Online Methods

Theoretical TFBS abundance

We estimated the abundance of TFBSs in random DNA by analyzing the information contents (ICs) of known motifs associated with yeast TFs[18]. The IC of a motif (IC_{motif}) is proportional to the frequency (f_{motif}) with which that motif is expected to be found on either strand of random DNA with the following relationship, where IC_{motif} is expressed in bits:

$$f_{motif} = 2^{-(IC_{motif}-1)}$$

The number of instances present in a library of a given TFBS motif, assuming that binding sites are independent, is the number of positions in the library that could potentially contain a complete binding site multiplied by the expected frequency of the TFBS motif. For a library with a complexity of 10^7, comprised of 80 bp sequences, the number of possible TFBSs is $(80 – length_{motif} + 1) * 10^7$.

For **Fig. 1a**, we used the average motif length as the $length_{motif}$ for all motifs so that the *x* axis could include frequency and the expected number of binding sites. For this analysis, motifs for zinc cluster monomers were excluded, since these are abundant in the database[18] and are likely to represent only a half TFBS. Several TFBS motifs that are long, but generally have low IC content, were also excluded since they are unlikely to represent true TF specificities. The motifs used in this analysis are summarized in **Supplementary Table 1**.

Promoter library construction

For pTpA and Abf1TATA libraries, a single-stranded oligonucleotide pool was ordered from IDT containing the random 80 bp oligonucleotide flanked by arms complementary to the promoter scaffold for use with Gibson assembly[51]. These oligonucleotides were double stranded with a complementary primer sequence and Phusion polymerase master mix (NEB), gel purified and cloned into the dual reporter vector, ensuring a complexity of at least 10^8 for each library for libraries for which we measured expression, and 10^5 for libraries for which we only inspected the overall expression distribution (**Fig. 1c** and **Supplementary Fig. 1a**). The dual reporter vector yeast_DualReporter (AddGene: 127546) was modified from Sharon et al[7] to fix a mutation in the YFP ORF, and to include a multiple cloning site in the YFP promoter, facilitating promoter scaffold cloning and library construction.

The two promoter scaffold sequences used for GPRA were:

For pTpA:

(poly-T; distal)

GCTAGCAGGAATGATGCAAAAGGTTCCCGATTCGAACTGCATTTTTTTCACAT
C

(poly-A; proximal)

GGTTACGGCTGTTTCTTAATTAAAAAAAGATAGAAAACATTAGGAGTGTAAC
ACAAGACTTTCGGATCCTGAGCAGGCAAGATAAACGA (up to the theoretical TSS).

For Abf1TATA:

(Abf1 site; distal)

GCTAGCTGATTATGGTAACTCTATCGGACTTGAGGGATCACATTTCACGCAGTATAGTTC

(TATA-box; proximal)

GGTTTATTGTTTATAAAAATTAGTTTAAACTGTTGTATATTTTTTCATCTAACGGAACAATAGTAGGTTACGCTAGTTTGGATCCTGAGCAGGCAAGATAAACGA.

In both cases, 80 Ns were inserted in between distal and proximal regions.

We restricted the randomized region to 80 bp because an 80 bp window is short enough that a bound nucleosome would likely cover the entire region, simplifying modeling of accessibility, and because the entire region could be sequenced with a 150 cycle kit, with overlap in the middle, which is necessary because the promoter sequence is unknown until we sequence it. We inserted the N80 oligonucleotide into a region corresponding to ~ (-170):(-90) bp relative to the TSS because this is where most TFBSs lie[52], and because randomizing the region more proximal to the TSS might alter TSS location and translation of the YFP reporter.

For the scaffold library (sequences in **Supplementary Table 2**), the library was cloned in two stages. In the first, the promoter scaffolds (synthesized by microarray synthesis) were amplified and cloned using Gibson Assembly. The resulting library had a common restriction site into which the N80 was cloned by ligation.

