Supplemental materials for
The origin and evolution of a distinct mechanism of transcription initiation in yeasts

Zhaolian Lu[1] and Zhenguo Lin[1]*

[1] Department of Biology, Saint Louis University, St. Louis, MO 63104

*Corresponding author:

E-mail: zhenguo.lin@slu.edu

Keywords: transcription initiation, core promoter, evolution, yeast

Running Title: Evolution of transcription initiation mechanisms in yeasts

Supplemental Table S1. Yeast strains and resources used in this study

Species name	Abbreviation	Strain	Source	Mapping reference
Saccharomyces cerevisiae	*S. cerevisiae*	BY4741	Dr. Zhenglong Gu lab, Cornell University	R64-2-1, SGD
Saccharomyces paradoxus	*S. paradoxus*	N17	Dr. Justin Fay lab, University of Rochester	N17, Justin Fay
Saccharomyces mikatae	*S. mikatae*	unknown	Dr. Zhenglong Gu lab, Cornell University	v7-Aug2012, YGOB
Saccharomyces bayanus	*S. bayanus*	IFO11022	NBRP--YGRC	GCA_000167035.1_ASM16703v1, NCBI
Naumovozyma castellii	*N. castellii*	IFO1992	NBRP--YGRC	GCF_000237345.1_ASM23734v1, NCBI
Kluyveromyces lactis	*K. lactis*	KB101	Dr. Zhenglong Gu lab, Cornell University	GCA_000149225.1_ASM14922v1, NCBI
Lachancea kluyveri	*L. kluyveri*	MATα	NBRP--YGRC	v7-Aug2012, YGOB
Kluyveromyces waltii	*K. waltii*	IFO1666	NBRP--YGRC	v7-Aug2012, YGOB
Candida albicans	*C. albicans*	unknown	Dr. Zhenglong Gu lab, Cornell University	GCF_000182965.3_ASM18296v3, NCBI
Yarrowia lipolytica	*Y. lipolytica*	MU4	NBRP—YGRC	GCF_000002525.2_ASM252v1, NCBI
Schizosaccharomyces pombe	*Sch. pombe*	JB1171	NBRP—YGRC	PomBase
Schizosaccharomyces japonicus	*Sch. japonicus*	yFS275	NBRP—YGRC	GCA_000149845.2, NCBI

Supplemental Table S2. Summary of CAGE sequencing and mapping results

Samples	Total # of reads	Overall alignment rate	Uniquely mapped reads	Unique mapping rate
S. cerevisiae -1	29,740,675	93.63%	17,334,431	58.29%
S. cerevisiae -2	30,005,902	93.95%	20,182,126	67.26%
S. paradoxus -1	41,471,738	88.88%	21,039,161	50.73%
S. paradoxus -2	43,638,383	83.13%	20,160,182	46.20%
S. mikatae -1	28,709,901	95.60%	26,510,250	92.34%
S. mikatae -2	42,236,831	95.92%	39,162,149	92.72%
S. bayanus -1	36,374,870	80.45%	18,525,455	50.93%
S. bayanus -2	32,652,793	80.51%	16,282,777	49.87%
S. castellii -1	36,763,702	96.25%	29,365,574	79.88%
S. castellii -2	37,560,080	96.06%	29,540,910	78.65%
L. kluyveri -1	26,800,269	68.96%	16,983,527	63.37%
L. kluyveri -2	32,170,247	67.46%	20,011,758	62.21%
K. waltii -1	28,644,338	95.11%	24,927,545	87.02%
K. waltii -2	28,330,602	95.34%	23,131,422	81.65%
K. lactis -1	34,812,967	96.16%	28,365,001	81.48%
K. lactis -2	35,553,041	95.92%	31,714,767	89.20%
C. albicans -1	43,896,423	89.40%	37,438,890	85.29%
C. albicans -2	37,818,481	89.25%	32,268,142	85.32%
Y. lipolytica -1	32,372,731	78.12%	24,283,938	75.01%
Y. lipolytica -2	36,749,734	67.80%	24,144,938	65.70%
Sch. pombe -1	41,068,801	90.87%	32,673,739	79.56%
Sch. pombe -2	35,638,178	90.11%	28,826,177	80.89%
Sch. japonicus -1	32,185,138	94.91%	26,170,714	81.31%
Sch. japonicus -2	33,428,849	95.02%	27,978,646	83.70%
Total	838,624,674		617,022,219	

Supplemental Table S3. Numbers of TSS and core promoter identified based on CAGE sequencing in each species

Sample	Uniquely mapped reads	# of TSS	# of Core promoters
S. cerevisiae	37,516,557	261,475	10,175
S. paradoxus	41,199,343	399,673	13,441
S. mikatae	65,672,399	239,156	8,684
S. bayanus	34,808,232	231,965	9,000
S. castellii	58,906,484	239,429	9,032
L. kluyveri	36,995,285	177,991	8,519
K. waltii	48,058,967	262,191	11,379
K. lactis	60,079,768	322,306	10,735
C. albicans	69,707,032	346,672	13,283
Y. lipolytica	48,428,876	410,348	14,480
Sch. pombe	61,499,916	259,555	11,879
Sch. japonicus	51,170,260	286,433	10,964

Supplemental Table S4. Numbers of TATA-containing promoters and genes identified based on two versions of TATA box consensus sequences

	TATAWAWR		TATAWAW	
Species	**Promoter #**	**Gene #**	**Promoter #**	**Gene #**
S. cerevisiae	1556	1492	1989	1989
S. paradoxus	1622	1539	2015	2015
S. mikatae	1403	1342	1863	1863
S. bayanus	1034	998	1312	1312
S. castellii	1323	1285	1763	1763
L. kluyveri	899	878	1190	1190
K. waltii	928	884	1358	1358
K. lactis	1542	1445	1889	1889
C. albicans	1538	1438	1974	1974
Y. lipolytica	1023	977	1288	1288
Sch. pombe	1290	1222	1850	1850
Sch. japonicus	643	621	878	878

