

# Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding

Daniel D. Le<sup>a,1</sup>, Tyler C. Shimko<sup>a,1</sup>, Arjun K. Aditham<sup>b,c</sup>, Allison M. Keys<sup>c,d</sup>, Yaron Orenstein<sup>e</sup>, and Polly M. Fordyce<sup>a,b,c,f,2</sup>

<sup>a</sup>Department of Genetics, Stanford University, Stanford, CA 94305; <sup>b</sup>Department of Bioengineering, Stanford University, Stanford, CA 94305; <sup>c</sup>Stanford ChEM-H, Stanford University, Stanford, CA 94305; <sup>d</sup>Department of Chemistry, Stanford University, Stanford, CA 94305; <sup>e</sup>Department of Electrical and Computer Engineering, Ben-Gurion University, Beer-Sheva, Israel, POB 653; <sup>f</sup>Chan Zuckerberg Biohub, San Francisco, CA 94158

This manuscript was compiled on September 25, 2017

**Transcription factors (TFs) are primary regulators of gene expression in cells, where they bind specific genomic target sites to control transcription. Quantitative measurements of TF-DNA binding energies can improve the accuracy of predictions of TF occupancy and downstream gene expression *in vivo* and further shed light on how transcriptional networks are rewired throughout evolution. Here, we present a novel sequencing-based TF binding assay and analysis pipeline capable of providing quantitative estimates of binding energies for more than one million DNA sequences in parallel at high energetic resolution. Using this platform, we measured the binding energies associated with all possible combinations of 10 nucleotides flanking the known consensus DNA target for two model yeast TFs, Pho4 and Cbf1. A large fraction of these flanking mutations change overall binding energies by an amount equal to or greater than consensus site mutations, suggesting that current definitions of TF binding sites may be too restrictive. By systematically comparing estimates of binding energies output by deep neural networks (NN) and biophysical models trained on these data, we establish that dinucleotide specificities are sufficient to explain essentially all variance in observed binding behavior, with Cbf1 binding exhibiting significantly more epistasis than Pho4. NN-derived binding energies agree with orthogonal biochemical measurements and reveal that dynamically occupied sites *in vivo* are both energetically and mutationally distant from the highest-affinity sites.**

protein-DNA binding | transcription factor binding | flanking sequence  
| protein-DNA specificity | transcriptional regulation

Gene expression is extensively regulated by transcription factors (TFs) that bind genomic sequences to activate or repress transcription of target genes (1). The strength of binding between a TF and a given DNA sequence at equilibrium depends on the change in Gibbs free energy ( $\Delta G$ ) of the interaction (2–5). Thermodynamic models that explicitly incorporate quantitative estimates of binding energies more accurately predict occupancies, rates of transcription, and levels of gene expression *in vivo* (4, 6–11). In addition, binding energy measurements for TF-DNA interactions can provide insights into the evolution of regulatory networks. Unlike coding sequence variants that manifest at the protein level to influence fitness, non-coding TF target site variants affect phenotype by modulating the binding energies of these interactions to affect gene expression (12, 13). Understanding how TFs identify their cognate DNA target sites *in vivo* and how these interactions change during evolution therefore requires the ability to accurately estimate binding energies for a wide ranges of sequences.

Most high-throughput efforts to develop accurate models

of TF binding specificities have focused on the effects of mutations within known TF target sites that dramatically change binding energies. However, even subtle changes in binding energies can have dramatic effects on both occupancy and levels of transcription (14–17). Sequences surprisingly distal from a known consensus motif can affect affinities and levels of transcription (18–21), and genomic variants in regulatory regions outside of known TFBSs may be subject to non-neutral evolutionary pressures (22). Therefore, understanding the fundamental mechanisms that regulate transcription requires the ability to not only measure binding energies for a large number of sequences, but to do so at sufficient resolution to resolve even small effects.

Despite the utility of comprehensive binding energy measurements, existing characterization methods often lack the energetic resolution and scale required to yield such datasets. Currently, three dominant technologies are used to query DNA specificities: methods based on systematic evolution of ligands by exponential enrichment (SELEX) (23–28), protein binding microarrays (PBMs) (29, 30), and mechanically-induced trapping of molecular interactions (MITOMI) (31, 32). SELEX-based methods require repeated enrichment and amplification cycles followed by sequencing of the enriched TF-bound material. Thus, these methods are optimized to identify the

## Significance Statement

Transcription factors (TFs) are key proteins that bind DNA targets to coordinate gene expression in cells. Understanding how TFs recognize their DNA targets is essential for understanding how variations in regulatory sequence disrupt transcription to cause disease. Here, we develop a novel high-throughput assay and analysis pipeline capable of measuring absolute binding energies for over one million sequences with high resolution, and apply it towards understanding how nucleotides flanking DNA targets affect binding energies for 2 model yeast TFs. Through systematic comparisons between models trained on these data, we establish that considering dinucleotide interactions is sufficient to accurately predict binding, and further show that sites used by TFs *in vivo* are both energetically and mutationally distant from the highest-affinity sequence.

D.D.L., T.C.S. and P.M.F. conceived of the study and designed experiments. D.D.L., T.C.S. and P.M.F. wrote the manuscript. D.D.L., T.C.S., A.K.A. and A.M.K. performed experiments and analysis. Y.O. contributed to analytic algorithm development.

The authors declare no conflicts of interest.

<sup>1</sup>D.D.L. and T.C.S. contributed equally to this work.

<sup>2</sup>To whom correspondence should be addressed. E-mail: pfordyce@stanford.edu

highest affinity substrates from extremely large random populations, but fail to measure moderately- or weakly- bound sequences, much less measure their affinities. Although PBMs quantify the binding of TFs to DNA microarrays using a fluorometric readout with a broad dynamic range, the precise relationship between measured fluorescence intensities and binding energies is unclear because PBMs require wash steps that disrupt binding equilibrium. The MITOMI platform, based on mechanical trapping of molecular interactions by microfluidic valves, enables high-resolution measurements of concentration-dependent to yield absolute affinities, but is limited to characterization of several hundred sequences. Recent iterations of MITOMI-based binding assays have addressed these throughput limitations by employing massively parallel sequencing to increase sequence space coverage, but at the cost of resolving binding energies (28, 33). Other techniques such as high-throughput sequencing-fluorescent ligand interaction profiling (HiTS-FLIP) couple massively parallel sequencing with the ability to perform concentration-dependent binding measurements; however, adoption of this technology has been limited by the requirement for extensively customized sequencing hardware (34). Taken together, these TF-DNA binding assays can sample vast sequence spaces, but it remains challenging to simultaneously measure binding energies at the scale and resolution necessary to derive complete binding energy landscapes.

The energetic resolution of direct binding energy measurements is fundamentally limited to  $\sim 1$  RT (or  $\sim 0.5$  kcal/mol) by intrinsic thermal noise. However, measuring binding to a large library of sequences can yield even higher precision estimates of binding energies for individual sequences by training specificity models on large amounts of noisy data. The most popular and widely used models represent TF specificities as a position weight matrix (PWM), in which each nucleotide at each position contributes additively and independently to overall binding energies (35). These mononucleotide models are easily implemented, visualized, and interpreted, and provide useful approximations of binding specificity for the majority of studied TFs (9, 36–38). However, PWM-based models fail to capture epistasis between nucleotides, which can lead to inaccurate predictions, particularly for low-affinity sites (31, 39). This approximation can be further refined by including contributions of higher-order sequence features, such as dinucleotides or longer k-mers (40–49). Several recently developed mechanistic models predict binding based on biophysical properties of DNA, such as the local shape of DNA sequences (*e.g.* minor groove width, propeller twist, helical twist, and roll) (50–54). However, these biophysical variables are predicted directly from primary sequence, rendering the relationship between the two somewhat degenerate. Deep neural network (NN)-based models have shown considerable success in learning complex patterns from large datasets across a variety of applications, including predicting the function of non-coding genomic sequences (55). Training NN models on large sets of binding data therefore has the potential to yield more accurate and higher-resolution estimation of binding behaviors at a per-sequence level, revealing local topography on of binding energy landscapes.

