# 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] and Raluca Gordân[1,3,5]

[1]*Center for Genomic and Computational Biology, Duke University, Durham, North Carolina 27708, USA;* [2]*Program in Computational Biology and Bioinformatics, Duke University, Durham, North Carolina 27708, USA;* [3]*Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina 27708, USA;* [4]*Department of Molecular Biomedical Sciences, North Carolina State University, Raleigh, North Carolina 27606, USA;* [5]*Department of Computer Science, Department of Molecular Genetics and Microbiology, Duke University, Durham, North Carolina 27708, USA*

Most eukaryotic transcription factors (TFs) are part of large protein families, with members of the same family (i.e., paralogous TFs) recognizing similar DNA-binding motifs but performing different regulatory functions. Many TF paralogs are coexpressed in the cell and thus can compete for target sites across the genome. However, this competition is rarely taken into account when studying the in vivo binding patterns of eukaryotic TFs. Here, we show that direct competition for DNA binding between TF paralogs is a major determinant of their genomic binding patterns. Using yeast proteins Cbf1 and Pho4 as our model system, we designed a high-throughput quantitative assay to capture the genomic binding profiles of competing TFs in a cell-free system. Our data show that Cbf1 and Pho4 greatly influence each other's occupancy by competing for their common putative genomic binding sites. The competition is different at different genomic sites, as dictated by the TFs' expression levels and their divergence in DNA-binding specificity and affinity. Analyses of ChIP-seq data show that the biophysical rules that dictate the competitive TF binding patterns in vitro are also followed in vivo, in the complex cellular environment. Furthermore, the Cbf1-Pho4 competition for genomic sites, as characterized in vitro using our new assay, plays a critical role in the specific activation of their target genes in the cell. Overall, our study highlights the importance of direct TF-TF competition for genomic binding and gene regulation by TF paralogs, and proposes an approach for studying this competition in a quantitative and high-throughput manner.

[Supplemental material is available for this article.]

Transcription factor (TF) proteins recognize specific DNA targets across the genome to regulate gene expression. In order to control precise cellular functions, TFs cooperate and compete with one another, forming complex gene regulatory networks (Zhou and O'Shea 2011; Jolma et al. 2015; Morgunova and Taipale 2017). Cooperative interactions between TFs, which are typically driven by direct contacts between compatible protein domains, have been extensively studied (Wotton et al. 1994; Jolma et al. 2015; Morgunova and Taipale 2017). TF competition, however, is still poorly understood, as few studies have directly addressed competitive interactions between TFs and the role of competitive binding in gene regulation (Miyamoto et al. 1997; Noro et al. 2011; Zhou and O'Shea 2011; Aow et al. 2013). TFs can compete for DNA binding whenever their target sites have partial or complete overlap. In the case of paralogous factors, that is, TFs from the same protein family, competition is especially important. TF paralogs arose from gene duplication and divergence during evolution (Fig. 1A) and are often associated with increased organismal complexity (Laudet et al. 1999; Banerjee-Basu and Baxevanis 2001; Amoutzias et al. 2007; Nitta et al. 2015; Murre 2019). Being conserved in their DNA-binding domains (DBDs), paralogous TFs have similar DNA binding specificities and share a large fraction of their putative target sites (Chen and Rajewsky 2007; Berger et al. 2008; Noyes et al. 2008; Singh and Hannenhalli 2008; Badis et al. 2009; Wei et al. 2010; Nakagawa et al. 2013; Weirauch et al. 2014; Shen et al. 2018). Most eukaryotic TFs belong to large protein families (Henikoff et al. 1997; Lin et al. 2008; Lambert et al. 2018), and coexpression of paralogs is common (see Discussion).

Whenever two or more paralogs are present in the nucleus at the same time, they may compete for DNA binding at their common target sites, with the potential for competitive binding being maintained throughout evolution (see Discussion). This also implies that the TFs' in vivo binding patterns, as assayed by chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-seq) and related techniques (Johnson et al. 2007; Rhee and Pugh 2012; He et al. 2015; Skene et al. 2018), are implicitly capturing the effects of TF-TF competition, although these effects are rarely studied explicitly on a genome-wide scale. Nevertheless, the important role that competition can play in gene regulation has been investigated for certain TFs and focusing on particular genomic sites. A prominent example is that of Hox proteins in *Drosophila*, where paralogs with slightly different DNA-binding specificities, driven by cofactor interactions, have been shown to compete at regulatory sites and tune gene expression during development (Noro et al. 2011; Slattery et al. 2011;

**Corresponding author: raluca.gordan@duke.edu**

Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.275145.120. Freely available online through the *Genome Research* Open Access option.

31:1216–1229 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/21; www.genome.org

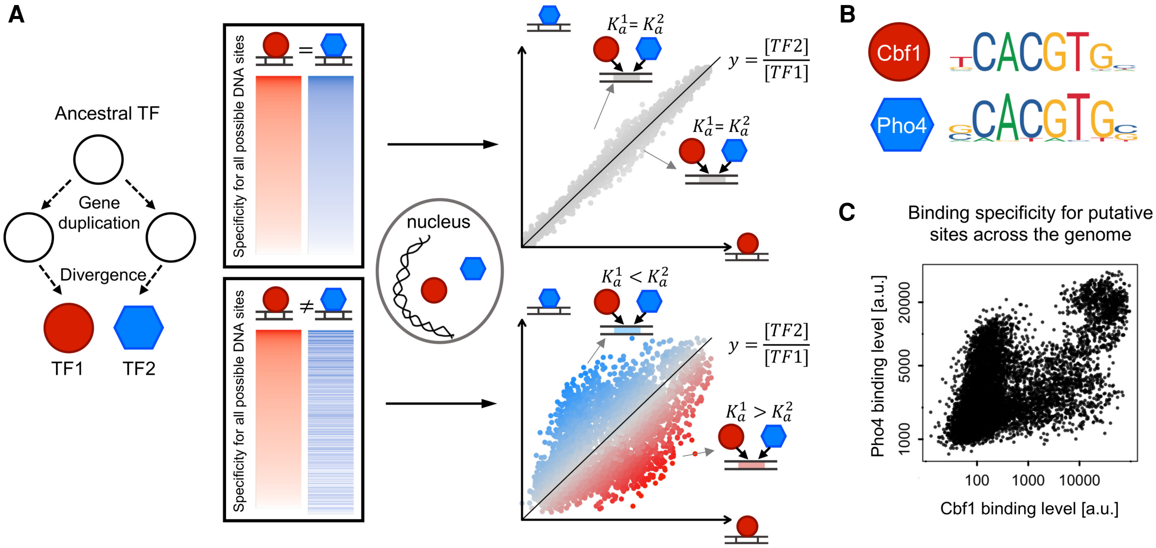
A
Ancestral TF
Gene duplication
Divergence
TF1
TF2
Specificity for all possible DNA sites
=
Specificity for all possible DNA sites
≠
nucleus
$K_a^1 = K_a^2$
$y = \frac{[TF2]}{[TF1]}$
$K_a^1 = K_a^2$
$K_a^1 < K_a^2$
$y = \frac{[TF2]}{[TF1]}$
$K_a^1 > K_a^2$
B
Cbf1
CACGTG
Pho4
CACGTG
C
Binding specificity for putative sites across the genome
Pho4 binding level [a.u.]
1000
5000
20000
100
1000
10000
Cbf1 binding level [a.u.]

**Figure 1.** Paralogous TFs compete for DNA binding. (*A*) Schematic showing different TF-TF competition scenarios. If paralogous TFs have identical DNA-binding preferences, then their binding is determined by the levels of the TFs in the nucleus (*top* panels). However, most paralogs have diverged in specificity, binding differently at different sites even in the absence of other proteins (Berger et al. 2008; Badis et al. 2009; Wei et al. 2010; Gordân et al. 2013; Shen et al. 2018); this divergence leads to complex patterns of competitive binding, which depend on the TFs' affinities and concentrations (*bottom* panels). (*B*) *S. cerevisiae* proteins Cbf1 and Pho4 s have similar, although not identical, DNA-binding specificities, as reflected by their position weight matrix (PWM) models (Sandelin et al. 2004). (*C*) Direct comparison between the in vitro binding levels of Cbf1 and Pho4 at their putative genomic binding sites, measured by genomic-context PBM (Gordân et al. 2013). Each data point corresponds to a 36-bp genomic region centered on a CACGTG site. Plot shows the fluorescence intensities from PBM assays, which are proportional to the level of bound TF at each genomic site. We note that fluorescence intensities are generally not directly comparable between PBM experiments for different proteins (see Supplemental Discussion). However, for all proteins tested, here and in prior studies (Berger et al. 2006; Siggers et al. 2011; Shen et al. 2018; Afek et al. 2020), the PBM fluorescence intensities correlate quantitatively with binding energies and equilibrium dissociation constants.

Crocker et al. 2015). The mammalian nuclear hormone receptor superfamily is another example, where peroxisome proliferator-activated receptor (PPAR) and thyroid hormone receptor (THR) proteins directly compete for binding to response elements involved in regulating lipid metabolism, cell growth, and differentiation (Miyamoto et al. 1997). In fungi, a well-known example of competitive TF-DNA binding is that of *Saccharomyces cerevisiae* basic helix-loop-helix (bHLH) proteins Cbf1 and Pho4, which perform different regulatory functions in the cell despite having similar DNA-binding specificities (Zhou and O'Shea 2011; Aow et al. 2013).

The result of competitive DNA binding between two paralogs can be difficult to predict. In the trivial case where their DNA-binding domains remain highly conserved during evolution and thus their specificity and affinity for DNA remain unchanged, we would expect the relative genomic occupancies of the paralogs to be proportional to their concentrations in the nucleus (Fig. 1A, upper panel), with any in vivo deviations from this pattern being due to the nuclear environment. Often, however, the DBDs of paralogous TFs accumulate mutations over time and the TFs start to diverge in specificity, especially at medium- and low-affinity target sites (Berger et al. 2008; Badis et al. 2009; Wei et al. 2010; Gordân et al. 2013; Shen et al. 2018). The differences in specificity and/or affinity between paralogs, which are intrinsically encoded in the DNA sequence, can lead to complex patterns of competitive binding (Fig. 1A, lower panel) even in the absence of additional effects from the nuclear environment. Currently, though, our understanding of competitive TF binding and its role in gene regulation is limited, and we lack the ability to predict how competition influences the genomic binding of individual TFs, both in vitro and in vivo.

Here, we use *S. cerevisiae* bHLH proteins Cbf1 and Pho4 as a model system to develop a high-throughput approach for characterizing TF-TF competition in vitro and exploring its role in TF binding and gene regulation in vivo. The bHLH domain is an essential DNA-binding domain that is highly conserved across eukaryotes (Jones 2004). Genes encoding this domain arose in early eukaryotes and then duplicated and diversified to give rise to proteins involved in critical cellular processes such as proliferation, differentiation, metabolism, and environmental response (Sailsbery and Dean 2012; Murre 2019). The domain was first elucidated in animals, where six major bHLH groups (A–F) were identified (Atchley and Fitch 1997). Fungal bHLH proteins, including Cbf1 and Pho4 (Robinson and Lopes 2000), are most closely related to group B—which includes mammalian factors such as MYC, MLX, and MITF and is believed to have been present in the common ancestor of fungi and animals (Sailsbery and Dean 2012; Sailsbery et al. 2012). This group of bHLHs is characterized by a conserved BxR motif at positions 5, 8, and 13 in the basic region of the DNA-binding domain (where B = H or K, and x stands for any amino acid) (Supplemental Fig. S1; Supplemental Table S1; Atchley and Fitch 1997; Sailsbery and Dean 2012). Similarly to group B bHLHs from other eukaryotes, *S. cerevisiae* Cbf1 and Pho4 recognize canonical CAnnTG E-box binding sites, with a strong preference for CACGTG (Atchley and Fitch 1997; Harbison et al. 2004; Maerkl and Quake 2007; Badis et al. 2008; Zhu et al. 2009). However, their quantitative binding levels to individual sites are different, depending on the genomic sequence context (Fig. 1B,C; Supplemental Fig. S1B; Gordân et al. 2013).

