6-9-2023

Zinc cluster transcription factors frequently activate target genes using a non-canonical half-site binding mode

Pamela S Recio
Washington University School of Medicine in St. Louis
Nikhil J Mitra
Washington University School of Medicine in St. Louis
Christian A Shively
Washington University School of Medicine in St. Louis
David Song
Washington University School of Medicine in St. Louis
Grace Jaramillo
Washington University School of Medicine in St. Louis

See next page for additional authors

5006–5021 *Nucleic Acids Research, 2023, Vol. 51, No. 10* *Published online 1 May 2023*
https://doi.org/10.1093/nar/gkad320

Zinc cluster transcription factors frequently activate target genes using a non-canonical half-site binding mode

Pamela S. Recio [1,2], Nikhil J. Mitra[1,2], Christian A. Shively[1,2], David Song[1,2], Grace Jaramillo[1,2], Kristine Shady Lewis[1,2], Xuhua Chen[1,2] and Robi D. Mitra [1,2,3,*]

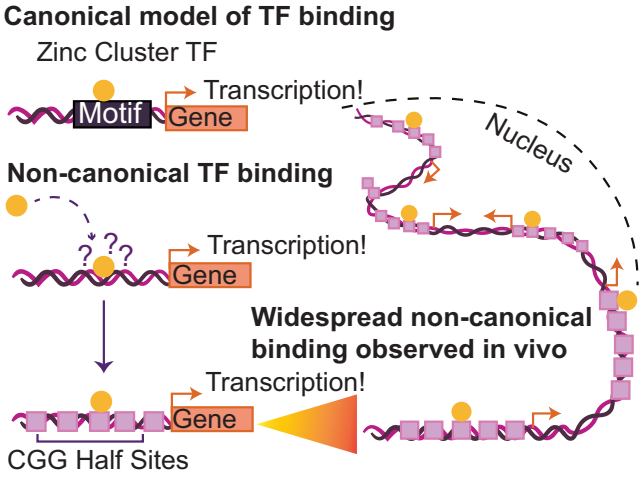
[1]Department of Genetics, Washington University School of Medicine in St. Louis, St. Louis, MO 63108, USA, [2]The Edison Family Center for Genome Sciences & Systems Biology, Washington University School of Medicine in St. Louis, St. Louis, MO 63108, USA and [3]McDonnell Genome Institute, Washington University School of Medicine in St. Louis, St. Louis, MO 63108, USA

Received December 09, 2022; Revised April 11, 2023; Editorial Decision April 13, 2023; Accepted April 14, 2023

ABSTRACT

Gene expression changes are orchestrated by transcription factors (TFs), which bind to DNA to regulate gene expression. It remains surprisingly difficult to predict basic features of the transcriptional process, including *in vivo* TF occupancy. Existing thermodynamic models of TF function are often not concordant with experimental measurements, suggesting undiscovered biology. Here, we analyzed one of the most well-studied TFs, the yeast zinc cluster Gal4, constructed a Shea–Ackers thermodynamic model to describe its binding, and compared the results of this model to experimentally measured Gal4p binding *in vivo*. We found that at many promoters, the model predicted no Gal4p binding, yet substantial binding was observed. These outlier promoters lacked canonical binding motifs, and subsequent investigation revealed Gal4p binds unexpectedly to DNA sequences with high densities of its half site (CGG). We confirmed this novel mode of binding through multiple experimental and computational paradigms; we also found most other zinc cluster TFs we tested frequently utilize this binding mode, at 27% of their targets on average. Together, these results demonstrate a novel mode of binding where zinc clusters, the largest class of TFs in yeast, bind DNA sequences with high densities of half sites.

GRAPHICAL ABSTRACT



INTRODUCTION

Transcription factors (TFs) orchestrate gene expression by recognizing and binding to specific DNA sequences that regulate gene expression. A mechanistic understanding of how transcription factors (TFs) identify and bind such sequences *in vivo* has been elusive (1). There has been recent progress using machine learning to predict patterns of *in vivo* binding (2–6), but such algorithms do not typically yield insights into the biophysical principles that govern TF specificity. A fruitful approach, therefore, is to generate specific mechanistic hypotheses from analyses of large scale *in vivo* and *in vitro* experiments (machine learning aided or otherwise) and then build biophysically motivated models to test them (6–8). If these biophysical models, which

*To whom correspondence should be addressed. Tel: +1 314 362 2751; Fax: +1 314 362 4227; Email: rmitra@wustl.edu
Present addresses:
David Song, Medical Education, University of Rochester School of Medicine, Rochester, NY 14642, USA.
Kristy Shady Lewis, Medical Education, University of Kentucky College of Medicine, Lexington, KY 40506, USA.

often have relatively few parameters and are interpretable, explain the data well, then support is provided for the proposed mechanism; however, failure of the model is often nearly as interesting because this forces the consideration of more baroque mechanisms that may violate standard assumptions about the mechanisms of TF action. Here, we report the latter. We constructed a Shea–Ackers thermodynamic model to predict the *in vivo* binding of the yeast TF Gal4 (9, 10). Our model incorporated knowledge about Gal4p binding preferences obtained from large scale *in vitro* binding experiments (11–14), biochemically measured homotypic cooperative interactions between Gal4p dimers (15), and correctly accounted for binding contributions from weak sites (16–18). We found that the model predictions correlated moderately well with *in vivo* Gal4p binding ($r = 0.73$), but for a large number of promoters, the model predicted no binding, yet substantial binding was observed. Further examination revealed these 'outlier' promoters all lacked canonical Gal4p binding motifs. This surprising discordance between theory and experiment motivated us to understand how Gal4p binds at these targets. We first confirmed that these genes were regulated by Gal4p by analyzing an orthogonal RNA-seq dataset. Subsequent investigation revealed Gal4p binds unexpectedly to promoters with high densities of its half site sequence, CGG. We estimate that Gal4p uses this novel mode of binding at 60% of its target promoters, and the *in vivo* binding of Gal4p was better predicted by CGG half-site density than by Gal4p's PWM, with areas under the curve (AUROCs) of 0.96 and 0.86, respectively. Furthermore, we found that a substantial fraction of yeast zinc cluster TFs employs this mode of binding. To dissect the molecular logic that governs this novel mode of TF binding, we performed Sort-seq on 6798 custom-designed yeast promoter sequences and measured their ability to drive reporter expression. These experiments revealed a linear relationship between CGG half site density and zinc cluster driven expression *in vivo.* We found that sequences with low densities of half sites (i.e. the expected frequency of CGG occurrence in the yeast genome) do not bind zinc cluster, whereas promoters with ∼10 half sites over a 200 bp region have the same transcriptional output as a single canonical Gal4p site. Furthermore, we found no simple relationship between half site orientation or spacing and the strength of gene activation. Together, our results demonstrate a novel mode of binding where TFs bind sequences with high densities of half sites. This binding mode is widespread, as it is frequently employed by the largest TF family in yeast to transcriptionally regulate their gene targets.

MATERIALS AND METHODS

Yeast strains

The Wild Type yeast strain used in this study was yRM1004 which was derived from matA_deltaSir4 and has the following genotype: *his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 Δsir4::KanMx Δtrp1::HygMx*. All TF knockout strains used were from the Yeast Knockout Collection (YKC) (19) and has the subsequent genotype: *MAT*a *his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*. Yeast strains used in transposon calling cards had the following genotype: *MAT*a *his3Δ0 leu2Δ0 met15Δ0 ura3Δ0 sir4Δ::KanMx (*see Supplemental Table 5 for all yeast strains used).

Library design and amplification

We designed a Sort-seq library that contained 6798 promoters, divided into five different TF (Transcription factors) sub-libraries, each sub-library was further divided into groups that addressed a specific feature of half site *in vivo* binding. Our Sort-seq library was designed and amplified using similar methods described in (20, 21), and constructed using array-based oligonucleotide synthesis. Each 230 bp oligonucleotide in the library contained five key sequence elements including: a 20 bp constant sequence that is homologous to the backbone plasmid, an 11 bp sequence unique to each sub-library to allow amplification of certain library elements, 170 bp user-defined variable synthetic promoter sequence, a 12 bp promoter barcode for identification of each promoter in Illumina sequencing, each promoter barcode has a hamming distance of 3. The last sequence element of each library sequence is a constant 17 bp sequence used for PCR amplification. Our library pool containing 6798 synthetic promoters was synthesized by Agilent. To amplify each TF sub-library, we divided 14.4 ng of the library DNA template in 96 wells for each sub library, in final 50 μl PCR reaction. Each 50 μl reaction included 0.2 mM dNTP mix, 0.5 μM forward primer, 0.5 μM reverse primer, 1X Herculase II reaction buffer, 1 M Betaine, 0.15 ng DNA template in water, 1 μl Herculase II polymerase (Agilent). The PCR cycling parameters were 95°C for 1 min, 16 cycles of 95°C for 30 s and 72°C for 2.5 min and then one cycle of 72°C for 7 min (see Supplemental Table 3 for list of primers use). PCR products from all 96 wells were combined for each sub library and concentrated using Amicon Ultra-0.5 centrifugal filter unit. Once concentrated, the DNA was purified using QIAGEN PCR MiniElute Purification Kit.

Library construction

Our Sort-Seq library was constructed using similar methods described in (20, 21). To prepare the vector plasmid pRM1806 for cloning (see Supplemental Table 4 for Addgene accession number), we linearized the plasmid with high fidelity restriction enzymes KpnI and ApaI (NEB) and purified the digested plasmid with a QIAGEN PCR MiniElute Purification Kit. We used a molar ratio of 3:1 plasmid to library DNA in a Gibson assembly reaction (NEB), following manufactures instructions. The Gibson assembly product was filtered by drop dialysis with a Nitrocellulose membrane (0.025 μm), following the Millipore Sigma protocol. The library product was electroporated into 7 0.1 cm (about 0.04 in) cuvette tubes, each containing 25 μl of *Escherichia coli* electrocompetent cells (Lucigen) and cells were then plated onto twenty-eight 15-cm Kanamycin containing plates. After 16 hours, plates containing 50 000 colonies each were scraped, and the plasmid DNA extracted using a Qiagen Maxiprep kit.

Yeast transformation

Yeast strains were transformed using the standard LiAc method with the library plasmid pRM1804, which carries

a *LEU2* auxotrophic selection marker, and plated on Synthetic Complete Glucose -Leu (SC-Leu) plates to select for the library plasmid. Sixteen transformations were executed in parallel per experiment to achieve a diverse population of sub-library sequences. After 2–3 days colonies were pooled together and grown in SC-Leu medium for 24 h. The yeast library cells were diluted in the desired medium (GAL4: Galactose -Leu, LEU3: SC– Leu, YRM1: SC-Leu, TEA1: SC-Leu) and grown for 6 hrs on the day of sorting.

