supplementary Table 1** and **Supplementary Fig. 1**), or biologically meaningful and spatially coherent combinations of chromatin marks, by performing a systematic genome-wide analysis based on a multivariate Hidden Markov Model (HMM). Multivariate HMMs are graphical probabilistic models that model multiple 'observed' inputs as generated by unobserved 'hidden' states, using transitions between hidden states to model spatial relationships (Online Methods).

Our model captures two types of chromatin information. The frequency with which different chromatin mark combinations are found with each other are captured by a vector of 'emission' probabilities associated with each chromatin state (**Fig. 2** and **Supplementary Figs. 2** and **3**) and the frequency with which different chromatin states occur in spatial relationships of each other along the genome are encoded in a 'transition' probability vector associated with each state. These spatial relationships capture both the spreading of certain chromatin domains across the genome, as well as the functional ordering of different states such as from intergenic regions to promoter regions and transcribed regions (**Supplementary Notes** and **Supplementary Figs. 4–6**). Biologically the genomic locations associated with a given chromatin state may correspond to specific types of functional elements, such as transcription start sites (TSS), enhancers, active genes, repressed genes, exons or heterochromatin, which can be inferred solely from the corresponding combinations of chromatin marks in their spatial context, even though no information about these annotations is given to the model as input.

We applied our model to the largest data set of chromatin mark information available, consisting of the genome-wide occupancy data for a set of 38 different histone methylation and acetylation marks and for the histone variant H2AZ, RNA polymerase II (PolII) and CTCF in human CD4 T-cells. The maps were previously obtained using chromatin immunoprecipitation followed by next generation sequencing (ChIP-seq) (Online Methods)[5,6]. To understand the biological importance of the resulting chromatin states, we undertook a large-scale, systematic data-mining effort, bringing to bear dozens of genome-wide data sets including gene annotations, expression information, evolutionary conservation, regulatory motif instances, compositional biases, genome-wide association data, transcription-factor binding, DNaseI hypersensitivity and nuclear lamina maps.

This work provides an unbiased and systematic chromatin-driven annotation for every region of the genome at a 200 base pair resolution, refining previously described epigenetic states and introducing

[1]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts, USA. [2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. Correspondence should be addressed to M.K. (manoli@mit.edu).

Published online 25 July 2010; doi:10.1038/nbt.1662

additional ones. Regardless of whether these chromatin states are causal in directing regulatory processes, or simply reinforcing independent regulatory decisions, these annotations should provide a resource for interpreting biological and medical data sets, such as genome-wide association studies for diverse phenotypes and could potentially help to identify new classes of functional elements.

RESULTS

Chromatin states model and comparison to previous work

Previous analyses have largely focused on characterizing the marks predictive of specific classes of genomic elements defined a priori such as transcribed regions, promoters or putative enhancers, and using the characterization to identify new instances of these classes[5–12].

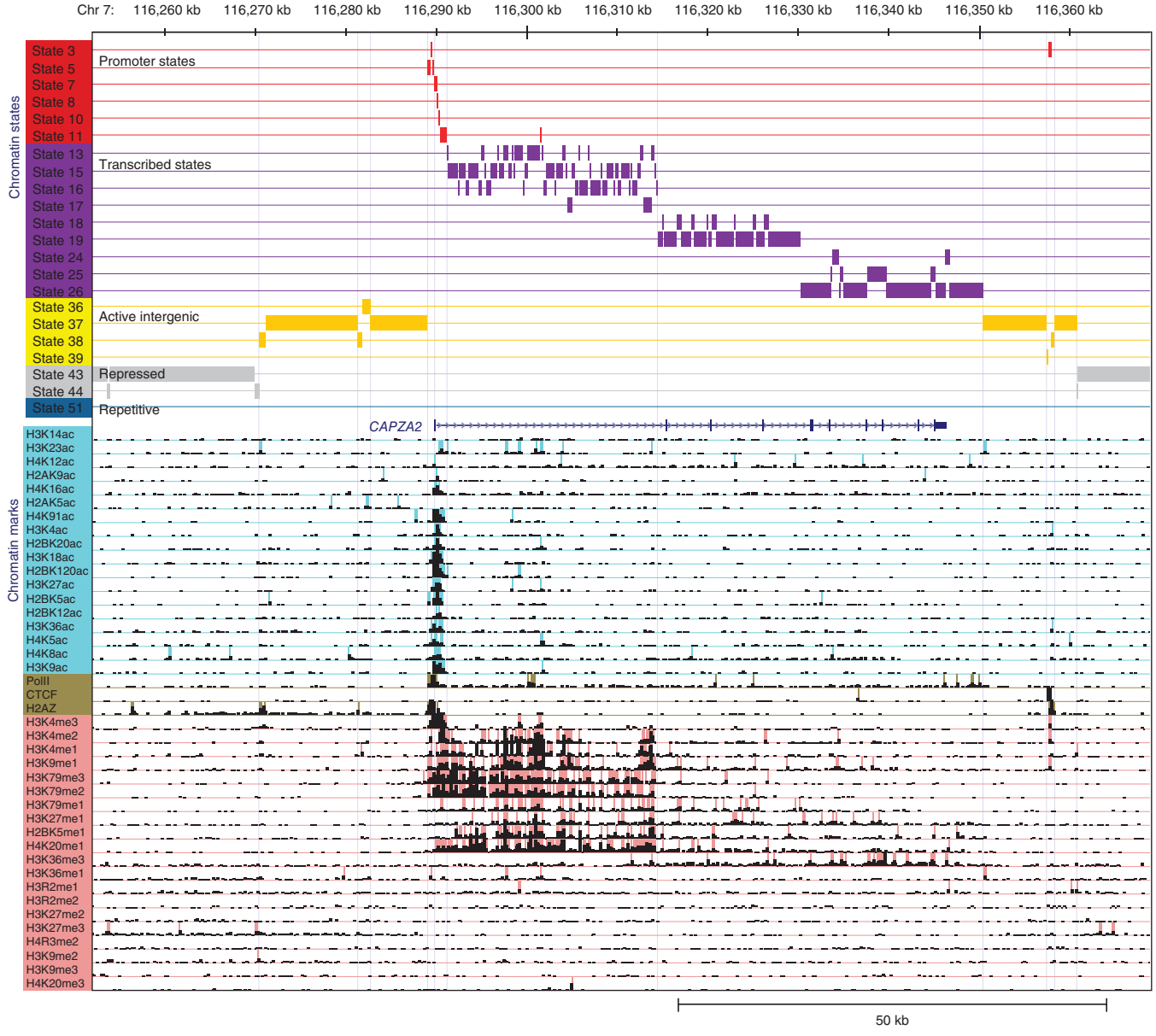


Figure 1 Example of chromatin state annotation. Input chromatin mark information and resulting chromatin state annotation for a 120-kb region of human chromosome 7 surrounding the *CAPZA2* gene. For each 200-bp interval, the input ChIP-Seq sequence tag count (black bars) is processed into a binary presence and/or absence call for each of 18 acetylation marks (light blue), 20 methylation marks (pink) and CTCF/PolI2/H2AZ (brown). The precise combination of these marks in each interval in their spatial context is used to infer the most probable chromatin state assignment (colored boxes). Although chromatin states were learned independently of any prior genome annotation, they correlate strongly with upstream and downstream promoters (red), 5′-proximal and distal transcribed regions (purple), active intergenic regions (yellow), repressed (gray) and repetitive (blue) regions (state descriptions shown in **Supplementary Table 1**). This example illustrates that even when the signal coming from chromatin marks is noisy, the resulting chromatin state annotation is very robust, directly interpretable and shows a strong correspondence with the gene annotation. Several spatially coherent transitions are seen from large-scale repressed to active intergenic regions near active genes, from upstream to downstream promoter states surrounding the TSS and from 5′-proximal to distal transcribed regions along the body of the gene. The frequent transitions to state 16 correlate with annotated Alu elements (57% overlap versus 4% and 25% for states 13 and 15, respectively). Transitions to state 13 are likely due to enhancer elements in the first intron of *CAPZA2*, a region where regulatory elements are commonly found and correlate with several enhancer marks. The maximum-probability state assignments are shown here, and the full posterior probability for each state in this region is shown in **Supplementary Figure 1**.

© 2010 Nature America, Inc. All rights reserved.

An unsupervised (without using prior knowledge) local chromatin pattern discovery method[13] first demonstrated that many of the patterns previously associated with promoters and enhancers could be discovered *de novo*, but did not discover patterns associated with broader domains and left the vast majority of the genome unannotated (**Supplementary Fig. 7**).

Unsupervised HMM approaches that modeled chromatin mark signal intensity levels using multivariate normals or nonparametric histograms[14–18] have been previously used, but in contrast we use a binarization approach that explicitly models the presence/absence frequency of each mark. Specifically, we make a local call of whether a mark was present in each 200-bp interval, and use a Bernoulli random variable to model the probability of detection of each mark in isolation, and a product of independent probabilities to model the probability of each combination of marks (Online Methods). Our approach has the advantage that the model parameters are directly interpretable as the frequencies of each mark and each mark combination, in contrast to previous approaches for which the biological significance of the parameters corresponding to varying signal intensity levels for each mark is often unclear. Moreover, the binarization also makes our

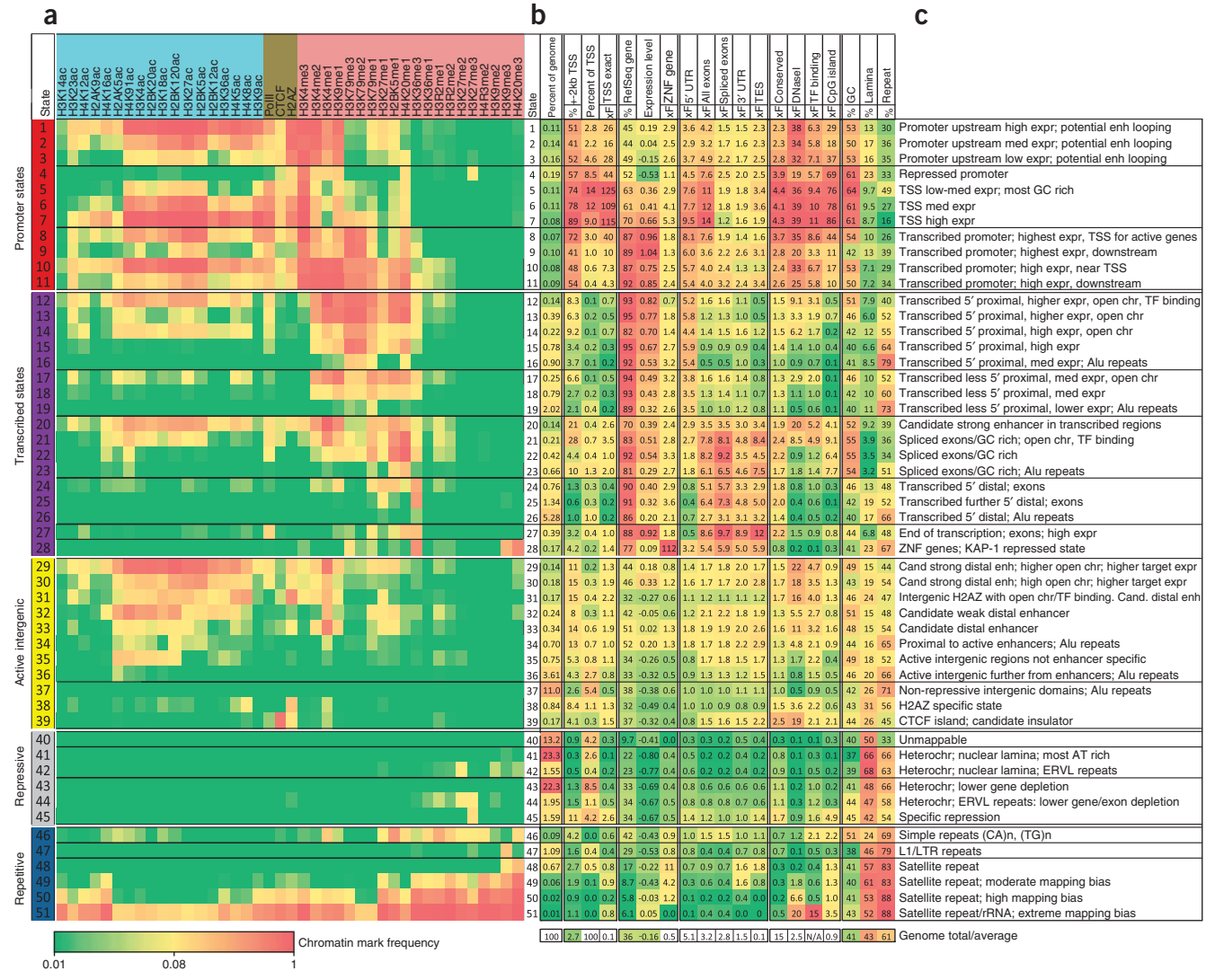


Figure 2 Chromatin state definition and functional interpretation. (**a**) Chromatin mark combinations associated with each state. Each row shows the specific combination of marks associated with each chromatin state and the frequencies between 0 and 1 with which they occur (color scale). These correspond to the emission probability parameters of the Hidden Markov Model (HMM) learned across the genome during model training (values shown in **Supplementary Fig. 2**). Marks and states colored as in **Figure 1**. (**b**) Genomic and functional enrichments of chromatin states. %, percentage; xF, fold enrichment. In order, columns are: percentage of the genome assigned to the state; percentage of state that overlaps a 200-bp interval within 2 kb of an annotated RefSeq TSS; percentage of RefSeq TSS found in the state; fold enrichment for TSS; percentage of state overlapping a RefSeq transcribed region; average expression level of genomic intervals overlapping the state; fold-enrichment for zinc-finger–named gene; fold-enrichment for RefSeq 5′ Untranslated Region (5′-UTR) exon and introns; fold enrichment for RefSeq exons; fold enrichment for spliced exons (2nd exon or later); fold enrichment for RefSeq 3′ Untranslated Region (3′-UTR) exons and introns; fold enrichment for RefSeq transcription end sites (TES); fold enrichment for PhastCons conserved elements; fold enrichment for DNaseI hypersensitive sites; median fold enrichment for transcription factor binding sites over a set of experiments (expanded in **Supplementary Fig. 23**); fold-enrichment for CpG islands; percentage of GC nucleotides; percent overlapping experimental nuclear lamina data; percent overlapping a RepeatMasker element (expanded in **Supplementary Fig. 31**). All enrichments are based on the posterior probability assignments. Genome total indicates the total percentage of 200 bp interval intersecting the feature or the genome average for expression and percent GC. (**c**) Brief description of biological state function and interpretation (chr, chromatin; enh, enhancer, full descriptions in **Supplementary Table 1**).

© 2010 Nature America, Inc. All rights reserved.

a

Gene GO category	Chromatin state at TSS of corresponding gene: 3	4	5	6	7	8
Cell cycle phase	2.70 (10^{-7})	0.57 (1.0)	1.61 (10^{-3})	1.45 (1.0)	1.15 (1.0)	1.51 (1.0)
Embryonic development	1.24 (1.0)	2.82 (10^{-22})	1.07 (1.0)	0.85 (1.0)	0.54 (1.0)	1.00 (1.0)
Chromatin	1.20 (1.0)	0.48 (1.0)	2.17 (10^{-7})	1.64 (1.0)	0.85 (1.0)	0.85 (1.0)
Response to DNA damage	1.20 (1.0)	0.35 (1.0)	1.55 (0.07)	2.13 (10^{-11})	1.97 (10^{-4})	0.84 (1.0)
RNA processing	0.49 (1.0)	0.26 (1.0)	1.31 (1.0)	1.91 (10^{-11})	2.64 (10^{-24})	2.46 (10^{-4})
T-cell activation	0.77 (1.0)	0.88 (1.0)	1.27 (1.0)	0.70 (1.0)	0.79 (1.0)	4.72 (10^{-7})

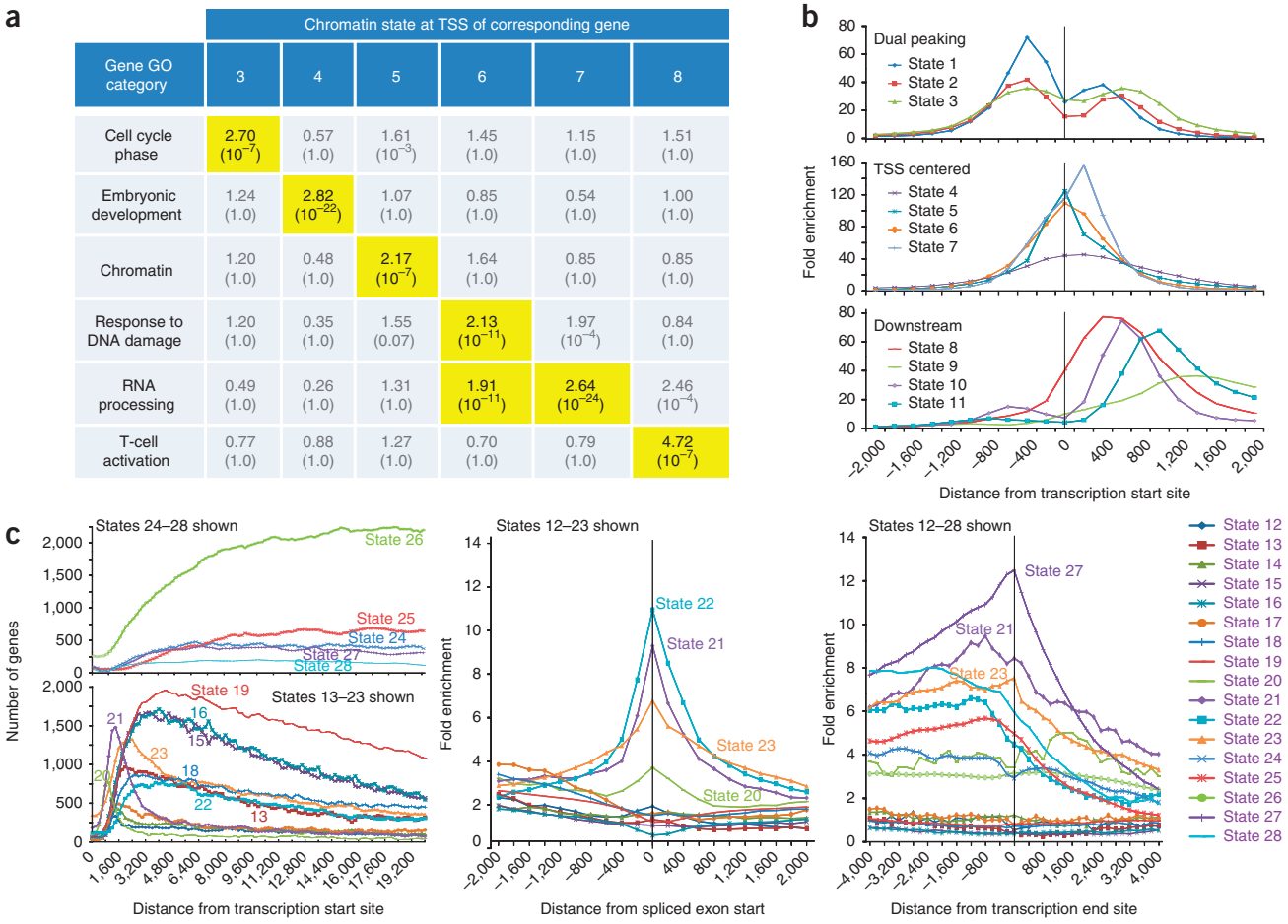


Figure 3 Promoter and transcribed chromatin states show distinct functional and positional enrichments. (**a**) Distinct Gene Ontology (GO) functional enrichments (fold and corrected *P*-values) found for genes associated with different promoter states at their TSS. For additional states and GO terms, see **Supplementary Figure 29**. (**b**) Distinct positional biases of promoter states with respect to nearest RefSeq TSS distinguish states peaking upstream, only downstream and centered at the TSS. (**c**) Positional biases of transcribed states with respect to TSS, nearest spliced exon start and transcription end sites (TES). These distinguish 5′-proximal states (12–23, left panel), 5′-distal states (24–28), states strongly enriched for spliced exons (middle panel, see also **Supplementary Fig. 24** for plot for states 24–28) and TES-associated states (with state 27 being particularly precisely positioned, right panel).

© 2010 Nature America, Inc. All rights reserved.

model less prone to forming states overfitting potentially insignificant variations in signal intensity levels. In contrast to models that use a multivariate normal distribution, our method avoids this strong parametric assumption, which is generally violated by the often relatively small discrete counts found in ChIP-seq experiments, enabling more robust models to be inferred. In comparison to the models previously inferred based on a nonparametric histogram strategy[18], our binarization approach uses an order of magnitude fewer parameters per state, further increasing model robustness and interpretability.

We developed a procedure for learning sets of chromatin states across a range of model complexities. For a given number of states and from a set of initial parameters, standard expectation maximization based procedures enable simultaneous local optimization of the state definitions (emission and transition probabilities) and the corresponding genome annotation consistent with the observed data. However the model inferred and its quality can depend on the initial set of parameters, which can confound comparing models with different number of states learned from independent initializations. We therefore used a two-stage process that first selected a 79-state model which had the highest complexity-penalized likelihood score across a large compendium of randomly-initialized models of varying complexity. We then pruned and optimized this model down to smaller numbers of states, leading to a model with 51 states that were relatively consistently recovered across the compendium of models, and that sufficiently captured all states found in larger models for which we could give a distinct biological interpretation (see Online Methods). This enabled us to maintain a relatively small number of states while capturing most of the unique biology uncovered across our compendium of randomly-initialized models. Put in other words, this procedure enabled us to maximize biological interpretability, while minimizing model complexity. We further ensured that general properties of the resulting model validated our approach, including robustness to varying thresholds and different background models, and independence of marks given a chromatin state (**Supplementary Notes**, **Supplementary Figs. 8–21** and **Supplementary Table 2**).

We next describe the likely biological functions of the 51 discovered chromatin states, divided into five large groups.

Promoter-associated states

The first group of states, states 1–11, all had high enrichment for promoter regions: 40–89% of each state was within 2 kb of a RefSeq TSS, compared with 2.7% genome-wide ($P < 10^{-200}$, for all states).

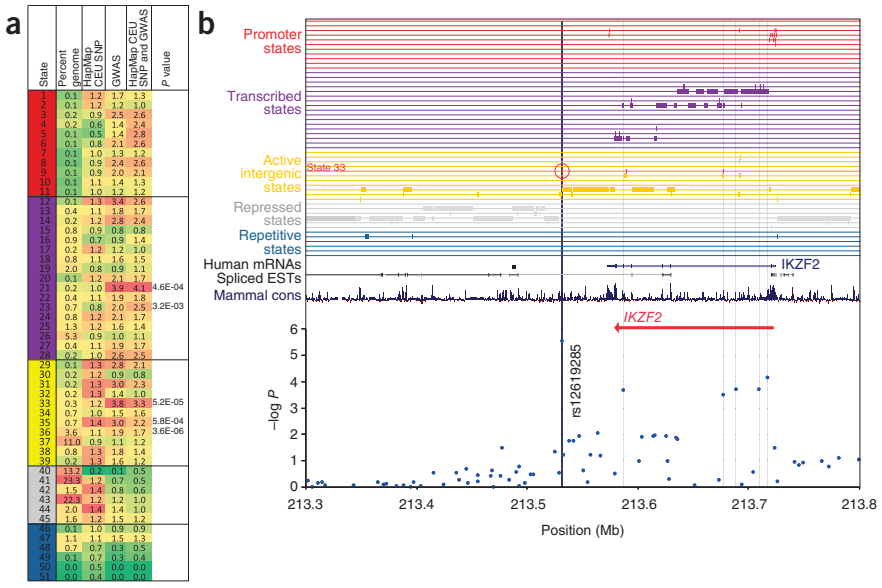


Figure 4 SNP and GWAS enrichments for chromatin states. (**a**) Several chromatin states show enrichments for disease association data sets. For each state is shown: genome percentage; fold enrichment for SNPs from the HapMap CEU population; fold enrichment from a collection of 1,640 GWAS SNPs associated with a variety of diseases and traits from numerous studies[25]; fold enrichment of GWAS SNPs relative to the HapMap CEU SNP enrichment; significance of GWAS SNPs relative to the underlying SNP frequency (when the corrected *P*-value < 0.01). (**b**) Example of intergenic SNP in GWAS-enriched state 33, found 40 kb downstream of the IKZF2 gene and associated with plasma eosinophil count levels[26]. SNP significance as reported[26] is shown for each SNP in the region (blue circles) and associated chromatin state annotation (similar to **Fig. 1**). Red circle denotes top SNP and its overlap with state 33. In addition to top SNPs, secondary SNPs were also frequently found at or near GWAS-enriched states in several cases.

These states accounted for 59% of all RefSeq TSS although they covered only 1.3% of genome. These states all had a high frequency of H3K4me3 in common, as well as significant enrichments for DNaseI hypersensitive sites, CpG islands, evolutionarily conserved motifs and bound transcription factors (**Fig. 2**). They differed however in the presence and levels of other associated marks, primarily H3K79me2/3, H4K20me1, H3K4me1/2 and H3K9me1, and of numerous acetylations leading to varying strength of the aforementioned functional enrichments, and varying expression levels of the downstream genes (**Supplementary Figs. 22** and **23**).

Promoter states differed in the enrichment of Gene Ontology (GO) terms of associated genes including cell cycle, embryonic development, RNA processing and T-cell activation (**Fig. 3a**). For instance, the term 'embryonic development' is specifically enriched in state 4, whereas the term 'T-cell activation' is specifically enriched in state 8. Promoter states also differed in their preferentially enriched positions with respect to the TSS of associated genes (**Fig. 3b**). States 4–7 were most concentrated over the TSS (showing upwards of 100-fold enrichment), states 8–11 peaked between 400 bp and 1,200 bp downstream of the TSS and corresponded to transcribed promoter regions of expressed genes and states 1–3 peaked both upstream and downstream of the TSS.

Transcription-associated states

The second large group of chromatin states consisted of 17 transcription-associated states. These are 70–95% contained within RefSeq-annotated transcribed regions compared to 36% for the rest of the genome (**Fig. 2b**, $P < 10^{-200}$, for all states). This group was not predominantly associated with a single mark, but instead defined by combinations of seven marks, H3K79me3, H3K79me2, H3K79me1, H3K27me1, H2BK5me1, H4K20me1 and H3K36me3 (**Fig. 2a**). Inspection of the transition frequencies between these states revealed subgroups of states that are associated with 5′-proximal or 5′-distal locations and with different expression levels (**Fig. 2c**, **Supplementary Notes**, **Supplementary Table 1** and **Supplementary Fig. 4**).

We observed several states strongly enriched for spliced exons (states 21–25 and 27–28 with 5.7- to 9.7-fold enrichments) (**Figs. 2b** and **3c** and **Supplementary Fig. 24**). Spliced exons were previously reported to be enriched in several individual marks[19–21]. In contrast to these previous studies, the combinatorial approach we have taken here shows that individual marks in spliced exonic states are also frequently detected in several other states that show only a modest 1.3- to 1.6-fold enrichment for spliced exons (e.g., states 12, 13, 14 and 17). This suggests that the chromatin signature of spliced exons is not solely defined by the presence of the previously reported H3K36me3, H2BK5me1, H4K20me1 and H3K79me1 marks, but their specific combinations and the absence of H3K4me2, H3K9me1 and H3K79me2/3.

State 27 showed a 12.5-fold enrichment for transcription end sites (TES) with its enrichment peaking directly over these locations (**Fig. 3c**). It was characterized both by the presence of H3K36me3, PolII and H4K20me1 and the absence of H3K4me1, H3K4me2 and H3K4me3, distinguishing it from other transcribed states with higher PolII or H3K36me3 frequencies. This suggests a distinct signature for 3′ ends of genes for which, to our knowledge, no specific chromatin signature had been described before. This was further validated by a 3.4-fold signal enrichment for the elongating form of PolII surveyed in an independent study[22] (**Supplementary Fig. 25**), even though our input data did not distinguish between the elongating and non-elongating form.

State 28 showed a 112-fold enrichment in zinc-finger genes, which comprise 58% of the state. This state was characterized by the high frequency for H3K9me3, H4K20me3 and H3K36me3 and relatively low frequency of other marks. This specific combination has been independently reported as marking regions of KAP1 binding, a zinc-finger–specific co-repressor, which also shows a specific 44-fold enrichment for state 28 (refs. 23,24). Although the association of H3K9me3 and H4K20me3 with zinc-finger genes has been previously reported[5], the *de novo* discovery of this highly specific signature of zinc-finger genes illustrates the utility of the methodology and also reveals the additional presence of H3K36me3 and lower frequency of other marks as complementing the signature of zinc-finger genes.

Active intergenic states

The third broad class of chromatin states consisted of 11 active intergenic states (states 29–39), including several classes of candidate enhancer regions, insulator regions and other regions proximal to expressed genes (**Supplementary Notes**). These states were associated with higher frequencies for H3K4me1, H2AZ, numerous acetylation marks and/or CTCF and with lower frequencies for other methylation marks (**Fig. 2a** and **Supplementary Figs. 2**

© 2010 Nature America, Inc. All rights reserved.

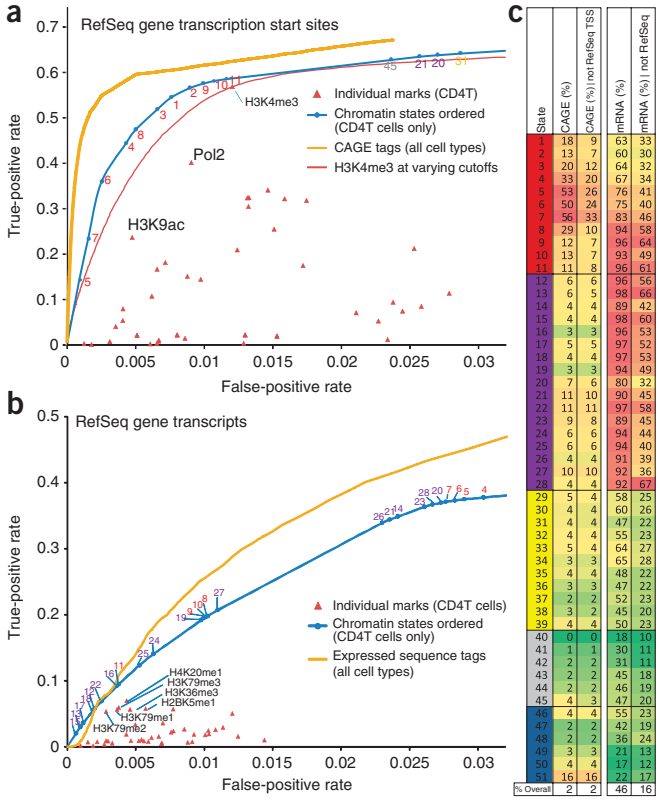


Figure 5 Discovery power of chromatin states for genome annotation. (**a**) Comparison of the power to discover TSS for individual chromatin marks (red), chromatin states (blue) ordered by their TSS enrichment and a directed experimental approach based on CAGE sequence tag data read counts from all available cell types[36] (gold), whereas the chromatin states and marks use only data from CD4 T-cells. Both chromatin states and CAGE tags are compared using a receiver operating characteristic (ROC) curve that shows the false-positive (*x* axis) and true-positive (*y* axis) rates at varying prediction thresholds or increasing numbers of states in the task of predicting if a 200-bp interval intersects a RefSeq TSS. Thin red curve compares performance of H3K4me3 mark at varying intensity thresholds. (**b**) Comparison of the power to detect RefSeq transcribed regions for chromatin states and marks as in **a**, and directed experimental information coming from EST data (gold) based on sequence counts from all available cell types[37,38]. (**c**) Independent experimental information provides support that a significant fraction of false positives in **a** and **b** are genuine unannotated TSS and transcribed regions currently missing from RefSeq. Percentage of each state supported by a CAGE tag (column 1), and the same percentage for locations at least 2 kb away from a RefSeq TSS (column 2), suggests that many promoter-associated state assignments outside RefSeq promoters are supported by CAGE tag evidence. Similarly, percentage of each state overlapping a GenBank mRNA (column 3), and the same percentage specifically outside RefSeq genes (column 4), suggest that transcription-associated state assignments outside RefSeq genes are supported by mRNA evidence. Similar support is found by GenBank ESTs and evolutionarily conserved, predicted new exons (**Supplementary Fig. 33**).

and **3**). They occurred primarily away from promoter regions (85–97% outside 2 kb of a TSS) and outside of transcribed genes (48–64% outside of RefSeq annotations, **Fig. 2b**). When they overlapped gene annotations, it was mainly in regions that were repressed or not highly expressed (see expression column in **Fig. 2b**).

States 29–33 were notable as they corresponded to smaller fractions of the genome specifically associated with greater DNaseI hypersensitivity, transcription factor binding and regulatory motif instances and are likely to represent enhancer regions (**Fig. 2** and **Supplementary Fig. 23**). Although these candidate enhancer states all shared higher H3K4me1 frequencies, they showed differences in the expression levels of downstream genes associated with subtle differences in their specific mark combinations (**Supplementary Fig. 22**). For instance, genes downstream of state 30 had a consistently higher average expression level than genes downstream of state 31 ($P < 0.001$ at 10 kb, two-sided *t*-test). The two states differed in the frequency of several acetylation marks (state 30 relative to 31 showed higher frequency for H2BK120ac, H3K27ac and H2BK5ac and lower frequency for H4K5ac, H4K8ac) and also in the level of H2AZ (higher in state 31 than 30), suggesting that these marks may be playing a more complex role than previously thought in enhancer regions.

Several active intergenic states showed significant enrichments for genome-wide association study (GWAS) hits (e.g., 3.3-fold for candidate enhancer state 33, **Fig. 4a**), based on a curated database of top-scoring single-nucleotide polymorphisms (SNPs) in a range of diseases and traits[25]. These states thus provide a likely common functional role and means of refining many intergenic SNPs even in the absence of other annotations. For example (**Fig. 4b**), a SNP reported to be strongly associated with plasma eosinophil count levels in inflammatory diseases (rs12619285)[26] and located 40 kb downstream of IKZF2 in an intergenic region devoid of annotations is in a section of the genome in the chromatin state 33, which is enriched for GWAS hits. In contrast, the surrounding region of the genome is assigned to other active or repressed intergenic states with no significant GWAS association.

Large-scale repressed states

The next group of states (40–45) marked large-scale repressed and heterochromatic regions, representing 64% of the genome. The two most frequently detected modifications in total for all the states in this group were H3K27me3 and H3K9me3. State 40, covering 13% of the genome, was essentially devoid of any detected modifications, states 41–42 (25% of the genome) had a higher frequency for H3K9me3 than H3K27me3, whereas states 43–45 (26% of the genome) had a higher frequency for H3K27me3. States 41–42 as compared to states 43–45 showed a stronger depletion for genes, promoters and conserved elements and stronger association with nuclear lamina regions[27] and the darkest-staining chromosomal bands[28]. It also had a higher frequency of A/T nucleotides (**Fig. 2b** and **Supplementary Figs. 26–28**).

State 45 likely corresponds to targeted gene repression. It showed the highest frequency for H3K27me3 and was unique among repressed states to show enrichment for TSS. The corresponding genes were enriched for development-related GO categories (**Supplementary Fig. 29**), similar to the repressed promoter state 4 marked by H3K4me3. However, in contrast to state 4, state 45 showed almost no change in acetylation levels in response to histone deacetylase inhibitor (HDACi) treatment (**Supplementary Fig. 30**), suggesting that state 4 is poised for activation whereas state 45 is stably repressed[29].

Repetitive states

The final group of six states (46–51) showed strong and distinct enrichments for specific repetitive elements (**Supplementary Fig. 31**). State 46 had a strong enrichment of simple repeats, specifically $(CA)_n$, $(TG)_n$ or $(CATG)_n$ (44, 45 and 302-fold, respectively), possibly due to sequence biases in ChIP-based experiments[30]. State 47 was characterized specifically by H3K9me3 and enriched for L1 and LTR repeats. State 48–51 all had higher frequencies of H4K20me3 and H3K9me3 and were heavily enriched for satellite repeat elements.

© 2010 Nature America, Inc. All rights reserved.

States 49–51 showed seemingly high frequencies for numerous modifications, but also strong enrichments in sequence reads from a nonspecific antibody (IgG) control[31] (**Supplementary Fig. 20**), suggesting these enrichments are due to a lack of coverage for the additional copies of these repeat elements in the reference genome assembly[32], thus illustrating the ability of our model to capture such potential artifacts by considering all marks jointly.

Predictive power for genome annotation

We next set out to study the predictive power of chromatin states for the discovery of functional elements. We focused on two classes of elements that benefit from ample experimental information independent of chromatin marks, TSS and transcribed regions. We found that chromatin states consistently outperformed predictions based on individual marks (**Fig. 5a**,**b**), emphasizing the importance of using

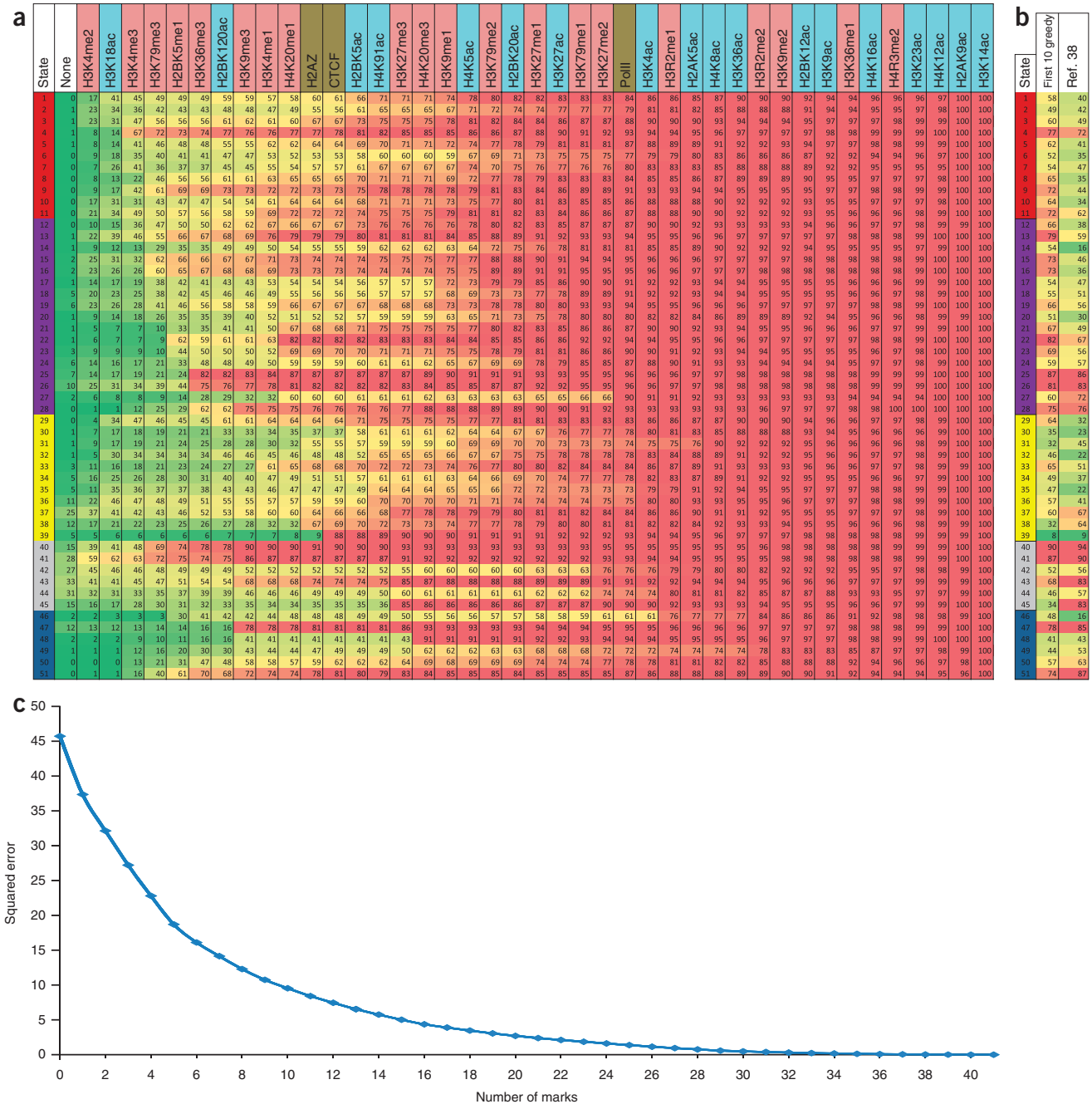


Figure 6 Recovery of chromatin states with subsets of marks. (**a**) The figure shows the ordering of marks (top, from left to right) based on a greedy forward selection algorithm to optimize a squared error penalty on state misassignments (Online Methods). Conditioned on all the marks to the left having already been profiled, the mark listed is the optimal selection for one additional mark to be profiled based on the target optimization function. Below each mark is the percentage of a state with identical assignments using the subset of marks. (**b**) Comparison of the percentage of each state recovered between the first ten marks based on the greedy method and the ten marks previously used[33] (**Supplementary Fig. 39**). The two columns after the state IDs are the proportion of the states recovered using the greedy algorithm and the set previously used[33]. (**c**) The figure shows a progressive decrease in squared error for state misassignment as a function of the number of marks selected based on the greedy algorithm.

© 2010 Nature America, Inc. All rights reserved.

mark combinations and spatial genomic information (**Supplementary Notes** and **Supplementary Fig. 32** for a comparison to *k*-means clustering and a supervised classifier). The prediction performance of chromatin states based on just CD4 T-cells was similar to that of cap analysis of gene expression (CAGE) tags and expressed sequence tags (ESTs) data, even though these were obtained across many diverse cell types. This was possible because active and inactive states together capture the information about genetic elements across cell type boundaries (**Fig. 5** and **Supplementary Figs. 33–35**). Moreover, based on our 51-state model, we could predict TSS and transcribed regions when applied to occupancy data obtained for a subset of ten chromatin marks in CD36 erythrocyte precursors and CD133 hematopoietic stem cells[33] (**Supplementary Fig. 36**).

We also found that chromatin states revealed candidate promoter and transcribed regions not in RefSeq, but further supported by independent experimental evidence. Candidate promoters overlapped with CAGE tags (**Fig. 5c**) and intergenic PolII (**Supplementary Fig. 37**), and candidate transcribed regions overlapped GenBank mRNAs (**Fig. 5c**) and EST data (**Supplementary Fig. 33**). A number of promoter and transcribed states outside known genes were also strongly enriched for not previously described protein-coding exons predicted using evolutionary comparisons of 29 mammals (Lin and M.K., unpublished data) (**Supplementary Fig. 33**). We note that some candidate promoters may represent distal enhancers, sharing promoter-associated marks potentially due to looping of enhancer to promoter regions[7].

Recovery of chromatin states using subsets of marks

As the large majority of chromatin states were defined by multiple marks, we next sought to specifically study the contribution of each mark in defining chromatin states. First, we found several notable examples of both additive relationships, such as acetylation marks in promoter regions, and combinatorial relationships, such as methylation marks associated with repressive and repetitive elements (**Supplementary Notes** and **Supplementary Fig. 38**). We also evaluated varying subsets of chromatin marks in their ability to distinguish between chromatin states (**Supplementary Notes** and **Supplementary Figs. 39–41**). More generally, we sought to provide guidelines for selecting subsets of chromatin marks to survey in new cell types that would be maximally informative.

As a proof of principle, we evaluated the recovery power of increasing numbers of marks in a greedy way, that is, selecting the best mark given all previous selected marks, weighing each state equally and penalizing mismatches uniformly (see Online Methods), which provided an initial unbiased recommendation of marks to survey for a new cell type (**Fig. 6**). We find that increasing subsets of marks rapidly converge to a fairly accurate annotation of chromatin states (**Fig. 6c**), providing cost-efficient recommendations for new cell types. In addition to an overall error score, this analysis provides information on the proportion of each state accurately recovered, and specific pairwise state misassignments. Such information could be incorporated in a modified scoring function to provide chromatin mark recommendations targeted to the subset of chromatin states that are of particular biological interest, or the particular state distinctions that are most important to each study.

DISCUSSION

The discovery and systematic characterization of chromatin states presented here reveals a diverse epigenomic landscape with 51 functionally distinct chromatin states. Although the exact number of chromatin states can vary based on the number of chromatin marks surveyed and the desired resolution at which state differences are studied, our results suggest that the genome annotation resulting from these states can extend the interpretable part of the human genome, especially outside protein-coding genes. The definition of the states themselves revealed numerous insights into the combinatorial and additive roles of chromatin marks, sometimes hinting at combinations of chromatin marks that were not previously described, and the genome-wide annotation of these states exposed many previously unannotated candidate functional elements.

We expect the usefulness of the methods presented here will increase as additional genome-wide epigenetic data sets become available, and as additional cell types are surveyed systematically. Chromatin states can be inferred with virtually any type of epigenetic and related information, including histone variants, DNA methylation, DNaseI hypersensitivity and binding of chromatin-associated and sequence-specific transcription factors. Although we focused on a single human cell type, the methods are generally applicable to any species and any number of cell types and even whole embryos, albeit in mixed cell populations mutually exclusive marks found in different subsets of cells could potentially be interpreted as co-occurring.

Specifically for understanding epigenomic dynamics, chromatin states can play a central role going forward, as they provide a uniform language for interpreting and comparing diverse epigenetic data sets, for selecting and prioritizing chromatin marks for additional cell types and for summarizing complex relationships of dozens of marks in directly-interpretable chromatin states. As several large-scale data production efforts are currently underway to map the epigenomes of many more cell types, exemplified by the ENCODE[34], modENCODE[35] and Epigenome Roadmap projects (http://www.roadmapepigenomics.org/), chromatin states will likely play a key role in the understanding of the human epigenome and its role in development, health and disease.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank P. Kheradpour for regulatory motif instances and M.F. Lin for predicted new exons. We thank M. Garber, A. Siepel, K. Lindblad-Toh, and E. Lander for use of comparative information on 29 mammals. We thank B. Bernstein, N. Shoresh, C. Epstein and T. Mikkelsen for helpful discussions. We thank L. Goff, C. Bristow, R. Sealfon and all members of the MIT CompBio Group for comments, feedback and support. This material is based upon work supported by the National Science Foundation under award no. 0905968 and funding from the US National Human Genome Research Institute (NHGRI) under awards U54-HG004570 and RC1-HG005334.

AUTHOR CONTRIBUTIONS

J.E. and M.K. developed the method, analyzed results and wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Bernstein, B.E., Meissner, A. & Lander, E.S. The mammalian epigenome. *Cell* **128**, 669–681 (2007).
2. Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693–705 (2007).
3. Strahl, B.D. & Allis, C.D. The language of covalent histone modifications. *Nature* **403**, 41–45 (2000).
4. Schreiber, S.L. & Bernstein, B.E. Signaling network model of chromatin. *Cell* **111**, 771–778 (2002).
5. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
6. Wang, Z. *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.* **40**, 897–903 (2008).

© 2010 Nature America, Inc. All rights reserved.

7. Heintzman, N.D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
8. Heintzman, N.D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
9. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
10. Hon, G., Wang, W. & Ren, B. Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Comput. Biol.* **5**, e1000566 (2009).
11. Wang, X., Xuan, Z., Zhao, X., Li, Y. & Zhang, M.Q. High-resolution human core-promoter prediction with CoreBoost_HM. *Genome Res.* **19**, 266–275 (2009).
12. Won, K.J., Chepelev, I., Ren, B. & Wang, W. Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC Bioinformatics* **9**, 547 (2008).
13. Hon, G., Ren, B. & Wang, W. ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. *PLOS Comput. Biol.* **4**, e1000201 (2008).
14. Day, N., Hemmaplardh, A., Thurman, R.E., Stamatoyannopoulos, J.A. & Noble, W.S. Unsupervised segmentation of continuous genomic data. *Bioinformatics* **23**, 1424–1426 (2007).
15. Jia, L. *et al.* Functional enhancers at the gene-poor 8q24 cancer-linked locus. *PLoS Genet.* **5**, e1000597 (2009).
16. Thurman, R.E., Day, N., Noble, W.S. & Stamatoyannopoulos, J.A. Identification of higher-order functional domains in the human ENCODE regions. *Genome Res.* **17**, 917 (2007).
17. Schuettengruber, B. *et al.* Functional anatomy of polycomb and trithorax chromatin landscapes in *Drosophila* embryos. *PLoS Biol.* **7**, e13 (2009).
18. Jaschek, R. & Tanay, A. Spatial clustering of multivariate genomic and epigenomic information. in *Proceedings of the 13th Annual International Conference on Research in Computational Molecular Biology* (ed. Batzoglou, S.) 170–183 (Springer, 2009).
19. Schwartz, S., Meshorer, E. & Ast, G. Chromatin organization marks exon-intron structure. *Nat. Struct. Mol. Biol.* **16**, 990–995 (2009).
20. Kolasinska-Zwierz, P. *et al.* Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat. Genet.* **41**, 376–381 (2009).
21. Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C. & Komorowski, J. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Res.* **19**, 1732–1741 (2009).
22. Schones, D.E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell.* **132**, 878–898 (2008).
23. Sripathy, S.P., Stevens, J. & Schultz, D.C. The KAP1 corepressor functions to coordinate the assembly of de novo HP1-demarcated microenvironments of heterochromatin required for KRAB zinc finger protein-mediated transcriptional repression. *Mol. Cell. Biol.* **26**, 8623–8638 (2006).
24. O'Geen, H. *et al.* Genome-wide analysis of KAP1 binding suggests autoregulation of KRAB-ZNFs. *PLoS Genet.* **3**, e89 (2007).
25. Hindorff, L.A., Junkins, H.A., Mehta, J.P. & Manolio, T.A. A catalog of published genome-wide association studies. <http://www.genome.gov/gwastudies> accessed July 22, 2009.
26. Gudbjartsson, D.F. *et al.* Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction. *Nat. Genet.* **41**, 342–347 (2009).
27. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948–951 (2008).
28. Furey, T.S. & Haussler, D. Integration of the cytogenetic map with the draft human genome sequence. *Hum. Mol. Genet.* **12**, 1037–1044 (2003).
29. Wang, Z. *et al.* Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell* **138**, 1019–1031 (2009).
30. Johnson, D.S. *et al.* Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome Res.* **18**, 393–403 (2008).
31. Zang, C. *et al.* A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **25**, 1952–1958 (2009).
32. Zhang, Y., Shin, H., Song, J.S., Lei, Y. & Liu, X.S. Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq. *BMC Genomics* **9**, 537 (2008).
33. Cui, K. *et al.* Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell* **4**, 80–93 (2009).
34. ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
35. Celniker, S.E. *et al.* Unlocking the secrets of the genome. *Nature* **459**, 927–930 (2009).
36. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38**, 626–635 (2006).
37. Karolchik, D. *et al.* The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.* **36**, D773–D779 (2008).
38. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Wheeler, D.L. GenBank: update. *Nucleic Acids Res.* **32**, D23–D26 (2004).

© 2010 Nature America, Inc. All rights reserved.

ONLINE METHODS

Input data for modeling. The initial unprocessed data were bed files containing the genomic coordinates and strand orientation of mapped sequence reads from ChIP-seq experiments[5,6]. There was a separate bed file for each of the 18 acetylations, 20 methylations, H2AZ, CTCF and PolII in CD4 T cells. We used the updated version of the H3K79me1/2/3 data, as reported[6], which differs from the version first reported[5].

To apply the model we first divided the genome into 200-base-pair nonoverlapping intervals within which we independently made a call as to whether each of the 41 marks was detected as being present or not based on the count of tags mapping to the interval. Each tag was uniquely assigned to one interval based on the location of the 5′ end of the tag after applying a shift of 100 bases in the 5′ to 3′ direction of the tag. The threshold, t, for each mark was based on the total number of mapped reads for the mark (**Supplementary Table 2**), and was set to be the smallest integer t such that $P(X>t)<10^{-4}$ where X is a random variable with a Poisson distribution with mean parameter set to the empirical mean of the number of tags per interval.

The probabilistic model. The probabilistic model is based on a multivariate instance of a Hidden Markov Model (HMM)[39]. The model assumes a fixed number of hidden states K. In each hidden state, the emission distribution, that is the probability distribution over each combination of marks, is modeled with a product of independent Bernoulli random variables. Formally, for each of the K states, and $M = 41$ input marks, there is an emission parameter $p_{k,m}$ denoting the probability in state k ($k = 1,\ldots,K$) that input mark m ($m = 1,\ldots,M$) has a present call. Let $c \in C$ denote a chromosome where C is the set of all chromosomes. Let c_t denote an interval on chromosome c where $t = 1,\ldots,T_c$ corresponds sequentially to the 200 bp intervals on chromosome c. c_1 is the interval corresponding to base pairs 1–200 on chromosome c and T_c is the number of nonoverlapping 200 bp intervals on chromosome c. Let $v_{c_t,m}$ be '1' if there is a present call for input mark m and '0' otherwise at location c_t. Denote the specific combination of marks at interval c_t as $v_{c_t} = (v_{c_t,1},\ldots,v_{c_t,m})$. Let b_{ij} denote the probability of transitioning from state i to j where $i = 1,\ldots,K$ and $j = 1,\ldots,K$. We also have parameters a_i ($i = 1,\ldots,K$), which denote the probability that the state of the first interval on the chromosome is i. Let $s_c \in S_C$ be an unobserved state sequence through chromosome c and S_C be the set of all possible state sequences. Let s_{c_t} denote the unobserved state on chromosome c at location t for state sequence s_c. The full likelihood of all of the observed data v for the parameters a, b and p can then be expressed as:

$$P(v|a,b,p) = \prod_{c\in C}\sum_{s_c\in S_c} a_{s_{c_1}}\left(\prod_{t=2}^{T_C} b_{s_{c_{t-1}},s_{c_t}}\right)\prod_{t=1}^{T_C}\prod_{m=1}^{M} p_{s_{c_t},m}^{v_{c_t,m}}(1-p_{s_{c_t},m})^{(1-v_{c_t,m})}$$

Model learning. We first used an iterative learning expectation-maximization approach to infer state emission and transition parameters that best summarize the observed genome-wide chromatin mark information using a fixed number of randomly-initialized hidden states, varying from 2 up to 80 states. To minimize the number of states and facilitate recovery of a robust and comparable set of states across models of varying complexity, we then used a nested initialization procedure that seeded parameters of lower-complexity models with states of higher-complexity models.

From an initial set of parameters we found a local optimum of the parameter values using a variant of the standard expectation-maximization based Baum-Welch algorithm for training HMMs[39]. Our variant after the first full iteration over all the chromosomes, used an incremental expectation-maximization procedure[40], which would update the parameters through a maximization step after performing an expectation over any chromosome. This allowed improved parameter estimates from the maximization step to be more quickly incorporated in the more time consuming expectation step. Also for computational considerations, if a transition parameter fell below 10^{-10} during training we set the parameter value to 0, which allowed faster training with virtually no impact on the final model learned. The transitions were initialized to be fully connected, and except for the 10^{-10} criterion there was no regularization forcing them closer to 0. We would terminate the training after 300 passes over all the chromosomes, which was sufficient for the likelihood to demonstrate convergence (**Supplementary Fig. 8**).

The procedure for determining the initial parameters used to learn the final set of HMMs was to first learn in parallel for every number of states from 2 to 80 three HMM models based on three different random initializations of the parameters. Each model was scored based on the log likelihood of the model minus a penalization on the model complexity determined by the Bayesian Information Criterion (BIC) of one-half the number of parameters times the natural log of the number of intervals. We then selected the model with the best BIC score among these 237 models, which had 79 states (**Supplementary Figs. 8** and **12**). We then iteratively removed states from this 79-state model. When removing a state the emission probabilities would be removed entirely, and any state that transitioned to it would have that transition probability uniformly redistributed to all the remaining states. This resulting set of models was then used as the initial parameters of the HMMs in the final model learning. During this final model learning, one HMM was learned for every number of states between 2 and 79 in parallel (**Supplementary Fig. 13**).

The criterion for selecting a state to remove from a model was based on first forming a set E containing all the emission vectors from all the 237 models learned from the random initializations. The procedure would then remove a state such that the elements in E had in total the least distance from their closest emission vector among the remaining states. Formally for a set of emission vectors C_n corresponding to states in a model the method would form a set C_{n-1} and corresponding model by removing r defined by

$$\underset{r}{\operatorname{argmin}} \sum_{e\in E} \min_{c\in C_n\setminus r} d(e,c)$$

where here we used $(1 - \rho)$ where ρ is the standard correlation coefficient as the distance d, though the method is general and could be used with other distance measures.

The entire procedure discovered models with comparable or superior likelihood scores to randomly initialized models, while also having sets of parameters that would be more directly comparable (**Supplementary Figs. 8** and **13**). The number of states for a model to analyze can then be selected by choosing the model trained from a nested initialization with the smallest number of states that sufficiently captures all states of offering distinct biological interpretations, which in our case was a 51-state model based on the recovery of the end of transcription state (State 27) (**Supplementary Notes**).

Associating genomic locations with states. After a model is learned, a posterior probability distribution over the state of each interval is computed using a forward-backward algorithm[39]. Unless otherwise noted, the analysis was based on the 'soft' state assignments of the posterior distribution. We also formed hard assignments of states to locations by using the maximum-posterior state assignment at a location. Both the full posterior and hard assignments are available on the website http://compbio.mit.edu/ChromatinStates/.

External fold enrichments and percentage overlaps. For a state the sum of the posterior probability over all 200 bp intervals was computed, denoted by a. For an external data source the total number of 200 bp intervals that it intersects in at least one base was computed, denoted by b. For the state and the external data source the total sum of the posterior for the state in intervals intersecting the external data source were computed, denoted by c. Also the total number of 200 bp intervals is denoted by d. The percentage of a state's overlap with an external data source is defined as $(c/a*100)$ whereas the fold enrichment is $(c/a)/(b/d)$. P-values of the overlap were computed based on the hypergeometric distribution.

The gene annotations used were the RefSeq annotations[41] as of December 14, 2008 obtained from the UCSC genome browser browser[37] and are based on hg18.

© 2010 Nature America, Inc. All rights reserved.

doi:10.1038/nbt.1662

The sequence data for computed nucleotide frequencies, CpG islands, repeats[42] and conservation data were also obtained from the UCSC genome browser. The conservation data were based on PhastCon conserved elements using the 44-way vertebrate alignment[43,44] (Lindblad-Toh, K. *et al.*, Broad Institute, unpublished data). Transcription factor binding enrichments were computed for 18 experiments from numerous publications (**Supplementary Fig. 23**), the median enrichment over all these experiments is reported in **Figure 2b**. The DNaseI hypersensitivity data was as described[45] obtained from the UCSC genome browser. The nuclear lamina data of human fibroblasts was obtained from ref. 27. The zinc-finger genes were defined as those that had 'ZNF' at the beginning of the gene symbol in the RefSeq gene table. For published coordinates that were in hg17 we converted them to hg18 using the liftover tool from the UCSC genome browser[46].

Expression, motif and gene ontology analyses. We obtained the processed CD4 T expression data from ref. 47 for both replicates. We then averaged the two replicates. After averaging the two replicates we performed a natural log transform of the average values. We then standardized all values by subtracting the mean log transformed value and then dividing by the s.d. of the log transform values. The genome coordinates of each probe set were obtained from the UCSC genome browser. Each 200 bp interval that overlapped a probe set obtained the transformed expression score. If multiple probe sets overlapped the same 200 bp then the average of the expression values associated with these were taken.

We generated transcription factor motif enrichments as described[48], extended for position-weight matrices (PWMs) (Kheradpour, P., MIT, and M.K., unpublished data) based on the hard state assignments.

Gene ontology enrichments were based on the hard state assignment of the interval containing the RefSeq annotated TSS of the gene. Enrichments were computed using the STEM software (v.1.3.4) and the Bonferroni corrected *P*-values are reported[49].

SNP and GWAS analysis. The HapMap CEU[50] data were downloaded from the UCSC genome browser. Significant GWAS hits were taken from ref. 25. SNPs listed as occurring multiple times were only counted once, and for the SNP set listed as a 17-marker haplotype only the first SNP was used giving 1,640 SNPs. In computing enrichment for HapMap and GWAS SNPs, if two SNPs mapped to the same interval, we counted them multiple times. To determine if the number of GWAS SNPs in a chromatin state was more significant than would be expected based on the general SNP frequency in the state we used a binomial distribution where n = 1,640 and p is the proportion of HapMap CEU SNPs assigned to the state. We applied a Bonferroni correction for testing multiple states and only reported those *P*-values significantly enriched with $P < 0.01$.

RefSeq TSS and gene transcripts discovery. The ROC curve for the CAGE data was based on the number of CAGE tags mapping to a 200 bp interval retrieved from the Fantom database and converted from hg17 to hg18 using the UCSC genome browser liftover tool[36]. The overlap with EST was based on those EST listed in the UCSC genome browser all_est table as of November, 29, 2009 (refs. 37,38). The overlap with GenBank mRNA is based on the overlap with the UCSC genome browser mRNA listed in the table as of October 31, 2009 (refs. 37,38). The novel exon predictions are from (Lin, M.F., MIT, and M.K., unpublished data).

Mark subset evaluation and selection. When evaluating the coverage of a specified subset of marks, first a posterior distribution over the states at each interval is computed using the model learned on the full set of marks, except that the marks not in the subset are omitted when computing emission probabilities. For an interval t we define here $s_{t,k}$ and $f_{t,k}$ to be the posterior assignment to state k at interval t based on the subset and full set of marks, respectively. The proportion of state k recovered with a subset of marks is defined as:

$$c_k = \frac{\sum_t \min(f_{t,k}, s_{t,k})}{\sum_t f_{t,k}}$$

where the sum is over all intervals t in the genome. The ordering of marks presented without any prior biological knowledge was based on a greedy forward selection algorithm designed to select marks that would minimize this function:

$$\sum_k (1 - c_k)^2$$

where the sum is over all states. At each step the algorithm would then choose the one additional mark, conditioned on all the other previously selected marks that would cause this function to be minimized. We note that this target function considers all nonidentical state assignments to have equal loss. An extension of this approach would be to apply target functions that weigh different misassignments differently. The proportion of state k with the full set of marks that is misassigned to state i using a subset of marks, $m_{k,i}$, as is presented in **Supplementary Figures 39** and **40**, is defined as:

$$m_{k,i} = \frac{\sum_t \left(\max(f_{t,k} - s_{t,k}, 0) \left(\frac{\max(s_{t,i} - f_{t,i}, 0)}{\sum_j \max(s_{t,j} - f_{t,j}, 0)} \right) \right)}{\sum_t f_{t,k}}$$

The first term in the sum in the numerator represents for an interval t the amount of posterior probability assigned to state k using the full set of marks not assigned using the subset of marks. The second term represents the portion of this posterior probability that will be credited to state i. The portion credited to state i is the proportion of the surplus posterior state i received with the subset of marks in the interval relative to the total surplus posterior all states received in the interval.

39. Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. *Biological Sequence Analysis* (Cambridge Univ. Press, 1998).
40. Neal, R.M. & Hinton, G.E. A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learn. Graph. Models* **89**, 355–368 (1998).
41. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, D61–D65 (2007).
42. Smit, A., Hubley, R. & Green, P. *RepeatMasker Open-3.0* 1996-2010 <http://www.repeatmasker.org>.
43. Miller, W. *et al.* 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res.* **17**, 1797–1808 (2007).
44. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
45. Boyle, A.P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).
46. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
47. Su, A.I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* **101**, 6062–6067 (2004).
48. Kheradpour, P., Stark, A., Roy, S. & Kellis, M. Reliable prediction of regulator targets using 12 *Drosophila* genomes. *Genome Res.* **17**, 1919–1931 (2007).
49. Ernst, J. & Bar-Joseph, Z. STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics* **7**, 191 (2006).
50. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).

© 2010 Nature America, Inc. All rights reserved.

Anal Bioanal Chem
DOI 10.1007/s00216-010-4093-x

ORIGINAL PAPER

Optimization of parameters for coverage of low molecular weight proteins

Stephan A. Müller · Tibor Kohajda · Sven Findeiß · Peter F. Stadler · Stefan Washietl · Manolis Kellis · Martin von Bergen · Stefan Kalkhof

Received: 3 June 2010 / Revised: 26 July 2010 / Accepted: 3 August 2010
© The Author(s) 2010. This article is published with open access at Springerlink.com

Abstract Proteins with molecular weights of <25 kDa are involved in major biological processes such as ribosome formation, stress adaption (e.g., temperature reduction) and cell cycle control. Despite their importance, the coverage of smaller proteins in standard proteome studies is rather sparse. Here we investigated biochemical and mass spectrometric parameters that influence coverage and validity of identification. The underrepresentation of low molecular weight (LMW) proteins may be attributed to the low numbers of proteolytic peptides formed by tryptic digestion as well as their tendency to be lost in protein separation and concentration/desalting procedures. In a systematic investigation of the LMW proteome of *Escherichia coli*, a total of 455 LMW proteins (27% of the 1672 listed in the SwissProt protein database) were identified, corresponding to a coverage of 62% of the known cytosolic LMW proteins. Of these proteins, 93 had not yet been functionally classified, and five had not previously been confirmed at the protein level. In this study, the influences of protein extraction (either urea or TFA), proteolytic digestion

Published in the special issue *Mass Spectrometry (DGMS 2010)* with Guest Editors Andrea Sinz and Jürgen Schmidt.

Electronic supplementary material The online version of this article (doi:10.1007/s00216-010-4093-x) contains supplementary material, which is available to authorized users.

S. A. Müller · M. von Bergen · S. Kalkhof (✉)
Department of Proteomics, UFZ,
Helmholtz-Centre for Environmental Research,
Permoserstraße 15,
04318 Leipzig, Germany
e-mail: Stefan.kalkhof@ufz.de

T. Kohajda · M. von Bergen
Department of Metabolomics, UFZ,
Helmholtz-Centre for Environmental Research,
Permoserstraße 15,
04318 Leipzig, Germany

S. Findeiß · P. F. Stadler
Bioinformatics Group, Department of Computer Science, and
Interdisciplinary Center for Bioinformatics, University of Leipzig,
Härtelstraße 16-18,
04107 Leipzig, Germany

P. F. Stadler
Institute for Theoretical Chemistry, University of Vienna,
Währingerstraße 17,
1090 Wien, Austria

P. F. Stadler
Max Planck Institute for Mathematics in the Sciences,
Inselstrasse 22,
04103 Leipzig, Germany

P. F. Stadler
Nomics Group,
Fraunhofer Institute for Cell Therapy and Immunology,
Deutscher Platz 5e,
04103 Leipzig, Germany

P. F. Stadler
Santa Fe Institute,
1399 Hyde Park Rd,
Santa Fe, NM 87501, USA

S. Washietl · M. Kellis
Computer Science and Artificial Intelligence Laboratory,
Broad Institute, Massachusetts Institute of Technology,
Cambridge, MA 02139, USA

Published online: 28 August 2010

(solely, and the combined usage of trypsin and AspN as endoproteases) and protein separation (gel- or non-gel-based) were investigated. Compared to the standard procedure based solely on the use of urea lysis buffer, in-gel separation and tryptic digestion, the complementary use of TFA for extraction or endoprotease AspN for proteolysis permits the identification of an extra 72 (32%) and 51 proteins (23%), respectively. Regarding mass spectrometry analysis with an LTQ Orbitrap mass spectrometer, collision-induced fragmentation (CID and HCD) and electron transfer dissociation using the linear ion trap (IT) or the Orbitrap as the analyzer were compared. IT-CID was found to yield the best identification rate, whereas IT-ETD provided almost comparable results in terms of LMW proteome coverage. The high overlap between the proteins identified with IT-CID and IT-ETD allowed the validation of 75% of the identified proteins using this orthogonal fragmentation technique. Furthermore, a new approach to evaluating and improving the completeness of protein databases that utilizes the program *RNAcode* was introduced and examined.

Keywords LTQ Orbitrap · Nano-HPLC · Nano-ESI-MS · MS · Proteomics · Low molecular weight proteome · *Escherichia coli*

Abbreviations

LMW	Low molecular weight (below 25 kDa)
CID	Collision-induced dissociation
ET(ca)D	Electron transfer (collision activation) dissociation
FDR	False discovery rate
FTICR MS	Fourier transform ion cyclotron resonance mass spectrometry
GO	Gene Ontology
HCD	Beam-type collision-activated dissociation
LB medium	Lysogeny broth medium
ORF	Open reading frame

Introduction

Escherichia coli (*E. coli*) is a Gram-negative bacterium of the family *Enterobacteriacae*. It is relatively easy to cultivate, fast growing, and allows for feasible genetic manipulation. Due to these characteristics, *E. coli* is omnipresent in molecular biology, biotechnology and gene technology, and it is one of the most intensively studied and best-characterized prokaryotes. Sequencing and analysis of the 4.6 Mb chromosome of the laboratory strain *E. coli* K12 coding for 4411 protein-coding genes was completed in 1997 [1].

In the last two decades, the *E. coli* proteome has been extensively analyzed by 2D gel electrophoresis (2D-GE) initially and then via LC/MS approaches. Besides investigations of numerous biological questions, the *E. coli* proteome has also been used to validate new technologies and methodologies, including sample prefractionation, protein enrichment and separation by 2D-GE or *n*-dimensional chromatography, and protein identification and quantification by MS [2].

The first proteome study was conducted using 2D-GE and resulted in the identification of 381 proteins [3]. By combining 2D-DIGE with biochemical prefractionation and the analysis of stationary and exponential growth phases, it was possible to detect and quantify 3199 protein species, among which 575 unique proteins could be identified [4]. In several gel-free approaches using *n*-dimensional LC for protein [5] or peptide separation [6–9], the number of proteins was successively increased further (Table 1). Most recently, in 2010, Iwasaki and coworkers used 1D-LC/MS/MS with a 350 cm long monolithic silica–C_{18} capillary column and 41 h of LC gradient time to identify 2602 proteins [10]. However, even with all of these different methods, the identification rate for LMW proteins of <25 kDa listed in the SwissProt protein database is usually below 25%, and is significantly lower than the average identification rate (Table 1).

Proteins that are essential in numerous biological functions, especially ribosome formation (e.g., 18 30S ribosomal protein subunits, 34 50S ribosomal protein subunits), transcription regulation, and stress response (cold shock proteins, universal stress proteins) are of LMW. Coverage of those functional proteins in proteomic studies is of great interest in systems biology in order to gain an in-depth understanding of the reactions of bacteria to external stresses [11], adaption to different substrates, and interdependencies in microbial bacterial communities in the new field of metaproteomics [12]. Furthermore, over 500 LMW proteins of *E. coli* are still classified as "functionally uncharacterized" according to the latest GO annotation database [13]. This number is astonishingly high given the limited genome of *E. coli* and the high feasibility of this organism for culturing and genomic manipulation.