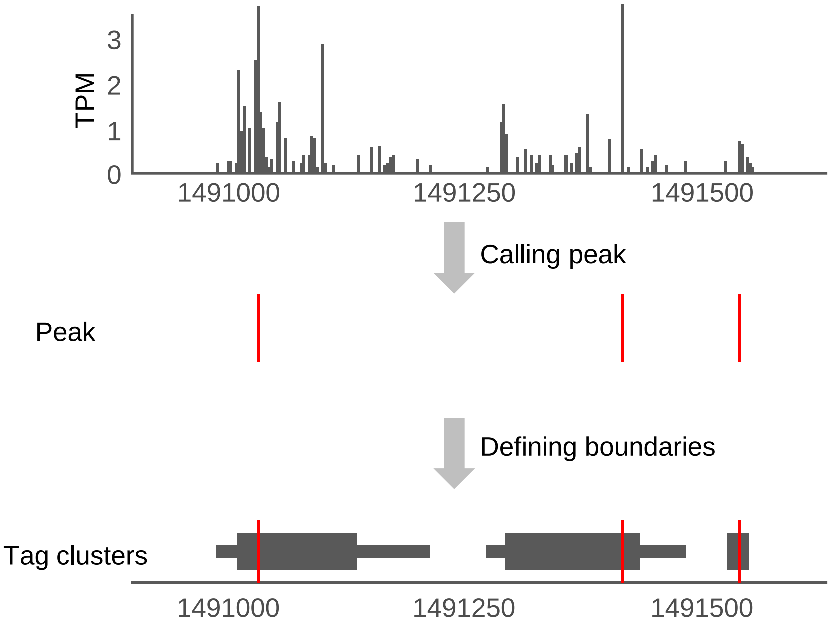
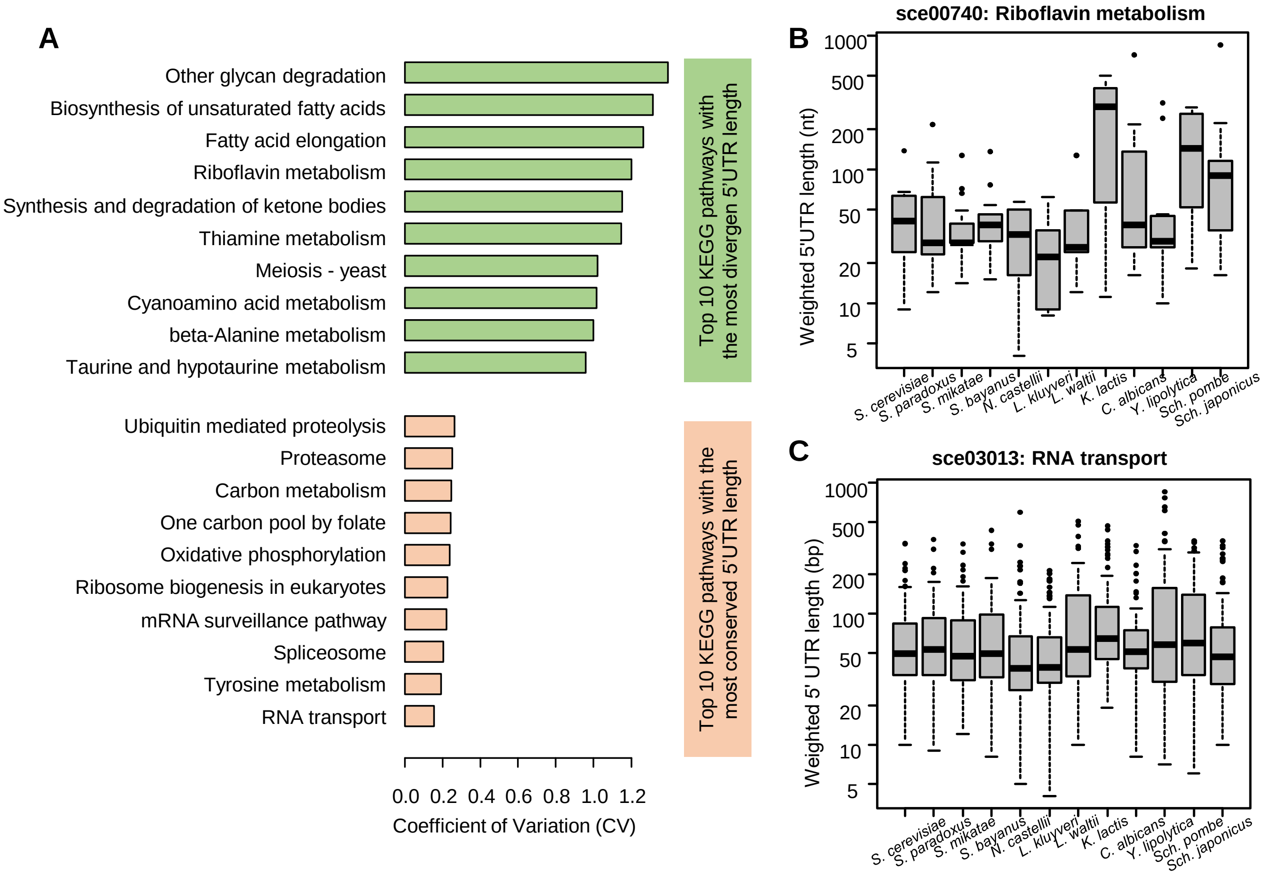
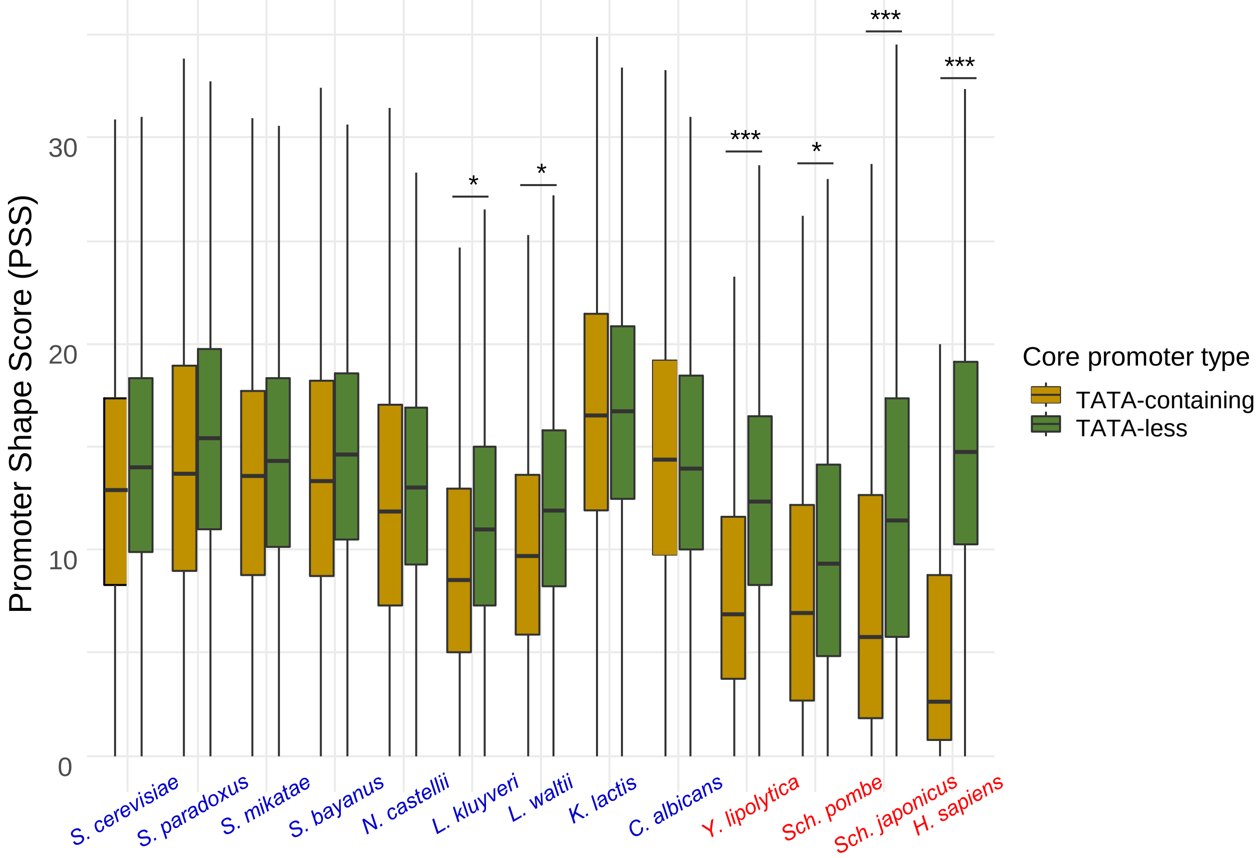


