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

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

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

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

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

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

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

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

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

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

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

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

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

Genome structure and evolution

Genome sequencing

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

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

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

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

Genome alignment

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

Genome evolution at large scale

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

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

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

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

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

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

Genome evolution at the nucleotide level

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

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

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

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

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

Identification of genes

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

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

Public yeast gene catalogue

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

Reading frame conservation test

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

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

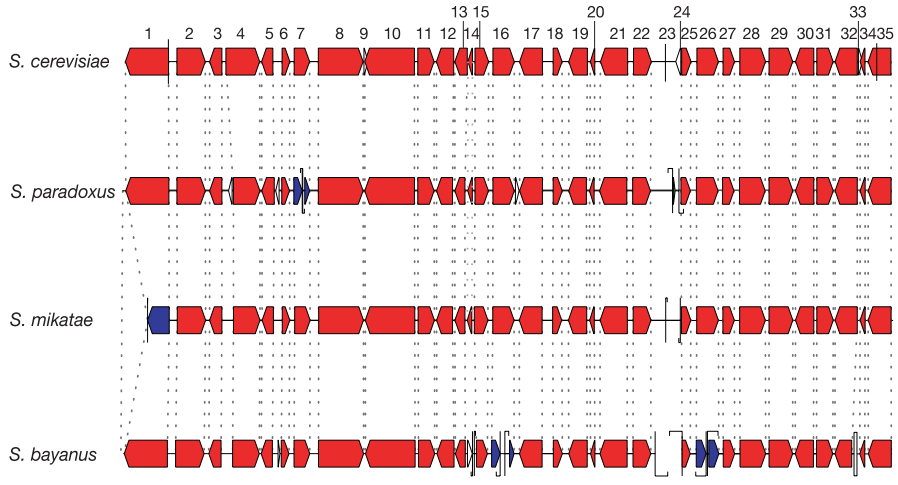
The RFC scores show a clearly bimodal distribution for each of the three species, allowing a decision threshold to be defined for each species to accept an ORF as biologically meaningful or reject an ORF as spurious; that is, occurring by chance. Each species thus 'votes' on the validity of each ORF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

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

Evaluating ORFs

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

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



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

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

to have accumulated enough frameshifts to allow definitive rejection.

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

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

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

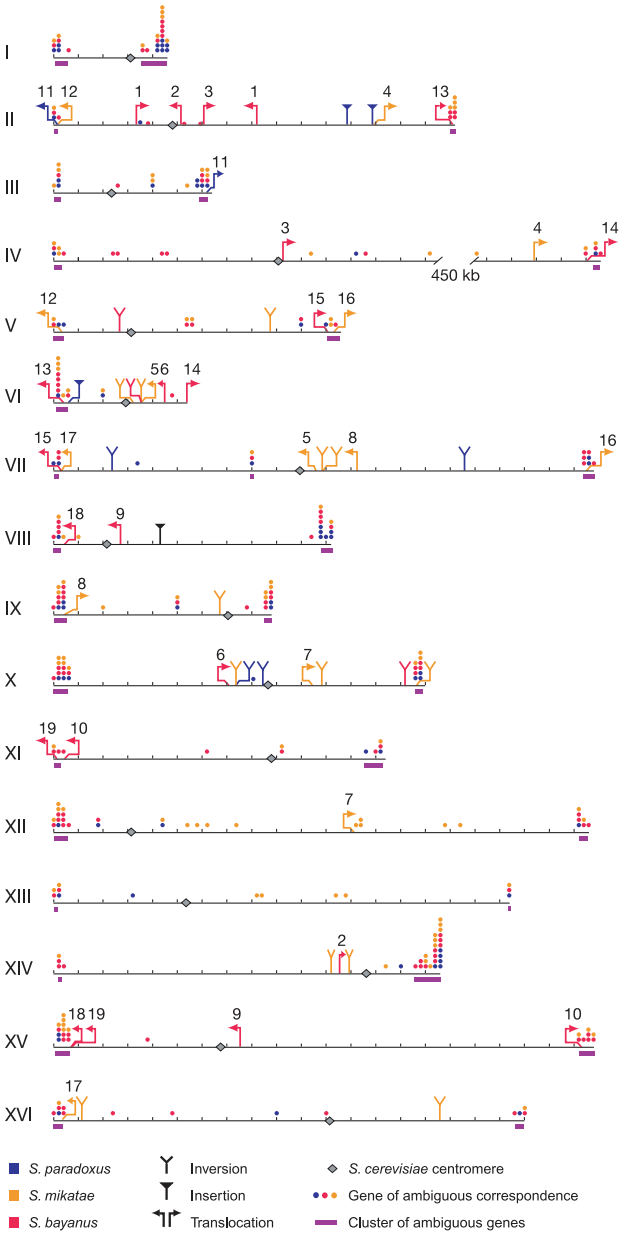


Figure 2 Genome evolution. Locations in the *S. cerevisiae* genome of chromosomal rearrangements in each of the three species are shown by vertical bars with the symbol at the top indicating the type of rearrangement (inversions, insertions and translocations) and colour indicating the species. Tick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