Sort-seq

Library expression measurements and calculations were performed as described in (20, 21). Cells were sorted into eight bins of 150 000 cells each, and then added to SC-Leu media to allow cells to recover and grow for 16 h. Cells from each bin were individually pelleted and DNA extracted with EZ Yeast Plasmid Prep kit (G Biosciences). Next, we conducted eight separate PCR reactions in parallel to amplify the desired regions of the Sort-seq sub-library. The reverse primers used in each reaction were indexed with a unique barcode to allow the reactions to be pooled together for sequencing. The PCR cycling parameters were 95°C for 1 min, 30 cycles of 95°C for 30 s and 72°C for 2.5 min and then one cycle of 72°C for 7 min.

Obtaining relative expression of promoter sequences

Relative expression values for each library sequence were calculated using the same method described in (21). Briefly, the mean expression values of each flow cytometer bin are calculated by estimating the expression of cells in each bin. The number of cells in each bin are determined by the number of sequencings reads that are mapped back to that promoter in the specific bin and the reads in each bin are then normalized to match the fraction of the bin in the entire population. Finally, the mean expression of each promoter sequence is described as the fraction of each sequence in a bin across all bins.

Yeast transposon calling cards

The calling card yeast assay was conducted using the same methods described in (22, 23). Briefly, the desired yeast strain for each experiment is created by co-transforming two plasmids, one plasmid contains a Sir4p-tagged TF regulated by the *ADH1* promoter with a *LEU2* and *URA3* auxotrophic marker, and the second contains a Ty5 transposon driven by the inducible *GAL1/10* promoter with a *HIS3* auxotrophic marker (see Supplemental Table 4 for Addgene accession number). After transformation, a single colony is picked and grown overnight (30°C) in SC liquid media containing the desired auxotrophic selection. Ty5 transposition is then induced by plating the liquid culture on galactose induction plates and grown at room temperature for 4–5 days. Plates were then replica plated to YPD (30°C for 4–5 days) and then SC-His with 5FOA (30°C for 2–3 days). Colonies were scraped, pooled together, and their genomic DNA extracted using standard methods. Genomic DNA was then divided into three separate enzyme digest reactions consisting of either HpaII, HinP1I or Taqα1 (NEB). Digested DNA was then purified with a QIAGEN PCR MiniElute Purification Kit, self-ligated using T4 DNA ligase (NEB), and purified with Amicon Ultra-0.5 centrifugal filter unit. Ty5 calling cards were recovered from genomic DNA in the self-ligated template by inverse PCR and purified products were submitted for next-gen Illumina sequencing (see Supplemental Table 3 for list of primers use).

Analysis of yeast transposon calling card sequencing reads for quantification of TF binding

To recognize unique insertions from DNA sequencing data we used the same methods described in (24), we first filtered for Read 1 sequences containing the correct 17 bp Ty5 3'LTR sequence and verified experiment-specific barcodes on both Read 1 (5 bp) and Read 2 (8 bp) matched. Once filtered, the 17 bp LTR and Read1 barcodes were removed from the 5' end and 80 bp of Read 1 genomic sequence were aligned to the *Saccharomyces cerevisiae* reference genome sacCer2 (R-61-1-1) via Novoalign. We defined promoters as all intergenic regions spanning 150 bp into the ORF of the upstream and downstream gene that were smaller than 5 kb in length. To measure TF binding at a given promoter, we normalized the number of Ty5 insertions into each promoter so that the total number of insertions recovered for each experiment was equal to 100 000 and assigned a Transposons per Hundred Thousand (TPH) value to the TF binding at each promoter. Statistically significant binding was determined by first calculating the expected number of insertions at each promoter under the null. Poisson statistics was then used to calculate the p-value for each promoter and significant binding was defined as those promoters having a P-value < 1e−5.

Shea-acker's model

To predict Gal4p binding, we implemented a Shea–Acker thermodynamic model, like the ones described by us previously (24) and by Segal *et al.* (25). Briefly, this model uses a free energy matrix to scan promoter sequences and calculate the free energy contribution if Gal4p were to bind at each promoter sequence. We then used a dynamic programming algorithm which utilizes Boltzman statistics to calculate the relative probabilities of the different binding configurations of Gal4p at each promoter. To parameterize this model, Gal4p nuclear concentration was estimated from Ghaemmaghami *et al.* (166 molecules/nucleus, final nuclear concentration, 27.5 nm). To model the known homotypic cooperative interaction between Gal4p dimers, we also included cooperative term between adjacent Gal4p molecules which was parameterized for optimal fit. See our Gitlab repository for the source code as well as a Jupyter notebook detailing the implementation and analysis.

Quantification of CGG half sites

We defined promoters as all intergenic regions spanning 150 bp into the ORF of the upstream and downstream gene that were smaller than 5 kb in length. Intergenic regions were manually inspected to find loci that were bound by Gal4p but had no canonical motif. Half

site promoter sequences were defined as regions spanning 170 bp that contained clusters of eight or more CGG half sites. CGG half site counts were then determined computationally by searching promoter sequences for all instances of CGG half sites and returning an array of the indices of positions corresponding to the half site. Overlapping half sites were counted as a single half site and half sites within a canonical motif counted towards the total number of half sites for that promoter sequence.

Processing PBM and RNA-seq data

Zinc cluster PBM data was acquired from the UniProbe database for estimating relative K_D 's using similar methods described in (26). Briefly, we processed raw Gal4p PBM data (14) by taking the ratio of Alexa 488 and Cy3 fluorescence signal and multiplied by 1000 for every sequence. A PWM (Position specific Weight Matrix) scan was conducted on each sequence to determine if a strong canonical Gal4p binding site (PWM > 13) was present, and those without a canonical motifs were grouped based on CGG half site number. To calculate the relative K_D for each sequence we used Equation (1). The K_D of each sequence was then normalized to the mean K_D of sequences with a canonical Gal4p motif.

Equation (1)

F_i: fluorescence value for a specific sequence

F_M: is the maximum fluorescence value in PBM data

P_T: is the minimum concentration of Zinc cluster TF added to each PBM experiment (25 nM)

P_B: concentration of protein bound to all sequences (0 nM)

$$K_d^i = \frac{\left(F_m^i - F^i\right)\left(P_T - P_B\right)}{F_i}$$

RNA-seq data from yeast cells grown in glucose and galactose induction conditions were obtained from (27) and used in permutation test.

TF motifs

For computational analyses requiring yeast TF motifs, we used the recommended PWMs collected by Spivak and Stormo and obtained from the SceTF database (stormo.wustl.edu/ScerTF). In all cases, scoring cutoffs used were those recommended by ScerTF, and were used to distinguish between the existence or absence of TF binding sites on DNA sequences.

ROC analysis

Receiver operator curves plot true positive rate (TPR) versus false positive rate (FPR), with a perfect classifier yielding an area under the curve (AUC) of 1.0. For Gal4p and all other zinc cluster TFs used binding target ROC curves, 'True positives' were defined as statistically significant binding measured by calling card assay (P-value cutoff < 0.0001, while 'false positives' were defined as loci that did not demonstrate statistically significant binding (P-value cutoff > 0.0001). Two scoring functions were used to classify zinc cluster TF targets. The first was the highest scoring motif in each intergenic promoter region using a 400 bp window, and the second, is the max number of CGG half sites in a 400 bp window.

RESULTS

Most gal4p-bound promoters do not contain canonical gal4p motifs.

To investigate the factors that specify a transcription factors (TF's) *in vivo* binding, we examined the well-studied zinc cluster protein, Gal4p, which binds to two inverted 'CGG' half sites separated by an 11 bp spacer (5'-CGG-N11-CCG-3'). We began by first mapping Gal4p's *in vivo* binding using transposon calling cards, a method that utilizes the Ty5 retrotransposon to map transcription factor binding sites in the yeast genome (22–24). In this method, transcription factors are fused to a fragment of the Sir4p heterochromatin protein known to interact with the Ty5 integrase. Thus, any TF can be made to recruit the Ty5 integrase and thereby direct Ty5 transposon insertion, leaving a 'Calling Card' whose location can be retrieved via second generation-sequencing. After sequencing, a binding score is calculated by taking the number of Ty5 insertion events at each promoter and normalizing to 100 000 total insertions (transpositions per hundred thousand, TPH). This method provides a quantitative measure of TF binding (20) and is highly concordant with other, orthogonal methods for identifying TF regulated genes (28). In order to predict Gal4p binding based on the known biophysical principles that influence this TF, we constructed a Shea–Ackers (10, 25, 29) thermodynamic model. This model estimated the free energy of binding to different DNA sequences from the position specific scoring matrix (PSSM) found to be most predictive of Gal4p binding (11, 12, 30). For each promoter sequence in the yeast genome, it used dynamic programming to compute the expected binding of Gal4p by using the Boltzmann distribution and then summing the contributions from all possible binding sites for each DNA sequence, in a manner that accounts for the contribution of weak binding sites, similar to previously published methods. (25) Our model also accounted for the known cooperative interactions between Gal4p dimers (see Methods; code available on Gitlab). (15) We compared the model's predictions to the transposon calling card data (Figure 1A,B). Although there was a reasonably good correlation ($r = 0.73$) between the observed and predicted occupancy, this was largely driven by a small number of canonical Gal4p target promoters, nearly all of which encode one or more strong Gal4p motifs (e.g. the GAL1_10, GAL2 and GAL7 promoters, Figure 1A, B). More surprisingly, we observed several outliers (Figure 1B) for which the model predicted essentially no binding, yet strong Gal4p binding was observed (53 total outliers, 7 of which were among the top 10 most strongly bound Gal4p targets). Manual inspection of these promoters revealed that the Gal4p consensus site (5'CGG-N11-CCG3') was not present in these promoters. In fact, none of these 'outlier' promoters bound by Gal4p contained a motif that passed the recommended PWM threshold for Gal4p (and which was previously shown to optimally separate bound and unbound sites (11)) (Figure 1C). This

