

Whole-genome Comparative Annotation and Regulatory Motif Discovery in Multiple Yeast Species

Manolis Kamvysselis^{1,2}, Nick Patterson¹, Bruce Birren¹, 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 (Kamvyselis 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 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

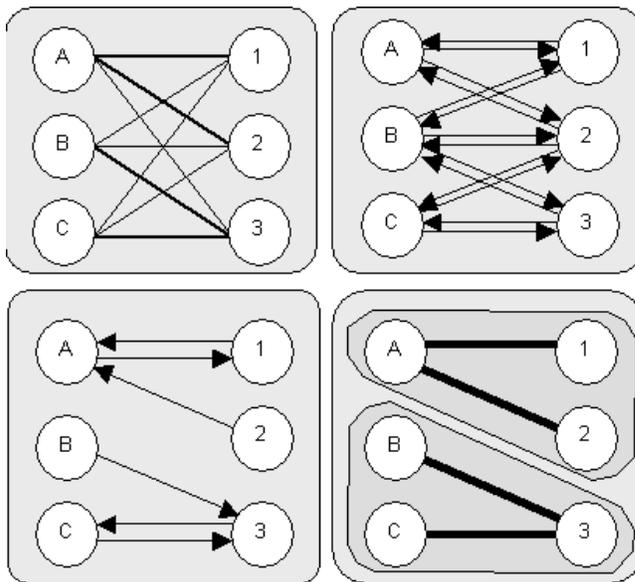


Figure 1. Overview of graph separation. We construct a bipartite graph based on the blast hits. We consider both forward and reverse matches for near-optimality based on synteny and sequence similarity. Sub-optimal matches are progressively eliminated simplifying the graph. We return the connected components 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 (x_1, y_1) and (x_2, y_2) were such that x_1 was physically near x_2 , and y_1 was physically near y_2 , we constructed a synteny block $B = (\{x_1, x_2\}, \{y_1, y_2\})$. Thereafter, for a gene x_3 that was proximal to $\{x_1, x_2\}$, if an outgoing edge (x_3, y_3) existed such that y_3 was proximal to $\{y_1, y_2\}$, we ignored other outgoing edges (x_3, y') if y' was not proximal to $\{y_1, y_2\}$.

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

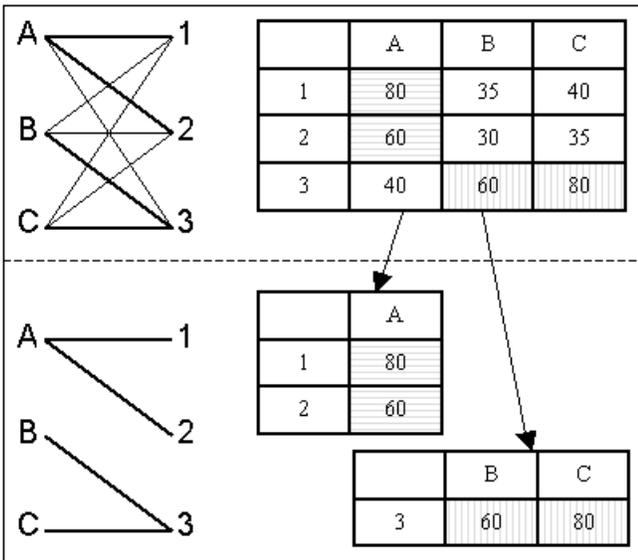


Figure 2. Best Unambiguous Subsets (BUS). A BUS is a set of genes that can be isolated from a homology group while preserving all potentially orthologous matches. Given the similarity matrix above and no synteny information, two such sets are $(A,1,2)$ and $(B,C,3)$.

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 .