To address the need for technologies capable of high-throughput thermodynamic measurements, we developed an integrated high-throughput sequencing assay and analysis

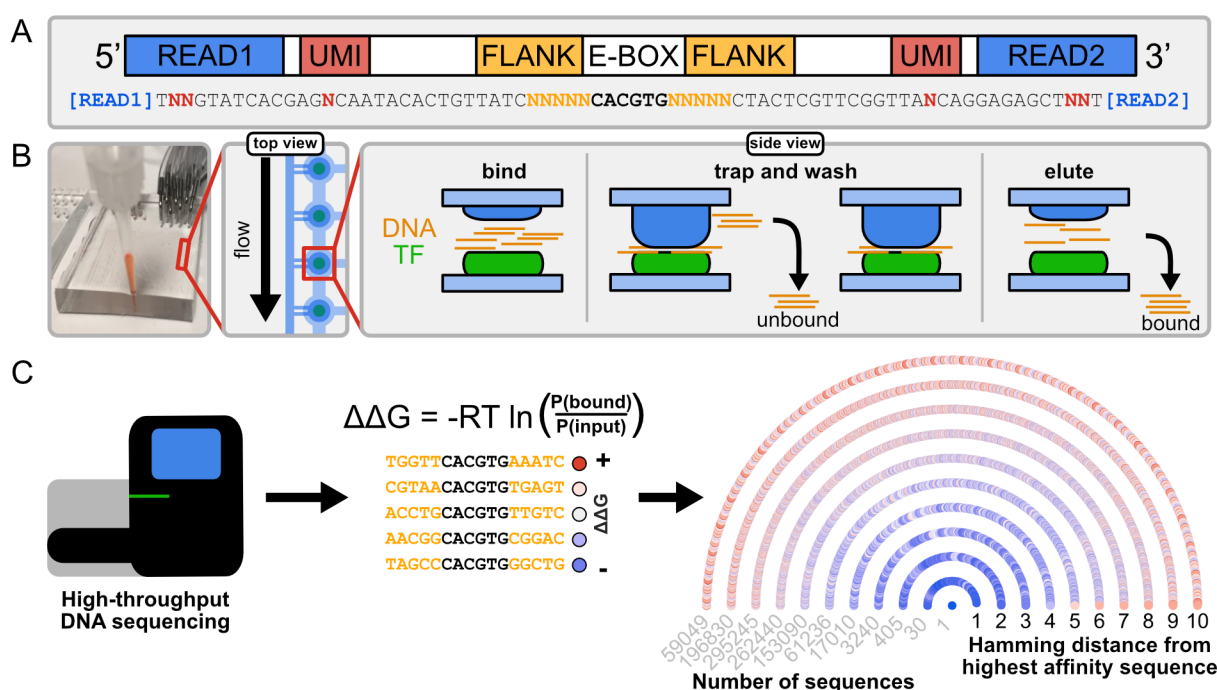
pipeline capable of estimating both relative and absolute binding energies ( $\Delta\Delta G$  and  $\Delta G$ , respectively) for  $>1$  million sequences in parallel, even for relatively small energetic differences. Using Monte Carlo simulations designed to mimic the effects of stochastic sampling noise on energetic resolution, we establish guidelines for the sequencing depth required to resolve accurate binding energies for libraries of different sizes and expected energy ranges. We then deploy this assay to measure comprehensive and quantitative binding energy landscapes for  $>1$  million mutations surrounding the known consensus motif for two model yeast TFs (Pho4 and Cbf1). Deep neural network (NN) models that incorporate all possible higher-order, non-additive contributions were then trained on these large datasets to yield high-resolution estimates of binding specificity for each sequence. Comparisons to orthogonal biochemical affinity measurements established that NN predictions are highly quantitative, accurately predicting measured binding energies over a range of 3 kcal/mol. A surprisingly large number of sequences tested have effects on binding energies as great or greater than mutations in the core, suggesting that current definitions of TFBSs are too restrictive and may limit accurate predictions of TF occupancy *in vivo*. Comparisons between NN-derived predictions and predictions derived from a series of biophysically motivated models reveal that dinucleotide specificity preferences are sufficient to explain nearly all observed binding behavior, with Cbf1 exhibiting significantly more epistasis than Pho4. Strikingly, most dynamically occupied target loci for both Pho4 or Cbf1 are mutationally distal from the energetically optimal flanking sequence representing the global minimum of the energy landscape, providing evidence of evolutionary molecular selection for near-neutral effects on binding energies. Taken together, these data demonstrate the utility of our high-throughput approach to measure relative binding energies and model determinants of substrate specificity required to understand biological behaviors. Furthermore, the experimental design and analysis frameworks presented here may be extended to a wide variety of currently incompletely characterized TFs, improving predictive models of TF-DNA affinities across species.

## Results

**A microfluidic approach employing high-throughput sequencing to derive comprehensive binding affinity landscapes.** We sought to develop an assay that significantly extends the scale at which TF-DNA interactions can be probed while maintaining the ability to quantitatively measure binding energies at high resolution. TF-DNA interactions can be considered a two-state system, such that the affinity of a given interaction can be determined by the equilibrium partitioning of sequences into bound and unbound states:

$$\Delta G = -RT \ln \left( \frac{[TF \cdot DNA_{bound}]}{[TF_{unbound}][DNA_{unbound}]} \right)$$

In recent work, several groups have established that molecular counting of individual DNA molecules via high-throughput sequencing can reliably measure bound and input concentrations for each species (10, 39, 56). These assays generally employ electromobility shift assays (EMSAs) to isolate bound material. However, TF-DNA complexes are not at chemical equilibrium during the electrophoresis step and complexes with particularly fast dissociation rates may be underrepresented



**Fig. 1.** DNA library design and assay overview. (A) Schematic of flanking sequence library design indicating locations of Illumina sequencing adapters (blue), unique molecular identifiers (UMIs, red), variable flanking regions (orange), and E-box consensus. (B) Photograph of MITOMI device (left) and schematic showing device operation (right). (C) Schematic showing downstream sample analysis. Counting of individual molecules within bound and input fractions allows calculation of relative binding energies for each sequence (left, middle) to yield a comprehensive thermodynamic binding affinity landscape (Pho4 example). Each color-coded point (blue =  $-\Delta\Delta G$ , red =  $+\Delta\Delta G$ ) represents a sequence, grouped by Hamming distance from the highest affinity sequence and alphabetically ordered in clockwise polar coordinates.

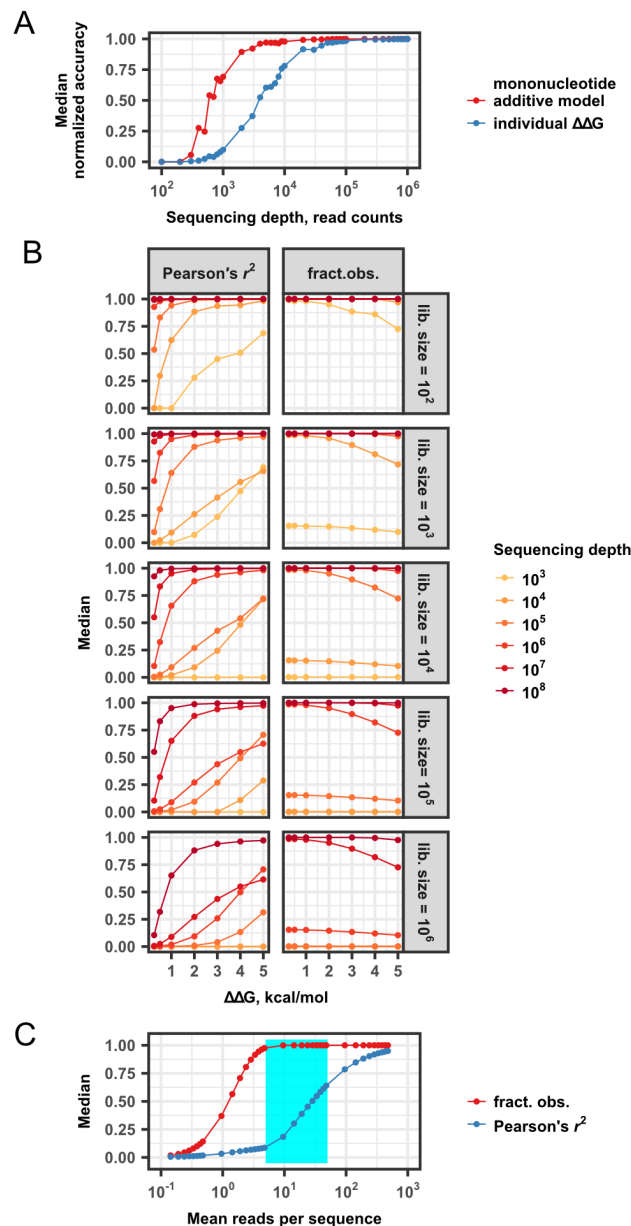
within the bound fraction, leading to a systematic underestimation of weak affinity interactions (57). To address this issue, we employed a microfluidic device incorporating pneumatic valves with fast (~100 ms) actuation times to mechanically “trap” DNA associated with TF proteins at equilibrium (28, 31–33) (Fig. 1A). This device requires very small amounts of both DNA substrate and expressed protein, eliminating the need for cell-based protein production. Antibody-patterned surfaces within the device capture meGFP-tagged TFs produced via *in vitro* transcription/translation prior to washing, effectively purifying the protein *in situ*. After TF capture, libraries of DNA sequences are introduced and allowed to interact with surface-immobilized TFs until equilibrium is reached. Mechanical valves then sequester TF-bound DNA sequences, making it possible to wash out unbound material without loss of weak interactions (31, 32, 58) (Fig. 1B). Isolated bound DNA species can then be eluted from the device and quantified using high-throughput sequencing (Fig. 1C). The concentration of DNA within the device (~1  $\mu\text{M}$ ) is in excess of the immobilized TF concentration (~30 nM), allowing approximation of unbound concentrations via sequencing of the input library.

