Deciphering cis-regulatory logic with 100 million synthetic 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 nearly 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 this data we learn 'billboard' models of transcriptional regulation that explain 93% of expression variation of test data, recapitulate the organization of native chromatin in yeast, and help refine *cis*-regulatory motifs. Analyzing the residual variation, we uncover more complex regulatory mechanisms, such as strand, position, and helical face preferences of TFs. 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 process 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 affecting common human traits and complex disease, most of which reside in regulatory sequences (reviewed in (1)).

Modeling *cis*-regulation is a long-standing challenge (reviewed in (2, 3)). In general, learning such a model 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. Such 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 when limited to the sequences present in the genome, in part due to the few examples of regulatory sequences (*e.g.*, the ~6,000 promoters in yeast) and their evolutionary origins. An alternative is to measure the expression output by synthetic promoters using either designed sequences (6) or designed elements (randomly-arranged; (7)). 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 such complex regulatory rules could 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 remained elusive.

An alternative approach would be to use random DNA sequence. 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 (8) and, in combination with high-throughput sequencing, can define the specificities (9) and affinities (10) of TFs. Random DNA has also been used to diversify regions of promoters (11) or peptide sequences (12), which can then be selected for function. More recently, random DNA has been used to explore translational regulation (13), and to determine that ~10% of random 100 bp sequences could serve as promoters in bacteria (14), 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. We demonstrate that random sequences include abundant functional TFBSs, which, when included in a promoter-like context, direct diverse expression levels. We measure the expression levels driven by ~100 million synthetic yeast promoters, in three growth conditions. Using this data, 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. Remarkably, 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. Analyzing the residual 7% of expression data that remained unexplained by the billboard models, we find that position, orientation, and helical face preferences are widespread among yeast TFs. Our approach enables cheap, precise, and accurate measurements of regulatory element libraries, providing 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 (15), TF motifs are expected to occur very frequently in random DNA (**Figure 1B**). For instance, the yeast Reb1 motif has a relatively high IC (14.59) and is predicted to occur once every 12,000 bp in random DNA. 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 about 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.

Random DNA yields diverse expression levels

We hypothesized that a library of random DNA, each sequence 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 (6) 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) 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 (16-18).

We created ten synthetic promoter scaffolds and one based on a native promoter sequence (the *ANP1* promoter) that each included 50-80 bp of constant sequence on either side of a cloning site into which we inserted 80 bp of random DNA (**Figure 1C and Figure 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**, **Figure S1**). 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 (**Figure 1C and Figure S1**). Each promoter scaffold had some specific characteristics. 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 modulating the specific expression level (**Figure S1**). This suggests that the TFBSs present in these promoter scaffolds induce 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 very diverse libraries of random promoters (~10^8), transformed them into yeast, and sorted the cells by the log(YFP:RFP) ratio 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**). Because the complexity of each promoter library (>10^8) was greater than the number of sorted cells (<10^8), many promoters will appear in only one bin, often representing a single observation of a single cell containing that promoter and thus yielding a discrete expression level. For those that appear in more than one bin (~22% of promoters), expression level is estimated as the weighted average of bins in which the promoter was observed.

We applied this strategy to the two promoter libraries (each complexity > 10^8) with the most diverse expression (**Figure 1C**, **Figure S1**) 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 sequenced promoters 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 S2A**); many promoters were observed from a single read. 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 S2A**). Altogether, we measured the expression output of nearly 100,000,000 promoters.

Learning 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 2C**). Our approach focused on the simple "billboard" model of TF regulation that stipulates that gene expression is a function of the total amount of TF binding (24), irrespective of the absolute or relative position or orientation of individual TFBSs and assuming no TF-TF interactions. The expression output is the sum of the predicted binding of each TF, weighted by its effect on transcriptional output (**Figure 2D**), and therefore assumes that TFs work independently and additively. We learned three 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), and transcriptional modulation. 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).

To learn the model (**Figure 2C**), we first scan 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. We consider all TFBSs, such that weak sites can also be influential, creating an affinity landscape for each TF across the region (26). The predicted occupancy of each TFBS is determined by the learned TF-specific concentration parameters ($Concentration_{TF}$), providing an initial estimate of TF occupancy of each promoter that does not yet consider chromatin state ($RawBinding_{TF,P}$). We learn TF-specific parameters for how much each TF can modulate the binding of other TFs ($Potentiation_{TF}$), which we assume is primarily driven by chromatin opening, since a promoter must be accessible for TFs to bind (27). Using the learned potentiation parameters, we estimate the probability that each promoter is accessible to TF binding ($Openness_P$) and scale the initial occupancy estimates by this value, yielding the amount of binding of each TF to each promoter ($Binding_{TF,B}$). 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). We calculate each TF's contribution to expression as the amount of binding for each TF multiplied by a TF-specific activity parameter, which can be either positive, for activation, or negative, for repression. To calculate expression, we sum the individual contributions to expression from each TF (**Figure 2D**). 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).

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.

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

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 an independent pTpA library of limited complexity (~100,000) 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**). All models performed very well on this data, with the highest predictive value for the pTpA+glucose model (r^2 = 0.922, **Figure 2E**). The galactose- and glycerol-trained pTpA models performed nearly as well as the glucose-trained model on this glucose test data (r^2 = 0.896 and 0.836, respectively), indicating 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 S2B**). 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 also compared the models' predictions to published expression measurements from another reporter system that used specific designed promoters based on modifications of native ones (6). Because this test data was measured in SC-Ura+Gal without most amino acids, we used the pTpA+galactose model (including TF activity saturation parameters; see below). There are many differences in the test system (**Methods**), including the core promoter sequence (pHIS3), the use of modified native promoters, and its design to test promoter features explicitly not captured by the billboard model, such as TFBS position, orientation, and relative arrangement. Nevertheless, our model predicted expression variation well *within* promoter contexts that share a common basal promoter sequence and test a similar variable (**Figure S2C**). Conversely, the model could not predict expression variation *between* promoter contexts (r^2 = 0.01) or when the variation focused on the organization of TFBSs (**Figure S2C**), which are not captured by a "billboard" model. Thus, these modified native promoter sequences contain features that overwhelm our model's ability to predict expression *across* promoter contexts, but, once these features are held constant, it can predict differences *within* a given context.

