

# **Transcription Factor Binding Dynamics and Spatial Co-localization in Human Genome**



**Xiaoyan Ma**

Downing College

University of Cambridge, Department of Genetics

This dissertation is submitted for the degree of Doctor of Philosophy.

24<sup>th</sup> August 2017

# Preface

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text

It does not exceed the prescribed word limit for the relevant Degree Committee.

# Abstract

Transcription factor (TF) binding has been studied extensively in relation to binding site affinity and chromosome modifications; however, the relationship between genome spatial organisation and transcription factor binding is not well studied. Using the recently available high resolution Hi-C contact map of human GM12878 lymphoblastoid cells [1], we investigated computationally the genome-wide spatial co-localization of transcription factor binding sites, for both within the same type and between different types.

First, we observed a strong positive correlation between site occupancy and homotypic TF co-localization based on Hi-C contacts, consistent with our predictions from biophysical simulations of TF target search. This trend is more prominent in binding sites with weak binding sequences and within enhancers, suggesting genome spatial organisation plays an essential role in determining binding site occupancy, especially for weak regulatory elements.

Furthermore, when investigating spatial co-localization between different TFs, we discovered two distinct co-localization networks of TFs in lymphoblastoid cells, one of which is enriched in lymphocyte specific pathways and distal enhancer binding. These two TF networks have strong biases for either the A1 or A2 chromosome sub-compartment, but nonetheless are still preserved within each, indicating a potential causal link between cell-type-specific transcription factor binding and chromosome subcompartment segregation. We called 40 pairs of significantly co-localized TFs according to the genome wide Hi-C contact map, which are enriched in previously reported, physical interactions, thus linking TF spatial network to co-functioning.

In addition to the above main project, I also worked on a side project to find compute-efficient ways in scaling binding site strength across different TFs based on Position-Weight-Matrices (PWM). While common bioinformatics tools produce scores that can reflect the binding strength between a specific TF and the DNA, these scores are not directly comparable between different TFs. We provided two approaches in estimating a scaling parameter  $\lambda$  to the PWM score for different TFs. The first approach uses

a PWM and background genomic sequence as input to estimate  $\lambda$  for a specific TF, which we applied to show that  $\lambda$  distributions for different TF families correspond with their DNA binding properties. Our second method can reliably convert  $\lambda$  between different PWMs of the same TF, which allows us to directly compare PWMs that were generated by different approaches.

## Acknowledgements

I would like to thank Dr. Boris Adryan first. When I was new to Cambridge and was very poor in expressing myself, it was him who guided me through step by step, making research plans and arming me with computational insights to solve various types of biological questions. He made me changed from an undergraduate student with little research experience to the one that is confident in conducting independent PhD project. He put great emphasis on research independence, giving us sufficient freedom to explore things we were interested in, but at the same time, was also keen to encourage collaborations, either within group or across institutions. Furthermore, I was very lucky to be able to work with great group members, especially Daphne Ezer and Dr. Nicolas Radu Zabet. Daphne and I worked on a project of transcription factor binding dynamics simulations together, a process I did enjoyed. She was always there to offer me help and eager for discussions related to any topic at any time, even after her graduation. Radu was very helpful in guiding me through facilitated diffusion as well as GRiP by the time I knew nothing about this field. By working together with Daphne and Radu, I had gradually learned how to plan things ahead and work efficiently on project.

I am very grateful to Prof. Alfonso Martinez-Arias who helped me out in the second half of my Ph.D and gave me a warm home of research. More importantly, he taught me about stem cell and embryogenesis. He is always very patient in listening to me and has provided me lots of helpful suggestions in my project. I especially appreciate his emphasis on investigating a system from multiple layers– from genes, cells, further to organoids; from basic physical mechanisms including diffusion reaction to morphogenesis. Though my project was quite restricted to a small part: the nuclei, I was constantly reminded that biological systems function as a whole. I was also deeply inspired by interesting works going around in the group, related to gastruloids and the mechanistic perspectives of development. I am grateful for time spent with all lab members, who are always very nice and helpful. Especially, I would like to thank Shlomit Edri, Meritxell Vinyoles and Penny Hayward for their help in molecular biology experiments; Vikas Trivedi for his useful discussions and also reading my

thesis; Tina Balayo for lots of help in equipment; Naomi Moris, Peter Baillie-Johnson and David Turner for their helpful suggestions.

In addition, I would like to say huge thanks to Dr. Tim J. Stevens at MRC Laboratory of Molecular Biology, who introduced me to Hi-C contact maps, deeply inspired me and guided me through to explore spatial co-localisation across different transcription factors. Because Tim was located quite far from our department, throughout the last two and a half years, most of day-to-day communication was via email, but it never stopped us exploring interesting biological questions together. When discussing questions face to face, he was always very patient in explaining everything and listening to my 'slow English', though sometimes it took extremely long in the tea room so that people around were all finished with tea and gone.

I would like to thank Prof. Steve Russell for being very supportive throughout my Ph.D. He kept encouraging me now and then and was always eager to chat with me about research progress, either good news or difficulties. In addition, I thank Rob Foy particularly for his help in R; Carmen Navarro, Yermek Aitenov, Jamie McGinn, Dr. Aylwyn Scally and Bettina Fischer for their generous help before; Miriam Duncumb for reading part of my thesis. Finally, I would like to offer special thanks to Edward Kinrade, who, although no longer with us, continues to be remembered by us for his hard-working and kindness.

I have also received lots of help on my Ph.D projects from many people outside of the department, some of them are even outside of Europe. I thank Dr. Steffen Rulands (Max Planck Institute, previously in Cavendish Laboratory) for his help in biophysical models, Dr. Srinjan Basu (Biochem) for his help in molecular cloning, Justin Malin and Prof. Sridhar Hannenhalli (University of Maryland) for their collaboration in developing simulation models related to molecular crowd-sourcing, Dr. James G. McNally (Helmholtz Center Berlin) for answering questions about protein diffusion coefficients and residence times, Dr. Erez Lieberman Aiden (Baylor College of Medicine) for his help in Hi-C maps.

Additionally, I would like to take this opportunity to thank Dr. Stein Aerts and Prof.

Steve Russell again for giving me a challenging but very helpful viva examination. Their comments and suggestions were very valuable to the improvement of my thesis and also to my future research.

I also thank CSC scholarship from Chinese Scholarship Council for their support in my Ph.D during the last three years.

Lastly, I would like to thank my parents and all my families for all their love and great support. Even though I was away from home for four years during my Ph.D, they cared about me, encouraged me and helped me out of difficulties all the time.

Xiaoyan Ma  
22nd Oct 2017

# Contents

<b>1</b>	<b>Literature review</b>	<b>12</b>
1.1	Genomic and bioinformatics approaches of evaluating transcription factor binding . . . . .	12
1.1.1	Transcription factors are key regulators of gene expression . .	12
1.1.2	Genome-wide mapping of TF binding sites . . . . .	14
1.1.3	Evaluating TF binding specificity: different approaches . . . .	16
1.1.4	Evaluating TF binding specificity:Position Weight Matrices .	17
1.1.5	Sequential clustering of TF binding sites . . . . .	19
1.2	Transcription factor binding dynamics: modelling and tracking . . . .	21
1.2.1	Facilitated diffusion mechanism of TF target site search . . . .	21
1.2.2	Towards consistent estimation of TF diffusion and binding kinetics parameters in live imaging . . . . .	24
1.3	Hi-C: an approach to study genome organisation . . . . .	27
1.3.1	Introduction to chromosome confirmation capture techniques .	27



1.3.2	Hi-C contact map normalization approaches . . . . .	29
1.3.3	Chromosome organisation and TADs structures revealed by Hi-C	31
1.3.4	Possible mechanisms of TADs and the associated loops formation	33
<b>2</b>	<b>Transcription factor binding dynamics and occupancy is influenced by co-localization of homotypic sites</b>	<b>37</b>
2.1	Introduction . . . . .	37
2.1.1	Discrepancy between the predicted and the experimentally-derived TF binding landscapes . . . . .	37
2.1.2	Binding dynamics simulations predicted accelerated target search associated with homotypic BS clusters . . . . .	38
2.1.3	Research aim . . . . .	40
2.2	Methods . . . . .	43
2.2.1	Biophysical model of TF 3D diffusion and jumping to BSs in spatial proximity . . . . .	43
2.2.2	Incorporation of TF jumping into Gillespie algorithm based simulations . . . . .	46
2.2.3	TF Binding sites Annotation . . . . .	48
2.2.4	Quantification of Spatial clustering of TF binding sites within the same type . . . . .	49
2.2.5	ChIP-seq NarrowPeak SignalValue comparison between paired binding sites . . . . .	51
2.3	results . . . . .	52

2.3.1	Simple scenarios of TF binding dynamics simulations in two binding site clusters of spatial proximity . . . . .	52
2.3.2	3D organizations of homotypic binding sites . . . . .	54
2.3.3	TF binding site occupancy has strong linear correlation with spatial homotypic BS clustering . . . . .	57
2.3.4	Homotypic BS co-localization lead to ChIP-seq SignalValue gain	66
2.4	Discussion . . . . .	69
<b>3</b>	<b>Trancription factor spatial co-localization networks in lymphoblas-</b>	
	<b>toid cells</b>	<b>74</b>
3.1	Introduction . . . . .	74
3.2	Methods . . . . .	78
3.2.1	Heterotypic co-localization score between TF pairs . . . . .	78
3.2.2	Calling significantly co-localized TF pairs . . . . .	79
3.2.3	TF BS conservation between two cell lines . . . . .	80
3.2.4	Calculation of integrated heterotypic co-localization score for a group of TF . . . . .	80
3.3	Results . . . . .	81
3.3.1	TFs were partitioned into two spatial co-localization networks in GM12878 . . . . .	81
3.3.2	TF spatial co-localization is closely related to TF-TF physical interaction . . . . .	88

3.3.3	TFCN1 and TFCN2 have markedly different binding preference in A1 and A2 chromosome subcompartments . . . . .	96
3.3.4	TFCN1 and TFCN2 exist independent of TF occupancy differences within A1/A2 subcompartments . . . . .	97
3.3.5	BSs in TFCN2 are less centred around TSS or CTCF . . . . .	101
3.3.6	Revisit the relationship between homotypic BS co-localization and BS occupancy with regards to TFCN1 and TFCN2 respectively . . . . .	102
3.3.7	Intra-TFCN heterotypic BS co-localization significantly enhances TF binding, while the role of inter-TFCN BS co-localization varies between enhancer and promoter regions . . . . .	103
3.3.8	TF binding sites co-localization in human embryonic stem cells	109
3.4	Discussion . . . . .	112
3.4.1	Cell-type specific TF co-localization network and its relationship with chromosome subcompartments . . . . .	113
3.4.2	TF binding in relation to heterotypic BS co-localization in two co-localization networks . . . . .	115
3.4.3	The influence of Hi-C protocol and sequencing depth on Hi-C contact map quality and data analysis . . . . .	116
<b>4</b>	<b>Proper scaling of Position Weight Matrices to enable TF binding strength comparison across different TFs</b>	<b>118</b>
4.1	introduction . . . . .	118
4.2	Methods . . . . .	121

4.2.1	PWM matrices for TFs for yeast, fly and vertebrates . . . . .	121
4.2.2	Simple equation to calculate $\lambda$ . . . . .	122
4.2.3	Estimating $\lambda$ of a new PWM matrix for the same TF based on the residence time landscape of the facilitated diffusion model	125
4.3	Results . . . . .	127
4.3.1	Estimating scaling parameter $\lambda$ for binding site affinity across different species and TF families based on Equation 6 . . . . .	127
4.3.2	Comparison of $\lambda$ values estimated with Equation 4.6 to $\lambda$ values derived from fitting ChIP-seq data . . . . .	131
4.3.3	Converting $\lambda$ between different PWM matrices of the same TF	133
4.4	Discussion . . . . .	135
<b>5</b>	<b>Conclusion, discussions and future directions</b>	<b>147</b>
5.1	Conclusion . . . . .	147
5.2	Discussions and future directions: Hypothesis for chromosome domain and subcompartment formations . . . . .	149
5.3	List of Publications . . . . .	152

# Chapter 1

## Literature review

### 1.1 Genomic and bioinformatics approaches of evaluating transcription factor binding

#### 1.1.1 Transcription factors are key regulators of gene expression

Eukaryotic genomes contain thousands of protein coding genes. Some are house keeping genes, while the rest are precisely controlled and are only expressed in a subset of tissues and cell lineages at certain time points or under specific stimuli [2, 3, 4]. Deciphering differential gene expression in development and disease requires identification and characterization of gene regulatory elements [5, 6, 7].

At the transcriptional level, differential gene expression has been shown to be achieved through the combinatorial binding of different regulatory proteins to specific genome regions, either near the transcription start sites or in distal genomic regions such as distal enhancers [8, 9]. There are certain regulatory proteins that are common to the transcription initiation machinery of most genes [10, 11], however, the majority

of regulatory proteins contribute to gene expression regulation in a genomic context specific manner [12, 13, 14]. Such proteins often recognize and bind to specific DNA sequences [15, 16, 14, 17], and are known as sequence-specific transcription factors (TFs). Each TF can be involved in gene regulation of many target genes, especially in higher eukaryotes. Deciphering complicated TF regulatory networks in different organisms has been extensively studied [13, 18, 14].

Various studies have tried to investigate the transcriptional response to differentiated TF binding in different model organisms [8, 19, 20]. In budding yeast, Kim *et al* used the PHO5 promoter to characterize the quantitative relationship between TF binding and gene expression output [8]. Their model, which involved variable interactions between TFs, nucleosomes and the gene promoter, successfully recovered the observed changes in gene expression due to variations in TF binding site composition, suggesting a deterministic role of TF binding in gene regulation.

In *Drosophila melanogaster*, Kaplan *et al* employed a thermodynamic model to describe the binding of 5 key TFs and further investigated their role in embryonic anterior-posterior patterning [19]. They demonstrated that the inclusion of chromatin accessibility landscape significantly improved their model prediction of TF binding. In fact most TFs, except known pioneer factors, can only bind to corresponding sites within open chromatin depleted of nucleosomes. These are often referred to as DNase-I hypersensitivity hotspots (DHS) [21, 22].

He *et al* developed statistical thermodynamics-based models of gene expression to study a number of mechanistic questions including the combined effect of multiple regulatory proteins and the action of repressors [23, 24]. In terms of how multiple activator sites contribute to expression, their model spoke in favour of synergistic activation, where the total effect of multiple sites is larger than the sum of their individual effects. In synergistic activation, individual statistical weights of activator sites are multiplied in the sense that all bound sites contribute to the energy term associated with basal transcription machinery. They observed a better fit to expression data using synergistic activation rather than additive effects [23]. They further introduced the concept of short range repression into thermodynamic models, where

repressors do not directly interact with the basal transcriptional machinery, but instead, act to switch chromatin from accessible to inaccessible when bound nearby, thereby blocking the binding of activators within several hundred base pairs. Their results clearly excluded the hypothesis of competitive binding being the main mechanism of repression, but were consistent with the hypothesis of short range repression for most *Drosophila melanogaster* TFs used in their simulations [23].

The binding of TFs to well-studied enhancer regions has also been shown to quantitatively correlate with gene expression in *Drosophila melanogaster*. Another study by Kim *et al* [25] tried to re-arrange some of the minimal stripe enhancers, namely MSE2 and MSE3, upstream of the gene promoter in the *even-skipped* locus by introducing or removing small spacers. They observed that associated gene expression in the 2/3 inter-stripe regions was enhanced by more than 10 folds[25]. By adopting a simple thermodynamic TF binding model, taking into account the activation and repression effects of known TFs as well as cooperative binding between different TFs, the observed changes in gene expression level can be reasonably well-explained. Kim *et al* further demonstrated that using the fitted parameters from the above two enhancers, their model was able to correctly predict other enhancer-driven gene expression levels involving the same set of TFs at other genes. This suggests similar underlying mechanisms exist for gene regulation across different target genes. It provided further support for the fact that binding of TFs to their target sites constitutes a sufficient set of regulatory input which determines gene expression output quantitatively, at least for well-studied genomic loci. Admittedly, for the vast majority of genes without sufficient annotation of cognate enhancers and well-characterized TF functions, how TF binding relates to gene transcription remains largely unexplored.

### 1.1.2 Genome-wide mapping of TF binding sites

TFs play a central role in controlling gene expression together with nucleosome-mediated mechanisms and DNA modifications [8, 19]. Therefore, mapping TF-DNA interactions genome-wide is an essential task in understanding transcription regula-

tion. Chromatin immunoprecipitation (ChIP) based methods [26, 27] and DNase-I digital footprint are the two main tools currently in use to achieve this [28].

Chromatin immunoprecipitation (ChIP) is a technique that captures DNA fragments to which a specific regulatory protein binds [29]. Previously, using microarray-based techniques, DNA fragments collected from ChIP were hybridized to a microarray to detect genome regions enriched in specific protein-DNA interactions, named ChIP-chip [30, 27, 31]. Some arrays were designed to cover the entire genome, while others were only focused on specific genome regions such as selected promoters or gene sets.

With the advent of next generation sequencing, chromatin immunoprecipitation followed by sequencing (ChIP-Seq) became possible [32, 26, 33, 34]. ChIP-Seq offers better resolution and greater coverage compared to ChIP-chip, because TF bound fragments can be sequenced directly [35, 36, 34]. In ChIP-seq, the TF of interest is first crosslinked to DNA using formaldehyde before the chromatin is sonicated into small fragments. Using specific antibodies to the TFs, the cross-linked TF-DNA complexes can be immunoprecipitated [35]. After reverse crosslinking and size selection of fragments, the immunoprecipitated DNA is sequenced and mapped back to the genome [37, 38]. Using certain peak calling algorithms, enriched genome regions can be identified as ChIP-seq peaks [39, 40, 41].

ChIP-seq allows the mapping of TF binding to certain chromatin regions down to several hundreds of base-pairs [34]. However, the exact position of binding cannot be revealed simply through the annotated peaks. A specific binding motif identification step is required to infer potential sequence motifs overlapping with ChIP-seq peaks [33, 42].

Alternatively, DNase-I digital footprint (DNase-seq) is capable of identifying putative binding sites (BSs) at single base-pair resolution [21, 43]. It makes use of the fact that bound molecules can protect DNA from being cleaved by DNase-I, thus a BS often appears to be a 5 - 20 bp region within two DNase-seq peaks. This is known as a footprint [21]. DNase-I digital footprint provides the precise position of a TF BS, however, there is no information about which type of molecule is bound directly.



Further computational analysis is essential to correct the DNase-seq signal for experimental artefacts such as DNase-I cleavage bias. Overlaying the identified significant footprints with the predicted TF binding sequence motifs is also required to infer the type of TF that is likely to bind [44, 45].

Both ChIP-seq and DNase-I digital footprint have their own caveats, some of which can be reduced to improve the accuracy of BS identification [44]. For example, in ChIP-seq, false-positive peaks can occur as a result of an artefact in the cross-linking procedure [39]. Genome regions that are not directly bound by TFs but somehow tethered together with real targets may also be sequenced. Subsequent TF motif finding steps can help to identify high-confidence real targets from the enriched regions [33]. Admittedly, there are other biases in ChIP-seq that are linked to peak-calling algorithms; fragment selection in library preparation and amplification etc. Those Bias can be removed or reduced using proper controls and peak callers [36, 34, 39].

### **1.1.3 Evaluating TF binding specificity: different approaches**

TF binding sites specificity need to be determined from a reasonable size of experimentally verified binding sites (BSs) collection. Two distinct classes of approaches have been employed to generate those BS collections. For a subset of TFs, especially the well-studied ones involved in model organisms development, sufficient number of BSs could be obtained from annotated cis-regulatory elements. Functional roles of those BSs had been checked carefully, so those BSs made up high-confidence sets of annotated BS pool [46, 47].