As a first application of this assay, we focused on two model TFs from *Saccharomyces cerevisiae*, Pho4 and Cbf1. Although Pho4 and Cbf1 are known to bind the same CACGTG variant of the 6-nucleotide enhancer-box (E-box) motif both *in vitro* and *in vivo* (15, 17, 27, 59–61), they bind largely non-overlapping sets of genomic loci and regulate distinct sets of target genes (18, 62). To comprehensively probe how flanking nucleotides affect Pho4 and Cbf1 binding affinities, we designed a library of 1,048,576 sequences in which the core E-box sequence was flanked by all possible random combina-

tions of five nucleotides upstream and downstream, embedded within a constant sequence empirically shown to exhibit negligible binding (32) (Fig. 1A). Because only sparse quantities of material are eluted from the device at the end of each experiment, eluted material must be amplified by PCR prior to Illumina sequencing. Constant sites at the 5' and 3' ends allowed simultaneous PCR amplification and incorporation of Illumina adapters, and unique molecular identifiers (UMIs) included within each library sequence allowed accurate counting of library species even in the presence of PCR bias (63). Each UMI barcode was segmented and interspersed along the library sequence to prevent the formation of an additional CACGTG consensus site. After sequencing, relative binding affinities ( $\Delta\Delta G$ s) were calculated for all sequences by considering relative enrichment of individual DNA species in the TF-bound fraction compared to the input library, thereby generating a comprehensive binding affinity landscape (example shown in Fig. 1C).

**Assay simulations determine sequencing depth requirements for binding affinity measurements.** Accurately estimating concentrations of DNA in TF-bound and input samples via sequencing requires that measured read counts reflect true abundances. However, read counts can be distorted by stochastic sampling error, particularly for low read count numbers (64, 65). To understand how stochastic sampling error depends on read depth, library size, and the expected range of binding energies across library sequences, we began by considering a previously published experiment that quantified interactions between the *E. coli* LacI repressor and a library of 1,024 binding site variants via deep sequencing (library R3.2) (39). Each sequence was sampled to an average depth





**Fig. 2.** Probing the relationship between assay accuracy, read depth, library size, and energy range. (A) Normalized median accuracy (Pearson's  $r^2 * f$ , where  $f$  = the fraction of observed species) scores comparing results for downsampled data with "true" values as a function of read depth. The mononucleotide model reference values were based on PWM predictions that assume perfect additivity (39). Individual  $\Delta\Delta G$  recovery (blue) and model recovery (red) shown. (B) Median squared Pearson's correlation coefficients between recovered and "true" values ( $r^2$ , left) and median fraction of observed species (right) as a function of binding affinity range for various library sizes (rows) sequenced to different depths. (C) Simulations parameterized to reflect putative flank library conditions. Median squared Pearson's correlation coefficient ( $r^2$ ) between recovered and true values (blue) and fraction of observed species (red) as a function of mean reads per sequence for ten replicate simulations; a cyan rectangle denotes the assay configurations presented here.

of roughly  $10^3$  reads per species, yielding  $\Delta\Delta G$  measurements with negligible sampling noise. To understand how read depth affects the recovery of accurate  $\Delta\Delta G$  measurements, we downsampled these data to simulate lower sequencing depths of  $10^2$ – $10^6$  reads, split evenly between bound and input fractions (ca. 0.05–5,000 reads per sequence). We then assessed the accuracy of recovered  $\Delta\Delta G$  values at these lower sequencing depths by calculating the squared Pearson's correlation coefficient ( $r^2$ ) between  $\Delta\Delta G$  values for each species calculated from downsampled data and published values calculated from the full data set. To minimize the effect of a few high accuracy values dominating the correlation statistic, each  $r^2$  was normalized by the fraction of observed species (Fig. 2A). For this 1,024-species library with binding energies that span  $\sim 3$  kcal/mol,  $\sim 2 \times 10^5$  total reads ( $\sim 100$  reads per sequence) were sufficient to recover highly-accurate  $\Delta\Delta G$  values for every sequence.

Measuring accurate  $\Delta\Delta G$ s for a 1,048,576 species library represents a 1000-fold increase in scale from these prior experiments. To understand more generally the determinants of  $\Delta\Delta G$  measurement accuracy, we generated a simulated test set of "true" relative binding energies and implemented Monte Carlo simulations designed to mimic stochastic sampling during high-throughput sequencing. For given combinations of library sizes, sequencing depths, and binding energy ranges, we calculated the median squared Pearson's correlation coefficient between calculated  $\Delta\Delta G$  values and "true" values, as before. As expected, accuracy improves and library coverage expands with increasing sequencing depth (Fig. 2B, S1A, S1B). As the difference in expected binding energies increases, accuracy improves but the fraction of sequences observed from the input library decreases. Nearly all existing motif discovery libraries used in SELEX-type experiments probe on the order of  $10^{18}$ – $10^{24}$  species with read depths of several thousand total reads. Extrapolating from the conditions considered in these simulations, such sparse sequencing depth relative to library size samples only the highest affinity sequences, representing an infinitesimal fraction of the input library. This limited set of observations likely misses smaller peaks within a binding energy landscape.

Previous observations of concentration-dependent Pho4 and Cbf1 binding to E-box motifs with mutations in the first flanking nucleotide revealed differences in affinities spanning  $\sim 1$  kcal/mol (31). To guide sequencing assays, we therefore examined in detail simulations sampling a 1,048,576-member library with this energy range at mean read depths per species ranging between  $10^{-1}$ – $10^3$  ( $10^3$ – $10^8$  total reads) (Fig. 2C). Although 95% of sequences can be recovered from as few as 4–5 reads per species, high read depths of  $\sim 10^2$  counts per species ( $10^8$  total reads per TF) are required to yield individual  $\Delta\Delta G$  measurements with accuracies of  $\sim 80\%$  and errors of  $\sim 0.2$  kcal/mol.

**Modeling specificity from noisy individual measurements improves assay resolution.** The requirement for high-depth sequencing may be cost-prohibitive for studies involving many TFs or when considering large DNA libraries. In those scenarios, modeling can be used to infer the determinants of binding specificity while minimizing stochastic sampling noise from sparse sequencing measurements. To illustrate the power of this approach, we again considered the published LacI repressor dataset (39). Although  $\sim 10^2$  reads per sequence were

required for accurate  $\Delta\Delta G$  estimates,  $10^1$  to  $10^2$ -fold fewer reads per sequence allowed the generation of additive mononucleotide PWM models with similar predictive power to those generated from the entire dataset (Fig. 2A). However, while PWMs yield reasonably good approximations of high affinity sequences, these models fail to explain the variance among lower affinity target sites that exhibit high sequence diversity (31, 39).

A neural network (NN) trained on millions of noisy individual per-sequence measurements can capture all measurable higher-order complexity, thereby yielding a high-resolution model capable of accurately predicting binding over a wide range of energies. However, this increased predictive power comes at the cost of interpretability. To improve the accuracy of our energetic estimates while preserving the ability to gain mechanistic insights, we applied a novel integrated measurement and modeling approach (Fig. 3A). First, we collected millions sequencing-based estimates of per-sequence  $\Delta\Delta G$ s. Next, we trained a NN model on these sequencing data to obtain high-resolution energetic predictions for each substrate that capture the effects of all possible higher-order epistatic interactions among nucleotides. Finally, we parsed and quantified the biophysical mechanisms responsible for observed TF-DNA binding behaviors by systematically comparing correlations between predictions made by the NN model and a series of biophysically-motivated linear models (mononucleotide, nearest-neighbor dinucleotide, and all dinucleotide models). This integrated scheme simultaneously yielded a novel binding energy landscape of unprecedented scale and energetic resolution and allowed dissection of the biophysical mechanisms responsible for Pho4 and Cbf1 specificity.

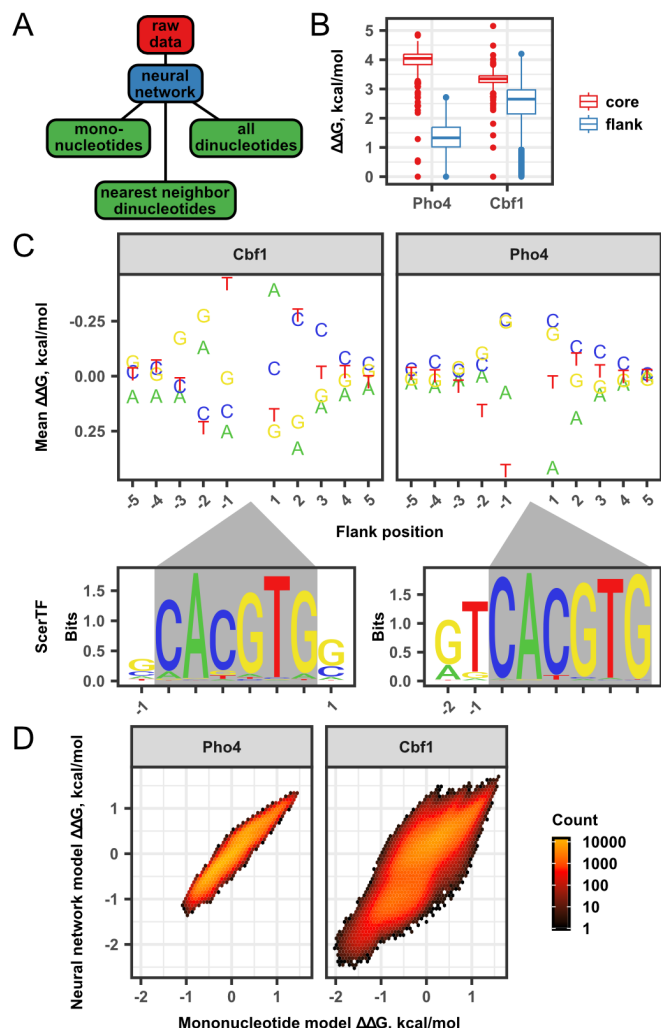
**High-throughput, comprehensive estimates of absolute binding affinities for Pho4 and Cbf1.** We used the assay and DNA library described above to acquire four replicate measurements of Pho4 and three of Cbf1 at sequencing depths ranging from approximately 5 to 50 million reads allocated to either bound or input samples (Tab. S1). For each experiment,  $\Delta\Delta G$ s were calculated for each sequence from the measured ratio of bound to input read counts (Fig. S2). As predicted, measured per-sequence  $\Delta\Delta G$ s between two experiments at low read depth (*ca.* 6–8 million limiting counts) show no correlation; conversely, at higher read depths ( $\sim 24$  million limiting counts), this correlation increases to  $r^2 = 0.67$  (Fig. S3, Tab. S2). To further improve resolution, we trained a NN regression model that predicted the measured  $\Delta\Delta G$  for each input flanking sequence. Accuracy of the network against both the observed training data and an unobserved validation dataset was recorded throughout training. Training of the network was stopped once accuracy against the validation dataset failed to improve, protecting against overfitting to the training data (Fig. S4). Predictions from NN models trained on the 2 high-depth Pho4 replicates showed excellent correlation ( $r^2 = 0.94$ ), validating the ability to apply such models to derive accurate, reproducible estimates of binding energies. For all analyses moving forward, we therefore use per-sequence  $\Delta\Delta G$  estimates output from the NN trained on a composite dataset of all replicates (Fig. S5).

