

Table S1

Percent nucleotide identity to *S. cerevisiae* of *Saccharomyces species* intergenic sequences

	% local similarity intergenic regions	% global similarity intergenic regions	% similarity synonymous codons
<i>S. mikatae</i>	71.0	66.8	57.5
<i>S. kudriavzevii</i>	69.0	62.9	54.6
<i>S. bayanus</i>	67.0	58.5	51.3
<i>S. castellii</i>	na	na	34.1
<i>S. kluyveri</i>	na	na	33.7

The identity in local alignments of intergenic sequences was calculated using WU-BLASTN (<http://blast.wustl.edu>); global sequence identity was calculated with GAP (GCG Corp.). Identity of synonymous codons (four-fold degenerate codons) was calculated based on amino acid sequence alignments. Identity in intergenic sequences is higher than in neutral sequences (synonymous third position sequences).

Table S2

Summary of genome sequence assembly.

	<i>S. mikatae</i>	<i>S. kudriavzevii</i>	<i>S. bayanus</i>	<i>S. castellii</i> *	<i>S. kluyveri</i>
# of contigs:	2,902	2,074	2,528	741	2,446
Ave. contig length:	3.4 kb	4.9kb	4.0kb	14.8kb	4.2kb
Total assembly length:	10.0Mb	10.6Mb	10.2Mb	11.1Mb	10.2Mb
#of reads in contigs:	54,027	68,429	60,813	78,133	75,312
%Phred 20 bases:	98.5	98.9	98.5	99.5	98.5
%Phred 40 bases:	92.4	94.7	92.9	97.4	93.5
Assembly redundancy: (Fold coverage)**	2.8	3.4	2.9	3.9	3.6

* Gaps spanning protein-coding regions were finished in *S. castellii*.

** Fold sequencing coverage of assembled contigs

Over 50 Mb of assembled sequence was generated. At least 92% of the assembled sequence has a Phred quality score of 40 or greater (estimated error rate of less than 1/10,000); more than 98% of the nucleotides have a Phred score of 20 or greater (estimated error rate of less than 1/100).

For comparison, the prefinished assembly information of two genomes is listed below:

S. mikatae: 4,033 contigs, 2.3kb/contig, 9.1Mb, 48,897 reads, P20=98.0, P40=90.7, 2.7X.

S. castellii: 2,050 contigs, 5.1/kb/contig, 10.5Mb, 73,683 reads, P20=99.2, P40=95.9, 3.8X

The gap-closing effort for *S. mikatae*, the genome with the lowest ‘shotgun’ coverage, yielded nearly 1 megabase of additional sequence and eliminated over 1,000 gaps (approximately 30% of the gaps are non-coding). In *S. castellii*, the genome with the highest ‘shotgun’ coverage, we closed both coding and non-coding gaps between linked contigs. This eliminated over 1,300 gaps (64% of the total). Of the resulting 741 contigs, approximately 400 are of high quality and are contained in approximately 150 scaffolds. The Sequences are available from Genbank (project accession numbers *S.kluyveri*, AACF000000000; *S. castellii*, AACG000000000; *S. bayanus*, AACI000000000; *S.mikatae*, AACF000000000; *S. kudriavzevii*, AACI000000000) and SGD (<http://genome-www.stanford.edu/Saccharomyces>) and our web site (http://www.genetics.wustl.edu/bclab/yeast_web/yeastseq_home.html).

Table S3

Distribution of exact n-mer sequences conserved in orthologous promoters of closely and distantly related species.

	Number of n-mers in:	
	Real alignments	shuffled alignments
		(\pm std. dev.)
10mers	242	.1 +- 0.3
9mers	631	1.5 +- 1.3
8mers	1,858	34.5 +- 7.2
7mers	7,590	1,002.2 +- 36.1
6mers	41,938	21,733.7 +- 148.7

Exact n-mers present in orthologous promoter sequences from both of the distantly related species and at least 3 *sensu stricto* species were counted and compared to the number of random n-mers present in 100 shuffled intergenic region sets.

Table S4

Categorization of the 160 most highly conserved sequence motifs in comparisons of 6 genomes

A. Candidate novel motifs (9)

Motif	Found in:	Similar motif with functional enrichment	enrichment category	P-value
TGTTGGAAAA	CAX4	TGTTGGNNAA	lipid metabolism (3 of 3)	0.0000201
AATTTTCTAG	HSP104	TTTTCTAG	protein folding (3 of 7)	0.0000083
GTAAATACTA	RPL5	GTAAATAC	protein synthesis (4 of 6)	0.00175
CACATACATT	TSR1	CACATAC	protein synthesis (4 of 11)	0.02469

No apparent enrichment in functionally similar genes:

