Summer 8-15-2021

Regulation of transcription factor binding specificity: from binding motifs to local DNA context

Jiayue Liu
Washington University in St. Louis

Recommended Citation

Liu, Jiayue, "Regulation of transcription factor binding specificity: from binding motifs to local DNA context" (2021). *Arts & Sciences Electronic Theses and Dissertations*. 2508.
https://openscholarship.wustl.edu/art_sci_etds/2508

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences

Computational and Systems Biology

Dissertation Examination Committee:
Robi Mitra, Chair
Douglas Chalker
Barak Cohen
Alex Holehouse
Gary Stormo

Regulation of Transcription Factor Binding Specificity: from Binding motifs to DNA Context

by
Jiayue Liu

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2021
St. Louis, Missouri

Table of Contents

List of Figures

List of Abbreviations

ANOVA	Analysis of Variance
AUC	Area Under the Curve
bHLH	Basic Helix Loop Helix
CCRA	Calling Cards Reporter Array
ChIP	Chromatin Immunoprecipitation
CSI	Cognate Site Identification
EMSA	Electrophoretic Mobility Shift Assay
FD	Facilitated Diffusion
LDC	Local DNA Context
LR	Logistical Regression
MITOMI	Mechanically Induced Trapping of Molecular Interaction
MPRA	Massively Parallel Reporter Assay
NBS	Normalized Binding Score
NuOc	Nucleosome Occupancy
PBM	Protein Binding Microarray
PRC	Precision Recall Curve
PWM	Position Weight Matrix
ROC	Receiver Operator Curve
SELEX	Systematic Evolution of Ligands by Exponential Enrichment
SPR	Surface Plasmon Resonance

TF	Transcription Factor
TFBS	Transcription Factor Binding Site
UMI	Unique Molecular Identifier
YFP	Yellow Fluorescence Protein

Acknowledgments

I would like to thank my mentor, Rob Mitra, for his kind support and encouragement when I am frustrated and his critical and valuable suggestions when I need improvement. Throughout my time in graduate school, Rob helped me a lot in conducting science and communicating science. This work would not be done without his guidance. I would like to thank my colleague Christian Shively; the discussion and collaboration with him always bring me more energy and ideas. I also want to thank my committee, Barak Cohen, Gary Stormo, Douglas Chalker and Alex Holehouse, for their insightful comments and suggestions for my thesis and always kind and positive spirit. I want to thank all members in Mitra lab for their warm support and interesting discussion. I want to thank CGS center for their ultra-fast and reliable sequencing service; the work would take much longer if not with their diligence. I want to thank my husband, Weijian Wang, for helping me with many coding related questions. Lastly, I would like to thank every staff in Washington University in Saint Louis, Division of Biology and Biochemical Sciences and Computational and Systems Biology for the friendly and efficient working environment.

Jiayue Liu

Washington University in St. Louis

August 2021

ABSTRACT OF THE DISSERTATION

Regulation of Transcription Factor Binding Specificity: from Binding Motifs to DNA Context

by

Jiayue Liu

Doctor of Philosophy in Biology and Biomedical Sciences

Computational and Systems Biology

Washington University in St. Louis, 2021

Robi Mitra, Chair

Regulation of transcription factor (TF) binding specificity lies at the heart of transcriptional control which governs how cells divide, differentiate, and respond to their environments. TFs are known to bind to DNA in a sequence specific manner, and such short sequence is known as transcription factor binding site (TFBS). However, the *in vivo* TF bound regions do not always contain a TFBS, and additionally, there are often excessive non-functional TFBSs with binding potential in the regulatory regions that are unbound for a given TF. This dissertation focuses on understanding the principles of TF binding specificity and is divided into two chapters: 1) developing a novel high throughput method that would facilitate the study of TF binding regulations and the resulting functional output; 2) analyzing the roles of local DNA context around TFBS in specifying TF localization.

In the first chapter of this dissertation, we report a tool, Calling Cards Reporter Arrays (CCRA), that measures transcription factor (TF) binding and the consequences on gene expression for hundreds of synthetic promoters in yeast. Using Cbf1p and MAX, we demonstrate that the CCRA method is able to detect small changes in binding free energy with a sensitivity comparable to *in vitro* methods, enabling the measurement of energy landscapes *in vivo*. We then demonstrate the quantitative analysis of cooperative interactions by measuring Cbf1p binding at synthetic promoters with multiple sites. We find that the cooperativity between Cbf1p dimers varies sinusoidally with a period of 10.65 bp and energetic cost of 1.37 K_BT for sites that are positioned “out of phase”. Finally, we characterize the binding and expression of a group of TFs, Tye7p, Gcr1p, and Gcr2p, that act together as a “TF collective”, an important but poorly characterized model of TF cooperativity. We demonstrate that Tye7p often binds promoters without its recognition site because it is recruited by other collective members, whereas these other members require their recognition sites, suggesting a hierarchy where these factors recruit Tye7p but not vice versa. Our experiments establish CCRA as a useful tool for quantitative investigations into TF binding and function.

In the second chapter of this dissertation, we seek out to investigate if predictive information is embedded in local DNA context (LDC) on a large collection of TFs in Saccharomyces cerevisiae. We identify there is a general preference for TFs to bind at CG rich sequences; we then analyze whether such preference is linked to intrinsic nucleosome binding preference and found the CG preference in LDC for TF binding was independent of nucleosome regulation. We next examine the possible mechanism by which LDC influence TFs binding site selection, through recruiting ‘licensing’ factors or kinetically assisting TF search for a target site. We show high CG LDC is preferred by TFs *in vitro* condition, which suggests such preference only involves TFs and DNA and directs us to TF search kinetics mechanism. CG rich feature in LDC may act as an energetical funnel to facilitate TF recognizing a target binding site, and we verify the theoretical validity of this hypothesis with Gillespie simulation. In the end, we reveal CG preference was also present in a large group of human TFs, indicating the usage of LDC is a general mechanism for TF binding specificity.

Chapter 1. Quantitative Analysis of Transcription Factor Binding and Expression Using Calling Cards Reporter Arrays

1.1 Introduction

Transcription factors (TFs) recognize and bind to specific sequences in regulatory DNA, called TF binding sites (TFBSs), and these events ultimately define the transcriptional programs that cells execute as they proliferate, develop, and respond to their environments (Accili & Arden, 2004; Simon, 2001; Vaquerizas, 2009). The principles that govern how TFs select functional binding sites *in vivo* are not well understood. For example, the *in vivo* occupancies of TFs cannot be predicted solely from their DNA binding preferences measured *in vitro*. Many TFs bind to only a small fraction of high-scoring TFBS in the genome, and, conversely, TF binding is often observed at loci without a nearby TFBS (Inukai, 2017; Villa, 2016; Yang, 1995). Explaining the binding of paralogous TFs is a related outstanding problem, as such factors often have nearly identical *in vitro* DNA binding preferences but regulate diverse sets of target genes and perform different cellular functions, even when expressed at the same time and in the same cell (Meyer, 2008; Dang, 2012; Shen, 2018). Finally, the relationship between TF binding and the resulting transcriptional consequences is also unclear, as it is difficult to predict whether a TF binding event will have any effect on the expression of a nearby gene or the directionality of such a change. Part of the reason for these difficulties is that TFs appear to act in a highly complex manner. Many TFs bind cooperatively (De Val, 2008; Fong, 2015; Frey, 2016; Hollenhorst, 2009; Zhou X., 2011; Wu, 1996), and we are far from having a complete description of which TFs interact with one another, or how they select their binding sites when they do interact. Even TFs that bind DNA independently may recruit transcriptional machinery in a combinatorial fashion after they bind to influence gene expression (Ong, 2011). Therefore, we need new experimental tools to study gene regulation that are quantitative, allow for the rapid analysis of many user-specified regulatory sequences, and can be easily multiplexed to study a number of different TFs.

High throughput methods such as Sort-Seq (Kinney, 2010; Sharon, 2012) and Massively Parallel Reporter Assays (MPRAs) (Maricque, 2017; White M. A., 2013) have emerged as important tools for investigations into the regulatory code, but these methods measure gene expression only, making it difficult to directly study the impact of TF binding on transcriptional regulation. Recent studies have performed ChIP-based binding measurements on libraries of promoter elements (Grossman, 2017; Zeigler, 2014); however, these studies were unable to quantitatively measure binding energies or analyze cooperative interactions, features which are critical for dissecting TF function. To study the complex nature of TF binding in a quantitative manner and correlate this binding with gene expression, we have developed Calling Cards Reporter Arrays (CCRA), a novel tool that builds on the previously reported Calling Card method (Wang H. J., 2007; Wang H. M., 2011; Shively, 2019). CCRA measures TF binding and the transcriptional consequences of this binding for hundreds of synthetic DNA sequences in the yeast, *Saccharomyces cerevisiae*. We first demonstrate that CCRA measures TF binding at synthetic promoters and gene expression from a downstream reporter in a sensitive, accurate, and reproducible manner. We then apply CCRA to study TF-DNA interactions and show that the

CCRA method is able to detect single nucleotide difference in the free energy of binding with a sensitivity that is comparable to *in vitro* methods. We then use CCRA to study how cooperativity dictates TFs binding *in vivo,* by analyzing the binding of the bHLH factor Cbf1p. We find that the cooperativity between Cbf1p dimers varies sinusoidally as the distance between two Cbf1p binding sites is changed, with an observed period of 10.65 base pairs. The helical phase of binding sites plays a major role in the cooperative binding of this factor, as “out of phase” sites incur an energetic cost of 3.40 kJ/mol (1.37 K_BT) relative to in-phase sites. Finally, we characterize the binding of a group of TFs that are thought to act together as a “TF collective”, a recently proposed model of cooperative binding (Junion, 2012; Spitz, 2012). Consistent with previous work (Shively, 2019), we find that one member of the group, Tye7p, is able to bind at promoters that do not encode its recognition sequence. Surprisingly however, the binding of other collective members, Gcr1p and Gcr2p, requires only their recognition sites, suggesting a hierarchy where these factors can recruit Tye7p but not vice versa. We further demonstrate that the expression of a reporter gene regulated by this collective can be best explained by considering the occupancy of all members of this complex. Together, these results establish CCRA as a useful tool for quantitative investigations into TF binding and function.

1.2 Results

1.2.1 Overview of Calling Cards Reporter Arrays (CCRA)

The CCRA method is designed to measure both TF binding and gene expression in parallel for hundreds of uniquely barcoded synthetic promoter sequences. To perform CCRA, the TF of interest is C-terminally fused to a short protein tag, so that the TF directs insertion of Ty5 retrotransposons (or “calling cards”) (Wang H. M., 2011; Wang H. J., 2007) near its binding sites (**Fig 1.1 upper and bottom panel**). For each CCRA assay, TF-directed insertions into the designed promoter library are recovered from yeast cells and the insertion locations and promoter sequence identities are determined via second-generation sequencing **(Fig 1.1 bottom panel)**. Each plasmid molecule in a CCRA library has a “library barcode” corresponding to a unique promoter sequence (**Fig 1.1 upper panel**), as well as a unique molecular identifier (UMI). The library barcode allows each transposon calling card to be assigned to the correct synthetic promoter sequence, and the UMI enables us to determine when multiple transposition have inserted into the same location in distinct copies of the same synthetic promoter sequence. By determining the number of independent transpositions inserted into each synthetic promoter and then normalizing by the promoter’s abundance in the library, we generate a normalized binding score (NBS), which is a quantitative measure of TF binding **(Fig 1.1 bottom panel)**.

Because the CCRA library is cloned upstream of a yellow fluorescence protein (YFP) reporter gene, it is also possible to measure the transcriptional output of each synthetic promoter in the library using Sort-Seq (Kinney, 2010; Sharon, 2012) (**Fig 1.1 middle panel**). To do so, the CCRA library is sorted by flow cytometry into subpopulations according to the ratio of YFP fluorescence to mCherry fluorescence. The mCherry gene is regulated by a constitutive promoter, allowing for normalization of the YFP signal to account for variation due to plasmid copy number, cell size, and other sources of extrinsic expression noise. Next, the sorted subpopulations of yeast cells are sequenced to quantify the abundance of each barcoded sequence in each subpopulation. Relative expression is then calculated by the proportion of each sequence in every binned library as per the standard Sort-Seq protocol (Kinney, 2010; Sharon, 2012). By combining aspects of both Calling Cards assay and Sort-Seq, CCRA allows us to quantitatively measure the binding of a TF to a library of regulatory sequences, and simultaneously measure the effect of that binding on gene expression.

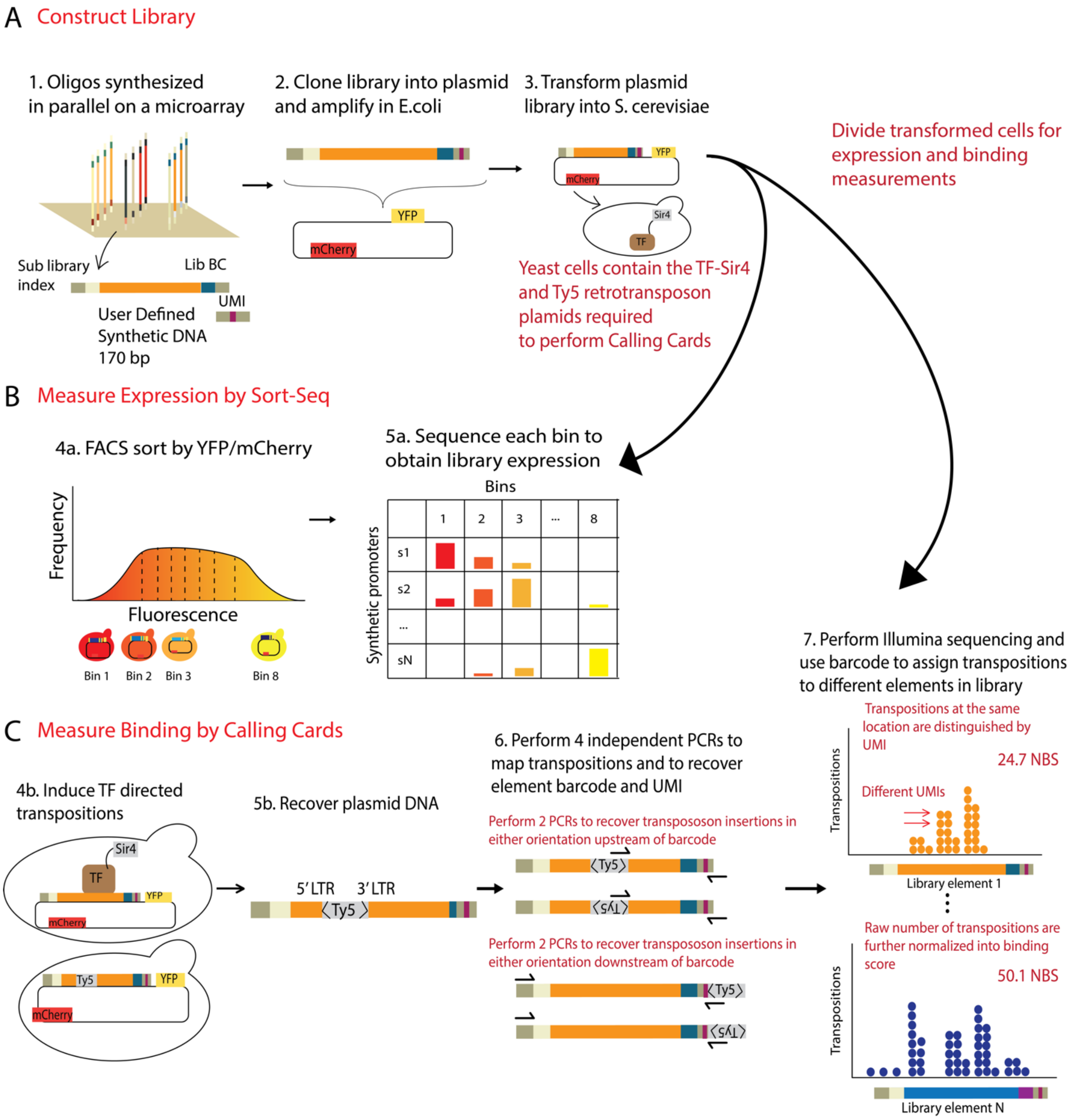

A Construct Library
1. Oligos synthesized in parallel on a microarray
2. Clone library into plasmid and amplify in E.coli
3. Transform plasmid library into S. cerevisiae
Divide transformed cells for expression and binding measurements
Sub library index
Lib BC
UMI
User Defined Synthetic DNA 170 bp
YFP
mCherry
Sir4
TF
Yeast cells contain the TF-Sir4 and Ty5 retrotransposon plamids required to perform Calling Cards
B Measure Expression by Sort-Seq
4a. FACS sort by YFP/mCherry
Frequency
Fluorescence
Bin 1
Bin 2
Bin 3
Bin 8
5a. Sequence each bin to obtain library expression
Bins
1
2
3
...
8
Synthetic promoters
s1
s2
...
sN
C Measure Binding by Calling Cards
4b. Induce TF directed transpositions
Sir4
TF
YFP
mCherry
Ty5
YFP
mCherry
5b. Recover plasmid DNA
5′ LTR
3′ LTR
Ty5
6. Perform 4 independent PCRs to map transpositions and to recover element barcode and UMI
Perform 2 PCRs to recover transpososon insertions in either orientation upstream of barcode
Ty5
Ty5
Perform 2 PCRs to recover transpososon insertions in either orientation downstream of barcode
Ty5
Ty5
7. Perform Illumina sequencing and use barcode to assign transpositions to different elements in library
Transpositions at the same location are distinguished by UMI
24.7 NBS
Different UMIs
Transpositions
Library element 1
Raw number of transpositions are further normalized into binding score
50.1 NBS
Transpositions
Library element N

Figure 1.1 Illustrations of CCRA experimental steps and binding results recovery. **a)** CCRA library sequences are synthesized on a microarray and cloned into plasmid and transformed into S. cerevisiae. The transformed cells are divided into two subpopulations for either binding measurements by Calling Cards method or expression measurements by Sort-Seq method. **b)** Because the promoter library is cloned upstream of YFP reporter gene, and the mCherry reporter is constantly expressed from the same vector for internal control, cells are sorted based on the ratio of YFP and mCherry fluorescence to estimate the relative strength of the library promoter sequences. **c)** Each element in the library is designed to contain a sub-library index that allows the user to assay a sub-population of the library, a unique barcode for identity, and a 4 bp randomized UMI to increase binding measurement capacity. TF-directed transpositions into barcoded library are fully recovered by four PCRs to account for insertions in either orientation and the relative position to barcode and UMI. PCR products are sequenced, and each library element is identified by barcode. Each dot represents a TF-directed transposition. The relative position of the insertion in the library sequence of each transposition is shown as X-axis. Multiple transpositions at the same position are distinguished by UMI. Raw number of transpositions are further normalized into a binding score (NBS) by correcting for the relative abundance of each element in the library as well as the total number of transpositions in one experiment to make accurate comparisons across experiments.

1.2.2 Binding and Expression Measurements are Sensitive, Accurate and Reproducible

To determine if CCRA can accurately and reproducibly measure TF binding in parallel, we first analyzed the binding of Cbf1p, a well-studied bHLH protein whose motif is strongly predictive of its *in vivo* binding pattern (Shively, 2019). To evaluate the sensitivity of the method for the detection of TF binding at weak sites, we created a library of 40 different sequences consisting of 10 synthetic promoters, each with 4 unique barcodes for replicates. Three of these sequences were taken from different endogenous yeast promoters previously shown to be bound by Cbf1p at a single recognition site (Shively, 2019). We also designed two synthetic promoters with nucleosome disfavoring sequences (Raveh-Sadka, 2012) that flanked a single Cbf1p consensus motif. As negative controls, we included five matched promoters with mutated Cbf1p binding sites. The binding of Cbf1p to a representative promoter, *OYE3/DAP1*, and its matched control is shown in **Fig 1.2**. Each symbol on the graph represents an independent calling card insertion. Cbf1p-directed transpositions appear to fit a Gaussian distribution centered at Cbf1p motif. Interestingly, the region directly over the motif contains few insertions, likely due to Cbf1p's footprint as binds to its recognition sequence. The wild-type *OYE3/DAP1* promoter is bound tightly by Cbf1p (70.1 NBS), but when the Cbf1p binding site is mutated, binding is greatly reduced (7.7 NBS, **Fig 1.2** bottom panel). Cbf1p's binding to all five pairs of promoters is summarized in **Fig 1.3**. In all instances, Cbf1p's binding was significantly stronger at promoters with intact Cbf1p sites than at the mutated promoters, demonstrating that the CCRA method can reliably detect TF binding even at relatively weak sites containing single motifs. It is interesting to note that although Cbf1p binding was significant at all five promoters with intact Cbf1p motifs, the binding was significantly stronger at the two promoters in which the Cbf1p binding sites were flanked by nucleosome disfavoring sequences.

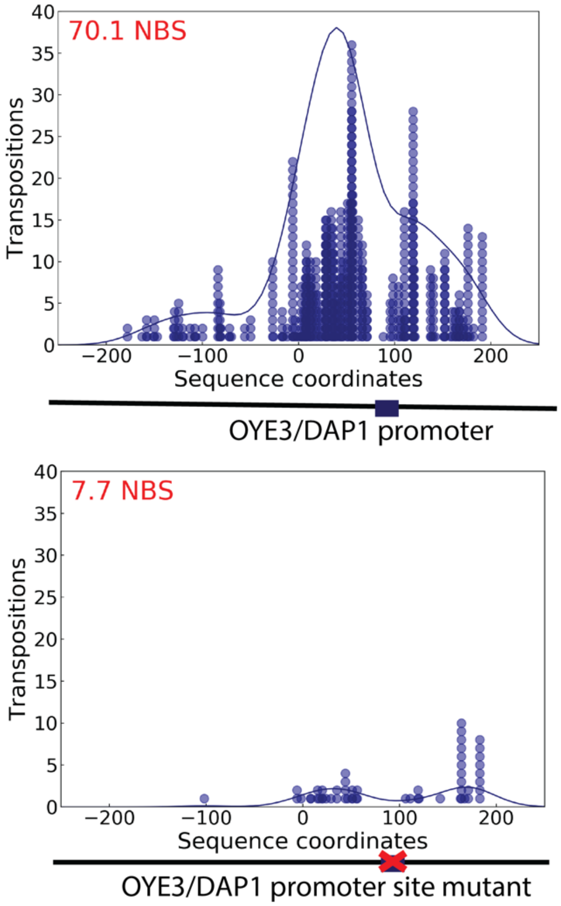


Figure 1.2 CCRA binding measurement on a Cbf1p target promoter. Cbf1p directed transpositions into the *OYE3_DAP1* intergenic region where only one E-Box motif is present. Each dot represents a unique TF-directed transposition along the sequence. The x-axis specifies the sequence coordinate to which a calling card insertion was mapped, whereas the y-axis specifies the number of independent insertions at each position. Transpositions at the same position are distinguished by a UMI. In general, transpositions follow a gaussian distribution center at the transcription factor binding site.

We next investigated the dynamic range of the CCRA assay. Since Cbf1p binding at regulatory elements is known to strongly depend on the number of Cbf1p sites present (Shively, 2019), we designed 183 synthetic promoters containing 0 to 6 sites and measured the binding of Cbf1p to this library. We observed a strong non-linear relationship between the normalized binding score (NBS), and the number of sites present in a given promoter (**Fig 1.4**). Importantly, we were able to measure Cbf1p binding across 3 orders of magnitude. These data demonstrate that CCRA technology can accurately measure TF binding across a large range of binding strengths.

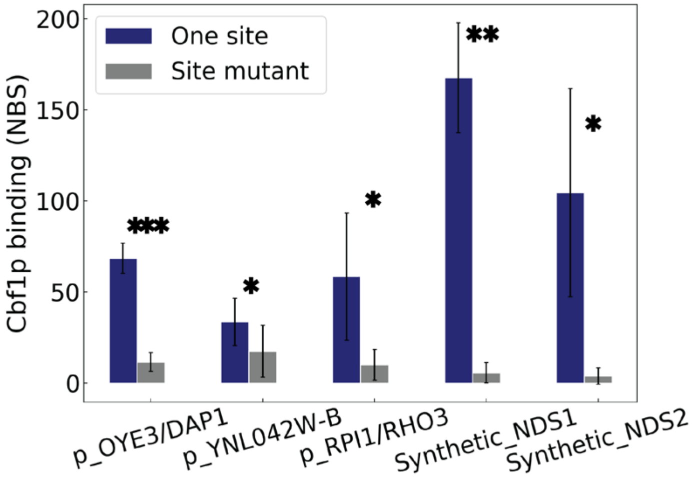


Figure 1.3 Binding measurement on five pairs of one motif containing promoters. Cbf1p binding measurements on three pairs of promoter regions and two pairs of synthetic sequences with one motif flanked by NDS. Blue bars represent sequences containing a motif, and gray bars represent the paired sequences with a mutated motif. The significance of binding detection on one motif is indicated by the number of stars. Three stars indicate a p-value of less than 0.0001 by paired t-test with four replicates, two stars, a p-value less than 0.001, and one star, a p-value less than 0.05.

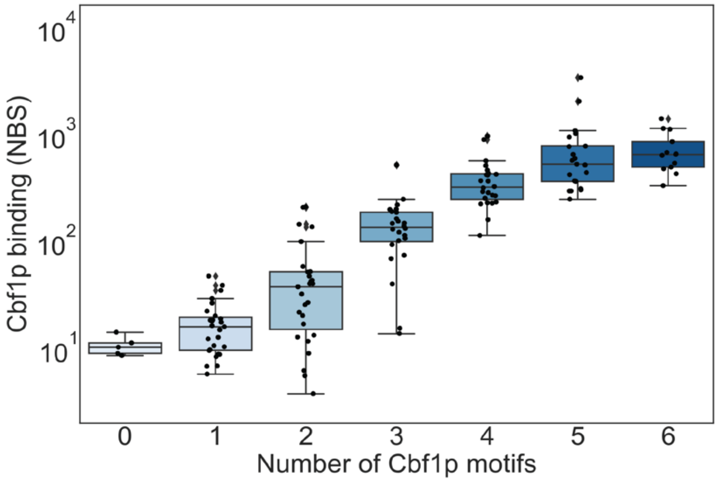


Figure 1.4 CCRA can measure TF binding with high dynamic range. Quantitative Cbf1p binding measurements on 183 sequences containing 0 to 6 motifs. Mean and standard deviation are indicated by the lines in each boxplot. The dynamic range spans over 3 orders of magnitudes.

Because the oligonucleotides used to create the synthetic promoters for CCRA are typically 170bp in length, we next sought to determine if TFs still bind *in vivo* with the same specificity as they do in their native genomic context. Therefore, we designed a 344-element library of genomic promoters derived from endogenous Gcn4p and Gal4p target promoters and used CCRA to measure the binding of these two TFs. We found that Gcn4p directed transpositions almost exclusively to synthetic promoters derived from Gcn4p targets whereas Gal4p directed transpositions to Gal4p targets (**Fig 1.5**), with each TF showing little non-specific binding to the other TF's set of target sequences. These results indicate that truncated genomic sequences in a plasmid-based system still retain their specificities and are not aberrantly bound by other TFs.

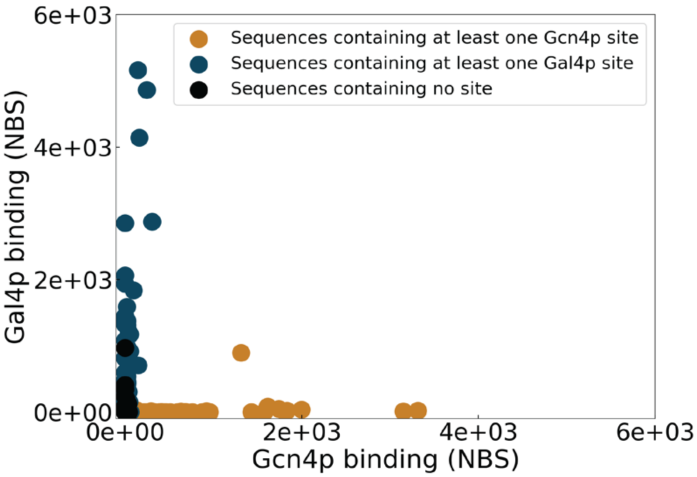


Figure 1.5 CCRAs measures binding with high specificity. Gcn4p and Gal4p were tested on a 344 elements library derived from Gcn4p or Gal4p naturally bound promoters. The library is categorized into three groups: sequences containing at least one Gcn4p site, sequences containing one Gal4p site and sequences containing no site. Most Gcn4p and Gal4p directed transpositions go to sequences containing at least one of either motif respectively, suggesting CCRA performs accurate binding measurement with little false positive.

