Supplemental Material for

Competition for DNA binding between paralogous transcription factors determines their genomic occupancy and regulatory functions

Yuning Zhang[1,2], Tiffany D. Ho[1,3], Nicolas E. Buchler[4], Raluca Gordan[1,3,5,*]

[1]Center for Genomic and Computational Biology, [2]Program in Computational Biology and Bioinformatics, [3]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, [4]Department of Molecular Biomedical Sciences, North Carolina State University, Raleigh, NC, [5]Department of Computer Science, Department of Molecular Genetics and Microbiology, Duke University, Durham, NC
*Email: raluca.gordan@duke.edu. Phone number: (919) 684-9881

SUPPLEMENTAL METHODS

Comparing competition PBM data across experiments

As listed in **Supplemental Table S1B**, in our competition PBM experiments we varied the concentration of the competitor TF over a wide range (from 0.05uM to 8uM), while keeping the concentration of the main TF constant (at 2uM), with the different experiments performed in different chambers of the same microarray slide. For example, we refer to the four competition chambers where Pho4 was the main TF as: 2uM Pho4 + 0.05uM Cbf1, 2uM Pho4 + 0.4uM Cbf1; 2uM Pho4 + 2uM Cbf1, and 2uM Pho4 + 8uM Cbf1. For brevity, we also refer to these chambers as (a), (b), (c) and (d). While we aimed to have a constant concentration of Pho4 in these four chambers, the effective concentration varied slightly, as reflected by the fluorescence intensity measurements at a subset of control DNA probes that are non-specifically bound by Cbf1 (i.e. within Cbf1 negative control range) but bound with low to medium-high affinities by Pho4 (**Supplemental Fig. S3C**). At these spots, we expect the Pho4 binding levels to be highly similar in all four competition chambers, and driven by the specific DNA binding of Pho4. Thus, we can use standard binding isotherm to describe the Pho4 occupancy as $p = \frac{[Pho4]/K_d}{1+[Pho4]/K_d}$. At two different concentrations of Pho4, we can write its occupancies as:

$$p_1 = \frac{[Pho4]_1/K_d}{1+[Pho4]_1/K_d} \text{ (1)}, \qquad p_2 = \frac{[Pho4]_2/K_d}{1+[Pho4]_2/K_d} \text{ (2)}$$

From equations (1) and (2) we can derive the relationship between Pho4 occupancies at two different concentrations:

$$p_1 = \frac{[Pho4]_1 \cdot p_2}{[Pho4]_2(1-p_2)+[Pho4]_1 \cdot p_2} = \frac{p_2}{conc_{ratio} \cdot (1-p_2)+p_2} \text{ (3)}, \quad \text{where } conc_{ratio} = [Pho4]_2/[Pho4]_1.$$

In the PBM assays, we observed the Pho4 occupancies $p^i = \frac{[DNA]^i_{bound}}{[DNA]_{total}} = \frac{[DNA \cdot TF]^i}{[DNA \cdot TF]^{max}} = \frac{F^i}{F^{max}}$, where F^i is the fluorescence intensity we observed at DNA probe i and F^{max} is the maximum fluorescence possible (Siggers et al. 2011). Focusing on the subset of probes that were non-specifically bound by Cbf1, we can then estimate the variable $conc_{ratio}$ between two different chambers that have slightly different Pho4 concentrations. Using nonlinear least square regression (minpack.lm package in R), we estimated that $conc_{ratio}$ is 0.492 between chambers (a) and (b), 1.085 between chambers (a) and (c), and 1.09 between chambers (a) and (d).

Next, based on the inferred effective concentrations, we adjusted the Pho4 occupancies at all DNA probes in chambers (b), (c), and (d), assuming the same Pho4 concentration as in chamber (a), while the concentration of Cbf1 remained unchanged. For any given DNA probe i, we can write the Pho4 occupancies at two Pho4 concentrations under the competition from same concentration of Cbf1 as:

$$p_1^{comp} = \frac{[Pho4]_1/K_{d,Pho4}}{1+[Pho4]_1/K_{d,Pho4}+[Cbf1]/K_{d,Cbf1}} \text{ (4)}, \quad p_2^{comp} = \frac{[Pho4]_2/K_{d,Pho4}}{1+[Pho4]_2/K_{d,Pho4}+[Cbf1]/K_{d,Cbf1}} \text{ (5)}$$

Using equations (4) and (5), we can derive the relationship between p_1^{comp} and p_2^{comp} as:

$$p_1^{comp} = \frac{p_2^{comp}}{conc_{ratio}\cdot(1-p_2^{comp})+p_2^{comp}} \text{ (6)}$$

where $conc_{ratio} = [Pho4]_2/[Pho4]_1$. Next, we plugged the estimated $conc_{ratio}$ into equation (6) to compute the binding levels of Pho4 at concentration $[Pho4]_1$, while the experimental effective concentration was actually $[Pho4]_2$. We set the concentration of Pho4 in chamber (a) as $[Pho4]_1$, and used the $conc_{ratio}$ values computed above to adjust the Pho4 binding levels in chambers (b), (c) and (d) so that their Pho4 concentrations are $[Pho4]_1$. This correction allowed us to make direct comparisons between different chambers where Pho4 was the main TF, despite slight differences in its concentration. We note that this adjustment was not needed when Cbf1 was treated as the main TF, as the $conc_{ratio}$ values compute using a similar procedure did not deviate significantly from 1.