Alternatively, for the rest of TFs, BS collection can be gained from high-throughput procedures to select high affinity BSs in vitro [48, 49, 50]. These procedures can generate collections of high affinity BSs, but also with sufficient sequence diversity. For instance, SELEX started from a pool of random DNA sequences and used the binding preference of a TF to select DNA fragments that were of higher binding affinity. After several rounds of selection and subsequent sequencing, DNA sequences

with strong binding preference could be highly enriched in the final product [51, 48].

Protein-binding micro-array (PBM) is another way to determine binding specificities of individual TFs [50]. An epitope tagged TF was purified and then bound directly to a double-stranded DNA microarray. With subsequent washes and labelling with fluorophore-conjugated antibody, the final read-out of micro-array fluorescence signal can be used to quantify TF binding specificity in a high-throughput way. In addition, yeast one-hybrid (Y1H) assay is another choice for detecting TF-DNA physical interactions [52]. By coupling Y1H system to a microfluidics-based protein-DNA interaction mapping assay, Y1H identified TF-DNA interactions can be further validated [53].

Different approaches have their own advantages and caveats, for example, PBM is only able to identify short DNA binding motifs of 8-mers due to the design limitations of their platform [50]. When quantifying the contribution from each nucleotide or di-nucleotide to the total binding specificity, SELEX sometimes gave a worse performance than PBM or Y1H, because the final product in SELEX reflects DNA fragment enrichment after multiple rounds of selection, which might amplify noises as well, rather than a direct representation of protein-DNA interaction strength [48].

#### **1.1.4 Evaluating TF binding specificity: Position Weight Matrices**

Given a repertoire of experimentally derived binding sites for a TF, either via high through-put in-vitro methods like SELEX and PBM, or experimentally validated functional sites collections stored in PAZAR [54] and RedFly [55], there needs to be a consistent and easily-adopted representation of DNA binding preference for this TF, which is referred to as a motif. The simplest representation would be a consensus sequence motif describing the most common bases, for example, AACNGT for a TF named Prd, where N stands for any base. However, most TFs do not just bind to a single DNA motif, but instead, can bind to a big repertoire of similar sequences with different binding strength. Even the mostly conserved nucleotides within consensus motifs could have alternatives in certain cases, albeit with declined binding affinity.

Therefore, quantitative modelling of TF binding preferences that can reflect specific degrees of flexibility at each position in the sequence motif is further required. Position Weight Matrices (PWMs) are the most widely adopted approach in modelling TF binding sequence preferences [56, 57, 58].

Proposed by Berg and von Hippel in 1987 based on statistical mechanics theory, the PWM model can be viewed as a probabilistic representation of binding sites [56]. They showed that the logarithms of the base frequencies is proportional to the binding energy contribution of the bases, which coincided with the idea of Relative Entropy in estimating binding strength from an information theory perspective. This nicely related the base-pair occurrence frequency in a set of known binding sequences to the binding free energy contributions from each base pairs, though it was based on an assumption that each position contributes independently and additively to the total binding energy [59].

Specifically, given a collection of known binding sites and the genome composition of the four bases (ATCG) in a specific organism, the Relative Entropy (also known as Information Content) at each position  $i$  of the motif can be described as [56, 58]

$$I_i = \sum_b g_{b,i} \log_2 \frac{g_{b,i}}{f_b} \quad (1.1)$$

where  $i$  is the position within the motif,  $f_b$  is the frequency of base  $b$  in the whole genome and  $g_{b,i}$  is the observed frequency of each base at position  $i$ .

Under the assumption of random background genome sequences and the additivity assumption of binding energy contribution from each base pair, Berg *et al* has showed that  $\log_2 \frac{g_{b,i}}{f_b}$  is a maximum probability estimate for the binding energy contribution of each base at each position, given the collection of known binding sites for TF [57]. The binding energy representation across all possible binding sites with a formula similar to partition function was involved in their original derivation of this maximum probability estimate. Therefore, the weight of  $\log_2 \frac{g_{b,i}}{f_b}$  is assigned to base  $b$  at position  $i$ , which gives the PWM score for each base at each position. In addition,  $\sum_i^L I_i$  is

the average binding energy of all validated binding sites ( $l$  represents sequence motif length).

One limitation of the PWM approach is the assumption that each base pair contribute to the total binding energy independently [56]. There are more complicated models taking into account the di-nucleotide or tri-nucleotide arrangement and various shape features of the binding motifs [60, 61, 62], but those models require more prior information and sometimes compute-intensive calculations. It has been argued that the di-nucleotide model is already sufficient to recover the in-vitro binding energy estimation made using protein-DNA interaction assay on a micro-fluid platform, which out-performed the simple PWM estimation for 3 TFs within the Helix-loop-helix family [63, 64]. However, due to the simplicity and the well-defined statistical physical energy term associated with it, PWM is still the most widely used approach to characterize TF binding specificity.

### 1.1.5 Sequential clustering of TF binding sites

TF binding sites are not randomly distributed in the genome, but rather prefer to cluster together sequentially [65, 66, 67]. Many studies have looked into one dimensional, sequential TF co-localization in the genome from different organisms, where ChIP-seq profiles were available. Co-occurrence of multiple types of TF BSs has been widely observed and used for predicting cis-regulatory elements and their potential functions [12, 68].

In particular, clustering of binding sites for the same type of TF (homotypic clusters) has been widely observed in both the *Drosophila* and human genome [69, 70, 65], with great enrichment in gene promoters and enhancers. The sequential homotypic clustering of BSs has several known advantages. First, it could arise from functional requirements of sequences [65, 71], since it can provide functional redundancy and regulatory robustness. Second, in terms of evolution, enhancers or promoters that harbour multiple sites of the same TF, especially weak BSs, are favoured by evolutionary sampling of the genotype-phenotype landscape [72]. He et al demonstrated

that more ways exist in their simulated evolutionary process to establish a fit genotype with many weak BSs than with a few strong sites, thus explaining why the occurrence of multiple weak BSs for the same TF within a single enhancer was widely adopted by different organisms. These weak BS clusters are also associated with deep evolutionary conservation as was revealed by the comparison between human, mouse and rat orthologous promoters, suggesting their positive selection during evolution [73]. In addition, from a perspective of TF binding and target search mechanisms, it facilitates TFs molecules to search for their target BSs via facilitated diffusion [74]. This will be discussed in more detail in the following sections.

In addition to homotypic BS clustering, specific types of TF BSs were observed to locate in proximity to each other within the genome much more frequently than expected by chance [67]. Using certain motif clustering algorithms and ENCODE ChIP-seq data sets, TF interaction networks based on overlapping ChIP-seq peaks and sequential proximal BS pairs have been derived [66, 67]. Their results identified many known TF-TF physical interactions, for example, the interaction between JUN, JUND and FOS, FOSL2. It suggested that TF BS sequential co-localization is closely linked to TF physical interactions, and was hypothesized to facilitate the co-binding of interacting TFs to recruit transcription machinery involving RNA PolII together [66]. Further, TFs known to function together in cis-regulatory modules have higher chances to cluster together along the genome [12, 75]. One of the typical examples is the *even-skipped* locus in the *Drosophila melanogaster* genome, where several different types of TFs including Bicoid, Hunchback, Kruppel, Knirps and Giant co-localize together to provide precise expression regulation in different embryonic body positions (*e.g.* stripe 3, 5 and 7) [25, 46]. The same TF may function as either activator or repressor in different genome context [25].

Therefore, the analysis of sequential TF co-localization has provided useful insights into TF physical interactions and co-function in gene regulation. However individual genome sequences do not give any information about BS spatial distribution. Given that the eukaryotic genome is organised into well-defined structures, whether TF binding sites are or are not clustered in three dimensions has not yet been investigated.

## 1.2 Transcription factor binding dynamics: modelling and tracking

### 1.2.1 Facilitated diffusion mechanism of TF target site search

Although there are millions of potential TF binding sites in the genome, true binding sites only constitute a minute fraction of the genome. In bacteria, certain types of TFs have less than 100 molecules within a cell[76]. In eukaryotes TFs are more abundant, estimated to be between a thousand to several hundred thousand molecules within nuclei according to different studies [77, 19]. Given the size of the eukaryotic genome, the chance of a TF protein incidentally meeting its target site is still extremely small. Decades ago, it was believed that DNA binding proteins located their target binding sites purely through diffusion driven mechanisms, where association rate follows the classical Smoluchowski limit. However, Riggs *et al* first observed that the rate at which the lac repressor identified its target site was much faster than one would predict by the classical model, and thus proposed a different mechanism which could account for this. [78].

Subsequently, with contributions from Winter and Berg, a comprehensive theoretical framework describing the TF target search process observed in E.coli was established: facilitated diffusion [79, 80]. In their model, TFs and other sequence-specific DNA binding molecules combine three-dimensional diffusion and one-dimensional random walk along the DNA to search for specific binding sites in the genome. Before the final arrival to a target BS, TFs alternate between the 3D diffusion phase and the 1D random walk phase, the rate of which depends on the dissociation constant from DNA for a specific TF. One of the most important biophysical parameters involved in this process is the so called the average sliding length; the typical length of DNA which a TF explores during one episode of random walk along the DNA [80, 56]. The average sliding length is intrinsic to TFs and needs to be determined via either Single Molecule Tracking (SMT) or in-vitro assay of single-molecule diffusion on noncognate DNA monitored by CCD camera [81]. Another study revealed that TFs tend to have higher

association rates to longer pieces of synthesized DNA, compared to shorter ones, while both types of synthesized DNA fragments all contain the same TF binding site in the middle [82]. It is because facilitated diffusion enables TFs to be transferred to the target site via sliding, even from positions far away from target BSs, thus the observed rate of TF association for the entire fragment of DNA is positively correlated with fragment length [83]. This work provided another piece of direct evidence supporting the existence of facilitated diffusion [82].

Both theoretical calculations and in-vivo tracking of TF molecules demonstrated that facilitated diffusion can speed up the TF target search process by 2 orders of magnitude less time than one would expect by simple unbiased protein diffusion in three-dimensional space [56, 76]. The proportion of time a TF spends on 3D diffusion and 1D random walk along the DNA can vary substantially between different type of TFs. This may be attributed to different structural properties within DNA binding domains across different families of TFs, or alternatively as a consequence of measurement bias across various studies [84, 85].

Analytical calculations based on several parameters could give an approximated order of the mean search time for a protein to find its target [86]. Such parameters include the 3D diffusion co-efficient of a specific type of TF in the nucleoplasm, the space volume of a target search, the 1D diffusion co-efficient along the DNA and the average sliding length along the DNA during each episode of 1D diffusion. Through dimensional analysis, one can roughly estimate the mean time a TF spends on 3D diffusion, DNA sliding, and also the mean number of 1D-3D search rounds. However, such analytical models often view DNA as either randomly coiled polymers or self-avoiding polymers, which is unlikely to be the real conformation of DNA within nuclei. These models also ignore inter-segmental transfers, in which the protein jumps directly between two pieces of DNA far apart in the genome, but close in physical proximity [87, 86].

Therefore, to better characterize the facilitated diffusion process with respect to BS distributions, protein inter-segmental transfers (jumping) and the interplay between TF and DNA chain conformation, several simulation models were developed to focus

on different aspects of the target search in attempt to correlate simulation results with experimental observations [88, 89]. Loverdo *et al* quantified the distribution of 3D re-location distances and time intervals of protein jumping on an ideal DNA chain. Zabet *et al* built a computational simulation frame work taking into account the real BS landscape in the genome and enabled user-defined properties of TF concentration, sliding length and specific and non-specific residence times *etc* [90]. Several studies further incorporated crowding molecules and road blocks surrounding real BSs, which better mimics the in-vivo case. Brackley *et al* found that macromolecular crowding can influence mechanistic features such as the proportion of time a TF spends on the 3D search versus 1D sliding, but the total average search time turned out to be very robust [91, 92].

Foffano *et al* has further investigated the facilitated diffusion process in the case of confined DNA, which is a similar scenario to inside compact nuclei [93]. Their results suggested that both confining geometry and chromatin elasticity contribute to searching efficiency. Foffano *et al* demonstrated that facilitated diffusion is most efficient when the confining volume is isotropic and chromatin fibre is flexible. This suggests that proteins search faster for their binding sites inside euchromatin regions where the chromosome is more flexible than in heterochromatin parts. Despite TF binding affinity differing between active and inactive chromatin, the target search favours open chromatin regions with more flexibility.

Another interesting observation is related to the effect of BS organisation on binding dynamics and occupancy. At present, most of the work done in this area focusses on BS sequential architecture. [94, 74]. In the context of facilitated diffusion, binding site clusters, either homotypic or heterotypic, may give rise to effects that cannot be reproduced by statistical thermodynamic models. For instance, Ezer *et al* simulated several simple but representative scenarios of BS building blocks. They found that two closely sited BSs for different TFs may function as barriers to each other. This is because the presence of another BS reduces the association rate of TF to the target site by blocking the TF 1D random walk from one direction [94]. In the case of homotypic BS clusters, a dual effect occurs: on one hand, it holds the molecule longer in the vicinity of the BSs by repeated sampling from the same sites during an episode



of 1D sliding, however simultaneously it contributes to the barrier effect described above. Thus a trade-off exists between BS spacing and affinity. Further, Sharon *et al* demonstrated via massive parallel experiments that sequential homotypic clusters influence both mean expression level and gene expression noise [95].

### 1.2.2 Towards consistent estimation of TF diffusion and binding kinetics parameters in live imaging

The modelling of facilitated diffusion of TF binding dynamics requires multiple crucial biophysical parameters, most of which are specific to TFs and even cellular systems. Thus, binding dynamic parameters obtained from live cells would be preferred over those from in vitro measures [76]. Important parameters include: the 3D diffusion co-efficient, the 1D diffusion co-efficient along the DNA, association and disassociation rates to non-specific DNA fragments, TF specific and non-specific residence times, bound proportion of molecules, inter-segmental transfer (jumping) probabilities and so on. Several methods exist to derive binding kinetic parameters in living cells. Single molecular tracking (SMT), fluorescence recovery after photo bleaching (FRAP) and fluorescence correlation spectroscopy (FCS) are three main approaches [96, 76, 97].

FRAP uses a high intensity laser to quickly bleach fluorophores in selected regions, thus generating dark spots without any fluorescence in microscopic images [97, 98]. With the Brownian motion of fresh fluorescing molecules in nearby regions, the dark spot can return to fluorescence gradually, which can subsequently be characterized using a standard form of the diffusion equation. By correctly fitting the fluorescence intensity observed over time into the equation of diffusion-limited fluorescence recovery, one can estimate the in-vivo diffusion co-efficient of a specific kind of molecule in a certain type of cell. By further fitting two or three components, kinetic models describing bound and freely diffusing states of TFs, residence time and the proportion of DNA bound TFs can be derived according to [99].

FCS can also be used to measure the diffusion co-efficient and residence time of TFs [96, 100]. FCS differs from FRAP in the sense that it utilizes the fluctuation of

the fluorescence intensity occurring naturally from both random noise and biological relevant effects rather than manual fluorophore bleaching. The information carried in these slight fluctuations can be extracted to investigate TF binding and diffusing properties. The core concept of FCS involves autocorrelation of time-dependent fluorescence intensity, which measures the self-similarity of the fluorescence signal over time and can be modelled quantitatively in 3D diffusion of molecules. If both the diffusion model and binding kinetic model are adopted appropriately, FCS and FRAP can yield a similar estimation of the TF binding parameters of interest. [85].

Another widely adopted approaches single molecular tracking (SMT) [101, 102]. Instead of measuring fluorescence intensity in a certain area or volume, SMT directly tracks single molecular trajectories [76]. By calculating the distances molecules jump at different time-lags, a series of histograms of displacement at different time points can be obtained. The histogram needs to be normalized to the total number of jumps measured at the shorted time interval and corrected for the effect of photo-bleaching [99]. The normalized histogram represents the probability of detecting displacement of a certain length within a given time, which can then be used to further derive diffusion or binding parameters of interest.

Through searching existing literature for the in-vivo diffusion co-efficient and estimated TF residence time needed for our TF binding dynamic model, it was noticed while the diffusion co-efficient estimations were relatively consistent with each other (3D diffusion co-efficient in a range of  $2 - 10 \mu m^2 s^{-1}$ ), the TF residence time estimations as well as bound fraction of molecules appeared to have dramatically different results. For instance, Elf *et al* [76] suggested a residence time of non-specifically bound Lac repressor on DNA based on SMT to be 0.3 to 5 ms (or 0.0003 to 0.005 s), while Chen *et al* suggested a non-specific binding time of Sox2 to be 0.8 s, while specific binding of Sox2 around 12s, based on their fitting for a two state binding kinetics model of SMT [84], which differed from Lac-I by more than 3 orders. In terms of the proportion of TF molecules bound to DNA, different studies also proposed very different predictions, ranging from 15% to 90% [101, 102, 76, 103]. Admittedly, there could be substantial differences among different TFs and cell lines. However, given the above variances, it is crucial to distinguish between real residence time differences

among TFs and measurement errors coming from either instrumental limitations, inappropriate normalization or model selection.

Mazza *et al* made direct evaluations and comparisons across different TF residence time measurement techniques [99]. They found that the accuracy of identifying the trajectory corresponding to truly bound molecules is crucial to SMT estimation of residence time. Confounding factors came from two aspects: 1) freely diffusing molecules can be judged as transiently bound if small movements occur below the threshold of displacement 2) a bound molecule can be viewed as diffusing because of the precision limit of localization. The first phenomenon if not controlled properly, may potentially lead to a much lower estimation of residence time by up to several orders of difference. Whereas in the second case, over-estimation for residence time is likely to occur [85, 99]. To resolve these problems and establish a reliable way of objective selection of SMT tracks reflecting either bound or diffusing molecules, Mazza *et al* suggested removing a fraction of short survival molecules within a certain number of consecutive frames when analysing particle jumping trajectories corresponding to DNA bound molecules. This was shown to effectively eliminate most contaminations from freely diffusing molecules. Further using the corrected SMT fit to guide the binding kinetic model selection for FRAP and FCS can greatly improve the consistency between these three different measurements. The resultant residence time for p53 was shown to be around 2s to 6s and the bound fraction of molecules tend to be within the range of 23% to 30% if using the model containing one bound and diffusion state [103, 99].

Therefore, SMT is the most direct measure of TF binding kinetics among the above three techniques, but great care is needed to distinguish bound and freely diffusing molecule tracks [102]. FRAP and FCS are both sensitive to background normalization methods including correction for photobleaching. Inappropriate binding kinetic model fitting can be another confounding factor, especially in cases when different kinetic models fit the data equally well. If so, SMT is recommended to guide the kinetic model selection for FRAP and FCS [85]. With proper control over technical variance and model fitting, these three approaches are able to reach a consensus on TF binding and diffusing kinetics.

## 1.3 Hi-C: an approach to study genome organisation

### 1.3.1 Introduction to chromosome conformation capture techniques

Chromosome conformation capture (3C) approach and its derived technologies have been widely adopted to study chromatin interactions in different organisms.

In 3C-based approaches, long-range interactions between pairs of loci are interrogated through spatial proximity based ligations. In the basic 3C approach, sample library is analysed using locus-specific PCR [104]. Further development of 3C includes the circularized chromosome conformation capture (4C) [105], which gives genome-wide interaction profiles for a single locus, and the multiplexed ligation-mediated amplification (5C), which is able to interrogate millions of interactions in parallel between two large sets of loci [106]. However, these techniques all need pre-selected genome regions of interests as target loci and are not able to give the genome-wide information of chromosome organisation properties in an unbiased way.

