Article

Surveying *Saccharomyces* Genomes to Identify Functional Elements by Comparative DNA Sequence Analysis

Paul F. Cliften,[1] LaDeana W. Hillier,[2] Lucinda Fulton,[2] Tina Graves,[2] Tracie Miner,[2] Warren R. Gish,[1,2] Robert H. Waterston,[1,2] and Mark Johnston[1]

[1]*Department of Genetics and* [2]*Genome Sequencing Center, Washington University School of Medicine, St. Louis, Missouri 63110, USA*

Comparative sequence analysis has facilitated the discovery of protein coding genes and important functional sequences within proteins, but has been less useful for identifying functional sequence elements in nonprotein-coding DNA because the relatively rapid rate of change of nonprotein-coding sequences and the relative simplicity of non-coding regulatory sequence elements necessitates the comparison of sequences of relatively closely related species. We tested the use of comparative DNA sequence analysis to aid identification of promoter regulatory elements, nonprotein-coding RNA genes, and small protein-coding genes by surveying random DNA sequences of several *Saccharomyces* yeast species, with the goal of learning which species are best suited for comparisons with *S. cerevisiae*. We also determined the DNA sequence of a few specific promoters and RNA genes of several *Saccharomyces* species to determine the degree of conservation of known functional elements within the genome. Our results lead us to conclude that comparative DNA sequence analysis will enable identification of functionally conserved elements within the yeast genome, and suggest a path for obtaining this information.

Identifying functional elements in DNA sequence is a significant challenge. It is difficult enough to predict correctly protein-coding genes; an even greater challenge is to identify functional sequences that do not code for protein, such as sequences regulating gene expression, sequences governing chromosome replication, structure and stability, and sequences of nonprotein-coding RNAs. This task is complicated by the diverse nature of these sequence elements. For example, sequences that regulate gene expression are usually short, often independent of orientation, and can reside at varying distances from their target gene. Genes encoding RNAs can be difficult to identify because they contain few hallmarks and because their folded structure, rather than their primary sequence, dictates their function.

Because functional sequences are maintained in evolution, they often can be recognized by their conservation among different organisms (Tagle et al. 1988; Hardison et al. 1997). This approach has most often been used to identify functional sequences in proteins, because they are relatively large and evolve relatively slowly, causing their functional elements to remain conserved in diverse organisms. In a few cases potential nonprotein-coding functional sequence elements have been identified by comparing DNA sequences of relatively diverged species (usually from different genera) (Venkatesh et al. 1997; Gellner and Brenner 1999; Wentworth et al. 1999; Gelfand et al. 2000; Kent and Zahler 2000; McGuire et al. 2000). The comparative approach for finding functional nonprotein-coding sequences will likely be more productive with DNA sequences of relatively closely related organisms. We sought to gain experience with this by comparing the DNA sequences of yeasts of the *Saccharomyces* genus because (1) the complete *Saccharomyces cerevisiae* genome sequence is available for comparison (Goffeau et al. 1996); (2) the genomes of *Saccharomyces* species are relatively small, with relatively compact noncoding sequences (~30% of the genome); (3) their phylogeny is well characterized, with many related species at various evolutionary distances; and (4) it is possible to test the functionality of sequences with relatively simple and incisive experiments.

The *Saccharomyces* genus is composed of three subgroups (Barnett 1992). The sensu stricto species are physiologically similar to *S. cerevisiae*, are capable of forming stable diploids with each other, and have a very similar karyotype. The sensu lato and petite-negative *Saccharomyces* species are quite diverged from *S. cerevisiae*. They have significantly different physiological characteristics, seldom form diploids with *S. cerevisiae*, and have a smaller number of chromosomes (Petersen et al. 1999). To identify which *Saccharomyces* species are optimally diverged from *S. cerevisiae* for comparative DNA sequence analysis, and to evaluate the number of sequences that must be compared to obtain useful information, we determined the sequence of randomly generated genomic clones of seven *Saccharomyces* species. In addition, we determined the DNA sequences of the regulatory regions of a few specific genes and of a few genes encoding nonprotein-coding RNAs from several *Saccharomyces* species to determine the degree of conservation of known functional elements in the genome. Our results suggest that comparative DNA sequence analysis will enable identification of many functional gene regulatory elements and other nonprotein-coding sequences, and suggest a path for obtaining this information.

Corresponding author.
E-MAIL mj@genetics.wustl.edu; FAX (314) 362-2985.
Article and publication are at www.genome.org/cgi/doi/10.1101/gr.182901.

RESULTS

Sequence Summary of *Saccharomyces* Species

We obtained sequence from ~1000 genomic DNA clones from each of four species of the sensu stricto group (*Saccharomyces bayanus*, *Saccharomyces paradoxus*, *Saccharomyces cariocanus*, and *Saccharomyces mikatae*) (Naumov et al. 2000) and from >2000 clones from each of two species of the sensu lato group (*Saccharomyces castellii* and *Saccharomyces unisporus*) and the petite-negative species *Saccharomyces kluyveri* (Table 1). Over 4.3 Mb of unassembled sequence was obtained, a sample sufficient for genomic exploration of the *Saccharomyces* genus.

Comparisons of the nucleotide sequences (using BLASTN, Table 1) and inferred protein sequences (using BLASTX, Table 2) to *S. cerevisiae* sequence are instructive for determining the relative genetic distance of the species from *S. cerevisiae* and for identifying which species are likely to yield the most data in sequence comparisons. Most (94.5%–98.2%) of the sensu stricto sequences align to *S. cerevisiae* coding or noncoding sequence using BLASTN. As expected, the sequences are most similar to *S. cerevisiae* DNA sequence in protein-coding genes, but the sequences of *S. paradoxus* and *S. cariocanus* are also highly conserved within nonprotein-coding regions. *S. mikatae* and *S. bayanus* have lower similarity and yield significantly fewer alignments to nonprotein-coding regions of the genome using the same BLASTN parameters (Table 1).

The sensu lato and petite-negative species seem to lie a similar evolutionary distance from *S. cerevisiae*, because the average identities of their protein-coding DNA (Table 1) and predicted proteins (Table 2) to *S. cerevisiae* sequences are similar for each sensu lato species. These species are more diverged from *S. cerevisiae* than we expected from ribosomal RNA comparisons. Many presumably orthologous DNA sequences (identified by aligning the protein-coding sequences using BLASTX) do not align to *S. cerevisiae* sequence with BLASTN, and many of those that do align are near the lower limits of detection by BLASTN.

Although most of the proteins of the sensu stricto group are >80% identical to their *S. cerevisiae* homologs, 29 alignments show <50% identity (Fig. 1A). These low similarity alignments may indicate rapidly evolving proteins or genes that duplicated in other *Saccharomyces* species (but not in *S. cerevisiae*), allowing one copy to diverge from its *S. cerevisiae* ortholog. A large number of proteins of the sensu lato species (>30%) are <50% identical to their *S. cerevisiae* homologs (Fig. 1B), emphasizing the substantial divergence of these species from *S. cerevisiae*.

We noted many discrepancies in open reading frame (ORF) boundaries in the different sensu stricto species (data not shown). Some of these are likely caused by sequencing errors (most likely in our sequence, but potentially in the *S. cerevisiae* sequence); some may possibly be attributable to real DNA sequence differences between these closely related species. ORF length polymorphisms are even more common in the sensu lato alignments, which is not surprising considering the relatively greater divergence of these species from *S. cerevisiae*.

Table 1. Summary of BLASTN Comparisons of *Saccharomyces* Species' DNA Sequences to *S. cerevisiae* DNA Sequence

Species	# of sequences	% identity CODING	% identity NON-CODING	% of sequences aligning CODING/NON
S. paradoxus	728	88.3	81.6	77.9/36.3
S. cariacanus	867	88.0	80.5	77.5/39.1
S. mikatae	1136	83.6	74.8	79.9/27.9
S. bayanus	851	79.7	73.6	65.7/19.5
S. castellii	2290	70.3	62.5*	48.9/1.7*
S. kluyveri	2145	70.2	72.8*	38.6/1.8*
S. unisporus	2357	69.2	67.6*	40.4/2.2*

Coding sequences are all ORFs known or predicted to encode protein. The remaining yeast genomic sequences (minus the sequences encoding RNA genes, transposable elements, and CEN sequences) were considered non-coding. BLASTN comparisons of sensu stricto yeast species used MATCH/MISMATCH parameters M = 5, N = 5 with the default gap penalties. This change was made to more accurately reflect the similarity of the sequences and equals a PAM distance of ~30 (States et al. 1991). Using these strict MATCH/MISMATCH parameters, only significant alignments were obtained. Default parameters were used for sensu lato and petite-negative comparisons to *S. cerevisiae* using a wordlength (W) of 5. The dust filter was used to mask low complexity nucleotide sequences. Alignments with a *P*-value <5.0e-06 were counted as significant.
*Many of the alignments are to repetitive sequences (ribosomal rDNA repeat alignments were not counted).