Another challenge is the de novo annotation of open reading frames (ORF) coding for small proteins on a genome-wide scale. In the past, computational gene-finding approaches excluded short ORFs with less than 40 or 50 amino acids. For such short ORFs, typical statistical signals in the sequence (ORF length and codon usage) are very weak, resulting in a high false-discovery rate (FDR). Thus, using standard methods with less stringent filters leads to the prediction of thousands of small ORFs, most of which are not likely to be translated [14]. The methods of choice to verify the existence of these small proteins are LC/MS

Table 1 Summary of total and LMW proteins detected in previous studies based on at least [a] four peptides, [b] two peptides, and [c]one peptide per protein

Study	Method	LMW	Complete proteome	LMW (%)	Complete proteome (%)	Reference
Lopez-Campistrous et al. (2005)	2D-PAGE after prefractionation in periplasm, inner membrane, and outer membrane	164	575	10	13	[4]
Geveart et al. (2002)	Diagonal 2D-LC-MS of methionine-containing peptides	187[c]	872[c]	11	20	[6]
Corbin et al. (2003)	1D-LC-MS with and w/o membrane fractionation (4 h per run)	218[a]– 331[c]	404[a]–1147[b]	13[a]-21[c]	26	[7]
Taoka et al. (2004)	2D-LC-MS (16 h per run)	401	1480	24	34	[8]
Ishihama et al. (2008)	More than 200 2D-LC-MS measurements after 1D-gel protein prefractionation	341	1103	20	25	[9]
Iwasaki et al. (2010)	LC-MS with a 3.5 m non-commercially available monolithic column (41 h per run)	737[b]–820[c]	2404[b]–2602[c]	44[b]-49[c]	60	[10]

approaches. Since these experimental methods are cost and time intensive, in silico methods are still required for efficient genome annotation. Recently, we developed *RNAcode*, a gene prediction program that uses the principle of comparative genomics [15] to detect protein-coding genes in multiple genome alignments [16]. Since *RNAcode* is based on evolutionary signatures, it can detect statistically significant signals—even in short ORFs—as long as sufficient phylogenetic information from related sequences is available. The fact that *RNAcode* is not based on the detection of complete ORFs also makes it applicable to incomplete data, such as fragments of transcriptome studies [17]. Thus, *RNAcode* fills a specific gap in the current repertoire of protein annotation software. To further investigate the applicability and power of RNAcode, we systematically analyzed the LMW of *E. coli* and compared these results with our proteome data.

The variation in the abundances of cytosolic proteins in *E. coli* ranges from less than 200 to more than 10^8 molecules per cell—in other words, more than six orders of magnitude [9]. The low abundances of some proteins certainly hamper their detection, and not all proteins will be expressed at the same time. Aside from these biological reasons for limited coverage, it has been discussed that losses during protein extraction [18], separation and purification [19], as well as the low number of detectable proteotypic peptides formed by proteolysis [19] are responsible for the low identification rate. Taking into account recent improvements in the coverage of LMW proteins, the best study achieved 49% coverage of LMW in *E. coli* (Table 1). It is obvious that there is plenty of scope for improvement. This can in principle be achieved by separation, fractionation or the complementary usage of multiple proteases, or on the LC/MS side. In order to get information on which strategy to start with in this study, key parameters associated with both prefractionation and LC/MS were tested. With respect to prefractionation and biochemical preprocessing, the following parameters were assessed for their influence on coverage: (i) protein extraction buffers, (ii) enrichment and separation, and (iii) enzymatic proteolysis. In terms of LC/MS, the crucial steps of (iv) the fragmentation procedure and (v) MS/MS data analysis were varied and evaluated with respect to identification rate, average sequence coverage, and validation of identifications.

Materials and methods

Cell culture

Cell lysates of *E. coli* strain K12 were analyzed to assess critical parameters for LMW proteome analysis. Analyses were performed in two (gel-based approach) and three (non-gel-based approach) independent biological replicates. Cells were grown in LB medium to stationary phase. Therefore, 1 l of fresh medium was inoculated with 100 ml of a preparatory culture grown under the same conditions. Cells were collected by centrifugation (10 min, 8,000×*g*, 4 °C).

Protein extraction and small protein enrichment

Cell pellets were resuspended in either urea lysis buffer (40 ml, 8 M urea, 10 mM DTT, 1 M NaCl, 10 mM Tris/HCl, pH 8.0) [20] or acidic lysis buffer (40 ml, 0.1% TFA) [21]. Cell disruption was performed by ultrasonification (5 min, 50% duty cycle, Branson Sonifier 250, Emerson, St. Louis, MO, USA). Undissolved material was removed by centrifugation (15 min, 10,000×*g*, 4 °C). High molecular weight proteins were depleted by centrifugation through a filter membrane (molecular weight cut-off: 50 kDa, Pall

Macrosep 50 K, Pall Life Science, Ann Arbor, MI, USA) [22]. The permeate was split into aliquots of 1.2 ml. TFA lysates were equilibrated to neutral pH with NH_4CO_3 (final concentration: 250 mM) and protein disulfide bonds were reduced by adding DTT (final concentration: 10 mM). Cysteines were alkylated by the addition of 2-iodoacetamide (final concentration: 51.5 mM) to both lysates and incubation for 45 min at room temperature in the dark. Proteins were desalted and concentrated by TCA precipitation (final concentration: 20% (w/v), incubation at 4 °C for 16 h, centrifugation at 20,000×*g* for 20 min).

Protein separation and protein digestion

For the non-gel approach, one protein pellet of every biological replicate was dissolved in 500 mM NH_4HCO_3 and the protein concentration was measured with a Bradford assay (Bradford Quick Start, Bio-Rad, Hercules, CA, USA) using bovine serum albumin for calibration. Pellets were redissolved in 100 μl 1.6 M urea in NH_4HCO_3 (100 mM). Trypsin (modified porcine trypsin, Sigma–Aldrich, Steinheim, Germany) was dissolved in 50 mM NH_4HCO_3 containing 10% acetonitrile to a concentration of 125 ng/μl. Trypsin solution was added to the dissolved protein pellets with a molecular weight ratio of 1:50 (trypsin:protein). Digestions were performed overnight at 37 °C and stopped by adding formic acid (final concentration: 4%). Digestion solutions were concentrated to 20 μL using vacuum centrifugation and reconstituted by adding 40 μL 1% formic acid.

For the gel separation, protein pellets were redissolved with SDS loading buffer (2% (w/v) SDS, 12% (w/v) glycerol, 120 mM DTT, 0.0024% (w/v) bromophenol blue, 70 mM Tris/HCl) and adjusted to neutral pH by adding 10× cathode buffer solution (1 M Tris, 1 M tricine, 1% (w/v) SDS, pH 8.25). GE was performed according to a modified protocol of Schaegger [23]. In brief, a 20% T, 6% C separation gel was used in combination with a 4% T, 3% C stacking gel. A prestained LMW protein standard (molecular weight range 1.7–42 kDa, multicolor low-range protein ladder, Fermentas, St. Leon-Rot, Germany) was applied as a molecular weight marker. For each experiment, three lanes were loaded with the LMW protein extract, among which one was stained with colloidal Coomassie. Nine gel slices from each of the two unstained lanes were excised in the molecular weight range 1–25 kDa and used for in-gel digestion.

The gel slices were washed twice with water for 10 min and once with NH_4HCO_3 (10 mM). In-gel digestion was performed by adding modified porcine trypsin (100 ng, Sigma–Aldrich) or endoproteinase AspN (100 ng, Sigma–Aldrich) in NH_4HCO_3 (10 mM, 30 μl volume) to the slices.

The digestions were performed overnight at 37 °C and stopped afterwards by adding formic acid (final concentration: 4%). The supernatant and the two gel elution solutions (first elution step: 40% (v/v) acetonitrile; second elution step: 80% (v/v)) were collected and mixed. The combined mixtures were dried using vacuum centrifugation. Peptides were reconstituted in 0.1% formic acid.

Analysis with nano-HPLC/nano-ESI-LTQ Orbitrap MS

LC/MS/MS analysis was performed on a nano-HPLC system (nanoAcquity, Waters, Milford, MA, USA) coupled to an LTQ Orbitrap mass spectrometer. Chromatography was conducted with 0.1% formic acid in solvents A (100% water) and B (100% acetonitrile).

In-solution digestion samples were injected by the autosampler and concentrated on a trapping column (nanoAcquity UPLC column, C18, 180 μm×2 cm, 5 μm, Waters) with water containing 0.1% formic acid at flow rates of 15 μL/min. After 10 min, peptides were eluted onto a separation column (nanoAcquity UPLC column, C18, 75 μm×150 mm, 1.7 μm, Waters). Peptides were eluted over 150 min with a 2–40% solvent B gradient (0 min, 2%; 3 min 2%;10 min, 6%;100 min, 20%; 150 min, 40%).

Scanning of eluted peptide ions was carried out in positive ion mode between *m/z* 300 and 1500, automatically switching to MS/MS mode for ions exceeding an intensity of 3,000. Precursor ions were dynamically excluded for MS/MS measurements for 3 min. Six runs with different MS/MS measurements were performed per biological sample. CID and ETD fragmentations were carried out with ion detection in the ion trap or the Orbitrap in separate runs. HCD fragmentations were detected in the Orbitrap. Additionally, a method with a decision tree between CID and ETD in the ion trap was performed.

In-gel digestion samples were injected and concentrated on a trapping column in an identical manner to the analysis of in-solution digestions. Peptides were eluted onto a separation column (nanoAcquity UPLC column, C18, 75 μm×250 mm, 1.7 μm, Waters) and separation was done over 30 min with a 2–40% solvent B gradient (0 min, 2%; 2 min 8%; 20 min, 20%; 30 min, 40%). Scanning of eluted peptide ions was carried out in positive ion mode in the range *m/z* 350–2000, automatically switching to CID-MS/MS mode for ions exceeding an intensity of 2,000. For CID-MS/MS measurements, a dynamic precursor exclusion of 3 min was applied.

Data analysis

Database searching was performed with *Proteome Discoverer* (version 1.0; Thermo Fisher Scientific, San Jose, CA, USA) using the MASCOT (version 2.2; Matrix Science,

London, UK) and SEQUEST (version 1.0.43.0; Thermo Fisher Scientific) algorithms that search through a target and decoy database containing all proteins of *E. coli* strain K12 in the SwissProt protein database. In-gel digestions with trypsin were searched with maximum of one missed cleavage, while two missed cleavages were allowed for in-gel digestion with AspN and in-solution digestions. For trypsin C-terminal cleavage to arginine and lysine, and for endoprotease AspN N-terminal cleavage to aspartic and glutamic acid were considered. MS/MS spectra were grouped with a precursor mass tolerance of 4.0 ppm and a retention time tolerance of 5 min. MASCOT and SEQUEST searched with a parent ion tolerance of 5.0 ppm. Fragment ion mass tolerances were specified as 0.5 Da when fragment ions were detected in the ion trap and 0.05 Da when detection was performed in the Orbitrap. Carbamidomethylation of cysteines was specified in MASCOT and SEQUEST as a fixed modification, and the oxidation of methionine as a variable modification. Additionally, deamidations of asparagine and glutamine were considered variable modifications for in-solution digestion samples.

SCAFFOLD (version SCAFFOLD_2_06_01_pre3; Proteome Software Inc., Portland, OR, USA) was used to validate MS/MS-based peptide and protein identifications. Peptide and protein identification parameters were adjusted to a false-positive rate of lower than 5% using the target and decoy database. False-positive rates were calculated as described by Elias et al. [24]. Peptide identifications were accepted if they could be established at a probability of greater than 70.0% as specified by the *Peptide Prophet* algorithm [25]. Peptide identifications were accepted by exceeding specific database search engine thresholds. MASCOT identifications required ion scores of greater than 10.0. SEQUEST identifications required deltaCn scores of greater than 0.10 and XCorr scores of greater than 1.7, 2.0, and 2.3 for doubly, triply and quadruply charged peptides. Protein identifications were accepted if they could be established at greater than 95.0% probability and contained at least two identified peptides. Protein probabilities were assigned by the *Protein Prophet* algorithm [26]. Proteins that contained similar peptides and which could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principles of parsimony. GO annotations were obtained with STRAP [27] from the EBI GO database (http://www.ebi.ac.uk/GOA/, version 05/07/2010).

ProtStat: protein statistics and peptide predictions

The software *ProtStat* is an in-house tool programmed with C# which calculates protein as well as proteolytic peptide properties. The program has three different modes: protein pre-statistics, protein post-statistics and peptide statistics.

For the protein statistics, various data can be obtained for every protein, including molecular weight, protein sequence, GRAVY score, protein database ID, protein description, and a calculation of the p*I* value. p*I* values are calculated using the advanced algorithm suggested by Kozlowski (http://isoelectric.ovh.org/) with a selectable set of amino acid p*K* increments according to EMBOSS, DTASelect, Solomon, Sillero or Rodwell.

The protein pre-statistic allows an in silico simulation of a proteolytic digestion by calculating the number and sequences of proteolytic peptides, the expected possible sequence coverage, and performing a comparison in terms of unique peptides and sequence coverage to other proteolytic digestions (e.g., those using other proteases). In terms of digestion parameters, several specific proteases as well as their combinations and fixed modifications are allowed.

In the protein post-processing mode, the same analysis is possible for a list of identified proteins, and this enables the comparison of experimental and theoretical LC/MS measurements.

The peptide statistics mode allows the calculation of inclusion or exclusion lists based on the results of a theoretical or experimental proteolytic digestion. Therefore, exact *m/z* values in a given *m/z* range were calculated for the charge states 1+ to 4+. Again, fixed protein modifications are taken into account. Additionally, p*I* values of all potential proteolytic peptides for every protein inside a protein FASTA database are calculated.

Prediction of protein coding regions in genome-wide alignments of nucleotide sequences by *RNAcode*

We used the Multiz pipeline [28] to align 54 fully sequenced enterobacteria species from GenBank (Electronic supplementary material Table S1). The alignments were screened using the default parameters of *RNAcode* (software available at http://wash.github.com/rnacode) and a *p*-value cutoff of 0.05. This resulted in 20,528 high-scoring coding segments. Multiple sequence alignments of such a high number of species tend to be fragmented into relatively small blocks. Therefore, high-scoring coding segments in the same reading frame and less than 15 nucleotides apart were combined. This reduced the number of high-scoring coding segments to 6,542.

The SwissProt protein database was downloaded (http://pir.uniprot.org/downloads, May 2010 release). For each registered *E. coli* protein, the ID, the type of evidence, and the amino acid sequence was extracted. In order to compare the *RNAcode* predictions, which are based on nucleotide alignments, with the protein sequences from SwissProt and

our peptide data, we blasted all peptide sequences (TBLASTN, *E*-value 10^{-3} and 98% identity) against the *E. coli* genome. Using this conservative method, 1574 proteins were mapped to 1605 distinct genomic loci.

Results and discussion

General experimental strategy

In this paper, our experiences relating to the large-scale identification of LMW proteins (molecular weights <25 kDa) using gel-based and gel-free approaches are summarized. By combining different methods, a total of 455 LMW proteins of *E. coli* were identified with high certainty (Electronic supplementary material Tables S2 and S3).

As a starting point for optimization, the procedure published in 2007 by Klein et al. [20] was used, as this study reported an identification rate of 35% of the LMW subproteome of *Halobacterium salinarum*. The outline of this study consisted of high molecular weight protein depletion, separation by 1D-GE using a modified protocol according to Schaegger [23], and ESI-LC/MS^3 analysis with FTICR MS.

Here we vary this strategy stepwise in order to estimate the influence of the critical parameters in (i) protein extraction, (ii) enrichment and separation, (iii) proteolysis, (iv) MS and MS/MS analysis, and (v) protein identification (Fig. 1).

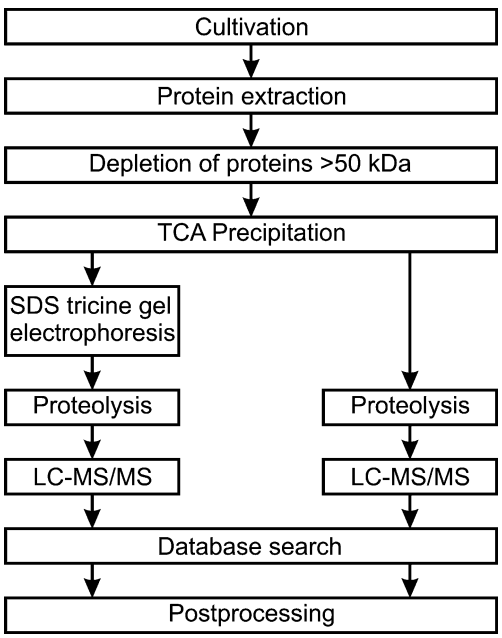


Fig. 1 Experimental workflow

Finally, the challenge of the de novo annotation of open reading frames (ORF) coding for small proteins on a genome-wide scale is addressed with the software *RNA-code*.

Optimization steps

Different protein extraction methods

To estimate the influence of the cell disruption and protein extraction methods, two different lysis buffers (a slightly basic ammonia buffer containing 8 M urea and an acidic buffer containing 0.1% TFA) were applied as a variant of the method described in Klein et al. [20]. Similar protein amounts were obtained with both buffers, which could not be increased by the successive usage of both extraction buffers (data not shown). After the depletion of higher molecular weight proteins using centrifugal filtration (molecular weight cut-off: 50 kDa), high enrichment in proteins <30 kDa was observed, with a maximum at approximately 15 kDa in terms of quantity (Fig. 2) and number of identifications (Fig. 3). The total protein amount determined after depletion and precipitation was approximately 2% for urea and 1% for TFA extracts. Proteins were separated using 1D SDS tricine GE, and the LMW range of each lane was cut into nine slices. Proteins were digested in gel with endoprotease AspN or trypsin, and the resulting peptides were subsequently analyzed by LC/MS.

The analysis resulted in a total of 333 and 223 protein identifications for extractions with urea and TFA, respectively. Interestingly, only 148±13 proteins were detected using both protocols, which represents 44% of all detected proteins (Fig. 4a).

The importance of an efficient cell disruption and protein extraction has already been pointed out in other studies [18,

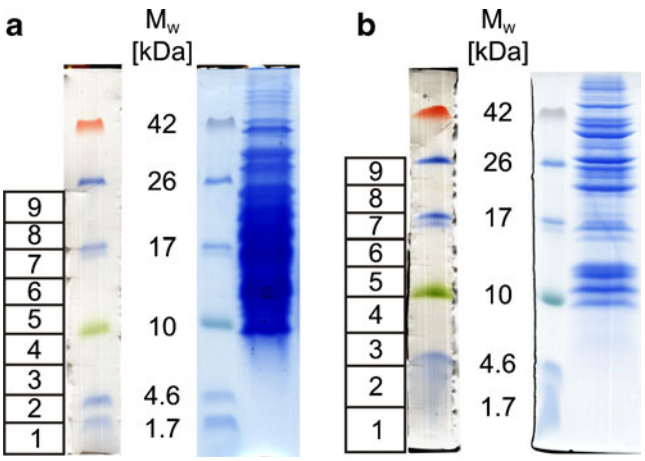


Fig. 2 SDS tricine gel after protein extraction with urea lysis buffer (**a**) and 0.1% TFA (**b**) and subsequent depletion of high molecular weight proteins. Excised bands of the unstained gel part are numbered

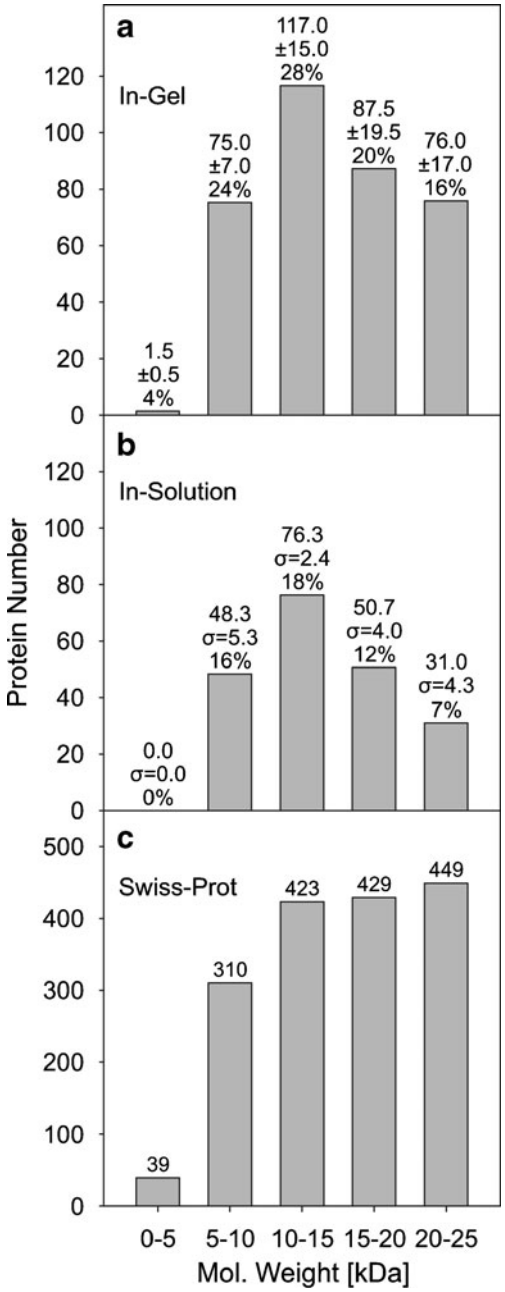


Fig. 3 Average mass distributions of the proteins identified using an in-gel (**a**) or in-solution (**b**) approach in comparison to the SwissProt protein database (**c**)

29]. Our results show that the choice of the extraction buffer can influence the number and type of identified proteins even more than the protease or the MS/MS fragmentation technique (discussed below).

For the proteins in the p*I* ranges of 5–7 and 11–14, the identification rate was higher with the urea than with the TFA lysis buffer (184 vs. 134 proteins, respectively, Fig. 5; Electronic supplementary material Figure S1). For very acidic proteins with a p*I* of <5, TFA lysis gives slightly better results than urea lysis (22 instead of 17 identified proteins).

Different protein separation methods

A 150 min gradient was used for the 1D-LC/MS analyses. However, a gel-based approach in which nine slices were analyzed by LC/MS using a 30 min gradient leads to a 49% increase (Fig. 3, Fig. 4b) in the identification rate. Thus, even though there are differences in terms of LC separation and measurement time, this indicates that investing time and effort in additional separation steps on the protein scale remains an efficient way of improving the proteome coverage. Nevertheless, some proteins may also be lost by additional separation steps. Eleven especially low-abundance (four proteins below 1000 copies/cell) or as-yet unquantified proteins (five proteins) were exclusively detected by the shorter LC/MS-based approach.

Proteolytic digestion

The possibility of increasing the protein identification rate as well as the average sequence coverage through the complementary application of more than one protease is a known strategy. Recently, Swaney and coworkers improved the coverage of the proteome of *Saccharomyces cerevisiae* by performing complementary proteolytic digestions with multiple enzymes and subsequently analyzing using LC/MS [19]. While the proteases trypsin, AspN, GluC, ArgC and LysC were used, the highest identification rate was obtained with trypsin. Nevertheless, the other proteases increased the identification rate by 18% (3908 instead of 3313 proteins) and—perhaps more importantly—the average sequence coverage increased from 24.5% to 43.4% as compared to that obtained with the exclusive use of trypsin.

In addition to trypsin, we used endoprotease AspN, which was predicted to create nearly the same number of proteolytic peptides in the molecular weight range 800–3,000 Da, and to present the highest orthogonality to trypsin in terms of sequence coverage for LMW proteins (Electronic supplementary material Table S4). Furthermore, the prediction showed that in a complementary analysis using both endoprotease AspN and trypsin, the number of unidentifiable LMW proteins would be reduced to 67 in comparison to the 233 not indentified when using trypsin as the only protease. For unequivocal identification, at least three detectable proteolytic peptides were required in this in silico digestion (Electronic supplementary material Table S4).

In summary, 292.5±76.5 proteins could be identified with trypsin, and 163.5±9.5 (46%) of these could be verified using endoprotease AspN (Figs. 3 and 4c). The average sequence coverage of proteins identified by both proteases was increased from 48.0% to 63.7% by combining

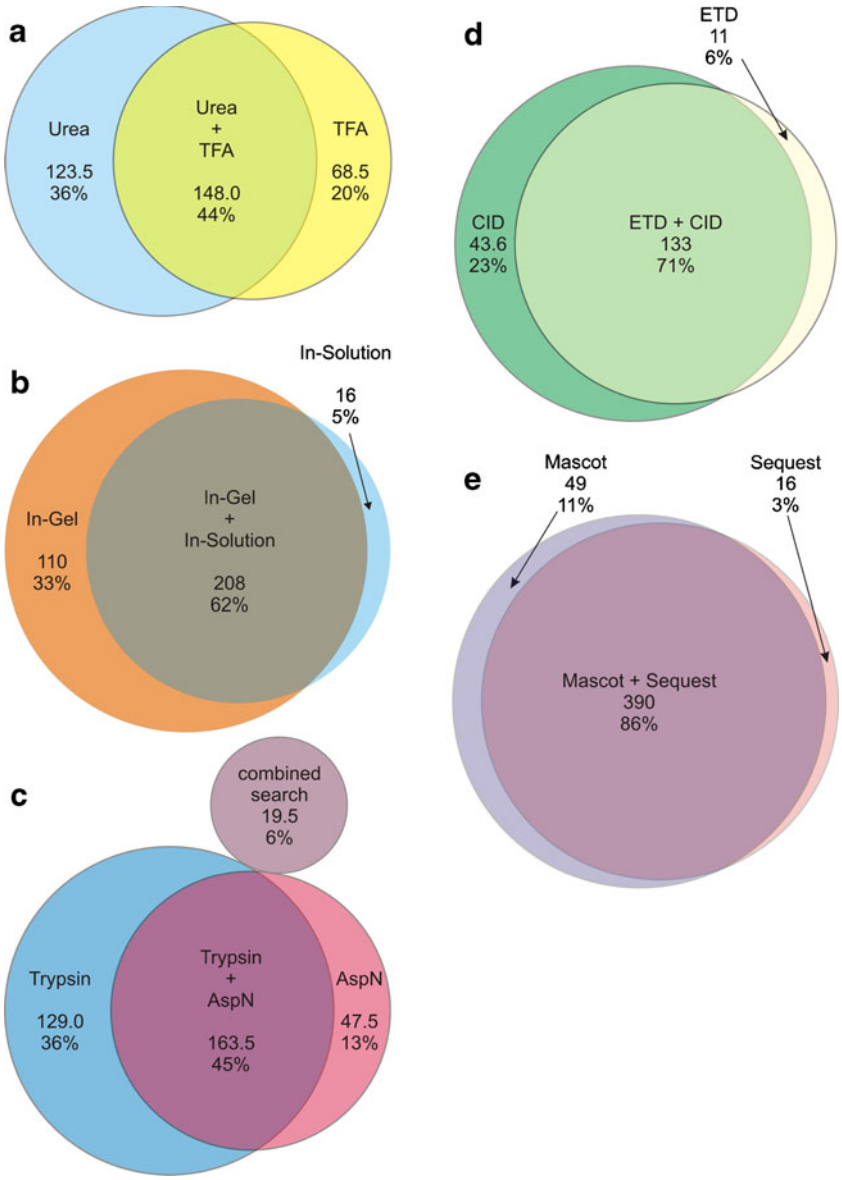


Fig. 4 Influence of different protocol variations. Comparison of average protein identifications after **a** protein extraction with urea lysis buffer or 0.1% TFA, **b** digestion with the in-solution or the in-gel approach, **c** digestion with trypsin or AspN, **d** MS/MS fragmentation and detection by IT-CID or IT-ETcaD, and **e** MS/MS database search using the MASCOT or SEQUEST search engines

the results obtained using trypsin with those obtained using endoprotease AspN (Table. 5). Furthermore, 47.5±25.5 (13%) proteins could only be identified after proteolysis with endoprotease AspN. According to Ishihama et al. [9], 21 of the 63 additionally identified proteins have copy numbers per cell of below 1000, whereas 28 were not covered by this study. Performing a database search by combining the LC/MS results obtained through digestion with trypsin and endoprotease AspN yielded 19.5±9.5 (6%) additional protein identifications. The abundance of at least several of these proteins was very low (7 were determined to be present with less than 1100 copies/cell), whereas 22 were not yet quantified.

In contrast to tryptic peptides (except C-terminal peptides), which always possess a "mass spectrometry friendly" C-terminal charge due to the occurrence of a C-terminal arginine or lysine, this is not necessarily the case for proteolytic peptides derived via cleavage with endoprotease AspN. This resulted in decreased spectral quality and thus in lower average MASCOT scores (C-terminal arginine or lysine: both 39, for N-terminal aspartic acid and glutamic acid: 30 and 31) and slightly lower SEQUEST scores (for lysine and arginine: 3.3 and 3.1; for acid and glutamic acid: 3.0 and 3.0). The cleavage efficiency of endoprotease AspN was lower for glutamic than for aspartic acid (1586 instead of 205 identified peptides).

Variation of fragmentation technique

The fragments created by ETD, CID and HCD can either be detected with high sensitivity and a short measuring time in the linear iontrap (IT-ETD and IT-CID) or with high

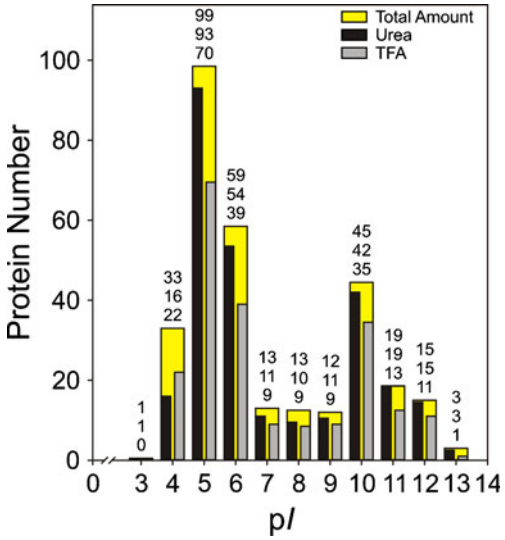


Fig. 5 p*I* distributions of the proteins identified with the in-gel approach after protein extraction with urea lysis buffer or 0.1% TFA in comparison with the total amount of identified proteins

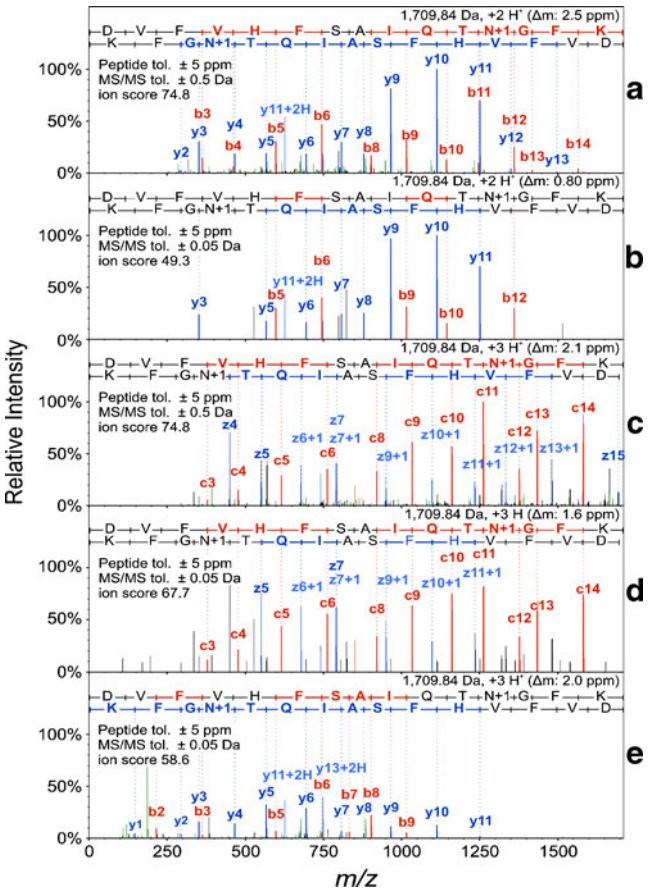


Fig. 6 Comparison of different fragmentation methods after in in-solution proteolysis, as exemplified by the peptide DVFVHFSAIQTnGFK from the cold shock-like protein cspE (**a** IT-CID, **b** FT-CID, **c** IT-ETD, **d** FT-ETD, **e** FT-HCD). *n* denotes an Asn that was found to be deamidated

accuracy and resolution in the Orbitrap analyzer (Orbitrap-ETD, Orbitrap-CID and HCD).

The benefits of using different analyzer types for MS/MS measurements as well as the different fragmentation techniques ETD, CID and HCD were evaluated with biological triplicates.

Using the linear ion trap as the mass analyzer for MS/MS detection, the three methods (a) CID, (b) ETD and (c) CID combined with ETD by a data-dependent decision tree provided an average of 177 (σ=19), 144 (σ=15) and 160 (σ=21) protein identifications with very high confidence. The overlap between the IT-ETD and IT-CID results was 71%, whereas only 6% more identifications were gained by using IT-ETD (Fig. 4d). However, since IT-ETD confirmed 75% of the proteins identified by IT-CID, this complementary fragmentation technique represents a useful method of independent validation. Moreover, the average sequence coverage and the average number of identified peptides per protein were increased by 5.5% and 21.7%, respectively (Table. 5).

Comparing the two different mass analyzers for MS/MS fragment ions, the Orbitrap offers highly accurate fragment ion mass measurements as well as enhanced signal-to-noise ratios for highly abundant peptides (Fig. 6). In contrast, due to its lower speed and sensitivity, about 50% fewer MS/MS spectra could be recorded per run, resulting in about 15% of the unique peptides being identified. On average, MS/MS analysis of the fragments created by CID, HCD or ETD in the Orbitrap resulted in the identification of only 27, 23 and 25 LMW proteins, respectively. This is also consistent with a recent in-depth study by Kim and coworkers, who analyzed *E. coli* lysates by CID fragmentation in the LTQ Orbitrap using different conditions for MS and MS/MS resolution [30]. However, the issue that the number of proteins identified is much lower due to the lower scanning speed and sensitivity of the techique may soon be overcome due to further improvements in the speed and sensitivity of the Orbitrap analyzer [31].

Influence of the MS analysis algorithm

There is still ongoing discussion about the quality of peptide MS/MS search engines [32, 33]. This issue is especially important here, due to the fact that the number of peptides per LMW protein formed by proteolysis is very limited. Additionally, the erroneous identification of a peptide could easily lead to wrong protein identification. Therefore, high sensitivity and accuracy is required during peptide identification. To address this issue with a special focus on LMW proteins, we performed searches with the two most widely used database search engines MASCOT and SEQUEST. After adjusting to 5% FDR using a decoy database, an overlap of 86% was observed (Fig. 4e). Here, MASCOT turned out to be more sensitive, resulting in the

unique identification of 49 unique proteins compared to the 16 discovered by SEQUEST. Furthermore, for the gel-based approach, the number of significant identifications performed by MASCOT, 1060±86 peptides (on average 5.4 peptides per protein), was higher than the 902±85 peptides (5.0 peptides per protein) identified with SEQUEST However, we decided to combine and re-evaluate the results obtained with both engines using SCAFFOLD in order to generate the final identification results.

Covered protein groups

According to the GO classification, the identified proteins were clustered using the GO terms "molecular function," "cell function," and "localization" [27]. Information about the copy number per cell was taken from Ishihama et al. [9].

Cellular localization of identified LMW proteins

With the protocol applied, we obtained good to excellent coverage for cytoplasmic (100 proteins, 45%), periplasmic (22 proteins, 52%) and ribosomal proteins (53 proteins, 98%). Not unexpectedly, the identification rate for inner membrane (43 proteins, 12%) and outer membrane proteins (12 proteins, 33%) was significantly lower (Table 2). However, it is possible to improve the coverage of membrane proteins by performing additional prefractionation [34, 35].

Protein abundance and molecular and cellular function

In order to estimate the copy numbers of a wide range of cytosolic proteins, Ishihama and coworkers [9] used label-free protein quantitation. The proteins identified in this and our study cover a dynamic range of six orders of magnitude. These proteins include highly abundant ribosomal proteins like the 50S ribosomal protein L33 (SwissProt entry: P0A7N9, 186,000,000 copies/cell) as well as rare proteins with less than 200 copies per cell such as Acyl-CoA thioesterase I (SwissProt entry: P0ADA1, 186 copies/cell). Furthermore, we identified about 100 proteins that are not covered by the study of Ishihama et al. (Electronic supplementary material Table S5).

According to the GO annotations of *E. coli*, neither the biological processes associated with nor the molecular functions of 846 proteins are characterized. Interestingly, 579 (i.e., 68%) of these proteins possess a molecular weight of <25 kDa (Tables. 2, 3 and 4). In our study, we were able to identify 93 of these uncharacterized proteins. The coverage of such proteins by proteome studies will subsequently allow protein quantification, and thus may ultimately contribute to the elucidation of their functional roles.

Detection and evaluation of proteins predicted at the DNA or transcriptome level using *RNAcode*

Among the 1723 individually predicted proteins, there are 837 (49%) LMW proteins that have not yet been validated at the proteome level. Of those 837 LMW proteins, 96 were detected in our study. However, 91 of these were recently covered by Iwasaki et al. [10], whereas, to our knowledge, the existence of the five remaining proteins has never been established before.

Aside from all the experimental challenges involved, an additional reason for the underrepresentation of LMW proteins in proteome studies is probably the inherent difficulty of the annotation process, which results in an significant number of either dubious or missing protein predictions [14, 36, 37]. In order to improve the prediction and annotation of LMW proteins, we used the recently developed *RNAcode* algorithm [16]. *RNAcode* performs a comparison of homolog sequences that show evolutionary conservation and has already been applied to transcriptome data [17].

In the present study, we show how *RNAcode* can revise existing annotations and also estimate their specificity by performing a comparison with our proteome data. Of 1605 mapped LMW SwissProt protein loci, at least 70% of the sequences of 1401 overlapped with segments that gave high scores in *RNAcode*. Ninety-five percent of the proteins with either proteome or transcriptome evidence listed in the SwissProt database are positively classified by *RNAcode*

Table 2 Gene ontology annotation according to localization

Localization	Cytoplasm		Ribosome		Membrane		Periplasmic space		Cell projection/ flagellum		Extracellular		Cell wall/cell membrane		Other/not assigned	
Swissprot *E.coli* K12 <25 kDa	219		55		356		42		36		3		36		995	
In gel	101	46.1%	53	96.4%	43	12.1%	21	50.0%	2	5.6%	1	33.3%	11	30.6%	213	21.4%
In solution	63	28.8%	48	87.3%	23	6.5%	13	31.0%	1	2.8%	1	33.3%	5	13.9%	114	11.5%
In gel + in solution	110	50.2%	53	96.4%	47	13.2%	22	52.4%	2	5.6%	1	33.3%	11	30.6%	229	23.0%

Table 3 Gene ontology annotation according to biological process

Biological process		Antioxidant activity		Binding		Catalytic activity		Enzyme regulator activity		Molecular transducer activity		Structural molecule activity		Transcription regulator activity		Translation regulator activity		Other/not assigned	
Swissprot *E.coli* K12	<25 kDa	8		636		400		12		21		59		65		4		851	
In gel		7	87.5%	228	35.8%	121	30.3%	7	58.3%	2	9.5%	54	91.5%	25	38.5%	3	75.0%	131	15.4%
In solution		6	75.0%	147	23.1%	63	15.8%	4	33.3%	1	4.8%	49	83.1%	16	24.6%	3	75.0%	73	8.6%
In gel + in solution		7	87.5%	242	38.1%	128	32.0%	8	66.7%	3	14.3%	54	91.5%	28	43.1%	4	100%	141	16.6%

(Electronic supplementary material Table S6). This indicates that there is a strong enrichment of experimentally supported proteins in *RNAcode* predictions. Among the 455 proteins identified in this study, 449 (99%) show a clear evolutionary signal for conservation at the nucleic acid level. Proteome or transcriptome evidence is also reported in the SwissProt database for 81% (365/449) of these. Thus, the proteins identified in our study and the *RNAcode* predictions are highly correlated.

On the other hand, of the proteins not covered in our study or which had already been validated experimentally or by sequence homology according to the SwissProt database, only 68% were supported by *RNAcode* predictions (Electronic supplementary material Table S6). This difference suggests that many but probably not all of the as-yet unverified reading frames in the SwissProt database are real protein-coding segments. Interestingly, 229 high-scoring protein-coding segments detected with *RNAcode* do not overlap with annotated genes. Thus, the existence of LMW proteins which are not included in the current version of the SwissProt database was indicated by *RNAcode* analysis [16].

This analysis clearly shows that the existing SwissProt protein database can be improved, specifically with respect to evolutionary conservation, by the novel in silico approach. Furthermore, the results of our LMW proteome analysis are supported by other experimental data and they show a good correlation with the protein coding signals predicted by *RNAcode* too (Electronic supplementary material Table S6).

In this study, 54 proteins were identified which were only predicted according to EXPASY SwissProt database information (http://expasy.org/sprot/). Furthermore, five of the identified proteins (SwissProt entries P76549, P21418, P0A703, A5A614, and P0AEG8; Electronic supplementary material Tables S2 and S3) have not yet been validated according to the latest large-scale studies by Iwasaki et al. [10] and Ishihama et al. [9]. By applying *RNAcode*, the corresponding gene regions were predicted to code for these LMW proteins with high probability (Fig. 7).

Validation is crucial when claiming newly detected proteins. We analyzed the samples after extraction with urea or TFA lysis buffer and digestion with the endoproteases AspN and trypsin, which produce complementary peptides. This enabled us to unambiguously confirm the existence of all of them by multiple detection with FDR probabilities of below 0.05. For example, for the protein P0AEG8, identification is based on two tryptic peptides and four proteolytic peptides created by the endoprotease AspN, so the sequence coverage was increased to 65% (Fig. 7). Additionally, the predicted proteins were found in independently processed biological replicates.

Table 4 Gene ontology annotations according to molecular function

Molecular function	Cellular process		Developmental process		Interaction with cells and organisms		Localization		Metabolic process		Regulation		Reproduction		Response to stimulus		Other/not assigned	
Swissprot *E.coli* K12 <25 kDa	606		1		63		146		110		205		5		100		807	
In gel	206	34.0%	0	0.0%	6	9.5%	29	19.9%	43	39.1%	71	34.6%	1	20.0%	45	45.0%	148	18.3%
In solution	143	23.6%	0	0.0%	5	7.9%	19	13.0%	26	23.6%	49	23.9%	1	20.0%	32	32.0%	73	9.0%
In gel + in solution	218	36.0%	0	0.0%	6	9.5%	31	21.2%	44	40.0%	76	37.1%	1	20.0%	48	48.0%	160	19.8%

Perspectives on LMW proteome analysis

However, even these improved identification rates (especially in the molecular weight range of 5–15 kDa), compared to state of the art standard proteome studies (Fig. 8), of 62% for cytosolic proteins and 27% for all known LMW proteins (including membrane proteins) still leave some room for further improvement. Aside from aiming for increased coverage through the additional prefractionation of membrane proteins, our results indicate that improving protein and/or peptide separation leads to significantly higher identification rates as well as enhanced average sequence coverage.

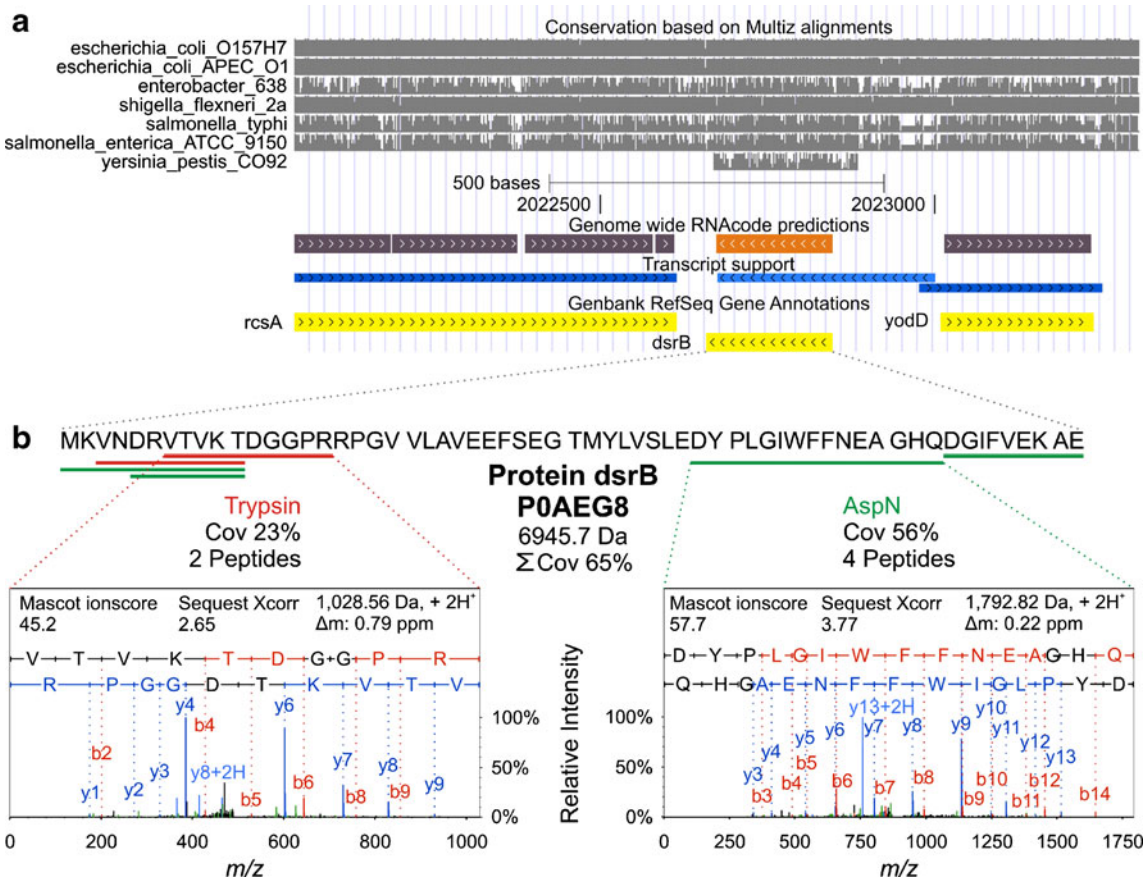


Fig. 7 Evaluation and validation of predicted proteins by **a** RNAcode and **b**. LC/MS/MS. **a** A UCSC screen shot of the genomic context around protein dsrB (Swiss Prot entry P0AEG8) is shown at the *top* with annotated protein coding genes (*yellow*), transcription units as defined by Cho et al. [41] (*blue*) and *RNAcode* high-scoring coding segments (*purple*). *Arrows within boxes* indicate the reading direction of the corresponding element. Marked in *light colors* are elements corresponding to protein dsrB. The *lower half* depicts the conservation of the *E. coli* region with respect to other enterobacteria. **b** Proteins were validated by LC/MS/MS analysis. Spectra and identification parameters of one of the peptides identified using the endoproteases trypsin or AspN are shown.

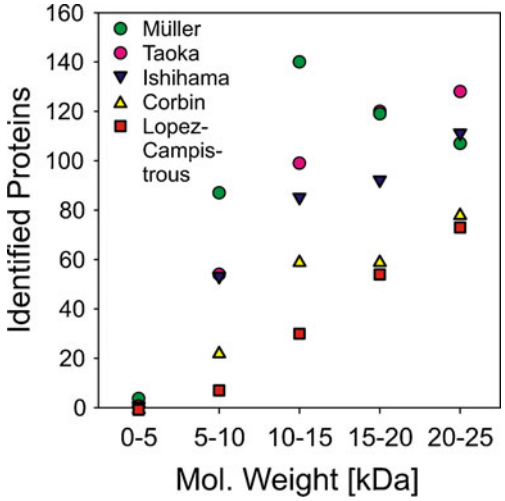


Fig. 8 Comparison of the total number of proteins identified here with the results of selected previous studies focusing on the coverage of the cytosolic proteome of *E. coli*

It was shown by Godoy et al. that near-complete proteome coverage is possible for yeast using *n*-dimensional protein and/or peptide separation prior to MS/MS analysis. However, these approaches are still very time intensive and require the analysis of several dozen proteolytic peptide fractions [38].

Recently, Iwasaki et al. used a non-commercially available 350 cm monolithic reversed-phase C_{18} column to achieve improved peptide separation for proteolytic peptide mixtures of whole *E. coli* cell lysates during a 41 h gradient. This approach allowed for the identification of 2602 proteins, of which 820 were LMW proteins (Table 1) [10]. However, even with this very powerful untargeted analysis, more than 50% of the LMW subproteome remained uncovered.

As a complement to the untargeted proteomics approaches, a targeted approach based on multiple reaction monitoring (MRM) has proven to be feasible for high-throughput proteomics studies [39]. The basic idea of this strategy is to optimize the detection of proteolytic peptides and to develop a sensitive and specific mass spectrometric assay. In a first step, these assays are developed based on specific precursor/fragment ion pairs called MRM transitions as well as LC retention time information by analyzing synthesized peptides corresponding to a proteolytic protein fragment. In a second step, proteins from real samples are identified and quantified by analyzing the real proteolytic peptides using the optimized MRM transitions. Using this approach, even proteins with very low abundances could be detected with a high success rate. However, synthesizing several hundreds to thousands of artificial proteolytic peptides as well as establishing suitable MRM transitions are relatively time- and cost-intensive processes. Nevertheless, especially for very sensitive, specific, and reproducible analyses of limited numbers of proteins, this strategy may be the best method currently available [40].

Summary

In conclusion (see also Table 5), there are various tailor-made strategies that can be used for LMW proteome analyses which vary in their aims and the technical equipment employed:

- For higher sequence coverage, employing a combination of enzymes can significantly increase the number of unique peptides per protein.
- In order to increase the identification rate, the use of an acidic extraction buffer may prove to be beneficial. Furthermore, sequential extraction using different extraction buffers may improve the identification rates, even if the total amount of extracted protein is not increased significantly (data not shown).
- To enhance the robustness of identifications based on an increased number of unique MS/MS spectra, the use of additional enzymes or complementary fragmentation methods like ETD represent efficient options.
- An easy and—with respect to measuring time—neutral way to improve the sensitivity and accuracy of peptide

Table 5 Gains in identification rate, sequence coverage and identification robustness obtained by performing a combined analysis rather than the standard procedure alone

Standard	Option	Proteins	Coverage**	Unique peptides**	Unique spectra**
Urea	TFA	+25.2%	+5.9%	+19.3%	+21.7%
Trypsin	AspN	+16.2% (+22.9%*)	+15.7%	+74.6%	+78.2%
IT-CID	IT-ETD	+6.2%	+5.5%	+21.7%	+30.1%
IT-CID	FT-CID	+0%	+0.7%	+2.6%	+2.4%
MASCOT	SEQUEST	+3.6 %	+1.4%	+3.6%	+4.3%

* Combined identification using trypsin and AspN results in one search

** Related to proteins identified in both experiments

identification is to combine multiple MS analysis algorithms. This is especially important for the identification of LMW proteins, which relies on a very limited number of proteotypic peptides.

- In terms of the efficient use of measurement time, analyzing different preparations of the same sample instead of multiple replicates or using extremely long gradients could be advantageous, as this can increase the total number of proteins identified, the sequence coverage, and the number of peptides per protein.

In conclusion, this study can be used as a guideline to improve the coverage of cytosolic LMW proteins, especially in the molecular weight range of 5–20 kDa.

Furthermore, in this study we investigated an automated protein-coding gene annotation tool. We analyzed the accuracy of *RNAcode* prediction in comparison to SwissProt protein database entries and proteins that we had experimentally verified. We found that the predictions made by *RNAcode* are highly correlated with experimentally validated proteins. Hence, there are 229 high-scoring protein-coding segments that do not overlap with annotated genes and which indicate the existence of additional putative small proteins in *E. coli*.

Acknowledgment Stephan A. Müller was supported by the German Research Council (Transregional Collaborative Research Centre 67, subproject Z4) and by the Helmholtz Impulse and Networking Fund through the Helmholtz Interdisciplinary Graduate School for Environmental Research (HIGRADE). Further financial support of the project was provided by the Helmholtz Alliance on Systems Biology and COST Action "Systems Chemistry" CM0703. Sven Findeiβ was supported by the DFG (grant No. STA 850/7-1 under the auspices of SPP-1258 "Small Regulatory RNAs in Prokaryotes"). Stefan Washietl was supported by an Erwin Schrödinger Fellowship from the Austrian Science Fund.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

1. Blattner FR, Plunkett G III, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–1462
2. Han MJ, Lee SY (2006) The *Escherichia coli* proteome: past, present, and future prospects. Microbiol Mol Biol Rev 70:362–439
3. Link A, Robison K, Church G (1997) Comparing the predicted and observed properties of proteins encoded in the genome of *Escherichia coli* K-12. Electrophoresis 18:1259–1313
4. Lopez-Campistrous A, Semchuk P, Burke L, Palmer-Stone T, Brokx SJ, Broderick G, Bottorff D, Bolch S, Weiner JH, Ellison MJ (2005) Localization, annotation, and comparison of the *Escherichia coli* K-12 proteome under two states of growth. Mol Cell Proteomics 4:1205–1209
5. Ihling C, Sinz A (2005) Proteome analysis of *Escherichia coli* using high-performance liquid chromatography and Fourier transform ion cyclotron resonance mass spectrometry. Proteomics 5:2029–2042
6. Gevaert K, Van Damme J, Goethals M, Thomas GR, Hoorelbeke B, Demol H, Martens L, Puype M, Staes A, Vandekerckhove J (2002) Chromatographic isolation of methionine-containing peptides for gel-free proteome analysis: identification of more than 800 *Escherichia coli* proteins. Mol Cell Proteomics 1:896–903
7. Corbin RW, Paliy O, Yang F, Shabanowitz J, Platt M, Lyons CE, Root K, McAuliffe J, Jordan MI, Kustu S, Soupene E, Hunt DF (2003) Toward a protein profile of *Escherichia coli*: Comparison to its transcription profile. Proc Natl Acad Sci USA 100:9232–9237
8. Taoka M, Yamauchi Y, Shinkawa T, Kaji H, Motohashi W, Nakayama H, Takahashi N, Isobe T (2004) Only a small subset of the horizontally transferred chromosomal genes in *Escherichia coli* are translated into proteins. Mol Cell Proteomics 3:780–787
9. Ishihama Y, Schmidt T, Rappsilber J, Mann M, Hartl FU, Kerner MJ, Frishman D (2008) Protein abundance profiling of the *Escherichia coli* cytosol. BMC Genomics 9:102
10. Iwasaki M, Miwa S, Ikegami T, Tomita M, Tanaka N, Ishihama Y (2010) One-dimensional capillary liquid chromatographic separation coupled with tandem mass spectrometry unveils the *Escherichia coli* proteome on a microarray scale. Anal Chem 82:2616–2620
11. Santos PM, Roma V, Benndorf D, von Bergen M, Harms H, Sa-Correia I (2007) Mechanistic insights into the global response to phenol in the phenol-biodegrading strain *Pseudomonas* sp. M1 revealed by quantitative proteomics OMICS. J Integr Biol 11:233–251
12. Benndorf D, Balcke GU, Harms H, von Bergen M (2007) Functional metaproteome analysis of protein extracts from contaminated soil and groundwater. ISME J 1:224–234
13. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25:25–29
14. Hemm MR, Paul BJ, Schneider TD, Storz G, Rudd KE (2008) Small membrane proteins found by comparative genomics and ribosome binding site models. Mol Microbiol 70:1487–1501
15. Hardison RC (2003) Comparative genomics. PLoS Biol 1:e58
16. Washietl S, Findeiß S, Mueller SA, Kalkhof S, von Bergen M, Hofacker IL, Stadler PF, Goldman N (2010) RNAcode: robust prediction of protein coding regions in comparative genomics data. www.bioinf.uni-leipzig.de/Publications/PREPRINTS/10-001.pdf
17. Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeisz S, Sittka A, Chabas S, Reiche K, Hackermuller J, Reinhardt R, Stadler PF, Vogel J (2010) The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature 464:250–255
18. De Mey M, Lequeux GJ, Maertens J, De Muynck CI, Soetaert WK, Vandamme EJ (2008) Comparison of protein quantification and extraction methods suitable for *E. coli* cultures. Biologicals 36:198–202
19. Swaney DL, Wenger CD, Coon JJ (2010) Value of using multiple proteases for large-scale mass spectrometry-based proteomics. J Proteome Res 9:1323–1329
20. Klein C, Aivaliotis M, Olsen JV, Falb M, Besir H, Scheffer B, Bisle B, Tebbe A, Konstantinidis K, Siedler F, Pfeiffer F, Mann M, Oesterhelt D (2007) The low molecular weight proteome of *Halobacterium salinarum*. J Proteome Res 6:1510–1518
21. Dai Y, Li L, Roser DC, Long SR (1999) Detection and identification of low-mass peptides and proteins from solvent suspensions of *Escherichia coli* by high performance liquid chromatography fractionation and matrix-assisted laser desorption/

ionization mass spectrometry. Rapid Commun Mass Spectrom 13:73–78
22. Harper RG, Workman SR, Schuetzner S, Timperman AT, Sutton JN (2004) Low-molecular-weight human serum proteome using ultrafiltration, isoelectric focusing, and mass spectrometry. Electrophoresis 25:1299–1306
23. Schagger H (2006) Tricine-SDS-PAGE. Nat Protoc 1:16–22
24. Elias JE, Haas W, Faherty BK, Gygi SP (2005) Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. Nat Meth 2:667–675
25. Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 74:5383–5392
26. Nesvizhskii AI, Keller A, Kolker E, Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 75:4646–4658
27. Bhatia VN, Perlman DH, Costello CE, McComb ME (2009) Software tool for researching annotations of proteins: open-source protein annotation software with data visualization. Anal Chem 81:9819–9823
28. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W (2004) Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 14:708–715
29. von der Haar T (2007) Optimized protein extraction for quantitative proteomics of yeasts. PLoS ONE 2:e1078
30. Kim M-S, Kandasamy K, Chaerkady R, Pandey A (2010) Assessment of resolution parameters for CID-based shotgun proteomic experiments on the LTQ-Orbitrap mass spectrometer. J Am Soc Mass Spectrom (in press)
31. Olsen JV, Schwartz JC, Griep-Raming J, Nielsen ML, Damoc E, Denisov E, Lange O, Remes P, Taylor D, Splendore M, Wouters ER, Senko M, Makarov A, Mann M, Horning S (2009) A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. Mol Cell Proteomics 8:2759–2769
32. Kapp EA, Schütz F, Connolly LM, Chakel JA, Meza JE, Miller CA, Fenyo D, Eng JK, Adkins JN, Omenn GS, Simpson RJ (2005) An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. Proteomics 5:3475–3490
33. Price TS, Lucitt MB, Wu W, Austin DJ, Pizarro A, Yocum AK, Blair IA, FitzGerald GA, Grosser T (2007) EBP, a program for protein identification using multiple tandem mass spectrometry datasets. Mol Cell Proteomics 6:527–536
34. Molloy MP, Herbert BR, Slade MB, Rabilloud T, Nouwens AS, Williams KL, Gooley AA (2000) Proteomic analysis of the *Escherichia coli* outer membrane. Eur J Biochem 267:2871–2881
35. Masuda T, Saito N, Tomita M, Ishihama Y (2009) Unbiased quantitation of *Escherichia coli* membrane proteome using phase transfer surfactants. Mol Cell Proteomics 8:2770–2777
36. Jäger D, Sharma CM, Thomsen J, Ehlers C, Vogel J, Schmitz RA (2009) Deep sequencing analysis of the *Methanosarcina mazei* Gö1 transcriptome in response to nitrogen availability. Proc Natl Acad Sci 106:21878–21882
37. Basrai MA, Hieter P, Boeke JD (1997) Small open reading frames: beautiful needles in the haystack. Genome Res 7:768–771
38. de Godoy LMF, Olsen JV, Cox J, Nielsen ML, Hubner NC, Frohlich F, Walther TC, Mann M (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature 455:1251–1254
39. Picotti P, Rinner O, Stallmach R, Dautel F, Farrah T, Domon B, Wenschuh H, Aebersold R (2009) High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. Nat Meth 7:43–46
40. Picotti P, Bodenmiller B, Mueller LN, Domon B, Aebersold R (2009) Full dynamic range proteome analysis of *S. cerevisiae* by targeted proteomics. Cell 138:795–806
41. Cho B-K, Zengler K, Qiu Y, Park YS, Knight EM, Barrett CL, Gao Y, Palsson BO (2009) The transcription unit architecture of the *Escherichia coli* genome. Nat Biotech 27:1043–1049

The NIH Roadmap Epigenomics Mapping Consortium

Bradley E. Bernstein[1,2], John A. Stamatoyannopoulos[3], Joseph F. Costello[4], Bing Ren[5], Aleksandar Milosavljevic[6], Alexander Meissner[1,7], Manolis Kellis[1], Marco A. Marra[8], Arthur L. Beaudet[6], Joseph R. Ecker[9], Peggy J. Farnham[10], Martin Hirst[8], Eric S. Lander[1], Tarjei S. Mikkelsen[1], James A. Thomson[11]

1. Broad Institute of Harvard and MIT, Cambridge, MA 02129
2. Howard Hughes Medical Institute, Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114
3. Departments of Genome Sciences and Medicine, University of Washington School of Medicine, Seattle, WA 98195.
4. Department of Neurosurgery, UCSF, San Francisco, CA 94158
5. Ludwig Institute for Cancer Research, UCSD School of Medicine, La Jolla, CA 92093
6. Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77381
7. Department of Stem Cell and Regenerative Biology at Harvard University, Cambridge, MA 02138
8. Genome Sciences Centre, BC Cancer Agency, Vancouver, BC Canada
9. Genomic Analysis Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92037
10. Genome Center, University of California at Davis, Davis, California 95616
11. University of Wisconsin School of Medicine and Public Health, Madison, WI 53706

The NIH Roadmap Epigenomics Mapping Consortium aims to produce a public resource of epigenomic maps for stem cells and primary *ex vivo* tissues selected to represent the normal counterparts of tissues and organ systems frequently involved in human disease. The maps will detail the genomewide landscapes of DNA methylation, histone modifications and related chromatin features. They are intended to provide a reference for studies of the genetic and epigenetic events that underlie human development, diversity and disease. Here we describe the organizational structure, goals and anticipated deliverables of the consortium.

Introduction

Recent years have seen remarkable progress in our understanding of human genetics, enabled by the availability of the human genome sequence and increasingly high-throughput technologies for DNA analysis[1]. Yet despite their breadth and comprehensiveness, purely DNA sequence-level investigations do not shed light on a crucial component of human biology: how the same genome sequence can give rise to over two hundred different cell types through remarkably consistent differentiation programs. This process of developmental specification, classically termed 'epigenesis', is now appreciated to involve differential regulation of genes and their products[2]. Aberrant regulation of such phenomena has been extensively linked to human diseases and, additionally, can be influenced by environmental inputs[3-5].

Gene regulation and genome function are intimately related to the physical organization of genomic DNA and in particular to the way it is packaged into chromatin, a complex nucleoprotein structure comprising histones, DNA binding factors, accessory protein complexes and non-coding RNAs[6-9] (**Figure 1**). Chromatin is a dynamic entity that is subject to modification of both its DNA and

protein components, with direct structural and functional consequences. The term ‘epigenome’ is used to describe the way in which these modifications and structural features are distributed across the genome in a given cell population. The epigenomic landscapes and the associated gene expression programs are maintained within a given cell lineage through complex processes that involve transcription factors, chromatin regulators, histone modifications and variants, and RNAs[10-12], but that remain poorly understood in mammals.

Although the mechanisms remain obscure, a now overwhelming body of evidence supports central roles for epigenomic changes in disease susceptibility and pathogenesis. Multiple disease processes including cancer are now well known to be associated with characteristic alterations in the patterns of chromatin, DNA methylation and gene expression[3,5]. In addition, epidemiological studies have linked early environmental exposures, such as *in utero* starvation, to long-term health consequences ranging from metabolic disorders to psychiatric diseases[13]. A causal role for epigenomic aberrations is supported by several lines of evidence, including mutations of genes encoding chromatin regulators in developmental disorders and cancer[4,14-16], and by the therapeutic efficacy of small molecule inhibitors of DNA methyltransferases and histone modifying enzymes[17].

Major epigenomic features can now be interrogated comprehensively by combining cellular, biochemical and molecular techniques with high-throughput sequencing. Production of genomewide maps of cytosine methylation, histone modifications, chromatin accessibility and RNA transcripts represents a powerful and general approach for surveying the regulatory state of the genome in a cell type of interest. The resulting data define the locations and activation states of diverse functional elements, including genes and their transcriptional control elements (promoters, enhancers, insulators, etc), non-coding transcripts and epigenetic effectors such as imprinting control regions[18-25]. More globally, such maps can provide insight into developmental state and potential, for example of a stem cell population, and shed light on aberrant regulatory programs in diseased tissues.

A Coordinated Study of Human Epigenomes

In 2008, the NIH Roadmap Epigenomics Mapping Consortium was launched with the goal of producing a public resource of human epigenomic data to catalyze basic biology and disease-oriented research (www.roadmapepigenomics.org). The consortium leverages experimental pipelines built around next-generation sequencing technologies to map DNA methylation, histone modifications, chromatin accessibility and RNA transcripts in stem cells and primary *ex vivo* tissues selected to represent the normal counterparts of tissues and organ systems frequently involved in human disease. The mapping of such normal epigenomes is being undertaken by four Epigenomics Mapping Centers and supported by a Data Analysis and Coordinating Center, which collectively coordinate experimental and analytical efforts in order to maximize consistency, data quality and overall coverage of the epigenomic landscape. Because the epigenomic landscape varies dramatically across tissue types (and between individuals) there is no single ‘reference’ epigenome. Rather, the consortium expects to deliver a collection of normal epigenomes for different tissues and individuals intended to provide a framework or ‘reference’ for comparison and integration within a broad array of future studies. A core goal of the consortium is to close the gap between data generation and its public dissemination by rapid release of raw sequence

data, profiles of epigenomic features, and higher-level integrated maps in coordination with the National Center for Biotechnology Information (NCBI). The consortium is also committed to the development, standardization and dissemination of protocols, reagents and analytical tools to enable the research community to utilize, integrate and expand upon this body of data (**Figure 2**).

Reference Maps for Major Epigenomic Features

The Epigenomics Mapping Centers have collaboratively established data collection pipelines to produce high-quality, comprehensive epigenomic maps. Specific data types have been prioritized that offer broad insight into genome regulation, are generally applicable to diverse cell populations and can be evaluated comprehensively and accurately by high-throughput sequencing. These include genomic maps for DNA methylation, histone modifications, chromatin accessibility and RNA expression. The Mapping Centers work with the Data Analysis and Coordination Center to evaluate, compare and integrate the different data types and formats in order to ensure data quality and standards that enable the larger community to build upon these data.

(i) DNA methylation is assayed by sequencing DNA that has been treated with sodium bisulfite (BS-seq), or enriched by methyl-cytosine pull-down (MeDIP-seq) or methylation-sensitive restriction enzymes (MRE-seq). BS-seq, applied either to whole genome or to reduced-representation samples, has been designated as a primary assay because it provides accurate and consistent nucleotide-resolution data. The consortium is implementing MeDIP-seq and MRE-seq on a more limited basis to benchmark and compare these widely applied approaches.

(ii) Histone modifications are assayed by sequencing DNA enriched by chromatin immunoprecipitation with modification-specific histone antibodies (ChIP-seq). The consortium has implemented rigorous specificity tests that utilize arrays of differentially-modified histone tail peptides to ensure antibody specificity. In addition, common cell sources are collectively profiled and compared to ensure consistency between the different data collection centers.

(iii) Chromatin accessibility is assayed by sequencing DNaseI cleavage sites in nuclear chromatin (DNase-seq). These assays are performed at high sequencing depth in order to provide a global survey of accessible regions as well as high-resolution information regarding the protein occupancy of specific sequences[24].

(iv) RNA expression is assayed by sequencing mRNAs or size-selected small RNA fractions to high depths (RNA-seq). These expression data are intended to augment and better appreciate the functional output of the epigenomic profiles.

Given its mandate to deliver epigenomic maps for hundreds of different cell populations, the consortium must balance breadth of cell coverage with the depth to which different epigenomic features are investigated. High-value cell types such as human embryonic stem cells will be subjected to deep

exploration of a very broad range of histone modifications and comprehensive, single nucleotide-resolution analysis of DNA methylation. Although it is not yet possible to specify a definitive set of features that represent a minimal epigenome, the consortium has initially identified DNA methylation, 6 major histone modifications, chromatin accessibility and RNA as essential features that will be assayed in most or all designated cell populations. This combination of deep and broad analysis is expected to maximize coverage of cellular diversity and disease-relevant human tissues, while ensuring that a broad range of epigenomic features is explored.

Prioritized Cells and Tissues

The consortium will investigate a diverse collection of cell and tissue models, including (i) embryonic and adult stem cells and their differentiated progeny; (ii) induced pluripotent stem cells; and (iii) primary *ex vivo* human fetal and adult tissues. These cells and tissues were prioritized based on broad scientific and biomedical interest, tractability, phenotypic diversity and under-representation in other collaborative projects.

Because of their biomedical importance, human embryonic stem cells and major lineage derivatives have been selected for intensive investigation. The resulting data will offer insight into the distributions, dynamics and inter-relationships among epigenomic features, and catalyze study of their functions in development, epigenetic control and genome regulation. The consortium will also target additional stem cell models, including mesenchymal and neural stem cells, and reprogrammed cells as *in vitro* models of development with particular relevance to regenerative medicine.

Broader coverage of human cellular diversity will be achieved through study of primary cells and tissues relevant to metabolic and cardiovascular disease, cancer, neuropsychiatric disease, aging and other leading health issues. These will be acquired from primary sources and sorted or otherwise manipulated to obtain suitably homogeneous cell populations that will be directly channeled to data collection pipelines. Prioritized cell types include sorted hematopoietic lineages, liver, muscle, adipose as well as selected cell types from breast and neural tissues. In addition, fetal tissues will be analyzed for insight into epigenomic landscapes of early development. Maps for such primary *ex vivo* tissues are urgently needed, since most of our current knowledge has come from either transformed cell lines or cultured cells, both of which experience dramatic non-physiologic changes to their chromatin environment, including aberrant DNA hyper-methylation and loss of heterochromatin integrity. Collectively, profiles for these diverse cell models should offer unprecedented insight into the breadth and dynamics of human epigenomes and provide a durable framework for future explorations of epigenomic alterations associated with human disease.

Integration and Dissemination of Human Epigenomes

The consortium aims to provide the scientific community ready access to a critical mass of high-quality epigenomic data for cells and tissues representative of normal human biology. These data will comprise multiple levels of information, from raw sequencing data and epigenomic profiles for an individual epigenomic feature in a single cell or tissue type, to integrated epigenomic maps that represent a

composite of multiple epigenomic profiles for an individual cell type or, alternatively, that capture biological variation of such features across different cell types. The consortium will also develop and disseminate software tools and algorithms to facilitate utilization of these results by the community, for example through the ability to search for epigenomic signatures common across genes or loci, or to identify distinguishing features of cell lineages, developmental stages, cellular environments or derivation history. The latter may also be used to classify disease states or to identify aberrant epigenomic features or regulatory programs that underlie human pathology.

The primary web portal for the consortium (www.roadmapepigenomics.org) offers detailed descriptions of the overall project, target cell and tissue types, and epigenomic assays utilized by the consortium. The portal links to companion sites managed by NCBI (www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/) and the Data Analysis and Coordination Center (www.epigenomeatlas.org) that provide access to raw and processed consortium data along with tools for visualization, analysis and integration of epigenomic data.

Progress and Challenges

The use of established technologies and approaches enabled the consortium to rapidly initiate data production. Notable progress during the initial phase included production of comprehensive DNA methylomes for a human embryonic stem cell line (H1) and primary fibroblasts[26]; hundreds of datasets for major histone modifications, targeted DNA methylation analysis, RNA expression and chromatin accessibility representing dozens of cell types, including multiple stem cell lines and ex *vivo* adult and developing tissues. These datasets are now available for download and viewing at the web portals referenced above. Guided by other NIH genomics projects[27], the consortium has adopted a data release policy under which users will have immediate access to the data, but are expected to abide by a moratorium on submission or presentation of works that incorporate these data for the 9 months following their release.