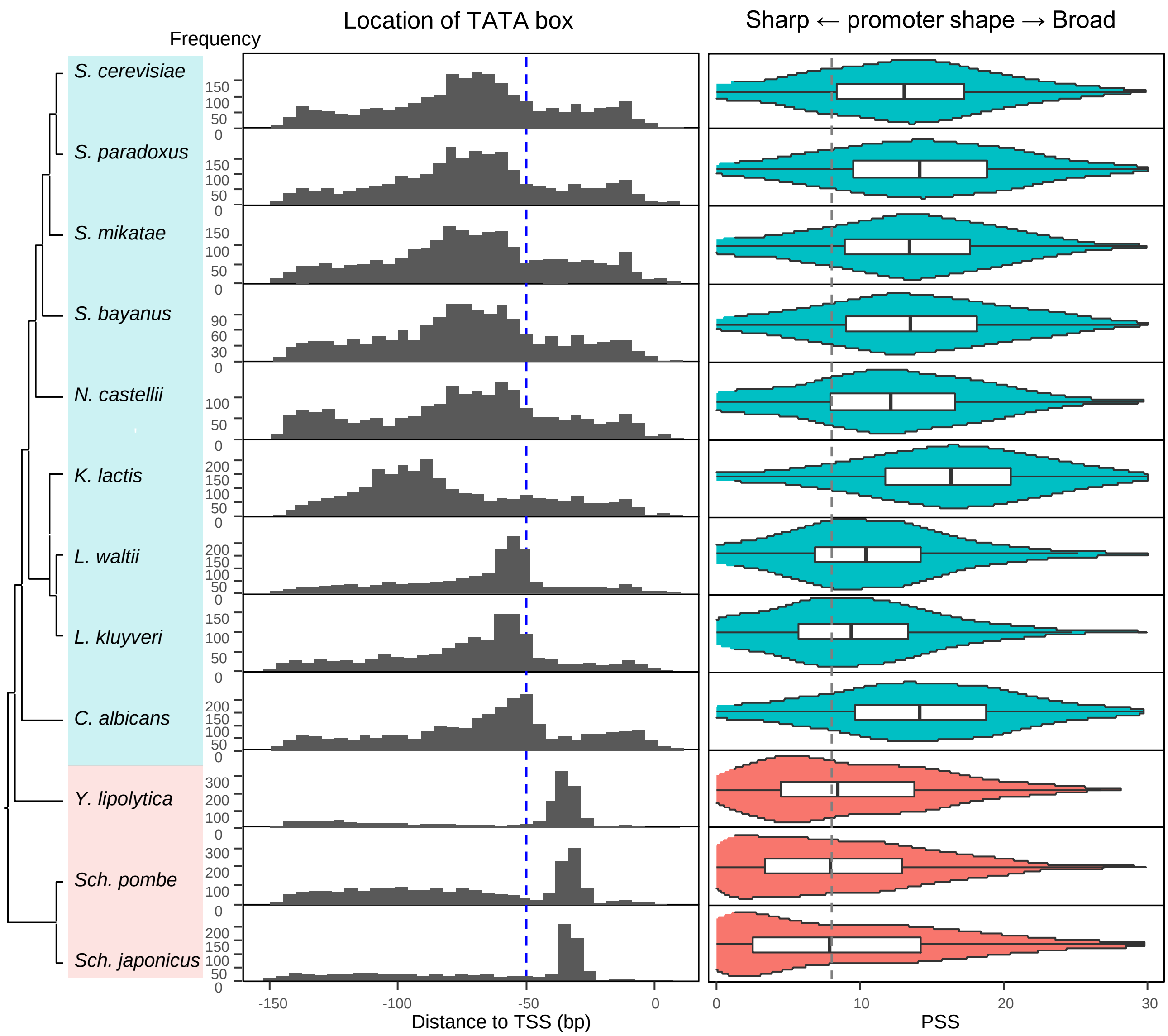
Figure S1. Schematic illustration of peak-based clustering algorithm "Peakclu". Peakclu first uses a sliding-window approach to identify peak signals (TSSs with highest TPM values within a genomic region) in a given window. A peak is considered as the dominant TSS in a tag cluster (TC) or core promoter. The window size is defined by user, and the default size is 100 bp, which means that only one peak or one core promoter can be present in a 200 (±100) bp region. In the next step, Peakclu aggregates TSSs near a peak to form a TC based on the user-defined maximum TSS distance. For example, with a defined maximum TSS distance of 30 bp, Peakclu search for neighboring TSSs both upstream and downstream of a peak, if a TSS is located within 30 bp, the TSS will be grouped with the peak and the boundary of this TC is extended to this TSS. Peakclu continues to group TSSs with 30 bp of its new boundaries until no more TSS is found within this range. Because some outliers may significantly increase the range of a TC (as indicated by a thinner box in the bottom track), Peakclu uses the positions of the 10th and 90th percentile of CAGE signals as the boundaries of a TC (as indicated by a thinker box in the bottom track). The source code for "Peakclu" and other R scripts for data analysis is available on GitHub (https://github.com/Linlab-slu/TSSr).



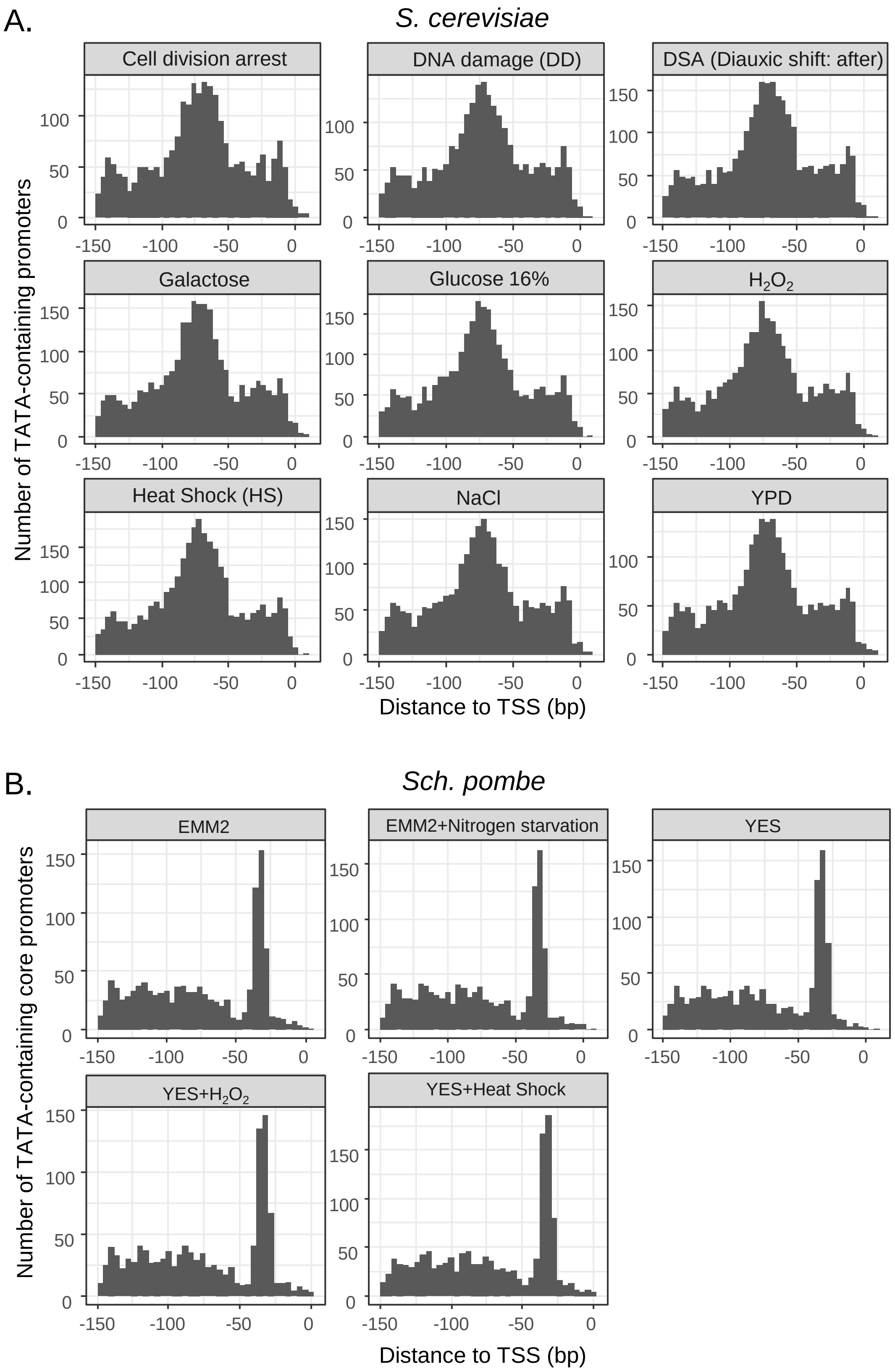
Supplemental Figure S2. Different evolutionary patterns of 5' UTR length among KEGG pathways. (A) The top ten KEGG pathways with the most divergent (with highest CV values) and most conserved (with lowest CV values) 5' UTR length. CV values are displayed as the length of bars. (B) An example of the KEGG pathway (Riboflavin metabolism pathway) with the most divergent 5' UTR length. The distributions of 5' UTR lengths in each species are illustrated by boxplots. (C) An example of the KEGG pathway (RNA transport pathway) with the most conserved 5'UTR length.