Table 2. Summary of BLASTX Comparisons between *Saccharomyces* Species' Protein Sequences and *S. cerevisiae* Protein Sequence

Species	# of sequences	% of sequences aligning	% sequence identity
S. paradoxus	728	78.3	91.2
S. cariacanus	867	78.8	90.3
S. mikatae	1136	81.8	87.4
S. bayanus	851	83.2	83.2
S. castellii	2290	71.0	60.6
S. kluyveri	2145	63.7	59.8
S. unisporus	2357	55.0	59.6

BLASTX comparisons of protein coding sequences included only the highest scoring pair in most cases. Additional HSPs were included only if the *S. cerevisiae* encoded proteins were adjacent on a chromosome and could be spanned by a sequence read (123 of the sensu stricto sequences and 104 of the sensu lato and petite-negative sequences aligned to adjacent proteins). Sensu stricto comparisons were performed with a PAM matrix of 40 whereas sensu lato and petite-negative comparisons used the default BLOSUM62 matrix. Several additional alignments could be generated using alternative matrices (such as PAM120) and altered GAP penalties, but essentially equal numbers of alignments were generated using less stringent matrices since these parameters often miss short alignments near the ends of the proteins. The seg and xnu filters were used to mask low complexity protein coding regions. Essentially all of the alignments are significant when the filters are used.

Species-Specific Sequences

A small percentage of the sequences of sensu stricto species have no significant similarity to any *S. cerevisiae* sequence at the nucleotide or amino acid levels, regardless of the BLAST parameters that were used. Several of these sequences are pre-

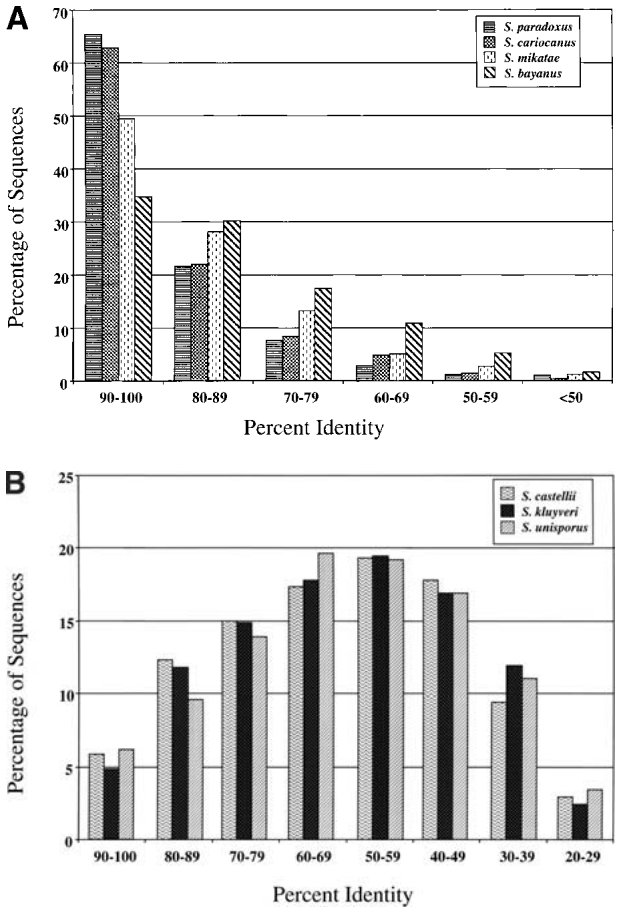

Figure 1 Summary of protein sequence identities between *Saccharomyces* sequences and their *S. cerevisiae* homologs. (*A*) Sensu stricto BLASTX comparisons to *S. cerevisiae* using a PAM40 matrix. (*B*) Sensu lato and petite-negative comparisons using the BLOSUM62 matrix.

dicted to encode proteins similar to those found in other species. For example, we identified a Ty5 transposon protein that is unique to *S. paradoxus*, and a gene in *S. bayanus* and *S. cariocanus* predicted to encode a protein related to an amidase of *S. pombe*. Many sensu lato and petite-negative sequences are predicted to encode proteins similar to those in species other than *S. cerevisiae* (Table 3). Similar findings have been made in *Kluyveromyces lactis* (Ozier-Kalogeropoulos et al. 1998) and other hemiascomycete species (Souciet et al. 2000).

To verify that these DNA sequences are specific to certain *Saccharomyces* species, we designed oligonucleotide primers to the amidase-encoding gene identified in *S. cariocanus*, and to two other unique sequences confined to *S. mikatae* and *S. paradoxus* that do not appear to encode protein. The sequences were amplified by PCR from genomic DNA of the corresponding species and produced products of the expected size. (Three other species-specific sequences produced PCR products larger than expected.) The amidase PCR product from *S. cariocanus* hybridized to a specific DNA fragment of *S. cariocanus* and *S. paradoxus* DNA in Southern blots (data not shown). The *S. mikatae*-specific sequence hybridized to genomic DNA from *S. mikatae*, but not to genomic DNA from the other sensu stricto species. Because we determined the DNA sequence of only about 1/32nd of the four sensu stricto species, it is likely that several more *Saccharomyces* sequences not present in *S. cerevisiae* remain to be identified.

Identification of Small Protein-Coding Genes

Some of the intergenic regions in the *S. cerevisiae* genome likely encode small proteins (<100 amino acids) that have not been annotated. Comparisons of *S. cerevisiae* intergenic sequences to those of the other *Saccharomyces* species using TBLASTX revealed many potential protein-coding sequences. Such comparisons of the sensu stricto sequences are not very informative because their relatively high degree of similarity to *S. cerevisiae* sequence produces many spurious alignments, but the sequences of the sensu lato and petite-negative species are well suited for identification of small ORFs (smORFs, <100 codons) because their significant divergence from *S. cerevisiae* sequence yields relatively few TBLASTX alignments outside of protein-coding sequences. Most of the high-scoring TBLASTX alignments to *S. cerevisiae* sequence are extensions or fusions of known ORFs, but we identified 11 alignments of smORFs that potentially encode a protein (TBLASTX *P* value <1.0e-05; Table 4). Two of the smORFs were predicted previously by searching the *S. cerevisiae* genome for transcripts that originate from large intergenic regions of the genome (Olivas et al. 1997); two smORFs are similar to small ORFs in other yeast species (Table 4).

Identifying Functional Non-Protein-Coding Sequences

Our primary goal is to use sequence conservation to identify functional nonprotein-coding sequences. However, much of the similarity of the nonprotein-coding sequences of the sensu stricto species is likely caused simply by an inadequate amount of evolutionary time for the accumulation of sequence changes. Which species are sufficiently diverged so that functional nonprotein-coding sequences are apparent? How many different sequences need to be compared to reveal regions of sequence similarity that are functionally significant? We began to address these questions by searching for nonprotein-coding RNA genes and promoter regulatory sequences in non protein-coding portions of the *S. cerevisiae* genome.

Identification of Non-Protein-Coding RNAs

Many highly conserved sequences in intergenic regions were identified that are unlikely to encode proteins because no ORF was apparent in the *S. cerevisiae* genome sequence. In the few cases where we obtained sequences from more than one species that are similar to the same *S. cerevisiae* sequence, the conserved sequence elements readily stand out from surrounding sequences. Because there are only a few regions of the genome for which we have sequences from multiple species, we determined the DNA sequence of a few genes encoding nonprotein-coding RNAs from many different *Saccharomyces* species to compare how quickly they diverge across the genus, and to help us decide which *Saccharomyces* species are best suited for identifying these genes by comparative sequence analyses. We amplified the sequences by PCR using oligonucleotide primers chosen from *S. cerevisiae* flanking sequences that seemed likely to be conserved. We determined the sequence of *SNR39*, encoding a C/D

