ARTICLES
https://doi.org/10.1038/s41587-019-0315-8
nature biotechnology

Deciphering eukaryotic gene-regulatory logic with 100 million random promoters

Carl G. de Boer[1*], Eeshit Dhaval Vaishnav[1,2], Ronen Sadeh[3], Esteban Luis Abeyta[4], Nir Friedman[1,3] and Aviv Regev[1,2*]

How transcription factors (TFs) interpret *cis*-regulatory DNA sequence to control gene expression remains unclear, largely because past studies using native and engineered sequences had insufficient scale. Here, we measure the expression output of >100 million synthetic yeast promoter sequences that are fully random. These sequences yield diverse, reproducible expression levels that can be explained by their chance inclusion of functional TF binding sites. We use machine learning to build interpretable models of transcriptional regulation that predict ~94% of the expression driven from independent test promoters and ~89% of the expression driven from native yeast promoter fragments. These models allow us to characterize each TF's specificity, activity and interactions with chromatin. TF activity depends on binding-site strand, position, DNA helical face and chromatin context. Notably, expression level is influenced by weak regulatory interactions, which confound designed-sequence studies. Our analyses show that massive-throughput assays of fully random DNA can provide the big data necessary to develop complex, predictive models of gene regulation.

Control of gene expression by DNA-binding regulatory proteins, known as *cis*-regulatory logic, is a fundamental determinant of cell phenotype and cell-state transitions. Constructing models of *cis*-regulatory logic generally requires a training set of sequences and associated expression levels. Analysis of the expression of natural regulatory sequences has shown some success[1,2], but their limited diversity and homology mean that models are easily overfit[2], even when the sequences are diversified by mutagenesis[3]. This is likely to be a problem in human cells as well, where there may be ~100,000 active regulatory elements in any given cell type[4–6]. Alternatively, measuring the expression of synthetic promoters, as in a massively parallel reporter assay, using either designed sequences[7] or randomly arranged designed elements[8], allows arbitrary hypothesis testing, but DNA synthesis is both limited in scale and costly. Consequently, TF binding sites (TFBSs) are often tested only in select affinities, contexts, positions and orientations, leading to uncertain generalizability and limiting the hypotheses that can be tested. Overall, the space of possible regulatory sequences far exceeds what has been explored to date. For example, testing all pairwise TF–TF interaction spacings only once each would require ~10^7 sequences. Learning complex regulatory rules might require far more sequences than exist in the genome or have previously been assayed[9]. Given the limited scale of previous work, predictive models of expression level from sequence alone remain elusive.

We hypothesized that fully random DNA could be used to test regulatory sequences at a much larger scale than has been studied previously. Although many sequences in the full space of possibilities may not exist in any organism, the increased scale could allow us to learn complex models of gene regulation. Past experiments that have used random DNA to study gene regulation support our hypothesis. In vitro selection (or systematic evolution of ligands by exponential enrichment (SELEX)) can define the specificities[10] and affinities[11] of TFs by isolating the high-affinity TFBSs that are present by chance in a random DNA pool[12]. Random DNA has also been used to diversify regions of native promoters[13], to explore translational regulation[14] and to show that ~10% of random 100-base-pair (bp) sequences could serve as bacterial promoters[15].

Although TF motifs are expected to occur frequently by chance in random DNA[16], it is often tacitly assumed that functional TFBSs are rare: most TF motif instances are neither evolutionarily conserved nor bound by experimental assays, and it remains unclear whether TFBSs require additional factors to function (for example, site clustering or interactions with neighboring factors)[17]. Thus, it was unclear whether random DNA sequences could drive reproducible expression levels and span a sufficient dynamic range from which to uncover regulatory rules. Moreover, no in vivo experiments have been conducted on the massive scale required to learn the complexities of *cis*-regulatory logic that can both (1) predict expression given any arbitrary sequence and (2) explain how that sequence generated the expression level with interpretable features reflecting mechanisms of gene regulation.

Here, we test our hypothesis, by developing the Gigantic Parallel Reporter Assay (GPRA) to measure the expression level associated with each of tens or hundreds of millions of random DNA sequences per experiment, and use these to learn models of *cis*-regulatory logic in the yeast, *Saccharomyces cerevisiae*, grown in each of three well-characterized carbon sources. We validate our findings in the context of a rich body of knowledge, and show that GPRA is a powerful approach to decipher gene regulation.

Results

Random DNA includes many TFBSs, yielding diverse expression. We first computationally predicted that random DNA sequences contain abundant yeast TFBSs. Consistent with previous models[16], the information content (IC) of TF motifs can be used to quantify

[1]Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Howard Hughes Medical Institute and Koch Institute of Integrative Cancer Research, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]School of Computer Science and Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel. [4]Initiative for Maximizing Student Development Program, University of New Mexico, Albuquerque, NM, USA. *e-mail: carlgdeboer@gmail.com; aregev@broadinstitute.org

their expected frequency in DNA uniformly sampled from the four bases ('random DNA'), without the need to define a 'match' (Fig. 1a and see Methods). For example, of the 221 motifs from YeTFaSCo[18] that we expect to represent true specificities of yeast TFs (see Methods), 58% are expected on average to occur every 1,000 bp or less, and 92% to occur every 100,000 bp or less. Consequently, 80 bp of random DNA is expected to have, on average, ~138 partly overlapping TFBS instances, representing ~68 distinct factors. Thus, a library of 10^7 80-bp random promoter sequences (the minimum assayed per experiment; see "A 'gigantic' parallel reporter assay of random DNA") is expected to include >10,000 distinct examples of each TFBS for >90% of yeast TFs, with orders of magnitude more examples for most TFs (Fig. 1a).

Next, we experimentally demonstrated that random DNA yields diverse expression levels in a yeast promoter library. To robustly quantify promoter activity, we used a previously described[7] episomal dual reporter system expressing a constitutive red fluorescent protein (RFP) and a variable yellow fluorescent protein (YFP). log(YFP/RFP) measured using flow cytometry (see Methods) reports an expression signal that is integrated over several generations and normalized for extrinsic noise (for example, plasmid copy number and cell size)[3,19,20] (Fig. 1b and see Methods). We created ten synthetic promoter scaffolds and one based on the native *ANP1* promoter sequence, each consisting of 50–80 bp of constant scaffold sequence flanking 80 bp of random DNA (−170 to −90, relative to the presumed transcription start site (TSS); Fig. 1c, Supplementary Fig. 1a and see Methods). In all instances, the random 80-mer libraries yielded diverse expression levels, up to ~50-fold expression range, while individual promoter clones yielded distinct expression levels (Fig. 1c and Supplementary Fig. 1a, left). When we also randomized the scaffold sequences (from −289 to −25, relative to the TSS; see Methods), ~83% of random promoter sequences yielded measurable expression (Supplementary Fig. 1b). Thus, random DNA frequently contains functional TFBSs and can modulate a range of gene expression.

A 'gigantic' parallel reporter assay of random DNA. We implemented GPRA as a robust assay that quantifies the promoter activity of tens of millions of sequences per experiment. To facilitate validation, we tested rich media growth conditions with different carbon sources (glucose, galactose and glycerol; see Methods), where regulation is well studied. We created libraries of ~10^8 random promoters, transformed them into yeast and sorted the cells by log(YFP/RFP) into 18 bins of equal intervals (Fig. 2a and see Methods). We re-grew the yeasts from each bin, and measured their expression distributions by flow cytometry, reproducing the original expression measurement (Fig. 2b and see Methods). We sequenced the promoters in each bin and estimated each promoter's expression level by its read distribution (see Methods). Because the complexity of each promoter library (>10^8) was greater than the number of sorted cells (<10^8), 78% of promoter sequences appear in only one bin, often representing one observation (read) from one cell containing that promoter. While this leads to ~24% error in our promoter expression estimates, as assessed on held-out test data (Supplementary Fig. 2), the many more examples produced with this approach outweigh this challenge, and yield highly informative data from which to learn rules of *cis*-regulation, as we show below.

Altogether, across five experiments, we measured the expression output of 102,371,025 promoter sequences with GPRA. These spanned two primary promoter libraries (each complexity >10^8) containing a random 80-mer with either: (1) an upstream poly-T sequence and downstream poly-A sequence (pTpA; Fig. 1c); or (2) an upstream Abf1 site and a downstream TATA box (Abf1TATA; Supplementary Fig. 1a). We assayed both libraries with glucose as a carbon source, and the pTpA library also with either galactose or glycerol as alternate carbon sources. We sequenced 15–31 million