Hi-C, instead, takes advantage of high-throughput sequencing to deal with the DNA proximity ligation samples, which generates a genome wide contact map of chromosomes. A unique step included in Hi-C is that the staggered DNA ends generated by restriction digestion can be filled in with biotinylated nucleotides, which enables the specific purification of ligation junctions [107, 108]. It gives an efficient way to probe unbiased interactions across the entire genome.

Hi-C protocol has gone through several rounds of improvements. The initial version of the protocol is the so-called 'diluted Hi-C', which might introduce severe disruption to the genome structure within the nucleus, because harsh conditions were applied to permeate cell membranes including the nuclear membrane before the proximity ligation step [107]. The improved protocol tried to preserve the nuclear structure as much as possible after crosslinking and before the ligation step, which is named in

situ Hi-C. In situ Hi-C aims to maintain the intact nuclear structure, so they checked the state of nuclei under the microscope after each key steps before ligation to ensure most of nuclei remain intact. I summarized briefly below the key steps in the pipeline for Hi-C inside intact nuclei using the improved protocol from Rao *et al*, 2014 [1].

First, 1% to 2% freshly made formaldehyde are used to cross-link cells for 10 min with mixing, then glycine solution was used to quench the reaction.

Next, to permeabilize the cell membrane and extract nuclei, ice cold lysis buffer containing Igepal CA630 and protease inhibitors is used (pH=8.0). After washing steps, nuclei pellets are resuspended in low concentration SDS (0.5%) and incubated at 62C for 5-10 minutes before Trinton X-100 is used to quench SDS. This step permeabilizes the nuclear membrane and partly removes chromatin-associated proteins in a very mild condition and facilitates the following restriction digestion step. Restriction digestion is performed using Mbol restriction enzyme (4-cutter) [1]. It creates a 5 overhanging end on each piece of DNA.

Then, the 5 overhangs are filled and also marked with biotinylated residues (biotin-14-dATP), which gives blunt-end fragments. Proximity ligation is performed using T4 DNA Ligase for 4 hours followed by de-cross-linking. The ligation products in the final sample should be highly enriched in DNA fragments that were located proximal to each other in the nucleus. Owing to the existence of the biotin marks in the junction of ligation products, DNA fragments in spatial proximity can be selectively purified using streptavidin beads [107].

After DNA shearing and size selection steps (300-500bp) using AMPure XP beads, DNA fragments containing biotinylated ligation junctions are selected using Streptavidin beads pull-down [1].

Using standard protocol for Hi-C library preparation with Illumina indexed adapter, the sequencing library is subject to PCR amplification, purification and finally Illumina sequencing. Sequencing reads are mapped back to the genome, checked through quality controls and analysed further using various available software [1, 109, 110, 38].

There are several variants to the above Hi-C protocol. For instance, in situ Hi-C can also be done without DNA cross-linking [1, 111]. In order to best preserve nuclear structure, uncrosslinked nuclei can be embedded in agar plugs, and the global results looked similar compared to the cross-linked ones, albeit with some differences in the resulting local interaction frequency for certain regions. It may indicate certain biases with formaldehyde cross-linking or slightly disrupted nuclear structure in sample preparation without cross-linking [1, 111].

One potential drawback of Hi-C due to the nature of the protocol is that it is only able to identify two DNA fragments in close proximity at one time, but not three or more. There could be certain amount of information loss due to this disadvantage, because it can be the case that more than two pieces of DNA need to be brought together in gene regulation at the same time in the same cell, i.e. promoter together with multiple enhancers [112]. The inability to identifying multiple interactions at once gives rise to ambiguity in interpreting Hi-C contact maps, especially when multiple loops identified in Hi-C map are all attached to the same genome region. It could be the case that those loops form simultaneously in all cells, or alternatively, each of them only appear in a subset of cells. Beagrie *et al* proposed a very unique way to solve this: instead of using proximity ligation, they measured chromatin co-localization patterns through sequencing DNA from a large collection of thin nuclear sections [113]. They named this technique genome architecture mapping (GAM) and used it to infer both two-way and three-way chromatin contacts genome-wide. They found enriched contacts between highly transcribed regions and super-enhancers in three-way contacts, suggesting the existence of transcription factory like organisations.

### 1.3.2 Hi-C contact map normalization approaches

The raw reads in Hi-C contact map can be influenced by various factors. Some are known systematic biases such as GC content, sequence mappability, cross-linking efficiency, restriction site distribution and cutting efficiency differences between different regions with distinct DNA accessibility and so on [114]. Some other biases could be

random and thus difficult to classify systematically. As a result, different genomic regions tend to show different visibility in terms of associated Hi-C contacts. Those confounding factors make it difficult to directly analyse the proximity ligation results and identify genome positions with higher than expected probability to be together [115].

Available Hi-C contact map normalization methods fall into two categories: one group of methods explicitly consider potential biases from different levels and put them into models explicitly one by one by calculating their effects additionally [110, 109]. For instance Yaffe et al decomposed restriction fragment level biases into factors including fragment length, GC content and sequence mappability [110]. However, those methods are only able to effectively remove known influences in sample preparation and sequencing. Due to the requirement for parameter estimation related to different biases components, it often involves compute-intensive machine learning processes, especially for large datasets with higher resolutions.

Another group of methods eliminate the need to decompose biases from biological systems at different levels. Most of approaches within this class rely on the assumption that all loci should have equal visibility in the processed contact maps after normalization [116]. Lieberman *et al* and Rao *et al* [107, 1] used several matrix balancing methods which further assume that the existing biases are scalar and multiplicative. The simplest way is Vanilla coverage normalization [107]. It calculates the row and column-specific normalization terms respectively, simply by summing the raw reads number in either a row or a column and taking the reciprocal. A more advanced matrix balancing algorithm proposed by Knight and Ruiz (KR normalization) [117] was adopted by Rao et al [1]. Matrix balancing aims to get a matrix with all rows and columns summing to the same value. Their method is similar to the old Sinkhorn-Knopp algorithm [118], in which VC normalization is repeatedly applied until reaching convergence, but with a much faster convergence [117]. KR normalization can produce a balanced matrix when the original matrix is not too sparse. There might only be issues with the highest resolution, where discarding around 5% of rows associated with too few reads is necessary.

Maxim Imakaev *et al* [116] proposed another approach to iteratively correct biases collectively for all factors, either known or unknown, which can potentially affect visibility. Based on the assumption of factorizable biases, i.e. biases for contacts detection between two loci can be viewed as the product of individual biases associated with two loci. They have showed that the equation used in their iterative correction is consistent with the maximum likelihood estimation of the relative contact probability, if assuming the observed contact counts were drawn from a big class of underlying probability distribution including Poisson distribution and exponential distribution. Their approach aims to generate a uniform coverage profile with rows and columns summing up to the same value in L1 norm, which is the same as matrix balancing [116]. One special feature of their algorithm is that the iterative correction procedure can be extended to include single-sided mappable reads when considering the coverage profile, which are abundant in regions near centromeres.

Both the KR normalization [117] and the iterative correction methods mentioned above [116] are data-driven and can adapt to slightly different experiment protocols better as they do not have specific assumptions for the sources of biases. Further, they are all relatively compute-efficient, so I chose to use methods from this class in my Hi-C data analysis.

### **1.3.3 Chromosome organisation and TADs structures revealed by Hi-C**

Based on the genome-wide contact maps ranging from budding yeast to human, chromosomes are partitioned into two distinct compartment, namely A and B compartments, which is enriched in active or inactive chromatin respectively [107, 119]. Active regions tend to associate with other active chromatin, while inactive parts interact more frequently with inactive regions. At a megabase to a few hundreds kilobase scale, chromosomes form domains with elevated interaction frequency, which are known as topologically associating domains (TADs). The interaction frequency within TADs were shown to be roughly 2-fold higher than the contact frequency



between adjacent TADs [120].

Though the identification of TADs can be influenced by the choice of algorithms and also the resolution of contact maps, most of algorithms yields similar or at least, highly correlated results [107, 121]. With higher resolution contact maps, for instance, the ones described in Rao *et al*, 2014, sub-structures within TADs could also be detected. They used the Arrowhead algorithm in calling those self-associating domains, most of which were within 500kb (around 200-300kb length in average). Their domain calling algorithm resulted in reporting hierarchical domains, which was a bit different from the non-overlapping TADs generated in other studies [120]. Nevertheless, the existence of the hierarchical structure in chromosome domain organisation has been confirmed even using contact maps of much lower resolutions and in other cell lines as well. One of the typical studies was done by Fraser *et al* [122]. They studied higher-order interactions among multiple TADs in the process of neuronal differentiation and observed a hierarchical tree structure which they named metaTADs. They showed that the metaTAD were closely linked to epigenomic profiles and gene expression. The rearrangements of the metaTADs tree during cell differentiation were also correlated with gene expression changes during cell lineage specification [122].

Though the higher-order metaTAD structures vary in different cell types, the TADs themselves are to a large extent consistent across different tissues [121]. Thus, several studies viewed TADs as the invariant building blocks of chromosomes [119]. Nora *et al* showed that during ES cell differentiation, genes within the same TAD have significantly higher correlation in terms of expression than gene pairs located in adjacent TADs [123].

Relocating genes near TAD boundaries can result in alteration to gene expression patterns. Lupianez *et al* used CRISPR/Cas system to rearrange the genomic position of the limb enhancers and the *wnt6* gene near a TAD boundary. They observed an inappropriate upregulation of *wnt6* in certain limb tissues in which the limb enhancers appear to be active [124]. It suggested that the long range interaction between enhancers and target genes can have major impact on gene regulation. TAD structures are vital to the maintenance of spatial and temporal gene expression patterns during

development [125, 124].

### 1.3.4 Possible mechanisms of TADs and the associated loops formation

Central to the understanding of genome organisation, the underlying mechanisms for TADs formation have been hypothesized and modelled quantitatively in different studies [126, 112, 127]. Rao *et al* first observed clear loops associated with chromosome domain boundaries, in which a large proportion overlapped with CTCF motifs [1]. Interestingly, CTCF sites preferentially have an inward orientation when flanking a TAD [1, 128], which indicates an interesting relationship between TADs and the CTCF motifs marking the boundaries.

The studies by Fudenberg *et al* and Sandborn *et al* [126, 112] proposed a hypothesis of loop formation called loop-extrusion, which can well explain the observed pattern of TADs boundary-associated loops. They hypothesized that some loop-extruding factors, for instance cohesins, associate with the chromatin fibre first and then progressively generate larger loops until they encounter some obstacles including CTCF near TADs boundaries. Sandborn *et al* further assumed that the interactions with CTCF can effectively stabilize a moving cohesin complex and make it less likely to dissociate from the chromosome. This can partly explain the enriched loops marking TADs boundaries. Moreover, if the function of CTCF to stall cohesins depends on the appropriate orientations of CTCF relative to cohesin, then the observation that only CTCF sites with inward orientation serve as TAD boundaries can also be understood [112]. Though there is still no direct evidence that cohesin complexes are able to make orientation-specific loops, early studies confirmed that cohesin complexes can move along the DNA after being loaded at certain positions [129] and Kim *et al* further demonstrated that SMC-containing complexes do slide along the DNA in vitro [130].

However, whether CTCF is essential for loop generation is still illusive. Kubo *et al* recently argued that upon acute loss of CTCF, chromatin domains and A/B compartment can be nearly preserved, though lamina associated domains were affected

[131]. They observed that despite the reduction of interactions associated with loop anchors, the majority of TADs boundaries still remain intact. It raised the possibility that CTCF is not the only factor that marks TADs boundaries and the driving force for TADs formation is independent of CTCF, which contradicted the loop extrusion model by Sandborn *et al* [112].

Even though loop extrusion is one of the contributing factors to local chromosome domain formation, it cannot explain the A/B compartment and further A1/A2 sub-compartment segregation [107, 1]. Other studies proposed alternative explanations for the formation of chromosome compartments and also provided insights into TAD specification. Barbieri *et al* showed that TADs could arise from preferential interactions between at least two types of loci [132], *i.e.* loci of the same type can interact via certain kinds of bridging molecule dynamically, while different types of loci do not have specific interactions. Each TAD could be represented as a chunk of genome region associated with a single type of loci under this simple model. At the same time, this simple model can lead to the segregation of compartment-like structures genome wide: *e.g.* A and B compartments in response to two different types of interacting loci. One observation favouring this mechanism is that TAD boundaries often coincide with the alternation between chromosome compartment or subcompartment, even though there is no loop-like structure marking the boundary of TADs (or saying chromosome domains called by the Arrowhead algorithm) according to Rao *et al* [1].

Additionally, DNA supercoiling could also possibly give rise to genome segmentation and the supercoiling domain boundaries were shown to have significantly higher chances to overlap with TADs boundaries [133]. Although supercoiling domains are much smaller than the observed TADs (130kb in average compared to Mega-base-scale TADs), it is possible that TADs are further comprise of smaller transcriptional units locally. Gene transcription mediated by processing RNA polymerase is able to generate over-wound DNA ahead of the transcription machinery and under-wound behind instead [134]. Some chromosome remodellers have also been shown to produce hundreds of base pairs of under-wound DNA loops in vitro, including proteins containing SNF2p-related ATPase [135]. The boundaries of supercoil domains could be determined by either insulators like CTCF and cohesin, similar to the ones men-

tioned in the loop-extrusion model, or alternatively, low-density gene regions in which supercoiled DNA dissipates gradually [134].

Fluorescence in situ hybridisation has showed that under-wound DNA supercoiling decompacts chromosome domains rather the over-wound supercoiling [133]. This effect relies on processing transcription machineries and it would be lost upon the inhibition of transcription or in presence of DNA-nicking reagents. It is consistent with the observation that active chromosome domains within A compartments are less compacted than inactive regions or polycomb complex associated domains [136].

Supercoiling has also been shown to enhance the interactions between gene promoters and the corresponding enhancers via Brownian dynamics simulations [137]. For DNA without bridging proteins connecting specific loci pairs, the effect of supercoiling in helping DNA compaction and overcoming electrostatic repulsion only present in low-ionic-strength buffers [138]. However, when assuming TFs or other mediator proteins can bridge enhancers and cognate promoters, super-coiling can significantly stabilize enhancer promoter interactions within the same super-coiling domain and slightly decrease inter-domain interactions [137]. Supercoiling also facilitates the re-binding of dissociated enhancer-promoter pairs, further helping to maintain specific enhancer-promoter interactions according to biophysical simulations [139, 137]. Those predictions for the role of DNA supercoiling in chromosome domain formations as well as enhancer-promoter interactions remain to be tested in vivo on mammalian chromosomes.

Although various mechanisms have been proposed to account for the observed TAD organisation, mechanisms ranging from loop extrusion to supercoiling mostly contribute to local chromosome folding, capable of generating and preserving TADs or intra-TAD organisations within a mega-base scale [126, 112, 133, 137]. However, when it comes to a larger scale more than 5 Mb extending to the whole chromosome, or even to the arrangement of chromosome territories [140], very few models can give a compelling explanation, except the ones proposed by Barbieri *et al* and Brackley *et al*, which argued that the multi-type DNA bridging complexes could induce chromosome domain and compartment segregation [132, 141]. Despite their intriguing simulation

results, evidence supporting the existence of different groups of DNA bridging complexes corresponding to different chromosome compartments or subcompartments are scant in higher eukaryotes. DNA bridging complexes in inactive chromatin could be polycomb complexes and so on, which are distinct from the ones enriched in actively transcribed regions like mediator proteins. However, similarly, whether there are any subgroups of potential regulatory proteins that correspond to chromosome subcompartments within active chromatin is unknown. This motivated us to investigate the potential link between regulatory protein spatial grouping and chromosome subcompartment segregation, particularly within active chromatin which is associated with enriched TF binding. We will further investigate if the spatial networks correspond to tissue specificity, protein physical interactions and co-functions.

# Chapter 2

## Transcription factor binding dynamics and occupancy is influenced by co-localization of homotypic sites

### 2.1 Introduction

#### 2.1.1 Discrepancy between the predicted and the experimentally-derived TF binding landscapes

Transcription factor (TF) binding site (BS) occupancy is closely related to gene expression regulation [142]. Computational prediction of transcription factor binding sites can help elucidate gene regulatory networks in eukaryotes [8]. Most TFs bind to specific DNA sequences in the genome, but the potential BS candidates predicted by DNA sequences, along with DNA accessibility information, often greatly outnumber the in-vivo BSs identified by either ChIP-seq profiles or DNaseI digital footprints in

different tissues [43, 142]. The possibility for putative BSs being bound (which we here refer to as occupancy) have been shown to be influenced by DNA methylation and histone marks [143, 144]. However, even taking epigenetic features into account, it still cannot fully explain the discrepancy between computationally predicted and real TF binding landscapes gained from ChIP-seq or DNase-I foot-printing across the genome [145]. Current available bioinformatics tools to predict BS occupancy based on DNA binding sequence motif information and epigenetic features often yield results contradicting the in-vivo binding profiles in certain genome regions. However good correlation can be achieved from a subset of well-defined cis-regulatory modules [146, 147, 148, 149].

Both the ChIP-seq and DNase-I digital footprint approaches have certain biases in detecting TF binding regions due to technical limitations [34, 21]. However, it is likely that other factors exist which may drive TF binding preferences. The discrepancy between the predicted and actual TF binding landscape must be reconciled by other influencing factors which play essential roles in determining BS occupancy. For instance, some TFs are known to rely on the binding of other partner TFs [150, 84] *e.g.* bZIP, bHLH and nuclear hormone receptor TF families. These are known as heterodimers when binding to the DNA [151, 152]. Apart from direct dimer-formation, increasing evidence exists from the exploration of TF binding mechanisms and dynamics, to suggest that BS organisation in the genome—both linearly along the genome and further spatial arrangement may also influence TF binding dynamics and occupancy [153, 92].

### **2.1.2 Binding dynamics simulations predicted accelerated target search associated with homotypic BS clusters**

Although there is no direct evidence from in-vivo assay to date showing the effect of spatial organisation of BSs on the binding of TFs, few studies exist using biophysical simulations such as Brownian dynamics of facilitated diffusion, to predict the possible impact for DNA conformation during the TF target search [93, 154, 86].

A typical example is the study from Brackley [141], which modelled bacterial DNA as a semi-flexible polymer with homotypic BSs evenly distributed along the polymer chain. Instead of treating DNA as a self-avoiding random polymer [93, 89], they gave the polymer chain specific topological conformations, for instance, a string of rosettes, each comprising multiple loops with loop points anchored together [141]. They investigated a range of TF-DNA interaction energies, demonstrating that when the TF binding affinity is above a certain threshold, a string of rosettes conformation significantly improved TF target search efficiency by facilitating direct TF inter-segmental transfers in the presence of homotypic clusters. In addition, in terms of BS sequential arrangement, without the effect of DNA looping, they investigated different patterns of homotypic BS distribution along the polymer chain e.g. random distribution versus funnel-like landscape, where a single high affinity BS is surrounded by multiple low affinity, non-specific BSs. They observed a severe slow down in the target search when the distribution of low affinity BSs was randomly distributed, as low affinity sites construct traps for TF diffusion. However, in a funnel-like landscape with traps around the high affinity BS, the search process becomes an order faster compared to the situation without traps. This is because low-affinity traps around the high-affinity target sites helps to decrease the chance in which TFs diffuse away from the vicinity of high-affinity sites, even in the case when binding affinity of those traps is very low [141]. Other work revealing similar mechanisms include the early studies from Leonid Mirny and Johanna Weindl et al [155, 153].