Reporter assay

Libraries were transformed into yeast (strain Y8205[53]) using the lithium acetate method[54], starting with 1L of yeast harvested at an OD of 0.3-0.4, ensuring at least 10^8 cells were transformed (with the exception of the high-quality pTpA library, where a dilution series was performed to achieve the desired lower complexity). The yeast were then grown in SC-Ura for two days, diluting the media by 1:4 three times during this period. Media was then either changed to YPD, growing for at least 5 generations prior to cell sorting, or to YPGly and YPGal, with culture grown for at least 8 generations (due to the different carbon source). In the final 10 hours of growth prior to cell sorting, all cultures were allowed to grow continuously in log phase, never achieving an OD above 0.6, by diluting in fresh media. All cultures were grown in a shaker incubator, at 30°C and approximately 250 RPM.

Prior to sorting, yeast were spun down, washed once in ice-cold PBS, and then suspended in ice-cold PBS and kept on ice until cell sorting. Cells were sorted by $\log_2$(RFP/YFP) signal (using mCherry and GFP absorption/emission) on a Beckman-Coulter MoFlo Astrios, using the constitutive RFP under pTEF2 regulation to control for extrinsic noise. Cells were sorted into 18 uniform bins, done in three batches of six bins each, with the exception of the scaffold library, which was sorted into non-uniform bins to account for the higher variance at low expression levels and the larger dynamic range of the library. The FACS configuration varied between experiments (*e.g.*, different laser intensities), resulting in different baseline expression values. Post sort, cells were spun down and resuspended in SC-Ura (supplemented with 1% Gal for Gal sort), grown for 2-3 days, shaking at 30°C. The plasmids were then isolated, the promoter region amplified, Nextera adaptors and multiplexing indices added, and the resulting libraries sequenced with 2 x 76 bp, paired-end reads, using 150 cycle kits on an Illumina NextSeq sequencer, achieving complete coverage of the promoter, including overlap in the center. Libraries were not sequenced to saturation. For example, the pTpA+glucose experiment was sequenced with 155 million reads, yielding 31 million promoter sequences, but doubling the number of reads is projected[55] to only have yielded a further 8.5 million promoter sequences (30%; **Supplementary Fig. 3**).

Promoter sequence consolidation and expression level estimation

The paired end reads representing both sides of the promoter sequence were aligned using the overlapping sequence in the middle, constrained to have 40 (+/-15) bp of overlap for pTpA and Abf1TATA libraries and 16 (+/-10) bp for the scaffold library, and discarding any reads that failed to align well within these constraints. Note that only ~0.3μg of N80 DNA was received from IDT, and only ~10^8 of these were successfully cloned; these are only a vanishingly small portion of the possible 4^{80} sequences in N80 (which would weigh ~10^{26} kg even with just one copy of each possible molecule). Thus, any very similar sequences we observe represent the same source promoter with high probability, with minor differences likely corresponding to PCR or sequencing errors. To collapse related promoters into a single representative sequence, we aligned the sequences observed in each library to themselves using Bowtie2 (version 2.2.1)[56], creating a Bowtie database containing all unique sequences observed in the experiment (default parameters), and aligning these same sequences, allowing for multimapping reads (parameters included "–N 1 -L 18 -a -f --no-sq --no-head -5 <N5> -3 <N3>", where

<N5> and <N3> are the lengths constant termini of the sequences, excluded from the alignment (e.g. "-5 17 –3 13" for pTpA)). Any sequences that align to each other were assigned to the same cluster. Sequences within each cluster were merged, using the sequence with the most reads as the "true" promoter sequence for each cluster. We note that it is impossible to guarantee that the data within an experiment contains no related sequences; this is addressed by using an independently-created high-quality experiment as test data.

Expression levels for each promoter sequence were estimated as the weighted average of bins in which the promoter was observed. For those observed only once, the expression level was the center of the observed bin. Although the high-quality pTpA+glucose dataset theoretically had ~100,000 promoter sequences in it, we restricted our analysis to only those ~10,000 promoter sequences that had sufficient coverage (>100 reads each).