SUPPLEMENTAL DISCUSSION

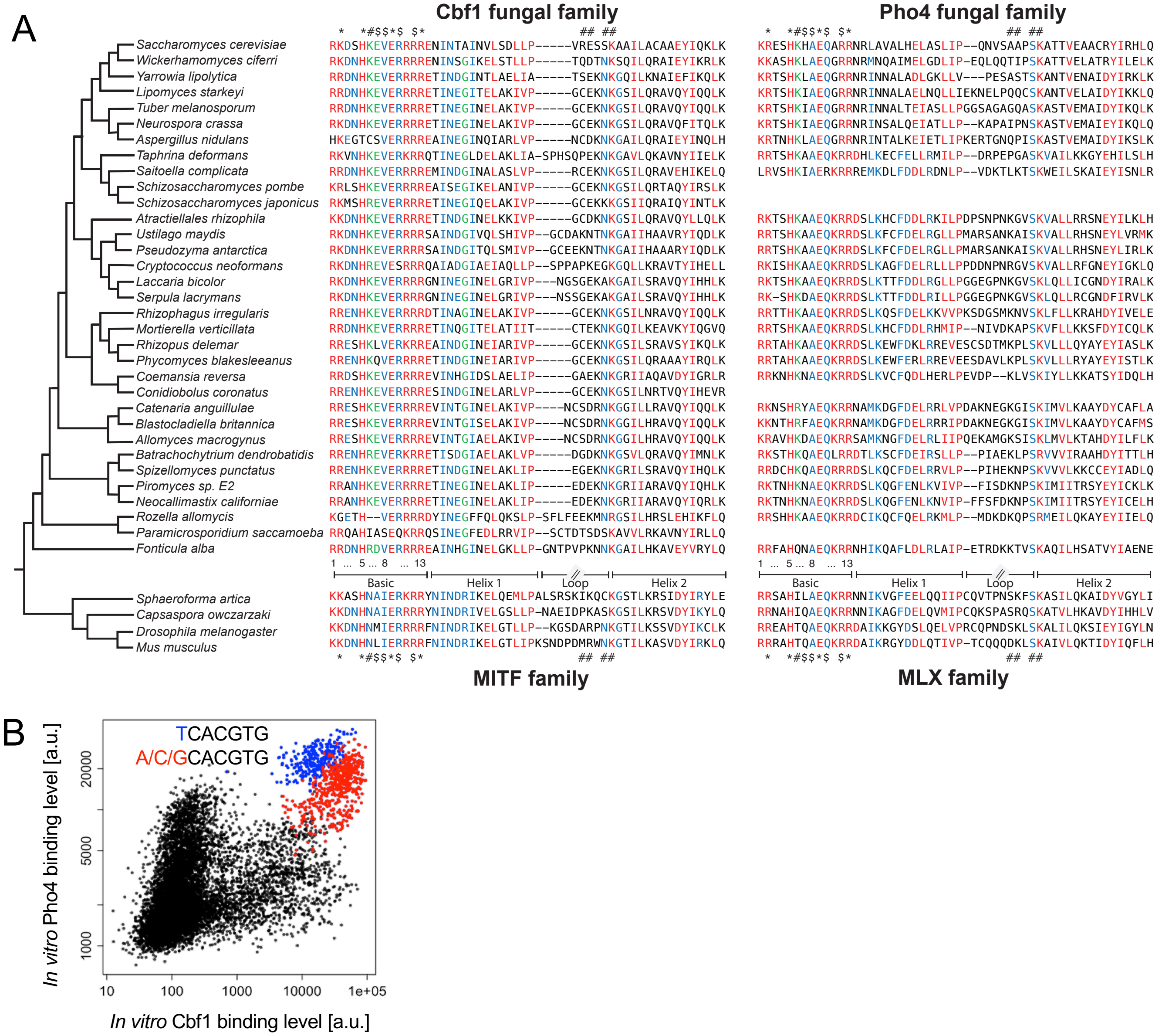
Interpretation of fluorescence intensities in PBM experiments

In PBM experiments, the level of DNA-bound protein at each spot on the microarray is assessed by measuring the fluorescence intensity at that spot (Berger et al. 2006; Berger and Bulyk 2009). As shown in our recent work (Figure 1f, Extended Data Figure 3e,f in Afek et al. (Afek et al. 2020)), PBM fluorescence intensities correlate very well with equilibrium dissociation constants (K_d) measured using alternative lower-throughput techniques such as EMSA (electrophoretic mobility shift assay), FA (fluorescence anisotropy), SPR (surface plasmon resonance), MITOMI (mechanically induced trapping of molecular interactions), and k-MITOMI (kinetic MITOMI), across a wide range of TFs. Measurements using these techniques can be used to calibrate the PBM fluorescence intensities and thus map them to Kd values and/or binding energies. In addition, as previously shown by Siggers et al. (Siggers et al. 2011), PBM data can be used to derive binding energies (ΔΔG), which are also in excellent agreement with independently measured binding energies, as shown here for Cbf1 and Pho4 (**Fig. 2E, Supplemental Fig. S4B**).