TTTCTTGAAAG	RPL27A
CAGAACGTTT	ASC1
AGTCAATTGA	TAL1
TAGTCAATTG	TAL1
TCGTTTCTTTT	YNL201C

B. Putative UTR sequences (11)

TTTGGAAGAC	RPL1B
TAAAGAAATAGTC	RPP1A
GATTATCTCTA	RPP1A
AAGGAATAGTC	RPP2B.
CTGTCTGAAGAAGTCA	RPS12
AATCAGTCAAA	RPS13
AAAAGGTATAT	RPS20
TTGCCGACAAGCCA	RPS31
AGAAAGAAGCA	RPS31
TGCAAGGAACTTGAA	SNU13
TGTCGATTGAAAGTTACCTACATCAACTTTCCGTG	RPS2

Supplemental Table 4, continued

C. AT rich sequences (46)

AAAAAAAAAAA (4 occurrences)
AAAAAAAAAAG (2 occurrences)
AAAAAAACT
AAAAAAGAAA (2 occurrences)
AAAAAGAAAA
AAAAGAAAAA
AGAAAAAAAAA
ATTTTTTTTT (3 occurrences)
GAAAAAAAAAAA
GAAAAAAAAAA (5 occurrences)
GAAAAGAAAA
TTTCTTTTTT
TCTTTTTTTT
TTTCTTTTTT
TTTCTTTTTT
TTTTCTTTTT (2 occurrences)
TTTTTTTTCA
TTTTTTTCTT
TTTTTTTTTC (4 occurrences)
TTTTTTTTTTTC (2 occurrences)
TTTTTTTTTTT (9 occurrences)
TTTTTTTTTTTT

D. Known Motifs (94)

RRPE (54 occurrences)
PAC (18 occurrences)
PACE (10 occurrences)
TBP
Mbf1 (2 occurrences)
Gcr1
Rgt1
Bas2
Rap1
Ume6
SCB
Cst6
Ndt80
Hcm1

Figure S1. Profile of the average sequence conservation upstream of 75 ribosomal protein coding genes illustrating highly conserved sequence immediately upstream of the ATG start codon.

Figure S2. Distribution of conserved n-mer sequences present in CLUSTALW alignments of *sensu stricto* species' intergenic regions and n-mers present in shuffled (randomized) alignments of the same sequences. Intergenic sequences with 40% or less identity over the length of the 4-way alignments (2377 of 3523 alignments) were used to eliminate the most highly conserved sequences. Non-overlapping n-mers (ranging in length from 6 to 30) were extracted from CLUSTALW alignments of sequences of four *sensu stricto* species, and from alignments where the individual columns (all four sequences) had been randomly shuffled (thereby maintaining the exact percent of sequence identity). The distribution of 6 to 16-mers is shown here. Data from the real promoters are shown in blue; and data from the shuffled promoters is shown in purple.

Figure S3. Histogram of the number of 8,873 n-mers that match one of 71 known regulatory sequence motifs. Only the 25 most frequent sequence motifs are shown. The remaining 35 sequence motifs are conserved less than 20 times in the 2,377 sequence alignments of *sensu stricto* species.