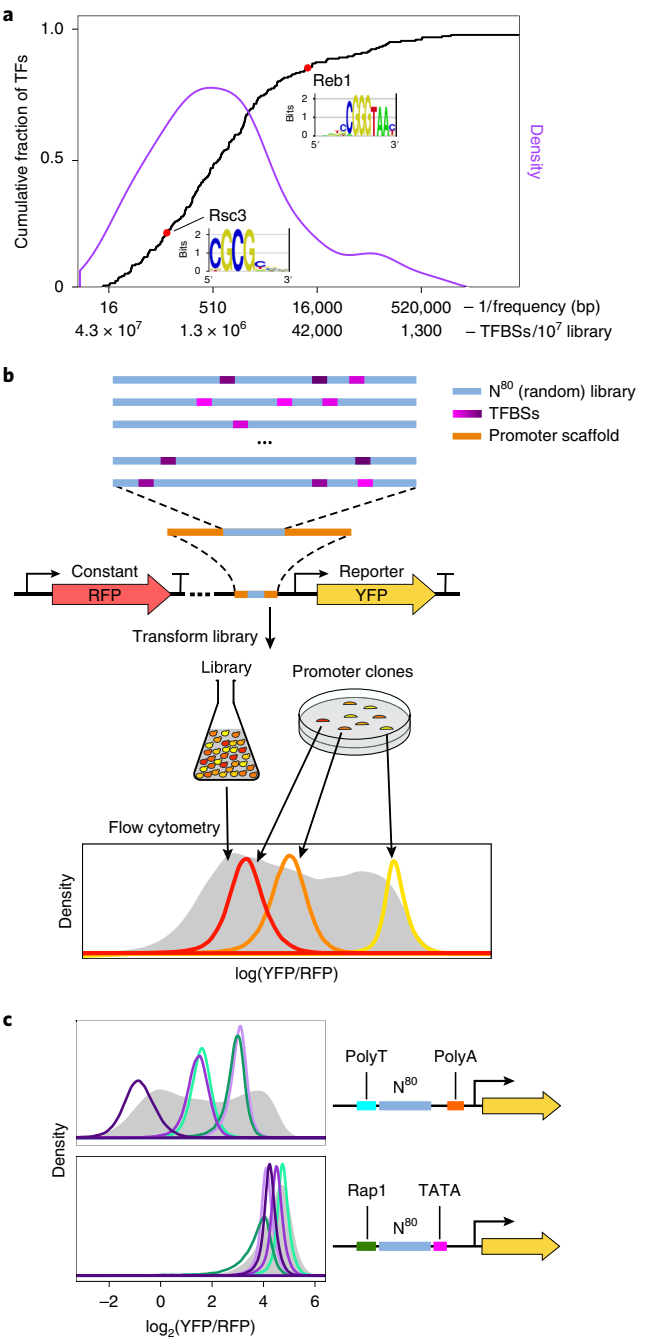

Fig. 1 | GPRA. a, TFBSs are common in random DNA. Cumulative distribution function (black) and density (purple) of the expected frequency of yeast TF motifs in random DNA. The expected number of TFBSs in a library of 10^7 random 80-bp promoters corresponding to each frequency is also indicated on the *x* axis. For instance, the relatively high IC (IC = 14.59) yeast Reb1 motif is expected to occur on average once every ~12,000 bp in random DNA, while Rsc3 (IC = 7.78) should occur every ~110 bp. **b**, GPRA overview. From top, a library of random DNA sequences (N^{80} here, blue) is inserted within a promoter scaffold (orange) in front of a reporter (yellow arrow). By chance, the random sequences include many TFBSs (purple). When grown in yeast, the library would yield a broad distribution of expression levels (gray, bottom) as measured by flow cytometry, where each promoter clone would have a distinctive expression distribution (red, orange, yellow). **c**, Random DNA yields diverse expression levels. For each promoter scaffold (right) shown are the expression distributions measured by flow cytometry (left) for the entire library (gray filled curves) and for a few selected clones, each from a different single promoter from each library (colored line curves).

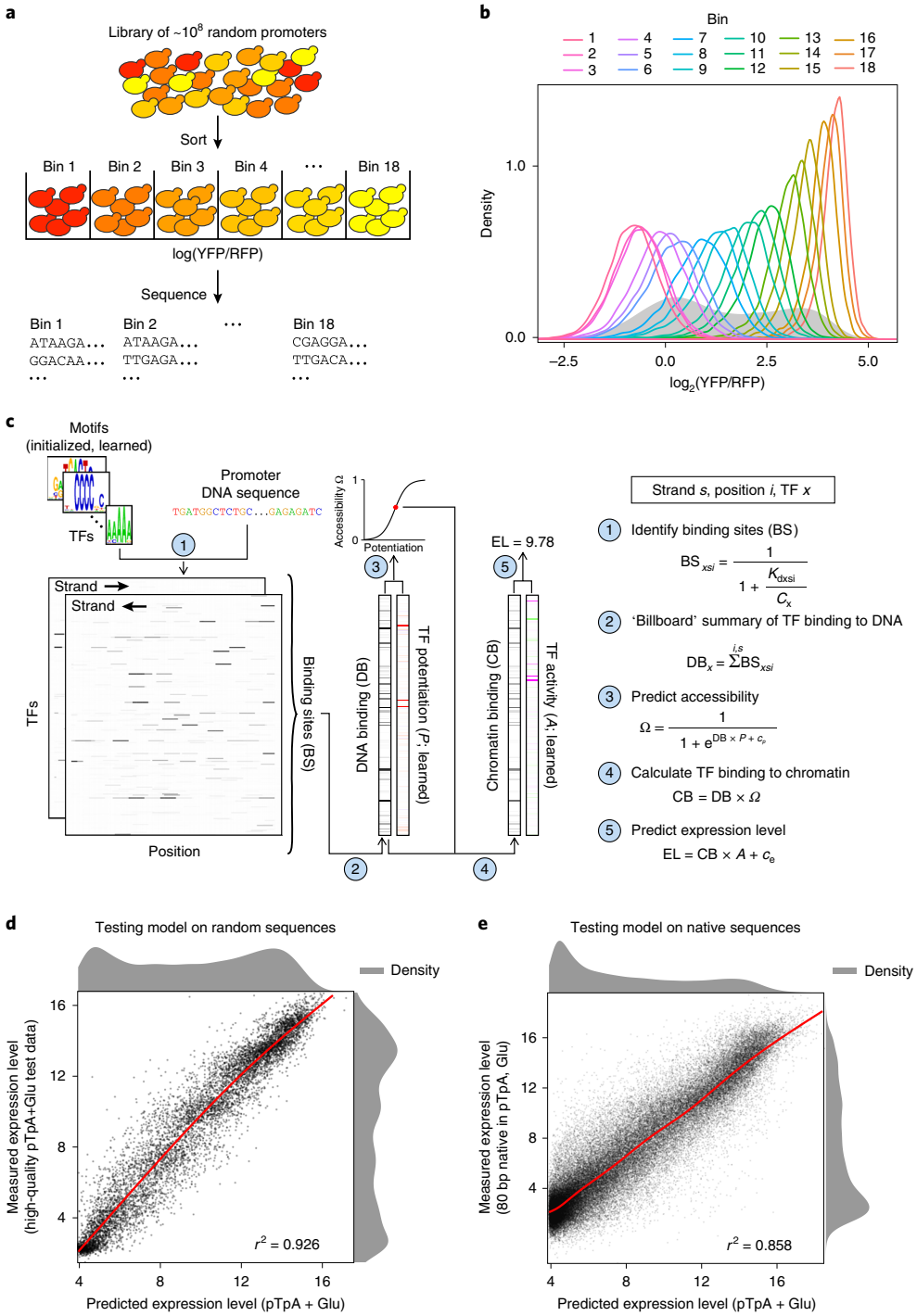

Fig. 2 | Expression models learned from a GPRA of 10^8 random promoters are highly predictive. a, Experimental strategy. Yeast GPRA library is sorted into 18 bins by the YFP/RFP ratio of the reporter (top) and the GPRA promoters in each bin are sequenced (bottom). **b**, Reproducibility of expression levels. Expression distributions ($\log_2$(YFP/RFP)) for cells from each bin (color code, top), after sorting as in **a**, which were regrown and re-assayed by flow cytometry. Expression distribution maintains the initial bin ranking. **c**, Computational 'billboard' model. Shown is a real example of the pTpA + glucose model predicting expression on a real DNA sequence (binding sites are smoothed over 8 bp for visualization purposes). Left, the model first scans each promoter DNA sequence with each PWM motif (1) to estimate a K_d for each TF at each strand and position (K_{dxsi}) and, through Michaelis–Menten binding using a learned concentration parameter (C_x), it estimates TF occupancy for every position and DNA strand. Next (2), it sums across positions and strands to estimate a single DNA binding amount per TF. Middle, the model learns a potentiation value for each TF (3), which, by pairwise multiplication with the estimated DNA binding and addition of a bias term (c_p), is used to infer the accessibility of each DNA sequence (Ω). The DNA binding vector is re-scaled (4) by the accessibility to estimate TF binding in chromatin. Right, chromatin binding is pairwise multiplied by learned activity parameters (5), capturing how the binding of each TF alters expression, and summed, including a bias term (c_e), to yield an estimated expression level for the promoter. **d**,**e**, Accurate prediction of expression from new random DNA and native yeast promoter sequences. Model-predicted expression (EL; pTpA + Glu; *x* axis) versus actual expression level (*y* axis; log(YFP/RFP) sorting bins) for high-quality random 80-bp test data in the pTpA promoter scaffold, grown in glucose (**d**), and native yeast promoter sequences, divided into 80-bp fragments and tested in the pTpA promoter scaffold, grown in glucose (**e**) ($n = 9{,}982$ and 70,924 promoters for **d** and **e**, respectively). Pearson's r^2 shown at bottom right. Red lines, generalized additive model lines of best fit.

