Supplementary Information

Molecular dissection of the genetic mechanisms that underlie expression conservation in yeast ribosomal promoters

Danny Zeevi[1,2,3,#], Shai Lubliner[1,#], Maya Lotan-Pompan[1,2], Eran Hodis[1,2], Rita Vesterman[1], Adina Weinberger[1,2], Eran Segal[1,2,*]

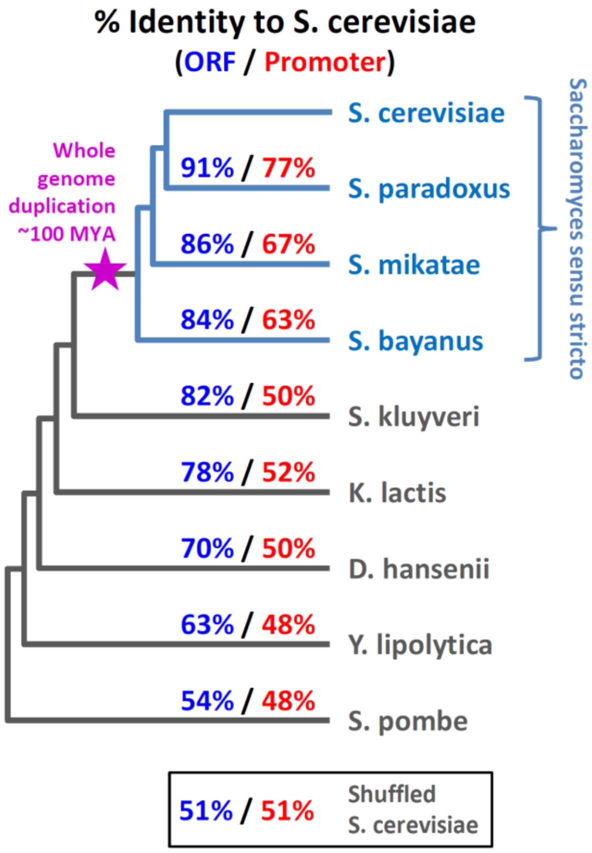
[1]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel.

[2]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel.

[3]Present address: Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA, USA.

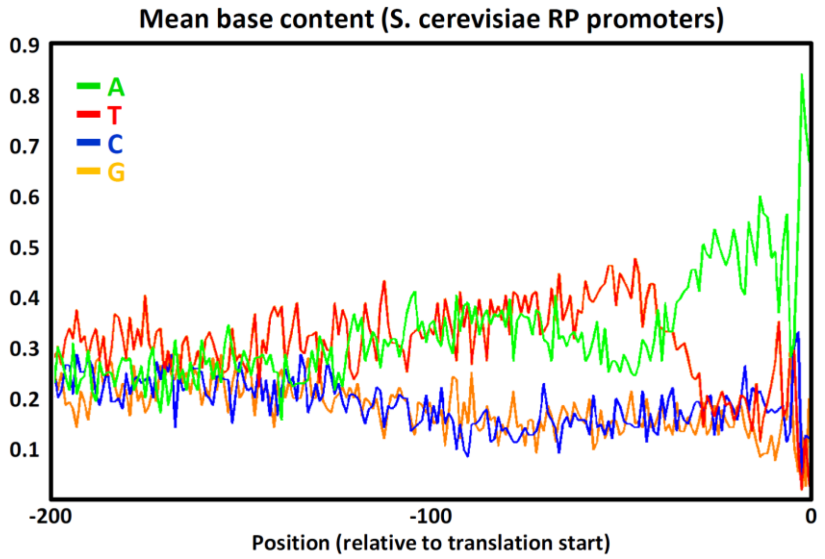
[#]These authors contributed equally to this work

[*]Corresponding author: eran.segal@weizmann.ac.il



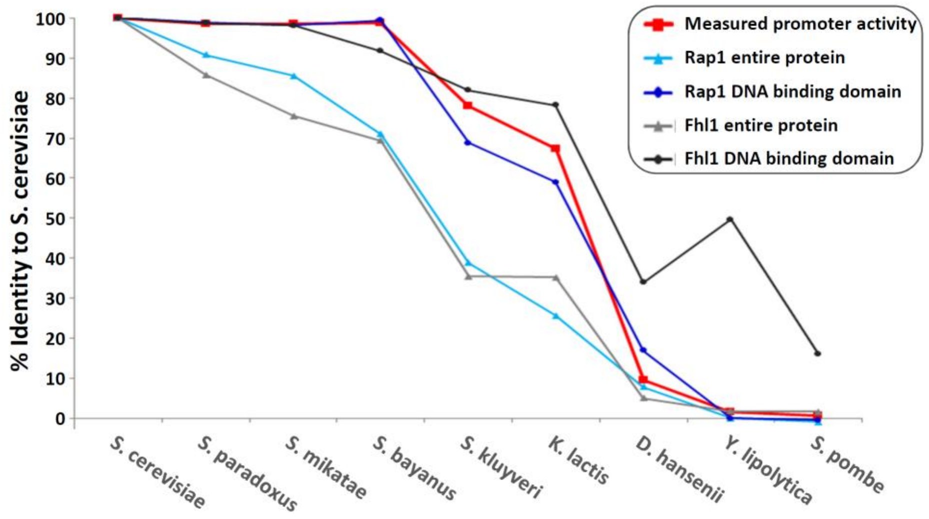
Supplementary Figure 1

Mean identity of RP ORFs and promoters of *S. cerevisiae* and of other yeast species. Phylogeny tree of all 9 species from which RP promoters were included in our native RP promoters library. For each species the mean percent identities of its RP ORFs and promoters (the 600 bps upstream of the translation start site) to those of *S. cerevisiae* are detailed in blue and in red, respectively.



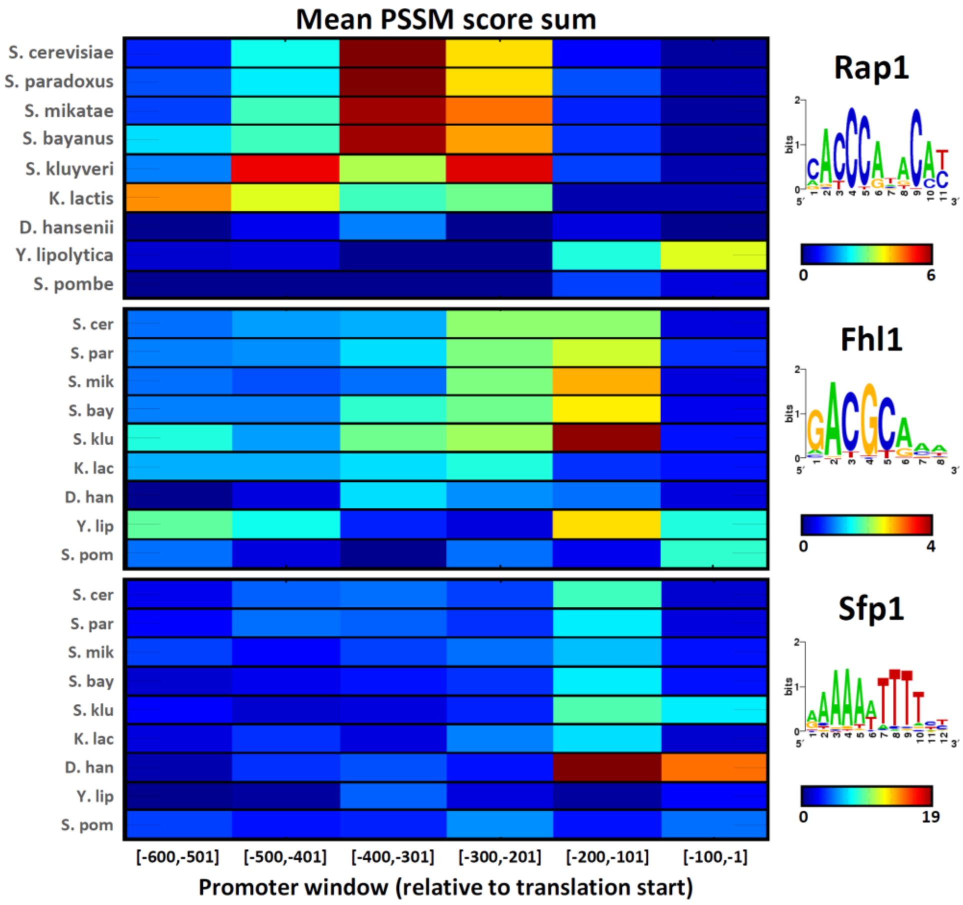
Supplementary Figure 2

***S. cerevisiae* RP promoters are A-rich upstream of the translation start site.** A-richness within the [-40,-1] region upstream of the translation start is mainly due to transcription start site (TSS) related sequence signals (Lubliner et al. 2013). Extreme A-richness within the last few bases fall within the 5'UTR and facilitates efficient translation initiation.