We note, however, that fluorescent intensities, as absolute values, do not necessarily have a direct interpretation, and they are generally not directly comparable between PBM experiments for different proteins (especially when different epitopes, antibodies, or fluorophores are being used). In addition, the specific binding properties of the TFs themselves can affect the ranges of intensities observed in PBM experiments. In the case of Cbf1 and Pho4, we generally observed a higher level of background signal in the PBM fluorescence intensity measurements, i.e. higher intensities at the negative control sites (see, for example, **Supplemental Fig. S2E**). Our observation is consistent with PBM data from previous studies (Gordan et al. 2013), including studies where Cbf1 and Pho4 were tested in universal

PBM experiments (Zhu et al. 2009; Gordan et al. 2011). The PBM data thus suggests that Pho4 has a higher overall level of non-specific DNA-binding compared to Cbf1, consistent with the MITOMI binding data showing generally lower binding energies (ΔΔG) for Pho4 at non-CACGTG-containing sites (Maerkl and Quake 2007). For example, in the CACNNN library tested by Maerkl and Quake (Maerkl and Quake 2007), the median ΔΔG at sites other than CACGTG, most of which are expected to be non-specifically bound, is 3.6 kcal/mol for Cbf1 (**Fig. 2E** left panel; full data in **Supplemental Table S2I**), versus 1.9 kcal/mol for Pho4 (**Supplemental Fig. S4B**, bottom left panel; full data in **Supplemental Table S2I**), with lower ΔΔG values corresponding to higher binding affinity.

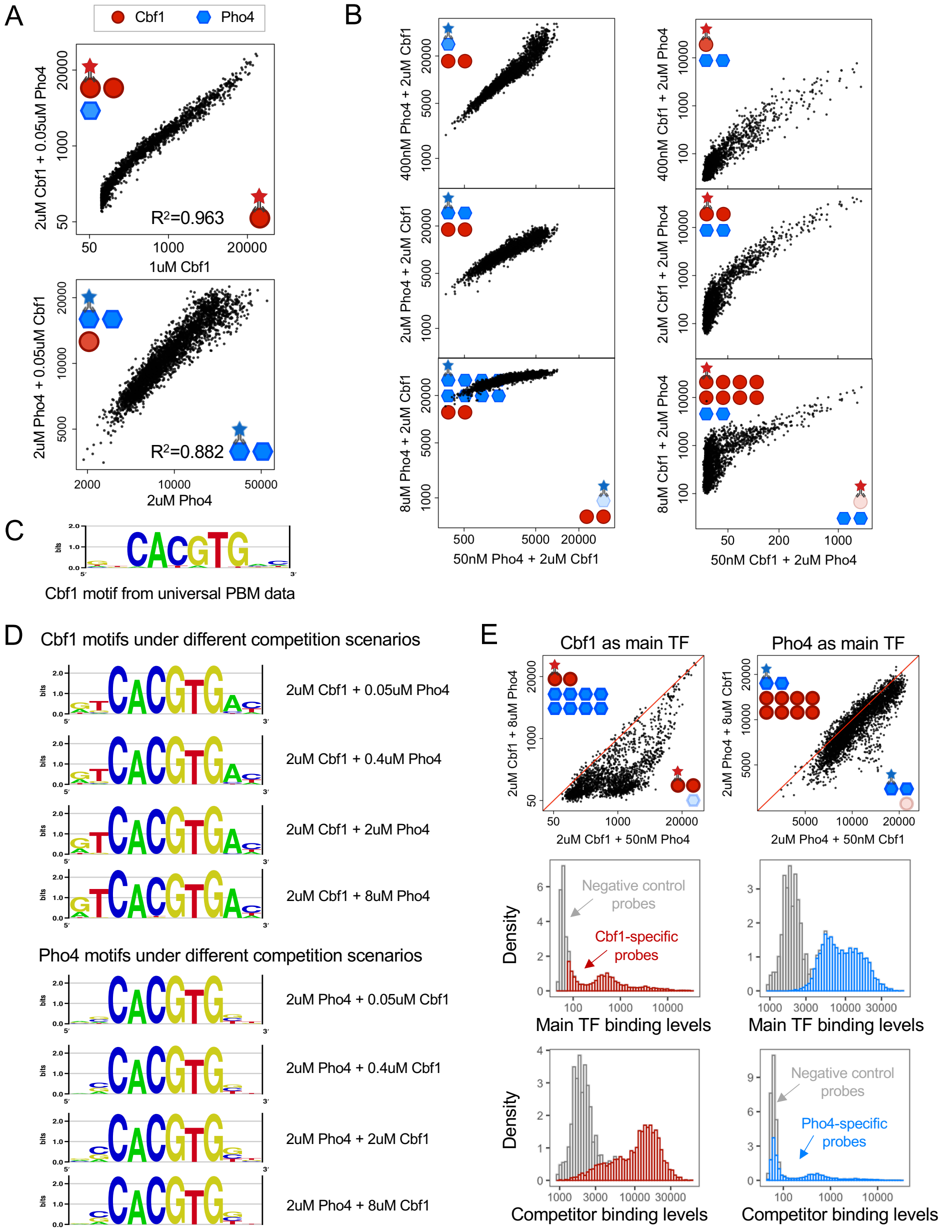
SUPPLEMENTAL FIGURES



Supplemental Figure S1.