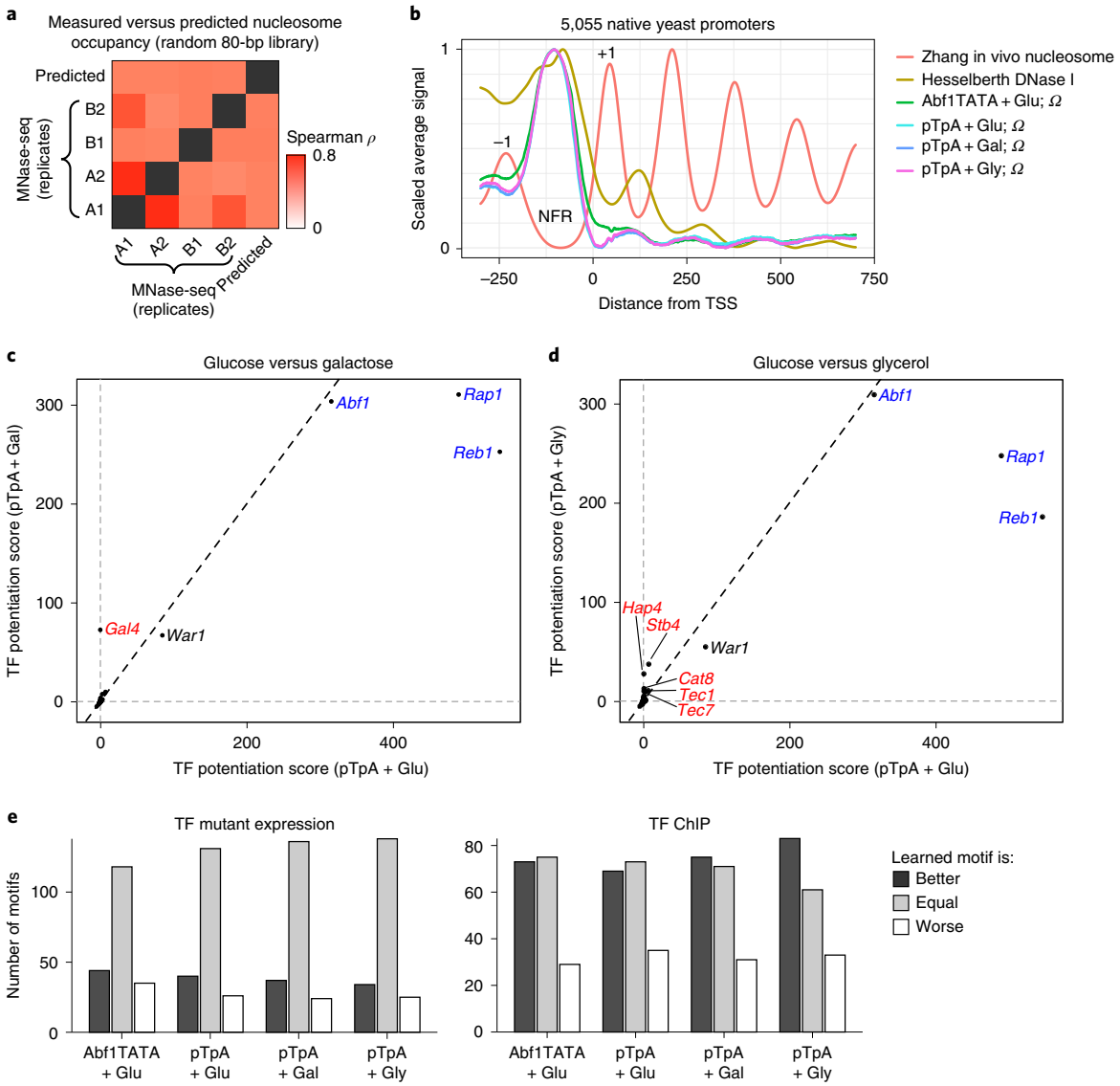

Fig. 3 | Billboard models learn biochemical activities of TFs. a,**b**, Model correctly predicts chromatin accessibility. **a**, Pairwise Spearman correlations (color) between model-predicted nucleosome occupancy ($1-\Omega$) and in vivo nucleosome occupancy measured by MNase-seq ($n=4$ biological replicates of $n=2$ independent library subsets). **b**, Average in vivo nucleosome occupancy[26] (Zhang), DNase I hypersensitivity[27] (representing accessibility; Hesselberth) and model-predicted accessibility (Ω) for each of the four billboard models surrounding the TSS. Each dataset is scaled. +1 and −1 nucleosome positions, and promoter nucleosome-free region, are indicated. **c**,**d**, TFs with predicted chromatin-opening ability. Shown is the predicted chromatin opening (potentiation) ability for each TF (dot) for pTpA models trained in glucose (*x* axes) versus either galactose (**c**) or glycerol (**d**) (*y* axes). Blue, GRFs with known chromatin-opening ability in all conditions; red, known and putative carbon source-specific regulators. **e**, Models improve TF motifs. The number of TFBS motifs (*y* axis) for which the model-refined motif predicted gene expression changes (TF mutant, left) or TF binding (ChIP, right) are better (dark gray), worse (white) or equal (light gray) to the original motifs, for each of the four models (*x* axis), where 'better' and 'worse' motifs are reproducibly so in at least 95% of random subsamples of the data (see Methods).

unique promoter sequences per experiment (<30% of sorted cells; <21% of promoter sequences theoretically in each library) and 50–155 million reads per experiment, and did not reach saturation (Supplementary Fig. 3).

TF-specific effects were captured well by GPRA. Even though each specific promoter sequence is typically associated with a single observed read, aggregating signal across the library revealed relationships between binding strength and observed expression. For each yeast TF, we used position weight matrices (PWMs)[18] to predict its occupancy of each promoter sequence[21]. Some TFs had a strong effect on expression, but explained only a small percentage of overall expression variation (for example, Abf1, a relatively rare motif in random DNA; Supplementary Fig. 4a, left; Pearson's correlation coefficient $r=0.10$). Others, including many zinc cluster monomeric motifs, correlated very strongly with expression (for example, Rsc30 $r=0.57$; Supplementary Fig. 4a, middle). Overall, the sum of the individual motif effects (348%) is much greater than what a simple linear model combining the motifs can explain (~47% of held-out training data; Pearson's r^2), suggesting that there is substantial redundancy between motifs. Moreover, cases where related motifs have distinct behaviors (for example, Rsc30 and Ume6; Supplementary Fig. 4a) further highlight the need to jointly analyze TFs.

A highly predictive 'billboard' model of *cis*-regulation. As a more faithful joint model of TF activity, we pursued an interpretable

'billboard' model[22] that captures the independent actions of all TFs, but does not model their positions or pairwise interactions (Fig. 2c). This model linearly relates TF occupancy to expression[23,24], as well as capturing interactions between TFs and nucleosomes (Fig. 2c and see Methods). Since nucleosomes can prevent TF binding[25], the model aims to infer promoter accessibility, which is used to scale the predicted occupancy of each TF (for example, a good TFBS will remain unbound if inaccessible). However, some TFs can displace nucleosomes, indirectly modulating the binding of other TFs. This potentiation can be learned from cases where a TFBS alters expression only in the presence of another binding site that 'potentiates' the first. Since we assume potentiation is driven primarily by chromatin opening, we model potentiating TFs as contributing to a global 'accessibility' value (Ω) for each promoter sequence. We scale TF binding with accessibility to reflect binding in the context of accessible chromatin, and calculate expression levels using these chromatin binding estimates and a linear model weighted by learned TF activities (Fig. 2c). To prevent TF functions from being apportioned amongst related motifs, we regularized the model to favor fewer and less information-rich motifs, and fewer potentiating and active TFs (see Methods). Once these parameters are learned, we also refine TF sequence specificities (see Methods).

When trained on our GPRA data, these models explained up to 92.6% of expression variation in independent, high-quality test data (Fig. 2d). We learned a separate model for each of the four high-complexity promoter datasets: pTpA in glucose, galactose and glycerol, and Abf1TATA in glucose. We tested each model's ability to predict expression in an independent set of ~10,000 pTpA promoters, measured with high coverage in glucose. On these high-quality test data, the pTpA + glucose model predicted expression best ($r^2 = 0.926$; Fig. 2d), but the galactose- and glycerol-trained pTpA models performed nearly as well ($r^2 = 0.904$ and 0.843, respectively). This indicates that the primary contributors to gene expression in the context of random DNA are not regulated by carbon source. As further validation, we generated 1,000 random sequences that were in silico predicted by the pTpA + glucose model to have a range of expression levels, synthesized these sequences and measured their expression, showing strong agreement between prediction and measurements across a ~50-fold range ($r^2 = 0.897$; Supplementary Fig. 4b). Overall, a remarkably high proportion of the variation in random promoter expression is explained by a billboard model.

Moreover, our models trained on random DNA data from GPRA predicted over 85% of the variation in expression driven by sequence fragments derived from native yeast promoters (Fig. 2e). To this end, we segmented each yeast promoter into 80-bp fragments from −480 to the TSS, and assayed these in the pTpA promoter scaffold in glucose media. Surprisingly, random DNA included more high-expressing sequences than most of the native promoter fragments tested (Fig. 2d,e), except for sequences from the −120 to −40 and −160 to −80 regions (Supplementary Fig. 5). The pTpA + glucose billboard model, which was trained on random DNA, predicted the expression of these native yeast sequences with high correlation (Pearson $r^2 = 0.858$, Fig. 2e). This shows the power of models trained on random DNA and indicates that nonbillboard regulatory mechanisms are either not predominant in yeast promoters, or are context-dependent.