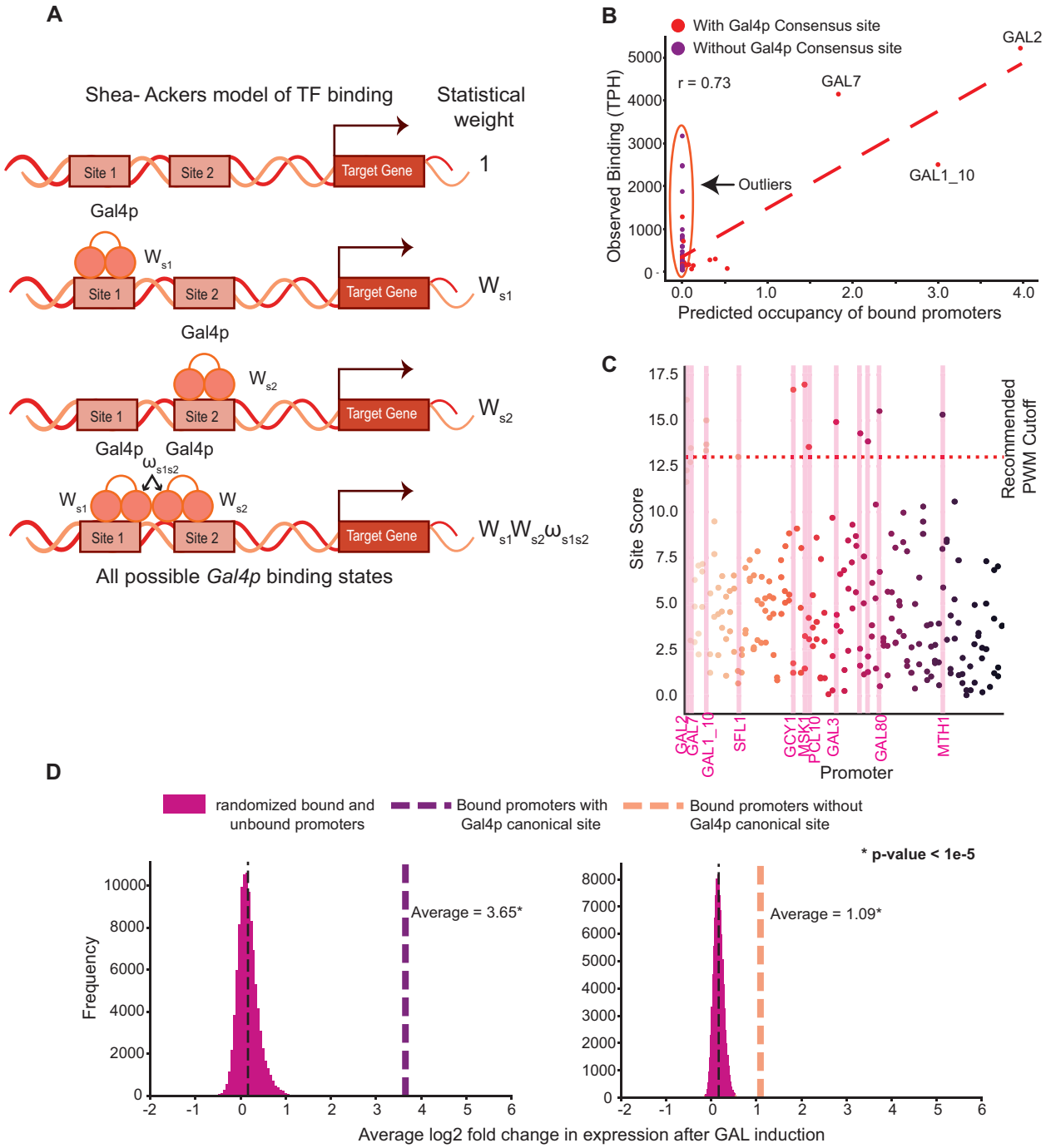


Figure 1. The canonical Gal4p motif is not present at many promoters where Gal4p is bound. (**A**) All possible molecular states of a Gal4p regulatory sequence with two binding sites, that is either bound or not bound by Gal4p. Each state's respective statist statistical weight W is shown to the right. (**B**) The correlation between observed Gal4p binding in the transposon calling card data and the predicted occupancy generated by a thermodynamic model. Promoters where Gal4p unexpectedly binds are circled in red and labeled as outliers. (**C**) Experimentally determined Gal4p targets are shown on the x-axis and their respective Gal4p site scores are plotted. All sites meeting the recommend PWM cutoff are colored highlighted in pink and those with no high scoring Gal4p motif are below the red dotted line. Most Gal4p targets do not have a binding motif above the recommended PWM cutoff. (**D**) A simulation showing the $\log_2$ fold change (x-axis) in expression of all yeast genes after galactose induction. Randomized promoters are shown in magenta, the mean $\log_2$ fold change in expression for bound promoters with a Gal4p motifs are shown in purple, and the mean $\log_2$ fold change in expression for bound promoters without a Gal4p canonical site are shown in orange.

suggested to us that Gal4p might bind to these promoters a novel fashion. We first sought to confirm that this motif-independent binding was not an experimental artifact but instead represented functionally important Gal4p binding. To do so, we analyzed RNA-seq profiles (27) of yeast cells grown in glucose vs. galactose conditions and determined the mean and median fold change in expression for genes regulated by Gal4p-bound promoters with and without a Gal4p site. As a negative control, we performed an identical analysis for unbound promoters. We found that genes regulated by promoters with unexpected binding displayed, on average, a 2.1-fold increase in gene expression, whereas genes regulated by promoters with canonical Gal4p motifs changed gene expression by 12.5-fold, on average. We conducted a permutation test to determine if the observed mean (or median) fold change was exceeded in 100 000 randomly selected gene sets of matched size and never observed a comparable mean (or median) fold change (Figure 1D, $P < 1e{-}5$; median fold change in Supplemental Figure 1S). These results indicate that the unexpected Gal4p binding is not an experimental artifact, but that the promoters in question are functional and activate gene expression in a galactose dependent manner.

Anomalous gal4p binding is not driven by cooperative interactions

We next sought to understand how Gal4p was binding to promoters lacking its canonical motif. We first considered the hypothesis that Gal4p might participate in cooperative interactions with other TFs, resulting in recruitment to these loci. We identified candidate TFs that bind to the same loci by using a Fisher's exact test to compute the overlap in TF binding in the transposon calling card data (Supplementary material Appendix 1). Notably, most of these candidates were zinc cluster TFs. We tested this hypothesis by knocking out each candidate cooperative binding partners and measuring Gal4p binding with calling cards, a strategy that has proven effective in the past (24). However, we observed no change in Gal4p binding suggesting that Gal4p was not binding with cooperative factors and there must be another mode of TF binding (Supplemental Figure 2S).

CGG half-site density predicts gal4p binding at promoters without canonical gal4p motifs

We noticed that promoters to which Gal4p unexpectedly binds encode for large numbers of CGG half sites. Previous *in vitro* gel-shift experiments have shown that Gal4p does not bind at a single CGG half site, nor does it bind to sequences encoding two half sites unless they are separated by precisely 11 bp (31). Therefore, we hypothesized that high densities of half sites might be required for Gal4p binding. To determine if there was a relationship between CGG half-sites and Gal4p binding, we investigated the ability of CGG half-sites to predict *in vivo* binding targets. The model was scored by scanning each promoter sequence with a 400 bp window and counting the maximum number of CGG half sites in that window. Interestingly, a receiver-operator curve (ROC) analysis of a Gal4p-binding model based solely on CGG half-site density outperformed a model based on Gal4p's PWM, with areas under the curve (AUROCs) of 0.96 and 0.86, respectively (Figure 2A). This suggests half-site density is highly predictive of Gal4p *in vivo* binding. To further examine the relationship between CGG half-sites and Gal4p binding, we used our calling card data to separate Gal4p transposon calling data into target and non-target sites based (using p-value cutoff of 1e-5). We then counted the number of CGG half-site occurrences for target and non-target sites and plotted the distribution shown in (Supplemental Figure 3S). We found that Gal4p targets contained significantly more CGG half-sites (mean = 9) than non-target promoters (mean = 4, $P = 4.8e{-}27$, by Welch's *t*-test). Similar results were obtained when we removed Gal4p's canonical motif and plotted the distribution of CGG half-sites, implying that CGG half-sites might promote binding at Gal4p target sites that lack canonical binding sites (Figure 2B). We next asked if the CGG density is uniquely predictive of Gal4p binding amongst all triplet sequences. To do so, we computed the densities of all 3mer sequences at yeast intergenic regions and attempted to predict Gal4p binding as we previously did for CGG half sites. We computed the AUC for all triplet sequences (Figure 2C). Strikingly, the highest AUC was for the CGG half site and the next most predictive triplets were C/G rich, whose densities would clearly be correlated with CGG half site density. These data further support the hypothesis that sequences with a high density of CGG half-sites are bound by Gal4p *in vivo*. Finally, we wanted to determine if there was a quantitative relationship between the strength of Gal4p binding *in vivo* and the number of CGG half-sites at a promoter. This was done by calling TF binding peaks using the same method described in (32) and correlating the number of insertions under calling card peaks with the number of CGG half-sites at each promoter. We obtained a good correlation ($r = 0.63$, $P = 7.24e{-}12$), indicating a strong association between Gal4p binding *in vivo* and CGG half-site number (Figure 2D). Together, these results validated our observation that Gal4p binds *in vivo* to clusters of CGG half-sites and this phenomenon explains its binding at promoters lacking the canonical Gal4p motif.

There is a relationship between CGG half sites and zinc custer TF binding

Like Gal4p, many other TFs in the zinc cluster protein family recognize and bind to palindromic CGG half sites separated by unique numbers of base pairs (33–35). Therefore, we next asked if zinc cluster TFs other than Gal4p bind at promoters with high half site densities. We performed transposon calling card assays for 11 additional zinc cluster TFs and asked how many significantly bound promoters lacked a canonical motif. For all 11 TFs, we found that a substantial fraction of bound promoters did not encode for the canonical binding motif (Figure 3A); for 10 of the 11 TFs more than half of the bound promoters lacked binding sites. We then compared the predictive power of PWM score to CGG half site density for the identification of *in vivo* binding targets. For 10 out of the 11 zinc cluster TFs, the CGG half site density performed significantly better than the PWM (Figure 3B, *P*-value < 0.001 by paired *z* test using a method described (36); code available on Gitlab). Only for

