Supplementary Information

Core promoter sequence in yeast is a major determinant of expression level

Shai Lubliner[1,#], Ifat Regev[1,2,#], Maya Lotan-Pompan[1,2],
Sarit Edelheit[3], Adina Weinberger[1,2,*], Eran Segal[1,2,*]

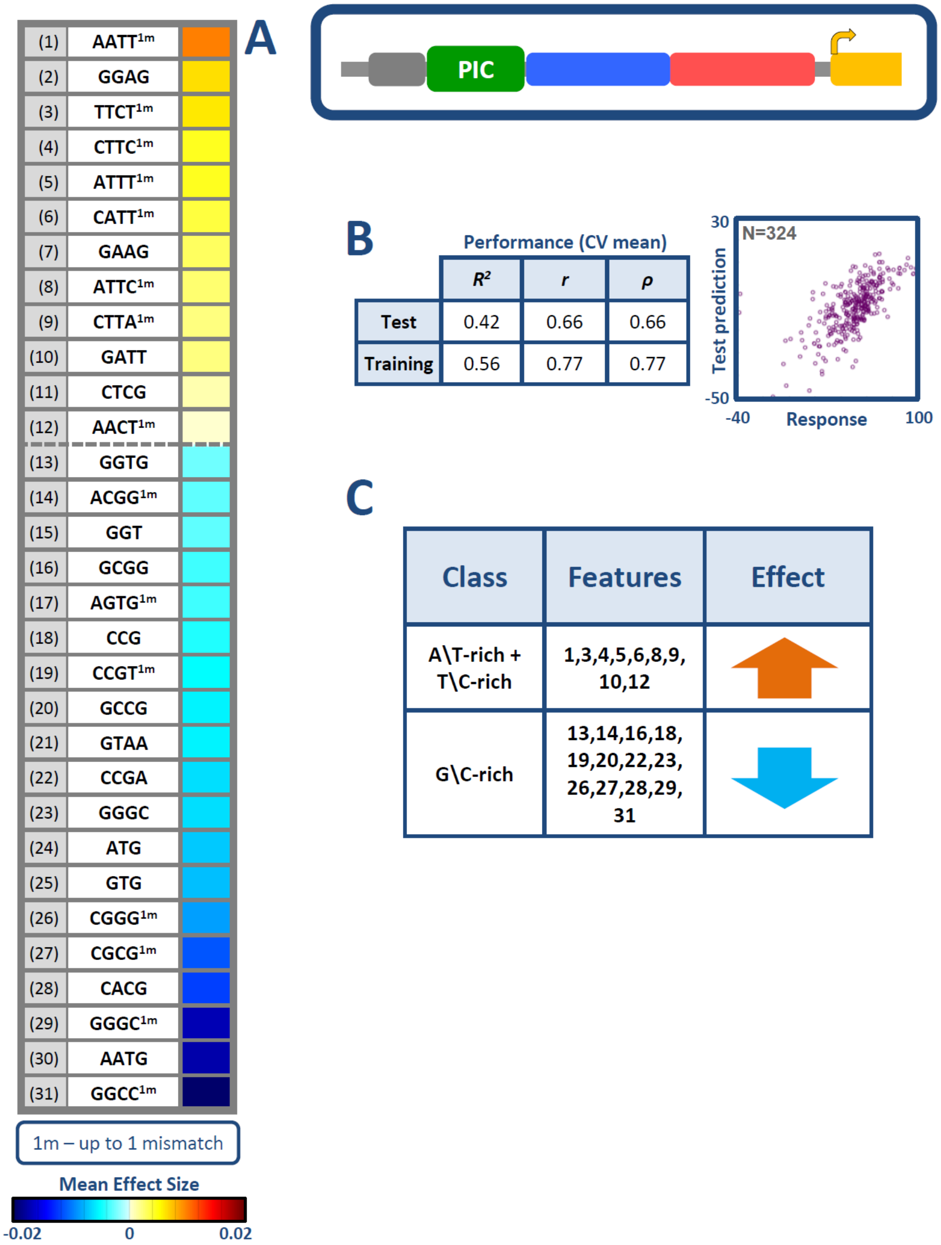
[1]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel.

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

[3]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel.

[#]These authors contributed equally to this work

[*]Corresponding authors: eran.segal@weizmann.ac.il, adina.weinberger@weizmann.ac.il



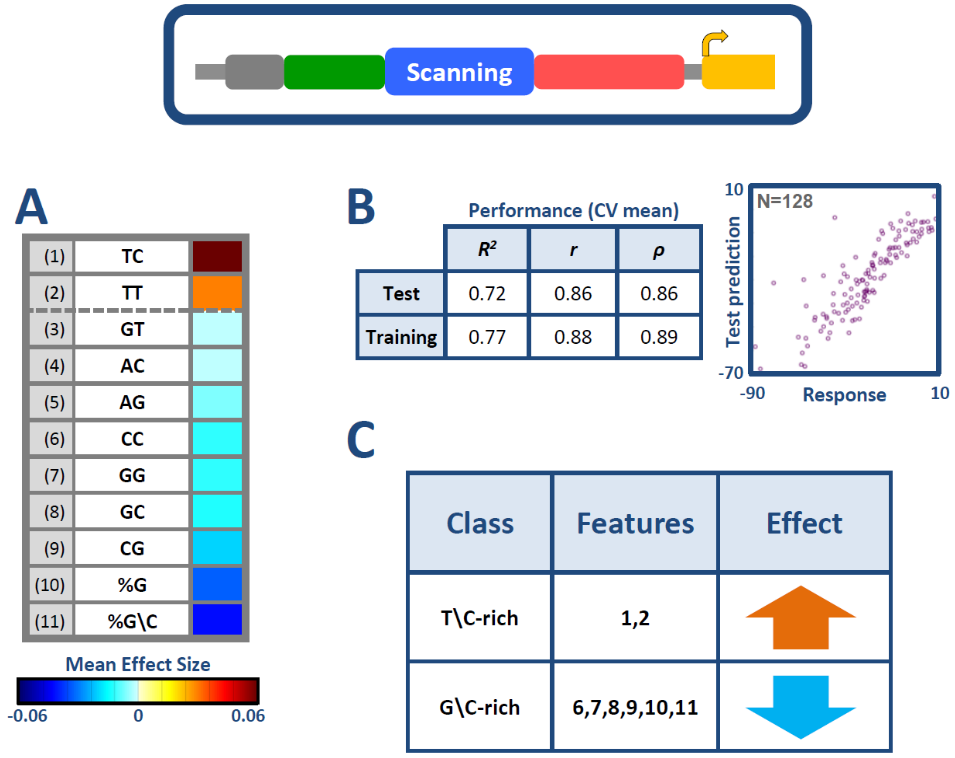
(1)	AATT^{1m}	🟧
(2)	GGAG	🟨
(3)	TTCT^{1m}	🟨
(4)	CTTC^{1m}	🟨
(5)	ATTT^{1m}	🟨
(6)	CATT^{1m}	🟨
(7)	GAAG	🟨
(8)	ATTC^{1m}	🟨
(9)	CTTA^{1m}	🟨
(10)	GATT	🟨
(11)	CTCG	🟨
(12)	AACT^{1m}	🟨
(13)	GGTG	🟦
(14)	ACGG^{1m}	🟦
(15)	GGT	🟦
(16)	GCGG	🟦
(17)	AGTG^{1m}	🟦
(18)	CCG	🟦
(19)	CCGT^{1m}	🟦
(20)	GCCG	🟦
(21)	GTAA	🟦
(22)	CCGA	🟦
(23)	GGGC	🟦
(24)	ATG	🟦
(25)	GTG	🟦
(26)	CGGG^{1m}	🟦
(27)	CGCG^{1m}	🟦
(28)	CACG	🟦
(29)	GGGC^{1m}	🟦
(30)	AATG	🟦
(31)	GGCC^{1m}	🟦