Any effort of this scope inevitably faces challenges and obstacles. Major issues have revolved around cell type selection and acquisition, assay standardization and developing the infrastructures for integration and dissemination of epigenome-scale datasets.

(i) Cell type selection and acquisition. A major ongoing challenge relates to the identification and prioritization of cells and tissues by the consortium. Ideally, models are selected on the basis of pervasive biological and medical importance. However, the decisions are confounded by issues of tractability. Many high value primary tissues are available in limited quantities that push the detection boundaries of current technologies. In addition, isolating relatively homogeneous populations from certain complex tissues can involve extensive preparative steps that may themselves effect changes to the epigenome. Finally, our relatively crude understanding of inter-individual epigenomic variation leaves an open question as to how many samples of a given tissue type must be analyzed to yield a representative map. These challenges highlight the importance of technology development, including effective procedures for isolating homogeneous cell populations, interrogating small samples and increasing the throughput of the assays.

(ii) Standardization of assays. The consortium is implementing the latest epigenomic technologies based on next-generation sequencing technology. Since these technologies continue to evolve and are inherently dependent on preparative steps there is an ongoing need to benchmark and validate assays. In the case of histone modification assays, substantial resources must be committed to procurement and validation of high-quality antibody reagents, including confirmation of biochemical specificity and ChIP-seq efficacy. In the case of DNA methylation, there is a need to benchmark and standardize different assay types, including BS-seq applied to either reduced-representations or the whole genome as well as various enrichment methods in widespread use by the scientific community[28].

(iii) Data integration and dissemination. Several challenges have emerged at the level of data handling and analysis. First, a clearer understanding of the underlying datasets in terms of sensitivity, specificity and precision is needed and is being pursued as a joint effort among the centers. Second, the sheer volume and complexity of consortium-generated data has pushed the limits of existing analytical and visualization tools. Hence, the development of a new generation of tools for integration, dissemination and interpretation of epigenomic data is vital to the overall success of the program.

Future and Context

The long-term goal of epigenomics research is a fuller understanding of how global alterations in diverse functional features superimposed on the human genome sequence contribute to cellular phenotypes in health and disease. This is a complex and ambitious undertaking, the realization of which will ultimately require systematic dissection and analysis of tissues, characterization of disease models, and detailed exposition of regulatory mechanisms through model organism studies. The efforts of the Roadmap Epigenomics Mapping Consortium to establish an expansive resource of epigenomic maps of normal cell and tissue phenotypes represents an important step in this direction. By catalyzing subsequent mechanistic studies of chromatin, DNA methylation, and transcription, these efforts should provide a springboard for disease-focused studies such as those currently being pursued under the parallel Roadmap program Epigenomics of Human Health and Disease. These Roadmap efforts will also be complemented by other major initiatives such as the International Human Epigenome Consortium which was established to accelerate and coordinate epigenomics research worldwide (see accompanying paper by Satterlee et al).

More broadly, the consortium aims to foster synergistic interactions with related collaborative projects, including the ENCyclopedia Of DNA Elements (ENCODE) Consortium[18], and the International HapMap and 1000 Genomes Projects[29]. The Epigenomics Mapping Consortium is distinguished from these efforts by the broad set of normal primary tissues and stem cell-derived developmental models that it will survey. As such, it will provide a highly complementary resource through which the *in vivo* state and behavior of DNA elements catalogued under ENCODE or implicated in studies of genome variation may be understood. Such information will be essential for appreciating the relevance of perceived genomic elements and variants to normal development and human disease.

In the coming years, the Roadmap Epigenomics Program and other complementary efforts should engender a vastly greater appreciation of the organization of the human epigenome and how it varies across tissues, individuals and disease states – information that may translate directly into the identification of aberrant epigenetic events that underlie susceptibility to specific diseases and environmental exposures.

Acknowledgements

We are grateful to Robert Waterland, Charles Epstein, Noam Shoresh and all consortium members, as well as to the NIH Epigenomics Implementation Group for thoughtful discussions and feedback in the drafting of this document. We also thank Bang Wong for his Illustration of the Epigenome.

References

1 Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science (New York, N.Y* **322**, 881-888 (2008).
2 Bird, A. Perceptions of epigenetics. *Nature* **447**, 396-398 (2007).
3 Feinberg, A. P. Phenotypic plasticity and the epigenetics of human disease. *Nature* **447**, 433-440 (2007).
4 Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature genetics* **33 Suppl**, 245-254 (2003).
5 Jones, P. A. & Baylin, S. B. The Epigenomics of Cancer. *Cell* **128**, 683-692 (2007).
6 Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693-705 (2007).
7 Bernstein, B. E., Meissner, A. & Lander, E. S. The mammalian epigenome. *Cell* **128**, 669-681 (2007).
8 Fraser, P. & Bickmore, W. Nuclear organization of the genome and the potential for gene regulation. *Nature* **447**, 413-417 (2007).
9 Zaratiegui, M., Irvine, D. V. & Martienssen, R. A. Noncoding RNAs and gene silencing. *Cell* **128**, 763-776 (2007).
10 Schwartz, Y. B. & Pirrotta, V. Polycomb silencing mechanisms and the management of genomic programmes. *Nat Rev Genet* **8**, 9-22 (2007).
11 Grewal, S. I. & Moazed, D. Heterochromatin and epigenetic control of gene expression. *Science (New York, N.Y* **301**, 798-802 (2003).
12 Henikoff, S. Nucleosome destabilization in the epigenetic regulation of gene expression. *Nat Rev Genet* **9**, 15-26 (2008).
13 Jirtle, R. L. & Skinner, M. K. Environmental epigenomics and disease susceptibility. *Nat Rev Genet* **8**, 253-262 (2007).
14 Hess, J. L. Mechanisms of transformation by MLL. *Critical reviews in eukaryotic gene expression* **14**, 235-254 (2004).
15 Hansen, R. S. *et al.* The DNMT3B DNA methyltransferase gene is mutated in the ICF immunodeficiency syndrome. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 14412-14417 (1999).
16 Dalgliesh, G. L. *et al.* Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature* **463**, 360-363 (2010).
17 Batty, N., Malouf, G. G. & Issa, J. P. Histone deacetylase inhibitors as anti-neoplastic agents. *Cancer letters* (2009).
18 Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).
19 Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* (2009).
20 Eckhardt, F. *et al.* DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature genetics* **38**, 1378-1385 (2006).

21 Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766-770 (2008).
22 Cokus, S. J. *et al.* Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452**, 215-219 (2008).
23 Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-837 (2007).
24 Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature methods* **6**, 283-289 (2009).
25 Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-560 (2007).
26 Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315-322 (2009).
27 Birney, E. *et al.* Prepublication data sharing. *Nature* **461**, 168-170 (2009).
28 Suzuki, M. M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* **9**, 465-476 (2008).
29 Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861 (2007).

Figures

Figure 1. Layers of genome organization. Genome function and cellular phenotypes are influenced by DNA methylation and the protein-DNA complex known as chromatin. In mammals, DNA methylation occurs on cytosine bases, primarily in the context of CpG dinucleotides. Accessible chromatin that is hyper-sensitive to DNase I digestion marks promoters and functional elements bound by transcription factors or other regulatory proteins. Histone modifications, associated proteins such as Polycomb repressors, and non-coding RNAs constitute an additional layer of chromatin structure that impacts genome function in a context-dependent manner.

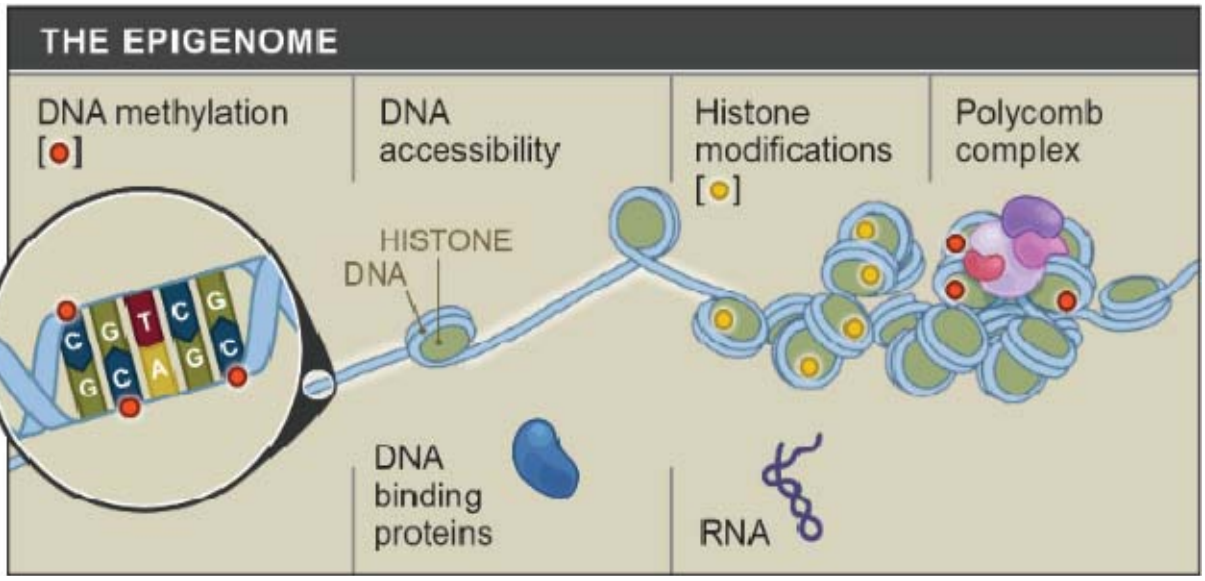
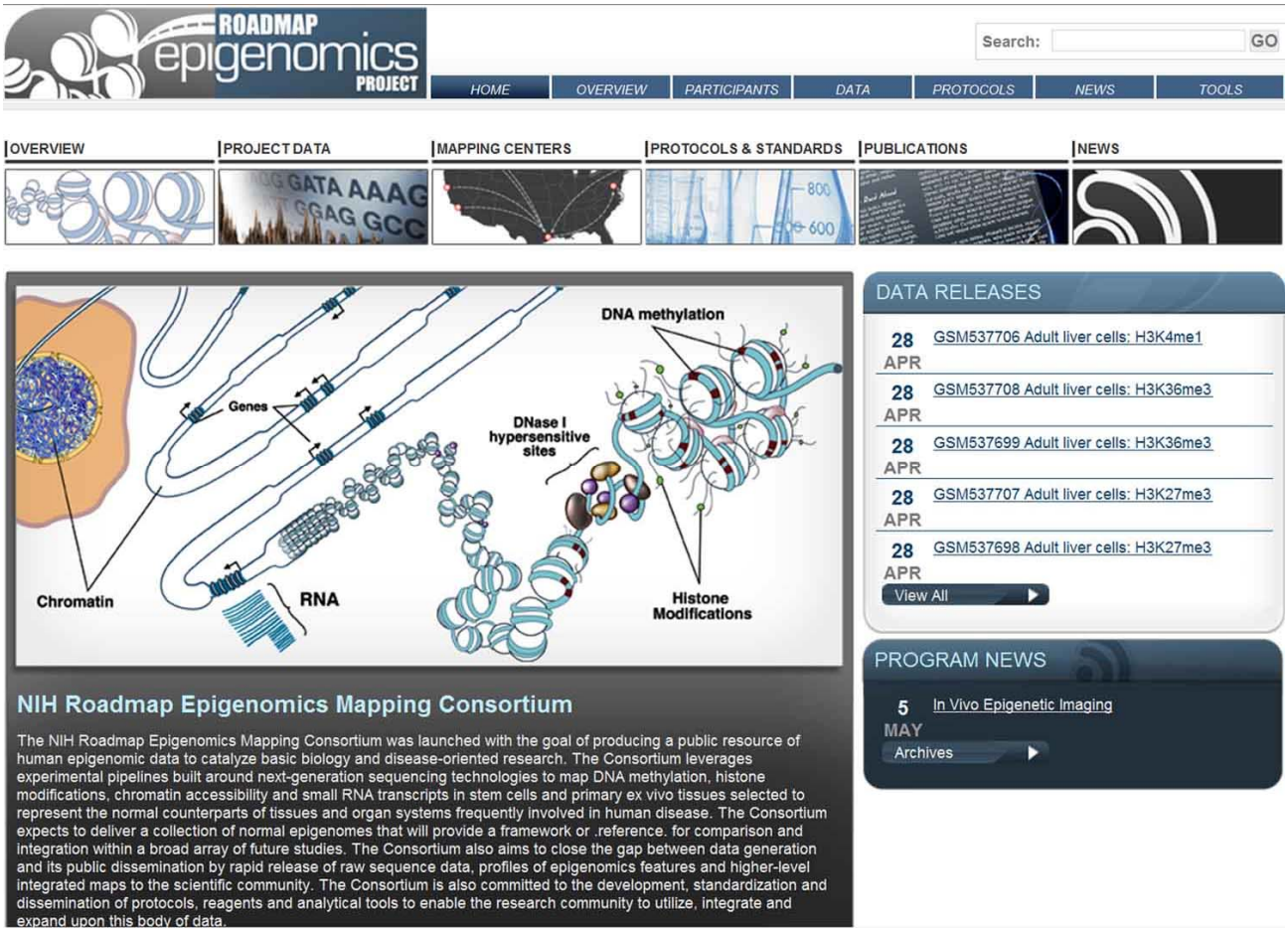


Figure 2. Portal for the NIH Roadmap Epigenomics Mapping Consortium. A public portal (www.roadmapepigenomics.org) provides general information about the consortium and its participants, along with links to experimental protocols, consortium data, and interfaces for visualizing epigenomic maps.



SubMAP: Aligning metabolic pathways with subnetwork mappings

Ferhat Ay*†‡ Manolis Kellis ‡ Tamer Kahveci†

Abstract

We consider the problem of aligning two metabolic pathways. Unlike traditional approaches, we do not restrict the alignment to one-to-one mappings between the molecules (nodes) of the input pathways (graphs). We follow the observation that in nature different organisms can perform the same or similar functions through different sets of reactions and molecules. The number and the topology of the molecules in these alternative sets often vary from one organism to another. With the motivation that an accurate biological alignment should be able to reveal these functionally similar molecule sets across different species, we develop an algorithm that first measures the similarities between different nodes using a mixture of homology and topological similarity. We combine the two metrics by employing an eigenvalue formulation. We then search for an alignment between the two input pathways that maximizes a similarity score, evaluated as the sum of the similarities of the mapped subsets of size at most a given integer k, and also does not contain any conflicting mappings. Here we prove that this maximization is NP-hard by a reduction from *Maximum Weight Independent Set (MWIS)* problem. We then convert our problem into an instance of MWIS and use an efficient vertex-selection strategy to extract the mappings that constitute our alignment. We name our algorithm **SubMAP** (**Sub**network **M**appings in **A**lignment of **P**athways). We evaluate its accuracy and performance on real datasets. Our empirical results demonstrate that SubMAP can identify biologically-relevant mappings that are missed by traditional alignment methods and, that it is scalable for metabolic pathways of arbitrary topology, including searching for a query pathway of size 70 against the complete KEGG database of 1,842 pathways.

Keywords. *metabolic pathway alignment, subnetwork mappings, one-to-many mappings, alternative reaction sets, maximum weight independent set*

†Computer and Information Science and Engineering, University of Florida, Gainesville, FL

‡MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA

*Correspondece author. E-mail: fay@cise.ufl.edu

**Implementation in C++ available at http://bioinformatics.cise.ufl.edu/SubMAP.html

1 Introduction

Biological pathways show how different molecules interact with each other to perform vital functions. In the literature, the terms *network* and *pathway* are used interchangeably for different types of interaction data. Metabolic pathways, an important class of biological pathways, represent how different compounds are transformed through various reactions. Analyzing these pathways is essential in understanding the machinery of living organisms.

The efforts on analyzing pathways can be classified into two types. The first type takes one pathway into account at a time and explores its important properties such as its robustness [1], steady states [2], modular structure [3, 4], network motifs [5, 6] as well as its representation [7, 8]. The second type is the comparative approach which considers multiple pathways to identify their frequent subgraphs [9, 10] and their alignments [11–21]. Alignment is a fundamental type of comparative analysis which aims to identify similar parts between pathways. For metabolic pathways, these similarities provide insights for drug target identification [22, 23], metabolic reconstruction of newly sequenced genome [24], phylogenic reconstruction [25, 26] and enzyme cluster and missing enzyme identification [27, 28].

In the literature, alignment is often considered as finding one-to-one mappings between the molecules of two pathways. In this case, the global/local pathway alignment problems are GI/NP complete as the graph/subgraph isomorphism problems can be reduced to them in polynomial time [29]. A number of studies have been done to systematically align different types of biological networks. For metabolic pathways, Pinter *et al.* [14] devised an algorithm that aligns query pathways with specific topologies by using a graph theoretic approach. Tohsato *et al.* proposed two algorithms for metabolic pathway alignment one relying solely on Enzyme Commission (EC [30]) numbers of enzymes and the other considering only the chemical structures of compounds of the query pathways [19, 20]. Latterly, Cheng *et al.* developed a tool, *MetNetAligner*, for metabolic pathway alignment that allows a certain number of insertions and deletions of enzymes [21]. However, these methods focus on a single type of molecule and the alignment is driven only by the similarities of these molecules (e.g., enzyme similarity, compound similarity). Furthermore, some of these methods limit the query pathways to certain topologies, such as trees, non-branching paths or limited cycles, which degrades their applicability to complex pathways. Recently, Singh *et al.* [15] and Ay *et al.* [17, 18] combined both topological features and homological

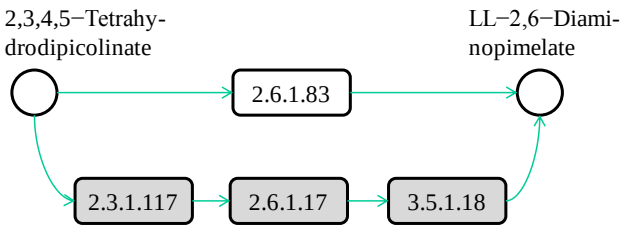


Figure 1: A portion of Lysine biosynthesis pathway. Each reaction is represented by the Enzyme Commission (EC) number of the enzyme that catalyze it. Circles represent compounds (intermediate compounds are not shown). *E.coli* and *H.sapiens* (human) use the path colored by grey with three reactions, whereas plants and *Chlamydia* achieve this transformation directly through the path with a single reaction shown in white.

similarity of pairwise molecules to find the alignments of protein interaction networks and metabolic pathways respectively. These algorithms showed that this combination increases the accuracy of the alignment. Another advantage of these two methods is that they do not restrict the topologies of query pathways and hence are applicable to arbitrarily complex pathways.

All the methods discussed above limit the possible molecule mappings to only one-to-one mappings. As also pointed out by Deutscher *et al.* [31] considering each molecule one by one fails to reveal its function(s) in complex pathways. This restriction prevents all the above methods from identifying biologically relevant mappings when different organisms perform the same function through varying number of steps. As an example, there are alternative paths for LL-2,6-Diaminopimelate production in different organisms [23, 32]. LL-2,6-Diaminopimelate is a key intermediate compound since it lies at the intersection of different paths on the synthesis of L-Lysine. Figure 1 illustrates two paths both producing LL-2,6-Diaminopimelate starting from 2,3,4,5-Tetrahydrodipicolinate. The upper path represents the shortcut used by plants and *Chlamydia* to synthesize L-Lysine. This shortcut is not an option, for example, for *E.coli* or *H.sapiens* due to the lack of the gene encoding LL-DAP aminotransferase (2.6.1.83). *E.coli* and *H.sapiens* have to use a three step process shown with grey path in Figure 1 to do this transformation. Thus, a meaningful alignment should map the two paths when, let say, the lysine biosynthesis pathways of human and a plant are aligned. However, since these two paths have different number of reactions traditional

alignment methods, limited to one-to-one mappings, fail to identify this mapping.

Our aim in this paper is to design an algorithm that can accurately identify such biologically relevant mappings by allowing one-to-many mappings of molecules. Note that, in Figure 1 both reaction sets form linear paths. It is possible to have reaction sets with different topologies performing a certain function (See Figure 4 for examples). Therefore, we use the term *subnetwork* to include all types of topologies. Also, since we only consider the sets of reactions that are connected, we will simply use the term *subnetwork* instead of *connected subnetwork*.

Problem definition: We consider the problem of aligning two metabolic pathways. Unlike traditional alignment approaches, we allow aligning a molecule of one pathway to a connected subnetwork of the other. More formally, let $\mathcal{P}$ and $\bar{\mathcal{P}}$ be two query pathways and k be a positive integer. We want to find the mapping between the molecules of $\mathcal{P}$ and $\bar{\mathcal{P}}$ with the largest alignment score, such that (1) each molecule in $\mathcal{P}$ ($\bar{\mathcal{P}}$) can map to a subnetwork of $\bar{\mathcal{P}}$ ($\mathcal{P}$) with at most k molecules and (2) each molecule can appear in at most one mapping.

The first condition above allows one-to-many mappings. The reason why we want to have one-to-many mappings in our alignment is not only that they capture functionally similar parts, but also they enable us to construct many-to-many mappings of arbitrary sizes. Identifying many-to-many mappings of molecule sets of different sizes is essential and is not possible with only one-to-one mappings as their combinations enforce both sides of a many-to-many mapping to be of the same size. The second condition enforces *consistency*. That is, if a molecule is already mapped alone or as a part of a subnetwork, it cannot map to another molecule. We elaborate on consistency and the problem definition later in Section 2. *Note that, allowing one-to-many mappings in alignment introduces new computational challenges (e.g., exponential increase in the problem size, conflicting mappings) that cannot be addressed using existing methods and hence novel methods are needed to tackle this problem.*

Contributions: In this paper, we propose a novel algorithm that finds subnetwork mappings in alignment of pathways. SubMAP accounts for both the effect of pairwise similarities (homology) and the organization of pathways (topology). This combination is motivated by its successful applications on pathway alignment by Singh *et al.* [15] and Ay *et al.* [17, 18]. However, al-

lowing one-to-many mappings makes it impossible to trivially extend these methods to our problem. To address this challenge, we map our problem to an eigenvalue problem. We solve this eigenvalue problem using an iterative technique called power method. The result of the power method converges to a principal eigenvector. This eigenvector defines a weighted bipartite graph where each node corresponds to a molecule or a subnetwork. The edges are only between two nodes from different pathways and their weights define the similarity of these nodes. Unlike the case of only one-to-one molecule mappings, the nodes on the same side of the bipartite graph can be intersecting as the same molecule can appear in more than one subnetwork. If two such intersecting nodes on one side are mapped to two different nodes of the other side, it will create inconsistent mappings for the elements of the intersection. We term such pair of mappings as *conflicting mappings*. Our aim is to extract a set of mappings that has no conflicting pairs and maximizes the sum of the similarity scores of the mappings. We prove that this maximization is NP-hard (See Theorem 1). We then construct a vertex weighted *conflict graph* with nodes representing a mapping of two subnetworks one from each pathway and edges representing a *conflict* between two mappings. The weights of the nodes in this graph are the edge weights of the bipartite graph from the earlier step. Our alignment problem is now equivalent to finding *maximum weight independent subset (MWIS)* of the conflict graph. To extract an independent set from the conflict graph, we use a vertex-selection strategy proposed for MWIS problem. We report the mappings that correspond to the selected set of nodes as the *alignment* of the query pathways. *Our experiments on the metabolic pathways from KEGG [33] database suggest that SubMAP finds biologically meaningful alignments efficiently. Also, SubMAP is scalable for subnetwork sizes up to three or four as it can align two real size metabolic pathways in a about a minute.*

The rest of the paper is organized as follows. Section 2 describes our algorithm. Section 3 presents experimental results. Section 4 concludes the paper.

2 Our Algorithm: SubMAP

In this section, we present our algorithm for pairwise metabolic pathway alignment that allows one-to-many molecule mappings. We begin by introducing some notation that we use throughout this section. Then, we formally

Table 1: Commonly used symbols in this paper

$\mathcal{P}$, $\bar{\mathcal{P}}$	Query metabolic pathways
k	Parameter for the largest subnetwork size
r_i, $\bar{r_j}$	Reactions of query pathways
R_i, $\bar{R}_j$	Subnetworks of query pathways
$\mathcal{R}_k$, $\bar{\mathcal{R}}_k$	Sets of all subnetworks with size at most k
n, m	Numbers of the reactions in query pathways
N_k, M_k	Numbers of all subnetworks of size at most k
φ_k	Set of all possible one-to-many mappings for a given k
$\lvert\varphi_k\rvert$	Number of all possible one-to-many mappings for a given k
S	Support matrix
H	Homological similarity vector
G_c	Conflict graph
α	Parameter adjusting relative weights of homology and topology

state the problem and describe the SubMAP algorithm in detail.

Let, $\mathcal{P}$ be a pathway which is represented by a directed unweighed graph $G = (V, E)$. Here, we only use the reactions of the pathway in graph representation. Hence, the vertex set $V = \{r_1, r_2, \ldots, r_n\}$ is the set of all reactions of $\mathcal{P}$. We include a directed edge e_{ij} from r_i to r_j in E if and only if at least one output compound of r_i is an input compound of r_j. We call r_i a *backward neighbor* of r_j and r_j a *forward neighbor* of r_i if $e_{ij} \in E$. Note that reactions can be reversible (bi-directional) and hence both e_{ij} and e_{ji} can exist.

A *subnetwork* of a pathway is a subset of its reaction set such that the induced undirected graph of the elements of this subset forms a connected graph. Let $R_i \subseteq V$ be such a subnetwork of $\mathcal{P}$. We define $\mathcal{R}_k$, the set of all subnetworks of $\mathcal{P}$ that have at most k reactions, as $\mathcal{R}_k = \{R_1, R_2, \ldots, R_{N_k}\}$ where $|R_i| \leq k$ for all $i \in [1, N_k]$. Here, $|R_i|$ denotes the cardinality of the reaction set R_i. Using this notation, we define a binary relation that maps a reaction of a query pathway to a subnetwork of the other as follows:

Definition 1 *Let $\mathcal{P}$ and $\bar{\mathcal{P}}$ be two pathways and k be a positive integer. Also, let $\mathcal{R}_k = \{R_1, R_2, \ldots, R_{N_k}\}$ and $\bar{\mathcal{R}}_k = \{\bar{R}_1, \bar{R}_2, \ldots, \bar{R_{M_k}}\}$ be the sets of subnetworks with size at most k of $\mathcal{P}$ and $\bar{\mathcal{P}}$ respectively. We define a binary relation φ between $\mathcal{R}_k$ and $\bar{\mathcal{R}}_k$ that allows one-to-many reaction mappings as $\varphi : \varphi \subseteq \varphi_k = (\mathcal{R}_1 \times \bar{\mathcal{R}}_k) \cup (\mathcal{R}_k \times \bar{\mathcal{R}}_1)$.* ∎

The number of all possible one-to-many mappings is:

$$|\varphi_k| = nM_k + mN_k - nm \tag{1}$$

where n, m are the number of reactions of $\mathcal{P}$ and $\bar{\mathcal{P}}$ respectively. The *alignment* of $\mathcal{P}$ and $\bar{\mathcal{P}}$ is a binary relation φ that is a subset of all these possible mappings and satisfies certain criteria that we describe next.

Recall that for a mapping $(R_i, \bar{R}_j) \in \varphi$ one of the R_i or $\bar{R}_j$ can contain more than one reaction. Reporting this mapping as a part of our alignment implies that all the reactions of the subnetwork with multiple reactions are aligned to a single reaction of the other. To have a *consistent alignment* none of the reactions of these subnetworks can be included in any other mapping. Next, we formally define the term *conflict* to characterize this property.

Definition 2 *Let φ be a binary relation and $R_i, R_u \in \mathcal{R}_k$ and $\bar{R}_j, \bar{R}_v \in \bar{\mathcal{R}}_k$. The distinct pairs $(R_i, \bar{R}_j) \in \varphi$ and $(R_u, \bar{R}_v) \in \varphi$* ***conflict*** *if and only if $(R_i \cap R_u) \cup (\bar{R}_j \cap \bar{R}_v) \neq \varnothing$.* ∎

Conflicts can cause inconsistencies about which reaction subset of one pathway should be aligned to the one of the other pathway. If φ has a conflicting pair of elements, we say φ is *inconsistent*. Since this is not a desirable property, *we limit our alignment to the consistent relations only.*

As well as discarding conflicting mappings we also need to use a meaningful scoring score in order to gather biologically relevant alignments. One standard scoring scheme for this purpose incorporates the homology of the aligned molecules with their topologies [15, 17, 18]. Here, we generalize this scheme to one-to-many mappings. We will elaborate on this similarity score later in Section 2.4. Next, we state our problem formally.

Problem formulation: Given k and two pathways $\mathcal{P}$ and $\bar{\mathcal{P}}$, let $\mathcal{R}_k$ and $\bar{\mathcal{R}}_k$ be the sets of subnetworks with size at most k of $\mathcal{P}$ and $\bar{\mathcal{P}}$ respectively. We want to find the *consistent* binary relation $\varphi \subseteq (\mathcal{R}_1 \times \bar{\mathcal{R}}_k) \cup (\mathcal{R}_k \times \bar{\mathcal{R}}_1)$ that *maximizes* the summation of the similarity scores of the aligned subnetworks.

In the following, we present our algorithm SubMAP. Section 2.1 explains the enumeration of the subnetworks of query pathways. Section 2.2 and 2.3 discuss homological and topological similarities respectively. Section 2.4 describes the eigenvalue formulation that combines these similarities and explains the extraction of subnetwork mappings.

2.1 Enumeration of Connected Subnetworks

The first step of SubMAP is to create the sets of all connected subnetworks of size at most k for each query pathway. Here, we describe the enumeration process for a single query pathway. Let $G = (V, E)$ represent a pathway and k be a positive integer. We construct the set of subnetworks $\mathcal{R}_k$ as follows. For $k = 0$, $\mathcal{R}_k = \mathcal{R}_0 = \varnothing$ and for $k = 1$, $\mathcal{R}_k = \mathcal{R}_1 = V$. For $k > 1$ we define $\mathcal{R}_k$ recursively by using $\mathcal{R}_{k-1}$. At each recursive step we check for each reaction in V if it can be added to already enumerated subnetworks of size $k-1$ to create a new connected subnetwork of size k. This way the kth recursive step takes $O(|V|.(|\mathcal{R}_{k-1}| - |\mathcal{R}_{k-2}|))$ time.

The size of the set $\mathcal{R}_k$ can be exponential in k when G is dense. However, metabolic pathways are usually sparse (on the average there are 2.5 forward neighbors per reaction). We observe that the number of subnetworks of the metabolic pathways in our dataset for $k = 3$ is around $5|V|$ and for $k = 4$ it is $10|V|$ on the average. In Section 3.2, we provide a detailed discussion of how $|\mathcal{R}_k|$ changes with different pathway sizes and different k values.

2.2 Homological Similarity of Subnetworks

Recall that the relation φ maps a reaction to a subnetwork that can contain multiple reactions. This necessitates computing the similarity between reaction sets. Since reactions are defined by their input and output compounds (i.e., substrates and products) and the enzymes that catalyze them, we measure the homological similarity between reactions using the similarities of these components.

In the literature, there are alternative pairwise similarity scores for compounds, enzymes and reactions. Particularly, two well known measure are information content similarity for enzyme pairs [14] and SIMCOMP [34] for compound pairs. We denote these measures by $SimE$ and $SimC$ respectively. We defer the readers to Ay *et al.* [18] for details on computing these similarities. Here, we utilize these similarity measures to compute the homological similarity between two reaction sets. To calculate this, we first construct the sets of the unions of input compounds (I_i), output compounds (O_i) and enzymes (E_i) of the reactions in each subnetwork R_i. For instance, in Figure 1 if we take upper path as the subnetwork R_i, then $E_i = \{2.3.1.117, 2.6.1.17, 3.5.1.8\}$. Let $\gamma_e, \gamma_i, \gamma_o$ denote the relative weights of

the similarities of enzymes, input and output compounds respectively. Then,

$$SimRSet(R_i, \bar{R}_j) = \\ \gamma_e W(E_i, \bar{E}_j, SimE) + \gamma_i W(I_i, \bar{I}_j, SimC) + \gamma_o W(O_i, \bar{O}_j, SimC) \quad (2)$$

Here W denotes the sum of edge weights of the pairs returned by the maximum weight bipartite matching (MWBM) of the two sets. MWBM finds an assignment between the nodes of two sets such that maximizes the sum of the weights of these assignments specified by the similarity score. We use $\gamma_i = \gamma_o = 0.3$ and $\gamma_e = 0.4$ as they provide a good balance between enzymes and compounds.

We calculate *SimRSet* for all possible one-to-many mappings between the subnetworks of two pathways. Therefore, we do this calculation $|\varphi_k|$ times in total. This way, we assess the homological similarities between all possible subnetwork mappings. Even though this scoring can be considered a good measure of similarity, relying solely on this score ignores the topology similarity which we explain next.

2.3 Topological Similarity of Subnetworks

The motivation for utilizing topological similarity is that the induced topologies of two aligned subnetworks should also be similar. In other words, if R_i is mapped to $\bar{R}_j$, then their neighbors in the corresponding pathways should also be similar. Motivated by this, we first extend the *neighborhood* definition of reactions to reaction subnetworks. Then, we introduce the notion of *support* between two mappings. Using these definitions, we formally describe how we calculate *the support matrix* and show it on an illustrative example.

Definition 3 *Let R_i, $R_u \in \mathcal{R}_k$. Then, R_u is a **forward neighbor** of R_i ($R_u \in FN(R_i)$) if and only if there exists $r_a \in R_i$ and $r_b \in R_u$ such that r_b is a forward neighbor of r_a or $R_i \cap R_u \neq \varnothing$. R_i is a **backward neighbor** of R_u ($R_i \in BN(R_u)$) if and only if R_u is a forward neighbor of R_i.* ∎

Definition 4 *Let $R_i, R_u \in \mathcal{R}_k$ and $\bar{R}_j, \bar{R}_v \in \bar{\mathcal{R}}_k$. The mapping ($R_i$, $\bar{R}_j$) **supports** the mapping (R_u, $\bar{R}_v$) if and only if both $R_j \in FN(R_i)$ and $\bar{R}_v \in FN(\bar{R}_u)$ or both $R_j \in BN(R_i)$ and $\bar{R}_v \in BN(\bar{R}_u)$.* ∎

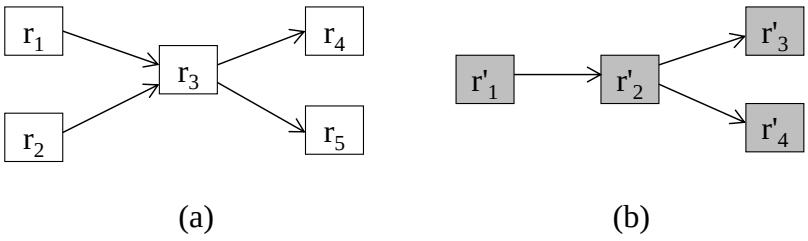


Figure 2: Reaction based representation of two hypothetical metabolic pathways. An arrow from r_i to r_j represents that an output of r_i is used as an input in r_j.

Definition 5 *Let $\mathcal{P}$, $\bar{\mathcal{P}}$ be two pathways and $\mathcal{R}_k$, $\bar{\mathcal{R}}_k$ be the sets of their subnetworks that have at most k reactions.* ***The support matrix*** *S is a $|\varphi_k| \times |\varphi_k|$ matrix with each entry $S[i,j][u,v]$ identifying the fraction of the total support provided by $(R_u, \bar{R}_v)$ mapping to $(R_i, \bar{R}_j)$ mapping. Let $\mathsf{N}(u,v) = |BN(R_u)||BN(\bar{R}_v)| + |FN(R_u)||FN(\bar{R}_v)|$ denote the number of all possible mappings between backward neighbors of R_u and $\bar{R}_v$ plus the ones between their forward neighbors. Then, each entry of S is computed as:*

$$S[i,j][u,v] = \begin{cases} \frac{1}{\mathsf{N}(u,v)} & \text{if } (R_u, \bar{R}_v) \text{ supports } (R_i, \bar{R}_j) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

■

Definition 4 states that the mapping of R_i to $\bar{R}_j$ favors all possible mappings of forward (backward) neighbors of R_i to those of $\bar{R}_j$. To see it on an example, let us consider the case when $k = 1$ and focus on the mapping $(\{r_3\}, \{r_2^{'}\})$ in Figure 2. We see that $FN(\{r_3\}) = 2$, $FN(\{r_2^{'}\}) = 1$, $BN(\{r_3\}) = 2$ and $BN(\{r_2^{'}\}) = 2$. Then, by Definition 5 we distribute the support of the mapping $(\{r_3\}, \{r_2^{'}\})$ to $2 \times 1 + 2 \times 2 = 6$ other mappings by placing $\frac{1}{6}$ in the corresponding entries of S.

There can be cases when one mapping does not provide support to any others. In such cases, we simply distribute its support equally to all possible mappings ($|\varphi_k|$). Notice that, by construction, the entries in each column of S sums up to 1. This is important as it ensures the stability and convergence of our algorithm as we explain in Section 2.4. Interested reader can find detailed description of the properties of support matrix in a previous work of ours [18].

Before moving into describing how we use the support matrix, it is important to explain how we create it in our implementation. Trivial but costly way of doing this is to check each mapping against all the others to calculate the support values. However, such an exhaustive strategy will require computing a huge matrix S of size $|\varphi_k| \times |\varphi_k|$. Since the creation of S will incur prohibitive computational costs, we do not construct this matrix literally. Instead, for each mapping $(R_i, \bar{R}_j)$, we take the sets $FN(R_i)$, $FN(\bar{R}_j)$ and $BN(R_i)$, $BN(\bar{R}_j)$ to generate only the pairs supported by $(R_i, \bar{R}_j)$. In other words, we use the sparse matrix form of the support matrix S.

2.4 Aligning Two Pathways

2.4.1 Combining Homology and Topology

Both the homological similarities of subnetworks and their topological organization provide us significant information for the alignment of metabolic pathways. A good alignment algorithm needs to combine these two factors in an efficient and accurate way. Here, we describe how we achieve this combination in SubMAP by using an iterative technique called *power method*.

Let k be a given parameter and $\mathcal{P}$, $\bar{\mathcal{P}}$ be two pathways with connected subnetwork sets $\mathcal{R}_k = \{R_1, R_2, \ldots, R_{N_k}\}$ and $\bar{\mathcal{R}}_k = \{\bar{R}_1, \bar{R}_2, \ldots, \bar{R_{M_k}}\}$ respectively. We represent the homological similarity of all subnetwork pairs by the column vector H of size $|\varphi_k|$. Each entry of H denotes the homological similarity between two subnetworks one from each pathway which corresponds to a mapping.

Let S be the $|\varphi_k| \times |\varphi_k|$ support matrix as described in Section 2.3. Given a parameter $\alpha \in [0, 1]$ to adjust the relative weights of homology and topology, we combine homology and topology through power method iterations as follows:

$$H^{i+1} = \alpha S H^i + (1 - \alpha) H^0 \tag{4}$$

In this equation, $H^0 = \frac{H}{||H||}$. We iterate this equation till $H^{i+1} = H^i$ (i.e., it converges). The resulting vector H^i is the principal eigenvector of the matrix $\alpha S + (1 - \alpha) H^0 e$ where e is a row vector of size $|\varphi_k|$ with all entries equal to 1. This system converges to a unique principal eigenvector when the matrices S and $H^0 e$ are both column stochastic. We ensure this since each column of S adds up to one and H^0 is normalized. Each entry of H^i gives us a combination of homological and topological similarities for the corresponding mapping. We use $\alpha = 0.6$ in this paper since in our previous

work we observed that this value provides a good combination of the two similarities [17, 18].

2.4.2 Extracting Subnetwork Mappings

Recall that our aim is to find an alignment that *maximizes* the summation of the similarity scores defined by H^i while preserving the consistency between different mappings. Here we first show that this maximization in NP-hard by a reduction from *maximum weight independent set (MWIS)* problem in bounded degree graphs. MWIS problem, even for graphs with largest degree 3, is NP-hard [35] and there is no constant factor approximation to the optimal solution in polynomial time [36, 37]. We then describe how we construct a conflict graph from our alignment problem and utilize a vertex-selection heuristic in order to extract the set of mappings that generates the alignment.

Theorem 1 (FINDING THE BEST ALIGNMENT IS NP-HARD) *Let $\mathcal{P}$ and $\bar{\mathcal{P}}$ be two pathways with reaction sets $\{r_1, \ldots, r_n\}$ and $\{\bar{r_1}, \ldots, \bar{r_m}\}$ respectively. Let $\mathcal{R} = \{R_1, R_2, \ldots, R_N\}$ and $\bar{\mathcal{R}} = \{\bar{R_1}, \bar{R_2}, \ldots, \bar{R_M}\}$ be all possible reaction subsets of $\mathcal{P}$ and $\bar{\mathcal{P}}$ with size at most a given positive integer k. Also, let $w : (\mathcal{R}, \bar{\mathcal{R}}) \rightarrow [0, 1] \cup \{-\infty\}$ be a similarity function defining the score for each mapping and the conflicts between mappings are defined according to Definition 2. Then, finding a set of mappings (i.e., alignment) that maximizes the sum of mapping scores (i.e., alignment score) and has no conflicting pairs is NP-hard.*

Proof:

We show this by a reduction from MWIS problem in bounded degree graphs. Let $G = (V, E, w)$ be a vertex weighted undirected graph with largest degree $k - 1$ (i.e., $k = \max_{i=1,\ldots,|V|} deg(v_i) + 1$). Let us set $n = |V| + |E|$, $m = |V|$. Then, $N = \sum_{i=1}^{k} \binom{n}{i}$ and $M = \sum_{i=1}^{k} \binom{m}{i}$ are the numbers of all possible subsets of size at most k for the sets with sizes n and m respectively. Since k is bounded N and M are polynomial in n and m. We want to construct a new graph $G' = (V', E', w')$ that is of polynomial size in the size of G through a polynomial time reduction such that the solution for MWIS problem on G is equivalent to the set of mappings extracted from G' that gives us the best alignment of $\mathcal{P}$ and $\bar{\mathcal{P}}$.

First, let us create a new vertex $v'_i \in V'$ for each $v_i \in V$ such that $v'_i = (R_i, \bar{R_i})$ where initially each $R_i = \{r_i\}$, each $\bar{R_i} = \{\bar{r_i}\}$ for all i from 1

to $|V|$. We ensure such r_i and $\bar{r_i}$ exist since $n = |V| + |E|$ and $m = |V|$. So far, we have not created any conflicts (i.e., edges) between any node pair (v'_i, v'_j) since $R_i \cap R_j = \bar{R}_i \cap \bar{R}_j = \varnothing$ for all $i \neq j$. Also, $R_i \in \mathcal{R}$ and $\bar{R}_i \in \bar{\mathcal{R}}$ since $|R_i| = |\bar{R}_i| = 1 \leq k$ for all i from 1 to $|V|$.

Next, for each $e_{ij} \in E$ we add $r_{h(i,j)}$ to both R_i ($R_i = R_i \cup r_{h(i,j)}$) and R_j ($R_j = R_j \cup r_{h(i,j)}$) so that we get a conflict between v'_i and v'_j. In order to ensure that we are only creating a conflict between the desired node pair (v'_i, v'_j) and no other pairs, $h(i, j)$ has to be a perfect hash function mapping each (i, j) to a distinct index between $|V| + 1$ and $n = |V| + |E|$. For this purpose, it suffices to simply choose $h(i, j) = |V| + (i - 1)|V| + j$. Also, for each vertex $v_i \in V$ we set $w'(v'_i) = w(v_i)$.

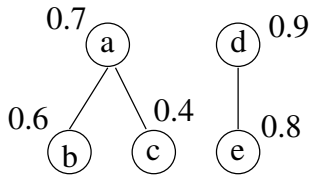
At this point, we have our graph G' isomorphic to G with their vertex weights having a one-to-one correspondence. Also, we have $\bar{R}_i \in \bar{\mathcal{R}}$ as $|\bar{R}_i| = 1 \leq k$ for each i from 1 to $|V|$. To have each $R_i \in \mathcal{R}$ as well, we need to make sure that $|R_i| \leq k$ for all i. Our construction starts with $|R_i| = 1$ for each i and adds one reaction to it for each undirected edge of v_i. Since G is a bounded degree graph with largest degree $k - 1$, we have at most $k - 1$ such additions and hence $|R_i| \leq k$.

Finally, we check for all possible mappings of the subsets, one from $\mathcal{R}$ and one from $\bar{\mathcal{R}}$, whether $(R_i, \bar{R}_j)$ exists as a node of G' or not. If this node does not exist, then we add it as an isolated node in G' with its weight set to $-\infty$ (i.e., $w'(R_i, \bar{R}_j) = -\infty$). After adding all such nodes, G' has $|V'| = NM$ vertices and $|E'| = |E|$ edges. Since k is bounded and N, M are polynomial in $n = |V| + |E|$ and $m = |V|$ respectively, the size of G' as well as the overall reduction is polynomial in the size of G. Furthermore, it can be seen from the construction that the maximum weight independent subset of G corresponds to the maximum weight independent subset of G', hence to the set of non-conflicting mappings that maximizes the overall alignment score. Therefore, finding the best alignment is NP-hard. ∎

Now that we proved extracting the best alignment is NP-hard, the next part is to describe how we tackle this problem. In the first step, we use the scores of mappings represented by H^i and the definition of conflict (Definition 2) to create a vertex weighted undirected graph $G_c = (V_c, E_c, w)$, which we name as *the conflict graph* as follows. We define a one-to-one correspondence from the mappings $(R_i, \bar{R}_j)$ to the vertices in V_c. We also set the weight of each vertex $a = (R_i, \bar{R}_j) \in V_c$ (i.e., $w(a)$) to the similarity between R_i and $\bar{R}_j$ as computed in H^i. We draw an undirected edge between two

	Subnetwork 1 ($\mathcal{P}$)	Subnetwork ($\bar{\mathcal{P}}$)	H^i
a:	$R_1 = \{r_1, r_2\}$	$\bar{R}_1 = \{\bar{r_1}\}$	0.7
b:	$R_2 = \{r_1\}$	$\bar{R}_2 = \{\bar{r_2}\}$	0.6
c:	$R_3 = \{r_3\}$	$\bar{R}_3 = \{\bar{r_1}\}$	0.4
d:	$R_4 = \{r_4\}$	$\bar{R}_4 = \{\bar{r_3}, \bar{r_4}, \bar{r_5}\}$	0.9
e:	$R_5 = \{r_4, r_5\}$	$\bar{R}_5 = \{\bar{r_5}\}$	0.8

(a)



(b)

Figure 3: (a) Each row corresponds to a possible mapping between subnetworks from two hypothetical metabolic pathways. The first column is the unique label for each mapping. Second and third columns are the reactions in the two subnetworks that can be mapped. The last column is the similarity between two subnetworks. (b) The conflict graph G_c for the mappings in (a).

vertices $a = (R_i, \bar{R}_j)$ and $b = (R_u, \bar{R}_v)$ if $(R_i \cap R_u) \cup (\bar{R}_j \cap \bar{R}_v) \neq \varnothing$ (i.e., a and b conflict). For instance, in Figure 3 there is an edge between a and b representing that they conflict since reaction r_1 is common to both a and b.

In the second step, we explain the greedy vertex-selection strategy we adopt from Sakai *et al.* [38] in order to extract the MWIS of G_c as our alignment. Let $N(v)$ denote the set of vertices that are connected to $v \in V_c$. At each iteration of this algorithm, we pick a vertex v that maximizes $f(v) = \sum_{\forall u_i \in N(v)} \frac{w(v)}{w(u_i)}$. This strategy implies that a vertex is more likely to be picked if the mapping it represents has large similarity score and conflicts with small number of other mappings with small similarity scores. After picking a vertex v, we put v into the result set and remove v and all the vertices connected to it (i.e., $v \cup N(v)$). We also remove all the edges incident to at least one of the removed vertices. When there are no more vertices to remove from G_c, the result set contains a maximal weight independent set. For our alignment problem, this vertex set corresponds to a list of non-conflicting subnetwork mappings. As an example, in Figure 3, d is the first vertex to be picked. Then, we remove d and $e \in N(d)$ from the graph and put d in the result set. Next, we pick the vertex b as $f(b) = \frac{0.6}{0.7} > f(a) = \frac{0.7}{0.6+0.4} > f(c) = \frac{0.4}{0.7}$. We remove b and $a \in N(b)$ and put b in the result set. Finally, only c is left and taking it into our result set, we have our alignment as the mappings $b = (r_1, \bar{r_2})$, $c = (r_3, \bar{r_1})$ and $d = (r_4, \{\bar{r_3}, \bar{r_4}, \bar{r_5}\})$.

3 Experimental Results

In this section, we experimentally evaluate the performance of SubMAP.

Dataset: We use the metabolic pathways of 20 organisms taken from the KEGG database. Our dataset contains 1,842 pathways in total. The average number of reactions per pathway is 21 and the largest pathway has 72 reactions. We also combined 12 different pathways under the metabolism of cofactors and vitamins for 10 different organisms. The average number of reactions of the largest connected components of these 10 pathways is 98 and the biggest one has 130 reactions.

3.1 Alternative Subnetworks

Different organisms can perform the same function through different subnetworks. We name such altered parts that have similar functions as *alternative subnetworks*. An accurate alignment should reveal alternative subnetworks in different pathways. In our first experiment we evaluate whether SubMAP can find them in real metabolic pathways. We align the pathway pairs which are known to contain functionally similar parts with different reaction sets and topologies. Table 2 presents a subset of reaction subnetwork mappings that are found by our algorithm. Figure 4 visualizes the topologies of these mappings by using an enzyme based representation.

The first row of Table 2 corresponds to alternative subnetworks in Figure 4(a) (also in Figure 1). The reaction R07613 represents the top path in Figure 4(a) that plants and Chlamydia use to produce LL-2,6- Diaminopimelate from 2,3,4,5- Tetrahydrodipicolinate. This path is discovered and reported as a shortcut on the L-Lysine synthesis path for plants and Chlamydia which is not present in humans or *E.coli* [23, 32]. Also, Watanabe *et al.* [23] suggested that since humans lack the catalyzer of the reaction R07613, namely LL-DAP aminotransferase (EC:2.6.1.83), this is an attractive target for the development of new drugs (antibiotics and herbicides). When we aligned the Lysine biosynthesis pathways of *H.sapiens* and *A.thaliana* (a plant), our algorithm mapped the reaction R07613 of *A.thaliana* to the three reactions that *H.sapiens* has to use to transform 2,3,4,5- Tetrahydrodipicolinate to LL-2,6- Diaminopimelate (R02734, R04365, R04475). In other words, SubMAP successfully identified the alternative subnetworks of different size (1 for *A.thaliana* and 3 for *H.sapiens*) that perform the same function.

Table 2: Alternative subnetworks that produce same or similar output compounds from the same or similar input compounds in different organisms. [1] Main input compound utilized by the given set of reactions. [2] Main output compound produced by the given set of reactions. [3] Reactions mappings that corresponds to alternative paths. Reactions are represented by their KEGG identifiers.

Pathway	Organisms	Input Comp.[1]	Output Comp.[2]	Reaction Mappings[3]
Lysine biosynthesis	***A.thaliana*** / ***H.Sapiens***	2,3,4,5-Tetra-hydrodipico.	LL-2,6-Di-aminopimelate	**R07613 ⇔ R02734 + R04365 + R04475**
Lysine biosynthesis	***A.thaliana*** / ***H.sapiens***	L-Saccharo. / meso-2,6-Di.	L-Lysine	**R00451 + R00715 + R00716 ⇔ R00451**
Pyruvate metabolism	***E.coli*** / ***H.sapiens***	Pyruvate	Oxaloacetate	**R00199 + R00345 ⇔ R00344**
Pyruvate metabolism	***E.coli*** / ***H.sapiens***	Oxaloacetate	Phosphoenol-pyruvate	**R00341 ⇔ R00431 + R00726**
Pyruvate metabolism	***T.acidophilum*** / ***A.tumefaciens***	Pyruvate	Acetyl-CoA	**R01196 ⇔ R00472 + R00216 + R01257**
Glycine,serine, threonine met.	***H.sapiens*** / ***R.norvegicus***	Glycine	Serine / L-Threonine	**R00945 ⇔ R00751 + R00945 + R06171**
Fructose and mannose met.	***E.coli*** / ***H.sapiens***	L-Fucose	L-Fucose 1-p / L-Fuculose 1-p	**R03163 + R03241 ⇔ R03161**
Citrate cycle	***S.aureus N315*** / ***S.aureus COL***	Isocitrate	2-Oxoglutarate	**R00268 + R01899 ⇔ R00709**
Citrate cycle	***H.sapiens*** / ***A.tumefaciens***	Succinate	Succinyl-CoA	**R00432 + R00727 ⇔ R00405**
Citrate cycle	***H.sapiens*** / ***A.tumefaciens***	Isocitrate / Citrate	2-Oxoglutarate / Oxaloacetate	**R00709 ⇔ R00362**

Another interesting example is the second row that is extracted from the same alignment described above. In this case, the three reactions that can independently produce L-Lysine for *A.thaliana* are aligned to the only reaction that produces L-Lysine for *H.sapiens* (Figure 4(b)). R00451 is common to both organisms and it utilizes meso-2,6- Diaminopimelate to produce L-Lysine. The reactions R00715 and R00716 take place and produce L-Lysine in *A.thaliana* in the presence of L-Saccharopine [39].

For the alignment of Pyruvate metabolisms of *E.coli* and *H.sapiens*, the third and fourth rows show two mappings that are found by SubMAP. The first one maps the two step process in *E.coli* that first converts Pyruvate to Orthophosphate (R00199) and then Orthophosphate to Oxaloacetate (R00345) to the single reaction that directly produces Oxaloacetate from Pyruvate (R00344) in *H.sapiens* (Figure 4(c)). The second one shows another mapping in which a single reaction of *E.coli* is replaced by two reactions of *H.sapiens* (Figure 4(d)). The first two rows for Citrate cycle also report similar mappings for other organism pairs (Figure 4(e)).

Note that all the above examples (also the rows from 6 to 9, Figure 4(f)-(i)) are one-to-many reaction mappings and hence a merit of the new al-

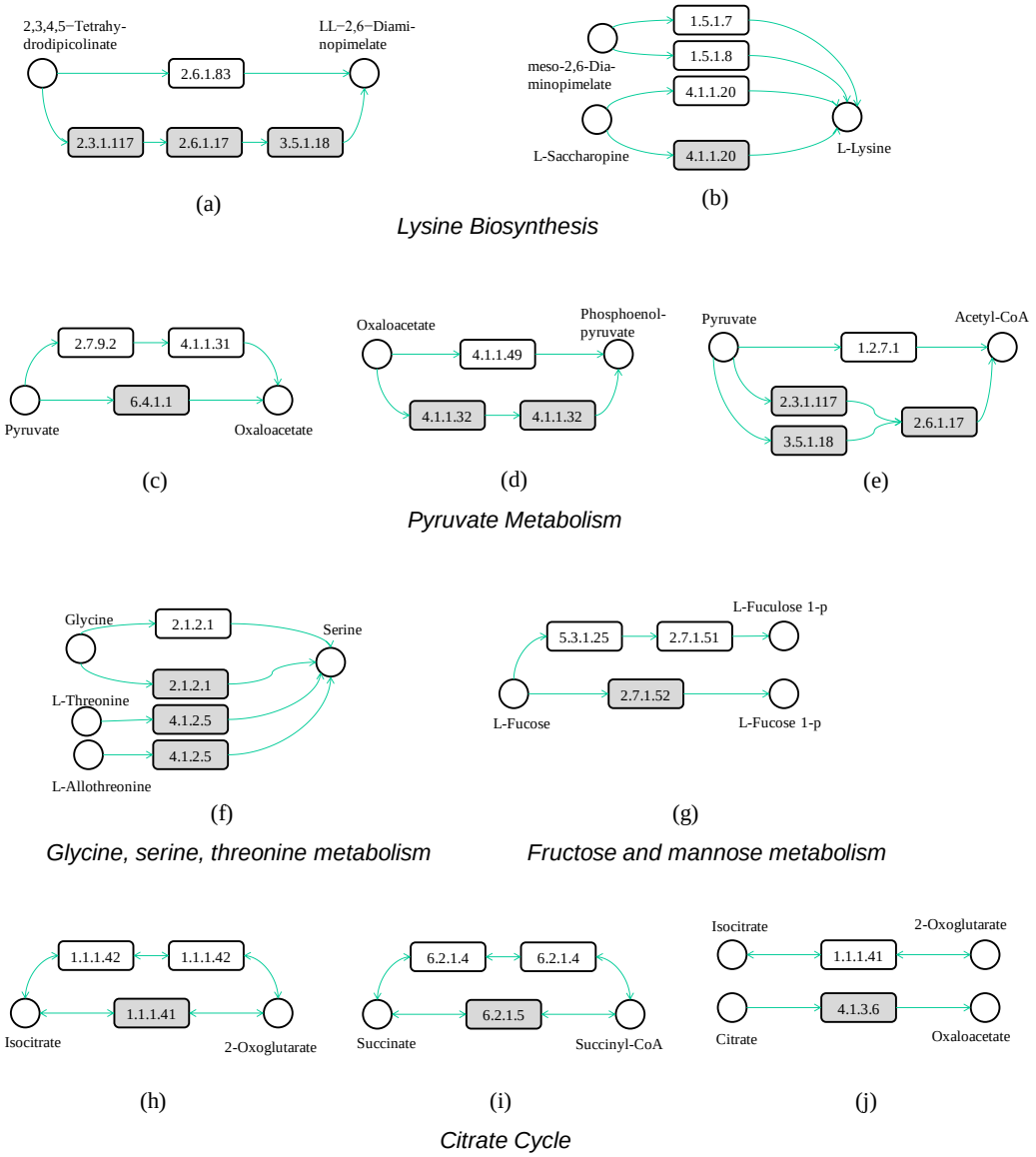


Figure 4: Visual representations of subnetwork mappings reported in Table 2. Figures (a) through (j) correspond to rows 1 through 10 of Table 2. Enyzmes are represented by their Enzyme Commission (EC) numbers [30].

gorithm we propose here. Our algorithm SubMAP also reports one-to-one mappings. The last row of Table 2 is an example in which one reaction of an organism is replaced by exactly one reaction of another organism. Aligning Citrate cycles of *H.sapiens* and *A.tumefaciens* reveals that even though both the input and output compounds of two reactions R00709 and R00362 are different SubMAP maps these reactions (Figure 4(j)). Also, if we look at the EC numbers of the enzymes catalyzing these reactions (1.1.1.41 and 4.1.3.6) their similarity is zero (see Information content enzyme similarity [18]). If we were to consider only the homological similarities, these two reactions could not have been mapped to each other. However, both these reactions are the neighbors of two other reactions R01325 and R01900 that are present in both organisms. The mappings of R01325 to R01325 and R01900 to R01900 support the mapping of their neighbors R00709 to R00362. Therefore, by incorporating the topological similarity our algorithm is able to find meaningful mappings with similar topologies and distinct homologies. An algorithm not considering pathway topologies would fail to identify such mappings.

These results suggest: (i) By allowing one-to-many mappings, our method identifies functionally similar subnetworks even if they have different number of reactions. (ii) Incorporation of topological similarity makes it possible to find mappings that can be missed by only considering homological similarity.

3.2 Number of Connected Subnetworks

Given the parameter k, our algorithm enumerates all connected reaction subnetworks of size at most k for each query pathway. One question that we need to answer is: How many such subnetworks exist? Figure 5 plots this average for different pathway sizes in our dataset. When $k = 1$, the figure shows the number of reactions. For $k > 1$ the results demonstrate that the number of subnetworks increase exponentially with k. However, the increase is significantly lower than the theoretical worst case $\sum_{i=1}^{k} \binom{n}{i}$ (i.e., n choose i). For instance, the number of subnetworks we obtained for $n = 72$ and $k = 5$ is around 750 times less than the theoretical worst case. For $n = 130$ and $k = 5$ the number of subnetworks is 0.027% of the worst case.

The figure also suggests that the number of subnetworks increase linearly with the size of the pathway. This is mainly because the average number of edges (i.e., neighbors) of a node (i.e., subnetwork) remains roughly same as the size of the network increases. As a result, we conclude that for $k \leq 4$, we can enumerate and store all the subnetworks of the pathways in KEGG

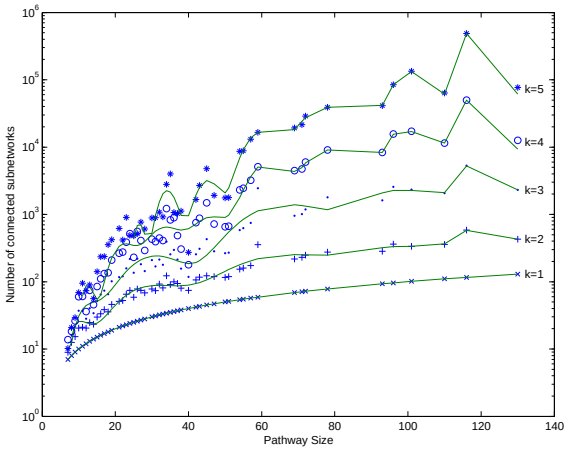


Figure 5: The number of subnetworks with at most k reactions for pathways of different sizes.

dataset. In practice it is unlikely for a single reaction to replace a subnetwork with more than 3 or 4 reactions. Therefore, we expect that using $k \leq 4$ would be sufficient to find most of the biologically relevant alternative subnetworks.

3.3 One-to-many Mappings within and across Major Clades

In Section 3.1, we demonstrated that our algorithm can find alternative subnetworks on a number of examples. An obvious question that follows is: How frequent are such alternative subnetworks and what are their characteristics? In other words, is there really a need to allow one-to-many mappings in alignment. In this experiment we aim to answer these questions.

We conduct an experiment as follows. We first pick 9 different organisms 3 from each major phylogenic clade. These organisms are:

T.acidophilum, *Halobacterium sp.*, *M.thermoautotrophicum* from Archaea;
H.sapiens, *R.norvegicus*, *M.musculus* from Eukaryota;
E.coli, *P.aeruginosa*, *A.tumefaciens* from Bacteria.

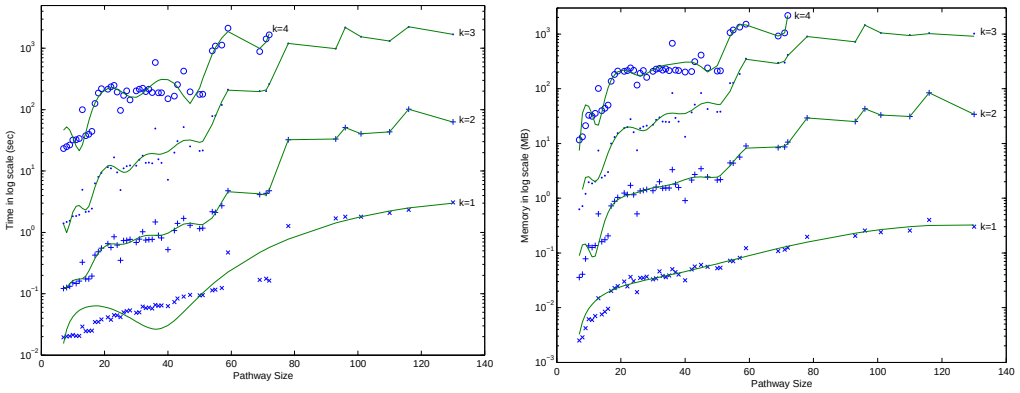
We then extract 10 common pathways for these 9 organisms from KEGG. For each of these common pathways, we choose all possible pairs of the 9 organisms ($\binom{9}{2} = 36$) and align that specific pathway for all organism pairs. In these alignments we exclude the self alignments and the alignment with

Table 3: Percentages of 1-to-1, 1-to-2, 1-to-3 and 1-to-4 mappings in between and across three major clades. (**A**: Archaea, **E**: Eukaryota, **B**: Bacteria).

	1-to-1	**1-to-2**	**1-to-3**	**1-to-4**
E-E	89.6	8.8	1.1	0.5
B-B	80.1	16.0	3.1	0.8
A-A	78.3	15.7	4.7	1.3
B-E	69.1	23.1	6.3	1.5
A-B	60.5	28.3	8.5	2.7
A-E	55.8	31.0	10.4	2.8

parameter $k = 1$ since those will definitely incur a bias favoring the number of one-to-one alignments. We computed all possible alignments ($10 \times 36 = 360$) for $k = 2$, 3 and 4 ($360 \times 3 = 1{,}080$ alignments in total). Finally, we calculated the number of four possible types of subnetwork mappings which are 1-to-1, 1-to-2, 1-to-3 and 1-to-4. We hypothesize that the metabolisms of the organisms within a clade will tend to perform the same function through the same (or similar) sized sets of reactions while those across different clades will perform from alternative subnetworks of varying sizes.

Table 3 summarizes the results of this experiment. The percentages of each mapping type between two clades is shown as a row in this table. The first three rows corresponds to alignments within a clade and the last three represents alignments across two different clades. An important outcome of these results is that there are considerably large number of one-to-many mappings between organisms of different clades. In the extreme case (last row), nearly half of the mappings are one-to-many. The results also support our hypothesis that one-to-one mappings is more frequent for alignments within the clades compared to across clades due to high similarity between the organisms of the same clade. For instance, for both the first and last row one side of the query set is the Eukaryota. However, going from first row to last, we see around 40% decrease in the number of one-to-one mappings and 250%, 850% and 450% increase in the number of 1-to-2, 1-to-3 and 1-to-4 mappings respectively. Considering Archaea are single-cell microorganisms (e.g., Halobacteria) and Eukaryota are complex organisms with cell membranes (e.g., animals and plants), these jumps in the number of one-to-many mappings suggest that the individual reactions in Archaea are replaced by a number of reactions in Eukaryota. These results have two



(a) (b)

Figure 6: Average running time and memory use of SubMAP when a query pathway is aligned with a database containing 50 pathways. Pathway size is measured by the number of reactions.

major implications. *(i) One-to-many mappings are frequent in nature. To obtain biologically meaningful alignments we need to allow such mappings. (ii) The characteristics of the alterative subnetworks can help in inferring the phylogenic relationship among different organisms.*

3.4 Running Time and Memory Utilization

SubMAP allows one-to-many mappings to find biologically relevant alignments. This however comes at the expense of increased computational cost. Theoretically, this increase can be exponential in k. The worst case happens when the pathway is highly connected. Metabolic pathways however are sparse and their connectivity follows power law distribution [40]. In order to understand the capabilities and limitations of our method we examine its performance on real datasets in terms of its running time and memory usage.

We evaluate the performance of our method for querying a database of pathways as follows. We create a query set by selecting 50 pathways of varying sizes from KEGG and adding it the combined pathways for cofactors and vitamins metabolism of 10 different organisms as described at the beginning of this section. We then select another 50 pathways of different sizes to use as our database set for this experiment. We pick the latter 50 pathways such that the average reactions per pathway is 21.4, which is very close to that of

the entire database. We then align each of the 60 query pathways with all the database pathways one by one for different values of k. We measure the average running time and the average memory usage for each query pathway and k value combination. Note that we do not present any performance comparison with an existing method as the existing methods do not allow one-to-many mappings. However, our results for $k = 1$, shows the performance of our algorithm when we restrict it to one-to-one mappings and can be compared with traditional alignment methods.

Figure 6(a) shows the average running time of SubMAP for query pathways with increasing number of reactions. When $k = 1$ (i.e., only one-to-one mappings as in existing methods), it runs in less than a few seconds even for the largest query pathway in our query set. As k increases, the running time increases significantly. This is because the number of subnetworks and the average number of forward and backward neighbors of subnetworks increase with k. We observe that our method can perform alignments in practical time even when $k = 4$ for pathways with around 70 reactions. Hence, it is scalable for all the individual pathways in KEGG. However, for the 10 combined pathways of cofactors and vitamins metabolism, the running time and memory use becomes a bottleneck when we consider $k > 3$.

We also measure the actual memory usage of our algorithm for real pathways of varying sizes and k values in Figure 6(b). For $k = 1$ or 2, the memory usage is negligible (100 MB or less) for all pathways. Although the memory usage increases quickly with k, it remains feasible for query pathways with around 70 reactions for $k = 4$. For $k = 3$, the biggest pathway of 130 reactions requires 1GB of memory per query on the average. *These results show that, SubMAP can run on a standard computer for aligning real-sized metabolic pathways for* $k \leq 4$.

4 Conclusion

In this paper, we considered the problem of aligning two metabolic pathways. The distinguishing feature of our work from the literature is that we allow mapping one molecule of one pathway to a set of molecules of the other. To address this problem, given two metabolic pathways $\mathcal{P}$ and $\bar{\mathcal{P}}$ and an upper bound k on the size of the connected subnetworks, we developed the SubMAP algorithm that can find the consistent mapping of the subnetworks of $\mathcal{P}$ and $\bar{\mathcal{P}}$ with the maximum similarity. We transformed the alignment problem to an eigenvalue problem. The solution to this eigenvalue problem

produced a good mixture of homological and topological similarities of the subnetworks. Using these similarity values, we constructed a vertex weighted graph that connects conflicting mappings with an edge. Then, our alignment problem is transformed into finding the maximum weight independent subset of this graph. We employed a heuristic method that is used to solve MWIS problem. The result of this method provided us an alignment that has no conflicting pair of mappings (i.e., consistent). Our experiments on real datasets suggested that our method can identify biologically relevant mappings of alternative subnetworks that are missed by traditional alignment methods. Furthermore, even though SubMAP does not restrict the topologies of query pathways, it is still scalable for real size metabolic pathways when the reaction subsets of size at most four are considered.

References

[1] Edwards JS and Palsson BO. Robustness analysis of the Escherichia coli metabolic network. *Biotechnology Progress*, 16:927–939, 2000.

[2] Ay F, Xu F, and Kahveci T. Scalable Steady State Analysis of Boolean Biological Regulatory Networks. *PLoS ONE*, 4(12):e7992, 2009.

[3] Schuster S, Pfeiffer T, Koch I, Moldenhauer F, and Dandekar T. Exploring the Pathway Structure of Metabolism: Decomposition into Subnetworks and Application to Mycoplasma pneumoniae. *Bioinformatics*, 18:351–361, 2002.

[4] Ay F, Dinh TN, and Thai MT Kahveci T. Finding dynamic modules of biological regulatory networks. In *BIBE*, 2010.

[5] Wernicke S and Rasche F. FANMOD: a tool for fast network motif detection. *Bioinformatics*, 22(9):1152–3, 2006.

[6] Grochow JA and Kellis M. Network motif discovery using subgraph enumeration and symmetry-breaking. In *RECOMB*, pages 92–106, 2007.

[7] Michal G. On representation of metabolic pathways. *Biosystems*, 47(1-2):1–7, 1998.

[8] Babur O, Dogrusoz U, Demir E, and Sander C. ChiBE: interactive visualization and manipulation of BioPAX pathway models. *Bioinformatics*, 26(3):429–31, 2010.

[9] Koyuturk M, Grama A, and Szpankowski W. An efficient algorithm for detecting frequent subgraphs in biological networks. In *ISMB*, pages 200–207, 2004.

[10] Qian X and Yoon B. Effective Identification of Conserved Pathways in Biological Networks Using Hidden Markov Models. *PLoS ONE*, 4(12):e8070, 2009.

[11] Berg J and Lassig M. Local graph alignment and motif search in biological networks. *PNAS*, 101(41):14689–94, 2004.

[12] Berg J and Lassig M. Cross-species analysis of biological networks by Bayesian alignment. *PNAS*, 103(29):10967–72, 2006.