Estimates of absolute binding energy and dissociation constants ( $\Delta G$  and  $K_d$ , respectively) allow direct comparison between different TFs and across experimental platforms, and further enable quantitative predictions of TF occupancy *in vivo*

under known cellular conditions. However, sequencing-based measurements of  $\Delta G$ s from sparse datasets can underestimate the true affinity range due to systematic undersampling of bound reads for low-affinity sequences. In addition, given that the NN is trained only on relative binding affinities ( $\Delta\Delta G$ s), it cannot return estimates of absolute energies ( $\Delta G$ s). NN-derived  $\Delta\Delta G$  estimates can be projected onto an absolute scale by calibrating to a set of high-resolution biochemical measurements of  $\Delta G$ s with a linear scaling factor and offset. To generate a set of high-confidence  $\Delta G$ s, we applied the traditional fluorometric MITOMI assay to measure concentration-dependent binding behavior for surface-immobilized Pho4 and Cbf1 TFs interacting with all single-nucleotide variants of AGACA\_TCGAG, a medium affinity reference flanking sequence (where the underscore indicates the CACGTG core motif) (Fig. S6A, S6B). For each oligonucleotide, observed binding was fit to a single-site binding model, yielding both  $K_d$ s and  $\Delta G$ s for each sequence (Tab. S3). The entire set of NN values was then scaled by fit parameters returned from a linear regression between NN predictions and experimental measurements for this set sequences. The median equilibrium dissociation constants for all flanking library sequences were 100 and 63 nM for Pho4 and Cbf1, respectively, in agreement with prior work (31). Strikingly, sequence variation in the flanking region around the consensus motif of Pho4 and Cbf1 can modulate  $K_d$  values by over two orders of magnitude, ranging between 11–1036 nM and 1–866 nM, respectively. In some cases, the magnitude of these effects exceeds that of mutations within the CACGTG core consensus (Fig. 3B, S7), demonstrating the importance of flanking sequences to specificity.

**Mononucleotide models reveal that Pho4 and Cbf1 flanking preferences extend far beyond the known consensus sequence.** To understand the biophysical features that contribute to the predictive performance of the NN model, we generated PWMs (70), which estimate the mean energetic contribution of each nucleotide at each position, from the full set of scaled NN-predicted  $\Delta\Delta G$  values (Fig. 3C). While the assumption of additivity fails to explain all specificity, these models often offer a close approximation (11, 43) and PWMs are easily visualized and interpreted. These mononucleotide model results confirm that positions proximal to the E-box core motif exhibit the largest mean effect on binding affinity, in agreement with PWMs generated by orthogonal techniques (66–68) (Fig. 3C). However, these results indicate that nucleotides up to 4 and 5 positions from the consensus contribute to specificity for Pho4 and Cbf1, respectively, significantly farther than previously reported.

To quantitatively assess the degree to which mononucleotide features dictate binding behavior, we determined the proportion of NN-derived per-sequence  $\Delta\Delta G$  variance that is explained by a simple PWM (Fig. 3D). If mononucleotide models capture all determinants of observed specificity, PWM predictions would explain all the variance in NN-derived  $\Delta\Delta G$  values. Conversely, discrepancies may indicate the presence of higher-order epistatic interactions not captured by the PWM. PWMs were capable of explaining a majority of the variance in NN predictions ( $r^2 = 0.92$  and  $r^2 = 0.70$  for Pho4 and Cbf1, respectively) (Tab. S4), consistent with the sentiment in the field that PWMs provide good approximations of sequence specificity (9, 11). Intriguingly, PWMs explain a significantly



**Fig. 3.** Modeling and interpretation of binding specificity based on mononucleotide features. (A) Data analysis flowchart shows neural network modeling from raw data, followed by interpretation based on linear combinations of sequence features. (B) Comparison of Pho4 and Cbf1 flanking sequence  $\Delta\Delta G$  measurements with those derived from core mutations (31). (C) DNA base letters represent measured Pho4 and Cbf1 mean mononucleotide  $\Delta\Delta G$  measurements as a function of flanking sequence position, compared to previously derived sequence logos from the ScerTF database (66–69). Gray triangles and gray boxes represent position of the consensus E-box motif CACGTG. (D) For Pho4 and Cbf1, density scatter plots show the correlation between NN model estimates and mononucleotide additive model predictions based on those estimates.

smaller proportion of observed Cbf1 measurement variance, suggesting that Cbf1 recognition may rely on higher-order determinants of specificity.

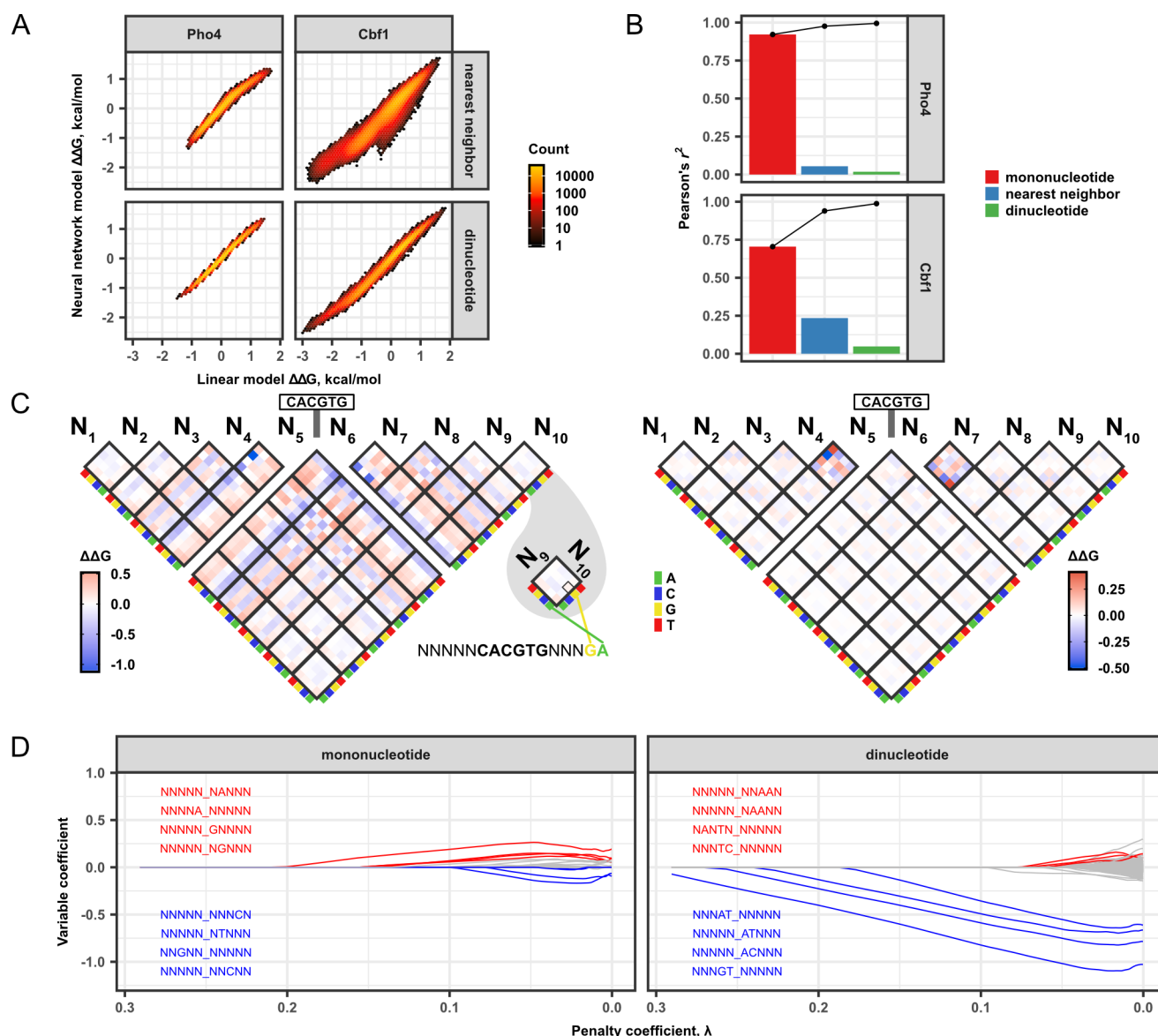
To evaluate assay reproducibility, we generated individual PWMs from each of the four Pho4 technical replicates and three Cbf1 technical replicates [Fig. S8]. Linear model coefficients for mononucleotides at each position were strongly correlated between replicates of a single TF (Pho4  $r^2 = 0.95$ – $0.97$  and Cbf1  $r^2 = 0.79$ – $0.87$ ) and uncorrelated between TFs (Tab. S5, Fig. S9); a meGFP negative control protein exhibited no sequence specificity (Fig. S10). Using the fraction of unexplained variance ( $1-r^2$ ) as a precision metric, the expected error range in NN-derived mean mononucleotide  $\Delta\Delta G$  values for Pho4 is 0.02–0.04 kcal/mol and that of Cbf1 is 0.09–0.16 kcal/mol. This analysis highlights the robustness of binding specificity models derived from the assay and data presented here.