When using our models to predict expression from all native yeast promoters (**Methods**), they explain only up to 16% of variance in either mRNA synthesis (28) (**Figure 2F,G**) or RNA-seq levels (29) (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. Nonetheless, these models are useful for learning the relevant biochemical activities of each TF, as we show next.

The model accurately captures 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 $2x10^{-5}$; **fig S3A**), consistent with open chromatin being more active. The models predicted that, of all TFs, most opened rather than closed chromatin (*i.e.*, had positive potentiation scores; 64-66%) and that most TFs were predicted activators rather than repressors (53-55%), although most TFs in all four experiments were predicted to have very little activity, consistent with many TFs being inactive in rich media (30). The model-predicted activity for each TF only 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 (31, 32) 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 (33); Cat8 activates gluconeogenic genes in ethanol and during the diauxic shift (34, 35) and Tye7 regulates glycolysis (36), which are the two endpoints of glycerol metabolism (37); Tec1 is known to regulate pseudohyphal growth (38, 39), which occurs constitutively in glycerol (40); 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.

The model correctly predicts 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 (**Methods**). The correlation between model predictions and individual experiments (Spearman ρ = 0.54-0.55; **Figure 3C** and **Figure 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**).

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 a predicted averaged meta-gene profile of chromatin openness by our model's predictions across all yeast promoters to meta-gene profiles from DNase I-seq (41) or *in vivo* nucleosome occupancy (42) (**Figure 3D, Methods**). The model accurately predicts the nucleosome free region and -1 and +1 nucleosomes, and even (weakly) predicts 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. This indicates that the models correctly learned aspects of how certain TFs regulate chromatin structure, even though they were trained to predict gene expression and were provided no prior information about chromatin state.

The model substantially refined 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 (43) and which yeast genes would change in expression when the cognate TF is perturbed (44) (**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 2D**). We considered the relationship between predicted TF binding

and the measured expression level or the residual expression level (actual expression minus expression predicted by the model; **Figure 4A**; **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 being notable exceptions, which had a strong negative relationship between binding strength and residual expression (Abf1, Reb1, and Rap1; **Figure 4B, Figure S4A**; Gal4 in galactose, **Figure S4C**). This reflects saturation in the impact 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 share this behavior.) When we allowed the model to learn a saturation parameter on a TFs' activity (**Methods**), 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 residual relationship was eliminated. Since strong binding sites may be more likely to occur *in vivo* than in random sequence, this is an important addition to the model.

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 (45), 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 held constant, without further optimization. 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.

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 (46, 47) 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 (48) (like the

second ranking CGG-variant; **Figure 5B**), open chromatin (48), and RSC3 is essential (49). 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

Some regulatory mechanisms, including the effects of motif position and orientation are present in random DNA at some frequency, but are not captured by our billboard model. Even if motifs at specific positions of the promoter are relatively rare in random DNA, (and thus the billboard model fits the data well overall), there could still be a sufficient number of instances in our large dataset from which to study these mechanisms. We reasoned that the residuals, after the fit by the billboard model, should highlight these effects, since most of the variance attributable to other (possibly confounding) factors has been eliminated. To identify such effects, for each TF, we identified all promoters that contained a cognate TFBS predicted to be bound at least 5% of the time (**Methods**), partitioned these promoters into bins by the TFBS position and orientation, and examined the distribution of expression residuals for promoters with the TFBS at each position and orientation bin (**Figure 6A**). Finally, we clustered the median residuals for all TFs at every position within the promoter and for both orientations (**Figure 6B**). We focused the analysis on the pTpA glucose dataset, where we had the largest number of promoters.

We found evidence for strong position and strand preferences (**Figure 6B**), as well as for helical face preference (**Figure S6**). Many TFBSs are associated with a higher-than-expected expression level when the TFBS is distal within the promoter (*e.g.*, ABF1, PHD1, RSC3). 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). Some TFBSs showed strong periodicity along the length of the promoter (*e.g.*, MCM1, PHD1, RSC3). We hypothesized these could reflect preference for a DNA helical face. To test this, we first removed large-scale preferences using loess regression, leaving only short-scale trends (**Figure S6**), and calculated the Spearman correlation to a 10.5 bp sine wave (**Methods**). The correlations were significantly higher than with randomized data (**Figure 6C**, rank sum $p<2x10^{-16}$; AUROC=0.82), suggesting that helical face preferences are commonplace.

The observed helical preference (periodicity) in TF activity tends to be proximal to the TSS (downstream of -150, relative the TSS), while the region that is most active when TFBSs are included is distal within the promoter (upstream of -150, relative to the TSS). Interestingly, 150 bp is the approximate persistence length of dsDNA (50), 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.

One notable exception to the proximal periodicity preference is the poly-A motif, recognized by the chromatin remodeler RSC (51, 52), which has a higher activity when minus strand motifs (*i.e.*, poly-Ts) are located distally within the promoter and when plus strand motifs (*i.e.*, poly-As) are proximal (**Figure 6B, bottom left – green curve**). Only

the distal poly-T motifs (-170 to -130) have a strong periodicity in activity (**Figure 6B, bottom left – blue curve**). These preferences are consistent with the proximal poly-A positioning the -1 nucleosome (and hence having little helical preference in our expression assay) and the distal poly-T positioning the +1 nucleosome (which may affect the transcriptional pre-initiation complex) (53).