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

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

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

Revised yeast gene catalogue

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

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

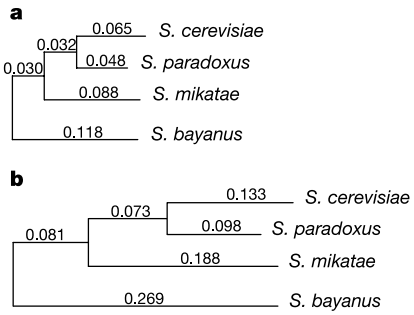


Figure 3 Evolutionary tree of the four yeast species. Branch lengths denote number of substitutions per site in genic (**a**) and intergenic (**b**) sequence. Pairwise distances were calculated using the Kimura two-parameter model and trees were constructed using the UPGMA method. The relative rate of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

Analysis of small ORFs

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

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

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

Defining gene structure

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

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

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

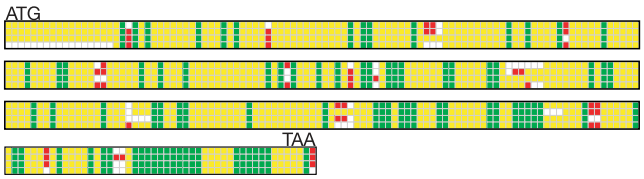


Figure 4 Spurious ORF rejected by RFC test. Schematic representation of the multiple sequence alignment of ORF YDR102C. Aligned nucleotides across the four species are shown as stacked squares (*S. cerevisiae*, S. *paradoxus*, *S. mikatae* and *S. bayanus*, respectively) coloured by their conservation: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

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

New introns

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

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

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

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

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

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

Rapid and slow evolution of genes

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

Species-specific genes

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

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

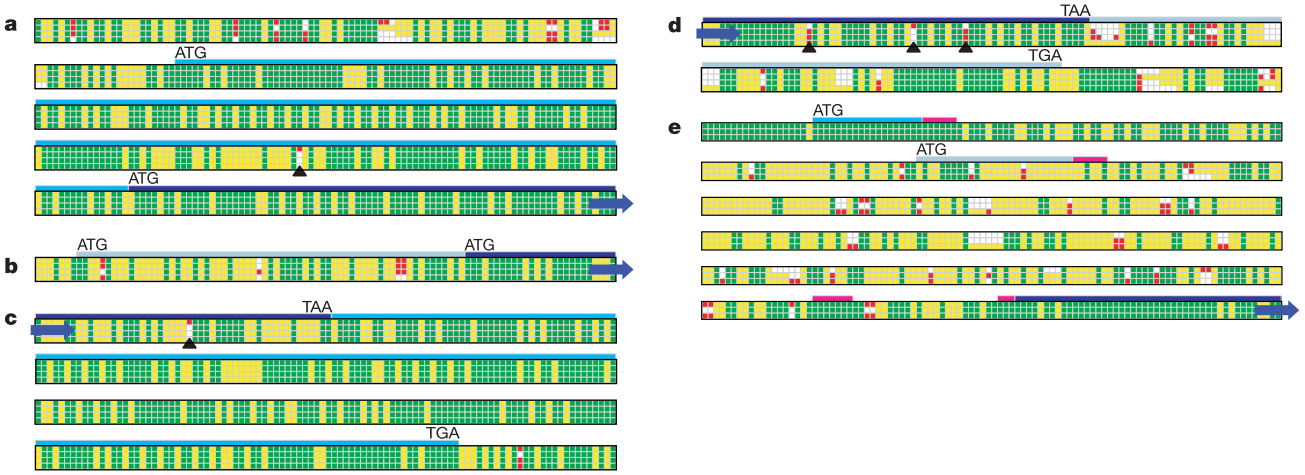


Figure 5 Examples of proposed changes in gene structure. Representation of alignments as in Fig. 4 (see Supplementary Information for full nucleotide alignments). Exon locations are indicated by overhead bars. Dark blue bars indicate portions of annotated exons unchanged in our analysis, and dark blue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