Billboard model correctly learns TF's biochemical features. Because our models are biologically interpretable ('white box'), we could next assess mechanistic features, such as TF function or chromatin organization, that underlie their predictions. For example, our models, trained on expression levels of random DNA, also accurately predict chromatin accessibility in the libraries themselves and the yeast genome. First, the model predicted the experimentally measured nucleosome occupancy in the libraries (Sequencing of micrococcal nuclease digested DNA, MNase-seq; see Methods; Spearman's correlation coefficient (ρ) = 0.54–0.55) comparably to the agreement between experimental replicates (Fig. 3a and Supplementary Fig. 6a,b). Moreover, the pattern of model-predicted accessibility when applied to the yeast genome sequence agrees well with previously measured in vivo nucleosome occupancy[26,27] (Supplementary Fig. 6c). On average, the models accurately predict the promoter nucleosome-free region, and −1 and +1 nucleosomes (Fig. 3b and see Methods). Thus, random sequences and expression measurements generated by GPRA are of sufficient quality to correctly infer how TFs regulate chromatin structure, without directly measuring chromatin.

The models also accurately captured biochemical TF activities, including chromatin remodeling and activator versus repressor function. The general regulatory factors (GRFs; Abf1, Reb1 and Rap1), which can displace nucleosomes[28–31], were predicted to open chromatin (positive potentiation scores) in all conditions tested (Fig. 3c,d). Moreover, the galactose-specific regulator Gal4 was correctly[32,33] predicted to open chromatin only in galactose (Fig. 3c). TFs predicted to open chromatin only in glycerol included Hap4, Stb4, Cat8, Tec1 and Tye7 (Fig. 3d). These molecular roles are new predictions, but with strong physiological support: both Hap4 and Cat8 are over-expressed in glycerol compared with glucose[34]; Hap4 is a global regulator of nonfermentative media such as glycerol[35]; Cat8 activates gluconeogenesis[36,37] and Tye7 regulates glycolysis[38], which are the two endpoints of glycerol metabolism[39]; Tec1 regulates pseudohyphal growth[40,41], which is constitutive in glycerol[42]; and predicted Stb4 targets are enriched for having 'oxidoreductase activity'[18], consistent with nonfermentable carbon source metabolism. Furthermore, Hap4 and Tec1 physically interact with the Swi/Snf chromatin remodeler[43,44], supporting their putative chromatin-remodeling role. For glucose-trained models, model-predicted TF activities weakly agreed with gene ontology (GO)-annotated activator/repressor status (Supplementary Fig. 7a; one-tailed hypergeometric P values, 0.02 and 0.04), while there was no association for either galactose ($P = 0.34$) or glycerol ($P = 0.79$). The lack of distinction in GO annotations between activation by opening chromatin and activation by other means could explain this weak agreement. Consistent with open chromatin being more active, potentiation scores (model-predicted chromatin opening and closing ability) significantly distinguished GO-annotated activators and repressors for all models (see Methods; one-tailed hypergeometric P values, 10^{-3} to 2×10^{-5}; Supplementary Fig. 7b). Thus, random sequences contain sufficient TFBSs to identify how TFs affect gene expression and chromatin, even for relatively rare motifs (for example, GRF motifs).

Furthermore, we could learn the specificities and activities of TFs without any TFBS training data. By initializing with random PWMs and learning the motifs de novo, we trained a model (pTpA + glucose data) that was highly predictive (Pearson's $r^2 = 94.6\%$; see Methods). This model—with >120,000 parameters—learned many motifs that closely resemble those of known factors (for example, the most potent chromatin-opening motifs closely resemble the GRFs; Supplementary Fig. 8a). However, it is more difficult to interpret this model, since the identities of the TFs recognizing each learned motif are unknown, and each TF may be represented by multiple motif variations or not at all.

Consequently, we allowed the models initialized with known motifs to optimize the TF motifs, yielding an improved model with refined motifs that better predict independent data. In particular, motif refinement (including adding more bases of specificity) improved the models' predictive accuracy on the independent high-quality test data by 9–12 percentage points (for example, from $r^2 = 80.3\%$ to 92.2% for pTpA + glucose). The four models often modified the original motifs in similar ways, suggesting that the revised motifs more faithfully represent the TFs' specificities (Supplementary Fig. 8b). Many of the refined motifs performed

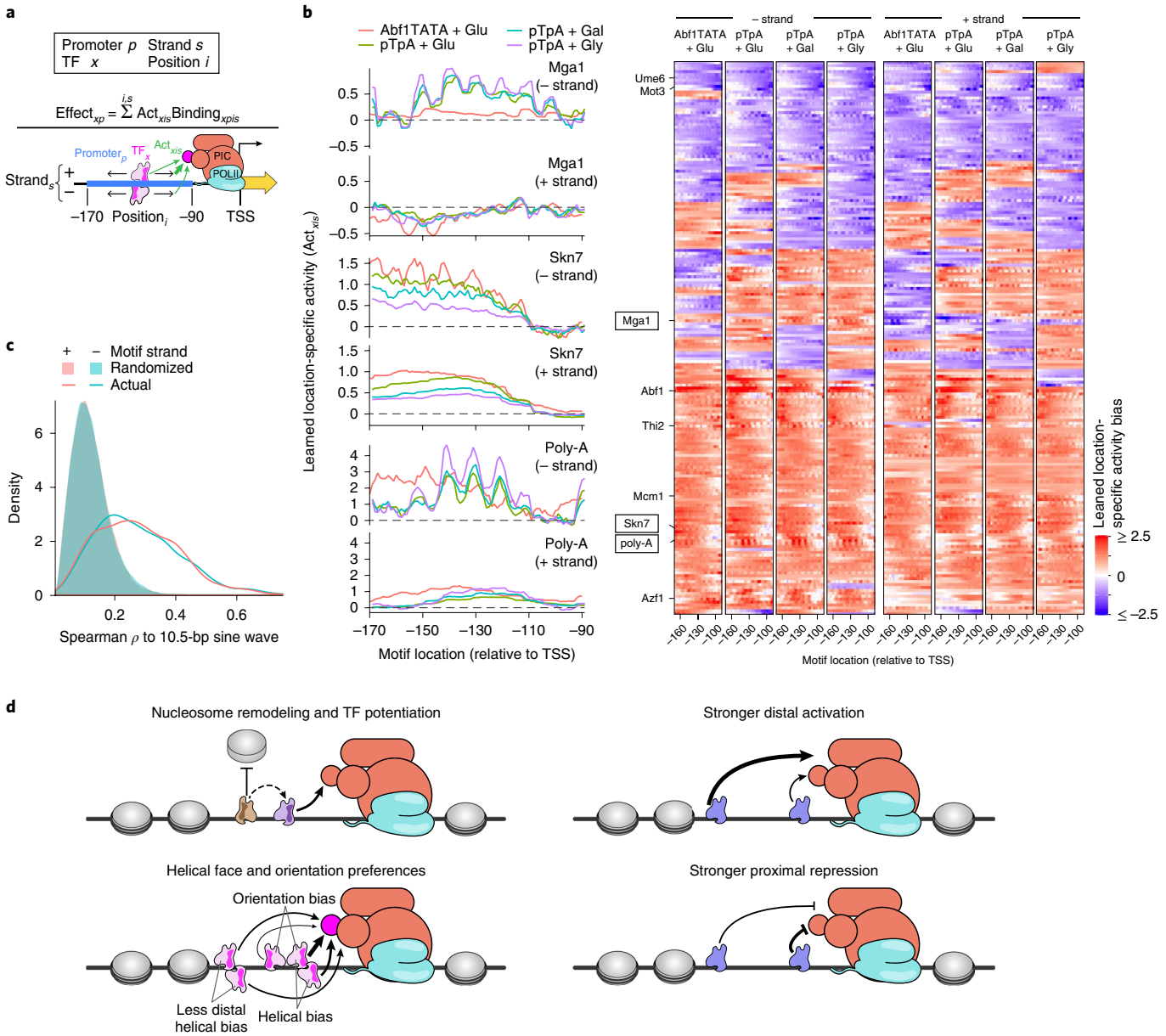

Fig. 4 | Position, orientation and helical face preferences among yeast TFs. a, Model with position- and orientation-specific activities. For each TF (*x*), the model learns parameters for how much binding site position (*i*) and strand (*s*) within the promoter affect transcriptional activity (Act_{xis}). The total effect of a TF ($Effect_{xp}$) is thus the sum of products of the position-specific activities (Act_{xis}) and TF occupancies ($Binding_{xpis}$) at the promoter (*p*), across all positions and both strands. For example, this could reflect the TF's ability to contact the transcriptional pre-initiation complex (PIC). **b**, Motif position and orientation effects on expression. Left, each plot shows the learned activity parameter values (*y* axis) for motifs in each position (*x* axis) and strand orientation (upper and lower panels) for each model (colors). Right, position-specific activity biases (color) for each TF (rows) at each position (columns) for minus (left half) and plus (right half) strand orientations for each of the four models (four subpanels). Only TFs for which all models retained the motif are shown. **c**, Helical face preferences. Distribution of Spearman's ρ between a 10.5-bp sine wave and the learned position-specific activity weights (as in Supplementary Fig. 13a) for plus strand (pink line) and minus strand (blue line) or with corresponding randomized data (pink and blue shaded areas) for all four models. **d**, Model of *cis*-regulatory logic. TFs display a variety of activity types. Some TFs potentiate the activity of other TFs by modulating nucleosome occupancy (upper left). Activators tend to have a greater effect on transcription when bound distally within the promoter (upper right), while repressors have the greatest effect when bound proximally (lower right). Many TFs show differential activity depending on the helical face or orientation of the TFBS, presumably through interaction with other factors bound nearby (lower left).

better than the originals in predicting, from DNA sequence alone, in vivo genomic binding of the cognate TF by chromatin immunoprecipitation (ChIP)[45], and gene expression changes resulting from cognate TF perturbation[46] (Fig. 3e, Supplementary Fig. 8c,d and see Methods). Many motifs were indistinguishable from the originals, suggesting that they maintained their cognate TFs. Of those that differed, the vast majority were improved (Fig. 3e), including when predicting ChIP data, despite many of the original motifs being derived from this same ChIP data[18]. This suggests that the refined motifs often more closely represent their cognate TF specificities.