Discussion

Here, we used a massive-throughput approach to measure the expression output of nearly 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, helped refine TFBSs, correctly predicted chromatin organization, and identified factors that can remodel chromatin, including condition-specific regulators. By analysis of the model's residuals we also uncovered regulatory features present in the data – but not captured in a "billboard" model – including strand, orientation and helical face preferences.

Random DNA has several key advantages for the study of *cis*-regulatory logic. The ease of generating very large 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.

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 (strong sites are less likely to occur in random DNA). The activities of most individual TFs were fit well by having a single parameter for their effect on transcription and a second 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, 31, 32, 54), 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 correspond to specific biochemical processes. There are many pathways through which TFs can affect transcription (55) and it is likely that incorporating these into future models will help glean further insights.

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. 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 (14). Here, we show that yeast promoter sequences also occur very frequently by chance: over 80% of promoter sequences appeared to be at least minimally active in glucose, in the context of the pTpA promoter scaffold. 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 (15). This is also consistent with the observed fast evolutionary turnover of regulatory DNA, while overall expression programs are conserved (56). 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. Over evolutionary time, further mutations can optimize the specificity and effect of these new regulatory sequences.

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. Thus, our results suggest that 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 7% 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 very 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. Third, the billboard model does not capture certain features that might particularly affect endogenous gene regulation, as these sequences have been selected by evolution. These include genomic context (63-65), which 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), and TFBS position and orientation preferences (6). Indeed, position and orientation-specific activity were commonplace according to our analysis of residuals.

In using GPRA, researchers will have to consider the scale needed for their question of interest. First, signal-to-noise increases as data quantity increases, but in a manner that depends on each TFBS's frequency (*e.g.*, **Figure 1B**). Second, some parameters can be learned with relatively little data: in particular, activity and potentiation parameters converge in models within the first 10% of the data. Conversely, an increase in data is important for learning motifs and for finding position and orientation-specific activities. As noted above, since pairs of 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), 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, Ray Jones, Luca Mariani, and Karthik Shekhar 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.

Figures

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 instances of 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).

FIGURE 1

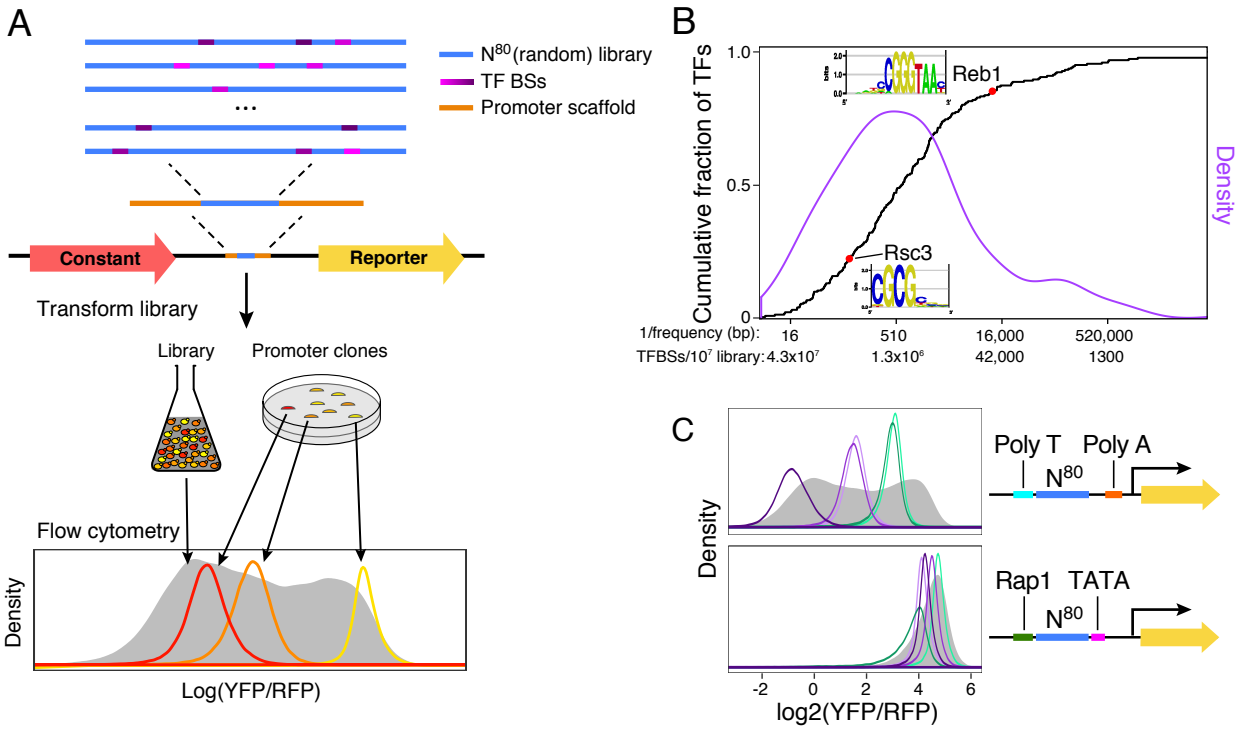

A
N80(random) library
TF BSs
Promoter scaffold
Constant
Reporter
Transform library
Library
Promoter clones
Flow cytometry
Density
Log(YFP/RFP)
B
Cumulative fraction of TFs
Density
Reb1
Rsc3
1.0
0.5
0
1/frequency (bp): 16 510 16,000 520,000
TFBSs/10^7 library: 4.3x10^7 1.3x10^6 42,000 1300
C
Density
Poly T
Poly A
N80
Rap1
TATA
N80
-2 0 2 4 6
log2(YFP/RFP)