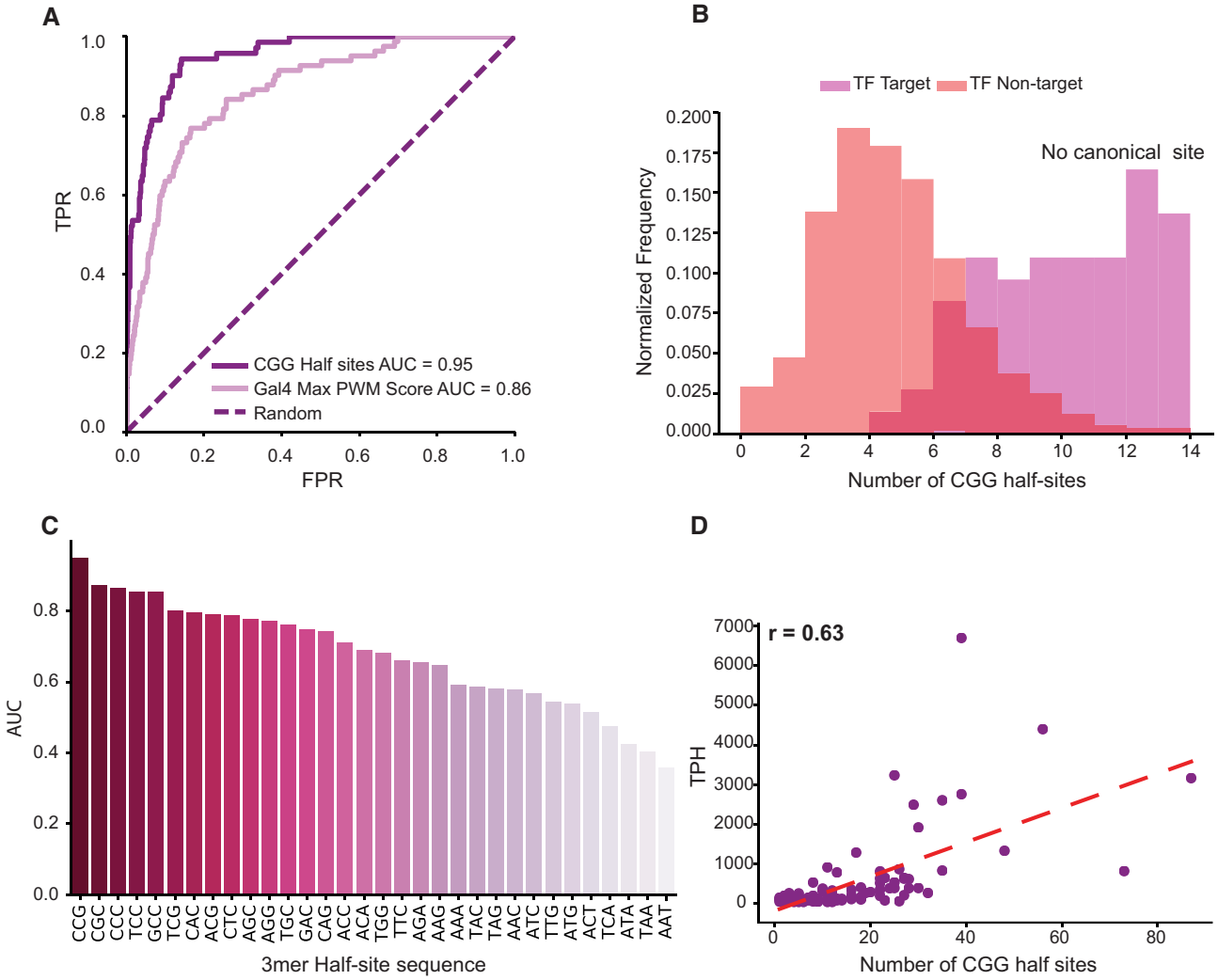


Figure 2. Relationship between CGG half sites and Gal4p binding. (**A**) Gal4p-bound loci can be differentiated from unbound loci by using CGG half sites. ROC showing the ability of Gal4p PWM score to distinguish between bound and unbound promoters (light purple), compared to a model using CGG half sites alone (dark purple). (**B**) Distribution of CGG half sites for bound and unbound promoters. Shown in pink are non-target sites, in purple Gal4p targets, and the overlap between the two is colored in red. Non target sites have low levels of CGG half sites while Gal4p targets show a high level of CGG half sites. (**C**) AUC for every possible 3mer half site. CGG half sites and C/G rich 3mers can differentiate between bound and unbound Gal4p promoters. (**D**) Quantitative correlation between Gal4p *in vivo* binding strength (TPH) and CGG half site number. A high correlation between TPH and CGG half site number is shown.

Lys14p did the PWM perform significantly better at classifying *in vivo* binding targets (Figure 3B). These results suggest that CGG half sites are predictive of zinc cluster TF *in vivo* binding. We next investigated the quantitative relationship between CGG half-site density and promoter occupancy for all 11 zinc cluster TFs. As before, we called calling card peaks, calculated a normalized binding score (TPH) and correlated this value with CGG half site density at intergenic regions lacking canonical motifs. For all 11 zinc cluster TFs, a positive association was observed between CGG half sites and binding strength, with 6 out of the 11 TFs having a correlation coefficient greater than Gal4p (R between 0.70 and 0.94) with 5 out of the 11 having R values less than Gal4p (R between 0.31 and 0.58) (Figure 3C). This result suggests that for many zinc cluster TFs, significant binding occurs at promoters with a high density of CGG half sites.

In vitro TF binding assays support the hypothesis that zinc clusters bind at sequences with high half-site densities

We next hypothesized that if sequences with high densities of CGG half sites were truly able to recruit Gal4p, we would observe this relationship in *in vitro* TF binding datasets. Therefore, we examined protein microarray datasets (PBM), a high throughput method widely used to determine TF binding specificities, for Gal4p (13, 37, 38). Briefly, microarrays for PBM experiments are designed with overlapping k-mers binding sites. The TF of interest is then expressed with an epitope tag, purified, and added to the microarray. Next, the amount of DNA-bound TF is quantified using a fluorescently labeled antibody. Some key strengths of this method include the ability to test tens of thousands of sequences in parallel and measure DNA-protein interactions with extremely high sensitivity. We

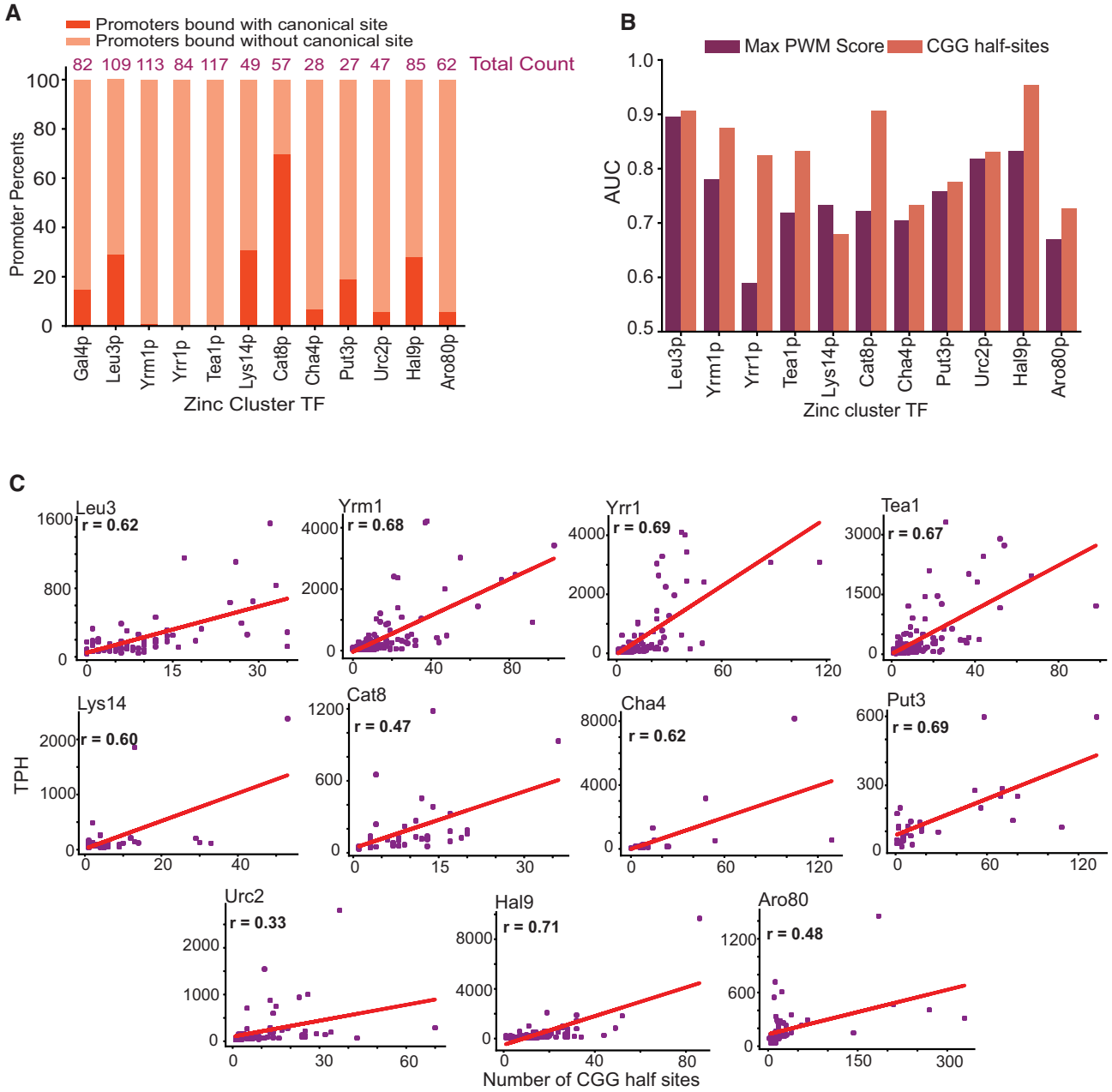


Figure 3. Relationship between CGG half sites and Zinc cluster TF binding. (**A**) Total number of significantly bound promoters is showed in pink. Percent of bound promoters with motif shown in dark orange and percent of bound promoters without motif in light orange. (**B**) Zinc cluster TF-bound promoters can be distinguished from unbound promoters using a CGG half site model. AUC comparison showing the ability of their recommended PWM to differentiate between bound and unbound promoters (purple) to a model that uses CGG half sites (orange). (**C**) Quantitative correlation between zinc cluster TFs *in vivo* binding strength (TPH) and CGG half site number. A high correlation between TPH and CGG half site number is shown.

analyzed PBM sequences by scanning each sequence for CGG half sites and counted the number of occurrences in each sequence. We then grouped sequences based on the max number of half sites and computed the mean binding score for each group, and again, observed increased Gal4p binding at CGG half-sites (Figure 4A). We asked if the same trends could be observed *in vitro*, using previously published PBM data (Figure 4A) available for 10 yeast zinc cluster TFs. For 9 of the 10 zinc clusters, we observed a significant correlation between binding and CGG half site density (Figure 4A). As a control, we conducted the same analysis on 3 bHLH TFs and observed no over-representation of half-sites at sequences lacking canonical binding sites (Supplementary Figure 4S), suggesting that half site binding is a unique feature of zinc cluster TFs. We next sought to determine the absolute binding affinity that Gal4p for has for sequences with CGG half sites by calculating the K_D of each PBM sequence using similar methods described in (26). We found that the K_D decreased with the number of half sites, and that sequences with seven half sites had a