	R^2	r	ρ
Test	0.42	0.66	0.66
Training	0.56	0.77	0.77

Class	Features	Effect
A\T-rich + T\C-rich	1,3,4,5,6,8,9, 10,12	↑
G\C-rich	13,14,16,18, 19,20,22,23, 26,27,28,29, 31	↓

Supplementary Figure 1

Linear model learning results for PIC region mutations. We used a 5-fold cross validation scheme (described in the Supplementary Note), learning 5 linear models that predict mutational effects on core promoter activity from features of sequence difference between mutated and native core promoters. (A) A table listing the *k*-mer counts features that were included in at least 3 of the 5 models, along with their mean effect size over the 5 models (color coded). (B) The table details mean model performance measures (R^2, Pearson correlation r, Spearman correlation ρ) over the training and held-out test data. The dot-plot is the same as in Figure 3A. (C) A manual classification of most of the features shown in (A). Classes of features that lead to higher (lower) predicted expression are marked with an up (down) arrow.



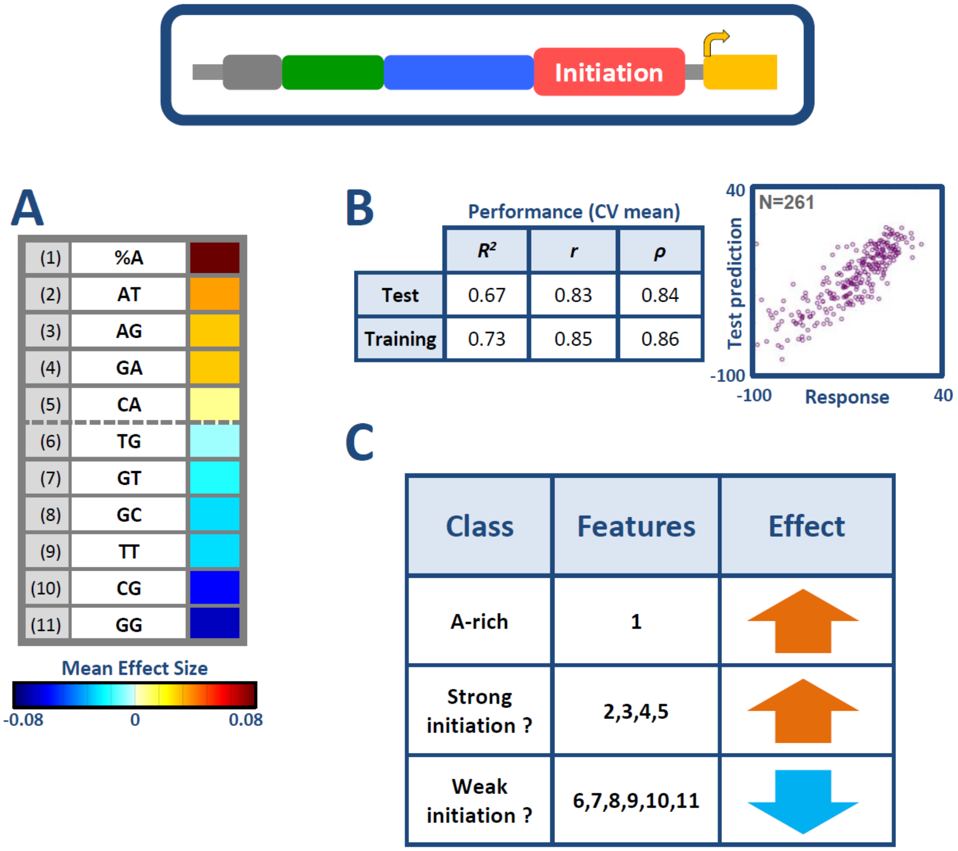
(1)	TC	🟥
(2)	TT	🟧
(3)	GT	🟦
(4)	AC	🟦
(5)	AG	🟦
(6)	CC	🟦
(7)	GG	🟦
(8)	GC	🟦
(9)	CG	🟦
(10)	%G	🟦
(11)	%G\C	🟦

	R^2	r	ρ
Test	0.72	0.86	0.86
Training	0.77	0.88	0.89

Class	Features	Effect
T\C-rich	1,2	⬆
G\C-rich	6,7,8,9,10,11	⬇

Supplementary Figure 2

Linear model learning results for Scanning region mutations. We used a 5-fold cross validation scheme (described in the Supplementary Note), learning 5 linear models that predict mutational effects on core promoter activity from features of sequence difference between mutated and native core promoters. (A) A table listing the *k*-mer counts and base content features that were included in at least 3 of the 5 models, along with their mean effect size over the 5 models (color coded). (B) The table details mean model performance measures (R^2, Pearson correlation r, Spearman correlation ρ) over the training and held-out test data. The dot-plot is the same as in Figure 3B. (C) A manual classification of most of the features shown in (A). Classes of features that lead to higher (lower) predicted expression are marked with an up (down) arrow.