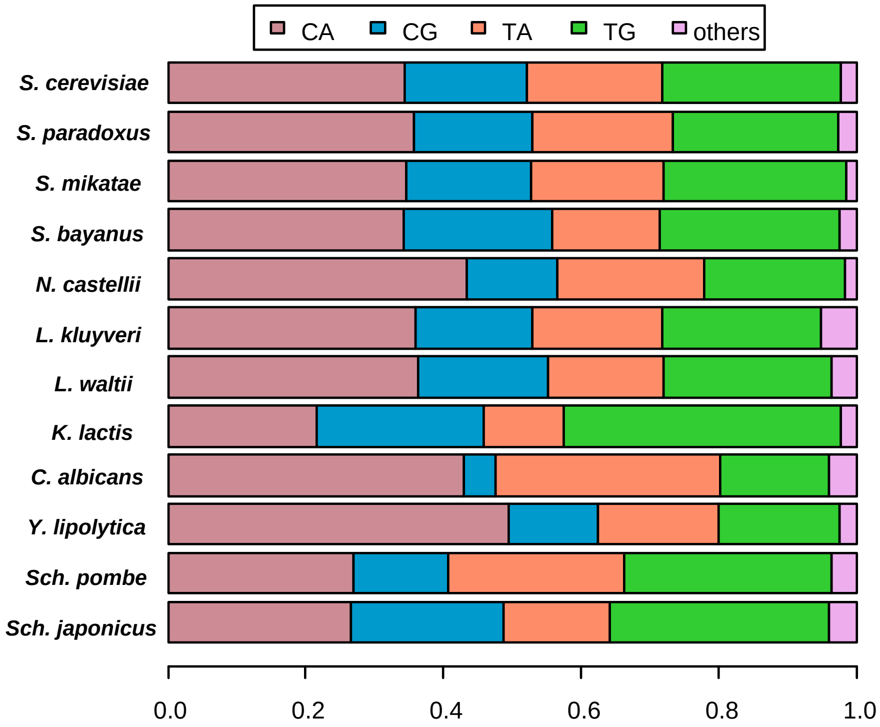
Supplemental Figure S3. Distribution of promoter shape score (PSS) of TATA-containing and TATA-less promoters in 12 yeast species and human. The names of "classic model" species are in red, and the names of "scanning model" species are in blue. In all "classic model" species, the TATA-containing promoters have significantly lower PSS values (sharper promoter shape) than TATA-less promoters. *: $p < 0.01$; **: $p < 0.001$; ***: $p = 0$, t-test.



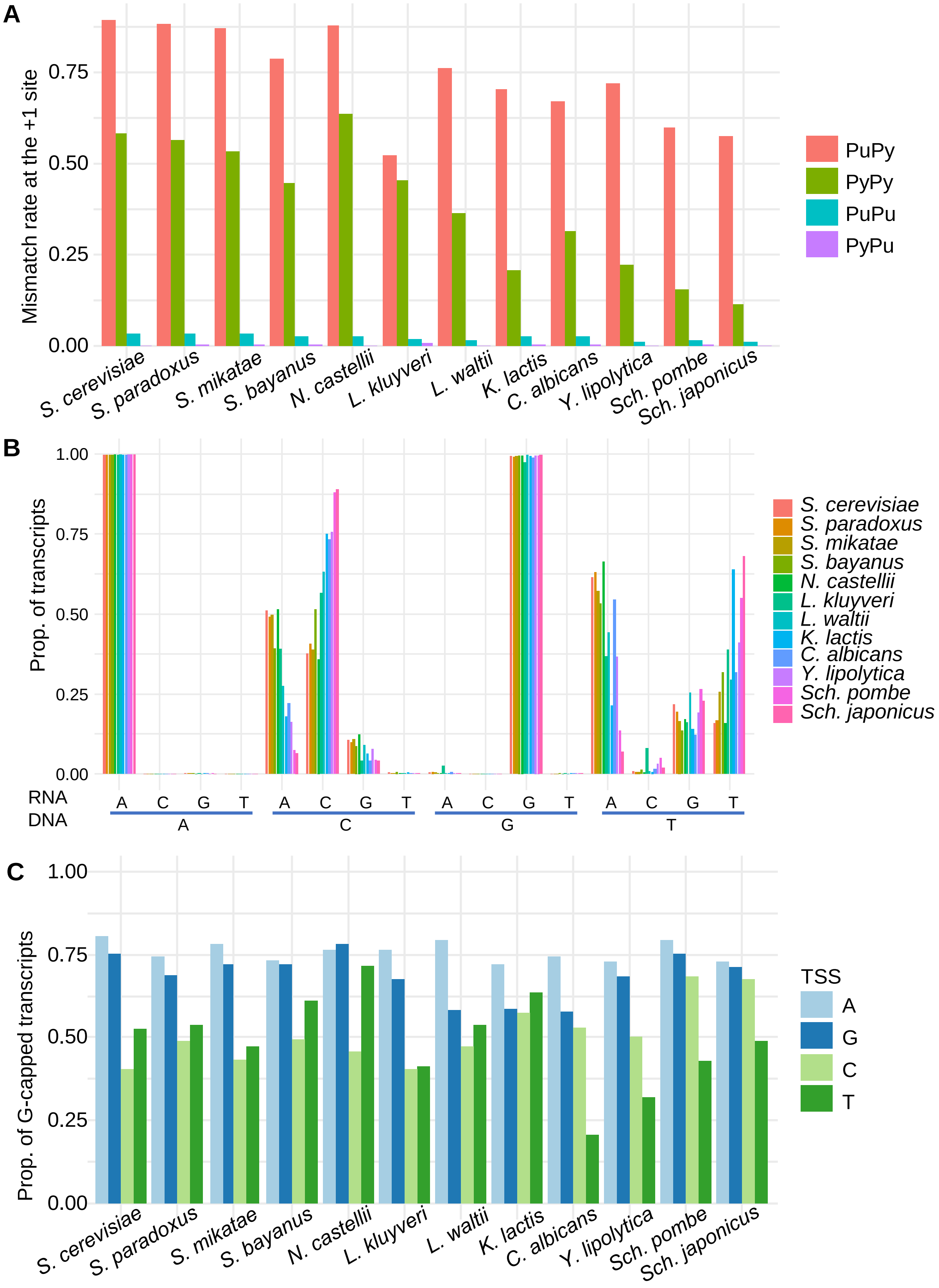
Supplemental Figure S4. Evolution of transcription initiation mechanisms from TATAWAW-containing promoters. The left panel displays the phylogenetic relationships of the 12 yeast species. The “scanning model” are shaded in cyan, and species names of “classic model” are shaded in orange. TATA box motifs were identified based on a consensus sequence of “TATAWAW”. The middle panel shows the distributions of distances between TATA box and TSSs in each species. Blue dashed line refers to the site of -50. The right panel shows the promoter shape score (PSS) of TATA box-containing core promoters. The grey dashed line indicates the median PSS value in *Y. lipolytica*, *Sch. pombe*, and *Sch. japonicas*.