(A) Fungal Cbf1 and Pho4 families are evolutionarily-related to animal MITF and MLX sub-families. Alignment shows the DNA-binding domains of yeast bHLH proteins Cbf1 (left) and Pho4 (right) against orthologous proteins across the Fungal lineage (see Methods). The fungal species tree was taken from (Gomes-Vieira et al. 2018) and modified to include new fungal genomes and diverse holozoans (e.g., animals) as an outgroup. We analyzed each proteome for the presence of orthologs from different sub-families using HMMs, as described in Methods. All orthologs were aligned to the canonical bHLH domain, which consists of a Basic region (Position 1-13), Helix 1 (Pos. 14-28), variable Loop (Pos. 29-49), and Helix 2 (Pos. 50-64); notation from (Atchley et al. 1999). Green amino-acids are conserved, yet specific to each fungal Cbf1 and Pho4 sub-family. Blue amino-acids are conserved, yet specific to an evolutionarily-related family (Cbf1/MITF or Pho4/MLX). Red amino-acids are common to all the sub-families. * = amino-acids in Pho4 known to make specific contacts to core 5'-CAnnTG-3' nucleotide bases (Shimizu et al. 1997). # = amino-acids in Pho4 known to make non-specific contacts to the DNA backbone (Aditham et al. 2021). $ = amino-acids in Pho4 shown to affect specificity to base-pairs immediately flanking the E-box (Fisher and Goding 1992; Aditham et al. 2021).

(B) Cbf1 and Pho4 have different preferences towards bases flanking the CACGTG core site. Scatter plot shows the *in vitro* DNA binding levels of the two TFs for 10,772 genomic sequences centered at CACGTG sites, according to genomic-context PBM data (Gordan et al. 2013). The data is consistent with previous studies showing that Cbf1 has a strong preference for T upstream of the CACGTG core site, while Pho4 prefers G, C, or A (Fisher and Goding 1992; Aditham et al. 2021).



Supplemental Figure S2.

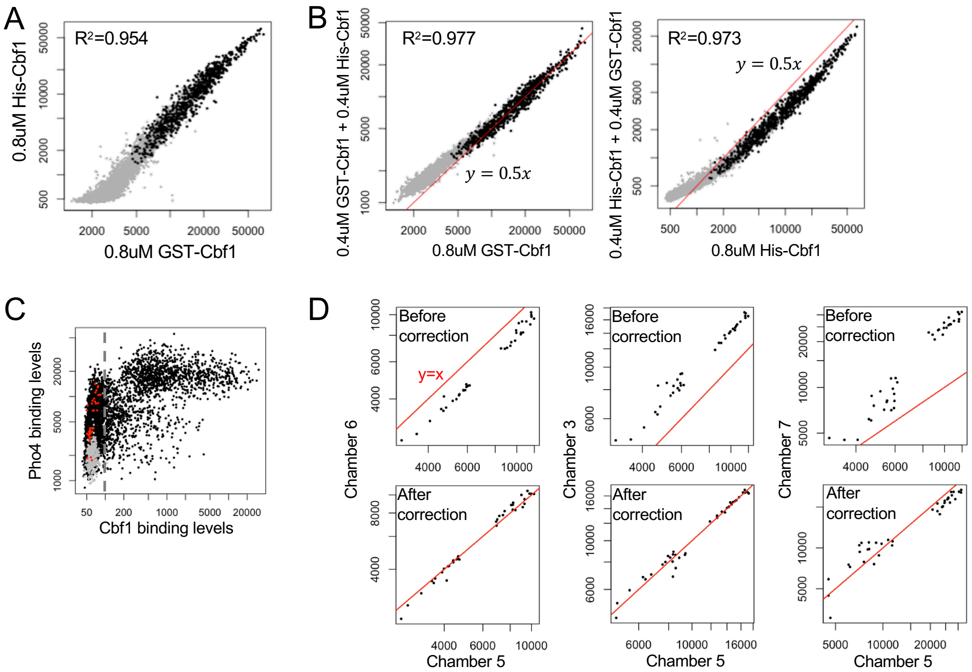
(A) Direct comparison of individual TF binding levels (x-axes) versus binding levels in the presence of very low concentrations of the competitor TF (y-axes). At this low concentration of the competitor TF, very little competition is observed. These data sets have not been normalized to each other, and thus the ranges of fluorescence intensities vary slightly between experiments. Nevertheless, the data sets have high correlations, with R^2 = 0.882-0.963.

(B) Binding levels of the competitor TFs in the competition PBM assays, corresponding to the 6 competition experiments shown in **Fig. 2B, C**.

(C) PWM motif logo for Cbf1, derived from universal PBM data (Zhu et al. 2009), where Cbf1 was measured individually.

(D) PWM motif logos for Cbf1 and Pho4, derived from all concentration combinations in the competition PBM assay.

(E) Histograms show the distributions of Cbf1 and Pho4 binding levels (tested as individual TFs) at different subsets of DNA sequences that correspond to the competition PBM data shown in the top scatter plots. The intrinsic DNA-binding preferences of Cbf1 and Pho4 are consistent with our observation from the competition assay that Cbf1 outcompetes Pho4 at few sites, while Pho4 outcompetes Cbf1 at a large number of sites. Left: many of the Cbf1-specific sites (defined as sites with Cbf1 binding signal above the 99th percentile of negative controls; Methods) are also bound well by Pho4, as illustrated by the small overlap between the red and grey histograms in the bottom left panel. Right: for the Pho4-specific sites (defined as sites with Pho4 binding signal above the 99th percentile of negative controls; Methods), we found that many of them are bound by Cbf1 with affinities in the negative control range, as illustrated by the large overlap between the blue and grey histograms in the bottom right panel.