Having established that the CCRA assay measures TF binding with high sensitivity and specificity, we next sought to benchmark the method's reproducibility. To do so, we performed replicate CCRA experiments using a 531-element synthetic promoter library and found that the NBS measured for each library member was highly reproducible (Pearson r = 0.92, p-value = 4.23e-216, Spearman r = 0.63, p-value = 3.21e-59. **Fig 1.6**).

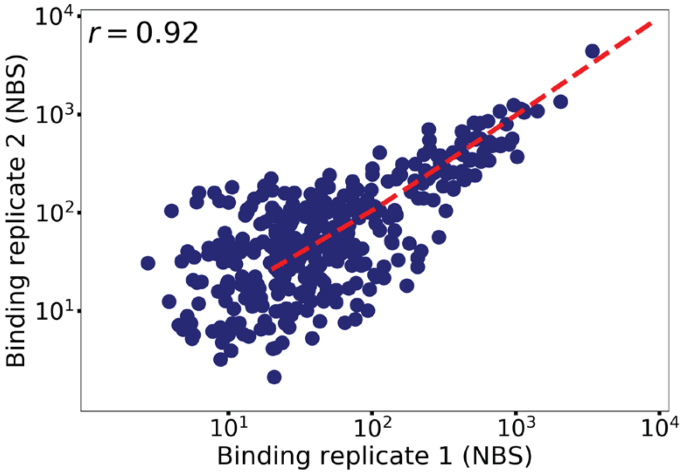


Figure 1.6 CCRA measures binding with high reproductivity. Showing binding reproducibility from two binding experiments with Cbf1p on 531-element library. Pearson r = 0.92, p-value = 4.23e-216; Spearman r = 0.63, p-value = 3.21e-59.

We next sought to establish that the CCRA method could accurately and reproducibly measure expression of the YFP reporter driven by a synthetic promoter library. To determine accuracy, we performed Sort-Seq to measure reporter expression for each member of a library containing sequences derived from Gcn4p and Gal4p promoters. We then cloned 24 of these library members and individually measured their expression levels by flow cytometry. We observed excellent agreement between the two measurements; the Pearson correlation coefficient was 0.95 (Pearson p-value = 6.59e-13, Spearman r = 0.96 and p-value = 7.91e-14), indicating that CCRA methodology accurately measures promoter activities from a library of synthetic sequences (**Fig 1.7**). To further investigate the accuracy of the method using a functional approach, we evaluated reporter expression as a function of the number of TF recognition sites for Gcn4p in an amino acid starvation growth condition and for Gal4p in galactose (Yan, 2018; Klar, 1974; Griggs, 1991; Hinnebusch A. G., 2002; Hinnebusch A. G., 1990). For both factors, reporter expression increased with the number of motifs, as expected from the known mechanism of action for these

TFs (**Appendix 1.2**). Finally, we also showed that expression measurements are highly reproducible between two biological replicates (Pearson r = 0.97 and p-value = 1.51e-228, Spearman r =0.95 and p-value = 1.29e-177 **Fig 1.8**).

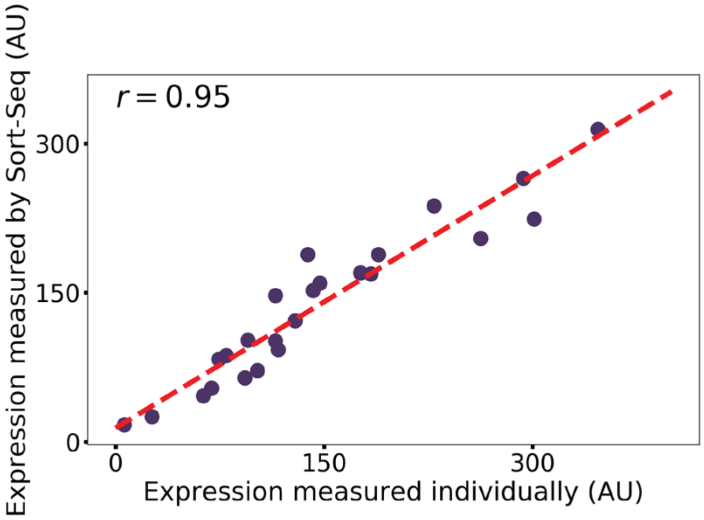


Figure 1.7 Expression measurement is highly accurate. 24 clones were measured by Flow cytometry individually and compared to the expression measured by Sort-Seq with Pearson correlation coefficient of 0.95. Pearson p-value = 6.59e-13; Spearman r = 0.96 and p-value = 7.91e-14.

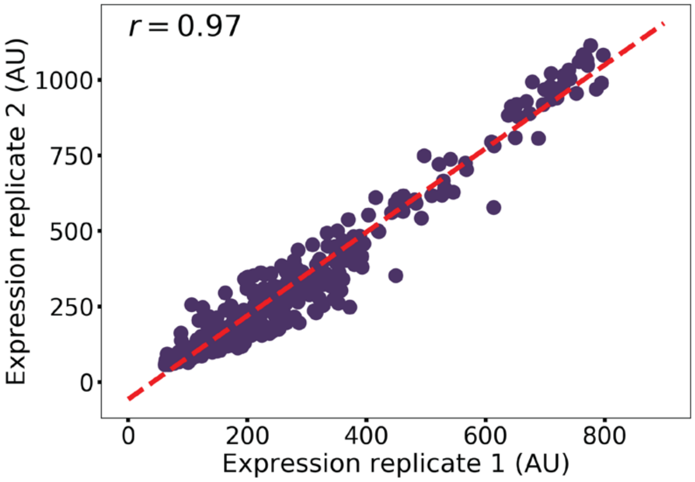


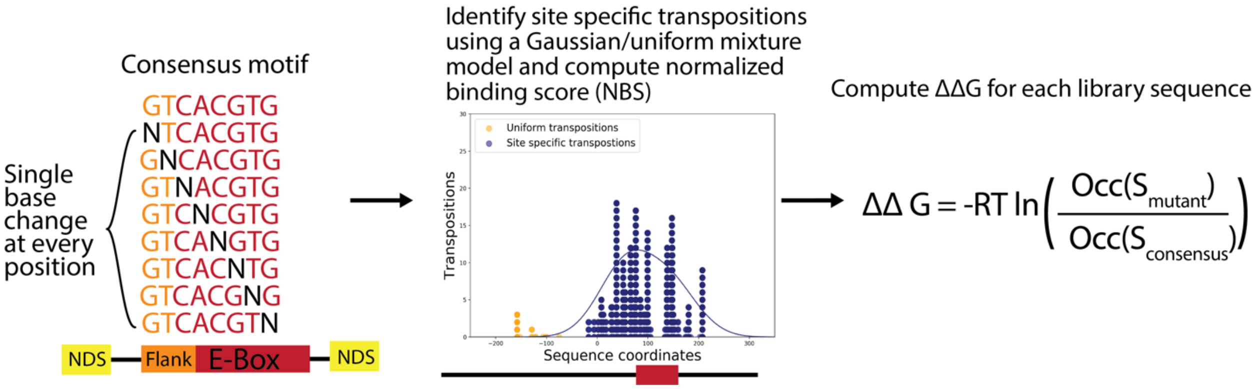
Figure 1.8 Expression measurement is reproducible. Showing expression reproducibility on a 344-element library derived from Gcn4p and Gal4p binding targets. Pearson r = 0.97 and p-value = 1.51e-228, Spearman r =0.95 and p-value = 1.29e-177.

The CCRA assay requires that the TF of interest be fused to a fragment of the Sir4p protein. This can be achieved by tagging the TF at its endogenous locus or by expressing the fusion from a plasmid, which is more convenient for many experiments. To investigate whether TF fusions expressed from plasmids binds to CCRA libraries in a similar manner as TF fusions expressed at their endogenous loci, we measured the binding for each using the same 531 synthetic promoter library and observed a high concordance (r=0.84, **Appendix 1.1**). We also confirmed that transcription factors tagged with the Sir4p fragment do not influence Sort-Seq expression measurements as they are highly correlated with measurements made using untagged proteins (r = 0.94 for Gal4p, r = 0.99 for Gcn4p, **Appendix 1.3 A, B**). Tagging TFs with Sir4p also does not appear to affect their functions (**Appendix 1.3 C-F**). Taken together, these results demonstrate that the CCRA method accurately and reproducibly measures the TF binding and expression consequences to a library of synthetic promoters.

1.2.3 Quantitative and High-throughput Measurement of the Binding Energy Landscapes of Transcription Factors *in Vivo*

Quantitative measurement of TF binding affinities to different DNA sequences is critical for understanding how TFs function *in vivo*. Because several studies have shown that minute variation in binding site affinity can specify alternative transcriptional or functional programs (Tanay, 2006; Bradley, 2010), it is important to be able to determine not only a TF's consensus binding sequence, but also its binding energy landscape (i.e. the TF's affinity for alternative binding sites). There are several methods that measure binding energy landscapes *in vitro*, such as MITOMI, PBM, Spec-seq, HT-SELEX, Bind-n-Seq, SPR, CSI and EMSA (Fordyce, 2010; Maerkl, 2007; Geertz, 2010; Stormo, Zuo, & Chang, 2015; Majka, 2007; Carlson CD, 2010;

Zhao, 2009; Garner, 1981), and these have proven invaluable for understanding TF-DNA interactions. However, there is currently no method to accurately discriminate the small changes in free energy needed to generate binding energy landscapes *in vivo*. Such landscapes may differ from those measured *in vitro* due to the effects of nucleosomes and other chromatin-associated proteins on DNA shape and binding site accessibility. Therefore, we sought to determine whether CCRA could measure binding energy landscapes *in vivo*.



Consensus motif
GTCACGTG
NTCACGTG
GNCACGTG
GTNACGTG
GTCNCGTG
GTCANGTG
GTCACNTG
GTCACGNG
GTCACGTN
Single base change at every position
NDS
Flank
E-Box
NDS
Identify site specific transpositions using a Gaussian/uniform mixture model and compute normalized binding score (NBS)
Uniform transpositions
Site specific transpostions
Transpositions
30
25
20
15
10
5
0
−200
−100
0
100
200
300
Sequence coordinates
Compute ΔΔG for each library sequence

$$\Delta\Delta\, G = -RT\ln\left(\frac{Occ(S_{mutant})}{Occ(S_{consensus})}\right)$$

Figure 1.9 Binding energy on alternative motif measurements scheme. A CCRA library was designed containing all possible alternative E-box motifs that are one base away from the consensus sequence and flanked with a nucleosome disfavoring site and analyzed for Cbf1p binding. Cbf1p directed transpositions were further processed using an expectation maximization algorithm. The change of free energy was then calculated using the binding occupancy of the alternative motif and the consensus.

We measured the binding of two basic helix loop helix (bHLH) factors, Cbf1p and MAX, to their consensus motifs and all sequences that differ by one base pair from the consensus (**Fig 1.9**). The TF binding sites were flanked by two intrinsic nucleosome disfavoring sequences to facilitate comparison to the *in vitro* binding landscapes previously determined (Raveh-Sadka, 2012). In order to accurately measure small changes in TF affinity, we used an expectation maximization algorithm to distinguish TF-directed transpositions from background insertions by assuming that

TF-directed transpositions follow a Gaussian distribution centered at the consensus motif whereas non-specific transpositions follow a uniform distribution across the full synthetic promoter (**see Methods**). Cbf1p and MAX occupancies at their consensus binding sites and at all possible one base substitution are shown in **Figure 1.10** and **1.12** respectively. As expected, both factors bound most strongly to their consensus sites. The changes in occupancies at non-consensus sites were strongly dependent on the position of the alteration and the identity of the substituted nucleotide. Some positions are crucial, such as the first position of core E-box motif, in the sense that any alternation resulted in completely abolished binding, whereas some positions such as flanking bases next to the core motif are more flexible when changed into other nucleotides. In general, Cbf1p binding appeared to be less tolerant to substitutions in its consensus motif than MAX, in agreement with previous *in vitro* measurements (Maerkl, 2007). We calculated the change of binding energy ($\Delta\Delta G$) from consensus site to the alternative site as follows (see Methods for a detailed derivation):

$$\Delta\Delta G = \Delta G(Sconsensus) - \Delta G(Smutant) = \text{-RTln}\left(\frac{Occ(Smutant)}{Occ(Sconsensus)}\right) \qquad (1.6)$$

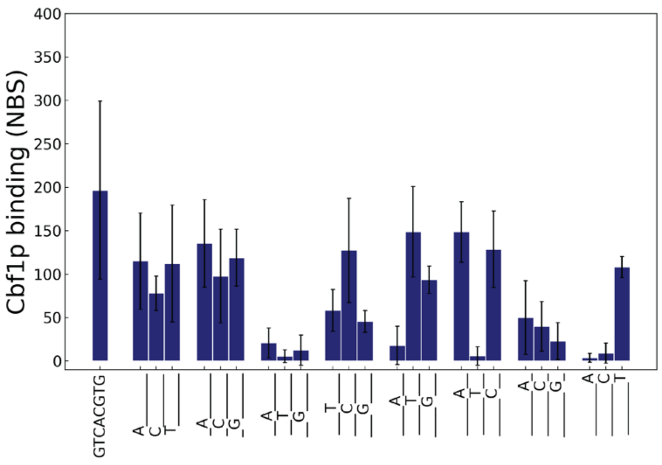


Figure 1.10 Cbf1p binding measurement on all alternative E-Box motif with four replicates. Standard deviation is indicated by the error bar.

To determine whether the measurements performed by CCRA are concordant with the binding energy landscapes of Cbf1p and MAX as measured by well-established *in vitro* methods, we compared our results to MITOMI and PBM (**Fig 1.11** and **Fig 1.13**). Both methods generated energy landscapes that were highly correlated to our CCRA measurements (For Cbf1p the correlation between CCRA and MITOMI: Pearson r of 0.75 and p-value = 1.90 e-5, Spearman r of 0.70 and p-value = 9.40e-5; the correlation between CCRA and PBM: Pearson r of 0.72 and p-value = 5.29e-5, Spearman r of 0.73 and p-value =3.97e-5. For MAX the correlation between CCRA and MITOMI: Pearson r of 0.72 and p-value = 1.42e-4, Spearman r of 0.74 and p-value = 9.06e-5; the correlation between CCRA and PBM: Pearson r of 0.80 and p-value =7.24e-6, Spearman r of 0.77 and p-value = 2.36e-5). Since the correlations between the measurements made by the two *in vitro* methods are similar in magnitude (For Cbf1p, the correlation between MITOMI and PBM: Pearson r = 0.73 and p-value of 3.35e-5, Spearman r = 0.78 and p-value = 4.57e-6. For MAX, the correlation between MITOMI and PBM: Pearson r of 0.79 and p-value = 1.28e-5, Spearman r = 0.79 and p-value = 1.25e-5 **Appendix 1.4**), these results demonstrate

CCRA measures binding energy landscapes *in vivo* with an accuracy comparable to *in vitro* methods.

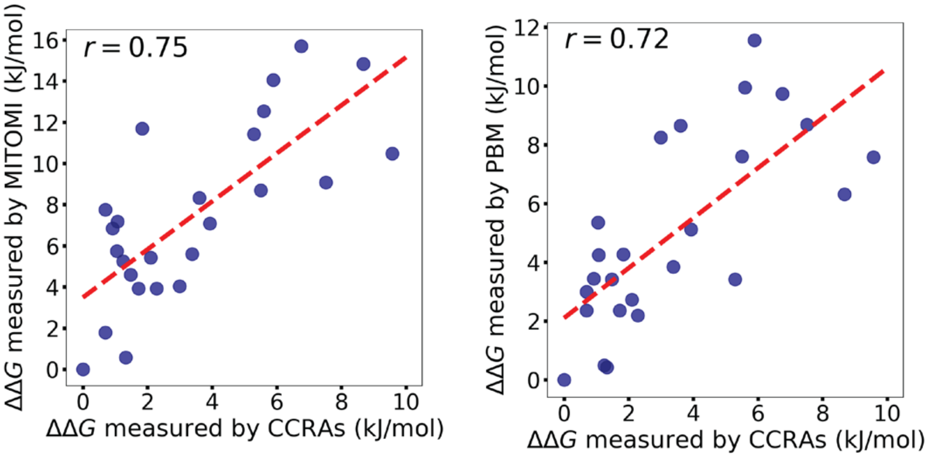


Figure 1.11 The measured change of free energy for each alternative TF motif for Cbf1p compared to the measurement by MITOMI and PBM. Pearson r of 0.75 and p-value = 1.90 e-5, Spearman r of 0.70 and p-value = 9.40e-5 for MITOMI comparison. Pearson r of 0.72 and p-value = 5.29e-5, Spearman r of 0.73 and p-value =3.97e-5 for PBM comparison.

The reported binding constant (K) for Cbf1p and MAX is (6.2 ± 1.4) × 10^7 M^{-1} at 20 °C (K_d = 1.6 nM) and (7.8 ± 2.6) ×10^6 M^{-1} (K_d = 130 nM) respectively (Park S. C., 2004; Kanaya, 1999), and therefore the binding energy ΔG for Cbf1p is about -45 kJ/mol (-18 K_BT) and -39 KJ/mol (-16 K_BT) for MAX. Given the largest $\Delta\Delta G$ calculated from the consensus to the mutant motif, Cbf1p loses $\frac{1}{4}$ of its binding energy with one nucleotide difference (e.g. $\Delta\Delta G$ is 9.6 KJ/mol from GTCACGTG to GTCACGT<u>A</u>) and therefore the K_d on the mutated motif GTCACGT<u>A</u> becomes 71 nM, a 40 fold increase relative to the consensus motif. MAX loses $\frac{1}{12}$ of its binding energy with one nucleotide difference *in vivo* (e.g. $\Delta\Delta G$ is 3.4 KJ/mol from CACGTG to CACTTG), and therefore the K_d on the mutant motif is 500 nM.

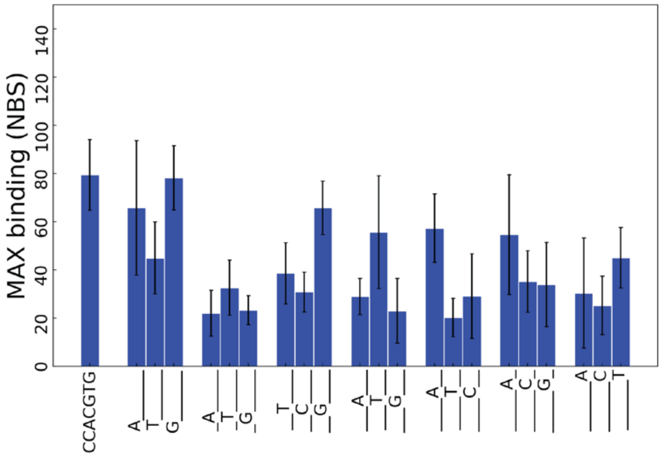


Figure 1.12 Cbf1p binding measurement on all alternative E-Box motif with four replicates. Standard deviation is indicated by the error bar.

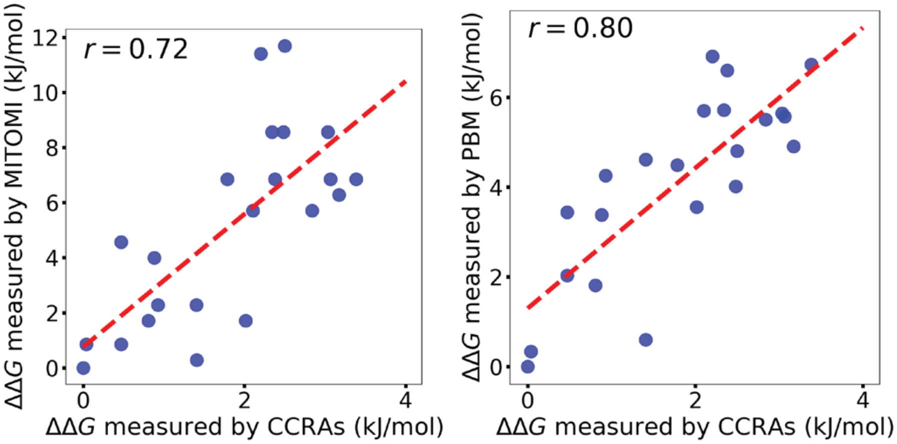


Figure 1.13 The measured change of free energy for each alternative TF motif for MAX compared to the measurement by MITOMI and PBM. Pearson r of 0.72 and p-value = 1.42e-4, Spearman r of 0.74 and p-value = 9.06e-5 for MITOMI comparison. Pearson r of 0.80 and p-value =7.24e-6, Spearman r of 0.77 and p-value = 2.36e-5 for PBM comparison.

1.2.4 Quantitative Measurement of the Cooperative Binding of Cbf1p

Understanding the mechanisms by which TFs select their targets *in vivo* will likely require more than just a characterization of their cognate DNA binding preferences, since it has been shown that many TFs achieve binding specificity through cooperative interactions with other DNA-binding proteins (De Val, 2008; Fong, 2015; Frey, 2016; Hollenhorst, 2009; Zhou X., 2011). Investigations into the cooperative interactions that occur between TFs are usually performed *in vitro*, under conditions that may not reflect the actual cellular environment (e.g. the lack of histones). *In vivo* investigations, which are less common, typically involve genome editing followed by quantitative binding measurement *in vivo,* which is experimentally challenging and time consuming (Shively, 2019; Kim, 2017; Wakabayashi, 2016). Given that CCRA is able to measure small changes in the free energy of TF binding, we sought to extend this approach to analyze TF-TF cooperativity. We focused on a pair of paralogous bHLH proteins, Cbf1p and Tye7p, both of which recognize the E-box motif CACGTG *in vitro* but bind to two distinct sets of target genes through different types of cooperative interactions.

We first set out to investigate Cbf1p, which has been shown to bind with homotypic cooperativity when two or more sites are present (Shively, 2019). This cooperativity was demonstrated by analyzing Cbf1p binding at mutated versions of the *IDH1_NCE103* divergent promoter, which normally contains three Cbf1p binding sites. This study showed that Cbf1p occupancy at the wild-type promoter was much stronger than the sum of the binding occupancies at three mutated promoters, each containing only a single Cbf1p binding site, demonstrating that Cbf1p binding is not additive but instead cooperative at this locus. However, in this study, Cbf1p's cooperativity was investigated at only a single promoter, so it is unclear to what extent this result can be generalized. We therefore sought to use CCRA to determine if this phenomenon occurs at other loci. We selected seven promoters with two or three Cbf1p sites including *IDH1_NCE103_pr* and designed a CCRA library in which these promoter sequences contained either zero, one, or two mutated Cbf1p sites. If Cbf1p binds cooperatively at these loci, we expect that, for each series of synthetic promoters, the sum of the binding scores from sequences with single Cbf1p sites will be significantly less than the binding at the "wild type" promoter sequence with multiple Cbf1p sites. In all seven cases, we found that Cbf1p binding at the wild type promoter was significantly higher than would be expected under an additive binding model (**Fig 1.14**), suggesting that Cbf1p binds cooperatively at all target promoters that contain multiple recognition sites.

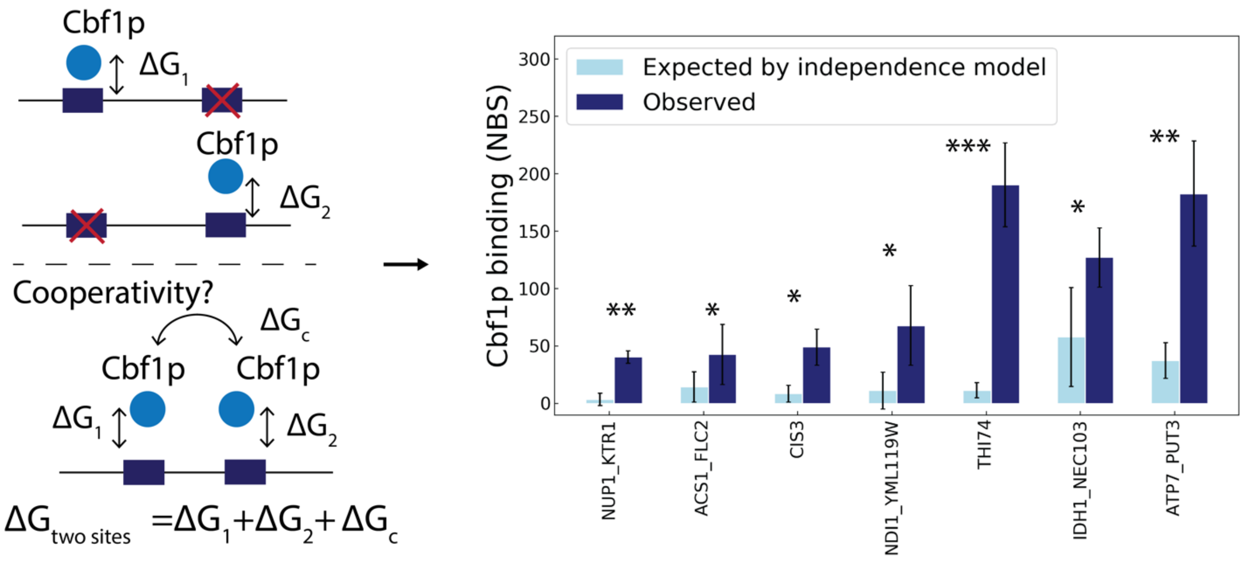


Figure 1.14 Experimental strategy to test if cooperativity exists between Cbf1p molecules when bind to sequences with multiple sites. Higher binding occupancy is expected than the sum of single site occupancy if cooperativity exists. Right Panel: Seven wild type promoters with either two or three motifs that are bound by Cbf1p were mutated such that only one motif was left. Binding on mutated sequences

was combined and then compared to the binding on the wild type sequence to verify the existence of cooperativity between two Cbf1p molecules. The light blue bar represents the sum of the binding from individual motif, and the dark blue bar represents the observed binding on the wild type sequence. Error bar is the standard deviation across three biological replicates. One star indicates p-value less than 0.05, two stars indicate p-value less than 0.01 and three stars indicate p-value less than 0.001 by T-test.

We next sought to characterize the relationship between the strength of Cbf1p cooperative binding and the distance between binding sites. Because the DNA double helix is thought to be rigid over length scale less than ~140 bp due to vertical base-stacking interactions and intra-helix phosphate charge repulsion (Mills, 2004; Wang J. C., 1979), one might expect that Cbf1p dimers would be unable to bind cooperatively at promoters with two recognition sites in close proximity. However, Cbf1p has been shown to sharply bend DNA upon binding (Palmieri, 1999; Shultzaberger, 2007; Harteis, 2014), and, furthermore, DNA is clearly malleable to some proteins, as it is tightly wrapped around nucleosomes and can be twisted and untwisted during replication and transcription (Allemand, 1998; Dickerson, 1989; Ussery, 2002). To investigate the relationship between Cbf1p cooperativity and the distance between recognition sites, we designed synthetic promoters where we varied the distance between two Cbf1p consensus motifs from 9 base pairs (bp) to 41 base pairs with two bp intervals. We used CCRA to measure Cbf1p binding on these synthetic sequences and plotted binding occupancy as a function of the distance between two sites. We found that the strength of Cbf1p binding at these synthetic promoters varied periodically with the distance between the binding sites (**Fig 1.15**). We observed strong binding at the shortest distance of 11 bp, and we observed additional peaks at 22 bp, 32 bp and 41bp apart. These distances are all shorter than the persistence length of DNA, and at the longest distance investigated, 41bp, the binding sites are separated by more than 65 Å, so it seems unlikely that the interaction between Cbf1p dimers could be explained by protein domain flexibility. Therefore, these results suggest that Cbf1p's ability to bend DNA allows the two dimers to interact with one another. We next hypothesized that the observed periodicity could be explained by the fact that Cbf1p makes its base pair contacts in the major groove of DNA so that at some motif distances, contact between Cbf1p dimers would require the rotation of the major groove around the axis of the double helix, incurring an energetic penalty. To test this, we fitted the binding to a cosine function. The calculated period was 10.65 bp, almost exactly the number of base pairs required for DNA to make one complete helical turn about its axis. We evaluated the fit of this model using Analysis of variance (ANOVA) and obtained a p-value 1.4e-6, indicating that the data follows the assumed model significantly better than expected by chance. This result suggested to us Cbf1p dimers that are not bound on the same side of the DNA helix must twist the DNA and incur an energetic cost. In contrast, two Cbf1p molecules on the same face of the helix are able to achieve the optimal cooperative binding efficiency. We next sought to compute the free energy cost associated with twisting the DNA double helix. Since we observed a 3.8-fold difference between the highest and the lowest occupancy, we calculated that the free energy lost due to twisting is 3.40 kJ/mol (1.37 K_BT). Compared to $\Delta\Delta Gs$ calculated for the consensus to mutant motif from the previous section, the energic cost of DNA twisting is comparably to a mild nucleotide change in the E-box motif (e.g. from GTCACGTG to GTC<u>T</u>CGTG). Interestingly, over the distance range examined in this experiment, the amplitude of the periodic function did not change appreciably, suggesting that, in contrast to twisting, Cbf1p bends DNA efficiently, with little energetic cost.