Their study on funnel-like binding landscapes provided insight into the effect of 1D sequential homotypic BS clustering on TF target search efficiency. Their simulation of DNA conformation in a string of rosettes incorporating multiple homotypic BSs inspired us to further explore the possible influences of genome architecture on TF binding dynamics and to seek evidence and support from in vivo measures in eukaryotic systems. However, as Brownian dynamic simulations are quite computationally intensive, it is not suitable for large scale simulations with user-defined scenarios [86, 141, 91]. The number of diffusing molecules also needs to be limited to a small number, due to non-linear increase of memory demands in relation to the number of tracked particles. Due to the above reasons Browning dynamic simulations are not



the first choice in our study to investigate BS clustering.

Furthermore, the polymer models often put more emphasis on the TF target search process, measuring the mean search time, as their strength is in characterising the dynamic process and estimating physical parameters related to different search phases e.g. fast diffusing versus sliding. Polymer models also require precise knowledge of various parameters relating to chromosome stiffness (Kuhn length), protein-DNA interacting energy and elasticity of chromosomal polymer chains etc [93, 89]. These parameters are difficult to measure in eukaryotic nuclei and can vary greatly across different genomic regions according to chromosome accessibility and epigenetic marks. The estimation of target search time using FRAP, FCS or SMT is also not available for every TF, given the limited availability of fluorescent tagged proteins [99, 96, 98]. Instead, the information that is widely-available in biological systems would be TF occupancy from ChIP-seq data [29, 26]. Therefore, in conjunction with the available BS occupancy profiles, we aimed to develop a robust approach to modelling TF binding dynamics being more occupancy-based rather than target-search-time-focused. We wanted to find some computational-efficient simulation methods to quantify the impact of homotypic BS organisation on TF binding dynamics, enabling the combination of 1D sequential BS distribution with 3D BS architectures in a user-defined way. In addition, since the presence of chromosome loops have been shown to be dynamic and may only appear in a subset of cells [112], we wished to explore BSs in spatial proximity in a more general context, not limited to the scenario of loops.

### **2.1.3 Research aim**

Previously, our research group has built a computational simulation framework, named FastGRiP. FastGRiP is capable of performing stochastic simulations that model TF binding and unbinding events with user-defined binding sites, sequential arrangements and other parameters thus characterizing the facilitated diffusion of a specific TF [94]. It is an computational efficient simulation tool in comparison to other software [156, 86, 89], as it focuses on modelling occupied and unoccupied state transitions

for an array of BSs, rather than tracking the movement of individual molecules. It makes use of the Gillespie algorithm incorporating TF association, disassociation and translocation to adjacent BSs following certain probabilities according to the facilitated diffusion mechanism [93]. FastGRiP is capable of dealing with multiple kinds of BS sequential arrangements, either for two adjacent BSs forming a switch or a barrier, or homotypic BS clusters. However, the spatial arrangement of BSs is not taken into account [94]. Therefore, using FastGRiP we made further extensions to include BS 3D arrangements in order to model the TF spatial target search process in a computationally efficient way. This piece of work was done in collaboration with Daphne Ezer, in which I modelled TF 3D diffusion and jumping probabilities in Matlab whilst Daphne further incorporated the 3D search results into her previous Java codes of FastGRiP and to enable this 3D extension.

In addition to biophysical simulations of TF binding dynamics, we sought to seek evidence and support from *in vivo* assays *e.g.* ChIP-seq TF binding profiles. Despite the mechanistic advantages provided by spatial homotypic TF clustering [76, 154], little attention has been paid to study this phenomenon using genomic approaches. Malin et al observed significantly higher TF BS occupancy derived from DNase-I digital footprinting in a set of bioinformatically inferred regulatory archipelago enhancers [157]. This was hypothesized to be a consequence of spatial homotypic clustering of BSs made possible by the spatial grouping of correlated enhancers. However, there is no direct evidence that those bioinformatically inferred enhancer clusters do appear to be in spatial proximity and no information about TF type can be obtained from DNase-I digital footprinting profiles. Their studies focused on special types of enhancers, but there has not been a direct investigation so far of the quantitative relationship between the level of homotypic BS co-localization and BS occupancy on a genome wide scale. It is also not clear if similar rules hold for genomic regions with different epigenetic features and for different kinds of TFs.

Chromosome conformation capture techniques including Hi-C, paved the way for us to systematically investigate the influence of spatial homotypic clustering of binding sites (BSs) on TF binding efficiency [119, 158, 120]. DNA proximity information obtained from genome-wide contact maps can be used to quantify the level of homotypic BS

co-localization.

Admittedly, there would be other alternatives to Hi-C contact maps, for instance, capture Hi-C contact data. However, capture Hi-C data only utilises a small subset of pre-defined genome loci, such as promoters or certain enhancers, as viewpoints to interrogate their interactions with all other genome regions [159, 160]. Their contact profiles were difficult to normalise and to remove various confounding factors that influence contact read distributions. Matrix balancing or other iterative normalization procedures based on the genome-wide contacts were not applicable to capture Hi-C [107, 116]. In addition, due to the scarcity of reads in distal regions, reliable peak calling was hard to achieve. Although there are several available software programs trying to solve this [159, 161], their peak calling algorithms are mostly based on certain distribution assumptions such as Gamma distribution, which might be violated in reality due to the complexity of the biological system.

Thus, we chose to use the genome-wide intra-chromosomal Hi-C contact map of GM12878 cells, with the improved protocol that well-preserved intact nuclei structures [1]. It reached the resolution of 5kb in the final binned contact reads profiles, achieved by super-deep sequencing. Based on the Hi-C contact map, we defined a metric of quantitative representation for the homotypic BSs co-localization level around each genomic loci. We found a strong linear correlation between BS occupancy and BSs homotypic co-localization, consistent with biophysical simulations of transcription factor binding site searching. This trend is more pronounced for BSs with weak binding motifs and with the enhancer state.

## 2.2 Methods

### 2.2.1 Biophysical model of TF 3D diffusion and jumping to BSs in spatial proximity

Our lab has previously developed a semi-analytical model, which was named fast-GRiP, to simulate TF target search process based on the facilitated diffusion mechanism [94]. It made use of a continuous time Markov chain, in which every state represented a specific bound or unbound configuration of an BS array. State transition rates were defined according to the probability of single event of TF binding, unbinding to DNA or relocation to their adjacent BSs within a sequential BS cluster. The fast computation of BS occupancy configuration was achieved by mathematical approximations of TF sliding and non-specific residence time calculation associated with target regions which we called 'sliding windows' around each sequence-specific BS [94]. The sliding window we defined is based on the observation that a given type of TF can perform 1D random walk on non-specific binding regions. TFs can explore an average length of  $s$  base pairs before disassociating from the DNA [76], where  $s$  is specific to different TFs and is related to protein structural properties within DNA binding domain. Thus, by approximation, if the target BS is located within  $s$  bp of the landing point of a TF molecule, the TF is then likely to reach this target via 1D sliding [156, 94].

FastGRiP enabled the incorporation of user-defined genomic binding affinity landscape into binding dynamic simulations, which is an useful feature absent in most of other simulation tools [94, 89]. However, the spatial arrangements of BSs were not taken into consideration initially in FastGRiP. In the original simulation, we assumed that unbound TFs are equally likely to bind to any other BSs, regardless of their spatial position. However, in fact, a recently dissociated TF is more likely to bind to a spatial nearby site than a far away site by 3D diffusion. The spatial arrangement of BSs cannot be simply reflected by Chemical Master Equations used in the initial simulations, but further requires quantifications of re-binding probabilities to BSs

located within different distances. We refer to the above event of TF disassociating and re-binding to a spatial nearby sites as TF jumping, similar to [93, 154].

The physics of jumping to spatial proximal BSs can be captured by a simple diffusion model assuming that 1) there is an absorbing sphere of radius  $r$  around each target BS and 2) TFs can be viewed as point particles diffusing freely with an effective diffusion coefficient  $D_{eff}$ , which is adjusted for DNA crowding.

The equation that describes the probability of a point molecule distance  $r$  away to reach the absorbing sphere at time  $t$  has been previously derived [162] [163]. It is equal to the probability flux into the absorbing sphere given the standard diffusion with the corrected diffusion co-efficient and assuming a transient point source of diffusion particles. Note in the online paper of Paramanathan *et al*, 2014, there was a typo in one of their equations describing TF absorbing probabilities on Page 8, equation 7 (the power of  $t$  in the denominator should be  $3/2$  instead of  $1/2$ ). The corrected equation is written below. I have informed the author to correct this typo in their on-line paper, while the master equation used to derive this formula originally in [162] was correct.

In the following equation,  $s$  is the radius of the absorbing sphere, and  $D_{eff}$  is the effective diffusion coefficient.

$$\phi(r, t) = \frac{s(r - s)}{2r\sqrt{\pi D_{eff}t^3}} \exp\left(-\frac{(r - s)^2}{4D_{eff}t}\right) \quad (2.1)$$

In analogy to the sliding window defined in FastGRiP [94], we adjust the diameter of the absorbing sphere  $s$  to 30nm for absorbing TFs from outside of the target, in which the target can either be a single BS or a BS sequential homotypic cluster located within 1000 bp (within the Kuhn length of eukaryotic DNA) [86]. In the case of internal jumps within sequential homotypic clusters, the absorbing sphere is set to be 2nm (the approximated size of a TF BS), representing directly reaching the binding site from 3D diffusion [164]. It is because fastGRiP has already incorporated the sliding of TFs between neighbouring binding sites, and we must be careful not

to double-count this effect. We calculate the diffusion coefficient  $D_{eff}$  using the following equation, as previously described [76]. We note the in vitro TF 3D diffusion co-efficient cannot be simply applied, because the nucleus is a crowded environment with various kinds of macro-molecule including DNA, histones and other crowding agents, which drastically slow down the diffusion of TFs compared to the invitro case.

$$D_{eff} = (1 - a)D + a\frac{D_1}{3} \quad (2.2)$$

where  $D_{eff}$  is the effective diffusion coefficient,  $a$  is the proportion of time the TF spends sliding on the DNA non-specifically in TF target search,  $D$  is the 3-dimensional diffusion coefficient measured in vitro without macro-molecules crowding, and  $D_1$  is the 1D diffusion coefficient of the TF on DNA. Therefore, the effective diffusion co-efficient is the observed equivalent diffusion co-efficient taking into account both TF free diffusion in 3D and the time delay caused by DNA crowding. In the following analysis, we choose  $D$  to be  $3\mu m^2/s$  and  $D_1$  to be  $0.046\mu m^2/s$ , as estimated by single molecule tracking (SMT) of LacI in live *E.coli* [76]. In eukaryotic nuclei,  $D$  and  $D_1$  for several TFs have been shown to be in the same order as Lac-I, though in certain cases, the precise estimation of  $D$  and  $D_1$  appeared to lack consistency across different studies for the same TF in eukaryotes [101, 102, 103]. The estimations of the fraction of time for TF non-specific association to DNA vary from 20% to 90% across different TFs and in different studies using either SMT, FRAP or FCS, which could reflect real differences across different families of TFs, or alternatively, biases due to inappropriate experimental design using either SMT or FRAP, as discussed in Chapter 1 Section 1.2.2.  $a$  equals 90% is only used as an illustrative example in our simulation, taking into account the data usage consistency with diffusion co-efficients from [154, 76]. The value of  $a$ , as well as  $D$  and  $D_1$ , can all be adjusted easily in our updated FastGRiP simulation tool according to fit into different scenarios based on TF choices. The functions to obtain the biophysical simulation results of TF diffusion and jumping probabilities were written in Matlab and were deposited in github: <https://github.com/ezer/DiffusionMarkovModelJumping> under my name.

In the original simulation, we assumed that any unbound TF is equally likely to bind

to any other binding site, so when a TF dissociates from a binding site it enters a *pool of TFs*. In the updated version, a recently dissociated TF is more likely to bind to a nearby site than a far away site, as shown by the probability density functions depicted in Figure 2.1B. Figure 2.1B illustrates that after 0.1 seconds, the probability density functions for TFs jumping between DNA strands that are 100nm, 200nm, and 2000nm apart nearly converge. After 10 seconds, the probability of the TF binding to a DNA strand 100nm, 200nm, or 2000nm away becomes less than 1% for all cases, so we replace a TF that is still free floating after 10 seconds into the TF pool in our simulation.

The other parameters we used were identical to those described by Ezer et al, 2014. We set  $\tau_0 = 3.3$ ,  $cn = 100$ , and the distance between binding sites in a cluster to 5 bp, unless otherwise stated.

## 2.2.2 Incorporation of TF jumping into Gillespie algorithm based simulations

In our updated fastGRiP, we incorporate the jumping probability from one strand to another by combining pre-computed diffusion probability look-up table with the Gillespie algorithm.

The Gillespie algorithm is able to compute a random time when the next event takes place and select the event that is most likely to happen next, given a set of known events. The thing to note is that the Gillespie algorithm has a core assumption that the probability of a reaction event is time-independent and must follow an exponential distribution, which was the case for TF association, disassociation and 1D translocation that was characterized in FastGRiP before. However, the re-absorbing or jumping probability of a TF from one DNA strand to another is time dependent, as described by the probability density function in 2.1. Selecting the time of the next reaction requires sampling a value from the averaged probability density function of the reaction times for all of the possible reactions. The exponential distribution, involved in the Gillespie algorithm, is the simplest one to deal with in this case,

because of its mathematical beauty that ensures the additivity of reaction rates. However, other more complicated distributions, especially the one encountered here, would require time consuming steps tackling numerical integrations and averaging custom functions to generate the integrated probability distribution for all possible events, and then sampling values from the above complicated distribution to select the next possible reaction event and its associated time. If this needs to be done for individual TF jumping associated event, it would require very long compute time and thus, drastically slow down our simulation.

Instead, we modified fastGRiP as follows to allow diffusion between DNA strands to be incorporated without substantially slowing the simulation. In the earlier version of fastGRiP, once a TF dissociated, it entered a pool where the TF is equally likely to bind to any location along the DNA, which ignored the spatial arrangement of TF BSs. Now, when a TF dissociates, it enters another pool of *diffusible TFs*. It samples the time of its next expected jump from a 100,000 element pre-computed lookup table generated in Matlab by me. All of the possible TF jumps are then read in and stored in a PriorityQueue in Java, which was coded by Daphne Ezer. When the Gillespie algorithm reaches the step to select the time of the next TF association, dissociation or intra-cluster translocation reaction, it first checks the pool of diffusible TFs to see if any TF jumping events have happened in the meantime, and updates the state of the system accordingly. Sometimes, a TF jump event can no longer occur, because that DNA binding site is already occupied by the time the new TF diffused to it. In these cases, we recomputed a new location for the TF to diffuse to and add it to the PriorityQueue again. If the time of a TF jump event is longer than 10 seconds, we do not store this TF in the pool of diffusible TFs, because it has nearly equal likelihood of diffusing to any binding site, as was illustrated in Figure 2.1 , and we place the TF in the original TF pool. This strategy allows us to model TF jump events with complicated, non-exponential probability density function in a compute-efficient way. The code for this modification is available at <https://github.com/ezer/DiffusionMarkovModelJumping>, where my Matlab code and D. Ezer's Java code was put together.



### 2.2.3 TF Binding sites Annotation

Given the biophysical simulations in respect to the influence of spatial homotypic BS co-localization on BS occupancy, we further explored this in real biological systems, where the degree of BS spatial proximity is represented by Hi-C contacts and BS occupancy is derived from ChIP-seq peaks. We choose human GM12878 lymphoblastoid cells as a model cell line, due to the availability of the highest resolution Hi-C contact map down to 5kb [1] and also sufficient number of TFs with ChIP-seq profiles in ENCODE[67].

ChIP-seq NarrowPeak profiles for TFs of GM12878 were obtained from ENCODE [67]. Position Weight Matrices (PWM) for TF motifs were collected from HOCCOMOCO [165], SwissRegulon [166] and JASPER [167] where available. To ensure the quality of sequence motif used for putative binding site (BS) identification, we only used PWMs that are derived from more than 30 validated binding sites and have a minimum motif length of 8 base-pairs. TFs without a suitable PWM motif were excluded from further analysis. After the above filtering, we recovered total of 40 TFs which both have ChIP-seq profiles and well-defined sequence motifs in GM12878 cell line.

Putative TF binding sites were defined as PWM motif matches of a certain transcription factor via FIMO motif scanning [168] in DNase Hypersensitivity Hotspots (DHS) with p-value threshold equals to  $10^{-4}$  (default setting) [9, 168]. Chromosome X was excluded from our analysis because of its specialised chromosome organisation [169, 170]. Each individual ChIP-seq peak from ENCODE profiles was mapped to the best scoring sequence motif which overlaps with it. Occupancy is then defined by the ratio of the number of ChIP-seq identified BSs and the number of total putative binding sites in specific groups of genome regions.

To account for potential influences from histone marks and chromosome sub-compartment on TF binding, we further grouped DHS regions and ChIP-seq identified BSs according to 1) chromosome sub-compartment annotation reported by Rao et al [1], 2) whether associated with H3K27Ac, H3K27Me3, H3K4Me1, H3K4Me3, H3K9Me3 and H3K9Ac, 3) ENCODE consensus chromatin states [171] 4) in addition, for DHS

regions located in gene promoter regions, we further classified them into strong promoters (highly active promoters) or weak promoters according to whether H3K36Me3 is present or not downstream of the gene body [172, 173]. For histone marks, we used ENCODE BroadPeak profiles from the Broad Institute for H3K27Me3, H3K9Me3 and H3K36Me3 marks due to the dispersing nature of the histone marks themselves, while NarrowPeak profiles were used in all other cases. Confounding genomic regions with both H3K27Ac and H3K27Me3 or both H3K9Ac and H3K9Me3 were removed from further analysis. In addition, methylated genome regions [174, 67] were excluded from our analysis to avoid potential influence of DNA methylation on TF binding patterns. The above filtering resulted in recovering 76% of DHS regions subject to subsequent analysis.

#### 2.2.4 Quantification of Spatial clustering of TF binding sites within the same type

We use the normalized intra-chromosome Hi-C contact frequency as an indicator of the strength of co-localization between any paired loci in each chromosome. We aim to establish a metric for each genomic locus that can represent how likely this locus interacts with any potential BSs of a certain TF in distal genomic regions, in parallel to the spatial clustering of BSs simulated in our above model. Ideally, real 3D distances are the most relevant measure, but it is currently something difficult to derive in a high-throughput way via microscopy. Hi-C contact frequency can also be related to 3D distance somehow, but the conversion itself depends on the choice of chromosome structure model and can give highly variable results depending on the fractal dimension within a certain chromosome region, which has been shown to vary substantially according to histone marks and gene activities [136].