### Dinucleotide models reveal that flanking nucleotides exhibit significant epistasis for Cbf1.

The remaining unexplained variance observed between NN-derived values and PWM-predicted values (~8% and ~30% for Pho4 and Cbf1, respectively) could indicate the presence of higher-order epistatic interactions governing specificity, or could simply represent experimental noise (40) (Tab. S4). To probe for higher-order interactions, we fit two different dinucleotide models to the NN-derived scaled  $\Delta\Delta G$  values: a nearest-neighbor model that explicitly considers contributions from adjacent dinucleotides and a more complex model that considers contributions from all dinucleotide combinations, including non-adjacent pairs (45).

Direct comparisons between nearest-neighbor dinucleotide model-predicted values and NN-derived binding energies showed increased correlation for both Pho4 and Cbf1, with associated  $r^2$  values of 0.98 and 0.94 (Tab. S4, Fig. 3A, 3B). These improvements, corresponding to ~5% and 24% increases in explanatory power over respective mononucleotide models, are consistent with the potential for physically interacting nucleotides to affect binding energies through local structural distortions. Considering all possible dinucleotide features accounts for nearly all of the remaining variance in NN-derived binding energies (improvements of approximately 1% and 5% for Pho4 and Cbf1, respectively). These findings highlight the differential degree to which epistasis defines binding even amongst structurally related TFs, which ultimately determines the predictive power and accuracy of widely-used PWMs.

To visualize and interpret binding energy contributions of dinucleotides alone, we calculated the mean residual  $\Delta\Delta G$  from the linear regression against PWM-predicted  $\Delta\Delta G$  values for all possible dinucleotides both within and across flanking sequences (Fig. 4C). Nucleotide interactions that exhibit measured  $\Delta\Delta G$ s lower than expected based on considering the linear combination of individual mononucleotides are positively epistatic; conversely, interactions that exhibit negative epistasis increase measured  $\Delta\Delta G$ s more than expected. The largest magnitude epistasis is observed for dinucleotides immediately upstream or downstream of the E-box (N4/N5 or N6/N7 pairs), with absolute energetic differences among combinations spanning ~0.5 kcal/mol (approximately 1 RT), and epistatic interactions primarily occur within flanks rather than between them. Both inter- and intra-flank dinucleotides exhibited palindromic arrangements near the core motif, consistent with the expectation of binding site symmetry in homodimeric



**Fig. 4.** NN model interpretation using dinucleotide features. (A) For Pho4 and Cbf1, density scatter plots show the correlation between NN model estimates and dinucleotide (nearest neighbor = only adjacent pairs) additive model predictions based on those estimates. (B) Variance in Pho4 and Cbf1 NN model estimates, captured by squared Pearson's  $r^2$  compared to linear models predictions, attributed to sequence feature groups. (C) Cbf1 heatmap of mean energetic contribution for all dinucleotide configurations (left) and that of mean residual energetic contributions when mononucleotide effects are removed (right). Color scale: blue = high affinity, red = low affinity. (D) Cbf1 LASSO regression penalization traces arranged by type of model feature. Lines represent magnitude of individual sequence feature coefficients as a function of penalty coefficient  $\lambda$ . Labels show flanking sequence feature. Color scale: blue = four most persistent negative energetic contributions, red = four most persistent positive energetic contributions.



TFs like Pho4 and Cbf1 (61, 71). For Cbf1, TT and TG dinucleotides upstream of the motif (and the corresponding downstream palindromes) exhibit large magnitude positive and negative epistasis, respectively. Although the overall magnitude of epistasis is significantly smaller for Pho4, a GG dinucleotide downstream of CACGTG shows strong synergistic effects. Interestingly, the Pho4 crystal structure reveals direct contacts between the Arg2 and His5 residues and this same GG dinucleotide, providing a potential structural basis for this observation (61).

**Incorporating weight constraints into dinucleotide models confirms that Cbf1 interactions are significantly more epistatic.** Models that incorporate additional free parameters should always increase explanatory power. While mononucleotide models attempt to describe all 1,048,576 observed measurements using only 40 free parameters (4 nucleotides per position across 10 positions), the nearest-neighbor dinucleotide model adds another 128 free parameters (16 pairs across 8 positions), and the all dinucleotides model includes 720 free parameters representing the contribution from all combinations of nucleotide identities [ $4^2=16$ ] and positional pairs [ $\binom{10}{2}=45$ ]. In most cases, dinucleotide coefficients are near zero (Fig. 4C), meaning that they contribute little explanatory power. To systematically identify a minimal set of important features that define sequence specificity in an unbiased fashion, we used least absolute shrinkage and selection operator (LASSO) regression to develop parsimonious linear models with weight constraints (72). Briefly, non-zero coefficients in the model are penalized, leading to inclusion of only the most explanatory variables with respect to reduction in squared error (Fig. S11). The regression explores a range of penalization stringencies to distinguish important sequence features based on differential coefficient minimization rates.

We applied LASSO regression to systematically select a parsimonious set of sequence features with optimal explanatory power among all possible dinucleotide features. From the selected Cbf1 features, four nearest neighbor dinucleotides exhibited relatively large initial coefficient magnitudes and persisted as model variables throughout most of the penalization regime. These four sequence features group into two pairs of palindromic dinucleotides spanning the core motif: (NNNAT\_NNNNN, NNNNN\_ATNNN) and (NNNGT\_NNNNN, NNNNN\_ACNNN) (Fig. 4D). Strikingly, this set of dinucleotide features exhibits model coefficients that are up to 3-fold greater in magnitude than that of the most significant mononucleotide feature (Fig. 4D, S12), highlighting the importance of dinucleotides to Cbf1 binding specificity in particular. Among the selected dinucleotide features for both Pho4 and Cbf1, nearest neighbor pairs exhibited the largest coefficient magnitudes (Fig. S12). Short-range nucleotide interactions contribute most to higher-order substrate specificity, consistent with the feature preferences of yeast TFs determined by ChIP-chip (45).

**Orthogonal *in vitro* biochemical measurements confirm results obtained via high-throughput sequencing.** To confirm that NN model predictions provide accurate per-sequence estimates of true binding energies, we quantitatively compared titration-based  $\Delta\Delta G$  values with unprocessed measurements and NN predictions. Using traditional fluorometric MITOMI, we determined the  $\Delta\Delta G$ s for Pho4 and Cbf1 binding to single-

site variants of the ACAGA\_TCGAG flanking sequence (Tab. S3, Fig. S6A, S6B). In addition, we correlated NN predictions with a set of previously reported  $\Delta\Delta G$  measurements of CACGTG flanking site mutations (17, 31). Consistent with the Monte Carlo simulations,  $\Delta\Delta G$  values calculated directly from raw sequencing data showed essentially no correlation to direct measurements, with  $r^2$  values ranging between 0.07–0.16 and 0–0.24 for Pho4 and Cbf1, respectively (Fig. S13). NN-predicted values showed remarkable agreement, with  $r^2$  values ranging between 0.76–0.94 and 0.61–0.69 for Pho4 and Cbf1. In all cases, predictions agreed with observations within  $\sim 1$  kcal/mol. Taken together, these results establish that the Pho4 and Cbf1 NN models presented here yield a complete and accurate measurements of binding energies for  $>1$  million TF-DNA interactions with similar resolution to ‘gold-standard’ biochemical measurements.

### High-resolution *in vitro* affinity measurements can be used to identify biophysical mechanisms underlying *in vivo* behavior.

The role of transcriptional activators *in vivo* is not simply to bind DNA, but rather to bind specific genomic loci and regulate transcription of downstream target genes. The high resolution of these comprehensive binding energy measurements makes it possible to quantitatively estimate the degree to which measured binding affinities explain differences in measured TF occupancies, rates of downstream transcription, and ultimate levels of induction.

First, we compared NN-modeled  $\Delta G$  values with measured rates of transcription and fold change induction for engineered promoters containing CACGTG Pho4 consensus sites with different flanking sequences driving the expression of fluorescent reporter genes (16, 17). As reported previously, rates of transcription and induction scaled with measured  $\Delta G$  values that spans  $\sim 1.5$  kcal/mol (Fig. S14). Next, we compared NN-derived binding energies with reported levels of TF occupancy *in vivo* at CACGTG consensus sites in the *S. cerevisiae* genome for both Pho4 and Cbf1 (62). Large magnitude TF enrichment induced by phosphate starvation was observed at loci with measured  $K_d$  values of around 100 nM or lower (Fig. 5A). While TF enrichment roughly correlated with binding energy, very high affinity sequence showed strikingly low enrichment. The observed nonlinearities may indicate the degree to which other regulatory mechanisms, such as cooperation and competition among TFs or changes in DNA accessibility due to nucleosome positioning, contribute to reported TF enrichment (62). Alternatively, these nonlinearities may reveal the need for higher resolution *in vivo* measurements to test the degree to which binding energies alone dictate occupancies.