Estimating the proportion of active random promoter sequences

We also created a library of scaffolds that included 3,811 scaffolds that were random but for the restriction site required to ligate in a random 80 mer, and the proximal 50 bp was ensured to be free of ATGs (to avoid out-of-frame reporter translation). Each scaffold included fixed distal and proximal promoter regions (-298:-195 and –103:-33, relative to the theoretical TSS, respectively) surrounding a variable 80 bp random oligonucleotide (-189:-109 regions). Each random scaffold was tested with ~660 random 80-mers, yielding approximately 2.5x10^6 distinct random promoter sequences total. This scaffold library was sequenced with a 300 cycle kit using a 190 bp read 1 and 112 bp read 2.

Promoter sequences were first clustered into those sharing a common scaffold, using Bowtie2 to align to the known scaffold sequences (using the following parameters: -L 18 -p 4 -f --no-sq --no-head --np 0 --n-ceil C,100). Promoter sequences were then sub-clustered within each scaffold using the sequences of the random 80-mers using CD-HIT (version 4.6.5, using the following parameters: -g 1 -p 1 -r 0 -c 0.96 -uS 0.05 -uL 0.05 -mismatch -1) [57], yielding a single consensus sequence for each promoter.

We estimated the proportion of random promoter sequences that were expressed at detectable levels using the empirical log(YFP/RFP) distributions of regrown, previously-sorted, cells (as in **Fig. 2b**). We considered any bin above the lowest expression bin to be "expressed", but since some cells might end up in this lowest expression bin upon re-sorting, we attempted to estimate the number of cells that would remain expressed upon resorting. AUROC statistics were calculated to estimate how well the cells sorted into each bin can be distinguished from those sorted into the not-expressed bin. Here, each AUROC is equivalent to the probability that a cell sorted into the corresponding expressing bin is expressed higher than a randomly selected cell from the not-expressed bin. Thus, cell proportions in expressing bins were weighted by the corresponding AUROC for that bin to get an estimate of the number of expressing random promoters, 83%.

Testing native yeast promoters by GPRA

To test native yeast promoters in the GPRA system, the promoter sequences from the S288C reference genome (v64; TSS coordinates given in **Supplementary Table 3**) were segmented into 80 bp fragments (from the TSS to -480), overlapping by 40 bp, for a total of 11 fragments per promoter and 62,897 promoter fragments overall. We also included 8,027 random promoter sequences originally assayed within the high-quality pTpA N80 glucose experiment for use as controls (these were excluded from analyses evaluating model performance on native promoter sequences). The sequences were created by pooled oligonucleotide synthesis (Twist Biosciences), including ends complementary to the pTpA scaffold. The fragments were amplified by PCR and cloned into the pTpA vector by Gibson assembly. The resulting library was transformed into yeast (S288C *ura3*Δ) and assayed as described above, with two replicates. We combined the two replicates, which showed some non-linearities resulting from differences in FACS binning procedures, using loess regression (span=0.1) to remove the non-linear relationship between one replicate and the average of the two replicates. After combining the replicates, the Pearson r^2 between expression measurements in the combined replicates and the expression values originally measured for the high-quality random promoter sequences (from the high-quality pTpA N80 glucose experiment) was 0.977.

Linear transcription model

TF motifs (**Supplementary Table 1**) were taken from the YeTFaSCo database[18] and supplemented with the poly-A motif (AAAAA), which we initialized to 100% A at all five positions. Motifs were trimmed to fill 25 bp 1-d convolutional filters, centering the motif if it was less than 25 bp, and, where motifs were longer than 25 bp, trimming off the least informative bases until it was 25 bp.

To identify dissociation constants, K_d, for each TFBS motif and each potential binding site instance, motif filters were applied to DNA sequences of each promoter (DNA_p) and their reverse complements by scanning them with the TFBS motif position weight matrix for each yeast TF (PWM). Binding to each site in the DNA was determined by the GOMER method using a fixed TF concentration (C_x) that corresponds to the minimum K_d possible with the motif (and therefore a perfect match corresponds to 50% occupancy)[21]. We considered all TFBSs, such that weak sites can also be influential, creating an affinity landscape for each TF across the region[58], and summed the predicted occupancy at each site, to obtain the expected occupancy for each TF of each sequence.