Binding position, strand and helical face alter TF activity. We next adapted the model to capture how transcriptional activity is altered by TFBS position. Motif position and orientation can affect

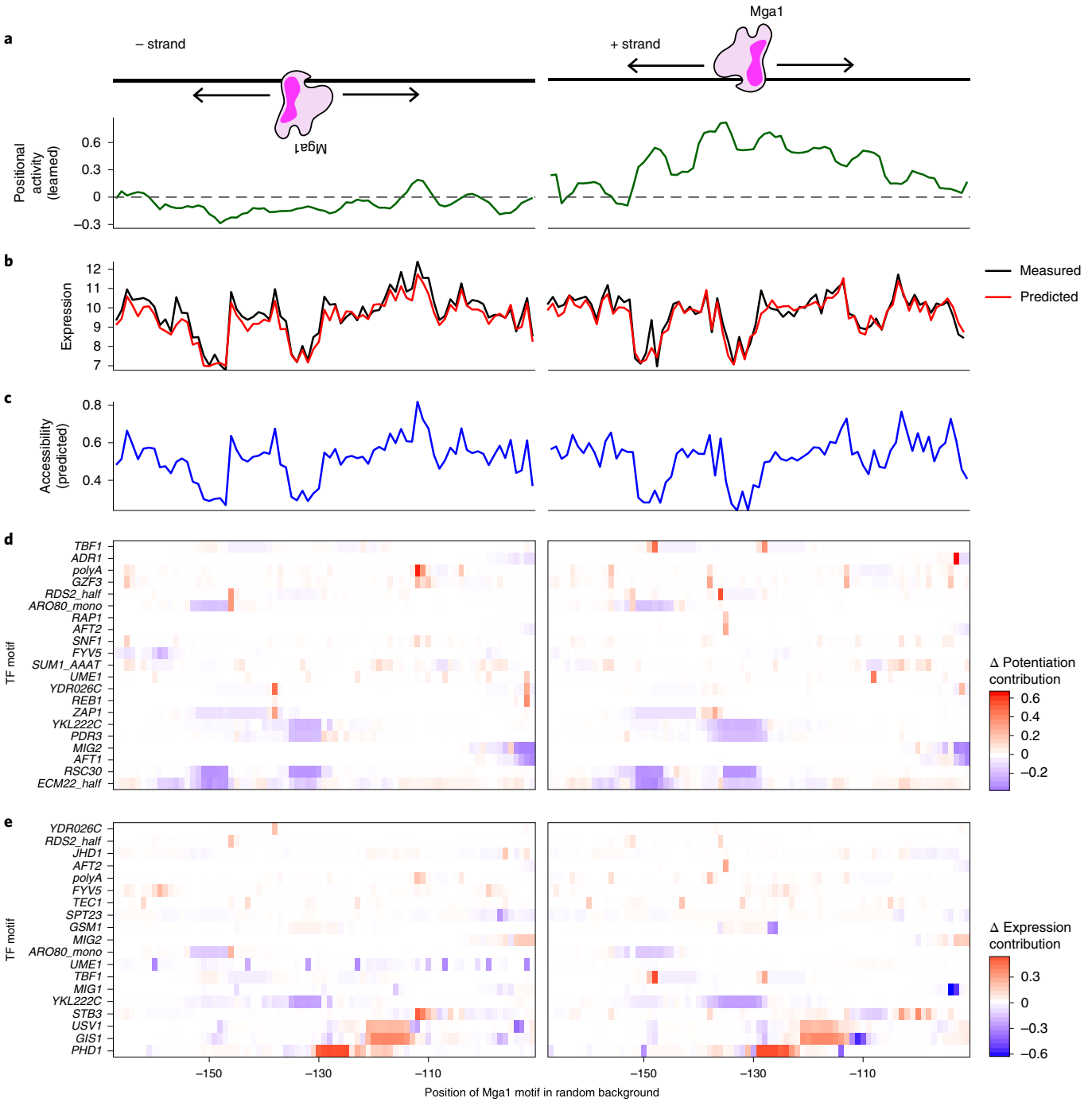

Fig. 5 | Inadvertent perturbation of abundant secondary TFBSs confounds TFBS tiling experiments. a–**e**, Mga1 motifs were inserted into a common background sequence at every possible position (common *x* axis) for both the −strand (left) and +strand (right). **a**, Position-specific activity parameters (*y* axis) learned for the Mga1 motif by the pTpA + glucose model (that is, how the Mga1 motif alters expression based on the location of its binding site). **b**, Model correctly predicts expression despite little correspondence to the position-specific activity of the Mga1 motif. Measured (black) and predicted (red) expression levels for Mga1 motif-tiling sequences. **c**, Most expression differences between sequences are attributed to changes in accessibility. Predicted accessibility (*Ω*; *y* axis) for Mga1 motif-tiling sequences. **d**,**e**, Expression changes are explained by perturbation of prevalent TFBSs when tiling the motif. Changes in potentiation score (**d**) and expression (**e**) attributable to perturbed TF binding for numerous diverse factors (rows) when tiling the Mga1 motif at each position (*x* axis). The dissimilarity between the rows indicates minimal redundancy between factors.

TF function, for instance, by modifying the TF's ability to contact its biochemical target. We thus extended the billboard model with localized activity bias terms (see Methods), allowing TFs to have different activities for every binding position and orientation (Fig. 4a). To encourage parameter parsimony, we added a regularization term that favored no positional preferences (see Methods). This model had up to ~220 activity parameters per TF instead of one in the billboard model, ~110 locations (including flanking constant regions) and two DNA strands (Supplementary Fig. 9), adding ~55,000 parameters overall. Fitting such complex models with data of more traditional scale would be unreliable, but the examples in our dataset still outnumber parameters ~360:1. Subsampling

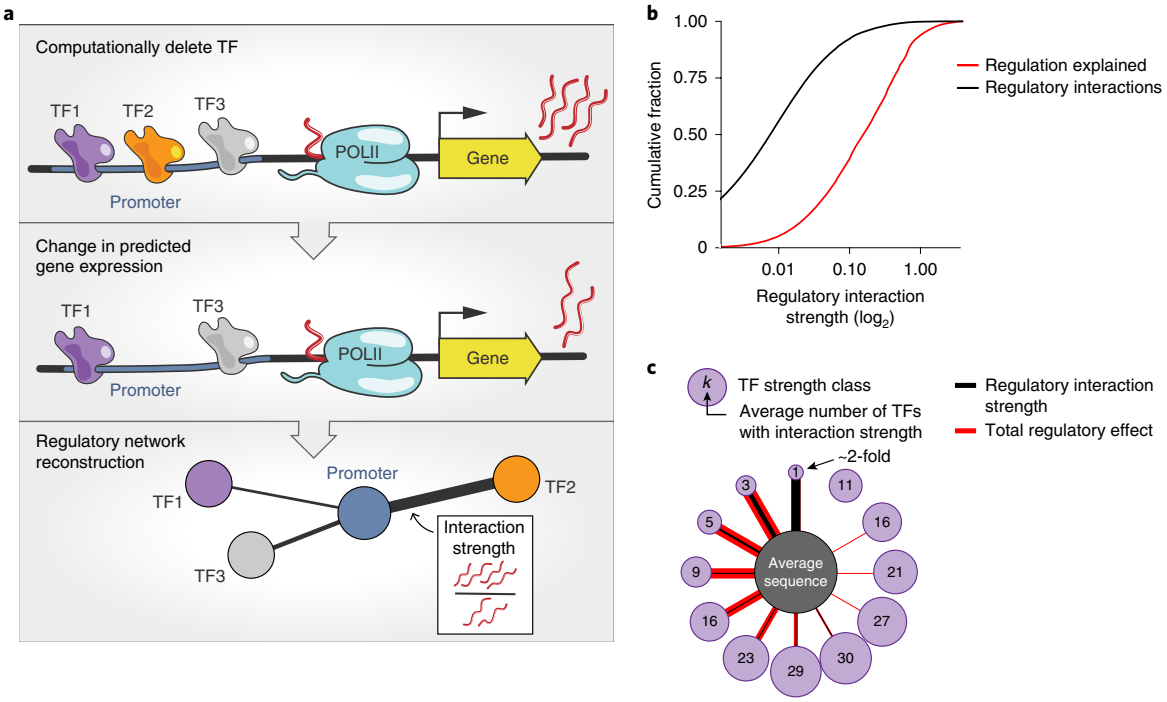

Fig. 6 | Abundant weak regulatory interactions explain most of expression level. a, Analysis overview. A computational 'TF knock-out experiment' is performed with the learned *cis*-regulatory model for each TF: we use the complete model (pTpA + Glu positional; top) and that model with that TF 'deleted' (setting its concentration parameter to 0; middle) to predict expression for each 80-bp fragment of native yeast promoter DNA. Bottom, the resulting difference in predicted expression is used to define a regulatory interaction strength (edge) between that TF and DNA sequence; these are used to build regulatory networks for all sequences and TFs. **b**,**c**, Aggregation of weak regulatory effects contributes more to expression than strong interactions. **b**, Cumulative distributions (*y* axis) of the number of regulatory interactions (black) and fraction of regulation explained (that is, fraction of the cumulative sum of all interaction strengths; red) for each regulatory interaction strength (*x* axis). The magnitude (and not the sign) of the interaction strength is considered. Because the *y* axis is scaled to 1, this is equivalent to the average distribution across all native sequence fragments. **c**, Regulatory interaction network summary for an 'average' sequence. Regulatory interactions were grouped by the strength of the regulatory interaction (thickness of black edges) into different strength classes (purple nodes), with the average number of TFs in that class indicated in the circle. The overall effect on expression, accounting for all TFs in each regulatory interaction strength class, is indicated in red (thickness of red edges). Although there are >2-fold regulatory interactions, these are too rare to be shown here (<1 per sequence).

analysis suggests that we minimally require millions of random sequences to learn these parameters without over-fitting, and additional data improve performance (Supplementary Fig. 10).