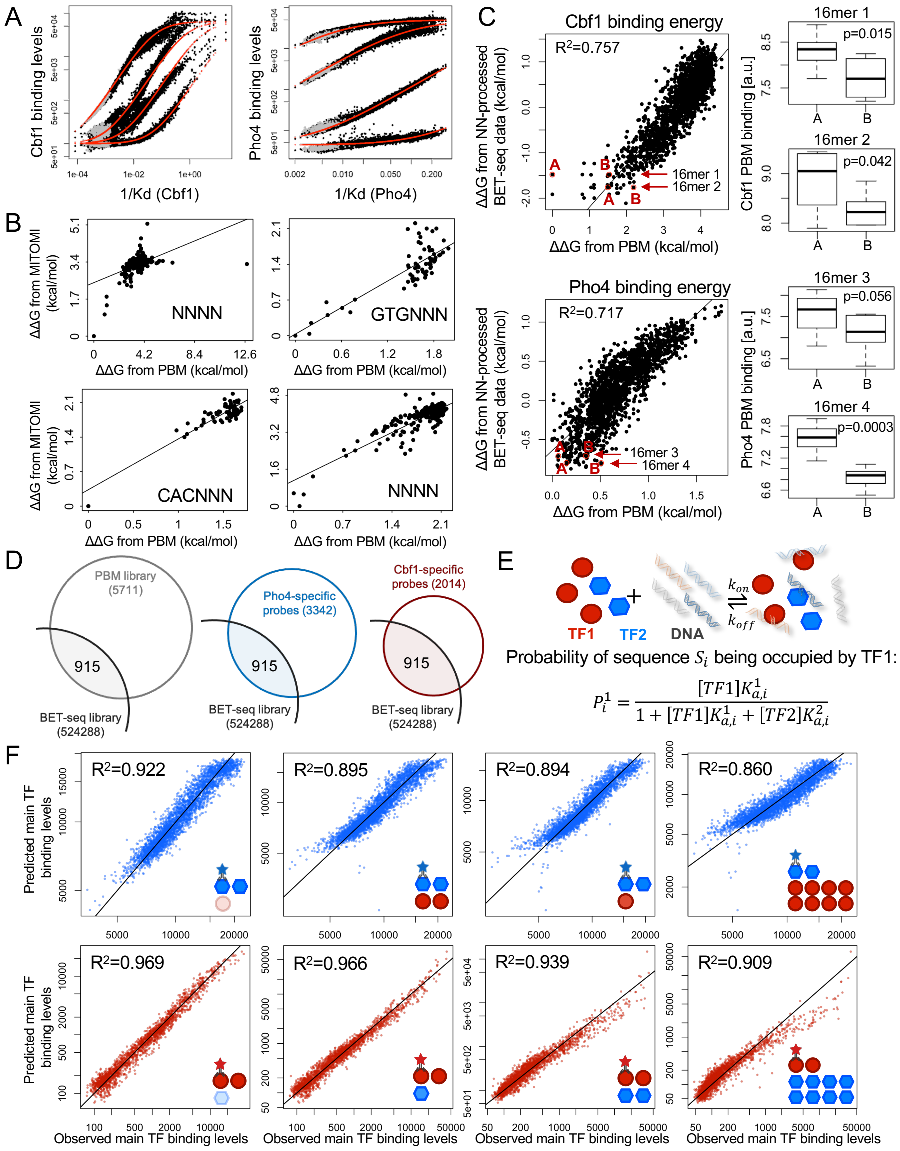


Supplemental Figure S3.

(A,B) Control experiments to determine whether the choice of epitope tags (GST and His in this study) might affect the outcome of competition PBM experiments. A: Direct comparison between the binding levels of GST-Cbf1 and His-Cbf1 at negative control sites (grey) and putative genomic binding sites for Cbf1 (black). See Methods for details. B: GST-Cbf1 and His-Cbf1 compete with each other with a linear trend. The red line (y=0.5x) is the predicted fitting line assuming that epitope tags would not influence Cbf1 binding or competition.

(C) Direct comparison of Cbf1 and Pho4 binding levels (when the proteins were tested individually at 1uM and 2uM, respectively) for the DNA probes tested in the competition PBM assays. Grey points are negative control sequences, i.e. they are non-specifically bound by Cbf1 and Pho4. Red points are the probes selected to normalize between competition PBM experiments where Pho4 was the main TF and Cbf1 was the competitor TF; these probes are non-specifically bound by Cbf1 but bound at different levels by Pho4. Grey dotted line shows the upper bound of the non-specific range of Cbf1 binding. See Methods for details.

(D) Scatter plots show the correlations between Pho4 fluorescence intensities in PBM chambers containing different concentrations of competitor Cbf1 protein (**Supplemental Table S2B**), for a subset of probes selected to be non-specifically bound by Cbf1, i.e. the red probes shown in panel C. As Cbf1 is not expected to compete specifically with Pho4 at these probes, we used them to apply a correction to the measured Pho4 intensities in the different chambers (Methods). After correction, the binding levels were comparable across chambers. Red line: y=x. Chamber 5: 2uM Pho4 + 0.05uM Cbf1; chamber 6: 2uM Pho4 + 0.4uM Cbf1; chamber 3: 2uM Pho4 + 2uM Cbf1; chamber 7: 2uM Pho4 + 8uM Cbf1 (**Supplemental Table S2B**).



Supplemental Figure S4.