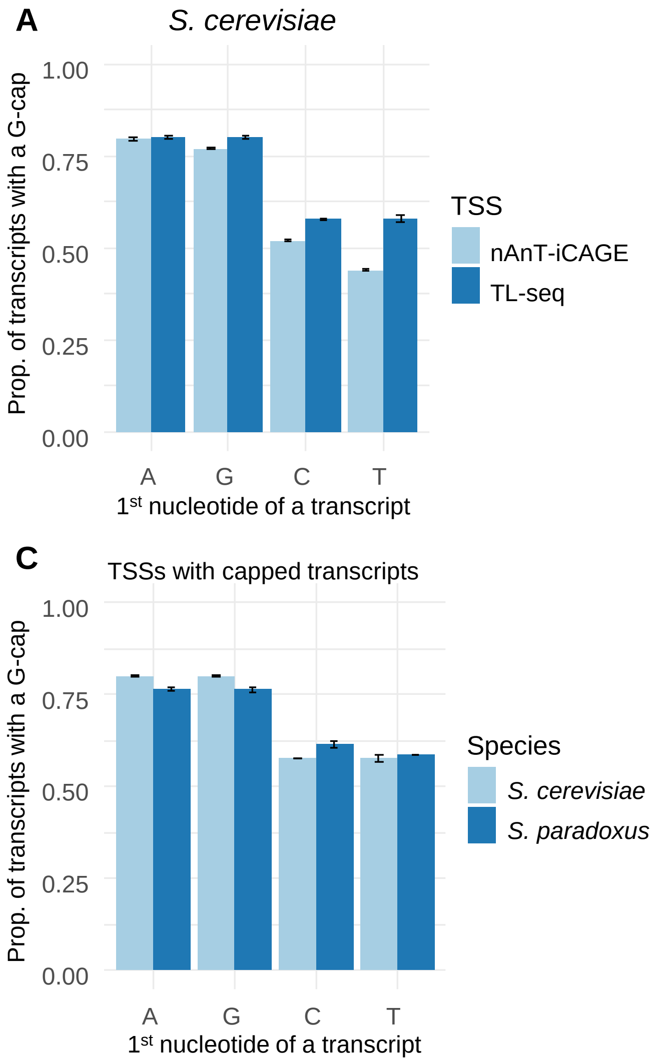
Supplemental Figure S5. Distributions of TATA box locations are independent of physiological regulation in both "scanning model" and "classic model" species. The TSS maps of (A) *S. cerevisiae* of nine growth conditions, and (B) *Sch. pombe* of five growth conditions were retrieved from (Lu and Lin 2019) and (Thodberg et al. 2019), respectively. TATA box motifs were identified based on a consensus sequence of "TATAWAW". YPD: yeast extract peptone dextrose; EMM2: Edinburgh minimal medium; YES: yeast extract with supplements.



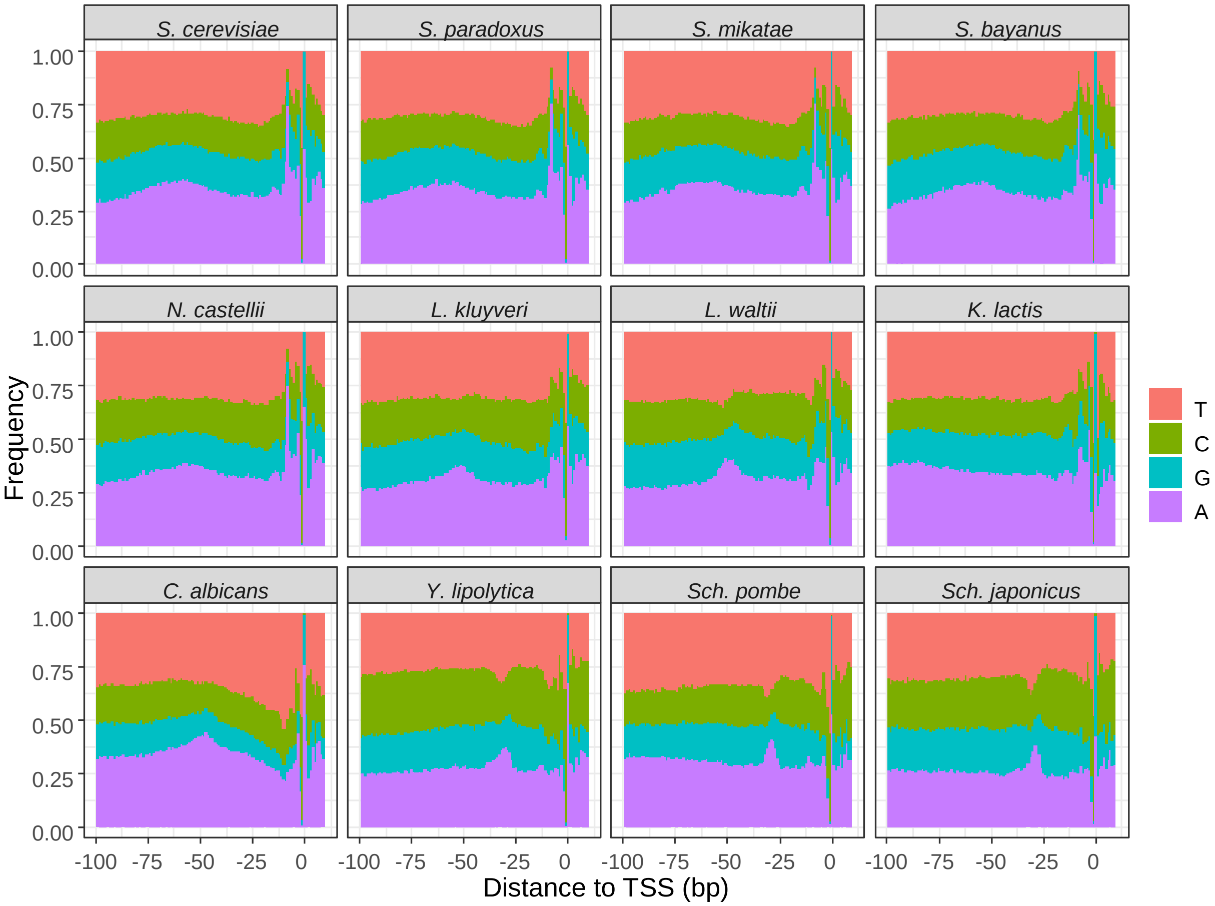
Supplemental Figure S6. Distributions of PyPu and all the other dinucleotides at the [−1/+1] sites in the 12 yeast species. The [−1/+1] dinucleotides examined are the dominant TSSs of all core promoters with TPM >1 in each species.


A
Mismatch rate at the +1 site
0.00
0.25
0.50
0.75
S. cerevisiae
S. paradoxus
S. mikatae
S. bayanus
N. castellii
L. kluyveri
L. waltii
K. lactis
C. albicans
Y. lipolytica
Sch. pombe
Sch. japonicus
PuPy
PyPy
PuPu
PyPu
B
Prop. of transcripts
0.00
0.25
0.50
0.75
1.00
RNA
DNA
A
C
G
T
A
C
G
T
A
C
G
T
A
C
G
T
A
C
G
T
S. cerevisiae
S. paradoxus
S. mikatae
S. bayanus
N. castellii
L. kluyveri
L. waltii
K. lactis
C. albicans
Y. lipolytica
Sch. pombe
Sch. japonicus
C
Prop. of G-capped transcripts
0.00
0.25
0.50
0.75
1.00
S. cerevisiae
S. paradoxus
S. mikatae
S. bayanus
N. castellii
L. kluyveri
L. waltii
K. lactis
C. albicans
Y. lipolytica
Sch. pombe
Sch. japonicus
TSS
A
G
C
T