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

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

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

Rapidly evolving genes

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

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

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

Evidence of rapid protein change

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

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

Slowly evolving genes

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

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

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

Genome-wide identification of regulatory elements

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

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

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

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

Conservation of the Gal4-binding site

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

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

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

Catalogue of known motifs

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

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

Methodology for genome-wide motif discovery

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

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

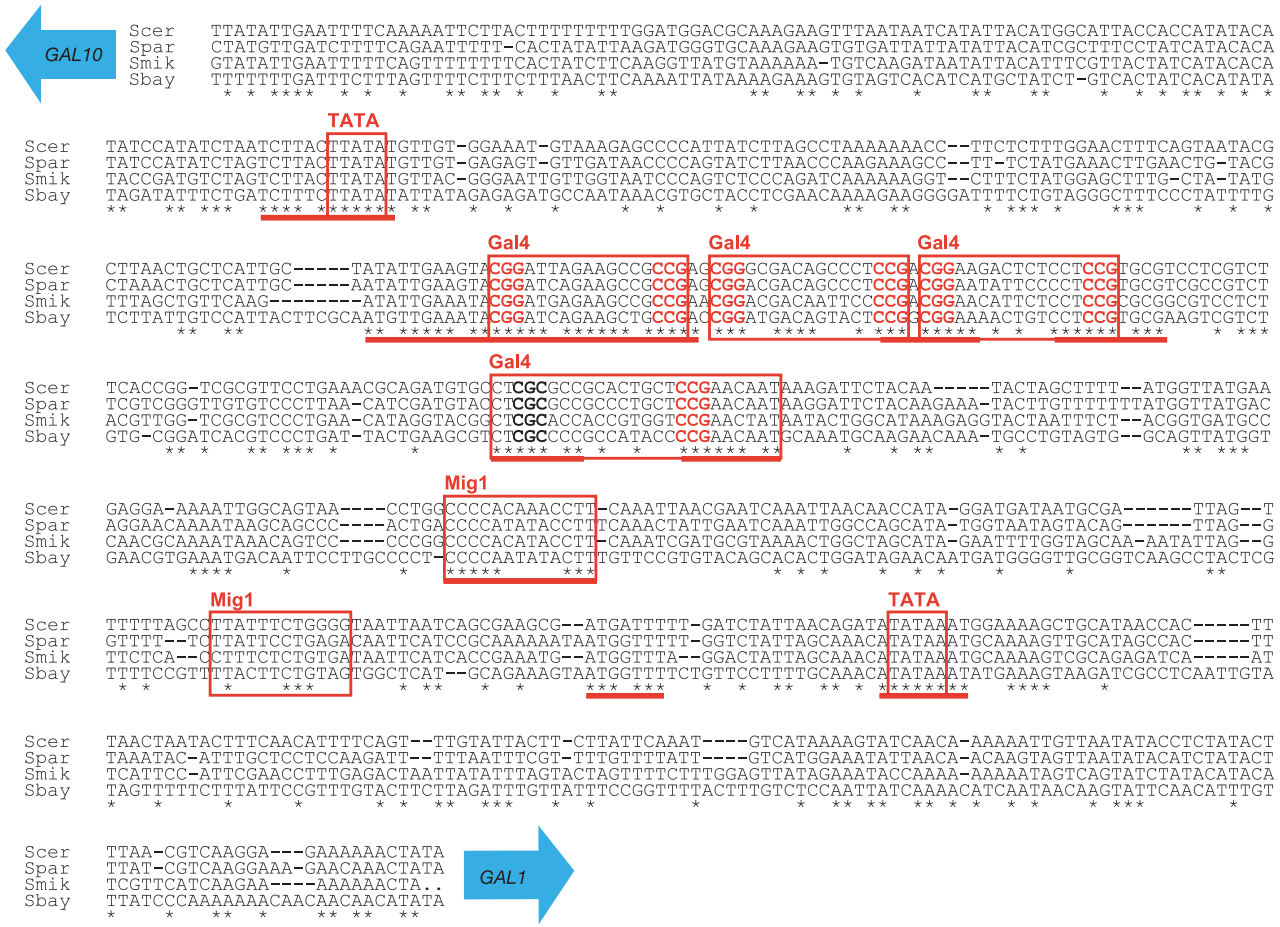


Figure 6 Conservation in the *GAL1–GAL10* intergenic region. Multiple alignment of the four species shows a strong overlap between functional nucleotides and stretches of conservation. Asterisks denote conserved positions in the multiple alignment. Blue arrows denote the start and transcriptional orientation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