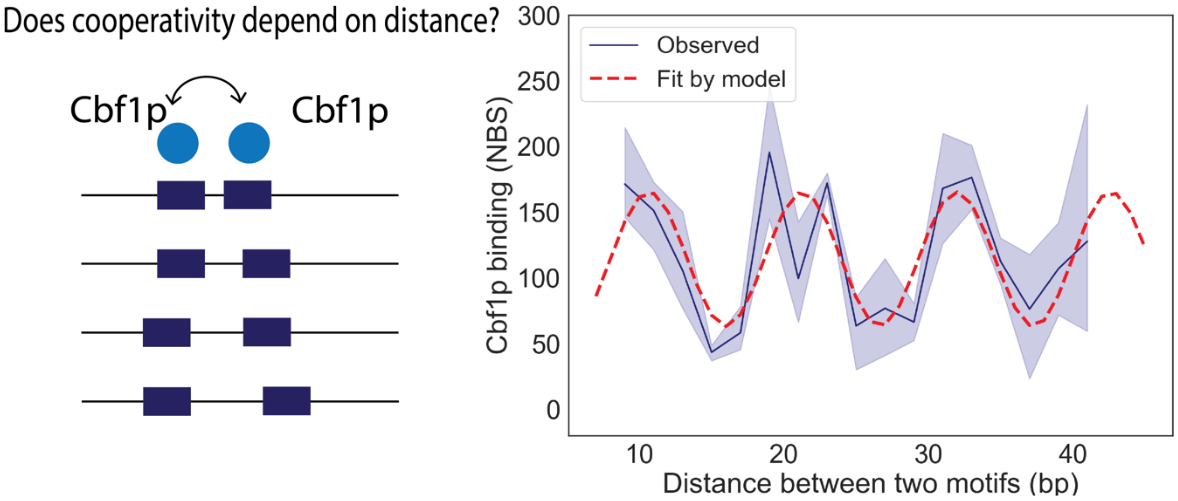


Figure 1.15 Cbf1p binding measured for two Cbf1p motifs were positioned from 9 bp to 41 bp apart in 2 bp intervals with four replicates. A trigonometric function model was used to fit the observed data, and the period obtained was 10.65 bp. An ANOVA test was performed to assess the learned parameter with a p-value of 1.4e-6.

We next asked if the phase of Cbf1p binding sites influenced the binding of this transcription factor at native genomic loci. We took published genome wide Cbf1p Calling Cards data (Shively, 2019) and grouped all intergenic regions with two Cbf1p binding sites within 100 bp according to the relative phase of the two sites. We found that promoters containing two Cbf1p binding sites separated by a multiple of 10.5 bp (i.e. with major grooves on the same side of the DNA helix) were bound significantly more tightly by Cbf1p than promoters with binding sites whose major grooves were on opposite sides of the DNA helix (**Figure 1.16** , p = 0.007). This result demonstrates that the periodicity in cooperative binding that we observed in our CCRA experiments also influences Cbf1p binding in the yeast genome.

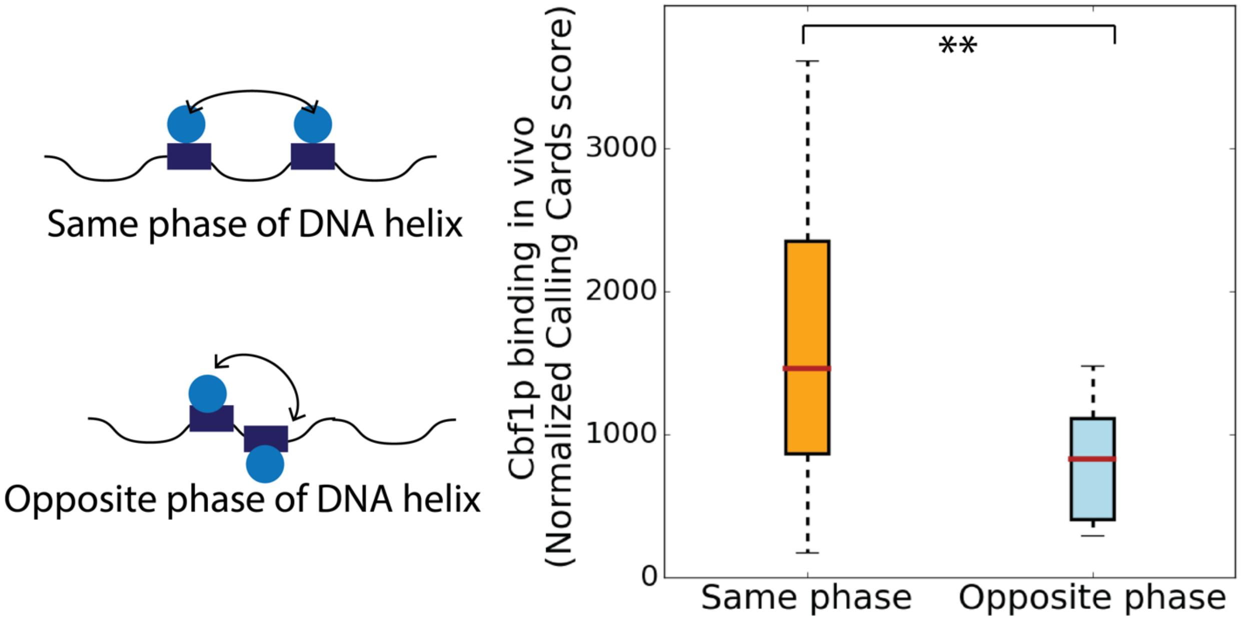


Figure 1.16 Cbf1p cooperativity on genomic regions. Genomic loci with two Cbf1p sites within 100 bp of each other were grouped according to whether they occur on the same side or opposite sides of the DNA helix (i.e. either separated by a multiple of 10.5 bp or by a multiple of 15.5bp). Genomic Calling Cards score was compared between two groups, and a T-test was performed with p-value of 0.007.

1.2.5 The Binding Logic of the Tye7p/Gcr1p/Gcr2p/Rap1p TF Collective

Unlike Cbf1p, many of the promoters bound by Tye7p do not encode an E-box, this factor's preferred binding motif (Shively, 2019). It has previously been shown that Tye7p binds cooperatively with the Gcr1p/Gcr2p/Rap1p complex and that by taking into account the DNA binding preferences of these proteins, the *in vivo* binding of Tye7p can be more accurately predicted (Shively, 2019). However, the biophysical principles that govern the binding of this complex are still unclear. For example, the binding of this complex does not appear to follow either of the two most well-studied models for TF binding, the Enhancesome model or the Billboard model (Panne, 2007; Kulkarni, 2003), because these models both posit a one-to-one correspondence between the binding of a TF and the presence of its recognition site. Instead, Tye7p binding appears to be consistent with the recently described TF collective model, in which a group of TFs bind together, but the motif positioning and composition at target sites is flexible (Junion, 2012; Spitz, 2012). However, the TF collective model is ambiguous with regard to the mechanistic details of binding, so important questions about the function of the Tye7p/Rap1p/Gcr1p/Gcr2p collective remain.

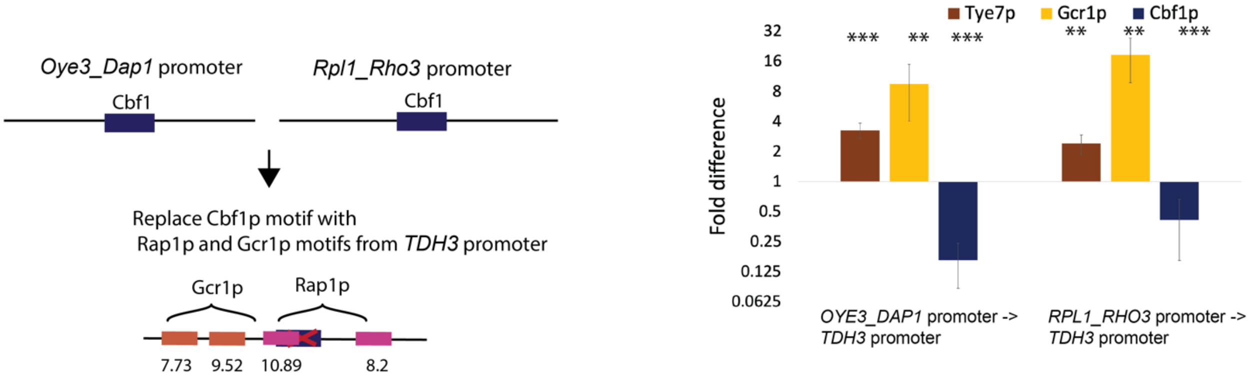


Figure 1.17 Tye7p is able to bind without its motif through protein-protein interactions with Gcr1/2p and Rap1p. To test if Tye7p is able to bind through other helpers, the Cbf1p motifs on the *Oye3_Dap1* and *Rpl1_Rho3* intergenic regions were mutated and two Gcr1p and two Rap1p motifs from *TDH3* promoter were added. Binding measurements were performed on the wild type and reprogrammed sequences for Tye7p, Gcr1p and Cbf1p. Tye7p bound to both reprogrammed promoters at significantly higher levels than the wild type *Oye3_Dap1* and *Rpl1_Rho3* sequences, as did Gcr1p. Cbf1p binding was abolished on these regions after mutation. T test was performed to assess the significance, and two stars indicate p-value less than 0.01 and three stars indicate p-value less than 0.001.

We first assessed the predictive power of the collective model by attempting to reprogram yeast promoters that normally bind Cbf1p, a Tye7p paralog, into promoters that bind Tye7p. To do so, we took two promoters, *OYE3_DAP1_pr* and *RPL1_RHO3_pr,* that are normally bound by Cbf1p, and removed their E-boxes (i.e. Cbf1p/Tye7p binding sites), and added Gcr1/2p and Rap1p sites with a design based on the *TDH3* promoter, which is bound by Tye7p. We then assessed the binding of Tye7p to these reprogrammed promoters using CCRA. Both showed significant decreases in Cbf1p binding (6.1-fold and 2.4-fold respectively) and significant increases in Tye7p (3.3-fold and 2.4-fold respectively) (**Fig 1.17**). We also observed an increase in Gcr1p binding at these reprogrammed promoters. Since neither of these reprogrammed promoters contain a consensus Tye7p binding site, we conclude that Tye7p binding is consistent with the collective model and that this TF can be recruited to promoters via cooperative interactions with Gcr1/2p and Rap1p.

Next, we wanted to better understand the molecular logic by which this collective binds. While Tye7p clearly does not require its motif to be present at a regulatory target, is this true for other members of the collective? When more than one binding site is present for a single TF, do the additional sites contribute to complex stability, or is one site sufficient and the others redundant? How is transcriptional output correlated with binding of each TF member? To answer these questions, we took a Tye7p bound promoter, *BMH1_pr*, which contains one Tye7p site, three Gcr1/2p sites and one Rap1p site, made every possible combination of mutated sites, and measured Tye7p binding using CCRA. Since Tye7 does not require its recognition sequence for binding, we first wanted to know if its motif made any energetic contribution to stabilize this factor. We divided the mutated sequences into two categories, those with and without a Tye7p motif. Sequences without a recognition site were still significantly bound by Tye7p (**Fig 1.18, middle group**), consistent with previous observations, but Tye7p binding at the wild-type *BMH1_pr* is reduced by 45% when the Tye7p recognition site is mutated (p-value = 0.012). Furthermore, when the 16 pairs of *BMH1_pr* mutants are compared across groups, we observe a significant reduction in Tye7p when the recognition motif is mutated (p-value =0.010). These results demonstrate that while the Tye7p motif is not required for Tye7p binding, it makes an energetic contribution when present. Notably, the positional distributions of Tye7p insertions across the *BMH1_pr* were essentially unaffected by the presence or absence of its cognate motif (**Appendix 1.6**), suggesting that the recruitment of Tye7p may be largely mediated by Gcr1/2p and Rap1p, even though the presence of a Tye7p binding site clearly makes an energetic contribution. Consistent with this hypothesis, we found that Tye7p binding is strongly dependent on Gcr1/2p and Rap1p sites (**Fig 1.18 middle group**). In general, we observed a gradual decrease in binding as more collective sites are mutated, and we did not observe large decrease in binding (>2 fold) upon the removal of any one site, suggesting that no single binding site is necessary for Tye7p binding at this promoter, but instead that all sites contribute to the binding affinity of this TF. Based on this observation, we reasoned that Tye7p binding might be predicted by the total free energy from all sites combined on a promoter. Therefore, we performed a regression analysis to understand how well the total sites information explains Tye7p binding (**Fig 1.18 right group)**. Given that PWM scores reflect the binding energy of TF to specific DNA sequences, we used the sum of PWM scores for all sites present on the promoters for the analysis and we found that the combined sites information correlates well with Tye7p binding (Pearson r = 0.69 and p-value = 1.07e-5, Spearman r = 0.63 and p-value = 1.04e-4).

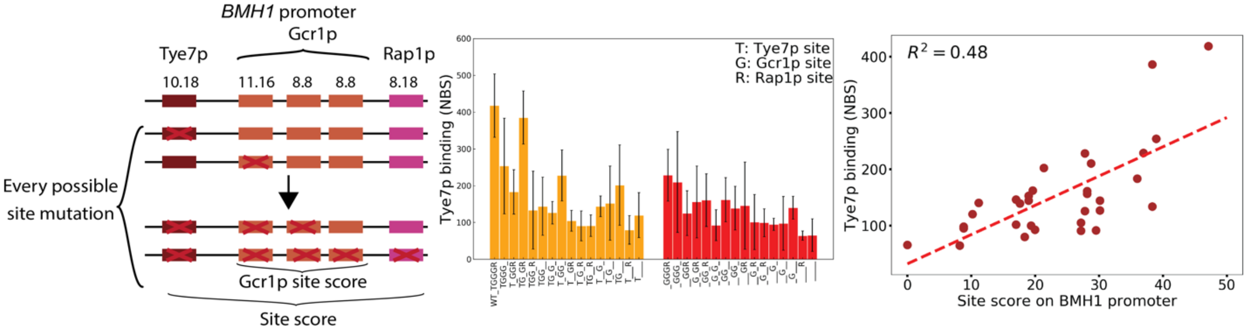


Figure 1.18 Tye7p binding measured at *BMH1* promoter. Left) The *BMH1* promoter, bound by Tye7p, contains one Tye7p motif, three Gcr1/2p motif and one Rap1p motif; a CCRA library was created in which all combinations of sites were mutated to create 32 sequences, including the wild-type sequence.

Middle) Tye7p binding was measured on these sequences and plotted. Intact sites are indicated as the x-axis label. All 32 sequences were classified into two sections, those with and without the Tye7p motif. Error bars represents the variation between four biological replicates. Right) The total binding free energy on each sequence based on the PWM score of the remaining sites was correlated with Tye7p binding result, and the total free energy of binding to DNA for the binding collective predicts Tye7p binding with $R^2 = 0.48$, Pearson r = 0.69 and p-value = 1.07e-5, Spearman r = 0.63 and p-value = 1.04e-4.

We then measured Gcr1p and Gcr2p occupancy on this promoter library. As before, we divided the mutated promoters into two categories based on whether they contained a Gcr1/2p motif. In contrast to what was observed for Tye7p, we found that neither Gcr1p nor Gcr2p was able to bind at any promoters without their shared recognition site (**Fig 1.19 & Appendix 1.7 A**), suggesting that these factors bind independently from the rest of the collective. To confirm this, we regressed Gcr1p and Gcr1p binding against the free energy of binding of Gcr1/2p or the full collective. We found that only Gcr1p/2p sites are required to explain Gcr1p and Gcr2p binding and that incorporating information from the other TF in the collective weakens the predictive power (**Fig 1.20 & Appendix 1.7 D for Gcr1p and Appendix 1.7 B & Appendix 1.7 C for Gcr2p**). Thus, the binding of the Gcr1/2p complex appears to be solely dependent on the presence and the number of Gcr1/2p sites. Furthermore, Gcr1/2p binding appears to saturate at two sites. Our Gcr2p binding measurements were more variable and weaker than our Gcr1p measurement, especially at sequences with only one Gcr1/2p motif, which might be due to the fact that Gcr2p is known to bind DNA indirectly through Gcr1p and depends on Gcr1p to function (Uemura, 1992; Baker, 1991).

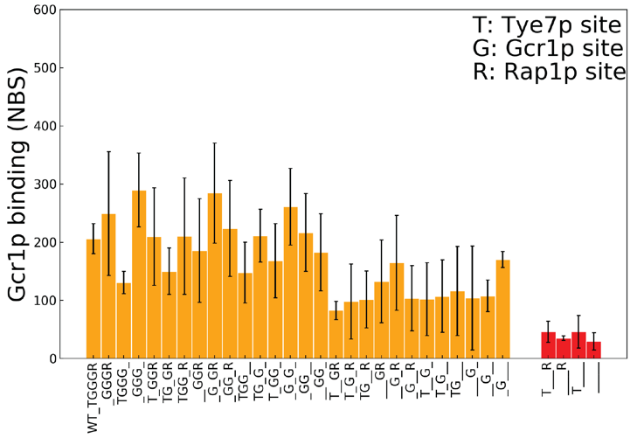


Figure 1.19 Gcr1p binding measured at *BMH1* promoter. The same as **Figure 1.18 (middle group)** but with Gcr1p, and these 32 sequences are classified into with and without any Gcr1/2p motif.

We next sought to investigate the relationship between the binding of the Tye7p collective and its transcriptional output. To do so, we performed Sort-Seq to measure the reporter gene expression from this library. We regressed reporter gene expression against the sum of the free energies of the binding sites (**Appendix 1.7 E**). We observed a good correlation, and we found that expression level correlated with the combined TF occupancy (**Fig 1.21**, Pearson r = 0.70 and p-value = 8.34e-6, Spearman r = 0.67 and p-value = 2.95e-9), suggesting that transcriptional output is determined by the whole complex. Similar analysis was done for *TDH3* promoter containing two Gcr1/2p sites and two Rap1p sites but no Tye7p site, and again the combined Tye7p, Gcr1p and Gcr2p occupancy correlated well with the expression (**Appendix 1.7 F & Appendix 1.7G**).

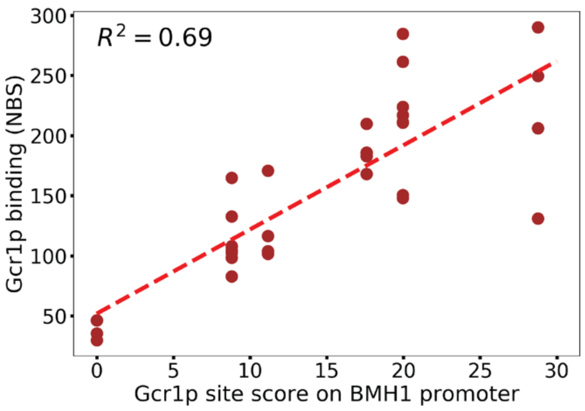


Figure 1.20 Correlation between Gcr1p binding and site score on *BMH1* promoter. PWM score of Gcr1/2p sites remained on the sequences was correlated with Gcr1p binding result, and Gcr1/2p sites alone predicts Gcr1p binding with R^2 of 0.69, Pearson r = 0.83 and p-value = 3.15e-9, Spearman r = 0.83 and p-value = 3.59e-9.

Rap1p binding was not measured in this study due to its inability to be tagged by Sir4p. However, Rap1p has been shown to interact with Gcr1p and Gcr2p as an activating complex (Menon, 2005; Tornow, 1993). With expression we measured on both *BMH1* and *TDH3* promoters, we compared sequence pairs that are with and without Rap1p site (**Appendix 1.7 H**). We performed a paired T-test on these sequence in terms of expression, and the p-value is 0.018, indicating Rap1p motif is contributing the genetic regulation.

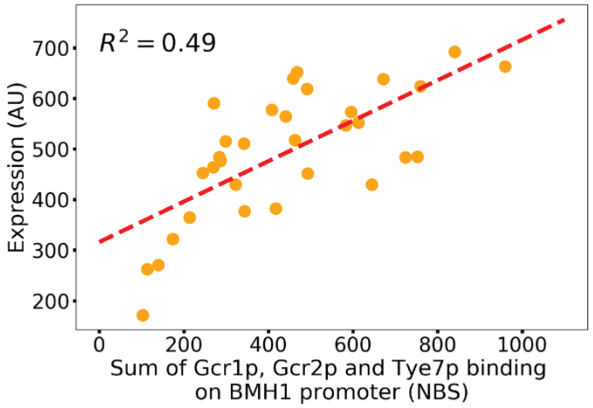


Figure 1.21 Correlation expression and site score on *BMH1* promoter. Expression was measured for all mutated sequences derived from *BMH1* promoter and was correlated with the summation of Gcr1/2p and Tye7p binding results. The binding of three factors from the collective predicts the expression with R^2 of 0.49, Pearson r = 0.70 and p-value = 8.34e-6, Spearman r = 0.67 and Spearman p-value = 2.95e-5.

Taken together, our experiments suggest that Tye7p is recruited to promoters by Gcr1p/Gcr2p/Rap1p complex and that Tye7p binding often occurs in the absence of its recognition site. However, it appears that Tye7p binding is stabilized by the presence of its motif. In contrast, the Gcr1/2p recognition site is necessary and sufficient for the binding of these proteins, suggesting a hierarchy in which these factors can recruit Tye7p but not vice versa (**Fig 1.22**). The transcriptional output at promoters bound by this complex correlates with the combined occupancy of all TFs, suggesting that each TF in the collective aides in the recruitment of the RNA Polymerase II holoenzyme.

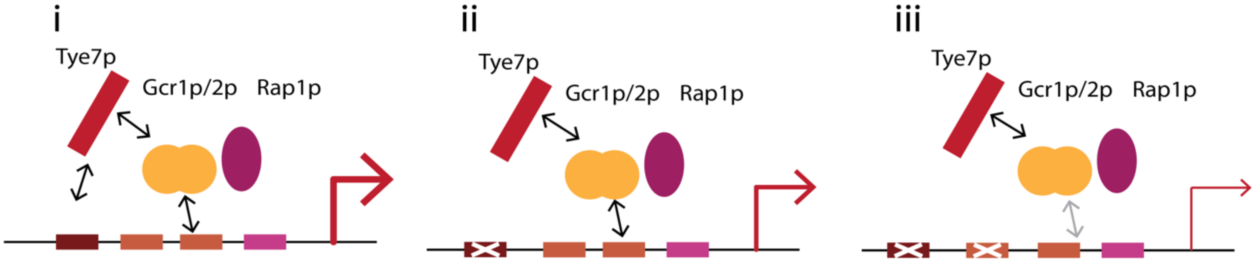


Figure 1.22 The suggested model for Tye7p/Gcr1p/Gcr2p/Rap1p binding collective. i) Tye7p is recruited to promoters by Gcr1/2p and the Tye7p motif, and the expression output is the strongest when all sites are available; ii) Tye7p can be recruited in the absence of a Tye7p motif via a protein-protein interaction with Gcr1/2p, but Tye7p binding occupancy is lowered and the overall expression output is lowered as well; iii) Gcr1/2p occupancy and Tye7p occupancy are lowered with fewer Gcr1/2p motifs, and the overall expression output is further reduced.

1.3 Discussion

In this study, we demonstrated that the CCRA method is a useful tool to study many different aspects of TF binding *in vivo*. Using CCRA, we first measured the DNA binding energy landscapes for Cbf1p and MAX, and we showed that the free energy differences measured by CCRA are strongly correlated with those measured by PBM and MITOMI, suggesting CCRA is a quantitative measure of equilibrium binding. This is likely because the rate of transposon insertion is slow relative to the typical on rates and off rates for TF binding to DNA; in contrast, crosslinking based methods may capture transient TF-DNA binding events as TFs sample weak binding sites (Park P. , 2009), and thus the measured occupancies may reflect a combination of on-rate and equilibrium binding. Next, we set out to understand TF cooperativity by studying a pair of paralogues bHLH TFs, Cbf1p and Tye7p; we observed that Cbf1p binding occupancy is dependent on the DNA helix turn, revealing the biophysical relations between DNA structure and a homotypic cooperative TF; Finally, we characterized the molecular binding logic of Tye7p, which is Tye7p finds its targets via protein-protein interaction with Gcr1/2p and Rap1p without requiring its own motif, further delineating the collective binding model.

Transcription factors orchestrate the gene expression changes that lie at the heart of most biological processes; however, the principles by which TFs locate their target genes and the functional consequences of binding are not well understood. Detailed investigations into the molecular mechanisms that govern TF binding have traditionally used *in vitro* methods (Maerkl, 2007; Fordyce, 2010; Bulyk, 2007; Berger M. F., 2009; Berger M. F., 2006; Stormo, Zuo, & Chang, 2015; Zhao, 2009; Zykovich A, 2009; Majka, 2007; Warren, 2006), which provide limited insights into TF binding *in vivo*, or employ genome editing (Shively, 2019; Kim, 2017; Wakabayashi, 2016), which is slow and costly. Due to these difficulties, many studies that have tried to understand the rules of TFs binding and function have focused on a finite set of loci and a limited number of genetic alternations (Shively, 2019; Kim, 2017; Wakabayashi, 2016). Recently, powerful high-throughput methods, such as Sort-Seq (Kinney, 2010; Sharon, 2012) and barcoded MPRAs (Maricque, 2017; White M. A., 2013), have been developed to allow more comprehensive investigations into the regulatory code, but these rely solely on reporter gene expression and must indirectly infer TF binding and its impact on gene expression. Two recent studies have coupled ChIP-based binding measurement with parallel reporter assays to reveal the correlations between chromatin marks and TF binding (Grossman, 2017) and to examine the predictive power of thermodynamically motivated models of gene expression (Zeigler, 2014). These studies demonstrated the parallel measurement of TF binding on synthetic promoters and represent an important advance; however, neither demonstrated the ability to quantitatively measure binding energies or to analyze cooperative interactions, which are critical measurements for understanding how TFs function. Methods in which TFs direct transposon insertion (Wang H. M., "Calling cards" for DNA-binding proteins in mammalian cells. , 2012; Wang H. M., 2011; Kaya-Okur, 2019; Wang H. J., 2007) or the enzymatic cleavage of DNA (Skene, 2017; Zentner, 2015) show promise for going beyond a qualitative description of TF binding. Here we demonstrate that CCRA is able to quantitatively measure TF binding and reporter gene expression on synthetic sequences in a high-throughput manner. It is a sensitive and accurate method that is amenable to the analysis of complexes of TFs. Therefore, CCRA should be a useful tool to better understand the regulatory principles of TFs localization and functionality.

When designing a CCRA library, certain considerations should be accounted for in order to ensure the accurate quantification of TF binding. It is important to collect enough transpositions events in each experiment relative to the size of the CCRA library. Although chip-based oligonucleotide synthesis allows for very large libraries (up to 244,000 unique oligos) to be synthesized in a cost-effective manner, we have found that it is advantageous to design the library so that smaller subsets (e.g. 100-1000 sequences) can be amplified with unique primer pairs. Since we typically collect 10,000-50,000 transpositions for each CCRA experiment (using 10 yeast plates), limiting the sub-libraries to this size ensures high statistical power for each experiment, while still allowing for the analysis of different TFs or the testing of different hypotheses in a single experiment. The optimal number of transpositions for a particular CCRA experiment will also depend on the transcription factors to be analyzed and the specifics of the library design (e.g. a library consisting of many high affinity sequences may yield more transpositions than library consisting of many low affinity sequences). In our experience, CCRA libraries with 500 or fewer unique sequences yield high-quality binding results, but this could be easily scaled by using more plates or through future improvements to the method. In the future, it should be possible to analyze multiple TFs simultaneously with CCRA technology by adding different TF barcodes during the first amplifying step and then transforming the barcoded libraries into different yeast strains, each containing a different TF-Sir4p fragment fusion.