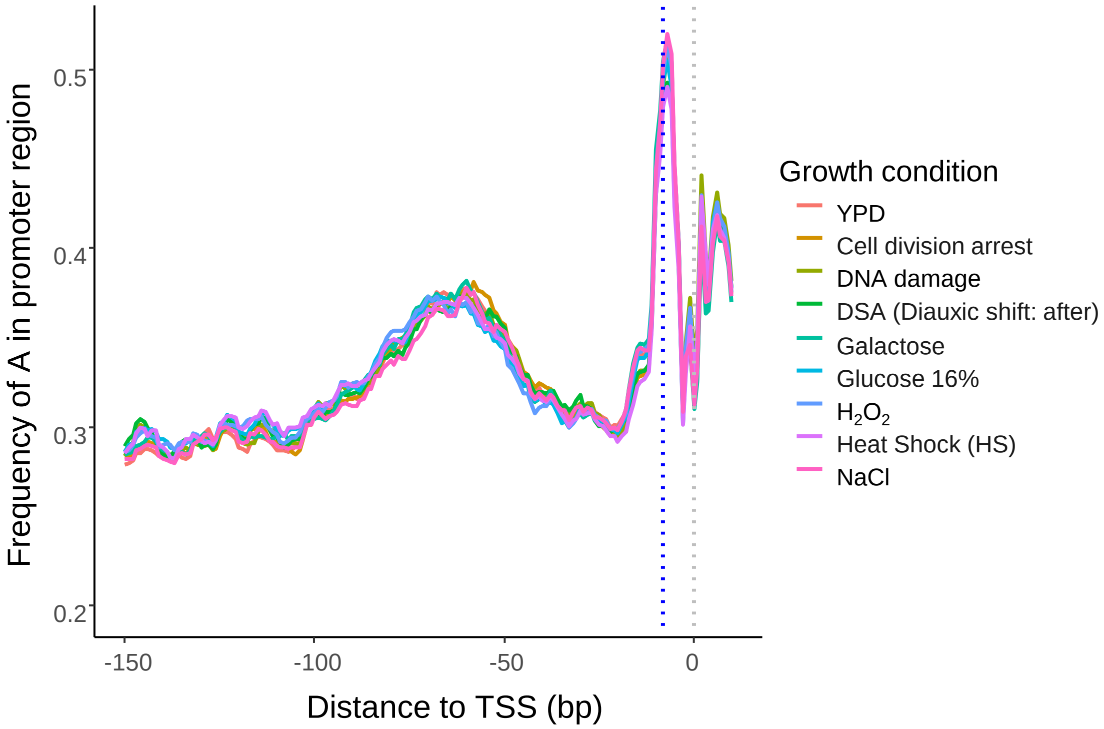
Supplemental Figure S7. Transcription initiation from a purine at the +1 site reduces mismatch rates. (A) Transcription from TSS with PyPu at the [−1/+1] sites has a significantly lower mismatch rate than other dinucleotides in each species. (B) The proportion of each type of nucleotides recruited by Pol II to the +1 site with the sense strand of DNA with A, C, G and T, at the TSS position, respectively. Mismatches in transcripts that are initiated with a purine at the sense strand are nearly depleted. A large proportion of purines are recruited even if the sense strand DNA has a pyrimidine at the +1 site. (C) Transcripts with a purine at the +1 site have a significantly higher chance of being capped with m7G than others in each species.



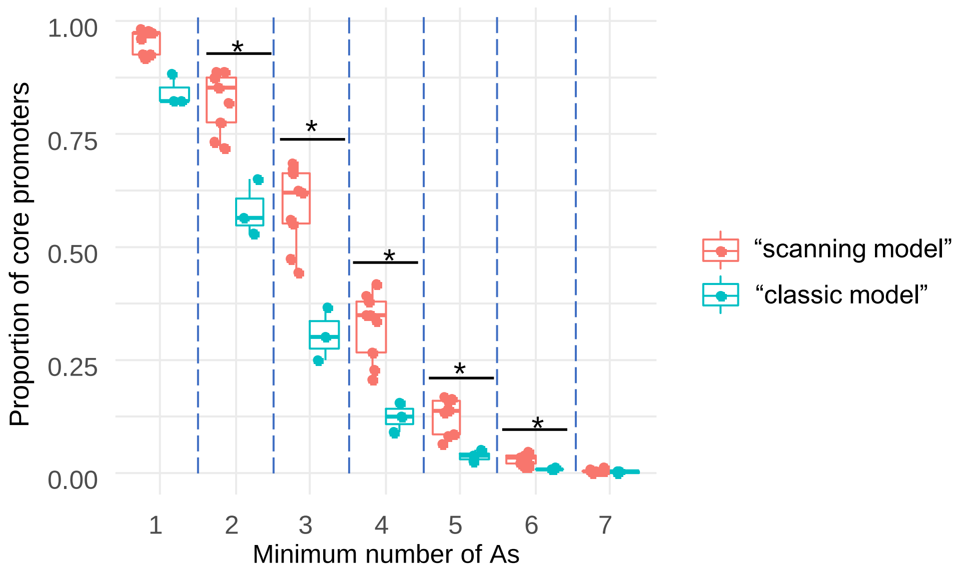
Supplemental Figure S8. Comparison of capping rates based on TSSs identified by techniques or criteria. A and B: Capping rates of the four different 1st nucleotide of transcripts initiated from TSSs identified by CAGE in this study, or by both of CAGE and TL-seq. The raw sequencing data of TL-seq in (A) *S. cerevisiae* and (B) *S. paradoxus* were obtained from (Spealman et al. 2018). The TSSs identified by TL-seq with TPM ≥ 1 were selected for calculation of G-capped rates (light blue bars). The G-capped rates of TSSs identified by both studies were calculated based on nAnT-iCAGE reads. C: Capping rates of the four different 1st nucleotide of transcripts initiated from high-confident TSSs in *S. cerevisiae* and *S. paradoxus*. The high-confident TSSs were identified as TSSs within TC and are supported by at least one capped-transcript based on our CAGE data



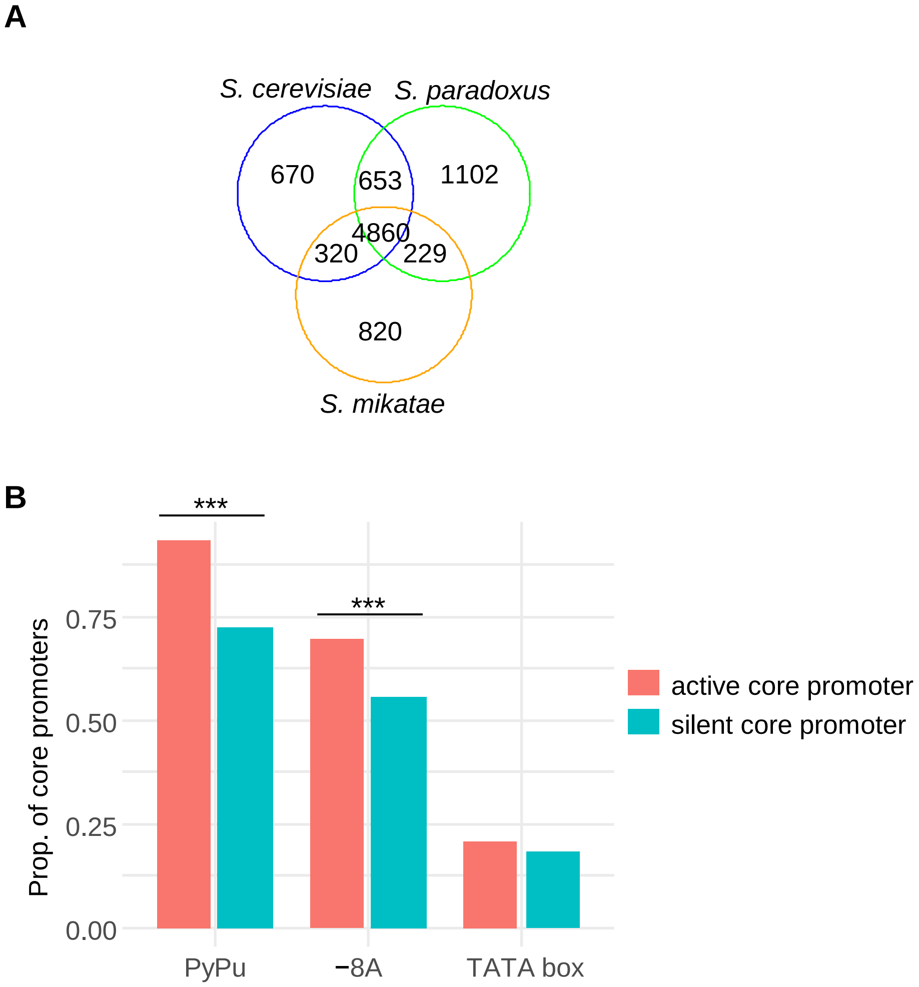
Supplemental Figure S9. The frequency of A,T,G and C in promoter region around the TSS (from − 100 to +10) in the 12 yeast species. The dominant TSS of each core promoter was used as distance 0. Only core promoters with TPM > 1 were used for this analysis.