Table 3. *Saccharomyces* Genes not Present in *S. cerevisiae*

Species	Predicted protein	Closest homolog	*P*-value	Accession no.
S. bayanus	putative amidase	*S. pombe*	2.3e-24	pir||T39112
S. cariocanus	putative amidase	*S. pombe*	8.9e-46	
S. paradoxus	Ty-5 associated protein	*S. paradoxus*	8.6e-40	gb|U19263.1
S. castellii	Squalene synthase	*C. glabrata*	2.6e-07	dbj|BAB12207.1
	NAD(P)H dehydrogenase	*P. aeruginosa*	3.1e-40	gb|AAG04613.1
	hypothetical protein	*C. albicans*	6.2e-35	emb|CAA21971.1
	putative p150	*H. sapiens*	5.1e-25	gb|AAA88038.1
	unnamed protein	*H. sapiens*	5.3e-09	dbj|BAA91193.1
	hyphally regulated protein	*C. albicans*	4.7e-08	pir||S58135
S. kluyveri	hypothetical protein	*X. fastidiosa*	1.8e-08	gb|AAF83652.1
	hypothetical protein	*S. pombe*	5.9e-30	pir||T42242
	RNA binding protein, pumilio-family	*S. pombe*	1.3e-09	pir||T41065
	probable dipeptidase	*S. pombe*	3.0e-26	pir||T41665
	WSC4 homologue	*K. lactis*	6.5e-11	emb|CAB50897.1
	hypothetical protein jhp1324	*H. pylori*	6.9e-15	pir||G71821
	coiled-coil protein	*S. Pombe*	9.1e-13	pir||T40559
	sphingomyelin phosphodiesterase	*M. musculus*	1.2e-17	pir||S27393
	neuronal thread protein AD7c-NT	*H. sapiens*	1.6e-14	gb|AAC08737.1
	ser/arg-rich protein specific kinase	*M. musculus*	3.9e-06	pir||JC5930
	hypothetical protein SPAC630.13c	*S. pombe*	2.9e-19	pir||T38991
	purine nucleoside permease	*C. albicans*	1.8e-41	gb|AAC64324.1
	beta-glucosidase	*K. lactis*	2.1e-59	emb|CAA29353.1
	stearoyl-CoA desaturase	*C. albicans*	2.1e-06	pir||T18228
	iron/ascorbate dependent oxidoreductase	*C. jejuni*	7.6e-08	emb||CAB73453.1
	N-carbamoyl-amidohydrolase	*P. aeruginosa*	1.5e-07	gb|AAG03833.1
	Nef attachable protein	*H. sapiens*	2.9e-08	dbj|BAA95214.1
	hypothetical protein B0252.2	*C. elegans*	1.5e14	pir||T15291
	unnamed protein product	*H. sapiens*	7.2e-20	dbj|BAA91131.1
	CLIP-170	*R. norvegicus*	6.5e-06	emb|CAB92974.1
	putative transaminase	*L. lactis*	1.5e-19	emb|CAB97148.1
	splicing factor u2af 35 kd subunit	*S. pombe*	1.6e-18	pir||T39243
	cytoskeleton assembly control protein	*C. albicans*	3.7e-09	gb|AAC62773.1
	Rbt2p (repressed by TUP1)	*C. albicans*	2.0e-08	gb|AAG09788.1
	uric acid-xanthine permease	*E. nidulans*	1.7e-10	emb|CAA50681.1
S. unisporus	hypothetical protein SPAC1687.07	*S. pombe*	3.4e-11	pir||T37750
	homolog of co-factor B Alp11p	*S. pombe*	1.5e-10	pir||T38860
	polyprotein	*O. sativa*	2.1e-32	gb|AAD27547.1

Alignments having a *P* value <1.0e-05 are shown. Sequences that had high scoring alignments to known genes were not included if they also had significant similarity to *S. cerevisiae* rDNA sequence.

box snoRNA required for methylation of the 25s ribosomal subunit (Kiss-Laszlo et al. 1996) and *SNR44*, encoding an H/ACA box snoRNA involved in pseudouridinylation of ribosomal RNA subunits (Ganot et al. 1997), both of which are located within introns of well-conserved ribosomal protein genes.

All eight *SNR39* sequences align well using `CLUSTALW`, with 58% of the 93 nucleotides of *SNR39* being conserved among eight *Saccharomyces* species. The C and D boxes as well as the guide sequence are conserved perfectly throughout the genus (Fig. 2). The gene is even readily distinguishable from surrounding intron sequence in most two-way alignments with *S. cerevisiae* sequence, although it is indistinguishable in an alignment of *S. cerevisiae* sequence to *S. paradoxus* (too similar), and is difficult to discern in an alignment to *S. exiguus* sequence (too diverged).

The *SNR44* gene (encoding an H/ACA snoRNA) sequences also align using `CLUSTALW`, but are less well conserved than *SNR39* (Fig. 3). The alignment reveals four highly conserved blocks, ranging from 5 to 15 nucleotides in length, which, surprisingly, do not include the ACA sequence (AAA in the sensu stricto species) or the H-box (Fig. 3), known functional elements in this snoRNA

Identification of Gene Regulatory Sequences

We expect potential gene regulatory sequences to be manifested as short blocks of sequence similarity in intergenic regions of the genome. These are difficult to recognize in the random DNA sequences of the *Saccharomyces* species for two reasons. On one hand, the sequences of the sensu stricto species are usually so similar to *S. cerevisiae* sequence that few isolated runs of identical nucleotides are found. In rare cases where we can align DNA sequence of two or more species, the background sequence similarity is reduced and potential regulatory elements begin to stand out (e.g., Fig. 4), but these alignments never extend over the entire promoter. On the other hand, the DNA sequences of the sensu lato species are generally too different from *S. cerevisiae* sequence to align with local alignment algorithms (such as `BLASTN` or `BESTFIT`). Some sensu lato sequences can be anchored to their presumed orthologs in *S. cerevisiae* gene using `BLASTX`, then

Table 4. Small *S. cerevisiae* ORFs (smORFs) with Homologs in Other *Saccharomyces* species

Location	Length	Species	% ID	Similar proteins
Chr2 363049-362771	92#	*S. unisporus*	25/36 (69%)	
Chr2 419123-418869*	85	*S. kluyveri*	52/77 (67%)	
Chr2 684936-685216+	94	*S. castellii*	51/91 (56%)	
Chr3 100839-101343	142#	*S. unisporus*	67/124 (54%)	*S. pombe*
Chr4 603805-603590+	72	*S. kluyveri*	35/81 (43%)	*C. elegans*
Chr4 691007-691204	66	*S. kluyveri*	34/69 (49%)	
Chr5 261045-260935*	37	*S. kluyveri*	25/45 (55%)	
Chr5 551117-550863	85	*S. unisporus*	42/66 (63%)	
Chr7 836659-836384	92	*S. castellii*	23/61 (37%)	
Chr10 159545-159324	74	*S. kluyveri*	40/50 (80%)	
Chr10 316571-316377+	65	*S. castellii*	38/63 (60%)	
		S. kluyveri	20/25 (80%)	

Location refers to SGD nucleotide coordinates. smORFs with an asterisk correspond to smORFs identified by Olivas and Parker (1997). Amino acid lengths followed by a pound sign indicate that a frame shift or an RNA splice is required to create the reading frame in the current annotation of the *S. cerevisiae* sequence. Species encoding sequences similar to the smORFs are listed (*S. pombe* accession no. T37750, protein information resource; *C. elegans* accession no. AAF59588.1, GenBank. A (+) refers to smORFs that are conserved in other hemiascomycetes yeast species (Blandin et al. 2000).

extended into the promoter region. Of the 4296 sensu lato sequences that align to *S. cerevisiae* proteins with BLASTX, 866 extend >30 nucleotides into the promoter region. We were able to align only 398 of them to their *S. cerevisiae* ortholog (using BLASTN, with a word length of five and only 111 of the alignments extend >100 bp into the promoter region. Many potential regulatory elements are apparent in these alignments. For instance, we can clearly see conservation of DNA-binding sites for the Alpha1 and Mcm1 proteins in the *MFA2* and *STE3* promoters of *S. castellii* (Fig. 5A,B). A consensus Hap2 binding site (ACCAATNA; Svetlov and Cooper 1995) is included in one of the three conserved runs of sequence present in the promoter of *ATP14* (Fig. 5C), a gene that is plausibly regulated by Hap2. Many other alignments contain short runs of conserved sequence that are potential gene regulatory elements.

To assess *Saccharomyces* species more broadly for their suitability for identifying regulatory sequences in gene promoters by comparative DNA sequence analysis, we amplified by PCR and determined the sequences of the promoters of three well-characterized genes from many *Saccharomyces* species (see Methods for details). Although the sequences of the promoter region between *GAL1* and *GAL10* of the sensu stricto species align well using CLUSTALW, the sensu lato species' sequences do not align. We searched for Gal4 and Mig1 protein-binding sites by a simple pattern search [FINDPATTERNS] and, as expected, they are apparent in the *GAL1–10* promoter of all 10 species (Fig. 6). The spacing of these binding sites is well conserved in the sensu stricto species, with only a few sites missing from some of the species. The spacing and number of Gal4 and Mig1 binding sites are not conserved in the sensu lato species. Of these, only the *S. castellii* sequence aligned with the *S. cerevisiae* sequence using local alignment algorithms (BLASTN or BESTFIT).