In addition to their highly conserved DNA-binding domain and their similar DNA-binding preferences, Cbf1 and Pho4 are an ideal system for our study because their competition for DNA binding has been shown to be important for Pho4's function in the cell. In particular, Pho4 plays an important role in the phosphate-responsive (*PHO*) signaling pathway (Ogawa et al. 2000; Zhou and O'Shea 2011). Pho4 is generally phosphorylated and located in the cytoplasm; when inorganic phosphate (Pi) becomes limited, Pho4 is dephosphorylated and translocated into the nucleus, where it binds a subset of CACGTG sites previously bound by Cbf1, leading to activation of downstream genes, most of which belong to the *PHO* regulon (Schneider et al. 1994; O'Neill et al. 1996; Zhou and O'Shea 2011). It remains unclear, though, why Pho4 competes differently with Cbf1 at different genomic sites and whether their direct competition for DNA binding can explain their binding patterns genome-wide. Insights into this system will be relevant to the many other bHLH proteins that have duplicated and evolved into different subfamilies with distinct functions, yet have maintained similar DNA-binding preferences.

In this study, we aimed to decipher the competitive DNA-binding patterns of TF paralogs Cbf1 and Pho4 by directly measuring their competition for thousands of genomic binding sites using a quantitative assay based on the protein-binding microarray (PBM) technology (Berger and Bulyk 2009; Siggers et al. 2011; Gordân et al. 2013; Shen et al. 2018). Furthermore, by comparing our in vitro competitive binding measurements against ChIP-seq and gene expression data, we aimed to determine whether the direct competition for DNA binding between TF paralogs is relevant for the TFs' genomic occupancies and gene regulatory patterns in the cell.

## Results

### Measuring the direct competition for DNA binding between paralogous TFs using "competition PBM"

The genomic binding profiles of TFs in the cell are typically assessed using chromatin immunoprecipitation coupled with high-throughput sequencing (Johnson et al. 2007), as well as related assays such as ChIP-exo (Rhee and Pugh 2012) and Cut&Run (Skene et al. 2018). Because these assays measure TF binding in the cell, they implicitly capture any in vivo effects of TF-TF competition. However, from ChIP data alone it is not possible to deconvolve the effects of competition from those of cofactors, DNA accessibility, and other cellular factors that influence TF binding. In order to isolate the effects of TF-TF competition and understand its contribution to genomic occupancy and gene regulation, we developed and used a controlled cell-free system where competitive binding can be easily quantified and modeled.

We developed a new assay, called "competition PBM," that leverages the quantitative and high-throughput nature of chip-based assays known as protein-binding microarrays (Berger et al. 2006), in order to measure the competitive binding profiles of paralogous TFs (Fig. 2A). Our assay uses DNA libraries containing tens of thousands of putative genomic binding sites for the competing TFs of interest (here, Cbf1 and Pho4), selected from in vivo-bound regions (here, ChIP-seq peaks [Johnson et al. 2007; Zhou and O'Shea 2011]), based on the rationale that these are the sites where TF-TF competition is most likely to occur. As in previous work (Gordân et al. 2013; Shen et al. 2018), we selected candidate genomic binding sites using universal PBM data, which contains comprehensive binding specificity measurements for all possible 8-mers (Berger and Bulyk 2009), and we used a loose cutoff for calling binding sites in order to cover a wide range of binding affinities (Methods). Given that paralogs differ in specificity mostly at medium- and low-affinity sites, we expect competitive binding to be different at these sites compared to the high-affinity sites that are bound similarly by TF paralogs (Shen et al. 2018).

After selecting a large set of genomic sites where competition between the TFs of interest is likely, we synthesized the DNA library on a chip and incubated it with the two proteins at different concentrations relative to each other (Fig. 2B,C; Supplemental Table S2B). To facilitate the interpretation of the data, we kept the concentration of one TF paralog constant (we henceforth refer to this protein as the "main TF") and varied the concentration of the competitor. Using a fluorophore-conjugated antibody specific for the main TF, we measured its DNA-binding level in the presence of various concentrations of the TF competitor, thus directly probing the effects of competition.

### Competition between Cbf1 and Pho4 determines their in vitro DNA-binding patterns

We performed competition binding assays using Cbf1 as the main TF and Pho4 as the competitor TF, which corresponds to their physiological scenario: Pho4 is typically present at very low (although detectable) levels in the nucleus; in phosphate-limited conditions, Pho4 is translocated into the nucleus where it competes with Cbf1 for binding to E-box CACGTG sites (O'Neill et al. 1996; Zhou and O'Shea 2011). Keeping the concentration of Cbf1 constant at 2 μM, we measured its DNA binding in the presence of Pho4, with the competitor present at four different concentrations: 0.05 μM, 0.4 μM, 2 μM, and 8 μM (Fig. 2B; Supplemental Table S2C). The lowest concentration of Pho4 was chosen to mimic the wild type Pi-rich conditions in yeast, where Pho4 is present in the nucleus at a much lower level than Cbf1 and is not expected to significantly compete with Cbf1 at any genomic site (O'Neill et al. 1996; Komeili and O'Shea 1999; Zhou and O'Shea 2011). Indeed, our in vitro data confirms that Pho4 has minimal effects on Cbf1 binding when the two proteins are at 0.05 μM and 2 μM concentrations, respectively (Supplemental Fig. S2A). Ideally, we would perform the competition PBM assays at protein concentrations similar to those of active Cbf1 and Pho4 in the yeast nucleus. Because this information is not available, we chose a concentration for the main TF that leads to moderate DNA-binding levels, whereas the concentration of the competitor TF was set to cover a wide dynamic range in order to capture various competition scenarios. As expected, increasing the concentration of the competitor TF (Pho4) leads to decreased DNA binding by the main TF (Cbf1). In fact, most genomic binding sites, especially in the medium- to low-affinity range, show a decreased level of Cbf1 binding as the concentration of Pho4 increases (Fig. 2B), with a corresponding increase in Pho4 binding (Supplemental Fig. S2B, left panel).

The decrease in Cbf1 binding is different at different DNA sites. At some genomic sites (such as the one marked with blue circles and arrows in Fig. 2B), Cbf1 is efficiently outcompeted by Pho4, as illustrated by a large decrease in Cbf1 binding level as the Pho4 concentration increases. At the highest concentration of competitor (8 μM), the binding level of Cbf1 is only 2.6% of its binding level at the lowest competitor concentration (0.05 μM), illustrating the magnitude of the competition effects. In contrast, at the genomic site marked with red arrows in Figure 2B, even high levels of Pho4 competitor have nonsignificant effects on Cbf1-DNA binding ($P$-values > 0.37) (Supplemental Table S2D). Overall, our data show that Pho4 competes with Cbf1 differently

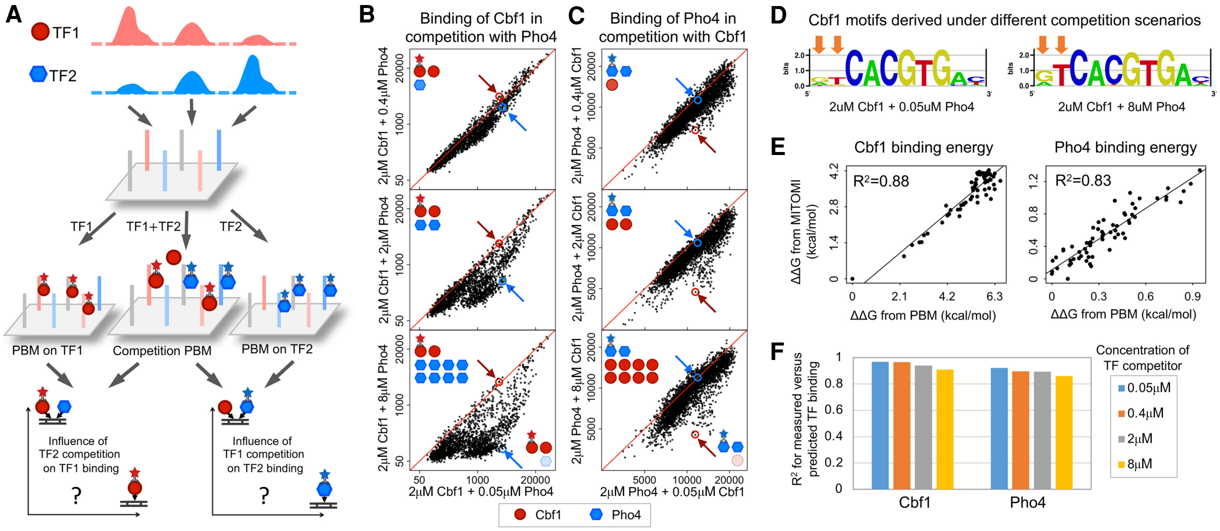

**Figure 2.** Characterizing the DNA-binding patterns of Cbf1 and Pho4 using "competition PBM." (*A*) Schematic of the competition PBM assay. Genomic DNA sites are selected from the ChIP-seq peaks of the TFs of interest and synthesized on a DNA chip, similarly to previous work (Siggers et al. 2011; Shen et al. 2018). The chip is then incubated with the TF paralogs of interest, alone or in competition, and the binding is quantified using fluorophore-conjugated antibodies. The effects of TF-TF competition are then quantified by comparing the binding of one TF under conditions where the concentration of the competitor is varied. See Methods for details. (*B*) DNA-binding levels for TF Cbf1, at 2 µM concentration, in competition with Pho4 at increasing concentrations: 0.05 µM (*x*-axis), 0.4 µM, 2 µM, and 8 µM (*y*-axes). The condition shown on the *x*-axis (which includes Pho4 at the low concentration of 0.05 µM) mimics the in vivo environment in rich media, where Pho4 levels in the nucleus are very low but still detectable (Schneider et al. 1994; O'Neill et al. 1996; Zhou and O'Shea 2011). Each of the 2014 data points corresponds to a putative Cbf1 binding site (defined as a site with Cbf1 binding intensity above negative controls) (Methods) in its native genomic sequence context. All sequences tested are 36 bp long, centered at the binding site. Data points *below* the diagonal demonstrate the influence of Pho4 competition, which results in decreased Cbf1-DNA binding. Blue circles and arrows point to a genomic site where Cbf1 binding decreases significantly with increasing Pho4 levels. Red circles and arrows point to a genomic site where Cbf1 binding is not significantly affected by increasing Pho4 levels. Data and statistics are available in Supplemental Table S2. (*C*) Similar to panel *B* but for Pho4 as the main TF and Cbf1 as the competitor TF. Each of the 3341 data points corresponds to a putative Pho4 binding site (defined as a site with Pho4 binding intensity above negative controls) (Methods). Red circles and arrows point to a genomic site where Pho4 binding decreases significantly with increasing Cbf1 levels. Blue circles and arrows point to a genomic site where Pho4 binding is not significantly affected by increasing Cbf1 levels. See also Supplemental Table S2. (*D*) Motif logos for Cbf1 derived from competitive binding data. Arrows mark the positions that are most different between the two motifs. See Supplemental Figure S2D for motifs derived from additional competitive binding conditions. (*E*) Comparing PBM-derived binding energies (ΔΔG) with binding energies derived from independent MITOMI experiments (Maerkl and Quake 2007). Each data point corresponds to one DNA sequence used in the MITOMI DNA libraries (the CACNNN library for Cbf1, where the lowest binding energy corresponds to CACGTG; and CACGTGNNN library for Pho4) (Methods; Maerkl and Quake 2007). See Supplemental Figure S4B for comparisons using additional MITOMI sequence sets. (*F*) Prediction accuracy for the biophysical model of competitive DNA binding by Cbf1 and Pho4. Bar plot shows the squared Pearson's correlation coefficients ($R^2$) between measured and predicted binding levels at various concentrations of competitor. See Supplemental Figure S4F for full comparisons.

at different genomic regions, even in a simple cell-free system where no other nuclear factors are present.