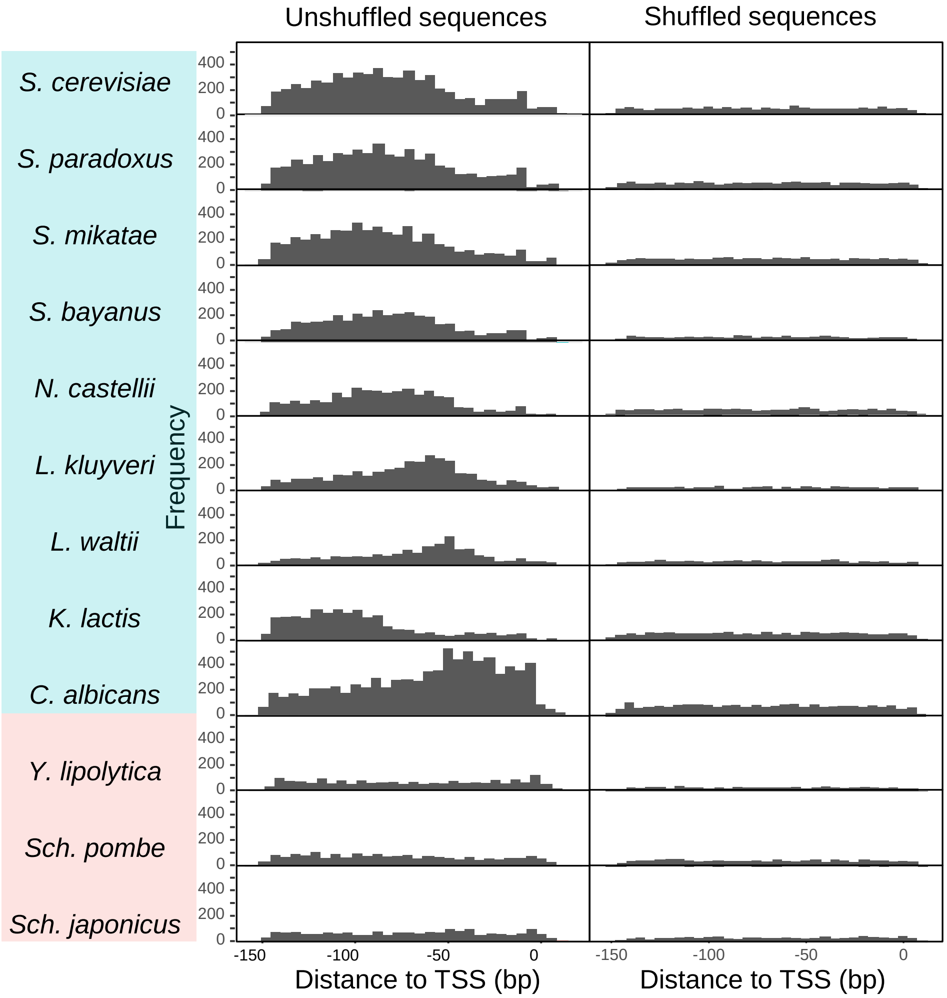
Supplemental Figure S10. A sliding window analysis of A frequency in the promoter regions of *S. cerevisiae*. The TSS maps of S. cerevisiae under nine different growth conditions were obtained from (Lu and Lin, 2019). Window size is 5 bp with a step size of 1 bp. The peak in the region [−100, − 50] indicates locations of the TATA box. Blue dashed line refers to the position 8 bp upstream of the TSS. The grey dashed line refers to the TSS position.



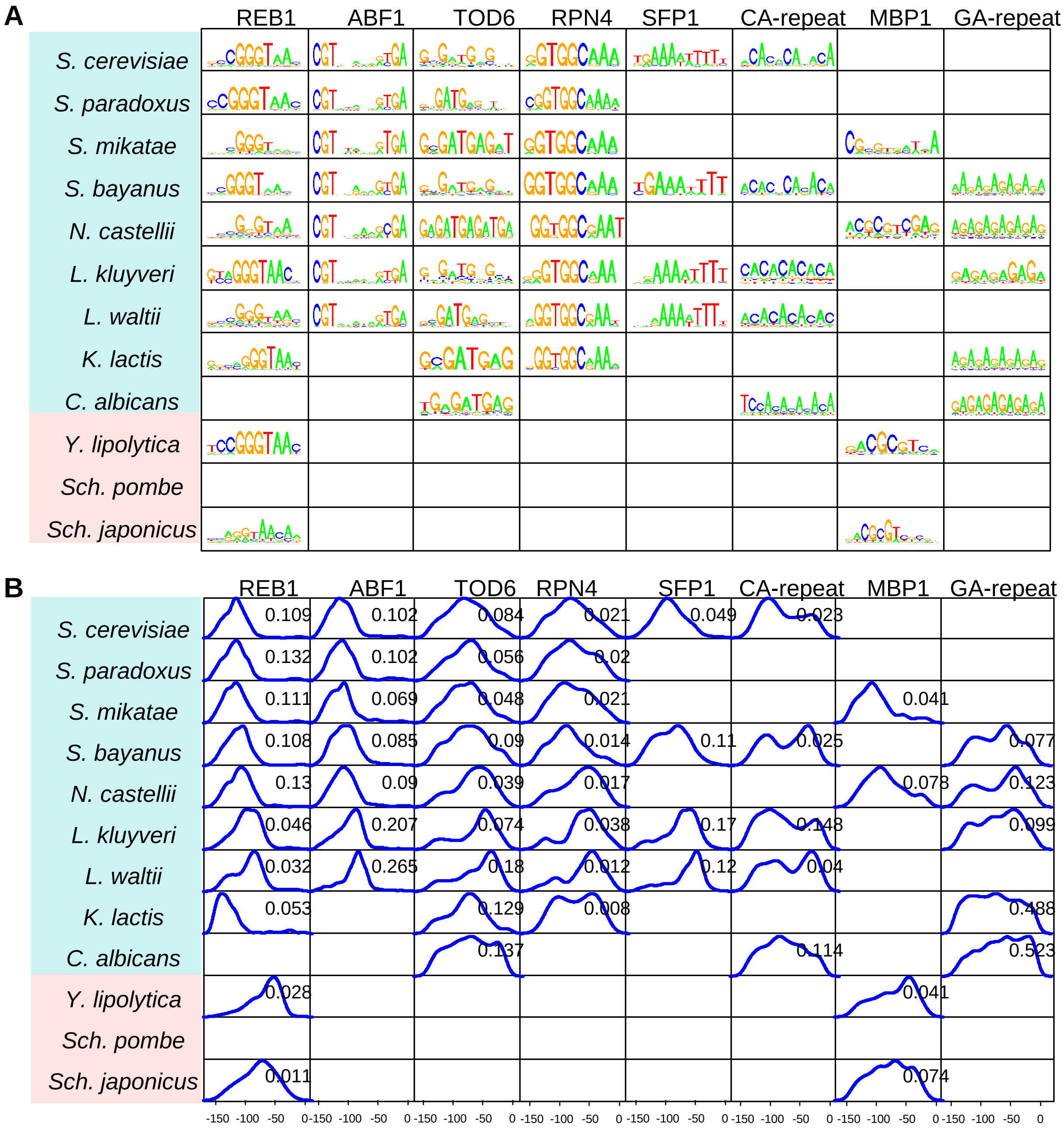
Supplemental Figure S11. Proportion of core promoters with a certain minimum number of As in the window from -9 to -3 bp. Each dot represent the proportion of core promoters with a certain minimum number of As in the window from −9 to −3 upstream of the TSS in a species. *: *p* < 0.01, t-test