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 $X \cup S$, to be a subset S of genes, such that $\forall x: x \in S \Leftrightarrow \text{best}(x) \subseteq S$, where $\text{best}(x)$ are the nodes incident to the maximum weight edges from x . We then constructed 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

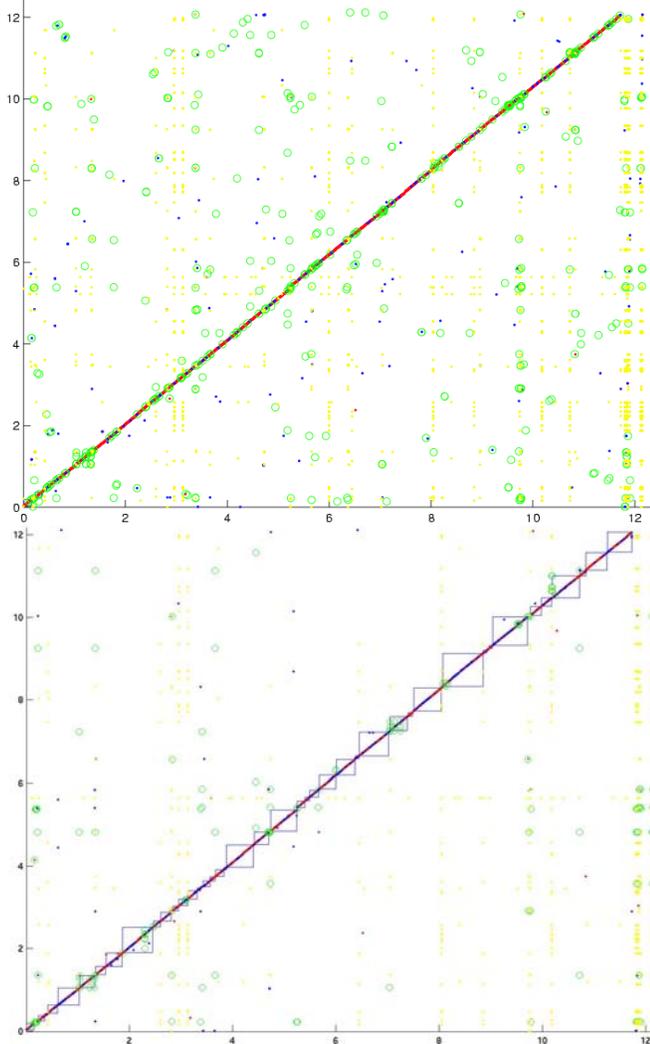


Figure 3. The effect of using synteny. Blocks of conserved gene order (blue squares) help resolve additional ambiguities.

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

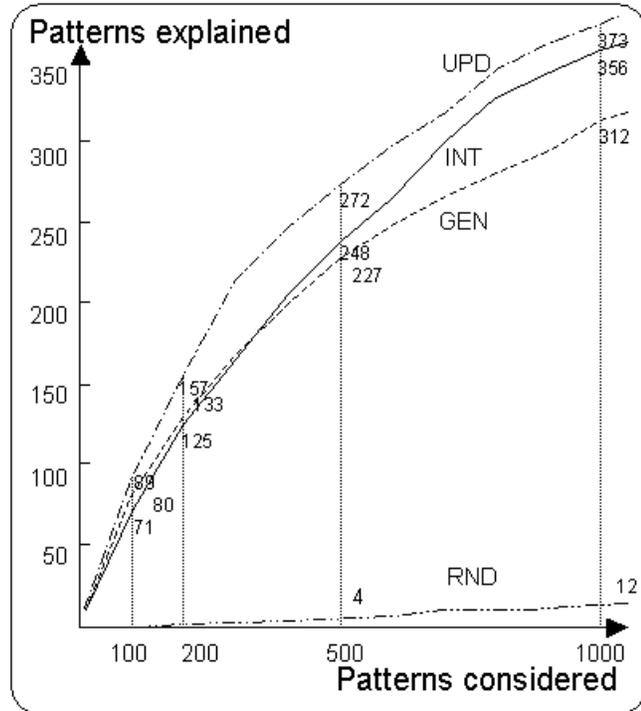


Figure 4. Selecting motif cores using three different tests. Patterns selected by one of the three tests (INT,GEN,UPD) correlate with function 90 times more frequently than randomly chosen motifs (RND).

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 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.

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 $1/4$ 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.