The expected binding (sum of all binding to all binding sites; DB_{px}), assuming Michaelis-Menten equilibrium binding occupancies for all possible binding sites (position i, strand s) for TF x in promoter p, where K_ds for each binding site are calculated from the position weight matrix:

$$DB_{px} = \sum^{\substack{\text{strand } s \\ \text{position } i}} \frac{1}{1+\frac{Kd_{pxsi}}{C_x}}$$

Correlations between predicted occupancy for each individual TF and expression level were done using these values (DB_{px}). We optimized a single “activity” weight for each TF (A_x), representing the ability of that TF to activate or repress transcription, as well as a constant (c_1).

$$EL_p = c_1 + \sum^{\text{TF } x} DB_{px} A_x$$

This model was implemented in Tensorflow, as described for the other models below, but without a regularization term.

Billboard model of transcription

The billboard model includes parameters for TF concentration (C_x), TF activity (A_x), TF potentiation (P_x), and TF activity limits (AL_x). Motifs were trimmed, as before, but filling 25 bp 1-d convolutional filters. As described above, we use these filters, the DNA sequence of each promoter (DNA_p), and the (now learned) TF concentration parameter to gain an initial estimate for DNA binding in the absence of chromatin (DB_{px}).

Some TFs can displace nucleosomes, so the model learns TF-specific parameters that capture the ability of each TF to modulate the binding of other TFs (P_x), which we assume is primarily driven by chromatin opening. Promoter accessibility is estimated as a logistic function on the potentiation-weighted DB_{px} estimates (including a constant c_p), yielding a probability of the DNA being accessible (Ω_p):

$$\Omega_p = \frac{1}{1+ e^{(c_p+\sum^{\mathrm{TF}\ x} DB_{px}P_x}}$$

Since nucleosomes can potentially prevent TF binding[25], the previous estimate of binding (DB_{px}) is then scaled with this value, yielding the expected binding of each TF to each promoter in the context of chromatin (CB_{px}):

$$CB_{px} = DB_{px} \times \Omega_p$$

Because our promoters are small, we can reasonably assume that a TF that opens chromatin would open it for the entire 80 bp variable region: if the promoter is open, all TFs can bind unimpeded; if the promoter is closed, no TFs can bind. For example, a promoter that is predicted to be 0% accessible will have no TF binding, regardless of the TFBSs present in the sequence (CB_{px} = 0 for all TFs x), while a promoter that is 100% accessible will have occupancy unchanged (DB_{px} = CB_{px}). Thus, the model learns which

TFs may, for example, open and close chromatin by their ability to potentiate the activity of other TFs (*i.e.*, TFBSs for TFs that affect transcription, but cannot open chromatin, only have an effect when "potentiated" by another factor, presumably by opening chromatin and allowing binding).

Finally, the predicted expression level (EL_p) is the sum of binding values for each TF x, weighted by their learned effect on expression (A_x), including a constant c_e, similar to the linear model described above:

$$EL_p = c_e + \sum^{\mathrm{TF}\,x} CB_{px}\, A_x$$

Here, the measured and predicted expression levels are in log space, corresponding to the log-space bins of YFP/RFP. One possible interpretation of the formulation above is that TF activities are proportional to how much the TF affects the zero-order rate constants for different steps of mRNA production, which would be multiplicative in linear space or additive (as above) in log space.

When activity limits for TFs (AL_x) were included as a learned parameter, the expression level was instead calculated as follows, putting an upper limit on TF activity:

$$EL_p = c_e + \sum^{\mathrm{TF}\,x} \begin{cases} \min\left(CB_{px}A_x, AL_x\right), \text{if } CB_{px}A_x \geq 0 \\ \max\left(CB_{px}A_x, AL_x\right), \text{otherwise} \end{cases}$$

