Methods in comparative genomics: genome correspondence, gene identification and motif discovery

Manolis Kellis[1,2], Nick Patterson[1], Bruce Birren[1], Bonnie Berger[2,3,5], Eric S. Lander[1,4,5]

manoli@mit.edu, nickp@genome.wi.mit.edu, bwb@genome.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 Computer Science and Artificial Intelligence Laboratory, 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 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*. 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 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.

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.

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 the genome-wide 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 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 human.

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 genome-wide comparison of four complete eukaryotic genomes. This paper focuses on the computational developments 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 the genome-wide 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 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 decided to use 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 should contain 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; Fitch 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 bi-directional matches will represent orthologous relations and incorrect one-to-one matches may be established.

Clusters of Orthologous Genes (COG) (Tatusov et al. 1997; Tatusov et al. 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 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 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 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 sub-optimal in a series of steps. As a pre-processing step, we eliminated all edges that are less than 80% of the maximum-weight edge both in amino acid identity and in length (Figure 2). 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 (Figure 3). 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 (Figure 4). These ensured that the bipartite graph is maximally separable, while maintaining all possibly orthologous relationships (Figure 5).

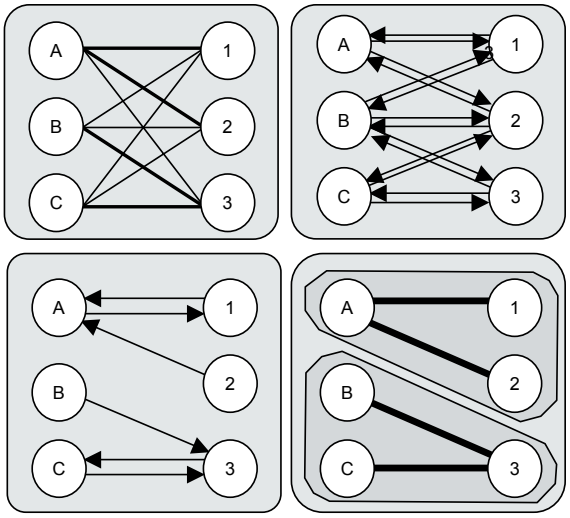

Figure 1. Overview of graph separation. We constructed a bipartite graph based on the BLAST hits between genes A,B,C in one genome and genes 1,2,3 in a related genome. We considered both forward and reverse matches for near-optimality based on synteny and sequence similarity. Sub-optimal matches were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

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 sequence of *S. paradoxus*, *S. mikatae* and *S. bayanus* was obtained at the MIT/Whitehead Institute Center for Genome Research. We used a whole-genome shotgun sequencing approach with paired-end sequence reads of 4kb plasmid clones, with lab protocols as described at www-genome.wi.mit.edu. We used ~7-fold redundant coverage, namely every nucleotide in the genome was contained on average in at least 7 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 sub-optimal matches

Let **G**=(**X**+**Y**,**E**) 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 ∈ **X** and y ∈ **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 non-overlapping 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 (Figure 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.

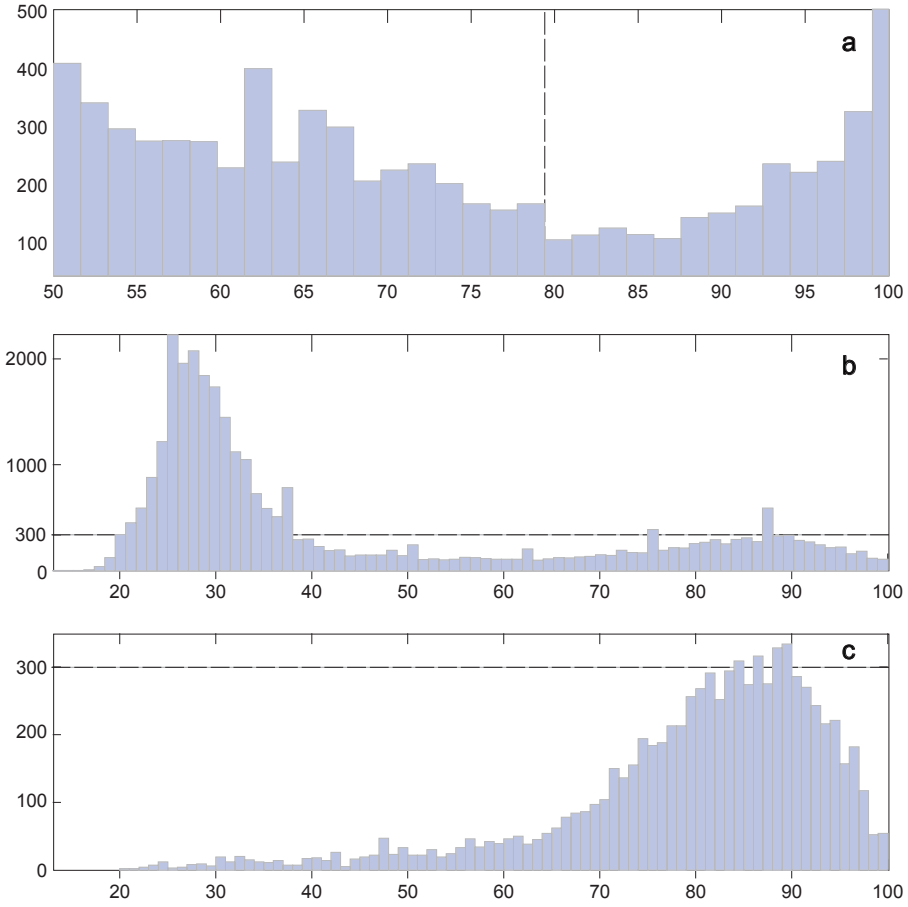

Figure 2. Pruning of sub-optimal matches. We used a relative threshold for filtering out BLAST hits that are unlikely to represent orthologous relationships. **a.** Histogram of amino-acid percent identity of secondary matches as compared to the strongest amino-acid percent identity found amidst matches 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.

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 (Figure 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 pre-processing 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 (Figures 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 super-family or proteins of similar function but whose separation well-precedes the divergence of the species.

The distribution of the remaining matches peaks between 70% and 95% amino-acid identity (Figure 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 only shows 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 pre-processing step increased the number of proteins with one-to-one correspondence by 60% (from 2790 to 4648 for *S. paradoxus*; from 2547 to 4113 for *S. mikatae*; from 2523 to 4052 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 (Figure 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 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}.