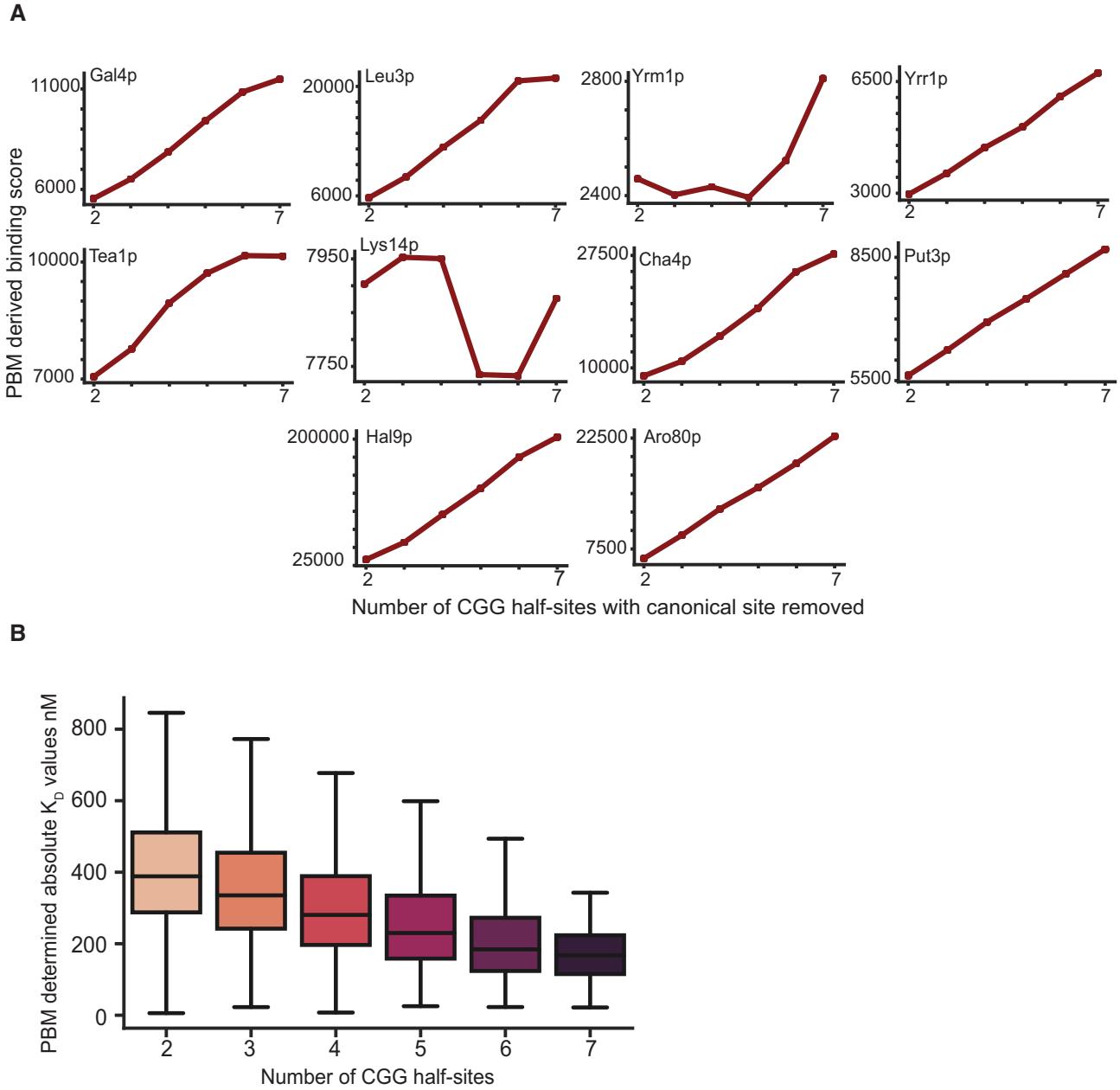


Figure 4. PBM validation. (**A**) Binding score increases with number of CGG half sites for many zinc cluster TFs. (**B**) Absolute binding affinity of Galp increases with CGG half site number.

K_D comparable to sequences with a Gal4p canonical binding motif (mean = 189.2, Figure 4B, mean = 120.9, Supplemental Figure 5S). Together, these results validated our observation that there is a relationship between Gal4p binding and CGG half sites.

CGG half site density influences binding and gene expression *in vivo*

Having established that zinc cluster TFs bind at sequences of high CGG density both *in vitro* and *in vivo*, we next sought to determine if such binding is functional – does binding drive changes in gene expression? To answer this question, we performed Sort-seq (20, 21, 39, 40) to quantify the gene expression of 6798 barcoded synthetic yeast promoter sequences that were generated using array-based oligonucleotide synthesis. Each synthetic promoter was synthesized with three different barcodes. The Sort-seq library was divided into sub-libraries, one for each of four zinc cluster TFs whose binding was correlated to half site density *in vivo* and *in vitro*: Gal4p, Leu3p, Yrm1p and Tea1p, and we analyzed 4–7 promoters for each zinc cluster TF. Each sub-library was cloned upstream of a YFP reporter gene on a plasmid that also constitutively expressed mCherry so that gene expression could be normalized and measured using Sort-seq. We amplified each library in

E. coli and then transformed into yeast and grew in the desired condition. As an additional control, we transformed each sub-library into the corresponding TF knockout strain. Yeast cells were then sorted into subpopulations using flow cytometry based on the ratio of YFP and mCherry fluorescence. These subpopulations were then sequenced and the proportions of each promoter barcode in each binned library was used to calculate the relative expression.

In this experiment, we used 200 bp synthetic promoter fragments, so to determine whether these truncated promoters were still regulated by their cognate zinc cluster TFs, we investigated gene expression driven by the unmutated promoters in zinc cluster TF knockout and WT yeast strains. In almost all cases, there was a significant reduction in promoter gene expression in the corresponding knockout strain (Supplemental Figure 6SA). Therefore, we conclude that these promoter fragments are good models for the native promoters in the context of our question.

We first asked if the activities of promoters without zinc cluster TF canonical motifs were governed by regions of high CGG density. To do so, we mutated promoters encoding at least 8 CGG half sites to either CAA or CTT in triplicate, each encoded by a unique barcode (for a total of six mutant sequence per promoter, Figure 5A). We analyzed 10 promoters bound by 4 different zinc-cluster TFs (Figure 5A). As a control, we also measured the expression of matched sequences where randomly selected non-CGG 3mers were mutated to CAA or CTT (Supplementary 6SB).

For all promoters tested, the mutation of CGG half sites to CAA or CTT significantly reduced the gene expression driven by the promoter (Figure 5A, $P < 0.05$). We observed 3.3-fold, 1.9-fold, 2.3-fold, 3.11-fold decreases in gene expression upon mutation of CGG half sites in the Gal4p, Leu3, Tea1 and Yrm1 promoters respectively. In contrast, we found that while randomly mutating triplet base pairs to CAA or CTT control promoter sequences reduced gene expression in some instances, the magnitude of the effects were much smaller than what was observed when CGG half sites were mutated (−0.75-fold, −0.47-fold, 0.13-fold, −0.43-fold change in gene expression, Supplemental Figure 6SB).

We then asked if we could observe this same relationship between CGG half site density and expression in an orthogonal *in vivo* method. We analyzed a dataset in which ∼200 different TFs were induced using an estradiol induction system and then a transcriptome-wide time series was collected at ∼8 different time points after induction (0, 5, 10, 15, 20, 30, 45 and 90 min), and the RNA from each induction experiment was hybridized to an Agilent microarray. This design allows one to identify the direct targets of the induced TF, as these genes are expressed very quickly—typically after 5 min—whereas indirect targets are induced at a slower rate. We compared the expression values across the 8 time points for sequences that were bound in our transposon calling cards CC data and contained a Gal4p canonical motif, were bound and had 10 or more half sites in a 200 bp window or were unbound in the CC data. We conducted a t-test between half site sequences and unbound sequences to determine if their mean expression values at different time points (30.0, 45.5, 90.0 min) were significantly different (Supplemental Figure 7S, $P < 0.01$). We found that genes regulated by promoter sequences containing clusters of CGG half sites were significantly upregulated relative to unbound Gal4p targets, suggesting half site promoters are bona fide Gal4p targets. Taken together, these results demonstrate that CGG half sites are necessary for the binding and activity of zinc cluster TFs in promoters without canonical binding sites.

Zinc cluster TF activity scales linearly with CGG half site density

We next sought to quantify the relationship between CGG density and gene expression. Does gene expression scale linearly with CGG density, or is there a sharp transition above some density threshold? To answer this question, for each zinc cluster TF, we took the regulatory targets with high CGG densities and mutated CGGs to create a series of mutant promoters with increasing numbers of CGGs (from 0 CGGs to the original number of CGGs in the WT promoter sequence). For each of the four zinc cluster TFs tested (Gal4p, Leu3p, Yrm1p and Tea1p), we observed a monotonic, nearly linear increase in gene expression with CGG half-site number (Figure 5B). For each TF, the slope of the increase was roughly the same at all promoters tested. Importantly, little to no linear increase was observed when the mutant promoter libraries were analyzed in TF knockout strains (Supplemental Figure 8S). From these results, we conclude that TF activity scales linearly with CGG half site density.

Promoters with mutated canonical motifs can be rescued with half-sites

We next asked if canonical zinc cluster binding sites could be replaced with half sites, and if so, how many half sites were required to drive gene expression to the same level as a canonical site. To answer this question, we took 2–4 WT promoters per TF with one or more strong canonical binding sites, mutated these sites, and then made a series of 'reprogrammed' promoters where we added 1–25 CGG half sites (Figure 5C, top left panel). Thus, for each promoter, we created 45 reprogrammed sequences and measured their expression using Sort-seq. We first analyzed the well-studied *GAL1-10* promoter, which contains four strong Gal4p consensus sites and drives extremely high levels of gene expression under galactose induction. It has been previously reported that Gal4p expression is highly sensitive to the mutation of the Gal4p consensus binding sites (41), and indeed, we also observed a 95.7% reduction in expression from the *GAL1* promoter when we mutated these sites. By adding 25 CGG half sites to the naked promoter, we were able to achieve close to half the levels of WT expression (Figure 5C, top right panel). Since the GAL1-10 promoter is controlled by 4 strong Gal4p binding sites, we wondered if a promoter regulated by a single Gal4p binding motif could be more effectively reprogrammed. Thus, we made a similar series of mutants for another Gal4p-regulated promoter, *RUP1_SFL1*. We observed a 77.5% reduction in gene expression upon mutation of the canonical binding site, and we found that adding CGG half sites increased expression