(1)	%A	🟥
(2)	AT	🟧
(3)	AG	🟨
(4)	GA	🟨
(5)	CA	🟨
(6)	TG	🟦
(7)	GT	🟦
(8)	GC	🟦
(9)	TT	🟦
(10)	CG	🟦
(11)	GG	🟦

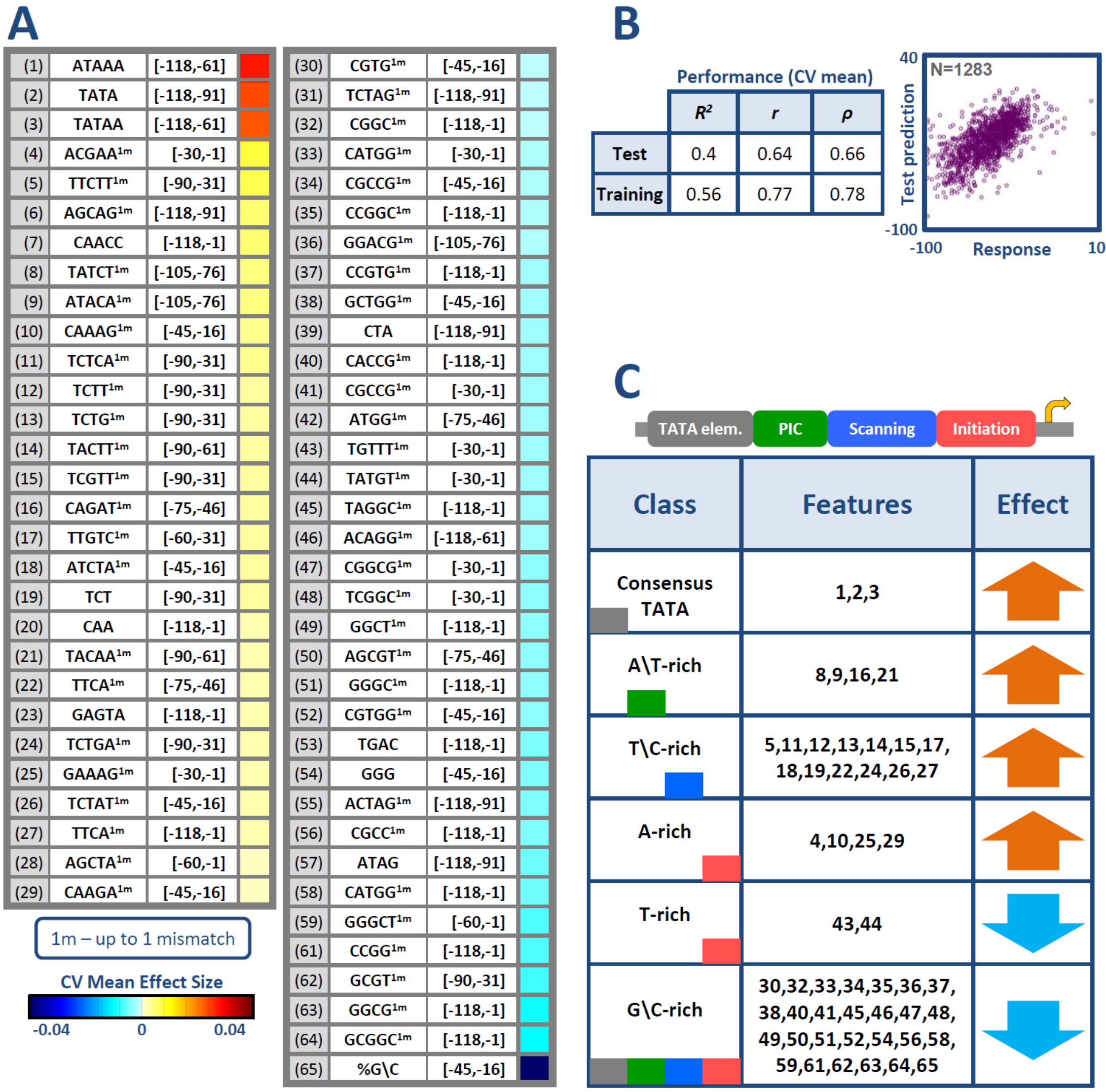
Performance (CV mean)

	R^2	r	ρ
Test	0.67	0.83	0.84
Training	0.73	0.85	0.86

Class	Features	Effect
A-rich	1	⬆
Strong initiation ?	2,3,4,5	⬆
Weak initiation ?	6,7,8,9,10,11	⬇

Supplementary Figure 3

Linear model learning results for Initiation region mutations. We used a 5-fold cross validation scheme (described in the Supplementary Note), learning 5 linear models that predict mutational effects on core promoter activity from features of sequence difference between mutated and native core promoters. (A) A table listing the *k*-mer counts and base content features that were included in at least 3 of the 5 models, along with their mean effect size over the 5 models (color coded). (B) The table details mean model performance measures (R^2, Pearson correlation r, Spearman correlation ρ) over the training and held-out test data. The dot-plot is the same as in Figure 3C. (C) A manual classification of the features shown in (A). Classes of features that lead to higher (lower) predicted expression are marked with an up (down) arrow.



(1)	ATAAA	[-118,-61]	🟥
(2)	TATA	[-118,-91]	🟧
(3)	TATAA	[-118,-61]	🟧
(4)	ACGAA^{1m}	[-30,-1]	🟨
(5)	TTCTT^{1m}	[-90,-31]	🟨
(6)	AGCAG^{1m}	[-118,-91]	🟨
(7)	CAACC	[-118,-1]	🟨
(8)	TATCT^{1m}	[-105,-76]	🟨
(9)	ATACA^{1m}	[-105,-76]	🟨
(10)	CAAAG^{1m}	[-45,-16]	🟨
(11)	TCTCA^{1m}	[-90,-31]	🟨
(12)	TCTT^{1m}	[-90,-31]	🟨
(13)	TCTG^{1m}	[-90,-31]	🟨
(14)	TACTT^{1m}	[-90,-61]	🟨
(15)	TCGTT^{1m}	[-90,-31]	🟨
(16)	CAGAT^{1m}	[-75,-46]	🟨
(17)	TTGTC^{1m}	[-60,-31]	🟨
(18)	ATCTA^{1m}	[-45,-16]	🟨
(19)	TCT	[-90,-31]	🟨
(20)	CAA	[-118,-1]	🟨
(21)	TACAA^{1m}	[-90,-61]	🟨
(22)	TTCA^{1m}	[-75,-46]	🟨
(23)	GAGTA	[-118,-1]	🟨
(24)	TCTGA^{1m}	[-90,-31]	🟨
(25)	GAAAG^{1m}	[-30,-1]	🟨
(26)	TCTAT^{1m}	[-45,-16]	🟨
(27)	TTCA^{1m}	[-118,-1]	🟨
(28)	AGCTA^{1m}	[-60,-1]	🟨
(29)	CAAGA^{1m}	[-45,-16]	🟨
(30)	CGTG^{1m}	[-45,-16]	🟦
(31)	TCTAG^{1m}	[-118,-91]	🟦
(32)	CGGC^{1m}	[-118,-1]	🟦
(33)	CATGG^{1m}	[-30,-1]	🟦
(34)	CGCCG^{1m}	[-45,-16]	🟦
(35)	CCGGC^{1m}	[-118,-1]	🟦
(36)	GGACG^{1m}	[-105,-76]	🟦
(37)	CCGTG^{1m}	[-118,-1]	🟦
(38)	GCTGG^{1m}	[-45,-16]	🟦
(39)	CTA	[-118,-91]	🟦
(40)	CACCG^{1m}	[-118,-1]	🟦
(41)	CGCCG^{1m}	[-30,-1]	🟦
(42)	ATGG^{1m}	[-75,-46]	🟦
(43)	TGTTT^{1m}	[-30,-1]	🟦
(44)	TATGT^{1m}	[-30,-1]	🟦
(45)	TAGGC^{1m}	[-118,-1]	🟦
(46)	ACAGG^{1m}	[-118,-61]	🟦
(47)	CGGCG^{1m}	[-30,-1]	🟦
(48)	TCGGC^{1m}	[-30,-1]	🟦
(49)	GGCT^{1m}	[-118,-1]	🟦
(50)	AGCGT^{1m}	[-75,-46]	🟦
(51)	GGGC^{1m}	[-118,-1]	🟦
(52)	CGTGG^{1m}	[-45,-16]	🟦
(53)	TGAC	[-118,-1]	🟦
(54)	GGG	[-45,-16]	🟦
(55)	ACTAG^{1m}	[-118,-91]	🟦
(56)	CGCC^{1m}	[-118,-1]	🟦
(57)	ATAG	[-118,-91]	🟦
(58)	CATGG^{1m}	[-118,-1]	🟦
(59)	GGGCT^{1m}	[-60,-1]	🟦
(61)	CCGG^{1m}	[-118,-1]	🟦
(62)	GCGT^{1m}	[-90,-31]	🟦
(63)	GGCG^{1m}	[-118,-1]	🟦
(64)	GCGGC^{1m}	[-118,-1]	🟦
(65)	%G\C	[-45,-16]	🟦