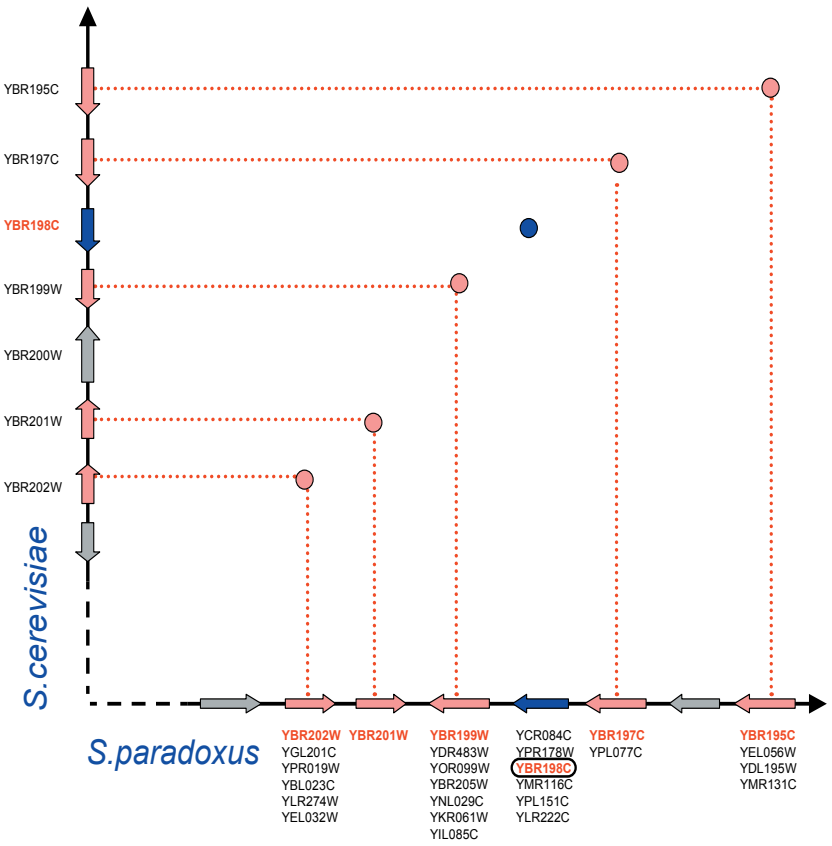

Figure 3. Using synteny information. Within blocks of conserved gene order (synteny), we preferentially keep those matches that conserve gene order. Annotated *S. cerevisiae* ORFs and standard names are shown along the y axis. Predicted *S. paradoxus* ORFs and their matches are shown along the x axis,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

We only considered 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 distance ***d*** for considering two neighboring genes as proximal to 20kb. This parameter was set to match the estimated density of unique syntenic anchors. Since *S. cerevisiae* genes are on average 2kb 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 only considered 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 non-maximal 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 **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 $\mathbf{D_{100}}$ (Figure 1, bottom panels).

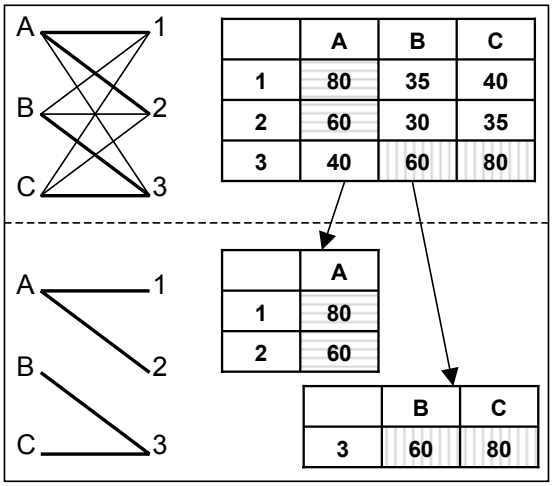

Figure 4. Example of Best Unambiguous Subsets (BUS). A BUS is a connected component of the graph constructed using only best matches, both forward and reverse. It defines a set of genes that can be isolated from the bipartite graph while preserving all potentially orthologous matches. Given the pairwise 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).

Intuitively, a BUS is a set of genes that starts as a Best Bi-directional 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 non-orthologous. 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.

Figure 5 illustrates the performance of the algorithm for the 6235 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 non-orthologous 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

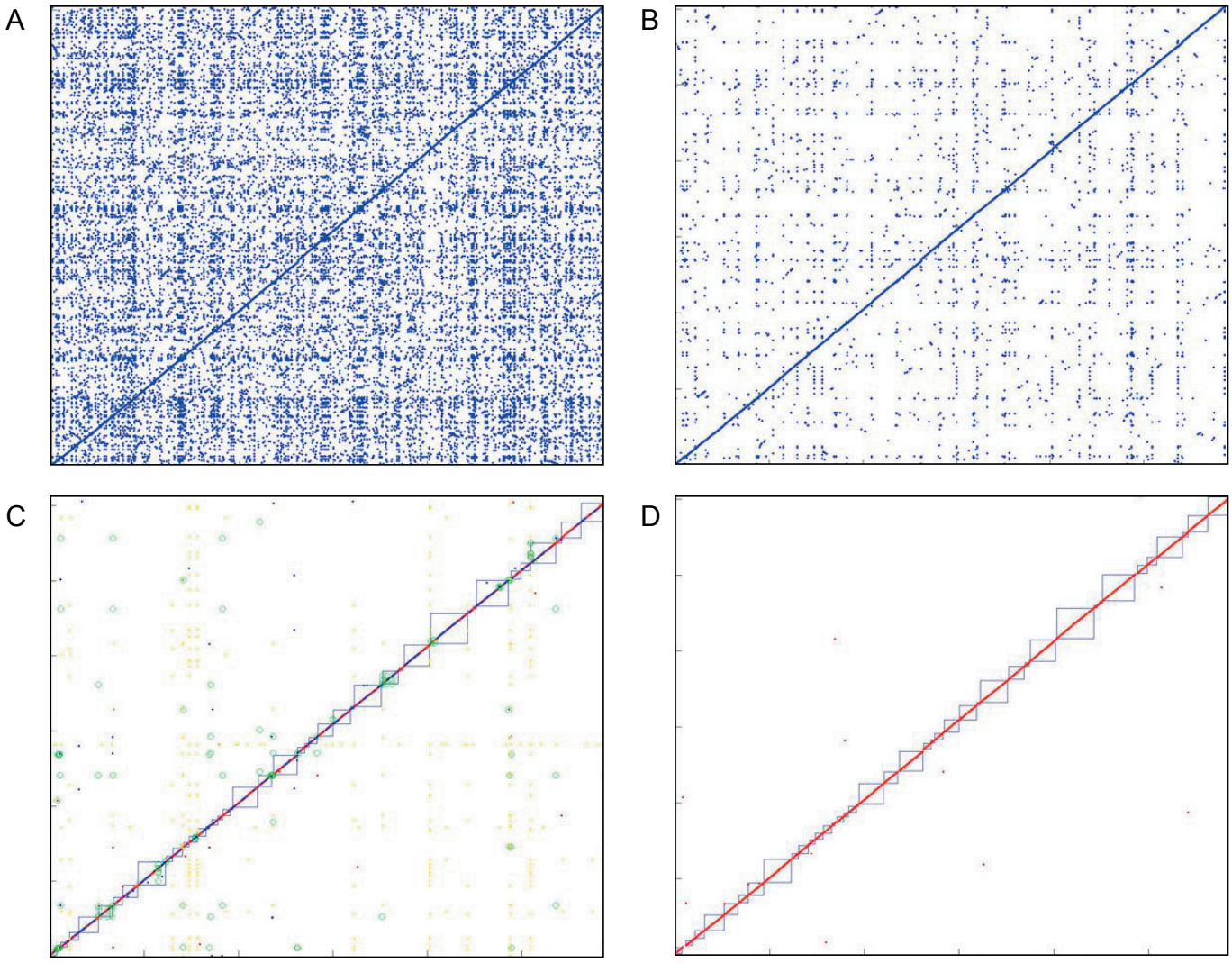

Figure 5. Performance of the algorithm. Dotplot representation of the bipartite graph. The 16 chromosomes of *S. cerevisiae* are stacked end-to-end along the y-axis and the scaffolds of *S. paradoxus* are shown along the x-axis. Every point (x, y) represents an edge between a *S. paradoxus* gene at position 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.