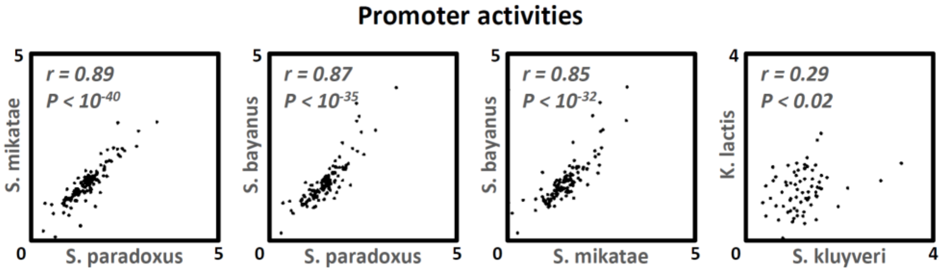
Supplementary Figure 3

Conservation of DNA binding motifs of RP regulators is higher than that of the entire protein sequences and is similar to the conservation of promoter activity. Median promoter activity shows extreme conservation between *sensu stricto* species (red, identical to red line in Fig. 2B, only here in values relative to *S. cerevisiae*). Celeste and grey mark the conservation of the entire Rap1 and Fhl1 protein sequences between different species and *S. cerevisiae*. Entire protein conservation drops rapidly even within the *sensu stricto* species. The blue and black circles mark the conservation of only the DNA binding domains of Rap1 and Fhl1 between the different species and *S. cerevisiae*. The DNA binding domain is more conserved than the entire protein, and in the case of Rap1 its conservation track (blue) resembles the conservation track of median promoter activity (red). Percent identity for the entire proteins was calculated using the Needleman-Wunsch algorithm, subtracting the background identity which was computed with a species shuffled sequence and normalized by *S. cerevisiae* (see Supplementary Note). The alignment of the DNA binding domains was calculated using the Smith-Waterman algorithm with a similar normalization scheme. *Y. lipolytica* has no Rap1 orthologous gene so its conservation levels were set to zero.



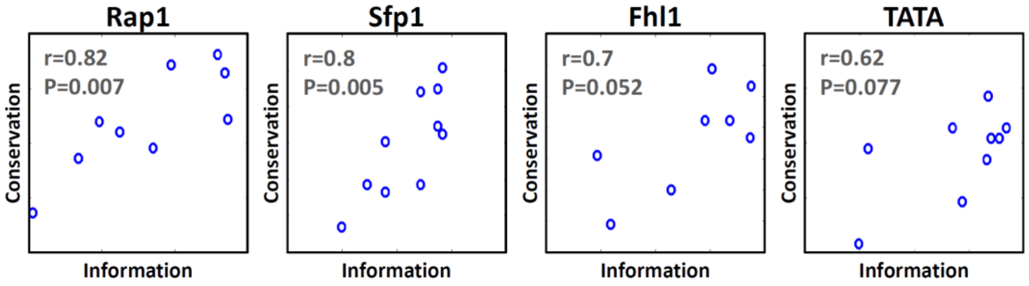
Supplementary Figure 4

Average TF binding site profiles of *S. cerevisiae* RP regulators along RP promoters of the 9 yeast species represented in our native RP promoters library. Hit scores (above genomic background) of the *S. cerevisiae* Rap1, Fhl1 and Sfp1 PSSMs (Pachkov et al. 2013) were computed along the RP promoters of our library that were taken from 9 yeast species, summed within each of several 100bp promoter windows, and the window sums were averaged for each specie over all of its promoters. Notably, the average TF binding site profiles are highly similar for the *sensu stricto* species, while are greatly diverged from that of *S. cereivisae* in more distant species.



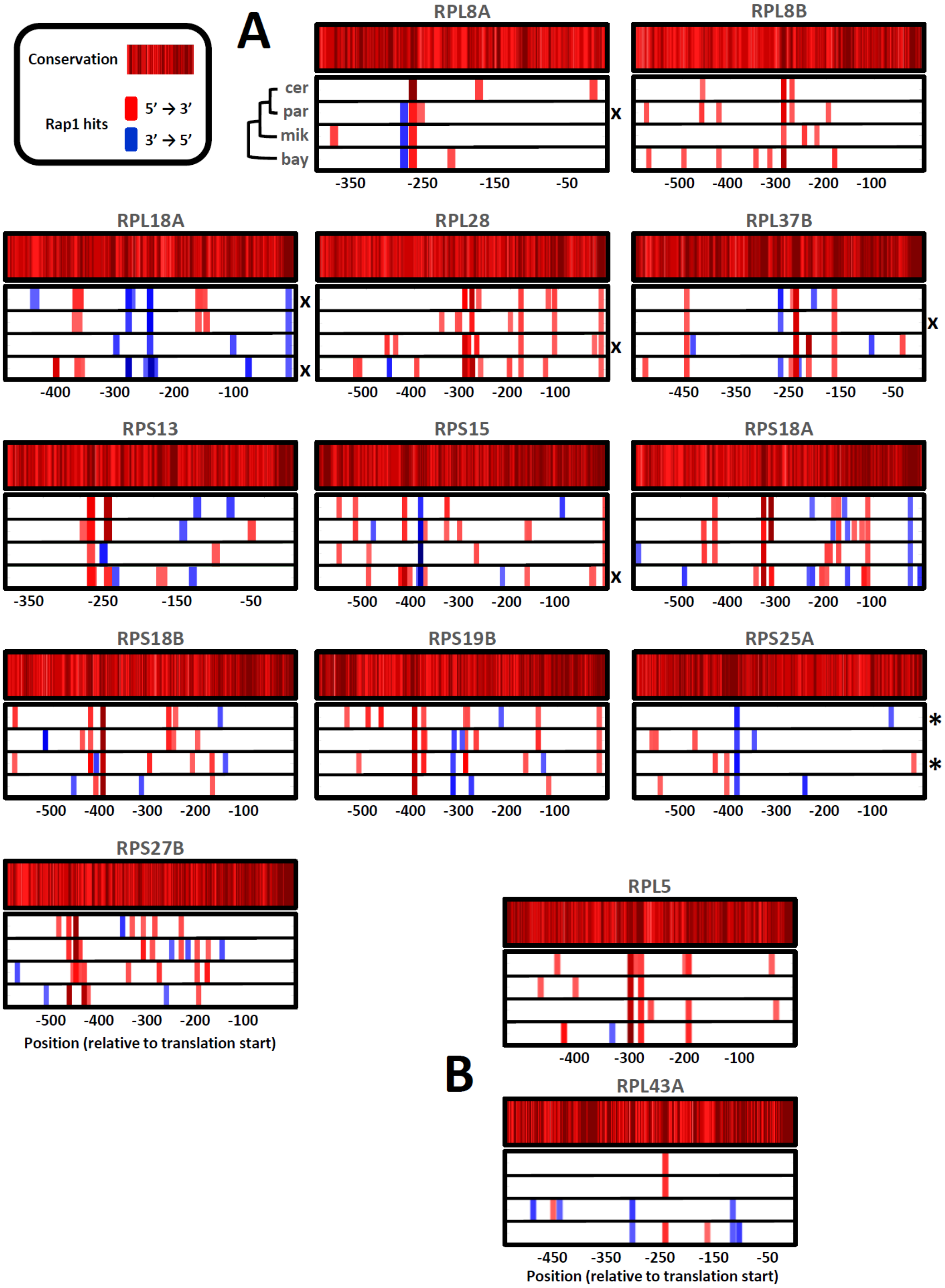
Supplementary Figure 5

Orthologous promoter activity comparisons. More dot plots (see Fig. 2C) comparing the measured promoter activities of orthologous RP promoters of pairs of species.