Position-specific activity model

Position-specific activity models (**Supplementary Fig. 9**) were built as an extension of the billboard model that included activity limits. Here, each potential TFBS position was allowed its own (learned) activity parameter. Position-specific TF binding in chromatin was estimated similarly to before, but accounting for the strand (s) and binding position (*i*) of each TF (*x*) to each promoter (*p*), again, weighted by the estimated accessibility of the DNA (Ω_p; calculated as before):

$$CB_{pxsi} = BS_{pxsi} \times \Omega_p$$

The activity contribution of each TF on each promoter (AC_{px}) was estimated using the position-specific activity parameters (A_{xsi}), which were implemented as a local scale of the overall TF activity (A_x) learned previously:

$$AC_{px} = \sum^{\substack{\text{strand } s \\ \text{position } i}} CB_{pxsi} A_{xsi}$$

We then re-implement the binding limits as follows:

$$EL_p = c_a + \sum^{\text{TF } x} \begin{cases} \min\left(AC_{px}, AL_x\right), \text{if } AC_{px} \geq 0 \\ \max\left(AC_{px}, AL_x\right), \text{otherwise} \end{cases}$$

Model learning

Parameters were learned iteratively, first learning TF activity and potentiation, then TF concentration, then allowing the motifs themselves to be changed, then including a parameter that limited the maximum activity of each TF, and finally learning position-specific activity parameters, each time, learning the new parameters and updating those previously included with a single pass through the data.

Transcriptional models were implemented in Tensorflow[59], minimizing the mean squared error between predicted and measured expression level using the AdamOptimizer and learning in batches of 1,024 promoter sequences. In all cases (except the linear model above), potentiation and activity parameters were regularized with an $L1$ penalty (0.00001), motifs were regularized with an $L2$ penalty (0.000001), and position-specific activity biases (when present) were regularized with an L2 penalty (0.00001) on the difference between adjacent (by location l) activity biases. Learning rate was set to 0.04 for the epoch learning activity and potentiation parameters, 0.01 when also learning concentration, and 0.001 when also learning motifs, activity limits, and position-specific activities. The model learning motifs *de novo* (on pTpA+glucose data) was initialized with 1,000 random motifs (PWM values normally distributed about mean=0, SD=1) of width 30 bp, potentiations and activities were initialized to 0.01, learning rate was set to 0.001, and the model was trained on 10 epochs of the training data. In addition to the canonical GRFs, this model also identified the Cbf1 motif as a potentiating motif, consistent with previous descriptions[60]. All learned parameters are included in **Supplementary Table 4**.

Applying models to native sequences

Since the models above were designed to operate on relatively short sequences (~110 bp), scanning the yeast genome (R64) was done in tiling windows of 110 bp each, spaced at 1 bp intervals, yielding expression and accessibility predictions for nearly all bases in the genome.

To compare to chromatin organization in native promoters, the accessibility predictions were averaged across all yeast promoter sequences to yield a metagene plot, as was done for DNase[27] and nucleosome occupancy[26] data.

Predictions on the 80 bp fragments of native promoters tested in the pTpA scaffold were done as with other pTpA-scaffold model predictions.

Comparing refined and original motifs

The original and model-refined motifs were evaluated for their ability to predict independent ChIP binding and TF mutant gene expression data. The GOMER method[21] was used to get a predicted binding occupancy of each sequence for the original and model-refined motifs. For ChIP data[45], ChIP-chip probes were scanned with the motifs, and their ability to predict ChIP binding for the corresponding TF was evaluated. For TF perturbation experiments[18,46] promoter sequences were scanned with motifs, and their ability to predict expression changes when the cognate TF is perturbed (mutated, over-expressed, or deleted) was evaluated. In both cases, there were often multiple experiments for the same TF. We repeatedly sampled the data from each experiment (50% of the data sampled randomly 100 times, without replacement), and with each sample calculated the Pearson correlation coefficient between motif-predicted binding and biological measurement (gene expression, ChIP intensity) for both model-refined and original motifs. If the model-refined motif had a Pearson r^2 greater than the original in at least 95% of samples, we considered the experiment to be predicted better by the refined motif. Conversely, if the original motif was better in at least 95% of samples, the experiment was considered to be predicted worse by the refined motif. A model-refined motif was considered to be better than the original if at least one experiment was predicted better and no experiment was predicted worse, while it was considered worse if at least one experiment was predicted worse and no experiment was predicted better. In all other cases, the motifs were considered equal. Motifs that were regularized out of the model (*i.e.* became neutral PWMs) were not considered in this analysis.