[13] Dandekar T, Schuster S, Snel B, Huynen M, and Bork P. Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochemistry Journal*, 343:115–124, 1999.

[14] Pinter RY, Rokhlenko O, Yeger-Lotem E, and Ziv-Ukelson M. Alignment of metabolic pathways. *Bioinformatics*, 21(16):3401–8, 2005.

[15] Singh R, Xu J, and Berger B. Pairwise Global Alignment of Protein Interaction Networks by Matching Neighborhood Topology. In *RECOMB*, pages 16–31, 2007.

[16] Koyuturk M, Grama A, and Szpankowski W. Pairwise Local Alignment of Protein Interaction Networks Guided by Models of Evolution. In *RECOMB*, pages 48–65, 2005.

[17] Ay F, Kahveci T, and de Crecy-Lagard V. Consistent alignment of metabolic pathways without abstraction. In *Computational Systems Bioinformatics Conference (CSB)*, volume 7, pages 237–48, 2008.

[18] Ay F, Kahveci T, and de Crecy-Lagard V. A fast and accurate algorithm for comparative analysis of metabolic pathways. *Journal of Bioinformatics and Computational Biology (JBCB)*, 7(3):389–428, 2009.

[19] Tohsato Y and Nishimura Y. Metabolic Pathway Alignment Based on Similarity of Chemical Structures. *Information and Media Technologies*, 3:191–200, 2008.

[20] Tohsato Y, Matsuda H, and Hashimoto A. A Multiple Alignment Algorithm for Metabolic Pathway Analysis Using Enzyme Hierarchy. In *ISMB*, pages 376–383, 2000.

[21] Cheng Q, Harrison R, and Zelikovsky A. MetNetAligner: a web service tool for metabolic network alignments. *Bioinformatics*, 25(15):1989–90, 2009.

[22] Sridhar P, Kahveci T, and Ranka S. An iterative algorithm for metabolic network-based drug target identification. In *Pacific Symposium on Biocomputing (PSB)*, volume 12, pages 88–99, 2007.

[23] Watanabe N, Cherney MM, van Belkum MJ, Marcus SL, Flegel MD, Clay MD, Deyholos MK, Vederas JC, and James MN. Crystal structure of LL-diaminopimelate aminotransferase from Arabidopsis thaliana: a recently discovered enzyme in the biosynthesis of L-lysine by plants and Chlamydia. *Journal of Molecular Biology*, 371(3):685–702, 2007.

[24] Francke C, Siezen RJ, and Teusink B. Reconstructing the metabolic network of a bacterium from its genome. *Trends in Microbiology*, 13(11):550–558, 2005.

[25] Clemente JC, Satou K, and Valiente G. Reconstruction of Phylogenetic Relationships from Metabolic Pathways Based on the Enzyme Hierarchy and the Gene Ontology. *Genome Informatics*, 16(2):45–55, 2005.

[26] Heymans M and Singh A. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinformatics*, 19:138–146, 2003.

[27] Ogata H, Fujibuchi W, Goto S, and Kanehisa M. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Research*, 28:4021–8, 2000.

[28] Green ML and Karp PD. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, 5:76, 2004.

[29] Damaschke P. Graph-Theoretic Concepts in Computer Science. *Lecture Notes in Computer Science*, 484:72–78, 1991.

[30] Webb EC. *Enzyme nomenclature 1992* . Academic Press, 1992.

[31] Deutscher D, Meilijson I, Schuster S, and Ruppin E. Can single knockouts accurately single out gene functions? *BMC Systems Biology*, 2:50, 2008.

[32] McCoy AJ, Adams NE, Hudson AO, Gilvarg C, Leustek T, and Maurelli AT. L,L-diaminopimelate aminotransferase, a trans-kingdom enzyme shared by Chlamydia and plants for synthesis of diaminopimelate/lysine. *PNAS*, 103(47):17909–14, 2006.

[33] Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, and Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27(1):29–34, 1999.

[34] Hattori M, Okuno Y, Goto S, and Kanehisa M. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *Journal of the American Chemical Society (JACS)*, 125(39):11853–65, 2003.

[35] Lovasz L. Stable set and polynomials. *Discrete Mathematics*, 124:137–53, 1994.

[36] Austrin P, Khot S, and Safra M. Inapproximability of Vertex Cover and Independent Set in Bounded Degree Graphs. In *IEEE Conference on Computational Complexity*, pages 74–80, 2009.

[37] Berman P and Karpinski M. On some tighter inapproximability results. *Lecture notes in Computer Science*, 1999.

[38] Sakai S, Togasaki M, and Yamazaki K. A note on greedy algorithms for the maximum weighted independent set problem. *Discrete Applied Mathematics*, 126:313–22, 2003.

[39] Saunders PP and Broquist HP. Saccharopine, an intermediate of aminoadipic acid pathway of lysine biosynthesis. *Journal of Biological Chemistry*, 241:3435–40, 1966.

[40] Jeong H, Tombor B, Albert R, Oltvai ZN, and Barabasi AL. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–4, 2000.

Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes

Michael F. Lin,[1,2] Pouya Kheradpour,[1,2] Stefan Washietl,[2] Brian J. Parker,[3] Jakob S. Pedersen,[3,5] and Manolis Kellis[1,2,4,6]

[1]*Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA;* [2]*Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA;* [3]*Department of Biology, University of Copenhagen, Ole Maaloes Vej 5, DK-2200 Copenhagen, Denmark;* [4]*The Broad Institute, Cambridge, Massachusetts 02139, USA*

The degeneracy of the genetic code allows protein-coding DNA and RNA sequences to simultaneously encode additional, overlapping functional elements. A sequence in which both protein-coding and additional overlapping functions have evolved under purifying selection should show increased evolutionary conservation compared to typical protein-coding genes—especially at synonymous sites. In this study, we use genome alignments of 29 placental mammals to systematically locate short regions within human ORFs that show conspicuously low estimated rates of synonymous substitution across these species. The 29-species alignment provides statistical power to locate more than 10,000 such regions with resolution down to nine-codon windows, which are found within more than a quarter of all human protein-coding genes and contain ~2% of their synonymous sites. We collect numerous lines of evidence that the observed synonymous constraint in these regions reflects selection on overlapping functional elements including splicing regulatory elements, dual-coding genes, RNA secondary structures, microRNA target sites, and developmental enhancers. Our results show that overlapping functional elements are common in mammalian genes, despite the vast genomic landscape.

[Supplemental material is available online at http://www.genome.org.]

It is often assumed that synonymous sites within protein-coding open reading frames (ORFs) evolve neutrally, since mutations in them do not change the amino acid translation. But, in fact, ORFs in many species simultaneously encode additional functional sequence elements within the codon sequence, often with strong evolutionary constraint on the synonymous sites (Chamary et al. 2006; Itzkovitz and Alon 2007). For example, mammalian ORFs are known to encode exonic splicing enhancers and silencers (Chen and Manley 2009), microRNA target sites (Lewis et al. 2005; Hurst 2006), A-to-I recoding sites (Rueter et al. 1999; Bass 2002), and transcriptional enhancers (Lang et al. 2005; Nguyen et al. 2007; Lampe et al. 2008; Tümpel et al. 2008; Dong et al. 2010). Several previous studies have observed strong genome-wide trends toward increased evolutionary constraint on such overlapping functional elements, by averaging across many loci pooled together (Baek and Green 2005; Xing and Lee 2005; Chen et al. 2006; Down et al. 2006; Goren et al. 2006; Parmley et al. 2006; Robins et al. 2008; Kural et al. 2009). However, these averaging approaches generally did not have the power to locate individual overlapping functional elements within specific genes.

In this study, we use the unprecedented discovery power provided by alignments of 29 mammalian genomes to provide a systematic annotation of individual functional elements embedded within protein-coding regions throughout the human genome. Since the average codon site in these multiple sequence alignments shows about four synonymous substitutions, we predict that overlapping functional elements will individually stand out as short, localized regions with exceptionally few synonymous substitutions—in much the same way that widely used methods such as GERP, phastCons, phyloP, and SiPhy locate conserved functional elements within a background of neutral nucleotide-level sequence evolution (Cooper et al. 2005; Siepel et al. 2005; Margulies et al. 2007; Garber et al. 2009; Pollard et al. 2010).

However, detecting overlapping evolutionary constraints within protein-coding ORFs is more difficult than detecting general nucleotide-level constraints, for two main reasons. First, since the majority of nucleotide sites in a typical human ORF are already highly conserved among mammals due to their protein-coding function, we must expect less statistical power to detect increased conservation for overlapping sequence elements of a given length. Second, it is important to account precisely for the protein-coding constraints on each nucleotide site by modeling the evolutionary process at the codon level, rather than analyzing the conservation of individual nucleotide sites independently of those surrounding them.

To address these challenges, we present a novel adaptation of statistical phylogenetic codon models widely used in evolutionary analysis of protein-coding genes (Anisimova and Kosiol 2009; for recent review, see Delport et al. 2009), which locates short windows within alignments of known human ORFs showing significantly reduced rates of synonymous substitution. Applying this new method to the 29-species alignments, we confidently locate more than 10,000 such regions, typically with 70%–90% reduced synonymous rates, down to a resolution of just nine codons. These putative "synonymous constraint elements" contain only ~2% of

[5]Present address: Department of Molecular Medicine, Aarhus University Hospital, Skejby, Brendstrupgaardsvej 100, DK-8200 Aarhus N, Denmark.
[6]Corresponding author.
E-mail manoli@mit.edu.
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.108753.110.

all synonymous sites, but are found within more than a quarter of all human protein-coding genes. We also present numerous lines of evidence that they indeed play diverse functional roles in several biological processes such as splicing and translational regulation, dual-coding regions, RNA secondary structures, miRNA targeting, and developmental enhancers.

A few previous studies have also sought to locate individual overlapping functional elements in human genes based on their increased conservation, but their power was limited compared to what should now be possible by comparing the 29 mammals. Schattner and Diekhans (2006) analyzed pairwise alignments of human and mouse ORFs to identify about 200 regions of at least 60 codons tolerating at most one synonymous substitution. Similarly, many of the "ultraconserved elements" of Bejerano et al. (2004), stretches of at least 200 bp perfectly conserved between human and mouse, overlap known coding regions. Such pairwise species comparisons allow for much simpler statistical models, but their resolution of 180–200 nt seems so long (considerably longer than the typical exon length of ~120 nt) that many shorter overlapping functional elements were probably averaged out, even if they are highly conserved. Other related studies, including some multispecies approaches, have analyzed only a small fraction of mammalian genes (Hurst and Pal 2001; Chen and Blanchette 2007; Parmley and Hurst 2007b; Lin et al. 2008), while we undertake a comprehensive genome-wide analysis. Lastly, a few methods have been developed to identify examples of certain known classes of overlapping functional elements with predictable evolutionary signatures, including dual-coding ORFs (Chung et al. 2007; Ribrioux et al. 2008) and RNA secondary structures (Pedersen et al. 2004a,b). These are complementary to the rate-based approach we take here.

Estimating synonymous substitution rates in short windows within open reading frames

Our method uses phylogenetic codon models to find short windows within multi-species alignments of known ORFs that exhibit unusually low rates of synonymous substitution as measured by d_S, a composite rate commonly used to summarize the relevant parameters of such models (Yang and Bielawski 2000). Specifically, our method analyzes any window of adjacent codon sites within an alignment to compute the maximum likelihood estimate of a parameter λ_s, which is a scale factor on d_S, indicating how much slower or faster synonymous substitutions have occurred in that window relative to a null model representing typical protein-coding sequence evolution. For example, a particular window with $\lambda_s = 0.5$ is estimated to have evolved with a synonymous substitution rate only one-half that of the null model average (window $d_S = \lambda_s \times$ average d_S) (Fig. 1A). Our method is designed to estimate λ_s for short windows of adjacent sites—in this study, nine to 30 codons. (The method also estimates λ_n, the analogous parameter for non-synonymous substitutions, which we use to exclude potentially misaligned regions.)

Furthermore, we can associate a statistical significance with our estimate of λ_s for any window, using standard techniques for testing the goodness-of-fit of statistical phylogenetic models. Specifically, we can perform a likelihood ratio test (LRT) to assess precisely whether a small estimate of λ_s explains an observed alignment window better than the average rates assumed by the

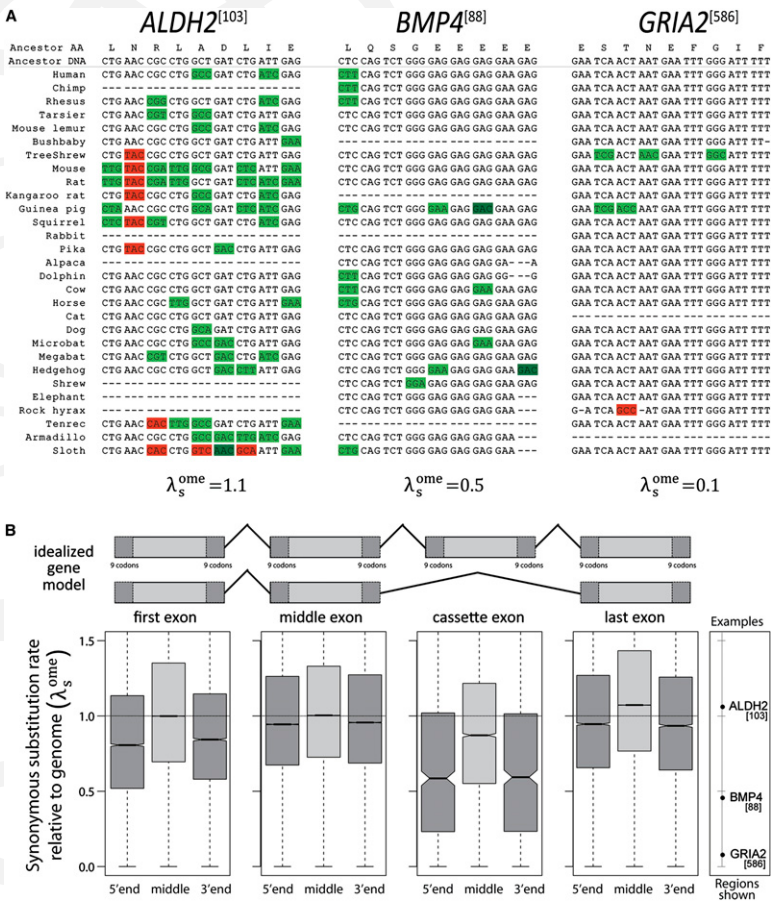


Figure 1. (*A*) Examples of local synonymous rate variation in alignments of 29 placental mammals for short nine-codon windows within the open reading frames (ORFs) of three known human protein-coding genes—*ALDH2*, *BMP4*, and *GRIA2*—with brackets denoting starting codon position within each ORF of shown alignment. (Bright green) Synonymous substitutions with respect to the inferred ancestral sequence; (dark green) conservative amino acid substitutions; (red) other nonsynonymous substitutions. The estimated parameter λ_s^{ome} denotes the rate of synonymous substitution within these selected windows relative to genome-wide averages. For example, the nine-codon window starting at codon 88 of the *BMP4* ORF shows $\lambda_s^{ome} = 0.5$, corresponding to an estimated synonymous substitution rate 50% below the genome average. (*B*) Variation in the estimated synonymous rate at different positions with respect to exon boundaries and translation start and stop, across all CCDS ORFs. For each class of regions, box-and-whisker plots show the observed distribution of λ_s^{ome}, including the median (*middle* horizontal bars), middle 50% range (boxes), extreme values (whiskers), and whether medians differ with high statistical confidence (nonoverlapping notches between two boxes). Estimated synonymous rates tend to be significantly reduced at the 5′ and 3′ ends of exons, and dramatically reduced in alternatively spliced exons, likely reflecting widespread splicing regulatory elements embedded within protein-coding regions.

Fig(s). 1 live 4/C

null model. Then, if a window has a significantly reduced λ_s according to this test, we infer that its synonymous sites have probably been constrained by natural selection acting on an overlapping functional element. This is very similar to likelihood methods for nucleotide-level constraint detection (Garber et al. 2009; Pollard et al. 2010), extended to codon models so that we can disentangle the different evolutionary pressures on synonymous and non-synonymous sites.

The LRT provides an elegant way to avoid certain potential pitfalls in detecting individual regions with reduced synonymous rates. For example, it accounts for the uncertainty in rate estimates based on the exact set of informant species aligned for each window. To illustrate, consider two windows, one with estimated λ_s = 0.5 with all 29 species aligned, and the other with λ_s = 0.1 but only a human/chimpanzee alignment available. Even though the estimated λ_s is lower in the second window, it is almost surely less significant by the LRT, because it is based on far less informative data. The LRT also accounts for the expected constraint at each individual site based on the amino acid it encodes. For example, consider a hypothetical window coding exclusively for conserved methionine and tryptophan residues, which are encoded by nondegenerate codons (ATG and TGG, respectively). By definition, this window does not exhibit any synonymous substitutions and might therefore appear to have a very low synonymous substitution rate. But a reduced estimate of λ_s in this window would not be considered significant, because it does not provide a better explanation for the conservation of the nondegenerate sites.

We also designed our method to control for background variation in sequence composition and evolutionary rates across the genome (Lercher et al. 2001; Williams and Hurst 2002; Fox et al. 2008), as well as the possibility of selection on diffuse effects that can constrain all or most of an ORF, such as transcript structural stability or codon bias for translation efficiency (Chamary et al. 2006). To account for these, we evaluated each window against two null models, one representing the average sequence composition and evolutionary rates of the entire "ORFeome" (to obtain λ_s^{ome}), and the second estimated specifically from the individual ORF containing each window (to obtain λ_s^{ORF}). By identifying statistically significant rate reductions with respect to both null models, we required windows of interest to be exceptional with respect to genome-wide averages, on one hand, and also not explained by local biases in composition, rates, and codon usage, on the other hand.

Results

To annotate likely overlapping functional elements in human genes, we applied our local synonymous rate estimation procedure to sliding windows in all open reading frames in the human Consensus Coding Sequence (CCDS) catalog, a conservative set containing the ~85% of human gene annotations that are unanimously agreed on by the major gene catalogs (Pruitt et al. 2009). The vast majority of CCDS ORFs are aligned across at least 15 of the 29 placental mammals used in this study, with an average of four synonymous substitutions per codon site (Supplemental Fig. 1).

Genome-wide trends in synonymous rate variation

Before attempting to locate individual regions of synonymous constraint, we were immediately able to confirm a few notable genome-wide trends in synonymous rate variation that have been observed previously using different methods. Specifically, we observed marked reductions in average synonymous rates at the boundaries of coding exons, suggesting widespread evolutionary constraint on overlapping translation and splicing regulatory elements (Smith and Hurst 1999; Baek and Green 2005; Xing and Lee 2005; Chen et al. 2006; Parmley et al. 2006; Parmley and Hurst 2007a). For example, the median λ_s^{ome} estimate for the first nine codons following the start codon in each ORF is only 0.81, indicating that the first several codons in a typical mammalian gene appear to have tolerated synonymous substitutions at a rate 19% below average ($P < 10^{-15}$, Mann-Whitney U test) (Fig. 1B). The median estimated synonymous rate is also reduced by 13% in nine-codon windows spanning exon–exon junctions ($P < 10^{-15}$). Strikingly, the first nine to 12 codons of alternatively spliced cassette exons annotated in CCDS typically have an estimated synonymous rate 41% below average ($P < 10^{-15}$). We will revisit the possible translation and splicing regulatory roles suggested by these overall trends after establishing the statistical significance of the synonymous constraint in each individual window.

More than one-third of CCDS ORFs contain short windows with statistically significant synonymous constraint

We next applied LRTs to identify individual windows that show statistically significant evidence of reduced synonymous substitution rates (Fig. 2). Using three window sizes of nine, 15, and 30 codons, and sliding across each ORF by one-third of the window length, we selected windows passing three likelihood ratio tests, for the hypotheses that λ_s^{ome} is significantly below one, that λ_s^{ORF} is also significantly below one, and that λ_n^{ome}, the relative rate of nonsynonymous substitutions, is not significantly above one. We applied appropriate corrections for multiple testing (see Methods), and simulation and permutation benchmarks confirmed the robustness of our approach for detecting significantly reduced synonymous rates (Supplemental Material S3).

F2

At the intermediate window length of 15 codons, 1.7% of the windows in CCDS ORFs meet these criteria. Overlapping significant windows collapse into 10,757 separate regions throughout human ORFs, covering 2.8% of all approximately 28 million CCDS coding nucleotide positions. More than one-third of CCDS genes (6033/16,939) contain at least one such region. Notably, although the test threshold is $\lambda_s^{ome} < 1$, the median λ_s^{ome} among the windows passing the test is only 0.23, corresponding to a 77% reduced rate of synonymous substitution compared to the genome-wide average. Furthermore, the estimates of λ_s^{ORF} in significant windows are also very low (best-fit line $\lambda_s^{ORF} = 0.78\lambda_s^{ome} + 0.06$ with $R^2 = 0.81$), confirming that these regions are generally not explained by ORF- or region-specific variation in sequence composition or evolutionary rates. Finally, the locally estimated synonymous and non-synonymous rates are not strongly correlated (Pearson coefficient between λ_s^{ORF} and λ_n^{ORF} of 0.04 in all windows, and 0.06 in significant windows), suggesting that our method largely succeeds in disentangling evolutionary pressures on synonymous and nonsynonymous sites, when controlling for regional biases in composition and rates. (Due to such biases, the estimates of λ_s^{ome} and λ_n^{ome} do correlate somewhat, with Pearson coefficient 0.24 in all windows and 0.22 in significant windows.)

Windows with significant synonymous constraints are not unusually enriched on any of the individual human autosomes, although they are about twofold depleted on chromosome 19 (Supplemental Table 1), which frequently stands out in genome-wide analyses owing to several unusual properties. In particular, the apparent depletion of synonymous constraints on this chromosome may be due to a calibration bias in our method arising

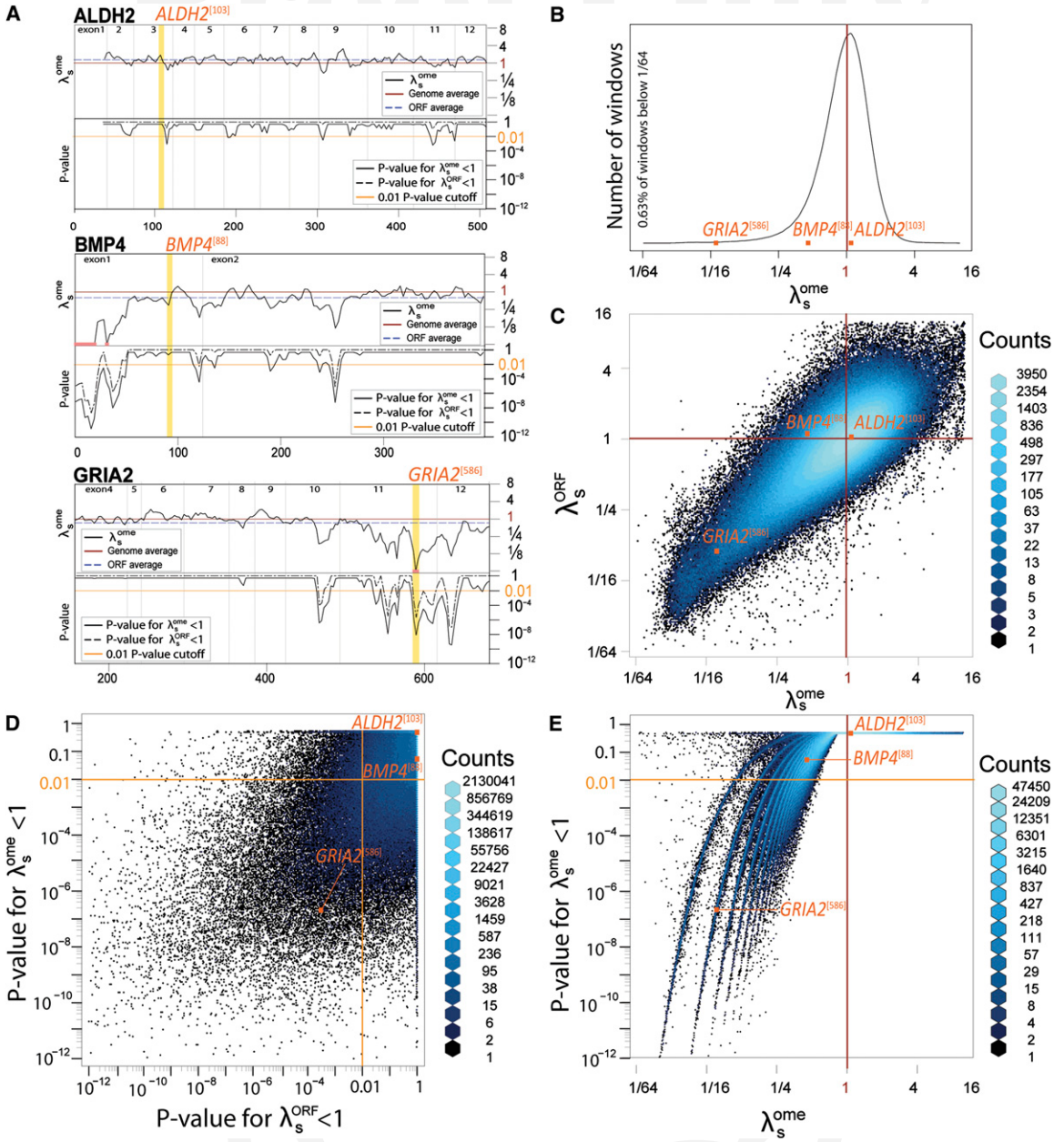


Figure 2. Identifying individual windows with statistically significant synonymous constraint. (*A*) Estimated synonymous rate relative to genome average (λ_s^{ome}) and corresponding *P*-value for the hypothesis $\lambda_s^{ome} < 1$ evaluated in nine-codon windows along the entire protein-coding regions of *ALDH2*, *BMP4*, and *GRIA2*, highlighting the windows corresponding to the three examples in Figure 1. For each plot, the *top* portion shows the λ_s^{ome} estimate for each window (black curve), the genome average (red line at $\lambda_s = 1$), and the ORF average (blue dashed line). The *bottom* portion shows the statistical significance of the reduction in the synonymous rate estimate in each window, accounting for evidence in the cross-species alignments, using a likelihood ratio test for the hypothesis $\lambda_s^{ome} < 1$ (continuous black curve, using the genome average as the null model), and for the hypothesis $\lambda_s^{ORF} < 1$ (dashed black curve, using the ORF average as the null model). (Vertical gray lines) Exon boundaries; (orange) regions where λ_s^{ome} drops below 1/16th toward the 5′ end of BMP4 and the 3′ end of GRIA2. (*B*) Overall distribution of λ_s^{ome} estimates for all nine-codon windows across all CCDS genes. Heavy *left* tail indicates an excess of windows with very low estimated synonymous rates, shifting the mean ($\lambda_s^{ome} = 1$) to the *left* of the distribution mode, which likely represents neutral rates. (*C*) Comparison of synonymous rates estimated relative to genome-wide (λ_s^{ome}) and ORF-specific (λ_s^{ORF}) null models, each point denoting one nine-codon window, and density of overlapping points denoted by color. Joint distribution shows that low λ_s^{ome} estimates also usually correspond to low λ_s^{ORF} estimates, and therefore that the heavy tail observed in *B* does not reflect regional or ORF-wide deceleration, but instead localized constraints in small windows within each ORF, also visible in the three examples of *A*. (*D*) Comparison of *P*-values for synonymous rate reduction with respect to genome-wide (*y*-axis) and ORF-specific (*x*-axis) null models. Candidate synonymous constraint windows are selected when synonymous rate reductions are significant at $P < 0.01$ with respect to both null models (orange lines). Note that many windows are significant with respect to one null model but not the other. (*E*) Correspondence between λ_s^{ome} and the associated significance estimate for the each nine-codon window. The visible stripes in this plot arise from windows that are perfectly conserved except for one, two, three, or more synonymous substitutions observed in the extant species, while the position along each stripe reflects variation in the λ_s^{ome} estimate and its significance, determined by the species coverage, codon composition, and observed codon substitutions in each window. (*B–E*) The three example regions highlighted in *A* are shown in each distribution and density plot, with horizontal and vertical axes aligned. The orange line in plots *A*, *D*, and *E* denotes the statistical significance cutoff of $P < 0.01$, and the red line in plots *A*, *B*, *C*, and *E* denotes the genome-wide average $\lambda_s^{ome} = 1$ and $\lambda_s^{ORF} = 1$ for *B*. The *ALDH2*[103] synonymous rate is not significantly reduced either relative to the genome or to the ALDH2 ORF; *BMP4*[88] is reduced relative to the genome but not relative to its ORF, which shows an overall reduced rate; *GRIA2*[586] is >80% reduced relative to both the genome and its ORF, resulting in significant *P*-values for both.

Fig(s). 2 live 4/C

from the chromosome's above-average G+C content and mutation rates (Lercher et al. 2001; Castresana 2002), or it could reflect a genuine biological tendency related to the chromosome's unusually large complement of genes from a few tandem families (Grimwood et al. 2004). We analyzed ORFs on the X chromosome using a null model estimated from coding sites on that chromosome only and found that its resulting proportion of significant windows is lower than most autosomes, but still much greater than chromosome 19. We did not analyze the fewer than 100 protein-coding genes on the Y chromosome due to their eccentric, fast-evolving properties (Hughes et al. 2005, 2010; Kuroki et al. 2006).

In the longer windows of 30 codons, our method has increased statistical power to detect synonymous constraints since it combines evidence from more sites, and a larger proportion of windows reach significance with somewhat higher typical synonymous rate estimates, although they collapse into fewer separate regions. Conversely, a smaller proportion of the shorter nine-codon windows reach significance, with even lower estimated synonymous rates (Table 1). We also attempted our analysis with even smaller windows of six and three codons, but vanishingly few reached significance. Evidently, in these very short windows, even perfect conservation across the available species is usually not adequate to infer synonymous constraints using our current methodology and alignments.

Since the three window sizes lead to different trade-offs between resolution and discovery power, it is reasonable to expect them to identify somewhat different sets of regions as significant. In fact, of the regions obtained by collapsing overlapping significant windows at the 15-codon resolution, 24% are not detected at either the longer or shorter resolution. Similarly, 33% of the 30-codon regions and 28% of the nine-codon regions are detected only at those resolutions. As expected, the intermediate 15-codon resolution has the most overlap with the others, including 67% of the 30-codon and 72% of the nine-codon regions.

Hereafter, we refer to the collapsed significant regions as "synonymous constraint elements" (SCEs) and undertake numerous downstream analyses to show that they correspond to overlapping functional elements with diverse biological functions. We will perform most of these analyses based on SCEs identified at the 15-codon resolution, since they include most of the other sets, but we will also use the nine-codon and 30-codon resolutions based on the expected length of different types of overlapping functional elements. Similarly, we expect that each can be useful in different contexts for future follow-up studies.

Table 1. Sliding windows in CCDS ORFs were tested for significantly reduced synonymous substitution rate estimates at different resolutions and stringencies

Resolution (window size)	Short (9 codons)	Intermediate (15 codons)	Long (30 codons)
No. of windows tested	2,915,773	1,727,202	842,475
Proportion of windows significant	1.04%	1.72%	2.75%
Median significant λ_s^{ome}	0.1361	0.2299	0.3584
Median significant λ_s^{ORF}	0.1443	0.2572	0.4120
Maximum significant λ_s^{ome}	0.5497	0.6323	0.7134
Maximum significant λ_s^{ORF}	0.4682	0.5805	0.6520
No. of nonoverlapping synonymous constraint elements	11,882	10,757	8933
Proportion of 27,812,282 CCDS nucleotide positions within a synonymous constraint element	1.79%	2.82%	4.48%
Proportion of 16,939 CCDS ORFs containing a synonymous constraint element	35.8%	35.6%	33.3%

Sequence composition and codon usage in SCEs

Since a major design goal for this study was to control for the specific codon sequence in each window, we thoroughly examined the composition of the regions reported as significant (Supplemental Material S6). Briefly, the SCEs exhibit certain biases in nucleotide, dinucleotide, and amino acid composition, for example, slightly below-average G+C content, but these biases are within the range of variation seen between protein-coding regions from different parts of the genome. Other compositional properties allowed us to rule out certain possible artifactual explanations for the low divergence in SCEs, including tandem and microsatellite repeats, biased conversion in recombination hotspots, and codon usage bias. In particular, the effective number of codons (ENC) (Wright 1990; Fuglsang 2006) shows that there is slightly less codon usage bias in SCEs compared to other coding regions, indicating that our method excluded regions explained by this effect, as intended. Overall, our analysis of sequence composition and codon usage did not suggest any debilitating shortcomings of our overall approach, and it is likely that the compositional differences that are seen largely reflect the sequence-dependent biological nature of the overlapping functional elements encoded by the SCEs.

Characteristics of genes containing SCEs

We next studied overall properties of the 6033 genes containing SCEs (at 15-codon resolution). Compared to the remaining CCDS genes, the typical gene containing an SCE has a much longer ORF (median 558 vs. 356 codons). This is not actually longer than expected based on drawing genes randomly weighted by their ORF length, but the genes containing SCEs also have more introns (median nine vs. five), lengthier individual introns (1727 nt vs. 1261 nt), and they span much larger genomic regions (36,000 nt vs. 11,000 nt), suggesting that the overall length distribution is entangled with the well-established correlations between gene length and other relevant characteristics including conservation, functional categories, and expression levels (Supplemental Material S7; Castillo-Davis et al. 2002; Urrutia and Hurst 2003; Stanley et al. 2006; Pozzoli et al. 2007). The genes containing SCEs also appear to be under stronger purifying selection on their amino acid sequences, as the median estimate of $\omega = d_N/d_S$ measured across each complete ORF is 0.068, much lower than the 0.138 for other genes, despite containing regions with greatly reduced d_S.

Next, we analyzed Gene Ontology (GO) annotations for the genes containing SCEs. While 36% of CCDS genes contain a 15-codon SCE, they include 70% genes annotated with the term "chromatin modification," a twofold enrichment (Bonferroni-corrected hypergeometric $P < 7.9 \times 10^{-13}$). Additionally, they include most of the genes in these and related categories: "ubiquitin-protein ligase activity" (1.8-fold, $P < 2.9 \times 10^{-6}$), "ion channel complex" (1.7-fold, $P < 5.6 \times 10^{-5}$), "nervous system development" (1.6-fold, $P < 6.8 \times 10^{-9}$), "transcription factor activity" (1.6-fold, $P < 3.5 \times 10^{-6}$), and "RNA splicing" (1.5-fold, $P < 8.4 \times 10^{-4}$). These enrichments remain strongly significant when controlling for the varying ORF lengths (Supplemental Table 4) and suggest a few interesting hypotheses

about genes that encode overlapping functional elements. For example, the enrichment for genes encoding chromatin modification and RNA splicing functions could suggest the existence of autoregulatory circuits for many such genes, similar to known examples such as ADARB1, which edits its own pre-mRNA and causes a change in its splicing (Rueter et al. 1999), and DGCR8, which binds its own mRNA and causes it to be cleaved by Drosha (Han et al. 2009). Also, the enrichment for ion channel genes suggests a connection with A-to-I editing, since several such genes are known targets of this recoding mechanism (Bass 2002); we explore this further below.

More than one-third of short SCEs can be provisionally assigned roles in transcript splicing or translation initiation

As expected based on the aforementioned general trends in synonymous rate variation (above; Fig. 1B), many SCEs can be provisionally classified as possible splicing regulatory elements based on their location within gene models. In particular, 34.7% of the nine-codon SCEs span an exon–exon junction, compared to only 20.3% of a set of random control regions placed uniformly throughout CCDS ORFs, with matching length distribution and total number.

Interestingly, the introns flanked by these SCEs tend to have weaker 3′ splice acceptor sites (nine-codon resolution; $P = 3.9 \times 10^{-7}$, Mann-Whitney *U* test), based on analysis of the sequence information content of the surrounding nucleotides (Yeo and Burge 2004). This is consistent with the hypothesis that while "strong" splice sites are constitutively recognized by the splicing machinery, the activity of weaker splice sites is more reliant on additional nearby *cis*-regulatory sequences, increasing the possibilities for their combinatorial and condition-specific regulation (Fairbrother et al. 2002; Chen and Manley 2009). The SCEs spanning exon–exon junctions also show an increased frequency of exonic splicing enhancer motifs compared to other coding regions, although this is difficult to distinguish from correlated compositional biases (Supplemental Material S8).

We also examined individual exons with multiple alternative acceptor or donor sites, for which the coding sequence of the longest exon isoform additionally encodes splice sites for shorter isoforms and perhaps additional splicing regulatory elements (e.g., Fig. 3A). Of 551 such alternative donor sites in RefSeq transcripts encoded within the CCDS exons we analyzed, 84 (15.2%) fall within SCEs, compared to only 1% in the random regions ($P < 10^{-21}$). Similarly, 57 of 576 (9.9%) alternative acceptor sites lie within SCEs (1.7% random; $P < 10^{-21}$).

SCEs are also enriched for elements potentially involved in translation initiation. The first nine-codon window (beginning at the first site following the start codon) in 3.9% of CCDS ORFs is found to be under synonymous constraint, a strong enrichment compared to 1.04% of all windows. Additionally, there are 744 RefSeq-annotated internal translation initiation sites encoded within longer coding exons of CCDS ORFs, of which 5.4% fall within SCEs (e.g., Fig. 3B) compared to 1.6% in the random regions ($P < 10^{-9}$).

Taken together, these provisional classifications of possible splicing and translation regulatory elements account for slightly more than one-third of the SCEs (at nine-codon resolution). Conversely, nearly two-thirds are not found in locations that directly suggest such roles, although it should be noted that some splicing regulatory sequences may act at considerable distances (Parmley and Hurst 2007a; Parmley et al. 2007).

Synonymous constraint at an alternate translation start site in BRCA1

One noteworthy example of an SCE with a possible role in translation initiation is found within the tumor suppressor gene *BRCA1*. Hurst and Pál (2001) first observed that an extended region within this ORF shows unusually low synonymous substitution rates, based on sliding windows of 100 codons in pairwise comparisons among the human, mouse, and rat orthologs. Recently, however, Schmid and Yang (2008) raised certain issues with their statistical methods and argued that their result was artifactual.

Our analysis is based on much more data than both previous studies and strongly supports the original conclusion of Hurst and Pál (2001). Within the most significant 15-codon window in *BRCA1*, the estimated rate of synonymous substitution in placental mammals is reduced by 80% ($\lambda_s^{ome} = 0.20$, $P < 9.9 \times 10^{-8}$), ranking among the slowest 1% of windows in human ORFs, and slower than average for SCEs at this resolution. Like all SCEs, the window also has a very low synonymous rate estimate compared to the *BRCA1* ORF specifically ($\lambda_s^{ORF} = 0.23$, $P < 2.3 \times 10^{-5}$). Due to the controversy over this region in particular, we also performed auxiliary permutation tests that further confirmed its statistical significance (Supplemental Material S4).

Furthermore, the much higher resolution of our analysis precisely localizes the significant region of synonymous constraint to the annotated translation initiation site of a CCDS-supported alternative splice form of *BRCA1* (Fig. 3B), immediately suggesting a hypothesis for an overlapping biological function—namely, a role in regulating translation initiation in this splice form. This highly suggestive positional association may have been much less clear at the 100-codon resolution used by both previous studies. Indeed, the synonymous constraint is not significant at the 30-codon resolution according to our analysis, possibly corroborating the statistical concerns raised by Schmid and Yang (2008), while nonetheless confirming and extending the main conclusion of Hurst and Pál (2001).

Enrichment of miRNA target sequences in SCEs

We next sought evidence that SCEs capture embedded miRNA target sites, since previous studies have demonstrated trends toward their preferential conservation in mammalian ORFs (Lewis et al. 2005; Hurst 2006; Kural et al. 2009). Although the main sequence determinant of miRNA targeting is a "seed" of only ~7 nt, which is much shorter than our present resolution, it is still reasonable to expect SCEs to capture at least some of these sites, due to additional flanking positions influential in targeting, or to multiple closely spaced target sites acting synergistically (Grimson et al. 2007).

Indeed, SCEs show a small but significant enrichment for the 7-nt target seed sequences of known human miRNAs: 8.9% of positions in SCEs start with one of these 7-mers, compared to 8.6% in all coding regions ($P < 10^{-7}$; nine-codon resolution). They are even more frequent in SCEs lying within the last coding exon of each ORF (9.3% of positions; $P < 10^{-5}$), consistent with the trend for miRNA target sites to appear toward the 3′ end of animal transcripts (Lewis et al. 2005). Matched random control motifs showed weaker or no enrichment, indicating that the observed enrichment is not explained by correlated sequence composition effects (Supplemental Table 5). Overall, while the excess of miRNA seeds in SCEs provides further evidence that many mammalian ORFs encode conserved target sites, greater power and resolution will be needed to precisely annotate them.

Proof Only

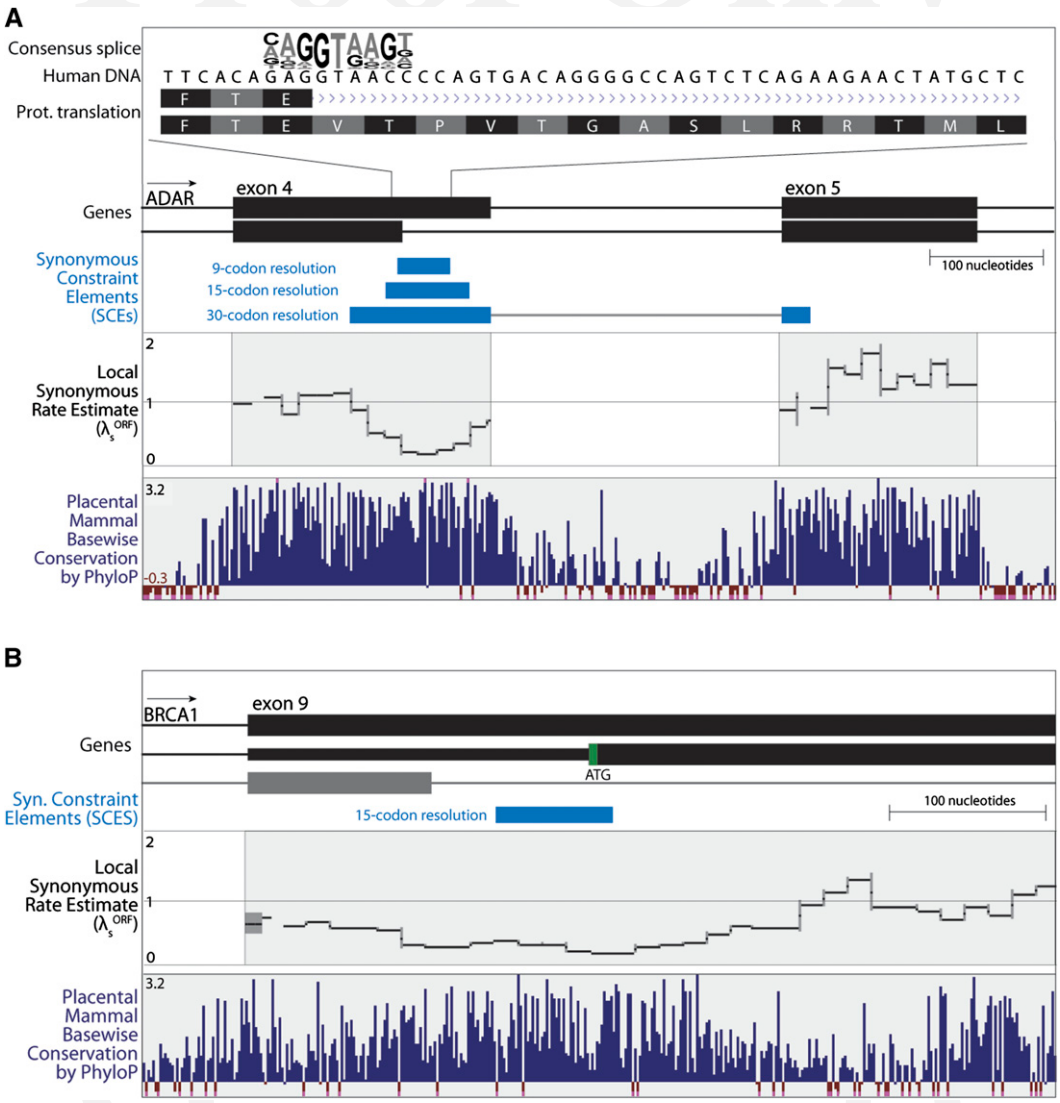


Figure 3. Examples of candidate synonymous constraint elements (SCEs) with likely roles in splicing and translation regulation. (*A*) Predicted SCEs (light blue) overlapping two isoforms of *ADAR* exon 4 (black) arising from an alternative splice donor site encoded within the longer exon variant. With increasing resolution, the SCE is more precisely localized to the region of overlap with the alternative splice site (motif logo for human donor sites rendered by WebLogo) (Crooks et al. 2004). The localization of the synonymous constraint to the splice site is also seen in the local synonymous rate estimate λ_s^{ORF} (relative to the ORF average). Note that the significant reduction in the synonymous rate is not obvious from the nucleotide-level conservation measure (dark blue, *bottom* panel). The extent of the predicted SCE may suggest the presence of additional splicing regulatory elements downstream from the alternative splice site. (*B*) Predicted SCE (light blue) overlapping an alternate translation initiation site (green) in *BRCA1* encoded within exon 9 of a longer isoform. Synonymous constraint ranges from shortly upstream to immediately downstream of the alternate start codon, suggesting this region may be involved in regulating translation initiation at the alternate site. The region just upstream of the predicted SCE also shows a reduced synonymous rate (black curve) overlapping an alternative splice donor site for a third *BRCA1* isoform (gray), although this reduction is not statistically significant and the third isoform is weakly supported. Annotation visualizations in Figures 3 and 4 are based on the UCSC Genome Browser (Kent et al. 2002).

SCEs in known and novel dual-coding genes

Genomic sequences can simultaneously encode different amino acid sequences in multiple reading frames, a common phenomenon in viral genomes but rare in animal genomes. Such "dual-coding" regions can involve ORFs on the same strand, but in an alternate "shifted" reading frame, which can be mediated by alternative splicing, internal translation initiation, or ribosomal frameshifting. Of the six long human dual-coding gene structures for which likely biological functions have been demonstrated (Sharpless and DePinho 1999; Klemke et al. 2001; Yoshida et al. 2001; Hameed et al. 2003; Poulin et al. 2003; Ahmed et al. 2008), all show at least some evidence of overlapping evolutionary constraint in our analysis: *XBP1*, *GNAS*, and the *ANKHD1/EIF4EBP3* fusion transcript contain SCEs in their dual-coding regions; the dual-coding 3′ end *IGF1* narrowly missed our threshold with a 47% reduced synonymous rate; and the dual-coding regions of *CDKN2A* and *LRTOMT* have very low synonymous rates, but were excluded from SCEs because of elevated nonsynonymous rates. Aside from these individually studied examples, CCDS annotates 237 other exons as protein-coding in multiple reading frames of the same strand, 24 (10.1%) of which contain SCEs, compared to six (2.5%) containing random control regions. This lower overlap might

Proof Only

Fig(s). 3 live 4/C

Proof Only

suggest dual-coding regions not under selection across placental mammals.

Alternatively, both strands of the genomic DNA can encode different protein sequences, expressed in "sense" and "antisense" transcription units. At least one such case has been thoroughly studied: an ~200-nt sense/antisense dual-coding sequence of the convergent transcription units for *THRA* and *NR1D1* (Hastings et al. 2000). Indeed, we find synonymous constraints in both ORFs precisely coinciding with the known dual-coding region (Fig. 4A). We similarly detect synonymous constraints in 10 of 44 individual exons that CCDS annotates on both strands (five in random regions). F4

In addition to these known examples, we found 19 candidate novel dual-coding ORFs in alternate reading frames of known CCDS ORFs, which contain one or more of our 30-codon SCEs, are longer than expected by chance, and also appear to be depleted of

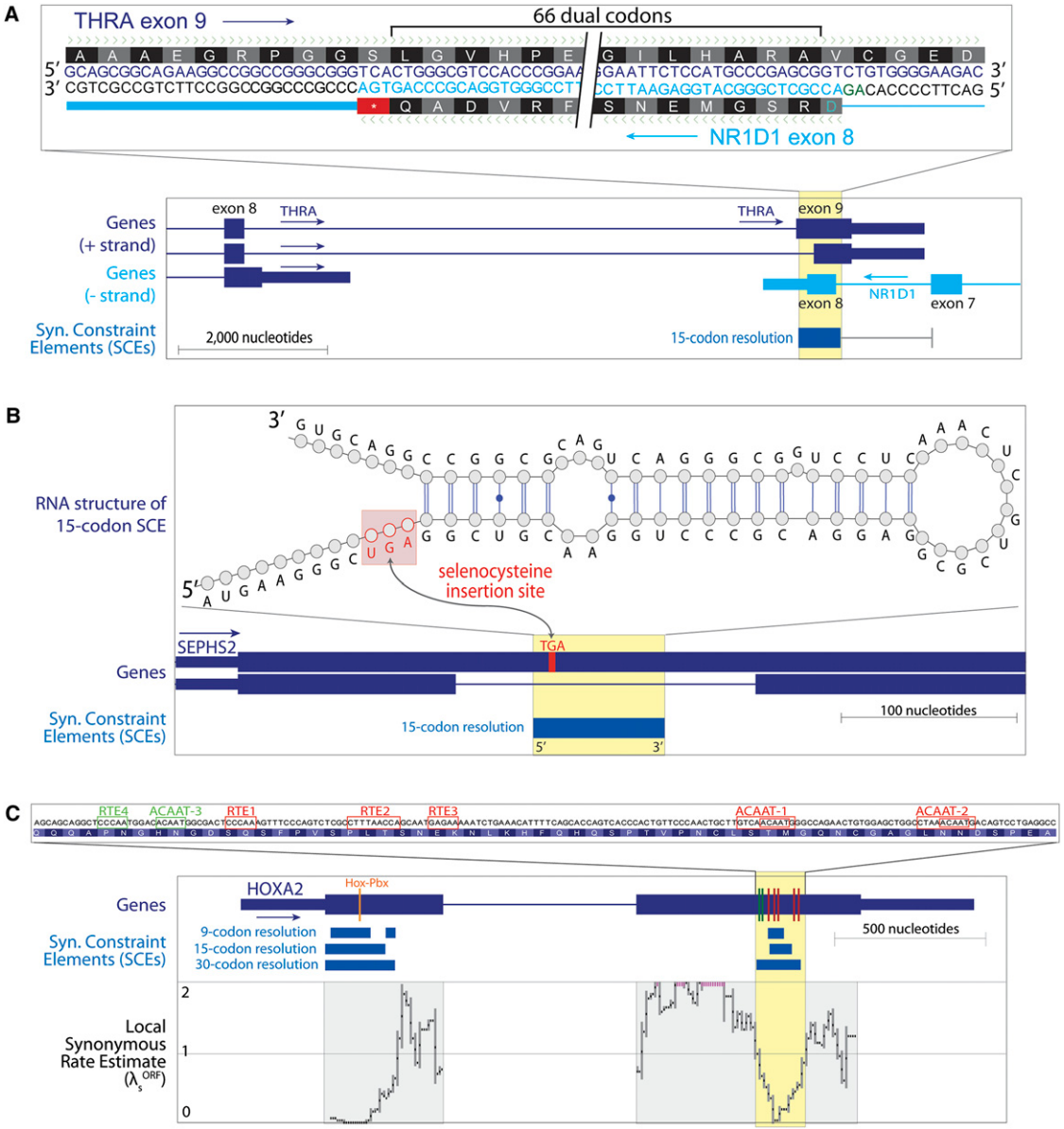


Figure 4. Synonymous constraint elements (SCEs) corresponding to dual-coding, selenocysteine insertion, and expression enhancer functions. (*A*) A large SCE (blue) fully encompasses a 66-codon sense/antisense dual-coding region in the convergent transcripts of *THRA* and *NR1D1*. The SCE is specifically localized to the overlapping exons, while upstream exons of each gene are excluded. (*B*) A predicted SCE in the selenoprotein-encoding gene *SEPHS2* encompasses the selenocysteine insertion site (red) and a predicted RNA hairpin structure (minimum free energy fold rendered by VARNA) (Darty et al. 2009) immediately downstream from the selenocysteine codon. Inferred structure is similar to a hairpin known to stimulate selenocysteine recoding in *SEPN1* (Howard et al. 2005). (*C*) Two SCEs are found within the *HOXA2* ORF, each corresponding to a different enhancer element regulating expression of the mouse ortholog in distinct segments of the developing hindbrain. The 5′ element encodes a Hox-Pbx responsive element and drives expression in rhombomere 4 (Lampe et al. 2008), and the 3′ element encodes SOX2 binding sites and drives expression in rhombomere 2 (Tümpel et al. 2008). The 3′ element includes several RTE and ACAAT motif instances that were investigated by site-directed mutagenesis in the previous study (red), as well as two additional upstream instances (green). SCEs are also found within most other *HOX* genes.

Proof Only

Fig(s). 4 live 4/C

stop codons in the other mammals. Twelve of these 19 are encoded on the same strand as the CCDS gene structure, some spanning multiple exons, and the remaining seven are found antisense to individual CCDS exons. A few are further supported by transcript cDNA evidence or similarity to known proteins (Supplemental Material S9). This preliminary assessment suggests that the SCEs probably capture several additional long dual-coding mammalian gene structures, although specialized methods for detecting the evolutionary signatures unique to dual-coding regions (Chung et al. 2007; Ribrioux et al. 2008) would probably have more power applied to the new set of 29 species.

SCEs capture most known A-to-I recoding sites

A-to-I editing is a recoding mechanism in which certain adenosine (A) bases in RNA transcripts are edited to inosine (I), which is read as a guanosine (G) (Bass 2002). This mechanism is essential for normal development of the mammalian nervous system, and, because the enzymes that mediate the reaction target double-stranded RNA, known A-to-I recoding sites conserved between human and mouse transcripts typically show extensive conservation of flanking sequence, presumably reflecting the interlocking constraints of encoding amino acids and pairing with another portion of the transcript (Aruscavage and Bass 2000; Hoopengardner et al. 2003).

Indeed, 10 of 14 known human A-to-I recoding sites in CCDS ORFs lie within SCEs (15-codon; none within random regions) (Supplemental Table 7), although this is not surprising since exceptional conservation was one signature originally used to identify many of the known sites. A recent human-specific study used a high-throughput sequencing approach not biased for highly conserved regions (Li et al. 2009) to identify 40 new edited sites within CCDS ORFs, only three of which also lie within our SCEs (two amino acid changing sites in *CADPS*, *FLNB*, and a synonymously edited site in *GRIA2*) (Supplemental Table 8). Therefore, consistent with that study's report that the 37 remaining experimentally identified sites lack extensive nucleotide-level conservation, we do not find specific evidence for synonymous constraints in other mammals.

SCEs and RNA secondary structures

In addition to the paired structures associated with A-to-I editing sites, the SCEs include several other striking examples of RNA secondary structures embedded within mammalian ORFs. For example, the selenoprotein-encoding gene *SEPHS2* contains a hairpin immediately downstream from its selenocysteine insertion site (Fig. 4B). This structure is likely to stimulate selenocysteine recoding, based on similar known structures in other genes (Howard et al. 2005; Pedersen et al. 2006). An SCE within *TTN*, the human gene with the most exons (313) encoding a protein with numerous functions in striated muscle and associated with several diseases, contains a hairpin showing both compensatory double substitutions and a compensatory deletion that preserve the paired structure in other vertebrates (Supplemental Fig. 2), strong evidence of selection on the RNA secondary structure. An SCE within *QKI*, which encodes an RNA-binding protein, appears to contain an instance of the protein's own binding motif followed by a hairpin showing a compensatory insertion, perhaps suggesting an autoregulatory mechanism (Supplemental Fig. 3).

We also attempted to study overall mutual enrichments between SCEs and computationally predicted RNA secondary structures using EvoFold (Pedersen et al. 2006; BJ Parker, Moltke, and JS Pedersen, in prep.) and RNAz (Gruber et al. 2010), but unfortunately this was confounded by correlations in sequence composition and conservation influencing both types of predictions (Supplemental Material S11). Careful further investigation of this topic is warranted, given previous studies suggesting that many additional RNA structures (Chamary and Hurst 2005; Shabalina et al. 2006; Tuller et al. 2010) and perhaps RNA–RNA interaction sites (Wang et al. 2008) are embedded within protein-coding regions.

Possible roles in exclusion of nucleosomes from certain exons

It was recently shown that nucleosomes preferentially localize within human exons compared to surrounding intronic regions (Schwartz et al. 2009; Tilgner et al. 2009). Two compositional properties of the SCEs suggest a possible relationship with this phenomenon. First, the general enrichment of nucleosomes within exons was shown to positively correlate with G+C content, while the SCEs have slightly lower G+C content than other coding regions (50.3% vs. 52.0%). Second, it is known that nucleosomes especially avoid contiguous stretches of adenine:thymine base pairs, and synonymous codon usage in many species is biased to avoid such "poly(dA:dT) tracts" (Cohanim and Haran 2009; Segal and Widom 2009). The SCEs have a 20% higher frequency of poly(dA:dT) tracts of 5 bp or longer compared to other coding regions.

Based on these properties, we would predict that exons containing SCEs tend to have lower nucleosome occupancy than other exons. We analyzed a high-throughput sequencing data set for nucleosome occupancy in human CD4+ T-cells (Schones et al. 2008) and found that the exons containing SCEs are, indeed, depleted for reads of nucleosome-bound DNA compared to other CCDS exons ($P = 1.0 \times 10^{-11}$, Mann-Whitney U test) (Supplemental Material S12).

Since the biological significance of nucleosome positioning within exons is not yet well understood, we cannot exclude the possibility that this relative depletion could just be a side effect of compositional biases in the SCEs. It is also possible, however, that some SCEs are under selection for sequence-dependent roles in excluding nucleosomes from certain exons, not unlike the "nucleosome-free regions" thought to facilitate chromatin access for regulatory factors near promoters (Schones et al. 2008; Warnecke et al. 2008; Washietl et al. 2008).

SCEs lie within most *HOX* genes and include two known developmental enhancers

Many of the lengthiest and most strikingly conserved SCEs are found within 27 of the 40 genes in the four *HOX* clusters. For example, the first 60 codons of *HOXB5* exhibit absolutely no synonymous substitutions in any of the species in our alignment of its ORF. More generally, we identify SCEs in eight of the 11 genes in the *HOXA* cluster, seven of nine *HOXB* genes, seven of nine *HOXC* genes, and three of nine *HOXD* genes, as well as the *EVX1* and *EVX2* homeobox-encoding genes adjacent to the *HOXA* and *HOXD* clusters. Lin et al. (2008) also noted striking regions of synonymous constraint in *HOX* genes, which they defined as stretches of at least 40 codons without any synonymous substitutions in pairwise comparisons. Our results confirm their findings while also providing much greater power and resolution, locating several additional shorter and/or less extremely conserved SCEs.

Remarkably, the two SCEs found within *HOXA2* correspond to known tissue-specific enhancers that regulate expression in distinct segments of the developing mouse hindbrain (Fig. 4C). A lengthier region (~200 bp) in the upstream exon encodes a Hox-Pbx

responsive element and drives *Hoxa2* expression in rhombomere 4 (Lampe et al. 2008), and a shorter region (~75 bp) in the downstream exon encodes SOX2 binding sites and drives expression in rhombomere 2 (Tümpel et al. 2008). Considering these two examples, the SCEs could suggest the existence of a largely unknown regulatory network relying on nucleotide sequence elements embedded within the ORFs of most of these key developmental genes (Woltering and Duboule 2009).

Discussion

In this study, we showed that thousands of human protein-coding genes contain short regions with conspicuously low estimated synonymous rates across placental mammals. We located between 9000 and 12,000 SCEs, depending on the resolution chosen, within one-third of all CCDS ORFs (Table 1), or well over one-quarter of all human protein-coding genes. Our preliminary results implicate many of these regions in biological roles including tissue-specific developmental enhancers, translation regulatory elements, RNA structures involved in A-to-I editing or selenocysteine incorporation, and different classes of post-transcriptional regulatory elements including miRNA targeting and alternative splicing. Still, these provisional explanations do not even account for half of all the SCEs, suggesting that there may be many other overlapping biological roles yet to be elucidated. Therefore, just as our view of nucleotide-level conserved elements throughout the genome has been greatly refined since the first human/mouse comparisons nearly a decade ago, we expect that this initial survey of SCEs can motivate many additional computational and experimental studies to reveal their biological functions.

In addition to identifying candidate overlapping functional elements, our extensive annotation of synonymous sites under selection in placental mammals can help refine many types of evolutionary and functional analyses that typically assume they are neutral. For example, widely used methods for detecting positive and negative selection on amino acid sites are based the ratio of nonsynonymous and synonymous substitution rates ($\omega = d_N/d_S$). Since the extreme drop in d_S observed in SCEs would naturally tend to elevate local estimates of ω, any claims of positive selection on the amino acid sites encoded within SCEs ought to be regarded with caution (Xing and Lee 2006; Parmley and Hurst 2007b). Our results can also inform future disease association studies and other types of population genetics analyses, which frequently ignore synonymous SNPs (Chamary et al. 2006). The SCEs we identified represent specific regions in which synonymous SNPs may well have significant consequences and should therefore be included in such analyses.

One major challenge in the design of our study, shared with nucleotide-level constraint detection methods, laid in the estimation of null models. Nucleotide-level methods typically calibrate their null models to presumptively neutral regions such as ancestral repeats or fourfold degenerate sites, but these were obviously not suitable for our purposes because our null models must capture the typical evolutionary rates of both synonymous and non-synonymous sites in coding regions. Therefore, we simply calibrated our null models to the ORFeome-wide or ORF-level background, averaging in any unusually evolving sites. If we assume that purifying selection is much more common than positive selection in synonymous sites (Resch et al. 2007), then this approach leads to a somewhat conservative test for synonymous constraint, providing one of our countermeasures against the possibility of background rate variation leading to spurious inferences of selection. More accurate null model calibration is an important direction for future investigation, perhaps by explicitly modeling statistical distributions of synonymous rates (Pond and Muse 2005; Rodrigue et al. 2008).

Still, despite our prudently conservative null model calibration, we were able to achieve far greater discovery power than previous efforts to locate regions of synonymous constraint in mammalian genes (Schattner and Diekhans 2006; Parmley and Hurst 2007b), which identified at most ~2% of our SCEs, and at much lower resolution. This is attributable both to the many informant species now available and to the rigorous phylogenetic methodology we devised to take advantage of them, based on maximum likelihood estimates of the synonymous substitution rates in short windows and formal statistical tests for their reduction. Naturally, this methodology can accommodate additional sequenced genomes and improved assemblies and alignments as they become available—perhaps eventually enabling systematic resolution of lineage-specific overlapping functional elements and individual binding sites for miRNAs and regulatory factors.

Methods

Genome annotations and alignments

This study was based on ORF annotations from the 2009-03-27 build of CCDS (Pruitt et al. 2009) for NCBI version 36 of the reference human genome assembly. When CCDS annotates multiple isoforms of a single locus (as defined by multiple CCDS IDs in overlapping chromosomal regions sharing the same HGNC gene symbol), only the isoform with the longest coding sequence was analyzed. We extracted the alignments for these ORFs from the MULTIZ whole-genome alignments of 44 vertebrate species, generated by UCSC Genome Bioinformatics and used throughout the initial analysis of the 2× mammals data set. Only the "rows" of the MULTIZ alignments corresponding to the 29 available placental mammals were used, and the whole-genome alignments were "spliced" as necessary to produce an alignment of the complete human ORF.

Estimating codon substitution rates in short windows

Our relative rate estimation procedure uses, as a parameter, any standard phylogenetic codon model $M = \langle T,Q \rangle$ where T specifies the topology and branch lengths of a phylogenetic tree and Q is a reversible 61×61 rate matrix describing codon evolution as a stationary, homogeneous, continuous-time Markov process, such that the transition probability matrix for any branch with length t is given by $P = \exp(Qt)$. Given such a model, the probability of any alignment of extant sequences can be computed using Felsenstein's algorithm, assuming independence of the codon sites and using the equilibrium frequencies of the codons implicit in Q as the prior distribution over the root. Nonaligned, gapped, or stop codons in any informant species are marginalized out so that they are irrelevant to the probability, using standard techniques for statistical phylogenetic models (Felsenstein 2004).

To analyze a given alignment window of several codons, we wish to obtain maximum likelihood estimates of the synonymous (and nonsynonymous) rates relative to the null model M. Our approach is to hold T fixed and estimate a new window-specific rate matrix Q^{wnd} by numerically maximizing the probability of the given alignment window jointly over two nonnegative parameters λ_s and λ_n, where the entries of Q^{wnd} relative to the entries of Q are given by

$$q_{ij}^{\mathrm{wnd}}\begin{cases}\lambda_s \times q_{ij} \text{ if } i \neq j \text{ and } aa_i = aa_j\\ \lambda_n \times q_{ij} \text{ if } i \neq j \text{ and } aa_i \neq aa_j\\ -\sum_{k \neq i} q_{ik}^{\mathrm{wnd}} \text{ otherwise } (i=j)\end{cases},$$

where aa_i denotes the amino acid translation of codon i. Thus, λ_s represents a scale factor on the synonymous rates specified by Q, and similarly for λ_n on the nonsynonymous rates. Importantly, while Q is typically normalized to unity mean rate of replacement at equilibrium, we do not renormalize Q^{wnd}. Since T is held fixed, this allows λ_s and λ_n to control the absolute synonymous and nonsynonymous rates, respectively. Assuming Q is reversible, it is easy to verify that Q^{wnd} is reversible with the same equilibrium frequencies π_j as Q, by considering the decomposition $q_{ij} = \pi_j \times s_{ij}$ for symmetric "exchangeabilities" s_{ij} and noting that λ_s and λ_n scale the entries symmetrically.

Using this approach, we reduced the parameter estimation problem in each window to a mere two-dimensional optimization by reusing ORF- or ORFeome-wide estimates of many other phylogenetic model parameters, for which we probably could not obtain reliable joint estimates based only on a few codon sites (Anisimova et al. 2001; Suzuki and Nei 2002; Schmid and Yang 2008; Nozawa et al. 2009). This is similar to the approach used by the sitewise likelihood ratio method of Massingham and Goldman (2005), which reuses ORF-wide parameter estimates while estimating $\omega = d_N/d_S$ at individual sites. We reuse common parameter estimates while effectively estimating both d_N and d_S in windows of several sites.

Null models

The relative rate estimation procedure does not make any assumptions about how the null model was originally estimated (except for reversibility of Q). In fact, we explored two different ways to estimate Q before providing it to this procedure. The first uses a parameterization equivalent to the M0 model of PAML, based on estimates of ω and κ, the transition/transversion rate ratio (Goldman and Yang 1994; Yang et al. 2000). The second is an empirical codon model (ECM) that essentially amounts to an independent estimate for every entry in the 61×61 rate matrix (under the reversibility constraint), not restricted to single-nucleotide instantaneous substitutions (Kosiol et al. 2007). Comparing the two approaches, we found the ECM parameterization to be clearly superior to M0 for our purposes: it achieved better fit to the training data based on BIC and AIC scores, led to tighter distributions of λ_s and λ_n, and resulted in a more conservative overall test for synonymous constraint (slightly fewer rejections of the null hypothesis at any significance level). The ECM approach also has the advantage that it accounts for the CpG hypermutability effect and any other sequence-specific rate biases, to the extent possible under the assumption of independence between codon sites. The results described in the main text are all based on the ECM parameterization. The parameterizations and estimation procedures are described in greater detail in Supplemental Material S2.

With both parameterizations, the ORFeome-wide null model was fit to a random sample of 5% of the codon sites in autosomal CCDS ORFs. A separate chromosome-specific null model was used for genes on chromosome X, based on all codon sites on that chromosome. The ORF-specific null models were estimated from the complete alignment of each ORF. The topology of the mammalian species tree proposed in the 2× mammals analysis (2× Mammals Sequencing and Analysis Consortium, in prep.) was used in all models, while the branch lengths were re-estimated for each ORF in the ORF-specific models.

Likelihood ratio tests

To test the significance of the adjusted rate estimates in any alignment window, we evaluated the likelihoods of several models:

1. The null model M ($\lambda_s = 1$; $\lambda_n = 1$)
2. $\lambda_s = 1$; λ_n estimated by maximum likelihood
3. $\lambda_s = 0.5$; λ_n estimated by maximum likelihood
4. λ_s and λ_n jointly estimated by maximum likelihood ($0 \leq \lambda_s \leq 1$)
5. λ_s and λ_n jointly estimated by maximum likelihood ($0 \leq \lambda_s \leq 0.5$)
6. $\lambda_s = 1$; λ_n estimated by maximum likelihood ($\lambda_n > 1$)

The likelihood ratios of nested models can then be used to perform the different significance tests we described. Specifically, the primary tests for $\lambda_s^{\mathrm{ome}} < 1$ and $\lambda_s^{\mathrm{ome}} < 0.5$ compare model 4 to model 2 and model 5 to model 3, respectively, with M being the genome-wide null model. The secondary test for $\lambda_s^{\mathrm{ORF}} < 1$ compares model 4 to model 2 with M being the ORF-specific null model. The tertiary test for $\lambda_n^{\mathrm{ome}} \leq 1$ (actually, not significantly greater than 1) compares model 6 to model 1.

Following the standard approach for statistical phylogenetic model comparison, we assume that, when the null model holds, the statistic $-2 \times$ lods converges in distribution to the χ^2 distribution with one degree of freedom, where lods is the log-likelihood ratio of the two models under comparison. We then report a P-value for each window by halving the χ^2 distribution tail probability corresponding to lods (Ota et al. 2000). AU6

The test statistic exhibits some artifactual discretization behaviors in windows with a small number of substitutions (reflected in the stripes seen in Fig. 2E), likely violating the asymptotic assumptions justifying these significance estimates (Whelan and Goldman 1999). To ensure the robustness of the test, we performed additional benchmarks with simulated and permuted data, which are described in Supplemental Material S3.

Sliding windows and multiple testing correction

We applied LRTs to windows of a designated length in every CCDS ORF, beginning at the first site following the start codon and sliding by one-third of the window length. This yielded a list of millions of P-values for each window size and stringency threshold, which has a complex internal dependency structure owing to the overlapping windows and to latent rate correlations between nearby windows. We corrected the primary tests on λ_s^{ome} using the Benjamini and Hochberg false discovery rate (FDR)-controlling method (Benjamini and Hochberg 1995), which tolerates local positive correlations (Benjamini and Yekutieli 2001; Storey and Tibshirani 2003), requiring estimated FDR < 0.01. The secondary tests on λ_s^{ORF} were Bonferroni-corrected for the number of windows tested in each ORF, requiring corrected $P < 0.01$. Again, benchmarks with simulated and permuted data confirmed the robustness of this strategy (Supplemental Material S3).

Acknowledgments

We thank Matt Rasmussen, Loyal Goff, Nick Goldman, Katie Pollard, Kerstin Linblad-Toh, and the anonymous reviewers for helpful advice and discussions. Funding for this work was provided by the National Institutes of Health (U54 HG004555-01) and the National Science Foundation (DBI 0644282).

References

Ahmed ZM, Masmoudi S, Kalay E, Belyantseva IA, Mosrati MA, Collin RWJ, Riazuddin S, Hmani-Aifa M, Venselaar H, Kawar MN, et al. 2008.