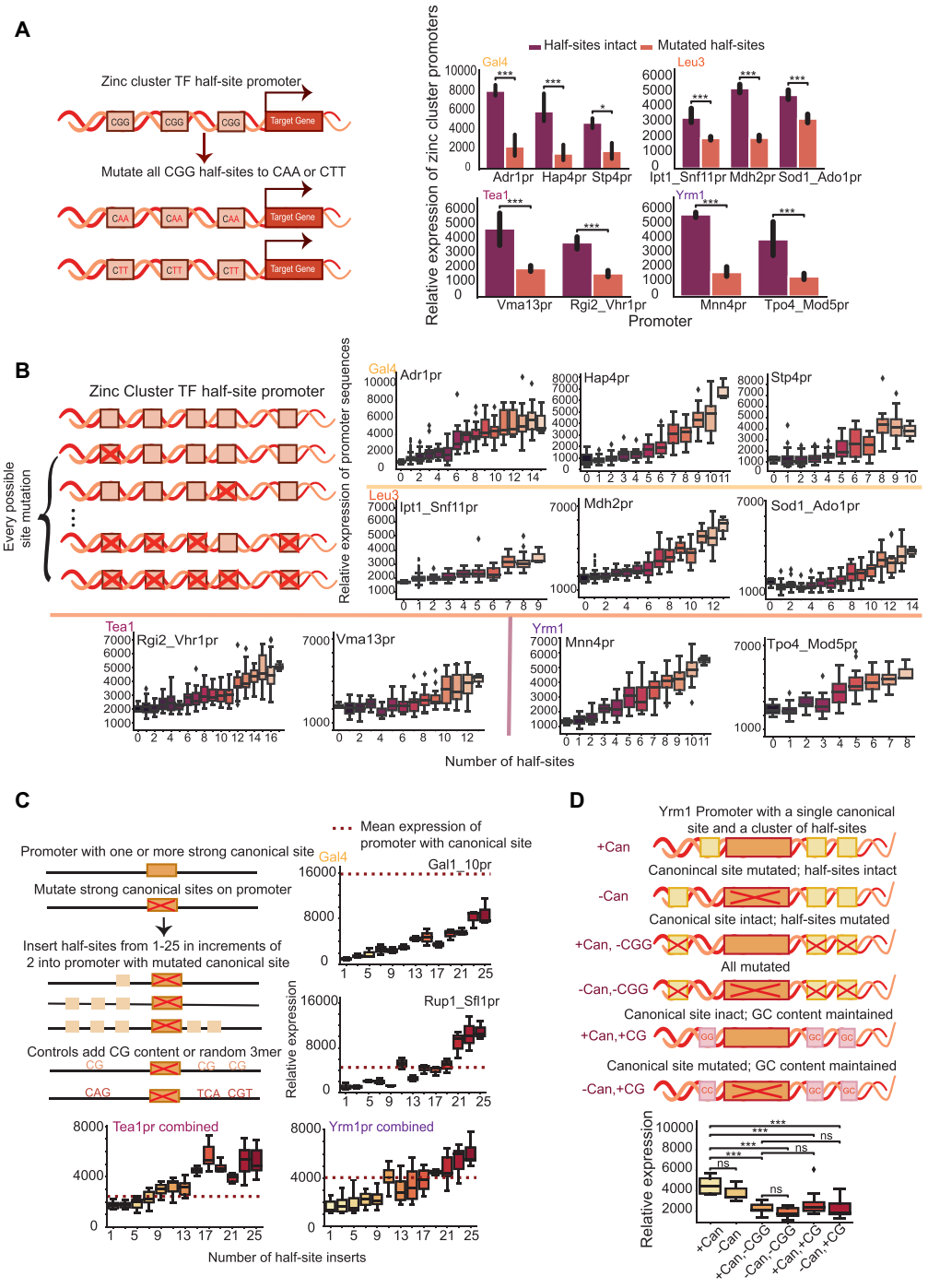


Figure 5. Enrichment of CGG half sites influences binding and expression *in vivo*. (**A**) To test the effect of CGG half sites on binding and expression *in vivo* all CGG half sites were mutated in the half site promoters and their expression measured using Sort-seq. There is a noticeable difference in expression between WT half site promoters (purple) and mutant half site promoters (orange). A two-way ANOVA and post-hoc test were performed to test the significance, one star indicates a *P*-value <0.05 and three stars indicates a *P*-value < 0.001. (**B**) Every possible half site mutation combination was generated, and expression was measured using Sort-seq. A linear increase in expression with half site number is shown for all CGG half site promoters. (**C**) To determine how many half sites are needed in a promoter to observe full site binding, up to 25 CGG half sites were inserted in promoters with a mutated canonical binding motif, and expression measured using Sort-seq. Except for the Gal1_10pr, all other promoters were able to achieve the same levels of expression as a WT promoter with a strong binding motif. (**D**) To determine if a high density of CGG half sites increases binding and expression of promoters with a canonical binding site we took Yrm1p promoters with a high density of CGG half sites and a canonical site and created three types of mutant sequences including: sequences where the canonical site is mutated but half sites remain intact, sequences were the canonical site is intact but all half sites are mutated, and promoter sequences with their canonical site and half sites mutated. Two CG controls were generated by taking a naked promoter and adding the same number of CGs as the WT promoter and taking a promoter with the canonical site intact, but half sites mutated and adding in an equivalent amount of CGs. Expression of all mutant promoters were measured using Sort-seq. There was a noticeable decrease in expression for mutant sequences with no CGG half sites and increasing CG content had no real effect on expression. A Two-way ANOVA and post-hoc test were performed to test the significance, ns indicates no statistical significance and three stars indicates a *P*-value <0.001.

levels in an approximately linear fashion. We found that 11 CGG half sites added to the naked promoter were able to restore WT levels of expression, and, when 25 half sites were inserted, WT levels of expression were exceeded by 148% (Figure 5C). We next asked if other zinc cluster TFs could be reprogrammed. We analyzed four promoters for the zinc cluster TF Yrm1p, and because these promoters all had very similar WT levels of relative expression, we plotted the average gene expression for all four promoters as we increased half site density (Figure 5C bottom right panel). As with Gal4p, deletion of Yrm1p's canonical binding site caused an average reduction of 75.1% from WT levels. We observed a linear increase in gene expression as the number of CGG half sites increased, and WT expression levels were restored when ∼11 CGG half sites were added. WT expression levels were significantly exceeded with 21–25 CGG half sites (Figure 5C, bottom right panel). Similar results were observed for Tea1p promoters: WT promoter expression levels were reached at around 7 CGG half sites, and the expression surpassed WT levels at around 13 CGG half sites (Figure 5C, bottom left panel). As controls, we increased the GC content of our Gal4p, Yrm1p and Tea1p naked promoters until they had equivalent levels of CG content as our naked promoters with CGG half sites inserted. We also included naked promoters with random 3mers inserted. Although some of our CG control promoters displayed increased gene expression, we did not observe a monotonic increase in expression and rarely observed expression values that surpassed the WT promoter (Supplemental Figure 9S), so we ascribed the occasional increases as the result of the serendipitous creation of a binding site for another yeast TF. Additionally, our random 3mer controls looked as expected, with no monotonic increase in expression and no values that were equivalent or exceeded the relative expression of the WT promoter sequences (Supplemental Figure 9S). Taken together, these results demonstrate that CGG half sites can be used to replace canonical TF motifs, and, if inserted at a high enough density, can drive gene expression at levels equivalent to canonical zinc cluster binding sites.

Half sites help recruit zinc cluster TFs to weak or singleton canonical motifs.

We next asked if a high density of CGG half sites enhance TF binding and gene expression at promoters with single or weak binding sites. We looked for zinc cluster promoters with a single canonical motif and clusters of CGG half sites nearby. We selected three Yrm1p-bound promoters that contained isolated motifs (PWM scores of 9.19, 9.47, 7.40) and a high density of CGG half sites nearby (9–17 half sites). We then generated three different types of mutant promoters by first taking the WT promoters (labeled + Can in Figure 5D) and mutating their canonical binding sites while keeping the CGG half sites intact (labeled -Can in Figure 5D). Second, we mutated all CGG half sites of each promoter while preserving the canonical motif (labeled WT-CGG in Figure 5D). Third, we mutated all CGG half sites and the single canonical motif (labeled -Can, -CGG). We found that mutating the canonical motif alone resulted in only a modest, borderline significant decrease in gene expression ($P = 0.07$, by two-way ANOVA; Figure 5D). Remarkably, when we mutated all CGG half sites but left the canonical motif intact, we saw a much larger decrease in gene expression ($P < 0.001$, two-way ANOVA, Figure 5D). Mutating the canonical motif in addition to these half sites further reduced gene expression; again, this reduction was modest, but significant ($P < 0.001$, two-way ANOVA). This reduction in gene expression was not the result of changes in GC content due to half-site removal, because when random CGs were added back into the mutated sequences, restoring average GC content, expression was not rescued (Figure 5D, +Can + CG, −Can + CG). We then asked if we could increase expression at weak singleton promoters by adding in CGG half sites. We took weak Gal4p and Leu3p singleton promoters with no CGG half sites nearby and created two types of mutant sequences by first mutating their canonical motifs (Supplemental Figure 10S, −Can) and then adding in 16 CGG half sites (Supplemental Figure 10S, −Can + CGG). We found that adding in 16 CGG half sites increased expression of promoters with a mutated canonical motif ($P < 0.001$, two-way ANOVA, Supplemental Figure 10S). Conversely, we saw no increase in expression at mutant promoters with increased CG content ($P =$ ns, two-way ANOVA, Supplemental Figure 10S). Taken together, these results indicate that at some zinc cluster promoters, CGG half sites help recruit TFs to canonical binding motifs.

Effects of half-site spacing and orientation on TF binding and expression *in vivo*

Gal4p binding at canonical motifs obeys strict spacing rules, as early experiments showed that changing the spacing between the CGG half sites of the consensus motif by even 1 bp greatly reduced the binding affinity of Gal4p (>500-fold reduction) (34). This sensitivity to spacer length was also observed for another zinc cluster TF, Ppr1p, which recognizes two palindromic CGG half sites separated by 6 bp (31). Zinc cluster binding is also constrained by sequence orientation, as a recent study measuring the effects of different systematic changes to consensus sites on gene expression found strict requirements on half-site orientation for Yrm1p and Tea1p (21). The strict grammar requirement at canonical TF binding sites led us to ask if similar spacing and orientation requirements applied to the half site mode of binding. To answer this question, we designed promoters for which the normal spacing between CGG half sites was increased in increments of 2–10 bp apart relative to the WT sequence. For each promoter we designed five sequences with mutated spacing and each sequence was tagged with three different barcodes. Due to our sequence length constraint of 170 bp, some CGG half sites were lost in the mutant sequences with large spacing between half sites. We next measured the expression of these sequences using Sort-seq. We observed no obvious trend in the relative expression of these mutant promoters as we increased the distance between CGG half sites (Figure 6A), suggesting that half-site distance is not a strong determinant of zinc cluster binding to half sites, at least over short distances. To determine if half site orientation influences gene expression, we changed the orientation of CGG half sites in each promoter to all face the forward orientation, the reverse