To better understand the relationship between binding energies and *in vivo* occupancies, we visualized the complete binding energy landscapes for both Pho4 and Cbf1 as a function of sequence space (Fig. 5B, S15). The single highest affinity sequence for each TF was placed at the center of a series of concentric rings, each of which includes all sequences at a given Hamming distance from the highest affinity sequence. Within each ring, points representing each sequence are arranged in alphabetical order, with the color of each point reporting the measured  $\Delta\Delta G$  for that sequence. As expected, the energetic landscape forms a somewhat rugged funnel, with binding energies increasing monotonically on average with mutational distance from the highest-affinity site (Fig. S16). Next, we projected flanking site occupancies from ChIP-seq



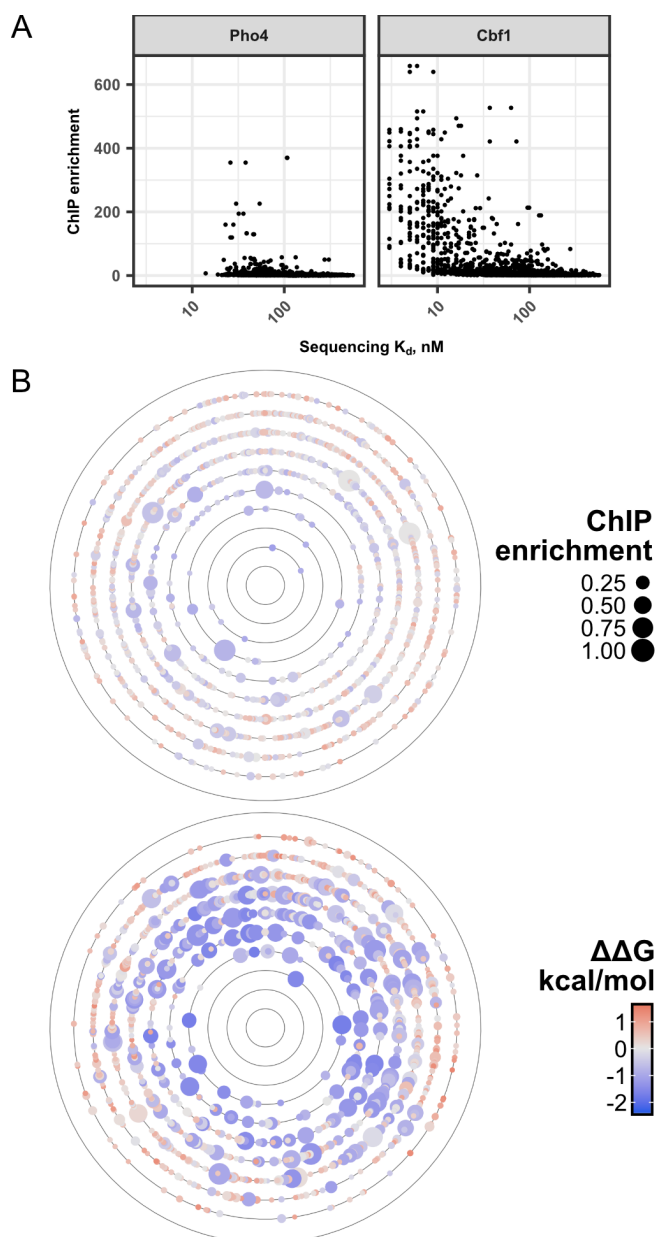
experiments (62) onto these binding affinity landscapes to yield a composite functional-energetic landscape (Fig. 5B). For both Pho4 and Cbf1 the majority of enriched genomic loci are greater than four mutational steps away from the the global minimum, corresponding to mean increases in binding energy of approximately 0.8 and 1.5 kcal/mol, respectively (Fig. S17). These quantitative comparisons between measured affinities and *in vivo* occupancies establish that even relatively small differences in  $\Delta\Delta G$  are associated with differential TF enrichment.

## Discussion

TFs play a central role in regulating levels of gene expression throughout development and allowing organisms to adapt to changing environmental conditions. The ability to quantitatively predict levels of TF occupancy *in vivo* from DNA sequence would therefore be transformative for our understanding of cellular function. Given that the probability of TF occupancy at a given locus includes an exponential dependence on the corresponding TF-DNA binding energy (73), accurate occupancy predictions require the ability to resolve even small (~1-2 kcal/mol) changes in binding energies at high resolution. Towards this goal, we developed a novel assay designed to provide comprehensive and quantitative measurements of near-neutral changes in binding energies caused by mutations in the flanking sequences surrounding TF consensus target sites. By training a neural network (NN) on estimates of binding energies for millions of sequences, we obtained a model that incorporates all higher-order complex interactions required for accurate binding energy estimates for each sequence.

We find that a wide variety of mutations in flanking nucleotides outside of “core motifs” (as defined by consensus PWMs) can change binding energies by an amount equal to or greater than mutations within the core consensus. For example, the difference in binding energy between a *TCCCCCACGTGCCCCA* sequence and a *AATTTACAGTGAAAAG* sequence is ~2.6 kcal/mol, equivalent to mutating the central consensus site from **CACGTG** to **CGTGTG**. However, current representations of TFBSs would predict a change in binding energy for only the core mutation. This discrepancy has several implications for our understanding of transcriptional regulation *in vivo*. First, flanking sequence effects may explain mysteries regarding chromatin immunoprecipitation (ChIP) data in which some genomic loci are occupied despite an apparent lack of a consensus site while other accessible regions containing consensus sites remain unoccupied. Second, many current efforts to infer the presence of bound TFs first analyze DNase-seq or ATAC-seq data to identify regions of accessible DNA, and then scan these regions for putative bound TFs by searching for sequence similarities to known TFBSs. Failing to consider the effects of flanking sequences could return a significant number of both false positives (in which a consensus site with unfavorable flanking nucleotides is considered to be bound) and false negatives (in which a mutated consensus site has strongly favorable flanks but is considered unoccupied).

In practice, measuring complete binding energy landscapes remains rare, with most assay development focused towards discovery of the highest-affinity sequences. The quantitative and complete binding energy landscapes presented here provide a unique opportunity to explore the mechanisms that drive



**Fig. 5.** Large-scale Pho4 and Cbf1 binding energies and *in vivo* activity. (A) Pho4 and Cbf1 equilibrium binding constants ( $K_d$ , in nM) compared to *in vivo* enrichment determined by ChIP-seq (62). (B) Functional-energetic landscapes of differential relative ChIP enrichment at dynamically regulated loci, relative to the measured highest affinity sequence in Hamming distance space. Point size: large = high enrichment, small = low enrichment. Color scale: blue = high affinity, red = low affinity.

evolution of transcriptional regulatory networks. It has been hypothesized that high affinity, but sub-maximal, TF binding sites may be evolutionarily favorable due to the potential for greater dynamic transcriptional control (74). Consistent with this hypothesis, we find that the most highly occupied sites *in vivo* are mutationally distant from the highest affinity flanking sequences. These large mutational distances could indicate an evolutionary buffer employed by an organism to avoid sequence proximity to a sub-optimal binding extreme. In addition, elevated levels of epistasis are thought to produce more rugged energetic landscapes compared to those created by additive binding interactions (75). Given that epistatic dinucleotide interactions play a larger role in determining Cbfl binding specificity, we speculate that Cbfl binding sites can traverse fewer non-deleterious evolutionary pathways than Pho4, ultimately rendering Pho4 binding sites more evolutionarily plastic than those of Cbfl.

Systematic comparisons between per-sequence estimates of binding energies output by a NN and those output by a series of linear models revealed the mechanistic features that drive specificity and quantified their contributions to observed binding energies. These results have relevance to recent debates surrounding the relative utility of DNA sequence-based models (PWMs) and DNA shape-based models representing TF specificity. Both models parameterize DNA binding preferences by a set of four values at each position: PWMs represent specificities based on preferences for each of the DNA bases (A, C, G, and T) (35), while shape-based models express preferences in terms of minor groove width, propeller twist, helical twist, and roll (50–54). While these models significantly enhance the ability to extract mechanistic determinants of specificity from sparse data, higher-order information is lost in the process. Here, we demonstrate that models based on nearest-neighbor dinucleotide preferences are sufficient to fully explain observed binding behavior, consistent with biophysical observations that local DNA structure is largely determined by base stacking interactions and inter-base pair hydrogen bonds in the major groove between adjacent base pairs (54, 76, 77). Such nearest-neighbor dinucleotide models require only a modest increase in the number of required free parameters relative to mononucleotide models. While the NN’s capacity to incorporate higher-order complexity ultimately proved unnecessary for accurately modeling Pho4 and Cbfl binding specificities, high-resolution predictions output by the NN were essential to quantify the degree to which simpler models could explain observed behavior. The high resolution of these measurements further allows direct quantification of the degree to which thermodynamic models based on the binding energies can and cannot predict behavior *in vivo*.