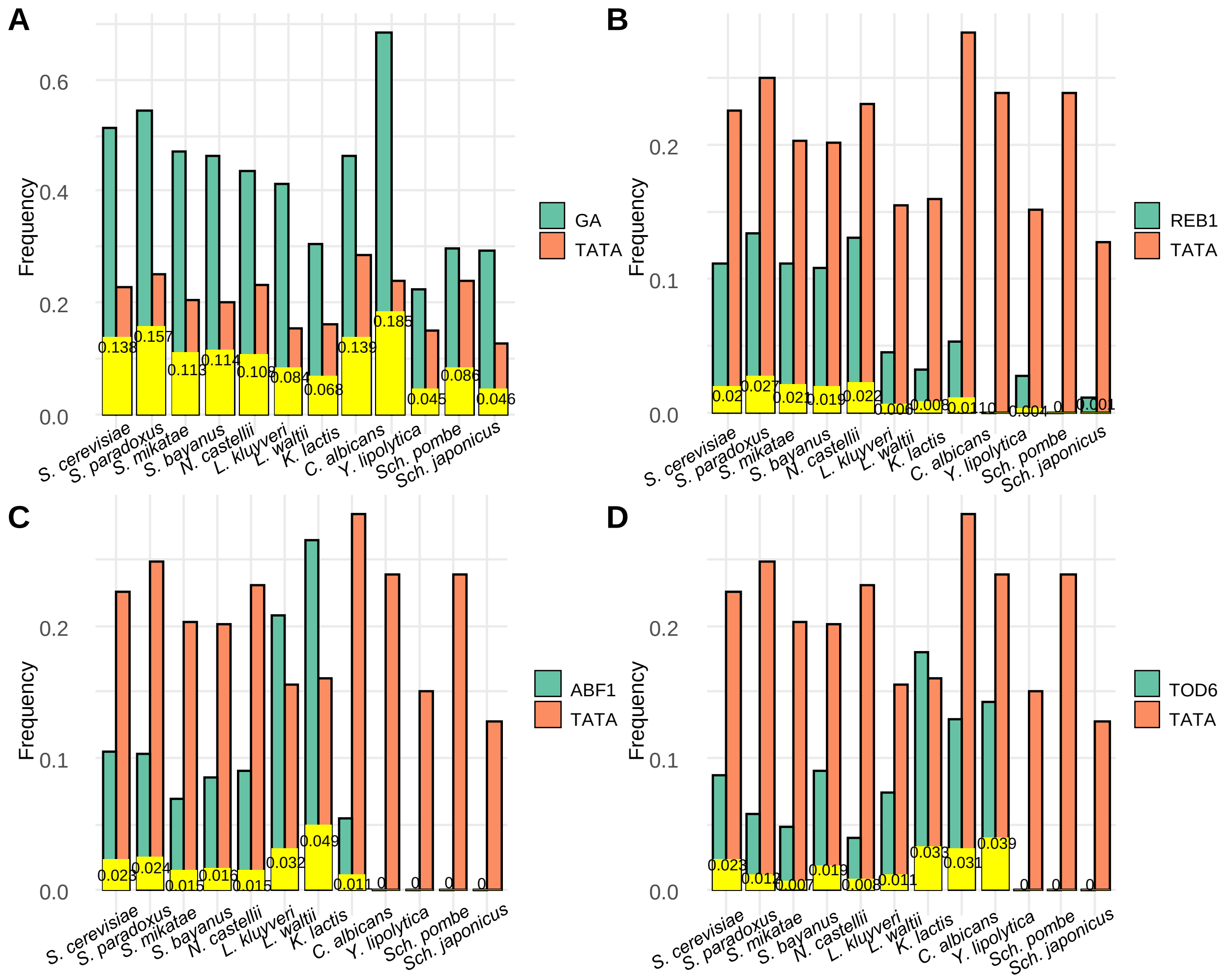
Supplemental Figure S12. Evolutionary divergence of core promoters in three closely related budding yeast species: *S. cerevisiae*, *S. paradoxus*, and *S. mikatae*. (A) Venn diagram shows the numbers of shared and species-specific core promoters among *S. cerevisiae*, *S. paradoxus*, and *S. makatea*. (B) Frequencies of PyPu, −8A, and TATA box in the “Turnover” group of core promoters. The “active core promoter” refers to core promoters with detectable transcriptional activities in the “Turnover” group. The “silent core promoter” refers to core promoters without detectable transcriptional activities in the “Turnover” group.



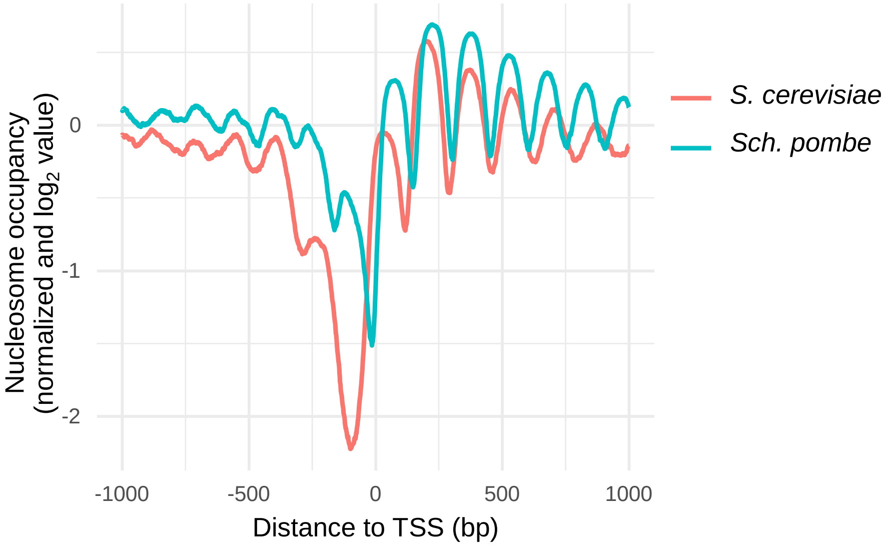
Supplemental Figure S13. Distribution of the GA element (GAE) in promoter regions. The GA element in promoter regions was identified by searching with a consensus sequence of GAAAAA with no mismatch allowed. The promoter regions include sequences from 150 bp upstream to 10 bp downstream of its dominant TSS. The left panel displays the distribution of the GA elements within this region based on unshuffled DNA sequences. The right panel shows the distribution of GA element after shuffling DNA sequences within this region.



Supplemental Figure S14. Putative core promoter motifs predicted by a *de novo* motif discovery approach. (A) Sequence logos of overrepresented sequence motifs identified in promoter regions [from −150 to +10 bp]. The sequence motifs that are significantly overrepresented in at least three yeast species are shown in this figure. (B) Density plots of distributions of these overrepresented motifs relative to dominant TSSs. The number in each plot indicates the proportion of promoters in a species containing the motif.



Supplemental Figure S15. The prevalence of four major motifs and their co-occurrences with TATA box in the 12 yeast species. The proportion of genes with the GA element (A), REB1 (B), ABF1 (C), and TOD6 (D) in each species. The proportions of TATA-containing genes in each species were shown as orange bars. The proportion of genes with co-occurrence of each motif and TATA box in their promoters are indicated by yellow bars with values displayed.



Supplemental Figure S16. Nucleosome occupancy pattern around TSSs of TATA box-containing core promoters in *S. cerevisiae* and *Sch. pombe*. Nucleosome occupancy was calculated with the total NCP (nucleosome center positioning) score of the redundant nucleosomes in the ±60 bp region of every genomic location, normalized by the average NCP score in the genome, and plotted with $\log_2$ value.

Supplemental Dataset S1. Genome annotation file of *S. cerevisiae* in GFF format

Supplemental Dataset S2. Genome annotation file of *S. paradoxus* in GFF format

Supplemental Dataset S3. Genome annotation file of *S. mikatea* in GFF format

Supplemental Dataset S4. Genome annotation file of *S. bayanus* in GFF format

Supplemental Dataset S5. Genome annotation file of *N. castellii* in GFF format

Supplemental Dataset S6. Genome annotation file of *L. kluyveri* in GFF format

Supplemental Dataset S7. Genome annotation file of *K. waltii* in GFF format

Supplemental Dataset S8. Genome annotation file of *K. lactis* in GFF format

Supplemental Dataset S9. Genome annotation file of *C. albicans* in GFF format

Supplemental Dataset S10. Genome annotation file of *Y. lipolytica* in GFF format

Supplemental Dataset S11. Genome annotation file of *Sch. pombe* in GFF format

Supplemental Dataset S12. Genome annotation file of *Sch. japonicus* in GFF format

Supplemental Dataset S13. A complete list of core promoters identified in this study

Supplemental Dataset S14. A complete list of orthologous groups identified by OrthoDB

Supplemental Dataset S15. Transcript abundance, PSS and weighted 5' UTR length of genes in the 12 yeast species

Supplemental Dataset S16. A list of orthologous core promoters in *S. cerevisiae*, *S. paradoxus*, and *S. mikatae*