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. Shown are the 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, showing reproducibility. (**C**) Computational "billboard" model. Model relates observed promoter DNA sequence (DNA_p) to expression ($Expression_p$) based on TF binding and activity. TF motifs (PWM_{TF}) are provided as input to the model, but can later be refined (**Methods**) and are used to calculate K_ds for each potential TFBSs. Three parameters are learned per TF (orange): $Concentration_{TF}$ determines the amount of binding to each TFBS given its K_d, $Potentiation_{TF}$ captures each TF's ability to open/close chromatin, and $Activity_{TF}$ captures how each TF impacts transcription. Latent variables (blue) are calculated directly from the inputs and learned parameters. (**D**) Model of TF activation. The model assumes linear activation of TFs, so 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 (Act_x). (**E**) The model accurately predicts the expression from new random DNA promoter sequences. Scatter plots show actual expression level (*y*-axis) for high-quality test data in the pTpA promoter scaffold grown in glucose *vs*. the predicted expression of the sequences by the pTpA+Glu model (*x* axis). (**F,G**) Weaker predictions of endogenous mRNA synthesis rates. Scatter plots shows inferred mRNA synthesis rates of yeast promoters (from (28)) (*y* axis) *vs*. the predicted expression of those promoters by the pTpA+Glu model (F) and the Abf1TATA model (G). Red lines: GAM lines of best fit (**Methods**).

FIGURE 2

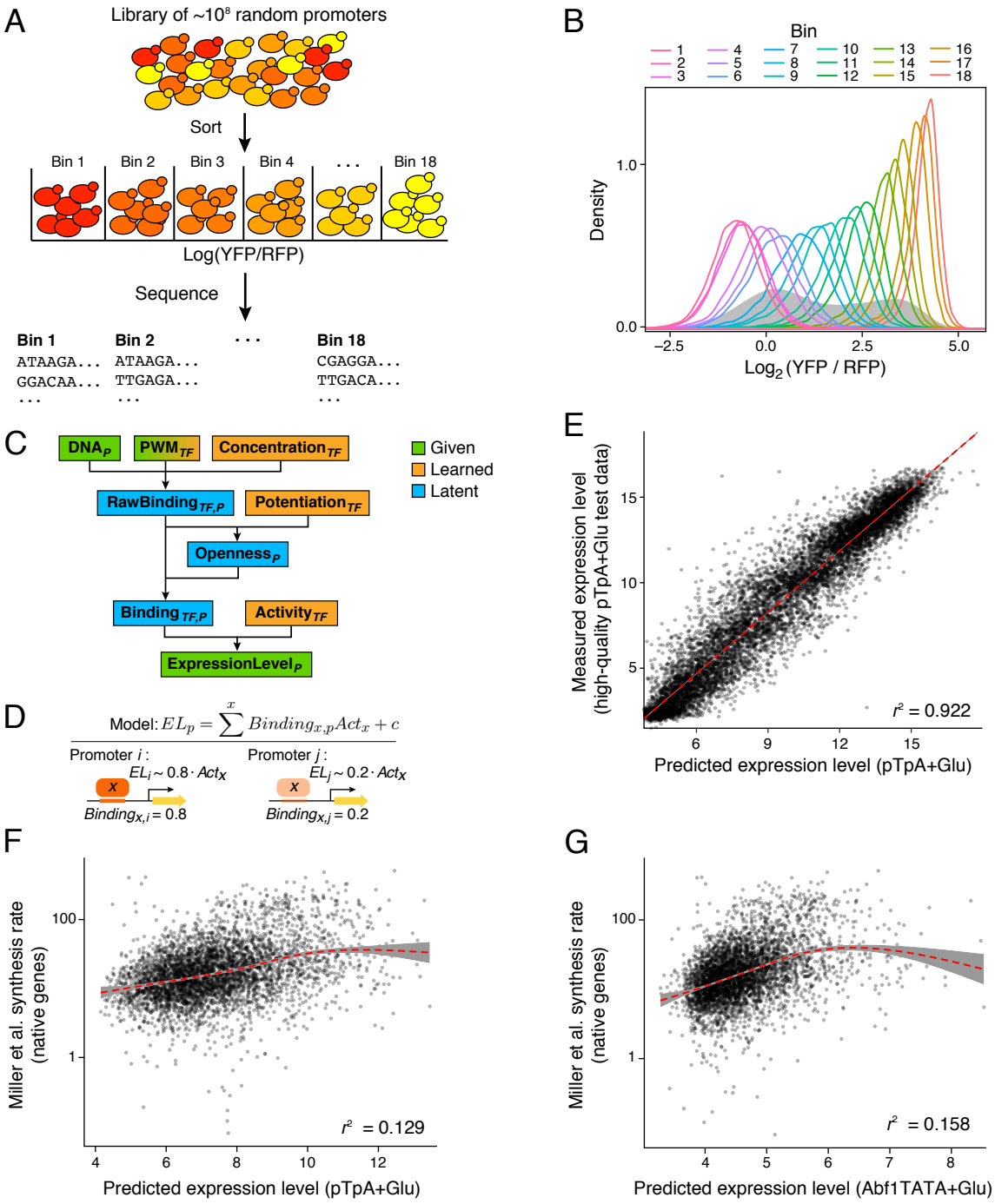

A
Library of ~10⁸ random promoters
Sort
Bin 1
Bin 2
Bin 3
Bin 4
. . .
Bin 18
Log(YFP/RFP)
Sequence
Bin 1
ATAAGA...
GGACAA...
...
Bin 2
ATAAGA...
TTGAGA...
...
. . .
Bin 18
CGAGGA...
TTGACA...
...
B
Bin
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Density
1.0
0.0
−2.5
0.0
2.5
5.0
Log2 (YFP / RFP)
C
DNAP
PWMTF
ConcentrationTF
RawBindingTF,P
PotentiationTF
OpennessP
BindingTF,P
ActivityTF
ExpressionLevelP
Given
Learned
Latent
D
Model: $EL_p = \sum^{x} Binding_{x,p} Act_x + c$
Promoter i :
$EL_i \sim 0.8 \cdot Act_X$
X
$Binding_{X,i} = 0.8$
Promoter j :
$EL_j \sim 0.2 \cdot Act_X$
X
$Binding_{X,j} = 0.2$
E
Measured expression level
(high-quality pTpA+Glu test data)
15
10
5
6
9
12
15
r² = 0.922
Predicted expression level (pTpA+Glu)
F
Miller et al. synthesis rate
(native genes)
100
1
4
6
8
10
12
r² = 0.129
Predicted expression level (pTpA+Glu)
G
Miller et al. synthesis rate
(native genes)
100
1
4
5
6
7
8
r² = 0.158
Predicted expression level (Abf1TATA+Glu)