The observed differential competition is consistent with the divergence in DNA-binding specificity and affinity between Cbf1 and Pho4 (Zhou and O'Shea 2011; Le et al. 2018). More broadly, recent studies have revealed that paralogous TF pairs have diverged in their intrinsic DNA-binding specificities, especially at medium- and low-affinity sites, with each paralog having individual preferences for a subset of DNA sequences (Shen et al. 2018). In the case of Cbf1 and Pho4, we would expect Pho4 to outcompete Cbf1 efficiently at Pho4-preferred sites (typically A/C/GCACGTG) but not at Cbf1-preferred sites (typically TCACGTG) (Supplemental Fig. S1B; Fisher and Goding 1992). Indeed, this is reflected in our in vitro competition PBM data. To illustrate how the individual preferences of Cbf1 and Pho4 play a role in their competitive binding, we derived position weight matrix (PWM) motifs for Cbf1 from its binding data under different competition scenarios (Fig. 2B; Methods). With competitor Pho4 at a low concentration (0.05 µM), Cbf1 has a motif logo that is highly similar to the one derived from universal PBM data where Cbf1 was tested individually (Supplemental Fig. S2C). However, when competitor Pho4 is present at a high concentration (8 µM), the core TCACGTG stands out in the Cbf1 motif due to Cbf1's preference for TCACGTG versus Pho4's preference for A/C/GCACGTG. Cbf1's preference for a T upstream of the core CACGTG is more and more evident as the concentration of Pho4 increases (Fig. 2D; Supplemental Fig. S2D), which is explained by the fact that A/C/GCACGTG sites are increasingly occupied by Pho4 and less by Cbf1.

We also performed Cbf1-Pho4 competition assays by considering Cbf1 as the competitor and studying its influence on Pho4-DNA binding (Fig. 2C). We kept the Pho4 concentration constant at 2 µM and measured its DNA binding levels with Cbf1 at four different concentrations (0.05 µM, 0.4 µM, 2—M, 8 µM). As expected, we observed different effects of Cbf1 competition on Pho4 binding at different genomic sites. In addition, we found that the overall pattern of competitive binding was different between Cbf1 (Fig. 2B) and Pho4 (Fig. 2C), with many Pho4 sites being only moderately affected by Cbf1 competition. These patterns are consistent with the intrinsic differences in DNA-binding preferences between Cbf1 and Pho4 (Supplemental Fig. S2E), with many Pho4-specific binding sites having Cbf1 binding affinities in the negative control range. Similarly to the Cbf1 binding motifs under different competition scenarios, the motif logos for Pho4 show that, as the concentration of the competitor increases, the preference of Pho4 for A, C, or G upstream of the core CACGTG becomes clearer (Supplemental Fig. S2D). Overall, our new data show that the direct competition for DNA binding between TF paralogs shapes their genomic binding profiles in a manner that depends directly on the intrinsic specificity differences between the paralogs.

In the current study, we expressed TFs as recombinant proteins with epitope tags (His or GST), and we used fluorescent antibodies for the tags to quantify TF-DNA binding. To ensure that the choice of tags did not influence TF binding or TF-TF competition, we also performed control experiments where one TF (Cbf1) was expressed with different tags. We did not observe any tag-specific effects (Methods; Supplemental Fig. S3A,B; Supplemental Table S2F–H).

## Modeling competitive DNA binding by Cbf1 and Pho4

Our competition PBM data provide direct evidence that Cbf1 and Pho4 compete for binding to their genomic sites in vitro. Next, we investigated the general principles that underlie this process, focusing on how differences in protein concentrations and binding affinities may lead to different competitive binding patterns. We tested whether a simple biophysical model of TF occupancy (Gerland et al. 2002; Djordjevic et al. 2003) can explain the in vitro competition data. Briefly, for a given DNA site *i*, the probability that the site is bound by the main TF can be written as

$$P_i = \frac{[TF^1]/K^1_{d,i}}{1 + [TF^1]/K^1_{d,i} + [TF^2]/K^2_{d,i}}, \qquad (1)$$

where $[TF^1]$ and $[TF^2]$ are the concentrations of free main TF and free competitor TF, respectively, and $K^1_{d,i}$ and $K^2_{d,i}$ are their equilibrium dissociation constants at site *i*.

To obtain affinity dissociation constants ($K_d$) for the sequences of interest, we used the approach of Siggers et al. (2011) to derive $K_d$ values from our custom PBM data by performing binding experiments at multiple protein concentrations. We performed such experiments for Cbf1 and Pho4, and we fitted the saturation curves to estimate the $K_d$ for each TF at each DNA site (Supplemental Fig. S4A). Next, we assessed the accuracy of our PBM-derived dissociation constants by using them to compute binding energies (ΔΔG) that we compared to the energetic binding measurements for Cbf1 and Pho4 obtained from mechanical trapping of molecular interactions (MITOMI) assays (Maerkl and Quake 2007). Such measurements are available for a few hundred artificial DNA sequences, which we included in our DNA library. We observed an excellent agreement between the two techniques ($R^2 = 0.83$–$0.88$), over a wide range of binding energies (Fig. 2E; Supplemental Fig. S4B; Methods). We also compared the PBM-derived ΔΔGs against binding energies predicted using a neural network model trained on Binding Energy Topography by sequencing (BET-seq) data, which are available for NNNNNCACGTGNNNNN sequences (Le et al. 2018). Similarly to our comparison against MITOMI data, we found a strong agreement between PBM-derived and BET-seq-derived binding energies ($R^2 = 0.72$–$0.74$) (Supplemental Fig. S4C). However, superior to previous studies, here, we extended our binding measurements to all CACGTG and non-CACGTG E-box genomic sites potentially bound by Cbf1 and Pho4 (Supplemental Fig. S4D), and we considered longer sequences flanking the E-box binding sites, which can significantly influence TF binding (Supplemental Fig. S4C).

Next, we incorporated the competition between paralogous TFs into the equilibrium thermodynamics model in Equation (1), and we expressed the occupancies of each TF paralog using standard binding isotherms (Methods; Supplemental Fig. S4E). Plugging in the PBM-derived $K_d$ values into Equation (1), we can then predict Cbf1 and Pho4 binding under any competition conditions. This standard biophysical model achieved high accuracy ($R^2 = 0.86$–$0.96$) in predicting the occupancies of Cbf1 and Pho4 under four different competition scenarios (Fig. 2F; Supplemental Fig. S4F), suggesting that, for simple systems of two competing paralogs without cobinding factors, the binding process is largely described by a standard biophysical model. In addition, our results highlight the value of PBM experiments for individual TFs, as $K_d$s derived from such experiments are sufficient (at least in the case of Cbf1 and Pho4) to accurately predict the equilibrium binding of the TFs in competition, even without having to perform competition experiments. For more complex systems, the competition model can be modified to account for additional interactions, such as dimerization partners and proteins cofactors (see Methods and Discussion for further details), enabling generalization of our competition study to other TF paralogs.

## In vivo TF binding data reflect the competitive binding patterns characterized in vitro

Next, we asked whether the DNA-binding patterns resulting from the competitive binding of Cbf1 and Pho4 in vitro are also reflected in their genomic occupancies in the complex environment of the cell. To assess the in vivo effects of Cbf1 on the binding patterns of Pho4, we leveraged available Pho4 ChIP-seq data from strains with constitutively nuclear-localized Pho4 (due to *PHO80* deletion) and with Cbf1 present versus absent, that is, yeast strains *pho80Δ* and *cbf1Δpho80Δ*, respectively (Fig. 3A; O'Neill et al. 1996; Zhou and O'Shea 2011). As expected, the Pho4 binding level is overall lower when Cbf1 is present versus absent (Supplemental Fig. S5A), consistent with our in vitro observations. In addition, Cbf1 acts differently on Pho4 binding at different genomic sites in vivo, which is again consistent with our in vitro competition data. To illustrate the differential competition, we introduce here the notion of "resilience" of one TF paralog to competition from another paralog, defined as the logarithm of the fold-change in DNA binding of the main TF when the competitor is present at a high versus a low concentration (Methods; Fig. 3B,F). Smaller values of the resilience indicate larger effects from the TF competitor.

We first computed Pho4's in vitro resilience to Cbf1 competition using two representative competition scenarios: 2 μM Pho4 + 0.05 μM Cbf1 versus 2 μM Pho4 + 2 μM Cbf1 (Fig. 3C; Supplemental Table S3), and we analyzed the in vivo binding of Pho4 and Cbf1 at genomic sites with high versus low resilience. For example, at a high-resilience Pho4 target site upstream of the *PHO84* gene (CCACGTGC), we found that the ChIP-seq signal was highly similar between the *pho80Δ* and *cbf1Δpho80Δ* strains, that is, in the presence versus the absence of Cbf1, consistent with Pho4's high in vitro resilience at this site (Fig. 3C, upper right panel). In contrast, at a genomic site with low Pho4 in vitro resilience (TCACGTGC, located upstream of the *SER33* gene), Pho4 shows virtually no in vivo binding signal, that is, no ChIP-seq peak, when Cbf1 is present but does show a ChIP-seq peak in the absence of Cbf1 (Fig. 3C, bottom right panel). Similarly to the resilience measure computed based on our in vitro competition data, we can use ChIP-seq data to compute in vivo resilience scores (Methods). We found that genomic sites with higher in vitro resilience also have higher resilience in vivo (Fig. 3D), demonstrating that

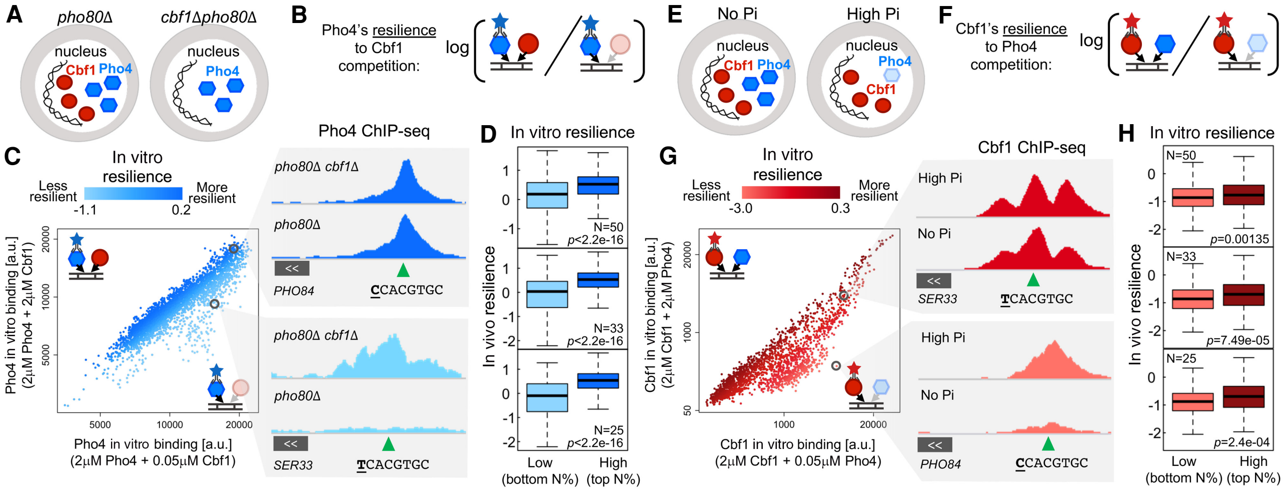

**Figure 3.** In vivo ChIP-seq data reflect the in vitro competitive binding patterns of Cbf1 and Pho4. (*A*) Data from *S. cerevisiae* strains *pho80Δ* and *cbf1Δpho80Δ* (Zhou and O'Shea 2011) were used to assess the effect of Cbf1 competition on Pho4. (*B*) Definition of the term "resilience" in the context of Pho4-Cbf1 competition. A smaller value of resilience indicates a larger impact from TF competition. (*C*) *Left*: Genomic Pho4 binding sites tested in the competition PBM assay, colored by Pho4's resilience to Cbf1 competition. *Right*: Genome browser tracks showing in vivo binding data at sites with high in vitro resilience (*upper* panel) versus low in vitro resilience (*lower* panel). The site with high resilience is less influenced by competition in vivo, whereas at the site with low resilience, Pho4 is efficiently outcompeted by Cbf1. (*D*) Comparisons between the in vivo resilience scores of genomic sites with low versus high in vitro resilience. Plots show comparisons between the top versus bottom N% of sites, sorted in decreasing order of in vitro resilience, for N = 50 (*top*), 33 (*middle*), and 25 (*bottom*). See Supplemental Figure S5B for a full comparison of resilience scores at individual binding sites. (*E*) Data from *S. cerevisiae* EY57 cells grown in media with no inorganic phosphate and high inorganic phosphate (Zhou and O'Shea 2011) were used to assess the effect of Pho4 competition on Cbf1. (*F*,*G*,*H*) Similar to panels *B*, *C*, and *D* but showing the effects of Pho4 competition on Cbf1 binding.

the DNA-binding patterns resulting from TF-TF competition in the cell are consistent with our in vitro observations, despite the complexities of the cellular environment.