Classifying TFs into activators and repressors by GO annotation

GO terms for yeast genes were downloaded from SGD[61] on Jan. 14, 2017. TFs annotated with a term containing any of "positive regulation of transcription", "transcriptional activator", "activating transcription factor binding", or "positive regulation of RNA polymerase II" were labeled as activators. TFs annotated with "negative regulation of transcription", "transcriptional repressor", "repressing transcription factor binding", or "negative regulation of RNA polymerase II" were labeled as repressors. Any annotated as both or neither were ignored for the purposes of testing for enrichment.

The models predicted that most TFs opened rather than closed chromatin (*i.e.*, had positive potentiation scores; 64-66%) and most were predicted activators rather than repressors (53-55%), although most TFs in all four experiments were predicted to have little activity, consistent with many TFs being inactive in rich media[62].

Promoter library MNase-Seq

Aliquots of the pTpA library, expected to correspond to ~100,000 (sample A) or ~200,000 (sample B) viable cells were each cultured in duplicate (Rep 1 and 2) in YPD for ~16 hours to an OD of ~0.4-1.0. For each sample, 0.5 mL of culture was pelleted and frozen to prepare input genomic DNA, and 3 mL of culture was crosslinked with 1% formaldehyde, washed twice with 1mL H_2O supplemented with a protease inhibitor cocktail, and the pellet frozen for MNase treatment. These pellets were next spheroplasted using zymolyase, and spheroplasts were lysed in NP buffer (10 mM Tris pH 7.4, 50 mM NaCl, 5 mM $MgCl_2$, 1 mM $CaCl_2$, and 0.075% NP-40, freshly supplemented with 1 mM β-mercaptoethanol, 500 μM spermidine, and EDTA-free protease inhibitor cocktail) at a concentration of 2 x 10^6 cells/ μl of NP buffer. 0.125 units of Worthington MNase were added per 10μl of lysed spheroplasts and MNase digestion was performed at 37°C for 20 minutes. MNase digestion was stopped by addition of equal volume of 2X MNase Stop Buffer (220 mM NaCl, 0.2% SDS, 0.2% sodium deoxycholate, 10 mM EDTA, 2% Triton X-100, EDTA-free protease inhibitor cocktail). MNased chromatin samples were treated with RNase A and proteinase K, reverse cross linked, separated on a 4% agarose gel and mononucleosome bands were isolated. Genomic DNA was prepared using the Masterpure Yeast Genomic DNA Preparation Kit (Epicenter). For both MNase and genomic DNA, the variable region of the promoter library was amplified, and adaptors added for sequencing using an Illumina NextSeq with 76 bp single-end reads.

Sequencing reads were mapped to all known promoters in any pTpA library using Bowtie2[56]. Only promoter sequences with at least 20 reads in the input DNA and 1 read in the MNase data were kept for subsequent analysis. Input and MNase counts were scaled within each sample to yield counts per million (CPM) per promoter sequence and the log ratio of MNase to input was compared between replicates and to the model's predicted occupancy, corresponding to log(1-predicted accessibility). To combine MNase replicates, the log ratio of MNase to input was averaged for promoter sequences present in both samples – those in only one sample were ignored. Similarly, pairwise correlations between samples in **Fig. 3a** reflect only the promoter sequences common to both samples, and all promoter sequences within the sample when comparing to the model's predictions. Spearman's rho was used to compare to model predictions, which is rank-invariant (unaffected by log-transformation).