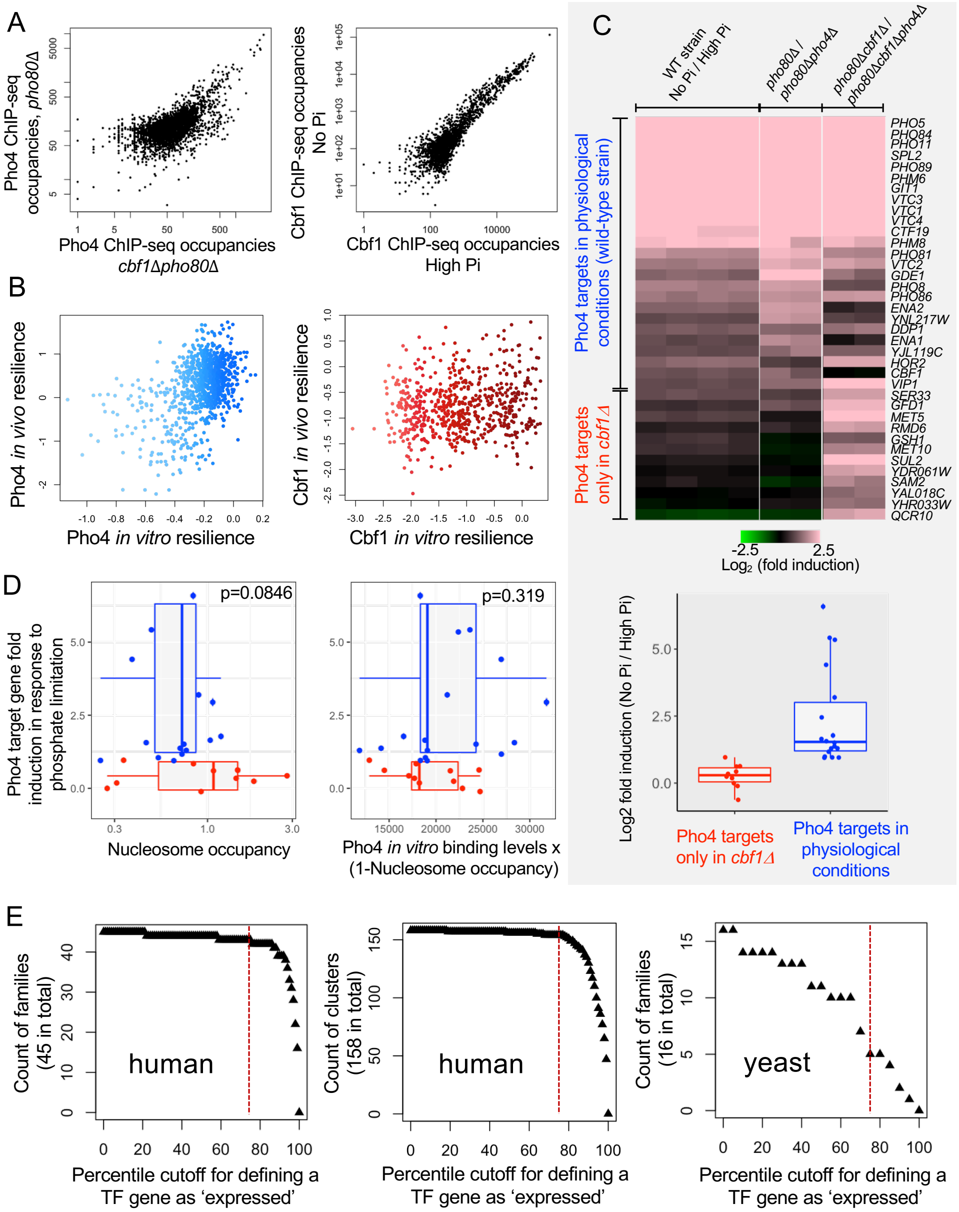
(A) Titration curves (black points) and fitted curves (red lines) with estimated Kds. Four concentrations were measured for each protein (Cbf1: 0.05uM, 0.2uM, 1uM, 8uM; Pho4: 0.1uM, 0.4uM, 2uM, 8uM). Black points are genomic regions; grey points are negative control sequences.

(B) Comparison between PBM-derived and MITOMI-derived binding energies (ΔΔG) for various DNA sequence sets from (Maerkl and Quake 2007). The NNNN library contains 256 TTGNNNNGTGGGTG sequences. The GTGNNN library contains 64 TTTTCACGTGNNNT sequences. The CACNNN library contains 64 TTGTCACNNNACTT sequences. All DNA sequences were measured simultaneously in our assay, but in different MITOMI experiments in (Maerkl and Quake 2007).