Considering Cbf1 as the main TF and Pho4 as the competitor, we analyzed available Cbf1 ChIP-seq data for the wild-type *S. cerevisiae* strain EY57 under two phosphate conditions ("no Pi" and "High Pi") where Pho4 is present at high versus low levels in the nucleus (Fig. 3E). As before, we computed the in vivo resilience from the ChIP-seq data and the in vitro resilience from representative in vitro competition scenarios (2 μM Cbf1 + 0.05 μM Pho4, vs. 2 μM Cbf1 + Cbf1 2 μM Pho). We again observed consistency between Cbf1's resilience to Pho4 competition in vitro and in vivo, both at individual sites (Fig. 3G) and genome-wide (Fig. 3H).

In summary, analysis of in vivo ChIP-seq data confirms that Cbf1 and Pho4 compete in the cell for genomic occupancy in a pattern consistent with our in vitro competitive binding data. The significant contribution of TF-TF competition to the overall binding profiles of paralogous TFs indicates that, when interpreting data from in vivo assays such as ChIP-seq, we should keep in mind that the measured binding levels will depend on the competition between the tested TF and its paralogs present in the cell nucleus.

## The Cbf1-Pho4 competition contributes to the differential regulation of Pho4 target genes genome-wide

TFs exert their regulatory functions through direct interactions with DNA sites across the genome. Because TF-TF competition is a critical determinant of TF-DNA binding, we asked whether its influence is also reflected at the level of gene expression regulation. In the Cbf1-Pho4 model system analyzed here, it is known that Pho4 functions as a transcriptional activator and that, out of all the potential Pho4 target genes, only a subset are actually activated in response to Pi starvation, that is, only a subset are bona fide Pho4 targets under physiological conditions. The remaining genes whose promoters contain putative Pho4 binding sites are activated only when Cbf1 is absent in the nucleus, that is, are Pho4 targets only in *cbf1Δ* (Zhou and O'Shea 2011). The differences in expression patterns between these two subsets of genes suggest that competition from Cbf1 might play a role in determining which genes are regulated by Pho4 in the cell.

To analyze the influence of Cbf1 competition on gene regulation by Pho4, we focused on 28 Pho4 target genes whose promoter regions contain a single Pho4 binding site (Methods). Out of the 28 genes, 18 are Pho4 targets in physiological conditions and 10 genes are Pho4 targets only in *cbf1Δ* (Supplemental Table S4A), with the two sets of genes showing distinct expression patterns in response to phosphate limitation when Cbf1 is present versus absent in the cell (Supplemental Fig. S5C; O'Neill et al. 1996; Zhou and O'Shea 2011). Comparing the in vitro Pho4 binding levels at the promoters of the two sets of genes, we found no significant difference (Fig. 4A, left panel), indicating that the intrinsic Pho4-DNA binding specificity cannot explain the differences in gene expression patterns. Next, we asked whether the two sets of promoters have different DNA accessibility levels, which could lead to differential Pho4 binding in vivo. However, nucleosome occupancy data (Methods) argued against this hypothesis, as the two sets of promoters have similar accessibility levels (Supplemental Fig. S5D). When taking into account the influence of Cbf1 competition, we found that the two groups of Pho4 targets are significantly different in their Pho4 binding levels, both in vitro (Fig. 4A, right plot) and in vivo (Fig. 4B, right plot). This indicates that direct competition from Cbf1 enables differences in Pho4 occupancies at binding sites with indistinguishable intrinsic Pho4 binding preferences, which subsequently contributes to the differential gene activation by Pho4 in the *PHO* signaling pathway. Our results are consistent with those of Aow et al. (2013), who

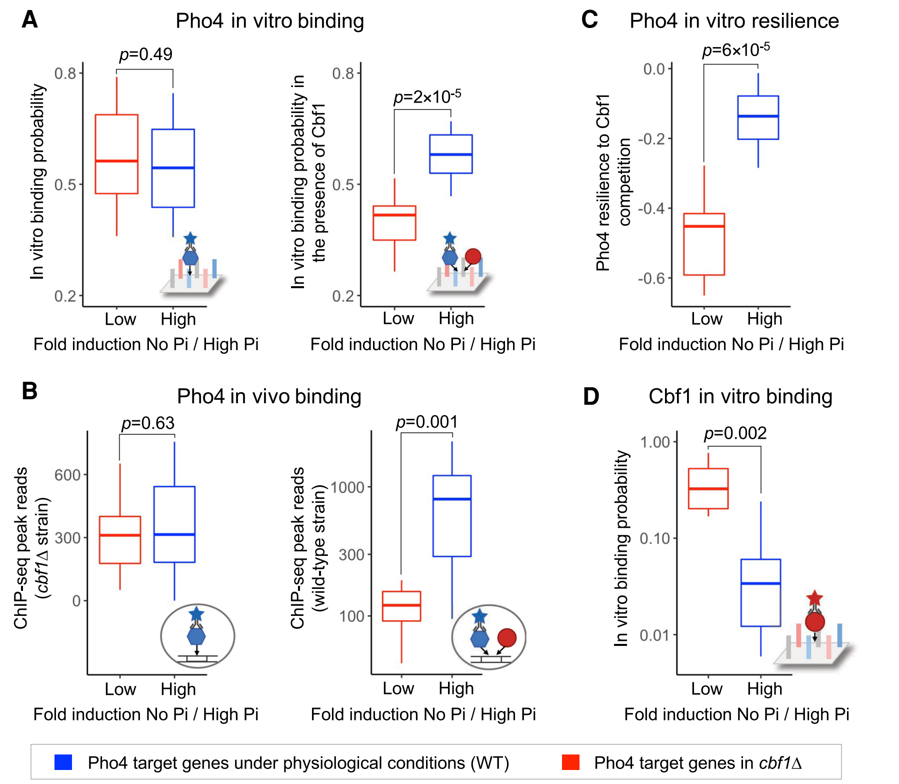

**Figure 4.** TF-TF competition contributes to differential gene activation. Box plots show the in vitro and in vivo TF binding data for sets of genes with low versus high fold induction in response to phosphate starvation (No Pi). Blue: Genes that are activated by Pho4 under physiological conditions, that is, in a wild-type strain where Cbf1 is present at physiological levels (Zhou and O'Shea 2011). Red: Genes that are activated by Pho4 only when Cbf1 is absent from the cell, that is, in a *cbf1*Δ strain (Zhou and O'Shea 2011). The two sets of Pho4 target genes were compared in terms of: (*A*) Pho4 in vitro binding (at 2 µM concentration) in the absence of Cbf1 (*left*) and in the presence of Cbf1 at 2 µM concentration (*right*), as measured by PBM and competition PBM, respectively; (*B*) Pho4 in vivo binding, as measured by ChIP-seq, in the absence of Cbf1 (*left*, *cbf1*Δ strain) and in the presence of Cbf1 (*right*, wild-type strain); (*C*) Pho4's in vitro resilience to Cbf1 competition (computed between competition PBM conditions: 2 µM Pho4 + 2 µM Cbf1 versus 2 µM Pho4 + 0.05 µM Cbf1) (Methods); and (*D*) Cbf1 in vitro binding (at 2 µM concentration), as measured by PBM. In vitro binding probabilities were computed from PBM or competition PBM data (Methods). In vivo binding levels are shown as read counts computed for Pho4 ChIP-seq peaks (Zhou and O'Shea 2011).

found that the competitive binding of Cbf1 and Pho4 at palindromic NNCACGTGNN sites explained the expression patterns of a reporter gene with high accuracy.

We also investigated Pho4's in vitro resilience to Cbf1 competition and found that it also distinguishes between the Pho4 target genes in physiological conditions (blue) versus the target genes unique to the Cbf1 knockout strain (red) (Fig. 4C). This implies that the Pho4 targets under physiological conditions have promoter sites resilient to Cbf1 competition, so that they are robustly activated by Pho4 in response to phosphate starvation. This set of genes includes *PHO84*, which encodes a high-affinity Pi transporter and whose promoter contains a C̲CACGTGC Pho4 binding site that is resilient to Cbf1 competition both in vitro and in vivo (Fig. 3C). In contrast, genes that are activated by Pho4 only in *cbf1*Δ are vulnerable to Cbf1 competition, so that the presence of Cbf1 in wild-type cells effectively prevents these genes from being activated by Pho4. A representative example is *SER33* (Fig. 3C), which is not part of the *PHO* regulon and whose promoter contains a T̲CACGTGC that is poorly bound by Pho4 when Cbf1 is present. Consistent with the results above, we found that the in vitro DNA binding probabilities of purified Cbf1 can also differentiate between the two sets of genes, with a trend opposite to the Pho4 binding probability (Fig. 4D). Thus, competition from Cbf1 effectively contributes to the specification of the functional Pho4 targets.

## Discussion

Despite the important role of TF-TF competition in gene regulation, studying this competition based on existing DNA-binding data is difficult. In vivo techniques that measure TF-DNA binding, such as ChIP-seq, ChIP-exo, and Cut&Run (Johnson et al. 2007; Rhee and Pugh 2012; He et al. 2015; Skene et al. 2018), reflect the genomic occupancy of one TF in a particular cellular context and thus in a very specific competition scenario. From individual ChIP-seq data sets, it is impossible to infer how/whether the competition for DNA binding among TF paralogs influences the genomic binding profile of the TF of interest. Data from carefully controlled experiments where a TF is ChIP'ed in the presence versus the absence of a competitor TF are rare, and even such data may not be quantitative enough or may not have the resolution to allow investigation of the TF-TF competition effects. To complement the in vivo data, we propose using high-throughput in vitro binding assays, such as the competition PBM approach introduced here, which leverages the quantitative nature of on-chip protein-DNA binding measurements (Berger and Bulyk 2009; Siggers et al. 2011; Gordân et al. 2013; Shen et al. 2018). By performing the competition experiments in a cell-free system where experimental variables are well controlled, we were able to generate highly quantitative data that directly reflect the influence of competition on TF binding to genomic sites in vitro. Next, using the in vitro competition data as reference, we reinterpreted the in vivo ChIP-seq data and found evidence of TF-TF competition in the cell, as well as confirming its role in gene regulation.

Overall, our results show that TF-TF competition is a sequence-specific process that translates the intrinsic differences in DNA specificity between paralogous TFs, which can be thoroughly and quantitatively characterized in vitro, into differential binding and gene regulation in the cell. Our findings are in great agreement with previous small-scale studies of TF paralogs, such as the Hox factors in *Drosophila* (Noro et al. 2011; Slattery et al. 2011; Crocker et al. 2015) and the POU homeodomain factors in mammals (Ferraris et al. 2011). Similar to what we found for the Cbf1/Pho4 system, the in vitro determinants of TF binding and competition—which, in the case of both Hox and POU factors, include cooperating proteins—were recapitulated in vivo (Ferraris et al. 2011; Noro et al. 2011). These results reaffirm that mechanistic in vitro studies can provide important insights into TF-TF competition and its role in gene regulation.

In the case of Cbf1 and Pho4, their potential for competitive DNA-binding seems to have been maintained throughout evolution. Cbf1 and Pho4 are closely related to "group B" bHLH proteins (Atchley and Fitch 1997; Sailsbery and Dean 2012), a large family of bHLHs in animals that also have similar binding preferences and likely compete with one another for binding to CAnnTG E-box sites. Similar to fungi, animal B-type bHLHs are involved in energy metabolism, lipid metabolism, cell growth, and proliferation (Jones 2004; Murre 2019). This functional conservation raises the question: Were competitive Cbf1-like and Pho4-like subfamilies already present in the common ancestor of fungi and animals, or did this feature evolve during the expansion of B-type bHLHs in fungi?