Mutations of LRTOMT, a fusion gene with alternative reading frames, cause nonsyndromic deafness in humans. *Nat Genet* **40:** 1335–1340.
Anisimova M, Kosiol C. 2009. Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Mol Biol Evol* **26:** 255–271.
Anisimova M, Bielawski JP, Yang Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol* **18:** 1585–1592.
Aruscavage PJ, Bass BL. 2000. A phylogenetic analysis reveals an unusual sequence conservation within introns involved in RNA editing. *RNA* **6:** 257–269.
Baek D, Green P. 2005. Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. *Proc Natl Acad Sci* **102:** 12813–12818.
Bass BL. 2002. RNA editing by adenosine deaminases that act on RNA. *Annu Rev Biochem* **71:** 817–846.
Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. *Science* **304:** 1321–1325.
Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Ser A Stat Soc* **57:** 289–300.
Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *Ann Stat* **29:** 1165–1188.
Castillo-Davis C, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short introns in highly expressed genes. *Nat Genet* **31:** 415–418.
Castresana J. 2002. Genes on human chromosome 19 show extreme divergence from the mouse orthologs and a high GC content. *Nucleic Acids Res* **30:** 1751–1756.
Chamary JV, Hurst LD. 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol* **6:** R75. doi: 10.1186/gb-2005-6-9-r75.
Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: Non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet* **7:** 98–108.
Chen H, Blanchette M. 2007. Detecting non-coding selective pressure in coding regions. *BMC Evol Biol* **7:** S9. doi: 10.1186/1471-2148-7-S1-S9.
Chen M, Manley JL. 2009. Mechanisms of alternative splicing regulation: Insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol* **10:** 741–754.
Chen F, Wang S, Chen C, Li W, Chuang T. 2006. Alternatively and constitutively spliced exons are subject to different evolutionary forces. *Mol Biol Evol* **23:** 675–682.
Chung WY, Wadhawan S, Szklarczyk R, Pond SK, Nekrutenko A. 2007. A first look at ARFome: Dual-coding genes in mammalian genomes. *PLoS Comput Biol* **3:** e91. doi: 10.1371/journal.pcbi.0030091.
Cohanim AB, Haran TE. 2009. The coexistence of the nucleosome positioning code with the genetic code on eukaryotic genomes. *Nucleic Acids Res* **37:** 6466–6476.
Cooper GM, Stone EA, Asimenos G, NISC Comparative Sequencing Program, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15:** 901–913.
Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: A sequence logo generator. *Genome Res* **14:** 1188–1190.
Darty K, Denise A, Ponty Y. 2009. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25:** 1974–1975.
Delport W, Scheffler K, Seoighe C. 2009. Models of coding sequence evolution. *Brief Bioinform* **10:** 97–109.
Dong X, Navratilova P, Fredman D, Drivenes O, Becker TS, Lenhard B. 2010. Exonic remnants of whole-genome duplication reveal *cis*-regulatory function of coding exons. *Nucleic Acids Res* **38:** 1071–1085.
Down T, Leong B, Hubbard TJ. 2006. A machine learning strategy to identify candidate binding sites in human protein-coding sequence. *BMC Bioinformatics* **7:** 419. doi: 10.1186/1471-2105-7-419.
Fairbrother WG, Yeh RF, Sharp PA, Burge CB. 2002. Predictive identification of exonic splicing enhancers in human genes. *Science* **297:** 1007–1013.
Felsenstein J. 2004. *Inferring phylogenies*. Sinauer, Sunderland, MA.
Fox AK, Tuch BB, Chuang JH. 2008. Measuring the prevalence of regional mutation rates: An analysis of silent substitutions in mammals, fungi, and insects. *BMC Evol Biol* **8:** 186. doi: 10.1186/1471-2148-8-186.
Fuglsang A. 2006. Estimating the "effective number of codons": The Wright way of determining codon homozygosity leads to superior estimates. *Genetics* **172:** 1301–1307.
Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. 2009. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25:** i54–i62.
Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* **11:** 725–736.
Goren A, Ram O, Amit M, Keren H, Lev-Maor G, Vig I, Pupko T, Ast G. 2006. Comparative analysis identifies exonic splicing regulatory sequences—the complex definition of enhancers and silencers. *Mol Cell* **22:** 769–781.
Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. 2007. MicroRNA targeting specificity in mammals: Determinants beyond seed pairing. *Mol Cell* **27:** 91–105.
Grimwood J, Gordon LA, Olsen A, Terry A, Schmutz J, Lamerdin J, Hellsten U, Goodstein D, Couronne O, Tran-Gyamfi M, et al. 2004. The DNA sequence and biology of human chromosome 19. *Nature* **428:** 529–535.
Gruber AR, Findeiss S, Washietl S, Hofacker IL, Stadler PF. 2010. Rnaz 2.0: Improved noncoding RNA detection. *Pac Symp Biocomput* **15:** 69–79.
Hameed M, Orrell RW, Cobbold M, Goldspink G, Harridge S. 2003. Expression of IGF-I splice variants in young and old human skeletal muscle after high resistance exercise. *J Physiol* **547:** 247–254.
Han J, Pedersen JS, Kwon SC, Belair CD, Kim YK, Yeom KH, Yang WY, Haussler D, Blelloch R, Kim VN. 2009. Posttranscriptional crossregulation between drosha and DGCR8. *Cell* **136:** 75–84.
Hastings ML, Ingle HA, Lazar MA, Munroe SH. 2000. Post-transcriptional regulation of thyroid hormone receptor expression by *cis*-acting sequences and a naturally occurring antisense RNA. *J Biol Chem* **275:** 11507–11513.
Hoopengardner B, Bhalla T, Staber C, Reenan R. 2003. Nervous system targets of RNA editing identified by comparative genomics. *Science* **301:** 832–836.
Howard MT, Aggarwal G, Anderson CB, Khatri S, Flanigan KM, Atkins JF. 2005. Recoding elements located adjacent to a subset of eukaryal selenocysteine-specifying UGA codons. *EMBO J* **24:** 1596–1607.
Hughes JF, Skaletsky H, Pyntikova T, Minx PJ, Graves T, Rozen S, Wilson RK, Page DC. 2005. Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. *Nature* **437:** 100–103.
Hughes JF, Skaletsky H, Pyntikova T, Graves TA, van Daalen SKM, Minx PJ, Fulton RS, McGrath SD, Locke DP, Friedman C, et al. 2010. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463:** 536–539.
Hurst LD. 2006. Preliminary assessment of the impact of microRNA-mediated regulation on coding sequence evolution in mammals. *J Mol Evol* **63:** 174–182.
Hurst LD, Pal C. 2001. Evidence for purifying selection acting on silent sites in BRCA1. *Trends Genet* **17:** 62–65.
Itzkovitz S, Alon U. 2007. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Res* **17:** 405–412.
Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The Human Genome Browser at UCSC. *Genome Res* **12:** 996–1006.
Klemke M, Kehlenbach RH, Huttner WB. 2001. Two overlapping reading frames in a single exon encode interacting proteins—a novel way of gene usage. *EMBO J* **20:** 3849–3860.
Kosiol C, Holmes I, Goldman N. 2007. An empirical codon model for protein sequence evolution. *Mol Biol Evol* **24:** 1464–1479.
Kural D, Ding Y, Wu J, Korpi A, Chuang J. 2009. COMIT: Identification of noncoding motifs under selection in coding sequences. *Genome Biol* **10:** R133. doi: 10.1186/gb-2009-10-11-r133.
Kuroki Y, Toyoda A, Noguchi H, Taylor TD, Itoh T, Kim D, Kim D, Choi S, Kim I, Choi HH, et al. 2006. Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. *Nat Genet* **38:** 158–167.
Lampe X, Samad OA, Guiguen A, Matis C, Remacle S, Picard JJ, Rijli FM, Rezsohazy R. 2008. An ultraconserved hox-pbx responsive element resides in the coding sequence of *Hoxa2* and is active in rhombomere 4. *Nucleic Acids Res* **36:** 3214–3225.
Lang G, Gombert WM, Gould HJ. 2005. A transcriptional regulatory element in the coding sequence of the human *bcl-2* gene. *Immunology* **114:** 25–36.
Lercher MJ, Williams EJB, Hurst LD. 2001. Local similarity in evolutionary rates extends over whole chromosomes in human–rodent and mouse–rat comparisons: Implications for understanding the mechanistic basis of the male mutation bias. *Mol Biol Evol* **18:** 2032–2039.
Lewis BP, Burge CB, Bartel DP. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120:** 15–20.
Li JB, Levanon EY, Yoon J, Aach J, Xie B, LeProust E, Zhang K, Gao Y, Church GM. 2009. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324:** 1210–1213.
Lin Z, Ma H, Nei M. 2008. Ultraconserved coding regions outside the homeobox of mammalian hox genes. *BMC Evol Biol* **8:** 260. doi: 10.1186/1471-2148-8-260.
Margulies EH, Cooper GM, Asimenos G, Thomas DJ, Dewey CN, Siepel A, Birney E, Keefe D, Schwartz AS, Hou M, et al. 2007. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res* **17:** 760–774.
Massingham T, Goldman N. 2005. Detecting amino acid sites under positive selection and purifying selection. *Genetics* **169:** 1753–1762.

Nguyen MQ, Zhou Z, Marks CA, Ryba NJP, Belluscio L. 2007. Prominent roles for odorant receptor coding sequences in allelic exclusion. *Cell* **131:** 1009–1017.
Nozawa M, Suzuki Y, Nei M. 2009. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proc Natl Acad Sci* **106:** 6700–6705.
Ota R, Waddell PJ, Hasegawa M, Shimodaira H, Kishino H. 2000. Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. *Mol Biol Evol* **17:** 798–803.
Parmley JL, Hurst LD. 2007a. Exonic splicing regulatory elements skew synonymous codon usage near intron–exon boundaries in mammals. *Mol Biol Evol* **24:** 1600–1603.
Parmley JL, Hurst LD. 2007b. How common are intragene windows with KA > KS owing to purifying selection on synonymous mutations? *J Mol Evol* **64:** 646–655.
Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol Biol Evol* **23:** 301–309.
Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. *PLoS Biol* **5:** e14. doi: 10.1371/journal.pbio.0050014.
Pedersen JS, Forsberg R, Meyer IM, Hein J. 2004a. An evolutionary model for protein-coding regions with conserved RNA structure. *Mol Biol Evol* **21:** 1913–1922.
Pedersen JS, Meyer IM, Forsberg R, Simmonds P, Hein J. 2004b. A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucleic Acids Res* **32:** 4925–4936.
Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D. 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* **2:** e33. doi: 10.1371/journal.pcbi.0020033.
Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20:** 110–121.
AU7
Pond SK, Muse SV. 2005. Site-to-site variation of synonymous substitution rates. *Mol Biol Evol* **22:** 2375–2385.
Poulin F, Brueschke A, Sonenberg N. 2003. Gene fusion and overlapping reading frames in the mammalian genes for 4E-BP3 and MASK. *J Biol Chem* **278:** 52290–52297.
Pozzoli U, Menozzi G, Comi GP, Cagliani R, Bresolin N, Sironi M. 2007. Intron size in mammals: Complexity comes to terms with economy. *Trends Genet* **23:** 20–24.
Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, et al. 2009. The Consensus Coding Sequence (CCDS) Project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* **19:** 1316–1323.
Resch AM, Carmel L, Marino-Ramirez L, Ogurtsov AY, Shabalina SA, Rogozin IB, Koonin EV. 2007. Widespread positive selection in synonymous sites of mammalian genes. *Mol Biol Evol* **24:** 1821–1831.
Ribrioux S, Brungger A, Baumgarten B, Seuwen K, John M. 2008. Bioinformatics prediction of overlapping frameshifted translation products in mammalian transcripts. *BMC Genomics* **9:** 122. doi: 10.1186/1471-2164-9-122.
Robins H, Krasnitz M, Levine AJ. 2008. The computational detection of functional nucleotide sequence motifs in the coding regions of organisms. *Exp Biol Med* **233:** 665–673.
Rodrigue N, Lartillot N, Philippe H. 2008. Bayesian comparisons of codon substitution models. *Genetics* **180:** 1579–1591.
Rueter SM, Dawson TR, Emeson RB. 1999. Regulation of alternative splicing by RNA editing. *Nature* **399:** 75–80.
Schattner P, Diekhans M. 2006. Regions of extreme synonymous codon selection in mammalian genes. *Nucleic Acids Res* **34:** 1700–1710.
Schmid K, Yang Z. 2008. The trouble with sliding windows and the selective pressure in BRCA1. *PLoS ONE* **3:** e3746. doi: 10.1371/journal.pone.0003746.
Schones DE, Cui K, Cuddapah S, Roh T, Barski A, Wang Z, Wei G, Zhao K. 2008. Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132:** 887–898.
Schwartz S, Meshorer E, Ast G. 2009. Chromatin organization marks exon–intron structure. *Nat Struct Mol Biol* **16:** 990–995.
Segal E, Widom J. 2009. Poly(dA:DT) tracts: Major determinants of nucleosome organization. *Curr. Opin. Struct. Biol.* **19:** 65–71.
Shabalina SA, Ogurtsov AY, Spiridonov NA. 2006. A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Res* **34:** 2428–2437.
Sharpless NE, DePinho RA. 1999. The *INK4A/ARF* locus and its two gene products. *Curr Opin Genet Dev* **9:** 22–30.
Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15:** 1034–1050.
Smith NGC, Hurst LD. 1999. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. *Genetics* **153:** 1395–1402.
Stanley S, Bailey T, Mattick J. 2006. GONOME: Measuring correlations between GO terms and genomic positions. *BMC Bioinformatics* **7:** 94. doi: 10.1186/1471-2105-7-94.
Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci* **100:** 9440–9445.
Suzuki Y, Nei M. 2002. Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino acid sites. *Mol Biol Evol* **19:** 1865–1869.
Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcarcel J, Guigo R. 2009. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol* **16:** 996–1001.
Tuller T, Waldman YY, Kupiec M, Ruppin E. 2010. Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci* **107:** 3645–3650.
Tümpel S, Cambronero F, Sims C, Krumlauf R, Wiedemann LM. 2008. A regulatory module embedded in the coding region of *Hoxa2* controls expression in rhombomere 2. *Proc Natl Acad Sci* **105:** 20077–20082.
Urrutia AO, Hurst LD. 2003. The signature of selection mediated by expression on human genes. *Genome Res* **13:** 2260–2264.
Wang P, Yin S, Zhang Z, Xin D, Hu L, Kong X, Hurst L. 2008. Evidence for common short natural trans sense–antisense pairing between transcripts from protein coding genes. *Genome Biol* **9:** R169. doi: 10.1186/gb-2008-9-12-r169.
Warnecke T, Batada NN, Hurst LD. 2008. The impact of the nucleosome code on protein-coding sequence evolution in yeast. *PLoS Genet* **4:** e1000250. doi: 10.1371/journal.pgen.1000250.
Washietl S, Machné R, Goldman N. 2008. Evolutionary footprints of nucleosome positions in yeast. *Trends Genet* **24:** 583–587.
Whelan S, Goldman N. 1999. Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Mol Biol Evol* **16:** 1292–1299.
Williams EJ, Hurst LD. 2002. Is the synonymous substitution rate in mammals gene-specific? *Mol Biol Evol* **19:** 1395–1398.
Woltering J, Duboule D. 2009. Conserved elements within open reading frames of mammalian Hox genes. *J Biol* **8:** 17. doi: 10.1186/jbiol116.
Wright F. 1990. The 'effective number of codons' used in a gene. *Gene* **87:** 23–29.
Xing Y, Lee C. 2005. Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proc Natl Acad Sci* **102:** 13526–13531.
Xing Y, Lee C. 2006. Can RNA selection pressure distort the measurement of K_a/K_s? *Gene* **370:** 1–5.
Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* **15:** 496–503.
Yang Z, Nielsen R, Goldman N, Pedersen AK. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155:** 431–449.
Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11:** 377–394.
Yoshida H, Matsui T, Yamamoto A, Okada T, Mori K. 2001. XBP1 mRNA is induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor. *Cell* **107:** 881–891.

Received April 5, 2010; accepted in revised form June 15, 2010.

Short Specialist Review

Gene finding using multiple related species: a classification approach

Manolis Kellis
MIT Broad Institute for Biomedical Research, Cambridge, MA, USA

1. Introduction

Ideally, we should be able to systematically discover all the functional genes in a newly sequenced genome from its sequence alone. Computational discovery methods rely both on the direct signals used by the cell to guide transcription, splicing, and translation, and also on indirect signals such as evolutionary conservation. In this paper, we summarize the principles of a classification-based approach for systematic gene identification, on the basis of comparative sequence information from multiple, closely related species. We first frame gene identification as a classification problem of distinguishing real genes from spurious gene predictions. We then present the Reading Frame Conservation (RFC) test, a new computational method implementing such a classification approach, on the basis of the patterns of nucleotide changes in the alignment of orthologous regions. We finally summarize our results of applying this method to reannotate the yeast genome, and the challenges of using related methods to discover all functional genes in the human genome.

2. Defining real genes

What is a real gene? This question is relatively easy to answer for those genes that are abundantly expressed, encode well-characterized proteins, and whose disruption affects a specific function in the organism. Beyond these, the distinction is much more subtle between functional genes and spurious gene predictions.

Experimentally, the definition of a functional gene comes largely as accumulating evidence of its usage, including gene expression (Velculescu, 1997), protein fragments (Rezaul *et al*., 2005), biochemical function (Jackman *et al*., 2003), protein interactions (Bai and Elledge, 1997), or the effect of its disruption (McAlister and Holland, 1982). It is important to note, however, that absence of experimental evidence does not imply that a gene annotation is spurious: a real

gene may be missing experimental evidence because it is not used in the particular conditions surveyed. Conversely, any individual report of gene usage could be due to experimental noise, cross-hybridization, or chance transcription events due to a basal level of intergenic transcription. Thus, experimental evidence alone is insufficient to distinguish between real genes and spurious gene predictions.

Computationally, the processes of transcription, splicing, and translation can be thought of as a series of decisions taken by the cell, based on signals in the genome. These signals include distant enhancer elements, regulatory elements surrounding the transcription start site, splicing enhancer and repressor signals in the message, and translation signals in the sequence or structure of mature mRNAs (Fairbrother *et al*., 2002; Wang, 2004; Thanaraj and Robinson, 2000; Yeo and Burge, 2004). The subset of these signals that we currently understand is insufficient to specify the set of known genes. Thus, in addition to the *direct* signals used by the cell, gene identification methods (Burge and Karlin, 1997; Majoros *et al*., 2004; Kulp *et al*., 1996; Krogh, 1997; Henderson *et al*., 1997; Stanke and Waack, 2003; Salzberg *et al*., 1998) routinely identify genes using additional *indirect* signals (reviewed in Fickett and Tung, 1992), which are not available to the cell, although they are generally good indicators of protein-coding genes. These include the frequency of each codon in protein-coding regions, the overall length of the translated protein product, and importantly, the evolutionary conservation of protein sequences across related species.

Evolutionary conservation is perhaps the strongest indicator that a predicted gene is functional. A gene that confers even a slight evolutionary advantage can be conserved across millions of years, regardless of the rarity of its usage. Hence, even if experimental methods fail to detect the usage of a gene in a given set of experimental conditions, evolutionary methods are able to detect *indirect* evidence of its usage, by observing the pressure to preserve its function over millions of years.

3. Gene finding as a classification problem

The challenge of using evolutionary conservation to define genes lies in the ability to reject spurious genes. Although it is generally well understood that genes conserved across large evolutionary distances are functional, lack of conservation is generally attributed to evolutionary divergence, thus providing no evidence toward either accepting or rejecting a gene prediction.

Yet, the comparison of related genomes contains information much richer than simply the presence or absence of a protein in a genome. By working with closely related genomes, we can define regions of conserved gene order, or *synteny blocks*, which span several genes, and potentially entire chromosomes. Defining conserved synteny blocks allows us to construct global alignments, spanning both well-conserved regions and regions of low conservation. In particular, these give us access to full nucleotide alignments for all predicted genes and for all intergenic regions within orthologous segments.

With the availability of orthologous alignments for both protein-coding and non-coding regions, we can study their distinct properties and build a classifier between the two types of region. The simplest and most commonly used such classifier

observes the overall level of nucleotide conservation in the alignment, and selects regions of high conservation that are likely to be functional. Other classifiers may observe more subtle signals, such as the number of amino acid changes per nucleotide substitution (Hurst, 2002), the frequency of insertions and deletions, the periodicity of mutations, and so on. By working with properties of the nucleotide alignment, rather than protein alignment, we are able to apply it uniformly to evaluate both genes and intergenic regions in the same test.

One particular conservation property unique to protein-coding segments is the pressure to preserve the reading frame of translation. Since protein sequences are translated every three nucleotides, the length of insertions and deletions (indels) is largely constrained to remain a multiple of three, thus preserving the frame of translation. Within coding regions, indels that disrupt the frame of translation are excluded, or compensated with nearby indels that restore the reading frame. In noncoding regions, the length of indels does not have this constraint, and short spacing changes are tolerated.

To evaluate this property quantitatively, we developed the RFC test. This test evaluates the pressure to preserve the reading frame in a fully aligned interval, by measuring the portion of nucleotides in this interval for which the reading frame has been locally conserved (Kellis *et al*., 2004). The RFC test provides a classifier between coding and noncoding regions that is completely independent of the traditional signals used in gene identification. It does not rely on start, stop, or splicing signals, nor does it rely on the conservation of protein sequence. It can, therefore, be combined with existing gene-finding tools and provide a highly informative score for any interval considered.

4. Reannotation of yeast

A classification approach is particularly well suited for the yeast genome. The general scarcity of introns makes it hard to rely on splicing signals to discover genes. Thus, the annotation of *Saccharomyces cerevisiae* has traditionally relied solely on the length of predicted proteins to annotate genes, resulting in 6062 annotated open reading frames (ORFs), which potentially encode proteins of at least 100 amino acids (aa). Additionally, a tentative functional annotation has been inferred for as many as 3966 of these ORFs, on the basis of classical genetic experiments and systematic genome-wide studies of gene expression, deletion phenotype, and protein–protein interaction. Together, the interval-based annotation and the large set of well-known genes make it possible to apply the RFC test systematically to evaluate the functional significance of the remaining ORFs.

To apply the test, we constructed multiple sequence alignments for every ORF and every intergenic region across four closely related species. We sequenced and assembled the genomes of *S. paradoxus*, *S. mikatae*, *and S. bayanus* (Kellis *et al*., 2003), and defined genome-wide synteny blocks with *S. cerevisiae*, on the basis of discrete anchors provided by unique protein blast hits (Kellis *et al*., 2004). Within these synteny blocks, we constructed global alignments of orthologous genes and intergenic regions across the four species using CLUSTALW (Higgins and Sharp, 1988), and systematically evaluated each alignment. We compared *S. cerevisiae* to

each species in turn, and every comparison cast a vote based on its overall RFC and a species-specific cutoff. A decision was then reached for each gene by tallying the votes from all comparisons.

We evaluated the sensitivity and specificity of the approach based on the 3966 genes with functional annotations ("known genes"), and 340 randomly chosen intergenic sequences with lengths similar to the annotated ORFs. The RFC test correctly accepted 3951 known genes (99.6%) and rejected only 15 known genes; upon manual inspection, these 15 are indeed likely to be spurious (most lack experimental evidence, and deletion phenotypes of the rest are likely due to their overlap with the promoter of other known genes). The method also correctly rejected 326 intergenic regions (96%), accepting only 14 intergenic regions (of which 10 appear to define short ORFs or extend annotated ORFs, suggesting that at most 1% of true intergenic regions failed to be rejected by the RFC test). In summary, the RFC test shows a very strong discrimination between genes and intergenic regions, with sensitivity and specificity values greater than 99%.

We then applied the RFC test to all previously annotated genes, leading to a major revisiting of the yeast gene set. For ORFs with lengths greater than 100 aa, our analysis accepted 5538 ORFs and rejected 503 (of which 376 were immediately rejected, 105 were rejected with additional criteria, and 32 were merged with neighboring ORFs); the classifier abstained from making a decision in 20 cases. The rejected ORFs show an abundance of frame-shifting indels across their entire length, in-frame stop codons, and low conservation of protein sequence, in addition to the low RFC score; their length distribution and atypical codon usage additionally suggest that they are likely occurring by chance (Goffeau, 1996; Dujon, 1994; Sharp and Li, 1987); furthermore, previous systematic experimentation showed no compelling evidence that these may encode a functional gene. Thus, it appears that more than 500 previously annotated ORFs in the yeast genome are spurious predictions.

Below the 100 amino acid cutoff, no previous systematic annotation or experimentation was available. Thus, to validate our method, we compared the results of the RFC test with an independent metric: the proportion of the *S. cerevisiae* ORF that was also free of stop codons in the other three species. Between 50 and 99 aa, we found that the method is still reliable and reports 43 candidate new genes at that length. As ORFs smaller than 50 aa were tested, we found that the specificity of the test decreased, since small intervals tend to be devoid of indels by chance rather than by presence of selective pressure. Thus, additional constraints, and additional species, will be needed to discover genes reliably at such short lengths.

In addition to the discovery of genes themselves, the comparative analysis allowed us to refine gene boundaries. Once an interval was determined to be under selective pressure for RFC, the boundaries of that interval were adjusted on the basis of the conservation of start/stop/splice signals and the boundaries over which the reading frame is conserved. This led to a large-scale reannotation of gene structure, which affected hundreds of genes (146 start codon changes, 67 intron changes, 32 merges of consecutive ORFs, and 45 changes of ORF ends). It is worth noting that in 134 cases, the inferred boundary changes pinpointed sequencing errors in the primary sequence of *S. cerevisiae*, ~50 of which were tested and

corrected by resequencing. These boundary changes reveal the true location of promoter regions, new protein domains in elongated ORFs, and previously overlooked functional relationships in the case of merged ORFs

In summary, the RFC test was able to reliably distinguish between genes and intergenic regions, leading to a systematic reannotation of the yeast genome. The comparative analysis led to the rejection of more than 500 previously annotated genes, and to the discovery of many novel genes. The results agree with similar comparative analyses carried out from a number of yeast species (Cliften *et al*., 2003; Blandin, 2000; Wood *et al*., 2001; Brachat, 2003). In addition, it allowed us to refine the gene structure of hundreds of genes, adjusting start and stop boundaries, merging consecutive ORFs, and discovering many new introns. Moreover, by using multiple species, the signals leading to gene identification were powerful enough to pinpoint sequencing errors in any individual species, including *S. cerevisiae*.

5. Implications for the human genome

The challenge of gene finding in the human genome is far greater than in yeast, due to the vastly larger intergenic regions, numerous and large introns, small exons, and alternatively spliced genes. Most approaches to gene finding in higher eukaryotes have relied on Hidden Markov Models (Burge and Karlin, 1997; Majoros *et al*., 2004; Kulp *et al*., 1996; Krogh, 1997; Henderson *et al*., 1997; Stanke and Waack, 2003), which inherently emphasize the importance of exon chaining and rely on knowledge of the expected length distributions for both exons and introns. These have been recently extended to use sequences from multiple species (Yeh *et al*., 2001; Siepel and Haussler, 2004a; Siepel and Haussler, 2004b; Batzoglou *et al*., 2000; Korf *et al*., 2001; Dewey *et al*., 2004; Rinner and Morgenstern, 2002; Parra, 2003; Meyer and Durbin, 2002; Novichkov *et al*., 2001). These approaches have limitations however, in cases of alternative splicing (Brett, 2000; Mironov *et al*., 1999; Cawley and Pachter, 2003), differences in splicing between species (Nurtdinov *et al*., 2003), widely varying exon and intron lengths, and noncanonical splice sites.

An exon-based classification approach to gene finding can help overcome these limitations. By exhaustively enumerating and testing all candidate protein-coding intervals, classification approaches make exon detection independent of chaining. Splice site models can be applied to each species independently and then compared with each other, which is much stronger than the evidence in any one species alone. The intervals can be then tested on the basis of discriminating variables that distinguish genes from noncoding regions (Fickett and Tung, 1992; Goldman and Yang, 1994). New discriminating variables can be defined as alignments of each type of region are systematically compared, and these can be combined and weighted, leveraging traditional machine-learning techniques for feature selection and classification (Moore and Lake, 2003). Once relevant intervals have been identified, their boundaries can be adjusted for optimal chaining into complete genes. In particular, chaining can also leverage an inferred frame of translation for each exon, based on the higher mutation rate observed in largely degenerate third codon positions (Hurst, 2002). The exon-chaining step produces full gene models,

and is able to cope with alternative splicing and missing exons, since it is not constrained to a single optimal path through the exons.

By systematically observing alignment properties of large sequence regions, we can build new rigorous approaches for sequence analysis. These are widely applicable beyond coding exons, and similar classification-based approaches can be used to distinguish CpG islands, 5′- and 3′-untranslated regions, promoter regions, regulatory islands, and other functional elements. Through the lens of evolutionary selection, our ability to directly interpret genomes is revolutionized. Coupled with systematic experimentation and validation, these analyses can lead to a systematic catalog of functional elements in the human genome, forming the future foundations of biomedical research.

References

Bai C and Elledge SJ (1997) Gene identification using the yeast two-hybrid system. *Methods in Enzymology*, **283**, 141–156.

Batzoglou S, Pachter L, Mesirov JP, Berger B and Lander ES (2000) Human and mouse gene structure: Comparative analysis and application to exon prediction. *Genome Research*, **10**, 950–958.

Blandin G, Durrens P, Tekaia F, Aigle M, Bolotin-Fukuhara M, Bon E, Casaregola S, de Montigny J, Gaillardin C, Lepingle A, *et al*. (2000) Genomic exploration of the hemiascomycetous yeasts: 4. The genome of *Saccharomyces cerevisiae* revisited. *FEBS Letters*, **487**, 31–36.

Brachat S, Dietrich FS, Voegeli S, Zhang Z, Stuart L, Lerch A, Gates K, Gaffney T and Philippsen P (2003) Reinvestigation of the *Saccharomyces cerevisiae* genome annotation by comparison to the genome of a related fungus: *Ashbya gossypii*. *Genome Biology*, **4**, R45.

Brett D, Hanke J, Lehmann G, Haase S, Delbruck S, Krueger S, Reich J and Bork P (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Letters*, **474**, 83–86.

Burge C and Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, **268**, 78–94.

Cawley SL and Pachter L (2003) HMM sampling and applications to gene finding and alternative splicing. *Bioinformatics*, **19**(Suppl 2), II36–II41.

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA and Johnston M (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, **301**, 71–76.

Dewey C, Wu JQ, Cawley S, Alexandersson M, Gibbs R and Pachter L (2004) Accurate identification of novel human genes through simultaneous gene prediction in human, mouse, and rat. *Genome Research*, **14**, 661–664.

Dujon B, Alexandraki D, Andre B, Ansorge W, Baladron V, Ballesta JP, Banrevi A, Bolle PA, Bolotin-Fukuhara M and Bossier P (1994) Complete DNA sequence of yeast chromosome XI. *Nature*, **369**, 371–378.

Fairbrother WG, Yeh RF, Sharp PA and Burge CB (2002) Predictive identification of exonic splicing enhancers in human genes. *Science*, **297**, 1007–1013.

Fickett JW and Tung CS (1992) Assessment of protein coding measures. *Nucleic Acids Research*, **20**, 6441–6450.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, *et al*. (1996) Life with 6000 genes. *Science*, **274**, 546, 563–567.

Goldman N and Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, **11**, 725–736.

Henderson J, Salzberg S and Fasman KH (1997) Finding genes in DNA with a hidden Markov model. *Journal of Computational Biology*, **4**, 127–141.

Higgins DG and Sharp PM (1988) CLUSTAL: A package for performing multiple sequence alignment on a microcomputer. *Gene*, **73**, 237–244.

Hurst LD (2002) The Ka/Ks ratio: Diagnosing the form of sequence evolution. *Trends in Genetics*, **18**, 486.

Jackman JE, Montange RK, Malik HS and Phizicky EM (2003) Identification of the yeast gene encoding the tRNA m1G methyltransferase responsible for modification at position 9. *RNA*, **9**, 574–585.

Kellis M, Patterson N, Birren B, Berger B and Lander ES (2004) Methods in comparative genomics: Genome correspondence, gene identification and regulatory motif discovery. *Journal of Computational Biology*, **11**, 319–355.

Kellis M, Patterson N, Endrizzi M, Birren B and Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.

Korf I, Flicek P, Duan D and Brent MR (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics*, **17**(Suppl 1), S140–S148.

Krogh A (1997) Two methods for improving performance of an HMM and their application for gene finding. *Proceedings/International Conference on Intelligent Systems for Molecular Biology*, **5**, 179–186.

Kulp D, Haussler D, Reese MG and Eeckman FH (1996) A generalized hidden Markov model for the recognition of human genes in DNA. *Proceedings/International Conference on Intelligent Systems for Molecular Biology*, **4**, 134–142.

Majoros WH, Pertea M and Salzberg SL (2004) TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics*, **20**, 2878–2879.

McAlister L and Holland MJ (1982) Targeted deletion of a yeast enolase structural gene. Identification and isolation of yeast enolase isozymes. *The Journal of Biological Chemistry*, **257**, 7181–7188.

Meyer IM and Durbin R (2002) Comparative *ab initio* prediction of gene structures using pair HMMs. *Bioinformatics*, **18**, 1309–1318.

Mironov AA, Fickett JW and Gelfand MS (1999) Frequent alternative splicing of human genes. *Genome Research*, **9**, 1288–1293.

Moore JE and Lake JA (2003) Gene structure prediction in syntenic DNA segments. *Nucleic Acids Research*, **31**, 7271–7279.

Novichkov PS, Gelfand MS and Mironov AA (2001) Gene recognition in eukaryotic DNA by comparison of genomic sequences. *Bioinformatics*, **17**, 1011–1018.

Nurtdinov RN, Artamonova II, Mironov AA and Gelfand MS (2003) Low conservation of alternative splicing patterns in the human and mouse genomes. *Human Molecular Genetics*, **12**, 1313–1320.

Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW and Guigo R (2003) Comparative gene prediction in human and mouse. *Genome Research*, **13**, 108–117.

Rezaul K, Wu L, Mayya V, Hwang SI and Han DK (2005) A systematic characterization of mitochondrial proteome from human T leukemia cells. *Molecular and Cellular Proteomics*, **4**, 169–181.

Rinner O and Morgenstern B (2002) AGenDA: Gene prediction by comparative sequence analysis. *In Silico Biology*, **2**, 195–205.

Salzberg S, Delcher AL, Fasman KH and Henderson J (1998) A decision tree system for finding genes in DNA. *Journal of Computational Biology*, **5**, 667–680.

Sharp PM and Li WH (1987) The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, **15**, 1281–1295.

Siepel A and Haussler D (2004a) Combining phylogenetic and hidden Markov models in biosequence analysis. *Journal of Computational Biology*, **11**, 413–428.

Siepel A and Haussler D (2004b) Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Molecular Biology and Evolution*, **21**, 468–488.

Stanke M and Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19**(Suppl 2), II215–II225.

Thanaraj TA and Robinson AJ (2000) Prediction of exact boundaries of exons. *Briefings in Bioinformatic*, **1**, 343–356.

Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE Jr, Hieter P, Vogelstein B and Kinzler KW (1997) Characterization of the yeast transcriptome. *Cell*, **88**, 243–251.

Wang Z, Rolish ME, Yeo G, Tung V, Mawson M and Burge CB (2004) Systematic identification and analysis of exonic splicing silencers. *Cell*, **119**, 831–845.

Wood V, Rutherford KM, Ivens A, Rajandream M-A and Barrell B (2001) A re-annotation of the *Saccaromyces cerevisiae* genome. *Comparative and Functional Genomics*, **2**, 143–154.

Yeh RF, Lim LP and Burge CB (2001) Computational inference of homologous gene structures in the human genome. *Genome Research*, **11**, 803–816.

Yeo G and Burge CB (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biology*, **11**, 377–394.

A Phylogenomic Approach to the Evolutionary Dynamics of Gene Duplication in Birds

Chris L. Organ[1*], Matt Rasmussen[2*], Maude W. Baldwin[1], Manolis Kellis[2], Scott V. Edwards[1]

[1]*Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, MA 02138.* [2]*Massachusetts Institute of Technology, Center for Genome Research, 320 Charles Street, Cambridge, MA 02141*

*These authors contributed equally

Corresponding author contact information:
Chris Organ
Museum of Comparative Zoology
Department of Organismic and Evolutionary Biology
26 Oxford Street
Harvard University
Cambridge, MA 02138
Tel (617) 496-9389
Fax (617) 495-5667

Keywords: gene duplication, phylogenomics, comparative genomics, birds

Running title: The Evolutionary Dynamics of Gene Duplication in Birds

Abstract: Gene duplication is a fundamental aspect of genome evolution that produces large and small gene families of (usually) related function. We perform a phylogenomic analysis of gene duplication in the chicken (*Gallus gallus*) to characterize the dynamics and evolution of gene duplication on the evolutionary line to birds. In *Gallus*, the distribution of the number of paralogs per gene family is heavily skewed towards small families. This finding is in accord with other studies that find gene family size typically follows a power-law distribution in animals, a pattern thought to be produced by differential rates of pseudogenization among families. We also test for within-family evolutionary rate variation in *Gallus*, finding that the vast majority of gene families exhibit substantial rate variation among lineages. This rate variation probably stems from two sources: natural deviations in the clock as commonly found, for example, in phylogenetic analyses of different species; and bursts of adaptive evolution among newly evolved gene family members. The age of gene duplications in *Gallus* are distributed exponentially, with most duplications occurring recently, a pattern consistent with analyses on other eukaryotes. Taken together, these results begin to reveal the dynamics of gene family evolution in birds, the most speciose group of living amniotes, though whole genome data are required from more bird and reptile species to fully understand patterns of gene gain and loss in this group.

Introduction

New genes are thought to primarily arise through a process of gene duplication. Genes that are homologous as a result of divergence across lineages via speciation are said to be orthologous, whereas genes that are homologous as a result of gene duplication are paralogous (Li 2006). Paralogous genes that are functionally redundant and selectively nearly neutral can result in one copy being mutated into a functionless sequence called a pseudogene, or they can be deleted altogether. On the other hand, some duplicated genes can be beneficial from their time of origin because of dosage effects (Kondrashov, Rogozin et al. 2002) and may ultimately be important for speciation. In Passeriformes (perching birds) this may be the case for growth hormone (GH) paralogs, which have undergone differential selection since their divergence (Yuri, Kimball et al. 2008). In a process called subfunctionalization, each paralog adopts partial function of their ancestral gene (Nowak, Boerlijst et al. 1997; Lynch and Force 2000). Changes in gene expression immediately following duplication as a result of subfunctionalization appear to be common (Gu, Nicolae et al. 2002). Duplicated genes can also diverge to produce novel functions in a process known as neofunctionalization (Zhang 2003). For example, some duplicated members of the *RNaseA* gene superfamily in primates evolved a novel antibacterial function that was not present in the common ancestral gene or its descendants (Zhang, Rosenburg et al. 1998). Through the acquisition of novel functions, gene duplication plays a vital role in generating diversity at both molecular (Eirin-Lopez, Gonzalez-Tizon et al. 2004) and organismal (Hittinger and Carroll 2007) levels and can increase the phenotypic complexity and diversity of animals (Ohno 1970; Maniatis and Tasic 2002). Protein coding genes, regulatory genes, and RNA non-coding genes are all subject to gene duplication. Gene duplication occurs in all three major domains of life; Bacteria, Archaea, and Eukarya (Zhang 2003), and comparative analyses have suggested an average origin rate of new gene duplicates on the order of 1 per gene per 100 million years (Lynch and Conery 2000).

The ubiquity of gene duplication and its power to generate material on which selection may act is an especially interesting topic in birds because of their uniquely structured genomes. Birds have smaller genomes than any other amniote group (Gregory 2002), but they also have a substantially reduced density of active repetitive elements (Shedlock 2006), and segmental duplications and pseudogenes (Hillier, Miller et al. 2004). Birds may also have fewer protein-coding genes in their genomes than mammals, with roughly 18,000 in chicken compared with approximately 22,000 in human (Hillier, Miller et al. 2004). By comparing the genomes of chicken with those of human and puffer fish (*Fugu*), the difference in gene count can be explained by significantly less lineage-specific gene duplication along the lineage of birds. For example, gene duplication by retroposition appears to be rare in birds compared with mammals, which have roughly 300 times the number of retrotransposed gene duplicates than in chicken (over 15,000 compared with 51) (Hillier, Miller et al. 2004). Of the 51 duplicates detected in the chicken genome by Hillier, Miller et al. (2004), 36 appear to be pseudogenes. This is probably due in part to the abundance of long interspersed nuclear elements (LINEs) in mammal genomes, which are likely responsible for the reverse transcription of retrotransposed gene duplicates. Chicken repeat 1 (CR1) is the dominate active transposable element in chicken and other archosaurs (Shedlock et al. 2007), a feature that may account for the lack of retrotransposed gene duplicates due to this element's inability to copy polyadenylated mRNA (Haasa, Grabowskia et al. 2001; Hillier, Miller et al. 2004).

The whole-genome sequencing of the chicken (*Gallus gallus*) provided an unprecedented window into the architecture of bird genomes (Hillier, Miller et al. 2004). This study revealed a number of important details of gene duplication. For example, most of the expansions in gene families within chicken are associated with the immune system and host defense against parasites (Ota and Nei 1994; Nei et al. 1997). The chicken genome project found an expansion of the Scavenger Receptor Cysteine-Rich (SRCR) domain (Hillier, Miller et al. 2004), a highly conserved protein module involved in the innate immune system (Sarrias, Grønlund et al. 2004). Linked with the major histocompatibility complex (MHC) class I gene cluster in humans are certain olfactory receptors which also underwent a lineage-specific expansion within the chicken. These olfactory receptor genes, related to two orthologs in human (*OR5U1* and *OR5BF1*), appear to have expanded within birds relatively recently to constitute the majority of the over 200 olfactory receptor genes in the chicken genome. The expansions of these genes may be part of the genetic mechanism linking the immune and olfactory systems with mate choice, kin recognition, and social interactions in birds (Zelano and Edwards 2002). Other gene expansions were also likely important for key innovations in birds, such as expansions in the keratin gene family, which are the proteinaceous building blocks of feathers and therefore vital for thermoregulation, sexual display, and flight. However, even in the absence of selection (most gene duplicates are thought to be pseudogenized during evolution) they may still play a critical role in creating postmating reproductive barriers that aid in speciation (Lynch and Conery 2000).

Age of gene duplications within lineages is also of fundamental interest in the study of gene family evolution. Analysis of animal, plant and fungal genomes has shown that the majority of duplications are recent, because most duplicated genes are thought to be pseudogenized shortly after the duplication event (Lynch and Conery 2000). It is currently unknown if the small, streamlined genomes of birds deviate from this pattern, given the fewer number of total genes, paucity of transposable elements, and highly recombinant microchromosomes found in chicken.

Here we explore the evolution of chicken gene families within the larger context of amniote evolution. A specific analytical strategy called phylogenomics (Eisen 1998) attempts to systematically reconstruct the phylogeny of gene families across multiple complete genomes so as to deduce putative protein homologies and functions as well as the specific gene duplications and losses responsible for gene family diversity. The phylogenomic approach to multigene family evolution has shown recent success in many different clades, such as vertebrates (Zmasek and Eddy 2002; Storm and Sonnhammer 2003; Li, Coghlan et al. 2006; Huerta-Cepas, Dopazo et al. 2007), and 16 fungi species (Wapinski, Pfeffer et al. 2007). Recently Rasmussen and Kellis (2007) deduced a pattern of substitution rates within and among *Drosophila* and yeast species by using machine-learning to compare the gene trees in multigene families to the presumed species tree. In addition, Heger and Ponting (2007) developed a comprehensive database of orthologous and paralogous gene sets for several clades, including amniotes. Here we follow a complementary approach to characterize the dynamics and evolution of gene duplication within birds. Whereas many previous approaches to recognizing paralogs use reciprocal BLAST hits and pairwise comparison of orthologs, we apply Ensembl phylogenies and a variety of phylogenetic and comparative analyses within a clade of two dozen sequenced amniote species to analyze the dynamics of gene duplication in the complete chicken (*Gallus gallus*) genome.

Computational Approach

For our analysis, we obtained 25,363 gene trees from Ensembl's (v50) gene tree database. Ensembl's gene trees were produced with their custom pipeline of computational analysis. Initially, family clusters were defined by single-linkage clustering of best reciprocal BLAST hits between amino acid translations of all genes within the Ensembl database (Hubbard, Aken et al. 2007). Each cluster was then processed by Ensembl to create a multiple amino acid alignment using the alignment algorithm MUSCLE (Edgar 2004) and a maximum likelihood phylogenetic tree using PHYML (Guindon and Gascuel 2003). Our database of gene trees assembled from Ensembl contains 611,441 genes from 39 species, including 25 mammals, *Gallus gallus*, *Xenopus tropicalis*, five fish, six invertebrates, and *Saccharomyces cerevisiae*. There are 17,487 annotated chicken (*Gallus gallus*) genes in the Ensembl database, 13,649 (78%) of which belong to one of the 7,785 gene trees.

To study gene duplication along the bird lineage, we identified bird/mammalian duplications in each gene tree (Figure 1 and Figure 2). This event is represented as a node that is parental to one clade of chicken-specific genes and another clade of mammalian-only genes. Although we describe the chicken paralogs as chicken-specific, we do so only because chicken is the only bird in our analysis; presumably, sampling of additional genomes within Reptilia would reveal many of these genes and gene duplication to be shared with other birds and non-avian reptiles. If duplications occurred prior to the common ancestor of birds and mammals, a single family may contain multiple gene divergences. In addition, some of these gene duplications can be partially obscured when there is complete gene loss in either the bird or mammalian lineage. We were able to identify 12,094 unambiguous amniote duplications (using *Gallus* and *Homo*). For each duplication, we isolated the subtree of amniote genes rooted at the amniote ancestor for further analysis. Outgroup species (fish; *Fugu*) were used to help position the subtree root. We expect this protocol to yield chicken gene families that differ from those circumscribed, for example, solely by reciprocal BLAST hits; sometimes our definition of chicken gene families include genes not detected by non-phylogenetic methods, whereas in other instances our approach will miss some genes that manual inspection and curation would have revealed. On the other hand, our approach has the advantage of being objective and repeatable, and can easily be extended to study the dynamics of gene duplication in other taxa.

Dynamics of Chicken-Specific Gene Duplication

We found that the distribution of the number of paralogs per chicken gene family was heavily skewed towards small families (Figure 3). Whereas nearly 30 gene families had three members, only six families had more than nine. These figures focus only on those gene families that duplicated after the divergence of chicken from the amniote ancestor (by comparing *Gallus* with *Homo*). Thus very ancient gene families – many potentially with large numbers of family members – do not figure into this calculation. Still, we were surprised by the small number of very large (> 20) gene families in the chicken. Gene family size has been found to follow a power-law distribution in animals, a pattern thought to be produced by differential rates of pseudogenization among families (Huynen and van Nimwegen 1998; Hughes and Liberles 2008). Moreover, birth-death models and purifying selection appear to account for much of the conservation seen within lineage-specific paralogs (Ota and Nei 1994; Piontkivska, Rooney et al. 2002; Piontkivska and Nei 2003; Eirin-Lopez, Gonzalez-Tizon et al. 2004).

Substitution rates within amniote lineages are known to be quite variable, and we expect similar rate variation among paralogous members of chicken gene families. We estimated dates of divergence for each node in our chicken multigene family trees by using a penalized likelihood model, as implemented in the r8s rate analysis program (Sanderson 2003), and by using 310 million years as an estimate for the *Gallus*/*Homo* common ancestor (Benton and Donoghue 2007). The penalized likelihood model finds the optimal trade-off between maximizing the likelihood of a Poisson process for nucleotide substitution and a penalty term for rate variation between neighboring branches. The weighting of these two terms is determined by a coefficient λ, such that higher values of λ greatly penalize rate variation in favor of a clock-like model, and lower values allow a large amount of rate variation. Using a cross-validation procedure (Sanderson 2003), we determined the optimal choice of λ from the possible values $10^{-2} - 10^{3}$. Within our amniote gene trees we found that over 54 % of them were best explained by a λ value of 10^{-2}, indicating that the vast majority of gene families exhibit substantial rate variation among lineages (Figure 4). This rate variation probably stems from two sources: natural deviations in the clock as commonly found, for example, in phylogenetic analyses of different species; and bursts of adaptive evolution among newly evolved gene family members. Under the first hypothesis, it might be expected that larger gene families would exhibit more rate variation than small ones, and that the incidence of rate variation would increase with family size. However, we did not find this trend when we regressed λ on gene family size (Figure 4). Many of the gene families best explained by a log λ value of -2 showed levels of divergence that suggests duplication since the Cenozoic era, especially since the Neogene period. For this reason, we suspect that much of the rate variation among gene family members may in fact be due to adaptive bursts, because generation time effects among different lineages of birds are only expected to influence rate variation for those gene families that duplicated prior to the chicken's divergence from other lineages. Of the other categories of substitution rate variation among gene family members, the class best explained by $\lambda = \log(3)$ was the next most common. Substitution rates among gene family members in this category are fairly clock-like.

The age of gene duplications in chicken are distributed exponentially, with most duplications occurring recently (Figure 5). This pattern is consistent with previous analyses (Lynch and Conery 2000) and suggests that, assuming a relatively constant rate of gene duplication, most genes are pseudogenized or eliminated from the genome soon after duplication. This pattern could also suggest that concerted evolution might be more common among chicken gene families than in other groups. For example, concerted evolution among major histocompatibility complex (MHC) gene copies in birds is thought to occur more frequently, and over a shorter time scale, than in mammals (Hess and Edwards 2002), and the phylogenetic scale over which MHC orthologs can be identified in birds is probably much smaller than in mammals.

Examples of families with chicken–specific duplications

Gene family composition is shaped both by gene gain and loss, yet as other researchers have noted (Furlong 2005), gene family expansion is easier to detect, especially when annotation is not complete and gaps still remain in recent genome builds. We examined several gene families containing lineage-specific expansions in the chicken, using the amniotic subtree rooted at the chicken-human divergence. In Table 1, we describe the dynamics of five representative families: Toll-like receptors, hemoglobin, ovalbumin-related serpins, four sub-families of olfactory receptors, and keratin. These families were selected for their variety in size, age and

function, and because the annotation and family membership could be at least partially cross-validated with recent studies.

Toll-like receptors

Temperley and colleagues (2008) describe the evolutionary history and chromosomal location of chicken toll-like receptors (TLRs), a family that is part of the innate immune system and is characterized by an ancient, highly conserved pathogen-recognition domain that triggers an inflammatory response. These authors discovered that while chickens and humans both have ten receptors, only four genes in chicken maintain one-to-one orthology with mammalian genes; much gain and loss has occurred in every lineage. In the chicken, a duplication event estimated at 67 million years ago (mya) gave rise to *TLR2A* and *TLR2B*, orthologs to the single *TLR2* in mammals. Three other genes, two that duplicated in tandem (*TLR1LA* and *TLR1LB*, estimated duplication time 147 mya), as well as *TLR15* have no mammalian counterpart. Other mammalian members have been pseudogenized or fully lost in chicken. Using our automated phylogenetic approach, we are able to analyze one of these chicken-specific expansions: the duplication event that gave rise to *TLR2A* and *TLR2B* in chickens. We obtained nearly the same estimate of duplication time (67 mya). As with majority of these five families that we looked at in detail, there was considerable rate variation, as log λ was -2. The amount of sequence evolution (point mutations) in the Toll-like receptor family is greater than 50.8% of sequence evolution in other chicken gene families. We could not analyze the other duplication in chicken, as our dataset from Ensembl was missing one of the chicken-specific genes (*TLR1LB*). *TLR15*, another gene unique to birds, had no mammalian ortholog, so it also was not included in our amniote subtree.

Ovalbumin-related serpins

Another gene family with documented chicken-specific expansions are the ov-serpins, also called the ovalbumin-related serpins, or clade B serpins. Benarafa and O'Donnell (2005) examine the phylogenetic relationship between the chicken members (some of which function as egg-white storage proteins) and their mammalian counterparts (involved in diverse roles like embryogenesis, inflammation regulation, and angiogenesis). The initial duplication is thought to have occurred very early in the vertebrate lineage; and, like TLRs, the family is also marked by recent lineage-specific expansions and losses. Chickens have 10 members and humans have 13 members. Three genes in chicken-- *ovalbumin*, and ovalbumin-like genes *X* and *Y* -- are paralogs and lack a human ortholog. Another gene, with a single human ortholog, seems to have duplicated to produce the chicken genes *Serpinb10* and *MENT* (Mature Erythrocyte Nuclear Termination state-specific protein). The remaining family members from chicken each have single human orthologs. Among the two subfamilies with chicken-specific expansions, rate variation is substantial (log $\lambda = -2$). Moreover, the amount of sequence evolution (point mutations) in the subfamily containing *ovalbumin* and ovalbumin-like genes *X* and *Y* is greater than 69.6% of sequence evolution in other chicken gene families that have duplicated since the chicken-mammalian split while one of the other ovalbumin subfamilies is similarly divergent (77.6%; *Serpinb10* and *MENT*).

Hemoglobin

Metabolic rate is an important trait that governs many organismal characters, from growth strategies to sustained physical activity. In amniotes, an elevated metabolism

(endothermy) has only evolved within two extant groups, birds and mammals, although paleontologists suspect that many extinct dinosaurian lineages possessed endothermy (de Ricqlès, Padian et al. 2001; Horner, Padian et al. 2001). Whereas the typical mammalian and avian condition is homeothermy (roughly constant body temperature), some birds, such as swifts, hummingbirds, and nightjars are facultatively poikilothermic, a condition in which their usually elevated body temperature can vary over a wider range than that seen in mammals. The hemoglobin multigene family is closely associated with metabolism and the respiratory system.

Hemoglobin, a multi-domain protein, has rapidly diversified with vertebrate lineages (Gribaldo, Casane et al. 2003). For example, α-globin underwent rapid duplication and deletion in mammals (Hoffmann, Opazo et al. 2008). Based on an analysis of the platypus genome, which incorporated information from flanking loci, a recent model (Patel, Cooper et al. 2008) proposes that the β-globin paralogs arose from a single transposition in the amniote ancestor followed by independent duplication in birds and mammals. From our dataset, the β-globin paralogs β^H, ρ and ε appear as a chicken-specific expansion, consistent with this model. Rate diversity is high in this family as well and sequence evolution (point mutations) in β-globin paralogs is similar to that seen in the Toll-like receptors family (greater than 50.8% of sequence evolution in other chicken gene families).

Olfactory Receptors

Olfaction has recently gained much recognition as an important sensory modality for birds (Nevitt, Losekoot et al. 2008; O'Dwyer, Ackerman et al. 2008; Steiger, Fidler et al. 2008); Historically, birds were assumed to communicate primarily via the visual or auditory systems, but behavioral and genomic data suggest that chemosensory perception plays a larger role. The chicken genome paper remarked on the surprisingly large group of chicken-specific olfactory receptors: 218 genes were identified, representing an avian expansion orthologous to the human receptors *OR5U1* and *OR5BF1*. Our bioinformatic approach detected 196 genes belonging to this subfamily, the discrepancy is perhaps in part due to different genome builds. Similar to recent work (Lagerström, Hellström et al. 2006), we also identified other families of olfactory receptors with small expansions in chicken (two to three genes in our analysis). These include a cluster associated with the previously identified *COR1-6* genes (Chicken Olfactory Receptor genes) on chromosome 5, a second cluster on chromosome 10, related to *COR7*, and a third cluster on chromosome 1. Interestingly, the subfamilies have differing amount of sequence evolution ranging from 26.7% (family containing *COR7)* to 88.4% (family containing *COR1-6).* The latter gene family also had a more clock-like rate than other families, indicating that it is among the oldest gene duplications (238 mya) in chicken and, like the ovalbumin clade B serpins (MENT, serpinb10) and the large olfactory receptor family, arose in non-avian reptilian ancestors (see Table 1 and Figure 5).

Keratin

The evolution of feathers in theropod dinosaurs was a major innovation that likely provided insulation for metabolically active animals, ornamentation for display, and in one lineage transformed arms into wings (Ji, Currie et al. 1998; Zhang and Zhou 2000; Currie and Chen 2001; Norell, Ji et al. 2002; Sawyer and Knapp 2003). β-keratins are the basic structural elements of feathers and therefore a gene family vital for the success of birds. In the publication of the first genome draft, the International Chicken Genome Sequencing Consortium (2004)

noted the large expansion of the avian-specific keratin gene family, estimated at around 150 members. This avian keratin family, which encodes proteins forming feathers and scales (Sawyer, Glenn et al. 2000), is functionally and evolutionarily distinct from the mammalian hair-specific alpha- keratin, which, like another main component of hair, the keratin-associated proteins (KRTAP) have no members in the chicken genome (Wu, Irwin et al. 2008). Within non-avian reptiles, β-keratins likely duplicated by retrotransposition, resulting in the loss of introns in some paralogs; in birds all paralogs have lost introns. Unequal crossover is also thought to expand and contract keratin gene arrays in birds, resulting in a tandem organization of multiple paralogs (Toni, Dalla Valle et al. 2007). We find that the amount of sequence evolution in keratin is very high (98% larger than other chicken gene families), likely owing to the absence of other reptilian comparisons, but also to the adaptive significance of these proteins within birds.

Prospectus and Conclusion

Reptilia, including birds and non-avian reptiles, is the sister group of mammals, and as such holds an important phylogenetic position for shedding light on patterns of gene duplication in amniotes. Reptiles are arguably more diverse than mammals in many traits; with ~17,000 species (~10,000 in birds and 7,000 non-avian reptiles) they are substantially more species-rich than mammals (~5,000 species), and possess a greater diversity of sex chromosome and sex determination systems (Organ and Janes 2008). The chicken genome is currently the sole member of Reptilia with a draft genome, and as such provides the only point of comparison of genome dynamics between mammals and their sister group. A greater understanding of genome and multigene family dynamics in mammals will undoubtedly require greater genome sampling and characterization in Reptilia.

Gene duplication and the families they produce are vital for generating the thread with which evolution weaves new adaptations and species. We have developed a pipeline for phylogenomic analysis of gene duplication in the chicken lineage, but our approach can be applied easily to any particular clade of interest. Our approach rests on the assumption that gene orthology and paralogy are best identified through phylogenetic analysis, and we delimit chicken-specific gene duplications by an approach (Figure 1) that combines initial identification and collection of gene copies across many vertebrates that show significant sequence similarity in Ensembl, followed by phylogenetic analysis of these gene sets; identification of particular nodes in these gene trees that correspond to gene duplications, in our case the mammal-bird divergence; identification of those gene clusters that diversify from these particular nodes; and statistical analysis of the collected gene trees. Many of the duplications we have identified here as “chicken-specific” in fact will be found to have duplicated in ancestors of the chicken, since orthologs of many chicken genes will no doubt be discovered in other reptile genomes as they emerge. Nonetheless, using an approximate time scale (Figure 5) we can predict which chicken gene duplicates might be found in upcoming reptilian genome projects based on their estimated timing of duplication relative to the divergence times of species whose genomes are being compared. Our approach has the advantage of providing an objective means of identifying chicken-specific gene duplications, but of course when conducted on a genome-wide scale, it will miss some gene family members that manual curation will identify; we have illustrated this with some specific examples. The loss of detail for some gene families is offset by the ability to study genome-wide distributions of multigene family dynamics; both approaches are required to provide an informed view of the dynamics of multigene family evolution in birds and relatives.

Phylogenomic approaches such as those presented here have only just begun to provide a window into the dynamics and importance of gene duplication within organisms. For example, non-protein coding RNA paralogs are dispersed throughout the chicken genome; this, along with an unusual paucity of non-protein coding RNA pseudogenes, suggests that they may not undergo the same processes of duplication (unequal crossover and retrotransposition) that characterize protein coding genes (Hillier, Miller et al. 2004). Currently available data are insufficient to address this and other hypotheses because as of the time of this writing the genome of only one reptile species has been sequenced. But progress is quickly being made with the imminent release of the zebra finch (*Taeniopygia guttata*) and anole lizard (*Anolis carolinensis*) genomes. An increase in the number of genomes will permit more detailed quantitative comparison of the evolutionary dynamics of gene duplication in amniotes and other lineages, and will help clarify the role of these gene duplications in organismal diversification.

References

Benarafa, C. and E. Remold-O'Donnell (2005). "The ovalbumin serpins revisited: Perspective from the chicken genome of clade B serpin evolution in vertebrates." Proceedings of the National Academy of Sciences of the United States of America 102(32): 11367-11372.

Benton, M. J. and P. C. J. Donoghue (2007). "Paleontological evidence to date the tree of life." Molecular Biology and Evolution 24(1): 26–53.

Currie, P. J. and P.-J. Chen (2001). "Anatomy of Sinosauropteryx prima from Liaoning, northeastern China." Canadian Journal of Earth Sciences 38(12): 1705-1727.

de Ricqlès, A. J., K. Padian, et al. (2001). The bone histology of basal birds in phylogenetic and ontogenetic perspectives. New Perspectives on the Origin and Early Evolution of Birds: Proceedings of the International Symposium in Honor of John H. Ostrom. J. Gauthier and L. F. Gall. New Haven, CT, Peabody Museum of Natural History: 411-426.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Research 32: 1792-1797.

Eirin-Lopez, J. M., A. M. Gonzalez-Tizon, et al. (2004). "Birth-and-death evolution with strong purifying selection in the histone H1 multigene family and the origin of orphon H1 genes." Molecular Biology and Evolution 21(10): 1992-2003.

Eisen, J. A. (1998). "Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis." Genome Research 8: 163-167.

Furlong, R. F. (2005). "Insights into vertebrate evolution from the chicken genome sequence." Genome Biology 6(2).

Gregory, T. R. (2002). "A bird's-eye view of the C-value enigma: genome size, cell size, and metabolic rate in the class Aves." Evolution 56(1): 121-130.

Gribaldo, S., D. Casane, et al. (2003). "Functional divergence prediction from evolutionary analysis: A case study of vertebrate hemoglobin." Molecular Biology and Evolution 20(11): 1754-1759.

Gu, Z., D. Nicolae, et al. (2002). "Rapid divergence in expression between duplicate genes inferred from microarray data." Trends in Genetics 18(12): 609-613.

Guindon, S. and O. Gascuel (2003). "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood." Systematic Biology 52(5): 696-704.

Haasa, N. B., J. M. Grabowskia, et al. (2001). "Subfamilies of CR1 non-LTR retrotransposons have different 50 UTR sequences but are otherwise conserved." Gene 265: 175-183.

Heger, A. and C. P. Ponting (2007). "Evolutionary rate analyses of orthologs and paralogs from 12 *Drosophila* genomes." Genome Research 17(12): 1837-1849.

Hess, C. M. and S. V. Edwards (2002). "The evolution of the major histocompatibility complex in birds." Bioscience 52(5): 423-431.

Hillier, L. W., W. Miller, et al. (2004). "Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution." Nature 432(7018): 695-716.
Hittinger, C. T. and S. B. Carroll (2007). "Gene duplication and the adaptive evolution of a classic genetic switch." Nature 449: 677-681.
Hoffmann, F. G., J. C. Opazo, et al. (2008). "Rapid rates of lineage-specific gene duplication and deletion in the alpha-globin gene family." Molecular Biology and Evolution 25(3): 591-602.
Horner, J. R., K. Padian, et al. (2001). "Comparative osteology of some embryonic and perinatal archosaurs: developmental and behavioral implications for dinosaurs." Paleobiology 27(1): 39-58.
Hubbard, T. J. P., B. L. Aken, et al. (2007). "Ensembl 2007." Nucleic Acids Research 35(Database issue): D610-D617.
Huerta-Cepas, J., H. Dopazo, et al. (2007). "The human phylome." Genome Biology 8: R109.
Hughes, T. and D. A. Liberles (2008). "The power-law distribution of gene family size is driven by the pseudogenisation rate's heterogeneity between gene families." Gene 414(1-2): 85-94.
Huynen, M. A. and E. van Nimwegen (1998). "The frequency distribution of gene family sizes in complete genomes." Molecular Biology and Evolution 15(5): 583-589.
Ji, Q., P. J. Currie, et al. (1998). "Two feathered dinosaurs from northeastern China." Nature 393: 753-761.
Kondrashov, F. A., I. B. Rogozin, et al. (2002). "Selection in the evolution of gene duplications." Genome Biology 3(2): 1-9.
Lagerström, M. C., A. R. Hellström, et al. (2006). "The G protein - Coupled receptor subset of the chicken genome." Plos Computational Biology 2(6): 493-507.
Li, H., A. Coghlan, et al. (2006). "TreeFam: a curated database of phylogenetic trees of animal gene families." Nucleic Acids Research 34: D572-D580.
Li, W.-H. (2006). Molecular evolution. Sunderland, MA, Sinauer Associates.
Lynch, M. and J. S. Conery (2000). "The evolutionary fate and consequences of duplicate genes." Science 290: 1151-1155.
Lynch, M. and A. Force (2000). "The probability of duplicate-gene preservation by subfunctionalization." Genetics 154: 459-473.
Maniatis, T. and B. Tasic (2002). "Alternative pre-mRNA splicing and proteome expansion in metazoans." Nature 418(6894): 236-243.
Nevitt, G. A., M. Losekoot, et al. (2008). "Evidence for olfactory search in wandering albatross, *Diomedea exulans*." Proceedings of the National Academy of Sciences 105(12): 4576-4581.
Norell, M. A., Q. Ji, et al. (2002). "'Modern' feathers on a non-avian dinosaur." Nature 416: 36-37.
Nowak, M. A., M. C. Boerlijst, et al. (1997). "Evolution of genetic redundancy." Nature 388(6638): 167-171.
O'Dwyer, T. W., A. L. Ackerman, et al. (2008). "Examining the development of individual recognition in a burrow-nesting procellariiform, the Leach's storm-petrel." Journal of Experimental Biology 211(3): 337-340.
Ohno, S. (1970). Evolution by gene duplication. Heidelberg, Germany, Springer-Verlag.
Organ, C. L. and D. E. Janes (2008). "Evolution of sex chromosomes in Sauropsida." Integrative and Comparative Biology 48(4): 512-519.
Ota, T. and M. Nei (1994). "Divergent Evolution and Evolution by the Birth-and-Death Process in the Immunoglobulin V-H Gene Family." Molecular Biology and Evolution 11(3): 469-482.
Patel, V. S., S. J. Cooper, et al. (2008). "Platypus globin genes and flanking loci suggest a new insertional model for beta-globin evolution in birds and mammals." BMC Biology 6(34).
Piontkivska, H. and M. Nei (2003). "Birth-and-death evolution in primate MHC class I genes: Divergence time estimates." Molecular Biology and Evolution 20(4): 601-609.
Piontkivska, H., A. P. Rooney, et al. (2002). "Purifying selection and birth-and-death evolution in the histone H4 gene family." Molecular Biology and Evolution 19(5): 689-697.

Rasmussen, M. D. and M. Kellis (2007). "Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes." Genome Research 17(12): 1932-1942.
Sanderson, M. J. (2003). "r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock." Bioinformatics 19(2): 301-302.
Sarrias, M. R., J. Grønlund, et al. (2004). "The Scavenger Receptor Cysteine-Rich (SRCR) domain: an ancient and highly conserved protein module of the innate immune system." Critical Reviews in Immunology 24(1): 1-37.
Sawyer, R. H., T. Glenn, et al. (2000). "The expression of beta (β) keratins in the epidermal appendages of reptiles and birds." American Zoologist 40(4): 530-539.
Sawyer, R. H. and L. W. Knapp (2003). "Avian skin development and the evolutionary origin of feathers." Journal of Experimental Zoology Part B-Molecular and Developmental Evolution 298B(1): 57-72.
Shedlock, A. M. (2006). "Phylogenomic investigation of CR1 LINE diversity in reptiles." Systematic Biology 55(6): 902-911.
Steiger, S. S., A. E. Fidler, et al. (2008). "Avian olfactory receptor gene repertoires: evidence for a well-developed sense of smell in birds? ." Proceedings of the Royal Society (B) 275(1649): 2309-2317.
Storm, C. E. V. and E. L. L. Sonnhammer (2003). "Comprehensive analysis of orthologous protein domains using the HOPS database." Genome Research 13: 2353-2362.
Temperley, N. D., S. Berlin, et al. (2008). "Evolution of the chicken Toll-like receptor gene family: A story of gene gain and gene loss." BMC Genomics 9(62).
Toni, M., L. Dalla Valle, et al. (2007). "Hard (Beta-) keratins in the epidermis of reptiles: composition, sequence, and molecular organization " Journal of Proteome Research 6(9): 3377 -3392.
Wapinski, I., A. Pfeffer, et al. (2007). "Natural history and evolutionary principles of gene duplication in fungi." Nature 449: 54-61.
Wu, D. D., D. M. Irwin, et al. (2008). "Molecular evolution of the keratin associated protein gene family in mammals, role in the evolution of mammalian hair." BMC Evolutionary Biology 8(241).
Yuri, T., R. T. Kimball, et al. (2008). "Duplication of accelerated evolution and growth hormone gene in Passerine birds." Molecular Biology and Evolution 25(2): 352-361.
Zelano, B. and S. V. Edwards (2002). "An Mhc component to kin recognition and mate choice in birds: predictions, progress, and prospects " American Naturalist 160: S225 -S237.
Zhang, F. and Z. Zhou (2000). "A primitive enantiornithine bird and the origin of feathers." Science 290(5498): 1955-1959.
Zhang, J. (2003). "Evolution by gene duplication: an update." Trends in Genetics 18(6): 292-298.
Zhang, J., H. F. Rosenburg, et al. (1998). "Positive Darwinian selection after gene duplication in primate ribonuclease genes." Proceedings of the National Academy of Sciences 98: 3708–3713.
Zmasek, C. M. and S. R. Eddy (2002). "RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs." BMC Bioinformatics 3: 14.

Tables

Table 1: Summary of properties discussed in the text for specific gene families in the amniote-bird lineage. Number of amniote paralogs is defined as the number of genes across all species in the amniote tree (25 mammals and chicken; see Ensembl for exact species). Mean branch length is the average path length (in substitutions/site) between chicken paralogs (tips) and their common ancestor (the first chicken duplication). Across all families this length is on average 0.235 substitutions/site. Estimated duplication time is the time of first chicken duplication in millions of years. Sequence divergence before the duplication in chicken is given in substitutions/site between the root of the amniote tree (*Gallus/Homo* common ancestor) and the first chicken duplication. λ is a molecular rate variation parameter (low values are highly variable rates; high values are clock-like).

Family name	Number of amniote paralogs	Number of chicken paralogs	Mean branch length from tips to duplication in chicken	Estimated duplication time (mya)	Sequence divergence before chicken duplication after amniote divergence	Log λ
Toll-like-receptors (*TLR2A* and *TLR2B*)	18	2	0.075	67	0.260	-2
Ovalbumin B serpins (gene *X*, gene *Y*, and *ovalbumin*)	73	3	0.141	107	0.227	-2
Ovalbumin B serpins (*MENT*, *serpinb10*)	24	2	0.183	203	0.094	-2
Hemoglobin β chain (β^H,ρ, ε)	69	3	0.075	122	0.086	-2
Olfactory receptors (orthologous to *OR5U1* and *OR5BF1* in *Homo*)	511	196	0.205	281	0.320	-2*
Olfactory receptors (related to *COR7*)	12	3	0.014	15	0.394	-2
Olfactory receptors (related to *COR 1-6*)	60	4	0.275	238	0.084	2
Olfactory receptors (small cluster on chromosome 1)	207	3	0.049	30	0.421	-2
β Keratin	117	117	0.635	NA	NA	NA

* This λ is calculated from *Homo* and *Gallus* for computational limitations due to large family size.