Capturing positional preferences significantly increased performance. Adding positional activities to the pTpA + glucose model decreased the error by ~20% for both the high-quality test data (94.3% versus 92.6%) and the 80-bp native promoter sequences (88.6% versus 85.8%; $P < 10^{-21}$ and 10^{-107}, respectively, Fisher's r to z transformation; Supplementary Fig. 11a,b). Predicted accessibility, which cannot be impacted by TFBS location (Supplementary Fig. 9), remained a dominant factor, explaining 90.8% of expression variation (high-quality test data; Pearson's r^2; Supplementary Fig. 11c). Adding positional activities decreased the prediction error twofold more than nonpositional activities (38% versus 19.6%), highlighting their importance.

The parameters learned by our model indicated that many TFs have strong position, orientation and helical face preferences, and the similarity between different models suggested that they are robustly learned (Fig. 4b and Supplementary Fig. 12). Predicted activators are often stronger when located distally to the TSS (for example, Abf1, Skn7 and Mcm1; Fig. 4b and Supplementary Fig. 12a,b), while many predicted repressors are most repressive when located proximally (for example, Ume6 and Mot3; Supplementary Fig. 12c,d). Many TFBSs are strand-specific, often with a lower-than-expected distal activity for one motif orientation (for example, Azf1, Mga1 and Thi2; Fig. 4b and Supplementary Fig. 12e,f). Rarely, TFBSs can be both activating and repressing in different positions (for example, Mga1 in minus versus plus strand; Fig. 4b).

Some TFBSs showed strong periodicity along the promoter's length (for example, Mcm1, Thi2, poly-A and Azf1; Fig. 4b and Supplementary Fig. 12b,e,f), consistent with DNA helical face preferences. This was widespread: the correlations between a 10.5-bp sine wave and the learned positional biases were significantly higher than with randomized data for each model (rank sum $P < 10^{-120}$; AUROC = 0.84–0.87; Fig. 4c, Supplementary Fig. 13a and see Methods). Helical preferences tend to be strongest when TFBSs are proximal to the TSS (downstream of −150, relative the TSS). Since 150 bp is the approximate persistence length of double-stranded DNA[47], this could reflect physical promoter constraints, where the rigidity of DNA prevents interactions between proximal TFs and the adjacent transcriptional machinery, but flexibility increases with distance, relieving this effect. Models trained on different scaffolds sometimes learned distinct positional parameters (for example, Mga1, Skn7, poly-A and Mcm1; Fig. 4b and Supplementary Fig. 12b), suggesting that the surrounding context can modify positional preferences. Adding positional biases sometimes worsened the models' ability to generalize between scaffolds, but always improved performance within a scaffold in another condition (Supplementary Fig. 13b). Overall, many TFs are predicted to have strong positional preferences (Fig. 4d).

Prevalent weak regulatory interactions explain expression. Finally, we leveraged our interpretable position-aware model to revisit an open question from previous studies that examined the impact of placing the same motif in different positions. A seminal study from Sharon et al.[7] attempted to identify positional preferences of TFs by tiling each TFBS across one or few background sequences in a massively parallel reporter assay, but found that, with few exceptions, expression changes were largely inexplicable, depending on both the embedded motif and the background sequences[7]. We therefore replicated this by tiling each of six motifs with strong model-predicted positional preferences (Azf1, Mga1, Mot3, Skn7, Ume6 and the poly-A motif) at base-pair resolution in either orientation within three random sequences predicted to have intermediate expression levels (Fig. 5a) and measured expression using sorting and sequencing (Fig. 2a).

Our model predicted the measured expression levels well (Pearson $r^2 = 0.919$; positional pTpA + glucose model, all motifs and background sequences; Fig. 5b and Supplementary Fig. 14), with accessibility, again, predicted to play a dominant role (Fig. 5c). In spite of this, expression showed few obvious trends across embedding contexts, motifs or motif positions (Fig. 5b and Supplementary Fig. 14), consistent with Sharon et al.[7]. There was also no clear relationship between the expression level resulting from a motif's embedded position (Fig. 5b) and the corresponding model-learned motif activity (Fig. 5a). Instead, deeper inspection revealed that, according to the model, most changes in expression result from the destruction and creation of many secondary TFBSs as each motif is tiled (Fig. 5d,e), with the effect of the tiled motif overwhelmed by these abundant secondarily perturbed TFBSs. This highlights how seemingly complex regulation can result from prevalent TFBSs and a simple *cis*-regulatory logic, rather than underlying complex mechanisms.

Following this observation, we queried our positional *cis*-regulatory model to assess how many TFs regulate a given native promoter. We quantified each TF's contribution to the expression of each native promoter fragment by performing in silico TF 'deletion' experiments, setting the concentration parameter for each TF to 0, and inspecting the predicted expression change (Fig. 6a). Although the number of regulators per promoter sequence varied, most were regulated by a surprisingly large number of factors. Strong regulatory interactions were rare: only 0.1% of possible regulatory interactions were predicted to alter expression by ≥2-fold (Fig. 6b,c, black; for example, see Supplementary Fig. 15a,b). Although these rare strong regulatory interactions explained a disproportionate amount of expression, 94% of expression was attributed to the much more prevalent weak (<2-fold) regulatory interactions (Fig. 6b,c, red, and Supplementary Fig. 15c).

Discussion

We showed that measuring the expression output of random DNA sequences can provide data at a radically larger scale, surpassing the complexity of the human genome. This scale allows us to learn complex interpretable models with remarkable predictive power and determine the roles played by the cell's entire complement of TFs with a simple and inexpensive experiment. Using these data, we refined models of TF specificities, and identified activators, repressors, chromatin-remodeling TFs and condition-specific regulators. Most TFs have strand, location and helical face preferences (Fig. 4d), which can be modified by the surrounding sequence/chromatin context (Fig. 4b), demonstrating that *cis*-regulatory logic can be highly complex.

Transferring the rules learned from such reporter assays to arbitrary contexts will be a subject of future studies. We expect that more will be learned from our data: for example, a deep convolutional neural net trained on GPRA data explained ~96% of expression variation of our high-quality test data (E.D.V., unpublished results), a 30% reduction in error.

Using random DNA to study *cis*-regulatory logic in vivo is a highly accessible approach, which facilitates assaying massive libraries at unprecedented scale, and learning complex models with many parameters. While designed sequences can be used to test specific hypotheses, random DNA is useful for analyzing anything that occurs reasonably often by chance, even if uncommon in the genome or not anticipated in advance. For instance, one could ask how G-quadruplex motifs affect expression (we saw no effect; data not shown). Further, learning an element's effect from thousands of examples with diverse affinities, positions, orientations and surrounding sequence contexts is likely to be more generalizable than the 'designed' approach, where a few elements are introduced into several locations.

In particular, we show that the common 'controlled' experiment of modifying one particular *cis*-regulatory parameter (for example, TFBS location) is inadvertently confounded by introducing or destroying many secondary elements whose combined effect can mask the element being studied (Fig. 5). Furthermore, any trend observed with this designed strategy could be explained by the action of another TF with a related specificity. Such inadvertently included TFBSs may also confound sequences designed for engineering purposes. In contrast, random DNA provides the diverse examples needed to learn complex regulatory logic: jointly modeling the many variables that simultaneously affect expression can separate each variable's effect, and selecting a random DNA sequence with the desired predicted expression level provides an alternative for promoter design.

Our results suggest that regulatory networks are more interconnected than previously assumed (Fig. 6). We showed that random DNA has diverse expression levels (Fig. 1) that can be explained by TF binding (Fig. 2), which regulate expression primarily through weak interactions (Fig. 6) that, in turn, can easily be perturbed when tiling a motif across a sequence (Fig. 5). Although low-affinity TFBSs have been shown in aggregate to alter expression[48] and the prevalence of TFBSs was predicted by biochemistry and information theory[16], weak regulatory interactions have largely been ignored. Most studies focus on the strongest interactions that explain most gene expression variation. To explain expression levels we must also account for these abundant weak effects, which, individually, are likely easily masked by experimental noise and secondary effects when studying endogenous gene regulation. Regulatory variants may contribute to phenotype by cascades of regulatory changes through highly interconnected networks[49,50], and abundant weak regulatory interactions suggest a mechanism for this interconnectedness. Although a highly predictive yeast model underlies our results, the human genome encodes more similarly low-IC TFs and has more regulatory DNA (promoters and enhancers), providing more opportunity for weak interactions.