The CCRA method is able to analyze a number of user-defined sequences in parallel, providing quantitative and well-controlled measurements that would be difficult to obtain using genome-wide methods. For example, the free energy binding landscape we described for Cbf1p was generated by analyzing all 1bp substitutions to this factor's consensus motif in exactly the same sequence context, a design which enabled the detection of small free energy changes. In contrast, small changes in binding energy cannot be inferred from genome-wide calling card measurements of Cbf1 (**Appendix 1.5**), although the broad trends are generally the same. This is likely due to the fact that while all 1bp substitutions to Cbf1p's consensus binding sequence are indeed present in the genome, they exist in different local sequence contexts, so the measurements are not well controlled. For example, in the yeast genome, one Cbf1p binding site might compete with a nucleosome, while another binding site may not, so the different local contexts confound the accurate measurement of binding energies. Indeed, we observed in our CCRA experiments that when a Cbf1p binding site is flanked with a nucleosome disfavoring sequence, Cbf1p binding consistently increases (**Fig 1.3**). The ability to make well-controlled measurements likely also contributed to our ability to detect the periodic phase dependence of Cbf1p's cooperativity. This phase dependence is an interesting phenomenon, and to our knowledge cooperative binding of a transcription factor complex has not been previously shown to be influenced by helical phase. However, an important related result was found by Kosuri and colleagues where they found that the expression output of a reporter gene depended on the helical phase between the transcription start site and the binding site of a transcriptional activator (Davis, 2019).

We envision CCRA will be broadly applied to study three different aspects of TF binding: 1) quantitative investigations into TF-DNA interactions in the native cellular environment; for example, mapping TF binding energy landscapes *in vivo* or evaluating the effect of flanking sequences on motif recognition; 2) studies into the mechanisms by which TFs bind cooperatively; for example, evaluating the energetic contributions of different TF binding sites to the binding of a TF complex; 3) dissection of the relationship between TF occupancy and transcriptional output. Furthermore, it is likely that CCRA can be extended to multicellular eukaryotic systems in the future using the appropriate transposon machinery. The Calling Card method has been applied to study mammalian TFs such as SP1 and BAP1 with PiggyBac transposon (Wang H. M., "Calling cards" for DNA-binding proteins in mammalian cells. , 2012; Yen, 2018), so this transposon system is an excellent candidate for performing CCRA in mammalian cells. Such investigations should ultimately lead to a better understanding of the roles that TFs play in orchestrating the transcriptional networks that allow cells to carry out their diverse functions.

1.4 Materials and Methods

1.4.1 Library Design and Amplification

CCRA libraries are created by array-based oligonucleotide synthesis (Agilent). Each element of the library is a distinct 230 bp oligonucleotide comprised of 5 different sequence regions. The first region is a 20 bp constant sequence that is homologous to the backbone plasmid to support Gibson cloning. The next (downstream) 11 bp sequence is unique to each sub-library to enable the amplification of subsets of the library elements that are synthesized in each batch. This allows for the analysis of different TFs or the testing of different hypotheses using a single oligonucleotide synthesis. The third region is the 170 bp user-defined variable synthetic promoter sequence. This region is followed by 12 bp "promoter" barcode that identifies the corresponding promoter sequence at Illumina sequencing step. Each promoter barcode is designed to be at least 3 bp different than all other barcodes to control for synthesis, PCR and sequencing errors. The last region of each library element is a constant 17 bp sequence used for PCR amplification. The library pool was synthesized by Agilent as 10 pmol of lyophilized nucleic acid. To amplify the library, we used 0.15 ng of library DNA template in a final 50 µL PCR reaction. In each 50 µL reaction, we used 0.2 mM dNTP mix, 0.5 µM forward primer, 0.5 µM reverse primer, 1X Herculase II reaction buffer, 1M Betaine, 0.15 ng DNA template in water, 1 µL of Herculase II polymerase (Agilent). The PCR reaction was cycled as follows: 95 degrees for 1 min, 16 cycles of 95 degrees for 30 secs and 58 degrees for 2.5 mins and then 72 degrees for 4 mins. PCR products were purified by AMPure XP beads from Beckman coulter with 1:1.6 of PCR sample to magnetic particles ratio according to manufacturer's instructions. Typically, we obtained 5 to 10 ng/µL of DNA in a final volume of 15 µL.

1.4.2 CCRA Library Construction

Plasmid pRS414 was used as the backbone to create library plasmid pRM1806. To clone library sequences into the pRM1806 backbone, we linearized the plasmid with high fidelity KpnI and SacI (NEB), and then performed gel extraction using the Qiagen DNA extraction kit. We used 0.03 pmol of the linearized plasmid and 0.12 pmol of purified PCR product in a Gibson assembly reaction (NEB), following the manufacturer's instructions. Nitrocellulose membrane (0.025 µm) was used to filter Gibson assembly product by drop dialysis following the Millipore Sigma protocol. The library was electroporated into 10G SUPREME Electrocompetent cells (Lucigen) using 0.1 cm cuvette and cells were plated on to Kanamycin containing LB plates after 1-hour recovery in SOC. After 16 hours of growth, over 50,000 colonies were scraped, and the plasmid DNA was extracted using Qiagen Miniprep Kit.

1.4.3 Calling Cards Induction and Promoter Library Recovery

The yeast strain used in this study was yRM1004, which is derived from matA_deltaSir4, and has the following genotype: his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 Δsir4::KanMx Δtrp1::HygMx. Induction of TF directed transposition was performed using a modified calling cards protocol (Wang H. J., 2007; Wang H. M., 2011). Briefly, plasmid containing a Sir4p (amino acids 951-1200) tagged TF driven by *ADH1* promoter with LEU2 auxotrophic marker was transformed into yeast cells (yRM1004) together with the plasmid pRM1804 which contains the URA3 marker and a galactose inducible Ty5 transposon with an artificial intron inside of His3 gene that is inside of Ty5 gene body for the purpose of selecting transposition positive cells in the next step (Zou, 1996). After transformation, cells were plated onto a Glu-Ura-Leu plate to select for cells carrying both the TF-sir4p fusion plasmid and Ty5 transposon plasmid. Next, a single colony was picked for library plasmid transformation. The library plasmid pRM1806 carries the TRP auxotrophic selection marker, so after the yeast cells were transformed with the library plasmid, they were plated onto a Glu-Ura-Leu-Trp plate to select for all three plasmids. Multiple parallel transformations were performed to obtain a diverse population of library sequences. We typically obtained over 10,000 colonies for each sub library. All colonies were pooled and plated to Gal-Ura-Leu-Trp to induce Ty5 transposition on 10 plates to increase the number of transpositions. Cells were allowed to grow on galactose plates for four days at room temperature. After galactose induction, we replica plated cells to Glu-His-Trp to select for yeast with Ty5 transpositions and that carry the library plasmid. After 2-3 days, colonies were scraped, and plasmid extraction was performed using the Yeast Plasmid Mini Kit (Omega).

1.4.4 Preparation of Illumina Libraries for Calling Cards Mapping

We performed four independent PCRs to recover transpositions that were inserted into synthetic promoters in either of two possible orientations and upstream or downstream of the barcodes and UMI. We performed an additional PCR to measure the relative abundance of elements in the library for normalization. For these four PCRs, one primer of each pair is specific to either 3' LTR of Ty5 transposon sequence or 5' LTR of Ty5 transposon sequence, and the other primer is specific to a constant region either upstream or downstream of the inserted library sequence on the plasmid. For the additional PCR, one primer is specific to an upstream constant region of the inserted library sequence on the plasmid, and the other primer is for the downstream constant region. All 5 PCR products were pooled together for sequencing.

In each PCR reaction, we used 1X RedTaq buffer, 0.2 mM dNTP mix, 1M Betaine, 0.5 µM forward primer, 0.5 µM reverse primer, 4 µL RedTag DNA polymerase (Sigma-Aldrich), 1 µg of the purified plasmid DNA and the corresponding amount of water to reach a final volume of 50 µL. The PCR parameters were set to be 93 degrees for 2 mins, 24~28 cycles of 93 degrees for 30 secs and 62 degrees for 6 mins, and 62 degrees for 6 mins. The PCR products were then purified with Qiagen PCR purification kit before sequencing.

1.4.5 Measuring Reporter Expression in CCRA Libraries by Sort-Seq

After transforming the library plasmid into yeast, we divided the cells for either Calling cards or Sort-Seq. For expression measurement, we followed the experimental procedures as well as promoter expression calculation described in (Kinney, 2010; Sharon, 2012). We sorted cells into 8 bins of 100,000 cells each, and then added yeast culture media to grow the cells for 16 hours. Cells from each bin were then pelleted separately and the plasmids were extracted with Yeast Plasmid Mini Kit (Omega) for sequencing.

Next, we performed a separate PCR reaction for each sorted bin. The primer sequences are listed in supplemental table 1, and they target the constant regions upstream and downstream of the CCRA library. In each of the 8 PCR reactions, the reverse primer was indexed with unique barcode to allow the reactions to be sequenced together. The PCR amplification conditions used were identical to those used for calling cards recovery.

1.4.6 Analysis of Sequencing Reads for Quantification of TF Binding

To quantify TF binding to CCRA libraries, we analyze Illumina paired end sequencing reads to count all unique insertions into each library member. A transposition is unique if it can be distinguished by its insertion coordinate relative to the library reference or contains a unique UMI in instances where multiple insertions have landed at the same position across four independent PCRs. To identify unique insertions from the sequencing data, we first filter for reads containing the appropriate 12 bp library barcode and 6 bp TF barcode. Filtered reads are then divided into five categories: reads from synthetic promoters where the Ty5 transposon inserted in the forward direction upstream of the promoter barcodes, reads where Ty5 inserted in reverse direction upstream of barcodes, reads where the Ty5 inserted in forward direction downstream of barcodes, reads where the Ty5 inserted in reverse direction downstream of barcodes, and reads from synthetic promoters without insertion. This categorization is achieved by analyzing the first 20bp of read 1 and read 2. The next 12 bp are used to map the precise location of the transposon insertion into the synthetic sequence. We used the 4 bp UMI to resolve events when multiple calling cards are deposited at the same base pair in a given synthetic sequence. Finally, we use the number of full-length sequences recovered for each library element as a normalization factor to control for the variation in abundance between library members. The total number of independent insertions for each library member is normalized by the relative abundance of each element in the library to compute a normalized binding score (NBS) of TF binding to each synthetic sequence.

1.4.7 Using an Expectation Maximum Algorithm to Distinguish TF-directed Insertions from Background

For experiments in which changes in binding energies are measured, it is important to measure TF binding strength as accurately as possible. Therefore, we used an expectation maximization algorithm to resolve TF-directed transpositions which occur near TF recognition sites from background transpositions which occur uniformly across the synthetic promoter. Since the distribution of TF directed insertions is approximately Gaussian with the distribution centered at the TF recognition site, we assumed that TF directed insertions can be modelled with this distribution while background insertions follow a uniform distribution. We then used an expectation maximum algorithm to estimate, for each synthetic promoter, the variance of the Gaussian distribution (the mean value is determined by the location of the TF recognition sequence) and the fraction of insertions that were the result of a TF-directed or background transposition. For each library element, we iterate each independent insertion for maximum of 1000 times or until the parameters no longer change. The estimated fraction of TF-directed insertions is used to multiply the raw number of insertions at each promoter to remove insertions due to non-specific transposition. This background correction step removes 0~20% of non-specific insertions, which is important for calculating small changes in binding energy; however, incorporating this step does not impact other analysis is not used for sequences where the Gaussian assumption is not appropriate (e.g. for sequences with multiple TF sites or for TFs whose recognition sequence is not well-characterized). Therefore, we performed this background correction only for the generation of binding energy landscapes.

1.4.8 Binding Energy Difference Calculation

To quantitatively compare CCRA with PBM and MITOMI in terms of binding affinity, we calculated the change of binding energy ($\Delta\Delta G$) from consensus site to the alternative site as follows:

Under binding equilibrium, [TF] and [sequence] associate at the same rate that the bound complex [TFS] disassociates:

$$[\mathrm{TF}] + [\mathrm{S}] <-> [\mathrm{TFS}] \qquad (1.1)$$

The Gibbs free energy ΔG is related to the binding constant K as follows:

$$\mathrm{K(S)} = \frac{[TF][S]}{[TFS]} = e^{\Delta G/RT} \qquad (1.2)$$

$$\Delta G = \mathrm{RTln(K(S))} \qquad (1.3)$$

The binding occupancy on a sequence is defined as the fraction of bound sequence to the total sequence in solution. Replace [TFS] with [TF][S]/K according to 2a, and by approximation that K(S) is much greater than [TF] as the affinity of these sequences are high, we get:

$$\mathrm{Occ(S)} = \frac{[TFS]}{[TFS]+[S]} = \frac{[TF]}{[TF]+K(S)} \approx \frac{[TF]}{K(S)} \qquad (1.4)$$

$$\mathrm{K(S)} = \frac{[TF]}{Occ(S)} \qquad (1.5)$$

Therefore, the change of binding energy equals:

$$\Delta\Delta G = \Delta G(Sconsensus) - \Delta G(Smutant) = \text{-RTln}\ (\frac{Occ(Smutant)}{Occ(Sconsensus)}) \qquad (1.6)$$

1.4.9 Test for Binding Cooperativity

To determine if Cbf1p binds cooperativity at various synthetic promoters, we compared the observed occupancy to expected occupancy assuming independent binding, and we derived this test by the following:

[Cbf1p] + [DNA with two free sites] $\overset{k1}{\leftrightarrow}$ [Cbf1p-DNA with one free site] + [Cbf1p] $\overset{k2}{\leftrightarrow}$ [2*Cbf1p-DNA with both sites occupied]

To simplify: [P]+[S] $\overset{k1}{\leftrightarrow}$ [PS]+[M] $\overset{k2}{\leftrightarrow}$ [P_2S] (1.7)

$$\mathrm{Occ(P)} = \frac{2*K1*P+2*K1*K2*P^2}{1+2*K1*P+K1*K2*P^2} \quad (1.8)$$

If Cbf2 binds additively, then K1 = K2 = K;

$$\mathrm{Occ(P)} = \frac{2*K*P+2*k^2*P^2}{1+2*K*P+k^2*P^2}=\frac{2*K*P(1+K*P)}{(1+K*P)^2}= 2*\left(\frac{K*P}{1+K*P}\right) \quad (1.9)$$

And so, the null expectation for binding occupancy is simple twice the observed binding to a single recognition site.

1.4.10 TF Motifs and NDS Definition

For yeast TF motifs, we used the recommended PWMs compiled by Spivak and Stormo in the ScerTF database(stormo.wustl.edu/ScerTF). The ScerTF recommended PWM cutoff scores were used to define the presence or absence of TF sites on DNA sequences. The binding motif of MAX, the human bHLH factor, was obtained from factorbook (v1.factorbook.org/mediawiki/index.php/MAX). The NDS sequences used for this study were taken from a study by Raveh-Sadka (Raveh-Sadka, 2012); the NDS1 and NDS2 sequences in this work correspond to the v1 and v37 sequences from that study, respectively.

1.4.11 Processing PBM and MITOMI Data

Cbf1p PBM data was obtained from UniProbe database, and we used dataset UP00397 for calculating free energy changes. We searched for each motif variant in PBM data, all the sequences that contains the same motif variant are grouped together, and the average PBM score was used to reflect the binding affinity for that variant. MITOMI data was obtained from the study by Maerkl (Maerkl, 2007) and the K_d for each relevant variant reported in the original publication was used for the calculation directly.

Chapter 2: CG Rich Sequences Act as a Kinetic Funnel to Specify Transcription Factor Binding

2.1 Introduction

Transcription factors (TFs) are critical elements in determining various cellular regulations, and they function by binding to regulatory DNA, called TF binding sites (TFBSs) to either activate or repress expression (Stormo G. , 2000; Accili & Arden, 2004; Vaquerizas, 2009; Simon, 2001). TFBS are generally short DNA sequences ranging from 5bp to 15bp long; and TFs often can tolerant a few mismatches to the consensus TFBS, both of which in turn result in excessive non-functional TFBSs with binding potential in the regulatory regions that are unbound for a given TF. The principle that TFs follows in selecting the actual functional TFBSs accurately is an intriguing problem (Slattery, 2014). Most efforts in studying the regulations of TF binding can be categorized into trans-factor related such as cooperativity with other co-factors and competition with nucleosomes (Liu X. L., 2006; Zhou X. &., Integrated approaches reveal determinants of genome-wide binding and function of the transcription factor Pho4., 2011; Mirny, 2009) and cis-factor related such as the affinity and structural features of the TFBS (Tanay, 2006; Bradley, 2010; Fordyce, 2010; Bulyk, 2007; Berger M. F., 2009; Berger M. F., 2006; Stormo, Zuo, & Chang, 2015; Warren, 2006).

Attempts to predict TF binding in the aspect of cis-factor have focused on almost exclusively on nucleotide sequences at, or immediately flanking (2-4bp) TF binding sites (Zhou T. S., 2015;

Mathelier, 2016; Zeiske, 2018). However, it has recently become appreciated that, in some instances, the local DNA context (LDC) in which a TF binding motif resides (the flanking 50-200 bp) can have an influence on TF binding. One possible explanation is the ability of these flanking sequences to recruit or exclude nucleosomes (Struhl, 2013; Raveh-Sadka, 2012; Levo, 2015), but in some instances, the effect appears to be independent of nucleosome occupancy. For example, White and colleagues used a plasmid based massively parallel reporter array system to show that only 84bp of DNA flanking high scoring CRX motifs determined whether a binding stie was transcriptionally active or not (White M. A., 2013). Similarly, Hartl and colleagues found that flanking sequences enhanced the probability that an enhancer was active and increased the probability of TF binding (Hartl D, 2019). Moreover, a study that examined both *in vivo* and *in vitro* binding data showed that distinct sequence composition and the similarity to the core binding motif on the environment DNA for TF bound regions (Dror, 2015).

These studies and others (White M. A., 2013; Hartl D, 2019; Dror, 2015) highlight the important role local sequence context plays in specifying TF function. However, many unanswered questions remain about this phenomenon: do flanking sequences influence the binding of all TFs, or just a select few? Are different TFs influenced by different flanking sequences, or are there universal sequences that affect all TFs? How strong is the influence of flanking bases on TF binding relative to better-characterized factors such as motif strength or nucleosome occupancy? Most importantly, what is the mechanism by which flanking bases influence TF binding?

In this study, we sought out to answer these questions by investigating if predictive information is embedded on local DNA sequence on various TFs in *Saccharomyces cerevisiae*. We discovered there was a general preference for TFs to bind at CG rich sequences; we then analyzed whether such preference was linked to intrinsic nucleosome binding preference and found the CG preference in LDC for TF binding was independent of nucleosome regulation. We next examined the possible mechanism by which LDC influence TFs binding site selection, through recruiting ‘licensing’ factors or kinetically assisting TF search for a target site. We showed high CG LDC was preferred by TFs *in vitro* condition, which suggested such preference only involves TFs and DNA and pointed us to TF search kinetics. CG rich feature in LDC may act as an energetical funnel to facilitate TF recognizing a target binding site, and we verified the theoretical validity of this hypothesis with simulation with Gillespie algorithm. In the end, we revealed CG preference was also present in a large group of human TFs, indicating the usage of LDC is a general mechanism for TF binding specificity.

2.2 Results

2.2.1 Local Sequence Context Predicts Motif Binding for Yeast TFs

Extensive research has been focused on TF-DNA interactions at the binding motif and bases immediately flanking the motif (Berger M. F., 2009; Berger M. F., 2006; Fordyce, 2010; Stormo G. , 2000; Tanay, 2006; Zhou T. S., 2015). The subtle variation within the motif and the flanking sequences alters the affinity of TF binding, which changes the strength or the residence time of the binding. The affinity of DNA sequence is of great importance for TF-DNA interaction and ultimately determines the binding potential of a site. However, there are many high binding potential DNA sequences on the genome that are not bound by TFs (See **Appendix 2.1** for intersection between binding peaks and motifs). In this study, we define TFBS as any DNA sequences with binding potential according to the score calculated based on position weight matrix (PWM). TF binding motifs are considerably short with regard to the genome size; and it is reasonable to anticipate that many other larger scale factors may contribute to the specificity of TF localizing in addition to TF binding motif. It has been suggested that chromatin structure, histone regulations and nucleosome occupancy can influence TF localization (Wang J. Z., 2012; Shai R Joseph, 2017; Zhou X., 2011; Liu X. L., 2006). In this study, we focus on the connection of local DNA sequence in the process of TF in searching for target binding sites. It has been proposed that the LDC can influence TF binding in relation to the intrinsic regulations of nucleosome occupancy (i.e., nucleosome disfavoring sequences) (Raveh-Sadka, 2012) and motif combinations of trans-factor (Liu J. S., 2020; Shively, 2019; Panne, 2007; Junion, 2012) however, limited investigation has been done to understand, on pure cis regulation level, if LDC contributes to TF localizing target binding sites and how strong is such impact.

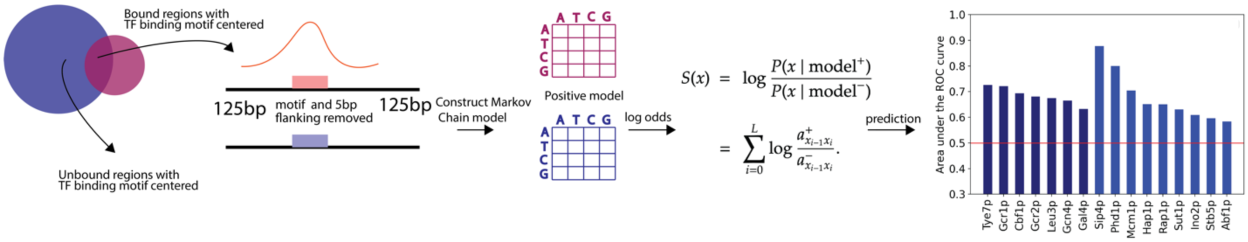


Figure 2.1 The flowchart for modelling LDC for TF binding prediction. The total TFBS on intergenic regions were intersected with binding peaks and were divided into bound and unbound sets. 125bp flanking DNA of either side of TFBS from the bound and unbound set was modeled by first order Markov Chain with 5-fold cross validation and applied to test sequences to generate a log-odds score for estimating the likelihood of being bound for the given TF. The binding prediction was evaluated with ROC curve and the results for all TFs with well-defined motifs were summarized into a bar chart shown in the right panel.

First, we decided to test if flanking sequences contain predictive information for TF binding. For each TF with a well-defined motif, we searched for all binding motifs that are present on intergenic regions and divided these motifs into bound and unbound set according the TF binding data by either Calling Cards or ChIP-exo method (Shively, 2019; Wang H. M., 2011; Rhee, 2011). For both sets, we took 125 bp upstream and downstream as LDC for analysis with the motif itself and 5bp immediately flanking the motif removed to exclude the motif strength and DNA shape effect on TF binding. To understand if these local DNA sequences alone can distinguish bound TFBS from unbound TFBS, we performed supervised learning on these two sets using first order Markov Chain model to preserve the sequential nucleotide information in a parsimonious way. With 5-fold cross validation, we showed that modelling the local DNA sequences can improve binding prediction for all TFs with area under the receiver operator curve (AUROC) from 60% - 90% (**Fig 2.1**), suggesting local DNA sequences alone contain predictive information for TF binding.

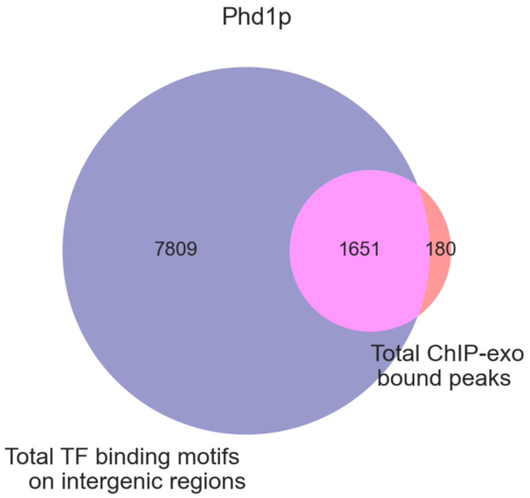


Figure 2.2 The overlap between ChIP-exo peaks and total Phd1p TFBS on intergenic regions. Phd1p TFBS was searched on yeast intergenic regions with one log score lower than the recommended PWM

score from ScerTF database and the coordinates of TFBS were intersected with all ChIP-exo binding peaks.

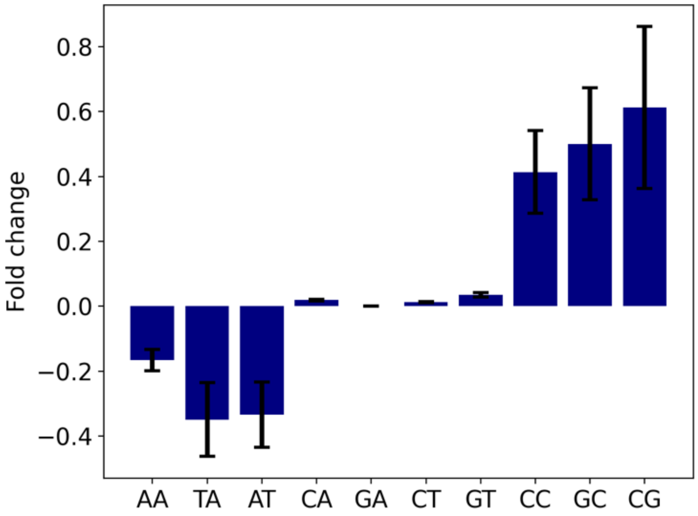


Figure 2.3 Dinucleotide fold change in LDC between Phd1p TFBS bound and unbound sets. All possible dinucleotides were counted in LDC on Phd1p TFBS bound and unbound sets and compared in terms of fold change. The standard deviation across all sequences were shown as error bars.

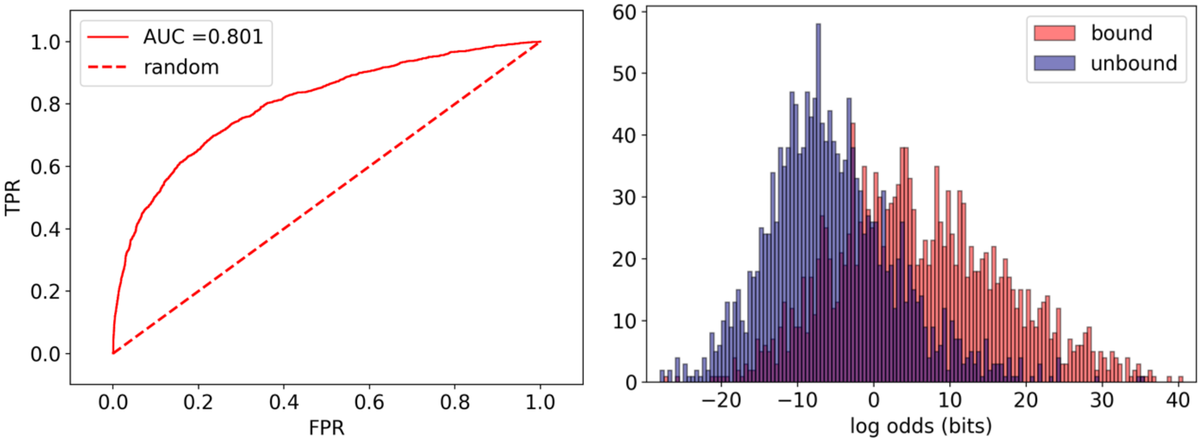


Figure 2.4 LDC modeled by first order Markov Chain can improve Phd1p binding prediction. Left) ROC curve using the calculated log odds score for each Phd1p TFBS containing sequence. Right) Histogram view of the calculated LDC score for Phd1p TFBS bound and unbound sets.