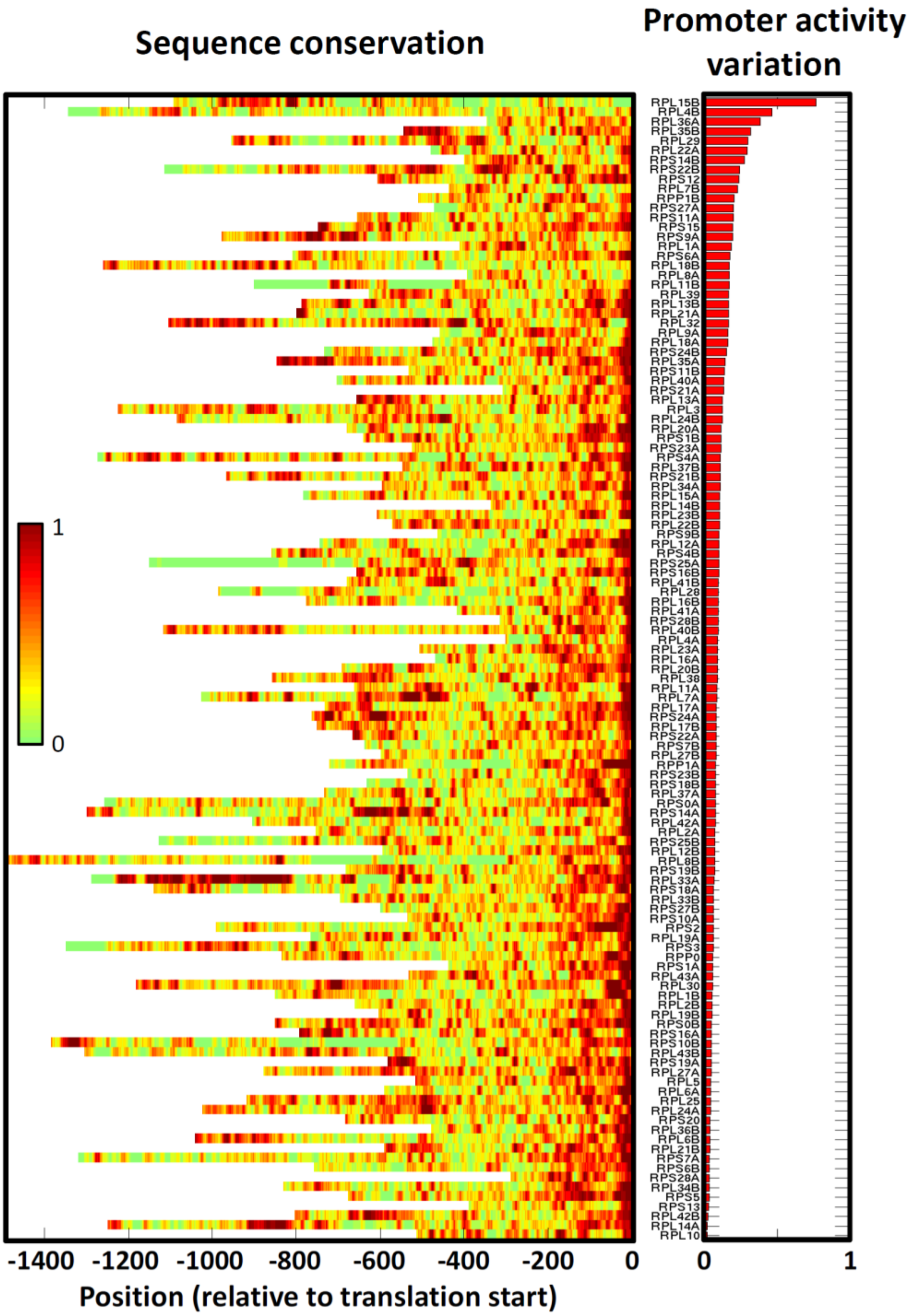
Supplementary Figure 6

PSSM motif position-wise information content and mean conservation within hits are correlated. For each of the 4 PSSM models shown in Fig. 3A, a dot plot shows the information content of its different positions versus the mean *sensu stricto* conservation per position within its high scoring hits. For all 4 PSSMs, position information content and conservation are highly correlated, although only for Rap1 and Sfp1 the correlation p-values are below 0.05.



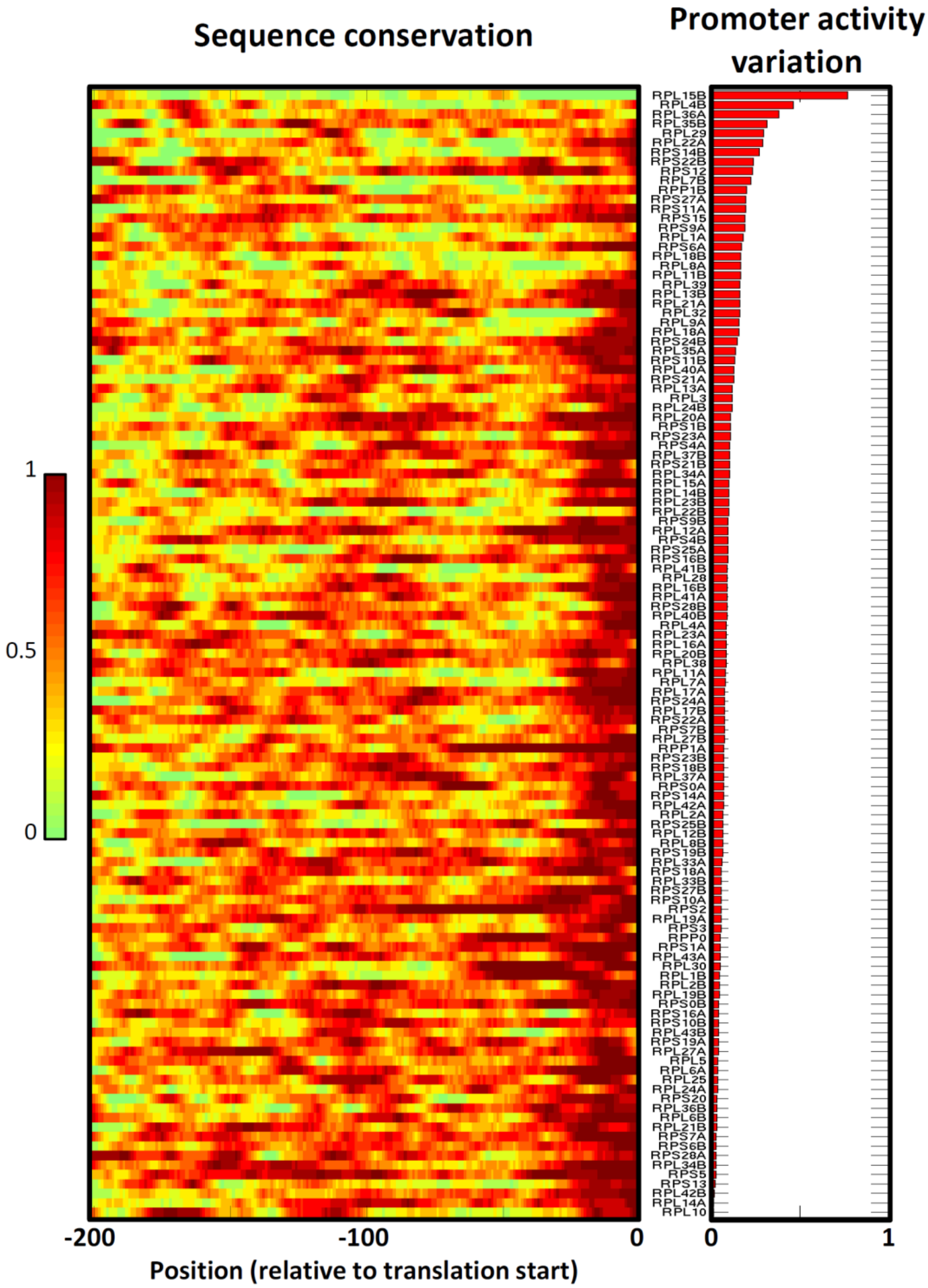
Supplementary Figure 7

Cases of Rap1 sites variation in 14 RP promoters. For each RP, the promoter conservation between the 4 orthologs is shown in the top track (in different shades of red, darker is higher). Below it, Rap1 PSSM (Pachkov et al. 2013) hits (above a low score threshold) are plotted on the multiple sequence alignment of the 4 orthologous promoters. The aligned tracks follow the order of the known phylogeny, schematically shown on the top left. Rap1 hits on the '+' strand are red, while hits on the '-' strand are blue (the darker the color the higher the score). Tracks of orthologs with promoter activity significantly different from the ortholog mean (±10% or more) were marked (on their right) with an 'x'. Tracks of orthologs with missing promoter activity measurements were marked with an asterisk. (A) 12 cases with tandem pairs of Rap1 sites, where one of the sites was lost in at least one of the orthologous promoters. (B) The *RPL5* and *RPL43A* cases of Rap1 sites variation.



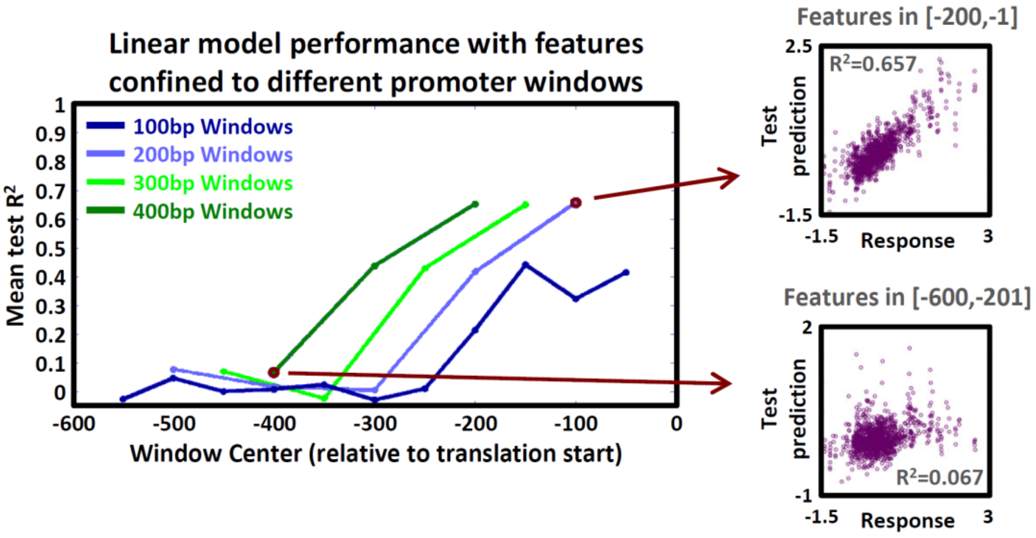
Supplementary Figure 8

Same as Fig. 4A, but with the entire conservation tracks shown and with the gene names listed.