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 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 dotplot 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 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.

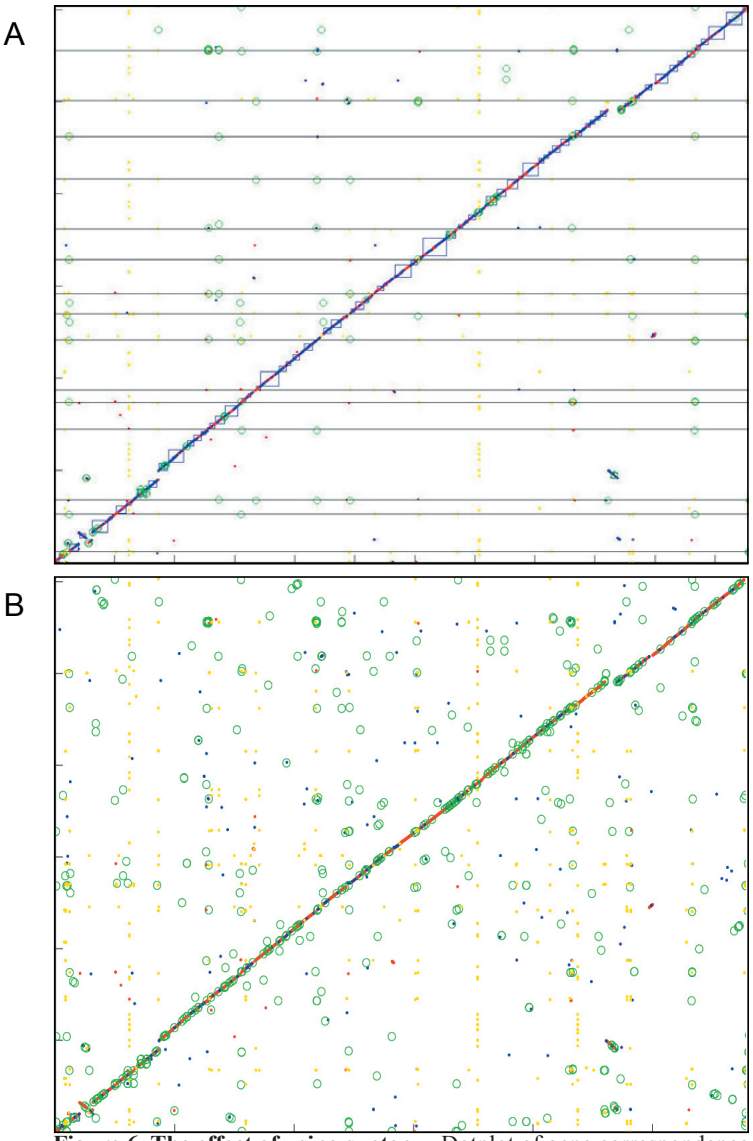

Figure 6. The effect of using synteny. Dotplot of gene correspondence between *S. cerevisiae* (y-axis) and *S. bayanus* (x-axis) with and without using synteny information. Blocks of conserved gene order (blue squares) help resolve additional ambiguities. Ambiguities remaining when the use of synteny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

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. yarrowii*, it successfully uncovered the correct gene correspondence between the two species that are more than 100 million years apart.

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 bi-directional 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 non-orthologous, 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 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 non-functional 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 (Intl_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 6062 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; Intl_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 genome-wide 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 and 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 genome-wide 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 piece-wise alignments adding gaps where appropriate such that the sequences in common between these alignments contain identical sets of gaps.

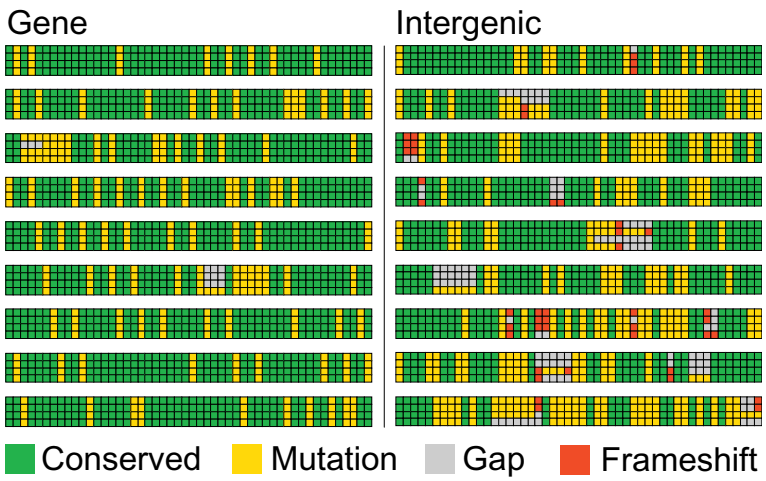

Figure 7. Patterns of change in genes and intergenic regions. Schematic representation of multiple sequence alignments in ORF YMR017W (left) and neighboring intergenic region (right). Aligned nucleotides across the four species are shown as stacked squares, colored by their conservation: green for conserved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

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 non-synonymous 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 (Figure 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 non-conserved 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 ~2-fold. The difference becomes much greater when one considers the gapped positions in alignments, representing insertion and deletion events (indels). The proportion of indels 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 (Figure 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.

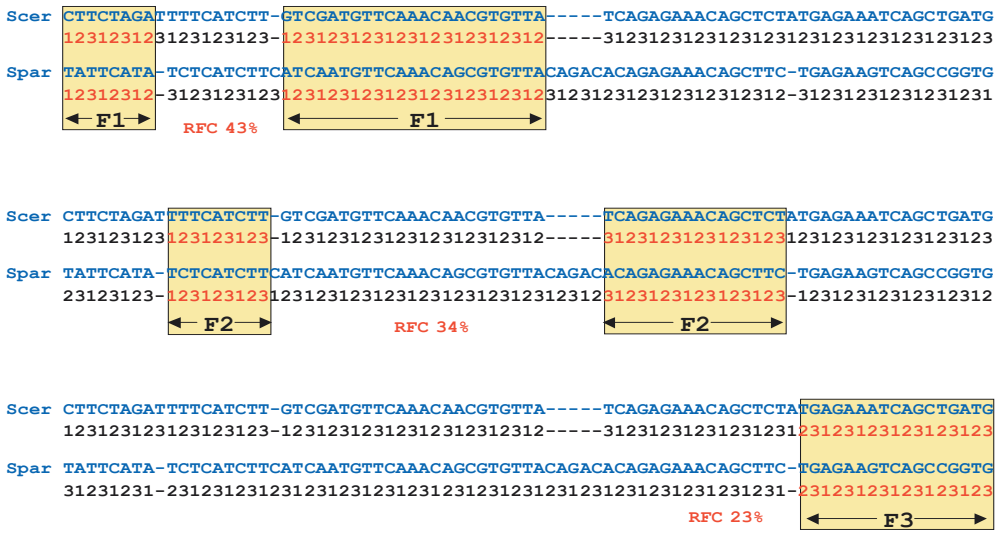

Figure 8. Reading Frame Conservation (RFC). Sample alignment between *S. cerevisiae* and *S. paradoxus* and frame numbering for the three possible frame numbering offsets. For each offset, matching frame indices are shown in red, and the portion of the alignment with agreeing reading frame is boxed. 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.