The simulation-guided assay design and experimental assay presented here should allow a much broader diversity of labs to make comprehensive and high-resolution measurements of binding energy landscapes. While the assay is deployed here for a specific use case (high-resolution measurements of near-neutral effects over a small energy range), these simulations can guide choice of sequencing depths to resolve absolute binding energies across a variety of applications and platforms (9, 11, 28, 78, 79), including target site discovery efforts. The experimental assay itself further offers the resolution of traditional MITOMI or HiTS-FLIP fluorescence-based assays while requiring significantly less equipment and infrastructure.

Traditional MITOMI fluorescence assays require both a DNA microarray printer to deposit DNA on slide surfaces and either a high-cost fluorescence scanner or fully automated microscope capable of quantitative fluorescence imaging of a slide with a microfluidic device attached; HiTS-FLIP assays require access to a customized Illumina GAIIX sequencing platform. The use of a sequencing-based readout eliminates these requirements, allowing any laboratory with access to educational or commercial deep sequencing services to measure energies at this scale and resolution. Moreover, the valving required on the microfluidic device itself is significantly simpler than for traditional MITOMI assays, reducing the pneumatics infrastructure required to operate devices.

Finally, this assay provides unique opportunities in future work to probe additional control mechanisms that influence TF binding *in vivo*. Introduction of synthesized DNA libraries containing modified bases involved in epigenetic regulation (*e.g.* 5-methylcytosine, 5-hydroxymethylcytosine, 5-formylcytosine, 5-carboxylcytosine) could allow systematic investigation of how these modifications affect TF specificities. The assay should further be compatible with DNA libraries assembled into nucleosomal arrays *in vitro*, facilitating direct and quantitative investigation of how competition between TFs and nucleosomes dictates occupancies, the degree to which particular TFs act as “pioneer factors”, and how site-specific histone modifications (*e.g.* methylation, acetylation, phosphorylation, ubiquitinylation) (80–82) influence this competition. It may also be possible to measure relative binding energies of TF interactions with extracted DNA or chromatin derived from cells under defined stimulation (*e.g.* drug response, disease) (83). Taken together, these areas of study highlight the complex interplay of chromatin landscape and TF binding, aspects of which may be characterized in the future using a variant of the assay and analysis pipeline presented in this work. In addition, the simulations presented here provide guidelines for the development of sequencing-based assays designed to measure binding energies for a wide variety of macromolecular interactions, including both protein-RNA and protein-protein interactions.

## Materials and Methods

**Binding energy calculations.** Using the model proposed by Djordjevic, *et al.* (56) that was refined by Stormo and colleagues (10, 39), we consider the binding of DNA sequences  $S_i$  among many  $S_j$  competitor substrates at equilibrium with a transcription factor. The probability of being bound to the transcription factor as a function of sequence identity, represented by  $P(\text{bound} | S_n)$ , is given by:

$$P(\text{bound} | S_n) = \frac{e^{-\Delta G_n}}{e^{-\mu} + e^{-\Delta G_n}}$$

Where  $\Delta G_n$  is the binding free energy and  $\mu$  is the chemical potential, equal to the natural log of the transcription factor concentration. Both terms are in units of RT.

Sequencing of TF-bound substrates yields  $P(S_i | \text{bound})$ , which is the inclusion probability of  $S_i$  among the bound substrate distribution. From the partition probability of substrates between bound and unbound,  $P(S_i | \text{bound})$ , application of Bayes’ theorem and the law of total probability yields:

$$P(S_i | \text{bound}) = \frac{P(\text{bound} | S_i)P(S_i)}{\sum_j P(\text{bound} | S_j)P(S_j)}$$

Where  $P(S_n)$  describes the probability of a given species within the input substrate distribution. The combination of these two

equations returns  $P(S_i | bound)$  as a function of binding energies and input probabilities.

Next, we make the following assumptions:

- The transcription factor concentration is substantially lower than the total DNA concentration.
- As the TF concentration is minimized,  $\mu$  approaches negative infinity. In effect, this causes the denominator in equation 1 to be dominated by the  $e^{-\mu}$  term, which we can approximate with a constant  $C \simeq e^{-\mu} + e^{-\Delta G_i}$ .
- The sum of  $P(S_j)$  in a large population is  $\sim$  equal to one.

Applying these assumptions to equation 2, it becomes possible to isolate  $\Delta G_i$  as a function of  $P(S_i | bound)$  and  $P(S_i)$ . This equality is given in equation 3, in units of RT:

$$\Delta G_i = -\ln\left(\frac{P(S_i | bound)}{P(S_i)}\right) + \Delta G_j$$

Where  $\Delta G_j$  equals the total binding energy excluding contribution from  $S_i$ . Setting this value to zero yields the relative binding affinity of  $S_i$ , represented by  $\Delta\Delta G_i$  in units of RT, which is given in equation 4:

$$\Delta\Delta G_i = -\ln\left(\frac{P(S_i | bound)}{P(S_i)}\right)$$

**Neural network binding models.** The input for the neural network was defined as a flattened 40-element vector derived from a 4x10 one-hot encoded matrix. Within the initial matrix, each row represents a nucleotide species (A, C, G, or T) and each column represents a position within the ten nucleotide sequence flanking the E-box motif. The output of the neural network was defined as a scalar corresponding to the predicted  $\Delta\Delta G$  value for the species of interest. The network consisted of three hidden layers of size 500, 500, and 250 units, respectively. All weights were initialized with Xavier initialization (84) and all layers used batch normalization (85) and ReLU activation. The entire available sequence space was randomly divided into training (60%), validation (10%), and test (30%) datasets. The networks were trained on the training portion of the data using stochastic gradient descent until the validation set root mean squared error failed to decrease for three consecutive epochs. At this point, the learning rate (initialized at  $10^{-3}$ ) was decreased in 10-fold increments, and training continued until error failed to improve for a further two epochs when training was halted to prevent overfitting (Fig. SXX).

**Linear binding models.** Energetic models of TF substrate specificity use linear combinations of sequence features to explain variation in binding energy (70). All linear binding models presented in this work were training on the binding predictions. Our mononucleotide model includes sequence features consisting of DNA bases present at each flanking position. We parameterized the nearest neighbor dinucleotide models using all mononucleotide features plus all possible combinations of adjacent nucleotide pairs. The full dinucleotide model builds upon the nearest neighbor model with the addition of all non-adjacent (gapped) dinucleotide combinations. All linear binding models were trained using the same 60% of the sequence space as for the neural network models. Reported accuracies are calculated with respect to the held-out 40% of the sequence space.

**ACKNOWLEDGMENTS.** We thank Justin Kinney for library design discussions, Hua Tang for discussions of epistasis and statistics, Anshul Kundaje for neural network discussions, and Daniel Herschlag and Rhiju Das for helpful comments on the manuscript. T.C.S. acknowledges support from an NSF Graduate Research Fellowship. This work was supported by NIH/NIGMS grant R00GM09984804. P.M.F. also acknowledges support from the Sloan Research Foundation, and a McCormick and Gabilan faculty fellowship. P.M.F. is a Chan Zuckerberg Biohub Investigator.