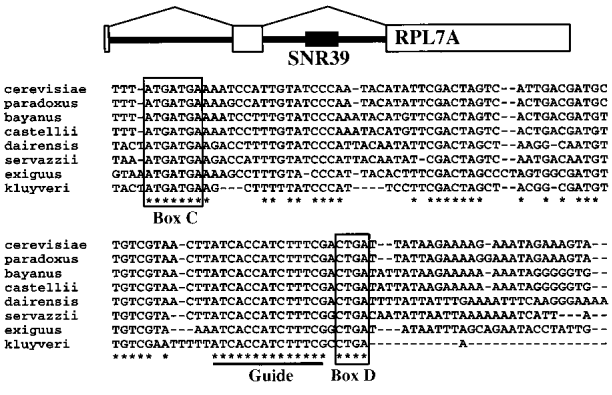

Figure 2 A CLUSTALW alignment of *SNR39* sequences (encoding a snoRNA) from eight different *Saccharomyces* species. Box C and Box D are known functional elements; the "guide" is the sequence complementary to rRNA sequence adjacent to the methylation site. The structure of the *RPL7A* transcript including the intronic *SNR39* gene is shown above the sequence alignment.

We obtained from each species the DNA sequence of at least one of the two nearly identical copies of the divergently transcribed *HHT* and *HHF* genes (*HHT1–HHF1* and *HHT2–HHF2*), encoding histones H3 and H4, respectively. In some cases we obtained the sequence of both copies (though they are so diverged from the *S. cerevisiae* sequence in the sensu lato species that it was usually difficult to distinguish between the two copies using BLASTN alignments). Although conserved blocks of sequence begin to emerge from 3-way CLUSTALW alignments of sequences of the sensu stricto species (data not shown), addition of a sensu lato sequence makes these elements clearly stand out (Fig. 7A). Both (TATA) boxes are conserved, as are sequences within and near CCA boxes 1 and 2 that were defined previously as regulatory elements of this promoter (Freeman et al. 1992). One of the conserved sequences (enclosed in dashed lines in Fig. 7A) is present only in one of the *HHT–HHF* copies in each species, so the two pairs of genes may be regulated differently. Other potential CCA boxes (CCA box 3 and 4) that were not recognized previously seem apparent. Using the pattern recognition program AlignAce (Roth et al. 1998), we identified many variations of the CCA box within the histone promoters (60 different motifs in the 13 promoter sequences) some of which were not identified by full-length sequence alignments (Fig. 7B); the logo consensus (Schneider and Stephens 1990) CCA box sequence is shown in Figure 7C. Each copy of the divergent promoter contains three to six of the relaxed CCA box sequence motifs.

We succeeded in amplifying the *GAL4* promoter from only three *Saccharomyces* species (presumably because the flanking protein-coding sequences are not well conserved). One of these sequences (from *S. paradoxus*) is too similar to the *S. cerevisiae* sequence to be useful, but alignments of the other two sequences (one from the sensu stricto species *S. bayanus*, the other from the petite-negative species *S. kluyveri*) seem very informative (Fig. 8). The promoter elements that were previously defined genetically (Griggs and Johnston

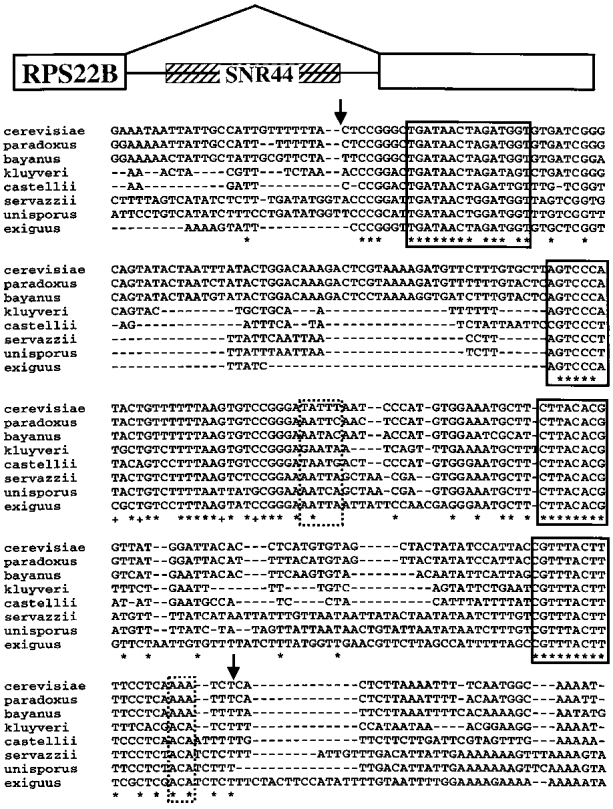

Figure 3 A CLUSTALW alignment of *SNR44* sequences encoding an H/ACA box snoRNA. The positions of the start and end of *S. cerevisiae SNR44* are marked with arrows. Highly conserved sequences are boxed. The H and ACA boxes are enclosed in dashed lines. The plus symbols in the consensus line indicates that seven of eight nucleotides in the alignment are conserved and were added to highlight conservation of sequence upstream of the H box. The structure of the *RPS22B* transcript including the intronic *SNR44* gene is shown above the sequence alignment.

1993) clearly stand out as conserved in a CLUSTALW alignment of these sequences with the *S. cerevisiae* sequence.

DISCUSSION

We have investigated the feasibility of using comparative DNA sequence analysis to identify functional sequences in the genome of *S. cerevisiae*, with the goal of identifying regulatory sequences and sequences specifying nonprotein-coding RNAs. We are confident that promoter regulatory sequences,

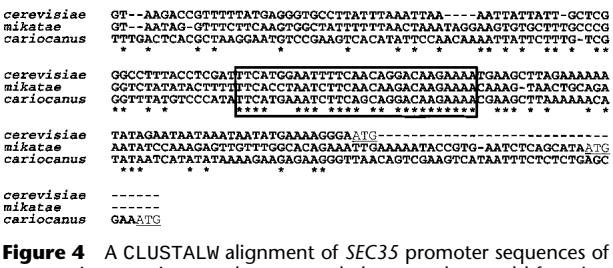

Figure 4 A CLUSTALW alignment of *SEC35* promoter sequences of sensu stricto species reveals conserved elements that could function as regulatory elements. The conserved sequences are boxed; the *S. cerevisiae* ATG start codon and putative start codons for the other species are underlined.

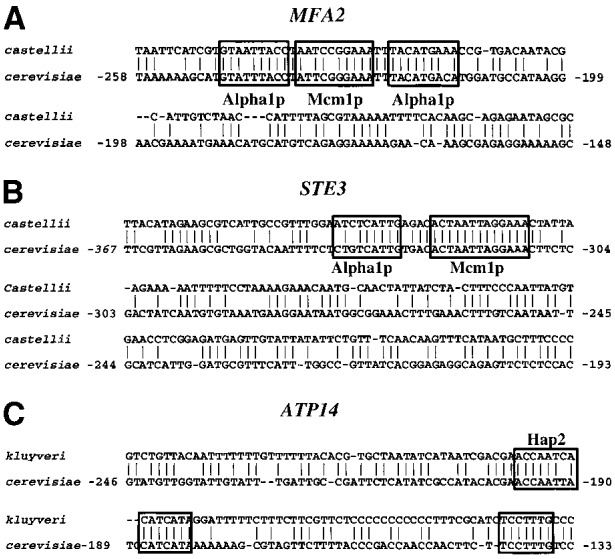

Figure 5 Known functional elements are conserved in BLASTN alignments between *S. cerevisiae* promoters and orthologous sequences of sensu lato species. (*A*) An alignment of *MFA2* promoter sequences from *S. castellii* and *S. cerevisiae* with known binding sites for Alpha1 and Mcm1 boxed. (*B*) An alignment of *STE3* promoter sequences from *S. castellii* and *S. cerevisiae* with known binding sites for Alpha1 and Mcm1 boxed. (*C*) An alignment of the *ATP14* promoter sequences from *S. kluyveri* and *S. cerevisiae* showing three conserved runs of sequence, one of which includes a consensus Hap2 (ACCAATNA) binding site. Numbers refer to the position of the *S. cerevisiae* sequence relative to the ATG start codon.

RNA genes, and smORFs could be identified by comparisons of orthologous *Saccharomyces* sequences.

Analysis of Protein-Coding Genes

Before discussing our analysis of non-protein-coding sequence, a few observations regarding protein-coding genes

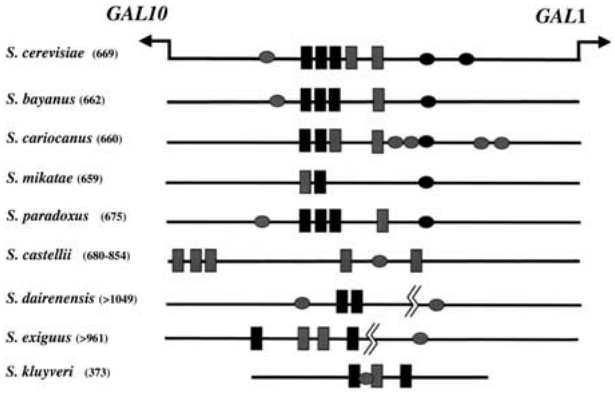

Figure 6 The location of Gal4 and Mig1 regulatory elements in the *GAL1–GAL10* promoter sequences from many *Saccharomyces* species. The length of the promoter (where known) is shown in parentheses. Gal4 binding sites are shown as black boxes with gray boxes indicating putative binding sites that are altered by one nucleotide from the consensus site ($CGGN_{11}CCG$). Potential Mig1 binding sites are shown as ellipses. Black ellipses refer to experimentally verified Mig1 sites (in *S. cerevisiae*) or Mig1 sites that clearly align to *S. cerevisiae* sites in pairwise alignments. Gray ellipses refer to putative Mig1 sites (CCCC followed by AT-rich sequence).