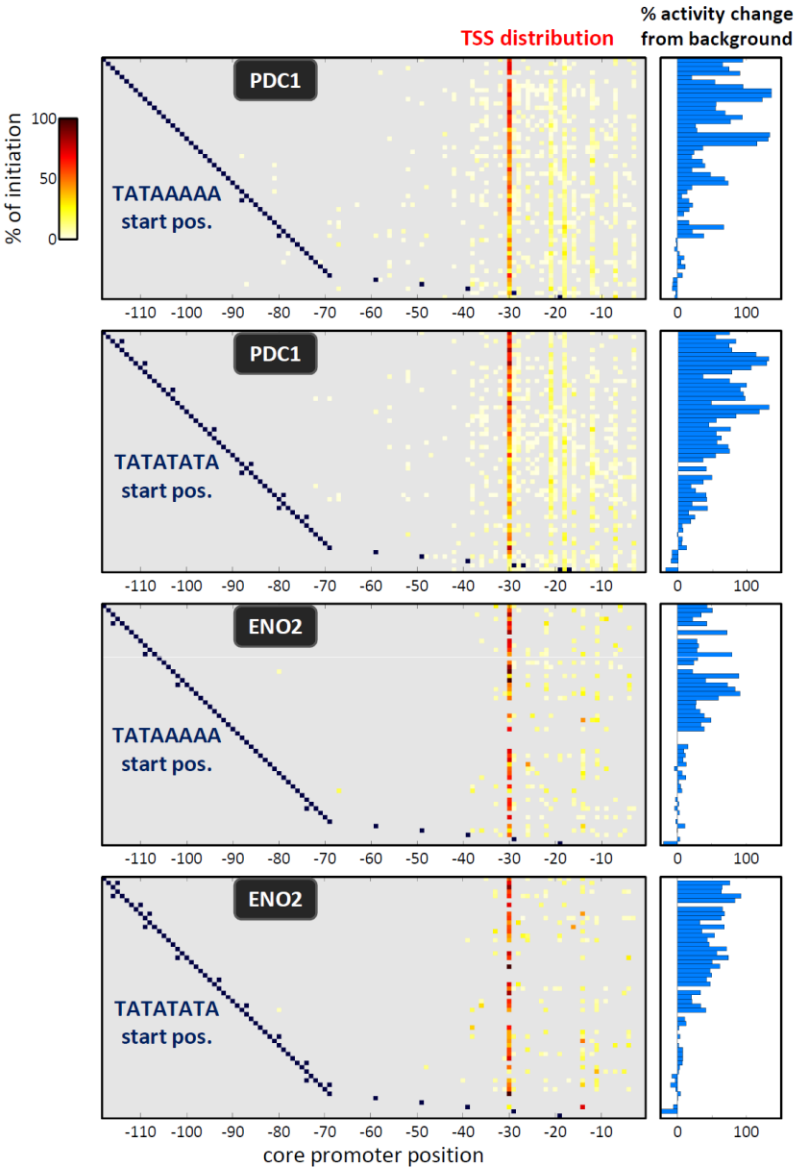
Performance (CV mean)

	R^2	r	ρ
Test	0.4	0.64	0.66
Training	0.56	0.77	0.78

Class	Features	Effect
Consensus TATA ◼	1,2,3	⬆
A\T-rich 🟩	8,9,16,21	⬆
T\C-rich 🟦	5,11,12,13,14,15,17, 18,19,22,24,26,27	⬆
A-rich 🟥	4,10,25,29	⬆
T-rich 🟥	43,44	⬇
G\C-rich ◼🟩🟦🟥	30,32,33,34,35,36,37, 38,40,41,45,46,47,48, 49,50,51,52,54,56,58, 59,61,62,63,64,65	⬇