Figure S1

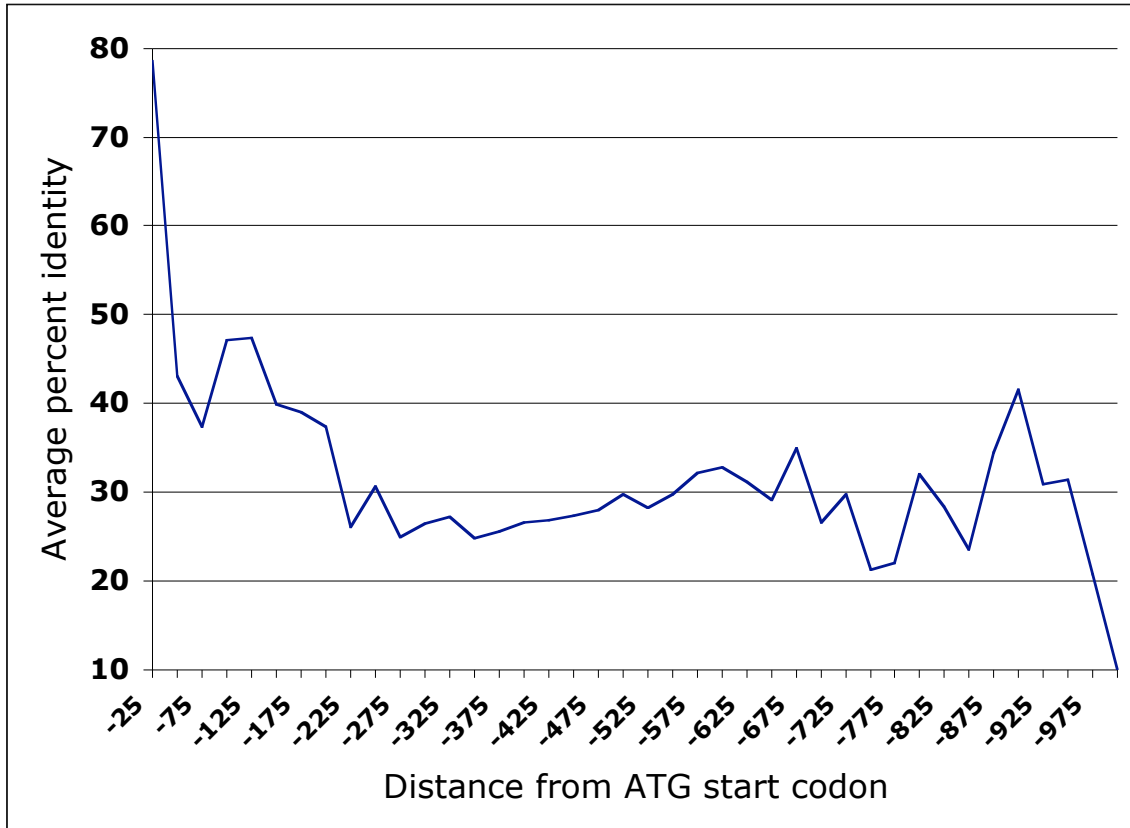


Figure S2.

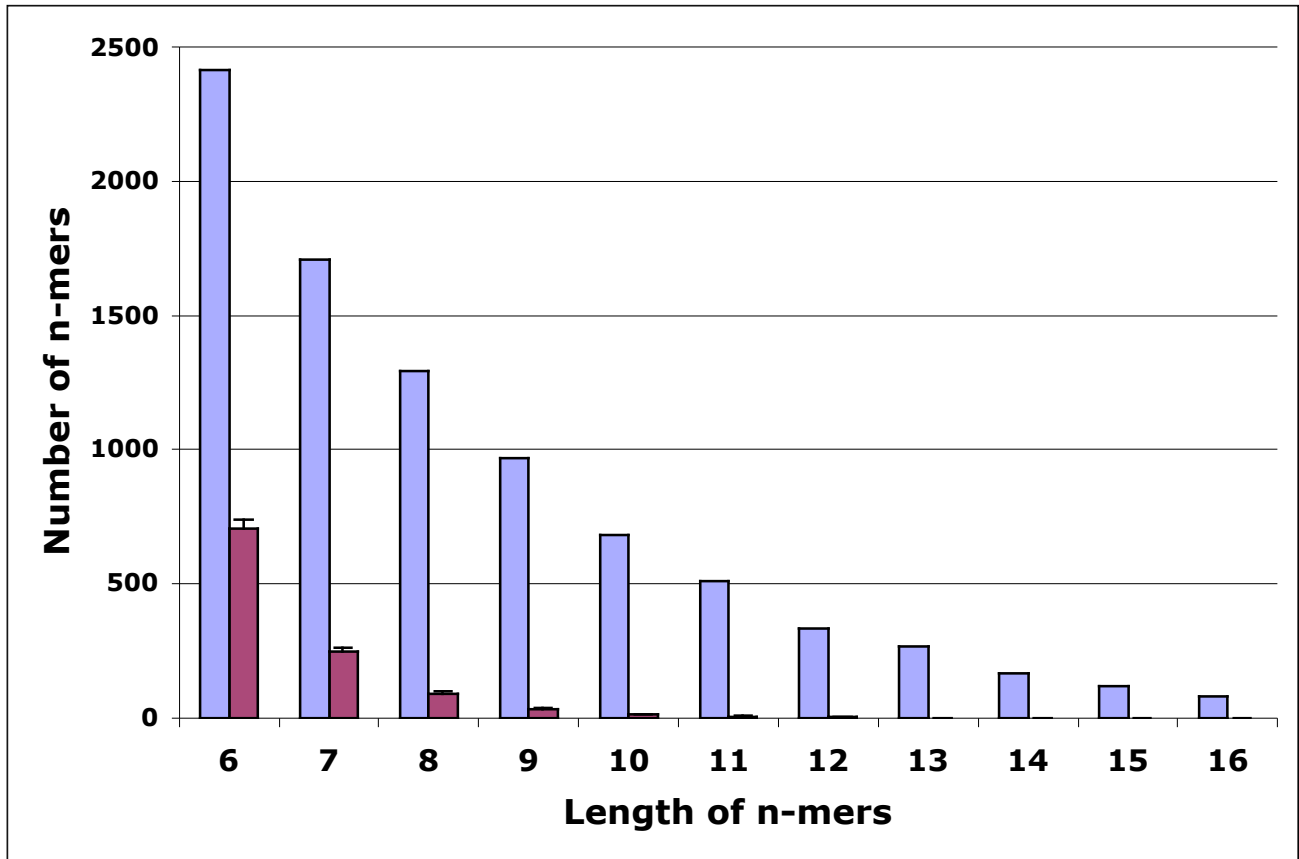


Figure S3.