Taken Phd1p and Gcr1p as examples, there are five folds more unbound motifs than bound ones (**Fig 2.2 & Fig 2.5**). To have a general view of the DNA characteristics of the flanking sequences, we took LDC from both bound motifs and unbound motifs and counted all possible dinucleotide in both sets and plotted the frequency fold change (**Fig 2.3& Fig 2.6**). CG preference on the LDC between bound set and unbound set was shown for both TFs significantly in **Figure 2.3 and Figure 2.6,** and the CG rich feature was observed for all other 14 TFs analyzed. With 5-fold cross validation, we compressed the local DNA information from training data into first order Markov Chain model, and we evaluated the model prediction performance with testing data by receiver operator curve (ROC). The area under the curve (AUC) was 0.801 for Phd1p and 0.723 for Gcr1p with Mann Whitney statistical test p-value 0 and 2.05e-9 respectively (**Fig 2.4 Left& Fig 2.7 Left**), and we used histogram to show the model performance in distinguishing bound and unbound sequences (**Fig 2.4 Right & Fig 2.7 Right).**

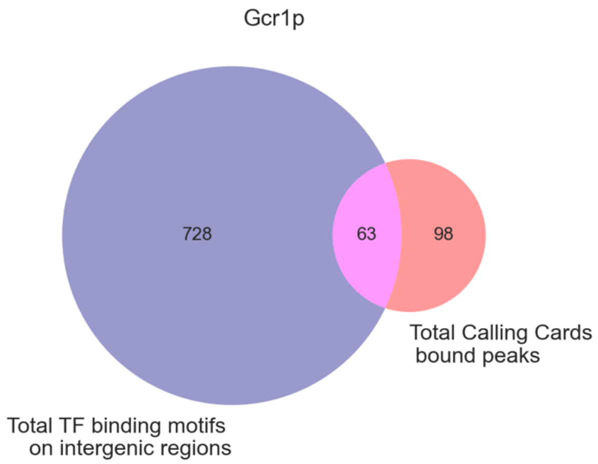


Figure 2.5 The overlap between Calling Cards peaks and total Gcr1p TFBS on intergenic regions. The same as Figure 2.2, Gcr1p TFBS was searched on yeast intergenic regions with one log score lower than the recommended PWM score from ScerTF database and the coordinates of TFBS were intersected with all Calling Cards binding peaks.

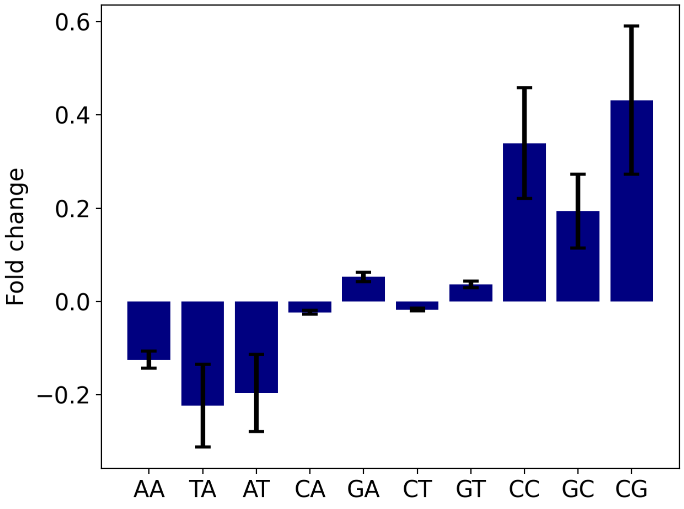


Figure 2.6 Dinucleotide fold change in LDC between Gcr1p TFBS bound and unbound sets. The same as Figure 2.3, all possible dinucleotides were counted in LDC on Gcr1p TFBS bound and unbound sets and compared in terms of fold change. The standard deviation across all sequences were shown as error bars.

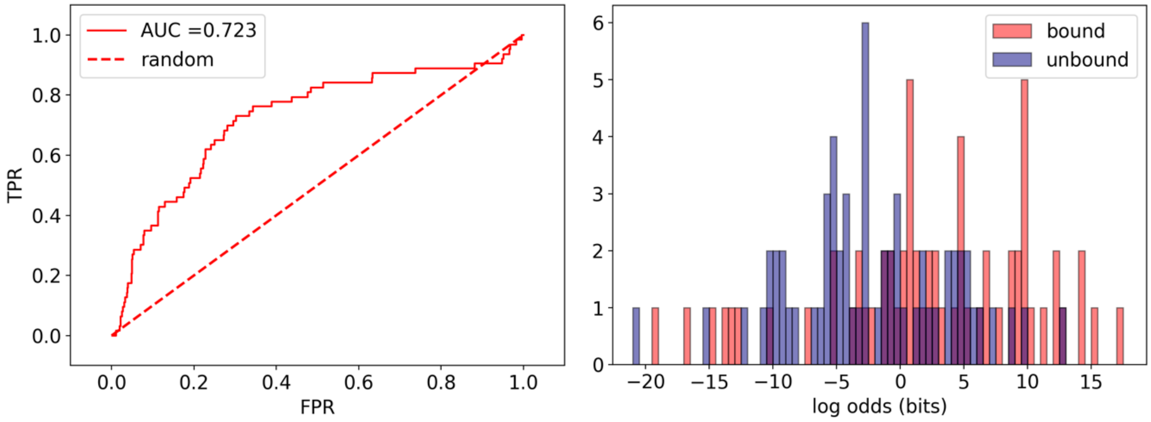


Figure 2.7 LDC modeled by first order Markov Chain can improve Gcr1p binding prediction. Left) ROC curve using the calculated log odds score for each Gcr1p TFBS containing sequence. Right) Histogram view of the calculated LDC score for Gcr1p TFBS bound and unbound sets.

2.2.2 A Universal Dinucleotide Signature in Flanking Bases Predicts TF Binding for TFs with Motif and for TFs without Well-defined Motif

From the analysis in the previous section, we noticed the characteristics of the dinucleotide frequency are very similar across all the TFs with well-defined motifs (i.e., higher CG content and lower AT content), which made us wonder if the preferred features on local DNA are shared across TFs. For each TF, we trained the Markov Chain model with LDC from all other TFs (i.e., all TFBS containing LDC from other 15 TFs with well-defined motifs categorized into bound and unbound set by TF binding data), and we compared the prediction result of using the model trained using other TFs to the result of using their own LDC as training and found that improvement of binding prediction in AUC for every tested TF remained to a similar level (**Appendix 2.2 A**). Moreover, with combining all TFs, we showed dinucleotides that both nucleotides are C or G are significantly enriched and dinucleotides that are A or T are significantly depleted (**Appendix 2.2 B**). Both evidence suggested to us the high CG feature is a general pattern shared by all analyzed TFs that facilitate TF binding, and therefore, we constructed a Universal Markov Chain model using all local sequences from these 16 TFs (See **Appendix 2.2 C & D** for the full model).

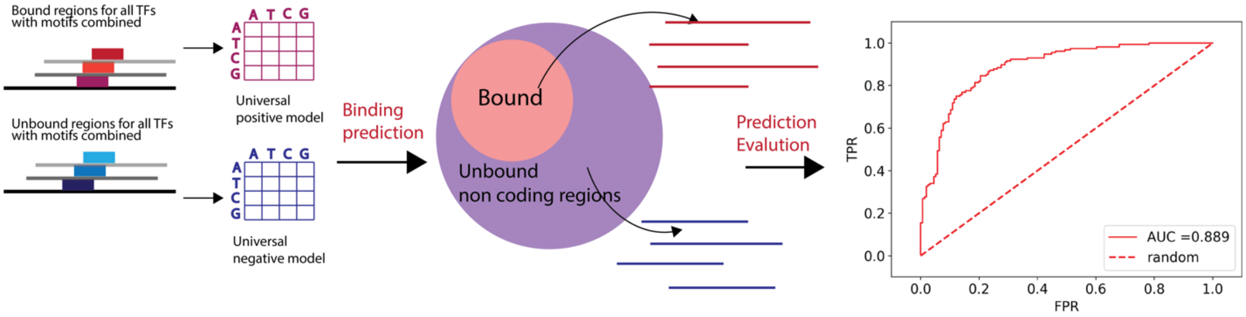


Figure 2.8 Universal LDC model predicts binding for TFs without well-defined motifs. LDC from bound and unbound TFBS for all TFs were combined and made into the Universal Markov Chain model to make binding predictions for TFs without specific motifs. As no TFBS information can be used for these TFs, all bound regions were treated as positive set and all unbound intergenic regions were treated as negative set. The ROC curve for Sef1p binding prediction was shown on the right.

There are many TFs lacking specific and informative motifs or may not yet have established binding motifs, which creates the difficulty in identifying possible TF binding locations in the genome. Therefore, we considered all intergenic regions with binding potential, and categorized all yeast intergenic regions into bound and unbound regions given TF binding peaks by either Calling Cards or ChIP-exo method. With the Universal Markov Chain model, we can now test on TF without specific motifs if the preferred local DNA feature is the same on TFs with well-defined motifs. Specifically, for each TF, we took their bound peaks and all other intergenic regions and asked whether this Universal model was able to classify them apart using ROC as prediction evaluation (**Fig 2.8**, binding prediction for Sef1p was shown). To our surprise, we were able to obtain AUROC from 65% to 90% for all 18 TFs with significant Mann-Whitney p values (**Fig 2.10**), suggesting there is a general preference on the LDC that facilitate TF binding. Given the fact the ROC compares all possible intergenic regions and true binding regions with one Universal model, the improvement in binding prediction for all TFs indicates potentially 'hot spot' and inactive regions for TF regulations. Furthermore, consistent with improvement on binding prediction, the dinucleotide features also remained similar as was observed for TFs with specific motifs, significantly higher CG content and lower AT content (**Fig 2.9**).

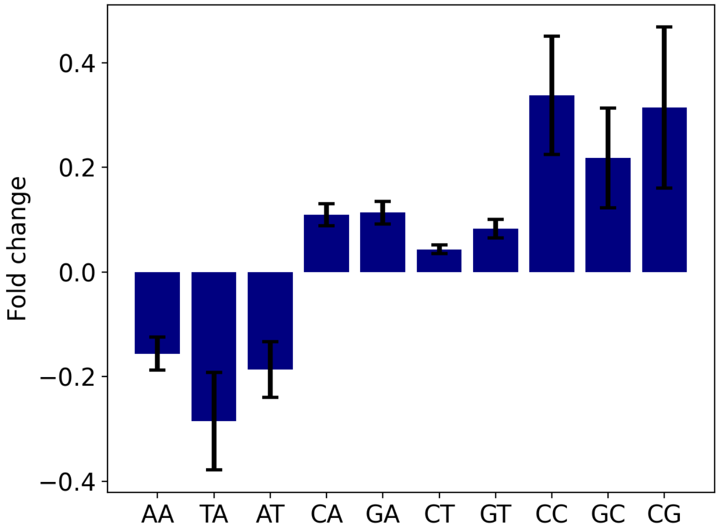


Figure 2.9 Overall dinucleotide fold change for TFs without specific motifs between binding peaks and unbound intergenic regions. All possible dinucleotides were counted in LDC bound and unbound sets for all TFs combined and compared in terms of fold change. The standard deviation across all sequences were shown as error bars.

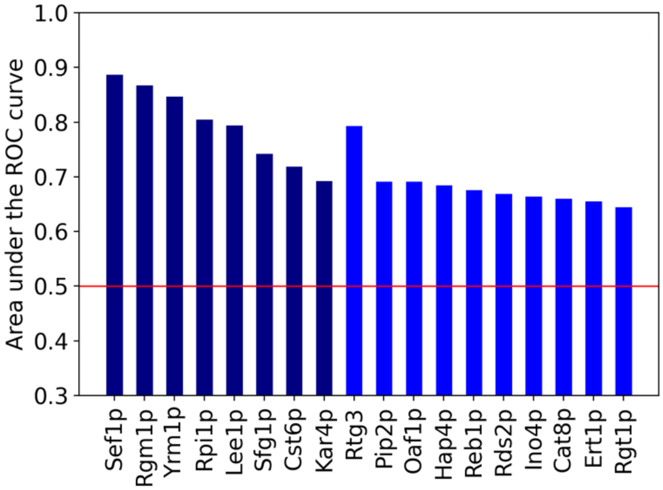


Figure 2.10 Binding prediction AUROC using universal LDC model for all TFs without known motifs. Dark blue bar represents TF binding data collected by Calling Cards method and light blue bar represents TF binding data measured by ChIP-exo.

2.2.3 Local Sequence Context Provides Information Independent of Nucleosome Occupancy

Nucleosome occupancy has been linked to TF binding regulation (Liu X. L., 2006; Zhou X., 2011); it has been observed that nucleosome can compete with TFs for binding sites and as a result exclude TFs to bind at the potential motif. It is reasonable to question if this high CG observation feature that improves TF binding prediction is simply a result of intrinsic nucleosome disfavoring characteristic. To test if this hypothesis is true, we identified binding motifs that are free of nucleosome, and divided them into bound and unbound sets for every TF with a well-defined motif. If it is true that the Universal Markov Chain model which we constructed in the previous section merely captures intrinsic nucleosome free sequence feature, the binding prediction improvement would be lost by comparing the bound and unbound that are both at nucleosome free regions (**Fig 2.11**).

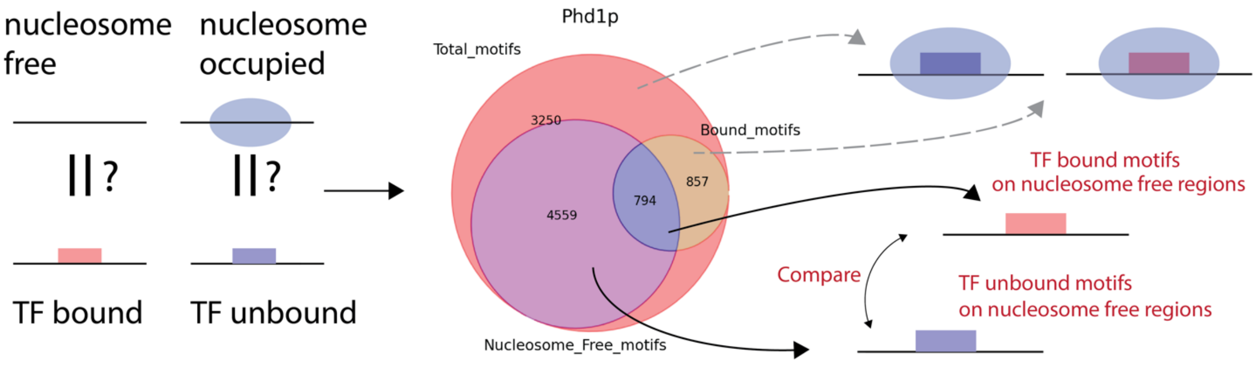


Figure 2.11 Flowchart of assessing if the predictive information in LDC is related intrinsic nucleosome disfavoring sequences. To understand if the bound TFBS region share the same feature of nucleosome free region, we compared flanking DNA sequences from bound TFBS to unbound TFBS that both locate at nucleosome free regions. If the LDC of bound TFBS is the same as intrinsic nucleosome disfavoring, the predictive power would be lost when we compare both sets at nucleosome free regions.

By applying the Universal model to the local DNA sequences, we evaluated the binding prediction with ROC, and we showed that the improvement in terms of AUC for differentiating TF bound motifs and unbound motifs both on nucleosome free regions remained to a similar level (**Fig 2.12**) and even higher for some TFs when we do not limit LDC to nucleosome free regions. Moreover, for TFs lacking specific motifs, we intersected the peak center with nucleosome free regions to identify bound regions that are free of nucleosome; and we compared those bound and nucleosome free regions to all other intergenic regions that are free of nucleosome. Similarly, the AUROC improvement was the about the same as the previous section (**Fig 2.13**). These results demonstrated that the intrinsic nucleosome preference if there is any does not reflect the predictive power on the local DNA sequences on TF binding specificity *in vivo*.

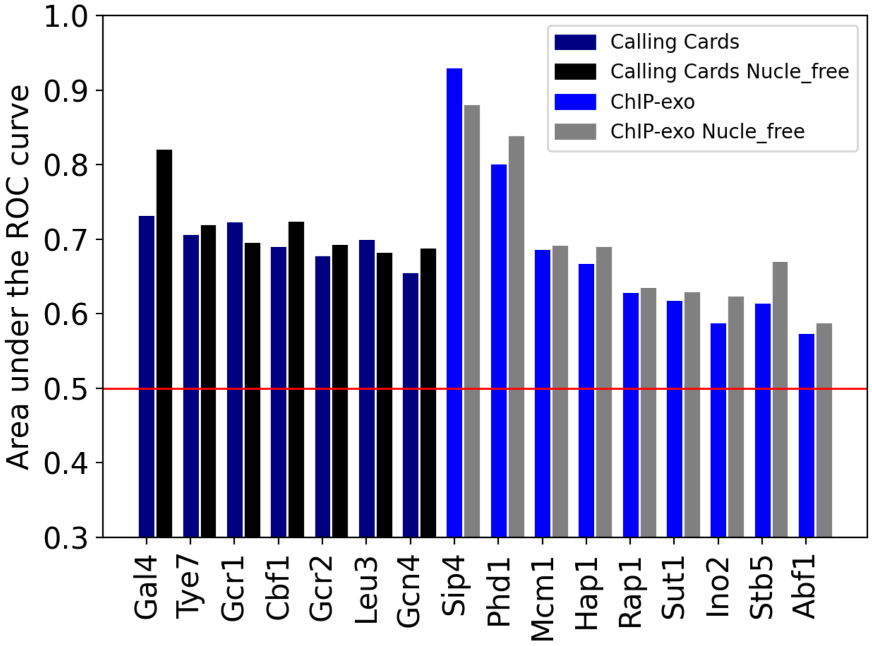


Figure 2.12 AUC comparison between using all LDC or LDC at nucleosome free regions for prediction for TFs with specific TFBS. Binding prediction were performed on bound and unbound TFBS sequences that are filtered based on nucleosome occupancy (only nucleosome free regions used) for all TFs and the AUC was compared to binding prediction with using all sequences for each TF.

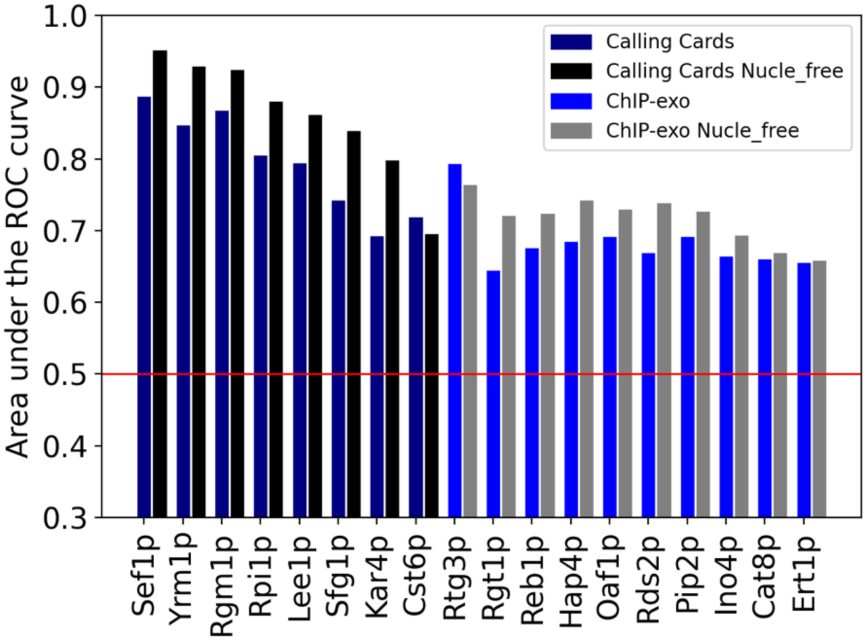


Figure 2.13 AUC comparison between using all LDC or LDC at nucleosome free regions for prediction for TFs without specific TFBS. Binding prediction were performed on bound and unbound sequences that are filtered based on nucleosome occupancy (only nucleosome free regions used) for all TFs and the AUC was compared to binding prediction with using all sequences for each TF.

2.2.4 A Universal Dinucleotide Signature can be Combined with Motif Information to Improve TF Binding Prediction

TF binding motif that is stored in the format of position weight matrix (PWM) has been the most common and informative predictor for TF binding (Stormo G. , 2000). As we have shown that the universal dinucleotide signature can help classifying TF bound and unbound regions, we further investigated if we could incorporate the LDC preference with the motif information in the PWM format to improve TF binding prediction. To have a relative estimation of the extent of improvement in prediction, we compared the incorporation of LDC to the incorporation of nucleosome occupancy which has been studied more extensively and shown to influence TF binding (Liu X. L., 2006).

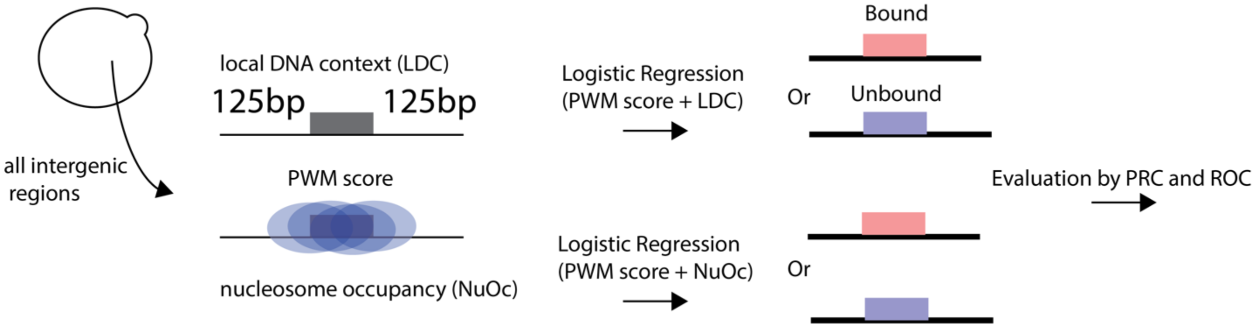

all intergenic regions
local DNA context (LDC)
125bp
125bp
PWM score
nucleosome occupancy (NuOc)
Logistic Regression (PWM score + LDC)
Logistic Regression (PWM score + NuOc)
Or
Or
Bound
Unbound
Evaluation by PRC and ROC

Figure 2.14 Flowchart of incorporating LDC score with PWM score to improve binding prediction. All intergenic regions in S. cerevisiae were included and searched for TFBS with requiring the minimal affinity for the given TF (i.e. PWM score of zero and above). The highest PWM scored TFBS were used to represent the PWM score of each intergenic region and the coordinates of such TFBS were intersected with nucleosome occupancy information for obtaining the NuOc score at the site. The LDC score was calculated using 125bp flanking DNA either side of the TFBS. Logistic regression was performed on either the combination of PWM score and LDC score or the combination of PWM score and NuOc value, and the prediction results were evaluated using PRC and ROC.

For every TF with a well-defined motif, we searched for every intergenic region with PWM score at threshold of zero, affinity higher than random sequences, to include all possible binding sites, and we used the highest score being the motif predictor of this promoter. Next, for every intergenic region, we obtained a LDC score (i.e., the log odds given the universal bound and unbound state model) with the range of LDC defined as 125bp up and downstream of the highest scored motif. Similarly, we obtained a nucleosome occupancy (NuOc) value using the summation of the normalized nucleosome occupancy surrounding the highest scored motif (**Fig 2.14**).

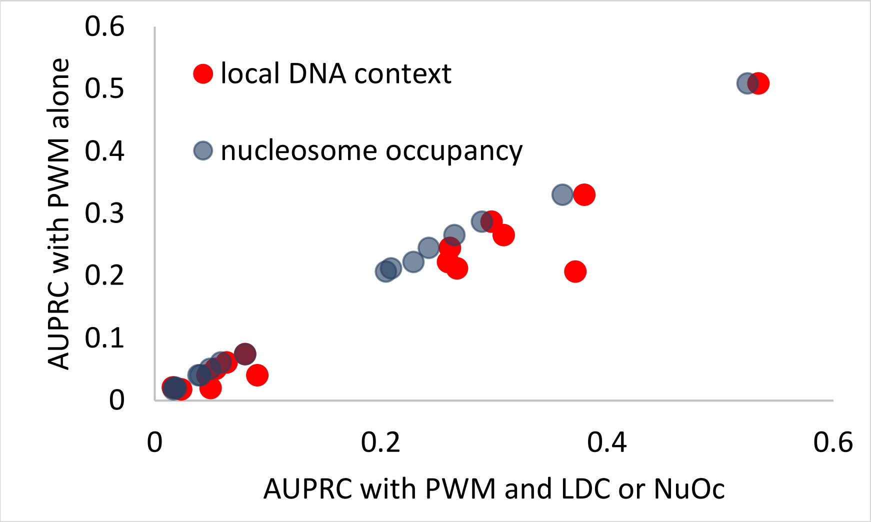


Figure 2.15 AUPRC comparison between using PWM score alone and using PWM score together with LDC score or Nucleosome occupancy (NuOc) score, with red marker representing LDC incorporating to the model and gray marker representing NuOc score incorporating to the model.

To have a generalized model, we performed a simple logistic regression with PWM score together with either the LDC score or NuOc value with 5-fold cross validation. We evaluated the prediction performance with Precision Recall Curve (PRC) and Receiver Operator Curve (ROC), and we compared the results of using both predictors to the results of using the highest PWM alone as predictor in understand how much the improvement is with additional information. With the incorporation of LDC score feature, we showed that binding predictions assessed by both PRC and ROC are better for most of these TFs with improvement ranging from 5% - 160% for PRC and 1%-15% for ROC (except for Sip4p; we reasoned the worse prediction result with including additional LDC score is due to the limited number of true binding targets in this data set which results in weakened model for prediction). The incorporation of LDC produced higher prediction improvement than incorporating additional NuOc value in the model, suggesting a greater *in vivo* contribution of TF binding specificity (**Fig 2.15 & Fig 2.16**).

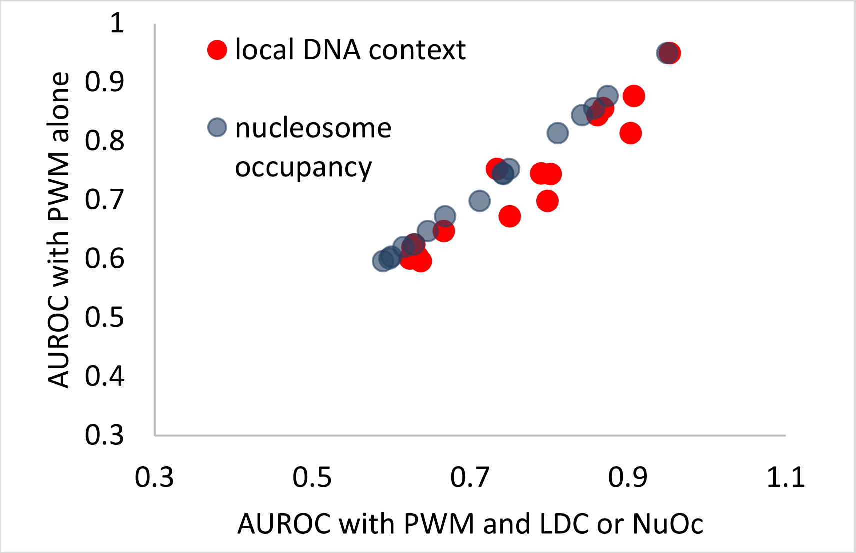


Figure 2.16 AUROC comparison between using PWM score alone and using PWM score together with LDC score or Nucleosome occupancy (NuOc) score, with red marker representing LDC incorporating to the model and gray marker representing NuOc score incorporating to the model.

2.2.5 Local Sequence Context Influences *in vitro* TF Binding

Next, we wanted to understand the mechanism by which LDC influences TF binding. There are a number of possible mechanisms – one or more "licensing" factors that bind at flanking sequences and recruit TFs or influence DNA structure so as to enable TF binding, flanking sequences recruit loci to transcription factories, or it could be that the local flanking sequences aid in TF search kinetics. These hypotheses can be distinguished by determining whether local sequence context has an influence on motif utilization *in vitro*. Therefore, we decided to analyze Dip-ChIP binding data where only the pure protein and naked DNA are present. The advantage of using Dip-ChIP data is that 1) the naked DNA come from the actual genomic sequences whereas other *in vitro* analysis such as PBM utilizes universal artificial DNA sequences that aims to identify short binding motif (Philippakis, 2008); 2) the average length of DNA sequence is about 600bp in Dip-ChIP (Rhee, 2011), which is much longer than common *in vitro* methods, and therefore the results of Dip-ChIP experiments are more suitable for the purpose of studying LDC of binding.