Previous work in higher fungi showed that there are 12 distinct phylogenetic bHLH subgroups, F1–F12, with Cbf1 belonging to F3 and Pho4 belonging to F6 (Sailsbery et al. 2012). Cbf1 and orthologs from the F3 group are involved in chromosome segregation and methionine biosynthesis, whereas Pho4 and orthologs from the F6 group are involved in phosphate uptake and sexual/asexual development. Some bHLHs, such as Cbf1, are strongly conserved in higher fungi and were likely present in the fungal ancestor. Pho4 is not universal in all fungi; however, the Pho4-mediated phosphate regulation network is conserved and functional in early-diverging fungi, such as *Blastocladiella emersonii* (Gomes-Vieira et al. 2018). To further investigate the evolutionary origin of fungal Pho4 and Cbf1, and their relationship to animal B-type bHLHs, we performed a bioinformatic analysis of Pho4 and Cbf1 in early-diverging fungi and animals (Methods). We found that the progenitors of the Pho4-like and Cbf1-like subfamilies were both present in the ancestor of fungi and animals (Supplemental Fig. S1A). We identified Max-like protein X (MLX) and Microphthalmia-associated transcription factor (MITF) subfamilies in animals as the likely descendants of these Pho4-like and Cbf1-like progenitors, respectively. The conservation of specific amino acids in each subfamily DNA-binding domain (Supplemental Fig. S1A) suggests that binding preferences of these TFs and their competition could be ancestral and conserved. A more recent example of competition between TF paralogs that may have been maintained throughout evolution can be found in the SP family of Cys2His2 zinc finger proteins in birds and mammals. The SP3 and SP4 subfamilies are paralogs and DNA-binding competitors of SP1, and they have accumulated convergent substitutions at homologous positions to SP1, several times during evolution, presumably to maintain competitive binding (Yokoyama and Pollock 2012).

Paralogous TFs are often coexpressed in the cell. The yeast *S. cerevisiae* genome encodes ~250 TF proteins belonging to 30 structural families (Weirauch et al. 2014; Ho et al. 2018). Excluding TFs with unknown structural families and zinc finger proteins, which represent a special family with complex and diverged patterns of specificity, we estimate that ~31% of yeast TFs (Supplemental Table S4B) are potentially competing with their paralogs for binding to genomic target sites (in this analysis, we considered a TF gene as "expressed" if the level of the corresponding TF protein was above the 75th percentile of all proteins) (Supplemental Fig. S5E; Methods). In higher eukaryotes, we expect this fraction to be even higher. Indeed, an analysis of the expression profiles of human TFs across 37 tissue types (Methods) revealed that, out of 58 TF families that contain more than one paralog (Lambert et al. 2018), 43 of them have at least two family members coexpressed in at least one tissue type (Supplemental Fig. S5E; Supplemental Table S4B). MITF and MLX, which are the mammalian bHLH proteins most closely related to *S. cerevisiae* Cbf1 and Pho4, are among the coexpressed family members, with TFs from the MLX and MITF orthologous group (Huerta-Cepas et al. 2019) being coexpressed in 15 out of the 37 tissues tested (Supplemental Table S4C). In a more stringent analysis, where TFs from each family were further separated into clusters of proteins with highly similar DNA-binding motifs (Lambert et al. 2018; Methods), we found that 154 out of 158 TF clusters with two or more members (~97%) contained at least two TF paralogs coexpressed in at least one cell type (Supplemental Fig. S5E; Supplemental Table S4D). These data reinforce that the potential for direct competition for DNA binding between paralogous TFs is widespread in human cells.

Our study is most closely related to Zhou and O'Shea (2011), who focused on the determinants of Pho4 genomic binding and function, including competition from Cbf1, and Aow et al. (2013), who investigated the differential binding of Pho4 and Cbf1 and its role in activating reporter gene expression, focusing on the 16 palindromic NNCACGTGNN sites. Our results extend and complement these studies by providing a comprehensive and quantitative view of the direct competition between Cbf1 and Pho4 genome-wide, using a PBM-based approach that can easily be applied to other regulatory systems. Moreover, our approach does not rely on existing TF-DNA binding affinity measurements, which is one limitation of the Aow et al. (2013) study. The PBM technology is straightforward to implement, and it uses commercially available DNA chips that are both cost-effective and high-quality. Furthermore, as shown in this study and in previous work by us and others (Siggers et al. 2011; Gordân et al. 2013; Afek et al. 2020), PBM data can be used to infer protein-DNA binding energies and equilibrium dissociation constants that are highly correlated with measurements from independent, small-scale assays. As shown here, the PBM-derived affinities can then be used directly in biophysical models to infer the occupancies of competing TFs at various concentrations. This alleviates the need to perform high-throughput assays in order to measure the competitive binding patterns of the TFs of interest and makes our approach easy to generalize to systems with more than two competitors.

We also found great agreement between our PBM-derived binding energies ($\Delta\Delta G$) and the $\Delta\Delta G$ values predicted for NNNNNCACGTGNNNNN sites using a deep neural network model trained on BET-seq data (Le et al. 2018), which is based on a combination of microfluidics and high-throughput sequencing (Supplemental Fig. S4C). In contrast to the BET-seq data, which covers a large number of artificial DNA sites, our PBM measurements focus on genomic sites targeted by Cbf1 and Pho4 in the cell, which include variants of the CAnnTG E-box not tested by BET-seq (Supplemental Fig. S4D). These variants are typically lower-affinity sites. However, previous studies have shown that paralogous TFs diverged in specificity mainly at medium- and low-affinity sites (Slattery et al. 2011; Crocker et al. 2015; Shen et al. 2018), making these sites particularly relevant for TF-TF competition. In addition to including lower-affinity CAnnTG variants, the DNA sequences used in our PBM library extend beyond the five neighboring bases of the E-box core binding sites (which were tested by BET-seq) to include 15 bp of genomic DNA on each side of the CAnnTG E-box. The additional genomic context can affect the binding affinity of Cbf1 and Pho4, as shown in our analyses (Supplemental Figs. S4C, S6A) and previous work

(Gordân et al. 2013) and can thus influence the competitive binding of the two TFs. Similarly to our previous work on the DNA binding specificity of Cbf1 (Gordân et al. 2013), here, we also observed that the flanking regions of the E-box binding site have a significant influence on Pho4 binding specificity, likely exerted through DNA shape (Supplemental Fig. S6). For both Cbf1 and Pho4, models of DNA-binding specificity benefit significantly from including features that reflect the DNA shape of the flanking regions (Supplemental Fig. S6A,B). However, in the case of Pho4, flanking shape features have a smaller effect on the overall accuracy of our binding models (Supplemental Fig. S6B), consistent with the findings of Le et al. (2018) that the magnitude of nonadditivity is smaller for Pho4 than Cbf1 binding sites. Nevertheless, given that DNA shape readout contributes, albeit to different extents, to the binding affinities of both Cbf1 and Pho4, it implicitly plays a role in their competitive binding.

Our in vitro approach can be expanded beyond two competing TFs to incorporate more of the in vivo factors relevant for TF binding and regulation. The functionality of many TFs involves interactions with other regulatory proteins. In our model system, both Cbf1 and Pho4 can interact with other factors: Cbf1 forms a regulatory complex with Met4 and Met28 to regulate sulfur metabolism genes (Kuras et al. 1997), whereas Pho4 cooperates with Pho2 in the regulation of the genes in the *PHO* regulon (Ogawa et al. 2000; Zhou and O'Shea 2011). Although in our current study we found that the in vivo competitive DNA binding of Cbf1 and Pho4 is largely determined by their intrinsic preferences for DNA, it is possible that interactions with cofactors will further refine the genomic targeting of Cbf1 and Pho4 in the cell. In other systems, cofactors may have even larger effects on paralogous TF competition, especially in cases where cofactors enable latent specificities of TF paralogs (Slattery et al. 2011; Crocker et al. 2015). To account for influences from cofactors, our protocols can be modified to incubate the competing paralogs with all their contributing cofactors, as long as the cofactors can be expressed and purified as recombinant proteins. In addition, the recent development of nextPBM (Mohaghegh et al. 2019) makes it feasible to perform protein-DNA binding assays in the endogenous nuclear environment, also facilitating the study of cellular contributors to competition. We expect such extensions to be critical for deciphering the competitive binding of TF paralogs in eukaryotes.

## Methods

### Protein expression and purification

Full-length *S. cerevisiae CBF1* and *PHO4* genes, cloned into the Gateway pDEST15 expression vectors (Invitrogen) were obtained from Gordân et al. (2013). Using the LR Clonase reaction (Thermo Fisher Scientific Gateway cloning system), the *CBF1* gene was transferred into the pDEST17 vector, for expression of N-terminal His-tagged Cbf1 protein. For *PHO4*, the pDEST15 vector was used to express N-terminal GST-tagged Pho4 protein. The Cbf1 protein was also expressed with a GST tag using the pDEST15 vector, for use in control experiments of GST-Cbf1 versus His-Cbf1 competition (see section "Competition PBM assay" below for details). Bacterial cells (BL21-CodonPlus [DE3]-RIL; Agilent 230245) were grown in LB culture to an $OD_{600}$ of 0.8–1.2 and induced with 1 mM IPTG overnight at 20°C for protein expression. Next, the cells were pelleted and lysed with lysozyme (Millipore 71110). The proteins were purified from the soluble portion of the lysate using His resin or GST resin (GE Healthcare FF affinity column) according to the manufacturer's instructions.

### Design of DNA library for PBM assays

Our DNA library consists of: (1) yeast genomic regions containing putative DNA binding sites for Cbf1 and Pho4; (2) negative control sequences not bound specifically by either Cbf1 or Pho4; and (3) DNA sequences that were used in previous MITOMI experiments to measure equilibrium dissociation constants and/or binding energies for Cbf1 and Pho4 (Maerkl and Quake 2007). Six replicate DNA spots were used for each probe, randomly distributed across the array surface. Microarrays using our custom DNA library were synthesized de novo by Agilent in 8 × 60 k format (8 chambers, 60,000 DNA spots per chamber).

#### *Probes containing genomic sites*

We analyzed publicly available Cbf1 and Pho4 ChIP-seq data to identify all sites in the yeast genome where Cbf1 and Pho4 may bind and compete. Cbf1 ChIP-seq data in wild-type strain EY 57 (K699 *MAT***a** *ade2-1 trp1-1 can1-100 leu2-3,112 his3-11,15 ura3*) under high phosphate condition and Pho4 ChIP-seq data in the same strain under no phosphate condition were downloaded from the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE29506 (Zhou and O'Shea 2011). Comprehensive, unbiased, 8-mer E-score data from universal PBM assays (Berger and Bulyk 2009; Gordân et al. 2011) were used to scan the Cbf1 and Pho4 ChIP-seq peaks and identify putative binding sites, called at a lenient E-score cutoff of 0.33 for two or more consecutive 8-mers, similarly to our previous work (Gordân et al. 2013; Shen et al. 2018). The cutoff was chosen to be lenient so that we include as many putative genomic sites as possible in our DNA library. Using as guidance the results of Berger and Bulyk (2009), who reported that a false discovery rate of 0.01 typically corresponds to E-scores of 0.32–0.36, we started our library design with an E-score cutoff of 0.36, and then we relaxed the cutoff to include more DNA probes until we reached the capacity of the microarray, which occurred at a cutoff of 0.33. Next, the selected genomic DNA sequences were aligned using PWM models in order to center the putative Cbf1/Pho4 binding sites. A total of 5424 genomic regions were selected using this procedure. In addition, we identified all CACGTG sites in the yeast genome and found 287 genomic sites that were not included in the ChIP-seq peaks, possibly due to occlusion by nucleosomes or other influences from the cellular environment. We manually added these additional CACGTG sites to our DNA library. In addition, we designed 300 negative control DNA probes that served as a reference for nonspecific Cbf1/Pho4 binding signals, similarly to our previous work (Gordân et al. 2013; Shen et al. 2018). The negative control sequences were 36 bp long and were selected randomly from accessible genomic regions in *S. cerevisiae* (according to DNase-seq data from GEO data set GSM1705337), excluding the ChIP-seq peaks of Cbf1 and Pho4 in order to exclude regions bound in vivo by our TFs of interest. The negative control probes also satisfied the criterion that all 8-mers in the 36-bp regions had E-scores < 0.33, to ensure that these probes were unlikely to contain specific sites for Cbf1 or Pho4. The final DNA probes were 60 bp long, consisting of 36-bp genomic regions followed by a constant 24-bp sequence (5′-GTCTTGATTCGCTTGACGCTGCTG-3′) that was complementary to the DNA primer. The primer was used to double-strand the DNA on the microarray by primer extension, as previously described (Berger and Bulyk 2009).