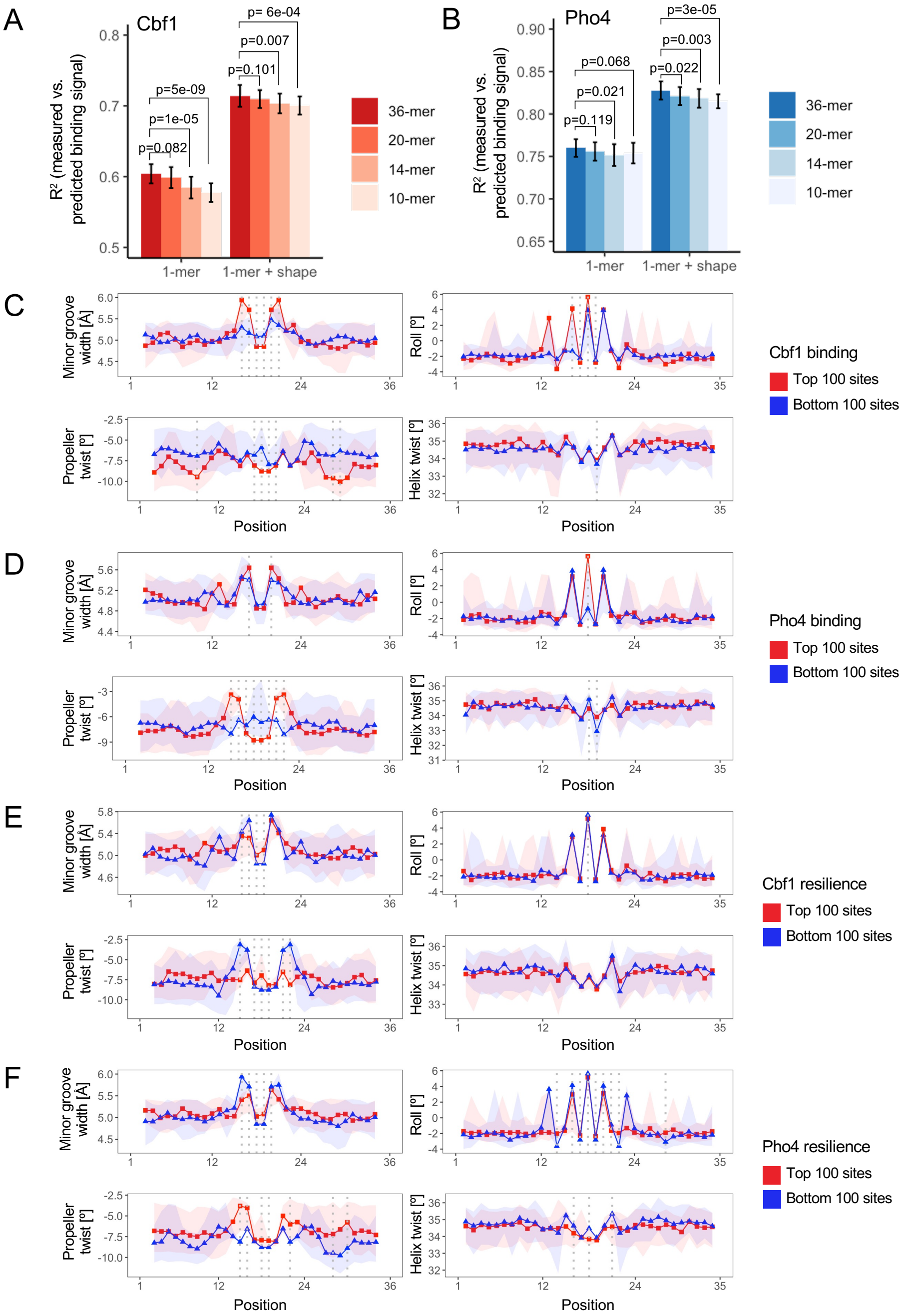
(C) Comparison between PBM-derived binding energies (x-axis) vs. binding energies predicted by a deep neural network (NN) trained on BET-seq data (y-axis) (Le et al. 2018). Each data point corresponds to one occurrence of CACGTG in the yeast genome. PBM-derived ΔΔGs were computed for 36-bp sequences centered on CACGTG. BET-seq-derived ΔΔGs were computed for 16-bp sequences centered on CACGTG. Red circles correspond to examples of genomic sites with identical 16-mers (i.e. single measurements in BET-seq) but different distal flanking regions (which resulted in different probes in our PBM assay). Boxplots show that the distal flanking regions included in our PBM probes have a significant effect on TF binding, assessed based on the difference in PBM-measured TF binding levels at the 6 replicate probes tested for each genomic site (p-vals are according to a one-sided Mann-Whitney *U* test). 16mer1: CGAGTCACGTGACGAA; 16mer2: CGAGTCACGTGATGGA; 16mer3: ACGGCCACGTGGGTAA; 16mer4: ACGGCCACGTGGCCGC.

(D) Venn diagrams showing the overlap between the DNA sequences included in our competition PBM library and the ones included in BET-seq (Le et al. 2018).

(E) Theoretical model of competitive binding between TF1 and TF2. **(F)** Observed competitive binding levels versus predicted competitive binding levels, shown for all the putative binding sites for each protein. Blue: Pho4. Red: Cbf1.



Supplemental Figure S5.
(A) Comparisons between ChIP-seq data under non-competing conditions (x-axes) versus competing conditions (y-axes) for Pho4 and Cbf1. Values represent ChIP-seq read pile-ups (Methods).
(B) Scatter plot of *in vitro* resilience versus *in vivo* resilience for Pho4 (left) and Cbf1 (right).
(C) Gene expression data (Zhou and O'Shea 2011) shows differential activation of two sets of Pho4 target genes. The *pho80Δ* strain has the *PHO* responsive pathway constitutively on, thus recapitulating the no phosphate condition of the wilt-type (WT) strain. The *pho80Δpho4Δ* strain has *PHO4* knocked out, thus recapitulating the high phosphate condition. As expected, the gene expression patterns for "WT stain No Pi / High Pi" and "*pho80Δ* / *pho80Δpho4Δ*" are similar. Individual columns show replicate experiments.
(D) Nucleosome occupancy by itself (left) and Pho4 *in vitro* binding levels corrected based on nucleosome occupancy (right) cannot explain how the two groups of genes are differently regulated in response to phosphate starvation.
(E) Expression patterns of human (left and center) and yeast (right) TF families. Plots show the number of TF families (or TF clusters – center plot) that have at least two TFs expressed simultaneously in at least on cell type (human) or strain (yeast). In human, a TF gene was considered 'expressed' if its mRNA level was higher than a given percentile of all genes, according to RNA-seq data (Lambert et al. 2018). In yeast, a TF gene was considered 'expressed' if its protein level was higher than a given percentile over all genes, according to (Ho et al. 2018). Dotted red line indicates 75th percentile.



