Deciphering *cis*-regulatory logic with 100 million random promoters

Carl G. de Boer[1], Ronen Sadeh[2], Nir Friedman[1,2], Aviv Regev[1,3]

[1] Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge MA 02142

[2] School of Computer Science and Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem 9190401, Israel

[3] Howard Hughes Medical Institute and Koch Institute of Integrative Cancer Research, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02140 USA

Abstract

Predicting how transcription factors (TFs) interpret regulatory sequences to control gene expression remains a major challenge. Past studies have primarily focused on native or engineered sequences, and thus remained limited in scale. Here, we use random sequences as an alternative, measuring the expression output of over 100 million synthetic yeast promoters comprised of random DNA. Random sequences yield a broad range of reproducible expression levels, indicating that the fortuitous binding sites in random DNA are functional. From these data we learn models of transcriptional regulation that explain over 94% of expression variation of test data, recapitulate the organization of native chromatin in yeast, characterize the activity of TFs, and help refine *cis*-regulatory motifs. We find that strand, position, and helical face preferences of TFs are widespread and depend on interactions with neighboring chromatin. Such high-throughput regulatory assays of random DNA provide the large-scale data necessary to learn complex models of *cis*-regulatory logic.

Introduction

Cis-regulatory logic, the code by which transcription factors (TFs) interpret regulatory DNA sequence to control gene expression levels, is a key component of gene regulation. Understanding *cis*-regulatory logic would allow us to predict how gene expression is affected by changes to *cis*-regulatory sequences or regulatory proteins. This is important for both our basic understanding of this fundamental process and for determining the impact of genetic variants associated with common human traits and complex disease, most of which reside in regulatory sequences (reviewed in (1)).

Deciphering *cis*-regulation is a long-standing challenge (reviewed in (2, 3)). In general, learning such models requires a training set of *cis*-regulatory sequences and the expression levels associated with them. One approach has been to use natural sequences in the genome and the related gene expression profiles. Quantitative and semi-quantitative models relating DNA sequence to gene expression level have met with some success when learning on native sequences (4, 5). However, the rules learned often fail to generalize (5) and it is easy to overfit models when limited to the few sequences present in the genome (*e.g.*, the ~6,000 promoters in yeast) and their constrained evolutionary origins. In some cases, native sequences have been further diversified by mutagenesis (6). Alternatively, the expression of synthetic promoters can be measured, using either designed sequences (7) or designed elements (randomly-arranged; (8)). Although models learned from such data met some success, they are limited by available technologies for DNA synthesis, which currently allow the creation of at most ~100,000 sequences. In contrast, the space of possible sequences or of combinatorial of TF-TF interactions is vast. For example, approximately 10^7 sequences would be required to individually test all pairwise interactions between TF binding sites (TFBSs) with specific spacing and orientation constraints. Learning complex regulatory rules might require far more sequences than exist in the genome or have previously been assayed (3), such that predictive models of expression level from sequence alone remain elusive.

An alternative approach would be to use random DNA sequences. Past experiments have used random DNA as a cheap source of highly diverse sequences with which to study some aspects of gene regulation. *In vitro* selection (or SELEX) relies on the fact that high-affinity TFBSs are present by chance in random DNA to select oligonucleotides from a random pool that are bound by a protein of interest (9) and, in combination with high-throughput sequencing, can define the specificities (10) and affinities (11) of TFs. Random DNA has also been used to diversify regions of native promoters (12) or peptide sequences (13), which can then be selected for function. More recently, random DNA has been used to explore translational regulation (14), and to determine that ~10% of random 100 bp sequences could serve as promoters in bacteria (15), presumably due to fortuitous inclusion of regulatory elements recognized by the endogenous transcription machinery. Using massive numbers of random sequences to study eukaryotic *cis* regulation *in vivo* is compelling because it could readily produce data at a large enough scale to learn complex models of gene regulation. However, one would first need to demonstrate that random DNA can indeed drive reproducible expression levels, at a sufficient dynamic range, and is indeed sufficient to uncover the rules of regulation.

Here, we develop 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. Using a straightforward, large-scale, and cost-effective assay, we measure the expression levels driven by over 100 million synthetic yeast promoters, in three

growth conditions. We demonstrate that random sequences include abundant functional TFBSs, which, when included in a promoter-like context, direct diverse expression levels. Even though any specific sequence assayed by GPRA is very sparsely sampled (often in only one cell and/or by a single sequencing read), we can determine from these sequences and measured expression levels how TFBSs affect expression. We learned biochemically-inspired quantitative gene expression "billboard" models, which assume that TFs act independently and that the positions and orientations of TFBSs are irrelevant. These models successfully predicted known and novel chromatin-opening TFs, correctly determined DNA accessibility, and improved our knowledge of the DNA binding specificities of many TFs. These models explained nearly 93% of the variation in gene expression in random sequences, but only ~16% of native yeast genes, suggesting a large role for more complex (non-billboard) regulatory mechanisms. By incorporating positional preferences for TFs into our model, we find that binding location, orientation, and helical face preferences are widespread among yeast TFs, but depend on the surrounding sequence. Our approach enables cheap, precise, large-scale, and accurate measurements of regulatory element libraries that provides, in a single experiment, the "big data" needed to learn much more complex modes of *cis*-regulatory logic.

Results

Random sequence contains abundant TFBSs

To estimate the prevalence of yeast TFBSs in DNA randomly sampled from the four bases (hereafter referred to as "random DNA"), we calculated their expected frequency using the information contents (IC) of their motifs (**Methods**). Consistent with previous models (16), TF motifs are expected to occur frequently in random DNA (**Figure 1B**). For instance, the yeast Reb1 motif with a relatively high IC (14.59) should occur on average once every 12,000 bp in random DNA, while Rsc3, with an IC of just 7.78, should occur every 110 bp. More generally, 58% of motifs are expected to occur at least once every 1,000 bp and 92% to occur at least once every 100,000 bp. Thus, in a library of 10^7 promoter sequences, each with a different 80 bp random oligonucleotide (as we create below), we expect, on average, that at least 90% of yeast TFs will have over 10,000 distinct instances of their respective TFBSs included, and most yeast TFs will have far more TFBSs instances (**Figure 1B**). Consequently, a random 80 bp section of DNA is expected to have on average ~138 yeast TFBS instances, comprised of partly overlapping sites for ~68 distinct factors. As a result, any such random oligonucleotide cloned into a regulatory (*i.e.*, promoter scaffold) context is likely to contain many potential yeast TFBSs by chance alone.

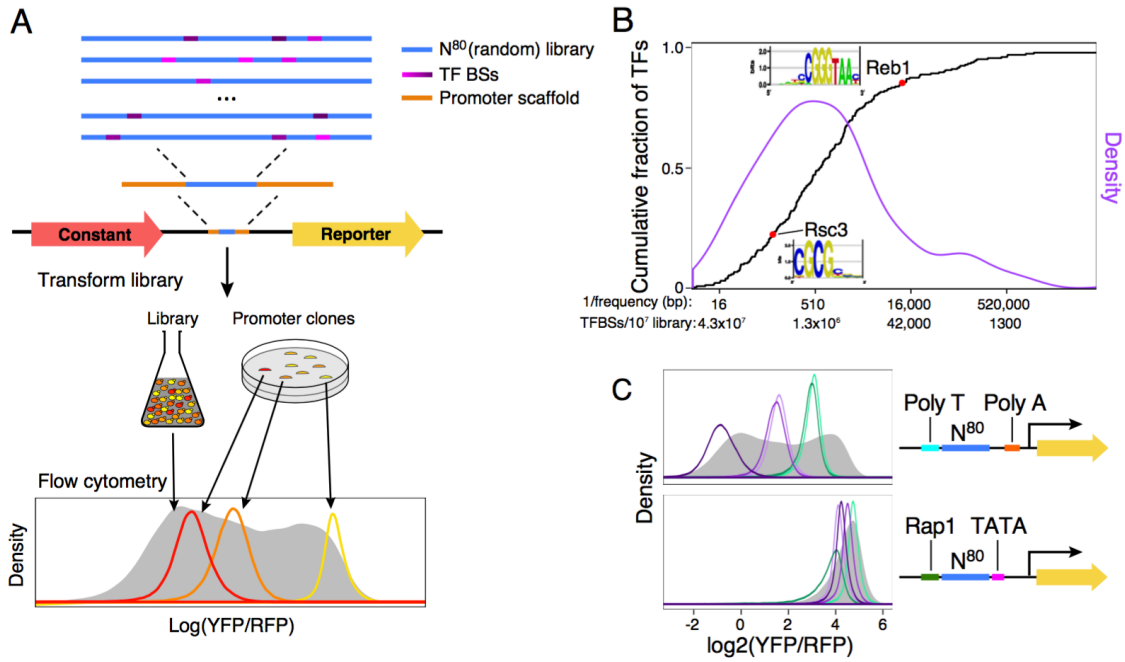

Figure 1: GPRA. (**A**) Overview. From top: A library of random DNA sequences (N^{80} here, blue) is inserted within a promoter scaffold (orange) in front of a reporter (yellow arrow). By chance, the random sequences include many TFBSs (purple). When grown in yeast, the library would yield a broad distribution of expression levels (grey, bottom) as measured by flow cytometry, whereas each promoter clone would have a distinctive expression distribution (red, orange, yellow). (**B**) TFBSs are common in random DNA. Shown is the cumulative distribution function (CDF; 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. (**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 the library (colored line curves).

Random DNA yields diverse expression levels

We hypothesized that a library of random DNA sequences, each containing a random assortment of TFBSs, will be associated with diverse expression levels. To test this hypothesis, we designed a system to robustly quantify promoter activity (**Methods, Figure 1A**). Using a previously described episomal dual reporter system (7) expressing RFP and YFP, we cloned a random 80 bp oligonucleotide into a promoter scaffold sequence regulating YFP, whereas RFP is under the control of a constitutive TEF2 promoter. We measured normalized expression levels as log(YFP/RFP) (**Methods**) using flow cytometry; this controls for sources of extrinsic noise, such as variation in plasmid copy number and cell size, similar to previously used strategies (6, 17, 18), and reports a gene expression signal that is integrated over several generations.