In analyzing the PBM data, we defined "specific" Cbf1 binding sites as sites with a Cbf1 binding level (i.e., fluorescence intensity signal) larger than the 99th percentile of negative control probes, based on experiments where Cbf1 was tested individually (i.e., not in competition with Pho4). Similarly, "specific" Pho4 binding sites were defined as sites with a Pho4 binding level larger than the 99th percentile of negative controls, based on experiments where Pho4 was tested individually. These sites were used in the analyses shown in Figure 2B,C.

### *MITOMI-based probes*

From the DNA libraries used in Maerkl and Quake (2007), we selected the NNNNGTG, CACNNN, and GTGNNN libraries for our PBM design. To the DNA sequences in the original MITOMI libraries, which were 14 bp long, we added random 11-bp flanks on each side to obtain 36-bp DNA sequences centered at the Cbf1/Pho4 binding sites, similar to the genomic sequences described above. The random flanks were generated using a uniform probability distribution over the four nucleotides. Given that the added 11-bp flanks could influence the binding specificity/affinity of Cbf1 and Pho4, we designed 10 different random flanks. When processing the data, we used the median measurements over the 10 flanks.

## "Competition PBM" assay

"Competition PBM" experiments were carried out following the standard PBM protocol (Berger et al. 2006; Berger and Bulyk 2009) but incubating the competing TFs (here, Cbf1 and Pho4) simultaneously with the double-stranded DNA molecules synthesized on the array. Briefly, after performing the primer extension step (Berger and Bulyk 2009) to double-strand the DNA probes on the microarray, each chamber on the array was blocked with 2% milk for 1 h. After mild washing, the array was incubated for 1 h with protein binding mixtures, at the Cbf1 and Pho4 final concentrations shown in Supplemental Table S2B. Alexa Fluor 488-conjugated anti-GST antibody (Invitrogen A-11131) and Alexa Fluor 647-conjugated anti-His antibody (Qiagen 35370) were used for Pho4 and Cbf1, respectively. After mild washing (Berger and Bulyk 2009), the array was scanned using a GenePix 4400A scanner (Molecular Devices) at 2.5-micron resolution. Standard analysis scripts (Berger et al. 2006; Berger and Bulyk 2009) were used to extract and normalize the florescence intensity data, and then median values over replicate DNA spots were computed for the unique DNA sequences. Previous studies have shown that this design strategy results in highly reproducible PBM data, with $R^2 = 0.92–0.98$ between duplicate experiments (Shen et al. 2018; Penvose et al. 2019; Afek et al. 2020). We use the term "binding level" to refer to the fluorescence intensity signal observed at a DNA spot, which results from the fluorophore-tagged antibody bound to the protein bound to the DNA at that spot.

To ensure that the choice of epitope tags did not influence the intrinsic binding of TFs or their competition, we performed a control experiment with GST-Cbf1 and His-Cbf1 in competition. We used a DNA oligonucleotide array in 4 × 44 k format (Agilent AMADID 029393) that contains putative Cbf1 binding sites (Gordân et al. 2013). Similar to the competition assay described above, we incubated the double-stranded array with protein binding mixture, with GST-Cbf1 and His-Cbf1 at different concentrations between 0.2 μM and 0.8 μM (Supplemental Table S2F). Alexa Fluor 488-conjugated anti-GST antibody (Invitrogen A-11131) and Alexa Fluor 488-conjugated anti-His antibody (Qiagen 35310) were used to target GST-Cbf1 and His-Cbf1, respectively, and competition binding data were collected as described above. As expected, when GST-Cbf1 and His-Cbf1 were tested by themselves, we saw an excellent agreement between their binding intensity levels (Supplemental Fig. S3A). In addition, for each competition scenario tested, we found that GST-Cbf1 and His-Cbf1 competed with each other in a linear pattern consistent with changes in their concentration (Supplemental Fig. S3B). These data indicate that no bias was introduced due to the epitope tags and emphasize the high reproducibility of custom PBM experiments ($R^2 = 0.95–0.98$) (Supplemental Fig. S3A,B), even in the case of proteins tagged with different epitopes.

Given the high reproducibility of custom PBM data, as described above, the high correlations between custom PBM data and independently measured binding energies and $K_d$ values ($R^2 = 0.83–0.88$ [Fig. 2E]; $R^2 = 0.84–0.99$ [Shen et al. 2018; Afek et al. 2020]), as well as the inclusion of replicate spots within our DNA libraries, in this study we did not perform duplicate PBM experiments.

## Comparing competition PBM data across experiments

To directly compare the fluorescence signal intensities between different competition PBM experiments, that is, between different chambers of the PBM arrays, we process the data as follows. As listed in Supplemental Table S2B, for competition PBMs we varied the concentration of the competitor TF over a wide range (0.05–8 μM), while keeping the concentration of the main TF constant (2 μM). The rationale for this design was to enable direct comparisons between different chambers with different concentrations of competitors when the concentration of the main TF remained constant. In practice, due to pipetting noise introduced during the binding steps and/or the dilution of the protein samples, it is possible that different chambers on the same microarray have slightly different concentrations of the main TF, which is what we observed for Pho4. To determine whether the concentration of Pho4 was the same across chambers, we used a subset of DNA probes that were nonspecifically bound by Cbf1 but bound with low- to medium-affinities by Pho4; these probes, which we call "Pho4-specific," were chosen randomly among those with Cbf1 signal in the negative control range (Supplemental Fig. S3C). If the Pho4 concentrations were the same (i.e., 2 μM) in all four chambers, then we would expect the Pho4-specific probes to have similar binding levels across these chambers. However, in practice, we observed deviations from this expected trend (Supplemental Fig. S3D), consistent with small differences in the effective Pho4 concentrations. To alleviate this problem, we used the standard binding isotherms to derive the correlation between Pho4 binding levels at different Pho4 concentrations. Based on this correlation, we estimated the effective concentrations of Pho4 to be 2 μM, 0.98 μM, 2.17 μM, and 2.18 μM, respectively, in the array chambers with 0.05 μM, 0.4 μM, 2 μM, and 8 μM Cbf1. Next, we adjusted the Pho4 binding levels in all chambers based on a 2 μM concentration, while keeping the concentration of Cbf1 unchanged (see Supplemental Methods for details). After this correction, the Pho4 binding signals for "Pho4-specific" probes showed excellent agreement between different chambers (Supplemental Fig. S3D), allowing us to use the competition data to directly assess the effects of Cbf1 competition on the probes bound specifically by both TFs. We applied a similar procedure to evaluate the concentration of Cbf1 across chambers and found that the Cbf1 effective concentration was 2 μM, after rounding, in all chambers. Thus, we did not apply any correction to the Cbf1 binding signals.

## Resilience to TF competition

The resilience of Cbf1 to in vitro competition from Pho4 was defined as the log fold-change of Cbf1 binding levels (fluorescence intensity, FI) when the concentration of the competitor increased, for example,

$$log\left(\frac{Cbf1\ FI \text{ at } 2\,\text{uM}\ Cbf1 + \ 8\,\text{uM}\ Pho4}{Cbf1\ FI \text{ at } 2\,\text{uM}\ Cbf1 + \ 0.05\,\text{uM}\ Pho4}\right).$$

Similarly, Pho4's in vitro resilience to Cbf1 binding was defined as

$$log\left(\frac{Pho4\ FI\ \text{at}\ 2\,\text{uM}\ Pho4\ +\ 8\,\text{uM}\ Cbf1}{Pho4\ FI\ \text{at}\ 2\,\text{uM}\ Pho4\ +\ 0.05\,\text{uM}\ Cbf1}\right).$$

In vivo, the resilience of Cbf1 was defined as

$$log\left(\frac{Cbf1\ ChIP-seq\ pileup\ in\ no\ Pi}{Cbf1\ ChIP-seq\ pileup\ in\ high\ Pi}\right),$$

whereas the resilience of Pho4 was defined as

$$log\left(\frac{Pho4\ ChIP-seq\ pileup\ in\ \Delta pho80}{Pho4\ ChIP-seq\ pileup\ in\ \Delta pho80\Delta cbf1}\right).$$

To compare in vitro versus in vivo resilience scores, we sorted the medium- and high-affinity binding sites of Cbf1/Pho4 (defined as sites with fluorescence intensity in the upper half of the intensity range, where ratios of intensity signals are not significantly affected by noise) in decreasing order of their in vitro resilience scores, and we compared the distributions of in vivo resilience scores between sets of sites with high versus low in vitro resilience (Fig. 3D,G; Supplemental Table S3A,B). Comparisons between sets of sites were performed using a one-sided *t*-test.

## PWM motif derivation

To derive motifs for the main TF at different concentrations of the competitor TF, we ranked DNA sequences by their binding levels (i.e., by the fluorescence intensities of the main TF) for each competition scenario. Next, we selected the top 200 DNA sequences, weighted them by their TF binding levels, and used the weighted counts to construct position frequency matrices. Motif logos were generated from the frequency matrices using enoLOGOS (Workman et al. 2005).

## DNA shape analyses

Similarly to our previous work (Gordân et al. 2013; Zhou et al. 2015; Shen et al. 2018), we analyzed the extent to which the shape of genomic regions flanking the Cbf1/Pho4 core E-box binding site contributes to TF-DNA binding specificity. We used DNAshape (Zhou et al. 2013) to predict the minor groove width, roll, propeller twist, and helix twist for all DNA sequences in our PBM library. Next, we asked whether these DNA shape features significantly improve the accuracy of Cbf1 and Pho4 DNA-binding specificity models when added to mononucleotide (1-mer) features. Least squares estimation, as implemented in the R "stats" package (R Core Team 2018), was used to train linear regression models of DNA-binding specificity from the PBM data. The models were trained and tested on sequences on different lengths, from 10 bp (which includes only the E-box site and the immediate 2-bp flanks) to 36 bp (which includes the full genomic context tested in our assays). For each length, we performed fivefold cross-validation to evaluate the model accuracy, assessed as the squared Pearson's correlation ($R^2$) between measured and predicted binding levels. We repeated each cross-validation test 25 times using 25 random splittings of the data. A Mann–Whitney *U* test was applied to the $R^2$ values over the 25 runs in order to compare the accuracy of different models.

## Analysis of ChIP-seq and nucleosome mapping data

ChIP-seq data for transcription factors Cbf1 and Pho4 were retrieved as raw reads from the GEO database, entry GSE29506 (Zhou and O'Shea 2011). We used the ChIP-seq data for Cbf1 in yeast strain EY57 (K699 *MAT***a** *ade2-1 trp1-1 can1-100 leu2-3,112 his3-11,15 ura3* [Zhou and O'Shea 2011]) under two physiological conditions: no Pi (sample "Cbf1_ChIP_NoPi") and high Pi (sample "Cbf1_ChIP_HighPi"), and ChIP-seq data for Pho4 in two mutant yeast strains: a strain with constitutively expressed *PHO4* ("Pho4_ChIP_dPHO80") and a strain with constitutively expressed *PHO4* and knocked-out *CBF1* ("Pho4_ChIP_dPHO80dCBF1"). Raw sequencing files were aligned to the yeast genome (sacCer2) using BWA (parameters -q 5 -l 32 -k 2 -t 4) (Li and Durbin 2009). The read coverage across the entire genome was then computed using BEDTools (Quinlan and Hall 2010). For comparisons with PBM data, the TF genomic occupancies at each genomic site of interest were calculated as the ChIP-seq read pileups within the central 6-bp window, where the centers of the binding sites were located. Nucleosome data were downloaded from the supplemental files of Zhou and O'Shea (2011) as processed nucleosome occupancy probabilities.