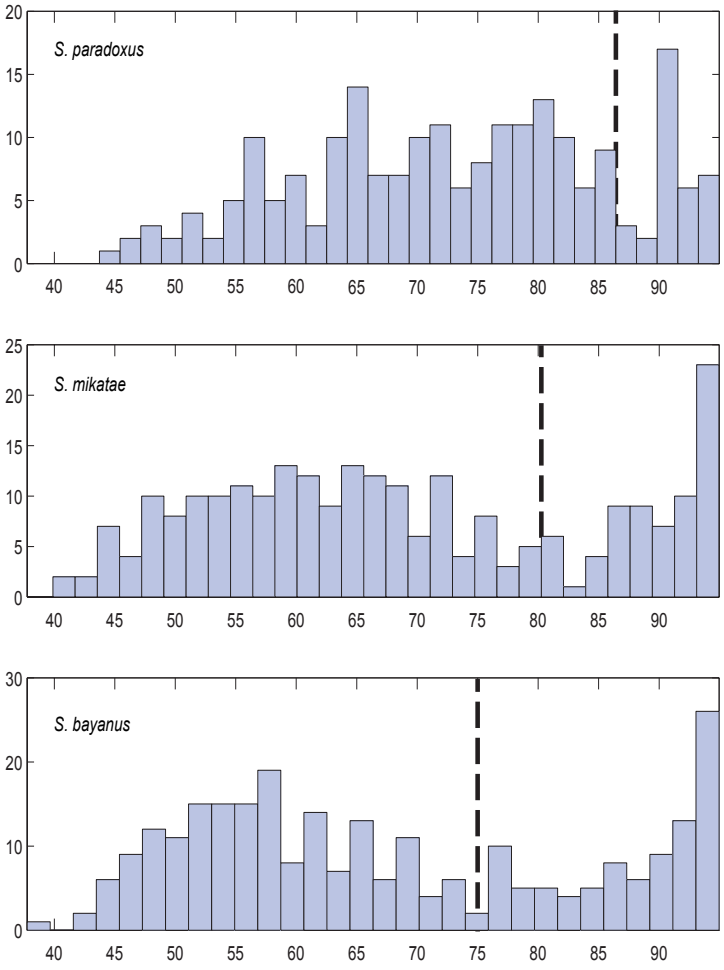

Figure 9. Bimodality of RFC test for hypothetical ORFs. Distribution of RFC scores for hypothetical ORFs in each of the species. The bimodality of the score distribution allows us to select a cutoff for each species above which an ORF is accepted as biologically meaningful and below which it is rejected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

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: 85% for *S. paradoxus*, 80% for *S.mikatae* and 75% for *S.bayanus* (Figure 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 and ORF, when information was available, or 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 3966 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 3966 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 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 seems likely to be explained by the fact that the ORF overlaps the promoters of other known genes.

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

Table 1. Performance of the RFC test. 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.

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 percent 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

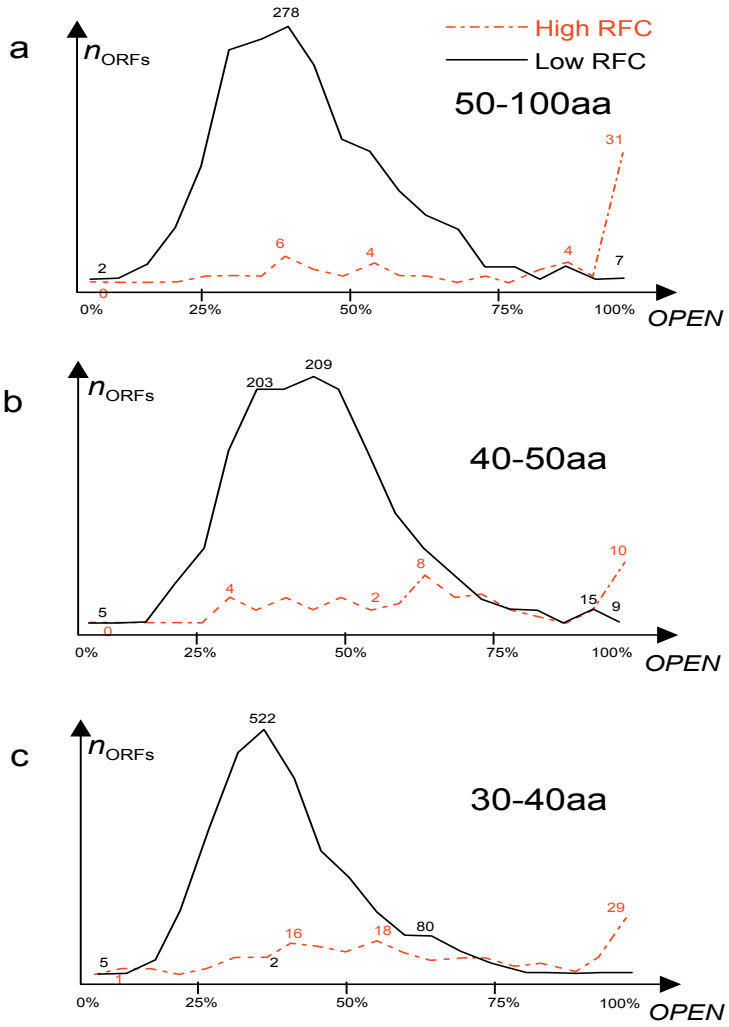

Figure 10. Power of RFC test to identify small ORFs. ORFs of decreasing lengths are tested using the RFC test. To evaluate the validity of the tested ORFs, we measure *OPEN*, namely the minimum percentage of the segment that is free of stop codons across the four species. The histogram of *OPEN* values 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 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.

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 (Figure 10a), and we inferred that the test has 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 (Figure 10b and 10c), and we concluded that the lack of indels within the small interval considered is not due to selective pressure, but instead 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 re-annotation 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*

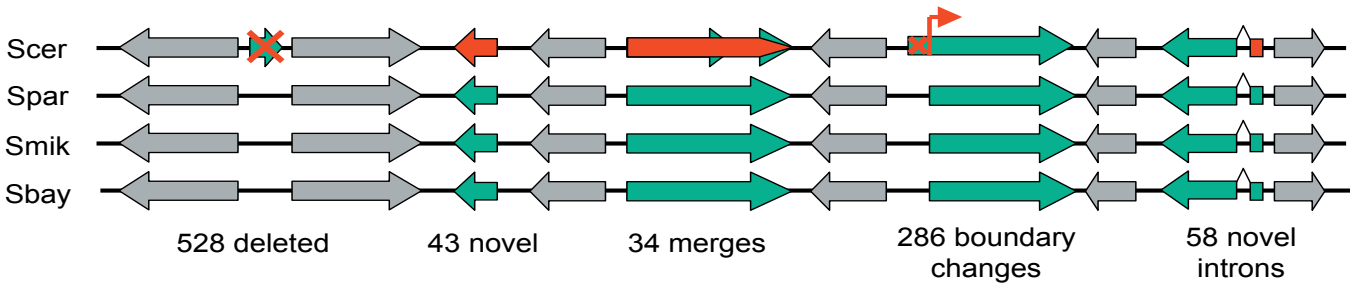

Figure 11. Revised gene catalogue. Our comparative analysis has revised the yeast gene annotation, affecting nearly 15% of all currently annotated genes. We found that 528 previously annotated ORFs are not real genes, identified 43 novel short ORFs that were previously overlooked, proposed merges of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

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 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 for conserved splicing signals. We conducted our search using 10 variants of splice donor signals (6-7bp) and 8 variants of branch site signals (7bp) 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 10bp of each other. We additionally required that branch and donor signals be no more than 600bp apart, which is the case for 90% of known *S. cerevisiae* introns. We then evaluated the multiple alignment surrounding the conserved 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 5538 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 mergeing). 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 (Figure 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 genome-wide 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 in (Stormo 2000). These were pioneered by Lawrence and coworkers (Lawrence et al. 1993), and made popular in software programs like 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 genome-wide 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 non-functional intergenic regions. This is largely due to the small number of functional instances of regulatory motifs, as compared to the large number of non-functional nucleotides. The discovery of regulatory motifs still relies heavily on extensive experimentation.