We created ten synthetic promoter scaffolds and one based on a native promoter sequence (the *ANP1* promoter). Each included 50-80 bp of constant sequence on either side of a cloning site into which we inserted 80 bp of random DNA (**Figures 1C and S1, Methods**). Each tested library had a complexity of at least 10^5 different sequences. To allow TFs to access the random DNA, we designed the synthetic promoter scaffolds to prevent nucleosome formation with either nucleosome disfavoring sites (poly-dA:dT tracts) (19, 20) or binding sites for the General Regulatory Factors (Abf1, Reb1, Rap1) (21-23), and sometimes with a TATA box (**Figure 1C** and **S1, right**). In each core promoter, the random 80 base pair oligonucleotide occupied the region from about –170 to –90, relative to the TSS.

In all instances, the random 80-mer libraries yielded diverse expression levels when measured by flow cytometry, and individual promoter clones from each library yielded distinct expression levels that fall within the range of the corresponding library (**Figures 1C and S1, left**). Each promoter scaffold showed some distinguishing behavior. The libraries containing an upstream poly-T sequence or Abf1 binding site spanned a ~50-fold range of expression levels, with a nearly uniform expression distribution. This indicates that random DNA contains functional TFBSs that modulate gene expression. Conversely, libraries based on the native pANP1 scaffold or on the synthetic scaffolds containing Rap1 or Reb1 sites were constitutively active, with the random DNA further modulating this basal expression level (**Figures 1C and S1**). This suggests that the TFBSs present in these promoter scaffolds drive expression and dominate the transcriptional outcome, which the random 80-mers further modulate.

High-throughput reproducible quantification of promoter activity

We next designed a system to readily and robustly assay the regulatory activity of tens of millions of random sequences in a single experiment (**Figure 2A, Methods**). We created diverse libraries of random promoters (~10^8), transformed them into yeast, and sorted the cells by log(YFP/RFP) into 18 bins of equal intervals. We regrew the yeast from each bin, and measured their expression distributions by flow cytometry, observing excellent reproducibility (**Figure 2B, Methods**). We sequenced the promoter libraries derived from each bin and collapsed related sequences that likely arose from errors in library amplification or sequencing (**Methods**). Expression level for each promoter was estimated as the weighted average of bins in which the promoter was observed, but because the complexity of each promoter library (>10^8) was often greater than the number of sorted cells (<10^8), most promoters (~78%) appear in only one bin, often representing one observation (read) from one cell containing that promoter and yielding a discrete expression level.

To determine the effect of promoter scaffold on expression, we applied this strategy to a library comprised of approximately 15,000 promoter *scaffolds*. Each scaffold included fixed distal and proximal promoter regions (-298:-195 and –103:-33, relative to the theoretical TSS, respectively) surrounding a variable 80 bp random oligonucleotide (-189:-109 regions, **Methods**). The 15,000 scaffolds comprised of: 4,985 native yeast promoter sequences, 3,811 specific sequences of random design, and 6,204 scaffolds that were random, but included various TFBSs spiked in either just upstream or just downstream of the random 80-mer. Overall, the library had ~10 million sequences: ~660 random 80-mers per scaffold on average. Scaffolds based on native promoters generally yielded higher expression than random scaffolds (**Figure S2A**, top). Overall, ~91% of promoters in a native scaffold and ~83% of promoters in a random scaffold had detectable expression (**Methods**). As before, when the scaffold itself drove a high mean expression level, the random 80-mer was less able to modulate expression level, regardless of whether the promoter scaffold was based on a native or random sequence (**Figure S2A**). We thus focused our in-depth analysis on the promoter scaffolds that yielded the most diverse expression (**Figure 1C**, **Figure S1**).

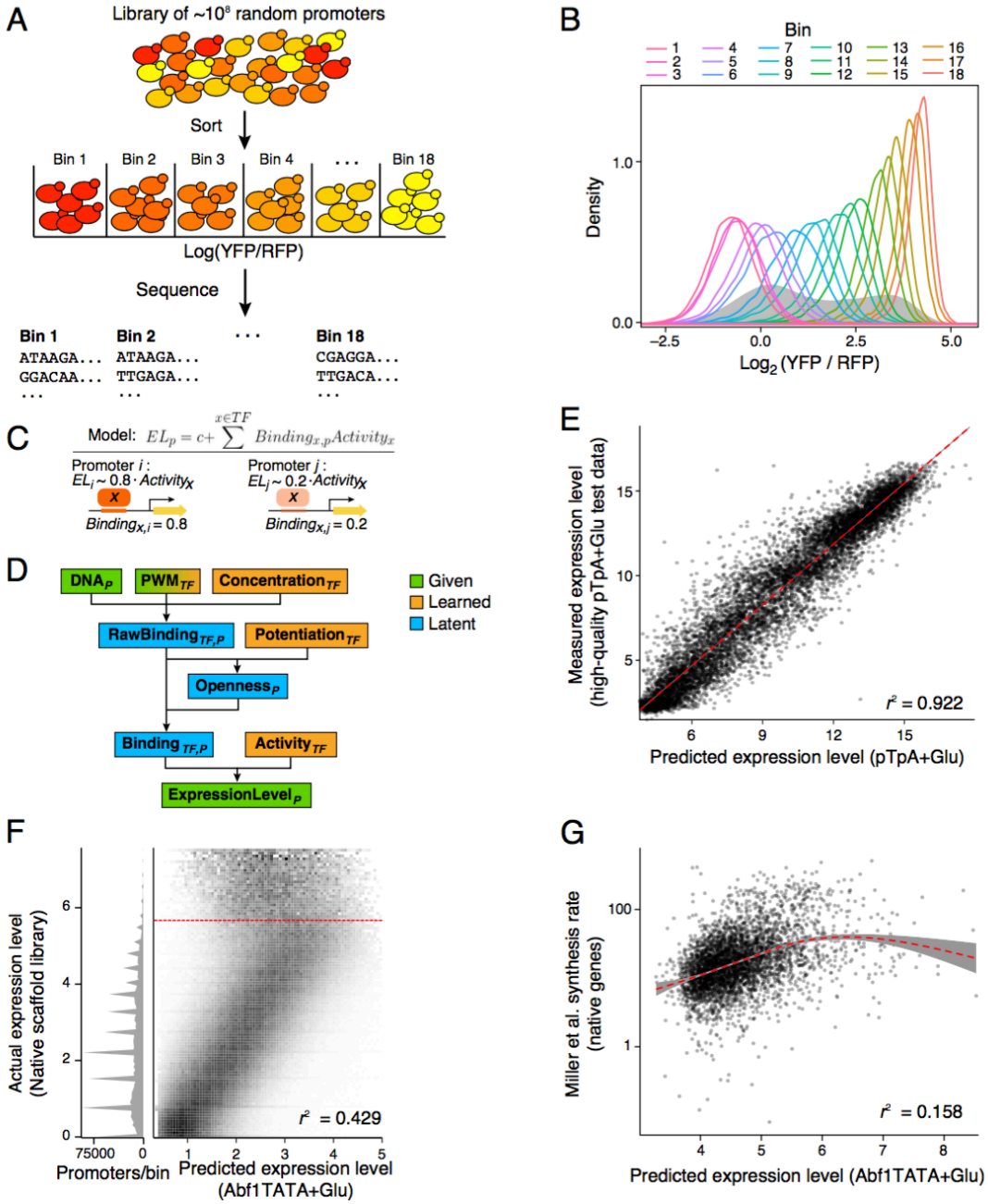

Figure 2: Learning an expression model from a GPRA of 10^8 random promoters. (**A**) Experimental strategy. Yeast GPRA library is sorted into 18 bins by the YFP/RFP ratio of the reporter (top) and the GPRA promoters in each bin are sequenced. (**B**) Reproducibility of expression levels. Expression distributions ($\log_2$(YFP/RFP)) for cells from each bin after sorting as in (A) (color code, top) which were regrown and reassayed by flow cytometry. Expression distribution maintains the initial bin ranking. (**C**) Linear model of TF activation. An increase in binding of a TF ($Binding_{x,p}$) results in a proportional change in expression (EL_p), scaled by the activity of that factor ($Activity_x$). (**D**) Computational "billboard" model. Observed promoter DNA sequence (DNA_p) are related to expression ($ExpressionLevel_p$) by TF binding and activity. TF motifs (PWM_{TF}) are provided as input, but can be refined and are used to calculate K_ds for each potential TFBSs. Three parameters are learned per TF (orange): $Concentration_{TF}$ (the amount of binding to each TFBS given its K_d), $Potentiation_{TF}$ (each TF's ability to open/close chromatin), and $Activity_{TF}$ (each TF's impact on transcription). Latent variables (blue) are calculated from the inputs and learned parameters. (**E**) Accurate prediction of expression from new random DNA. Actual expression level (*y* axis) for high-quality test data in the pTpA promoter scaffold grown in glucose *vs*. the predicted expression the pTpA+Glu model (*x* axis). (**F**) Partial prediction of hybrid native-random promoters. Right: Actual expression (*y* axis) *vs*. predicted (*x* axis) expression levels for native scaffolds containing central random 80-mers. Promoter density in the scatter plot (right) is scaled by the number of promoters (*x* axis, left) for each actual expression level bin (*y* axis). Red dashed line: approximate upper expression limit in the original GPRA experiments. (**G**) Weaker predictions of endogenous mRNA synthesis rates. Inferred mRNA synthesis rates of yeast promoters (from (24), *y* axis) *vs*. their predicted expression by the Abf1TATA model. Red line: GAM line of best fit (**Methods**).

TF specific effects are captured well by GPRA

We applied GPRA to two 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); or (**2**) an upstream Abf1 site and a downstream TATA box (Abf1TATA). We tested both libraries in glucose and the pTpA library also in galactose and glycerol, with 15-31 million unique promoters sequenced per experiment (<30% of the cells sorted; <21% of the theoretical number of promoters present in the libraries). Each library was sequenced to a depth of 50-155 million reads and did not reach saturation (**Figure S2B**). For example, the pTpA+glucose experiment was sequenced with 155 million reads, yielding 31 million promoters, but doubling the number of reads is projected to only have yielded a further 8.5 million promoters (30%; **Figure S2B**). Altogether, we measured the expression output of 102,371,025 promoters.