The prevalence of functional TFBSs in random DNA and its demonstrated ability to modulate expression has evolutionary implications. In some cases when genes are created, the DNA-encoded regulatory program must arise de novo. Random sequences have been shown to yield functioning bacterial promoters ~10% of the time[15]. In yeast, we found ~83% of promoter sequences with both random scaffold and insert expressed. Therefore, evolving regulatory sequences from previously nonregulatory DNA may be comparatively straightforward. Creating new mammalian enhancers may be similarly likely since mammalian TFs have, on average, even less specificity than in yeast[16]. Over evolutionary time, further mutations can optimize the specificity and effect of these novel regulatory sequences.

When using GPRA, researchers will have to consider the scale needed for their question of interest. Since TFBSs occur with different frequencies (Fig. 1a), more data are needed for rare TFBSs. The activity and potentiation parameters for each TF converged with ~100,000 promoter examples (Supplementary Fig. 10). Conversely,

millions of promoter examples were required for refining or learning new motifs, and for finding position- and orientation-specific activities (Supplementary Fig. 10). Since arbitrary pairs of specific TFBSs are inherently rare in random DNA, learning all possible TF–TF interactions with GPRA, especially when considering competition (where both binding sites must be high-affinity), may require much larger datasets. Although mammalian gene regulation is more complex, GPRA could provide the 'big data' that would allow learning models to explain how genetic variation affects gene expression and disease risk.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information, details of author contributions and competing interests, and statements of code and data availability are available at https://doi.org/10.1038/s41587-019-0315-8.

Received: 24 January 2019; Accepted: 16 October 2019;
Published online: 2 December 2019

References

1. Beer, M. A. & Tavazoie, S. Predicting gene expression from sequence. *Cell* **117**, 185–198 (2004).
2. Yuan, Y., Guo, L., Shen, L. & Liu, J. S. Predicting gene expression from sequence: a reexamination. *PLoS Comput. Biol.* **3**, e243 (2007).
3. Kinney, J. B., Murugan, A., Callan, C. G. Jr. & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl Acad. Sci. USA***107**, 9158–9163 (2010).
4. van Arensbergen, J. et al. Genome-wide mapping of autonomous promoter activity in human cells. *Nat. Biotechnol.* **35**, 145–153 (2017).
5. Muerdter, F. et al. Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat. Methods* **15**, 141–149 (2018).
6. Wang, X. et al. High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat. Commun.* **9**, 5380 (2018).
7. Sharon, E. et al. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30**, 521–530 (2012).
8. Gertz, J., Siggia, E. D. & Cohen, B. A. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature* **457**, 215–218 (2009).
9. Hughes, T. R. & de Boer, C. G. Mapping yeast transcriptional networks. *Genetics* **195**, 9–36 (2013).
10. Jolma, A. et al. DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013).
11. Nutiu, R. et al. Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nat. Biotechnol.* **29**, 659–664 (2011).
12. Oliphant, A. R., Brandl, C. J. & Struhl, K. Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Mol. Cell. Biol.* **9**, 2944–2949 (1989).
13. Horwitz, M. S. & Loeb, L. A. Promoters selected from random DNA sequences. *Proc. Natl Acad. Sci. USA* **83**, 7405–7409 (1986).
14. Cuperus, J. T. et al. Deep learning of the regulatory grammar of yeast 5' untranslated regions from 500,000 random sequences. *Genome Res.* **27**, 2015–2024 (2017).
15. Yona, A. H., Alm, E. J. & Gore, J. Random sequences rapidly evolve into de novo promoters. *Nat. Commun.* **9**, 1530 (2018).
16. Wunderlich, Z. & Mirny, L. A. Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet.* **25**, 434–440 (2009).
17. Arnosti, D. N. & Kulkarni, M. M. Transcriptional enhancers: intelligent enhanceosomes or flexible billboards? *J. Cell. Biochem.* **94**, 890–898 (2005).
18. de Boer, C. G. & Hughes, T. R. YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. *Nucleic Acids Res.* **40**, D169–D179 (2012).
19. Kosuri, S. et al. Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **110**, 14024–14029 (2013).
20. Shalem, O. et al. Systematic dissection of the sequence determinants of gene 3' end mediated expression control. *PLoS Genet.* **11**, e1005147 (2015).
21. Granek, J. A. & Clarke, N. D. Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biol.* **6**, R87 (2005).
22. Kulkarni, M. M. & Arnosti, D. N. Information display by transcriptional enhancers. *Development* **130**, 6569–6575 (2003).
23. Bussemaker, H. J., Li, H. & Siggia, E. D. Regulatory element detection using correlation with expression. *Nat. Genet.* **27**, 167–171 (2001).
24. Conlon, E. M., Liu, X. S., Lieb, J. D. & Liu, J. S. Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl Acad. Sci. USA* **100**, 3339–3344 (2003).
25. Liu, X., Lee, C. K., Granek, J. A., Clarke, N. D. & Lieb, J. D. Whole-genome comparison of Leu3 binding in vitro and in vivo reveals the importance of nucleosome occupancy in target site selection. *Genome Res.* **16**, 1517–1528 (2006).
26. Zhang, Z. et al. A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. *Science* **332**, 977–980 (2011).
27. Hesselberth, J. R. et al. Global mapping of protein–DNA interactions in vivo by digital genomic footprinting. *Nat. Methods* **6**, 283–289 (2009).
28. Bernstein, B. E., Liu, C. L., Humphrey, E. L., Perlstein, E. O. & Schreiber, S. L. Global nucleosome occupancy in yeast. *Genome Biol.* **5**, R62 (2004).
29. Hartley, P. D. & Madhani, H. D. Mechanisms that specify promoter nucleosome location and identity. *Cell* **137**, 445–458 (2009).
30. Ganapathi, M. et al. Extensive role of the general regulatory factors, Abf1 and Rap1, in determining genome-wide chromatin structure in budding yeast. *Nucleic Acids Res.* **39**, 2032–2044 (2011).
31. Levo, M. et al. Systematic investigation of transcription factor activity in the context of chromatin using massively parallel binding and expression assays. *Mol. Cell* **65**, 604–617 e606 (2017).
32. Axelrod, J. D., Reagan, M. S. & Majors, J. GAL4 disrupts a repressing nucleosome during activation of GAL1 transcription in vivo. *Genes Dev.* **7**, 857–869 (1993).
33. Morse, R. H. Nucleosome disruption by transcription factor binding in yeast. *Science* **262**, 1563–1566 (1993).
34. Roberts, G. G. & Hudson, A. P. Transcriptome profiling of *Saccharomyces cerevisiae* during a transition from fermentative to glycerol-based respiratory growth reveals extensive metabolic and structural remodeling. *Mol. Genet. Genomics* **276**, 170–186 (2006).
35. Forsburg, S. L. & Guarente, L. Identification and characterization of HAP4: a third component of the CCAAT-bound HAP2/HAP3 heteromer. *Genes Dev.* **3**, 1166–1178 (1989).
36. Hedges, D., Proft, M. & Entian, K. D. CAT8, a new zinc cluster-encoding gene necessary for derepression of gluconeogenic enzymes in the yeast *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **15**, 1915–1922 (1995).
37. Haurie, V. et al. The transcriptional activator Cat8p provides a major contribution to the reprogramming of carbon metabolism during the diauxic shift in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **276**, 76–85 (2001).
38. Sato, T. et al. TheE-box DNA binding protein Sgc1p suppresses the gcr2 mutation, which is involved in transcriptional activation of glycolytic genes in *Saccharomyces cerevisiae*. *FEBS Lett.* **463**, 307–311 (1999).
39. Grauslund, M. & Ronnow, B. Carbon source-dependent transcriptional regulation of the mitochondrial glycerol-3-phosphate dehydrogenase gene, GUT2, from *Saccharomyces cerevisiae*. *Can. J. Microbiol.* **46**, 1096–1100 (2000).
40. Madhani, H. D. & Fink, G. R. Combinatorial control required for the specificity of yeast MAPK signaling. *Science* **275**, 1314–1317 (1997).
41. Gavrias, V., Andrianopoulos, A., Gimeno, C. J. & Timberlake, W. E. *Saccharomyces cerevisiae* TEC1 is required for pseudohyphal growth. *Mol. Microbiol.* **19**, 1255–1263 (1996).
42. Cullen, P. J. & Sprague, G. F. Jr. Glucose depletion causes haploid invasive growth in yeast. *Proc. Natl Acad. Sci. USA* **97**, 13619–13624 (2000).
43. Neely, K. E., Hassan, A. H., Brown, C. E., Howe, L. & Workman, J. L. Transcription activator interactions with multiple SWI/SNF subunits. *Mol. Cell. Biol.* **22**, 1615–1625 (2002).
44. Kim, T. S., Kim, H. Y., Yoon, J. H. & Kang, H. S. Recruitment of the Swi/Snf complex by Ste12-Tec1 promotes Flo8-Mss11-mediated activation of STA1 expression. *Mol. Cell. Biol.* **24**, 9542–9556 (2004).
45. Harbison, C. T. et al. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004).
46. Hibbs, M. A. et al. Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics* **23**, 2692–2699 (2007).
47. Bednar, J. et al. Determination of DNA persistence length by cryo-electron microscopy. Separation of the static and dynamic contributions to the apparent persistence length of DNA. *J. Mol. Biol.* **254**, 579–594 (1995).
48. Tanay, A. Extensive low-affinity transcriptional interactions in the yeast genome. *Gen. Res.* **16**, 962–972 (2006).
49. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
50. Liu, X., Li, Y. I. & Pritchard, J. K. Trans effects on gene expression can drive omnigenic inheritance. *Cell* **177**, 1022–1034 e1026 (2019).