NA = these statistics cannot be computed because of chicken-only expansions.

Figures

Figure 1: Diagram showing the phylogenetic pattern that gene duplication takes in multi-species comparisons.

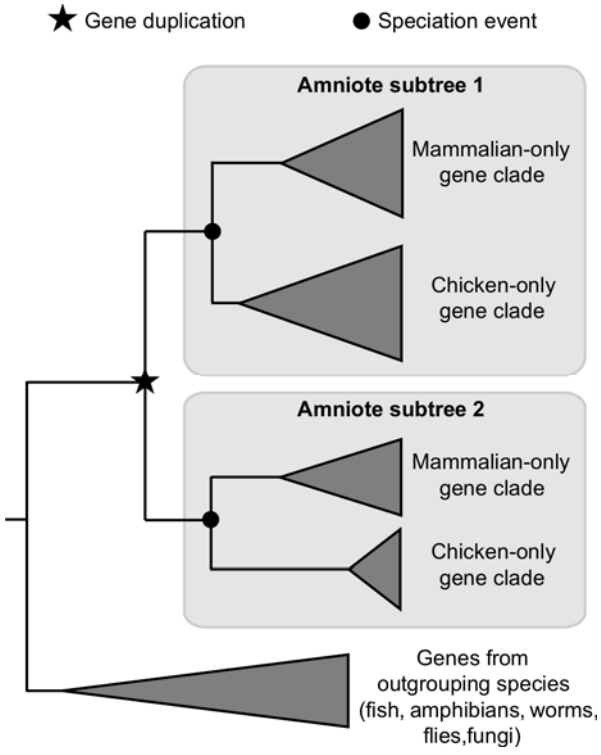


Figure 2: The β-keratin gene family tree as defined by single-linkage clustering of best reciprocal BLAST hits between amino acid translations within the Ensembl database. This is an example of chicken gene family tree used in various analyses throughout this chapter. The tip names are designated as Ensembl proteins.

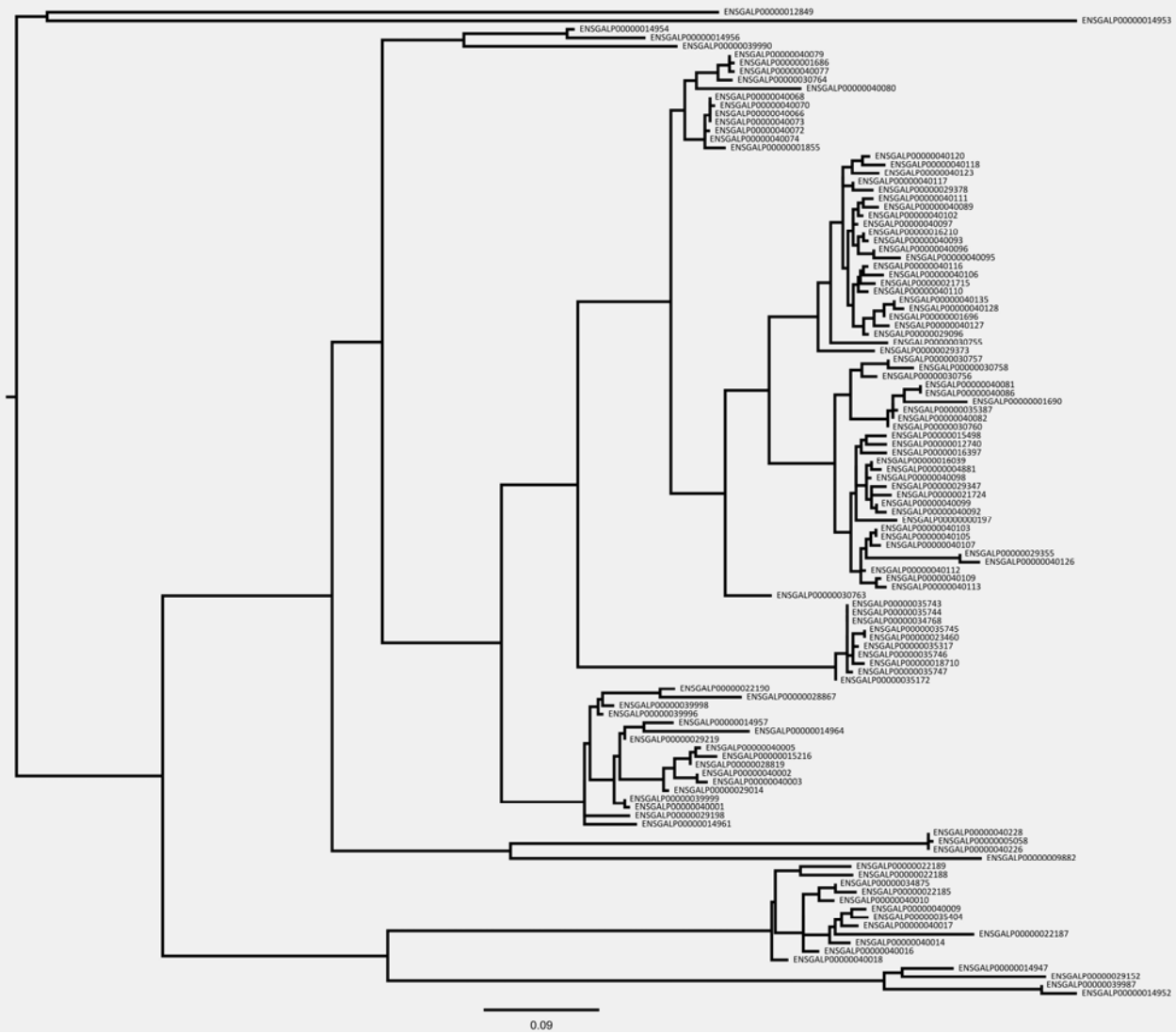


Figure 3: Count of gene families vs. the number of chicken genes in amniote subtrees.

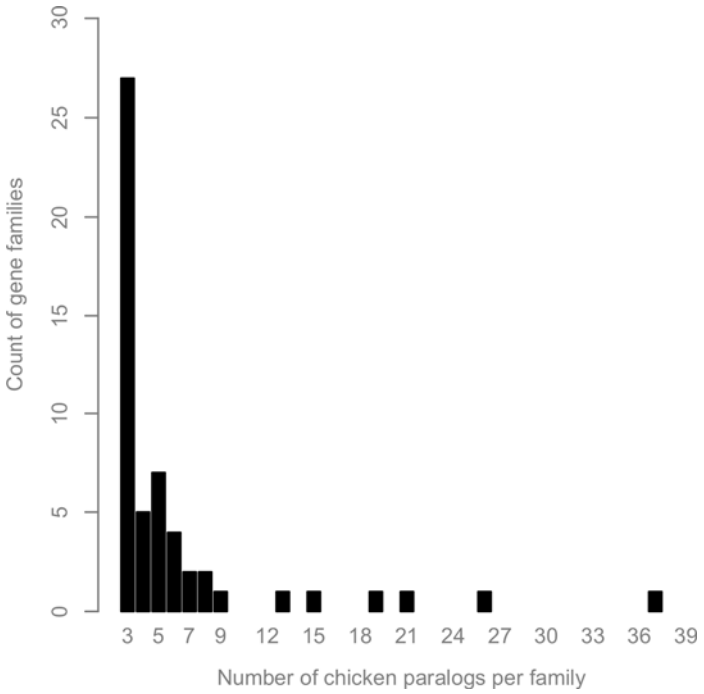


Figure 4: Branch length rate variation (molecular clock; λ) for amniote gene trees (25 mammals and chicken). **A**, the distribution of branch length rate variation binned in different λ values. Low values of λ represent highly variable rates and high values of λ represent clock-like rates. **B**, the best fit (by likelihood criterion) λ by tree size. r^2 relating tree size with log λ is equal to 0.0002843 (p-value = 0.08, n = 6,901). Values of λ were dithered for display purposes to illustrate density.

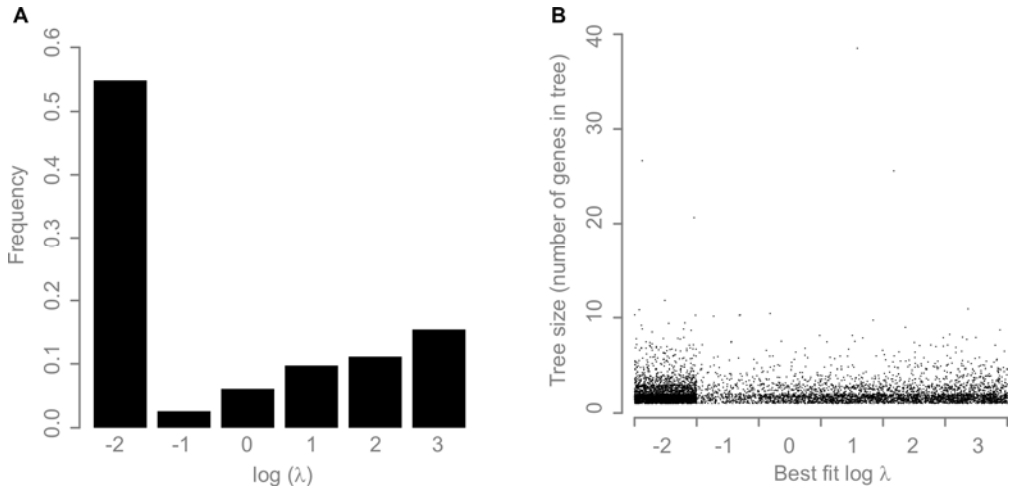
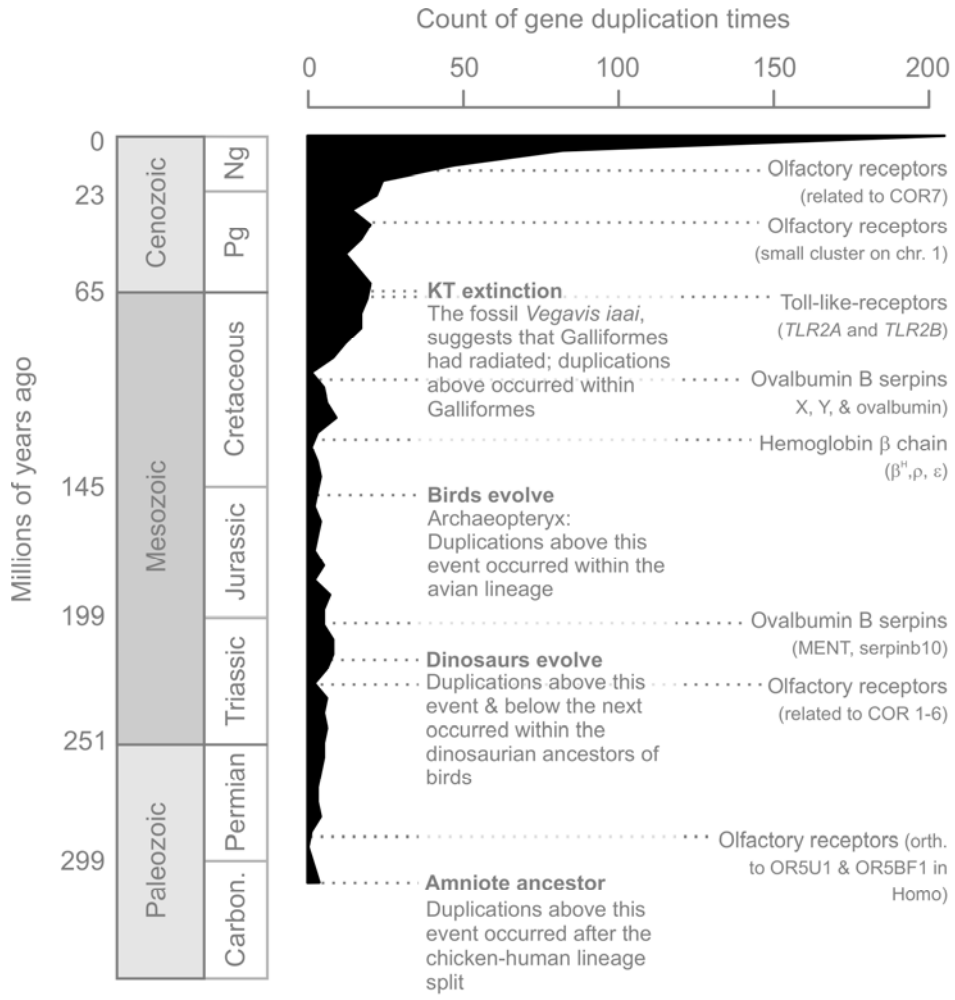


Figure 5: Depth of paralogs on the lineage leading to *Gallus gallus* that evolved after the divergence between *Gallus* and *Homo*. Pivotal events during the evolution of this lineage are noted on the figure. Pg stands for Paleogene and Ng stands for Neogene.



A New Voronoi-Based Surface Reconstruction Algorithm

Nina Amenta* Marshall Bern Manolis Kamvysselis†
UT - Austin Xerox PARC M.I.T.

Abstract

We describe our experience with a new algorithm for the reconstruction of surfaces from unorganized sample points in $\mathbb{R}^3$. The algorithm is the first for this problem with provable guarantees. Given a "good sample" from a smooth surface, the output is guaranteed to be topologically correct and convergent to the original surface as the sampling density increases. The definition of a good sample is itself interesting: the required sampling density varies locally, rigorously capturing the intuitive notion that featureless areas can be reconstructed from fewer samples. The output mesh interpolates, rather than approximates, the input points.

Our algorithm is based on the three-dimensional Voronoi diagram. Given a good program for this fundamental subroutine, the algorithm is quite easy to implement.

Keywords: Medial axis, Sampling, Delaunay triangulation, Computational Geometry

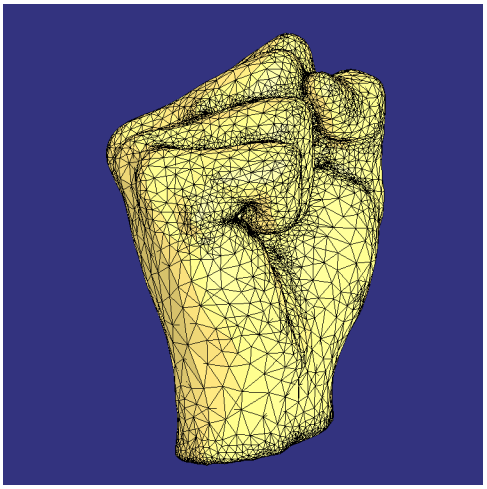


Figure 1. The fist mesh was reconstructed from the vertices alone. Notice that the sampling density varies. Our algorithm requires dense sampling only near small features; given such an input, the output mesh is provably correct.

1 Introduction

The process of turning a set of sample points in $\mathbb{R}^3$ into a computer graphics model generally involves several steps: the reconstruction of an initial piecewise-linear model, cleanup, simplification, and perhaps fitting with curved surface patches.

We focus on the first step, and in particular on an abstract problem defined by Hoppe, DeRose, Duchamp, McDonald, and Stuetzle [14]. In this formulation, the input is a set of points in $\mathbb{R}^3$, without any additional structure or organization, and the desired output is a polygonal mesh, possibly with boundary. In practice, sample sets for surface reconstruction come from a variety of sources: medical imagery, laser range scanners, contact probe digitizers, radar and seismic surveys, and mathematical models such as implicit surfaces. While the most effective reconstruction scheme for any one of these applications should take advantage of the special properties of the data, an understanding of the abstract problem should contribute to all of them.

The problem formulation above is incomplete, since presumably we should require some relationship between the input and the output. In this and a companion paper [2], we describe a simple, combinatorial algorithm for which we can prove such a relationship. A nontrivial part of this work is the fitting of precise definitions to the intuitive notions of a "good sample" and a "correct reconstruction". Although the actual definition of a good sample is rather technical, involving the medial axis of the original surface, Figure 1 gives the general idea: dense in detailed areas and (possibly) sparse in featureless ones.

The algorithm is based on the three-dimensional Voronoi diagram and Delaunay triangulation; it produces a set of triangles that we call the *crust* of the sample points. All vertices of crust triangles are sample points; in fact, all crust triangles appear in the Delaunay triangulation of the sample points.

The companion paper [2] presents our theoretical results. In that paper, we prove that given a good sample from a smooth surface, the output of our reconstruction algorithm is topologically equivalent to the surface, and that as the sampling density increases, the output converges to the surface, both pointwise and in surface normal.

Theoretical guarantees, however, do not imply that an algorithm is useful in practice. Surfaces are not everywhere smooth, samples do not everywhere meet the sampling density conditions, and sample points contain noise. Even on good inputs, an algorithm may fail to be robust, and the constants on the running time might be prohibitively large. In this paper, we report on our implementation of the algorithm, its efficiency and the quality of the output.

Overall, we were pleased. The program gave intuitively reasonable outputs on inputs for which the theoretical results do not apply. The implementation, using a freely available exact-arithmetic Voronoi diagram code, was quite easy, and reasonably efficient: it can handle 10,000 points in a matter of minutes. The main difficulty, both in theory and in practice, is the reconstruction of sharp edges.

*Much of this work was done while the author was employed by Xerox PARC, partially supported by NSF grant CCR-9404113.

†Much of this work was done while the author was an intern at Xerox PARC.

2 Related work

The idea of using Voronoi diagrams and Delaunay triangulations in surface reconstruction is not new. The well–known α-*shape* of Edelsbrunner et al. [9, 10] is a parameterized construction that associates a polyhedral shape with an unorganized set of points. A simplex (edge, triangle, or tetrahedron) is included in the α-shape if it has some circumsphere with interior empty of sample points, of radius at most α (a circumsphere of a simplex has the vertices of the simplex on its boundary). The *spectrum* of α-shapes, that is, the α-shapes for all possible values of α, gives an idea of the overall shape and natural dimensionality of the point set. Edelsbrunner and Mücke experimented with using α-shapes for surface reconstruction [10], and Bajaj, Bernardini, and Xu [4] have recently used α-shapes as a first step in the entire reconstruction pipeline.

An early Delaunay-based algorithm, similar in spirit to our own, is the "Delaunay sculpting" heuristic of Boissonnat [6], which progressively eliminates tetrahedra from the Delaunay triangulation based on their circumspheres. In two dimensions, there are a number of recent theoretical results on various Delaunay-based approaches to reconstructing smooth curves. Attali [3], Bernardini and Bajaj [5], Figueiredo and Miranda Gomes [11] and ourselves [1] have all given guarantees for different algorithms.

A fundamentally different approach to reconstruction is to use the input points to define a signed distance function on $\mathbb{R}^3$, and then polygonalize its zero-set to create the output mesh. Such *zero-set* algorithms produce approximating, rather than interpolating, meshes. This approach was taken by Hoppe et al. [14, 13] and more recently by Curless and Levoy [8]. Hoppe et al. determine an approximate tangent plane at each sample point using least squares on k nearest neighbors, and then take the signed distance to the nearest point's tangent plane as the distance function on $\mathbb{R}^3$. The distance function is then interpolated and polygonalized by the marching cubes algorithm. The algorithm of Curless and Levoy is tuned for laser range data, from which they derive error and tangent plane information. They combine the samples into a continuous volumetric function, computed and stored on a voxel grid. A subsequent hole-filling step also uses problem-specific information. Their implementation is especially fast and robust, capable of handling very large data sets.

Functionally our crust algorithm differs from both the α-shape and the zero-set algorithms. It overcomes the main drawback of α-shapes as applied to surface reconstruction, which is that the parameter α must be chosen experimentally, and in many cases there is no ideal value of α due to variations in the sampling density. The crust algorithm requires no such parameter; it in effect automatically computes the parameter locally. Allowing the sampling density to vary locally enables detailed reconstructions from much smaller input sets.

Like the α-shape, the crust can be considered an intrinsic construction on the point set. But unlike the α-shape, the crust is naturally two-dimensional. This property makes the crust more suitable for surface reconstruction, although less suitable for determining the natural dimensionality of a point set.

The crust algorithm is simpler and more direct than the zero-set approach. Zero-set algorithms, which produce approximating rather than interpolating surfaces, inherently do some low-pass filtering of the data. This is desirable in the presence of noise, but causes some loss of information. We believe that some of our ideas, particularly the sampling criterion and the normal estimation method, can be applied to zero-set algorithms as well, and might be useful in proving some zero-set algorithm correct.

With its explicit sampling criterion, our algorithm should be most useful in applications in which the sampling density is easy to control. Two examples are digitizing an object with a hand-held contact probe, where the operator can "eyeball" the required density, and polygonalizing an implicit surface using sample points [12], where the distribution can be controlled analytically.

3 Sampling Criterion

Our theoretical results assume a *smooth surface*, by which we mean a twice-differentiable manifold embedded in $\mathbb{R}^d$. Notice that this allows all orientable manifolds, including those with multiple connected components.

3.1 Geometry

We start by reviewing some standard geometric constructions. Given a discrete set S of sample points in $\mathbb{R}^d$, the *Voronoi cell* of a sample point is that part of $\mathbb{R}^d$ closer to it than to any other sample. The *Voronoi diagram* is the decomposition of $\mathbb{R}^d$ induced by the Voronoi cells. Each Voronoi cell is a convex polytope, and its vertices are the *Voronoi vertices*; when S is nondegenerate, each Voronoi vertex is equidistant from exactly $d+1$ points of S. These $d+1$ points are the vertices of the *Delaunay simplex*, dual to the Voronoi vertex. A Delaunay simplex, and hence each of its faces, has a circumsphere empty of other points of S. The set of Delaunay simplices form the *Delaunay triangulation* of S. Computing the Delaunay triangulation essentially computes the Voronoi diagram as well. See Figure 5 for two-dimensional examples.

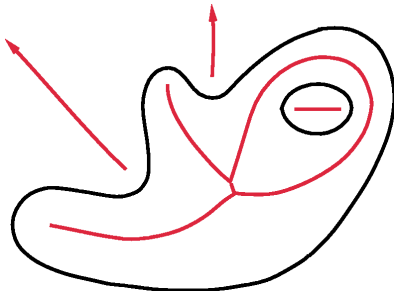


Figure 2. The red curves are the medial axis of the black curves. Notice that components of the medial axis lie on either side of the black curves.

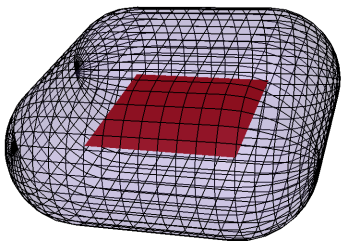


Figure 3. In three dimensions, the medial axis of a surface is generally a two-dimensional surface. Here, the square is the medial axis of the rounded transparent surface. A nonconvex surface would have components of the medial axis on the outside as well, as in the 2D example of Figure 2.

The *medial axis* of a $(d-1)$-dimensional surface in $\mathbb{R}^d$ is (the closure of) the set of points with more than one closest point on the surface. An example in $\mathbb{R}^2$ is shown in Figure 2, and in $\mathbb{R}^3$ in Figure 3. This definition of the medial axis includes components on the exterior of a closed surface. The medial axis is the extension to continuous surfaces of the Voronoi diagram, in the sense that the

Voronoi diagram of S can be defined as the set of points with more than one closest point in S.

In two dimensions, the Voronoi vertices of a dense set of sample points on a curve approximate the medial axis of the curve. Somewhat surprisingly—a number of authors have been misled—this nice property does not extend to three dimensions.

3.2 Definition

We can now describe our sampling criterion. A good sample is one in which the sampling density is (at least) inversely proportional to the distance to the medial axis. Specifically, a sample S is an *r-sample* from a surface F when the Euclidean distance from any point $p \in F$ to the nearest sample point is at most r times the distance from p to the nearest point of the medial axis of F.

The constant of proportionality r is generally less than one. In the companion paper [2], we prove our theorems for small values of r such as $r \leq .06$, but the bounds are not tight. Hence the theoretical results apply only when the sampling is very dense.

We observe that in practice $r = .5$ generally suffices. Figure 4 shows a reconstruction from a dense sample, and from a sample thinned to roughly $r = .5$. We did not compute the medial axis, which can be quite a chore. Instead, we used the distance to the nearest "pole" (see Section 4.2) as a reasonable, and easily computed, estimate of the distance to the medial axis.

Figure 4. The sampling spacing required to correctly reconstruct a surface is proportional to the distance to the medial axis. On the left is a surface reconstructed from a dense sample. The color represents estimated distance to medial axis—red means close. On the right, we use the estimated distance to thin the data to a .5-sample (meaning that the distance to the nearest sample for any point on the surface is at most half the distance to the medial axis), and then reconstruct. There were about 12K samples on the left and about 3K on the right.

Notice that our sampling criterion places no constraints on the distribution of points, so long as they are sufficiently dense. It inherently takes into account both the curvature of the surface—the medial axis is close to the surface where the curvature is high—and also the proximity of other parts of the surface. For instance, although the middle of a thin plate has low curvature, it must be sampled densely to resolve the two sides as separate surfaces. In this situation an r-sample differs from the distribution of vertices typically produced by mesh simplification algorithms, which only need to consider curvature.

At sharp edges and corners, the medial axis actually touches the surface. Accordingly, our criterion requires infinitely dense sampling to guarantee reconstruction. Sharp edges are indeed a problem in practice as well, although the reconstruction errors are not noticeable when the sampling is very dense. We discuss a heuristic approach to resolving sharp edges in Section 6, and propose a stronger theoretical approach in Section 7.

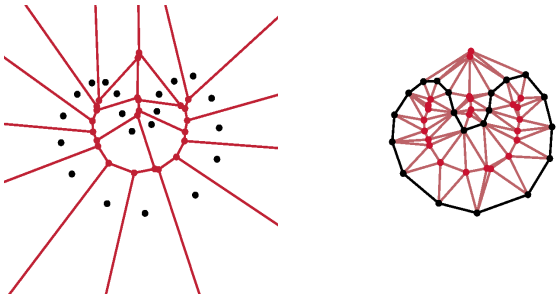


Figure 5. The two-dimensional algorithm. On the left, the Voronoi diagram of a point set S sampled from a curve. Just as S approximates the curve, the Voronoi vertices V approximate the medial axis of the curve. On the right, the the Delaunay triangulation of $S \cup V$, with the crust edges in black. Theorem 1 states that when S is an r-sample, for sufficiently small r, the crust edges connect only adjacent vertices.

4 The crust algorithm

4.1 Two Dimensions

We begin with a two-dimensional version of the algorithm [1]. In this case, the crust will be a graph on the set of sample points S. We define the crust as follows: an edge e belongs to the crust if e has a circumcircle empty not only of all other sample points but also of all Voronoi vertices of S. The crust obeys the following theorem [1].

Theorem 1. *The crust of an r-sample from a smooth curve F, for $r \leq .25$, connects only adjacent sample points on F.*

The medial axis provides the intuition behind this theorem. An important lemma is that for any sample S, an edge between two nonadjacent sample points cannot be circumscribed by a circle that misses both the medial axis and all other samples. When S is an r-sample for sufficiently small r, the Voronoi vertices approximate the medial axis, and any circumcircle of an edge between nonadjacent samples contains either another sample or a Voronoi vertex. An edge between two adjacent samples, on the other hand, is circumscribed by a small circle, far away from the medial axis and hence from all Voronoi vertices.

The definition of the two-dimensional crust leads to the following simple algorithm, illustrated in Figure 5. First compute the Voronoi diagram of S, and let V be the set of Voronoi vertices. Then compute the Delaunay triangulation of $S \cup V$. The crust consists of the Delaunay edges between points of S, since those are the edges with circumcircles empty of points in $S \cup V$. Notice that the crust is also a subset of the Delaunay triangulation of the input points; adding the Voronoi vertices filters out the unwanted edges from the Delaunay triangulation. We call this technique *Voronoi filtering*.

4.2 Three Dimensions

This simple Voronoi filtering algorithm runs into a snag in three dimensions. The nice property that all the Voronoi vertices of a sufficiently dense sample lie near the medial axis is no longer true. Figure 6 shows an example. No matter how densely we sample, Voronoi vertices can appear arbitrarily close to the surface.

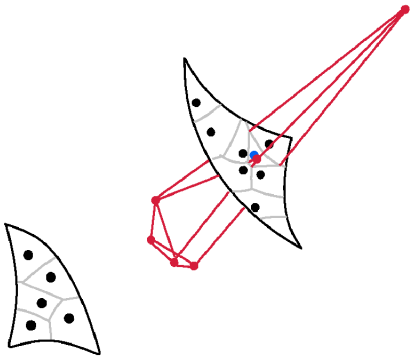


Figure 6. In three dimensions, we can use only a subset of the Voronoi vertices, since not all Voronoi vertices contribute to the approximation of the medial axis. Here, one sample on a curved surface is colored blue, and the edges of its three-dimensional Voronoi cell are drawn in red. One red Voronoi vertex lies near the surface, equidistant from the four samples near the center. The others lie near the medial axis, near the center of curvature on one side and halfway to an opposite patch of the surface on the other.

On the other hand, many of the three-dimensional Voronoi vertices *do* lie near the medial axis. Consider the Voronoi cell V_s of a sample s, as in Figure 6. The sample s is surrounded on F by other samples, and V_s is bounded by bisecting planes separating s from its neighbors, each plane nearly perpendicular to F. So the Voronoi cell V_s is long, thin and roughly perpendicular to F at s. V_s extends perpendicularly out to the medial axis. Near the medial axis, other samples on F become closer than s, and V_s is cut off. This guarantees that some vertices of V_s lie near the medial axis. We give a precise and quantitative version of this rough argument in [2].

This leads to the following algorithm. Instead of using all of the Voronoi vertices in the Voronoi filtering step, for each sample s we use only the two vertices of V_s farthest from s, one on either side of the surface F. We call these the *poles* of s, and denote them p^+ and p^-. It is easy to find one pole, say p^+: the farthest vertex of V_s from s. The observation that V_s is long and thin implies that the other pole p^- must lie roughly in the opposite direction. Thus in the basic algorithm below, we simply choose p^- to be farthest vertex from s such that sp^- and sp^+ have negative dot-product. Here is the basic algorithm:

1. *Compute the Voronoi diagram of the sample points S*
2. *For each sample point s do:*
   (a) *If s does not lie on the convex hull, let p^+ be the farthest Voronoi vertex of V_s from s. Let n^+ be the vector $sp+$.*
   (b) *If s lies on the convex hull, let n^+ be the average of the outer normals of the adjacent triangles.*
   (c) *Let p^- be the Voronoi vertex of V_s with negative projection on n^+ that is farthest from s.*
3. *Let P be the set of all poles p^+ and p^-. Compute the Delaunay triangulation of $S \cup P$.*
4. *Keep only those triangles for which all three vertices are sample points in S.*

Notice that one does not need an estimate of r to use the crust algorithm; the basic algorithm requires no tunable parameters at all. The output of this algorithm, the *three-dimensional crust*, is a set of triangles that resembles the input surface geometrically. More precisely, we prove the following theorem [2].

Theorem 2. *Let S be an r-sample from a smooth surface F, for $r \leq .06$. Then 1) the crust of S contains a set of triangles forming a mesh topologically equivalent to F, and 2) every point on the crust lies within distance $5r \cdot d(p)$ of some point p on F, where $d(p)$ is the distance from p to the medial axis.*

The crust, however, is not necessarily a manifold; for example, it often contains all four triangles of a very flat "sliver" tetrahedron. It is, however, a visually acceptable model.

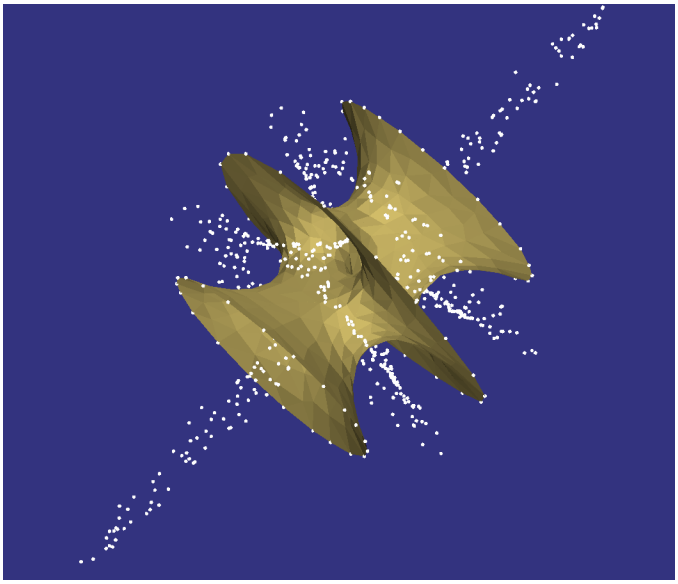


Figure 7. The crust of a set of sample points and the poles (white points) used in its reconstruction. Each sample selects the two vertices of its Voronoi cell that are farthest away, one on either side of the surface, as poles. The poles lie near the medial axis of the surface, sketching planes separating opposite sheets of surface that degenerate to one-dimensional curves where the cross-section of the surface is circular.

4.3 Normal Estimation and Filtering

Additional filtering is required to produce a guaranteed piecewise-linear manifold homeomorphic to F, and to ensure that the output converges in surface normal as the sampling density increases.

In fact, whatever the sampling density, the algorithm above may output some very thin crust triangles nearly perpendicular to the surface. We have an important lemma [2], however, which states that the vectors $n^+ = sp^+$ and $n^- = sp^-$ from a sample point to its poles are guaranteed to be nearly orthogonal to the surface at s. The angular error is linear in r. The intuition (put nicely by Ken Clarkson) is that the surface normal is easy to estimate from a point far away, such as a pole p, since the surface must be nearly normal to the largest empty ball centered at p.

We can use these vectors in an additional *normal filtering* step, throwing out any triangles whose normals differ too much from n^+ or n^-. When normal filtering is used, the normals of the output triangles approach the surface normals as the sampling density increases. We prove in [2] that the remaining set of triangles still contains a subset forming a piecewise-linear surface homeomorphic to F.

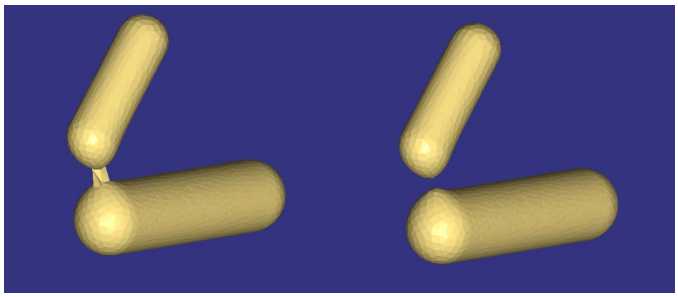


Figure 8. The crust of points distributed on an implicit surface (left). The additional normal filtering step is needed to separate the two connected components (right), which are undersampled at their closest point. Triangles are deleted if their normals differ too much from the direction vectors from the triangle vertices to their poles. These vectors are provably close to the surface normals.

Normal filtering can be useful in practice as well, as shown in Figure 8. In the usual case in which r is unknown the allowable difference in angle must be selected experimentally. Normal filtering can be dangerous, however, at boundaries and sharp edges. The directions of n^+ and n^- are not nearly normal to all nearby tangent planes, and desirable triangles might be deleted.

We note that n^+ and n^-, our Voronoi-based estimates of normal direction, could be useful in the zero-set reconstruction methods, which depend on accurate estimation of the tangent planes. For the algorithm of Hoppe et al. [14], a Voronoi-based estimate could replace the estimate based on the k-nearest neighbors. The Voronoi-based estimate has the advantage that it is not sensitive to the distribution; whereas, for instance, on medical image data, all k nearest neighbors might lie in the same slice, and so would the estimated tangent plane. In the algorithm of Curless and Levoy [8], the Voronoi-based estimate could be checked against the bounds on normal direction derived from the laser-range scanner.

4.4 Manifold Extraction

After the normal filtering step, all the remaining triangles are roughly parallel to the surface. We can define a sharp edge as one which is adjacent to triangles only on one side of a plane through the edge and roughly perpendicular to the surface. Notice that an edge of degree one counts as a sharp edge. If the surface F is indeed a smooth manifold without boundary, we are guaranteed that the normal-filtered crust contains a piecewise-linear manifold homeomorphic to F. Any triangle adjacent to a sharp edge cannot belong to this piecewise-linear manifold, and can be safely deleted. We continue recursively until no such triangle remains. A piecewise-linear manifold can then be obtained by a *manifold extraction* step which takes the outside surface of the remaining triangles on each connected component. This simple approach, however, cannot be applied when F is not a smooth manifold without boundary. In that case we do not know how to prove that we can extract a manifold homeomorphic to F.

4.5 Complexity

The asymptotic complexity of the crust algorithm is $O(n^2)$ where $n = |S|$, since that is the worst-case time required to compute a three-dimensional Delaunay triangulation. Notice that the number of sample points plus poles is at most $3n$. As has been frequently observed, the worst-case complexity for the three-dimensional Delaunay triangulation almost never arises in practice. All other steps are linear time.

5 Implementation

5.1 Numerical Issues

Robustness has traditionally been a concern when implementing combinatorial algorithms like this one. Our straightforward implementation, however, is very robust. This success is due in large part to the rapidly improving state of the art in Delaunay triangulation programs. We used Clarkson's *Hull* program. *Hull* uses exact integer arithmetic, and hence is thoroughly robust, produces exact output, and requires no arithmetic tolerancing parameters. The performance cost for the exact arithmetic is fairly modest, due to a clever adaptive precision scheme. We chose *Hull* so that we could be sure that numerical problems that arose were our own and did not originate in the triangulation. Finding the exact Delaunay triangulation is not essential to our algorithm.

Hull outputs a list of Delaunay tetrahedra, but not the coordinates of their circumcenters (the dual Voronoi vertices) which always contain some roundoff error. Fortunately, the exact positions of the poles are not important, as the numerical error is tiny relative to the distance between the poles and the surface. We computed the location of each Voronoi vertex by solving a 4×4 linear system with a solver from *LAPACK*. The solver also returns the condition number of the coefficient matrix, which we used to reject unreliable Voronoi vertices. Rejected Voronoi vertices were almost always circumcenters of "slivers" (nearly planar tetrahedra) lying flat on the surface; for a good sample such vertices cannot be poles. It is possible that this method also rejects some valid poles induced by very flat tetrahedra spanning two patches of surface. We have not, however, observed any problems in practice. Presumably there is always another Voronoi vertex nearby that makes an equally good pole.

5.2 Efficiency

Running times for the reconstruction of some large data sets are given in the table below; the reconstructions are shown in Figure 9. We used an SGI Onyx with 512M of memory.

Model	Time (min)	Num. Pts.
Femur	2	939
Golf club	12	16864
Foot	15	20021
Bunny	23	35947

The running time is dominated by the time required to compute the Delaunay triangulations. *Hull* uses an incremental algorithm [7], so the running time is sensitive to the input order of the vertices. The triangulation algorithm builds a search structure concurrently with the triangulation itself; the process is analogous to sorting by incrementally building a binary search tree. When points are added in random order, the search structure is balanced (with extremely high probability) and the expected running time is optimal. In practice, random insertions are slow on large inputs, since both the search structure and the Delaunay triangulation begin paging. We obtained better performance by first inserting a random subset of a few thousand points to provide a balanced initial search structure, and then inserting the remaining points based on a crude spatial subdivision to improve locality.

Most likely much greater improvements in efficiency can be achieved by switching to a three-dimensional Delaunay triangulation program that, first, does not use exact arithmetic, and second, uses an algorithm with more locality of reference.

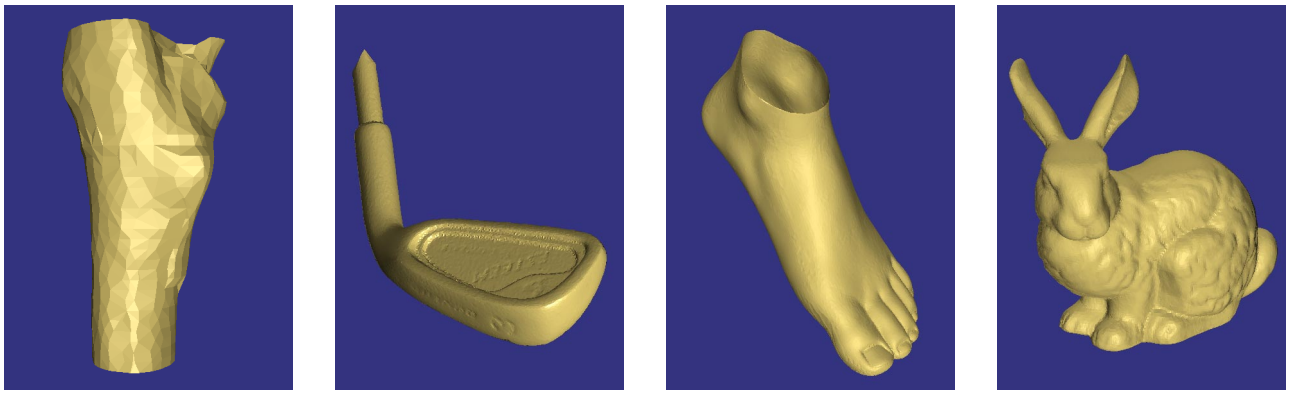


Figure 9. Femur, golf club, foot and bunny reconstructions. Notice the subtle "3" on the bottom of the club (apparently a 3-iron), showing the sensitivity of the algorithm. The foot, like all our reconstructions, is hollow. The bunny was reconstructed from the roughly 36K vertices of the densest of the Stanford bunny models in 23 minutes.

6 Heuristic Modifications

As we have noted, our algorithm does not do well at sharp edges, either in theory or in practice. The reason is that the Voronoi cell of a sample s on a sharp edge is not long and thin, so that the assumptions under which we choose the poles is not correct. For example, the Voronoi cell of a sample s on a right-angled edge is roughly fan-shaped. The vector n^+ directed towards the first pole of s might be perpendicular to one tangent plane at s, but parallel to the other. The second pole would then be chosen very near the surface, punching a hole in the output mesh.

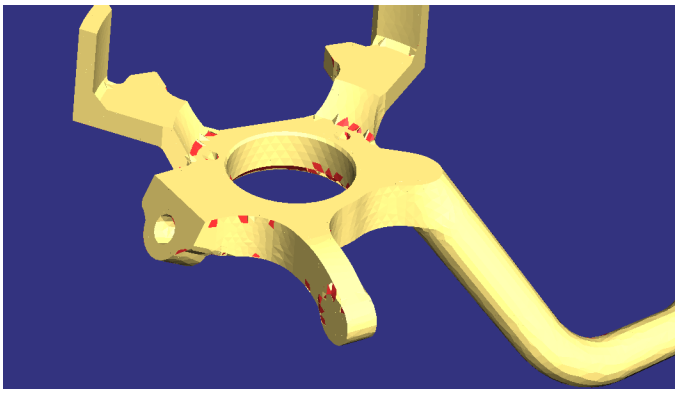


Figure 10. We resolve the sharp edges on this model of a mechanical part by using the two farthest Voronoi vertices as poles, regardless of direction. The basic algorithm forces the poles to lie in opposite directions, but is only guaranteed to work properly on a smooth surface. The red triangles do not appear in the reconstruction when using the basic algorithm.

We experimented with other methods for choosing the second pole. We found that choosing as p^- the Voronoi vertex with the greatest negative projection in the direction n^+ gave somewhat better results. This modification should retain the theoretical guarantees of the original algorithm. The best reconstructions, however, were produced by a different heuristic: choosing the farthest and the second farthest Voronoi vertices, regardless of direction, as the two poles (see Figure 10). This heuristic is strongly biased against choosing poles near the surface, avoiding gaps near sharp edges but sometimes allowing excess triangles filling in sharp corners. We believe that pathological cases could be constructed in which this fill causes a topologically incorrect reconstruction irrespective of the sampling density.

Boundaries pose similar problems in theory, but the reconstructions produced by the crust algorithm on surfaces with boundaries are usually acceptable. Figure 7 and the foot in Figure 9 are examples of perfectly reconstructed boundaries. When the boundary forms a hole in an otherwise flat surface, with no other parts of the surface nearby, the crust algorithm fills in the hole.

Undersampling also causes holes in the output mesh. For example, consider a sample in the middle of a a flat plate. Although its second pole lies in the correct direction, if there are two few sample points on the opposite side of the plate, the pole may fall near the surface on the opposite side and cause a hole. We experimented with heuristics to compensate for this undersampling effect, and for similar reconstruction errors in undersampled cylindrical regions. We found that moving all poles closer to their samples by some constant fraction allowed thin plates and cylinders to be reconstructed from fewer samples, while sometimes introducing new holes on other parts of the model. We were sometimes able to get a perfect reconstruction by taking the union of a crust made with this modification and one without.

7 Research Directions

We have identified a number of future research directions.

7.1 Noise

Small perturbations of the input points do not cause problems for the crust algorithm, nor do a few outliers. But when the noise level is roughly the same as the sampling density, the algorithm fails, both in theory and in practice. We believe, however, that there is a Voronoi-based algorithm, perhaps combining aspects of crusts and α-shapes, that reconstructs noisy data into a "thickened surface" containing all the input points, some of them possibly in the interior. See Melkemi [15] for some suggestive experimental work in $\mathbb{R}^2$.

7.2 Sharp Edges and Boundaries

We would like to modify the crust algorithm to handle surfaces with sharp edges and to provide theoretical guarantees for the reconstruction of both sharp edges and boundaries. Interpolating reconstruction algorithms like ours have an advantage here, since approximating reconstruction algorithms smooth out sharp edges. One important goal is to develop reliable techniques for identifying samples that lie on sharp edges or boundaries. As noted, the Voronoi cells of such samples are not long and thin. This intuition

could be made precise, and perhaps combined with more traditional filtering techniques.

7.3 Using Surface Normals

A variation on the problem is the reconstruction of surfaces from unorganized points that are equipped with normal directions. This problem arises in two-dimensional image processing when connecting edge pixels into edges. In three dimensions, laser range data comes with some normal information, and we have exact normals for points distributed on implicit surfaces. It should be possible to show that with this additional information, reconstruction is possible from much sparser samples. In particular, when normals are available, dense sampling should not be needed to resolve the two sides of a thin plate, suggesting that a different sampling criterion than distance to medial axis is required.

7.4 Compression

One intriguing potential application (pointed out by Frank Bossen) of interpolating, rather than approximating reconstruction, is that it can be used as a lossless mesh compression technique. A model created by interpolating reconstruction can be represented entirely by its vertices, and no connectivity information at all must be stored. A model which differs only slightly from the reconstruction of its vertices can be represented by the vertices and a short list of differences. These differences might be encoded efficiently using some geometrically defined measure of "likelihood" on Delaunay triangles. The vertices themselves could then be ordered so as to optimize properties such as compressibility or progressive reconstruction by an incremental algorithm. With the current best geometry compression method [16], most of the bits are already used to encode the vertex positions, rather than connectivity, but the connectivity is encoded in the ordering of the vertices. Allowing arbitrary vertex orderings could improve compression; we are experimenting with an octree encoding.

Our current crust algorithm is not incremental, and our implementation is too slow for real-time decompression, so this application motivates work in both directions.

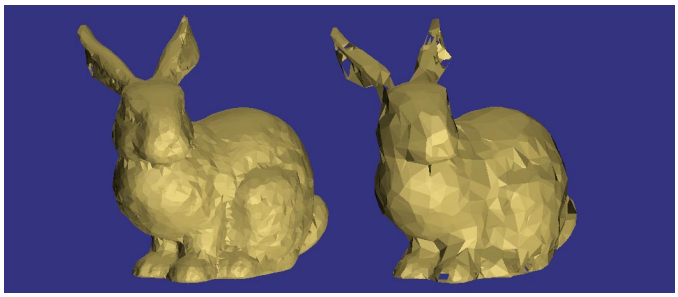


Figure 11. Reconstructions from subsets of the samples resemble the final reconstructions. The crust of the first 5 % of the points in an octree encoding of the bunny samples is still quite recognizable (right); the crust of 20 % of the points is on the left. Rough reconstructions like these could be shown during progressive transmission.

Acknowledgments

We thank David Eppstein (UC–Irvine) for his collaboration in the early stages of this research, and Frank Bossen (EPF–Lausanne) and Ken Clarkson (Lucent) for interesting suggestions. We thank Ping Fu (Raindrop Geomagic) for the fist and the mechanical part, Hughes Hoppe (Microsoft) for the head, the golf club and the foot, Chandrajit Bajaj (UT–Austin) for the femur, Paul Heckbert (CMU) for the hot dogs, and the Stanford Data Repository for the bunny. We thank Ken Clarkson and Lucent Bell Labs for *Hull*, and The Geometry Center at the University of Minnesota for *Geomview*, which we used for viewing and rendering the models.

References

[1] Nina Amenta, Marshall Bern and David Eppstein. The Crust and the β-Skeleton: Combinatorial Curve Reconstruction. To appear in *Graphical Models and Image Processing*.

[2] Nina Amenta and Marshall Bern. Surface reconstruction by Voronoi filtering. To appear in *14th ACM Symposium on Computation Geometry*, June 1998.

[3] D. Attali. r-Regular Shape Reconstruction from Unorganized Points. In *13th ACM Symposium on Computational Geometry*, pages 248–253, June 1997.

[4] C. Bajaj, F. Bernardini, and G. Xu. Automatic Reconstruction of Surfaces and Scalar Fields from 3D Scans. *SIGGRAPH '95 Proceedings*, pages 109–118, July 1995.

[5] F. Bernardini and C. Bajaj. Sampling and reconstructing manifolds using α-shapes, In *9th Canadian Conference on Computational Geometry*, pages 193–198, August 1997.

[6] J-D. Boissonnat. Geometric structures for three-dimensional shape reconstruction, *ACM Transactions on Graphics* 3: 266–286, 1984.

[7] K. Clarkson, K. Mehlhorn and R. Seidel. Four results on randomized incremental constructions. *Computational Geometry: Theory and Applications*, pages 185–121, 1993.

[8] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *SIGGRAPH '96 Proceedings*, pages 303–312, July 1996.

[9] H. Edelsbrunner, D.G. Kirkpatrick, and R. Seidel. On the shape of a set of points in the plane, *IEEE Transactions on Information Theory* 29:551-559, (1983).

[10] H. Edelsbrunner and E. P. Mücke. Three-dimensional Alpha Shapes. *ACM Transactions on Graphics* 13:43–72, 1994.

[11] L. H. de Figueiredo and J. de Miranda Gomes. Computational morphology of curves. *Visual Computer* 11:105–112, 1995.

[12] A. Witkin and P. Heckbert. Using particles to sample and control implicit surfaces, In *SIGGRAPH '94 Proceedings*, pages 269–277, July 1994.

[13] H. Hoppe. Surface Reconstruction from Unorganized Points. Ph.D. Thesis, Computer Science and Engineering, University of Washington, 1994.

[14] H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, and W. Stuetzle. Surface Reconstruction from Unorganized Points. In *SIGGRAPH '92 Proceedings*, pages 71–78, July 1992.

[15] M. Melkemi, $\mathcal{A}$-shapes and their derivatives, In *13th ACM Symposium on Computational Geometry*, pages 367–369, June 1997

[16] G. Taubin and J. Rossignac. Geometric compression through topological surgery. *Research Report RC20340*, IBM, 1996.

Whole-genome Comparative Annotation and Regulatory Motif Discovery in Multiple Yeast Species

Manolis Kamvysselis[1,2], Nick Patterson[1], Bruce Birren[1], Bonnie Berger[2,3,5], Eric Lander[1,4,5]

manoli@mit.edu, nickp@genome.wi.mit.edu, birren@wi.mit.edu, bab@mit.edu, lander@wi.mit.edu

(1) MIT/Whitehead Institute Center for Genome Research, 320 Charles St., Cambridge MA 02139

(2) MIT Lab for Computer Science, 200 Technology Square, Cambridge MA 02139

(3) MIT Department of Mathematics, 77 Massachusetts Ave, Cambridge MA 02139

(4) MIT Department of Biology, 31 Ames St, Cambridge MA 02139

(5) Corresponding author

ABSTRACT

In [13] we reported the genome sequences of *S. paradoxus*, *S. mikatae* and *S. bayanus* and compared these three yeast species to their close relative, *S. cerevisiae*. Genome-wide comparative analysis allowed the identification of functionally important sequences, both coding and non-coding. In this companion paper we describe the mathematical and algorithmic results underpinning the analysis of these genomes.

We developed methods for the automatic comparative annotation of the four species and the determination of orthologous genes and intergenic regions. The algorithms enabled the automatic identification of orthologs for more than 90% of genes despite the large number of duplicated genes in the yeast genome, and the discovery of recent gene family expansions and genome rearrangements. We also developed a test to validate computationally predicted protein-coding genes based on their patterns of nucleotide conservation. The method has high specificity and sensitivity, and enabled us to revisit the current annotation of *S.cerevisiae* with important biological implications.

We developed statistical methods for the systematic de-novo identification of regulatory motifs. Without making use of co-regulated gene sets, we discovered virtually all previously known DNA regulatory motifs as well as several noteworthy novel motifs. With the additional use of gene ontology information, expression clusters and transcription factor binding profiles, we assigned candidate functions to the novel motifs discovered.

Our results demonstrate that entirely automatic genome-wide annotation, gene validation, and discovery of regulatory motifs is possible. Our findings are validated by the extensive experimental knowledge in yeast, confirming their applicability to other genomes.

Categories and Subject Descriptors

J.3 [Life and medical sciences]: Biology and Genetics

General Terms: Algorithms.

Keywords: Computational biology, Comparative genomics, Genome annotation, Regulatory motif discovery.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

RECOMB '03, April 10-13, 2003, Berlin, Germany.

Copyright 2003 ACM 1-58113-635-8/03/0004…$5.00.

1. INTRODUCTION

With the availability of complete sequences for a number of model organisms, comparative analysis becomes an invaluable tool for understanding genomes. Complete genomes allow for global views and multiple genomes increase predictive power.

In [13] we used a comparative genomics approach to systematically discover the full set of conserved genes and regulatory elements in yeast. We sequenced and assembled three novel yeast species, *S.paradoxus*, *S.mikatae* and *S.bayanus* and compared them to their close relative *S. cerevisiae*. The work represented the first genome-wide comparison of four complete eukaryotic genomes. This paper focuses on the mathematical and algorithmic developments underpinning the work.

First, we describe our methods for resolving the gene correspondence between each of the newly sequenced species and *S. cerevisiae* to identify orthologous regions and validate predicted protein-coding genes. We then describe our methods to identify conserved intergenic sequence elements within these regions and to cluster them into a small number of regulatory motifs.

The gene correspondence method presented here was used for the automatic annotation of the three newly sequenced species, and correctly identified unambiguous orthologs for more than 90% of protein coding genes. It also correctly identified the evolutionary events that separate the four species, discerning segmental duplications and gene loss, while correctly resolving genes that duplicated before the divergence of the species compared.

The methods for regulatory motif discovery presented here do not rely on previous knowledge of co-regulated sets of genes, and in that way differ from the current literature on computational motif discovery. The motifs discovered include most previously published regulatory motifs, adding confidence to our method. Moreover, a number of novel motifs are discovered that appear near functionally related genes. We have used the extensive experimental knowledge in yeast to validate our results, thus confirming that the methods presented here are applicable to other species.

1.1 Comparative annotation: graph separation

The first issue in comparative genomics is determining the correct correspondence of functional elements across the species compared. We decided to use predicted protein coding genes as genomic anchors in order to align and compare the species. Resolving the correspondence between ~6000 predicted genes in each species requires an algorithm for comparative annotation that

accounts for gene duplication and loss, and ensures that the 1:1 matches established are true orthologs.

Previously described algorithms for comparing gene sets have been widely used for various purposes, but they were not applicable to the problem at hand. Best Bidirectional Hits (BBH) [6, 7] looks for gene pairs that are best matches of each other and marks them as orthologs. In the case of a recent gene duplication, only one of the duplicated genes will be marked as the ortholog without signaling the presence of additional homologs. Clusters of Orthologous Genes (COG) [22, 23] goes a step further and allows many-to-many orthologous matches. It is able to capture gene duplication events when both copies of a duplicated gene have the same best hit in two other species that are themselves orthologous. It still suffers though from having slight changes in similarity influence a hard decision of a single best match. Moreover, since *Saccharomyces* underwent a whole-genome duplication event [14] before the divergence of the species compared, individual COGs currently contain both copies of each duplicated pair of genes in a single cluster of orthology, and hence was not applicable in our pairwise comparative annotation.

The comparative annotation algorithm we developed has features that make it useful in many applications. It compares two genomes at a time, and hence can be applied at any range of evolutionary distances, without requiring a balanced phylogenetic tree. Moreover, at its core, it represents the best match of every gene as a set of genes instead of a single best hit, which makes it more robust to slight differences in sequence similarity. Also, it groups the genes into progressively smaller subsets, retaining ambiguities until later in the pipeline when more information becomes available. It progressively refines the synteny map of conserved gene order while resolving ambiguities, one task helping the other. When it terminates, it returns the one-to-one orthologous pairs resolved, as well as sets of genes whose correspondence remains ambiguous in a small number of homology groups.

We applied this algorithm to automatically annotate the assemblies of the three species of yeast. Our Python implementation terminated within minutes for any of the pairwise comparisons. It successfully resolved the graph of sequence similarities between the four species, and found important biological implications in the resulting graph structure. More than 90% of genes were connected in a one-to-one correspondence, and groups of homologous proteins were isolated in small subgraphs. These contain expanding gene families that are often found in rapidly recombining regions near the telomeres, and genes involved in environmental adaptation, such as sugar transport and cell surface adhesion [13]. Not surprisingly, transposon proteins formed the largest homology groups.

This algorithm has also been applied to species at much larger evolutionary distances, with very successful results (Kamvysselis and Lander, unpublished). Despite hundreds of rearrangements and duplicated genes separating *S.cerevisiae* and *K.yarowii*, it successfully uncovered the correct gene correspondence between the two species that are more than 100 million years apart.

Finally, the algorithm works well with unfinished genomes. By working with sets of genes instead of one-to-one matches, this algorithm correctly groups in a single orthologous set all portions of genes that are interrupted by sequence gaps and split in two or multiple contigs. A best bi-directional hit would match only the longest portion and leave part of a gene unmatched. Finally, since synteny blocks are only built on one-to-one unambiguous matches, the algorithm is robust to sequence contamination. A contaminating contig will have no unambiguous matches (since all features will also be present in genuine contigs from the species), and hence will never be used to build a synteny block. This has allowed the true orthologs to be determined and the contaminating sequences to be marked as paralogs.

This algorithm provides a good solution to comparative genome annotation, works well at a range of evolutionary distances, and is robust to sequencing artifacts of unfinished genomes.

1.2 Motif discovery: signal from noise

Having accounted for the evolutionary events that gave rise to the gene sets in each species, we can align orthologous genes and intergenic regions and use the multiple alignments to discover conserved features, and in particular regulatory motifs. This amounts to extracting small sequence signals hidden within largely non-functional intergenic sequences. This problem is difficult in a single genome where the signal-to-noise ratio is very small.

Traditional methods for regulatory motif discovery have addressed the signal-to-noise problem by focusing on small subsets of co-regulated genes whose promoter regions are enriched in regulatory motifs. A number of elegant algorithms have been developed to search for subtle sequence signals within unaligned sequences, pioneered by Lawrence and coworkers [15], and made popular in programs such as AlignACE [11, 20, 24], MEME [10] or BioProspector [17]. More recent work has presented additional statistical methods for motif discovery using phylogenetic footprinting [3, 12, 18, 26]. Computational methods have also been developed for finding groups of possibly co-regulated genes that share similar expression profiles in a number of experimental conditions [8]. Additional experimental methods to find co-regulated genes include genome-wide discovery of promoter regions bound by a tagged transcription factor in chromatin IP experiments [16, 21], proteins found in the same protein complex obtained by MS [9] and proteins involved in the same genetically defined pathway [19]. Together, these experiments have allowed the elucidation of a large number of regulatory motifs in yeast [28] that have been categorized in promoter databases [27, 29].

Known regulatory motifs are short and sometimes degenerate, and hence appear frequently throughout the genome, often by chance alone, other times with a functional role. Phylogenetic footprinting has been used to distinguish between functional and non-functional instances, by observing alignments of orthologous promoters across multiple genomes [4]. The functional sites are constrained to contain the motifs since their change disrupts regulation which is detrimental to the organism, whereas non-functional sites are free to change and accumulate mutations.

The use of comparative information thus provides additional information that can help us separate signal from noise. This, together with a genome-wide view of the complete set of aligned orthologous intergenic regions, allows us to approach motif discovery at the genome-wide level. We are no longer constrained to observing subsets of co-regulated genes, but can search for regulatory motifs in all 6000 intergenic regions simultaneously for those sequences that are preferentially

conserved. We can then provide a global view of regulatory sequences that is not constrained by the experimental conditions generated in the laboratory, but instead captures the entire evolutionary history since the divergence of the species compared.

Our motif discovery strategy consists of an exhaustive enumeration and testing of short sequence patterns to find unusually conserved motif cores, followed by a motif refinement and collapsing step that ultimately produces a small number of full motifs. We used three different genome-wide statistics of non-random conservation to select motif cores from a large exhaustive set of short sequence patterns. We extended these cores with correlated surrounding bases that are frequently conserved, and collapsed them hierarchically based on sequence similarity and genome-wide co-occurrence. The final list of 72 genome-wide motifs includes most previously published regulatory motifs, as well as additional motifs that correlate strongly with experimental data.

Our results provide a global view of functionally important regulatory motifs, and provides an important link between protein interaction networks, clusters of gene expression, and transcription binding profiles towards understanding the dynamic nature of the cell and the complexity of regulatory interactions.

2. COMPARATIVE ANNOTATION

The first step to comparative genomics is understanding the correspondence between genes and other functional features across the species compared. Each species is under selective pressure to conserve the sequence of functionally important regions. We can begin to understand these pressures by observing the patterns of change in the sequence of orthologous regions.

In presence of gene duplication however, some of the evolutionary constraints a region is under are relieved, and uniform models of evolution no longer capture the underlying selection for these sites. Hence, before any type of motif discovery, we needed to identify unambiguously all orthologous sequences across the four genomes as a guide to our subsequent work.

We used genes as discrete genomic anchors to construct a large-scale alignment. The anchors were then used to construct a nucleotide-level alignment of genes and flanking intergenic regions. With the full assemblies of the yeast species available, we predicted all Open Reading Frames (ORFs) of at least 50 amino acids in each of the newly sequenced species, and compared the predicted proteins to the annotated proteins of S. cerevisiae using protein BLAST [1]. Since every predicted protein typically matched multiple *S.cerevisiae* genes, we first had to resolve the resulting ambiguities.

We formulated the problem of genome-wide gene correspondence in a graph-theoretic framework. We represented the similarities between the genes as a bipartite graph connecting genes between two species (Figure 1). We weighted every edge connecting two genes by the sequence similarity between the two genes, and the overall length of the match. We separated this graph into progressively smaller subgraphs until the only remaining matches connected true orthologs. To achieve this separation, we eliminated edges that are sub-optimal in a series of steps. As a pre-processing step, we eliminated all edges that are not within 20% of the maximum-weight edge incident to each node. We then separated the resulting graph into connected components, and built blocks of conserved gene order (synteny) when neighboring genes in one species had one-to-one matches to neighboring genes in the other species. We used these blocks of conserved synteny to resolve additional ambiguities by preferentially keeping syntenic edges incident to a node, and eliminating its non-syntenic edges. We finally separated out subgraphs that were connected to the remaining edges by solely non-maximal edges as described in the Best Unambiguous Subsets (BUS) algorithm. When the set of edges for each node was no further reducible, we output the connected components of the final graph as the orthology groups between the two species. We finally marked the isolated genes as paralogs of their best match.

2.1. Initial pruning of sub-optimal matches

Let **G** be a weighted bipartite graph describing the similarities between two sets of genes **X** and **Y** in the two species compared (Figure 1, top left panel). Every edge e=(x,y) in E that connects nodes $x \in X$ and $y \in Y$ was weighted by the total number of amino acid similarities in BLAST hits between genes x and y. When multiple BLAST hits connected x to y, we summed the non-overlapping portions of these hits to obtain the total weight of the corresponding edge. We constructed graph **M** as the directed version of **G** by replacing every undirected edge **e**=(**x**,**y**) by two directed edges (**x**,**y**) and (**y**,**x**) with the same weight as **e** in the undirected graph (Figure 1, top right panel). This allowed us to rank edges incident from a node, and construct subsets of **M** that contain only the top matches out of every node.

This step drastically reduced the overall graph connectivity by simply eliminating all out-edges that are not near optimal for the node they are incident from. We defined $\mathbf{M_{80}}$ as the subset of **M** containing for every node only the outgoing edges that are at least 80% of the best outgoing edge. This was mainly a preprocessing

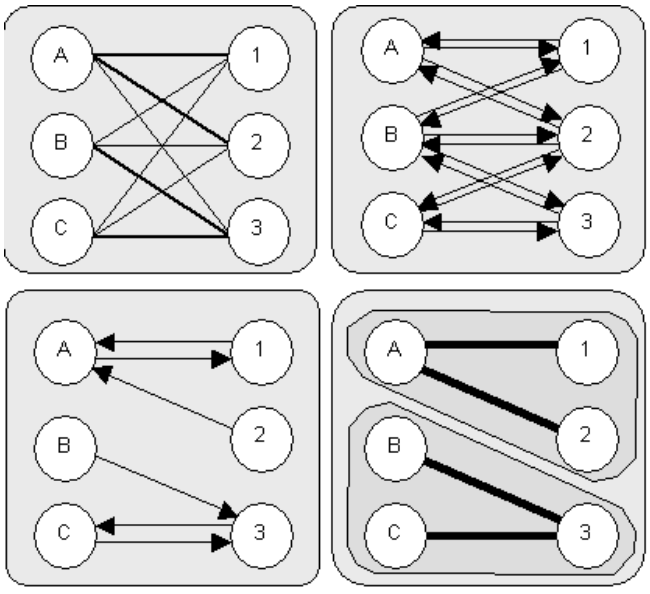


Figure 1. Overview of graph separation. We construct a bipartite graph based on the blast hits. We consider both forward and reverse matches for near-optimality based on synteny and sequence similarity. Sub-optimal matches are progressively eliminated simplifying the graph. We return the connected components of the undirected simplified graph.

step that eliminated matches that were clearly non-optimal. Virtually all matches eliminated at this stage were due to protein domain similarity between distantly related proteins of the same super-family or proteins of similar function but whose separation well-precedes the divergence of the species. Selecting a match threshold relative to the best edge ensured that the algorithm performs at a range of evolutionary distances. After each stage, we separated the resulting subgraph into connected components of the undirected graph (Figure 1, bottom right panel).

2.2. Blocks of conserved synteny

The initial pruning step created numerous two-cycle subgraphs (unambiguous one-to-one matches) between proteins that do not have closely related paralogs. We used these to construct blocks of conserved synteny based on the physical distance between consecutive matched genes, and preferentially kept edges that connect additional genes within the block of conserved gene order. Edges connecting these genes to genes outside the blocks were then ignored, as unlikely to represent orthologous relationships. Without imposing an ordering on the scaffolds or the chromosomes, we associated every gene x with a fixed position (s, start) within the assembly, and every gene y with a fixed position (chromosome, start) within *S. cerevisiae*. If two one-to-one unambiguous matches (x1, y1) and (x2, y2) were such that x1 was physically near x2, and y1 was physically near y2, we constructed a synteny block B=({x1, x2},{y1,y2}). Thereafter, for a gene x3 that was proximal to {x1, x2}, if an outgoing edge (x3, y3) existed such that y3 was proximal to {y1,y2}, we ignored other outgoing edges (x3, y') if y' was not proximal to {y1,y2}.

Without this step, duplicated genes in the yeast species compared remained in two-by-two homology groups, especially for the large number of ribosomal genes that are nearly identical to one another. We found this step to play a greater role as evolutionary distances between the species compared became larger, and sequence similarity was no longer sufficient to resolve all the ambiguities. We only considered synteny blocks that had a minimum of three genes before using them for resolving ambiguities, to prevent being misled by rearrangements of isolated genes. We set the maximum distance ***d*** for considering two neighboring genes as proximal to 20kb, which corresponds to roughly 10 genes. This parameter should match the estimated density of syntenic anchors. If many genomic rearrangements have occurred since the separation of the species, or if the scaffolds of the assembly are short, the syntenic segments will be shorter and setting *d* to larger values might hurt the performance. On the other hand if the number of unambiguous genes is too small at the beginning of this step, the genes used as anchors will be sparse, and no synteny blocks will be possible for small values of ***d***.

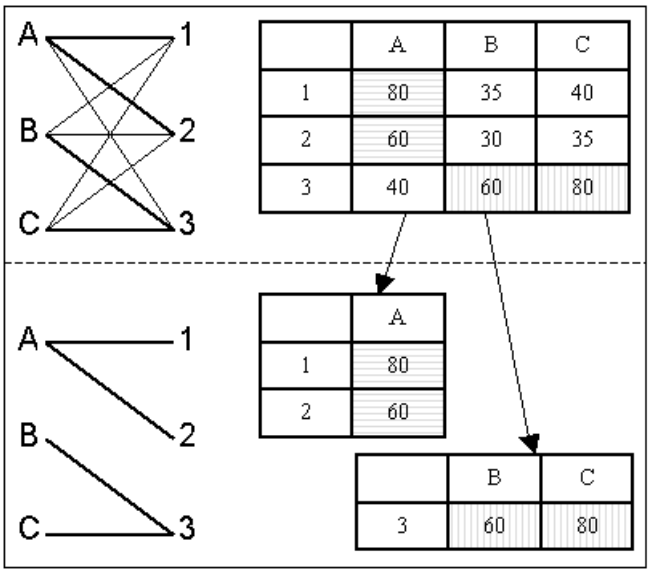


Figure 2. Best Unambiguous Subsets (BUS). A BUS is a set of genes that can be isolated from a homology group while preserving all potentially orthologous matches. Given the similarity matrix above and no synteny information, two such sets are (A,1,2) and (B,C,3).

2.3. Best Unambiguous Subsets (BUS)

To resolve additional orthologs, we extended the notion of a best bi-directional hit for sets of genes instead of individual genes. Moreover, we only constructed such a best subset when no gene outside the subset had its best match within the subset, hence when the best bi-directional subset was unambiguous. We defined a Best Unambiguous Subset (BUS) of the nodes of $\mathbf{X} \cup \mathbf{S}$, to be a subset **S** of genes, such that $\forall x: x \in S \Leftrightarrow best(x) \subseteq S$, where best(x) are the nodes incident to the maximum weight edges from x. We then constructed $\mathbf{M_{100}}$, following the notation above, namely the subset of **M** that contains only best matches out of a node. Note that multiple best matches were possible based on our definition. To construct a BUS, we started with the subset of nodes in any cycle in M_{100}. We augmented the subset by following forward and reverse best edges, that is including additional nodes if their best match was within the subset, or if they were the best match of a node in the subset. This ensured that separating a subgroup did not leave any node orphan, and did not remove the strictly best match of any node. When no additional nodes needed to be added, the BUS condition was met.

Figure 2 shows a toy example of a similarity matrix. Genes A, B, and C in one genome are connected in a complete bipartite graph to genes 1, 2 and 3 in another genome (ignoring for now synteny information). The sequence similarity between each pair is given in the matrix, and corresponds to the edge weight connecting the two genes in the bipartite graph. The set (A,1,2) forms a BUS, since the best matches of A, 1, and 2 are all within the set, and none of them represents the best match of a gene outside the set. Hence, the edges connecting (A,1,2) can be isolated as a subgraph without removing any orthologous relationships, and edges (B,1), (B,2), (C,1), (C,2), (A,3) can be ignored as non-orthologous. Similarly (B,C,3) forms a BUS. The resulting bipartite graph is shown. A BUS can be alternatively defined as a connected component of the undirected version of M_{100} (Figure 1, bottom panels).