Even though each specific promoter sequence is typically associated with a single observed read, aggregating signal across the library reveals that for TFBSs there is a relationship between the strength of a binding site and the observed level of expression. To this end, we first scanned the DNA sequences of each promoter (DNA_P) with position weight matrices (PWMs) (25), given for each yeast TF (PWM_{TF}), revealing potential binding sites and providing an estimate for the dissociation constant (K_d) for each site (26). We considered all TFBSs, such that weak sites can also be influential, creating an affinity landscape for each TF across the region (27), and summed the predicted occupancy at each site, to obtain the expected occupancy for each TF of each sequence. Some TFBSs (*e.g.*, Abf1) had a strong effect on expression (**Figure S2C**, left), and yet explained a small percent of expression overall (Pearson's $r = 0.10$), presumably because these binding sites are relatively rare. In contrast, the predicted occupancy for many motifs, often corresponding to zinc cluster monomeric motifs (see below), correlated very strongly with expression (*e.g.*, Rsc30 r=0.57; **Figure S2C**, middle). Finally, the predicted occupancy of some TFBSs, like Ume6 (a similar motif to Rsc30), is positively correlated with expression overall (r=0.14), but has a strong negative trend at high occupancies (**Figure S2C**, right). Thus, examining the effect of TFBSs in isolation may be misleading, especially due to similar motifs with distinct behaviors (like Rsc30 and Ume6).

Next, considering only the overall, additive contribution of TFs, as defined by this measure we can explain a substantial proportion (47%) of expression level variation We combined the expected occupancies for each TF as above with a linear model (**Figure 2C**), learned a single parameter for each TF's activity ($Activity_{TF}$) and used this to predict expression on held-out data. If each TF explained wholly distinct subsets of the data, the linear model should have explain the sum of the variances explained by each TF: an impossible 348%. Instead, it explains only 47%, pointing to significant redundancy or interactions between related TFBSs and further motivating the need for more faithful joint modeling, which we pursued next.

A billboard model of TF action

We formulated a computational model of *cis*-regulatory logic that takes as input DNA sequence and predicts expression level (**Figure 2D**). Based on the explanatory power of a simple additive model above (**Figure 2C**), we focused on a simple "billboard" model of TF regulation that stipulates that gene expression is a function of the total amount of TF binding (28), irrespective of the absolute or relative position or orientation of individual TFBSs and assuming no TF-TF interactions. As before, the

expression output is the sum of the predicted binding of each TF, weighted by its effect on transcriptional output (**Figure 2C**), and therefore assumes that TFs work independently and additively. We, however, extend this simple model by learning two additional parameters per TF: cellular concentration of active protein (which determines the amount of binding), the ability to potentiate activity of other factors (*e.g.* by chromatin opening). Framing the model in this way potentially captures two important aspects of *cis*-regulation: (**1**) chromatin can block TFs from binding, and (**2**) some TFs can open chromatin and allow other TFs to bind (potentiation).

As above, the billboard model (**Figure 2D**) begins with scanning the DNA sequences of each promoter (DNA_P) with the motif models for each yeast TF (PWM_{TF}) to estimate dissociation constants (K_d) for each TF binding each site, still considering all potential TFBSs, including weak ones. To tune the amount of predicted TF binding given the motif-predicted K_ds, the model learns TF-specific concentration parameters ($Concentration_{TF}$). We next use Michaelis-Menten equilibrium binding (**Methods**) to get an initial estimate of TF occupancy of each promoter that does not yet consider chromatin state ($RawBinding_{TF,P}$). However, since nucleosomes can potentially prevent TF binding (29), the model encodes promoter accessibility ($Openness_P$). 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 ($Potentiation_{TF}$), which we assume is primarily driven by chromatin opening. Promoter accessibility is estimated as a logistic function on the potentiation-weighted $RawBinding_{TF,P}$ estimates, yielding a probability of the DNA being accessible. The previous estimate of binding ($RawBinding_{TF,P}$) is then scaled with this value, yielding the expected binding of each TF to each promoter ($Binding_{TF,B}$). For example, a promoter that is predicted to be 0% accessible will have no TF binding, regardless of the TFBSs present in the sequence ($Binding_{TF,P} = 0$), while a promoter that is 100% accessible will have occupancy unchanged ($Binding_{TF,P} = RawBinding_{TF,P}$). Thus, the model learns which TFs may, for example, open and close chromatin by their ability to potentiate the activity of other TFs (*i.e.*, TFBSs for TFs that affect transcription, but cannot open chromatin, only have an effect when "potentiated" by another factor, presumably by opening chromatin and allowing binding). Finally, the predicted expression level is calculated as with the linear model (**Figure 2C**), but using the potentiated TF binding estimate. Here, the learned TF-specific activity parameter can be either positive, for activation, or negative, for repression. Once these parameters are learned, we also allow the model to optimize the PWMs representing the TF binding specificities and finally add a saturation parameter that bounds the maximal effect a TF can have on expression (**Methods** and below).

The model explains >92% of expression in random DNA, but only 16% in yeast DNA

We learned a separate model for each of the four high-complexity promoter datasets: pTpA in glucose, galactose, and glycerol, and Abf1TATA in glucose, withholding, in each case, 500,000 promoters (1.5 – 3% of each library) from the training process to serve as initial test data. Performance on the withheld data ranged from 59.1% to 71.7% (Pearson r^2). However, as we show next, this vastly underestimates model performance, likely due to the substantial experimental noise in the measurement of these test promoters (~24%, per estimate below), many of which are derived from a single observation of a single cell.

Testing on independent, high quality data, the models predict >90% of the variation in expression in the same biological condition, and >80% of the variance even when testing on a different biological condition. To generate high quality test data, we assayed a smaller (~100,000 promoters) independent

pTpA library in glucose, sorted it into bins, and estimated the mean expression of only those ~10,000 promoters that had sufficient coverage (>100 reads each) (**Methods**). The best predictive performance was achieved by the pTpA+glucose model (r^2 = 0.922, **Figure 2E**), but the galactose- and glycerol-trained pTpA models performed nearly as well as on this glucose test data (r^2 = 0.896 and 0.836, respectively). This indicates that the primary contributors to gene expression in the context of random DNA sequence are not regulated by carbon source. A different promoter context led to weaker predictions: the Abf1TATA+glucose model had a lower predictive power (r^2 = 0.776, Spearman ρ^2 = 0.832) and a sigmoidal relation with the observed test data (**Figure S2D**). The models also predicted correctly which promoters are *not expressed*: although we sorted cells into each of the 18 bins, the lowest mean expression of the high-quality test data corresponds to bin 3, and, consistently, the models' predictions were no lower than expression bin 4. Overall, a remarkably high proportion of the variance in expression of random promoters is explained by a billboard model.

We next tested each model's ability to predict the expression driven by random sequences embedded within each of the 4,985 different native promoter scaffolds above (**Figure S2A**,**E**). The models explain 19% (pTpA) and 21% (Abf1TATA) of the variance in expression using only the random 80-mer sequences embedded within native scaffolds. When including the flanking scaffold sequences (combining occupancy and expression predictions of five overlapping regions tiling the promoter with a linear model trained on a random promoter subset; **Methods**), we can explain 42% (pTpA; **Figure S2E**) and 43% (Abf1TATA; **Figure 2F**) of expression, indicating that the models partly generalize to other parts of the promoter. Although both models predict the expression variation elicited by the random 80-mers *within* each promoter scaffold (median Pearson r^2 of 43% for both models), variance in performance *between* scaffolds was large (**Figure S2F**). The models perform particularly poorly at predicting very high expression levels. exceeding the levels in the training data (**Figure 2F** and **S2E**; red dashed line). (We obtained similar results with published expression measurements (7) (data not shown)). The models performed similarly on the same promoter scaffold (Pearson r^2=0.93, **Figure S2F**). Thus, some native promoter sequences may impact the activity of TFs, or may use more complex regulatory mechanisms not captured by our billboard models, particularly at high expression levels (**Figure 2F** and **S2E**).

When using our models to predict expression directly on native yeast promoters (**Methods**), they explain only up to 16% of variance in either mRNA synthesis (24) (**Figure 2G** and **S2G**) or RNA-seq levels (30) (data not shown). Several factors may contribute to this performance, including that our model is trained on short (80 bp) sequences compared to the longer yeast promoters (**Discussion**), and that the billboard model does not account for TF-TF interactions and position/orientation effects, which may be more prominent in native promoters. Nonetheless, these models are useful for learning the relevant biochemical activities of each TF, as we show next.

Accurately capturing biochemical activities of TFs

We next assessed each of the key parameters and features learned by the models: (**1**) which TFs are activators and repressors; (**2**) which TFs can open chromatin; (**3**) DNA accessibility; and (**4**) refinement of TF binding motifs. Overall, the parameters learned are remarkably concordant between the models, with few notable exceptions, which correctly highlight biological distinctions in regulation across conditions.

In all four models, TFs annotated as activators were predicted to have positive potentiation scores (*e.g.*, may open chromatin) and TFs annotated as repressors had negative potentiation scores (*e.g.*, may close it) (**Methods**; hypergeometric P-value: 10^{-3} to $2x10^{-5}$; **Figure S3A**), consistent with open chromatin being more active. The models predicted that, most TFs opened rather than closed chromatin (*i.e.*, had positive potentiation scores; 64-66%) and most were predicted activators rather than repressors (53-55%), although most TFs in all four experiments were predicted to have little activity, consistent with many TFs being inactive in rich media (31). The model-predicted activity for each TF weakly agreed with known activator/repressor status for models trained on glucose data (**Figure S3B**; hypergeometric P-values: 0.02 and 0.04), while there was no association for either galactose (P=0.34) or glycerol (P=0.79). This could reflect environment-specific activity of TFs and the ascertainment bias for TFs in glucose (the most common carbon source used to study yeast).