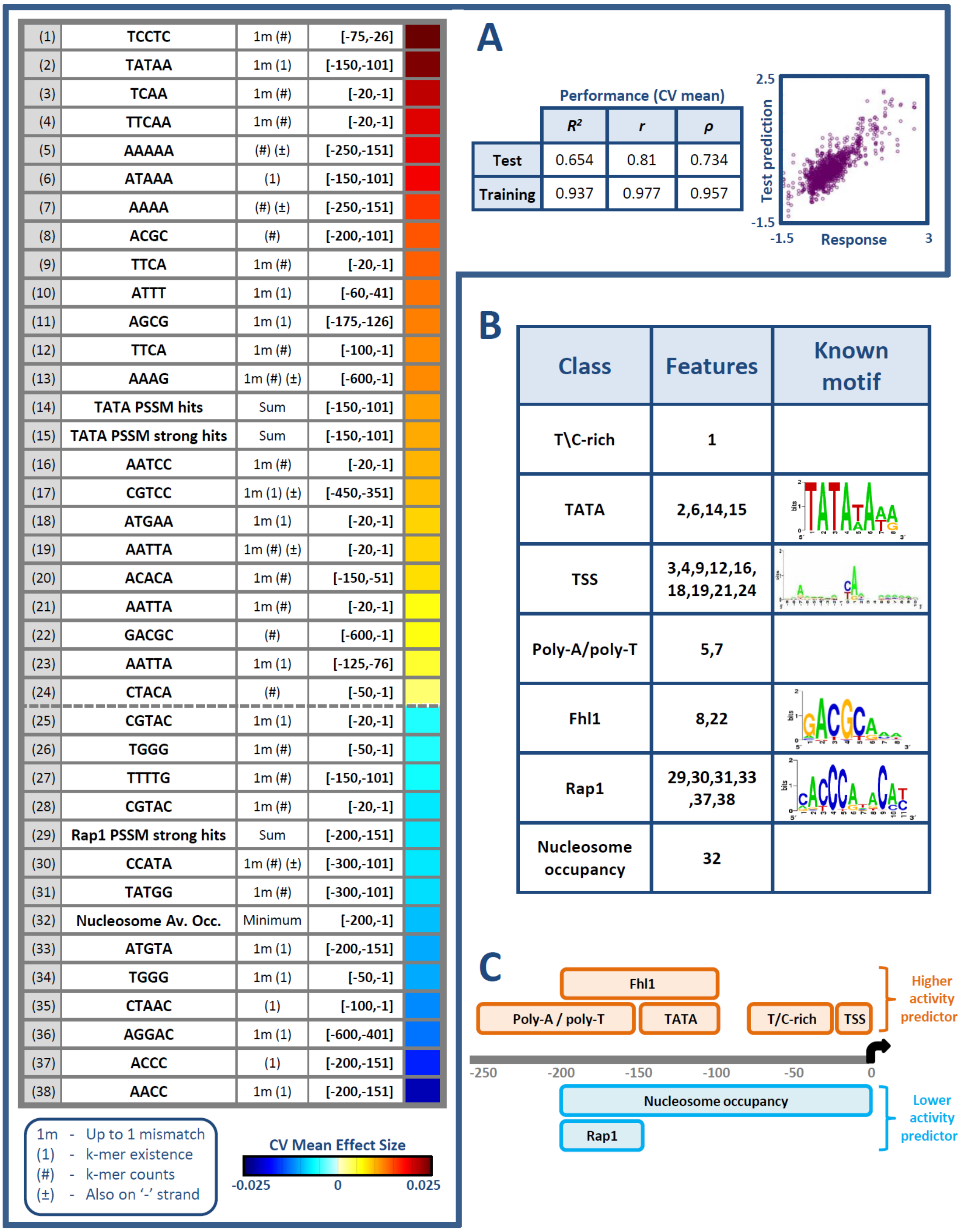
Supplementary Figure 9

Same as Fig. 4A, but larger and with the gene names listed.



Supplementary Figure 10

***Sensu stricto* RP promoter sequence features within 200bp of the ORF explain 65% of the variance in their promoter activity.** Following a 10-fold cross validation scheme (see Supplementary Note), we learned 10 linear models that predict promoter activity of native *sensu stricto* RP promoters from features of their promoter sequence. We repeated this learning scheme several times, each time using only features that fall within a certain window over the promoter to learn the models, and computed the mean R^2 statistic (quantifying the proportion of the promoter activity variance that is explained by the model) over held-out test data. For each such window, the mean test R^2 is plotted above the window center position. On the right, we further detail results for the [-200,-1] and the [-600,-201] windows (positions relative to the translation start site), showing dot-plots of the predicted promoter activities on test data (aggregated for all 10 test sets) versus their true value (the response). Notably, the mean test R^2 for all windows that contain the [-200,-1] region was ~0.65, showing that sequence features within this region (mostly consisting of the core promoter) can explain at least 65% of the variance in *sensu stricto* RP promoter activity. Conversely, windows that do not overlap with the [-200,-1] region can only explain a few percent of that variance.



(1)	TCCTC	1m (#)	[-75,-26]	🟥
(2)	TATAA	1m (1)	[-150,-101]	🟥
(3)	TCAA	1m (#)	[-20,-1]	🟥
(4)	TTCAA	1m (#)	[-20,-1]	🟥
(5)	AAAAA	(#) (±)	[-250,-151]	🟥
(6)	ATAAA	(1)	[-150,-101]	🟥
(7)	AAAA	(#) (±)	[-250,-151]	🟥
(8)	ACGC	(#)	[-200,-101]	🟧
(9)	TTCA	1m (#)	[-20,-1]	🟧
(10)	ATTT	1m (1)	[-60,-41]	🟧
(11)	AGCG	1m (1)	[-175,-126]	🟧
(12)	TTCA	1m (#)	[-100,-1]	🟧
(13)	AAAG	1m (#) (±)	[-600,-1]	🟧
(14)	TATA PSSM hits	Sum	[-150,-101]	🟧
(15)	TATA PSSM strong hits	Sum	[-150,-101]	🟧
(16)	AATCC	1m (#)	[-20,-1]	🟨
(17)	CGTCC	1m (1) (±)	[-450,-351]	🟨
(18)	ATGAA	1m (1)	[-20,-1]	🟨
(19)	AATTA	1m (#) (±)	[-20,-1]	🟨
(20)	ACACA	1m (#)	[-150,-51]	🟨
(21)	AATTA	1m (#)	[-20,-1]	🟨
(22)	GACGC	(#)	[-600,-1]	🟨
(23)	AATTA	1m (1)	[-125,-76]	🟨
(24)	CTACA	(#)	[-50,-1]	🟨
(25)	CGTAC	1m (1)	[-20,-1]	🟦
(26)	TGGG	1m (#)	[-50,-1]	🟦
(27)	TTTTG	1m (#)	[-150,-101]	🟦
(28)	CGTAC	1m (#)	[-20,-1]	🟦
(29)	Rap1 PSSM strong hits	Sum	[-200,-151]	🟦
(30)	CCATA	1m (#) (±)	[-300,-101]	🟦
(31)	TATGG	1m (#)	[-300,-101]	🟦
(32)	Nucleosome Av. Occ.	Minimum	[-200,-1]	🟦
(33)	ATGTA	1m (1)	[-200,-151]	🟦
(34)	TGGG	1m (1)	[-50,-1]	🟦
(35)	CTAAC	(1)	[-100,-1]	🟦
(36)	AGGAC	1m (1)	[-600,-401]	🟦
(37)	ACCC	(1)	[-200,-151]	🟦
(38)	AACC	1m (1)	[-200,-151]	🟦

1m - Up to 1 mismatch
(1) - k-mer existence
(#) - k-mer counts
(±) - Also on '-' strand