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* axis) *vs*. either galactose (**A**) or glycerol (**B**) (*y* axis). The GRFs, with known chromatin opening ability in all conditions, are indicated in blue, and known and putative carbon source-specific regulators are marked in red. (**C,D**) Prediction of chromatin accessibility. (**C**) Heatmap shows the 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 (41)), 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).

FIGURE 3

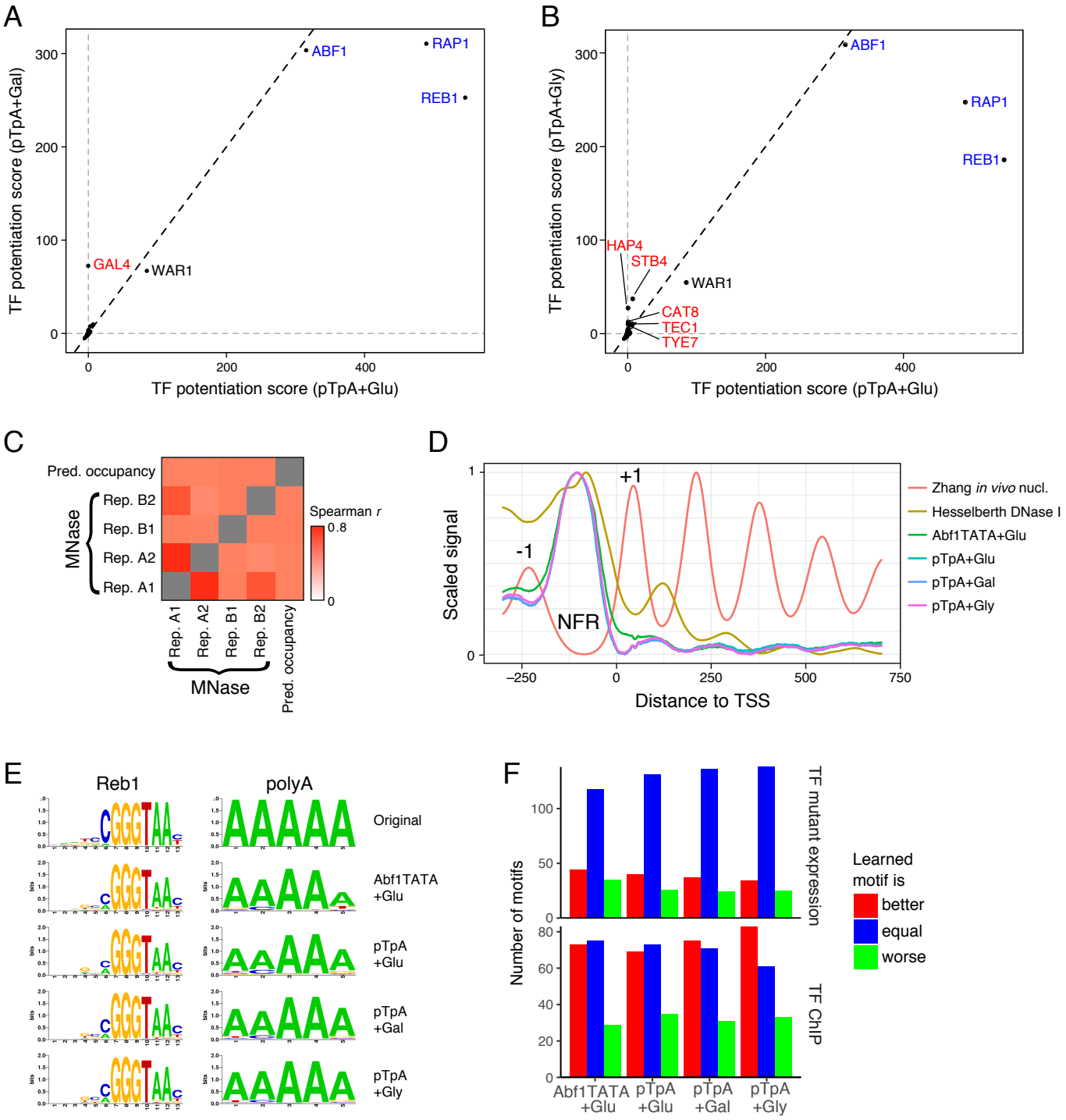

A
TF potentiation score (pTpA+Gal)
TF potentiation score (pTpA+Glu)
ABF1
RAP1
REB1
GAL4
WAR1
B
TF potentiation score (pTpA+Gly)
TF potentiation score (pTpA+Glu)
ABF1
RAP1
REB1
HAP4
STB4
WAR1
CAT8
TEC1
TYE7
C
Pred. occupancy
MNase
Rep. B2
Rep. B1
Rep. A2
Rep. A1
Spearman r
0.8
0
Rep. A1
Rep. A2
Rep. B1
Rep. B2
MNase
Pred. occupancy
D
Scaled signal
+1
-1
NFR
Distance to TSS
Zhang in vivo nucl.
Hesselberth DNase I
Abf1TATA+Glu
pTpA+Glu
pTpA+Gal
pTpA+Gly
E
Reb1
polyA
Original
Abf1TATA +Glu
pTpA +Glu
pTpA +Gal
pTpA +Gly
F
Number of motifs
TF mutant expression
TF ChIP
Abf1TATA +Glu
pTpA +Glu
pTpA +Gal
pTpA +Gly
Learned motif is
better
equal
worse

Figure 4. Only GRFs show nonlinear transcriptional activity. (A) Lingering expression relationships for well- and poorly-fit TFs. Shown are simulated relationships 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. (**B**) Most TFs binding is captured well, with the notable exception of the 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 have particularly poor fits.