Comparative genomics provides a powerful way to distinguish regulatory motifs from non-functional patterns based on their conservation. Over evolutionary time, mutations accumulate in non-functional 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 genome-wide 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 genome-wide level, simultaneously observing all conserved and non-conserved 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. Genome-wide 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 (mini-motifs) to find strongly conserved motif cores, followed by a motif refinement and collapsing step that ultimately produced a small number of full motifs (Figure 12). We introduce three different conservation criteria (CC) for testing candidate motifs: overall genome-wide intergenic conservation (CC1), preference for intergenic conservation over genic conservation (CC2), and differential conservation in upstream-only vs. downstream-only regions (CC3). Using these three criteria, we tested a large set of short sequence patterns (mini-motifs), and selected those with non-random conservation. We extended these mini-motifs, searching for neighboring positions with significant non-random conservation, and collapsed the extended motifs hierarchically based on sequence similarity and genome-wide co-occurrence. This resulted in a dictionary of regulatory elements, discovered solely based on their genome-wide conservation.

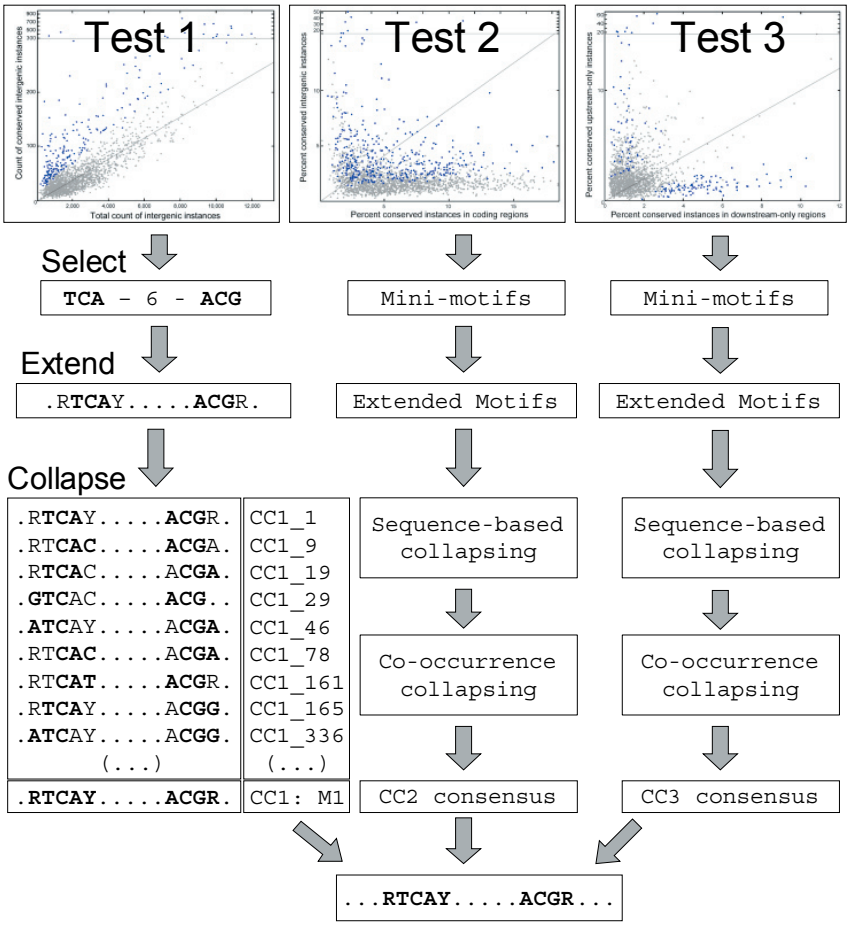

Figure 12. Genome-wide motif discovery. We select mini-motifs with significant non-random conservation according to three conservation criteria. We extended these by searching for neighboring bases that are preferentially conserved when a mini-motif is conserved. We collapsed mini-motifs with similar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

The final list of 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, 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. Genome-wide 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 non-random 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 known in yeast: either a single stretch of nucleotides typically 6 to 8 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 mini-motif 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 mini-motifs will serve as seeds in sequence space for constructing full motifs in the following section. Both Gal4 and Mbp1 can be expressed directly as mini-motifs, 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, mini-motifs allow us to explore motif space and follow strongly conserved seeds.

To explore sequence space, we enumerated all mini-motifs with gap size *m* between 0 and 21 nucleotides. For a given gap size there are 4^6=4096 possible sequences, but only 2080 unique mini-motifs, considering a motif and its reverse complement as the same motif. This resulted in a total of 45,760 distinct mini-motifs. We assumed that the large majority of these show a conservation typical of random sequences and thus mini-motifs 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 mini-motif to this basal conservation rate in order to evaluate the non-randomness of its conservation.

In order to derive conservation signatures that allow us to distinguish real motifs from non-functional 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 ~5-fold higher than for random mini-motifs (12.5% vs. 2.4%). Second, the Gal4 motif is preferentially conserved in intergenic regions rather than in genic regions (12.5% vs. 3%). By contrast, random motifs are less frequently conserved in intergenic regions than genic regions (3.1% vs. 7.0%), reflecting the higher overall level of conservation in genic regions. Thus, the relative conservation rate in intergenic vs. genic regions is ~11-fold higher for Gal4 than for random motifs. Third, the Gal4 motif shows a higher conservation rate in divergent vs. convergent intergenic regions (those that lie upstream vs. 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.

Evaluate conservation within:	Gal4	Controls	Test
(1) All intergenic regions	12.5%	2.4%	CC1
(2) Intergenic : coding	12:3	3:7	CC2
(3) Upstream : downstream	12:0	1:1	CC3

Table 2. Conservation properties of Gal4 motif inspire three tests for motif discovery. 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). **a**. We found that 12.5% of all intergenic instances of Gal4 were conserved, but only 2.4% for control motifs (a 6-fold enrichment). **b.** Moreover, coding instances of Gal4 diverged more frequently than for control motifs (3% vs. 7%), providing an 11-fold enrichment for Gal4 when comparing intergenic to coding conservation. **c.** Finally, Gal4 showed a preference for conservation in intergenic regions that are upstream of both flanking genes, but control motif show no such preference. We select motif cores (mini-motifs) based on three conservation criteria inspired by these conservation properties of known motifs.

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 mini-motifs that show similar conservation patterns by enumerating all 45,760 mini-motifs.