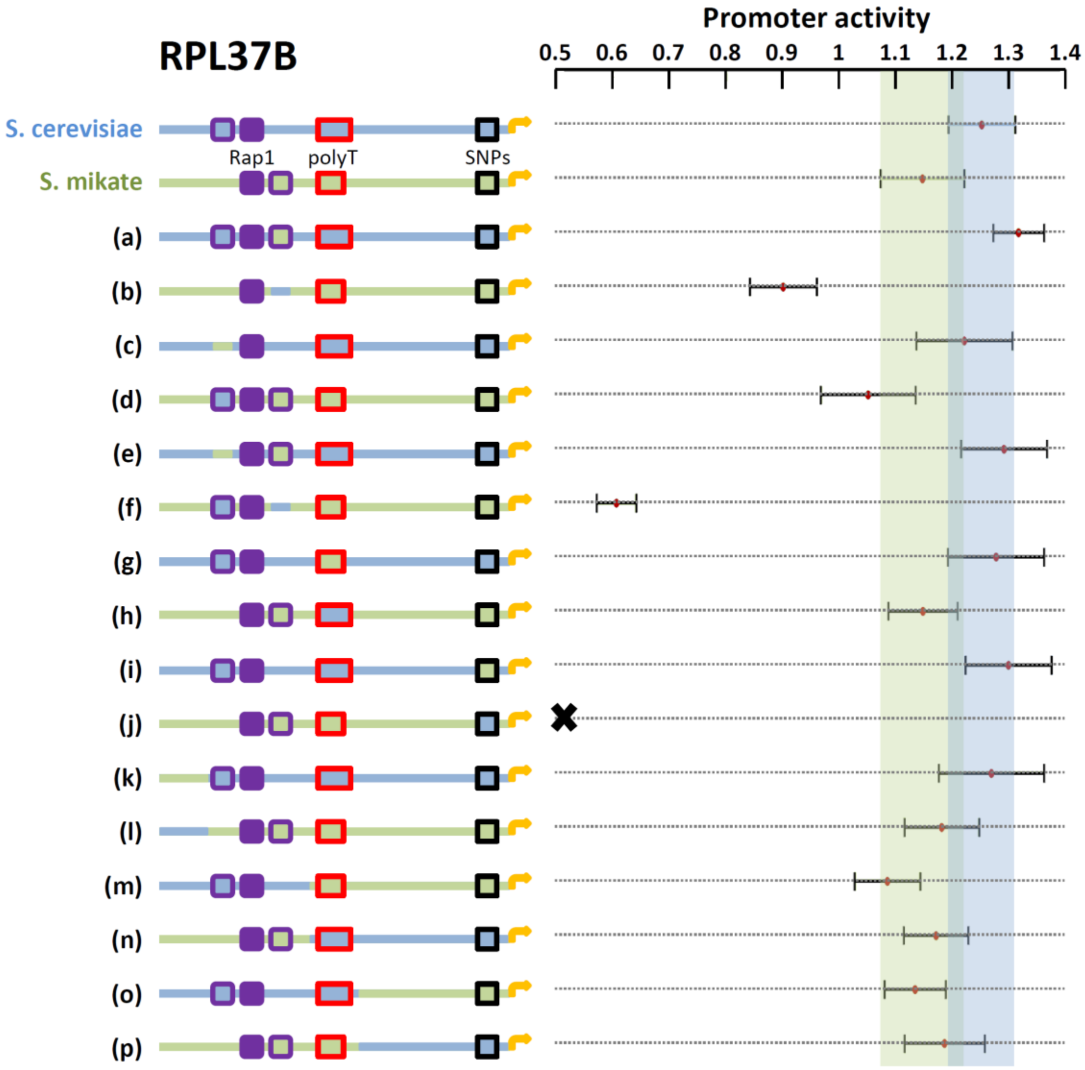
Performance (CV mean)

	R^2	*r*	*ρ*
Test	0.654	0.81	0.734
Training	0.937	0.977	0.957

Class	Features	Known motif
T\C-rich	1	
TATA	2,6,14,15	TATATAAA
TSS	3,4,9,12,16, 18,19,21,24	
Poly-A/poly-T	5,7	
Fhl1	8,22	GACGCA
Rap1	29,30,31,33 ,37,38	CACCCAACAT
Nucleosome occupancy	32	

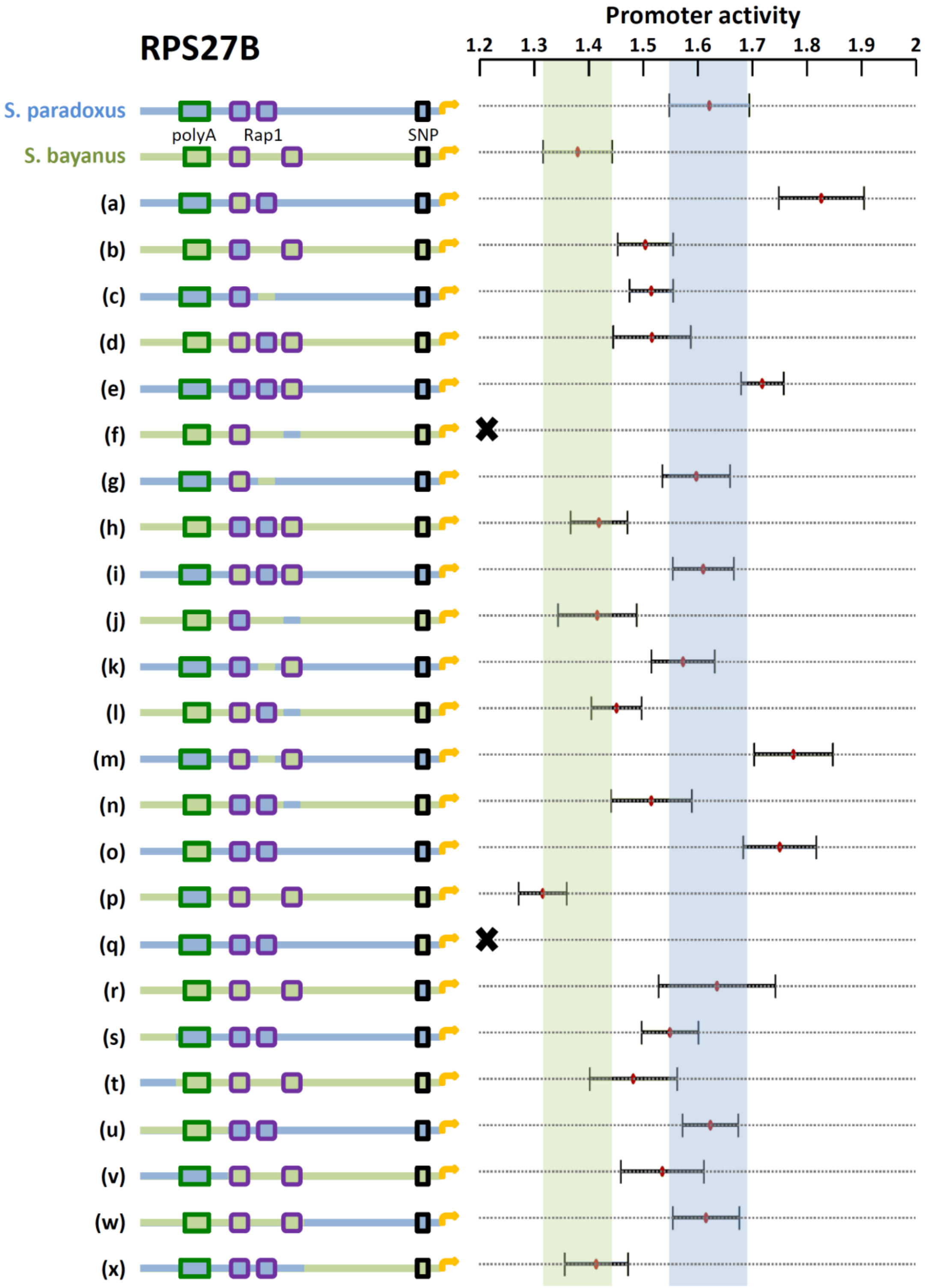
Supplementary Figure 11

Performance of linear models that predict *sensu stricto* RP promoter activity from promoter sequence features within 600bp of the ORF. 10 linear models were learned (for 10 different partitions of the data to training and held-out test sets, see Supplementary Note) from sequence features within the [-600,-1] region (relative to the translation start site) of orthologous *sensu stricto* RP promoters. (A) The table on the top right shows the mean model performance measures (the R^2 statistic; Pearson's correlation, *r*; Spearman's correlation, ρ) on both training and held-out test data. The dot plot shows the predicted promoter activities on test data (aggregated for all 10 test sets) versus their true value (the response). The table on the right details 38 robust features (that were included in at least 8 out of the 10 models). These features include k-mer existence and counts, features of hits of PSSMs of known RP regulators, and a feature of the predicted intrinsic nucleosome occupancy. Each feature was computed over a certain promoter window (positions relative to the translation start site). The mean (over the 10 models) effect size of each feature is color coded in the right column of the table. (B) 25 out of 38 robust features can be assigned to several groups of features. The TATA motif is that of (Basehoar et al. 2004). The TSS motif is that of (Zhang and Dietrich 2005). The Fhl1 and Rap1 motifs are those of (Pachkov et al. 2013). (C) A schematic representation of the groups of features in (B), showing their location on the promoter and their effect on expression.