This part of the algorithm allowed us to resolve the remaining orthologs, mostly due to subtelomeric gene family expansions, small duplications, and other genes that did not benefit from synteny information. In genomes with many rearrangements, or assemblies with low sequence coverage, which do not allow long-range synteny to be established, this part of the algorithm will play a crucial role. We have experimented running only BUS without the original pruning and synteny steps, and the results were satisfactory. More than 80% of ambiguities were resolved, and the remaining matches corresponded to duplicated ribosomal

proteins and other gene pairs that are virtually unchanged since their duplication. The algorithm was slower, due to the large initial connectivity of the graph, but a large overall separation was obtained. Figure 3 compares the dotplot of *S. paradoxus* and *S.cerevisiae* with and without the use of synteny. Every point represents a match, the x coordinate denoting the position in the *S.paradoxus* assembly, and the y coordinate denoting the position in the *S.cerevisiae* genome, with all chromosomes put end-to-end. Lighter dots represent homology containing more than 15 genes (typically transposable elements) and circles represent smaller homology groups (rapidly changing protein families that are often found near the telomeres). The darker dots represent unambiguous 1-to-1 matches, and the boxes represent synteny blocks.

2.4. Validating predicted protein-coding genes

Once we have resolved the pairwise species comparisons to *S.cerevisiae*, we build multiple alignments of both genes and flanking intergenic regions using CLUSTALW [25]. We can then

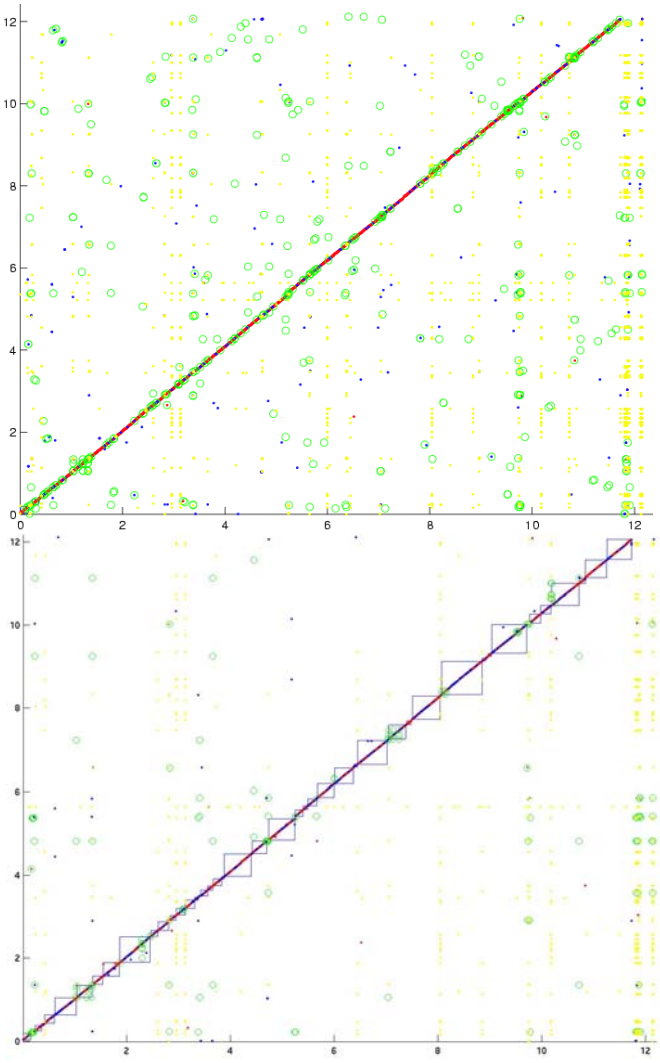


Figure 3. The effect of using synteny. Blocks of conserved gene order (blue squares) help resolve additional ambiguities.

observe the different patterns of nucleotide change in genes and intergenic regions. We find radically different types of conservation. Intergenic regions typically show short stretches between 8 and 10 bases of near-perfect conservation, surrounded by non-conserved bases, rich in isolated gaps. Protein-coding genes on the other hand are much more uniform in their conservation, and typically differ in the largely-degenerate third-codon position. Importantly, gaps are rare and when they do occur, they either happen in multiples of three, or are compensated by proximal gaps that restore the reading frame. This pressure for reading frame conservation can be used to discriminate protein-coding from intergenic regions, simply based on the pattern of gaps in the alignment.

To measure frame conservation between two aligned sequences, we label each non-gap nucleotide of the first sequence as 1, 2 or 3 cycling in order and starting at 1. We label the second sequence similarly, but once for every starting frame offset. We then simply count the percentage of aligned nucleotides that contain the same label for each of the three offsets. The offset with the maximum number of in-frame nucleotides is selected. To evaluate the frame conservation of a complete ORF, we average the percentages obtained in overlapping windows of 100bp. We obtain an average of 44% for intergenic regions (we should expect 33% at random), and an average of more than 99% for protein-coding genes. We applied a simple cutoff for each species, and tested all named S.cerevisiae ORFs, and as a control three hundred intergenic regions. We found that only 1% of intergenic regions pass the test, and less than 0.5% of named ORFs are rejected. The rejected ORFs show weak biological evidence and probably do not correspond to real genes [13].

Hence, comparative analysis can complement the primary sequence of a species and provide general rules for gene discovery that do not rely solely on known splicing signals for gene discovery. In the availability of comparative sequence information, this test provides a nice complement to programs such as GENSCAN that only look for signals in primary sequence, judging the predictions in the eye of evolution. The test presented can be used to test the validity of predicted genes in a wide range of sequenced species, even in absence of biological experimentation or known splicing signals. Even in well-studied species, this test can be used to discover additional genes that may not follow the typical rules of translation due to non-standard splicing signals, stop-codon read-through, post-transcriptional RNA editing, varying codon composition, or simply sequencing errors.

Thus, in a fully automated fashion, we have used comparative genomics to discover orthologs for virtually all protein-coding ORFs, and construct multiple alignments across the entire genome. We have used these alignments to judge the validity of protein coding genes. We now turn to the discovery of conserved regulatory motifs within the aligned intergenic regions.

3. REGULATORY MOTIF DISCOVERY

The traditional method for computational discovery of regulatory motifs has been to search within sets of co-regulated genes for enriched intergenic sequence patterns. We have undertaken a genome-wide discovery approach that should be applicable without previous knowledge of co-regulated sets. This approach is possible because the signal-to-noise ratio can be increased by comparing multiple species. Since mutations in transcription

factor binding sites may disrupt regulation, we expect regulatory motifs to be more strongly conserved than non-functional sequences that are free to diverge. Indeed, in four-way alignments of orthologous intergenic regions we observe that experimentally determined transcription factor binding sites correlate strongly with islands of sequence conservation. Moreover, the sequences of known regulatory motifs show a stronger genome-wide conservation as summed over all intergenic regions, as compared to random control patterns of the same degeneracy. Motivated by these results, we will search for motifs that show a strong genome-wide conservation.

We first exhaustively enumerated and tested the conservation of short sequence patterns to find unusually conserved motif cores. We then refine and combine these cores to construct full motifs.

3.1. Discovery of motif cores

We first enumerated all motif sequences of length 6, separated by a central gap between 0 and 21 nucleotides (mini-motifs). Each gap size consists of 2080 motifs, considering a motif and its reverse palindrome as the same motif. This results in a total of 45760 distinct mini-motifs. We assume that the large majority of these show a random conservation. We then look for those sequences that are unusually conserved as compared to a random population of mini-motifs. We use three different conservation tests.

3.1.1: Intergenic conservation (INT)

We first searched for motifs that show a significant conservation in all intergenic regions. For every mini-motif, we counted the number of perfectly conserved intergenic instances in all four species (**ic**), and the total number of intergenic instances in *S.cerevisiae* (**i**). The two counts seem linearly related for the large majority of patterns, which can be attributed to a basal level of conservation $\mathbf{r_i}$ given the total evolutionary distance that separates the four species compared. We estimate the typical ratio r_i as the log-average of non-outlier instances of ***ic/i*** within a control set of motifs. We then calculate for every motif the binomial probability p_i of observing ***ic*** successes out of ***i*** trials, given parameter r_i. We then assign a z-score S_i to every motif as the number of standard deviations away from the mean of a normal distribution that correspond to tail area p_i. This score is positive if the motif is conserved more frequently than random, and negative if the motif is diverged more frequently than random. We find that the distribution of scores is symmetric around zero for the vast majority of motifs. The right tail of the distribution however extends much further than the left tail, containing 1190 motifs more than 5 sigma away from the mean, as compared to 25 motifs for the left tail. By comparing the two counts, we estimate that 94% of these 1190 motifs are non-random in their conservation enrichment.

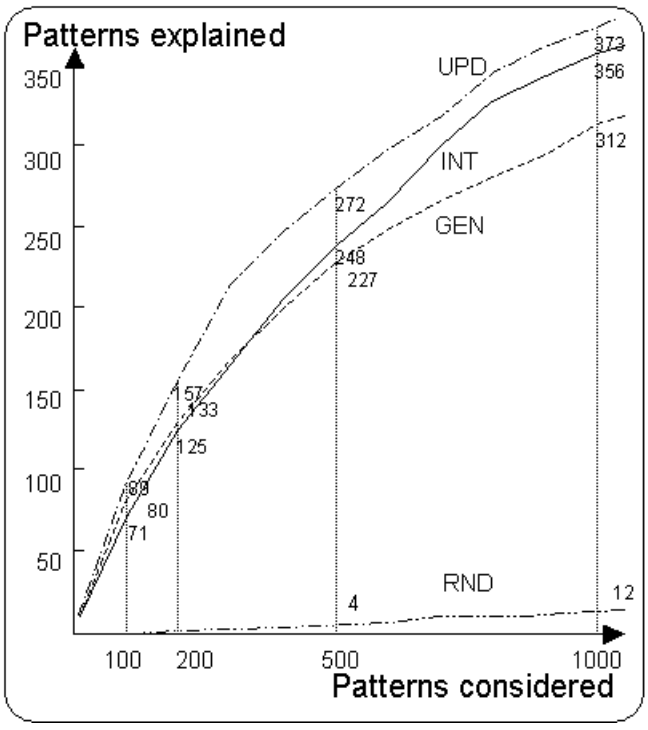


Figure 4. Selecting motif cores using three different tests. Patterns selected by one of the three tests (INT,GEN,UPD) correlate with function 90 times more frequently than randomly chosen motifs (RND).

3.1.2: Intergenic vs. coding conservation (GEN)

We then searched for motifs that are preferentially conserved in intergenic regions, as compared to coding regions. In addition to ***ic*** and ***i*** (see previous section), we counted the number of conserved coding instances ***gc***, and the number of total coding instances ***g***, for every mini-motif. We then compared the proportion of motif instances that were in intergenic regions for both conserved instances and total instances, namely ***a=ic/(ic+gc)*** and ***b=i/(i+g)***. On average, 25% of all motif instances are found in intergenic regions, which account for roughly ¼ of the yeast genome. However, only 10% of conserved motif instances appear in intergenic regions, since nucleotides in genes are more strongly conserved. For a population of motifs of similar GC-content, the ratio ***f =a/b*** remains constant. For a given motif, we calculate the enrichment in the proportion of intergenic instances, as the binomial probability of seeing at least ***ic*** successes, given ***ic+gc*** trials, given the probability ***p*** of success. To estimate ***p***, we use the proportion of total intergenic instances for that motif ***i/(i+g)***, and corrected by the log-average ***f*** of control motifs. We then score this motif by the standard deviations away from the mean of a standard normal distribution that correspond to this probability. The distribution of scores is again centered around zero for most motifs, but shows a heavier right tail. At 5 sigma, 1110 motifs are on the right tail, as compared to 39 motifs on the left tail. Hence at this cutoff, we expect 97% of motifs to be non-random.

3.1.3: Upstream vs. downstream conservation (UPD)

We finally searched for motifs that are conserved differently in upstream regions and downstream regions. We defined upstream-only intergenic regions as divergent promoters that are upstream of both flanking ORFs, and downstream-only regions as convergent 3' intergenic regions that are downstream of both flanking ORFs. We then counted ***uc*** and ***u***, the conserved and total counts in upstream-only regions, and similarly ***dc*** and ***d*** in downstream-only regions. Although upstream-only regions account for twice the total length of downstream-only regions, they show the same level of conservation, and the two ratios ***uc/u*** and ***dc/d*** are both similar to ***ic/i*** for the large majority of motifs. To detect a specificity in the upstream vs. downstream

conservation of a motif, we use a chi-square contingency test on the four counts (uc,u,dc,d). We find 1089 mini-motifs with a chi-square value of 10.83 or greater, which corresponds to a p-value of .001. We thus expect to see roughly 46 of the 45760 motifs with such a score by chance alone, and hence estimate that 96% of the 1089 mini-motifs chosen to be non-random.

3.1.4: Motifs found show category enrichment

The mini-motifs selected are indeed enriched in regulatory sequences. Many of them are at the core of well-known motifs such as Abf1, Reb1, Cbf1, Mbp1. Moreover, their conserved instances are enriched in functionally related genes. We calculated the hypergeometric enrichment score for each of these motifs against 358 functional categories, consisting of 146 sets of genes co-bound in chromatin immunoprecipitation experiments[16], 120 sets of GO molecular processes as annotated in SGD[2, 5], and 92 clusters of genes that are coordinately expressed [8]. We found that more than a third of these motifs show a significant enrichment (hypergeometric score of 10^{-5} or stronger). If we compare this result with that of a random collection of 1000 motifs, we find that only 1% show a category enrichment.

Figure 4 shows the number of motifs that show a significant category enrichment score for increasingly larger sets of top-ranked motifs in each test (INT,GEN,UPD), as compared to a random sorting (RND). From the top 100 motifs of each test, 71, 80, and 89 are explained by at least one category, as compared to only 1 for random motifs. This trend continues for the top 200, 500 and 1000 motifs. Naturally, the categories chosen here do not capture but a small fraction of the wealth of transcriptionally controlled molecular processes a cell coordinates, and hence we should not expect all motifs to show a category correlation. However, with respect to functional categories, our search shows a 90-fold enrichment in explained motifs as compared to random.

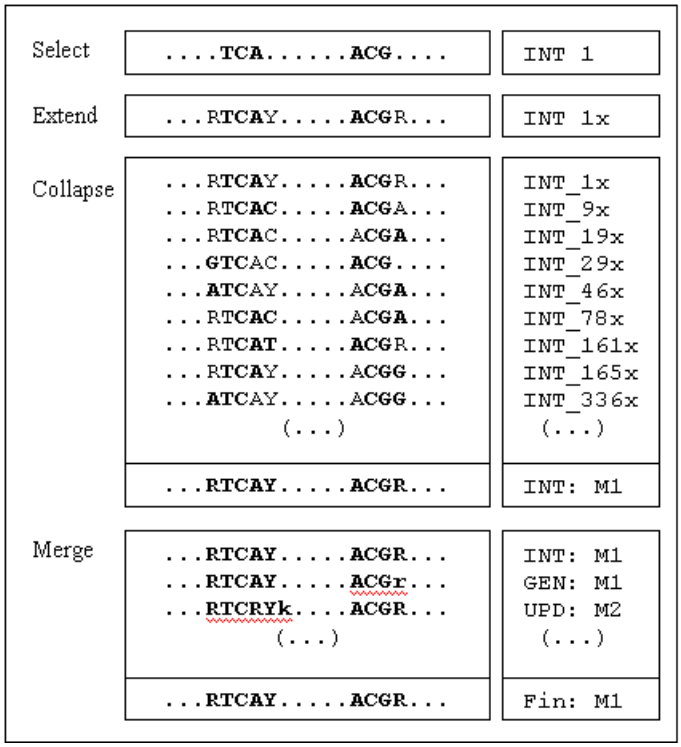


Figure 5. Overview of genome-wide motif discovery. We select motif cores by one of three tests, extend them to include additional conserved bases, and collapse together motifs with similar extension. We then merge motifs across multiple tests based on their co-occurrence in the same intergenic regions.

3.2. Constructing full motifs

We extend each of these mini-motifs by searching for surrounding bases that are preferentially conserved when the motif is conserved. We extend the motif iterative, one base at a time, by choosing, amidst the neighborhood of all conserved instances of the motif the base that maximally discriminates these from the neighborhood of non-conserved instances. The added base can be any of the fourteen degenerate symbols of the IUB code (A, C, G, T, S, W, R, Y, M, K, B, D, H, V). When no such symbol separates the conserved instances, the extension terminates. Figure 5 shows the top-scoring mini-motif found in the first test (INT_1), and the corresponding extension (INT_1x).

Many mini-motifs will have the same or similar extensions, and we group these based on sequence similarity. The similarity between two profiles is measured as the number of bits in common in the best ungapped alignment of the two profiles, divided by the number of bits contained in the profile with fewer bits. Based on the pairwise motif similarity matrix, we cluster the motifs hierarchically, until an average 70% similarity within a group is reached. This collapses the 1190 extended motifs discovered in test1 (INT) into 332 unique patterns, the 1110 motifs from test2 into 269, and the 1089 motifs from test 3 into 285 distinct patterns. The first 9 members of a cluster containing ABF1-like motifs from test1 are shown in figure 5, with mini-motif cores shown in bold, and the corresponding consensus INT_M1.

Finally, we merge motifs that co-occur in the same intergenic regions (Figure 5). The same motif will frequently be discovered across tests, or even multiple times within a test with slightly different sequences. These variations may prevent the sequence-based clustering from detecting an overlap, but the motifs will still typically occur in the same intergenic regions. To detect further overlaps, we compute a co-occurrence score between the conserved intergenic regions of each pair of collapsed motifs, and construct a consensus for the resulting group. We iterate this collapsing based on the newly constructed consensus and obtained fewer than 200 distinct motifs, of which 71 show a strong genome-wide conservation as compared to motifs of similar degeneracy.

These contain 30 known motifs, of which 28 correlate with functional categories, and an additional 41 'novel' motifs of which 61% correlate with at least one category (see [13]).

3.3. Category-based motif discovery

We further applied our motif discovery methods within functional categories. To select mini-motifs, we counted the conserved instances within the category (IN), and the conserved instances outside the category (OUT). We estimated the ratio IN/(IN+OUT) that we should expect for the category, based on the entire population of mini-motifs. We then calculated the significance of an observed enrichment as the binomial

probability of observing IN successes out of IN+OUT trials given the probability of success p. We assign a z-score to each mini-motif, as described in the genome-wide search, and similarly extended and collapsed the significant mini-motifs.

From the 106 profiled factors, 42 recognize a well-characterized motif. Of these however, only 25 show an actual enrichment in the published motif within the regions bound. In the remaining cases, the published motif may be incorrect or the ChIP experiment may be incorrect. For these 25 factors, we compared the published motif to the motif we discovered using our method, as well as the motif discovered by MEME and reported in Lee et al.

We identified short and concise motifs for all 25 factors, all of which agreed with the published consensus. On the contrary, the patterns produced by MEME typically contain additional bases that obscure the real binding site. By comparing multiple species, the signal therefore becomes stronger. It allows the search to focus on the conserved bases, eliminating most of the noise.

Table 1 summarizes the results. For each factor, we show the published motif, the hypergeometric enrichment score of the motif within the category (Hyper), the motif discovered by MEME and a quality assessment, the motif discovered by our method, as well as the corresponding category-based score and a quality assessment, and finally the comparison of our method to MEME. The performance of MEME degrades for less enriched motifs, but we consistently find the correct motif.

We then applied our methods to the complete set of 358 categories and discovered a total of 183 significant motifs. 109 categories gave rise to at least one motif, 46 gave rise to at least two motifs, and 16 gave rise to 3 motifs or more. The category-based motifs found are frequently shared across categories. After collapsing category-based motifs by sequence similarity, we obtain only 51 distinct motifs.

This overlap of the motifs discovered across categories is certainly to be expected between functionally related categories such as the chromatin IP experiment for Gcn4, the expression cluster of genes involved in amino acid biosynthesis, as well as the GO annotations for amino acid biosynthesis, all of which are enriched in the Gcn4 motif, the master regulator of amino acid metabolism.

Table 1: Category-based motif discovery. By searching for motifs that are both enriched in the category and evolutionarily conserved across the four species, we increase our sensitivity and specificity in category-based regulatory motif discovery. Here we compare known regulatory motifs to those discovered by MEME in a single genome and the ones we discover in conserved bases.

	Name	Motif	Hyper	MEME	Quality1	Our method	Score	Quality2	Comparison
1	**ABF1**	RTCRYnnnnnACG	91.4	**TRTCAY**T-Y--**ACG**RA	✓	**RTCAC_____ACG**A	14.6	✓	same
2	**GCN4**	ATGACTCAT	47.8	**TGAGTCA**Y	✓	**RTGACTCA**	10.9	✓	same
3	**REB1**	CCGGGTAA	44.7	**SCGGGTAA**Y	✓	**CCGGGTAA**C	8.7	✓	same
4	**MCM1a**	TTWCCnwwwwrGGAA	35.9	**TTTCC-AAW-RGGAA**A	✓	**TCC_____GGA**	4.4	✓	same
5	**RAP1**	ACACCCATACATTT	30.0	TTW**ACAYCCRTACAY-Y**	✓	**ACCCA.ACA**	8.7	✓	same
6	**Cbf1**	RTCACRTG	24.2	TR**GTCACGTG**	✓	**GTCACGTG**	10	✓	same
7	**FKH2**	TTGTTTACST	20.7	**TTGTTTAC**-TWTT	✓	**TGTTTTAC..TT**	8.3	✓	same
8	**SWI4**	CRCGAAAA	19.9	CSMRR**CGCGAAAA**	✓	CAA**CRCGAAAA**	8.1	✓	same
9	**MBP1**	ACGCGTnA	19.6	G-RR-A-**ACGCGT-R**	~	A**ACGCGT**CG	9.5	✓	better (+)
10	**STE12**	RTGAAACA	17.8	GSAASRR-**TGA**TR**A**WGYA		**YTGAAACA**	12.2	✓	better (+)
11	**Gal4**	CGGnnnnnnnnnnnCCG	16.1	**CGG**M---CW-Y--**CCCG**	~	**CGG_____CCGA**	7.8	✓	better (+)
12	**SWI6**	ACGCGT	15.6	**WCGCGT**CGCGTY-C	✓	**ACGCGT**	7.4	✓	same
13	**PHO4**	CACGTG	14.2	TTGTA**CACT**TYGTTT		CG**CACGTG**	4.6	✓	better (+)
14	**HSF1**	TTCTAGAA	14.1	TY**TTC**YA**GAA--TTC**Y	✓	G**TTC**TA**GAA__TTC__**G	9.6	✓	same
15	**Dig1**	RTGAAACA	13.6	CCY**TG-AY**TTCW-CTTC		**TGAAACR**	11.8	✓	better (+)
16	**INO4**	CATGTGAAat	13.4	G..G**CATGTGAAAA**	✓	G...**CATGTGAA**	6.8	✓	same
17	**FKH1**	TTGTTTACST	13.2	C**YTRTTTAY**-WTT	✓	**TGTTTAC**	6.5	✓	same
18	**Leu3**	CCGGNNCCGG	13.1	G**CCGG**TMM**CGS**YC--	✓	**CCGG___CGG**	6.6	✓	better (+)
19	**Bas1**	TGACTC	10.2	CS-CCAATGK--CS		**TGACTCT**A	9.5	✓	better (+)
20	**SWI5**	KGCTGR	9.2	CACACACACACACACACA		T**GCTGG**	6.1	✓	better (+)
21	**HAP4**	TnRTTGGT	8.5	YC**T-ATTSG**-C-GS	~	**TGATTGGT**	6.4	✓	better (+)
22	**RLM1**	CTAWWWWTAG	8.4	A-CTSGAAGAAATGCGGT		**CTA..TTTAG**	4.7	✓	better (+)
23	**INO2**	CATGTGAAat	7.4	G**CATGTGRAAA**	✓	**CATGTG**	4.4	✓	same
24	**MET31**	AAACTGTGGC	7.0	GCACGTGATS		**TGTGGC**	5.8	✓	same
25	**ACE2**	GCTGGT	5.2	GTGTGTGTGTGTGTG		T**GCTGGT**	7.4	✓	better (+)

More surprisingly however, different transcription factors often share the same binding specificity, and the same motif appears in multiple expression clusters and functional categories. For example, Cbf1, Met4, and Met31 share a motif, and so do Hsf1, Msn2 and Msn4; Fkh1 and Fkh2; Fhl1 and Rap1; Ste12 and Dig1; Swi5 and Ace2; Swi6, Swi4, Ash1 and Mbp1. Also, a single motif involved in environmental stress response is found repeatedly in numerous expression clusters, and in functional categories ranging from secretion, cell organization and biogenesis, transcription, ribosome biogenesis and rRNA processing.

Hence, the set of regulatory motifs that are specific to the categories analyzed seems limited. Only a small minority of the transcription factors probed show specificity to a concise sequence. This may be due to the cooperative nature of binding that hides the actual sequence elements used in each region. The expression clusters we have used, although constructed over an impressive array of experiments, are still limited to the relatively few experimental conditions generated in the lab. Finally, the functional categories we used are limited to the few well-characterized processes in yeast, and the molecular function of more than 3000 ORFs remains unknown.

Moreover, category-based computational identification of regulatory elements can be hampered by the fact that motifs are shared across categories. No category will be enriched in a single motif, and no motif will be enriched in a single category. By discovering in an unbiased way the complete set of conserved sequence elements, as well as their target intergenic regions, we will have the building blocks to subsequent analyses of regulatory interaction networks. Thus, a genome-wide approach is a new and powerful paradigm to understanding the dictionary of regulatory motifs.

4. CONCLUSION

Our results show that comparative analysis with closely related species can be invaluable in annotating a genome. It reveals the way different regions change and the constraints they face, providing clues as to their use. Even in a genome as compact as that of *S.cerevisiae*, where genes are easily detectable and rarely spliced, much remains to be learned about the gene content. We found that a large number of the annotated ORFs are dubious, adjusted the boundaries of hundreds of genes, and discovered more than 50 novel ORFs and 40 novel introns. Moreover, our comparisons have enabled a glimpse into the dynamic nature of gene regulation and co-regulated genes by discovering most known regulatory motifs as well as a number of novel motifs. The signals for these discoveries are present within the primary sequence of *S.cerevisiae*, but represent only a small fraction of the genome. Under the lens of evolutionary conservation, these signals stand out from the non-conserved noise. Hence, in studying any one genome, comparative analysis of closely related species can provide the basis for a global understanding of all functional elements.

5. ACKNOWLEDGEMENTS

We would like to acknowledge the continuous help of the SGD curators and in particular Mike Cherry, Kara Dolinski and Dianna Fisk. We thank Tony Lee, Nicola Rinaldi, Rick Young, Julia Zeitlinger for discussions and providing pre-publication chromatin immunoprecipitation data. We thank Mike Eisen and Audrey Gasch for discussions and providing pre-publication clusters of expression data. We thank Jon Butler, Sarah Calvo, Matt Endrizzi, James Galagan, David Jaffe, Joseph Lehar, Li-Jun Ma and all the people at the MIT/Whitehead Institute Center for Genome Research for their help and discussions. We thank Ziv Bar-Joseph, John Barnett, Tim Danford, David Gifford and Tommi Jaakkola in the MIT Lab for Computer Science for their help and discussions. We thank Gerry Fink, Ernest Fraenkel, Ben Gordon, Trey Ideker, Sue Lindquist and Owen Ozier in the Whitehead Institute for their help and discussions.

6. REFERENCES

1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. Basic local alignment search tool. *J Mol Biol*, *215* (3). 403-410.
2. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, *25* (1). 25-29.
3. Blanchette, M., Schwikowski, B. and Tompa, M. Algorithms for phylogenetic footprinting. *J Comput Biol*, *9* (2). 211-223.
4. Cliften, P.F., Hillier, L.W., Fulton, L., Graves, T., Miner, T., Gish, W.R., Waterston, R.H. and Johnston, M. Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res*, *11* (7). 1175-1186.
5. Dwight, S.S., Harris, M.A., Dolinski, K., Ball, C.A., Binkley, G., Christie, K.R., Fisk, D.G., Issel-Tarver, L., Schroeder, M., Sherlock, G., Sethuraman, A., Weng, S., Botstein, D. and Cherry, J.M. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res*, *30* (1). 69-72.
6. Fitch, W.M. Distinguishing homologous from analogous proteins. *Syst Zool*, *19* (2). 99-113.
7. Fitch, W.M. Uses for evolutionary trees. *Philos Trans R Soc Lond B Biol Sci*, *349* (1327). 93-102.
8. Gasch, A.P. and Eisen, M.B. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol*, *3* (11). RESEARCH0059.
9. Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. and Superti-Furga, G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, *415* (6868). 141-147.
10. Grundy, W.N., Bailey, T.L., Elkan, C.P. and Baker, M.E. Meta-MEME: motif-based hidden Markov models of protein families. *Comput Appl Biosci*, *13* (4). 397-406.
11. Hughes, J.D., Estep, P.W., Tavazoie, S. and Church, G.M. Computational identification of cis-regulatory elements

associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol*, *296* (5). 1205-1214.

12. Jiao, K., Nau, J.J., Cool, M., Gray, W.M., Fassler, J.S. and Malone, R.E. Phylogenetic footprinting reveals multiple regulatory elements involved in control of the meiotic recombination gene, REC102. *Yeast*, *19* (2). 99-114.
13. Kamvysselis, M., Patterson, N., Edrizzi, M., Birren, B. and Lander, E.S. submitted.
14. Keogh, R.S., Seoighe, C. and Wolfe, K.H. Evolution of gene order and chromosome number in Saccharomyces, Kluyveromyces and related fungi. *Yeast*, *14* (5). 443-457.
15. Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F. and Wootton, J.C. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, *262* (5131). 208-214.
16. Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K. and Young, R.A. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, *298* (5594). 799-804.
17. Liu, X., Brutlag, D.L. and Liu, J.S. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*. 127-138.
18. McCue, L., Thompson, W., Carmack, C., Ryan, M.P., Liu, J.S., Derbyshire, V. and Lawrence, C.E. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res*, *29* (3). 774-782.
19. Mewes, H.W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S. and Weil, B. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*, *30* (1). 31-34.
20. Roth, F.P., Hughes, J.D., Estep, P.W. and Church, G.M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol*, *16* (10). 939-945.
21. Simon, I., Barnett, J., Hannett, N., Harbison, C.T., Rinaldi, N.J., Volkert, T.L., Wyrick, J.J., Zeitlinger, J., Gifford, D.K., Jaakkola, T.S. and Young, R.A. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, *106* (6). 697-708.
22. Tatusov, R.L., Koonin, E.V. and Lipman, D.J. A genomic perspective on protein families. *Science*, *278* (5338). 631-637.
23. Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D. and Koonin, E.V. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*, *29* (1). 22-28.
24. Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and Church, G.M. Systematic determination of genetic network architecture. *Nat Genet*, *22* (3). 281-285.
25. Thompson, J.D., Higgins, D.G. and Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, *22* (22). 4673-4680.
26. Tompa, M. Identifying functional elements by comparative DNA sequence analysis. *Genome Res*, *11* (7). 1143-1144.
27. Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhauser, R., Pruss, M., Schacherer, F., Thiele, S. and Urbach, S. The TRANSFAC system on gene expression regulation. *Nucleic Acids Res*, *29* (1). 281-283.
28. Zhang, M.Q. Promoter analysis of co-regulated genes in the yeast genome. *Comput Chem*, *23* (3-4). 233-250.
29. Zhu, J. and Zhang, M.Q. SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics*, *15* (7-8). 607-611.

Phylogenetically and Spatially Conserved Word Pairs Associated with Gene Expression Changes in Yeasts

Derek Y. Chiang
Dept of Molecular and Cell Biology, UC Berkeley
dchiang@ocf.berkeley.edu

Alan M. Moses
Graduate Group in Biophysics, UC Berkeley
amoses@ocf.berkeley.edu

Manolis Kamvysselis
MIT Laboratory of Computer Science
manolis@mit.edu

Eric S. Lander
MIT/Whitehead Institute Center for Genome Research
lander@genome.wi.mit.edu

Michael B. Eisen
Dept of Molecular and Cell Biology, UC Berkeley
Division of Genome Sciences, Lawrence Berkeley National Lab
mbeisen@lbl.gov

ABSTRACT

Background. Transcriptional regulation in eukaryotes is often multifactorial, involving multiple transcription factors binding to the same transcription control region (*e.g.*, upstream activating sequences and enhancers), and to understand the regulatory content of eukaryotic genomes it is necessary to consider the co-occurrence and spatial relationships of individual binding sites. The identification of sequences conserved among related species (often known as phylogenetic footprinting) has been successfully used to identify individual transcription factor binding sites. Here, we extend this concept of functional conservation to higher-order features of transcription control regions involved in the multifactorial control of gene expression.

Results. We used the genome sequences of four yeast species of the genus *Saccharomyces* to identify sequences potentially involved in multifactorial control of gene expression. We found 1,117 potential regulatory "templates": pairs of hexameric sequences that are jointly conserved in transcription regulatory regions and also exhibit non-random relative spacing. Many of the individual sequences in these templates correspond to known transcription factor binding sites, and the sets of genes containing a particular template in their transcription control regions tend to be differentially expressed in conditions where the corresponding transcription factors are known to be active.

Conclusions. The incorporation of both joint conservation and spacing constraints of sequence pairs predicts groups of target genes that were specific for common patterns of gene expression. Our work suggests that positional information, especially the relative spacing between transcription factor binding sites, may represent a common organizing principle of transcription control regions.

Copyright 2003 Association for Computing Machinery. ACM acknowledges that this contribution was authored or co-authored by a contractor or affiliate of the U.S. Government. As such, the Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only.
RECOMB'03, April 10-13, 2003, Berlin, Germany.
Copyright 2003 ACM 1-58113-635-8/03/0004…$5.00.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and genetics

General Terms

Algorithms

Keywords

Phylogenetic footprinting, comparative genomics, multifactorial regulation, transcription regulation, promoter structure

1. INTRODUCTION

All organisms have evolved intricate signaling networks that sense and respond to their environment. At a cellular level, the activation of one or more signaling networks often leads to coordinated changes in gene expression, via the regulated activity and binding of transcription factors to transcription control regions (TCR's) of genes (e.g. enhancers and upstream activating sequences). In yeast and most other eukaryotes, the transcriptional regulation of individual genes is often multifactorial, as multiple transcription factors may bind to a single TCR [1] [2] [3]. In some cases, multiple transcription factors bind to a TCR and act independently of one another to alter gene expression in response to distinct cellular cues [4]; in other examples, multiple factors bind and/or act cooperatively to modulate gene expression via direct or indirect physical interactions with each other [5] [6] [7].

The challenges in understanding how regulatory information is encoded in genomes include both the identification of regulatory sequences in TCR's, and the elucidation of the constraints on productive multifactorial regulation. Many experiments have shown that specific pairs of factors must be bound near each other in order to act cooperatively [8] [9] [10], and it is on these spatial constraints that we focus here.

Previous computational work has been devoted to identifying putative transcription factor binding sites. A plethora of computational methods has been developed to find over-represented sequences in a subset of genes believed to contain a common transcription factor binding site (reviewed in [11]). The

rapid pace of genome sequencing has enabled a complementary approach – phylogenetic footprinting (reviewed in [12] [13]) – that recognizes that the conservation of sequences across related organisms often reflects evolutionary selection for their presence in TCR's. Several algorithms have been developed to perform phylogenetic footprinting analyses systematically [14] [15] [16].

After compiling a collection of putative binding sites, associations can be made between various binding site assortments and gene expression. Some recent approaches include Boolean logic [17], regression methods [18] [19] [20], spatial clustering [21], and multiple binding site matrix classifiers [22] [23] [24]. Spatial information on the relative locations of binding sites is ignored in all but the last two classes of approaches. Yet even these methods, which often search for fixed arrangements among the individual binding sites, may miss permutations of binding sites within TCR's that may still be bound and regulated by their corresponding transcription factors.

The primary aim of this work was to incorporate positional information and phylogenetic footprinting to identify sequence motifs that may regulate gene expression. Consequently, we expanded the focus of phylogenetic footprinting from the conservation of contiguous sequences to higher-order features of TCR's, namely the spatial organization of individual binding sites. Since transcription factors participating in multifactorial regulation may require physical proximity among their binding sites, we searched for groups of conserved sequences that were more closely spaced in TCR's than expected. We refer to these spatially organized sequences as conserved word templates. As a proof of principle, we started with the simplest example of such templates: pairs of conserved 6-bp words. Conservation was assessed using the genome sequences of three additional *Saccharomyces* species, which were chosen to be sequenced in order to elucidate regulatory sequences conserved among these closely related species [25]. To exploit this comparative genome data, we have devised a method that systematically tested sequence pairs for joint conservation across genomes and close spacing within individual TCR's. Since genes regulated by the same set of transcription factors often display similar gene expression patterns in certain experimental conditions, we identified conserved word pair templates whose gene targets were associated with common changes in gene expression. We adopted a group-by-sequence approach to first identify genes that contained the word pair templates and then to test for significant associations with expression levels of the identified genes [26]. Significant associations between conserved word pair templates and specific gene expression changes, the prevalence of known transcription factor binding sites, and the enrichment for common functional roles among gene groups, suggest that conserved word pair templates comprise sequences important for multifactorial regulation in yeast.

2. CONSERVED WORD PAIR TEMPLATE ALGORITHM

2.1 Overview

We present a method to find conserved higher-order sequence templates from related *Saccharomyces* genomes. Our method incorporates sequential statistical tests, with each step focusing on a distinct property of conserved sequence templates. The simplest instances of sequence templates involve word pairs and their relative spacing. First, word pairs that show enriched conservation as a unit were identified using a chi-square test for independence. Next, the relative spacing of conserved word pairs was assessed using a permutation test. Finally, those conserved word pairs with close spacing were verified for functional importance by testing for gene expression differences between matching genes and the rest of the genome. The output for this algorithm is a P × C data matrix, whose entries correspond to the strength of association with differential gene expression, i.e. the K-S significance level (see §2.5). Note that P is the number of significant conserved word pairs, and C is the number of gene expression conditions.

2.2 Datasets

Whole-genome shotgun sequencing of *Saccharomyces bayanus*, *Saccharomyces mikatae*, and *Saccharomyces paradoxus* has been previously described [25]. All of these organisms are highly related to *Saccharomyces cerevisiae*, as they are grouped within the *sensu stricto* branch of the *Saccharomyces* genus [47]. Intergenic regions were aligned using CLUSTALW as described [25] and are available from the Saccharomyces Genome Database [43]. A total of 4101 CLUSTALW alignments were analyzed. These alignments were filtered for orthologs in at least 3 genomes.

Gene expression measurements were obtained from the Stanford Microarray Database [48] and Rosetta [34]. The main experimental types among the 342 conditions examined include diauxic shift [27], cell cycle [29] [30], environmental stress response [28], DNA damage [31] [32], low phosphate [33], cadmium (N. Ogawa and P. O. Brown, unpublished data), and inhibition of ergosterol biosynthesis [34]. This data has been log-transformed (base 2), and each experimental condition has been median normalized.

2.3 Dependent Conservation of Word Pairs

To assess whether two words were co-conserved in the same intergenic regions, a chi-square test of independence was systematically conducted for all possible words of length six. We defined a word to include a 6-bp sequence and its reverse complement. Define a transcription control region (TCR) for a gene as the 600 base pairs upstream of its translation start site. TCR's shared between divergently transcribed genes less than 600 bp long were only counted once. A word was labeled conserved in a TCR if all six bases were identical among three or more genomes in the CLUSTALW alignment. For each word pair (W, V) whose overlap was less than 4, a contingency table $\boldsymbol{C}_{wv}$ was constructed. In this table, $\boldsymbol{C}_{wv}$ = # TCR($\boldsymbol{I}_w \cap \boldsymbol{I}_v$), where $\boldsymbol{I}_w$, $\boldsymbol{I}_v$ are indicator variables for the presence of each conserved word in a TCR. TCR's shared between divergently transcribed genes less than 600 bp long were only counted once. The expected counts $\boldsymbol{E}_{wv}$ were obtained from an independence assumption, *i.e.* the product of the individual word conservation probabilities, multiplied by the total number of TCR's. Thus the chi-square statistic with Yates continuity correction was computed according to the definition:

$$\chi^2_{wv} = \sum_{I_w=0}^{1} \sum_{I_v=0}^{1} \frac{(|C_{wv} - E_{wv}| - \frac{1}{2})^2}{E_{wv}} \qquad (1)$$

2.4 Spatial Proximity of Word Pairs

The second requirement for a conserved sequence template involved constraints on spatial arrangements between individual words. Any method that evaluates spacing distributions between word pairs must take into account positional biases that may be present for individual words (A. M. Moses, unpublished results). We used a permutation test to evaluate the significance of the average minimum distance, excluding overlaps, between conserved word pairs. By permuting the TCR labels for one of the words, but not the word positions themselves, we retained the positional biases of individual words within intergenic regions. Within any given TCR t, define $p_t(W) = \{p_t^1(W), \ldots, p_t^j(W)\}$ as a vector of positions in *S. cerevisiae* where word W is conserved. Suppose that words W and V were jointly conserved in TCR's $T_1 \ldots T_N$. Then the average minimum distance, $\overline{D}$, can be computed as:

$$\overline{D}_{wv} = \frac{1}{T}\sum_{t=1}^{T} \min_{j,k} \left| p_t^j(W) - p_t^k(V) \right| \qquad (2)$$

We used a permutation test to generate an empirical null distribution of $\overline{D}$ for all word pairs with $N \geq 10$. After randomly permuting the labels t for the position vectors of word V, a permutation test statistic, $\overline{D}^*$, can be calculated as above. By repeating this resampling procedure R times, an empirical null distribution $\overline{D}_{null} = \{\overline{D}^{*1}, \ldots, \overline{D}^{*R}\}$ can be obtained. The significance of the observed average minimum distance, $\overline{D}$, in the N promoters was calculated as its quantile in the empirical null distribution $\overline{D}_{null}$. We set an upper bound of $R = 10^6$, but stopped permutations early if 20 or more values in $\overline{D}_{null}$ were found less than $\overline{D}$.

Correction for multiple testing involved control of the proportion of false positives using a False Discovery Rate method [1]. This method has increased power over Bonferroni-type methods. Permutation quantiles for all N word pairs tested were sorted in non-decreasing order: $q_1 \leq \ldots \leq q_N$. Let $k = \max\left(i : q_i < \frac{0.05 i}{N} \right)$. Then the first k word pairs in the ordering had a corrected significance level of $q < 0.05$, *i.e.* the rate of false positives is approximately 5%.

2.5 Association between Template-Specified Gene Groups and Gene Expression Changes

So far we have identified word pairs with two properties: dependent conservation and spatial proximity among all TCR's in the whole genome. These word pairs can be viewed as sequence-based rules for selecting a subset of genes based on the conservation of an element of TCR architecture. In this stage, we would like to evaluate the transcriptional information associated with these rules by assaying for gene expression changes among genes that match these sequence constraints.

For each gene expression condition c in our dataset, $c \in \{1, \ldots, 342\}$, we tested the null hypothesis that a gene subset $G_{wv} \subseteq G$ selected by a conserved word pair (w, v) had the same distribution of gene expression ratios (E_{wv}^c) as the entire genome (E^c). The alternate hypothesis stated that the two gene expression distributions were significantly different. Any gene was an element of G_o if its corresponding TCR conserved both sequences in the word pair. Since the size N_o of gene subsets may be small and the distributions may not be normally distributed, we used the nonparametric Kolmogorov-Smirnov (K-S) test. The test statistic K compares the cumulative distribution functions F_{wv}^c and F^c corresponding to E_{wv}^c and E^c by the formula $K = \max_x \left| F_{wv}^c(x) - F^c(x) \right|$. The significance level of an observed value K^* can be obtained using a numerical approximation [51].

A gene subset determined by a word pair was deemed to have significantly different expression if its K-S p-value was less than a certain threshold. To correct for multiple testing, this threshold was established by controlling the False Discovery Rate. The significance levels p_i from each K-S test were ordered in ascending order. Let N represent the total number of K-S tests performed, i.e. the number of jointly conserved, closely spaced word pairs times the number of gene expression experiments). If k was the largest i such that $p_i < i\alpha / N$, then the first k word pairs in the ordering were deemed to have a significance level of $p < \alpha$.

We ensured that the K-S p-value for the conserved word pair subset G_o was more significant than subsets G_w or G_v comprised of only one conserved word by computing K for E_w^c vs. E_v^c, as well as for E_w^c vs. E^c. The marginal improvement of the joint word pair was defined as: K (F_o^c vs. F^c) – max(K (F_w^c vs. F^c), K (F_v^c vs. F^c)).

3. RESULTS

3.1 Identification of conserved word pair templates

We initialized our word list using all 2080 words of length six, treating a given word and its reverse complement as identical. For each TCR (consisting up to 600 bp upstream of an open reading frame), a word was labeled conserved if all six bases were identical in at least three of the four *Saccharomyces* genomes, based on the CLUSTALW alignment of that TCR. To systematically test whether words were conserved more often in the same intergenic regions of the *Saccharomyces* genomes than expected by independent conservation, a chi-square test was performed on all possible pairwise combinations of words (see §2.3). Pairs of words that overlapped each other by more than three nucleotides were excluded. A significant proportion of word pairs showed dependent conservation: among the 2.16 million word pairs tested, 8452 of them (~0.4%) had conservation c2 scores greater than 31.1. This threshold corresponds to a probability of 0.05 for obtaining one or more false positives after a Bonferroni correction for multiple testing.

Next, we selected word pairs that displayed closer physical spacing in intergenic regions than expected by chance. As a metric for the closeness in relative spacing between word pairs, the average minimum distance between two words in *S. cerevisiae*, $\overline{\ }$, was calculated based on the genes whose TCR's conserved both words. If two non-overlapping words were closely

spaced in all TCR's, we should find $\overline{D}$ to be smaller than expected by chance. This spacing was assessed using a permutation test by selecting the set of genes that contained a conserved word pair and then randomizing the assignment of one of the words to the genes containing that word (see §2.4). By permuting the TCR labels for one of the words, but not the word positions themselves, we retained the positional biases of individual words within intergenic regions.

After correcting for multiple testing, a total of 1117 out of 8452 word pairs (~13%) had significantly small values (FDR $q < 0.05$) for $\overline{D}$ (Figure 1). As a negative control, we also assayed a sample of word pairs that did not show dependent conservation (conservation $\chi^2 < 1$), yet were jointly conserved in at least 10 TCR's. Only 161 out of 42801 (~0.4%) random word pairs with non-dependent conservation ($\chi^2 < 1$) showed significantly small values for $\overline{D}$. Figure 2 illustrates the distributions of $\overline{D}$ for conserved word pair templates, jointly conserved word pairs, and randomly conserved word pairs. The medians of these distance distributions were 100 nucleotides, 116.5 nucleotides and 132 nucleotides, respectively. Notably, the median $\overline{D}$ for template pairs was significantly smaller ($p < 0.05$) than the median $\overline{D}$ for randomly conserved pairs. These results indicate that many of the word pairs that were conserved in the same intergenic regions of multiple *Saccharomyces* genomes also exhibited closer spacing in TCR's.

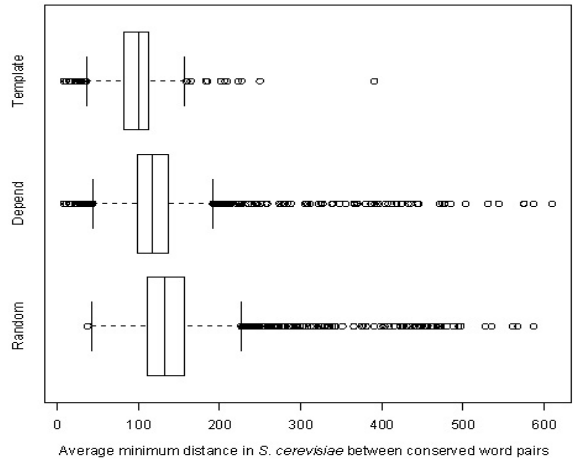


Figure 1) Word pairs in conserved word pair templates are closely spaced in *S. cerevisiae*

Template denotes closely spaced and jointly conserved word pairs ($\chi^2 > 31.1$, spacing $q < 0.05$, $N = 1117$). *Depend* denotes dependently conserved word pairs ($\chi^2 > 31.1$, $N = 8452$) and includes all of the word pairs in the template category. *Random* denotes a sample of randomly conserved word pairs ($\chi^2 < 1$, $N = 4667$). For each category, the distribution of average minimum distances is represented by a box-and-whisker plot.

3.2 Conserved word pair templates were significantly associated with gene expression

Our method identified conserved word pair templates that were statistically significant with respect to both co-conservation in multiple genomes and close spacing in *S. cerevisiae* TCR's. To evaluate the regulatory information in these templates, we assessed the statistical association between gene groups that shared a template and changes in gene expression. Similar to other group-by-sequence approaches for finding regulatory sequences, we expect that gene subsets defined by common TCR sequence rules should have gene expression patterns that are similar under conditions where the transcription factors are active, yet are different from the average expression of genes in the genome [26].

To assess the association between conserved word pair templates and differentially expressed genes, we identified gene subsets that contain both conserved words in the template within their TCR's and observed their expression patterns in S. cerevisiae in publicly available datasets ([27] to [34], see §2.5). We then conducted Kolmogorov-Smirnov (K-S) tests to evaluate for differential gene expression between each gene subset and the whole genome. A P × C matrix was computed: each conserved word pair in P was assigned a K-S p-value for each experimental condition observed in C. (see §2.5). Entries in this matrix (K-S p-values) were filtered out if the K-S *p*-value: (1) did not meet the threshold for multiple testing; or (2) was less than 10 times more significant than the K-S *p*-value for a gene subset associated with either word alone (see §2.5). The latter criterion discounts gene expression changes that are due predominantly to the action of a single transcription factor.

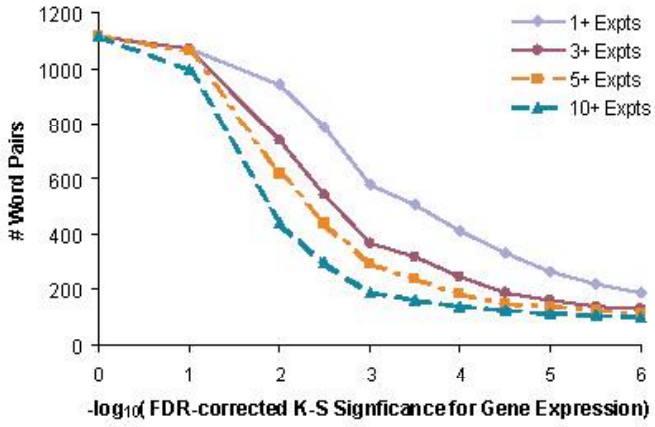


Figure 2) Total number of conserved word pair template associations at different K-S significance values

The horizontal axis shows different multiple testing-corrected significance levels for the K-S test (see §2.5). The number of closely spaced word pairs meeting this cutoff for different minimum numbers of expression conditions is shown on the vertical axis. Word pairs were also filtered for an improvement of 10× over the K-S significance from any single word.

Figure 2 displays the number of conserved word pair templates that were significantly associated with gene expression changes, for varying significance levels of the K-S test, which have been corrected for multiple testing (see §2.5). Each line indicates the number of gene subsets that were significant in a different minimum number of experimental conditions. Several hundred closely spaced word pairs were significantly associated with differential gene expression. For example, 293 word pairs met an FDR-corrected significance threshold of $p < 10^{-3}$ for 5 or more experimental conditions.

3.3 Many identified sequences represent known transcription factor binding sites

Conserved word pair templates that were most strongly associated with gene expression changes also agreed with prior experiments on transcription factors [35]. In all analyses described below, we used a set of 339 word pairs that had significant associations with gene expression changes at an FDR-corrected multiple testing threshold of 0.005 for 10 or more experiments. For visualization purposes, we organized the P × C matrix by hierarchically clustering the K-S p-values for the 339 word pairs.

Hierarchical clustering of this output matrix identified groups of word pairs with similar K-S p-values in specific subsets of experimental conditions (Figure 4). In many cases, the word pairs that clustered together also comprised overlapping hexamer sequences, suggesting that some of the hexamers in different pairs may represent a larger, somewhat variable sequence (Table 1). For example, group #13 in Figure 4 includes 8 word pairs. In each of these word pairs, one of the component words – such as TCACGT, GCACGT or CACGTG – matched part of the Cbf1p or Pho4p binding sites. The other component word in each pair – such as AACTGT, ACTGTG, CTGTGG, TGTGGC or GTGGCT – represented part of the known Met31/32p binding site (AAACTGTGG). Therefore, genes whose TCR's contained any word pair within this group likely contained a conserved Cbf1p or Pho4p binding site, along with a conserved Met31/32p binding site, and the distances between the conserved sites in these genes were also smaller than expected by chance. These results agree with the known interaction of Cbf1p and Met31/32p for the regulation of genes involved in sulfur utilization (see Discussion).

Table 1 shows a partial list of the 13 most significant groups of consensus sequences, which were assembled by joining adjacent word pairs in the clustered output matrix with overlapping sequences. Many of these consensus sequences matched transcription factor binding sites that had been biochemically verified. Several pairs of transcription factors, denoted by boldface in Table 1, were not previously known to participate in multifactorial regulation. Three of these pairs included new putative transcription factor binding sites. In group 8, the word ACAGCC is found in a template with the GATA motif. In group 9, the word CGGGCC is found in a template with the binding site for the stress-induced transcription factor, Msn2/4p. In group 2, one of the words in each word pair was the binding site for Swi4/6p, which regulates the expression of cell-cycle dependent genes. The other word in each pair was an invariant CGCCAA, which is highly similar to, though distinct from, the characterized Swi4/6 binding site CRCGAAA [35].

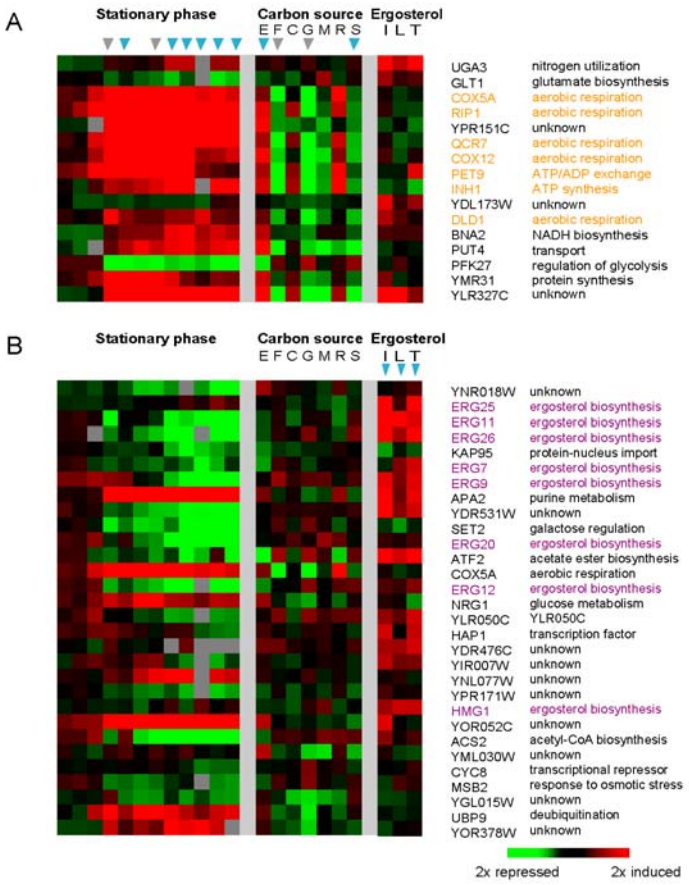


Figure 3) Multifactorial regulation of Hap1p target genes

Gene expression patterns are shown for genes whose TCR's contain binding sites for: (A) Hap1p (CCGATA) and Hap2/3/4/5p (CCAATC); or (B) Hap1p (CCGATA) and Ecm22p/Upc2p (TCGTTT). The genes are listed in ascending order of minimum distance between the two conserved words in the corresponding TCR of *S. cerevisiae*. Each row represents a given gene's expression pattern under the conditions shown in each column: progression into stationary phase (2 h, 4 h, 8 h, 12 h, 1 day, 2 days, 3 days, 5 days, 7 days, 13 days, 22 days, 28 days of growth) [28]; steady-state growth on different carbon sources: ethanol (E), fructose (F), galactose (C), glucose (G), mannose (M), raffinose (R) and sucrose (S) [28]; and growth in the presence of drugs that inhibit ergosterol biosynthesis: itraconazole (I), lovastatin (L) and terbinafine (T) [34]. A red color indicates that the gene's expression was induced under those conditions, while a green color indicates that the gene was repressed under those conditions; black indicates no detectible change in expression, and grey indicates missing data. Gene names highlighted in orange (A) or in purple (B) correspond to genes whose products are involved in respiration and ergosterol biosynthesis, respectively. Arrows above the columns indicate conditions in which the displayed gene groups show significant gene expression changes according to the Kolmogorov-Smirnov test, after False Discovery Rate correction for multiple testing at a *p*-value of 0.005 (blue) or 0.01 (grey).

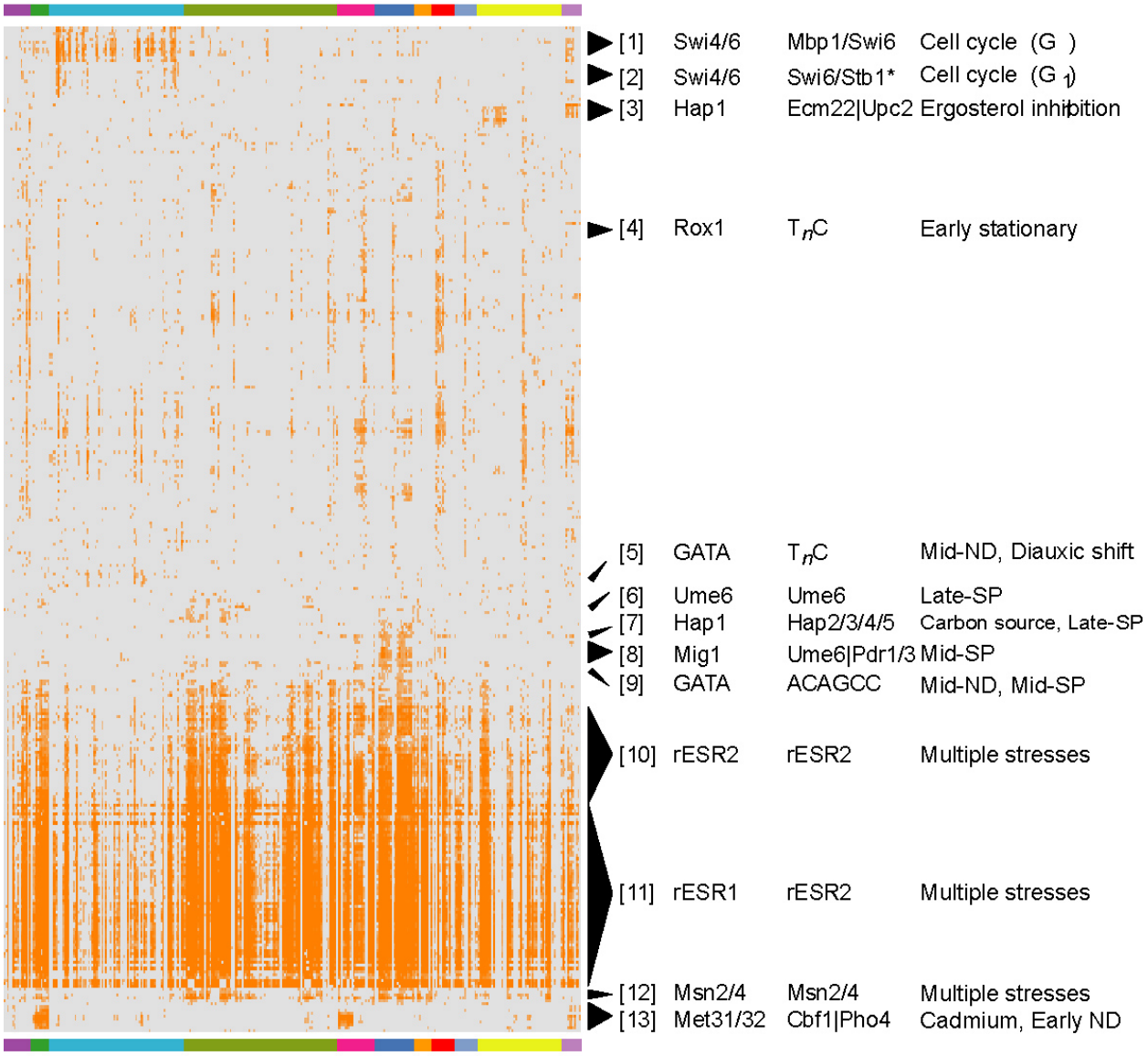


Figure 4) Specific patterns of gene expression changes associated with templates

The $P \times C$ matrix of K-S p-values was hierarchically clustered by rows and visualized with TreeView (http://rana.lbl.gov). Each row corresponds to a conserved word pair template, and each column represents a single gene expression experiment. The experimental conditions are indicated by the color bar above and below the figure, according to the key shown below. The value in each cell corresponds to the K-S p-value of gene expression changes in each condition (column) for a group of genes that contain the conserved word pair template (row) in their TCR's. An orange color denotes a K-S p-value below the FDR critical value of 0.005 for multiple testing, while grey represents values that were not significant. Word pairs that failed to meet a False Discovery Rate critical value of 0.005 for multiple testing in 10 or more experiments are not shown. Some of the most significant conserved word pair associations are labeled and annotated in Tables 1 and 2. Abbreviations for experimental conditions include: ND (nitrogen depletion), SP (stationary phase).

Table 1) Gene expression associations for most significant groups of word pairs

The output $\boldsymbol{P} \times \boldsymbol{C}$ matrix of word pairs ($\boldsymbol{P}$) that were significantly associated ($p < 0.005$) with at least 10 or more environmental conditions ($\boldsymbol{C}$) was ordered using hierarchical clustering. Numbers correspond to groups of overlapping word pairs indicated in Figure 4. Boldface denotes sequence pairs whose involvement in multifactorial regulation has not been previously reported. Consensus sequences were assembled from groups of word pairs that were found in adjacent rows in the ordering. Residues are shown in bold if it is contained in at least two hexamers. Numbers denote the groups that are indicated in Figure 4. IUPAC codes used: K (G or T); M (A or C); R (A or G); S (C or G); W (A or T).

	Most significant word pair in consensus group					
	Conserved Word Pairs (Consensus of overlapping words)	Known transcription factors or motifs	Conservation (χ^2, p-val via Bonferroni)	Avg. min dist $\overline{\boldsymbol{D}}$	# TCR	Expression conditions with significant gene subsets (FDR significance)
1	RCGAAA, RACGCG,	Swi4/6, Swi6/Mbp1	83.0 (7×10^{-13})	68.1	36	Cell cycle, G1 phase (10^{-6})
2*	**CACGAAA, CGCCAA**	**Swi4/6, Stb1 (putative)**	**55.6 (2×10^{-7})**	**78.2**	**25**	**Cell cycle, G1 phase (10^{-4})**
3*	**CCGATA, TC[GT]TTT**	**Hap1, Ecm22 | Upc2**	**36.2 (0.004)**	**81.7**	**30**	**Ergosterol inhibition (10^{-4}) MMS (DNA damage) (10^{-3})**
4*	**GATAAG, TTCTTT**	**GATA, T_nC**	**36.0 (0.005)**	**100.5**	**88**	**Nitrogen depletion 8h (10^{-5})**
5	GGCTAA CGGCGG	Ume6, Ume6	179.2 (2×10^{-34})	81.9	15	Late stationary phase (10^{-4})
6	CCGATA, CCAATC	Hap1, Hap2/3/4/5	35.0 (0.007)	88.6	16	Stationary phase (10^{-4}) Ethanol (10^{-4})
7	ACCCCA, CCGCCG	Mig1, Ume6 | Pdr1/3	66.7 (7×10^{-10})	70.5	16	Stationary phase (10^{-6}) Ethanol (10^{-5})
8*	**GATAAG, ACAGCC**	**GATA, Novel**	**39.5 (0.004)**	**75.5**	**21**	**Nitrogen depletion 8,12h (10^{-5}) Stationary phase 10h-2d (10^{-4})**
9*	**AAGGGG, CGGGCC**	**Msn2/4, Novel**	**33.4 (0.016)**	**79.6**	**14**	**Elutriation 2d, 4d, 6d (10^{-5}) Stationary phase 10h-3d (0.005)**
10	ANTGAAA, GAAAAWT	rESR2 (Overlap)	96.9 (2×10^{-16})	96.8	68	Repressed in multiple environmental stresses (10^{-6})
11	G[AC]GATGAG TGAAAATTTT	rESR1 motif, rESR2 motif	240.6 (10^{-49})	41.6	183	Repressed in multiple environmental stresses (10^{-6})
12	AWAAGG, AGGGG	Msn2/4 (Overlap)	94.7 (5×10^{-16})	99.0	29	Multiple stresses (10^{-3})
13	ACTGTGGC, [GT]CACGTG	Met31/32, Cbf1 | Pho4	47.5 (2×10^{-5})	43.5	22	Amino acid starv. (10^{-6}) Nitrogen depletion (10^{-6}) Cadmium (10^{-6})

Recent chromatin immunoprecipitation experiments suggested that this sequence may represent the binding site for Stb1, a transcription factor that binds Swi6 in vitro and is implicated in cell cycle regulation [36]. This sequence was found in several genes adjacent to intergenic regions bound by Stb1 *in vivo* [37].

Some groups of genes with shared word pair templates were enriched for known targets of transcription factors. Genes with a conserved half-site for the Hap1p transcription factor, as well as a conserved Hap2/3/4/5p binding site, in their TCR's were significantly associated with induction in late stationary phase (Figure 3A). In addition, many of these genes were more highly expressed in growth medium containing ethanol, relative to other carbon sources (Figure 3A). Many of these genes encode aerobic respiration enzymes, which are required for the switch from fermentation to respiration [38] [39]. Indeed, both the Hap1p transcription factor and the Hap2/3/4/5p transcription factor complex are known to regulate the expression of these genes in response to heme and/or oxygen availability and carbon source, respectively. By contrast, gene groups with both a conserved Hap1p binding site and a conserved Ecm22p or Upc2p binding site in their TCR's were only significantly associated with induction in the presence of a drug that inhibited ergosterol biosynthesis (Figure 3B). This group of 30 genes contained 8 ergosterol biosynthesis genes; this proportion represented an enrichment compared to the rest of the genome. The transcription factors Ecm22p and Upc2p have been shown to induce the expression of ergosterol biosynthesis genes in response to low intracellular concentrations of ergosterol, while Hap1p is known to regulate the expression of these genes according to the availability of heme and oxygen which are required for the pathway (see Discussion) [40]. Note that the gene groups shown in Figure 3 and Figure 3 showed significant gene expression changes in different sets of environmental conditions.

4. DISCUSSION

This work describes two principles for analyzing combinations of regulatory sequences. First, sequence conservation among closely related yeast species was used to find sequences that were more likely to be functionally important. Secondly, a template approach that considered joint positional distributions of word pairs increased the specificity of gene expression predictions using sequence-based rules. We have demonstrated that higher-order sequence features within TCR's were conserved across multiple *Saccharomyces* genomes. Closely spaced and jointly conserved word pairs were also more likely to be associated with gene expression changes. A large proportion of words contained in templates matched known transcription factor binding sites, and some of the uncharacterized words may represent novel regulatory sequences. In many cases, associations between templates and gene expression changes were significant in conditions when the corresponding transcription factors are known to be active. In addition, groups of genes that co-conserved both words in a template often were enriched for common functional roles. These results suggest that conserved word pair templates, which were discovered strictly based on higher-order properties of sequence conservation, also carry biological relevance.

Conserved word pair templates may be classified under several distinct classes of regulatory elements in TCR's. One possible interpretation of templates is that closely spaced sequence pairs may promote direct or indirect interactions between transcription factors by increasing the local concentrations of the individual factors. For example, the proximity of Cbf1p and Met31/32p binding sites may promote interaction between these factors in recruiting their common transcriptional activators, Met4 and Met28. Experimental studies on the TCR's of *MET3* and *MET28* have demonstrated that the binding of Cbf1p enhances the DNA binding affinity of Met31/32p [41]. Indeed, biochemical experiments suggest that all of these proteins interact at the TCR's of some sulfur utilization genes [41].

Another possible regulatory scheme for conserved, closely-spaced word pairs is that individual sequences found in templates may correspond to binding sites for transcription factors that bind independently under the same or separate conditions. The Hap1p and Hap2/3/4/5p transcription factors, whose binding sites were identified in a template, represent an example of multifactorial regulation in response to different environmental stimuli [42]. In some cases, templates could discern genes that shared binding sites for one transcription factor, but were differentially expressed under certain sets of conditions. Genes that conserved both Hap1p and Hap2/3/4/5p binding sites in their TCR's included genes encoding mitochondrial enzymes, as well as respiration proteins, that showed significant induction during growth in stationary phase and ethanol (Figure 3A). By contrast, genes that conserved both the Hap1p and Upc2p/Ecm22p binding sites in their TCR's were enriched for genes encoding ergosterol biosynthesis enzymes. Unlike the genes encoding mitochondrial enzymes, these genes showed no expression changes in response to different carbon sources, yet they were significantly induced under treatment with drugs that inhibit ergosterol biosynthesis: itraconazole, lovastatin and terbinafine (Figure 3B) [34]. A biochemical link between these two enzyme categories may explain their common regulation via Hap1p: the protein products of these genes all require the cofactor heme, whose intracellular levels are sensed by Hap1p [43] [44]. Our results suggest that Hap1p controls the expression of all of these genes in response to heme and/or oxygen levels. The expression of genes encoding mitochondrial and respiration enzymes may be controlled by Hap2/3/4p in response to nonfermentable carbon sources, whereas the expression of ergosterol biosynthesis genes may be regulated by Ecm22p and/or Upc2p in response to ergosterol levels. Whether these factors act cooperatively with, or independently of, Hap1p will require further biochemical investigation to elucidate.

Close spacing between word pairs may be important for reasons other than the promotion of transcription factor interactions. Different regions of TCR's at varying windows away from translation start sites may be more competent at recruiting or inhibiting RNA polymerase. These differences may be influenced by nucleosome accessibility, chromatin structure, or DNA physical properties, which can be correlated with local A/T content (see [45] for references). Notably, we have also found that the relative proportions of A and T nucleotides vary considerably within the 200 bp closest to translation start sites (A. M. Moses, M. B. Eisen and Audrey Gasch, unpublished results). Low-complexity words that contained 4 or more A's or T's could be found in many templates (denoted by TnC in Figure 4 and Table 1); these words may serve as surrogates for a distance window from translation start. Transcription factor binding sites that are closely spaced to these low-complexity words may be found in more transcriptionally competent regions of TCR's.

Since transcription factor binding sites often contain degenerate positions that reflect specificity, a key limitation of our approach is the use of exact words [11]. The known binding sites listed in Table 1 correspond to transcription factors with high sequence specificities. Since other known binding sites are poorly modeled by exact words (in that they bind sequences with relaxed specificity at certain positions in their binding sites), our method has failed to include them in conserved word pair templates. In addition, our method currently requires sequence identity for a word to be labeled as conserved. This strict requirement omits binding sites that may retain their function, despite mutations in degenerate positions that may have little impact on transcription factor binding.