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

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

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

Results of genome-wide motif discovery

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

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

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

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

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

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

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

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

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

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

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

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

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

Inferring function of genome-wide motifs

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

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

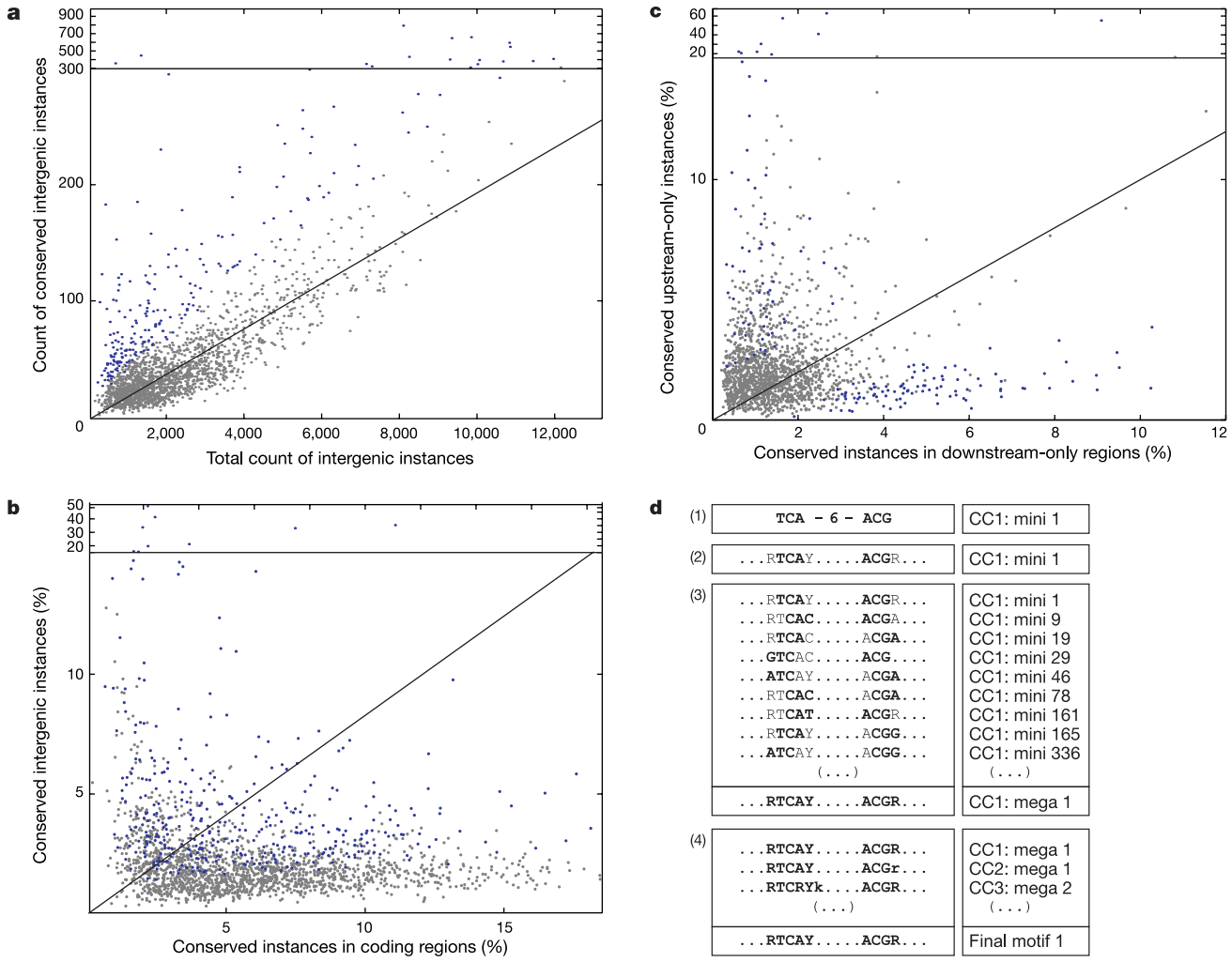


Figure 7 Distribution of motifs by conservation score. **a**–**c**, Points denote individual mini-motifs analysed with respect to the three conservation properties described in the text (CC1, intergenic conservation (**a**); CC2, intergenic–genic conservation (**b**); CC3, upstream–downstream conservation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

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

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

Table 3 **Discovered motifs**

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

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

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

Examples of new motifs

We describe a few of the new motifs below.

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

Category-based identification of regulatory elements

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

Known motifs

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

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

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

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

New motifs

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

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

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

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

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

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

Combinatorial control

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

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

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

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

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

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

Discussion

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

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

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

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

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

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

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

Methods

Strain information

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

Determining one-to-one orthologous ORFs

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

Motif conservation score

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

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

Conservation criteria

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

Constructing full motifs

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

Additional methods

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

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

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

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

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

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

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