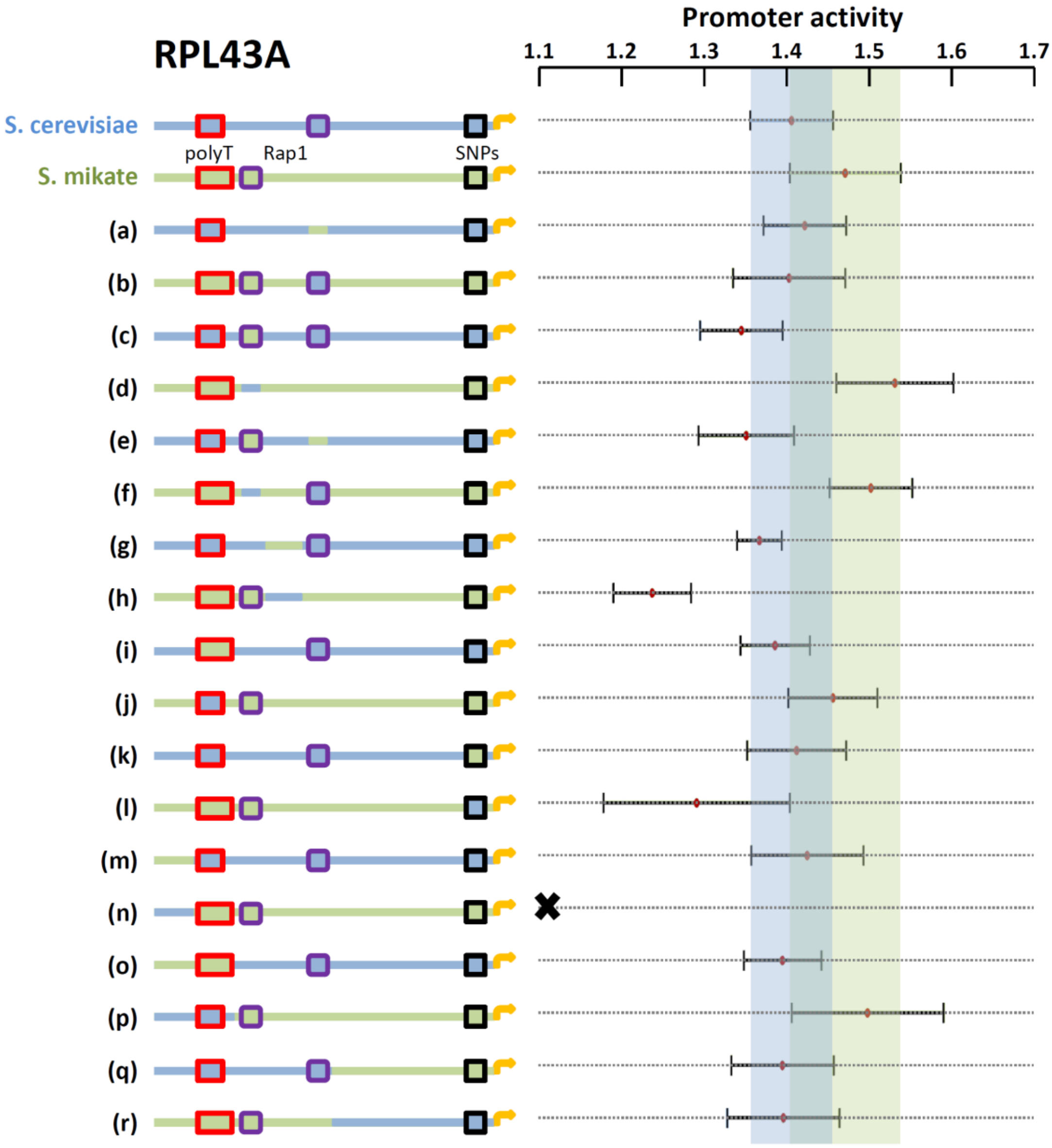
Supplementary Figure 12

Orthologous mutations of two native *RPL37B* promoters. On the left, a schematic representation of the *S. cerevisiae* (light blue) and the *S. mikatae* (light green) *RPL37B* promoters (highlighting functional elements within them, see main text), along with all of their orthologously mutated promoters (a)-(p). Promoter elements and regions are colored according to their native origin (light blue for *S. cerevisiae*, light green for *S. mikatae*). On the right, we show the measured promoter activities (red dots show the mean over several replicates), at the center of a 95% confidence interval (±2 standard errors). The light blue (light green) shaded range marks the 95% confidence interval for the native *S. cerevisiae* (*S. mikatae*) promoter. For technical reasons, we could not measure the promoter (j).



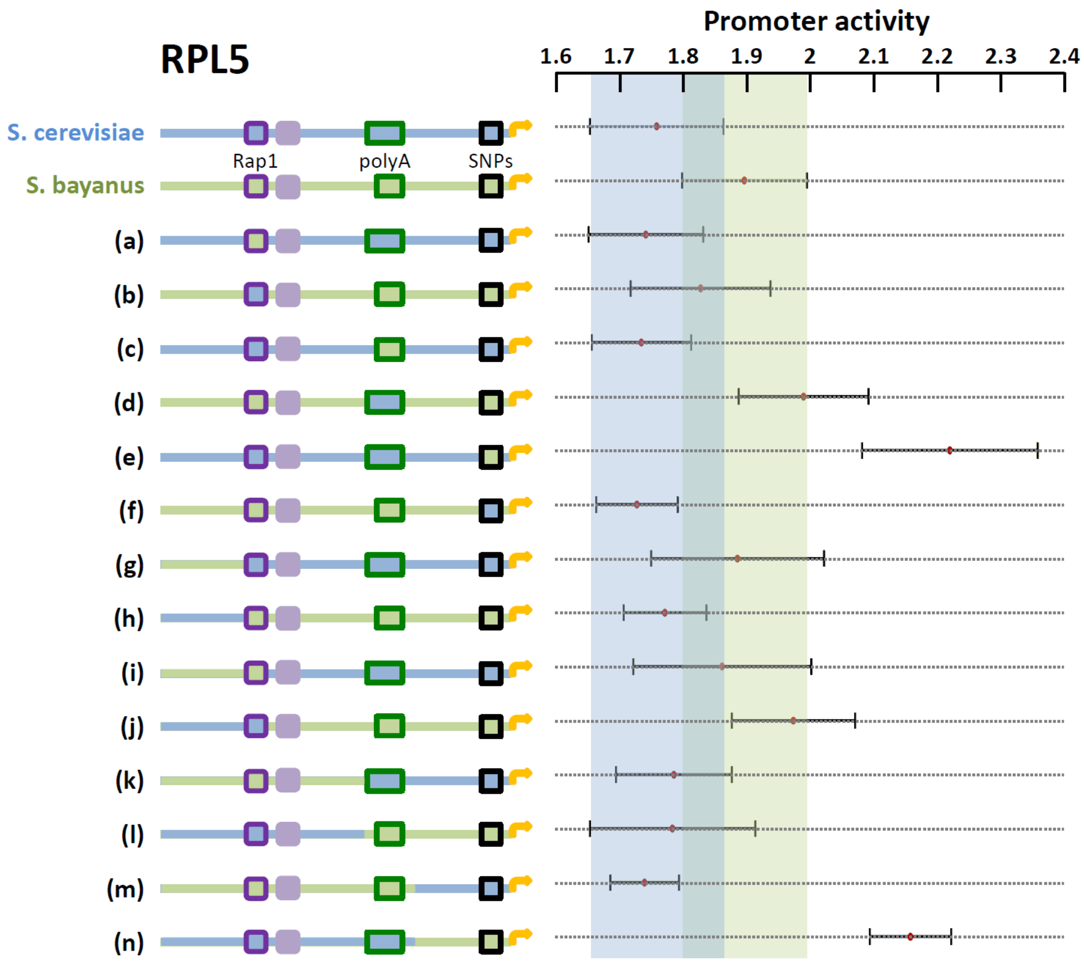
Supplementary Figure 13

Orthologous mutations of two native *RPS27B* promoters. On the left, a schematic representation of the *S. paradoxus* (light blue) and the *S. bayanus* (light green) *RPS27B* promoters (highlighting functional elements within them, see main text), along with all of their orthologously mutated promoters (a)-(x). Promoter elements and regions are colored according to their native origin (light blue for *S. paradoxus*, light green for *S. bayanus*). On the right, we show the measured promoter activities (red dots show the mean over several replicates), at the center of a 95% confidence interval (±2 standard errors). The light blue (light green) shaded range marks the 95% confidence interval for the native *S. paradoxus* (*S. bayanus*) promoter. For technical reasons, we could not measure the promoters (f) and (q).