The consideration of joint conservation and close spacing has provided insights into how TCR organization may influence the multifactorial regulation of gene expression in *Saccharomyces cerevisiae*. These criteria were motivated by experimental studies on the positional organization of individual binding sites within TCR's, with the hypothesis that this underlying architecture would be functionally conserved. Even more complicated higher-order sequence rules are apparent in the organization of cis-regulatory modules in *Drosophila melanogaster* [46]. Nevertheless, a common organizational theme of the TCR's in both of these organisms is the importance of relative spacing between transcription factor binding sites. The discovery of additional principles for TCR organization will further advance our understanding of how regulatory information is encoded in genome sequences.

5. ACKNOWLEDGEMENTS

We thank Audrey Gasch and Justin Fay for their critical reading of the manuscript and insightful suggestions during the development of this work; Peter Bickel and John Storey for their statistical advice; and Nabuo Ogawa and Patrick Brown for sharing gene expression data on cadmium conditions prior to publication. D.Y.C. is a Howard Hughes Medical Institute Predoctoral Fellow, and M.B.E. is a Pew Scholar in the Biomedical Sciences. This work was conducted under the US Department of Energy contract No. ED-AC03-76SF00098.

6. REFERENCES

[1] Composite regulatory elements: structure, function and classification. [http://www.gene-regulation.com/pub/databases/transcompel/compel.html]

[2] Wolberger C: Multiprotein-DNA complexes in transcriptional regulation. *Annu Rev Biophys Biomol Struct* 1999, 28:29-56.

[3] Kel OV, Romaschenko AG, Kel AE, Wingender E, Kolchanov NA: A compilation of composite regulatory elements affecting gene-transcription in vertebrates. *Nucleic Acids Res* 1995, 23:4097-4103.

[4] Gasch AP: The environmental stress response: a common yeast response to diverse environmental stresses. In: *Yeast Stress Responses* Edited by S Hohmann, WH Mager, vol. 1. pp. 11-70. Berlin: Springer; 2003: 11-70.

[5] Mead J, Bruning AR, Gill MK, Steiner AM, Acton TB, Vershon AK: Interactions of the Mcm1 MADS box protein with cofactors that regulate mating in yeast. *Mol Cell Biol* 2002, 22:4607-4621.

[6] Bhoite LT, Allen JM, Garcia E, Thomas LR, Gregory ID, Voth WP, Whelihan K, Rolfes RJ, Stillman DJ: Mutations in the Pho2 (Bas2) transcription factor that differentially affect activation with its partner proteins Bas1, Pho4, and Swi5. *J Biol Chem* 2002, 277:37612-37618.

[7] Verger A, Duterque-Coquillaud M: When Ets transcription factors meet their partners. *Bioessays* 2002, 24:362-370.

[8] Ambrosetti DC, Basilico C, Dailey L: Synergistic activation of the fibroblast growth factor 4 enhancer by Sox2 and Oct-3 depends on protein-protein interactions facilitated by a specific spatial arrangement of factor binding sites. *Mol Cell Biol* 1997, 17:6321-6329.

[9] Ludwig MZ, Patel NH, Kreitman M: Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. *Development* 1998, 125:949-958.

[10] Liu Z, Little JW: The spacing between binding sites controls the mode of cooperative DNA-protein interactions: implications for evolution of regulatory circuitry. *J Mol Biol* 1998, 278:331-338.

[11] Stormo GD: DNA binding sites: representation and discovery. *Bioinformatics* 2000, 16:16-23.

[12] Duret L, Bucher P: Searching for regulatory elements in human noncoding sequences. *Curr Opin Struct Biol* 1997, 7:399-406.

[13] Pennacchio LA, Rubin EM: Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet* 2001, 2:100-109.

[14] Blanchette M, Tompa M: Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res*. 2002, 12:739-748.

[15] Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM: rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res* 2002, 12:832-839.

[16] Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W: PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res* 2000, 10:577-586.

[17] Pilpel Y, Sudarsanam P, Church GM: Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.* 2001, 29:153-159.

[18] Bussemaker HJ, Li H, Siggia ED: Regulatory element detection using correlation with expression. *Nat. Genet.* 2001, 27:167-171.

[19] Keles S, van der Laan M, Eisen MB: Identification of regulatory elements using a feature selection method. *Bioinformatics* 2002, 18:1167-1175.

[20] Wang W, Cherry JM, Botstein D, Li H: A systematic approach to reconstructing transcription networks in Saccharomyces cerevisiae. *Proc Natl Acad Sci USA* 2002, 99:16893-16898.

[21] Wagner A: Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics* 1999, 15:776-784.

[22] Klingenhoff A, Frech K, Quandt K, Werner T: Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics* 1999, 15:180-186.

[23] Pavlidis P, Furey TS, Liberto M, Haussler D, Grundy WN: Promoter region-based classification of genes. *Proceedings of the Pacific Symposium on Biocomputing* 2001, 6:151-164.

[24] Kel-Margoulis OV, Ivanova TG, Wingender E, Kel AE: Automatic annotation of genomic regulatory sequences by searching for composite clusters. *Pac Symp Biocomput* 2002, 7:187-198.

[25] Kamvysselis M, Patterson N, Birren B, Berger B, Lander ES: Whole-genome comparative annotation and regulatory motif discovery in multiple yeast species. *Proceedings of the 7th International Conference on Research in Computational Molecular Biology 2003*, 7.

[26] Chiang DY, Brown PO, Eisen MB: Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles. *Bioinformatics* 2001, 17 Suppl 1:S49-S55.

[27] DeRisi JL, Iyer VR, Brown PO: Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 1997, 278:680-686.

[28] Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* 2000, 11:4241-4257.

[29] Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* 1998, 9:3273-3297.

[30] Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, et al: A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell.* 1998, 2:65-73.

[31] Gasch AP, Huang M, Metzner S, Botstein D, Elledge SJ, Brown PO: Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Mol Biol Cell* 2001, 12:2987-3003.

[32] Lee SE, Pellecioli A, Demeter J, Vaze MP, Gasch AP, Malkova A, Brown PO, Botstein D, Stearns T, Foiani M, et al: In: *Biological Responses to DNA Damage*, vol. 65. pp. 303-314. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2000: 303-314.

[33] Ogawa N, DeRisi J, Brown PO: New components of a system for phosphate accumulation and polyphosphate metabolism in Saccharomyces cerevisiae revealed by genomic expression analysis. *Mol. Biol. Cell* 2000, 11:4309-4321.

[34] Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai HY, He YDD, et al: Functional discovery via a compendium of expression profiles. *Cell* 2000, 102:109-126.

[35] Zhu J, Zhang MQ: SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics* 1999, 15:607-611.

[36] Ho Y, Costanzo M, Moore L, Kobayashi R, Andrews BJ: Regulation of transcription at the Saccharomyces cerevisiae start transition by Stb1, a Swi6-binding protein. *Mol Cell Biol* 1999, 19:5267-5278.

[37] Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, et al: Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 2002, 298:799-804.

[38] Gancedo JM: Yeast carbon catabolite repression. *Microbiol Mol Biol Rev* 1998, 62:334-361.

[39] Kwast KE, Burke PV, Poyton RO: Oxygen sensing and the transcriptional regulation of oxygen-responsive genes in yeast. *J Exp Biol* 1998, 201:1177-1195.

[40] Vik A, Rine J: Upc2p and Ecm22p, dual regulators of sterol biosynthesis in Saccharomyces cerevisiae. *Mol Cell Biol* 2001, 21:6395-6405.

[41] Blaiseau PL, Thomas D: Multiple transcriptional activation complexes tether the yeast activator Met4 to DNA. *EMBO J* 1998, 17:6327-6336.

[42] Guarente L, Lalonde B, Gifford P, Alani E: Distinctly regulated tandem upstream activation sites mediate catabolite repression of the CYC1 gene of *S. cerevisiae*. *Cell* 1984, 36:503-511.

[43] Burke PV, Poyton RO: Structure/function of oxygen-regulated isoforms in cytochrome c oxidase. *J Exp Biol* 1998, 201:1163-1175.

[44] Lees ND, Skaggs B, Kirsch DR, Bard M: Cloning of the late genes in the ergosterol biosynthetic pathway of Saccharomyces cerevisiae--a review. *Lipids* 1995, 30:221-226.

[45] Liao GC, Rehm EJ, Rubin GM: Insertion site preferences of the P transposable element in *Drosophila melanogaster*. *Proc Natl Acad Sci USA* 2000, 97:3347-3351.

[46] Berman BP, Nibu Y, Pfeiffer BD, Tomancek P, Celniker SE, Levine M, Rubin GM, Eisen MB: Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl. Acad. Sci. USA* 2002, 99:757-762.

[47] Cliften PF, Hillier LW, Fulton L, Graves T, Miner T, Gish WR, Waterston RH, Johnston M: Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res*. 2001, 11:1175-1186.

[48] Saccharomyces Genome Database. [http://genome-www.stanford.edu/Saccharomyces/]

[49] Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA, et al: The Stanford Microarray Database. *Nucleic Acids Res* 2001, 29:152-155.

[50] Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* 1995, 57:289-300.

[51] Press WH, Teukolsky SA, Vertterling WT, Flannery BP: *Numerical Recipes in C*, Second Edition. Cambridge: Cambridge University Press; 1992.

Network Motif Discovery Using Subgraph Enumeration and Symmetry-Breaking

Joshua A. Grochow and Manolis Kellis

Computer Science and AI Laboratory, M.I.T.
Broad Institute of M.I.T. and Harvard
joshuag@cs.uchicago.edu (current affiliation), manoli@mit.edu

Abstract. The study of biological networks and network motifs can yield significant new insights into systems biology. Previous methods of discovering network motifs – network-centric subgraph enumeration and sampling – have been limited to motifs of 6 to 8 nodes, revealing only the smallest network components. New methods are necessary to identify larger network sub-structures and functional motifs.
Here we present a novel algorithm for discovering large network motifs that achieves these goals, based on a novel symmetry-breaking technique, which eliminates repeated isomorphism testing, leading to an exponential speed-up over previous methods. This technique is made possible by reversing the traditional network-based search at the heart of the algorithm to a motif-based search, which also eliminates the need to store all motifs of a given size and enables parallelization and scaling. Additionally, our method enables us to study the clustering properties of discovered motifs, revealing even larger network elements.
We apply this algorithm to the protein-protein interaction network and transcription regulatory network of *S. cerevisiae*, and discover several large network motifs, which were previously inaccessible to existing methods, including a 29-node cluster of 15-node motifs corresponding to the key transcription machinery of *S. cerevisiae*.

1 Introduction

1.1 Network Motifs

In the past decade, new technologies have enabled the observation and study of networks of thousands and millions of nodes, such as social networks, computer networks, and, notably, *biological networks*, including protein-protein interaction networks [4–6], genetic regulatory networks [12, 18], and metabolic networks [7]. In order to extract meaningful information from these vast and sometimes noisy datasets, it is necessary to develop methods of computational analysis that are both efficient and robust to errors in the underlying data.

Network motifs – patterns of connectivity that occur significantly more frequently than expected – were introduced by Milo *et al.* [18] and provide one such robust property of biological networks. Network motifs also provide an important tool for understanding the modularity and the large-scale structure of networks

[8, 13, 20, 25]. The importance of network motifs as information-processing modules has been modeled theoretically [12, 21] and verified experimentally [8, 13, 20, 25]. Network motifs also have numerous other applications: they have been used to classify networks into "superfamilies" [17], they have been used in combination with machine learning techniques to determine the most appropriate network model for a given real-world network [16], and they have been used to determine which properties to use in parsimony models of phylogeny [19].

Unfortunately, all of these applications are hampered by the limited size of motifs discoverable by current methods. Exact counting methods have only been reported to find motifs up to 4 nodes [18] and motif generalizations up to 6 nodes [10]. Subgraph sampling methods have found motifs up to 7 [9] and 8 nodes [1, 16]. The statistical measures developed by Ziv *et al.* [26] are an important step towards larger network structures, but unfortunately lack a one-to-one correspondence with subgraphs, making them potentially difficult to interpret. Motif generalizations [10] are another important step towards these goals, although current methods are still limited to finding motif generalizations of only 6 nodes.

This current size limitation leaves many fundamental questions unanswered, and significant additional insight could be gained by exploring larger subgraphs and finding larger motifs. [1, 10]. We should not expect *a priori* that the building blocks of complex networks are as small as 4 nodes, or that the largest significant structures and pathways contain only 8 nodes. What are the fundamental building blocks? How do they combine to form larger structures? [1, 10] Do networks which share the same building blocks also share the same combinations of these blocks? [10] How can larger structures be used to distinguish between networks of different types, or between proposed models for a given network? [1]

In this paper, we present a new approach for discovering network motifs. The heart of our algorithm exhaustively assesses the significance of **a single query subgraph** as a potential motif. This can then be applied to all subgraphs of a given size to emulate the behavior of previous exhaustive algorithms, but with an exponential speed-up due to a **novel symmetry-breaking technique** (which is not feasible with previous methods). The symmetry-breaking technique also allows us to write instances of a subgraph to disk as they are found, **further eliminating limitations due to memory usage**. We are thus able to find motifs of up to 15 nodes, to find all instances of subgraphs of 31 nodes, and potentially even larger subgraphs. Although this work is motivated by biological networks, and this paper focuses on the protein-protein interaction (PPI) network and the transcription network of *S. cerevisiae*, our methods are applicable to any network – directed or undirected – and thus to many different fields, even outside the realm of biology.

In this section, we review previous work and give an overview of our algorithm, outlining several novel techniques which apply both to our approach and to previous approaches. In Sec. 2 we present our algorithm in detail. In Sec. 3 we present benchmarks comparing our approach to previous approaches. Additionally, we present data as to the effectiveness of the resulting improvements as

applied to both the transcription and PPI networks of *S. cerevisiae*. In Sec. 4, we present some of the larger subgraphs we have discovered. Finally, in Sec. 5 we discuss the significance of these contributions for the understanding of networks in general.

1.2 Limitations of Network-Centric Approaches for Motif Discovery

Two basic techniques have been proposed for identifying network motifs: exact counting [18] and subgraph sampling [1, 9, 16]. These methods attempt to determine the significance of all or many subgraphs of a given size by comparing their frequency in a given network to their frequency in a random ensemble of networks with similar properties to the original. To determine which subgraphs are motifs, subgraph sampling [1] is an effective and efficient approach, and has been used to evaluate the significance of larger subgraphs than can be evaluated by the exact counting method.

Most methods for finding DNA *sequence motifs* scan or sample a sequence pattern from a genome. Similarly, previous techniques for finding network motifs scan or sample subgraphs from a network, and count the number of occurrences of each subgraph encountered. (This process is then repeated for each network in a random ensemble resembling the initial network, and the counts are compared.) For the discovery of DNA sequence motifs, this general methodology is very efficient, because sequence motifs can be efficiently hashed based on their content. Thus a single linear scan of the genome suffices to exhaustively count all possible substrings of a given size, *regardless of the size of the substrings*.

In contrast, for the discovery of network motifs, enumerating all subgraphs of a given size is in general *exponential in the number of nodes of the subgraphs*. Additionally, there is no known efficient algorithm that correctly identifies two graphs as isomorphic or not. (The *graph isomorphism problem* is not known to be either in P or to be NP-complete.) This intrinsic difference in complexity between discovering *sequence* motifs and discovering *network* motifs makes traditional network-scanning methodologies inefficient for network motif discovery.

1.3 Distinguishing Features of the New Algorithm

To avoid these limitations of the traditional network-centric approaches, we have taken a motif-centric approach which has several attractive features, outlined here. Features 1-3 are specific to motif-centric methods, while features 4 and 5 can also benefit traditional network-centric methods.

(1) Searching for a single query graph. To avoid the increased complexity of subgraph enumeration (in the absence of an appropriate hashing scheme) our algorithm works by exhaustively searching for the instances of a *single query graph* in a network. (To find all motifs of a given size we couple this search with subgraph enumeration, using McKay's `geng` and `directg` tools [15]). Even though the *subgraph isomorphism problem* – finding a given graph as a subgraph of a larger network – is known to be NP-complete, several algorithmic improvements

enable this search to be carried out effectively in practice, even for subgraphs up to 31 nodes (and potentially even more).

(2) Mapping instead of enumerating. Rather than enumerating all connected subgraphs of a given size and testing to see whether each is isomorphic to the query graph, our algorithm attempts to map the query graph onto the network in all possible ways. We developed this technique for subgraph isomorphism independently, but subsequently identified a prior use [23].

(3) Taking advantage of subgraph symmetries. We introduce a technique that *avoids spending time finding a subgraph more than once* due to its symmetries. This technique improves the speed of our method by a factor exponential in the size of the query subgraph (Table 1). Moreover, since each instance is discovered exactly once, our algorithm can write instances to disk as they are found, greatly improving memory usage.

(4) Improved isomorphism testing. Our isomorphism test takes into account the degree of each node, and the degrees of each node's neighbors, leading to marked improvements over current motif-finding algorithms, which use exhaustive graph isomorphism tests.

(5) Subgraph hashing. When examining all subgraphs of a given size we hash the graphs based on their degree sequences, which leads to a significant improvement in the number of isomorphism tests needed. In a *directed* network, we group the query graphs based on their *undirected* isomorphism types, find all instances, and then go back to the directed network and divide these instances into their directed isomorphism types.

2 Description of the Algorithm

For clarity, we first present the basic mapping algorithm for subgraph isomorphism, without taking into account the symmetries of the query graph. In Sec. 2.2 we incorporate our symmetry-breaking technique into the algorithm. In the pseudo-code, we identify statements used solely for symmetry-breaking by enclosing them in square brackets. Finally, In Sec. 2.3 we incorporate our technique into two new methods of finding motifs.

Throughout this section, G will denote the network being searched and H will denote the query subgraph. We say that a node g of G can *support* a node h of H if we cannot rule out a subgraph isomorphism from H into G which maps h to g based on the degrees of h and g and the degrees of their neighbors. (Other constraints could also be used here, but these two proved effective and simple to implement.) This notion of support is used to exclude inconsistent candidate maps during the backtracking search.

2.1 Finding a Given Subgraph (Subgraph Isomorphism)

We start by presenting the algorithm without symmetry-breaking. Note that symmetry-breaking is not required for correctness of the algorithm.

FindSubgraphInstances(H,G):
Finds all instances of query graph H in network G

Start with an empty set of instances.
[Find Aut(H). Let H_E be the equivalence representatives of H.]
[Find symmetry-breaking conditions C for H given H_E and Aut(H).]
Order the nodes of G by increasing degree and
 then by increasing neighbor degree sequence.
For each node g of G
 For each node h of H [H_E] such that g can support h
  Let f be the partial map associating $f(h) = g$.
  Find all isomorphic extensions of f [up to symmetry]
   i.e. call IsomorphicExtensions(f,H,G[,$C(h)$]).
  Add the images of these maps to the set of all instances.
 Remove g from G.
Return the set of all instances.

FindSubgraphInstances includes the *images* of the maps in the list of instances, thus merging all maps which differ only by a symmetry of H. (Without symmetry-breaking, the algorithm spends additional time finding several distinct maps to a single subgraph.)

IsomorphicExtensions is a backtracking search to find all isomorphisms from H into G. As is standard in backtracking searches, the algorithm uses the most constrained neighbor to eliminate maps that cannot be isomorphisms: that is, the neighbor of the already-mapped nodes which is likely to have the fewest possible nodes it can be mapped to. First we select the nodes with the most already-mapped neighbors, and amongst those we select the nodes with the highest degree and largest neighbor degree sequence.

For each call to IsomorphicExtensions, f is extended by a single node. Each time an extension is made, the algorithm ensures that the newly mapped node is appropriately connected to the already-mapped nodes. Thus when IsomorphicExtensions returns a map, it is guaranteed to be an isomorphism.

We have effectively pushed the isomorphism testing of previous exhaustive methods into IsomorphicExtensions, which allows the isomorphism test to abort early. The ability to abort early when finding instances of a particular query graph presents significant savings over previous methods.

2.2 Exploiting Subgraph Symmetries to Speed Up the Search

Due to symmetries, a given subgraph of G may be mapped to a given query graph H multiple times. For example, the subgraph in Fig. 1 can be mapped to the same 6 nodes in 8 different ways. Thus a simple mapping-based search for a query graph will find each instance of the query graph as many times as the graph has symmetries. To avoid this, we compute and enforce several symmetry-breaking conditions, which ensure that there is a *unique* map from the query graph H to each instance of H in G, so that our search only spends time finding each instance once.

IsomorphicExtensions**(f,H,G[,C(h)]:**
Finds all isomorphic extensions of partial map $f : H \to G$ [satisfying C(h)]

Start with an empty list of isomorphisms.
Let D be the domain of f.
If $D = H$, return a list consisting solely of f. (Or write to disk.)
Let m be the most constrained neighbor of any $d \in D$
(constrained by degree, neighbors mapped, etc.)
For each neighbor n of $f(D)$
If there is a neighbor $d \in D$ of m such that n is *not* neighbors with $f(d)$,
or if there is a *non*-neighbor $d \in D$ of m such that n *is* neighbors with $f(d)$
[or if assigning $f(m) = n$ would violate a symmetry-breaking condition in $C(h)$],
then continue with the next n.
Otherwise, let $f' = f$ on D, and $f'(m) = n$.
Find all isomorphic extensions of f'.
Append these maps to the list of isomorphisms.
Return the list of isomorphisms.

The symmetries of a graph H are known as automorphisms (self-isomorphisms), and the group of automorphisms of H is denoted $\mathrm{Aut}(H)$. For a set A of automorphisms, two nodes are said to be "A-equivalent" if there is some automorphism in A which maps one to the other, or simply "equivalent" if $A = \mathrm{Aut}(H)$. We denote the A-equivalence of two nodes n_1, n_2 by $n_1 \sim_A n_2$. This equivalence relation partitions the nodes of H into equivalence classes. Since starting a map from two equivalent nodes is unnecessary and wasteful, FindSubgraphInstances uses a set consisting of one representative from each equivalence class of H.

The symmetry-breaking conditions are based on labellings of the nodes of H by integers, represented as maps from $H \to \mathbf{Z}$. Let $\ell : G \to \mathbf{Z}$ be a labelling of the nodes of G by *distinct* integers. Then each map $f : H \to G$ generates a labelling $L : H \to \mathbf{Z}$, given by $L(n) = \ell(f(n))$ for nodes $n \in H$. Thus, conditions on labellings of H translate into restraints on maps from H into G.

Given a set of conditions C, we say an automorphism α *preserves the conditions* C if, given a labelling L_1 of H which satisfies C, the corresponding labelling $L_2 : H \to \mathbf{Z}$ given by $L_2(n) = L_1(\alpha(n))$ also satisfies C. We are thus searching for conditions C such that the only automorphism which preserves C is the identity. This ensures there will be exactly one map from H onto each of its instances in G which satisfies the conditions.

To find these conditions, we pick an $\mathrm{Aut}(H)$-equivalence class $\{n_0, \ldots, n_k\}$ of nodes of H, and we impose the condition $L(n_0) < \mathrm{MIN}(L(n_1), \ldots, L(n_k))$. Any automorphism must send n_0 to one of the n_i, since these are all of the nodes equivalent to n_0. But to preserve this condition, an automorphism must send n_0 to itself. Then we continue recursively, replacing $\mathrm{Aut}(H)$ with the set A of automorphisms which send n_0 to itself. For example, see Fig. 1.

Because FindSubgraphInstances starts with a particular node, we can consider that node already fixed. (Note that the version of FindSubgraphInstances which uses symmetry-breaking only iterates over a set of equivalence

class representatives, and not over all nodes of H.) Thus for each representative used by FindSubgraphInstances, SymmetryConditions must generate a series of symmetry-breaking conditions which start by fixing that node.

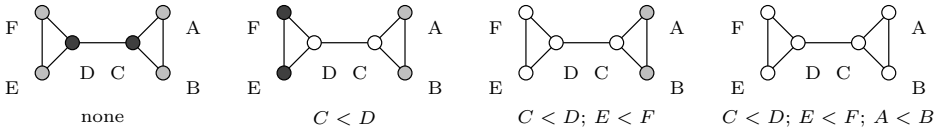


Fig. 1. Finding conditions that will break all the symmetries of a 6-node graph. White nodes are fixed by any automorphism preserving the indicated conditions, and other nodes are shaded according to their equivalence class under the automorphisms which preserve the indicated conditions.

To find the automorphisms of H, we use IsomorphicExtensions *without* symmetry-breaking, which returns an exhaustive list of all isomorphisms from H to itself. To find the automorphisms which fix a node or a set of nodes, the algorithm filters this list in a single pass.

Finding the automorphisms of a graph is thought to be computationally expensive[1], but in practice we have found this is far from the bottleneck in motif-finding algorithms. We were able to *exhaustively* find the automorphisms of all 11,117 8-node undirected graphs in under 30 seconds on a standard laptop, and McKay's tools [14] can find all the automorphisms of very large graphs very rapidly (e.g. some graphs with thousands of nodes and millions of automorphisms, in less than one second on a standard laptop).

SymmetryConditions:
Finds symmetry-breaking conditions for H given H_E, $\mathbf{Aut}(H)$

Let M be an empty map from equivalence representatives to sets of conditions.
For each $n \in H_E$
 Let C be an empty set of conditions.
 $n' \leftarrow n$, and $A \leftarrow \mathrm{Aut}(H)$.
 Do until $|A| = 1$:
  Add "$\textsc{label}(n') < \textsc{min}\{\textsc{label}(m) | m \sim_A n' \text{ and } m \neq n'\}$" to C.
  $A \leftarrow \{f \in A | f(n') = n'\}$.
  Find the largest A-equivalence class E.
  Pick $n' \in E$ arbitrarily.
 Let $M(n) = C$.
Return M.

[1] Finding graph automorphisms is at least as hard as determining if two graphs are isomorphic. Like the graph isomorphism problem, the graph automorphism problem is not known to be either in P or to be NP-complete.

2.3 Subgraph Evaluation and Network Motif Discovery

To find network motifs we enumerate candidate subgraphs H (exhaustively or by sampling), and evaluate each candidate based on its instances.

Evaluating candidate subgraphs. We find all instances of a query graph H in the network G, as well as in a random ensemble of networks with the same degree distribution and same distribution of 3-node subgraphs as G.[2] We evaluate the overrepresentation of the query graph H based on the z-score of its abundance in G against the distribution of its abundance in the random ensemble, as in [18, 21].

Exhaustive subgraph enumeration. Our method can be used to find all instances of subgraphs of a given size, similar to previous exhaustive methods. To do this, we generate all non-isomorphic graphs of a particular size using McKay's `geng` and `directg` tools [15]. Then for each graph, we evaluate its significance as above.

Subgraph sampling. Our method can also be used in combination with subgraph sampling. We sample connected subgraphs (usually relatively large, compared to previous network motifs: 10, 15, or 20 nodes) by picking a node at random, and taking a random walk until we have as many nodes as desired [16]. Then we assess the significance of this subgraph as above.

Sampling subgraphs to find anti-motifs. Some studies have also considered *anti*-motifs: subgraphs which are significantly *under*represented compared to randomized versions of the network. To use a sampling method to find anti-motifs, it might be more fruitful to sample initial subgraphs from the random ensemble rather than the network being studied. Anti-motifs will be more prevalent in the ensemble than in the target network, and thus are more likely to be discovered by sampling from the ensemble.

3 Results and Evaluation

We applied our algorithm to the PPI network (1379 nodes, 2493 edges) [4] and transcription network (685 nodes, 1052 edges) [2] of *S. cerevisiae* and compared its performance to previous methods of motif disccovery.

Comparison with previous methods: time. We compare the time requirements of our method to those of Milo *et al.* [18] (Fig. 2). We make this comparison on the undirected PPI network of *S. cerevisiae* [4], by exhaustively counting subgraphs up to 7 nodes.

We implemented both our algorithm and two versions of the Milo *et al.* algorithm [18]: both as originally presented [18], and also by additionally hashing subgraphs by their degree sequence (Sec. 1.3). Fig. 2 shows that our algorithm

[2] Although Shen-Orr *et al.*[21] use a model in which the distribution of $(n-1)$-node subgraphs is preserved when looking for n-node motifs, they only applied this to the case $n = 4$, and we have found it computationally infeasible to preserve this distribution for $n > 4$. Nonetheless, we have found it fruitful to preserve the distribution of 3-node subgraphs, regardless of n.

provides an *exponential* improvement in time, even compared to the modified version of the previous algorithm [18].

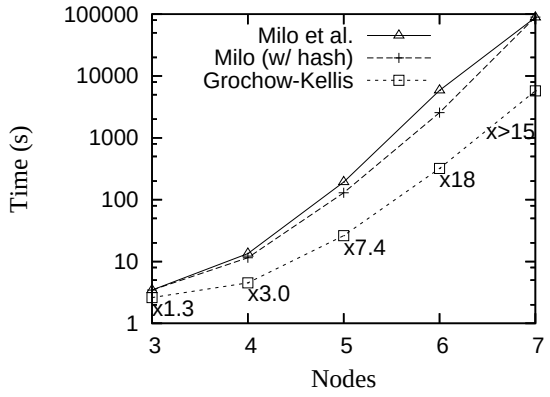


Fig. 2. The runtimes of the original algorithm of Milo *et al.* [18], an improved version of their algorithm, and our new algorithm, as applied to the undirected PPI network of *S. cerevisiae* [4]. The speed-up from the original algorithm of Milo *et al.* [18] to our algorithm is indicated. (Note: the values for 7 nodes for the two variants of Milo *et al.*'s algorithm are underestimates: the program ran out of memory before finishing.)

Comparison with previous methods: space. Our method gains an exponential memory advantage over previous exhaustive methods by not keeping a list of previously visited subgraphs. In the previous exact counting method [18], a list of the subgraphs encountered at each node is necessary in order to avoid duplication, even when the instances of the subgraphs are not desired as output. Thus the space required by the previous method is proportional to the number of subgraphs of a given size going through a given node, which can be exponential in the size of the subgraphs. Because our method does not need to keep such a list, its asymptotic memory requirements are determined by the maximum depth of recurion of IsomorphicExtensions, which is linear in the size of the query graph. Our method thus uses exponentially less space than previous exhaustive methods.

Disk usage. Furthermore, our algorithm uses no more memory to find a list of all instances than to simply count the instances. Since each instance is encountered exactly once, it can be written to disk and removed from active memory as soon as it is encountered, using effectively no additional memory.

Improvement due to symmetry-breaking. The main reason for these improvements is our novel symmetry-breaking technique. Symmetry-breaking ensures that each instance is discovered exactly once, so our method does not have to check a list of the subgraphs previously encountered at a node in order to avoid duplicate counting, while the previous method of exact counting does. Ta-

ble 1 quantifies this contribution explicitly, showing that the average number of automorphisms of graphs weighted by their occurences in the PPI network and regulatory network of yeast – i.e. the savings gained by symmetry-breaking – appears to grow exponentially.

Table 1. The number of subgraphs encountered by our algorithm with and without symmetry-breaking (including multiple encounters for the version without symmetry-breaking). The improvement factor is exactly the average number of automorphisms of subgraphs of the associated size.

	Undirected PPI Network			Directed Regulatory Network		
Nodes	Total Subgraphs Searched	With Symmetry-Breaking	Improvement	Total Subgraphs Searched	With Symmetry-Breaking	Improvement
3	3.7×10^4	1.1×10^4	$\times3.13$	2.6×10^4	1.3×10^4	$\times2.02$
4	4.0×10^5	7.0×10^4	$\times5.77$	9.7×10^5	1.8×10^5	$\times5.41$
5	4.4×10^6	4.1×10^5	$\times10.9$	4.4×10^7	2.5×10^6	$\times18.0$
6	5.1×10^7	2.3×10^6	$\times22.2$	2.3×10^9	3.2×10^7	$\times73.3$
7	5.7×10^8	1.2×10^7	$\times46.3$	1.3×10^{11}	4.0×10^8	$\times334$
8	6.4×10^9	6.6×10^7	$\times96.2$	—	—	—

4 Discovered Motifs and Their Biological Significance

Discovered motifs. We exhaustively evaluated all candidate motifs and anti-motifs up to 7 nodes in the PPI network of *S. cerevisiae*[4] (1379 nodes, 2493 edges). We used a random ensemble of networks with the same degree distribution and the same distribution of 3-node subgraphs as the PPI network.[3] The most significant subgraphs tend to be motifs rather than anti-motifs: using a z-score cutoff of 4.0, only 3 of the 54 significant subgraphs of size at most 7 were anti-motifs. Two of the motifs were trees, and the most dense motif had 18 edges. Most of the significant graphs were of moderate density: the mean number of edges for 7-node motifs and anti-motifs is 11.49 ± 2.89.

Large discovered motifs. We also discovered larger motifs by first sampling connected subgraphs from the PPI network of *S. cerevisiae*, and then assessing their significance using our method. We sampled approximately 100 connected subgraphs of 15 and 20 nodes, and found several motifs. One such 15-node motif (Fig. 3) represents a common connectivity pattern found within the transcriptional machinery of *S. cerevisiae* (see discussion below).

Clustering of discovered motifs and larger network structures. We noted that almost all of the larger subgraphs we evaluated have large numbers of overlapping instances, which become apparent since our method reports all network instances of a discovered motif. To quantify this property, we developed

[3] See footnote 2.

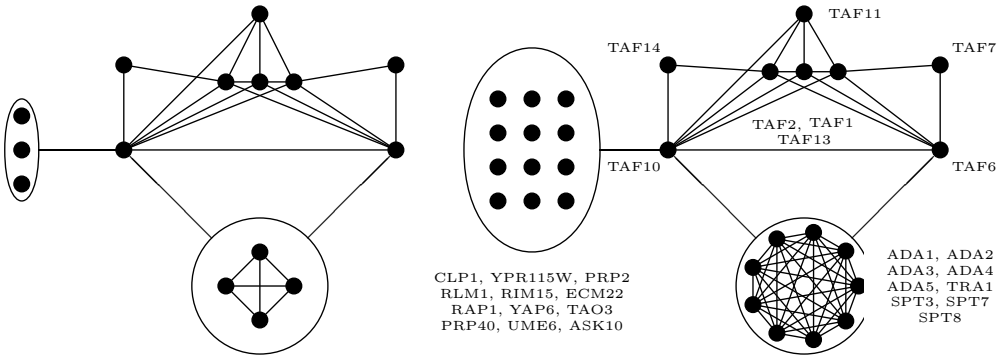


Fig. 3. A motif of 15 nodes and 34 edges (left). An edge from a group of nodes to a node n indicates that each node in the group is connected to n. This motif appears 27,720 times in the PPI network of *S. cerevisiae*[4], and does not appear at all in random ensembles which preserve the degree distribution and the distribution of 3-node subgraphs. All 27,720 instances are clustered into a total of 29 nodes (right), corresponding to the cellular transcription machinery.

a subgraph clustering score, based on the number of subgraph instances overlapping a given node, averaged over all nodes in any subgraph instance. We applied this score to evaluate the clustering properties of all discovered motifs, and we found that indeed some of the most abundant motifs show striking clustering properties.

The clustered instances frequently reveal important larger network structures. For example, the 15-node motif of Fig. 3 occurs 27,720 times in a single sub-network of 29 nodes, part of the core transcription machinery of *S. cerevisiae*. This includes a complete 11-node graph (including the two central hubs) corresponding to the SAGA complex, and consisting almost entirely of chromatin modification and histone acetylation factors an 8-node core (shared by all instances of the 15-node motif) corresponding to the TFIID complex, and 12 attachments, which are known activators and suppressors of these two complexes [11]. Similarly, the subgraph of 20 nodes shown in Fig. 4 occurs 5,020 times in a total of 31 nodes, enriched in cell-cycle regulation.

The role of combinatorial effects. The extreme clustering properties of the most abundant motifs appear to result from combinatorial connectivity patterns prevalent in larger network structures. For example, all 27,720 instances of the 15-node motif in Fig. 3 result by choosing 3 attachments from the left and 4 attachments from the the bottom of Fig. 3 ($\binom{12}{3}\binom{9}{4} = 27,720$), and similarly for the 5,020 instances of the 20-node subgraph in Fig. 4. Additionally, in the random ensemble, these combinatorially appearing structures occur either thousands of times, or not at all – they almost never occur just a few or a few hundred times.

Although motif clustering has previously been observed [3] and demonstrated analytically [24], previous motifs studied do not have enough nodes to exhibit the extreme combinatorial clustering we observed for large subgraphs (at least 15 nodes). The magnitude of this combinatorial clustering effect brings into question

the current definition of network motif, when applied to larger structures. We propose that additional statistics, either alone or in combination, might be well-suited to identify larger meaningful network structures: our subgraph clustering score, the total number of nodes covered by all instances of the query graph, the total number of edges, and the weighting of the number of nodes/edges based on the number of overlapping instances. All of these statistics can be easily calculated using our algorithm, since it finds and stores all motif instances, and these will be the subject of future studies.

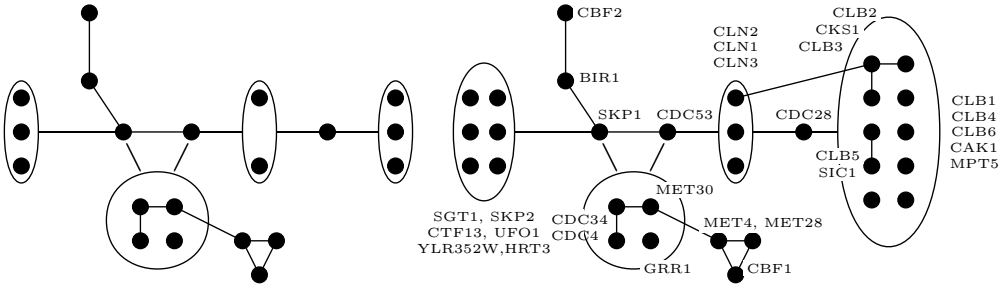


Fig. 4. A subgraph of 20 nodes and 27 edges (left). An edge from a group of nodes to a node n indicates that each node in the group is connected to n. This subgraph appears 5,020 times in the PPI network of *S. cerevisiae* [4]. All 5,020 instances are clustered into a total of 31 nodes (right), enriched in cell-cycle regulation.

5 Discussion

We presented a novel approach to the discovery of network motifs, based on a solution to the subgraph isomorphism problem that uses a new symmetry-breaking technique, an improved isomorphism test, and hashing based on degree sequences. Several of the techniques presented in our algorithm can also be used in previous algorithms, and lead to significant improvements.

We implemented our algorithm and used it to find significant structures of 15 and 20 nodes in the PPI network and the regulatory network of *S. cerevisiae*, where previous methods had been limited to motifs of 6 and 8 nodes. Using our approach to motif-finding, we re-discovered the cellular transcription machinery – as a 29-node cluster of 15-node motifs – based solely on the structure of the protein interaction network.

Previous methods of motif discovery were network-centric, and could therefore not take advantage of subgraph symmetries. By using a motif-centric algorithm instead, we are able to use symmetry-breaking to get an exponential improvement.

5.1 Applications and Advantages of the New Method

(1) Finding larger motifs. Our improvements have enabled the exhaustive discovery of motifs up to 7 nodes. To find even larger motifs, we sample a connected subgraph as in [16], and then find *all* its instances and assess its significance using our method. This technique has enabled us to find motifs up to 15 nodes and examine subgraphs up to 31 nodes.

(2) Querying a particular subgraph. Our method can be used to query whether a particular subgraph is significant, whereas previous methods can only do this by examining all subgraphs of the same size, which quickly becomes prohibitive for even moderate sizes. This technique could be used to explore *in silico* the prevalence of a subgraph of interest, identified experimentally (e.g. known pathways), computationally (e.g. motif generalizations [10]), or genetically.

(3) Exploring motif clustering. Because our algorithm finds all instances of a given subgraph, it can be used to explore how these instances cluster together to form larger structures. For example, after finding a 15-node motif, we were able to determine that all of its 27,720 instances clustered in 29 nodes (Fig. 3).

(4) Time and space. Our method, applied to all subgraphs of a given size, takes exponentially less time than previous methods, even when we implement the previous method with our hashing scheme (Sec. 3). Additionally, there are essentially no space limitations on our method: since each instance is found exactly once due to our symmetry-breaking technique, it can be written to disk and removed from active memory as soon as it is found.

(5) Parallelization. Our method is more easily parallelizable than previous motif-finding methods, since each subgraph can be counted on a separate processor. We have found this attribute to be very useful, and we believe other researchers will as well, as cluster computing becomes commonplace in the computational biology community.

5.2 Clustering Properties of Large Subgraphs and Motifs

We revealed that larger subgraphs tend to cluster together *combinatorially* – that is, all instances share a significant core of nodes, and each instance represents a choice of attachments to these core nodes. This combinatorial clustering brings into question the relevance of the standard definition of network motif for large subgraphs of 15 nodes or more. We proposed several statistics which may be more appropriate in this domain.

Finally, we mention that the statistics of Ziv *et al.* [26] may not suffer from these combinatorial effects. The main drawback of these statistics is their lack of one-to-one correspondence with subgraphs. In combination with our algorithm, however, the large subgraphs encompassed by these statistics could be further explored, allowing for a clearer interpretation of the most significant statistics.

Moving forward, we expect the network motifs and methodology presented here will open a window into the large structures and global organization of biological and other networks.

Acknowledgements. The authors would like to thank Pouya Kheradpour, Mike Lin, Matt Rasmussen, Alex Stark, and Radek Szklarczyk (all at the M.I.T. Computer Science and AI Laboratory) for many useful and interesting discussions.

All algorithms were implemented in Java using the Java Universal Networks and Graphs (JUNG) framework [22]. Our software is available on request. McKay's `geng` and `directg` tools [15] were used to enumerate all graphs of a given size. Much of the computational work was carried out on the compute cluster at the Broad Institute of M.I.T. and Harvard.

This work was supported in part by startup funds from Professor Kellis.

References

1. K. Baskerville and M. Paczuski. Subgraph ensembles and motif discovery using a new heuristic for graph isomorphism, 2006. arxiv.org:q-bio/0606023.
2. M. C. Costanzo, M. E. Crawford, J. E. Hirschman, J. E. Kranz, P. Olsen, L. S. Robertson, M. S. Skrzypek, B. R. Braun, K. L. Hopkins, P. Kondu, C. Lengieza, J. E. Lew-Smith, M. Tillberg, and J. I. Garrels. Ypd(tm), pombepd(tm), and wormpd(tm): model organism volumes of the bioknowledge(tm) library, an integrated resource for protein information. *Nucleic Acids Res.*, 29:75–79, 2001.
3. R. Dobrin, Q. K. Beg, A.-L. Barabási, and Z. N. Oltvai. Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network. *BMC Bioinformatics*, 5:10, Jan 2004.
4. J.-D. J. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. M. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995):88–93, Jul 2004.
5. A. Jaimovich, G. Elidan, H. Margalit, and N. Friedman. Towards an integrated protein-protein interaction network: a relational markov network approach. *J. Comp. Bio.*, 13:145–164, 2006.
6. H. Jeong, S. Mason, A.-L. Barabási, and Z. N. Oltvai. Centrality and lethality of protein networks. *Nature*, 411, 2001.
7. H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407, 2000.
8. S. Kalir, J. McClure, K. Pabbaraju, C. Southward, M. Ronen, S. Leibler, M. G. Surette, and U. Alon. Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. *Science*, 292(5524):2080–2083, Jun 2001.
9. N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20(11):1746–1758, Jul 2004. Evaluation Studies.
10. N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Topological generalizations of network motifs. *Phys. Rev. E*, 70:031909, 2004.
11. T. I. Lee and R. A. Young. Transcription of eukaryotic protein-coding genes. *Annu. Rev. Genet.*, 34:77–137, 2000.
12. S. Mangan and U. Alon. Structure and function of the feed-forward loop network motif. *PNAS*, 100(21):11980–11985, Oct 2003.
13. S. Mangan, A. Zaslaver, and U. Alon. The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *J. Mol. Biol.*, 334(2):197–204, Nov 2003.

14. B. D. McKay. Practical graph isomorphism. In *Proceedings of the Tenth Manitoba Conference on Numerical Mathematics and Computing, Vol. I (Winnipeg, Man., 1980)*, volume 30, pages 45–87, 1981. http://cs.anu.edu.au/∼bdm/nauty/.
15. B. D. McKay. Isomorph-free exhaustive generation. *J. Algorithms*, 26:306–324, 1998.
16. M. Middendorf, E. Ziv, and Chris H. Wiggins. Inferring network mechanisms: the Drosophila melanogaster protein interaction network. *PNAS*, 102(9):3192–3197, Mar 2005.
17. R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, Mar 2004.
18. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, Oct 2002.
19. T. M. Przytycka. An important connection between network motifs and parsimony models. In *RECOMB 2006*, pages 321–335, 2006.
20. M. Ronen, R. Rosenberg, B. I. Shraiman, and U. Alon. Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc Natl Acad Sci U S A*, 99(16):10555–10560, Aug 2002.
21. S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics*, 31(1):64–68, May 2002.
22. JUNG Framework Development Team. Jung: The java universal network/graph framework, 2005.
23. J. R. Ullman. An algorithm for subgraph isomorphism. *J. Assoc. Comp. Mach.*, 23(1):31–42, Jan 1976.
24. A. Vazquez, R. Dobrin, D. Sergi, J.-P. Eckmann, Z. N. Oltvai, and A.-L. Barabasi. The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *PNAS*, 101(52):17940–17945, Dec 2004.
25. A. Zaslaver, A. E. Mayo, R. Rosenberg, P. Bashkin, H. Sberro, M. Tsalyuk, M. G. Surette, and U. Alon. Just-in-time transcription program in metabolic pathways. *Nature Genetics*, 36(5):486–491, May 2004.
26. E. Ziv, R. Koytcheff, M. Middendorf, and C. Wiggins. Systematic identification of statistically significant network measures. *Phys. Rev. E*, 71:016110, 2005.

Information-Theoretic Inference of Gene Networks Using Backward Elimination

Patrick E. Meyer[1,2,3], **Daniel Marbach**[1,2], **Sushmita Roy**[1,2] **and Manolis Kellis**[1,2]
[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, MA, USA
[2]Broad Institute of MIT and Harvard, MA, USA
[3]Machine Learning Group, Université Libre de Bruxelles, Brussels, Belgium

Abstract— *Unraveling transcriptional regulatory networks is essential for understanding and predicting cellular responses in different developmental and environmental contexts. Information-theoretic methods of network inference have been shown to produce high-quality reconstructions because of their ability to infer both linear and non-linear dependencies between regulators and targets. In this paper, we introduce MRNETB an improved version of the previous information-theoretic algorithm, MRNET, which has competitive performance with state-of-the-art algorithms.*

MRNET infers a network by using a forward selection strategy to identify a maximally-independent set of neighbors for every variable. However, a known limitation of algorithms based on forward selection is that the quality of the selected subset strongly depends on the first variable selected. In this paper, we present MRNETB, an improved version of MRNET that overcomes this limitation by using a backward selection strategy followed by a sequential replacement. Our new variable selection procedure can be implemented with the same computational cost as the forward selection strategy.

MRNETB was benchmarked against MRNET and two other information-theoretic algorithms, CLR and ARACNE. Our benchmark comprised 15 datasets generated from two regulatory network simulators, 10 of which are from the DREAM4 challenge, which was recently used to compare over 30 network inference methods. To assess stability of our results, each method was implemented with two estimators of mutual information. Our results show that MRNETB has significantly better performance than MRNET, irrespective of the mutual information estimation method. MRNETB also performs comparably to CLR and significantly better than ARACNE indicating that our new variable selection strategy can successfully infer high-quality networks.

Keywords: mutual information, systems biology, network inference

1. Introduction

Transcriptional regulatory networks are networks of transcription factor (TF) proteins and their target genes, where each edge represents the regulatory activity of each TF on a target gene. These networks summarize the underlying regulatory circuitry by which precise, condition-specific patterns of gene expression are generated under different cellular contexts. Understanding these networks can not only provide insights into how cells function, but can lead to effective drug delivery for treating human diseases by specific network-informed intervention.

Network inference methods (also called *reverse engineering* methods) attempt to reconstruct the transcriptional network of a cell from gene expression data [1]. However, this is a challenging task because of the large amount of experimental and biological noise in expression data, and because of the high dimensionality and combinatorial nature of the problem [25]. Notwithstanding, the bioinformatics community has developed several successful network inference techniques in the last decade [8]. In particular, information-theoretic methods for network inference have been shown to scale to datasets with several thousand genes and a limited amount of samples [7], [13], [15].

In this paper, we present MRNET Backward (MRNETB), an improved version of the previous network inference method called MRNET (Minimum Redundancy NETwork) [15]. MRNET infers a network of interactions between TFs and target genes by using a forward selection strategy to identify a maximally independent set of neighbors for every variable. However, methods based on forward selection, including MRNET, rely on the correct choice of the first neighbor and suffer in performance if the first neighbor is chosen incorrectly. MRNETB overcomes this limitation by implementing a combination of a backward elimination and a sequential replacement procedure. Importantly, we implement our new neighbor selection procedure at the same computational cost as forward selection.

Our proposed technique is benchmarked against MRNET and two other state-of-the-art information-theoretic algorithms, namely CLR (Context Likelihood of Relatedness) [7] and ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) [13]. Our benchmark comprised fifteen datasets generated from two regulatory network simulators. Ten out of the fifteen datasets come from the DREAM4 challenge, a competition devoted to compare the performance of network inference methods [11]. To assess stability of our results, each method was implemented

with two different estimators of mutual information. Our results show that irrespective of the mutual information estimation used, MRNETB performs better than MRNET and ARACNE and performs comparably against CLR. Our new strategy thus successfully infers high-quality networks, without incurring additional computational costs.

The rest of the paper is organized as follows: Section 2 provides background on information-theoretic methods of network inference and the two network inference methods, CLR and ARACNE. Section 3 describes MRNET and our new approach, MRNETB. Section 4 describes the experimental framework and results and the conclusions are summarized in Section 5.

2. Mutual Information Network Inference

Mutual information network inference methods comprise a subcategory of network inference methods, which infer regulatory interactions between genes based on pairwise mutual information [15]. As a first step, these methods require the computation of the mutual information matrix (MIM), a square matrix whose mim_{ij} element is given by the mutual information between X_i and X_j:

$$\text{mim}_{ij} = I(X_i; X_j) \tag{1}$$

where X_i and X_j are random variables denoting the expression levels of genes i and j, respectively.

The main advantages of mutual information networks for inference of transcriptional regulatory networks are:

- The computational complexity is affordable. This results from the fact that only $\binom{n}{2}$ calls of mutual information, based on bivariate probability distributions, are required to compute the MIM [13].
- The number of required samples is rather low, since only bivariate distribution are to be estimated [17].

In the following, we first review two state-of-the-art network inference methods based on pairwise mutual information. We proceed by describing two commonly used mutual information estimators.

2.1 Mutual Information Estimation

Two popular ways of computing mutual information are: (1) discretizing variables with an equal frequency binning (so that marginal distributions are uniform), and (2) assuming normally distributed variables. We will now describe these two procedures.

2.1.1 Uniformly distributed discrete variables and empirical estimation

If X_i is a continuous random variable taking values between a and b, the interval $[a, b]$ can be discretized by partitioning it into $|\mathcal{X}_i|$ subintervals, called *bins*, where the symbol $\mathcal{X}_i$ denotes the bin vector. We use $\#(x_{i_k})$ to denote the number of data points in the kth bin of variable X_i, the symbol $m = \sum_k \#(x_{i_k})$ to denote the total number of samples and the symbol $\hat{p}(x_{i_k}) = \frac{\#(x_{i_k})}{m}$ to denote the empirical probability.

One possible binning strategy is to divide the interval $[a, b]$ of X_i such that each subinterval contains the same number of data points. It follows that subinterval sizes are typically different. This binning strategy is often referred as *equal frequency binning*. The number of subintervals should be chosen so that all bins contain a significant number of samples. Here, we set $|\mathcal{X}_i| = \sqrt{m}$, which is a common choice [28]. The entropy of X_i is $H(X_i) = \log |\mathcal{X}_i|$, since its distribution is uniform across the bins. As a result, the "empirical" or "plug-in" estimation of the mutual information becomes

$$\begin{array}{c} I(X_i; X_j) = \log(|\mathcal{X}_i||\mathcal{X}_j|) \\ + \sum_{x_{i_k} \in \mathcal{X}_i} \sum_{x_{j_l} \in \mathcal{X}_j} \hat{p}(x_{i_k}, x_{j_l}) \log \hat{p}(x_{i_k}, x_{j_l}) \end{array} \tag{2}$$

It should be noted that this entropy estimation tends to underestimate the true value of entropy, i.e. $E[H(p)] \leq H(E[p])$ [18].

2.1.2 Normally distributed continuous variables

Let X be a multivariate Gaussian random variable with mean μ and covariance C. The probability density function of X is

$$f(X) = \frac{1}{\sqrt{(2\pi)^n |C|}} \exp^{(-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu))}. \tag{3}$$

The mutual information between two Gaussian random variables, X_i and X_j, is then given by

$$I(X_i, X_j) = \frac{1}{2} \log \left(\frac{\sigma_{ii}\sigma_{jj}}{|C|} \right) \tag{4}$$

$$= -\frac{1}{2} \log(1 - \rho^2). \tag{5}$$

where $|C|$ is the determinant of the covariance matrix and ρ is the Pearson's correlation [9].

The Spearman rank correlation was shown to have better empirical performance for network inference than Pearson's correlation [17].

2.2 Context Likelihood of Relatedness (CLR)

The CLR algorithm [7] is an extension of the relevance network approach [4]. The latter approach has been introduced for gene clustering and successfully applied to infer relationships between RNA expression and chemotherapeutic susceptibility [3]. The relevance networks approach consists of inferring a network in which a pair of genes $\{X_i, X_j\}$ are linked by an edge if the mutual information $I(X_i; X_j)$ is larger than a given threshold θ. The complexity of the method is $O(n^2)$ since all pairwise interactions are considered.

The CLR algorithm derives a score from the empirical distribution of the mutual information for each pair of genes. In particular, instead of considering the information $I(X_i; X_j)$ between genes X_i and X_j, it estimates a score $w_{ij} = \sqrt{z_i^2 + z_j^2}$, where

$$z_i = \max\left(0, \frac{I(X_i; X_j) - \mu_i}{\sigma_i}\right) \quad (6)$$

The parameters μ_i and σ_i are the mean and the standard deviation of the empirical distribution of the mutual information values $I(X_i, X_k)$ of X_i with all other variables X_k ($k = 1, \ldots, n$). The CLR algorithm has a complexity in $O(n^2)$. It was successfully applied to decipher the *E. coli* transcriptional regulatory network [7].

2.3 Algorithm for Reconstruction of Accurate Cellular Networks (ARACNE)

ARACNE [13] is based on the Data Processing Inequality, which states that, if gene X_i interacts with gene X_j through gene X_k, then

$$I(X_i; X_j) \leq \min\left(I(X_i; X_k), I(X_j; X_k)\right)$$

ARACNE first assigns to each pair of nodes a weight equal to their mutual information. Then, all edges for which $I(X_i; X_j) < \theta$ are removed, where θ is a given threshold. Eventually, the weakest edge of each triplet is interpreted as an indirect interaction and is removed.

An extension of ARACNE removes the weakest edge only if the difference between the two lowest weights lies above a threshold η. Hence, η allows the number of pruned edges to be tunable.

If the network is a tree including only pairwise interactions, the method guarantees the reconstruction of the original network, once it is provided with the exact MIM [13]. ARACNE's complexity is $O(n^3)$ since the algorithm considers all triplets of genes. ARACNE successfully recovered components of transcriptional regulatory networks in mammalian cells and was shown to have favorable performance compared to Bayesian networks and relevance networks [13].

3. Minimum Redundancy Networks

MRNET infers a network by using a variable selection procedure called Maximum Relevance Minimum Redundancy (MRMR) for every random variable $X_j \in X$, [23], [19]. Assume X_j is the variable for which we need to select the predictor variables. The MRMR methods ranks a set $X_{S_j} \subseteq \{X \setminus X_j\}$ of the predictor variables according to the difference between the mutual information of $X_i \in X_{S_j}$ with X_j (the relevance) and the average mutual information with the selected variables in X_{S_j} (the redundancy). The rationale is that direct interactions should be ranked before indirect interactions (i.e. the ones with redundant information with the direct ones) by the method.

The MRMR method has been introduced together with a forward selection that starts by selecting the variable X_k that has the highest mutual information with the target X_j. The second selected variable X_i will be the one having a high information $I(X_i; X_j)$ with the target and at the same time a low information $I(X_i; X_k)$ with the previously selected variable. In the following steps, given a set X_{S_j} of selected variables, X_{S_j} is updated by choosing the variable, X_i^{MRMR}, that maximizes the MRMR score s_i:

$$\begin{gathered} X_i^{MRMR} = \arg\max_{X_i \in X_{-(i,j)}} (s_i) \\ s_i = I(X_i; X_j) - \frac{1}{|S_j|} \sum_{k \in S_j} I(X_i; X_k) \end{gathered} \quad (7)$$

At each step of the algorithm, the selected variable thus represents a trade-off between relevance and redundancy.

The generic principle of MRNET consists of identifying a subset X_{S_j} (for each variable X_j, $j = 1, 2.., n$) whose variables have maximal pairwise relevance with X_j and, at the same time, maximal pairwise independence among them. More formally,

$$\begin{gathered} X_{S_j} = \arg\max_{X_{S_j} \in X_{-j}} (u - r) \\ u = \frac{1}{|S_j|} \sum_{i \in S_j} I(X_i; X_j) \\ r = \frac{2}{|S_j|(|S_j|-1)} \sum_{i,k>i \in S_j} I(X_i; X_k) \end{gathered} \quad (8)$$

The network inference approach MRNET consists of repeating this selection procedure for each target gene $X_j \in X$. For each pair $\{X_i, X_j\}$, MRNET returns two (not necessarily equal) scores s_i and s_j according to (7). The score of the pair $\{X_i, X_j\}$ is defined as the maximum between s_i and s_j. A specific network can then be inferred by only keeping edges whose score lies above a given threshold θ (as in CLR). Thus, the algorithm infers an edge between X_i and X_j either when X_i is a well-ranked predictor of X_j ($s_i > \theta$), or when X_j is a well-ranked predictor of X_i ($s_j > \theta$). In practice, the selection of variables stops when the average redundancy term $\frac{1}{|S_j|} \sum_{k \in S_j} I(X_i; X_k)$ exceeds the relevance term $I(X_i; X_j)$.

MRNET has complexity of $O(n^3)$ since each variable selection is $O(n^2)$. MRNET was shown to have similar or better performance than alternative information-theoretic network inference methods [15], [10], [21], [17].

3.1 MRNET Backward (MRNETB)

In this section, we introduce MRNETB, an improved version of MRNET. This new algorithm overcomes a major limitation of MRNET coming from the use of forward selection as the subset search strategy. A known limitation of algorithms based on forward selection, including MRNET, is that the quality of the selected subset strongly depends on the first variable selected. If the first selected variable, i.e. the variable with the highest mutual information with the target variable X_j, is not a true neighbor of X_j, then minimizing

redundancy (or equivalently maximizing independency) with that wrongly selected variable is not desirable.

The optimisation problem of MRNET (8) is a binary quadratic optimization problem. Backward elimination combined with a sequential search is known to perform well on binary quadratic problems [2]. The backward elimination method starts with a set containing all the variables and then selects the variable X_i whose removal induces the highest increase of the objective function till the stopping criterion is fulfilled (i.e. a relevance term $\frac{1}{|S_j|}\sum_{i\in S_j} I(X_i;X_j)$ higher than redundancy term $\frac{2}{|S_j|(|S_j|-1)}\sum_{i,k>i\in S_j} I(X_i;X_k)$). The procedure is enhanced by an iterative sequential replacement which, at each step, swaps the status of a selected and a non selected variable such that the largest increase in the objective function is achieved. The sequential replacement is stopped when no further improvement is met.

Forward selection, backward elimination, and sequential replacement all have an algorithmic complexity of $O(n^2)$ [14]. In other words, the network built using a backward elimination followed by sequential replacement has the same asymptotic computational cost as the one based on a forward selection strategy. Backward elimination has been previously shown to outperform forward selection in variable selection [6]. In the next section, we show that the backward selection strategy also improves the performance in network inference using mutual information.

We have made MRNETB available in the latest version (2.5) of our bioconductor open-source package called minet [16].

4. Experiments

We compare the performance of our new approach MRNETB with the three existing network inference methods described above (ARACNE, CLR and MRNET) using a framework based on artificial (simulated) regulatory networks (Fig 1). In simulation, the ground truth is known and inferred networks can be systematically evaluated (which is typically not possible in vivo) [11].

The framework is composed of the following four steps:

1) Produce artificial gene expression data from networks of known structure
2) Compute the MIM using two different mutual information estimators for each produced dataset, namely the empirical and the Spearman based estimator.
3) Infer the network from each computed MIM
4) Assess the quality of the inferred networks using the area under the precision-recall (PR) curve.

4.1 Metrics

The area under the PR curve (AUPRC) is a standard metric used in the network inference community [20]. The PR curve plots precision (pre) against recall (rec) for different values of a threshold determining which pairs qualify as a predicted edge [22]. Precision, pre, is given by

$$pre = \frac{tp}{tp + fp} \tag{9}$$

Precision measures the fraction of real edges (tp) among the ones classified as positive ($tp + fp$). Recall, rec, also called true positive rate (tpr), is given by

$$rec = tpr = \frac{tp}{tp + fn} \tag{10}$$

It denotes the fraction of real edges that are correctly inferred.

In a network inference setting, the PR curve illustrates the trade-off between eliminating many low confidence edges by using a high threshold (low recall, high precision) versus keeping many arcs (high recall) with doubt on their significance (low precision). These quantities depend on the threshold chosen. A good network inference method maximizes both precision and recall.

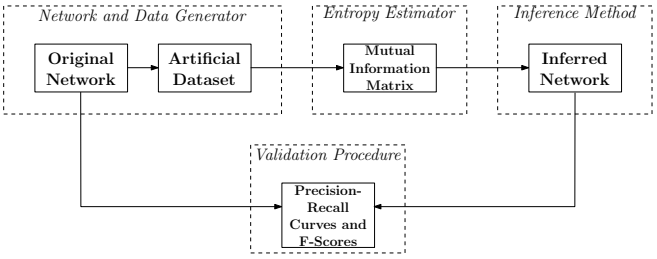


Fig. 1: Experimental framework used to assess network reconstruction quality

4.2 Datasets

We compared the different network inference methods on fifteen datasets (Table 1). These datasets come from two different generators, namely Syntren [27] and GeneNetWeaver (GNW) [12], [11]. Ten out of these fifteen datasets are from the DREAM4 challenge, which was recently used to compare over 30 network inference methods. Briefly, five of the DREAM4 datasets are from the so-called multifactorial challenge. These five datasets have been obtained by applying multifactorial perturbations to the original network. The other five datasets contain time courses capturing the response of the network to a perturbation, followed by gradual return to steady state upon removal of the perturbation. For a detailed description of these datasets, refer to the DREAM project website (http://wiki.c2b2.columbia.edu/dream/).

4.3 Results

We compared MRNETB against three state-of-the-art information theoretic methods for network inference (Tables 2 and 3) using the area under the PR curve (AUPRC) on the benchmark datasets, as described in the previous section. MRNETB has significantly higher AUPRC than both

Dataset	Source	Variables	Samples
syntren1	Syntren - E.Coli	300	100
syntren2	Syntren - E.Coli	300	100
syntren3	Syntren - E.Coli	300	100
syntren4	Syntren - E.Coli	300	100
syntren5	Syntren - E.Coli	300	100
multifact1	Dream4-Multifact1	100	100
multifact2	Dream4-Multifact2	100	100
multifact3	Dream4-Multifact3	100	100
multifact4	Dream4-Multifact4	100	100
multifact5	Dream4-Multifact5	100	100
dream1	Dream4-TS1	100	210
dream2	Dream4-TS2	100	210
dream3	Dream4-TS3	100	210
dream4	Dream4-TS4	100	210
dream5	Dream4-TS5	100	210

Table 1: Benchmark Datasets

MRNET and ARACNE (p-values were computed using a paired Wilcoxon test on the AUPRC, see Table 4), suggesting that MRNETB outperforms these methods. MRNETB outperforms CLR on the DREAM time series datasets, and is outperformed by CLR on the Syntren datasets, suggesting that MRNETB and CLR have at par performance. Note that the results for ARACNE are over-pessimistic because the method typically produces networks with high precision and low recall. As a result, for lower thresholds the area under ARACNE's PR-curve is much smaller than for the other methods.

The results are similar for both estimators of mutual information (i.e. empirical and Spearman based estimation). However, network inference based on Spearman's correlation has higher performance than inference based on empirical estimation of mutual information.

It is worth noting that MRNETB and CLR would have ranked third and fourth in the multifactorial DREAM4 challenge.

	CLR	ARACNE	MRNET	MRNETB
syntren1	**0.12**	0.04	0.09	0.11
syntren2	**0.12**	0.02	0.08	0.10
syntren3	**0.12**	0.02	0.08	0.10
syntren4	**0.11**	0.02	0.07	0.09
syntren5	**0.12**	0.02	0.08	0.10
multifact1	0.12	0.12	0.11	**0.13**
multifact2	0.12	0.11	0.11	**0.13**
multifact3	**0.23**	0.19	0.21	**0.23**
multifact4	**0.18**	0.14	0.14	**0.18**
multifact5	0.16	0.15	0.16	**0.17**
dream1	**0.11**	0.05	0.08	0.10
dream2	**0.11**	0.08	0.10	**0.11**
dream3	**0.18**	0.10	0.12	**0.18**
dream4	**0.14**	0.08	0.10	0.13
dream5	0.13	0.09	0.12	**0.14**
avg	0.138	0.082	0.11	0.133

Table 2: AUPRC Scores for the Empirical Estimation of Mutual Information. Best scores are in bold.

	CLR	ARACNE	MRNET	MRNETB
syntren1	**0.13**	0.04	0.10	0.12
syntren2	**0.13**	0.03	0.10	0.12
syntren3	**0.14**	0.04	0.11	0.12
syntren4	**0.13**	0.04	0.10	0.12
syntren5	**0.14**	0.03	0.09	0.12
multifact1	**0.22**	0.18	0.21	0.21
multifact2	0.21	0.16	0.22	**0.24**
multifact3	0.34	0.35	**0.38**	0.32
multifact4	0.33	0.29	**0.34**	0.31
multifact5	0.32	0.32	**0.36**	0.30
dream1	**0.09**	0.05	0.08	**0.09**
dream2	0.10	0.08	0.11	**0.12**
dream3	0.18	0.11	0.18	**0.21**
dream4	**0.12**	0.07	0.11	**0.12**
dream5	0.14	0.11	**0.17**	0.15
avg	0.181	0.126	0.177	0.178

Table 3: AUPRC Scores for the Squared Spearman Rho Correlation. Best scores are in bold.

MRNETB Losses/Ties/Wins	CLR	ARACNE	MRNET
Total	16/6/8	2/0/28	4/0/26
P-val MRNETB vs	CLR	ARACNE	MRNET
Total	0.11	5e-6	0.01

Table 4: Losses/ties/wins of MRNETB vs each method and p-values obtained with a paired Wilcoxon test.

5. Conclusion

This paper introduces MRNETB, a new information-theoretic method for inferring gene regulatory networks using gene expression data. Similar to other state-of-the-art information-theoretic methods, MRNETB relies on pairwise mutual information. As a result, this method can tackle datasets with large number of variables and low number of samples. An appealing aspect of the new method is the use of a better and more robust search for identifying a maximally informative subset of variables than a classic forward selection, without incurring additional computational cost. We compared MRNETB against three other approaches on fifteen datasets and the experimental results show that the proposed technique is competitive. Further research will focus on using MRNETB within an integrative framework to assess the relative information contribution of different data sources for improving network reconstruction.

References

[1] A. Ambesi-Impiombato and D. di Bernardo. Computational biology and drug discovery: from single-target to network drugs. *Current Bioinformatics*, 2006.

[2] A. Billionnet and F. Calmels. Linear programming for the 0-1 quadratic knapsack problem. *European Journal of Operational Research*, 92:310–325, 1996.

[3] A. J. Butte, P. Tamayo, D. Slonim, T.R. Golub, and I.S. Kohane. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences*, 97(22):12182–12186, 2000.

[4] A. J. Butte and I. S. Kohane. Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurments. *Pacific Symposium on Biocomputing*, 5:415–426, 2000.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley, New York, 1990.

[6] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research* **3** (2003) 1157–1182

[7] J.J. Faith, B. Hayete, J.T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J.J. Collins, and T.S. Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, 5, 2007.

[8] T. S. Gardner and J. Faith. Reverse-engineering transcription control networks. *Physics of Life Reviews 2*, 2005.

[9] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall International, 1999.

[10] F. M. Lopes, D. C. Martins, and R. M. Cesar. Comparative study of grns inference methods based on feature selection by mutual information. In *IEEE International Workshop on Genomic Signal Processing and Statistics*, 2009.

[11] D. Marbach, R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *PNAS*, 107(14):6286-6291, 2010.

[12] D. Marbach, T. Schaffter, C. Mattiussi, and D. Floreano. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of Computational Biology*, 16(2):229-239, 2009.

[13] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7, 2006.

[14] P. Merz and B. Freisleben. Greedy and local search heuristics for unconstrained binary quadratic programming. *Journal of Heuristics*, 8(2):1381–1231, 2002.

[15] P. E. Meyer, K. Kontos, F. Lafitte, and G. Bontempi. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP Journal on Bioinformatics and Systems Biology*, Special Issue on Information-Theoretic Methods for Bioinformatics, 2007.

[16] P. E. Meyer, F. Lafitte, and G. Bontempi. Minet: An open source r/bioconductor package for mutual information based network inference. *BMC Bioinformatics*, 2008.

[17] C. Olsen, P. E. Meyer, and G. Bontempi. On the impact of entropy estimation on transcriptional regulatory network inference based on mutual information. *EURASIP Journal on Bioinformatics and Systems Biology*, 2009.

[18] L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003.

[19] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[20] R. J. Prill, D. Marbach, J. Saez-Rodriguez, P. K. Sorger, L. G. Alexopoulos, X. Xue, N. D. Clarke, G. Altan-Bonnet, and G. Stolovitzky. Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS ONE*, 5(2):e9202, 2010.

[21] T. Shimamura, S. Imoto, R. Yamaguchi, A. Fujita, M. Nagasaki, and S. Miyano. Recursive regularization for inferring gene networks from time-course gene expression profiles. *BMC Systems Biology*, 3, 2009.

[22] M. Sokolova, N. Japkowicz, and S. Szpakowicz. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Proceedings of the AAAI'06 workshop on Evaluation Methods for Machine Learning*, 2006.

[23] G. D. Tourassi, E. D. Frederick, M. K. Markey, and Jr. C. E. Floyd. Application of the mutual information criterion for feature selection in computer-aided diagnosis. *Medical Physics*, 28(12):2394–2402, 2001.

[24] M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde T, H. Bartelink H, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medecine*, 347, 2002.

[25] E. P. van Someren, L. F. A. Wessels, E. Backer, and M. J. T. Reinders. Genetic network modeling. *Pharmacogenomics*, 3(4):507–525, 2002.

[26] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 406, 2002.

[27] T. Van den Bulcke, K. Van Leemput, B. Naudts, P. van Remortel, H. Ma, A. Verschoren, B. De Moor, and K. Marchal. Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, 7(1):43, 2006.

[28] Y. Yang and G. I. Webb. Discretization for naive-bayes learning: managing discretization bias and variance. Technical Report 2003/131 School of Computer Science and Software Engineering, Monash University, 2003.

[29] N. V. Zhukov and S. A. Tjulandin. Targeted therapy in the treatment of solid tumors: Practice contradicts theory. *Biochemistry Moscow*, 2008.