## Gene expression analysis

Gene expression data from Zhou and O'Shea (2011) were downloaded as normalized log ratios from GEO, accession number GSE23580, for samples: "Wild type no vs high Pi conditions", "*pho80Δ* vs *pho80Δpho4Δ* in high Pi conditions", and "*pho80Δcbf1Δ* vs *pho80Δpho4Δcbf1Δ* in high Pi". As described in Zhou and O'Shea (2011), comparing wild type no Pi versus high Pi conditions identifies genes induced in response to inorganic phosphate limitation. Here, we refer to this set of genes as the genes induced by Pho4 under physiological conditions. Comparing *pho80Δ* versus *pho80Δpho4Δ* in high Pi conditions identifies genes induced by Pho4 when the *PHO* signaling pathway is fully activated. We used this data to further validate the gene regulatory role of Pho4 in the presence of Cbf1. Comparing *pho80Δcbf1Δ* versus *pho80Δpho4Δcbf1Δ* in high Pi conditions identifies the influence of Cbf1 on the gene activation role of Pho4. We used these data to illustrate how Cbf1 helps specify the true target genes of Pho4 by competing for DNA binding sites at other genes with putative binding sites. Zhou and O'Shea (2011) reported the Pho4-regulated genes in the wild-type strain and the *pho80Δcbf1Δ* strain. We refer to the Pho4-regulated genes in physiological condition as "Pho4 targets in physiological condition" (Fig. 4, blue). Genes that are induced in the *pho80Δcbf1Δ* strain but not in wild type are referred to as "Pho4 targets only in *cbf1Δ*" (Fig. 4, red). We identified the Pho4 binding site(s) potentially responsible for the regulation of each target gene using the following criteria: (1) the site was located within 1000 bp upstream of the gene TSS; and (2) the in vitro Pho4-DNA binding level at the site was higher than the binding level at any of the negative control probes in the PBM experiment. Based on these criteria, there were nine (out of 37) genes with more than one Pho4 binding site in their upstream region, which made it difficult to explicitly associate individual Pho4 binding events with gene regulation. Because the focus of our study was not the interplay between multiple binding sites within the regulatory regions, for further analysis we only considered the genes with a single Pho4 site identified. The list of genes and the corresponding Pho4 regulatory sites can be found in Supplemental Table S4A.

## Estimation of equilibrium dissociation constants ($K_d$) from PBM data

In a PBM experiment, the measured fluorescence intensities linearly reflect the amount of TF protein bound at each DNA spot on the microarray slide, that is, the amount of TF-DNA complexes. Thus, we can write: $[TF \cdot DNA]^i = aF^i$, where $[TF \cdot DNA]^i$ is the concentration of TF-DNA complex at DNA spot *i*, $F^i$ is the fluorescence signal measured at spot *i*, and *a* is a constant (Siggers et al. 2011). We can then express the equilibrium dissociation constant as

$$\begin{aligned} K_d^i &= \frac{[DNA]_{total} - [DNA]^i_{bound}}{[TF \cdot DNA]^i}[TF]_{unbound} \\ &= \frac{F^{total} - F^i}{F^i}[TF]_{unbound} = \frac{F^{total} - F^i}{F^i}e^{\mu}, \qquad (2) \end{aligned}$$

where $F^{total}$ is the fluorescence intensity at saturated DNA spots (i.e., spots where all DNA molecules are bound by the TF), and $\mu = \ln([TF]_{total})$, as used in Zhao et al. (2009). Next, we can express the observed fluorescence signal $F^i$ as

$$F^i = \frac{F^{total}}{1 + e^{-\mu}K_d^i}\ . \qquad (3)$$

To infer equilibrium dissociation constants in high throughput, we performed PBM experiments for each TF of interest at four different total concentrations: 0.05 μM, 0.2 μM, 1 μM, and 8 μM for Cbf1, and 0.1 μM, 0.4 μM, 2 μM, and 8 μM for Pho4. In Equation (3), $F^i$ is observed, leaving $\mu_k$ ($k = 1, 2, 3, 4$), $F^{total}$, and $K_d^i$ undetermined. We estimated these parameters iteratively. In the initial round, $\mu_k\_r_0$ was set to $\mu_k = \ln([TF]_{total})$, and $F^{total}\_r_0$ was set to the highest fluorescence signal we observed on the entire slide. Next, the $K_d^i$ parameter in round 0 ($K_d^i\_r_0$) at each DNA spot *i* was estimated from the $F^i$ measurements at the four different concentrations using nonlinear least square estimation. Then, with $F^i$ observations at thousands of DNA spots and $K_d^i\_r_n$ available in our data (*n* is the iteration number), $F^{total}\_r_{n+1}$ and $\mu_k\_r_{n+1}$ ($k = 1, 2, 3, 4$) were estimated with nonlinear least square regression. $F^{total}\_r_{n+1}$ and $\mu_k\_r_{n+1}$ ($k = 1, 2, 3, 4$) were used again to estimate $K_d^i\_r_{n+1}$ iteratively until all the parameters converged. For both Cbf1 and Pho4, the parameters converged fast, within, at most, 30 rounds.

All the regression analyses were done using minpack in R. All loss functions were in natural log scale, as in log space the PBM measurement error does not correlate with the binding signal (Zhao et al. 2017). We compared our binding data against MITOMI data (Maerkl and Quake 2007) in terms of binding energies ($\Delta\Delta G$). Binding energies were computed from $K_d$ values as $\Delta\Delta G = RT \cdot ln\left(\frac{K_d}{K_{d,ref}}\right)$, where *T* is the temperature (298 K) and *R* is the gas constant. We chose the DNA probe with highest affinity (i.e., smallest $K_d$) as $K_{d,ref}$. When comparing our estimated $\Delta\Delta G$ values against MITOMI data (Maerkl and Quake 2007), the range of $\Delta\Delta G$ values was slightly different. For example, in Figure 2E, Cbf1 $\Delta\Delta G$ ranges from 0 to 6.5 kcal/mol in PBM, and 0 to 4.2 kcal/mol in MITOMI. Pho4 $\Delta\Delta G$ ranges from 0 to 0.94 kcal/mol in PBM, and 0 to 1.3 kcal/mol in MITOMI. These differences were not unexpected given that we are using different techniques and different protein samples. We found the best agreement for the MITOMI probe group CACNNN for Cbf1 ($R^2 = 0.88$) and GTGNNN for Pho4 ($R^2 = 0.83$), as shown in Figure 2E. All PBM-derived $K_d$ and $\Delta\Delta G$ values were computed over 36-bp sequences. In the comparisons against MITOMI data, for each 14-bp sequence tested by MITOMI, we used 10 random flanks in our DNA library and we computed the median $K_d$ and $\Delta\Delta G$ value over the 10 flanks.

### Biophysical modeling of competitive DNA binding

Using $K_d$ data for the main TF and the competitor TF (which we estimated in this study as described above), we can directly derive the probability that a DNA site *i* is bound by the main TF

$$P_i = \frac{[TF^{main}]/K_{d,i}^{main}}{1 + [TF^{main}]/K_{d,i}^{main} + [TF^{competitor}]/K_{d,i}^{competitor}}, \qquad (4)$$

where $[TF^{main}]$ and $[TF^{competitor}]$ are the concentrations of free main TF and free competitor TF, respectively, and $K_{d,i}^{main}$ and $K_{d,i}^{competitor}$ are their equilibrium dissociation constants at site *i*. Determining the concentrations of free proteins in any systems is not trivial. Thus, $[TF^{main}]$ and $[TF^{competitor}]$ parameters were unknown. Here, we first estimated these parameters from our data using nonlinear least square regression, similar to the $K_d$ estimation. Next, we used all the estimated parameters in Equation (4) in order to calculate the probabilities of binding under four different competition scenarios (Supplemental Table S2B). All loss functions were in natural log scale, because in log space the PBM measurement error does not correlate with the signal.

### Evolutionary conservation analyses

We searched for bHLH orthologs across fungi and animals. 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. Proteomes were downloaded from NCBI Genome (https://www.ncbi.nlm.nih.gov/genome/) or JGI MycoCosm (https://mycocosm.jgi.doe.gov/mycocosm/home), as indicated in Supplemental Table S1. We identified orthologs in each genome using HMMER profiles downloaded from eggNOG5, a database of annotated orthologs created via phylogenetic and functional analysis of thousands of genomes (Huerta-Cepas et al. 2019).

*Saccharomyces cerevisiae* Cbf1 is part of a large orthologous group (KOG1318) that includes Class B animal bHLHs, such as the MITF and USF subfamilies. Starting from *Saccharomyces cerevisiae* Cbf1 and *Mus musculus* MITF, we used HMMER to identify fungal and animal orthologs using the KOG1318 profile. During our analysis, it became clear that the Cbf1-like progenitor duplicated to create Cbf1 and Rtg3 subfamilies in fungi. Thus, we downloaded the HMMER profile for the Rtg3 subfamily (ENOG503P20B) and extended our analysis to identify both Rtg3 and Cbf1 orthologs; see Supplemental Table S1. Last, we aligned all proteins to the canonical bHLH domain (PF00010) and identified unique sequences features for each subfamily (Supplemental Fig. S1). Overall, our data suggest that fungal Cbf1 and animal MITF are likely descended from the same Cbf1-like progenitor that was present in the common ancestor of both fungi and animals.

*Saccharomyces cerevisiae* Pho4 is part of a yeast-specific ortholog group (ENOG502S1Z7), whereas the Pho4 orthologs in *Neurospora crassa* (Nuc-1) and *Aspergillus nidulans* (PalcA) are from a broader, fungal-specific ortholog group (ENOG502S7T4). We used the broad fungal Pho4 profile to identify both new and known orthologs (Gomes-Vieira et al. 2018). Fungal Pho4 has a highly conserved and unique H**AEQK motif in its basic domain; see alignment in Supplemental Table S1. We could not detect Pho4 in the animal outgroups, but fungal Pho4 had weak hits to animal orthologs in the MAX (KOG2483), Max-like protein X (MLX, KOG1319), and MLX-interacting protein (MLXIP, KOG3582) subfamilies. Upon closer inspection, only the MLX and MLXIP subfamilies have the H**AEQK motif. This suggests that fungal Pho4 and animal MLX/MLXIP are likely descendants from the same Pho4-like progenitor that was present in the common ancestor of both fungi and animals (Supplemental Fig. S1).

### Analysis of paralogous TFs' expression profiles

The list of TF families in the yeast *S. cerevisiae* was acquired from the Cis-BP database (Weirauch et al. 2014). The protein abundance of all yeast TFs were obtained from Ho et al. (2018), in units of molecules per cell. Next, we counted the number of families that have more than one TF expressed in the cell with abundance higher than the Xth percentile among all TFs, where X = 0, 0.05… 0.95, 1. The results are presented in Supplemental Figure S5E and Supplemental Table S4B.

The expression profiles of TF families in 37 human tissues were obtained from Lambert et al. (2018), as normalized transcripts per million (TPM). For each TF family, we counted the number of TFs that have an expression level higher than the Xth percentile among all TFs, where X = 0, 0.01, 0.02…0.99, 1. Next, we counted the number of families that have more than one TF expressed above the threshold in at least one tissue type (Supplemental Table S4B). The results of these analyses are shown in Supplemental Figure S5E, where we plotted the counts of TF families against the percentiles that we used as cutoffs. Next, we refined our analysis by focusing only on paralogs with highly similar DNA-binding motifs, as represented by PWM models and compiled by Lambert et al. (2018). Briefly, Lambert et al. used hierarchical clustering to group motifs into 585 clusters that covered 1211 TFs with known DNA-binding motifs. Out of the 585 clusters, 421 had only one TF protein, most commonly (in 82% cases) a zinc finger protein. Of the remaining 164 clusters, six contained proteins without available gene expression data, leaving 158 clusters for analysis; each of the 158 clusters contained two or more TF paralogs with highly similar DNA binding motifs, covering a total of 871 TFs from 36 families. Of these clusters of interest, 154 (~97%) contained at least two paralogs coexpressed in at least one cell type, at a percentile cutoff X = 75%. Results for all cutoffs between 0.01 and 1 are available in Supplemental Table 4D and Supplemental Figure S5E.

## Data access

All raw and processed PBM data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE163512.

## Competing interest statement

The authors declare no competing interests.

## Acknowledgments

The authors thank Dr. David MacAlpine and Dr. Lingchong You for critical reading of the manuscript. The work presented in this manuscript was supported by National Science Foundation grants MCB-1715589 (R.G.) and IOS-1755248 (R.G. and N.E.B.), and National Institutes of Health grant R01-GM117106 (R.G.).