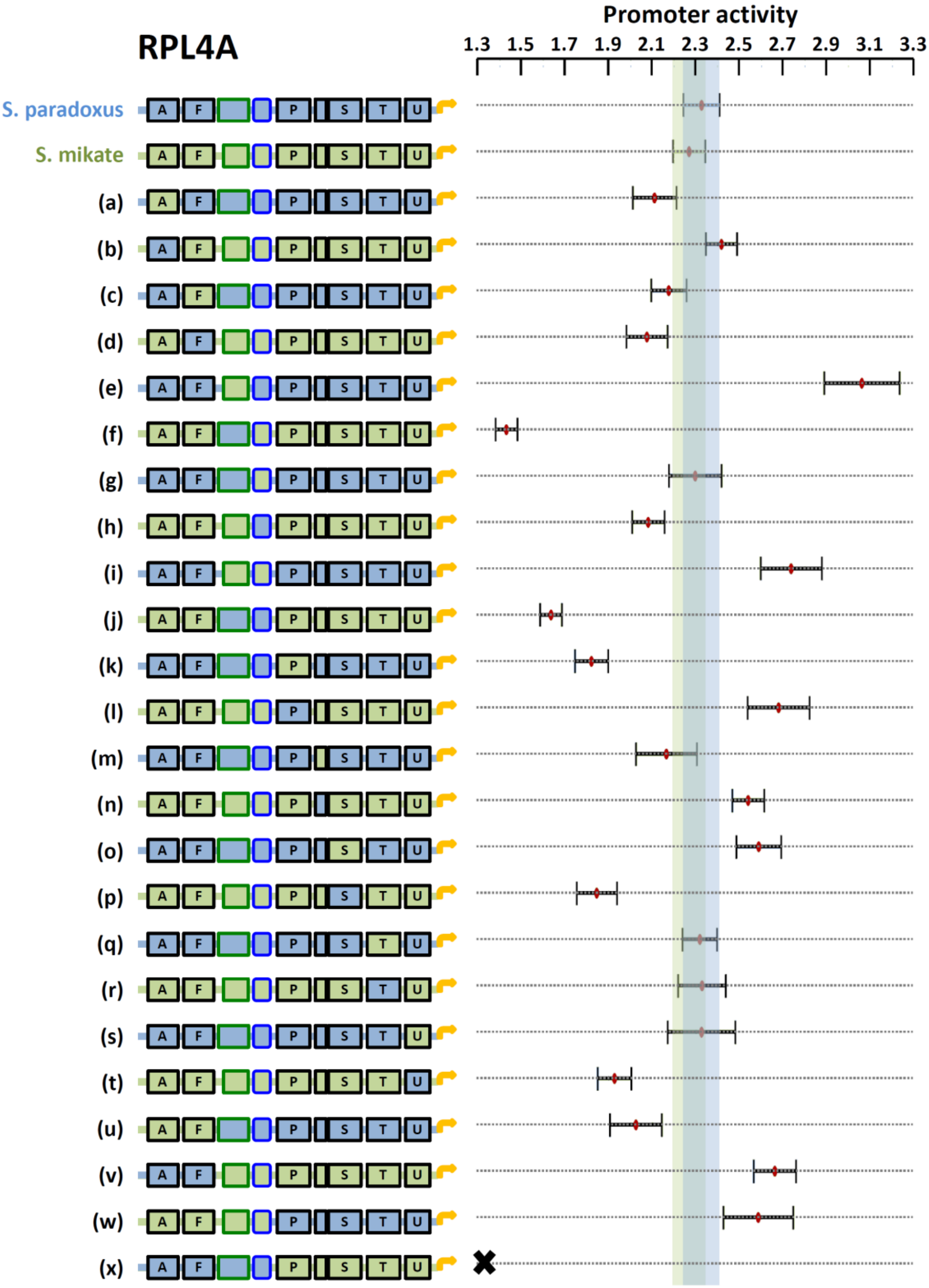
Supplementary Figure 14

Orthologous mutations of two native *RPL43A* promoters. On the left, a schematic representation of the *S. cerevisiae* (light blue) and the *S. mikatae* (light green) *RPL43A* promoters (highlighting functional elements within them, see main text), along with all of their orthologously mutated promoters (a)-(r). Promoter elements and regions are colored according to their native origin (light blue for *S. cerevisiae*, light green for *S. mikatae*). On the right, we show the measured promoter activities (red dots show the mean over several replicates), at the center of a 95% confidence interval (±2 standard errors). The light blue (light green) shaded range marks the 95% confidence interval for the native *S. cerevisiae* (*S. mikatae*) promoter. For technical reasons, we could not measure the promoter (n).



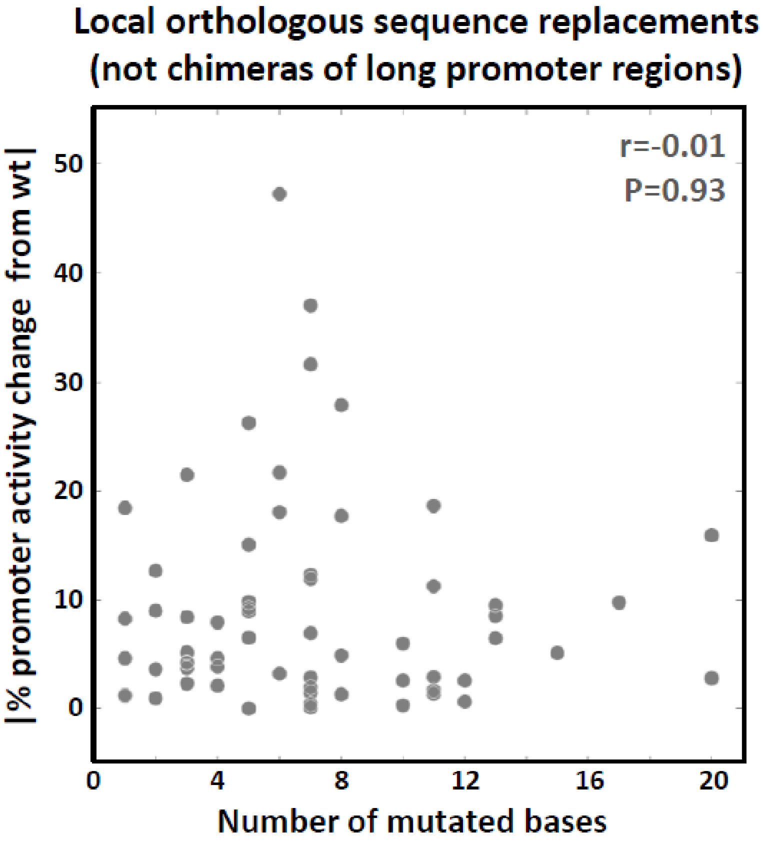
Supplementary Figure 15

Orthologous mutations of two native *RPL5* promoters. On the left, a schematic representation of the *S. cerevisiae* (light blue) and the *S. bayanus* (light green) *RPL5* promoters (highlighting functional elements within them, see main text), along with all of their orthologously mutated promoters (a)-(x). Promoter elements and regions are colored according to their native origin (light blue for *S. cerevisiae*, light green for *S. bayanus*). On the right, we show the measured promoter activities (red dots show the mean over several replicates), at the center of a 95% confidence interval (±2 standard errors). The light blue (light green) shaded range marks the 95% confidence interval for the native *S. cerevisiae* (*S. bayanus*) promoter.



Supplementary Figure 16

Orthologous mutations of two native *RPL4A* promoters. On the left, a schematic representation of the *S. paradoxus* (light blue) and the *S. mikatae* (light green) *RPL4A* promoters (highlighting different blocks within them, see main text), along with all of their orthologously mutated promoters (a)-(x). Promoter blocks and regions are colored according to their native origin (light blue for *S. paradoxus*, light green for *S. mikatae*). On the right, we show the measured promoter activities (red dots show the mean over several replicates), at the center of a 95% confidence interval (±2 standard errors). The light blue (light green) shaded range marks the 95% confidence interval for the native *S. paradoxus* (*S. mikatae*) promoter. For technical reasons, we could not measure the promoter (x).



Supplementary Figure 17

Effects of orthologous mutations on promoter activity are not due the degree of introduced variation. For each orthologous mutation, we compare the number of bases that vary from the wild type sequence (x-axis) to the absolute percent promoter activity change from the wild type promoter activity (y-axis). As can be seen, no correlation was observed between the two measures. We did not include here large chimeras, where whole promoter regions were replaced yet little effect on expression was observed.

Supplementary Note

Plasmid sequence context

Each of the measured promoters was inserted into a plasmid, upstream of a YFP reporter gene. For the complete plasmid sequence see (Zeevi et al. 2011)). Below we detail the 1000 bps long sequences that flanked each inserted promoter, as these were used in the process of promoter window features computation (for instance, nucleosome occupancy predictions were computed with flanks to avoid edge related errors).

The 1000 bps upstream of each promoter:

```
CTGCACAGAACAAAAACCTGCAGGAAACGAAGATAAATCATGTCGAAAGCTACATATAAGGAACGTGCTGCTACTCATCCTAGTCC
TGTTGCTGCCAAGCTATTTAATATCATGCACGAAAAGCAAACAAACTTGTGTGCTTCATTGGATGTTCGTACCACCAAGGAATTAC
TGGAGTTAGTTGAAGCATTAGGTCCCAAAATTTGTTTACTAAAAACACATGTGGATATCTTGACTGATTTTTCCATGGAGGGCACA
GTTAAGCCGCTAAAGGCATTATCCGCCAAGTACAATTTTTTACTCTTCGAAGACAGAAAATTTGCTGACATTGGTAATACAGTCAA
ATTGCAGTACTCTGCGGGTGTATACAGAATAGCAGAATGGGCAGACATTACGAATGCACACGGTGTGGTGGGCCCAGGTATTGTTA
GCGGTTTGAAGCAGGCGGCAGAAGAAGTAACAAAGGAACCTAGAGGCCTTTTGATGTTAGCAGAATTGTCATGCAAGGGCTCCCTA
TCTACTGGAGAATATACTAAGGGTACTGTTGACATTGCGAAGAGCGACAAAGATTTTGTTATCGGCTTTATTGCTCAAAGAGACAT
GGGTGGAAGAGATGAAGGTTACGATTGGTTGATTATGACACCCGGTGTGGGTTTAGATGACAAGGGAGACGCATTGGGTCAACAGT
ATAGAACCGTGGATGATGTGGTCTCTACAGGATCTGACATTATTATTGTTGGAAGAGGACTATTTGCAAAGGGAAGGGATGCTAAG
GTAGAGGGTGAACGTTACAGAAAAGCAGGCTGGGAAGCATATTTGAGAAGATGCGGCCAGCAAAACTAAAAAACTGTATTATAAGT
AAATGCATGTATACTAAACTCACAAATTAGAGCTTCAATTTAATTATATCAGTTATTACCCTGCGGTGTGAAATACCGCACAGATG
CGTAAGGAGAAAATACCGCATCAGGGTCGGGGTGTGTTGTTGGTGGGTTGGGTG
```