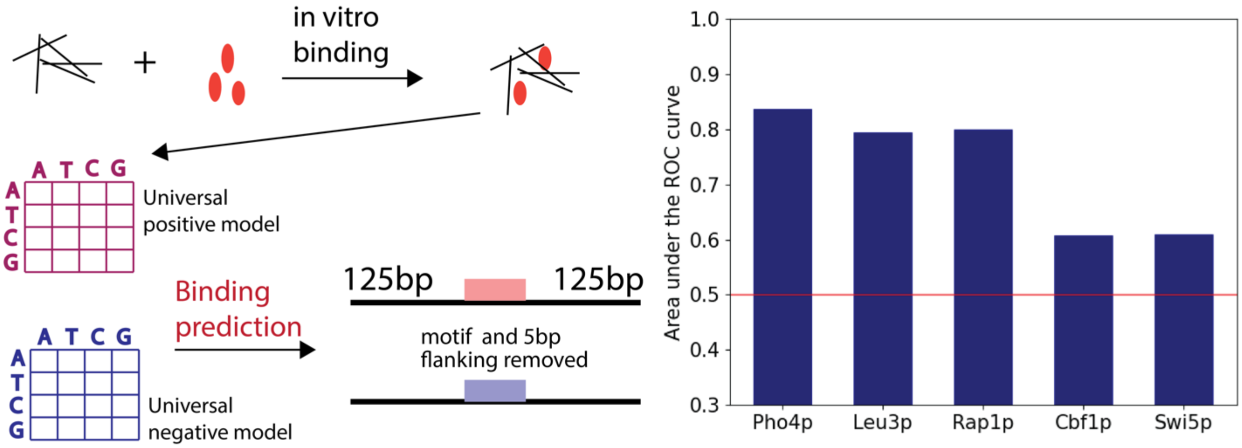


Figure 2.18 The contribution of LDC to TF binding prediction is present *in vitro* condition. The Universal Markov Chain model was applied to make binding prediction for Dip-ChIP data, and the prediction results were summarized in terms of AUC shown on the right.

We revealed that the CG preference is also present in Dip-ChIP data; the dinucleotide frequency characteristics of bound regions remains *in vitro* (**Fig 2.19**)*,* and the Universal Markov Chain model from previous sections learnt from *in vivo* data is also predictive for these *in vitro* TF binding data with AUC ranging from 0.6 to 0.85 (**Fig 2.18**). Moreover, three TFs in this analysis coincided with our *in vivo* analysis in Fig 2.1, and all of them showed similar level of improvement in binding prediction, suggesting consistency of local DNA preference in both conditions, and thus, we believe the CG rich sequence is preferred by TFs through purely biophysical protein-DNA interaction and such preference in LDC facilitates TF searching kinetics.

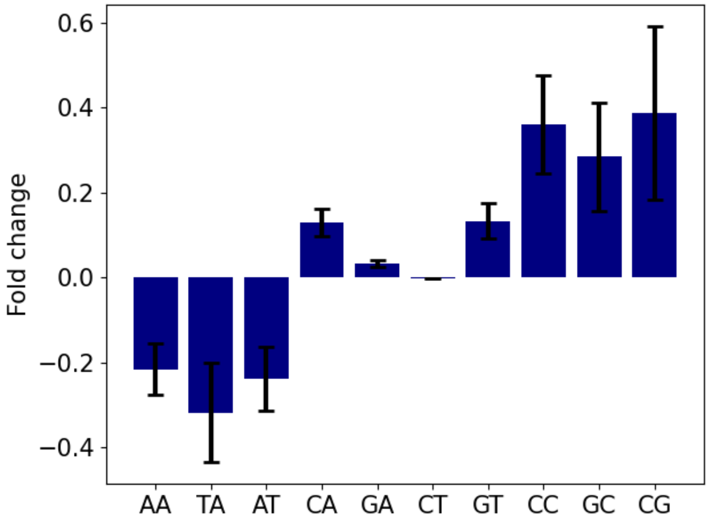


Figure 2.19 The overall dinucleotide signature in LDC for *in vitro* binding assay. All possible dinucleotides were counted in LDC of Dip-ChIP bound regions and unbound TFBS containing regions for all TFs combined and compared in terms of fold change. The standard deviation across all sequences were shown as error bars.

2.2.6 LDC Serves as a Kinetic Funnel for TF Binding

The genome size is considerably large with respect to the size of a TF; however, TF can rapidly locate its binding target with high accuracy. Facilitated diffusion (FD) by Berg and Von Hippel was proposed to explain the fast target search process (Berg OG, 1981); in FD, TFs can switch between two modes to search for a binding site while sliding on the DNA chain. This process was further characterized by Slutsky and Mirny; the TF-DNA complex undergoes confirmational changes to switch between two modes, a highly specific recognition mode and a weakly specific search mode, for fast and stable DNA exploration (Slutsky & Mirny, 2004). One important implication of the FD mechanism is that the local DNA environment may affect the kinetics in the searching process.

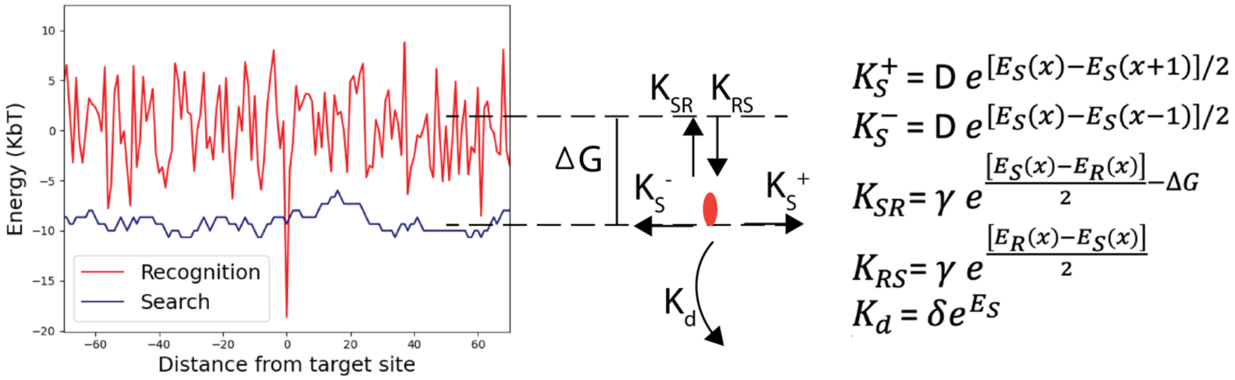

10
5
0
−5
−10
−15
−20
Energy (KbT)
Recognition
Search
−60
−40
−20
0
20
40
60
Distance from target site
K_{SR}
K_{RS}
ΔG
K_S^-
K_S^+
K_d
$K_S^+ = D\, e^{[E_S(x)-E_S(x+1)]/2}$
$K_S^- = D\, e^{[E_S(x)-E_S(x-1)]/2}$
$K_{SR} = \gamma\, e^{\frac{[E_S(x)-E_R(x)]}{2}-\Delta G}$
$K_{RS} = \gamma\, e^{\frac{[E_R(x)-E_S(x)]}{2}}$
$K_d = \delta e^{E_S}$

Figure 2.20 The FD kinetics of TF target search process. A graphical example of the binding energy for TF-DNA exploration in two modes, recognition mode colored in red and search mode colored in blue. The rate equations that characterize the two modes searching process shown on the right.

Following the FD hypothesis, we assumed two modes of exploration in TF searching process; in the recognition mode, TFs associate DNA tightly with the binding energy dependent on the preferred specific motif signature descried by PWM; in the search mode, TFs weakly bind to DNA with a combination of unspecific electrostatic attraction (Gerland, 2002; Halford, 2004) and sequence-dependent weak interaction (Slutsky & Mirny, 2004). The model and kinetics rates for the searching process is depicted in **Figure 2.20**.

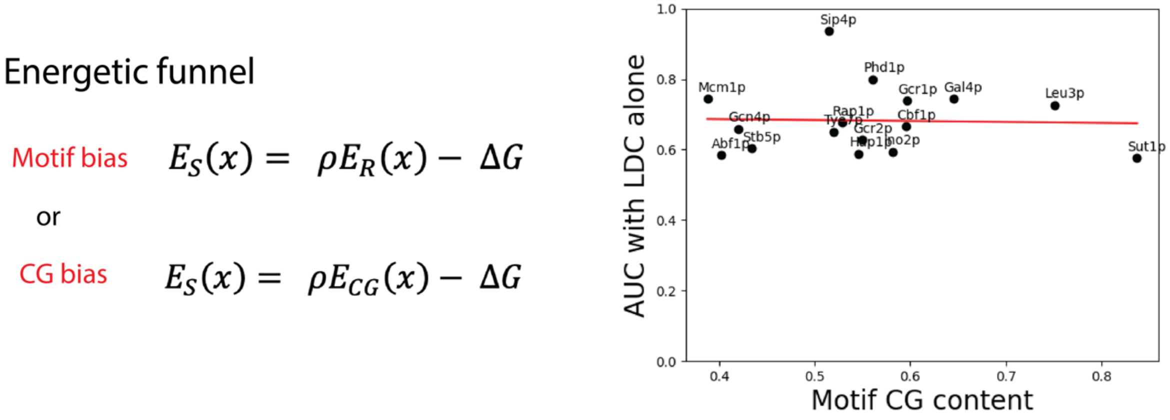


Figure 2.21 Two possible mechanisms, motif bias and CG bias, for TF search mode binding. No correlation was found between the CG content in TF motif and improvement in binding prediction with LDC information, suggesting the improvement in binding prediction with using LDC is not related to motif bias at searching step.

A study by Cencini showed the AT rich landscape at target binding site serves as energic funnel in *E.coli* (Cencini, 2018), and such funnel can increase the probability of TF binding to the target. In that study, the sequence-dependent contribution at search mode was assumed to be proportional to the specific binding at recognition mode. As TFs in *E.coli* are prone to bind at AT rich sequences, it is reasonable to model the energetic funnel relative to the AT bias within the TF motif. However, the underlying mechanism is not necessarily linked to the bias within the TF motif preference but is a rather general pattern for a large group of TFs at search process.

In our study, we showed CG richness preferred in LDC by TF binding in eukaryotic organism S. cerevisiae., and this phenomenon holds true both *in vivo* and *in vitro* conditions. To examine if this CG preference is a result of CG bias in the TF motif, we compared the binding prediction improvement with LDC information alone and the CG content within TF motif, and no correlation is found (**Fig 2.21**). Therefore, we concluded the energetic funnel in S. cerevisiae is not proportional to the TF motif bias at search mode as observed in (Cencini, 2018). Moreover, we compared the CG frequency bias landscape of the bound set to the unbound set for TFs with well-defined motifs, and we saw a general CG bias (see reference (Cencini, 2018) for the detailed b(r) calculation) with a funnel shape peaked at the target motif (**Fig 2.22**). Thus, we reformed the two modes FD model with the binding energy at searching mode as the CG content in a window size of the length for a given TF motif instead of letting search mode binding energy being a fraction of recognition mode binding energy.

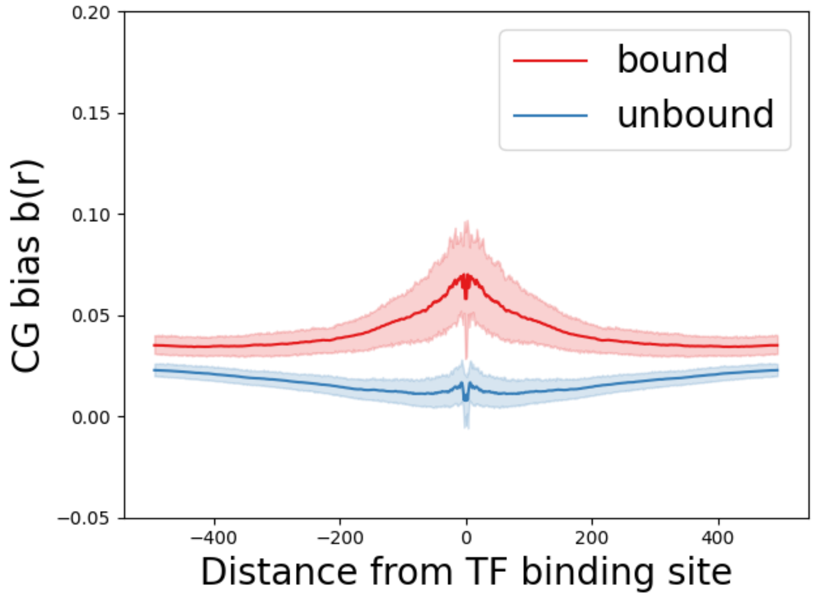


Figure 2.22 CG bias in LDC between TFBS bound and unbound sets. The CG bias around TFBS was approximated for overall bound and unbound sets with all TFs with well-defined motifs, and a funnel shaped CG bias was observed for the TFBS bound set.

We further simulated the two modes FD process with Gillespie algorithm (Gillespie, 1976) in order to verify the relationship between the CG bias in LDC and TF binding. For every simulation, we took 500bp native genomic region upstream and downstream of the target TFBS as the LDC, and we placed the TF within 250bp either side of the target site randomly to initiate the search process. The success rate is approximated with 100 simulations for every sequence, and the LDC score is calculated using the Universal Markov Chain model constructed in the previous section, and a significant positive correlation with r equals to 0.51 is shown (**Fig 2.23**), suggesting the CG bias in LDC can lead to higher probability to locate a TFBS.

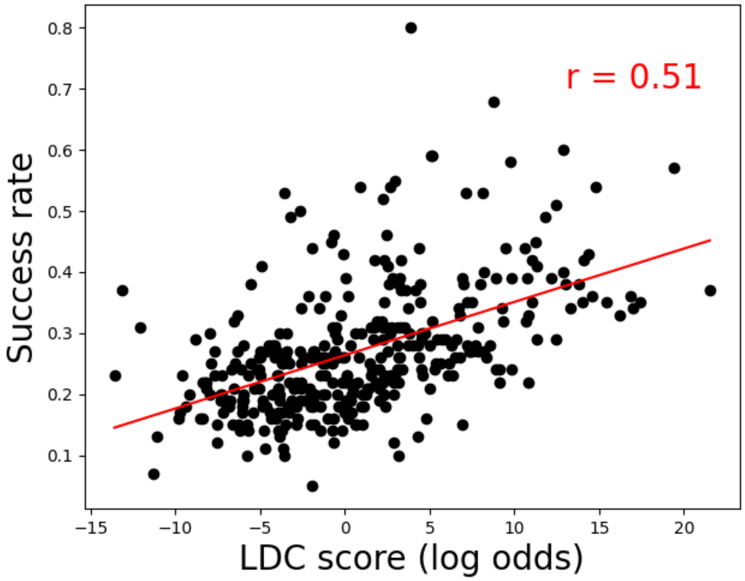


Figure 2.23 Correlation between LDC score and the success rate of TF finding a target TFBS simulated by Gillespie algorithm. Gillespie simulation was performed on TFBS containing sequences to verify the relationship between the LDC score of the flanking DNA sequences around TFBS from the Universal Markov Chain model and the success rate of TF locating the target binding site with the proposed FD mechanism. The test sequences were sampled from TFs with well-defined motifs with equal representation from all TFs.

2.2.7 LDC Improves Binding Prediction for Human TFs

All previous analysis was performed on TF binding data in S. cerevisiae, a simple unicellular eukaryotes organism, which only involves a small set of TFs for cellular regulations. To appreciate roles of LDC in a more complex environment, we expand our LDC analysis to human cells. From ENCODE consortium, we acquired 258 ChIP-seq data in K562 condition, and we further divided them into TFs with DNA binding motifs and TFs without known motifs according to CIS-BP database (**Fig 2.24**)

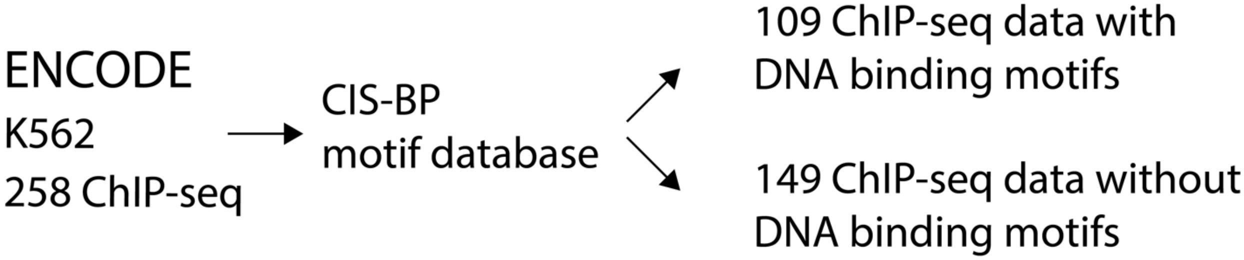


Figure 2.24 The overview of ENCODD human ChIP-seq analysis. 258 K562 ChIP-seq data were divided into TFs with motif information and TFs without known binding motif according to CIS-BP database.

To examine whether the LDC of TFBS in human can predict TF binding, we processed them the same way as was done for Figure 2.1 for TFs with well-defined motifs, where the presence and coordinates were identified and 125bp upstream and downstream LDC was taken for analysis. To obtain appropriate searching space for negative set, 2000bp upstream of coding genes were defined as promoter regions and searched with given motif information and took 125bp either side of TFBS for negative LDC. 5-folds cross validation was performed and evaluated with AUROC; a various level of improvement in binding prediction was observed for these TFs ranging from little advance to AUC of 0.9. For detecting the signature of LDC, we determined the dinucleotide fold change between the TF bound set and unbound set. CG rich feature of LDC is favored by most TFs; interestingly however, we also noticed there are some TFs prefer the opposite LDC feature, CG depleted and AT rich sequences. To understand the relationship of

LDC signature and binding prediction, the correlation between CG dinucleotide fold change and the AUC results from 5-fold cross validation was plotted (**Fig 2.25**); the more extreme of CG fold change in either direction was correlated with higher improvement in AUC. For TFs without known motif information, the centered 250bp binding peaks from ChIP-seq were treated as the positive dataset, and a random sampling of 250bp sequences from all promoter regions in human were treated as negative dataset. The same comparison for CG fold change in LDC and prediction results in AUC was shown for TFs without known motif information (**Fig 2.26**), and the results resembled the results for TFs with specific motifs, suggesting a diverse bias towards LDC for human TFs.

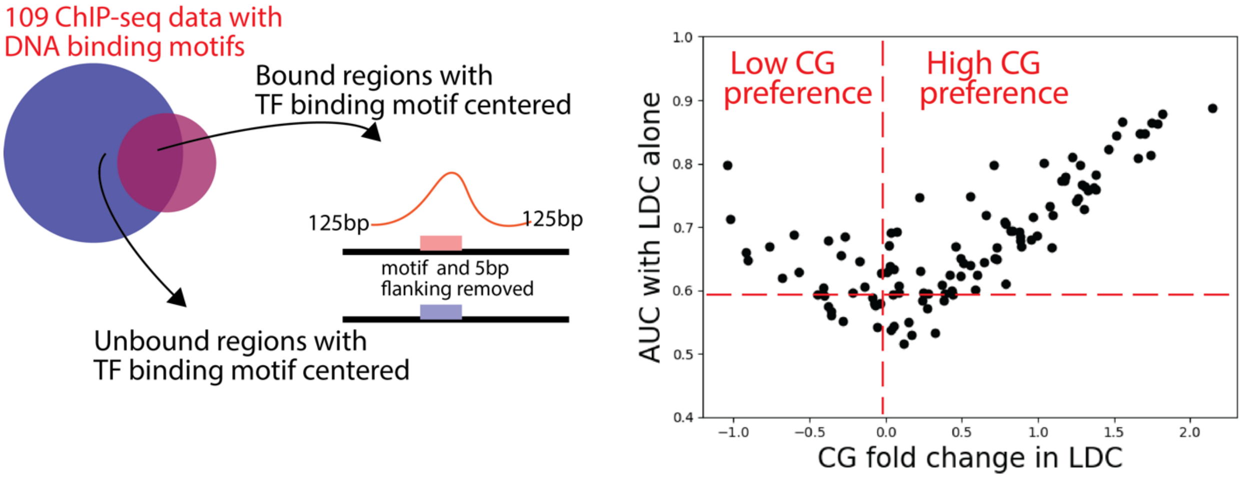


Figure 2.25 Comparison between CG dinucleotide fold change in LDC and AUC for binding prediction. For all ChIP-seq data with DNA binding motifs information, the TFBS were searched in the binding peaks and 125bp flanking DNA upstream and downstream of the site were taken as positive dataset for LDC analysis. TFBS for the given TF were also search in all promoter regions and the flanking DNA either side of the site were used as negative dataset. 5-fold cross validation were performed for constructing the Markov Chain model and binding prediction; the AUC result was compared to the CG dinucleotide fold change in LDC shown on the right.

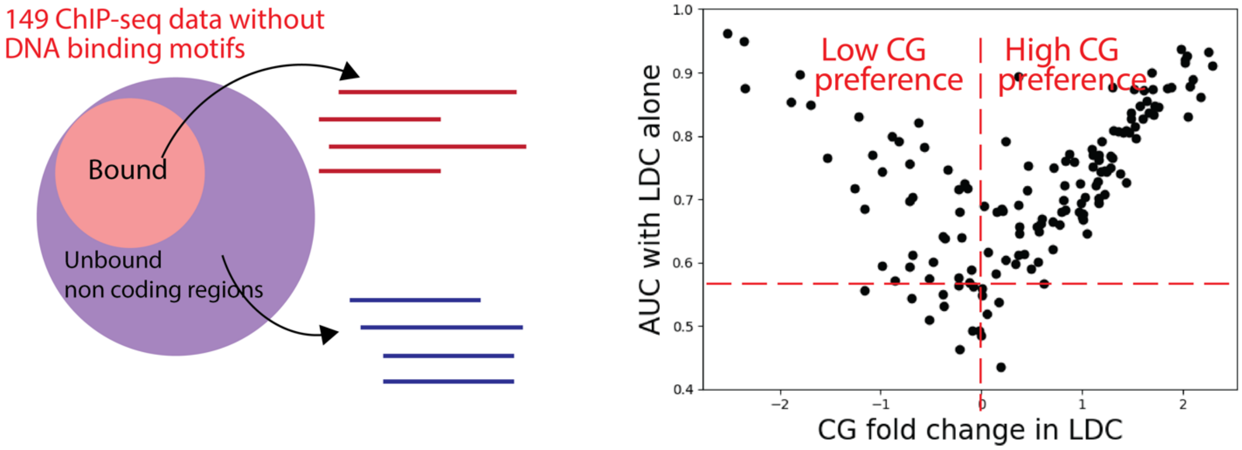


Figure 2.26 Comparison between CG dinucleotide fold change in LDC and AUC for binding prediction for TFs without DNA binding motifs. For all ChIP-seq data without DNA binding motifs information, the center 250bp binding peaks were treated as positive dataset for LDC analysis. A random sampling of equal sized promoter regions for the given TF were used as negative dataset. 5-fold cross validation were performed for constructing the Markov Chain model and binding prediction; the AUC result was compared to the CG dinucleotide fold change in LDC shown on the right.

We summarized the analysis for individual TF into three categories, TFs favoring CG rich LDC (168 TFs), TFs favoring CG depleted LDC (40 TFs), TFs with unclear LDC preference (50 TFs). The signature of LDC for TFs in these groups in the view of dinucleotide fold change between bound and unbound were shown (**Fig 2.27 & Appendix 2.3**). With grouping these TFs according their LDC preference, we made CG rich and CG depleted Universal Markov Chain model with combining all TFs in each category. We demonstrated the Universal models improve binding prediction with comparative level as single model made from individual TF, suggesting general LDC preference across TFs within groups (**Appendix 2.4**), with a few outliers which could be a result of misclassification for the TF.

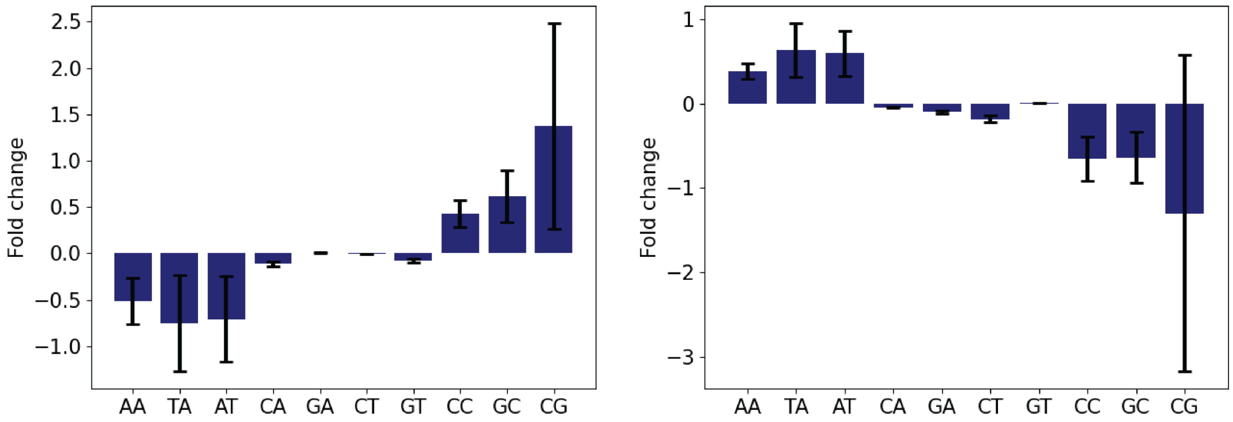


Figure 2.27 The overall dinucleotide fold change for the grouped TFs with opposite LDC preference. Left) Overall dinucleotide fold change between bound and unbound sets for TFs prefer high CG in LDC. Right) Overall dinucleotide fold change for TFs that prefer low CG in LDC.

2.3 Discussion

The direct and specific interaction between TF and the short DNA motif has been the focus of understanding TF binding specificity; in this study we revealed the local DNA environment can also contribute to the regulations of TF binding in the way of helping TF locate its target binding sites during searching process. In the beginning of the study, we first demonstrated the presence of predictive information embedded in LDC for TF binding. Such predictive information exists and the signature of LDC is coherent and independent of intrinsic nucleosome characteristics among all analyzed TFs. We showed the binding prediction with TFBS score can be further improved by incorporation with LDC score. To understand the role of LDC in TF binding mechanism, we investigated *in vitro* binding data and found the preference of the same LDC signature in purely protein-DNA interaction. Furthermore, we related the CG richness around

TFBS in LDC to energetic funnel in the TF search process and showed the CG richness can theoretically enhances the probability of TF recognizing a target site. Lastly, we expand our analysis to a large collection of human TFs and identified similar preferred LDC for most human TFs.

The connection of local DNA to regulations of TF binding has been discussed in some studies. For example, nucleosome disfavoring sequence in close relation to the functional TF binding site was implicated to increase TF occupancy (Segal, 2006; Raveh-Sadka, 2012; Levo, 2015). Moreover, the presence of additional TF motif, of the same kind or from other TFs, can enhance or alter the binding outcomes (Shively, 2019; Liu J. S., 2020; Panne, 2007; Levo, 2015). These examples conveyed the intricate the complex regulations of TF binding that involves DNA as the intermediate platform for protein-protein interactions. Nevertheless, we extend the scope of LDC in the mechanism of TF binding; the surrounding DNA environment can directly influence TF binding without recruiting trans-factors, but through the kinetics effect while TF searches for a target motif to bind. In our model, the gradient of CG content around the binding sites contributes to the weak sequence-dependent interaction with TFs at searching step; it serves as energetic funnel to guide TF in recognizing a functional motif by retaining the TF on DNA strand for longer residence time.

From study by Pal et al, where at the last cycle of HT-SELEX experiment (i.e., highest affinity oligos remain), there exists promiscuous ‘shapemers’ that are generally enriched across TFs regardless of TF families, which implicates the possibility of non-motif specific background binding (Soumitra, Jan, & Teresa, 2019). This finding aligns with our hypothesis that LDC facilitates TF binding by energetic funnel effect and reducing the chances of TF falling off the DNA strand. We analyzed those promiscuous ‘shapemers’ in terms of the CG content and we found those top promiscuous ‘shapemers’ have higher CG content compared to non-enriched ‘shapemers’ (**Appendix 2.5**). Additionally, the observation made by (Hartl D, 2019) suggested CpG island enhances TF binding independent of methylation, which is consistent with our study and extend the perspective of CpG island usage in TF binding.