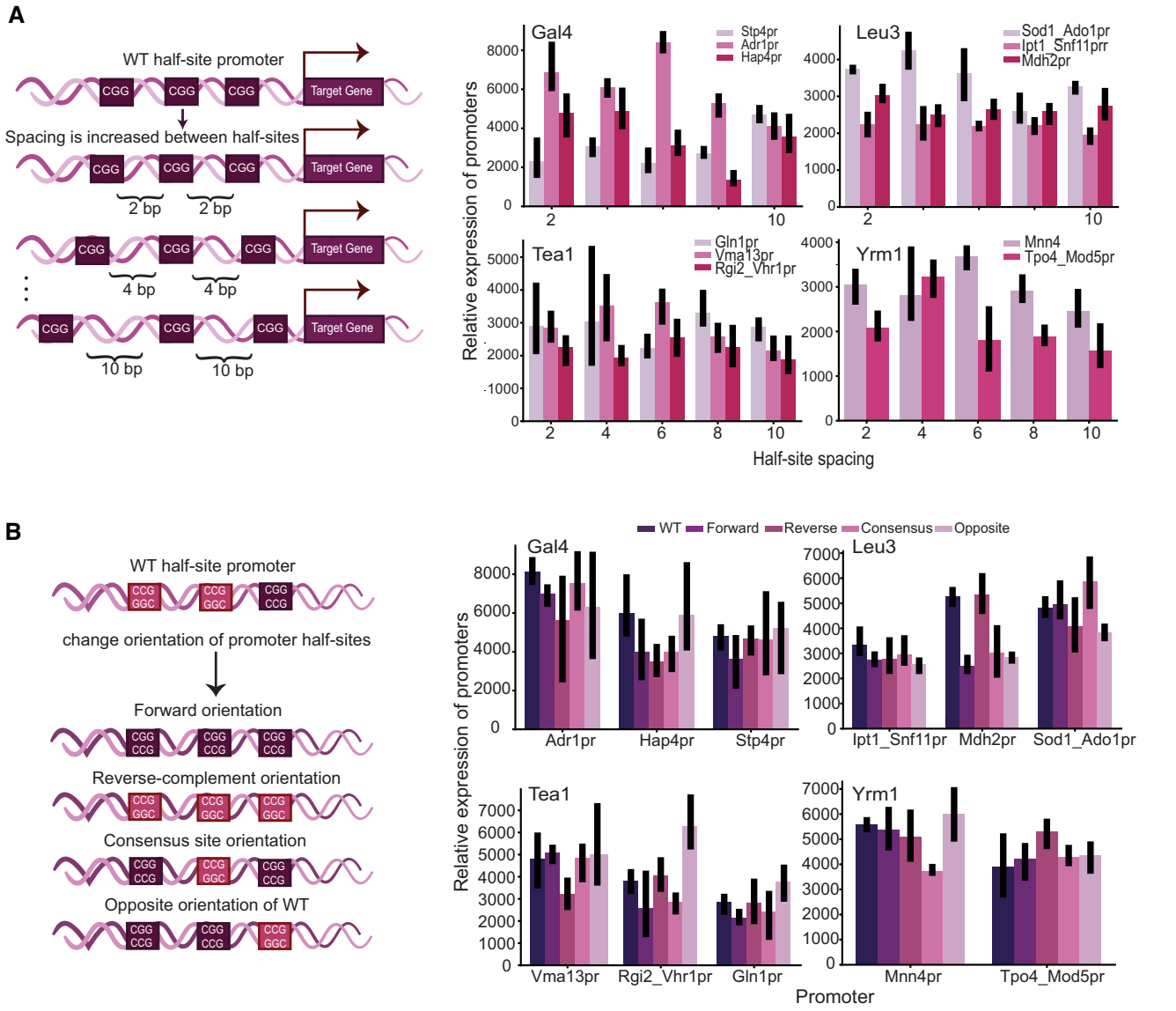


Figure 6. Half site grammar has no effect on binding and expression *in vivo*. (**A**) To investigate if half site promoters have a strict grammar requirement, the spacing between half sites was increased in increments of 2 until each half site had an additional 10bp between them and their expression was measured using Sort-seq. No significant trends in expression are shown. (**B**) Orientation of half sites were changed to the forward orientation, the reverse complement orientation, and the orientation opposite of the WT half site promoter. No significant trends in expression were observed when expression was measured using Sort-seq.

complement orientation, consensus site orientation, or an orientation exactly opposite of the WT half site promoter. We again saw no obvious trend in the gene expression changes as we modulated orientation (Figure 6B). Apart from one Leu3p promoter, we observed no trend in expression when we randomly changed the orientation of the CGG half sites across the promoter (Supplemental Figure 11S). We then asked if we could observe a relationship between half-site distance using the *in vitro* PBM data. We analyzed Gal4p PBM data to see if strength of binding correlated to half site spacing. We took all sequences containing seven half sites and determined the distance between each half site and observed no obvious trend in Gal4p binding strength. Similarly, we looked at sequences containing only two half sites and found a weak, but significant relationship where CGGs within 10 bp of one another were more likely to be bound then CGGs >20 bp apart ($r = 0.13$; Supplemental Figure 12S). Taken together, these results indicate that for several zinc cluster TFs, there is no simple relationship between half site spacing and orientation.

DISCUSSION

In this study, we tested the hypothesis that Gal4p and other zinc cluster TFs utilize a novel half site mode of binding to interact with sequences lacking canonical binding motifs. We have presented several lines of evidence to support this hypothesis that include: the ability of CGG half sites to predict zinc cluster TF *in vivo* binding, a clear relationship between CGG half site density and zinc cluster

binding across multiple *in vivo* and *in vitro* binding datasets, and a linear relationship between CGG half site density and zinc cluster driven expression *in vivo*. Additionally, we found that promoters encoding 10 half sites have the same transcriptional output as promoters with a single high-scoring canonical motif, and we found no clear effects of half site orientation and spacing on strength of activation. Notably, we found convincing evidence that for the zinc cluster TFs studied, 141 genes are regulated by half sites (25% of total targets, Supplemental Figure 13S). Taken together, these results demonstrate that this novel mode of half site binding is widespread across the yeast genome and is utilized by the largest TF family in yeast to regulate transcription of target genes.

Although Gal4p and other zinc cluster TFs have never been previously shown to bind to clusters of CGG half sites at promoters lacking canonical motifs, previous studies on the genome-wide binding of the bZIP yeast transcription factor Gcn4p showed its ability to weakly bind the half site sequence, ATGAC (42). Additionally, their study showed, G-SELEX peaks with more than three half sites have a higher occupancy than those with fewer half sites. This observation agrees with our finding that zinc cluster binding and expression *in vivo* scales linearly with half site number. Furthermore, nuclear hormone receptors (NRs), which are thought to be analogous to zinc cluster TFs, have been shown to bind DNA using a half site mode (38, 43, 44). Direct binding assays showed that NRs can engage in a half site mode of binding even when a full site is present. They also found that half site binding is not affected by the orientation of the half site and that high affinity binding to full sites with many different spacer lengths is predominantly mediated by a half site binding mode (38). These observations further support our findings that half site orientation and spacing has no effect on the strength of transcriptional activation of zinc cluster regulated promoters.

Despite the insights previously described, the methods used in our study have some limitations, the most significant of which is the limited (230 bp) length of the promoter regions in the Sort-seq library. Shortened promoters and enhancer do not always fully recapitulate the expression patterns driven by the full-length element. We mitigated these concerns by verifying our shortened promoters were regulated in a TF-specific manner by repeating our experiments in knockout strains. Furthermore, for two of the zinc cluster proteins analyzed, Gal4p and Leu3p, we confirmed promoter activity was induced in response to galactose and leucine respectively. Some of our observations were challenging to explain. For example, mutation of random 3mer sequences in half site promoters sometimes had a significant effect on gene expression. One possible explanation is that we mutated unknown regulatory features that contribute to TF specificity at half site promoters; several recent studies have shown the importance of flanking bases and GC composition on TF binding (45–48). Another possibility is that some of the introduced mutations inadvertently created binding sites for other yeast TFs. Future synthetic promoter library designs and Sort-seq studies could help clarify the observations specified above.

Our results raise an important question: if most zinc cluster TFs can bind at half sites, is there any specificity at half site promoters, and if so, how is this achieved? We performed a detailed analysis of 4 zinc cluster TFs and found a significant enrichment for combining at these half-site promoter targets (Figure14S, Supplemental Tables 1 and 2). However, for each of the four TFs, only about 25% of their half site targets overlap with one or more of the other three TFs, so there is clearly a mechanism by which these TFs achieve specificity at their half-site targets. Our MPRA experiments (Figure 6) investigated the effects of orientation and spacing and ruled these out as major factors contributing to the observed specificity. It is likely, then, that a combination of factors, such as the structural features of the protein, the local DNA context, and cooperative or competitive interactions, influence TF half site binding specificity. For example, the linker domain structure varies widely among zinc cluster TFs and has been shown to influence DNA-binding specificity with mutations in the liker region leading to changes in protein function and DNA binding (49), and these differences in linker structure between the zinc cluster TFs could be why some zinc clusters appear to recognize certain pattern of half site sequences. Another structural feature of TFs that could potentially play a role in determining specificity are long intrinsically disordered regions that are present in nearly all zinc cluster TFs. Recent studies in yeast highlighted the importance of IDRs in promoter selection (50,51). They demonstrated that IDRs contain scores of specificity determinants that influence and increase the speed at which a TF locates and binds to promoters it favors. Furthermore, it has long been established that DNA shape plays a crucial role in binding specificity. It is possible that the flanking sequences in half site promoters influence zinc cluster binding by modulating DNA shape. Finally, competitive, and cooperative interactions between zinc cluster TFs could also play a significant role in fine-tuning binding profiles at half site promoters. It has been shown that some zinc cluster paralogs that bind to the same motif compete for their preferred sites while others interact and recruit each other to their preferred sites (51). The mechanisms specified above are not mutually exclusive, and some combination of these likely explain the observed specificity in zinc cluster binding at half site promoters. We think this is an important question to investigate in future studies.

Transcription factors initiate patterns of gene expression that influence a wide array of biological processes, yet the molecular logic that governs how a TF locates its gene targets and the functional outcome of binding is not well understood. Our study, highlights that there are likely many yet uncharacterized mechanisms of DNA-TF binding. The work presented here demonstrates that many zinc cluster TFs use a half site mode of binding to regulate transcription of target genes. We hope to better characterize this novel mode of binding in future studies, specifically asking whether Gal4p and NRs share a common mechanism for half-site binding (38). Additionally, we would be interested in extending our study to a wider list of zinc cluster TFs as well homologous TFs (such as nuclear hormone receptors) in higher eukaryotes such as mice and humans; we expect this half site mode of binding to be a widely used mechanism across taxa. We postulate that these short half site sequences can evolve quickly and play a key role in the diversification of gene expression patterns, while longer

canonical sequences tend to be more conserved and are critical for maintaining specific regulatory programs. Such studies would increase our understanding of TF binding specificity and ultimately create a more complete map of gene regulatory networks.

DATA AVAILABILITY

Synthetic DNA library and the analyzed results are available as excel spreadsheet. Scripts and samples for analysis are provided in https://gitlab.com/pamelarecio1/halfsite_binding_codes. Raw sequencing reads are available in GEO with series number GSE216840.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the personnel at the DNA Sequencing Innovation and High Throughput Computing Facility at The Edison Family Center for Genome Sciences and Systems Biology of Washington University in St. Louis for their sequencing and computational expertise. We are grateful to Jim Skeath, Barak Cohen and Ting Wang for their valuable discussions and comments on the manuscript. P.S.R. and R.D.M. designed research; P.S.R., N.J.M., C.A.S., D.S., G.J., K.S.L., X.C. performed research; P.S.R., N.J.M., R.D.M. analyzed data; and P.S.R. and R.D.M. wrote the paper.