All models correctly identified factors known to open chromatin and predicted additional condition-specific chromatin-opening factors. The General Regulatory Factors (GRFs; Abf1, Reb1, and Rap1), which have known nucleosome displacing activity (21-23), were predicted by all models to open chromatin (positive potentiation scores) in all conditions tested (**Figure 3A,B**). In addition, only in galactose, the galactose-specific regulator Gal4 was correctly (32, 33) predicted to open chromatin (**Figure 3A**). TFs predicted to open chromatin only in glycerol included Hap4, Stb4, Cat8, Tec1, and Tye7 (**Figure 3B**). There is strong support for these predictions: Hap4 was previously described as a global regulator of non-fermentative media like glycerol (34); Cat8 activates gluconeogenic genes in ethanol and during the diauxic shift (35, 36) and Tye7 regulates glycolysis (37), which are the two endpoints of glycerol metabolism (38); Tec1 is known to regulate pseudohyphal growth (39, 40), which occurs constitutively in glycerol (41); and although little is known about Stb4, its motif occurs preferentially in promoters of genes annotated for "oxidoreductase activity" (25), consistent with a role in using non-fermentable carbon sources.

Correct prediction of accessibility in the libraries and in the yeast genome

There was a good correspondence between the model's predicted occupancy and the occupancy we experimentally measured by MNase-seq. We quantified enrichment of promoter sequences from limited-complexity subsets of the pTpA library among nucleosome-sized DNA fragments that were protected from MNase digestion, and compared this to model-predicted accessibility for these sequences (**Methods**). The correlation between model predictions and individual experiments (Spearman ρ = 0.54-0.55; **Figures 3C** and **S3E**) was similar to that between experimental replicates for pTpA promoters (**Figure 3C**), and was even higher when comparing to the average occupancy across experimental replicates (Spearman ρ = 0.80 and 0.67; **Figure S3F**).

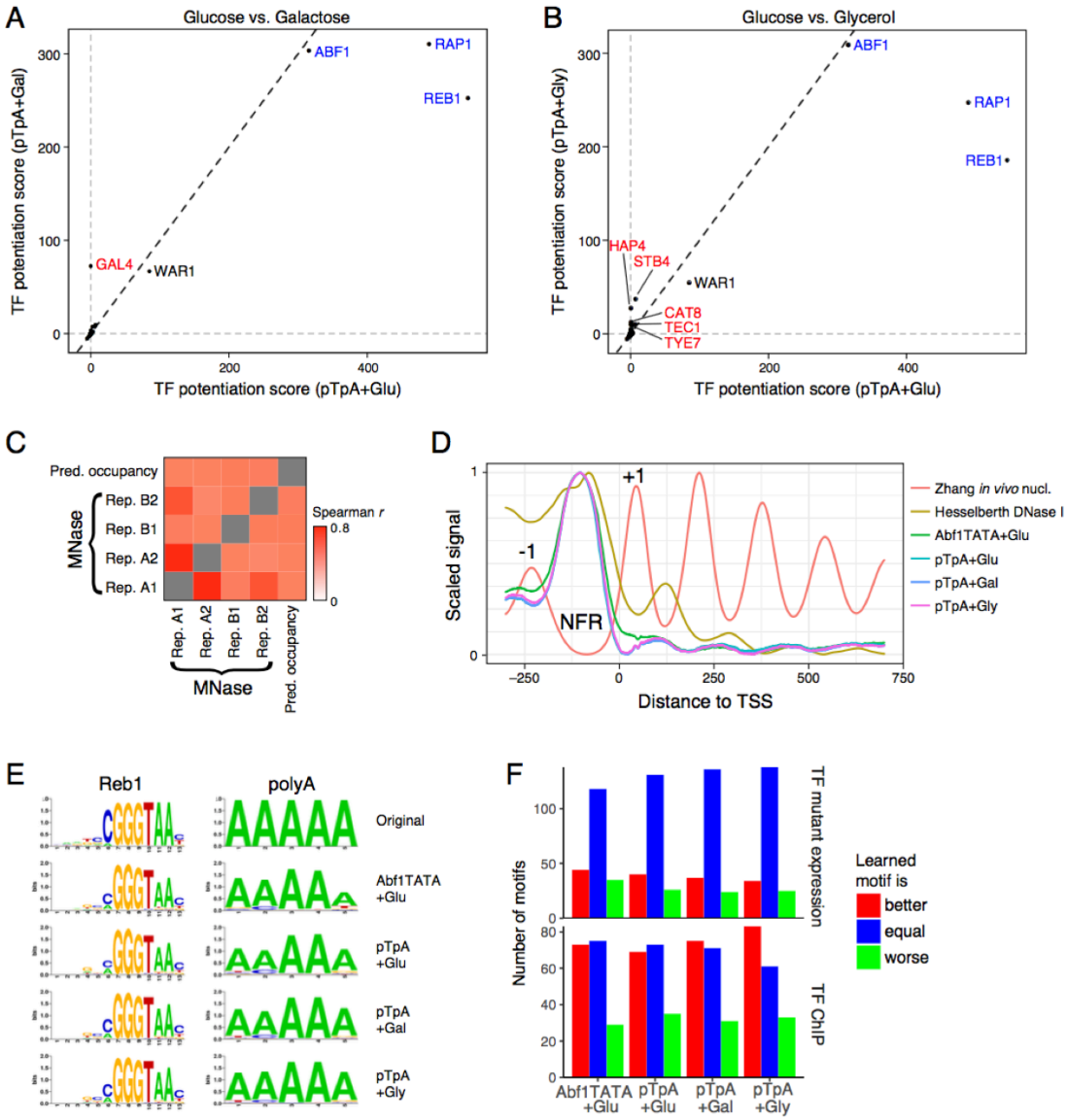

Figure 3: Billboard models learn biochemical activities of TFs. (**A,B**) Prediction of chromatin opening ability. Shown is the predicted chromatin opening ability for each TF (dot) for pTpA models trained in glucose (*x* axes) *vs*. either galactose (**A**) or glycerol (**B**) (*y* axes). Blue: GRFs with known chromatin opening ability in all conditions; red: known and putative carbon source-specific regulators. (**C,D**) Prediction of chromatin accessibility. (**C**) Pairwise Spearman correlations (color) between model-predicted nucleosome occupancy (1-predicted accessibility) and *in vivo* nucleosome occupancy measured by MNase (four replicates/conditions). (**D**) Metagene profile surrounding the TSS, based on *in vivo* nucleosome occupancy (Zhang (42)), DNase I hypersensitivity (representing accessibility; Hesselberth (43)), and model-predicted accessibility for each of the four billboard models. Each dataset is scaled. +1 and -1 nucleosome positions, and promoter Nucleosome Free Region (NFR) are indicated. (**E,F**) TFBS motif refinement by the model. (**E**) Similar refinement in independent models. Comparison of the original TFBS motif (top) and model-refined motifs from each of the four models for two example motifs. (**F**) Motif refinement improves experimental predictions. The number of TFBS motifs (*y* axis) for which the model-refined motif predicted gene expression changes (TF mutant, top) or TF binding (ChIP, bottom) better (red), worse (green), or equally well (blue) as the original motif, for each of the four models (*x* axis).

The pattern of nucleosome accessibility predicted by applying the models to the yeast genome also agrees well with previously measured endogenous nucleosome occupancy in yeast (**Figure 3D**). Specifically, we compared averaged meta-gene profiles of chromatin openness predicted by our models across all yeast promoter sequences to openness as measured by DNase I-seq (43) or to *in vivo* nucleosome occupancy (42) (**Figure 3D, Methods**). The models accurately predict the nucleosome free region and -1 and +1 nucleosomes, and even (weakly) predict the array of nucleosomes within the first

part of the gene body, indicating that this nucleosomal array is partly encoded in the DNA sequence, and read by TFs. Thus, billboard models trained on expression levels associated with random sequences correctly learned aspects of how certain TFs regulate chromatin structure.

Substantial refinement of TF binding motifs

The model is allowed to optimize the position weight matrices (PWMs) describing TF specificities (including by introducing additional bases of specificity), and doing so improved the predictive power (r^2) of the models by 9-12 percentage points. Although in principle the motifs could be altered to the point where they no longer represented the original TFBS, this was not generally the case: most motifs either (**1**) closely resemble the original ones, or (**2**) were not useful and so the PWMs were degraded to neutrality, such that they no longer specifically recognize any distinct sequence. The four models often made the same changes to the motif, suggesting that the revised motif may more faithfully represent the true specificity of the factor (**Figure 3E**).

Many of the refined motifs performed better than the original ones at the independent tasks of predicting which targets are bound by the cognate TF in the yeast genome by ChIP (44) and which yeast genes would change in expression when the cognate TF is perturbed (45) (**Figure 3F**, **Figure S3C,D**, **Methods**). While many motifs were indistinguishable from the originals (**Figure 3F**), of those that differed, the model-refinement improved the majority of motifs. For ChIP data, over twice as many motifs had improved as had worsened, even though many of the original motifs were learned from the same ChIP data (25). This suggests that the refined motifs often more closely represent their cognate TF specificities.

The activity of most TFs is proportional to their binding

We tested whether each TF's activity is directly proportional to its binding, as assumed by the model (**Figure 2C**). We considered the relationship between model-predicted TF binding and the measured expression level (**Figure 4A top**, red) or the residual expression level (actual expression minus expression predicted by the model; **Figure 4A bottom**, purple; **Methods**). If a TF's activity is correctly captured by the model, there should not be a lingering relationship with the residual because the model correctly incorporated the TF's effect on expression (**Figure 4A**, left). Alternatively, if a TF's activity is *not* faithfully represented by the model, a lingering relationship will exist (**Figure 4A**, right), and will be reflected as a non-zero slope for the line of best fit between predicted binding and residual expression level.

The activity of the vast majority of TFs was directly proportional to their binding (**Figure 4B**), with the GRFs (Abf1, Reb1, and Rap1) and Gal4 (in galactose) being notable exceptions. These had a strong negative relationship between binding strength and residual expression (**Figure 4B, Figure S4A**), reflecting saturation of the effect of the factor's binding on expression (**Figure S4B,C**). (Although all these factors are nucleosome displacing factors, other displacing factors, such as Rsc3 (**Figure S4D**) or Hap4 in glycerol (**Figure S4E)** did not exhibit this behavior.) We modified the model to allow it to capture such saturation, by introducing a parameter that enforces an upper limit for each TFs' activity (**Methods**). In this revised model, the activity of the GRFs was predicted to saturate at relatively low occupancies (4%, 5%, and 11% for Abf1, Rap1, and Reb1), and the model's predictive power improved

by only 0.6% (on the high-quality test pTpA+glucose data), but the lingering relationship between binding strength and residual expression level was eliminated. Since strong binding sites may be more likely to occur *in vivo* than in random sequence, this is an important addition.