To quantify the degree of spatial clustering of homotypic BSs around individual genome loci, we adopted the following formula (equation 2.3:

$$HCS_i = \sum_j \log \frac{obs_{i,j}}{exp_{i,j}} \quad (2.3)$$

where the  $HCS_i$  is the abbreviation for the Homotypic Clustering Score we defined for a specific genomic locus  $i$ ;  $obs_{i,j}$  refers to the observed Hi-C contact score between each genome loci  $i$  and each binding site  $j$  of a specific kind of TF; the  $exp_{i,j}$  is the expected average score accounting for the effect of genome distance, which is the empirical average of contact scores between a certain genome distance in each chromosome. We adopted the Hi-C contact score via the KR normalization method, the same as used in [1]. For each genomic locus of interest, we summed up the logarithm of the ratio for the observed versus expected scores across all ChIP-seq identified binding sites  $j$  of a specific kind of TF in the same chromosome, which measures how likely a genome locus is in contact with certain types of TF binding sites. Genome loci with the total associated raw reads less than one third of the median reads of each chromosome were deemed to be associated with insufficient reads and were removed from subsequent analysis. Diagonal elements of the Hi-C contact map were excluded as well as the adjacent 25kb regions left and right to avoid potential high levels of noise in Hi-C reads. Thus, the HCS defined above were mainly focused on the spatial co-localization of genome distal BSs, while the effect of local sequential clustering of BSs was not taken into account. (In the case of hESC contact map, given the basic resolution of 40kb, we only excluded the diagonal elements as well as the immediate neighbouring bins left and right. )

Notice that the average number of Hi-C reads between loci dropped quickly as genome distance increases— for distal loci, read number can be very low in the current resolution of 5kb (see figure 2.3 for an example and also [1]), which might lead to high noise in the above calculation. Therefore, we increased the bin size to be 25kb when two loci are more than 100kb apart by merging adjacent bins, and further to 55kb for loci more than 1Mb apart. (Similarly, in hESC contact map of 40kb resolution, we increased the bin size to be 120kb and 200kb for loci further than 400kb and 1Mb apart.) To avoid potential high noise in calculating the observed versus the expected ratio of contact pairs associated with low reads number, any contact pairs containing less than 20 raw reads were discarded.

HCSs for BSs of each TF were rank normalized, i.e. each score were replaced by its fractional rank, and further put into decile groups (10 groups) or grouped into high

(top 0.33), mid (0.33-0.67) or low (bottom 0.33) levels based on the requirement for each analysis.

### **2.2.5 ChIP-seq NarrowPeak SignalValue comparison between paired binding sites**

To account for potential variations of TF binding site occupancy caused by genomic DNA sequence motif differences, we made BS pairs with exactly the same DNA sequence motif, histone marks including H3K27Ac, H3K27Me3, H3K4Me1, H3K4Me3, H3K9Me3 and H3K9Ac and further made sure that both of the BSs were located within DNaseI hypersensitivity regions without any types of DNA methylation [67]. When evaluating the effect of BSs homotypic clustering, chromosome subcompartment of BS pairs were also kept the same, while in the study of the influence of A1/A2 subcompartment, the level of homotypic VBS clustering were kept the same. In order to make a reliable comparison of ChIP-seq SignalValues, we only used ChIP-seq NarrowPeaks that map to one unique DNA binding sequence from FIMO motif scanning [168].

ChIP-seq SignalValues of each TF were rank normalized with the highest value assigned the score of 1 and the lowest of 0. We performed Wilcoxon signed-rank test to the list of paired BSs for each TF to test if different categories of BSs show significant differences in ChIP-seq SignalValues. We first calculated the differences between the normalized SignalValues for each pair of BSs, then as a control, the two BSs in each pair were randomly shuffled to obtain the expected distribution of the SignalValue differences between pairs of BS. If there are more than one BS that can make match to a specific BS, all possible combinations were retained in our analysis.

## 2.3 results

### 2.3.1 Simple scenarios of TF binding dynamics simulations in two binding site clusters of spatial proximity

In our TF binding dynamics simulations with updated FastGRiP, we compare three scenarios 1) First, we look at a pair of homotypic clusters that are on two different DNA strands, as shown in Figure 2.1AII, and we vary the distance between two DNA strands 2) Then, in the same scenario, we adjust the distance between TF binding sites within the homotypic cluster (Figure2.1AIII). 3) We vary the number of TF binding sites within each BS sequential homotypic cluster located on the same strand (Figure2.1AIV).

In each case, we are interested in determining how these binding site organizations influence TF occupancy, which we define here as the average probability that each TF binding site is bound. For instance, TF occupancy of 0.05 means that on average each TF binding site is bound 5% of the total time. We note that the exact amount of binding time depends on the average residence time of sequence-specific BSs that is used as input to the simulation. The parameters we use here are the same as Ezer 2014 [94], but it could potentially vary according to different TF of interest.

In the first scenario with two binding site clusters on different strands located at different distances from each other, we see that the closer these two strands are in 3D, the higher average occupancy they have. It suggests that TF jumping between different strands significantly increases the average TF occupancy of the region (Figure2.1C).

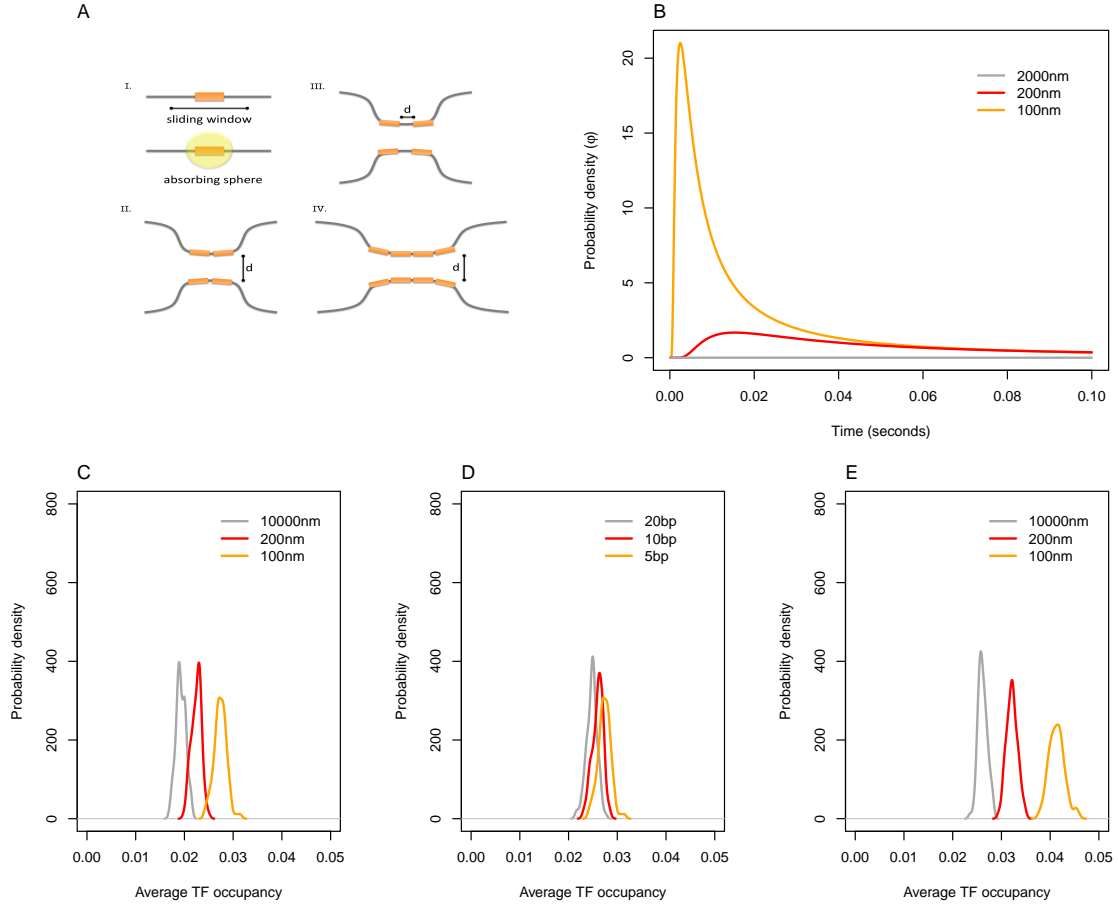


Figure 2.1: *Biophysical simulations of the effect of homotypic BS co-localization.* We evaluate the effect of spatial homotypic BS clustering on TF binding dynamics and occupancy using updated fastGRiP simulations. Subfigure AI demonstrates how fast-GRiP's sliding window concept is extended to an absorbing sphere when considering 3D diffusion. AII-AIV show the simulated scenarios of BS sequential and spatial arrangements. The shape of the probability density function  $\phi$  from equation 2 is shown in B. The results from the simulated scenarios AII-AIV are depicted in C-E. They show the probability density plots of TF occupancy, which is the probability for each TF binding site being bound. Note that the TF occupancy, as defined by fastGRiP, includes not only the time at which a binding site itself is occupied, but also the time when the TF is within the average sliding length of the binding site (here we choose to use 90 bp [76]).

Next, we vary the distance between binding sites within sequential homotypic clusters.

In contrast to the substantial variation of BS occupancy in relation to the distance between DNA strands, in this case, 1D distance between BSs only slightly influences overall TF occupancy, at least given the parameters that we simulated (Figure 2.1D). This result is a reflection that there are two opposite effects influencing TF binding site occupancy. On the one hand, there is increased translocation of TFs between two binding sites in a cluster when the distance between binding sites decrease. On the other hand, the absorbing spheres around each BS may intersect more if the two sites are very close together in a sequential homotypic cluster. Hence, when considering TF jumps from outside of the cluster (other DNA strands), the overall chance for a TF to reach each binding site is reduced. This is comparable to playing a game of darts with two dartboards that are partially overlapping. In this case, the chance of scoring is higher when there is less overlap between them. Therefore, the distance between BSs in sequential homotypic clusters might not have very much influence on TF occupancy.

Finally, we consider homotypic clusters with four binding sites. As shown in Subfigure E, Homotypic clusters with more BSs are more greatly influenced in the presence of TF 3D jumps between strands. There is a 58% improvement in TF occupancy when DNA strands are 100nm apart in the quadruple TF binding sites homotypic cluster case as opposed to a 42% improvement in the double TF binding site cluster case.

### 2.3.2 3D organizations of homotypic binding sites

Afterwards, we considered four more complicated 3D organizations of homotypic binding sites that would more closely resemble the case *in vivo*. We simulated a tetrahedron and a cube of homotypic clusters (Figure 2.2A and C, respectively). In both cases, there is either a single binding site or a pair of binding sites in a homotypic cluster in each corner of the shape. In the scenarios of homotypic clusters, the tetrahedron has 4 corners and 8 total binding sites and the cube case has 8 corners and 16 total binding sites.

We evaluated the impact of the presence of sequential homotypic clusters first. When

comparing the mean occupancy when DNA strands are 100nm apart in a tetrahedron organization, the cluster case (Figure2.2B) shows an increase of occupancy of 39% compared to the single site case (Figure2.2A). However, this percent increase in occupancy due to the presence of sequential homotypic clusters drops to 21% in the case of a cube organisation (Figure2.2C and D). It suggests that when inter-DNA strand jumping plays a more significant role, as is the case for the cube organisation, sequential homotypic clusters may have less impact on occupancy.

TF binding occupancy increases substantially when the DNA strands are close together, especially when the BSs are organized into more complicated configurations, due to the strong effect of inter-DNA strand jumping. When homotypic clusters are found in the corner of a tetrahedron (Figure2.2B), there is a 60% or 170% increase in TF occupancy when the DNA strands are 200nm or 100nm apart, respectively, compared to the extreme case in which the DNA strands are 10000nm apart, where there is negligible inter-stand jumping. In the cube case (Figure2.2D), there is a 118% or 277% improvement in TF occupancy when the the edge length of the cube is 200nm or 100nm. In summary, the spatial co-localization of multiple BSs of the same type of TF represents a potential strategy for increasing local TF occupancy.



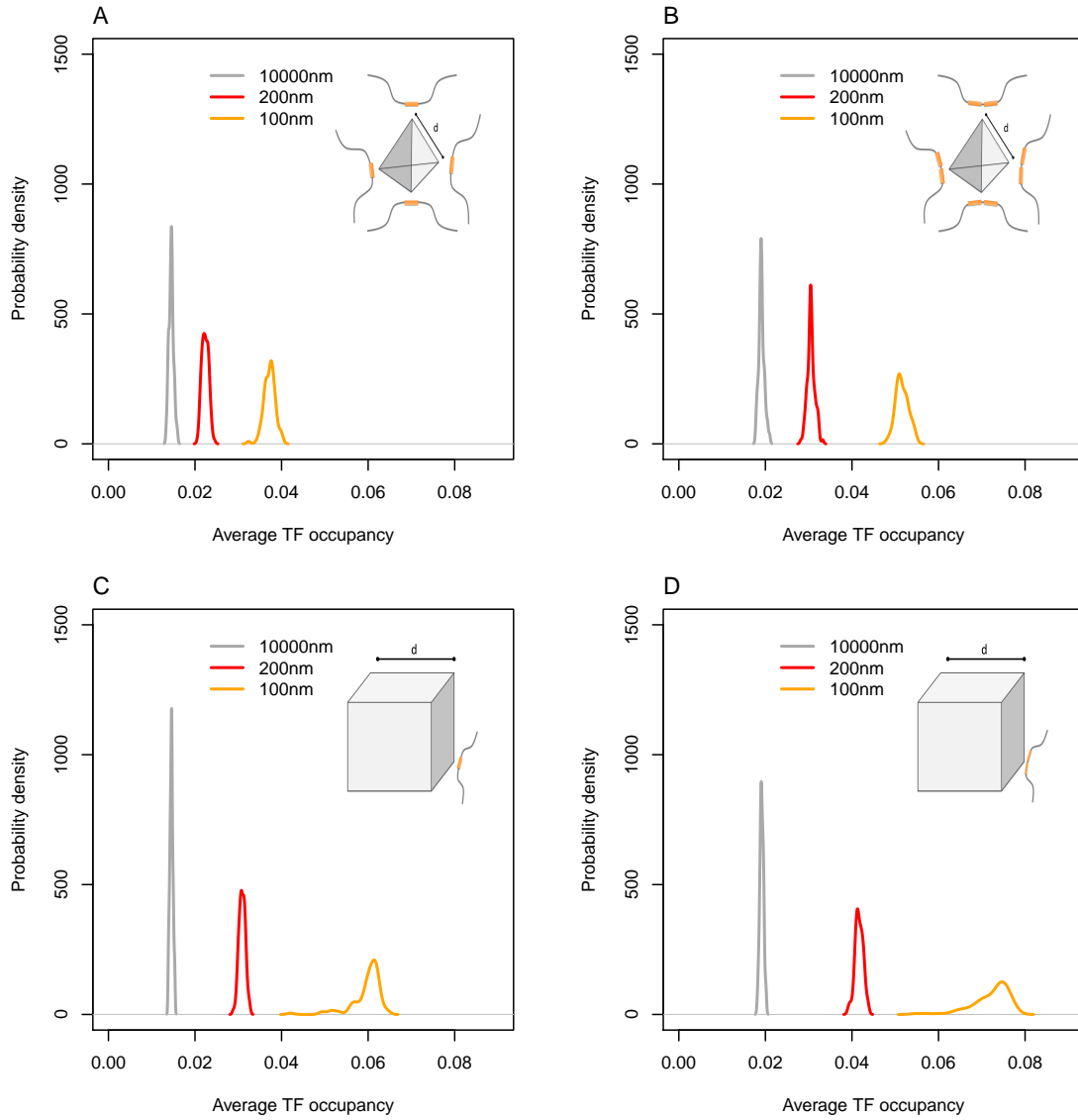


Figure 2.2: *Occupancy gain associated with complex binding site arrangements.* We simulate two more complex arrangements of homotypic clusters: a tetrahedron with one binding site in each corner(A), a pair of binding sites in a homotypic cluster in each corner(B), a cube with a single binding site(C) and double binding sites(D) in each corner.

### 2.3.3 TF binding site occupancy has strong linear correlation with spatial homotypic BS clustering

In our simulation of TF binding dynamics, the substantial increase in BS occupancy accompanied by homotypic BS spatial co-localization inspired us to further explore this *in vivo* with the available chromosome organisation data.

In Hi-C contact maps, the chromosome proximity ligation frequency can be viewed as an indicator of how likely two pieces of DNA can be spatially adjacent to each other. Given that the probability of observing a cis Hi-C contact has a strong dependence on the sequence separation between interacting sites, we used a measure of Hi-C contact enrichment above the background expectation (see Methods), as a way to quantify the strength of interaction between any paired loci.

With the availability of the high-resolution (5kb) intrachromosome Hi-C contact map in GM12878 cell line [1], we defined a quantitative measure for homotypic BSs co-localization levels around each genome locus, which we refer to as homotypic clustering score (HCS), as illustrated in Figure 2.3 and described by Equation 2.3.

Most sequence specific TFs can only recognize and bind to cognate DNA motifs within open chromatin, except very few pioneer factors which may also bind to closed chromatin [143]. Overall 96% of ChIP-seq peaks used in our analysis in human lymphoblastoid (GM12878) overlapped with DNase-I hypersensitivity hotspots (DHS). Therefore, we defined putative BSs by FIMO sequence motif scan ([168] within DHS, as described in Methods.

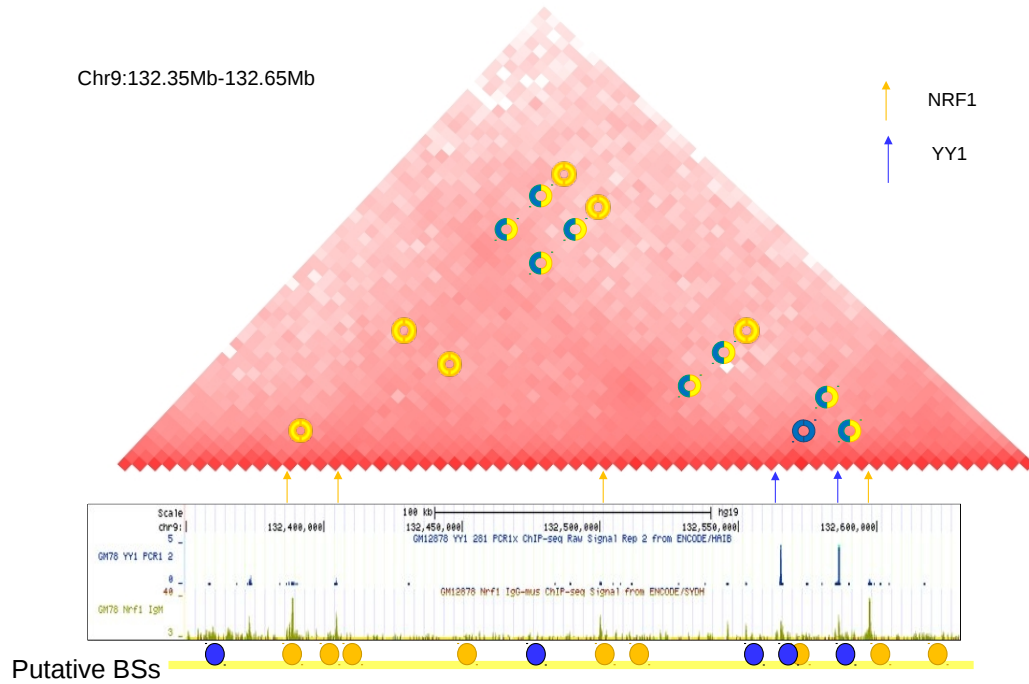


Figure 2.3: *Graphical overview of quantifying the spatial co-localization of TFs, both within the same type and between different types, using Hi-C contact data.* A section of the Hi-C contact map at 5 kb resolution (upper triangle) showing normalized contact counts of human lymphoblastoid GM12878 [1]. The darker colour corresponds to higher contact score. Binding sites of two TFs (YY1 and NRF1) identified by ChIP-seq are shown in the genome track below. Interactions between binding site pairs within the same type of TF are shown as either blue (YY1) or yellow (NRF1) circles, whereas interactions between two different TFs are shown with dual colours. Putative sites, identified by sequence motif scan, in DNase-I hypersensitivity regions for YY1 (blue) and NRF1 (yellow) are depicted in the bottom track.