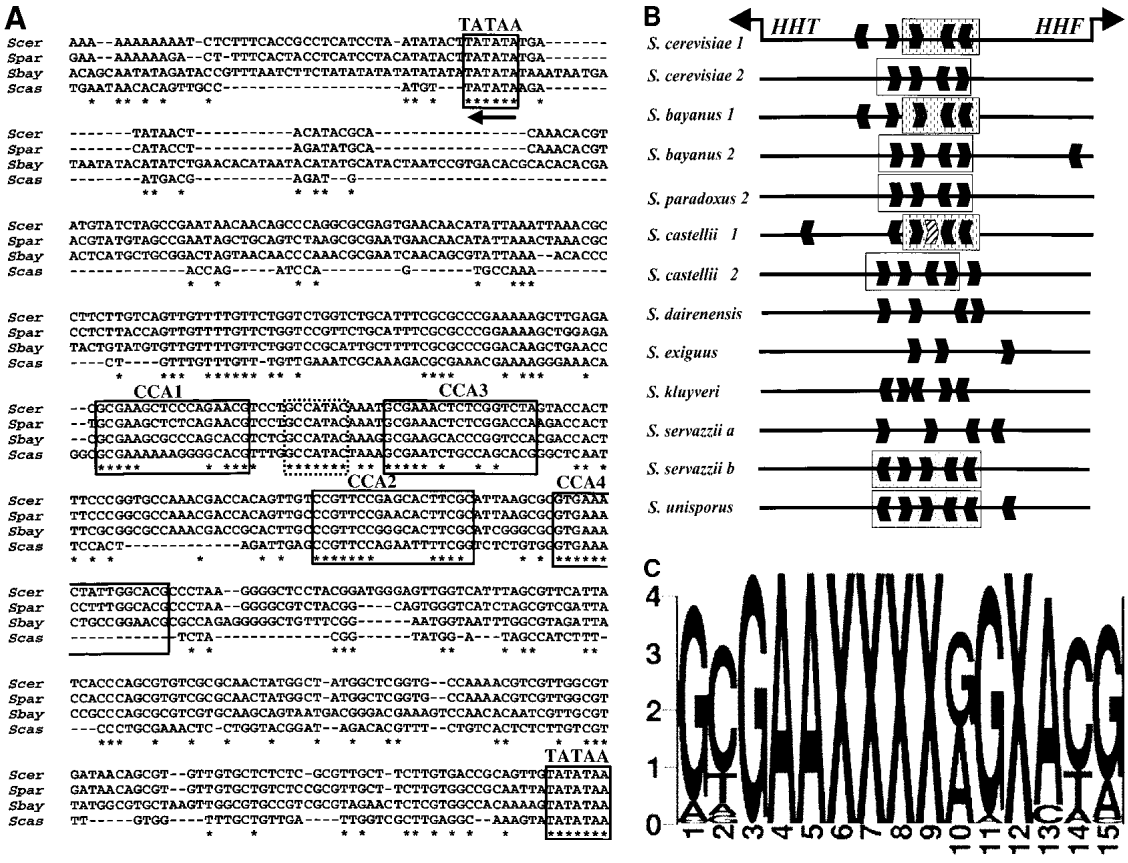

Figure 7 Conserved sequence elements of the *HHT2–HHF2* promoter. (*A*) A CLUSTALW alignment of promoter sequences from *S. cerevisiae*, *S. paradoxus*, *S. bayanus* and *S. castellii.* Putative TATA boxes are boxed and underlined with arrows, indicating the direction of transcription. A conserved heptamer is enclosed in a dashed box. CCA box 1 and 2 representing elements identified in *S. cerevisiae* (Freeman et al. 1992) are boxed and numbered. CCA box 3 and 4 are related to the CCA box. Only part of the *S. castelli* CCA box 4 is captured in this alignment. (*B*) The location and orientation of putative CCA boxes (identified by AlignAce, Roth et al. 1998) in *Saccharomyces* promoter sequences. Sequences clearly orthologous to *S. cerevisiae* Copy 1 or Copy 2 are labeled. Unclear cases are unlabeled (if one copy was obtained from the species) or labeled as a and b (if two copies were obtained). CCA box related elements are shown as black hexagons that depict the direction of the elements. The unfilled, stippled, and vertically dashed boxes depict closely related sequences where the CCA box elements clearly align by pairwise (BESTFIT) or multiple sequence alignments (CLUSTALW). Note the extra CCA box within the *S. castellii* sequence (striped hexagon). The gray hexagon (in *S. bayanus* sequence 1) depicts an excellent CCA box that contains a single nucleotide insertion and was not identified by AlignAce, but is apparent in multiple sequence alignments. (*C*) A "logo" consensus for the CCA box based on 60 related sites identified by AlignAce.

are worth noting. First, we identified several gene sequences not found in the *S. cerevisiae* genome (Table 3). These are genes that likely were lost from (or evolved rapidly in) *S. cerevisiae*, though it is possible they were acquired by the other species after they diverged from *S. cerevisiae*. Second, we noticed several ORF boundaries that seem to differ among the species. Some of these could be attributable to errors in the *S. cerevisiae* genome sequence, and some could be real, indicating variations in gene length in the different *Saccharomyces* species. More interestingly, we identified several small ORFs that likely encode protein, two of which show similarity to proteins found in distantly related organisms (Table 4). Based on the number of smORFs identified in the small amount of sequence we analyzed, we estimate that ⩾40 more smORFs will be identified in the *S. cerevisiae* genome by analysis of the complete genome sequences of a few *Saccharomyces* species.

Identification of Novel RNA Genes

Non-protein-coding RNA genes can be difficult to recognize in a DNA sequence or to identify experimentally (Eddy 1999), and many surely remain to be discovered. Searches of the yeast genome for potential RNA sequences have identified a few candidates for this type of gene (Olivas et al. 1997; Lowe and Eddy 1999). The comparative sequence approach should greatly speed their identification by revealing potential RNA sequences that can be experimentally evaluated. Indeed, we

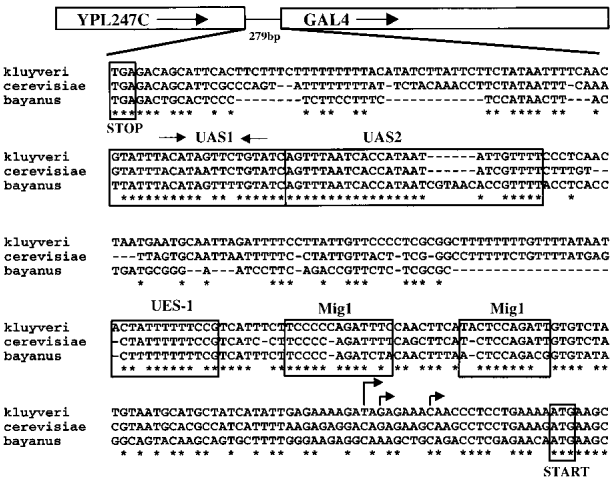

Figure 8 A three-way CLUSTALW alignment of *GAL4* promoter sequences reveals strong conservation of known regulatory elements. The entire promoter is shown in the alignment, from the stop codon of the upstream gene (*YPL247C*) to the ATG start codon of *GAL4*. Regulatory elements defined by Griggs and Johnston (1993) are boxed and labeled: UAS (upstream activator sequence), UES (upstream essential sequence), and Mig1 binding sites. Transcriptions start sites are depicted as arrows.

identified several conserved sequences that may encode functional RNAs and that are good candidates for experimentation to test this hypothesis (data not shown). This approach is validated by the clear conservation of snoRNAs across the *Saccharomyces* genus (Figs. 2 and 3) and by our ability to identify in the random sequences snoRNA-encoding sequences by BLASTN comparison to *S. cerevisiae* RNA genes (data not shown).