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

Select	... TCA ... ACG ...	INT 1
Extend	... RTCAY ... ACGR ...	INT 1x
Collapse	... RTCAY ... ACGR RTCAC ... ACGA RTCAC ... ACGA GTCAC ... ACG ATCAY ... ACGA RTCAC ... ACGA RTCAT ... ACGR RTCAY ... ACGG ATCAY ... ACGG ... (...)	INT_1x INT_9x INT_19x INT_29x INT_46x INT_78x INT_161x INT_165x INT_336x (...)
	... RTCAY ... ACGR ...	INT: M1
Merge	... RTCAY ... ACGR RTCAY ... ACGR RTCRYk ... ACGR ... (...)	INT: M1 GEN: M1 UPD: M2 (...)
	... RTCAY ... ACGR ...	Fin: M1

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.

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	TRTCAYT-Y--ACGRA	✓	RTCAC___ACGA	14.6	✓	same
2	GCN4	ATGACTCAT	47.8	TGAGTCAY	✓	RTGACTCA	10.9	✓	same
3	REB1	CCGGGTAA	44.7	SCGGGTAAAY	✓	CCGGGTAAAC	8.7	✓	same
4	MCM1a	TTWCCcnwwwrGGA	35.9	TTTCC-AAW-RGGAAA	✓	TCC___GGA	4.4	✓	same
5	RAP1	ACACCCATACATTT	30.0	TTWACAYCCRTACAY-Y	✓	ACCCA.ACA	8.7	✓	same
6	Cbf1	RTCACRTG	24.2	TRGTCACGTG	✓	GTCACGTG	10	✓	same
7	FKH2	TTGTTTACST	20.7	TTGTTTAC-TWTT	✓	TGTTTAC..TT	8.3	✓	same
8	SWI4	CRCGAAAA	19.9	CSMRRCGCGAAAA	✓	CAACRCGAAAA	8.1	✓	same
9	MBP1	ACGCGTnA	19.6	G-RR-A-ACGCGT-R	~	AACGCGTCG	9.5	✓	better (+)
10	STE12	RTGAAACA	17.8	GSAASRR-TGATRAWGYA		YTGAAACA	12.2	✓	better (+)
11	Gal4	CGGnnnnnnnnnnCCG	16.1	CGGM---CW-Y--CCCG	~	CGG_____CCGA	7.8	✓	better (+)
12	SWI6	ACGCGT	15.6	WCGCGTCGCGTY-C	✓	ACGCGT	7.4	✓	same
13	PHO4	CACGTG	14.2	TTGTACACTTYGTTT		CGCACGTG	4.6	✓	better (+)
14	HSF1	TTCTAGAA	14.1	TYTTCYAGAA--TTCY	✓	GTTCTAGAA_TTC_G	9.6	✓	same
15	Dig1	RTGAAACA	13.6	CCYTG-AYTTCW-CTTC		TGAAACR	11.8	✓	better (+)
16	INO4	CATGTGAAat	13.4	G..GCATGTGAAAA	✓	G...CATGTGAA	6.8	✓	same
17	FKH1	TTGTTTACST	13.2	CYTRTTTAY-WTT	✓	TGTTTAC	6.5	✓	same
18	Leu3	CCGGNCCGG	13.1	GCCGGTMMCGSYC--	✓	CCGG__CGG	6.6	✓	better (+)
19	Bas1	TGACTC	10.2	CS-CCAATGK--CS		TGACTCTA	9.5	✓	better (+)
20	SWI5	KGCTGR	9.2	CACACACACACACACACA		TGCTGG	6.1	✓	better (+)
21	HAP4	TnRTTGGT	8.5	YCT-ATTSG-C-GS	~	TGATTGGT	6.4	✓	better (+)
22	RLM1	CTA\WWWWTAG	8.4	A-CTSGAAGAAATGCGGT		CTA..TTTAG	4.7	✓	better (+)
23	INO2	CATGTGAAat	7.4	GCATGTGRAAA	✓	CATGTG	4.4	✓	same
24	MET31	AAACTGTGGC	7.0	GCACTGTGATS		TGTGGC	5.8	✓	same
25	ACE2	GCTGGT	5.2	GTGTGTGTGTGTGTG		TGCTGGT	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.