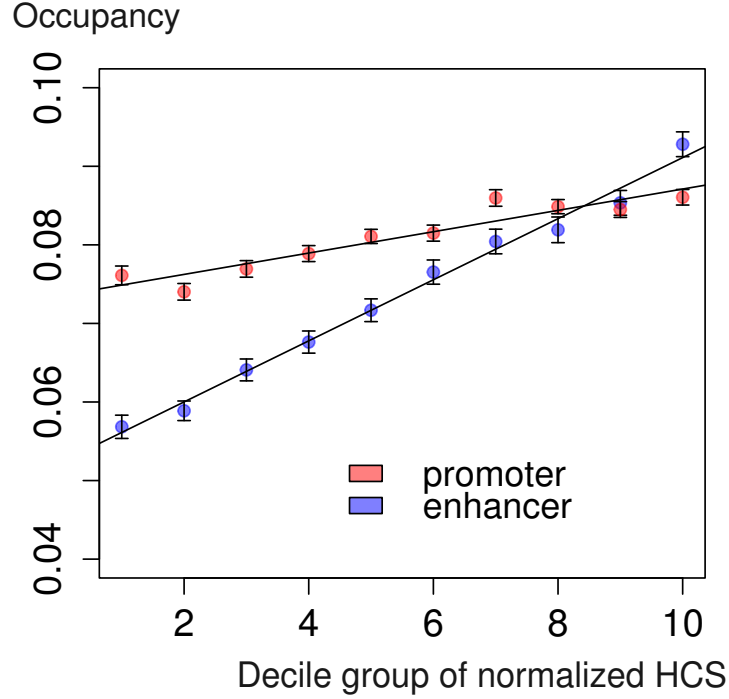


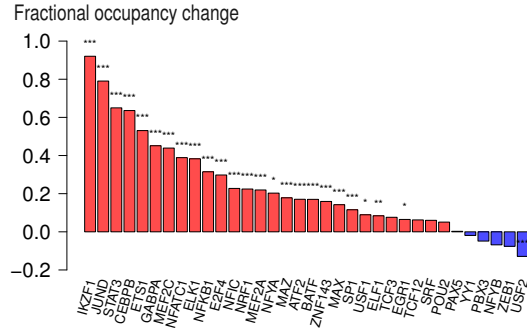
Figure 2.4: Correlation between BS occupancy and the co-localization of BSs of its own type as is revealed by Homotypic Clustering Score (Pearson’s  $R=0.98$  and  $0.91$ , respectively). The Homotypic Clustering Score (HCS) were derived from the Hi-C contact map for either the promoter or the enhancer state. BSs were grouped into 10 equal-sized bins (decile groups) according to HCS. (Error bars represent the standard deviation of calculated BS occupancy in each decile group using re-sampled BSs data by leaving  $1/3$  of BSs at each time and repeating 1000 times.

Combining the Hi-C contact map with the predicted genome-wide putative TF binding sites, we grouped the putative BSs for each TF into 10 groups of equal size(decile groups) according to rank normalized HCSs. By defining occupancy as the proportion of putative BSs occupied by ChIP-seq peaks, we observe a substantial gain of BS occupancy in relation to HCS in genome regions with either the enhancer or the promoter states (ENCODE consensus chromatin state [67]). However, the magnitude of occupancy increase is much larger in the enhancer state compared to that in the

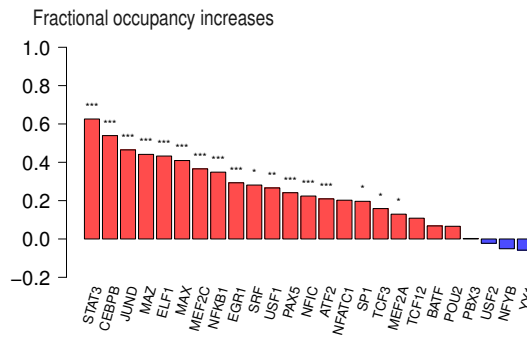
promoter state (Figure 2.4).

The above results demonstrate that among all TFs the amount of homotypic co-localisation is associated with TF occupancy, but it is unclear whether this rule holds for individual TFs. Thus, we grouped putative BSs for each TF according to their HCS into low (bottom 0.33 of HCS), mid and high (top 0.33 of HCS) groups. Triple grouping was done because of the limited number of ChIP-seq identified BSs for some TFs. For TFs with sufficient ChIP-seq data—at least 300 called ChIP-seq peaks in total—24 out of 34 TFs (71%) showed a significant increase in BS occupancy in the high HCS group compared to the low (G-test with William correction,  $p < 0.05$ , 22 of which have  $p < 0.01$ , Figure 2.5a), in contrast, only 1 TF (USF2) had significantly decreased occupancy in the high HCS group.

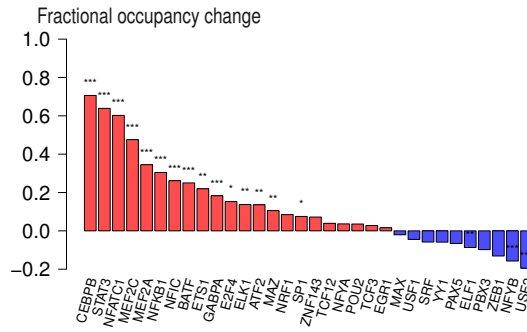
Next analysis was conducted for BSs that fell into certain chromatin states i.e. promoter or enhancer; TFs with less than 300 ChIP-seq peaks within each chromatin state group were not included in subsequent analysis. For the promoter state, 15 out of 32 TFs had a significant occupancy boost in the high HCS group (G-test with William correction,  $p < 0.05$ ), whereas 3 TFs (USF2, NFYB and ELF1) showed an occupancy decrease instead (Figure 2.5c). For the enhancer state, we observed a significant occupancy boost in 17 out of 25 TFs (G-test with William correction,  $p < 0.05$ ), while no TF displayed the reversed trend (Figure 2.5b), which indicates a more pronounced positive relationship between spatial homotypic BS co-localization level and BS occupancy in the enhancer state compared to the promoter state.



(a) All



(b) Enhancer regions

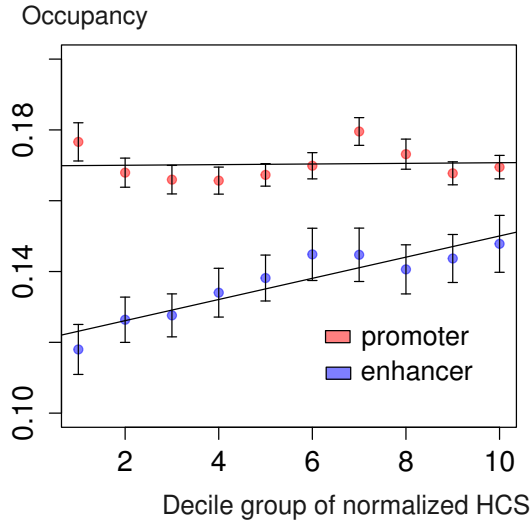


(c) Promoter regions

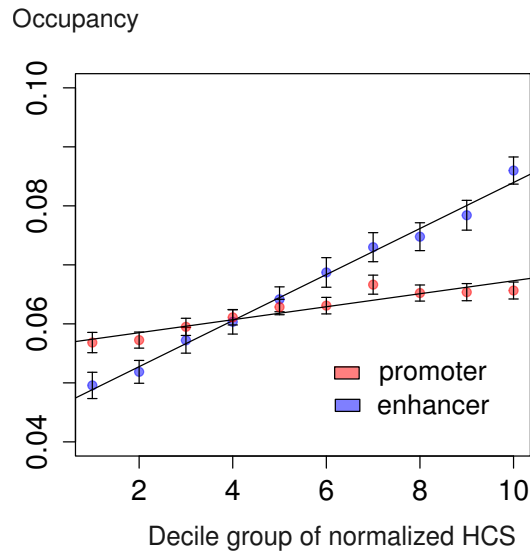
Figure 2.5: The fraction of BS occupancy increase in the high HCS group compared to the low group for each TF. Shown are the bar plots with respect to all BSs for each TF (a), BSs of the enhancer state (b) and the promoter state (c). Stars above the bar plots denote the level of significance (single star:  $0.01 < p < 0.05$ ; double stars:  $0.001 < p < 0.01$  and triple stars:  $p < 0.001$ ).

For instance, ELF1 had a significant occupancy boost in the high HCS group within the enhancer state ( $p = 3.2 \cdot 10^{-9}$ ), but an occupancy decrease in the promoter state ( $p = 2.2 \cdot 10^{-3}$ ). It suggests a more sophisticated binding dynamic may be involved in shaping its binding within the promoter state, which cannot be captured by simple diffusion driven binding dynamics. In addition, NFY and USF family TFs have been shown previously to have special DNA binding properties i.e. they can bind to motifs without "positive" histone marks or even containing H3K27me3 marks [175]. Furthermore, NFY co-associates with FOS and is stereo-positioned with growth-controlling transcription factors, which may also contribute to its binding occupancy complexity [175].

We wondered whether there are other factors that may contribute to the differences in the level of occupancy boost, such as the binding motif strength. Therefore, we classified putative BSs based on BS sequence motif strength, which is related to DNA binding affinity, into weak putative BSs ( $10^{-4} > p$  value reported by FIMO motif scan  $> 10^{-5}$ ) and strong putative BSs ( $p < 10^{-5}$ ). Interestingly, in the enhancer state, weak putative BSs showed 72% of occupancy boost when comparing the ones in the top 10% versus the bottom 10% in terms of HCS, while strong ones only have 23% of occupancy increase (Figure 2.6a,b). For the promoter state, the occupancy of weak putative BSs showed good correlation with their HCS (Pearson's  $R=0.93$ ,  $p = 9 \cdot 10^{-4}$ ), though there is only 18% difference between the top 10% versus the bottom 10% in terms of HCS; in contrast, the occupancy of strong putative BSs showed no significant correlation with HCS ( $p=0.54$ ).



(a) Sites with strong motifs

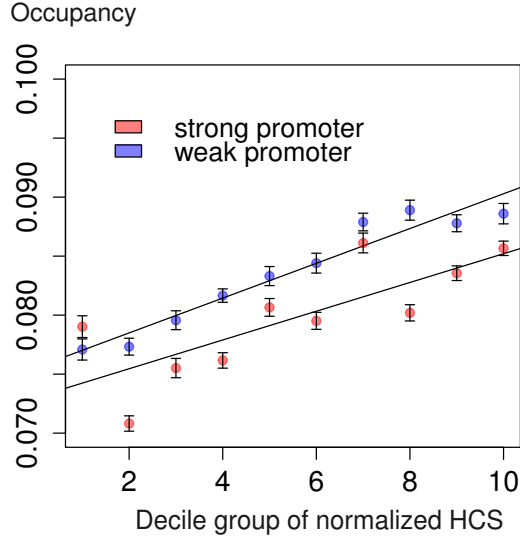


(b) Sites with weak motifs

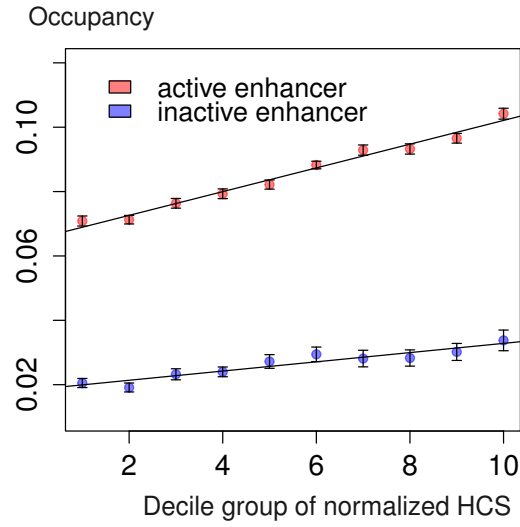
Figure 2.6: Relationship between BS occupancy and homotypic co-localization of BSs. BSs were further classified according to DNA binding sequence motif strength into weak (a) or strong BSs (b) based on their sequence motif strength.



The divisions into enhancer and promoter chromatin states were a reflection of enriched histone mark patterning; however, it was unclear whether a particular histone mark per se was responsible for the differences observed between these two chromatin states. In addition, there might be subgroups of chromatin states that may have different effects on spatially-lined occupancy boosts, which could be missed by our previous analysis. For instance, since H3K36Me3 was reported to be a sign of highly transcribed genes [172], it can be used in conjunction with other histone marks to represent the promoter strength. Therefore, for putative BSs in the promoter state within 2000bp upstream of TSS, we grouped those associated with H3K4Me3, H3K27Ac and also H3K36Me3 in 2000bp up- or down-stream of their TSS into a category named strong promoters (highly active promoters), in contrast, those only associated with H3K4Me3 and H3K27Ac but without H3K36Me3 are called weak promoters. Interestingly, there is a strong correlation between BS occupancy and HCS (Pearson'  $R=0.96$ ) (Figure 2.7a) in weak promoters, but the correlation becomes weaker in strong promoters ( $R=0.74$ ) (Figure 2.7a), though the overall binding occupancy remains at similar levels. There is no significant difference in the ratio between strong and weak BS number in the above two groups, suggesting that the differences between the two groups are not a consequence of differences in BS strength. These results indicate that spatial homotypic BSs co-localization plays a more essential role in boosting BS occupancy in weak promoters than strong promoters—BS occupancy in strong promoters might be strongly affected by other factors, beyond the diffusion related biophysical mechanisms, and not require such a strong TF support network. In addition, in the enhancer state, we grouped putative BSs based on whether they were found in active- or inactive-enhancers, according to whether H3K27Ac or H3K27Me3 marks were presented. Note that all of these regions also contained H3K4Me1, a typical mark presented in enhancers. We observed a good correlation between TF occupancy and the HCS groups in both active and inactive enhancers ( $R=0.97$  for active enhancer and  $R=0.94$  for inactive enhancer, Figure 2.7b), though the overall occupancy of active enhancer was more than 2 fold higher than that of inactive enhancer.



(a)

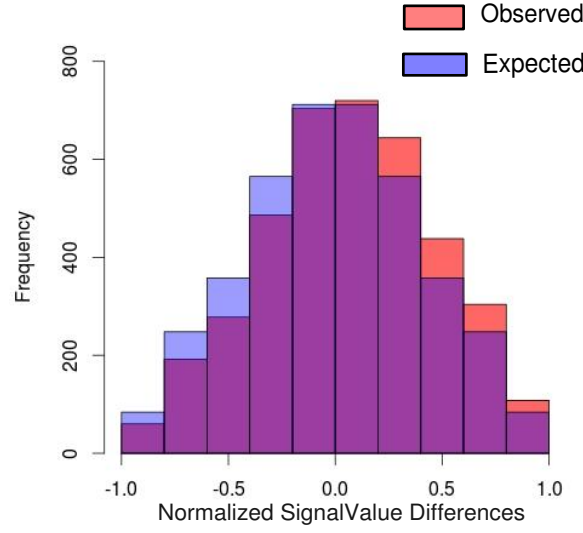


(b)

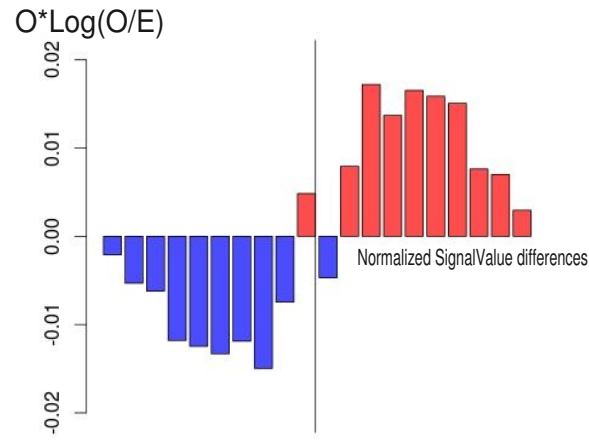
Figure 2.7: Linear correlation between BS occupancy and homotypic BS co-localization in weak promoters without H3k36me3 or strong promoters (highly active promoters) with H3k36me3 (a); active enhancers with H3k27ac or inactive enhancers with H3k27me3 (b). Note H3k4me3 and H3k27ac was required to be present in all promoter regions, while H3k4me1 was required in all enhancer regions.

### 2.3.4 Homotypic BS co-localization lead to ChIP-seq Signal-Value gain

The DNA sequence of TF binding motifs can have a strong impact on BS affinity and binding occupancy. In previous sections, we defined a putative BS as having a motif-score above a threshold within DHS, but we did not differentiate these BSs by their sequence composition in the subsequent analysis. To account for the influence of binding site affinity differences induced by BS sequence variations, we paired each BS within a region with high levels of spatial co-localisation (high HSC group) with a BS that had exactly the same DNA sequence but that was found in a DNA region with low levels of spatial clustering (low HSC group). In addition, we made sure to assign pairs of BS that had exactly the same histone marks, chromatin states, and chromosome sub-compartments. Using rank normalized ChIP-seq SignalValue (a measure for read enrichment in peak regions adopted by ENCODE uniform peak callers) [67] as the measure for peak signal strength, we observed a significant SignalValue increase in the high HCS group (Wilcoxon signed rank test,  $p = 1.3 \cdot 10^{-8}$ ). There were 16 TFs with at least 300 paired BSs– 10 out of the 16 TFs showed significant SignalValue increase in the high HCS group (Wilcoxon signed rank test,  $p < 0.05$ ), while only one of them showed decreased SignalValues (USF2). The SignalValue comparison between the paired BSs of high versus low HCS groups of a TF named NFIC is depicted in Figure 2.8a as an example, while the enrichment plot of the observed versus expected SignalValue differences is shown in Figure 2.8b and also in Figure 2.9 for other TFs with sufficient ChIP-seq peaks. The above comparison of ChIP-seq SignalValue between paired BSs of exactly the same sequence motif and chromosome environment provides additional support to the idea that homotypic BS clustering can significantly boost TF binding. The results show good consistency with BS occupancy analysis described before.

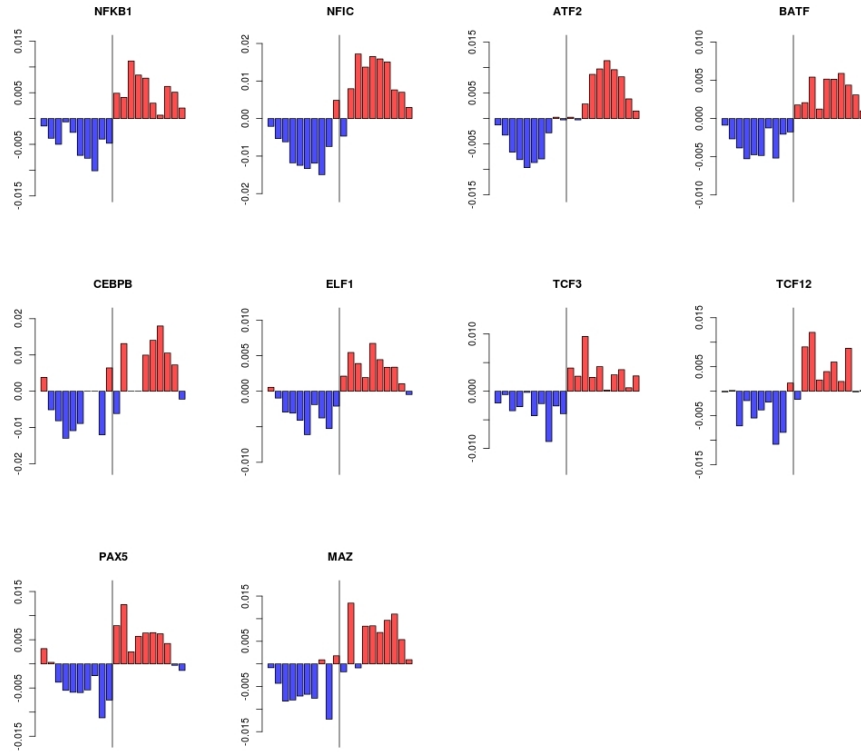


(a)

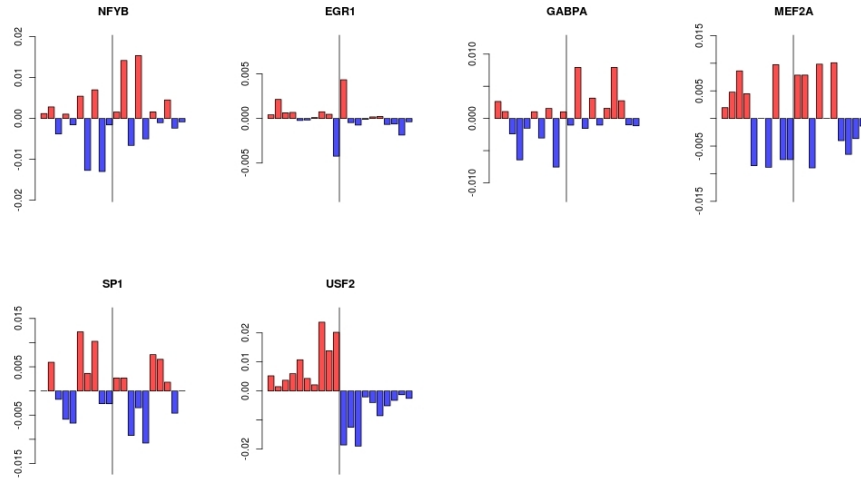


(b)

Figure 2.8: a: The histogram of the observed and the expected distribution of normalized ChIP-seq SignalValue differences between the paired BSs of high versus low HCS groups of a TF named NFIC as an example (Wilcoxon signed rank test,  $p = 9.1 \cdot 10^{-5}$ ). b: The enrichment plot of ChIP-seq SignalValue increase associated with high HCS for NFIC.