Identification of Regulatory Sequences

Finding gene regulatory sequences is a difficult task because they are often short, independent of orientation, and their position in a promoter can vary greatly. For these reasons, DNA sequences of presumably orthologous promoters often fail to align using tools such as BLASTN or CLUSTALW. Because most comparisons of nonprotein-coding DNA sequence have used relatively diverged species (such as human–mouse or human–puffer fish), we tested the use of closely related yeast species to facilitate identification of promoter elements. As expected, most of the DNA sequences of the closely related sensu stricto *Saccharomyces* species align to *cerevisiae* sequences, and known promoter regions are conserved in the alignments. Two species (*S. cariocanus* and *S. paradoxus*), are too closely related to *S. cerevisiae* (>80% identity in noncoding regions) to yield much information from the alignments (Table 1), but many *S. mikatae* and *S. bayanus* alignments show conservation of short runs of conserved sequence and are nearly as informative as the three-way alignment shown in Figure 8. We believe that many *S. cerevisiae* regulatory elements could be predicted using sequence alignments of *S. cerevisiae* sequence to *S. mikatae* and *S. bayanus*. Alignment of multiple sequences will be required to add statistical confidence to the predictions.

Less than 2% of the sensu lato and petite-negative species' sequences align to intergenic regions of *S. cerevisiae*, and many of them are repetitive. By anchoring the alignment to adjacent protein-coding sequence we were able to extend the alignments of 111 sequences (out of 579) ⩾100 bp into the promoter. In this way, ~20% of the sensu lato promoters can be aligned to their *S. cerevisiae* orthologs, and it appears that many regulatory elements are conserved between *Saccharomyces* species (e.g., Fig. 5). Sensu lato promoters that cannot be aligned accurately to their *S. cerevisiae* counterparts will require other analysis techniques, such as pattern searching (e.g., CONSENSUS [Hertz et al. 1990], Gibbs sampling [Lawrence et al. 1993], AlignAce [Roth et al. 1998]). Pattern search algorithms were useful for identifying known regulatory sequence elements in the *HHT–HHF* (Fig. 7) and *GAL1–GAL10* (Fig. 6) promoters, where they are conserved through the most diverged *Saccharomyces* species.

Choosing *Saccharomyces* Species for Comparative Analysis

Clearly, multiple sequences from several species of various degrees of divergence will be needed to identify conserved sequences. How many sequences will be required for informative sequence comparisons? Which species are optimal for this kind of analysis? The answers depend somewhat on the gene being analyzed, but we believe a few general principles guide the choice of species.

Number of Species for Genome Sequencing

We estimate that at least three sequences of varying degrees of similarity need to be aligned to *S. cerevisiae* sequence for short blocks of conserved sequence in non-protein-coding DNA to be significant. Consider an alignment of *S. cerevisiae* noncoding sequence to the two most diverged sensu stricto species (*S. mikatae* and *S. bayanus*). The percent identity of their noncoding sequence to *S. cerevisiae* sequence averages ~75% for the sequence reads that align with BLASTN (Table 1). Because some reads do not align readily, the overall identity is somewhat less; 70% identity seems like a conservative estimate. The chance of both sequences having the same nucleotide as in the *S. cerevisiae* sequence at any given position is then ~0.49 (.7 × .7, or ~0.5), so the chance of a hexamer aligning perfectly in all three sequences is ~0.015 (0.5^6), assuming the sequences are equally diverged from each other, which is approximately true. Thus, a conserved hexamer in this three-way alignment is not very significant, as we expect to find one every 67 base pairs on average. Adding a fourth sequence of similar divergence increases the significance of the alignments only modestly: The chance of a conserved hexamer in a four-way alignment is 0.0016 = $(.7^3)^6$, or one every 625 base pairs on average. However, if the third sequence added is only 40% identical to the *S. cerevisiae* sequence (probably a reasonable estimate for the sensu lato non-protein-coding sequences, because few of them align by BLASTN [Table 1]), the expected frequency of an exact hexamer alignment decreases significantly, to .000057, or one approximately every 17,600 nucleotides (1 in ~5000 nucleotides if the sensu lato sequence is 50% identical to *S. cerevisiae*).

Indeed, this is close to what we observed in alignments of randomly generated sequence: in many four-way CLUSTALW alignments of sequences (three of them 70% identical to each other; one 40% identical to these), hexamers appeared on average once every 5555 nucleotides; heptamer or longer runs appeared once every 25,000 nucleotides. Because the average intergenic region in *S. cerevisiae* is ~600 nucleotides in length,

a conserved hexamer is expected to occur by chance about once in every eight promoters on average. We realize that these "back of the envelope" calculations must be interpreted cautiously because it is unlikely that even a probable optimal multiple sequence alignment will correspond precisely to the biological events that generated the sequences being aligned (Thorne and Churchill 1995), but in any case, four-way alignments seem to be the minimum required for efficient recognition of conserved regulatory sequences.

Candidate Species for Comparative Analysis

Which species' sequences will yield the most information? Two considerations need to be balanced in making this decision. The sequences need to be similar enough so that most of them can be aligned with their *S. cerevisiae* ortholog, a requirement that favors the sensu stricto species, as more of their sequence reads can be aligned to *S. cerevisiae* sequence (Tables 1, 2). Conversely, if the sequences are too similar the functional elements do not stand out, a consideration that favors the sensu lato species. It seems clear that at least one of each group of species need to be compared.

Two of the sensu stricto species for which we obtained DNA sequence — *S. mikatae* and *S. bayanus* — seem like good candidates for large-scale sequence comparisons. *S. cariocanus* and *S. paradoxus* seem too similar to *S. cerevisiae* for this purpose (Tables 1, 2). Another good candidate for comparative sequence analysis is *S. kudriavzevii*, a recently described species (Naumov et al. 2000), that seems to lie between *S. bayanus* and *S. mikatae* in its evolutionary distance from *S. cerevisiae* (Fischer et al. 2000).

There are few differentiating characteristics for selecting one sensu lato species over another, and sequence information from any of these species will likely be useful. The species that we studied are a similar evolutionary distance from *S. cerevisiae*. *S. unisporus* is topologically much closer to *S. cerevisiae* than *S. kluyveri* in most phylogenies, but some show a much longer branch length for *S. unisporus* (James et al. 1997; Kurtzman and Robert 1991; Oda et al. 1997), consistent with our data. Two species, *S. servazzii* and *S. dairenensis* branch closely with species that we studied (*S. unisporus* and *S. castellii*, respectively). It is unlikely that any known sensu lato species are significantly closer to *S. cerevisiae* than the sensu lato species we studied (Petersen et al. 1999). We favor *S. kluyveri* and *S. castellii* over *S. unisporus* for comparative sequence analysis for a few reasons. First, we obtained a few more BLASTX alignments for these species (Table 2), and *S. castellii* has a significantly greater number of BLASTN alignments (Table 1) that will help in locating promoter sequences. Second, we identified more snoRNA sequences by BLASTN comparisons from *S. kluyveri* and *S. castellii* than from *S. unisporus* (data not shown). Finally, these two species are estimated to have the smallest genomes within the *Saccharomyces* genus (Vaughan-Martini et al. 1993; Petersen et al. 1999) thus reducing the amount of sequence that would need to be determined.

METHODS

Sequencing Methods

The DNA sequence of random genomic clones from different *Saccharomyces* species was determined by the Washington University Genome Sequencing Center (WUGSC). A brief description of each step follows. Detailed protocols are provided in Mardis (1997), and at http://genome.wustl.edu/gsc/Protocols.

Table 5. Yeast Strains Used in this Study

Strains	Name	Other designation	Obtained from
*S. bayanus**	623-6C		E. Louis[a]
*S. cariocanus**	50791		E. Louis
*S. mikatae**	1815		E. Louis
*S. paradoxus**	N17		E. Louis
S. bayanus	NRRL Y-12624	(CBS 380)	C. Kurtzman[b]
*S. castellii**	NRRL Y-12630	(CBS 4309)	C. Kurtzman
S. dairenensis	NRRL Y-12639	(CBS 421)	C. Kurtzman
S. exiguus	NRRL Y-12640	(CBS 379)	C. Kurtzman
*S. kluyveri**	NRRL Y-12651	(CBS 3082)	C. Kurtzman
S. paradoxus	NRRL Y-17217	(CBS 432)	C. Kurtzman
S. servazzii	NRRL Y-12661	(CBS 4311)	C. Kurtzman
*S. unisporus**	NRRL Y-1556	(CBS 398)	C. Kurtzman

An asterisk indicates species there were used for sequencing random genomic DNA.
[a]Dept. of Genetics, Univ. of Leicester, Leicester, UK
[b]USDA, Peoria, IL

Preparation of Genomic DNA

Derivatives of each yeast species lacking mitochondrial DNA were obtained by plating colonies on YPD agar media containing ethidium bromide (Slonimski et al. 1968) (except in the case of *S. kluyveri* which requires mitochondrial activity for viability). Genomic DNA was extracted from cells (50–100 mL cultures grown overnight to stationary phase in YPD). Cells were harvested and resuspended in 2 cell volumes of cold buffer (50 mM EDTA, 20 mM Tris at pH 8 containing 1 mg/mL ethidium bromide). Glass beads (0.5 mm, acid washed) were added to the cells and the suspension was vortexed for 5 min at 4°C. Sarkosyl (1% final concentration) was added and mixed by inversion. Cell debris was removed by centrifugation (2000 g for 10 min) and cesium chloride was added to the supernatant (1 g/mL). The mixture was centrifuged again (5000 g for 10 min) to remove additional debris. The supernatant was centrifuged at 100,000 G for 6 h in a Beckman TL-100 rotor.