There is some discrepancy in terms of the underlying mechanism that environment DNA utilize to help TFs locating their target sites between our study and other related studies (Dror, 2015; Cencini, 2018). The studies by Dror et al and Cencini et al showed nucleotide bias correlations between the core binding motif and the flanking DNA; however, we did not find this correlation. In our yeast TF analysis, there is a flat relationship between the nucleotide composition in the core motif and the increase of predictive power in LDC (**Figure 2.21**); and similarly, in our ENCODE ChIP-seq analysis, no correlation was found between the nucleotide composition in the human core binding motif and the predictive power in LDC or the CG dinucleotide fold change, a comparable parameter to the CG content in LDC (**Appendix 2.6**). This discrepancy could be a result of TF sampling between ours and others, yet a well-controlled experiment should be performed to dissect the relationship between the core motif and LDC usage. A distinctive result would be meaningful to understand the molecular basis of the kinetic mechanism at TF searching process. Specifically, interaction involving DNA binding domain can be inferred in the case of LDC nucleotide bias is correlated with core motif bias, whereas in the case of no such correlation is found would indicate other protein domain related mechanism. Intrinsic disordered domain (IDR) has been linked with many TF regulation processes (Erik W. Martin, 2020; Sabari, 2018) and TF binding specificity (Brodsky, 2020), and could be a possible protein domain that TFs employ while exploring LDC for target binding sites.

In the last section of the study, we investigated a large collection of human TFs, where we discovered that most TFs tend to bind CG rich sequences, yet there is also a group of TFs prefers to bind at CG depleted sequences, and such unique nucleotide composition preference is not directly linked to TF families or the similarity of the DNA binding domain in our analysis. This diverse observation raises a possible mechanism that there may be several regimes in terms of the preferred signature of LDC. TFs that share similar LDC preference are more likely to bind cooperatively and function together. We believe this discovery opens a new venue to identify groups of TFs that tend to bind in proximity.

2.4 Materials and Methods

2.4.1 Binding Data Collection and Preprocess

***In vivo* binding data for *S. cerevisiae*.** The *in vivo* binding data in this study comprises of 15 TFs from Calling Cards method (Wang H. M., Calling Cards enable multiplexed identification of the genomic targets of DNA-binding proteins, 2011; Shively, 2019) and 19 TFs from ChIP – exo method (Rhee, 2011); more detailed information is summarized in Figure 2.28. The data choice is on the basis of a study by Kang, et.al (Kang, 2020); according to the study, the binding data from both methods have higher correspondence to perturbation-response data than ChIP-chip binding data. The position weight matrices (PWM) for all TFs were obtained from ScerTF database complied by Spivak and Stormo (Spivak AT, 2012). One natural log below the ScerTF recommended PWM cutoff scores were used throughout this study to define the presence or absence of TF sites on DNA sequences.

Both Calling Cards data and ChIP-exo data were obtained directly from original publications in the format of genomic coordinates of binding events. For Calling Cards data, we applied Blockify (Moudgil, et al., 2020), a peak caller designed for Calling Cards experiments, to call binding peaks with default parameter setting. For ChIP-exo experiments, we merged significant binding locations within distance of 20 bp with bedtools (Quinlan & Hall, 2010) as a single binding peak in the analysis. For DNA sequence of S. cerevisiae, we used S288C reference in 2015 version downloaded from SGD (Cherry, 2012). All processed binding peaks for Calling Cards and ChIP-exo data are provided in the supplemental data.

We classified 34 TFs into two categories: 1) TFs with well-defined motifs and 2) TFs lack of specific motifs. To ensure the motif is informative and specific, we required the PWM for such motif to have information content greater than 8 and p-value less than 10e-5 from ScerTF database (Spivak AT, 2012) (**See Fig 2.28**).

	TFs with well-defined motifs	TFs lack of specific motifs	Publications
Calling Cards	Cbf1p, Leu3p, Gcr1p, Gcr2p, Gcn4p, Gal4p, Tye7p	Cst6p, Kar4p, Lee1p, Rgm1p, Rpi1p, Sef1p, Sfg1p, Yrm1p	Rhee and Pugh 2011

ChIP-exo	Abf1p, Hap1p, Ino2p, Mcm1p, Phd1p, Rap1p, Sip4p, Stb5p	Cat8p, Ert1p, Hap4p, Ino4p, Oaf1p, Pip2p, Reb1p, Rds2p, Rgt1p, Rtg3p	Wang et al. 2011; Shively et al. 2019

Figure 2.28. TF binding data categories and sources. The detailed information for the collection of TF binding data in this study.

In vitro* binding data for *S. cerevisiae. We analyzed *in vitro* binding data by Dip-chip method from (Noam Kaplan, 2009), and included Cbf1p, Leu3p, Pho4p, Rap1p and Swi5p in our data analysis, with Pho2p and Rox1p excluded from this study as these two factors do not contain specific motifs by our requirement.

To obtain binding peaks from Dip-chip data, we first took entries with binding signal two standard deviation higher than the mean, and we then merged those entries within 20bp distance with applying sum operation for binding signals from merged entries by bedtools. Finally, we kept merged coordinates with summed binding signals two standard deviation than median as the binding peaks for later analysis. The processed binding peaks are provided in the supplemental data.

Nucleosome occupancy data for *S. cerevisiae.* The nucleosome occupancy was attained from (Segal, 2006) which was measured by Mnase assay and reported as normalized values. To categorize genomic regions into nucleosome occupied and nucleosome free, we processed the data so that regions with consecutive negative nucleosome occupancy value were defined as nucleosome free regions, whereas regions with consecutive positive values were defined as nucleosome occupied regions. In Figure 2.15& Figure 2.16 where we wanted to know whether a TF motif is occupied by nucleosome, we identified the genomic coordinates of the TF motif and took the sum of the nucleosome occupancy value over the entire motif.

ChIP-seq data for Human TFs. Human TF ChIP-seq data were obtained directly from ENCODE (The ENCODE Project Consortium, 2012) in the format of bed narrowPeak. All ChIP-seq data were experimented on K562 cell line and assembled with hg19. All human TF motifs were acquired from CIS-BP database (Weirauch, 2014).

2.4.2 Local DNA Sequence Context Definition

For TFs with well-defined motifs, we searched for all binding motifs that are present on the intergenic regions of S. cerevisiae with requiring the motifs are at least one natural log below the recommended score provided by ScerTF. The local DNA context (LDC) is defined as 125 bp upstream and downstream of the motif; and to avoid any confounding effect such as DNA structural shape flanking the motif may have on TF binding, we further removed the motif itself with 5bp flanking on both sides. For all motifs of each TF, we divided them into TFBS that are bound and unbound by whether the motif is within the binding peaks by either Calling Cards assay or ChIP-exo assay.

For TFs without well-defined motifs, as there is no other known and predictive DNA feature to categorize intergenic regions into TF regulated regions or unregulated regions, we took the entire binding peaks as positive LDC, and the rest of all other intergenic regions as negative LDC. To obtain equivalent sets, we down sampled the negative sets to the same size of binding peaks. To make fair comparison to the group of TFs with well-defined motifs, only the 250 bp center for both positive binding peaks and negative intergenic regions were used for LDC score calculation.

2.4.3 Construction and Application of Markov Chain Model

To preserve the information embedded in the DNA context in a parsimonious way, we employed first order Markov Chain model. We calculated the transition frequency of every possible pair of nucleotides from the local DNA sequences defined at the previous section for both positive and negative datasets and constructed four by four matrices as the first order Markov Chain model. For a given unseen sequence with length(L), we can therefore calculate a relative likelihood of the given the sequence coming from the positive model, which is the sum of log odds for every consecutive dinucleotide (a_{xi-1xi}) between the positive and the negative Markov Chain model, with the higher value indicating higher probability (1).

$$S(x) = \log\frac{P(x|Model^{+})}{P(x|\, Model^{-})} = \sum_{i=0}^{L} log\,\frac{a^{+}_{xi-1xi}}{a^{-}_{xi-1xi}}$$

In Figure 2.1 and Figure 2.25 where we wanted to understand whether predictive information is in LDC, we performed 5-fold cross validation so that the Markov Chain model was made from training data and the model was evaluated using the unseen testing data. For the rest of the analysis, the LDC score was calculated from the Universal model that was constructed with all LDC combined from TFs with well-defined motifs in Figure 2.1 (see supplemental figures for the actual models).

2.4.4 ROC and PRC for Model Evaluation

To evaluate the performance of the Markov Chain model, standard receiver operator curve (ROC) which compares the ranking of the sequences from positive and negative datasets based on LDC score obtained from the Markov Chain model was used throughout the study, and Mann-Whiney test was performed to measure the significance of the prediction improvement. Both ROC and Precision recall curve (PRC) were applied for Figure 2.15 and Figure 2.16 which compares the probability of being bound returned by the logistical regression (LR). Sklearn python package was utilized for ROC and PRC calculation.

2.4.5 Incorporating LDC Score with PWM Score into Logistical Regression (LR) Model

In Figure 2.15 and Figure 2.16, the learning problem was defined to predict whether an intergenic region is bound by a TF or not, and we used LR algorithm for this classification problem with LDC score/Nucleosome occupancy (NuOc) and PWM score as predictors in the model. The detailed feature generation procedure is as follows: 1) for every intergenic region, we searched for all possible TFBS with PWM score of zero and above and took the highest score as the value for PWM score feature; 2) for LDC score, we identified the coordinates of the highest PWM, and then we applied the Universal Markov Chain model to 125bp upstream and downstream of the TFBS with 5bp directly flanking the motif removed to calculate LDC score; 3) for NuOc value, we summed over the normalized nucleosome occupancy from (Segal, 2006) for the highest scored TFBS on the intergenic region. Binding peaks from either Calling Cards or ChIP-exo were used to label each intergenic region as bound or unbound.

For this supervised learning problem, we performed 5-fold cross validation with LR algorithm to estimate the likelihood of a given intergenic region been bound or not, and we then used ROC and PRC that compares the probability of been bound produced by LR for model evaluation.

2.4.6 Kinetic Funnel Model and Gillespie Simulation

In the process of TF searching for a target site, we followed the facilitated diffusion (FD) hypothesis proposed by (Slutsky & Mirny, 2004; Berg OG, 1981) and assumed that TF can switch between recognition mode and search mode, and such switch is a result of conformational changes in TF-DNA complex. For rates equations, we maintained the fundamental structure and parameter setting of (Cencini, 2018) (equations 2.1 - 2.5). The change we made is the energy contribution in the search state; instead of letting the motif preference contributing proportional to sequence-dependent binding energy, we modulate the CG preference on LDC as the sequence-dependent contribution at search state (equation 2.6).

$$K_S^+ = \mathrm{D}\ e^{[E_S(x)-E_S(x+1)]/2} \qquad (2.1)$$

$$K_S^- = \mathrm{D}\ e^{[E_S(x)-E_S(x-1)]/2} \qquad (2.2)$$

$$K_{SR} = \gamma\ e^{\frac{[E_S(x)-E_R(x)]}{2}-\Delta G} \qquad (2.3)$$

$$K_{RS} = \gamma\ e^{\frac{[E_R(x)-E_S(x)]}{2}} \qquad (2.4)$$

$$K_d = \delta e^{E_S} \qquad (2.5)$$

$$E_S(x) = \rho E_{CG}(x) - \Delta G \qquad (2.6)$$

To approximate the rate of a TF recognizes a TFBS, we simulated this stochastic process using Gillespie algorithm (Gillespie, 1976). The target TFBS is centered with the native LDC of length 500bp either side. For each realization, the TF is initialized at search state and placed with uniform distribution in a region [-250, 250] with respect to the target site. The success of finding a target is defined as a TF switches to recognition mode at the target TFBS, and 100 simulations was done for estimating the success rate of each sequence. 320 sequences were sampled from 16 TFs with specific TFBS in Figure 2.23.

References

Accili, D., & Arden, K. C. (2004). FoxOs at the crossroads of cellular metabolism, differentiation, and transformation. *Cell*, 421-426.

Allemand, J. F. (1998). Stretched and overwound DNA forms a Pauling-like structure with exposed bases. *roceedings of the National Academy of Sciences of the United States of America*, 95(24), 14152–14157.

Baker, H. (1991). GCR1 of Saccharomyces cerevisiae encodes a DNA binding protein whose binding is abolished by mutations in the CTTCC sequence motif. *Proceedings of the National Academy of Sciences of the United States of America*, 88(21), 9443–9447.

Berg OG, W. R. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry*, 6929-48.

Berger, M. F. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology*, 1429–1435.

Berger, M. F. (2009). Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nature protocols*, 393–411.

Bradley, R. K. (2010). Binding Site Turnover Produces Pervasive Quantitative Changes in Transcription Factor Binding between Closely Related Drosophila Species. *Plos BIology*.

Bulyk, M. L. (2007). Protein binding microarrays for the characterization of DNA-protein interactions. *Advances in biochemical engineering/biotechnology*, 65–85.

Carlson CD, W. C. (2010). Specificity landscapes of DNA binding molecules elucidate biological function. *Proc Natl Acad Sci U S A.*, 107(10):4544-9.

Cencini, M. &. (2018). Energetic funnel facilitates facilitated diffusion. *Nucleic Acids Research*, 558-567.

Cherry, J. M. (2012). Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic acids research*, D700–D705.

Dang, C. V. (2012). MYC on the path to cancer. *Cell*, 22-35.

Davis, J. E. (2019). Multiplexed dissection of a model human transcription factor binding site architecture.

De Val, S. C. (2008). Combinatorial regulation of endothelial gene expression by ets and forkhead transcription factors. *Cell*, 1053–1064.

Dickerson, R. E. (1989). Definitions and nomenclature of nucleic acid structure components. *Nucleic acids research*, 17(5), 1797–1803.

Dror, I. G.-G. (2015). A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome research*, 25(9), 1268–1280.

Fong, A. P. (2015). Conversion of MyoD to a neurogenic factor: binding site specificity determines lineage. *Cell reports*, 1937–1946.

Fordyce, P. M. (2010). De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nature biotechnology*, 970–975.

Frey, F. S. (2016). Molecular basis of PRC1 targeting to Polycomb response elements by PhoRC. *Genes & development*, 1116–1127.

Garner, M. M. (1981). A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic acids research*, 9(13), 3047–3060.

Geertz, M. &. (2010). Experimental strategies for studying transcription factor-DNA binding specificities. *Briefings in functional genomics*, 362–373.

Gerland, U. M. (2002). Physical constraints and functional characteristics of transcription factor-DNA interaction. *Proceedings of the National Academy of Sciences of the United States of America*, 12015–12020.

Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 403-434.

Griggs, D. W. (1991). Regulated expression of the GAL4 activator gene in yeast provides a sensitive genetic switch for glucose repression. *Proceedings of the National Academy of Sciences of the United States of America*, 8597–8601.

Grossman, S. R. (2017). Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proceedings of the National Academy of Sciences of the United States of America*, E1291–E1300.

Halford, S. E. (2004). How do site-specific DNA-binding proteins find their targets? *Nucleic acids research*, 3040–3052.

Harteis, S. &. (2014). Making the bend: DNA tertiary structure and protein-DNA interactions. *International journal of molecular sciences*, 15(7), 12335–12363.

Hartl D, K. A. (2019). CG dinucleotides enhance promoter activity independent of DNA methylation. *Genome Research*, 554-563.

Hinnebusch, A. G. (1990). Transcriptional and translational regulation of gene expression in the general control of amino-acid biosynthesis in Saccharomyces cerevisiae. *Progress in nucleic acid research and molecular biology*, 195–240.

Hinnebusch, A. G. (2002). Gcn4p, a master regulator of gene expression, is controlled at multiple levels by diverse signals of starvation and stress. *Eukaryotic cell*, 22–32.

Hollenhorst, P. C. (2009). DNA specificity determinants associate with distinct transcription factor functions. *PLoS genetics*, 5(12), e1000778.

Inukai, S. K. (2017). Transcription factor-DNA binding: beyond binding site motifs. *Current opinion in genetics & development*, 110–119.

Junion, G. S. (2012). A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell* , 473–486.

Kanaya, E. N. (1999). Characterization of the transcriptional activator CBF1 from Arabidopsis thaliana. Evidence for cold denaturation in regions outside of the DNA binding domain. *The Journal of biological chemistry*, 274(23), 16068–16076.

Kang, Y. P. (2020). Dual threshold optimization and network inference reveal convergent evidence from TF binding locations and TF perturbation responses. *Genome Research*, 459-471.

Kaya-Okur, H. S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature communications*, 10(1), 1930.

Kim, Y. W. (2017). Deletion of transcription factor binding motifs using the CRISPR/spCas9 system in the β-globin LCR. *Bioscience reports*, 37(4), BSR20170976.

Kinney, J. B. (2010). Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences of the United States of America*, 9158–9163.

Klar, A. J. (1974). Studies on the positive regulatory gene, GAL4, in regulation of galactose catabolic enzymes in Saccharomyces cerevisiae. *Molecular & general genetics : MGG*, 203–212.

Kulkarni, M. M. (2003). Information display by transcriptional enhancers. *Development*, 130(26), 6569–6575.

Levo, M. Z.-P. (2015). Unraveling determinants of transcription factor binding outside the core binding site. *Genome research*, 25(7), 1018–1029.

Liu, J. S. (2020). Quantitative analysis of transcription factor binding and expression using calling cards reporter arrays. *Nucleic Acids Research*, e50.

Liu, X. L. (2006). Whole-genome comparison of Leu3 binding in vitro and in vivo reveals the importance of nucleosome occupancy in target site selection. *Genome research*, 1517-1528.

Maerkl, S. J. (2007). A systems approach to measuring the binding energy landscapes of transcription factors. *Science* , 233–237.

Majka, J. &. (2007). Analysis of protein-DNA interactions using surface plasmon resonance. *Advances in biochemical engineering/biotechnology*, 104, 13–36.

Maricque, B. B. (2017). A genome-integrated massively parallel reporter assay reveals DNA sequence determinants of cis-regulatory activity in neural cells. *Nucleic acids research*, 45(4), e16.

Mathelier, A. X. (2016). DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. *Cell system*, 278–286.

Menon, B. B. (2005). Reverse recruitment: the Nup84 nuclear pore subcomplex mediates Rap1/Gcr1/Gcr2 transcriptional activation. *Proceedings of the National Academy of Sciences of the United States of America*, 102(16), 5749–5754.

Meyer, N. &. (2008). Reflecting on 25 years with MYC. *Nature reviews. Cancer*, 976–990.

Mills, J. B. (2004). Origin of the intrinsic rigidity of DNA. *Nucleic acids research*, 32(13), 4055–4059.

Mirny, L. (2009). Nucleosome-mediated cooperativity between transcription factors. *nature precedings*.

Moudgil, A., Li, D., Hsu, S., Purushotham, D., Wang, T., & Mitra, R. D. (2020). The qBED track: a novel genome browser visualization for point processes. *bioRxiv*.

Noam Kaplan, I. K.-M. (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, 362–366.

Ong, C. T. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature reviews. Genetics*, 283–293.

Palmieri, M. S. (1999). Interaction of the nuclear protein CBF1 with the kappaB site of the IL-6 gene promoter. *Nucleic acids research*, 27(13), 2785–2791.

Panne, D. M. (2007). An atomic model of the interferon-beta enhanceosome. *Cell*, 1111–1123.

Park, P. (2009). ChIP–seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, 10, 669–680 .

Park, S. C. (2004). Determination of binding constant of transcription factor myc-max/max-max and E-box DNA: the effect of inhibitors on the binding. *Biochimica et biophysica acta*, 1670(3), 217–228.

Philippakis, A. A. (2008). Design of compact, universal DNA microarrays for protein binding microarray experiments. *ournal of computational biology : a journal of computational molecular cell biology*, 15(7), 655–665.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 841–842.

Raveh-Sadka, T. L.-P. (2012). Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nature genetics*, 743–750.

Rhee, H. S. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, 1408–1419.

Sabari, B. R. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. *Science* .

Segal, E. F.-M. (2006). A genomic code for nucleosome positioning. *Nature*, 772–778.

Shai R Joseph, M. P. (2017). Competition between histone and transcription factor binding regulates the onset of transcription in zebrafish embryos. *eLife*.

Sharon, E. K.-S. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature biotechnology*, 521–530.

Shen, N. Z. (2018). Divergence in DNA Specificity among Paralogous Transcription Factors Contributes to Their Differential In Vivo Binding. *Cell systems*, 470–483.

Shively, C. A. (2019). Homotypic cooperativity and collective binding are determinants of bHLH specificity and function. *Proceedings of the National Academy of Sciences*, 16143-16152.

Shultzaberger, R. K. (2007). Determining physical constraints in transcriptional initiation complexes using DNA sequence analysis. *PloS one*, 2(11), e1199.

Simon, I. B. (2001). Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, 697-708.

Skene, P. J. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife*, 6, e21856.

Slattery, M. Z. (2014). Absence of a simple code: how transcription factors read the genome. *rends in Biochemical Sciences*, 381-399.

Slutsky, M., & Mirny, A. L. (2004). Kinetics of Protein-DNA Interaction: Facilitated Target Location in Sequence-Dependent Potential. *Biophysical*, 4021-4035.

Soumitra, P., Jan, H., & Teresa, M. P. (2019). Co-SELECT reveals sequence non-specific contribution of DNA shape to transcription factor binding in vitro. *Nucleic Acids Research*, 6632-6641.

Spitz, F. &. (2012). Transcription factors: from enhancer binding to developmental control. *Nature reviews. Genetics*, 613–626.

Spivak AT, S. G. (2012). ScerTF: a comprehensive database of benchmarked position weight matrices for Saccharomyces species. *Nucleic Acids Res*, D162-8.

Stormo, G. (2000). DNA binding sites:representation and discovery. *Bioinformatics*, 16-23.

Stormo, G., Zuo, Z., & Chang, Y. (2015). Spec-seq: determining protein–DNA-binding specificity by sequencing. *Briefings in Functional Genomics*, 30–38.

Struhl, K. &. (2013). Determinants of nucleosome positioning. *Nature structural & molecular biology*, 267-273.

Tanay, A. (2006). Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Research*, 962-972.

The ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 57-74.

Tornow, J. Z. (1993). GCR1, a transcriptional activator in Saccharomyces cerevisiae, complexes with RAP1 and can function without its DNA binding domain. *The EMBO journal*, 12(6), 2431–2437.

Uemura, H. &. (1992). Role of GCR2 in transcriptional activation of yeast glycolytic genes. *Molecular and cellular biology*, 12(9), 3834–3842.

Ussery, D. W. (2002). DNA structure: A-, B- and Z-DNA HELIX FAMILIES. *Encyclopedia of Life Sciences*.

Vaquerizas, J. M. (2009). A census of human transcription factors: function, expression and evolution. *Nat Rev Genet*, 252-263.

Villa, R. S. (2016). PionX sites mark the X chromosome for dosage compensation. *Nature*, 244-248.

Wakabayashi, A. U. (2016). Insight into GATA1 transcriptional activity through interrogation of cis elements disrupted in human erythroid disorders. *Proceedings of the National Academy of Sciences of the United States of America*, 113(16), 4434–4439.

Wang, H. J. (2007). Calling cards for DNA-binding proteins. *Genome research*, 17(8), 1202–1209.

Wang, H. M. (2011). Calling Cards enable multiplexed identification of the genomic targets of DNA-binding proteins. *Genome research*, 748–755.

Wang, H. M. (2012). "Calling cards" for DNA-binding proteins in mammalian cells. . *Genetics*, 941–949.

Wang, J. C. (1979). Helical repeat of DNA in solution. *Proceedings of the National Academy of Sciences of the United States of America*, 200–203.

Wang, J. Z. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome research*, 1798–1812.

Warren, C. K. (2006). Defining the sequence-recognition profile of DNA-binding molecules. *Proceedings of the National Academy of Sciences of the United States of America*, 867-872.

Weirauch, M. T.-M. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 1431–1443.

White, M. A. (2013). Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proceedings of the National Academy of Sciences of the United States of America*, 11952–11957.

White, M. A. (2013). Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proceedings of the National Academy of Sciences of the United States of America*, 11952–11957.

Wu, Y. R. (1996). Quantitation of putative activator-target affinities predicts transcriptional activating potentials. *The EMBO journal*, 3951–3963.

Yan, C. C. (2018). Systematic Study of Nucleosome-Displacing Factors in Budding Yeast. *Molecular cell*, 294–305.

Yang, S. W. (1995). Comparison of protein binding to DNA in vivo and in vitro: defining an effective intracellular target. . *The EMBO journal*, 6292–6300.

Yen, M. Q. (2018). Transposase mapping identifies the genomic targets of BAP1 in uveal melanoma. *BMC Med Genomics* , 11, 97.

Zeigler, R. D. (2014). Discrimination between thermodynamic models of cis-regulation using transcription factor occupancy data. *Nucleic acids research*, 2224–2234.

Zeiske, T. B. (2018). Intrinsic DNA Shape Accounts for Affinity Differences between Hox-Cofactor Binding Sites. *Cell Reports*, 2221-2230.

Zentner, G. E. (2015). ChEC-seq kinetics discriminates transcription factor binding sites by DNA sequence and shape in vivo. *Nature* , 6, 8733.

Zhao, Y. G. (2009). Inferring binding energies from selected binding sites. *PLoS computational biology*, 5(12), e1000590.

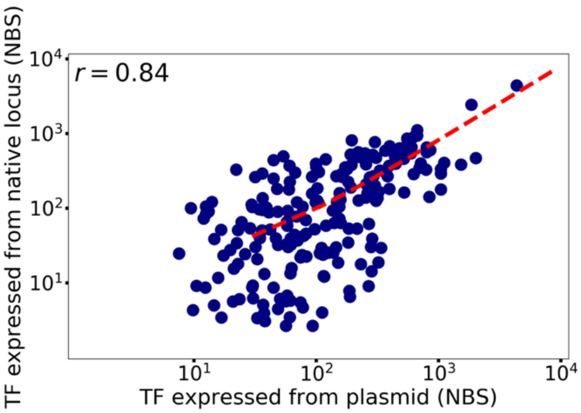
Zhou, T. S. (2015). Quantitative modeling of transcription factor binding specificities using DNA shape. *Proceedings of the National Academy of Sciences*, 4654-4659.

Zhou, X. &. (2011). Integrated approaches reveal determinants of genome-wide binding and function of the transcription factor Pho4. *Molecular cell*, 826–836.

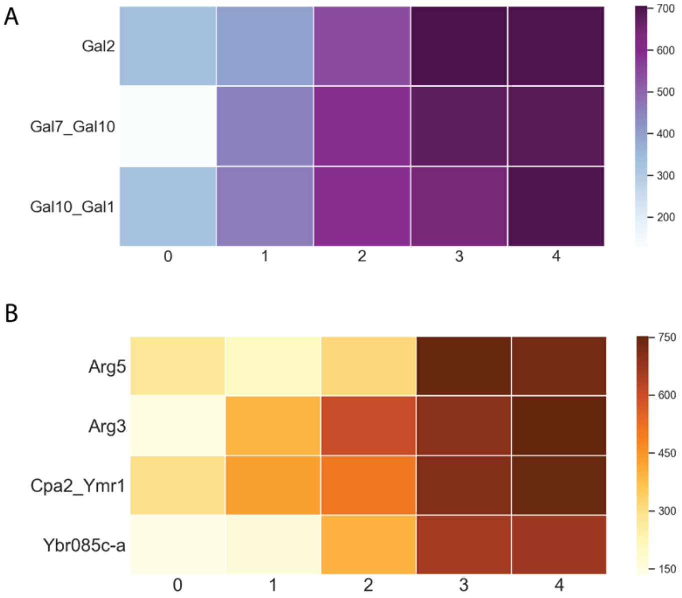
Zou, S. K. (1996). The Saccharomyces retrotransposon Ty5 integrates preferentially into regions of silent chromatin at the telomeres and mating loci. *Genes & development*, 10(5), 634–645.

Zykovich A, K. I. (2009). Bind-n-Seq: high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. . *Nucleic Acids Res*, 37(22):e151.