(a)



(b)

Figure 2.9: For each TF, the enrichment plots of ChIP-seq SignalValue increase associated with high HCS. a: TFs with significant SignalValue increase associated with high HCS; b: TFs without significant SignalValue increase or even with decreased SignalValue within the high HCS group (USF2).

## 2.4 Discussion

We investigated sequence-specific transcription factor DNA binding dynamics and occupancy in relation to spatial BS homotypic clustering using a combination of computational simulations and in-vivo binding profile analysis. Both of the approaches confirm the fact that spatial arrangement of TF BSs have strong influences on TF binding occupancy, and this holds for most of TFs we studied so far and for BSs within either the promoter or the enhancer state, albeit to different degrees. We note that the linear correlation we observed between BS occupancy and the HCS groups does not necessarily mean a linear response of occupancy to homotypic clustering, but rather a hint that homotypic Hi-C contact enrichment can be used to inform on TF binding efficiency. There are so many factors that collectively contribute to TF binding. Even in the simplest case only with respect to physical diffusion and binding kinetics, the non-linear component in binding dynamics is involved.

Our results suggest that genome organisation has a strong impact on the function of weak regulatory elements in terms of TF binding. Particularly, BSs in the enhancer state and in weak promoters witness stronger correlations and larger magnitude of occupancy increase when they are found in spatial proximity. Moreover, BSs with weak putative DNA binding sequences also showed much larger magnitude of occupancy increase in relation to homotypic BS clustering compared to those with strong sequence motifs, and this holds for either the promoter or the enhancer state.

Weak regulatory elements play an important role in modulating gene expression across cell types [5, 176]. There were several studies revealed that enhancer specificity depends on binding motifs having reduced binding affinities, especially for cell-type specific gene expression [177, 178]. Hentsch *et al* showed that the conversion of weak binding sites to strong canonical binding sites of several lymphocyte specific TFs including NfκB and AP-1 results in the induction of certain T-cell specific enhancers in non-T cells, thus disrupting gene expression patterns in other cell types [179]. The way by which cell-type dependent binding of weak BSs can be achieved is not well understood. Tissue-specific signal response and endogenous TF concentration has

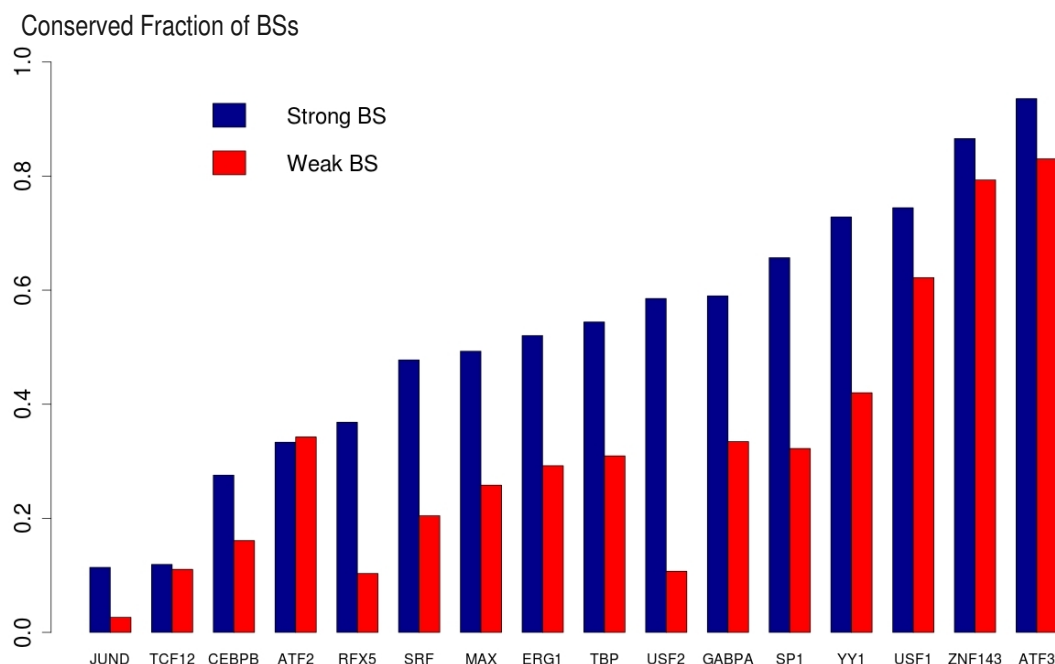


Figure 2.10: The fraction of conserved ChIP-seq peaks associated with either weak or strong BSs between GM12878 and hESC. Shown are TFs with ChIP-seq profiles available in both cell lines

been suggested to be potential factors that give rise to the fine-tuning of TF binding, especially for weak BSs [177]. There were other contradicting studies suggested that weak BSs in the genome may not be functional in terms of determining transcription output [180]. However, these weak BSs, especially the ones within enhancers and promoters, are often under strong purifying selections across different species [73], indicating that weak BSs may be of biological importance actually.

Based on the available ChIP-seq profiles for common TFs in GM12878 and h1-ES cells, weak BSs are much less conserved than strong BSs between these two cell lines (Figure 2.10, Wilcoxon signed rank test,  $p=3.2E-5$ ). As is mentioned above, there could be various reasons that lead to the less conserved binding associated with weak BSs across tissues [177]. Our analysis proposed another possible mechanism that spatial BS organisation may help weak BSs to convey tissue specific TF binding.

Chromosome organisation, especially inter-TADs interactions, has been shown to vary drastically between different cell types [122]. These chromosome structural variation can potentially give rise to variations in BS spatial arrangements. As TF binding occupancy in weak BSs are more sensitive to BS co-localization than strong BSs according to our results, chromosome architectural changes may serve to better fine-tune weak BSs occupancy in different tissues and throughout the time-line of embryonic development.

Most of TFs respond positively to spatial homotypic BS clustering, as is revealed by either BS occupancy or ChIP-seq SignalValues. However, there is one very unique TF, USF2, associated with both decreased occupancy and ChIP-seq SignalValue in genome regions with high level of homotypic BS clustering, opposing the trend of most other TFs. USF2 is a c-Fos interacting protein with a bHLH-zip domain that recognizes the CACGTG DNA motif (class I E-box). We noticed there was another TF in our analysis, USF1, which shares a very similar DNA binding domain structure with USF2 and both of them are in the same TF family [181], however, USF1 followed the trend of homotypic BS clustering induced occupancy gain well (Figure2.5). The structural difference between USF1 and USF2 comes from their transcriptional activation domains. While USF1 only contained an extended activation domain, USF2 contained both an activation domain and a negative regulatory region in N-terminal [182], which was absent in USF1. This negative regulatory domain, consisting of around 200 amino acids, functions as a dominant-negative regulator of class I E-box enhancer activity, which makes USF2 special, though the mechanism of its repression function has not been studied in relation to histone occupancy or protein-protein interactions to date. We hypothesize the presence of this domain might also contribute to its uniqueness in binding dynamics associated with high level of BS clustering.

We have provided the updated FastGRiP simulation tool to characterize TF binding dynamics and occupancy in presence of spatial homotypic BS clusters. We have showed that inter-strand jumping of TFs significantly boost overall BS occupancy, which is sensitive to the distances between DNA strands containing homotypic BSs. In our simulation, without the knowledge of specific sub-diffusive properties associated with molecules within different chromosome regions, we made the general assumption



that the diffusion behaviour of TFs can be captured by simple diffusion with adjusted diffusion co-efficient taking into account DNA crowding. However, *in vivo* diffusive properties of molecules are much more complicated and not fully studied for different TFs under different macromolecule crowding conditions. Anomalous diffusion with variable diffusion co-efficient depending on molecular concentration, pressure or other factors, in contrast to simple diffusion with a constant diffusion co-efficient, may be present in biological systems. For example, there are studies confirming the occurrence of anomalous diffusion in fractal organisation of crowding polymers, which can be characterized by fractal kinetics [183]. The key feature associated with fractal kinetics is that the diffusion co-efficient is time dependent according to  $t^{-a}$ , where  $a$  is the fractal exponent of the reaction related to the fractal architecture of the chromatin. They showed that the fractal kinetic model could reproduce the subdiffusive behaviour of fluorescence molecules observed in heterochromatin, which cannot fit into the reaction diffusion model assuming random organisation of obstacles in diffusion paths, especially in a short length scale of below 100nm. In contrast, in euchromatin and when it comes to larger scales of heterochromatin (above micrometer scale), the reaction diffusion model is already sufficient to explain the experiment observations [184]. Our diffusion model is mainly used for describing the TF diffusion behaviour in euchromatin, which is enriched in TF BSs. If it needs to be generalized to heterochromatin, the fractal kinetics should be taken into consideration.

In our TF binding dynamics simulation with updated FastGRiP, we have considered the effects of both sequential and spatial BS homotypic clusters. However, the effect of sequential clustering on BS occupancy is difficult to investigate precisely using the available ChIP-seq data, because each ChIP-seq peak could cover more than a few hundred base-pairs, while the sequential homotypic BS clustering occurs mostly within that scale. If you are interested in sequential homotypic BS clustering, please referred to the following established work concerning the role of sequential homotypic BS clusters [70, 72].

Admittedly, the determining factors to the *in vivo* binding landscape are far more complicated than the simple scenario of homotypic BS clustering we studied here. For instance, also from a perspective of spatial organisation of BSs, multiple types of TFs

could possibly co-localize together and that may influence the binding of each TF due to either TF collaborativity or the interplay between TFs and histones. Therefore, we wish to broaden our scope to investigate the spatial co-localization between different TFs and how that contribute to TF binding with respect to individual TFs in our next chapter.

# Chapter 3

## Transcription factor spatial co-localization networks in lymphoblastoid cells

### 3.1 Introduction

We have focused on the spatial clustering of BSs within the same type of TF in the last chapter, because homotypic BS clustering is easy to investigate and the mechanical advantages it provides can be well-explained by simple biophysical models. However, in vivo transcription regulation is achieved by complicated co-binding and co-function of a great variety of TFs, especially in higher eukaryotes [5][185]. Furthermore, the genome organisation has been shown to be closely linked to transcription regulation [122]. Enhancers and promoters tend to form loop like structures to facilitate gene transcription, as was observed in the mammalian alpha-globin loci and between Hox gene clusters [186, 187]. In those scenarios, looping of DNA is unlikely to be achieved by a single regulatory protein, but rather via regulatory protein complexes, which may involve a great diversity of molecules. Therefore, to gain a better understanding of transcription regulation in higher eukaryotes requires linking genome architecture

and multiple types of regulatory protein binding together.

Deciphering interaction and co-function across different types of regulatory proteins is not an easy task. Most studies probing physical interactions of proteins fall into the following two categories: (1) *in vivo* methods including protein pull-down and tandem affinity purification followed by Mass Spectrometry [188, 189] and (2) *in vitro* approaches including yeast two hybrid [190], peptide array [191] and GST fusion protein pull-down [192]. However, those strategies only focused on global protein-protein interactions, without the context of the chromosome. Moreover, the interaction partners detected using those methods required strong and stable physical interactions between the bait protein and their partners, while transient and dynamic interactions between proteins were missed. There are some approaches trying to explore TF binding in the context of local chromosome organisation, for instance, ChIA-PET [193], which involves the combination of immuno-precipitation of DNA binding proteins and capturing of chromosome contacts associated with TF binding regions. However, this technique has been shown to suffer from several technical drawbacks [194] and only a single TF was involved at each stage, such that it was not possible to probe interaction and co-localization between multiple TFs. Furthermore, in ChIA-PET, only a small subset of genome regions associated with a certain TF binding could be investigated at each stage, which made genome wide analysis not applicable.

There are very few studies directly probing the TF co-localization pattern spatially, despite the existence of extensive research focusing on one dimensional sequential TF co-localization in the genome [68, 73]. Live imaging of two interacting proteins, c-Fos and c-Jun, has revealed their molecular enrichment and dual binding map within the nucleus [100, 96]. Using fluorescent cross-correlation spectroscopy, Pernu et al found a strong correlation between TF diffusional mobility and the interaction between these two TFs. Dimerization of the TFs significantly slowed down their motion and enriched the co-localization level of these two TFs in the nucleus. Their results suggested that TF spatial co-localisation could be strongly influenced by TF-TF physical interaction, and studying TF spatial co-localization patterns can improve our understanding of TF interaction and co-function in gene regulation. However, due to limited number of channels and the availability of fluorescent tagged protein,

identifying TF co-localization patterns using direct imaging is still limited to a small number of TFs at one time. The extent to which TF interaction networks are also spatial networks, involving co-localization of TF BSs, is not yet known for most TFs.

TF binding global patterns from a spatial point of view was difficult to investigate quantitatively until very recently, with the advent of genome wide chromosome conformation capture techniques, including Hi-C [194, 1]. The chromosome contact maps produced DNA proximity information, which can be used to probe genomic interactions such as chromosome loops and promoters-enhancer interactions [107]. The contact maps revealed that chromosomes are segmented into topologically associating domains (TADs), within which physical contacts occur more frequently [107]. At the megabase scale, the contact map suggests the spatial segregation of open and closed chromatin, namely, A and B chromosome compartment [107]. Recently with the availability of higher resolution Hi-C maps (down to a kilo-base scale), the two compartments could be further partitioned based on their distinct patterns of long-range contacts. For instance, the inter-chromosome contact map of human GM12878 cells was subdivided into at least six different subcompartments [1], two of which were enriched in actively transcribed genes— namely the A1 and A2 subcompartments. Even though these two subcompartments were clearly distinct in the contact map, they were similarly enriched in active histone marks and open chromatin, though A2 was slightly more enriched in H3K9Me3, leaving the question of what makes these two subcompartments unique. Given the very similar epigenetic properties between A1 and A2 chromosome subcompartments and similar gene transcriptional activity, we wish to explore further if their spatial segregation is associated with differentiated TF binding or co-localization preferences, as TFs can contribute to chromosome organisation in addition to CTCF and cohesin complexes [1, 112].

The non-randomness of chromosome organisation has been hypothesised to be influenced by the combinatorial binding of different kinds of protein complexes using polymer dynamics simulations [132, 195]. Using a simple polymer model, named strings and binders switch model, Barbieri et al showed that a chromosome domain-like structure can be established through attachment of diffusible factors to their specific binding sites in the genome. In the simplest scenario, only two different kinds

of protein complexes, each with multiple DNA binding facets and well-segregated DNA binding regions, were sufficient to drive the segmentation of chromosomes into A/B compartment-like structures. The simple polymer model also identify several scaling properties of chromatin folding observed in vivo, including the fractal state of chromatin [132]. It also showed that the folding properties of chromatin can change in response to changes in binding site distribution, protein concentration, or binding affinity. However, those are only predictions from various kinds of models. To date, very few studies have provided direct evidence in vivo to support those hypotheses and elucidate the cause/effect between regulatory protein binding and chromosome compartment/subcompartment formation[132].

Furthermore, as we showed in the last chapter, there is a strong positive correlation between binding site occupancy and the level of homotypic BS spatial co-localization. Thus we want to further ask if heterotypic BS co-localization can similarly influence BS occupancy. Even though the mechanism cannot be simply explained by the facilitated diffusion properties of TF binding similar to homotypic BS clustering, complicated protein-protein interactions are prevalent among DNA binding molecules. Sequential co-localization of BSs of physically interacting TFs has already been shown to promote TF binding to DNA and the formation of stable protein complexes [196]. For instance, Sox2 and Oct4 dimerization is crucial for the binding of Oct4 to its target sites [84]. There was a substantial decrease in the mean Oct4 target search time and an enhancement of the long-lived DNA bound fraction of Oct4 molecules when Sox2 was present in the cell [84]. Though there is no direct evidence to date of binding enhancement due to spatial co-localization of BSs, mechanically, BSs of interacting partner TFs found in spatial proximity could be favoured by the TF target BS search, if assuming transient protein interactions act as transient traps for diffusing TF molecules. This motivated us to further examine the relationship between TF binding occupancy and BS spatial co-localization, particularly between different types of TFs.

Therefore, we wish to explore spatial co-localization between different TFs at a genome-wide scale, using computational analysis combining chromosome architecture information gained from Hi-C and TF binding profiles from ChIP-seq.

## 3.2 Methods

### 3.2.1 Heterotypic co-localization score between TF pairs

Similar to the HCS of a specific TF, the Hi-C heterotypic co-localization score between two TFs, namely A and B, is defined for each binding site  $i$  of TF A as:

$$heteroCS_{i,AB} = \sum_{j \in B} \log \frac{Obs_{i,j}}{Exp_{i,j}} (i \in A) \quad (3.1)$$

where  $heteroCS_{i,AB}$  stands for the heterotypic co-localization score between the two TFs for each site  $i$  of TF A, considering all possible interactions with TF B sites on the same chromosome. Note that this score has direction *i.e.*  $heteroCS_{i,AB}$  is calculated for each site of TF A, while  $heteroCS_{i,BA}$  is for each site of B.

In order to compare the observed  $heteroCS_{i,AB}$  score distribution of TF A to the expected, as control, we generated randomized TF A sites by permuting BSs of all available TFs for each chromosome 1000 times, while keeping TF B sites fixed. Note the number of BSs for each TF on each chromosome was kept the same in the above permutation. It gave us the expected  $heteroCS_{i,AB}$  score distribution. Similar procedure can be used with respect to  $heteroCS_{i,BA}$ . In addition, 25kb adjacent region (corresponding to 5 Hi-C map bins) left and right to the locus of interest was removed to avoid potentially high noise near the diagonal of the Hi-C contact map.

Since chromosome sub-compartments [1] may have potential influence on TF binding, instead of randomly shuffling all BSs on the same chromosome, we also constructed the control set in the way that BSs were randomly shuffled within each sub-compartment for each chromosome, which preserves the BS composition in each subcompartment. TFs with very low number of ChIP-seq identified BSs (less than 300) in either A1 or A2 subcompartment were not included in our subsequent analysis.