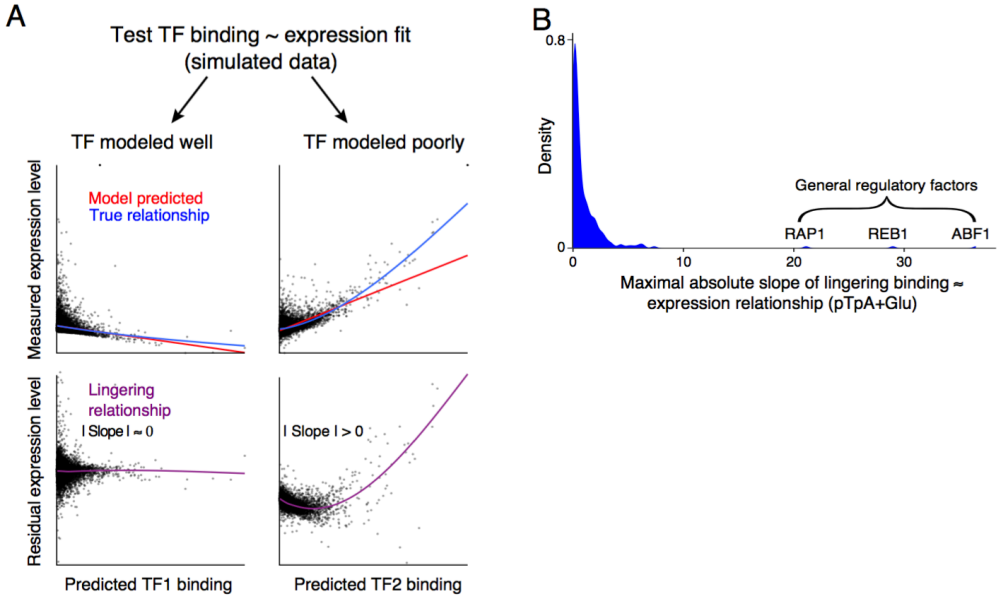

Figure 4: Only GRFs show nonlinear transcriptional activity. (**A**) Lingering expression relationships for well-fit (left) and poorly-fit (right) TFs. Shown are relationships on simulated data between predicted TF binding (*x* axis) and measured expression level (top, *y* axis) or residual expression level not explained by the model fit (bottom, *y* axis) for an example TF that is fit well (left) and another that is fit poorly (right). Blue: the true relationship between TF binding and expression; red: the model's learned linear fit; purple: Generalized Additive Model (GAM) line of best fit to residual and its slope. (Expression level and residual expression are expressed in log space, corresponding to the sorting bins, and TF binding is in linear space, representing expected occupancy in our analysis.) (**B**) Binding of most TFs is captured well, with the notable exception of GRFs. Distribution of maximal absolute slopes for the GAM lines of best fit between TF binding *vs*. residual expression (as in (A), bottom, purple curves) for the TFs in the pTpA+glucose model. The three GRFs with particularly poor fits are marked.

CGG-related motifs explain 57% of variation in expression in random DNA

Examining the effect of each TF motif across the libraries (considering both the number of promoters affected, and the effect size in each case; **Methods**), many monomeric motifs for zinc cluster TFs (CGG and related) had a large potentiation impact (*e.g.*, WAR1 in **Figure 3A,B**; **Figure S5A**). Zinc cluster TFs are generally thought to bind as dimers (46), but our result highlighted a monomeric motif. To assess the specific impact of these monomeric motifs, we learned a model whose input motif features included *only* the zinc cluster monomeric consensus (CGG/CCG) and its one base pair variants, which were not further optimized. The resulting model explained 57% of the variance in expression of the high-quality pTpA glucose test data (**Figure 5A,B**). By several tests (**Methods**), this is unlikely to merely reflect lower-order features, such as G+C-content or dinucleotide frequencies (**Figure S5B**). The large impact of these motifs is likely attributed in part to their high frequency: CGG is expected to occur approximately once every 32 bases in random DNA (50% G+C), and every 73 bases in the yeast genome (38% G+C). The activity of these CGG-variants could be due to either one or a few TFs binding the monomeric motifs, or the combined action of many TFs.

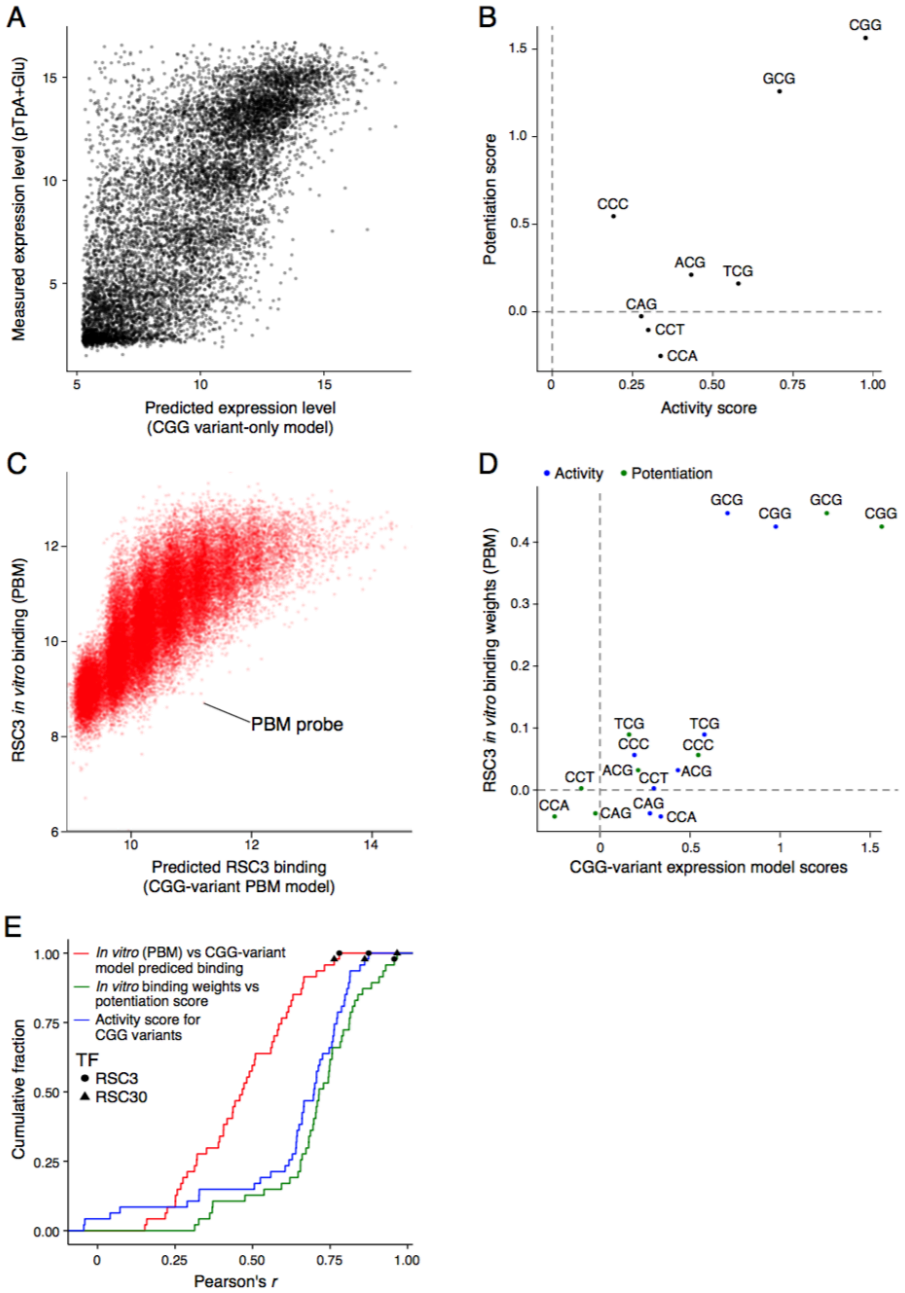

Figure 5: Rsc3/30 explain much of the variation in expression in random sequence. (**A**) A billboard model with only CGG variants explains a large portion of the variation in expression. Shown are measured expression levels (*y* axis) for each random sequence in the high-quality pTpA test data *vs*. the corresponding predictions (*x* axis) for these sequences based on a billboard model that can use only the consensus zinc cluster monomeric motif (CGG) and its 1 bp variants as motif features. (**B**) Potentiation and activity scores for CGG-variant motifs. Shown are the potentiation (*y* axis) and activity (*x* axis) scores for each of the CGG variants, learned by the CGG-variant model. (**C-E**) The role of Rsc3/30 is supported by comparison to protein binding microarrays (PBMs). (**C**) Predicting Rsc3 *in vitro* DNA binding only from CGG-variant motif abundance in DNA. Shown is the measured binding of RSC3 to different sequences in a PBM (*y* axis), *vs*. the model-predicted binding for a linear model trained on the same data, including only the abundance of CGG variants within each PBM probe as features (Pearson r = 0.78). (**D**) Agreement between model-predicted activity and Rsc3 *in vitro* binding weights. Shown is a comparison between the CGG-variant model's feature weights (as in B; *x* axis) for activity (blue) and potentiation (green), and the DNA binding weights learned for each CGG variant by a model trained to predict *in vitro* Rsc3 binding using only these CGG variants (*y*-axis) (the model as in the *x* axis of C). Pearson r = 0.87 and 0.96 for activity and potentiation, respectively. (**E**) Rsc3/30 best explain the activity of CGG-variants. CDFs show, for all zinc cluster TFs with PBM data in UniPROBE (47), the Pearson correlation coefficient r (*x* axis) for how well binding can be explained by CGG variants (as in C, red), and how well *in vitro* CGG-variant binding weights match activity (blue) and potentiation (green) scores (as in D) performed for each TF. Rsc3 and Rsc30 are marked within each distribution.