FIGURE 4

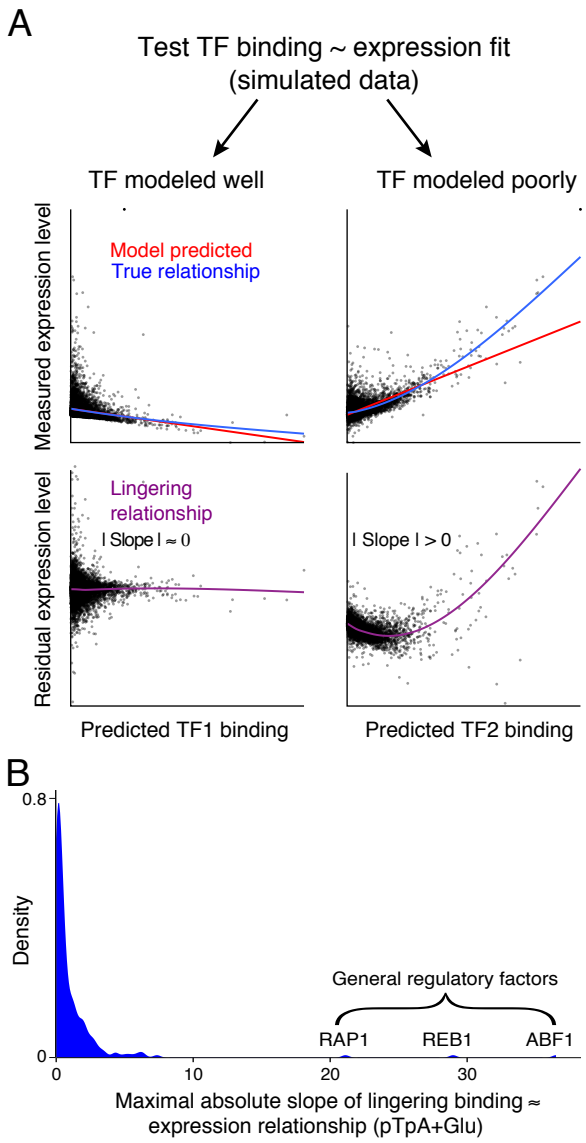

A
Test TF binding ~ expression fit
(simulated data)
TF modeled well
TF modeled poorly
Measured expression level
Model predicted
True relationship
Residual expression level
Lingering
relationship
| Slope | ≈ 0
| Slope | > 0
Predicted TF1 binding
Predicted TF2 binding
B
Density
0.8
0
0
10
20
30
General regulatory factors
RAP1
REB1
ABF1
Maximal absolute slope of lingering binding ≈
expression relationship (pTpA+Glu)

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 (46), 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.