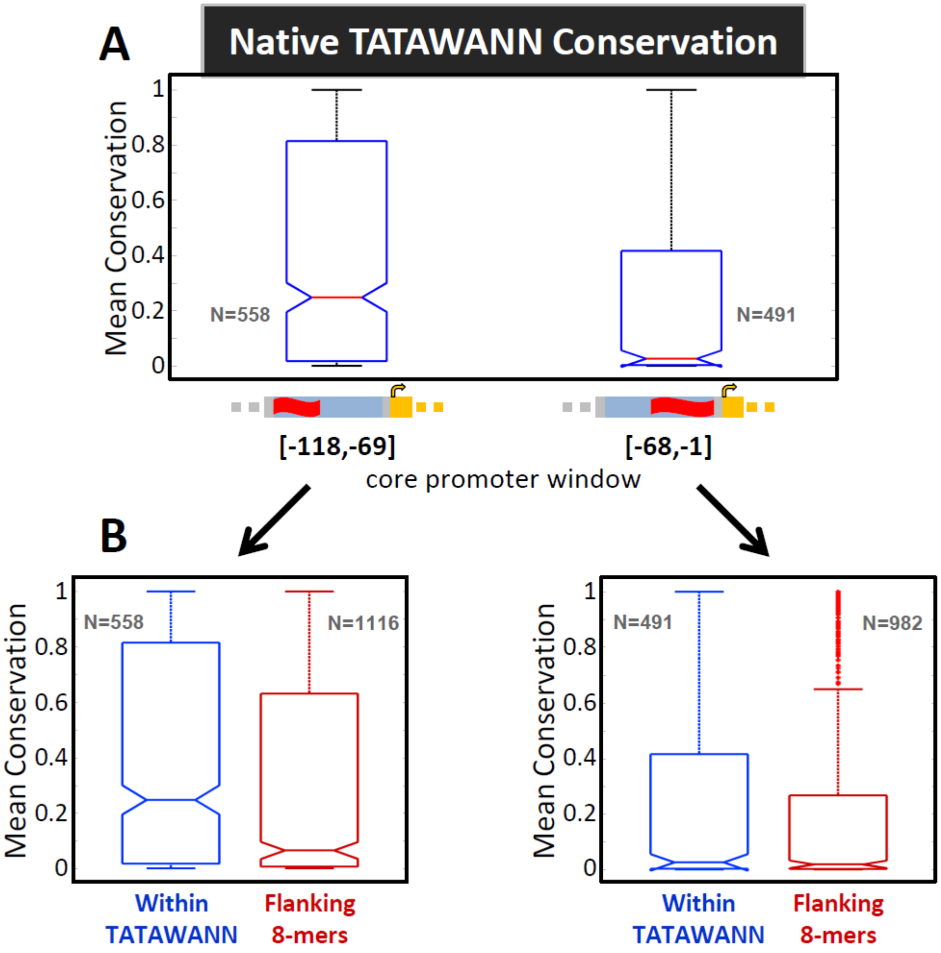
Supplementary Figure 4

Linear model learning results for sliding window mutations. We used a 10-fold cross validation scheme (described in the Supplementary Note), learning 10 linear models that predict mutational effects on core promoter activity from features of sequence difference between mutated and native core promoters. (A) A table listing the *k*-mer counts and base content features (with the core promoter windows in which they were computed) that were included in at least 6 of the 10 models, along with their mean effect size over the 10 models (color coded). (B) The table details mean model performance measures (R^2, Pearson correlation r, Spearman correlation ρ) over the training and held-out test data. The dot-plot is the same as in Figure 3D. (C) A manual classification of most of the features shown in (A). Classes of features that lead to higher (lower) predicted expression are marked with an up (down) arrow.



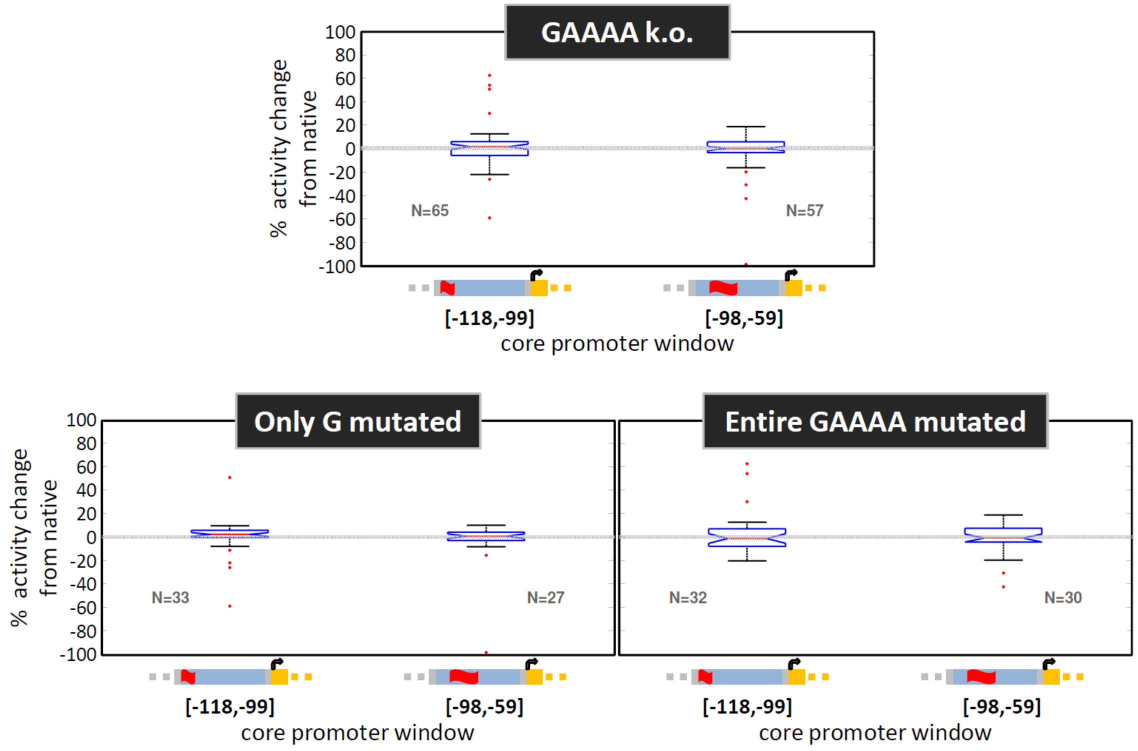
Supplementary Figure 5

TATA element location affects expression and TSS utilization. We inserted the TATA consensus 8-mers TATAAAAA and TATATATA into different positions along the *PDC1* and *ENO2* background sequences (see main text). Each row in the left panel heatmaps corresponds to one insertion case, with TATA start position marked in dark blue (in a few cases the insertion actually resulted in two overlapping TATA 8-mers, and then both start positions are marked), and the measured TSS distribution appears in red and yellow colors (see color bar on the top left). The effect of each insertion on core promoter activity is shown by the corresponding bar within the right panel. Note that there are instances with missing TSS or activity data. The top case of TATAAAAA insertions into the *PDC1* background is also shown in Figure 4.



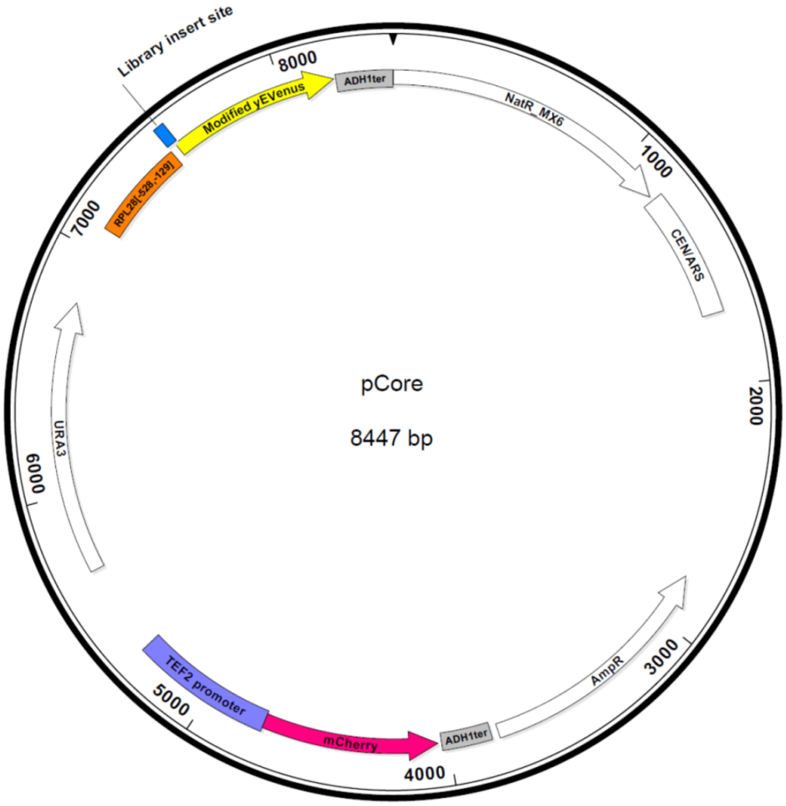
Supplementary Figure 6

Conservation of native TATA elements in different core promoter regions. (A) A comparison of mean position conservation within TATAWANN (W=A/T, N=A/C/G/T) 8-mers found within the [-118,-69] or the [-68,-1] regions of native *S. cerevisiae* promoters (positions are relative to the translation start site). A TATAWANN 8-mer was assigned to each region based on inclusion of its start position. Conservation tracks were downloaded from (http://hgdownload.soe.ucsc.edu/goldenPath/sacCer2/phastCons7way/). (B) A comparison of mean position conservation within the native TATAWANN 8-mers and their flanking 8-mers. The left panel is for the [-118,-69] TATAWANN 8-mers, and the right panel is for the [-68,-1] ones.