Methods

Theoretical TFBS abundance. We estimated the abundance of TFBSs in random DNA by analyzing the 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, composed of 80-bp sequences, the number of possible TFBSs is $(80 - length_{motif} + 1) \times 10^7$.

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 open reading frame, 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)

GCTAGCAGGAATGATGCAAAAGGTTCCCGATTCGA-ACTGCATTTTTTTCACATC

(poly-A; proximal)

GGTTACGGCTGTTTCTTAATTAAAAAAAGATAGAAAACATTAGGAGT-GTAACACAAGACTTTCGGATCCTGAGCAGGCAAGATAAACGA (up to the theoretical TSS).

For Abf1TATA:

(Abf1 site; distal)

GCTAGCTGATTATGGTAACTCTATCGGACTTGAGGGATC-ACATTTCACGCAGTATAGTTC

(TATA box; proximal)

GGTTTATTGTTTATAAAAATTAGTTTAAACTGTTGTATATTTTTTCAT CTAACGGAACAATAGTAGGTTACGCTAGTTTGGATCCTGAGCAGGCAAG ATAAACGA. In both cases, libraries of oligos containing 80 random bases (N80) 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 N^{80} oligonucleotide into a region corresponding to approximately −170 bp to −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 N^{80} was cloned by ligation.

Reporter assay. Libraries were transformed into yeast (strain Y8205 (ref. [53])) using the lithium acetate method[54], starting with 1 l yeast collected at an optical density of 0.3–0.4, ensuring that 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 yeasts were then grown in SC-Ura for 2 d, diluting the medium by 1:4 three times during this period. Medium was then either changed to YPD, growing for at least five generations before cell sorting, or to YPGly and YPGal, with culture grown for at least eight generations (due to the different carbon source). In the final 10 h of growth before cell sorting, all cultures were allowed to grow continuously in log phase, never achieving an optical density above 0.6, by diluting in fresh medium. All cultures were grown in a shaker incubator, at 30 °C and approximately 250 r.p.m.

Before sorting, yeasts 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 nonuniform 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 (for example, 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), then grown for 2–3 d, shaking at 30 °C. The plasmids were then isolated, the promoter region amplified, Nextera adapters and multiplexing indices added, and the resulting libraries sequenced with 2×76-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 bp (±15 bp) of overlap for pTpA and Abf1TATA libraries and 16 bp (±10 bp) for the scaffold library, and discarding any reads that failed to align well within these constraints. Note that only ~0.3 µg of N^{80} DNA was received from IDT, and only ~10^8 of these sequences were successfully cloned; this are only a vanishingly small portion of the possible 4^{80} sequences in N^{80} (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 of constant termini of the sequences, excluded from the alignment (for example, '-5 17 –3 13' for pTpA)). Any sequences that aligned 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 contain 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 were ensured to be free of ATGs (to avoid out-of-frame reporter translation). Each scaffold included fixed distal and proximal promoter regions (−298 to −195 and −103 to −33, relative to the theoretical TSS, respectively) surrounding a variable 80-bp random oligonucleotide (−189 to −109 regions). Each random scaffold was tested with ~660 random 80-mers, yielding approximately 2.5×10^6 distinct random promoter sequences in 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 subclustered 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 on re-sorting, we attempted to estimate the number of cells that would remain expressed on 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 N^{80} 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 in the "Reporter assay" Methods section, with two replicates. We combined the two replicates, which showed some nonlinearities resulting from differences in FACS binning procedures, using loess regression (span = 0.1) to remove the nonlinear 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 N^{80} 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 one-dimensional 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 the dissociation constant, 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 PWM for each yeast TF. Binding to each site in the DNA was determined by the "generalizable occupancy model of expression regulation" (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_d values for each binding site are calculated from the PWM:

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

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), which, summed, yielded the predicted expression level, EL_p.

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

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 one-dimensional convolutional filters. As described in the "Linear transcription model" section of the Methods, 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):

$$\Omega_p = \frac{1}{1 + e^{\left(c_p + \Sigma^{TF\,x} DB_{px} P_x\right)}}$$

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 (that is, 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:

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

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 messenger RNA 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:

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

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 described in the "Billboard model of transcription" Methods section):

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

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:

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

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 a penalty on the L_1 norm (0.00001), motifs were regularized with an L_2 penalty (0.000001) and position-specific activity biases (when present) were regularized with an L_2 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) were initialized with 1,000 random motifs (PWM values normally distributed about mean = 0, s.d. = 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 ten 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 with 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 (that is, 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 the Saccharomyces Genome Database[61] on January 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 (that is, 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 rep 2) in YPD for ~16 h to an OD of ~0.4–1.0. For each sample, 0.5 ml culture was pelleted and frozen to prepare input genomic DNA, and 3 ml culture was cross-linked with 1% formaldehyde, washed twice with 1 ml H_2O supplemented with a protease inhibitor cocktail and the pellet frozen for micrococcal nuclease (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 2×10^6 cells per µl NP buffer. Then, 0.125 units of Worthington MNase were added per 10 µl lysed spheroplasts and MNase digestion was performed at 37 °C for 20 min. MNase digestion was stopped by addition of an 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 and separated on a 4% agarose gel, and then 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 adapters 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 (ref. [56]). Only promoter sequences with at least 20 reads in the input DNA and one read in the MNase data were kept for subsequent analysis. Input and MNase counts were scaled within each sample to yield counts per million per promoter sequence, and the log ratio of MNase to input was compared between replicates and with 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 ρ was used to compare to model predictions, which is rank-invariant (unaffected by log-transformation).

Position- and orientation-specific TF activities. 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 with 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 with 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 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 (poly-A: 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 (that is, 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-versus-actual expression resulting from differences in the relative scaling of the bins between experiments. Reported Pearson r^2 values are on the raw data (without correction).

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.

Data availability

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

Code availability

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

References

51. Gibson, D. G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. 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* https://doi.org/10.17504/protocols.io.j4tcqwn (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. *Nat. 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. *Nat. Rev. Genet.* **10**, 443–456 (2009).

59. Abadi, M. et al. TensorFlow: large-scale machine learning on heterogeneous systems. *arXiv* 1603.04467 (2016).
60. Kent, N. A., Eibert, S. M. & Mellor, J. Cbf1p is required for chromatin remodeling at promoter-proximal CACGTG motifs in yeast. *J. Biol. Chem.* **279**, 27116–27123 (2004).
61. Cherry, J. M. et al. *Saccharomyces* Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* **40**, D700–D705 (2012).
62. Chua, G. et al. Identifying transcription factor functions and targets by phenotypic activation. *Proc. Natl Acad. Sci. USA* **103**, 12045–12050 (2006).

Acknowledgements

We thank R. Nelken, J. Weinstein, A. Dixit, B. Cleary, K. Shekhar and U. Eser for analysis advice; C. Muus, B. Cleary, A. Dixit, Y. Oren, T. Jones, L. Mariani, K. Shekhar, J. B. Kinney, D. M. McCandlish and J. Vierstra for feedback on the manuscript; T. Delorey, J. Pfiffner and C. Bashor for experimental advice; L. Gaffney and A. Hupalowska for help with figures; P. Rogers for cell sorting; and E. Segal for the dual reporter yeast vector. C.G.D. was supported by a Fellowship from the Canadian Institutes for Health Research and by the NIH (grant no. K99-HG009920-01). E.D.V. was supported by the MIT Presidential Fellowship. Work was supported by the Klarman Cell Observatory, the NHGRI Center of Excellence in Genome Science, the HHMI (A.R.) and the Israel Science Foundation ICORE on Chromatin and RNA in Gene Regulation (N.F.).

Author contributions

C.G.D. and A.R. drafted the manuscript, with all authors contributing. C.G.D. analyzed the data. C.G.D., E.D.V., E.L.A. and R.S. performed the experiments. A.R. and N.F. supervised the research.

Competing interests

A.R. is an SAB member of Thermo Fisher Scientific, Neogene Therapeutics, Asimov, and Syros Pharmaceuticals, an equity holder of Immunitas, and a founder of and equity holder in Celsius Therapeutics. All other authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-019-0315-8.

Correspondence and requests for materials should be addressed to C.G.d.B. or A.R.

nature research

Corresponding author(s): Regev A, de Boer CG

Last updated by author(s): Sep 30, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist .

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No specialized software was used for data acquisition.
Data analysis	Bowtie2 (version 2.2.1), R (3.3.1), Python (3.6.3), Tensorflow (1.1.0), custom scripts and programs (https://github.com/Carldeboer/CisRegModels)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

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

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were not predetermined. We aimed to get as many promoters as possible.
Data exclusions	No data were excluded, except for sorting bin EL=15 of the Abf1TATA+Glu library due to accidental inclusion of part of sorting bin EL=10.
Replication	Similar conclusions were derived from independently created datasets.
Randomization	Promoters were randomly synthesized and sampled, a random subset of yeast cells were transformed, and the order of training data were randomized prior to model training.
Blinding	The identities of the promoter sequences were unknown to the experimenter until after they were measured.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	S288C (ATCC), Y8205 (Charlie Boone Lab).
Authentication	Based on nutritional requirements (e.g. whether the yeast could grow in the presence/absence of certain nutrients). PCR of recombinant loci (Ura3).
Mycoplasma contamination	Yeast strains were not tested for mycoplasma.
Commonly misidentified lines (See ICLAC register)	Strains used in this study are not commonly misidentified.