Preparation of Clone Libraries

After quantification, the DNA was sheared by sonication, the ends repaired with mung bean nuclease, and fragments of 1–2 kb were selected by electrophoresis through a 0.8% agarose gel. The DNA fragments were excised and extracted from the gel, ligated to the plasmid (pBC) vector, and introduced by electroporation into competent *Escherichia coli* DH10B cells. DNA sequence of a representative sample of the resulting plasmid subclones was determined to assess library quality.

DNA Sequencing

Plating and sequencing of plasmid library subclones, as well as sample loading, data collection and processing was done as described by Mardis and Wilson, (1997) (description of methods also available at http://genome.wustl.edu/gsc/Protocols/protocols.shtml.). *E. coli* colonies carrying plasmid subclones were picked and their plasmid DNA prepared by a simple method that lyses cells with microwaves (Marra et al. 1999). DNA sequencing used "big dye terminator" chemistry in cycle sequencing reactions. Cycle sequencing reactions were ethanol precipitated, resuspended in loading buffer and loaded onto an ABI 3700 capillary sequencing machine. Statistics were monitored constantly, and representative traces were inspected for each run. The data was processed by XGASP (Wendl et al. 1998), a script that directs vector clipping, quality analysis, data transfer, and initial assembly. The traces

were processed with PLAN, which performs PHRED basecalling (Ewing and Green 1998; Ewing et al. 1998).

BLAST Sequence Comparisons

Sequence alignments were initially generated using WU-BLAST 2.0 (http://blast.wustl.edu). *S. cerevisiae* sequence databases were obtained from SGD (http://genome-www.stanford.edu/Saccharomyces). The databases included *S. cerevisiae* genomic DNA (all_sacchdb.dna), *S. cerevisiae* protein-coding sequences and genes (orf.trans.fasta and orf-_coding.fasta) and a library of intergenic DNA (NotFeature.fasta). This latter database consists of *S. cerevisiae* DNA from which was removed all genes (encoding both proteins and RNA), LTRs, and transposons. Another library was created by fusing the sequences of a 5′ UTR library (consisting of the 500 bp upstream of the ATG start codon) to the genes sequences in orf_coding.fasta. BLAST output was parsed and processed with ad hoc PERL scripts. Percent identities of BLAST alignments were taken from the highest scoring alignment, to avoid biasing the calculations with low scoring and/or multiple alternative alignments of the same sequence segment.

TBLASTX comparisons used default parameters. Sensu lato and petite-negative sequences were compared to *S. cerevisiae* "not-feature" DNA sequences (those not known or predicted to encode proteins or functional RNAs) to identify potential protein-coding sequences. High scoring alignments were manually selected from all TBLASTX alignments. *S. cerevisiae* sequences with interesting alignments were visualized using ACeDB (Eeckman and Durbin 1995) to look for potential new ORFs and extensions of known ORFs.

Multiple sequence alignments were created with CLUSTALW (Thompson et al. 1994). Some alignments were edited manually to investigate alternative equal scoring alignments. Local and global sequence alignments were generated using BESTFIT and GAP (GCG Corp.) with default parameters. FINDPATTERNS (GCG Corp.) was used to identify specific motifs within sequences.

Identification of Random Sequences Encoding Proteins Absent from *S. cerevisiae*

Random genomic sequences were compared with BLASTX to a nonredundant protein database similar to that maintained at NCBI. Because the database contains *S. cerevisiae* sequences, most of the identified homologs were from *S. cerevisiae*. In the sensu stricto species only five DNA sequences had significant hits to proteins not found in *S. cerevisiae*. The situation was more complex for the sensu lato and petite-negative sequences. In addition to finding genes not encoded in *S. cerevisiae*, we also found many sequences that had more similarity to sequences of species other than *S. cerevisiae*. Often the proteins with higher similarity are from closely related fungi, but occasionally they are from a distantly related organism, with the *S. cerevisiae* homolog showing much weaker similarity.

PCR Amplification and Southern Blotting of Species-Specific Sequences

Sequences were amplified by a standard PCR using the oligonucleotide primers listed below with genomic templates prepared as described by Hoffman and Winston (1987). PCR products were randomly labeled with ^{32}P (Ausubel et al. 1998) and used as probes for Southern blots of *Hin*dIII-digested genomic DNA from different species fractionated on a 1% agarose gel and transferred to a charged nylon filter.

S. cariocanus amidase
OM2186 AAACAACACAGCACCAGC
OM2187 TATCCTCAGACGCAGTCG

S. cariocanus unique sequences
OM2196 ATGGTGTGCGTTGTTATC
OM2197 TCAACATGTCTGCATTCG
OM2198 GCTAGTAGTTCCGTGGTG
OM2199 CATCCCGACTCTGTCTTG

S. paradoxus unique sequence
OM2191 GGTACTTGGAGTTCTGTG
OM2192 TGGTGCTTTGGCAACAAG

S. mikatae unique sequence
OM2194 CACGTCCACATTAACCTG
OM2195 GATGTGGACATCATGCTTG

S. bayanus unique sequence
OM2206 CAATAACACGGATGCTCAAC
OM2207 TGTCGGAGGTCACAGGAG

Cloning of Specific Genomic Regions

Yeast genomic DNA was prepared as described by Hoffman and Winston (1987). The sequences of oligonucleotide primers were chosen based on conserved protein-coding regions that flank the sites of interest. Nested primers were used in a second reaction to amplify the *GAL1–10* and the *GAL4* promoters from some species. Initial PCRs were diluted 1/250 in TE and 1 µL was used as template in the second PCR. PCR products were cloned into a TA cloning vector (pGEMT-EASY [Promega] or pCRII [Invitrogen]). In some cases the PCR products were ligated directly into the vector; in other cases the PCR product was first purified by preparative agarose gel electrophoresis and extracted with QIAGEN gel purification columns. Plasmids were isolated using QIAGEN miniprep columns, checked by restriction digest for inserts of the correct size, and the DNA sequence of the insert was determined on an ABI 377 automated sequencer using universal or T7/SP6 primers and Big-Dye Terminator reactions (ABI). Sequence reads were processed with PHRED (Ewing and Green 1998; Ewing et al. 1998), PHRAP, and CONSED (Gordon et al. 1998). Sequences from *S. bayanus* and *S. paradoxus* were obtained from strains NRRL Y-12624 and NRRL Y-17217, respectively.

PCR Primers (asterisks indicate nested primers):

26s rDNA
OM1833 GCATATCAATAAGCGGAGGAAAAG
OM1834 GGTCCGTGTTTCAAGACGG

GAL1–10
OM1975 AGCCGTCAGTTCAAAACATCACC
OM1866* CCATGTATCCAGCACCACC
OM1867 AGTCACAATAATCAATATGTTCACC

GAL4
OM1941 TTCCCACCAATAACATCATTTGACTGGAA
OM1920 TTCCACTTCTGTCAGATGTGCCCTAGTCAGCGG
OM1921* GACTCGAACAAAATCATTATTCTAGATATGAG
OM2208* GATAAACAACATTGCATGGAGGC

S. paradoxus: OM1920-OM1941

S. bayanus: OM1920-OM1941/ OM1920-OM1921

S. kluyveri: OM1920-OM2208

HHT–HHF
OM1837 CACCAGTGGACTTTCTTGCTGTTTG
OM1838 CCACCTTTACCTAGACCTTTACCACCTTTACC

SNR39
OM1959 GAAAAAATCTTGACCCCAGAATCTCAGTTGAAGA
OM1961 CTTGTTAATACCCTTGATTCTGACAACGAA

SNR44
OM1835 TCATCAAGTTTTTACAAGTTATGCAAAAGC
OM1836 CGACAATCTTACCAGATCTGTGGTC

Estimation of Hexamer Frequency in Multiple Sequence Alignments

Random DNA sequences were generated in the computer with Seq-Gen (Rambaut and Grassly 1997). One thousand groups of four sequences (500 nucleotides long) were generated. After the first sequence was generated, two sequences having 70% identity and one having 40% identity to the first sequence were generated. The separate groups were aligned with CLUSTALW and the number of aligned n-mers (n = 1–10) were computed using an ad hoc PERL script.