Further analysis suggests the paralogs Rsc3 and Rsc30 may be the main binders to these sites. To rank candidates among all zinc cluster TFs, we built models that predicted *in vitro* TF binding using protein binding microarray (PBM) data (47, 48) for each such TF by the occurrence of CGG-variant motifs in PBM probes (**Methods**), and then compared the CGG-variant weights for *in vitro* binding to those learned by our CGG-variant gene expression model (**Methods**). The highest correlation was for Rsc3 and Rsc30 (**Figure 5C-E**), whose binding in the PBM assay was also best explained by CGG-variants. Rsc3 and Rsc30 are part of the RSC chromatin remodeler, bind CG repeats (49) (like the second ranking CGG-variant; **Figure 5B**), open chromatin (49), and RSC3 is essential (50). Thus, ~57% of the variation in expression from random sequence promoters may be due to Rsc3/Rsc30 binding, although we cannot fully rule out contributions from other factors.

Widespread position, orientation, and helical face preferences

Motif position and orientation can affect TF function by modifying the TF's ability to contact its biochemical target within mediator, the transcriptional pre-initiation complex (PIC), or surrounding chromatin. We therefore extended the billboard model with localized activity bias terms (**Methods**), allowing each TF to have a different activity for each possible binding position and orientation (**Figure 6A**). This model fits ~224 parameters per TF (instead of only four in the original model), including ~110 location-specific activity parameters for both strands of DNA (**Figure 6B**). Including these parameters significantly increased predictive value. For example, the pTpA+Glu model now explained 94.3% of the high-quality test data, representing a 21% decrease in the error.

We found evidence for strong position and strand preferences (**Figure 6B** and **S6A-F**), as well as for helical face preference (**Figure 6C**). Many predicted activators are more active when binding distally within the promoter (*e.g.*, ABF1, SKN7, MCM1), while many predicted repressors appeared most repressive when binding proximal to the TSS (*e.g.* UME6, MOT3). Many others are strand-specific in their activity, often with a lower-than-expected activity distally, but for only one motif orientation (*e.g.*, AZF1, MGA1, THI2).

Some TFBSs showed strong periodicity along the length of the promoter (*e.g.*, MCM1, THI2, poly-A, AZF1), potentially reflecting a preference for a DNA helical face. To test this, we first removed long-distance positional preferences using loess regression, leaving only short-scale trends (**Figure S6G**), and calculated the Spearman correlation to a 10.5 bp sine wave (**Methods**). The correlations for each model were significantly higher than with randomized data (**Figure 6C**, rank sum $p<10^{-120}$; AUROC=0.84-0.87), suggesting that helical face preferences are commonplace. The observed helical preference (periodicity) in TF activity tends to be strongest when proximal to the TSS (downstream of -150, relative the TSS), while many TFs appear to be most active when bound distally within the promoter (upstream of -150, relative to the TSS). Interestingly, 150 bp is the approximate persistence length of dsDNA (51), and so this could indicate physical constraints of the promoter sequence, where a TF bound close to the TSS can only contact the transcriptional pre-initiation complex when bound to a particular helical face. Conversely, after a distance of ~150 bp, the DNA is flexible enough that TFs can regulate transcription efficiently regardless of the helical face on which they bind.

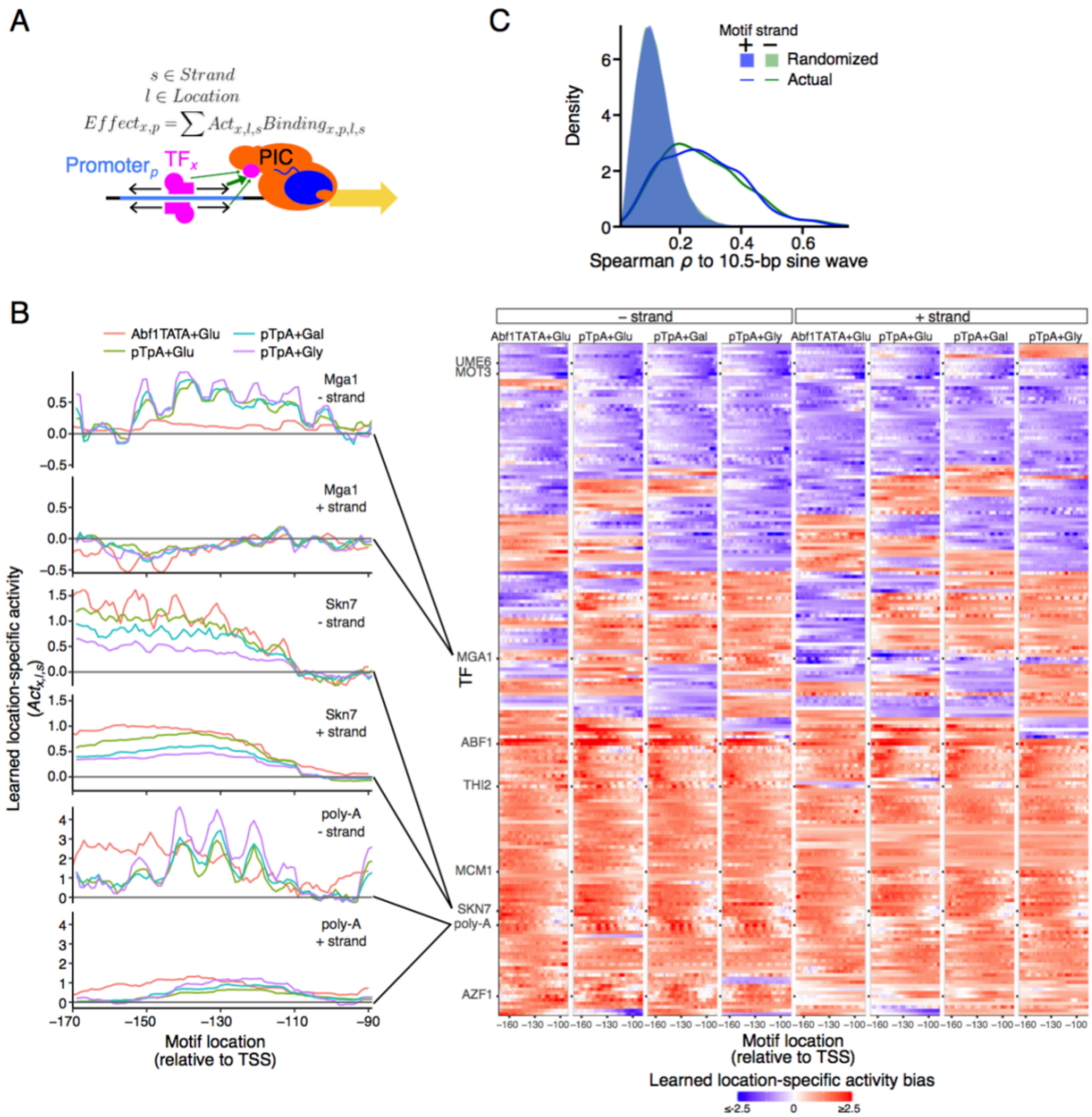

Figure 6: Position, orientation, and helical face preferences among yeast TFs. (**A**) Model with position and orientation-specific activities. For each TF, the model learns parameters for how much binding at each location within the promoter affects transcriptional activity. For example, this could reflect the TF's ability 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 orientation for each of the four models (four subpanels). Only TFs for which all models retained the motif are shown. (**C**) Helical face preferences. Shown is the distribution of Spearman ρ between a 10.5 bp sine wave and the learned position-specific activity weights (as in **Figure S6**) for plus strand (blue line) and minus strand (green line) or with corresponding randomized data (blue and green shaded areas) for all four models.

Helical preference varied between models learned from different scaffolds and conditions. Although each model learned a helical preference for minus strand poly-A motifs (poly-T), the pTpA scaffold (which already uses the poly-A element) has a much stronger helical preference than the Abf1TATA scaffolds (**Figure 6B** – bottom left). This is consistent with the chromatin remodeler RSC acting on proximal poly-A motifs to position the -1 nucleosome (and hence having little helical preference in our expression assay) and acting on distal poly-Ts to position the +1 nucleosome (which may affect the transcriptional pre-initiation complex) (52-54). Similarly, only pTpA models learned a strong strand-specific helical bias for Mga1 motifs (**Figure 6B** – top left), whereas only Abf1TATA-trained model

showed bias for Skn7 motifs (**Figure 6B** – middle left). Indeed, adding positional biases often worsened the performance of models trained using one scaffold on data generated in the other scaffold, but improved their performance on data generated using the same scaffold in another condition (**Figure S6H**). This suggests that the models correctly learned positional preferences of TFs, but these preferences can depend on the surrounding context.

Discussion

We used a massive-throughput approach to measure the expression output of over 100 million sequences, a radically different scale than prior studies, relying on random DNA. Through a regulatory "billboard" model, we explained the vast majority of expression variance of random DNA, improved our ability to predict where TFs will bind, correctly predicted chromatin organization, and identified factors that can remodel chromatin, including condition-specific regulators that can potentiate the activity of other TFs (**Figure 7A**, top left). We showed that the majority of TFs have strand, location, and helical face preferences, demonstrating that *cis*-regulatory logic can be exceedingly complex (**Figure 7A**). This exceptionally high-throughput approach allows researchers to determine the roles played the cell's entire complement of TFs with a simple and inexpensive experiment.