Position and orientation-specific TF activities

In order to identify the approximate fraction of TFs displaying a 10.5 bp helical activity bias, the position-specific activities across the variable promoter region were compared to a 10.5 bp sine wave. First, the overall positional activity bias was regressed out using loess regression (span=0.5; green curves in **Supplementary Fig. 13a**). These long-range trends were subtracted from the data, leaving only the short-range trends (blue curves in **Supplementary Fig. 13a**), which were then compared to a 10.5 bp sine wave for 100 possible alignments of the sine wave, taking the largest magnitude correlation for each TF and strand, and calculating Spearman's correlation coefficient, ρ. As background, the same procedure was performed after first shuffling the position-specific activity biases for 100 permutations of the data per TF. A P-value and AUROC were calculated describing the difference between the randomized and actual data for each model using Wilcoxon's rank sum test. Although we tried using a Fourier Transform for this analysis and its results were suggestive of a 10.5 bp period, the length of the region being studied (~80 bp) was too short to yield sufficient signal.

Testing designed sequences and motif tiling in random sequences by GPRA

We generated 1,000 random DNA sequences *in silico*, predicted expression for each using the pTpA+glucose positional model, found that they were predicted to span a wide range of expression levels, and included these sequences for synthesis (below). We further selected three of these corresponding to the 25, 50, and 75th percentiles for predicted expression, as background sequences in which to embed motifs. We then embedded a single consensus for each motif (polyA: AAAAA; Skn7: GTCTGGCCC; Mga1: TTCT; Ume6: AGCCGCC; Mot3: GCAGGCACG; and Azf1: TAAAAGAAA) at every possible position (with the motif contained completely within the 80 bp variable region) and orientation for each of the three background sequences, for a total of 2,658 sequences. We synthesized (Twist Biosciences), cloned, and assayed these sequences as described above, using the pTpA scaffold, and measuring expression in glucose. Data were processed as before, but considering only reads that were perfect matches to the sequences ordered (*i.e.*, no mismatches or indels). All sequences for which at least one read matching that sequence was observed were included in the expression estimates. For the plots in **Fig. 5** and **Supplementary Fig. 14**, we used loess regression to correct for a nonlinear relationship between the predicted-*vs*.-actual expression resulting from differences in the relative scaling of the bins between experiments. Reported Pearson r^2s are on the raw data (without correction).

Data availability Data are available at NCBI's GEO: GSE104903, GSE104878.

Code availability

Open source code for our transcriptional models is available at https://github.com/Carldeboer/CisRegModels

Statistics

All statistics were calculated in R, with the cor.test function for calculating Pearson and Spearman correlation coefficients and associated P-values (as indicated in the text), phyper for calculating hypergeometric P-values (one tailed), and pnorm and a custom Fisher's *r* to *z* function (function (r1,r2,n) (atanh(r1) - atanh(r2)) / ((1/(n-3))+(1/(n-3)))^0.5) for calculating the significance of differences in Pearson's *r*.

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Supplementary References

51. Gibson, D.G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature methods* **6**, 343-345 (2009).

52. Erb, I. & van Nimwegen, E. Transcription factor binding site positioning in yeast: proximal promoter motifs characterize TATA-less promoters. *PloS one* **6**, e24279 (2011).

53. Tong, A.H. & Boone, C. Synthetic genetic array analysis in Saccharomyces cerevisiae. *Methods Mol Biol* **313**, 171-192 (2006).

54. de Boer, C. High-efficiency S. cerevisiae lithium acetate transformation. *protocols.io* (2017).

55. Deng, C., Daley, T. & Smith, A.D. Applications of species accumulation curves in large-scale biological data analysis. *Quant Biol* **3**, 135-144 (2015).

56. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359 (2012).

57. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).

58. Segal, E. & Widom, J. From DNA sequence to transcriptional behaviour: a quantitative approach. *Nature reviews. Genetics* **10**, 443-456 (2009).

59. Abadi, M. et al. TensorFlow: Large-scale machine learning on heterogeneous systems. (2015).

60. Kent, N.A., Eibert, S.M. & Mellor, J. Cbf1p is required for chromatin remodeling at promoter-proximal CACGTG motifs in yeast. *The Journal of biological chemistry* **279**, 27116-27123 (2004).

61. Cherry, J.M. et al. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic acids research* **40**, D700-705 (2012).

62. Chua, G. et al. Identifying transcription factor functions and targets by phenotypic activation. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 12045-12050 (2006).