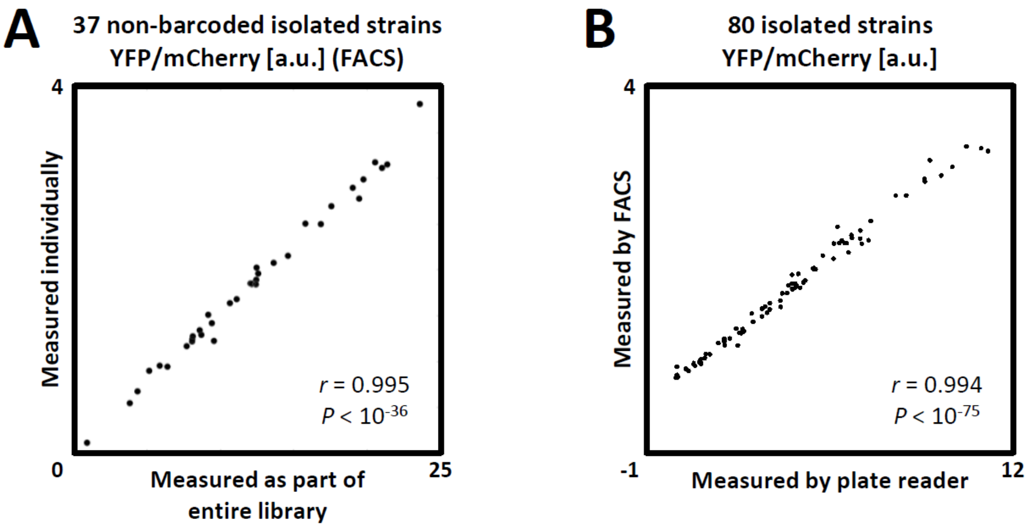
Supplementary Figure 7

Effects of knockout mutations of native GAAAA 5-mers. Box plots of the percent changes to core promoter activity caused by knockout mutations of native GAAAA 5-mers in two core promoter windows: [-118,-99] and [-98,-59]. Assignment to windows was based on the 5-mer start position. The top panel includes all mutations. The bottom left panel includes mutations only of the 'G' nucleotide. The bottom right panel includes mutations of the entire 5-mer.



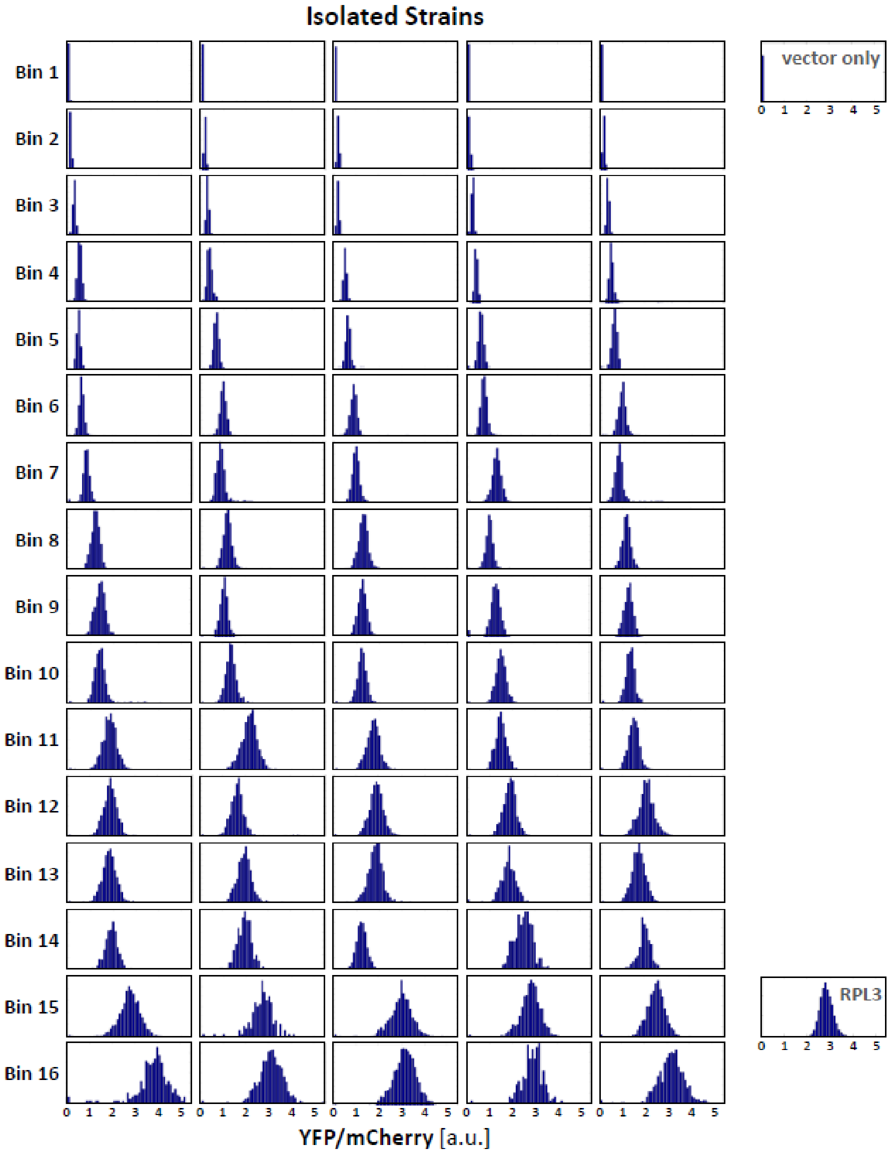
Supplementary Figure 8

Our library's pCore plasmid. The library insert site includes the SexAI and BstXI restriction sites separated by a short linker sequence. The BstXI site is also part of the modified yEVenus (*YFP*). The modified yEVenus sequence is missing the first 37 nucleotides (designed to be part of the inserted oligo), and was additionally modified with a few synonymous mutations in order to insert the BstXI site and also knock out a BstXI site originally found further downstream. See also main text and Supplementary Note.



Supplementary Figure 9

Quality controls using isolated strains. As described in the Supplementary Note, we isolated 80 strains, 5 from each of the 16 YFP/mCherry FACS bins, Sanger sequenced most of them and measured their YFP/mCherry levels using both FACS and plate reader. (A) Based on the Sanger sequencing we were able to detect the identity of 37 non-YFP-barcoded isolates, hence could compare their YFP/mCherry measured by flow cytometry individually (y-axis) and as part of the entire library (x-axis). The two measures were found to be almost perfectly correlated. (B) A comparison of the YFP/mCherry measured for all 80 isolates by flow cytometry (y-axis) and by plate reader (x-axis) reveals an almost perfect correlation between the two measures.