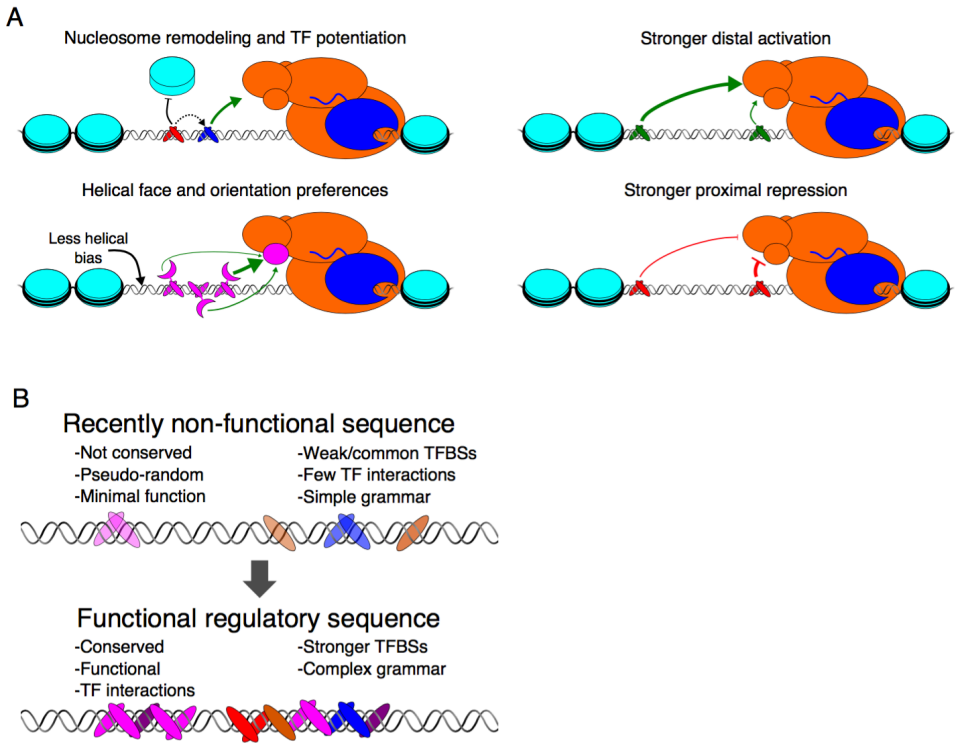

Figure 7: Model of *cis*-regulatory logic. (**A**) TFs display a variety of activity types. Some TFs potentiate the activity of other TFs by modulating nucleosome occupancy (upper left). Activators tend to have a greater effect on transcription when bound distally within the promoter (upper right), while 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). (**B**) Evolution of regulatory sequence. Previously non-functional intergenic sequence can be made functional by mutation or by a change in its context (e.g. nearby retroposition or deletion). At this point, the sequence, previously evolving neutrally, is pseudo-random and has a simple regulatory grammar consistent with a billboard model. Over evolutionary time, mutations can accumulate, creating stronger TFBSs and interactions between TFs that can result in tighter regulatory control, but leading to a complex regulatory grammar.

Random DNA has several key advantages for the study of *cis*-regulatory logic. The ease of generating massive libraries allows measurements of unprecedented scale, important for learning complex models from many independent examples of TFBSs in a variety of contexts and of diverse binding strengths. Conversely, the traditional approach of introducing the feature for study into a common background sequence can inadvertently affect binding sites for other TFs that partly overlap the one studied; indeed, such fortuitous introduction or destruction of secondary TFBSs is highly likely in designed studies.

The activities of most individual TFs were fit well by having one parameter for their effect on transcription and one for their effect on chromatin. The major exceptions to this were the GRFs, whose activity saturated, potentially reflecting the way in which they open chromatin (21-23, 32, 33, 55), since once the chromatin is open in all cells at all times, it cannot be opened further. While we aimed to capture specific biochemistry by including TF potentiation scores (which we generally have interpreted as chromatin opening), the TF activity scores learned by the model do not specifically correspond to any of the biochemical processes through which TFs are known to affect transcription (56). Furthermore, our simplifying "billboard" assumption ignores known biochemical phenomena (*e.g.* competition between factors, positioning of nucleosomes). It is likely that incorporating these into future models will help glean further insights.

Indeed, several known features of gene regulation were not incorporated in our modeling framework. We represented TFBSs by traditional position weight matrices, which assume independence between adjacent positions of the motif, and did not consider possible contributions from DNA shape features (*e.g.*, (57, 58)). We also did not allow a TF to simultaneously act as an activator and repressor in the same condition. However, we found that the activities of some TFs appeared to depend on the surrounding promoter scaffold (**Figure 6B** – right), and, in rare cases, TFs were predicted to activate when bound in some locations and orientations and repress in others (*e.g.,* Mga1 in **Figure 6B**). Thus, our results suggest that the majority of cases where a TF has seemingly different functions in different contexts result from interactions with other factors that alter, block, or render redundant the activity of the TF (59-61). Since binding sites for individual TFs are common in our dataset and only 6% of the data remain unexplained by the model, it is unlikely that these regulatory mechanisms contribute significantly to expression level in the context of independent TF action.

In contrast to the billboard model's successful predictions in random DNA, it explained less than 16% of the mRNA synthesis rates of native genes from their promoter sequences. We consider three possible reasons for the discrepancy. First, there are substantial limitations in the experimental techniques used to infer RNA synthesis rates, and different techniques for measuring mRNA decay rates (used to infer synthesis rates) correlate poorly (r ranges from -0.14 to 0.56) (62). Second, we only analyzed a portion of the promoter (from -170 to -90, relative to the TSS), and our model did not capture contributions to expression from the proximal promoter and upstream (distal) activating sequences, although our results suggest that our models at least partly generalize to these regions. Third, the billboard model does not capture genomic context (63-66), which can impact the expression of the same promoter up to ~15 fold (67), but is held constant in our promoter assay; TF-TF interactions (61), which are expected to occur comparatively infrequently in random DNA (and so have little impact on model performance); or TFBS position and orientation preferences (7). Indeed, we found that many TFs have position and orientation-specific activity, presumably via interactions with factors binding to the surrounding sequence and which we show can depend on the surrounding sequence. Although we cannot determine the relative

contributions of these factors in limiting the success of our model on native genes, this is consistent with regulatory sequences evolving to be increasingly complex as they are optimized over time (**Figure 7B**).

The prevalence of functional TFBSs in random DNA and its demonstrated ability to modulate gene expression has implications for the ways in which genes evolve. Since our billboard model explained the vast majority (93%) of the expression variance of random sequence, it provided strong support for the hypothesis that many weak sites can impact transcription additively (68) (strong sites are less likely to occur in random DNA). When a new gene is created by a mechanism like retroposition of an existing gene, the regulatory program, encoded by the DNA, must arise *de novo*. In bacteria, where there are no nucleosomes, random sequences have been shown to yield functioning promoters about 10% of the time (15). Here, we show that yeast promoter sequences also occur frequently by chance: ~83% of random promoter sequences appeared to be at least minimally active in glucose. Therefore, it may not be difficult to evolve basal gene regulatory sequences from previously non-regulatory DNA when a new gene is formed. Creating new enhancers in mammals may be similarly likely since mammalian TFs have, on average, even less specificity than those of yeast (16). This is also consistent with the observed fast evolutionary turnover of regulatory DNA, while overall expression programs are conserved (69). According to this hypothesis, newborn evolutionarily naive sequences will be primarily comprised of many weak TFBSs that have a comparatively weak effect on expression, potentially dominated by constitutive TFs with low specificity, like we show for Rsc3/30 (**Figure 7B**). Over evolutionary time, further mutations can optimize the specificity and effect of these new regulatory sequences.

In using GPRA, researchers will have to consider the scale needed for their question of interest. Signal-to-noise increases as data quantity increases, but in a manner that depends on each TFBS's frequency (*e.g.*, **Figure 1B**). In our analysis, the activity and potentiation parameters for each TF converged within the first 10% of the data. Conversely, an increase in data is important for refining or learning new motifs, and for finding position and orientation-specific activities. As noted above, since pairs of TFBSs are inherently rarer in random DNA, learning all possible TF-TF interactions with GPRA, especially when considering competition (where both binding sites must be high-affinity), will require much bigger datasets. Such truly "big data" will allow learning more elaborate models to address all facets of gene regulation.

Acknowledgments

We are grateful to Rani Nelken, Josh Weinstein, Atray Dixit, Brian Cleary, Karthik Shekhar, and Umut Eser for analysis advice; Christoph Muus, Brian Cleary, Atray Dixit, Yaara Oren, Thouis Jones, Luca Mariani, Karthik Shekhar, Justin B. Kinney, and David M. McCandlish for feedback on the manuscript; Toni Delorey, Jenna Pfiffner, and Caleb Bashor for experimental advice; Leslie Gaffney for help with figures; Patricia Rogers for cell sorting; and Eran Segal for the dual reporter yeast vector. CGD was supported by a Fellowship from the Canadian Institutes for Health Research. Work was supported by the Klarman Cell Observatory at the Broad Institute, NHGRI Center of Excellence in Genome Science (CEGS), HHMI (AR), and the Israel Science Foundation ICORE on Chromatin and RNA in Gene Regulation (NF). AR is a member of the Scientific Advisory Board of ThermoFisher Scientific, Syros Pharmaceuticals and Driver Group.

Author Contributions

Conceptualization and Writing – Original Draft, C.G.D. and A.R.; Methodology, Software, Formal Analysis, and Visualization, C.G.D.; Investigation, C.G.D. and R.S.; Writing – Review & Editing, C.G.D., A.R., N.F., and R.S.; Funding Acquisition, A.R., C.G.D., and N.F.; Supervision, A.R. and N.F.

Declaration of Interests

AR is an SAB member of ThermoFisher Scientific. All other authors declare no competing interests.