## References

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. doi:10.1038/s41586-020-2843-2

Amoutzias GD, Veron AS, Weiner J III, Robinson-Rechavi M, Bornberg-Bauer E, Oliver SG, Robertson DL. 2007. One billion years of bZIP transcription factor evolution: conservation and change in dimerization and DNA-binding site specificity. *Mol Biol Evol* **24:** 827–835. doi:10.1093/molbev/msl211

Aow JS, Xue X, Run JQ, Lim GF, Goh WS, Clarke ND. 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 Res* **41:** 4877–4887. doi:10.1093/nar/gkt210

Atchley WR, Fitch WM. 1997. A natural classification of the basic helix-loop-helix class of transcription factors. *Proc Natl Acad Sci* **94:** 5172–5176. doi:10.1073/pnas.94.10.5172

Badis G, Chan ET, van Bakel H, Pena-Castillo L, Tillo D, Tsui K, Carlson CD, Gossett AJ, Hasinoff MJ, Warren CL, et al. 2008. A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol Cell* **32:** 878–887. doi:10.1016/j.molcel.2008.11.020

Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X, et al. 2009. Diversity and complexity in DNA recognition by transcription factors. *Science* **324:** 1720–1723. doi:10.1126/science.1162327

Banerjee-Basu S, Baxevanis AD. 2001. Molecular evolution of the homeodomain family of transcription factors. *Nucleic Acids Res* **29:** 3258–3269. doi:10.1093/nar/29.15.3258

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. doi:10.1038/nprot.2008.195

Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW III, Bulyk ML. 2006. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* **24:** 1429–1435. doi:10.1038/nbt1246

Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Pena-Castillo L, Alleyne TM, Mnaimneh S, Botvinnik OB, Chan ET, et al. 2008. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133:** 1266–1276. doi:10.1016/j.cell.2008.05.024

Chen K, Rajewsky N. 2007. The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet* **8:** 93–103. doi:10.1038/nrg1990

Crocker J, Abe N, Rinaldi L, McGregor AP, Frankel N, Wang S, Alsawadi A, Valenti P, Plaza S, Payre F, et al. 2015. Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell* **160:** 191–203. doi:10.1016/j.cell.2014.11.041

Djordjevic M, Sengupta AM, Shraiman BI. 2003. A biophysical approach to transcription factor binding site discovery. *Genome Res* **13:** 2381–2390. doi:10.1101/gr.1271603

Ferraris L, Stewart AP, Kang J, DeSimone AM, Gemberling M, Tantin D, Fairbrother WG. 2011. Combinatorial binding of transcription factors in the pluripotency control regions of the genome. *Genome Res* **21:** 1055–1064. doi:10.1101/gr.115824.110

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. doi:10.1002/j.1460-2075.1992.tb05503.x

Gerland U, Moroz JD, Hwa T. 2002. Physical constraints and functional characteristics of transcription factor–DNA interaction. *Proc Natl Acad Sci* **99:** 12015–12020. doi:10.1073/pnas.192693599

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. doi:10.1016/j.fgb.2018.04.004

Gordân 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. doi:10.1186/gb-2011-12-12-r125

Gordân 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. doi:10.1016/j.celrep.2013.03.014

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, et al. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431:** 99–104. doi:10.1038/nature02800

He Q, Johnston J, Zeitlinger J. 2015. ChIP-nexus enables improved detection of *in vivo* transcription factor binding footprints. *Nat Biotechnol* **33:** 395–401. doi:10.1038/nbt.3121

Henikoff S, Greene EA, Pietrokovski S, Bork P, Attwood TK, Hood L. 1997. Gene families: the taxonomy of protein paralogs and chimeras. *Science* **278:** 609–614. doi:10.1126/science.278.5338.609

Ho B, Baryshnikova A, Brown GW. 2018. Unification of protein abundance datasets yields a quantitative *Saccharomyces cerevisiae* proteome. *Cell Syst* **6:** 192–205.e3. doi:10.1016/j.cels.2017.12.004

Huerta-Cepas J, Szklarczyk D, Heller D, Hernandez-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, et al. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* **47:** D309–D314. doi:10.1093/nar/gky1085

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316:** 1497–1502. doi:10.1126/science.1141319

Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, Enge M, Kivioja T, Morgunova E, Taipale J. 2015. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527:** 384–388. doi:10.1038/nature15518

Jones S. 2004. An overview of the basic helix-loop-helix proteins. *Genome Biol* **5:** 226. doi:10.1186/gb-2004-5-6-226

Komeili A, O'Shea EK. 1999. Roles of phosphorylation sites in regulating activity of the transcription factor Pho4. *Science* **284:** 977–980. doi:10.1126/science.284.5416.977

Kuras L, Barbey R, Thomas D. 1997. Assembly of a bZIP–bHLH transcription activation complex: formation of the yeast Cbf1–Met4–Met28 complex is regulated through Met28 stimulation of Cbf1 DNA binding. *EMBO J* **16:** 2441–2451. doi:10.1093/emboj/16.9.2441

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. doi:10.1016/j.cell.2018.09.045

Laudet V, Hänni C, Stéhelin D, Duterque-Coquillaud M. 1999. Molecular phylogeny of the ETS gene family. *Oncogene* **18:** 1351–1359. doi:10.1038/sj.onc.1202444

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* **115:** E3702–E3711. doi:10.1073/pnas.1715888115

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25:** 1754–1760. doi:10.1093/bioinformatics/btp324

Lin H, Ouyang S, Egan A, Nobuta K, Haas BJ, Zhu W, Gu X, Silva JC, Meyers BC, Buell CR. 2008. Characterization of paralogous protein families in rice. *BMC Plant Biol* **8:** 18. doi:10.1186/1471-2229-8-18

Maerkl SJ, Quake SR. 2007. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315:** 233–237. doi:10.1126/science.1131007

Miyamoto T, Kaneko A, Kakizawa T, Yajima H, Kamijo K, Sekine R, Hiramatsu K, Nishii Y, Hashimoto T, Hashizume K. 1997. Inhibition of peroxisome proliferator signaling pathways by thyroid hormone receptor. Competitive binding to the response element. *J Biol Chem* **272:** 7752–7758. doi:10.1074/jbc.272.12.7752

Mohaghegh N, Bray D, Keenan J, Penvose A, Andrilenas KK, Ramlall V, Siggers T. 2019. NextPBM: a platform to study cell-specific transcription factor binding and cooperativity. *Nucleic Acids Res* **47:** e31. doi:10.1093/nar/gkz020

Morgunova E, Taipale J. 2017. Structural perspective of cooperative transcription factor binding. *Curr Opin Struct Biol* **47:** 1–8. doi:10.1016/j.sbi.2017.03.006

Murre C. 2019. Helix-loop-helix proteins and the advent of cellular diversity: 30 years of discovery. *Genes Dev* **33:** 6–25. doi:10.1101/gad.320663.118

Nakagawa S, Gisselbrecht SS, Rogers JM, Hartl DL, Bulyk ML. 2013. DNA-binding specificity changes in the evolution of forkhead transcription factors. *Proc Natl Acad Sci* **110:** 12349–12354. doi:10.1073/pnas.1310430110

Nitta KR, Jolma A, Yin Y, Morgunova E, Kivioja T, Akhtar J, Hens K, Toivonen J, Deplancke B, Furlong EE, et al. 2015. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife* **4:** e04837. doi:10.7554/eLife.04837

Noro B, Lelli K, Sun L, Mann RS. 2011. Competition for cofactor-dependent DNA binding underlies Hox phenotypic suppression. *Genes Dev* **25:** 2327–2332. doi:10.1101/gad.175539.111

Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, Wolfe SA. 2008. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell* **133:** 1277–1289. doi:10.1016/j.cell.2008.05.023

Ogawa N, DeRisi J, Brown PO. 2000. New components of a system for phosphate accumulation and polyphosphate metabolism in *Saccharomyces cerevisiae* revealed by genomic expression analysis. *Mol Biol Cell* **11:** 4309–4321. doi:10.1091/mbc.11.12.4309

O'Neill EM, Kaffman A, Jolly ER, O'Shea EK. 1996. Regulation of PHO4 nuclear localization by the PHO80-PHO85 cyclin-CDK complex. *Science* **271:** 209–212. doi:10.1126/science.271.5246.209

Penvose A, Keenan JL, Bray D, Ramlall V, Siggers T. 2019. Comprehensive study of nuclear receptor DNA binding provides a revised framework for understanding receptor specificity. *Nat Commun* **10:** 2514. doi:10.1038/s41467-019-10264-3

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033

R Core Team. 2018. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Rhee HS, Pugh BF. 2012. ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Curr Protoc Mol Biol* **Chapter 21:** Unit 21.24. doi:10.1002/0471142727.mb2124s100

Robinson KA, Lopes JM. 2000. SURVEY AND SUMMARY: *Saccharomyces cerevisiae* basic helix-loop-helix proteins regulate diverse biological processes. *Nucleic Acids Res* **28:** 1499–1505. doi:10.1093/nar/28.7.1499

Sailsbery JK, Dean RA. 2012. Accurate discrimination of bHLH domains in plants, animals, and fungi using biologically meaningful sites. *BMC Evol Biol* **12:** 154. doi:10.1186/1471-2148-12-154

Sailsbery JK, Atchley WR, Dean RA. 2012. Phylogenetic analysis and classification of the fungal bHLH domain. *Mol Biol Evol* **29:** 1301–1318. doi:10.1093/molbev/msr288

Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32:** D91–D94. doi:10.1093/nar/gkh012

Schneider KR, Smith RL, O'Shea EK. 1994. Phosphate-regulated inactivation of the kinase PHO80-PHO85 by the CDK inhibitor PHO81. *Science* **266:** 122–126. doi:10.1126/science.7939631

Shen N, Zhao J, Schipper JL, Zhang Y, Bepler T, Leehr D, Bradley J, Horton J, Lapp H, Gordân R. 2018. Divergence in DNA specificity among paralogous transcription factors contributes to their differential in vivo binding. *Cell Syst* **6:** 470–483.e8. doi:10.1016/j.cels.2018.02.009

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. doi:10.1038/msb.2011.89

Singh LN, Hannenhalli S. 2008. Functional diversification of paralogous transcription factors via divergence in DNA binding site motif and in expression. *PLoS One* **3:** e2345. doi:10.1371/journal.pone.0002345

Skene PJ, Henikoff JG, Henikoff S. 2018. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nat Protoc* **13:** 1006–1019. doi:10.1038/nprot.2018.015

Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, Dror I, Zhou T, Rohs R, Honig B, Bussemaker HJ, et al. 2011. Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* **147:** 1270–1282. doi:10.1016/j.cell.2011.10.053

Wei GH, Badis G, Berger MF, Kivioja T, Palin K, Enge M, Bonke M, Jolma A, Varjosalo M, Gehrke AR, et al. 2010. Genome-wide analysis of ETS-family DNA-binding *in vitro* and *in vivo*. *EMBO J* **29:** 2147–2160. doi:10.1038/emboj.2010.106

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al. 2014. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158:** 1431–1443. doi:10.1016/j.cell.2014.08.009

Workman CT, Yin Y, Corcoran DL, Ideker T, Stormo GD, Benos PV. 2005. enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res* **33:** W389–W392. doi:10.1093/nar/gki439

Wotton D, Ghysdael J, Wang S, Speck NA, Owen MJ. 1994. Cooperative binding of Ets-1 and core binding factor to DNA. *Mol Cell Biol* **14:** 840–850. doi:10.1128/MCB.14.1.840

Yokoyama KD, Pollock DD. 2012. SP transcription factor paralogs and DNA-binding sites coevolve and adaptively converge in mammals and birds. *Genome Biol Evol* **4:** 1102–1117. doi:10.1093/gbe/evs085

Zhao Y, Granas D, Stormo GD. 2009. Inferring binding energies from selected binding sites. *PLoS Comput Biol* **5:** e1000590. doi:10.1371/journal.pcbi.1000590

Zhao J, Li D, Seo J, Allen AS, Gordân R. 2017. Quantifying the impact of non-coding variants on transcription factor-DNA binding. *Res Comput Mol Biol* **10229:** 336–352. doi:10.1007/978-3-319-56970-3_21

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. doi:10.1016/j.molcel.2011.05.025

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–W62. doi:10.1093/nar/gkt437

Zhou T, Shen N, Yang L, Abe N, Horton J, Mann RS, Bussemaker HJ, Gordân R, Rohs R. 2015. Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc Natl Acad Sci* **112:** 4654–4659. doi:10.1073/pnas.1422023112

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. doi:10.1101/gr.090233.108

*Received December 15, 2020; accepted in revised form May 6, 2021.*