For every mini-motif, we obtained the total count of intergenic instances ***i***, and the count of perfectly conserved intergenic instances ***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 mini-motifs ***r***, and use it to evaluate the non-randomness in the conservation of any one mini-motif. If we assume that the chance conservation of any one mini-motif 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 ***r***, the basal mini-motif conservation rate. The probability of observing ***ic*** successes out of ***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 ***ic***=12 conserved instances out of ***i***=96 total instances (12.5%). However, the 2080 mini-motifs 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 ***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 mini-motifs, to discover novel sequences with conservation patterns similar to Gal4. For each gap size, we estimated the basal conservation rate ***r*** in two steps. In the first step, we selected non-outlier mini-motifs as those within three standard deviations of the log-average ratio ***ic/i***. We then calculated for every motif the binomial probability ***p*** of observing ***ic*** successes out of ***i*** trials, given parameter ***r***. We finally assigned a z-score ***S*** to every mini-motif as the number of standard deviations away from the mean of a normal distribution that correspond to tail area ***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 mini-motifs with striking conservation scores. For example, the top-ranked mini-motif CGT-6-TGA was perfectly conserved in 397 intergenic regions (33% of all instances) and showed a z-score of 37 standard deviations. This mini-motif corresponds to the central portion of RTCRYnnnnnACG, the motif recognized by the Abf1 transcription factor. Mini-motifs 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 genome-wide conservation. Additionally, a number of novel motifs were discovered (see below).

To estimate the total number of significant mini-motifs, we studied the tails of the z-score distribution (Figure 13, top panel). We found that the right tail of the distribution (strongly conserved mini-motifs with positive z-scores) was much longer than the left tail (weakly conserved mini-motifs with negative z-scores). For a cutoff of 5 sigma, as many as 1190 mini-motifs showed z-scores of +5 sigma or greater, as compared to only 25 mini-motifs

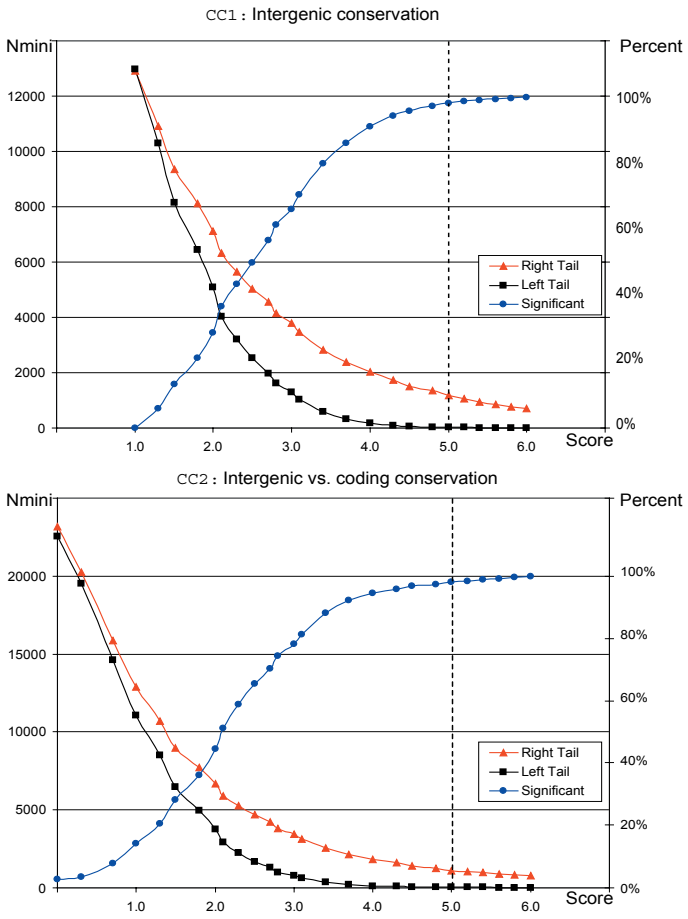

Figure 13. Estimating number of significant mini-motifs. We compare the right tail (R) and the left tail (L) of the distribution of mini-motif conservation scores. The ratio (R-L)/R allows us to the estimate the number of significant points at a given score cutoff. We find that 94% of scores 5 sigma away from the mean are on the right tail for intergenic conservation score *CC1*, and 97% for intergenic vs. coding conservation score *CC2*. The vast majority of mini-motifs selected at this cutoff are non-random.

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 mini-motifs in the left tail, and R is the count of mini-motifs at the right tail of the distribution. At a cutoff of 5 standard deviations, we therefore estimate that 97.5% of the 1190 mini-motifs are non-random in their conservation enrichment.

3.2.2: Conservation criterion 2: Intergenic vs. 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 mini-motif by comparing its conservation to the conservation of a control population of random mini-motifs.

We quantified the conservation preference for intergenic regions over coding regions based on counts of conserved and non-conserved 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 mini-motif. We searched for mini-motifs for which $\frac{ic}{i} >> \frac{gc}{g}$, which can alternatively be written as $\frac{ic}{gc} >> \frac{i}{g}$. Intuitively, these are mini-motifs 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 mini-motif instances are found in intergenic regions (as expected, since intergenic regions account for roughly ¼ of the yeast genome). Additionally, we found that motifs with similar GC content have similar ***b/a*** ratios. Hence, b_x/a_x for mini-motif ***x*** is similar to b_{avg}/a_{avg}, the average ratio for mini-motifs containing the same number of C or G bases as motif ***x***. Thus, for mini-motif ***x*** we can approximate $b_x \sim f_{avg} \cdot a_x$, where f_{avg} is the log-average ratio ***b/a*** for mini-motifs 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 mini-motif ***x***.

To quantify the observed imbalance between ***ic*** and ***gc***, as compared to the expected ratio b_x for a given mini-motif, we used a binomial test as previously. We considered b_x to be the probability that any conserved mini-motif instance 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 vs. 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 mini-motifs that showed a preference for intergenic conservation. The top-scoring mini-motifs 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 mini-motifs 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 anti-correlation between genic and intergenic conservation: mini-motifs with the strongest intergenic conservation (average ***ic/i***=20% for the top 50 mini-motifs, and average ***ic/i***=8% for the top 500 mini-motifs) show weak genic conservation (average ***gc/g***=4% for the two sets, as compared to 7% for random mini-motifs).

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 mini-motifs, but show a longer right tail (Figure 13, bottom panel). We compared the right and left tails to estimate the specificity of 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 1110 motifs, as compared to 39 motifs for the left tail (negative z-scores). Hence at this cutoff, we expect 97% of the selected 1110 mini-motifs to be non-random.

3.2.3: Conservation criterion 3: Upstream vs. 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 vs. downstream conservation of a mini-motif, we 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 rate of conservation. Thus, we found that the two ratios ***uc/u*** and ***dc/d*** are both similar to ***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 mini-motifs corresponding to transcription factors that are known to bind upstream of genes showed upstream-only conservation. For example, the Abf1 and Reb1 mini-motifs showed a strong conservation in upstream-only regions (***uc/u***=50%), and weak conservation in downstream-only regions (***dc/d***=2%).

Strikingly, we also discovered mini-motifs that showed a strong preference for downstream-only conservation rather than upstream-only. For example, the mini-motif TGT-1-AAT showed a strong conservation in downstream regions (***dc/d***=14%), and a weak conservation in upstream-only regions (***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 1089 mini-motifs 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 45760 tested motifs with such a score by chance alone, and hence estimate that 96% of the 1089 chosen mini-motifs to be non-random.

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 mini-motifs discovered, and created full motifs that better represent the full sequence specificity of regulatory motifs and transcription factors.