The 1000 bps downstream of each promoter:

```
ATGTCTAAAGGTGAAGAATTATTCACTGGTGTTGTCCCAATTTTGGTTGAATTAGATGGTGATGTTAATGGTCACAAATTTTCTGT
CTCCGGTGAAGGTGAAGGTGATGCTACTTACGGTAAATTGACCTTAAAATTGATTTGTACTACTGGTAAATTGCCAGTTCCATGGC
CAACCTTAGTCACTACTTTAGGTTATGGTTTGCAATGTTTTGCTAGATACCCAGATCATATGAAACAACATGACTTTTTCAAGTCT
GCCATGCCAGAAGGTTATGTTCAAGAAAGAACTATTTTTTTCAAAGATGACGGTAACTACAAGACCAGAGCTGAAGTCAAGTTTGA
AGGTGATACCTTAGTTAATAGAATCGAATTAAAAGGTATTGATTTTAAAGAAGATGGTAACATTTTAGGTCACAAATTGGAATACA
ACTATAACTCTCACAATGTTTACATCACTGCTGACAAACAAAAGAATGGTATCAAAGCTAACTTCAAAATTAGACACAACATTGAA
GATGGTGGTGTTCAATTAGCTGACCATTATCAACAAAATACTCCAATTGGTGATGGTCCAGTCTTGTTACCAGACAACCATTACTT
ATCCTATCAATCTGCCTTATCCAAAGATCCAAACGAAAAGAGAGACCACATGGTCTTGTTAGAATTTGTTACTGCTGCTGGTATTA
CCCATGGTATGGATGAATTGTACAAATAAGGCGCGCCACTTCTAAATAAGCGAATTTCTTATGATTTATGATTTTTATTATTAAAT
AAGTTATAAAAAAAATAAGTGTATACAAATTTTAAAGTGACTCTTAGGTTTTAAAACGAAAATTCTTATTCTTGAGTAACTCTTTC
CTGTAGGTCAGGTTGCTTTCTCAGGTATAGTATGAGGTCGCTCTTATTGACCACACCTCTACCGGCAGATCCGCTAGGGATAACAG
GGTAATATAGATCTGTTTAGCTTGCCTTGTCCCCGCCGGGTCACCCGGCCAGCG
```

Computing percent identities of orthologous protein sequences

For each species, we calculated the percent identity of a specific protein (e.g. Fhl1) to its *S. cerevisiae* ortholog by pairwise alignment of their amino acid sequences (Needleman–Wunsch algorithm). To avoid a bias in identity for species with different protein lengths, we defined a background identity level for each species by permuting its own amino acid sequence 100 times, aligning each permuted sequence to the one of *S. cerevisiae* and taking the average identity. We then subtracted this background level from the calculated identity of the actual protein and the *S. cerevisiae* protein. We computed this identity measure for each of the 9 species, including *S. cerevisiae* (vs. itself), and then normalized by dividing each by the *S. cerevisiae* measure.

TF binding sites annotation for conservation analysis

We annotated high scoring hits (with scores above 50% of the maximal score) of in-vitro derived PSSM models of known RP regulating TFs within the *S. cerevisiae* RP promoters that were included in our native RP promoters library. For Rap1 we used the PSSM of (Badis et al. 2008) over the [-600,-1] promoter region (upstream of the translation start). For Fhl1, Sfp1 and TATA (TATA binding protein, a.k.a. Spt15) we used truncated versions (flanking positions with negligible information content were removed) of the PSSMs of (Zhu et al. 2009), over the [-300,-1], [-300,-1] and [-200,-1] promoter regions, respectively.

TF binding sites variation measure

For each TF, we computed a measure of the *sensu stricto* sequence variation of its PSSM hits for each gene that had high scoring promoter hits. For each hit, we defined its mean sequence variation to be one minus its mean conservation (over hit positions). For each gene, its binding site variation measure was taken to be the minimal mean sequence variation (over any of its promoter hits).

Promoter conservation tracks

For each RP gene that had its *sensu stricto* promoters included in our native RP promoters library, we computed a 4-way alignment of these promoters, and from it the promoter's *sensu stricto* conservation track: each position in the alignment got a value of 1 if and only if it was identical in all 4 promoters, and the values were smoothed using a 5bp sliding window.

Learning linear models that predict promoter activity from promoter sequence features

For the purpose of model learning, we partitioned the *sensu stricto* RP promoters to training and held-out test sets as follows: for each of the 120 RP genes, we randomly chose one orthologous promoter for the test set, while the other three were added to the training set. A linear model was learned using the training set alone and its performance was assessed on the held-out test set. We repeated this data partitioning and model learning 10 times, adhering to a 10-fold cross validation scheme. This allowed us to compute mean performance measures of the 10 different models, and also highlight sequence features that were included in most models.

We used the *glmnet* software (http://www.stanford.edu/~hastie/glmnet_matlab) to run the elastic-net regression algorithm (Zou and Hastie 2005), with a mixing ratio of 1:1 between the L_1- and the L_2-regularization terms (*glmnet* parameter alpha=0.5). *glmnet* uses least angle regression (LARS) (Efron et al. 2004) to generate a grid of solutions on the regularization path of the model coefficients vector, between the 0-model and the non-regularized model. Each solution on the regularization path corresponds to a specific value of the regularization coefficient λ, with λ monotonically decreasing between the 0-model and the non-regularized model (where λ=0). To select the value of λ, we used a 10-fold cross validation scheme over the training data. For this purpose, the training set was randomly partitioned (10 different times) to an internal training set and a validation set. For each internal training set, we learned a grid of up to 1000 solutions (*glmnet* parameter nlambda=1000) on the regularization path, and took the value of λ of the solution that performed best on the held out validation set (in terms of the R^2 statistic). The final value of λ was taken to be the mean of the 10 selected λ values.

On the orthologous promoter pairs included in our orthologous mutations library

In the *RPL37B*, *RPS27B* and *RPL43A* orthologous promoter pairs, we investigated divergence of Rap1 binding sites (annotated using the Rap1 PSSM model from the SwissRegulon database (Pachkov et al. 2013)), including lost and gained sites, divergence of poly(dA)/poly(dT) tracts (Segal and Widom 2009; Raveh-Sadka et al. 2012) adjacent to Rap1 sites and single nucleotide polymorphisms (SNPs) located in the downstream part of the core promoter region.

In the *RPL5* promoter pair we investigated divergence of a single Rap1 binding site, a poly(dA) tract immediately upstream of a conserved TATA-like element, and SNPs in the downstream part of the core promoter.

For the last case we picked *RPL4A* promoters, that are highly expressed, relatively short (<300 bps) and lack Rap1 binding sites. Relying on our recently published study of yeast core promoters (Lubliner et al. 2013), we annotated orthologous blocks within the *RPL4A* core promoters (detailed in a 5'→3' order): a poly(dA) tract, a TATA box, the sequence (denoted by 'P') to which the pre initiation complex (PIC) is recruited (and initial unwinding of the DNA strands occurs), a 7bp duplication existing in *S.paradoxus*, the region (denoted by 'S') where RNA polymerase II (pol-II) starts its downstream scan of the template strand, a T-rich region (denoted by 'T') upstream of the transcription start sites (TSSs), and the downstream region (denoted by 'U') where the TSSs are located (in *S.cerevisiae* the TSSs of *RPL4A* were shown to be within 24 bps of the translation start site (Miura et al. 2006)). Upstream of the core promoter, we annotated two orthologous blocks, one (denoted by 'F') containing an Sfp1 site and another (denoted by 'A') containing an Abf1 site and a weak Fhl1 site (sites were annotated using the Sfp1, Abf1 and Fhl1 PSSM models from the SwissRegulon database (Pachkov et al. 2013)).

References

Badis G, Chan ET, van Bakel H, Pena-Castillo L, Tillo D, Tsui K, Carlson CD, Gossett AJ, Hasinoff MJ, Warren CL, et al. 2008. A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol Cell* **32**: 878–87.

Basehoar AD, Zanton SJ, Pugh BF. 2004. Identification and distinct regulation of yeast TATA box-containing genes. *Cell* **116**: 699–709.

Efron B, Hastie T, Johnstone I, Tibshirani R. 2004. Least Angle Regression. *Ann Stat* **32**: 407–499.

Lubliner S, Keren L, Segal E. 2013. Sequence features of yeast and human core promoters that are predictive of maximal promoter activity. *Nucleic Acids Res* **41**: 5569–81.

Miura F, Kawaguchi N, Sese J, Toyoda A, Hattori M, Morishita S, Ito T. 2006. A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc Natl Acad Sci U S A* **103**: 17846–51.

Pachkov M, Balwierz PJ, Arnold P, Ozonov E, van Nimwegen E. 2013. SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Res* **41**: D214–20.

Raveh-Sadka T, Levo M, Shabi U, Shany B, Keren L, Lotan-Pompan M, Zeevi D, Sharon E, Weinberger A, Segal E. 2012. Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat Genet* **44**: 743–50.

Segal E, Widom J. 2009. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr Opin Struct Biol* **19**: 65–71.

Zeevi D, Sharon E, Lotan-Pompan M, Lubling Y, Shipony Z, Raveh-Sadka T, Keren L, Levo M, Weinberger A, Segal E. 2011. Compensation for differences in gene copy number among yeast ribosomal proteins is encoded within their promoters. *Genome Res* **21**: 2114–28.

Zhang Z, Dietrich FS. 2005. Mapping of transcription start sites in Saccharomyces cerevisiae using 5’ SAGE. *Nucleic Acids Res* **33**: 2838–51.

Zhu C, Byers KJRP, McCord RP, Shi Z, Berger MF, Newburger DE, Saulrieta K, Smith Z, Shah M V, Radhakrishnan M, et al. 2009. High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res* **19**: 556–66.

Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *J R Stat Soc* **67**: 301–320.