Appendix

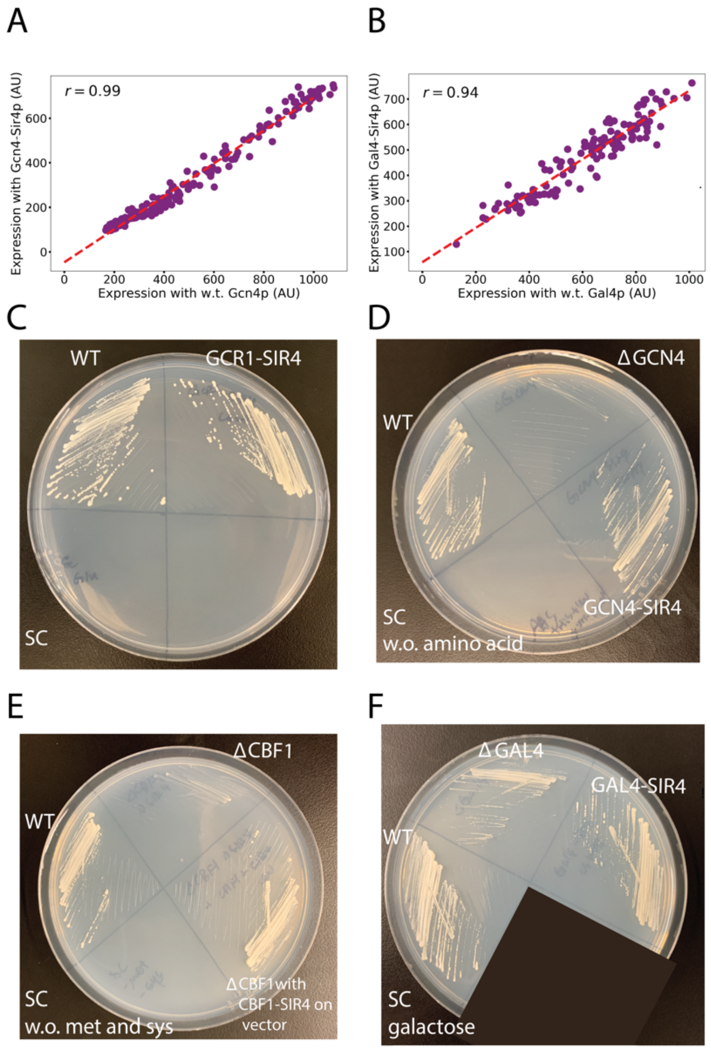


Appendix 1.1 Binding of Cbf1-Sir4p expressed from a plasmid is well-correlated with binding of Cbf1-Sir4p expressed from the native locus. The CCRA library used here was identical to the one used in **Figure 1.6** .



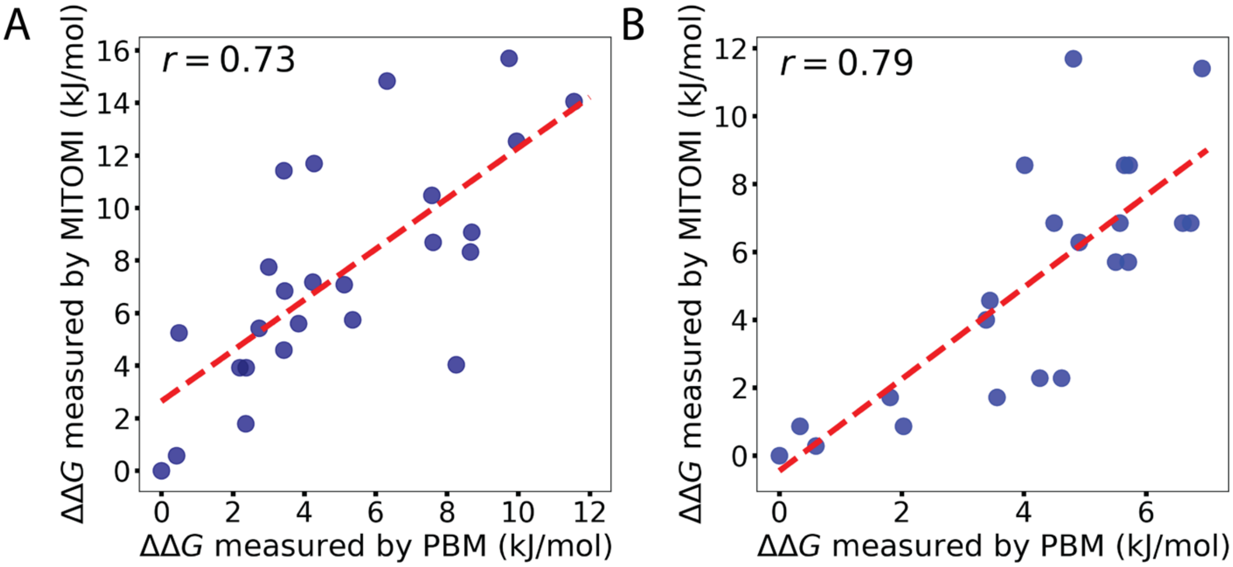
Appendix 1.2 Expression measurements were performed for a library consisting of synthetic promoters derived from Gal4p- and Gcn4p-regulated promoters. The number of corresponding motifs for each TF was varied in the library. **A)** reporter gene expression increases as the number of Gal4p motifs increases

under galactose condition. **B)** reporter gene expression increases as the number of Gcn4p motifs increases under amino acid starvation condition.

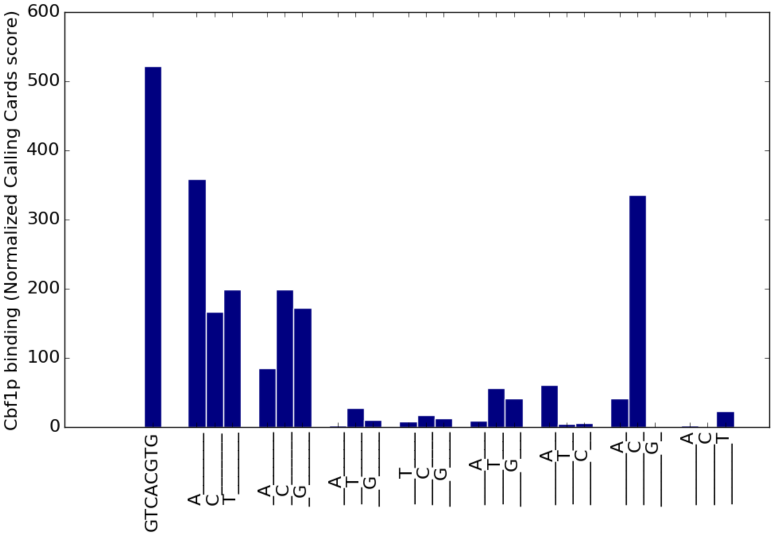


Appendix 1.3 Comparison of Sir4p tagged and untagged transcription factors. To determine if Sir4p tagged TFs produce the same Sort-Seq measurements of gene expression, we took the Gcn4p and Gal4p CCRA library and performed Sort-Seq in an untagged w.t. background and compared the results to those obtained with the tagged TFs. Expression measurements for wild-type and Sir4p-tagged **A)** Gcn4p, and

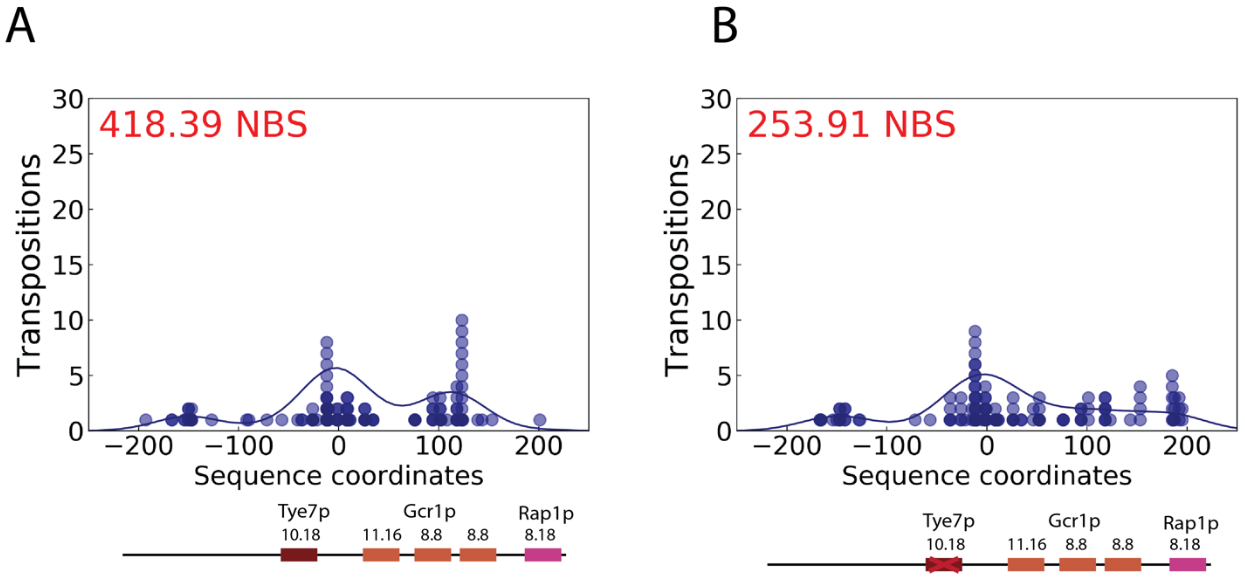
B) Gal4p were highly correlated; To determine whether the Sir4p affects TF function, we analyzed four different Sir4p-tagged TFs to see if they could rescue growth in a deletion strain grown under conditions where the TF is required. **C)** Gcr1p tagged with Sir4p is viable in yeast grown in SC; **D)** Gcn4p tagged with Sir4p is viable under amino acid starvation condition; **E)** Cbf1p tagged with Sir4p expressed from plasmid can rescue Cbf1p deletion strain under MET and CYS deficient condition; **F)** Gal4p tagged with Sir4p recovers the normal growth of yeast under galactose condition.



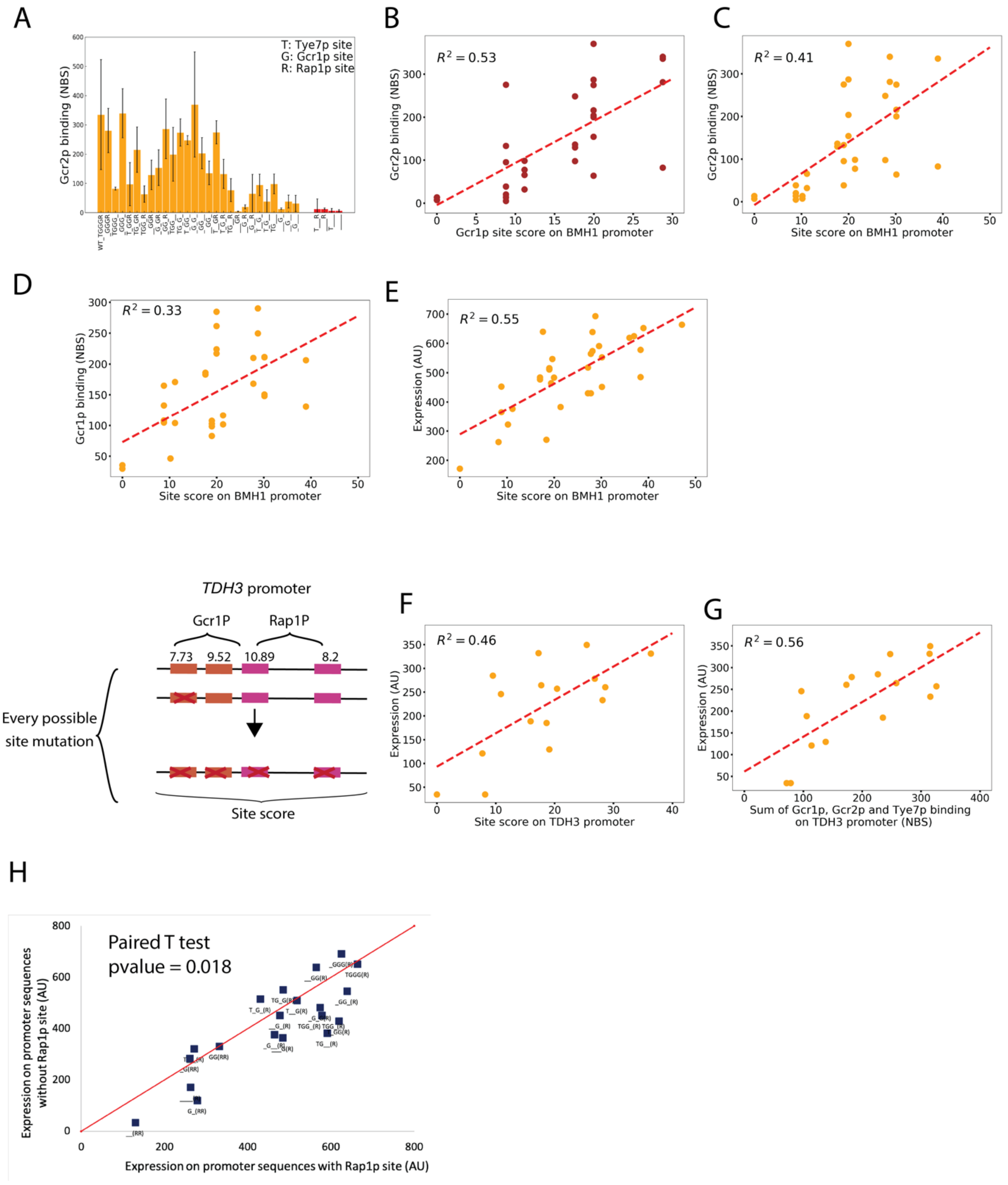
Appendix 1.4 A) Comparison of the change of binding energy measured by PBM and MITOMI for Cbf1p. **B)** The same as panel A, but with MAX transcription factor.



Appendix 1.5 Average *in vivo* genomic Cbf1p Calling Cards binding score on all alternative E-box motif. As expected from our CCRA binding energy landscape, mutations to the core CACGTG had a larger impact on Cbf1p binding than non-core motifs, but the effect was exacerbated in vivo, perhaps because of competition with nucleosomes. It is important to note that it would be impossible to generate accurate Cbf1p binding energies (benchmarked against *in vitro* measurements) solely from the *in vivo* binding data. This is because in the yeast genome, Cbf1p binding sites (and 1bp mutant sites) occur in a variety of different sequence contexts, whereas in the CCRA experiments, the Cbf1p sites were analyzed in precisely the same sequence context.

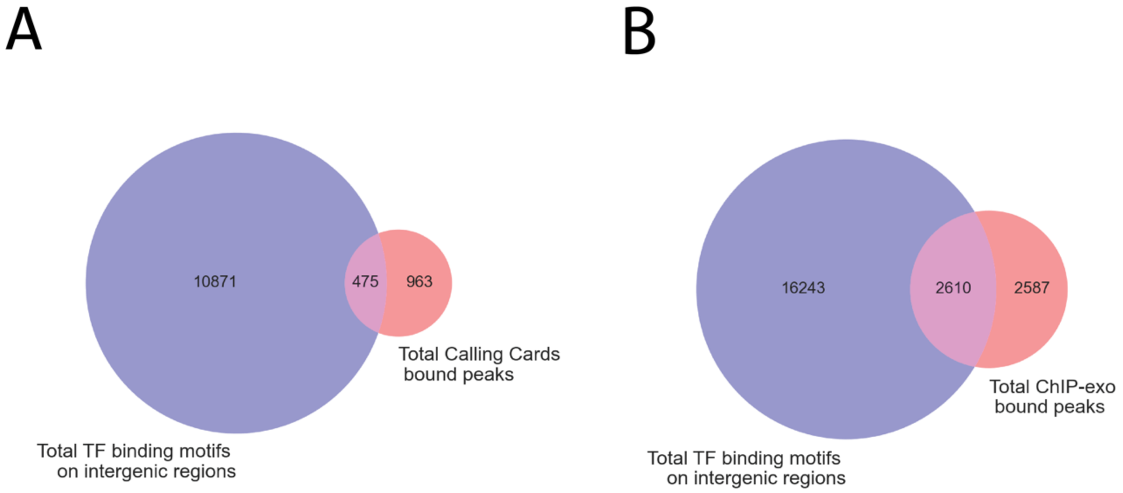


Appendix 1.6 Tye7p transposition distribution on **A)** w.t. *BHM1_pr* promoter and **B)** Tye7p motif mutated *BHM1_pr* promoter.

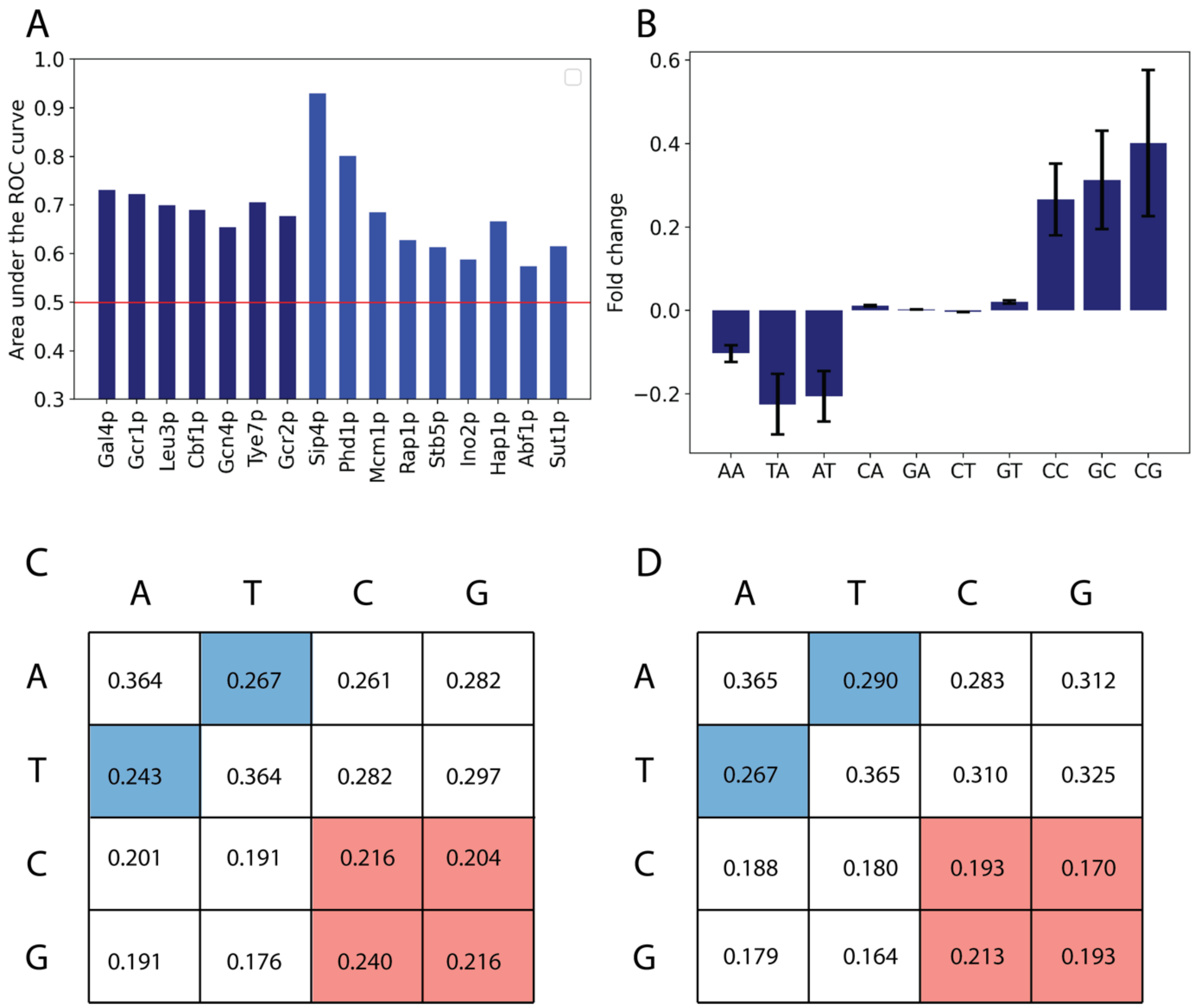


Appendix 1.7 A) The same as in **Figure 1.19** but with Gcr2p. **B)** The same as in **Figure 1.20** but with Gcr2p. **C)** Gcr2p binding was compared to the total PWM scores from all remaining sites, which has

weaker correlation than comparing to the total PWM scores from Gcr1/2p sites alone. **D)** The same as in **C)** but with Gcr1p. **E)** Expression was regressed against the PWM scores from the remaining sites on *BMH1* promoter. **F)** We mutated *TDH3* promoter the same as we did for *BMH1* promoter, and expression was regressed against site score on TDH3 promoter. **G)** Gcr1p, Gcr2p and Tye7p binding were also measured for *TDH3* promoter, and we regressed the expression against the sum of all binding. **H)** Expression for sequences from *BMH1* and *TDH3* promoters is divided pairs that is either with or without any Rap1p motif. The red line is a 1:1 diagonal line for clear visualization purpose. The expression for sequences with Rap1p motif is generally higher than those without any Rap1p motif. Paired T test was performed, and the p-value is 0.018.



Appendix 2.1 Intersection between TF binding peaks and TF motifs. **(A)** Blue circle represents the total number of TF binding motifs on intergenic regions from seven TFs, and red circle represents the total number TF bound peaks from these seven TFs by Calling Cards method. **(B)** The same as **(A)**, but with nine TFs measured by ChIP-exo method.



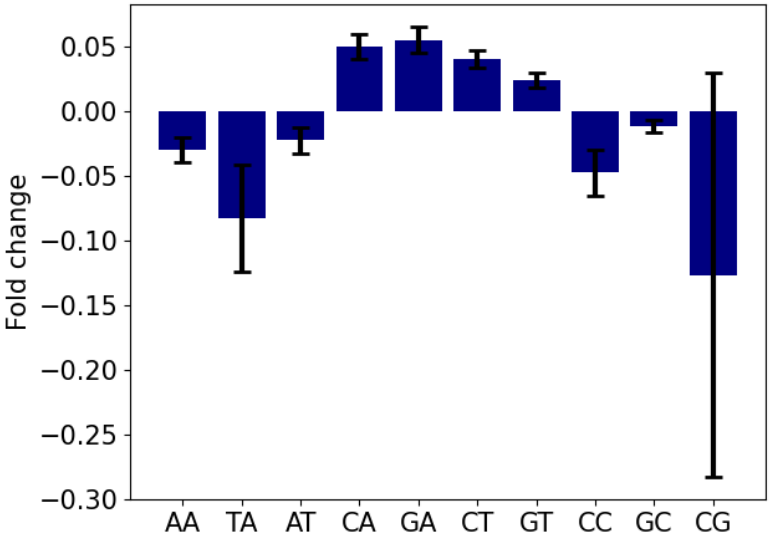
C

	A	T	C	G
A	0.364	0.267	0.261	0.282
T	0.243	0.364	0.282	0.297
C	0.201	0.191	0.216	0.204
G	0.191	0.176	0.240	0.216

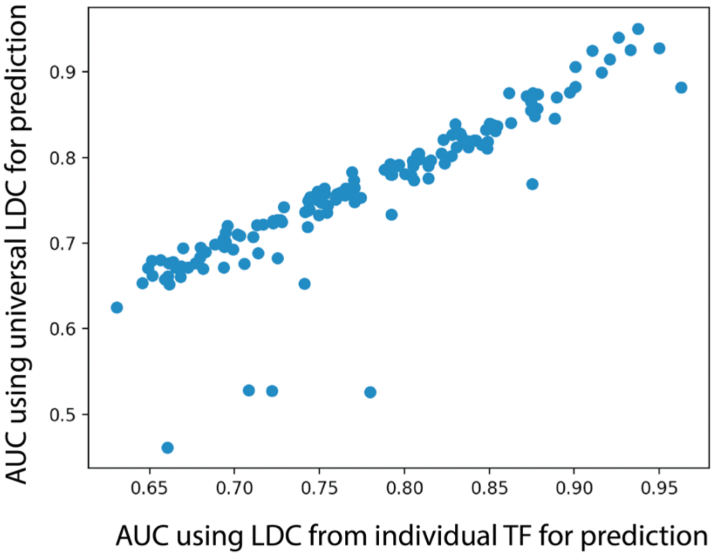
D

	A	T	C	G
A	0.365	0.290	0.283	0.312
T	0.267	0.365	0.310	0.325
C	0.188	0.180	0.193	0.170
G	0.179	0.164	0.213	0.193

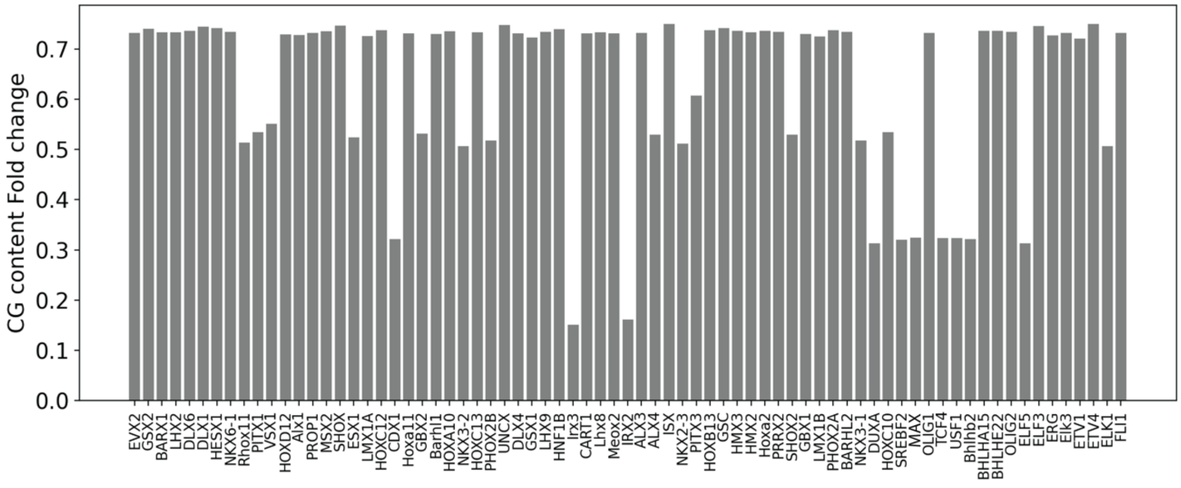
Appendix 2.2 **(A)** AUC for TF binding prediction. LDC from all other TFs was used to prediction for each TF. **(B)** Overall dinucleotide fold change in LDC from all 16 TFs combined. **(C)** The positive universal model (i.e., First order Markov Chain transition matrix). **(D)** The negative universal model.



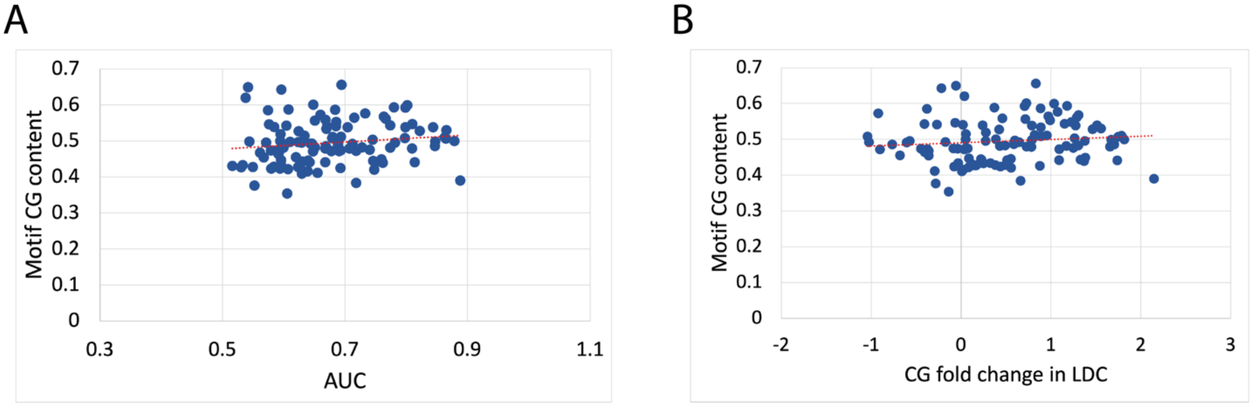
Appendix 2.3 Overall dinucleotide fold change between bound and unbound sets for TFs with unclear preference in LDC.



Appendix 2.4 ENCODE TFs AUC comparison between using LDC from individual TF and all TF combined by CG rich and CG depleted groups. Binding predictions with 5-fold cross validation on individual TF were compared to binding predictions with Universal model constructed from combined TFs with similar LDC preference.



Appendix 2.5 CG content fold change on promiscuous enriched 'shapemers' from HT-SELEX data. The CG content of top promiscuous 'shapemers' were compared to bottom 'shapemers', the fold change for each TF is shown.



Appendix 2.6 A) Comparison of nucleotide bias in motif and prediction for human TFs with well-defined motifs. **B)** Comparison of nucleotide bias in motif and CG dinucleotide fold change between bound and unbound TFBS in local DNA.