Supplemental Figure S6.

(A,B) Regression models of DNA-binding specificity for Cbf1 **(A)** and Pho4 **(B)** show that adding DNA shape features improves model accuracy. Models using only 1-mer features are based on the mono-nucleotide identities at each position in the binding site and flanking regions, as in our previous work (Gordan et al. 2013; Zhou et al. 2015; Shen et al. 2018) while "1-mer + shape" models used as features the mono-nucleotide identities as well as DNA shape features predicted using DNAshape (Zhou et al. 2013); see Methods for details. Plots show the prediction accuracy of models trained and tested on 10-mer, 14-mer, 20-mer, and 36-mer sequences centered on the E-box site, using 5-fold cross-validation. Each cross-validation test was run 25 times. Bar plots show the medians and standard deviations over the 25 runs. One-sided Mann-Whitney *U* tests were used to assess whether the accuracy of different models shows statistically significant differences. Overall, we found that 1-mer+shape models always perform better than 1-mer models (p-values < 8e-15), and models that include longer flanking regions have higher accuracy. R^2 = squared Pearson's correlation coefficient.

(**C-F**) Direct comparisons between the DNA shape profiles of the top 100 versus bottom 100 Cbf1 or Pho4 binding sites sorted according to Cbf1 binding levels as measured by PBM (**C**), Pho4 binding levels as measured by PBM (**D**), Cbf1 resilience to Pho4 competition (**E**), and Pho4 resilience to Cbf1 competition (**F**). Resilience scores were computed as in **Fig. 3B,F**.

REFERENCES

Aditham AK, Markin CJ, Mokhtari DA, DelRosso N, Fordyce PM. 2021. High-Throughput Affinity Measurements of Transcription Factor and DNA Mutations Reveal Affinity and Specificity Determinants. *Cell Syst* **12**: 112-127 e111.

Afek A, Shi H, Rangadurai A, Sahay H, Senitzki A, Xhani S, Fang M, Salinas R, Mielko Z, Pufall MA et al. 2020. DNA mismatches reveal conformational penalties in protein-DNA recognition. *Nature* **587**: 291-296.

Atchley WR, Terhalle W, Dress A. 1999. Positional dependence, cliques, and predictive motifs in the bHLH protein domain. *J Mol Evol* **48**: 501-516.

Berger MF, Bulyk ML. 2009. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat Protoc* **4**: 393-411.

Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, 3rd, Bulyk ML. 2006. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* **24**: 1429-1435.

Fisher F, Goding CR. 1992. Single amino acid substitutions alter helix-loop-helix protein specificity for bases flanking the core CANNTG motif. *EMBO J* **11**: 4103-4109.

Gomes-Vieira AL, Wideman JG, Paes-Vieira L, Gomes SL, Richards TA, Meyer-Fernandes JR. 2018. Evolutionary conservation of a core fungal phosphate homeostasis pathway coupled to development in Blastocladiella emersonii. *Fungal Genet Biol* **115**: 20-32.

Gordan R, Murphy KF, McCord RP, Zhu C, Vedenko A, Bulyk ML. 2011. Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights. *Genome Biol* **12**: R125.

Gordan R, Shen N, Dror I, Zhou T, Horton J, Rohs R, Bulyk ML. 2013. Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep* **3**: 1093-1104.

Ho B, Baryshnikova A, Brown GW. 2018. Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. *Cell Syst* **6**: 192-205 e193.

Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. 2018. The Human Transcription Factors. *Cell* **175**: 598-599.

Le DD, Shimko TC, Aditham AK, Keys AM, Longwell SA, Orenstein Y, Fordyce PM. 2018. Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding. *Proc Natl Acad Sci U S A* **115**: E3702-E3711.

Maerkl SJ, Quake SR. 2007. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315**: 233-237.

Shen N, Zhao J, Schipper JL, Zhang Y, Bepler T, Leehr D, Bradley J, Horton J, Lapp H, Gordan R. 2018. Divergence in DNA Specificity among Paralogous Transcription Factors Contributes to Their Differential In Vivo Binding. *Cell Syst* **6**: 470-483 e478.

Shimizu T, Toumoto A, Ihara K, Shimizu M, Kyogoku Y, Ogawa N, Oshima Y, Hakoshima T. 1997. Crystal structure of PHO4 bHLH domain-DNA complex: flanking base recognition. *EMBO J* **16**: 4689-4697.

Siggers T, Duyzend MH, Reddy J, Khan S, Bulyk ML. 2011. Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Mol Syst Biol* **7**: 555.

Zhou T, Shen N, Yang L, Abe N, Horton J, Mann RS, Bussemaker HJ, Gordan R, Rohs R. 2015. Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc Natl Acad Sci U S A* **112**: 4654-4659.

Zhou T, Yang L, Lu Y, Dror I, Dantas Machado AC, Ghane T, Di Felice R, Rohs R. 2013. DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res* **41**: W56-62.

Zhou X, O'Shea EK. 2011. Integrated approaches reveal determinants of genome-wide binding and function of the transcription factor Pho4. *Mol Cell* **42**: 826-836.

Zhu C, Byers KJ, McCord RP, Shi Z, Berger MF, Newburger DE, Saulrieta K, Smith Z, Shah MV, Radhakrishnan M et al. 2009. High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res* **19**: 556-566.