References

1. Albert FW & Kruglyak L (2015) The role of regulatory variation in complex traits and disease. *Nat Rev Genet* 16(4):197-212.
2. Bussemaker HJ, Foat BC, & Ward LD (2007) Predictive modeling of genome-wide mRNA expression: from modules to molecules. *Annu Rev Biophys Biomol Struct* 36:329-347.
3. Hughes TR & de Boer CG (2013) Mapping yeast transcriptional networks. *Genetics* 195(1):9-36.
4. Beer MA & Tavazoie S (2004) Predicting gene expression from sequence. *Cell* 117(2):185-198.
5. Yuan Y, Guo L, Shen L, & Liu JS (2007) Predicting gene expression from sequence: a reexamination. *PLoS computational biology* 3(11):e243.
6. Kinney JB, Murugan A, Callan CG, Jr., & Cox EC (2010) Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences of the United States of America* 107(20):9158-9163.
7. Sharon E*, et al.* (2012) Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature biotechnology* 30(6):521-530.
8. Gertz J, Siggia ED, & Cohen BA (2009) Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature* 457(7226):215-218.
9. Oliphant AR, Brandl CJ, & Struhl K (1989) Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Molecular and cellular biology* 9(7):2944-2949.
10. Jolma A*, et al.* (2013) DNA-binding specificities of human transcription factors. *Cell* 152(1-2):327-339.
11. Nutiu R*, et al.* (2011) Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nature biotechnology* 29(7):659-664.
12. Horwitz MS & Loeb LA (1986) Promoters selected from random DNA sequences. *Proceedings of the National Academy of Sciences of the United States of America* 83(19):7405-7409.
13. Winter G, Griffiths AD, Hawkins RE, & Hoogenboom HR (1994) Making antibodies by phage display technology. *Annu Rev Immunol* 12:433-455.
14. Cuperus JT*, et al.* (2017) Deep learning of the regulatory grammar of yeast 5' untranslated regions from 500,000 random sequences. *Genome research*.
15. Yona AH, Alm EJ, & Gore J (2017) Random Sequences Rapidly Evolve Into De Novo Promoters. *bioRxiv*.
16. Wunderlich Z & Mirny LA (2009) Different gene regulation strategies revealed by analysis of binding motifs. *Trends in genetics : TIG* 25(10):434-440.
17. Kosuri S*, et al.* (2013) Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America* 110(34):14024-14029.
18. Shalem O*, et al.* (2015) Systematic dissection of the sequence determinants of gene 3' end mediated expression control. *PLoS Genet* 11(4):e1005147.
19. Iyer V & Struhl K (1995) Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J* 14(11):2570-2579.
20. de Boer CG & Hughes TR (2014) Poly-dA:dT tracts form an in vivo nucleosomal turnstile. *PLoS One* 9(10):e110479.

21. Ganapathi M*, et al.* (2011) Extensive role of the general regulatory factors, Abf1 and Rap1, in determining genome-wide chromatin structure in budding yeast. *Nucleic acids research* 39(6):2032-2044.
22. Bernstein BE, Liu CL, Humphrey EL, Perlstein EO, & Schreiber SL (2004) Global nucleosome occupancy in yeast. *Genome biology* 5(9):R62.
23. Hartley PD & Madhani HD (2009) Mechanisms that specify promoter nucleosome location and identity. *Cell* 137(3):445-458.
24. Miller C*, et al.* (2011) Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Mol Syst Biol* 7:458.
25. de Boer CG & Hughes TR (2012) YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. *Nucleic acids research* 40(Database issue):D169-179.
26. Granek JA & Clarke ND (2005) Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome biology* 6(10):R87.
27. Segal E & Widom J (2009) From DNA sequence to transcriptional behaviour: a quantitative approach. *Nat Rev Genet* 10(7):443-456.
28. Kulkarni MM & Arnosti DN (2003) Information display by transcriptional enhancers. *Development* 130(26):6569-6575.
29. Liu X, Lee CK, Granek JA, Clarke ND, & Lieb JD (2006) Whole-genome comparison of Leu3 binding in vitro and in vivo reveals the importance of nucleosome occupancy in target site selection. *Genome research* 16(12):1517-1528.
30. Lipson D*, et al.* (2009) Quantification of the yeast transcriptome by single-molecule sequencing. *Nature biotechnology* 27(7):652-658.
31. Chua G*, et al.* (2006) Identifying transcription factor functions and targets by phenotypic activation. *Proceedings of the National Academy of Sciences of the United States of America* 103(32):12045-12050.
32. Axelrod JD, Reagan MS, & Majors J (1993) GAL4 disrupts a repressing nucleosome during activation of GAL1 transcription in vivo. *Genes Dev* 7(5):857-869.
33. Morse RH (1993) Nucleosome disruption by transcription factor binding in yeast. *Science* 262(5139):1563-1566.
34. Forsburg SL & Guarente L (1989) Identification and characterization of HAP4: a third component of the CCAAT-bound HAP2/HAP3 heteromer. *Genes Dev* 3(8):1166-1178.
35. Hedges D, Proft M, & Entian KD (1995) CAT8, a new zinc cluster-encoding gene necessary for derepression of gluconeogenic enzymes in the yeast Saccharomyces cerevisiae. *Molecular and cellular biology* 15(4):1915-1922.
36. Haurie V*, et al.* (2001) The transcriptional activator Cat8p provides a major contribution to the reprogramming of carbon metabolism during the diauxic shift in Saccharomyces cerevisiae. *The Journal of biological chemistry* 276(1):76-85.
37. Sato T*, et al.* (1999) The E-box DNA binding protein Sgc1p suppresses the gcr2 mutation, which is involved in transcriptional activation of glycolytic genes in Saccharomyces cerevisiae. *FEBS Lett* 463(3):307-311.
38. Grauslund M & Ronnow B (2000) Carbon source-dependent transcriptional regulation of the mitochondrial glycerol-3-phosphate dehydrogenase gene, GUT2, from Saccharomyces cerevisiae. *Can J Microbiol* 46(12):1096-1100.
39. Madhani HD & Fink GR (1997) Combinatorial control required for the specificity of yeast MAPK signaling. *Science* 275(5304):1314-1317.

40. Gavrias V, Andrianopoulos A, Gimeno CJ, & Timberlake WE (1996) Saccharomyces cerevisiae TEC1 is required for pseudohyphal growth. *Mol Microbiol* 19(6):1255-1263.
41. Cullen PJ & Sprague GF, Jr. (2000) Glucose depletion causes haploid invasive growth in yeast. *Proceedings of the National Academy of Sciences of the United States of America* 97(25):13619-13624.
42. Zhang Z*, et al.* (2011) A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. *Science* 332(6032):977-980.
43. Hesselberth JR*, et al.* (2009) Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature methods* 6(4):283-289.
44. Harbison CT*, et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* 431(7004):99-104.
45. Hibbs MA*, et al.* (2007) Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics* 23(20):2692-2699.
46. Todd RB & Andrianopoulos A (1997) Evolution of a fungal regulatory gene family: the Zn(II)2Cys6 binuclear cluster DNA binding motif. *Fungal Genet Biol* 21(3):388-405.
47. Hume MA, Barrera LA, Gisselbrecht SS, & Bulyk ML (2015) UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic acids research* 43(Database issue):D117-122.
48. Zhu C*, et al.* (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome research* 19(4):556-566.
49. Badis G*, et al.* (2008) A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol Cell* 32(6):878-887.
50. Akache B, Wu K, & Turcotte B (2001) Phenotypic analysis of genes encoding yeast zinc cluster proteins. *Nucleic acids research* 29(10):2181-2190.
51. Bednar J*, et al.* (1995) 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(4):579-594.
52. Krietenstein N*, et al.* (2016) Genomic Nucleosome Organization Reconstituted with Pure Proteins. *Cell* 167(3):709-721 e712.
53. de Boer C & Hughes TR (2015) The RSC complex may be the poly-A nucleosome turnstile mechanism. *figshare*.
54. de Boer C & Hughes TR (2015) Model for how poly-dA:dT sites act as nucleosome turnstiles. *figshare*.
55. Yu L & Morse RH (1999) Chromatin opening and transactivator potentiation by RAP1 in Saccharomyces cerevisiae. *Molecular and cellular biology* 19(8):5279-5288.
56. Hahn S & Young ET (2011) Transcriptional regulation in Saccharomyces cerevisiae: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. *Genetics* 189(3):705-736.
57. Rohs R*, et al.* (2009) The role of DNA shape in protein-DNA recognition. *Nature* 461(7268):1248-1253.
58. Mathelier A*, et al.* (2016) DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. *Cell Syst* 3(3):278-286 e274.
59. Voth WP*, et al.* (2007) Forkhead proteins control the outcome of transcription factor binding by antiactivation. *EMBO J* 26(20):4324-4334.

60. Turcotte B & Guarente L (1992) HAP1 positive control mutants specific for one of two binding sites. *Genes Dev* 6(10):2001-2009.
61. Zhou X & O'Shea EK (2011) Integrated approaches reveal determinants of genome-wide binding and function of the transcription factor Pho4. *Mol Cell* 42(6):826-836.
62. Geisberg JV, Moqtaderi Z, Fan X, Ozsolak F, & Struhl K (2014) Global analysis of mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast. *Cell* 156(4):812-824.
63. de Boer CG*, et al.* (2014) A unified model for yeast transcript definition. *Genome research* 24(1):154-166.
64. Kaplan CD, Laprade L, & Winston F (2003) Transcription elongation factors repress transcription initiation from cryptic sites. *Science* 301(5636):1096-1099.
65. Mazo A, Hodgson JW, Petruk S, Sedkov Y, & Brock HW (2007) Transcriptional interference: an unexpected layer of complexity in gene regulation. *J Cell Sci* 120(Pt 16):2755-2761.
66. Chen M, Licon K, Otsuka R, Pillus L, & Ideker T (2013) Decoupling epigenetic and genetic effects through systematic analysis of gene position. *Cell Rep* 3(1):128-137.
67. Chen X & Zhang J (2016) The Genomic Landscape of Position Effects on Protein Expression Level and Noise in Yeast. *Cell Syst* 2(5):347-354.
68. Tanay A (2006) Extensive low-affinity transcriptional interactions in the yeast genome. *Genome research* 16(8):962-972.
69. Weirauch MT & Hughes TR (2010) Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends in genetics : TIG* 26(2):66-74.
70. Tong AH & Boone C (2006) Synthetic genetic array analysis in Saccharomyces cerevisiae. *Methods Mol Biol* 313:171-192.
71. de Boer C (2017) High-efficiency S. cerevisiae lithium acetate transformation. *protocols.io*.
72. Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods* 9(4):357-359.
73. Li W & Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22(13):1658-1659.
74. Abadi M*, et al.* (2015) TensorFlow: Large-scale machine learning on heterogeneous systems.
75. Cherry JM*, et al.* (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic acids research* 40(Database issue):D700-705.
76. Deng C, Daley T, & Smith AD (2015) Applications of species accumulation curves in large-scale biological data analysis. *Quant Biol* 3(3):135-144.