We used the mini-motifs discovered according to the three genome-wide conservation criteria as seeds in sequence space, that we then used to discover full motifs. We first extended each mini-motif by searching for surrounding positions that show non-random conservation. We found that different mini-motifs sometimes showed similar extensions and we developed methods for merging similar extended motifs based on their sequence similarity and their genome-wide co-occurrence. We finally developed methods for testing the genome-wide 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 genome-wide conservation.

3.3.1. Motif extension

We saw previously that the mini-motifs discovered were frequently only a subset of the known binding site for a particular factor (for example, the mini-motif CCG-1-GTA is only part of **CCG**G**GTA**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 mini-motifs by including surrounding bases that show additional non-random 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 mini-motif or up to four nucleotides away on either side of the mini-motif. For each position, we selected the best character extension amidst the fourteen 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 mini-motif 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 mini-motifs with gap-size m=3, and 266 extensions for m=*11*).

We evaluated the non-randomness 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 mini-motif and its neighborhood as defined previously. The *control set* contained the alignments and neighborhood of all "one-off" instances of the mini-motif, namely all instances where one base was not diverged on each half of the mini-motif. For mini-motif XYZ-*m*-UVW, the control set contained all instances of (notX)YZ-*m*-(notU)VW, (notX)YZ-*m*-U(not V)W and so on. By using a control as 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 $\boldsymbol{a/A} >> \boldsymbol{b/B}$. 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-scores of 3 standard deviations, and reported the extended motif.

A special case arose in extending mini-motifs that are reverse palindromes. These are mini-motifs 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 mini-motif 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 mini-motifs 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 mini-motif seeds. Mini-motif CGT-6-TGA was extended to Y**CGT**NNNNNR**TGA**Y, successfully discovering the additional degenerate bases of the Abf1 motif. The mini-motifs 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 mini-motif 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 mini-motif 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 genome-wide 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

Mini-motifs 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 genome-wide co-occurrence. 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 $\boldsymbol{p[j]}=[\boldsymbol{p_A}, \boldsymbol{p_C}, \boldsymbol{p_G}, \boldsymbol{p_T}]$ to each motif position $\boldsymbol{m[j]}$, representing the probability of occurrence of each of the four bases at that position (for example W=[A|T]=[½, 0, 0, ½] and D=[A|G|T]=[⅓, 0, ⅓, ⅓]). We then assigned a similarity score between two motifs $\boldsymbol{m_1}$ and $\boldsymbol{m_2}$, 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 $\boldsymbol{m_1[j]}$ and $\boldsymbol{m_2[k]}$ at every offset of each orientation. At a given offset, we summed the bits in common at each position, corresponding to the probability $\boldsymbol{p_1[j] \cdot p_2[k]}$ 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 pre-computed similarity scores of motifs across the two clusters. For every motif $\boldsymbol{m_i}$ in cluster ***c1***, we scored $\boldsymbol{S_{1\to 2}(m_i)}$, the similarity to the closest point in ***c2***. Similarly, for every motif $\boldsymbol{m_j}$ in cluster ***c2***, we scored $\boldsymbol{S_{2\to 1}(m_j)}$, 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 $S_{12}=\frac{1}{(i+j)}[\sum_{m_i \in c1} S_{1\to 2}(m_i) + \sum_{m_j \in c2} S_{2\to 1}(m_j)]$. 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 $\boldsymbol{m_i \in c1}$ and $\boldsymbol{m_j \in c2}$. A high-scoring similarity between motifs $\boldsymbol{m_i}$ and $\boldsymbol{m_j}$ may not result in a merge if the clusters that contain the two motifs are too dissimilar, and conversely a low-scoring similarity between $\boldsymbol{m_i}$ and $\boldsymbol{m_j}$ 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 mini-motifs discovered. The 1190 extended motifs discovered in test1 (CC1) clustered into 332 unique patterns, the 1110 motifs from test2 (CC2) into 269, and the 1089 motifs from test 3 (CC3) into 285 distinct patterns. The first 9 members of a cluster containing ABF1-like motifs from test1 are shown in Figure 12, with mini-motif 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.

3.3.3. Motif collapsing by genome-wide co-occurrence

Finally, we merged motifs from the three conservation tests together when they co-occurred in the same intergenic regions. This steps 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 only shared 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 co-occurrence score $\boldsymbol{C_{12}}$ between motifs $\boldsymbol{m_1}$ and $\boldsymbol{m_2}$ based on the respective counts of intergenic regions $\boldsymbol{r_1}$ and $\boldsymbol{r_2}$ containing conserved instances of $\boldsymbol{m_1}$ and $\boldsymbol{m_2}$ respectively, and the count of intergenic regions $\boldsymbol{r_{1\&2}}$ that contain conserved instances of both motifs. Given a total count $\boldsymbol{n}$ of all intergenic regions, and $\boldsymbol{r_1}$ intergenic regions containing motif $\boldsymbol{m_1}$, we evaluated the probability $\boldsymbol{p_{1\&2}}$ that a random set of $\boldsymbol{r_2}$ intergenic regions contains at least $\boldsymbol{r_{1\&2}}$ items in common with $\boldsymbol{r_1}$. This chance probability of observing at least such an overlap is given by the hypergeometric sum $p_{1\&2} = \sum_{k \geq r_{1\&2}} \binom{r_1}{k}\binom{n-r_1}{r_2-k} \Big/ \binom{n}{r_2}$. 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 co-occurrence 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). We then constructed consensus motifs for each of these and evaluated their genome-wide conservation to select a final list of strongly conserved genome-wide motifs.

3.3.4. Genome-wide conservation score for complex motifs

We previously developed methods for evaluating the genome-wide conservation of a mini-motif by comparing its conservation to that of similar mini-motifs. 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 mini-motif conservation, for every motif ***M***, we first constructed a control set of motifs, then used these to estimate a basal rate of conservation ***r***, and finally evaluated the non-randomness of the conservation of ***M***, given the basal rate ***r.***

For a given motif ***M*** of length ***L***, we constructed random control motifs ***R*** of the same length and degeneracy. For each position ***R[i]***, we selected a unique, 2-fold, 3-fold, or 4-fold degenerate base, depending on the degeneracy level of ***M[i]***. For example, if the motif base ***M[i]*** is W=[A|T] then the degeneracy level for ***R[i]*** will be 2, and we will select a 2-fold degenerate base for ***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 ***R*** we selected a random intergenic sequence ***S*** of length ***L*** in the *S. cerevisiae* genome, regardless of conservation. The nucleotide ***S[i]*** was used to select the degenerate base at ***R[i]***, such that ***S[i]*** appears with non-zero probability in ***R[i]***. For example, if we seek to select a 2-fold degenerate base for ***R[i]*** and ***S[i]***=G, then ***R[***i***]*** can be one of R=[G|A], S=[G|C], K=[G|T], all of which are 2-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 $\boldsymbol{R_1}$-$\boldsymbol{R_{20}}$ for a given motif ***M***, that 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 non-trivial, since we could no longer count exact matches and require perfect conservation as for mini-motifs.

We instead used a probabilistic framework to detect conserved motif instances. We interpreted every genome-wide motif ***M*** of length ***L*** as a probabilistic model, generator of sequences of length ***L*** over the alphabet {A,C,G,T}. We then evaluated for every genome position, the probability that the sequence was generated by motif ***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 search 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 ***t*** and obtained the count of all intergenic regions containing conserved instances of the motif with score at least ***t***. We similarly obtained the count of all intergenic regions within *S. cerevisiae* containing instances of the motif above threshold ***t***.