Supplementary Figure 10

Raw flow cytometry measurements for the 80 isolated strains. From each of the 16 YFP/mCherry FACS bins we isolated 5 strains and measured their YFP/mCherry individually using flow cytometry. For comparison we also measured the signal coming from a strain containing our pCore plasmid without a library oligo inserted into it (top right, 'vector only'), and from another strain in which the *mCherry* and *YFP* expression cassette is genomically integrated with YFP expression driven by the *RPL3* promoter, a highly expressed constitutive RP promoter (bottom right, 'RPL3').

Supplementary Note

A few more notes on library oligos design

To distinguish sequencing reads of non-YFP-barcoded oligos from those of YFP-barcoded oligos (to compute mean YFP/mCherry, see below) the A-rich 10-mer upstream of the *YFP* was different between the two sets of oligos.

To clone each oligo into the plasmid required a downstream restriction site that would fit within bases 37-54 of the *YFP* sequence up to synonymous mutations (hence the BstXI was chosen). Since the performed mutations were more than 37 bp downstream of the translation start site, they were not expected to introduce substantial codon bias or mRNA folding related effects on translation efficiency (Kudla et al. 2009; Tuller et al. 2010), and pilot measurements supported this.

The 1000 bp upstream of each core promoter sequence variant

The SexAI restriction site is marked in purple. The [-528,-129] region of the *RPL28* promoter is marked in red. Rap1 binding sites are marked in light blue. Sfp1 binding sites are marked in dark blue. An Fhl1 binding site is marked in green.

GGTTTAGATGACAAGGGAGACGCATTGGGTCAACAGTATAGAACCGTGGATGATGTGGTCTCTACAGGA
TCTGACATTATTATTGTTGGAAGAGGACTATTTGCAAAGGGAAGGGATGCTAAGGTAGAGGGTGAACGT
TACAGAAAAGCAGGCTGGGAAGCATATTTGAGAAGATGCGGCCAGCAAAACTAAAAAACTGTATTATAA
GTAAATGCATGTATACTAAACTCACAAATTAGAGCTTCAATTTAATTATATCAGTTATTACCCTATGCGGT
GTGAAATACCGCACAGATGCGTAAGGAGAAAATACCGCATCAGGAAATTGTAAGCGTTAATATTTTGTTA
AAATTCGCGTTAAATTTTTGTTAAATCAGCTCATTTTTTAACCAATAGGCCGAAATCGGCAAAATCCCTTAT
AAATCAAAAGAATAGACCGAGATAGGGTTGAGTGTTGTTCCAGTTTGGAACAAGAGTCCACTATTAAAG
AACGTGGACTCCAACGTCAAAGGGCGAAAAACCGTCTATCAGGGCGATGGCCCACTACGTGAACCATCA
CCCTAATCAAGTTTTTTGGGGACCTGGACTATGAGCGTAAGCTAATGTTATAAAGAAACAAGCTATAATA
TTGTTAAATATAGTTGATCAACAGCATTGTAATGATTACAAGAGACGAGGTGGAATGAACCTTATGAAAT
GCGTATTATATATAAACTGTAATAAGAGCTAAGTTGAATTGAAATCTACGATACTTGATGTTGACATTATA
GCACTAGTTCCCAGGAAACCCTTTCGAAAAACACAGCAAAAACAAGAGTACTGTAACCAATGTAA**CATCT**
GTACACCAGG**GACCCACACAT**TACCAAAATC**AAAATTATT**TTTCTAATGCCTGTTATTTTTCCTATTTTTCC
TCTGGCGCGTGAATAGCCCGCAGA**GACGCAA****ACAATTTT**CCTCGCAGTTTTTCGCTTGTTTAATGCGTATT
TTCC**ACCTGGT**CTCTGCG

The 1000 bp downstream of each core promoter sequence variant

The constant 10-mer immediately downstream of the variant is marked in blue. The *YFP* sequence is marked in orange. The BstXI restriction site is marked in purple. Synonymously mutated bases (2 for inserting the BstXI site, 1 to delete a BstXI site further downstream) are marked in light blue.

```
TAAATAAAAAATGTCTAAAGGTGAAGAATTATTCACTGGTGTTGTCCCCATTTTGGTGGAATTAGATGGT
GATGTTAATGGTCACAAATTTTCTGTCTCCGGTGAAGGTGAAGGTGATGCTACTTACGGTAAATTGACCT
TAAAATTGATTTGTACTACTGGTAAATTGCCTGTTCCATGGCCAACCTTAGTCACTACTTTAGGTTATGGTT
TGCAATGTTTTGCTAGATACCCAGATCATATGAAACAACATGACTTTTTCAAGTCTGCCATGCCAGAAGGT
TATGTTCAAGAAAGAACTATTTTTTTCAAAGATGACGGTAACTACAAGACCAGAGCTGAAGTCAAGTTTG
AAGGTGATACCTTAGTTAATAGAATCGAATTAAAAGGTATTGATTTTAAAGAAGATGGTAACATTTTAGG
TCACAAATTGGAATACAACTATAACTCTCACAATGTTTACATCACTGCTGACAAACAAAAGAATGGTATCA
AAGCTAACTTCAAAATTAGACACAACATTGAAGATGGTGGTGTTCAATTAGCTGACCATTATCAACAAAA
TACTCCAATTGGTGATGGTCCAGTCTTGTTACCAGACAACCATTACTTATCCTATCAATCTGCCTTATCCAA
AGATCCAAACGAAAAGAGAGACCACATGGTCTTGTTAGAATTTGTTACTGCTGCTGGTATTACCCATGGT
ATTGATGAATTGTACAAATAAGGCGCGCCACTTCTAAATAAGCGAATTTCTTATGATTTATGATTTTTATTA
TTAAATAAGTTATAAAAAAAATAAGTGTATACAAATTTTAAAGTGACTCTTAGGTTTTAAAACGAAAATTC
TTATTCTTGAGTAACTCTTTCCTGTAGGTCAGGTTGCTTTCTCAGGTATAGTATGAGGTCGCTCTTATTGAC
CACACCTCTACCGGCAGATCCGCTAGGGATAACAGGGTAATATACGGGTCACCCGGCCAGCGACATTAA
GGCCCAGAAT
```

Isolating control strains

In our experimental system we introduced small synonymous changes to the *YFP* sequence, as well as depended on proper ligation of the library oligos to the plasmid in order to have an intact *YFP* sequence (see main text). To make sure that these changes did not have any undesirable effect on our ability to accurately measure expression, we isolated 80 strains, 5 from each of the 16 YFP/mCherry bins, Sanger sequenced them and measured their YFP/mCherry levels using both FACS and plate reader, as was done in (Sharon et al. 2012).

Out of the sequenced isolated strains, 37 were not YFP-barcoded thus we had their YFP/mCherry measured by FACS as part of the high throughput library measurements. We found that this measure was extremely correlated to that measured by FACS separately for each strain (r=0.995, $P<10^{-36}$, **Supplementary Fig. 9A**), proving the accuracy of our high throughput measurement. For further validation of accuracy, for all 80 isolated strains we found the YFP/mCherry FACS measurements to be extremely correlated to those measured by plate reader (*r*=0.994, *P*<10^{-75}, **Fig. 9B**).