FIGURE 5

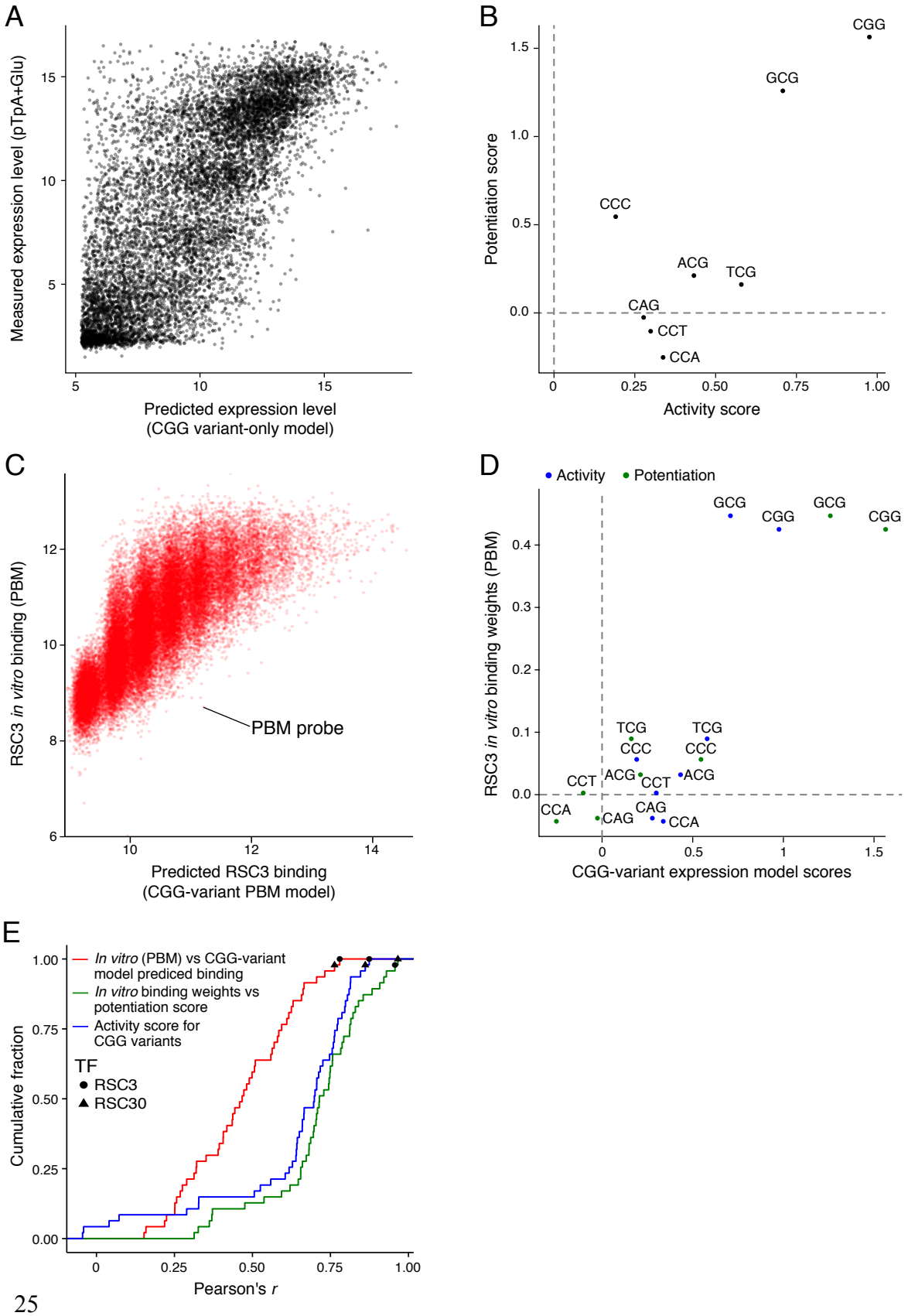
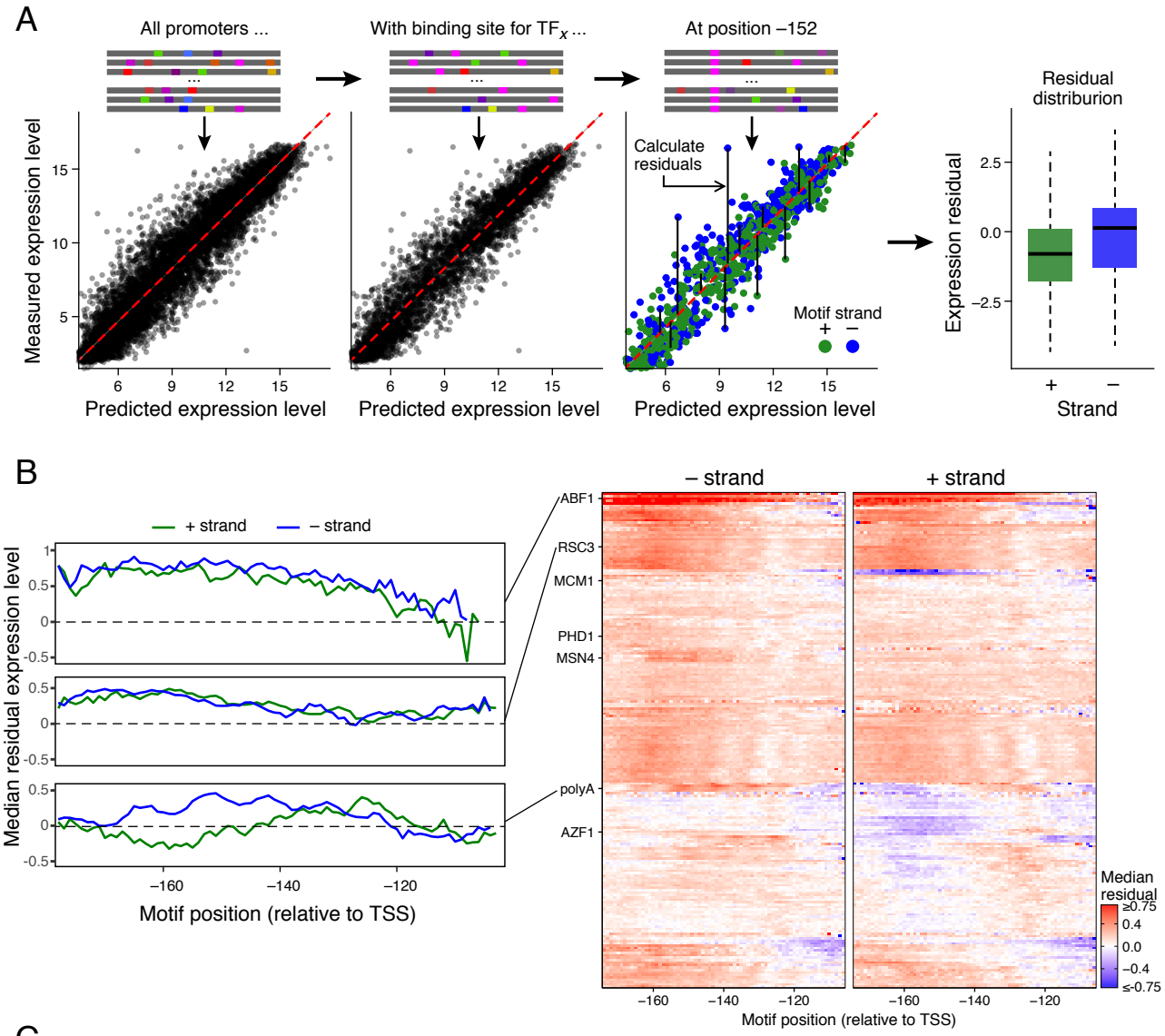
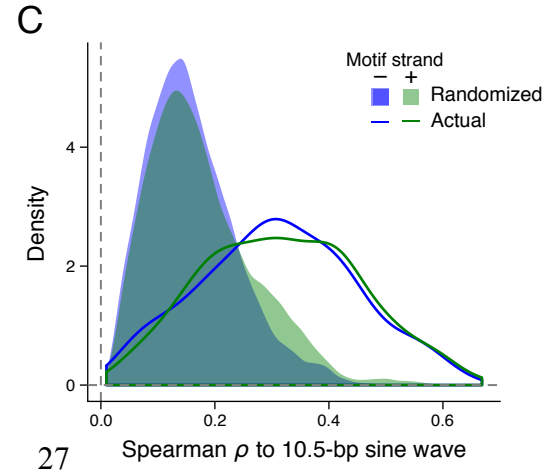