Depending on the threshold ***t***, the count of conserved and total instances of motifs ***M*** and $\boldsymbol{R_i}$ will vary. For a given threshold, we can evaluate the enrichment of genome-wide conservation of ***M*** as compared to the controls, but we had to choose a sensitive threshold. If the threshold ***t*** is too low, then too many instances will be found and the signal for ***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 lowest 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 genome-wide 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 genome-wide motifs (Kellis et al. 2003, Table 3)

3.4. Results of genome-wide motif discovery

We evaluated the performance of our methods by comparing our dictionary of 72 genome-wide motifs to a catalog of 55 previously known motifs and to a 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 genome-wide 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 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 genome-wide 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 mini-motifs from each conservation score and randomly-sorted mini-motifs. For a hypergeometric score of at least 10^{-5}, Figure 14 compares the number of mini-motifs that show category enrichment for increasingly larger sets of top-ranked mini-motifs in each test (CC1, CC2, CC3), as compared to a randomly sorted mini-motifs (RND). From the top 100 mini-motifs 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 1000 mini-motifs. Hence, for a cutoff of 10^{-5}, our search showed a 90-fold enrichment

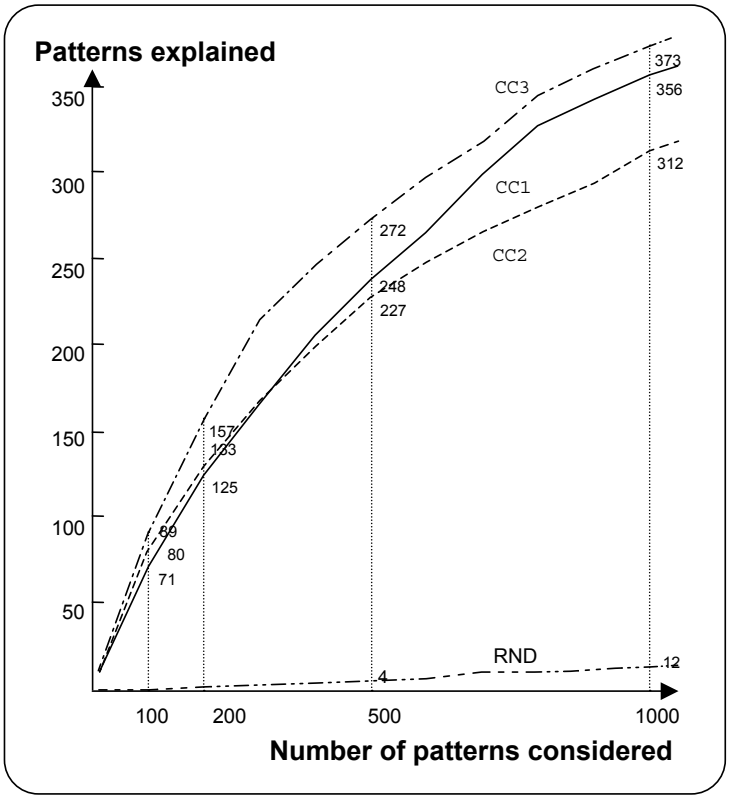

Figure 14. Selected mini-motifs show functional enrichment. Mini-motifs selected by any of the three tests (CC1,CC2,CC3) correlate with function 90 times more frequently than randomly chosen mini-motifs (RND).

in enriched mini-motifs as compared to random sequences.

We used this cutoff to assign a candidate function to the 72 genome-wide motifs. Most of the 36 genome-wide 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 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 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 post-transcriptionally. 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 mtDNA translational machinery and the mitochondrial outer membrane (Jacobs Anderson and Parker 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 related genes that we previously used to assign candidate functions to genome-wide 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 genome-wide search. We first enumerated mini-motifs, selecting the ones that show conservation specifically within the category. We then extended these mini-motifs 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 mini-motif by counting the conserved instances within the category (IN), and the conserved instances outside the category (OUT). We estimated the ratio r=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 r. We assigned a z-score to each mini-motif, as described in the genome-wide search (section 3.2.1).

We extended those mini-motifs 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 mini-motif instances within the category from conserved mini-motif instances outside the category, thus increasing our specificity.

We finally collapsed motifs of similar extension based on their sequence similarity. We did not use genome-wide co-occurrence as a collapsing criterion, since the motifs discovered were already biased to significantly co-occur 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 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 7 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 9 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 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.

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	TTWCCnwwwrGGAAA	35.9	T**TTCC-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**-TWTT	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**TR**A**WGYA		**YTGAAACA**	good	better (+)
Gal4	CGGnnnnnnnnnnCCG	16.1	**CGG**M---CW-Y--C**CCG**		**CGGnnnnnnnnnnCCGA**	good	better (+)
Swi6	ACGCGT	15.6	**WCGCGT**CGCGTY-C	good	**ACGCGT**	good	same
Pho4	CACGTG	14.2	TTGTA**CAC**T**T**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	CACACACACACACACA		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 (+)

Table 3. Category-based motif discovery. 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.

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 9 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 re-use across multiple categories

Category-based motif discovery contributes only a modest number of additional motifs beyond those found by genome-wide 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, 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 genome-wide 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 to 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.

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 non-conserved noise. Hence, in studying any genome, comparative analysis of closely related species can provide the basis for a global understanding of all functional elements.

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 Computer Science and Artificial Intelligence Laboratory 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.

REFERENCES

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

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

Bailey, T.L. and C. Elkan. 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., D.B. Jaffe, K. Stanley, J. Butler, S. Gnerre, E. Mauceli, B. Berger, J.P. Mesirov, and E.S. Lander. 2002. ARACHNE: a whole-genome shotgun assembler. *Genome Res* **12:** 177-189.

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

Blanchette, M., B. Schwikowski, and M. Tompa. 2002. Algorithms for phylogenetic footprinting. *J Comput Biol* **9:** 211-223.

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

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

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

Dujon, B., D. Alexandraki, B. Andre, W. Ansorge, V. Baladron, J.P. Ballesta, A. Banrevi, P.A. Bolle, M. Bolotin-Fukuhara, P. Bossier, and 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., M.A. Harris, K. Dolinski, C.A. Ball, G. Binkley, K.R. Christie, D.G. Fisk, L. Issel-Tarver, M. Schroeder, G. Sherlock, A. Sethuraman, S. Weng, D. Botstein, and J.M. Cherry. 2002. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic 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 M.B. Eisen. 2002. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol* **3:** RESEARCH0059.

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

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

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

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

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

Hughes, J.D., P.W. Estep, S. Tavazoie, and G.M. Church. 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.

Intl_Human_Genome_Sequencing_Consortium. 2001. Initial sequencing and analysis of the human genome. In *Nature*, pp. 860-921.

Intl_Mouse_Genome_Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-562.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Oeltjen, J.C., T.M. Malley, D.M. Muzny, W. Miller, R.A. Gibbs, and J.W. Belmont. 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 E.M. Rubin. 2001. Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet* **2:** 100-109.

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

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

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

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

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

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

Thompson, J.D., D.G. Higgins, and T.J. Gibson. 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.

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

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

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

Wood, V., K.M. Rutherford, A. Ivens, M.-A. Rajandream, and B. Barrell. 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 M.Q. Zhang. 1999. SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics* **15:** 607-611.