Computing mean YFP/mCherry

As described in **Figure 1**, following sequencing and read mapping we got a matrix of read counts where each row corresponded to one of the 7,536 non-YFP-barcoded core promoter sequences in our library and columns corresponded to the 16 YFP/mCherry bins. Columns were normalized by dividing each cell by the sum of reads in that column and then multiplying by the proportion of cells sorted into that bin. Then rows were normalized to give a distribution over the 16 bins (summing to 1).

Following that we applied a few filters (after each filter, nonzero rows were again normalized to sum to 1). First, cells that originally had less than 5 reads were zeroed, and so were entire rows that had less than 100 reads in total. Second, cells holding less than 2% of the weight of their row's distribution were zeroed. Third, in many cases we observed bi-modality (and sometimes even tri-modality) in the row distributions, with the left peak at the lowest YFP/mCherry bins. In these cases we took only the right most peak (from the higher YFP/mCherry bins) and zeroed the rest of the cells. If the remaining peak included cells that originally summed to less than 100 reads, or included less than 33% of the weight of the row distribution prior to applying this filter, we zeroed the row.

For each YFP/mCherry bin we computed, based on data measured during the cell sorting, its mean YFP/mCherry over all of its cells, to get the vector of bin YFP/mCherry means. After filtering the above matrix, for each core promoter sequence its mean YFP/mCherry level was computed by the dot product of its row in the above matrix with the vector of bin YFP/mCherry means.

The validity of our filtering is supported by the fact that for the above mentioned 37 non-YFP-barcoded isolated strains we got an extremely high correlation between their high-throughput FACS measure (computed using the filters described here) and the one measured separately for each strain (*r*=0.995, *P*<10^{-36}, **Supplementary Fig. 9A**). With respect to our third filter, we note that the FACS measurements of our 80 isolated strains (see above) did not show YFP/mCherry bi-modality (**Supplementary Fig. 10**). Since our library oligos included the first 54 bases of the *YFP* (see **Fig. 1**), and since they were PCR amplified, we believe that the right most peak represents the real YFP/mCherry levels, and other peaks represent cells that had mutations in their *YFP* sequence. In support of that, out of the 5 isolated strains from the lowest YFP/mCherry bin, 3 lacked an insert altogether and the other 2 had a *YFP* nonsense mutation.

Computing TSS distributions

As described in **Figure 1**, following sequencing and read mapping we got a matrix of read counts where each row corresponded to one of the 5,464 YFP-barcoded core promoter sequences and each column corresponded to a 5'UTR length between 1 and 100. We zeroed cells with less than 5 reads, and normalized each row to give the distribution of initiation events (TSS distribution, summing to 1). In our analyses we only used core promoters for which their respective row in the matrix summed to at least 50 (after zeroing small valued cells).

The above was computed based on our TAP treated sample. In addition we prepared a sample that was untreated with TAP and therefore included only RNA molecules that were not capped (e.g. degradation products). Compared to the TAP treated sample, from this sample we got ~5-fold less sequencing reads that could be mapped to core promoter positions, and these positions mostly did not overlap the ones from the TAP treated sample. We therefore concluded that the amount of noise in the TAP treated sample was low, justifying using it without normalization by the signal from the TAP untreated sample.

Linear model learning scheme

For the purpose of model learning, we randomly partitioned each mutated sequences set to *K* subsets of equal size (with *K*=5 or *K*=10, see main text). We then learned *K* linear models, each time using one of the *K* subsets as a held-out test set, and the union of the rest as a training set. This allowed us to compute mean performance measures for the *K* different models, and also highlight sequence features that were included in most of them.

For the purpose of learning each of the linear models we used the *glmnet* software (http://www.stanford.edu/~hastie/glmnet_matlab) to run the LASSO regression algorithm (Tibshirani 1996) (*glmnet* parameter alpha=1). *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 $\lambda=0$). To select the value of λ, we used a K-fold cross validation scheme over the training data. For this purpose, the training set was randomly partitioned (K 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 K selected λ values.

Core promoter sequence features predicting core promoter activity

Here we bring a short description of the classes of features found to be predictive of core promoter activity, shown in **Supplementary Figures 1-4** and summarized in **Figure 3E**.

Throughout the core promoter, G\C-rich sequence signals (base content, *k*-mers) predict lower expression. G\C content is known to be highly correlated to intrinsic nucleosome occupancy (Tillo and Hughes 2009), which in yeast is known to be highly correlated to in-vivo nucleosome occupancy (Kaplan et al. 2009). Hence, higher core promoter G\C content implies less accessibility of the transcriptional machinery to the DNA, thereby reducing expression.

At the upstream half of the core promoter, consensus TATA *k*-mers predict higher expression. This result adds up to our results showing that consensus TATA elements are functional in the upstream part of the core promoter and that consensus TATA 8-mers increase expression more than 8-mers with 1 mismatch from a consensus one.

At the PIC region, both A\T-rich and T\C-rich *k*-mers predict higher expression. The role of the A\T-rich *k*-mers may be double. They may be part of additional weak TATA elements found downstream of those we annotated, and may also allow more efficient melting of the DNA strands (Giardina and Lis 1993) due to their lower number of hydrogen bonds. The T\C-rich *k*-mers are probably signals related to Pol II scanning efficiency (see next paragraph).

At the Scanning region, T\C-rich *k*-mers predict higher expression. This result is in agreement with our recently published study where T-rich sequence signals upstream of the main TSS were shown to be predictive of higher maximal promoter activity (Lubliner et al. 2013). Indeed, these signals mainly consisted of T-rich *k*-mers that also contained a cytosine base. We hypothesize that these signals contribute to a more rapid scanning of the template strand by Pol II. While thymidine is apparently more common than cytosine in the Scanning region of highly expressed promoters it may very well be the case that both pyrimidines contribute equally to Pol II scanning efficiency, yet the additional constraint on lowering G\C-content in order to allow high expression (see above) introduces a bias in favor of thymidine.

At the Initiation region, A-rich sequence signals predict higher expression while T-rich *k*-mers predict lower expression. These results add up to our results on poly(dA)/poly(dT) inversions, and are in line with past studies (Maicas and Friesen 1990; Lubliner et al. 2013).

References

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

Giardina C, Lis JT. 1993. DNA melting on yeast RNA polymerase II promoters. *Science* **261**: 759–62.

Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, et al. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**: 362–366.

Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in Escherichia coli. *Science* **324**: 255–8.

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.

Maicas E, Friesen JD. 1990. A sequence pattern that occurs at the transcription initiation region of yeast RNA polymerase II promoters. *Nucleic Acids Res* **18**: 3387–93.

Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z, Weinberger A, Segal E. 2012. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol* **30**: 521–30.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *J R Stat Soc B* **58**: 267–288.

Tillo D, Hughes TR. 2009. G+C content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics* **10**: 442.

Tuller T, Waldman YY, Kupiec M, Ruppin E. 2010. Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci U S A* **107**: 3645–50.