1. Latchman DS (1990) Eukaryotic transcription factors. *The Biochemical journal* 270(2):281–289.

2. Kim HD, O'Shea EK (2008) A quantitative model of transcription factor-activated gene expression. *Nature structural & molecular biology* 15(11):1192–1198.
3. Kim HD, Shay T, O'Shea EK, Regev A (2009) Transcriptional regulatory circuits: predicting numbers from alphabets. *Science* 325(5939):429–432.
4. Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U (2008) Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature* 451(7178):535–540.
5. Raveh-Sadka T, Levo M, Segal E (2009) Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome research* 19(8):1480–1496.
6. Gertz J, Sigga ED, Cohen BA (2009) Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature* 457(7226):215–218.
7. Foat BC, Morozov AV, Bussemaker HJ (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics (Oxford, England)* 22(14):e141–9.
8. Riley TR, Lazarovici A, Mann RS, Bussemaker HJ (2015) Building accurate sequence-to-affinity models from high-throughput in vitro protein-DNA binding data using FeatureREDUCE. *eLife* 4:307.
9. Zhao Y, Stormo GD (2011) Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nature biotechnology* 29(6):480–483.
10. Zhao Y, Granás D, Stormo GD (2009) Inferring binding energies from selected binding sites. *PLoS Comput Biol* 5(12):e1000590.
11. Weirauch MT, et al. (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nature biotechnology* 31(2):126–134.
12. Mustonen V, Kinney J, Callan CG, Lässig M (2008) Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. *Proceedings of the National Academy of Sciences of the United States of America* 105(34):12376–12381.
13. Haldane A, Manhart M, Morozov AV (2014) Biophysical fitness landscapes for transcription factor binding sites. *PLoS Comput Biol* 10(7):e1003683.
14. Bintu L, Buchler NE, Garcia HG, Gerland U (2005) Transcriptional regulation by the numbers: models. *Current opinion in . . .*
15. Lam FH, Steger DJ, O'Shea EK (2008) Chromatin decouples promoter threshold from dynamic range. *Nature* 453(7192):246–250.
16. Aow JSZ, et al. (2013) Differential binding of the related transcription factors Pho4 and Cbf1 can tune the sensitivity of promoters to different levels of an induction signal. *Nucleic Acids Research* 41(9):4877–4887.
17. Rajkumar AS, Dénervaud N, Maerkl SJ (2013) Mapping the fine structure of a eukaryotic promoter input-output function. *Nature Genetics* 45(10):1207–1215.
18. Gordán R, et al. (2013) Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell reports* 3(4):1093–1104.
19. Levo M, et al. (2015) Unraveling determinants of transcription factor binding outside the core binding site. *Genome research* 25(7):1018–1029.
20. Atek A, Schipper JL, Horton J, Gordán R, Lukatsky DB (2014) Protein-DNA binding in the absence of specific base-pair recognition. *Proceedings of the National Academy of Sciences of the United States of America* 111(48):17140–17145.
21. Farley EK, Olson KM, Zhang W, Rokhsar DS, Levine MS (2016) Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. *Proceedings of the National Academy of Sciences of the United States of America* 113(23):6508–6513.
22. Atek A, Cohen H, Barber-Zucker S, Gordán R, Lukatsky DB (2015) Nonconsensus Protein Binding to Repetitive DNA Sequence Elements Significantly Affects Eukaryotic Genomes. *PLoS Comput Biol* 11(8):e1004429.
23. Jolma A, et al. (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome research* 20(6):861–873.
24. Slattery M, et al. (2011) Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* 147(6):1270–1282.
25. Tuerk C, Gold L (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 249(4968):505–510.
26. Ellington AD, Szostak JW (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature* 346(6287):818–822.
27. Zykovich A, Korf I, Segal DJ (2009) Bind-n-Seq: high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. *Nucleic Acids Research* 37(22):e151–e151.
28. Chen D, et al. (2016) SELMAP - SELEX affinity landscape MAPPING of transcription factor binding sites using integrated microfluidics. *Scientific reports* 6(1):33351.
29. Mukherjee S, et al. (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nature Genetics* 36(12):1331–1339.
30. Berger MF, et al. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology* 24(11):1429–1435.
31. Maerkl SJ, Quake SR (2007) A systems approach to measuring the binding energy landscapes of transcription factors. *Science* 315(5809):233–237.
32. Fordyce PM, et al. (2010) De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nature biotechnology* 28(9):970–975.
33. Isakova A, et al. (2017) SMILE-seq identifies binding motifs of single and dimeric transcription factors. *Nature methods* 14(3):316–322.
34. Nutiu R, et al. (2011) Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nature biotechnology* 29(7):659–664.
35. Stormo GD, Schneider TD, Gold L, Ehrenfeucht A (1982) Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Research* 10(9):2997–3011.
36. Stormo GD, Fields DS (1998) Specificity, free energy and information content in protein-DNA interactions. *Trends in biochemical sciences* 23(3):109–113.
37. Stormo GD, Hartzell GW (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proceedings of the National Academy of Sciences of the United States of America* 86(4):1183–1187.
38. Hertz GZ, Stormo GD (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics (Oxford, England)* 15(7-8):563–577.



39. Zuo Z, Stormo GD (2014) High-resolution specificity from DNA sequencing highlights alternative modes of Lac repressor binding. *Genetics* 198(3):1329–1343.
40. Bulky ML, Johnson PLF, Church GM (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Research* 30(5):1255–1261.
41. Mordelet F, Horton J, Hartemink AJ, Engelhardt BE, Gordán R (2013) Stability selection for regression-based models of transcription factor-DNA binding specificity. *Bioinformatics (Oxford, England)* 29(13):1117–25.
42. Mathelier A, Wasserman WW (2013) The next generation of transcription factor binding site prediction. *PLoS Comput Biol* 9(9):e1003214.
43. Zhao Y, Ruan S, Pandey M, Stormo GD (2012) Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics* 191(3):781–790.
44. Tomovic A, Oakeley EJ (2007) Position dependencies in transcription factor binding sites. *Bioinformatics (Oxford, England)* 23(8):933–941.
45. Siddharthan R (2010) Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. *PLoS one* 5(3):e9722.
46. Badis G, et al. (2009) Diversity and complexity in DNA recognition by transcription factors. *Science* 324(5935):1720–1723.
47. Annala M, Laurila K, Lähdesmäki H, Nykter M (2011) A linear model for transcription factor binding affinity prediction in protein binding microarrays. *PLoS one* 6(5):e20059.
48. Zhao X, Huang H, Speed TP (2005) Finding short DNA motifs using permuted Markov models. *Journal of computational biology : a journal of computational molecular cell biology* 12(6):894–906.
49. Sharon E, Lubliner S, Segal E (2008) A feature-based approach to modeling protein-DNA interactions. *PLoS Comput Biol* 4(8):e1000154.
50. Rohs R, et al. (2009) The role of DNA shape in protein-DNA recognition. *Nature* 461(7268):1248–1253.
51. Abe N, et al. (2015) Deconvolving the recognition of DNA shape from sequence. *Cell* 161(2):307–318.
52. Chiu TP, et al. (2016) DNASHapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics (Oxford, England)* 32(8):1211–1213.
53. Yang L, et al. (2017) Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Molecular systems biology* 13(2):910.
54. Zhou T, et al. (2013) DNASHape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Research* 41(Web Server issue):W56–62.
55. Quang D, Xie X (2016) DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Research* 44(11):e107–e107.
56. Djordjevic M, Sengupta AM, Shraiman BI (2003) A biophysical approach to transcription factor binding site discovery. *Genome research* 13(11):2381–2390.
57. Hellman LM, Fried MG (2007) Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nature protocols* 2(8):1849–1861.
58. Fordyce PM, et al. (2012) Basic leucine zipper transcription factor Hac1 binds DNA in two distinct modes as revealed by microfluidic analyses. *Proceedings of the National Academy of Sciences of the United States of America* 109(45):E3084–93.
59. Jones S (2004) An overview of the basic helix-loop-helix proteins. *Genome biology* 5(6):226.
60. Fisher F, Goding CR (1992) Single amino acid substitutions alter helix-loop-helix protein specificity for bases flanking the core CANTGT motif. *The EMBO journal* 11(11):4103–4109.
61. Shimizu T, et al. (1997) Crystal structure of PHO4 bHLH domain-DNA complex: flanking base recognition. *The EMBO journal* 16(15):4689–4697.
62. Zhou X, O'Shea EK (2011) Integrated approaches reveal determinants of genome-wide binding and function of the transcription factor Pho4. *Molecular Cell* 42(6):826–836.
63. Kivioja T, et al. (2012) Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods* 9(1):72–74.
64. Fu GK, et al. (2014) Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. *Proceedings of the National Academy of Sciences of the United States of America* 111(5):1891–1896.
65. Fu GK, Hu J, Wang PH, Fodor SPA (2011) Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proceedings of the National Academy of Sciences of the United States of America* 108(22):9026–9031.
66. Spivak AT, Stormo GD (2012) ScerTF: a comprehensive database of benchmarked position weight matrices for Saccharomyces species. *Nucleic Acids Research* 40(Database issue):D162–8.
67. Morozov AV, Siggia ED (2007) Connecting protein structure with predictions of regulatory sites. *Proceedings of the National Academy of Sciences of the United States of America* 104(17):7068–7073.
68. MacIsaac KD, et al. (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC bioinformatics* 7(1):113.
69. Wagih O (2017) ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics (Oxford, England)*.
70. Stormo GD, Schneider TD, Gold L (1986) Quantitative analysis of the relationship between nucleotide sequence and functional activity. *Nucleic Acids Research* 14(16):6661–6679.
71. Mellor J, et al. (1990) CPF1, a yeast protein which functions in centromeres and promoters. *The EMBO journal* 9(12):4017–4026.
72. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological)*.
73. Bintu L, et al. (2005) Transcriptional regulation by the numbers: models. *Current opinion in genetics & development* 15(2):116–124.
74. Crocker J, Noon EPB, Stern DL (2016) The Soft Touch: Low-Affinity Transcription Factor Binding Sites in Development and Evolution. *Current topics in developmental biology* 117:455–469.
75. Aguilar-Rodríguez J, Payne JL, Wagner A (2017) A thousand empirical adaptive landscapes and their navigability. *Nature Ecology & Evolution* 1(2):0045.
76. Olson WK, Gorin AA, Lu XJ, Hock LM, Zhurkin VB (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proceedings of the National Academy of Sciences of the United States of America* 95(19):11163–11168.
77. Yang L, et al. (2014) TFBSshape: a motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Research* 42(Database issue):D148–55.
78. Jolma A, et al. (2013) DNA-binding specificities of human transcription factors. *Cell* 152(1–2):327–339.
79. Tuğrul M, Paixão T, Barton NH, Tkačik G (2015) Dynamics of Transcription Factor Binding Site Evolution. *PLoS genetics* 11(11):e1005639.
80. Simon MD, et al. (2007) The Site-Specific Installation of Methyl-Lysine Analogs into Recombinant Histones. *Cell* 128(5):1003–1012.
81. Yang A, et al. (2016) A chemical biology route to site-specific authentic protein modifications. *Science* 354(6312):623–626.
82. McGinty RK, Kim J, Chatterjee C, Roeder RG, Muir TW (2008) Chemically ubiquitinated histone H2B stimulates hDot1L-mediated intranucleosomal methylation. *Nature* 453(7196):812–816.
83. O'Malley RC, et al. (2016) Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell* 166(6):1598.
84. Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. *Aistats*.
85. Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International Conference on Machine Learning*.