ACKNOWLEDGMENTS

We are indebted to Ed Louis for yeast strains, advice, helpful discussion, and enthusiastic support. We also thank members of the Gish Lab for programming assistance, and Sean Eddy, David States, and Gary Stormo for ongoing advice and for reviewing the manuscript. Linda Riles and Matt Curtis provided technical assistance that was instrumental in the initial phases of the project. This work was supported by funds provided by the James S. McDonnell Foundation. P.F.C. was supported in part by a NHGRI NSRA Fellowship (NIH #IF32HG00218).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A., and Struhl, K. 1998. In *Current protocols in molecular biology* (ed. V.B. Chanda), John Wiley and Sons, New York.

Barnett, J.A. 1992. The taxonomy of the genus *Saccharomyces Meyen* ex Reess: A short review for non-taxonomists. *Yeast* **8:** 1–23.

Blandin, G., Durrens, P., Tekaia, F., Aigle, M., Bolotin-Fukuhara, M., Bon, E., Casaregola, S., de Montigny, J., Gaillardin, C., Lepingle, A., et al. 2000. Genomic exploration of the hemiascomycetous yeasts: 4. The genome of *Saccharomyces cerevisiae* revisited. *FEBS Lett.* **487:** 31–36.

Eddy, S.R. 1999. Noncoding RNA genes. *Curr. Opin. Genet. Dev.* **9:** 695–699.

Eeckman, F.H. and Durbin, R. 1995. ACeDB and macace. *Meth. Cell Biol.* **48:** 583–605.

Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8:** 186–194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8:** 175–185.

Fischer, G., James, S.A., Roberts, I.N., Oliver, S.G., and Louis, E.J. 2000. Chromosomal evolution in *Saccharomyces*. *Nature* **405:** 451–454.

Freeman, K.B., Karns, L.R., Lutz, K.A., and Smith, M.M. 1992. Histone H3 transcription in *Saccharomyces cerevisiae* is controlled by multiple cell cycle activation sites and a constitutive negative regulatory element. *Mol. Cell. Biol.* **12:** 5455–5463.

Ganot, P., Caizergues-Ferrer, M., and Kiss, T. 1997. The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. *Genes & Dev.* **11:** 941–956.

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

Gellner, K. and Brenner, S. 1999. Analysis of 148 kb of genomic DNA around the wnt1 locus of *Fugu rubripes*. *Genome Res.* **9:** 251–258.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. 1996. Life with 6000 genes. *Science* **274:** 546, 563–567.

Gordon, D., Abajian, C., and Green, P. 1998. Consed: A graphical tool for sequence finishing. *Genome Res.* **8:** 195–202.

Griggs, D.W. and Johnston, M. 1993. Promoter elements determining weak expression of the GAL4 regulatory gene of *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **13:** 4999–5009.

Hardison, R.C., Oeltjen, J., and Miller, W. 1997. Long human-mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome. *Genome Res.* **7:** 959–966.

Hertz, G.Z., Hartzell, III, G.W., Stormo, G.D. 1990. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.* **6:** 81–92.

Hoffman, C.S. and Winston, F. 1987. A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformation of *Escherichia coli*. *Gene* **57:** 267–272.

James, S.A., Cai, J., Roberts, I.N., and Collins, M.D. 1997. A phylogenetic analysis of the genus *Saccharomyces* based on 18S rRNA gene sequences: Description of *Saccharomyces kunashirensis* sp. nov. and *Saccharomyces martiniae* sp. nov. *Int. J. Syst. Bacteriol.* **47:** 453–460.

Kent, W.J. and Zahler, A.M. 2000. Conservation, regulation, synteny, and introns in a large-scale *C. briggsae–C. elegans* genomic alignment. *Genome Res.* **10:** 1115–1125.

Kiss-Laszlo, Z., Henry, Y., Bachellerie, J.P., Caizergues-Ferrer, M., and Kiss, T. 1996. Site-specific ribose methylation of preribosomal RNA: A novel function for small nucleolar RNAs. *Cell* **85:** 1077–1088.

Kurtzman, C.P. and Robnett, C.J. 1991. Phylogenetic relationships among species of *Saccharomyces*, *Schizosaccharomyces*, *Debaryomyces* and *Schwanniomyces* determined from partial ribosomal RNA sequences. *Yeast* **7:** 61–72.

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

Lowe, T.M. and Eddy, S.R. 1999. A computational screen for methylation guide snoRNAs in yeast. *Science* **283:** 1168–1171.

Marra, M.A., Kucaba, T.A., Hillier, L.W., and Waterston, R.H. 1999. High-throughput plasmid DNA purification for 3 cents per sample. *Nucleic Acids Res.* **27:** e37.

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

Naumov, G.I., James, S.A., Naumova, E.S., Louis, E.J., and Roberts, I.N. 2000. Three new species in the *Saccharomyces* sensu stricto complex: *Saccharomyces cariocanus*, *Saccharomyces kudriavzevii* and *Saccharomyces mikatae*. *Int. J. Syst. Evol. Microbiol.* **50 Pt 5:** 1931–1942.

Oda, Y., Yabuki, M., Tonomura, K., and Fukunaga, M. 1997. A phylogenetic analysis of *Saccharomyces* species by the sequence of 18S-28S rRNA spacer regions. *Yeast* **13:** 1243–1250.

Olivas, W.M., Muhlrad, D., and Parker, R. 1997. Analysis of the yeast genome: Identification of new non-coding and small ORF-containing RNAs. *Nucleic Acids Res.* **25:** 4619–4625.

Ozier-Kalogeropoulos, O., Malpertuy, A., Boyer, J., Tekaia, F., and Dujon, B. 1998. Random exploration of the *Kluyveromyces lactis* genome and comparison with that of *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **26:** 5511–5524.

Petersen, R.F., Nilsson-Tillgren, T., and Piskur, J. 1999. Karyotypes of *Saccharomyces* sensu lato species. *Int. J. Syst. Bacteriol.* **4:** 1925–1931.

Rambaut, A. and Grassly, N.C. 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.* **13:** 235–238.

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

Schneider, T.D. and Stephens, R.M. 1990. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.* **18:** 6097–6100.

Slonimski, P.P., Perrodin, G., and Croft, J.H. 1968. Ethidium bromide induced mutation of yeast mitochondria: Complete transformation of cells into respiratory deficient non-chromosomal "petites". *Biochem. Biophys. Res. Commun.* **30:** 232–239.

Souciet, J.-L., Aigle, M., Artiguenave, F., Blandin, G., Bolotin-Fukuhara, M., Bon, E., Brottier, P., Casaregola, S., de Montigny, J., Dujon, B., et al. 2000. Genomic exploration of the hemiascomycetous yeasts: 1. A set of yeast species for molecular evolution studies. *FEBS Lett.* **487:** 3–12.

Svetlov, V.V. and Cooper, T.G. 1995. Review: Compilation and

characteristics of dedicated transcription factors in *Saccharomyces cerevisiae*. *Yeast* **11:** 1439–1484.
Tagle, D.A., Koop, B.F., Goodman, M., Slightom, J.L., Hess, D.L., and Jones, R.T. 1988. Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.* **203:** 439–455.
Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.
Thorne, J.L. and Churchill, G.A. 1995. Estimation and reliability of molecular sequence alignments. *Biometrics* **51:** 100–113.
Vaughan-Martini, A., Martini, A., and Cardinali, G. 1993. Electrophoretic karyotyping as a taxonomic tool in the genus *Saccharomyces*. *Antonie Leeuwenhoek* **63:** 145–156.
Venkatesh, B., Si-Hoe, S.L., Murphy, D., and Brenner, S. 1997. Transgenic rats reveal functional conservation of regulatory controls between the Fugu isotocin and rat oxytocin genes. *Proc. Natl. Acad. Sci.* **94:** 12462–12466.
Wendl, M.C., Dear, S., Hodgson, D., and Hillier, L. 1998. Automated sequence preprocessing in a large-scale sequencing environment. *Genome Res.* **8:** 975–984.
Wentworth, J.M., Schoenfeld, V., Meek, S., Elgar, G., Brenner, S., and Chatterjee, V.K. 1999. Isolation and characterisation of the retinoic acid receptor-alpha gene in the Japanese pufferfish, *F. rubripes*. *Gene* **236:** 315–323.
Wilson, R.K. and Mardis, E.R. 1997. Flourescence-based DNA sequencing. In *Genome analysis: A laboratory manual* (ed. B. Birren et al.) Vol. I, pp. 301–395. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Received February 1, 2001; accepted in revised form April 11, 2001.

Surveying *Saccharomyces* Genomes to Identify Functional Elements by Comparative DNA Sequence Analysis

Paul F. Cliften, LaDeana W. Hillier, Lucinda Fulton, et al.

Genome Res. 2001 11: 1175-1186
Access the most recent version at doi:10.1101/gr.182901