FUNDING

National Institute of Mental Health [RF1MH117070, RF1MH126723]; National Institute of General Medical Sciences [R01GM123203, 5R25GM103757-08, 5R25GM103757-09]. Funding for open access charge: Fund number is [28051].
Conflict of interest statement. None declared.

REFERENCES

1. Zeitlinger,J. (2020) Seven myths of how transcription factors read the cis-regulatory code. *Curr. Opin. Syst. Biol.*, **23**, 22–31.
2. Moerman,T., Aibar Santos,S., Bravo González-Blas,C., Simm,J., Moreau,Y., Aerts,J. and Aerts,S. (2019) GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, **35**, 2159–2161.
3. Weirauch,M.T., Cote,A., Norel,R., Annala,M., Zhao,Y., Riley,T.R., Saez-Rodriguez,J., Cokelaer,T., Vedenko,A., Talukder,S. *et al.* (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.*, **31**, 126–134.
4. Greenside,P., Shimko,T., Fordyce,P. and Kundaje,A. (2018) Discovering epistatic feature interactions from neural network models of regulatory DNA sequences. *Bioinformatics*, **34**, i629–i637.
5. Avsec,Ž., Weilert,M., Shrikumar,A., Krueger,S., Alexandari,A., Dalal,K., Fropf,R., McAnany,C., Gagneur,J., Kundaje,A. *et al.* (2021) Base-resolution models of transcription factor binding reveal soft motif syntax. *Nat. Genet.*, **53**, 354.
6. Karollus,A., Mauermeier,T. and Gagneur,J. (2023) Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biol.*, **24**, 56.
7. Avsec,Ž., Agarwal,V., Visentin,D., Ledsam,J.R., Grabska-Barwinska,A., Taylor,K.R., Assael,Y., Jumper,J., Kohli,P. and Kelley,D.R. (2021) Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods*, **18**, 1196–1203.
8. Novakovsky,G., Dexter,N., Libbrecht,M.W., Wasserman,W.W. and Mostafavi,S. (2022) Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nat. Rev. Genet.*, **24**, 125–137.
9. Klar,A.J.S. and Halvorson,H.O. (1974) Studies on the positive regulatory gene, GAL4, in regulation of galactose catabolic enzymes in Saccharomyces cerevisiae. *Mol. Gen. Genet.*, **135**, 203–212.
10. Shea,M.A. and Ackers,G.K. (1985) The OR control system of bacteriophage lambda: a physical-chemical model for gene regulation. *J. Mol. Biol.*, **181**, 211–230.
11. Spivak,A.T. and Stormo,G.D. (2012) ScerTF: a comprehensive database of benchmarked position weight matrices for *Saccharomyces* species. *Nucleic Acids Res.*, **40**, 161–168.
12. Stormo,G.D. and Fields,D.S. (1998) Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.*, **23**, 109–113.
13. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.
14. Zhu,C., Byers,K.J.R.P., McCord,R.P., Shi,Z., Berger,M.F., Newburger,D.E., Saulrieta,K., Smith,Z., Shah,M.V., Radhakrishnan,M. *et al.* (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res.*, **19**, 556–566.
15. Giniger,E. and Ptashne,M. (1988) Cooperative DNA binding of the yeast transcriptional activator GAL4. *Proc. Natl. Acad. Sci. U.S.A.*, **85**, 382–386.
16. Gertz,J., Siggia,E.D. and Cohen,B.A. (2009) Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature*, **457**, 215–218.
17. Segal,E., Raveh-Sadka,T., Schroeder,M., Unnerstall,U. and Gaul,U. (2008) Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature*, **451**, 535–540.
18. Granek,J.A. and Clarke,N.D. (2005) Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biol.*, **6**, 87.
19. Giaever,G. and Nislow,C. (2014) The Yeast Deletion Collection: a Decade of Functional Genomics. *Genetics*, **197**, 451.
20. Liu,J., Shively,C.A. and Mitra,R.D. (2020) Quantitative analysis of transcription factor binding and expression using calling cards reporter arrays. *Nucleic Acids Res.*, **48**, 50.
21. Sharon,E., Kalma,Y., Sharp,A., Raveh-Sadka,T., Levo,M., Zeevi,D., Keren,L., Yakhini,Z., Weinberger,A. and Segal,E. (2012) Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.*, **30**, 521–530.
22. Wang,H., Johnston,M. and Mitra,R.D. (2007) Calling cards for DNA-binding proteins. *Genome Res.*, **17**, 1202–1209.
23. Wang,H., Mayhew,D., Chen,X., Johnston,M. and Mitra,R.D. (2011) Calling Cards enable multiplexed identification of the genomic targets of DNA-binding proteins. *Genome Res.*, **21**, 748–755.
24. Shively,C.A., Liu,J., Chen,X., Loell,K. and Mitra,R.D. (2019) Homotypic cooperativity and collective binding are determinants of bHLH specificity and function. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 16143–16152.
25. Raveh-Sadka,T., Levo,M. and Segal,E. (2009) Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome Res*, **19**, 1480.
26. Siggers,T., Duyzend,M.H., Reddy,J., Khan,S. and Bulyk,M.L. (2011) Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Mol. Syst. Biol.*, **7**, 555.
27. Dalal,C.K., Zuleta,I.A., Mitchell,K.F., Andes,D.R., El-Samad,H. and Johnson,A.D. (2016) Transcriptional rewiring over evolutionary timescales changes quantitative and qualitative properties of gene expression. *Elife*, **5**, 87.
28. Kang,Y., Patel,N.R., Shively,C., Recio,P.S., Chen,X., Wranik,B.J., Kim,G., Scott McIsaac,R., Mitra,R. and Brent,M.R. (2020) Dual threshold optimization and network inference reveal convergent evidence from TF binding locations and TF perturbation responses. *Genome Res.*, **30**, 459–471.
29. Gertz,J., Siggia,E.D. and Cohen,B.A. (2008) Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nat. 2008 4577226*, **457**, 215–218.
30. Berg,O.G. and von Hippel,P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193**, 723–743.

31. Liang,S.D., Marmorstein,R., Harrison,S.C. and Ptashne,M. (1996) DNA sequence preferences of GAL4 and PPR1: how a subset of Zn2 Cys6 binuclear cluster proteins recognizes DNA. *Mol. Cell. Biol.*, **16**, 3773–3780.
32. Moudgil,A., Wilkinson,M.N., Chen,X., He,J., Cammack,A.J., Vasek,M.J., Lagunas,T., Qi,Z., Lalli,M.A., Guo,C. *et al.* (2020) Self-Reporting transposons enable simultaneous readout of gene expression and transcription factor binding in single cells. *Cell*, **182**, 992–1008.
33. Marmorstein,R., Carey,M., Ptashne,M. and Harrison,S.C. (1992) DNA recognition by GAL4: structure of a protein-DNA complex. *Nature*, **356**, 408–414.
34. Reece,R.J. and Ptashne,M.Determinants of binding-site specificity among yeast C$_6$ zinc cluster proteins. *Science*, **261**, 909–911.
35. Wu,Y., Reecel,R.J., Ptashne2,M., Wu,Y. and Reece,R.J. (1996) Quantitation of putative activator-target affinities predicts transcriptional activating potentials. *EMBO J.*, **15**, 3951–3963.
36. Hanley,J.A. and McNeil,B.J. (1983) A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, **148**, 839–843.
37. Mukherjee,S., Berger,M.F., Jona,G., Wang,X.S., Muzzey,D., Snyder,M., Young,R.A. and Bulyk,M.L. (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.*, **36**, 1331–1339.
38. Penvose,A., Keenan,J.L., Bray,D., Ramlall,V. and Siggers,T. (2019) Comprehensive study of nuclear receptor DNA binding provides a revised framework for understanding receptor specificity. *Nat. Commun.*, **10**, 2514.
39. Kinney,J.B., Murugan,A., Callan,C.G. and Cox,E.C. (2010) Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 9158–9163.
40. Ireland,W.T., Beeler,S.M., Flores-Bautista,E., McCarty,N.S., Röschinger,T., Belliveau,N.M., Sweredoski,M.J., Moradian,A., Kinney,J.B. and Phillips,R. (2020) Deciphering the regulatory genome of *Escherichia coli*, one hundred promoters at a time. *Elife*, **9**, e55308.
41. Liang,S.D., Marmorstein,R., Harrison,S.C. and Ptashne,M. (1996) DNA sequence preferences of GAL4 and PPR1: how a subset of Zn2 Cys6 binuclear cluster proteins recognizes DNA. *Mol. Cell. Biol.*, **16**, 3773–3780.
42. Coey,C.T. and Clark,D.J. (2021) A systematic genome-wide account of binding sites for the model transcription factor Gcn4. *Genome Res.*, **32**, 367–377.
43. Sandelin,A. and Wasserman,W.W. (2005) Prediction of nuclear hormone receptor response elements. *Mol. Endocrinol.*, **19**, 595–606.
44. Ma,B., Pan,Y., Zheng,J., Levine,A.J. and Nussinov,R. (2007) Sequence analysis of p53 response-elements suggests multiple binding modes of the p53 tetramer to DNA targets. *Nucleic Acids Res.*, **35**, 2986–3001.
45. L Mariani,K.W.A.V.L.B.M.B. (2017) Identification of human lineage-specific transcriptional coregulators enabled by a glossary of binding modules and tunable genomic backgrounds. *Cell Syst.*, **5**, 187–201.
46. Yella,V.R., Bhimsaria,D., Ghoshdastidar,D., Rodríguez-Martínez,J.A., Ansari,A.Z. and Bansal,M. (2018) Flexibility and structure of flanking DNA impact transcription factor affinity for its core motif. *Nucleic Acids Res.*, **46**, 11883.
47. Dror,I., Rohs,R. and Mandel-Gutfreund,Y. (2016) How motif environment influences transcription factor search dynamics: finding a needle in a haystack. *Bioessays*, **38**, 605.
48. Gordân,R., Shen,N., Dror,I., Zhou,T., Horton,J., Rohs,R. and Bulyk,M.L. (2013) Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep.*, **3**, 1093–1104.
49. MacPherson,S., Larochelle,M. and Turcotte,B. (2006) A fungal family of transcriptional regulators: the zinc cluster proteins. *Microbiol. Mol. Biol. Rev.*, **70**, 583–604.
50. Brodsky,S., Jana,T., Mittelman,K., Chapal,M., Kumar,D.K., Carmi,M. and Barkai,N. (2020) Intrinsically disordered regions direct transcription factor in vivo binding specificity. *Mol. Cell*, **79**, 459–471.
51. Gera,T., Jonas,F., More,R. and Barkai,N. (2022) Evolution of binding preferences among whole-genome duplicated transcription factors. *Elife*, **11**, 73225.