A
Measured expression level (pTpA+Glu)
Predicted expression level
(CGG variant-only model)
B
Potentiation score
Activity score
CGG
GCG
CCC
ACG
TCG
CAG
CCT
CCA
C
RSC3 in vitro binding (PBM)
Predicted RSC3 binding
(CGG-variant PBM model)
PBM probe
D
Activity
Potentiation
RSC3 in vitro binding weights (PBM)
CGG-variant expression model scores
E
In vitro (PBM) vs CGG-variant model prediced binding
In vitro binding weights vs potentiation score
Activity score for CGG variants
TF
RSC3
RSC30
Cumulative fraction
Pearson's r

Figure 6. Position, orientation, and helical face preferences among yeast TFs. (**A**) Identifying position and orientation-specific activities for TFs. A subset of promoters is identified by a specific TFBS at a specific position (horizontal arrows, top). The residuals from the model fit are calculated separately when the motif is on the plus or minus strand, and their medians are determined. (**B**) Motif position and orientation effects on expression. Left: The median residual (*y* axis, units are expressed in expression bins) for promoters with indicated motifs in each position (*x* axis) and orientation (green/blue: +/- strand). Right: Median residual (color) for each TFBS (rows) at each position (columns) for minus (left) and plus (right) strand orientation. (**C**) Helical face preferences. Shown is the distribution of Spearman ρs between a 10.5 bp sine wave with the median residual of TFBSs per position (as in **Figure S6B**) along the minus strand (blue line) and plus strand (green line) or with corresponding randomized data (blue and green shaded areas).

FIGURE 6


A
All promoters ...
With binding site for TF$_X$ ...
At position −152
Measured expression level
Predicted expression level
Calculate residuals
Motif strand
+ −
Residual distriburion
Expression residual
Strand
B
+ strand
− strand
Median residual expression level
Motif position (relative to TSS)
ABF1
RSC3
MCM1
PHD1
MSN4
polyA
AZF1
Median residual
≥0.75
0.4
0.0
−0.4
≤-0.75


C
Motif strand
− +
Randomized
Actual
Density
Spearman ρ to 10.5-bp sine wave

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. 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.
7. Gertz J, Siggia ED, & Cohen BA (2009) Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature* 457(7226):215-218.
8. 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.
9. Jolma A*, et al.* (2013) DNA-binding specificities of human transcription factors. *Cell* 152(1-2):327-339.
10. Nutiu R*, et al.* (2011) Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nature biotechnology* 29(7):659-664.
11. 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.
12. Winter G, Griffiths AD, Hawkins RE, & Hoogenboom HR (1994) Making antibodies by phage display technology. *Annu Rev Immunol* 12:433-455.
13. Cuperus JT*, et al.* (2017) Deep learning of the regulatory grammar of yeast 5' untranslated regions from 500,000 random sequences. *Genome research*.
14. Yona AH, Alm EJ, & Gore J (2017) Random Sequences Rapidly Evolve Into De Novo Promoters. *bioRxiv*.
15. Wunderlich Z & Mirny LA (2009) Different gene regulation strategies revealed by analysis of binding motifs. *Trends in genetics : TIG* 25(10):434-440.
16. 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.
17. 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.

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. Kulkarni MM & Arnosti DN (2003) Information display by transcriptional enhancers. *Development* 130(26):6569-6575.
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. Segal E & Widom J (2009) From DNA sequence to transcriptional behaviour: a quantitative approach. *Nat Rev Genet* 10(7):443-456.
27. 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.
28. Miller C*, et al.* (2011) Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Mol Syst Biol* 7:458.
29. Lipson D*, et al.* (2009) Quantification of the yeast transcriptome by single-molecule sequencing. *Nature biotechnology* 27(7):652-658.
30. 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.
31. 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.
32. Morse RH (1993) Nucleosome disruption by transcription factor binding in yeast. *Science* 262(5139):1563-1566.
33. 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.
34. 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.

35. 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.
36. 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.
37. 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.
38. Madhani HD & Fink GR (1997) Combinatorial control required for the specificity of yeast MAPK signaling. *Science* 275(5304):1314-1317.
39. Gavrias V, Andrianopoulos A, Gimeno CJ, & Timberlake WE (1996) Saccharomyces cerevisiae TEC1 is required for pseudohyphal growth. *Mol Microbiol* 19(6):1255-1263.
40. 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.
41. Hesselberth JR*, et al.* (2009) Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature methods* 6(4):283-289.
42. Zhang Z*, et al.* (2011) A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. *Science* 332(6032):977-980.
43. Harbison CT*, et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* 431(7004):99-104.
44. Hibbs MA*, et al.* (2007) Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics* 23(20):2692-2699.
45. 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.
46. 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.
47. Zhu C*, et al.* (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome research* 19(4):556-566.
48. 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.
49. Akache B, Wu K, & Turcotte B (2001) Phenotypic analysis of genes encoding yeast zinc cluster proteins. *Nucleic acids research* 29(10):2181-2190.
50. 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.

51. Krietenstein N*, et al.* (2016) Genomic Nucleosome Organization Reconstituted with Pure Proteins. *Cell* 167(3):709-721 e712.
52. de Boer C & Hughes TR (2015) The RSC complex may be the poly-A nucleosome turnstile mechanism. *figshare*.
53. de Boer C & Hughes TR (2015) Model for how poly-dA:dT sites act as nucleosome turnstiles. *figshare*.
54. Yu L & Morse RH (1999) Chromatin opening and transactivator potentiation by RAP1 in Saccharomyces cerevisiae. *Molecular and cellular biology* 19(8):5279-5288.
55. 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.
56. 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.
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.