We used Kullback-Leibler divergence to represent the extent to which the observed distribution differs from the expected, in other words, the spatial co-localization re-

relationships between TFs. The Kullback-Leibler distances (KL distance) between the observed co-localization score distribution and the expected distribution were calculated as follows, which is denoted by co-localization enrichment score (CES)

$$CES = (sign) \cdot \sum_k P_{obs,k} \log \frac{P_{obs,k}}{P_{exp,k}} \quad (3.2)$$

where  $CES$  represents the KL distance (with a sign) between the observed and the expected distribution of heteroCS3.1. The sign of the formula depends on the right (+) or left (-) shift of the observed mean co-localization scores from the control set.

We performed average-linkage hierarchical clustering of TFs based on TF co-localization scores defined above by using a distance measure below:

$$e^{-(CES(AB)+CES(BA))/2} \quad (3.3)$$

We adopted the R package DynamicTreeCut [197] and used the setting of DynamicTree mode to define clusters of TFs based on the dendrogram from the above hierarchical clustering.

By comparison, the average linkage hierarchical clustering as well as the Wards clustering methods [198] based on the squared-Euclidean distance between every pair of rows in the  $CES$ -score matrix were also performed. We noticed that the distance measure either from equation 3.3 or just simple squared-Euclidean distance gave similar results in most of cases, but equation 3.3 out-performs squared Euclidean distance when it comes to analysing the co-localization within A2 subcompartment, possibly because the information regarding to co-localization enrichment is used in a more direct way in the former approach.

### 3.2.2 Calling significantly co-localized TF pairs

Furthermore, we called significant co-localization of TF pairs based on the distribution of co-localization scores. For co-localized TF pairs, there would be an enrichment of



BSs with high spatial proximity, which are indicated by the high co-localization score groups in the right tail of the *heteroCS* distribution. For each TF pair, we calculated empirical p-values for the observed frequency of BSs compared to the randomly shuffled BS control sets (1000 times of permutation) in high co-localization score groups (the top 20%, 10% and 5% in the score distribution were examined). We called significantly co-localized TF pairs by using FDR threshold of 0.05 [199] and requiring significant enrichment of BSs within top score groups in both pairing directions (*heteroCS<sub>i,AB</sub>* and *heteroCS<sub>i,BA</sub>*). In comparison, we also called co-localization pairs within either A1 or A2 subcompartments similarly using heterotypic co-localization scores and randomly permuted control sets within each subcompartment.

### 3.2.3 TF BS conservation between two cell lines

To compared BSs between h1-ESC and GM12878, ChIP-seq peaks in h1-ESC were matched to corresponding GM12878 ChIP-seq peaks, defined as the h1-ESC ChIP-seq peak that overlapped with the center of the GM12878 ChIP-seq peak, such that the center-to-center distance of ChIP-seq peaks in the two cell lines was less than 300bp. The fraction of mapped ChIP-seq peaks in ESC was used as the indication of BS conservation level.

### 3.2.4 Calculation of integrated heterotypic co-localization score for a group of TF

Due to the presence of TF spatial co-localization groups, different TFs within the same group may have additive effects on creating a molecule crowd-sourcing environment. In order to investigate the behaviour of TFs within a specific co-localization group as a whole, we defined the integrated heterotypic co-localization score at position  $i$  for

TF A( $SumHeteroCS_{i,AG}$ ) in respect to group G composing  $k$  different TFs:

$$SumHeteroCS_{i,AG} = \sum_{B \in G}^k heteroCS_{i,AB} \quad (3.4)$$

where  $heteroCS_{i,AB}$  is the heterotypic co-localization score between TF A and B for each site  $i$ .

### 3.3 Results

#### 3.3.1 TFs were partitioned into two spatial co-localization networks in GM12878

Many TFs are known to have interactions or collaboration with other TFs to regulate gene activities [200, 201], but whether this can be seen from TF spatial proximity is unclear. Therefore, we make use of the high-resolution chromosome contact map to investigate the spatial co-localization properties between heterotypic pairs of TFs based on ChIP-seq identified binding sites. We defined a measure of spatial co-localization between two TFs (hetero-CS), based on the Hi-C contact map, in a manner similar to homotypic clustering score (Equation2.3); however, the observed-versus-expected score ratio was defined between BSs of two different TFs as is shown in Equation3.1 and depicted in Figure2.3). By expressing the heterotypic co-localization score distribution for different pairs of available TFs as a single enrichment value, representing whether pairs have more or less co-localisation than expected (see Equation3.1 and Equation3.2), we are able to investigate different clustering behavior across all possible TF pairs.

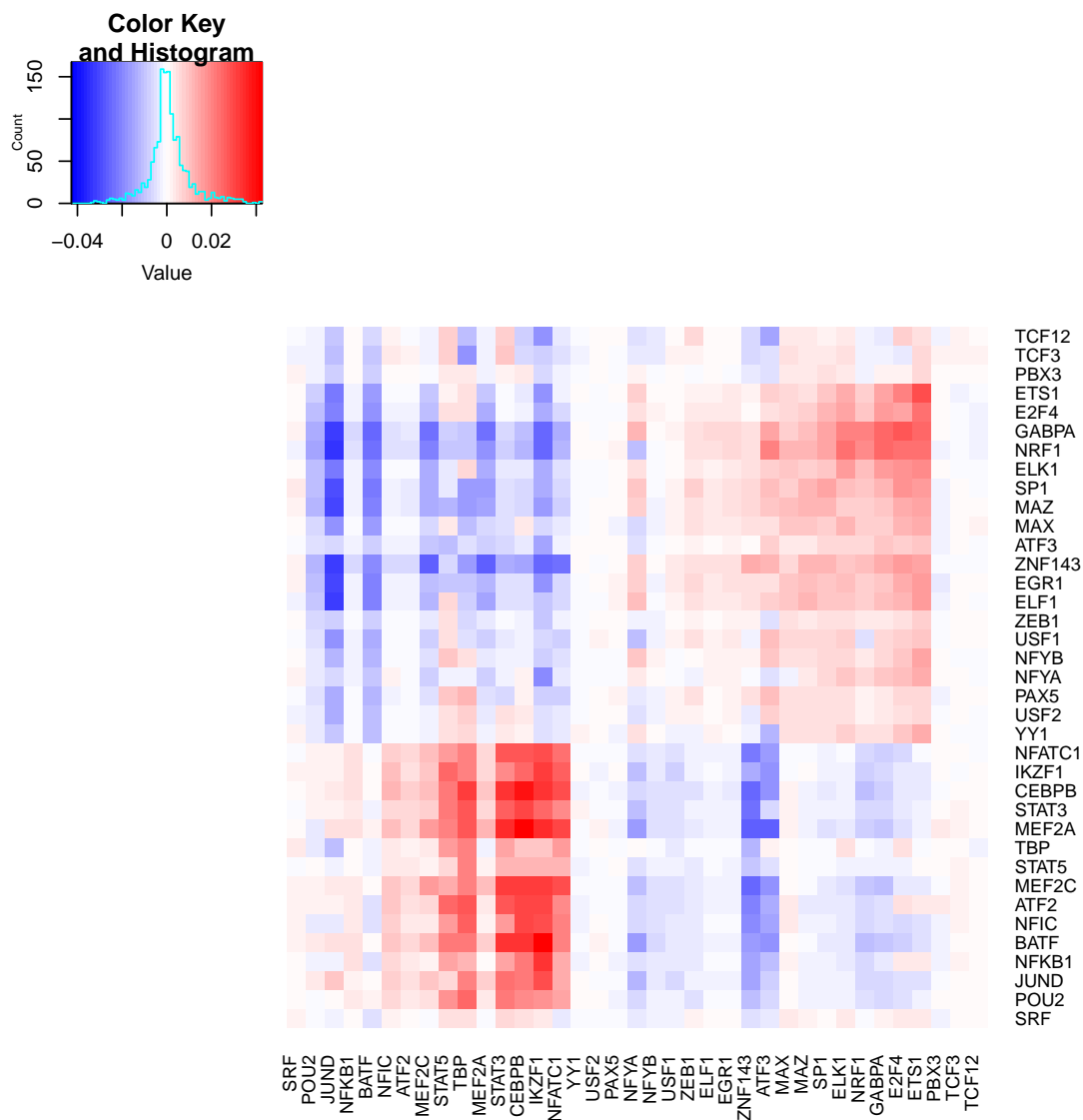


Figure 3.1: *TF co-localization enrichment map in human GM12878 cells.* Colours indicate the enrichment score of spatial co-localization between two TFs based on the Kullback-Leibler distances (KL distances) between the observed and the expected heterotypic co-localization score distribution for each TF pair, where red or blue color represents higher or lower than expected respectively. Hierarchical clustering was used to generate the layout of the colour map.

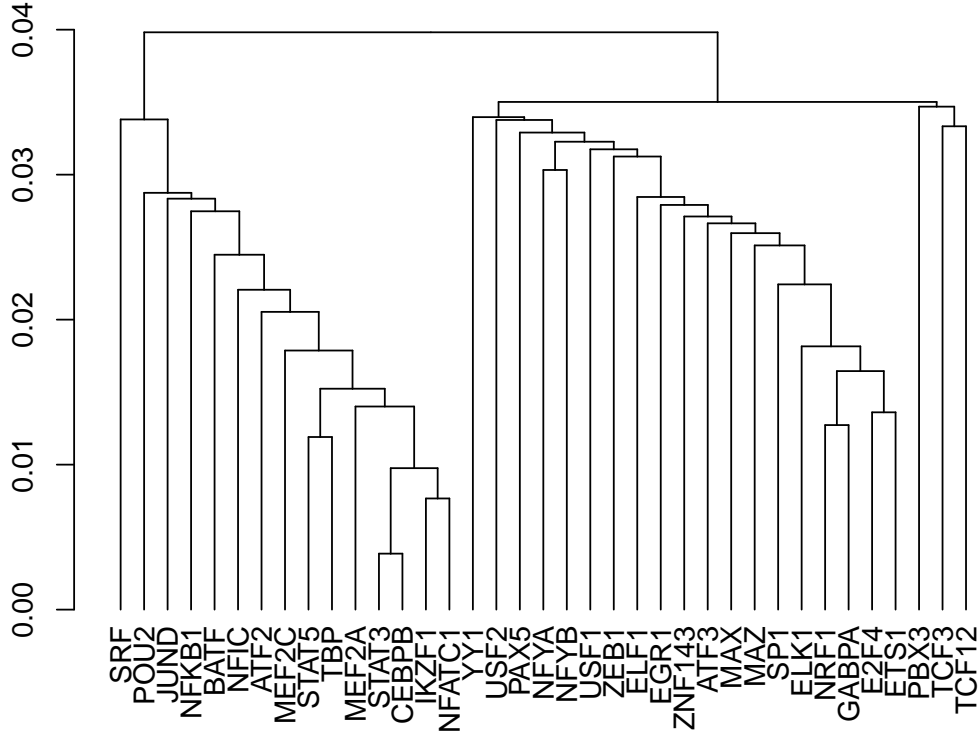


Figure 3.2: The dendrogram of the TF co-localization hierarchical clustering in Figure 3.1.

The grid of spatial co-localization enrichment scores (CES) for the available human lymphoblastoid TFs is plotted via hierarchical clustering in Figure3.1 with the dendrogram in Figure3.2. There are two distinct groups of TFs that show higher than expected intra-group co-localisation and less-than-expected inter-group co-localisation, namely TF co-localization network 1 and 2 (TFCN1 and TFCN2), while there are some other pairs of TFs (e.g YY1 and PAX5) for which the random expectation fits well. (See Table3.3 for the full list of TFs in TFCN1 and TFCN2). Note that in our definition of spatial co-localization, the effect of sequential clustering of BSs was not

taken into account, as 25kb adjacent regions left and right were removed to avoid high noise in Hi-C contacts near diagonal element, similar to[109]; instead, we only defined co-localization of TF BSs made possible by distal chromosome contacts.

Other hierarchical clustering metrics were also adopted (Wards and average linkage hierarchical clustering using squared euclidean distance) and similar results are shown in Figure 3.4 and Figure 3.5.

Scope of analysis	Genome-wide	Genome-wide	Within A1	Within A1	Within A2	Within A2
TF networks	TFCN1	TFCN2	TFCN1	TFCN2	TFCN1	TFCN2
TF names	ELK1_f1	JUND_f1	ELK1_f1	JUND_f1	ELK1_f1	NFKB1_f1
	MAX_f1	NFKB1_f1	MAX_f1	NFKB1_f1	NFYA_f1	STAT3_si
	NRF1_f1	STAT3_si	NRF1_f1	STAT3_si	NRF1_f1	NFIC_f2
	SP1	NFIC_f2	SP1	NFIC_f2	SP1	ATF2_f1
	E2F4_do	ATF2_f1	E2F4_do	ATF2_f1	E2F4_do	CEBPB_f1
	ATF3_f1	BATF_si	EGR1_f2	BATF_si	ELF1_f1	MEF2A_f1
	EGR1_f2	CEBPB_f1	ETS1_si	CEBPB_f1	ETS1_si	MEF2C_f1
	ELF1_f1	MEF2A_f1	GABPA_f1	MEF2A_f1	GABPA_f1	IKZF1
	ETS1_si	MEF2C_f1	ZNF143	MEF2C_f1	ZNF143	NFATC1
	GABPA_f1	IKZF1	MAZ_f1	IKZF1		STAT5
	ZNF143	NFATC1		NFATC1		
	MAZ_f1	POU2		POU2		
		STAT5		STAT5		
		TBP_f1				

Figure 3.3: The list of TFs within TFCN1 and TFCN2 respectively defined using DynamicTreeCut[197] from TF co-localization analysis genome-wide or within either A1/A2 subcompartment.



( $p = 4.0 \cdot 10^{-3}$ ), specifically, JAK-STAT cascade ( $p = 4.8 \cdot 10^{-2}$ ) and cellular defense response ( $p = 4.6 \cdot 10^{-2}$ ), whereas TFs of TFCN1 showed no enriched cell-type specific pathways other than general transcription activation.

Since TFCN2 seems to be more cell-type specific compared to TFCN1 according to GO term analysis, we wish to see if BS associated histone marks are also more conserved in TFCN1 compared to TFCN2. We chose four types of histone marks, namely, H3k4me1, H3k4me3, H3k27ac and H3k27me3, as well as DNase-I hypersensitivity (DHS) profiles, and compared those BS associated epigenetic profiles between GM12878 and h1-ESC. The proportion of BSs for each TF with consistent histone marks and DHS in the two cell lines is depicted in Figure3.6, colour coded by TF co-localization networks or ungrouped. Much higher levels of consistency for these two marks in TFCN1 was observed compared to those in TFCN2 (Wilcoxon rank sum test,  $p = 9 \cdot 10^{-8}$ ), while the ungrouped ones lie in between. Further, from the available ChIP-seq profiles for common TFs between the above two cell lines, TF BSs in TFCN2 are less shared between cell lines than those in TFCN1 (See Figure2.10). Taken together the GO term analysis and the above results, TFs and their target genome regions in TFCN2, are more cell-type specific, while TF binding regions in TFCN1 are more conserved across cell-types.

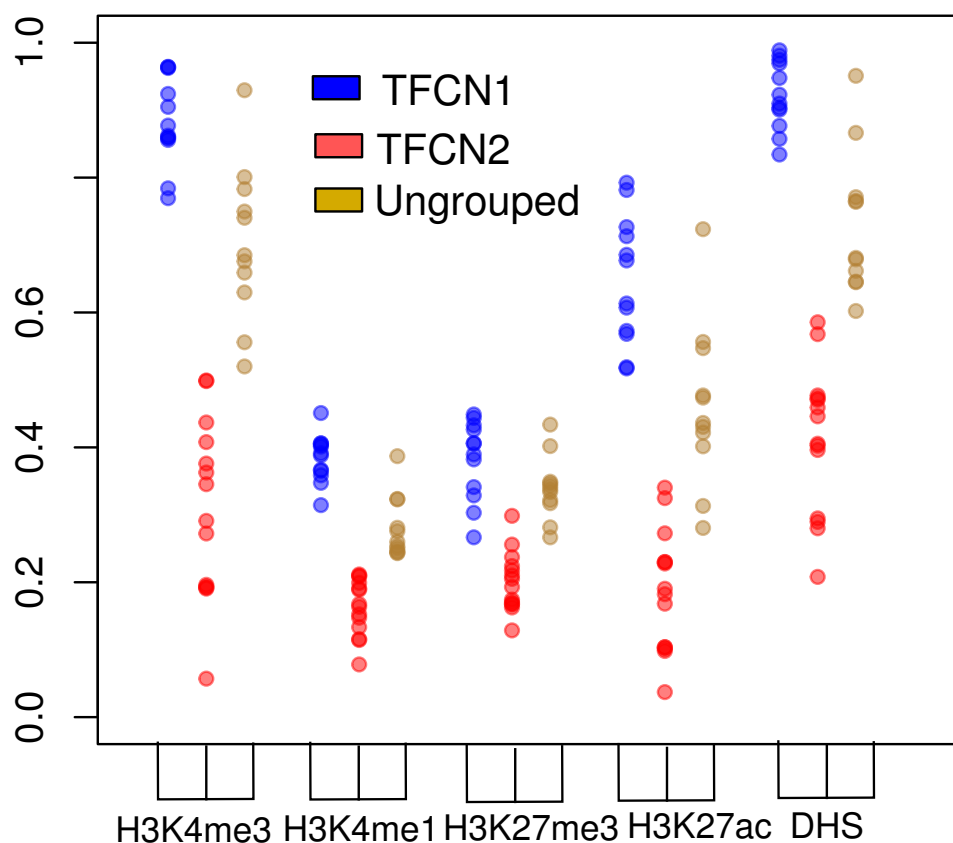


Figure 3.6: *The consistency of histone marks and DNase-I hypersensitivity between GM12878 and h1-ESC.* Each point represents the proportion of BSs for each TF that shows consistent H3k4me1, H3k4me3, H3k27ac, H3k27me3 or DHS between GM12878 and h1-ESC. TFs are colour coded according to the previously defined TF co-localization network TFCN1, TFCN2 or ungrouped.



### 3.3.2 TF spatial co-localization is closely related to TF-TF physical interaction

Based on the co-localization score distribution for each pair of TFs, we identified specific pairs of TFs with particularly significant co-localization. Such TF pairs would be good candidates for potential direct TF-TF interaction, or at least, the implication for co-functioning in transcription regulation. As is shown in Figure3.7, if two TFs are co-localized, we would expect a higher proportion of BSs to fall into the high co-localization score groups than expected. By comparing the observed versus the expected heterotypic co-localization score distributions for each TF pair, we obtained 40 pairs of significantly co-localized TFs out of the total 780 pairs of possible combinations when examining co-localization scores genome-wide (Table3.8). Similarly, when examining co-localization scores within either A1 or A2 chromosome subcompartment, 53 or 32 pairs of TFs were called respectively (See Table 3.10, 3.11 and 3.12). The called pairs genome-wide or within each subcompartment show good consistency with each other, in which 23 pairs are shared among all 3 analyses (Figure3.13). More than 94% of the co-localized TF pairs we called appear to be within the two TF co-localization networks. Interestingly, these TF co-localization pairs we identified have a significant overlap with previously reported TF-TF physical interaction pairs. There are at least 10 pairs of known physically interacting TFs [204, 205, 206, 207] that also appear in our list (Table3.14). Given more than half of TFs lack available data for direct TF-TF interactions, the overlaps between the co-localization pairs we called and the known interaction pairs are much higher than expected by chance.

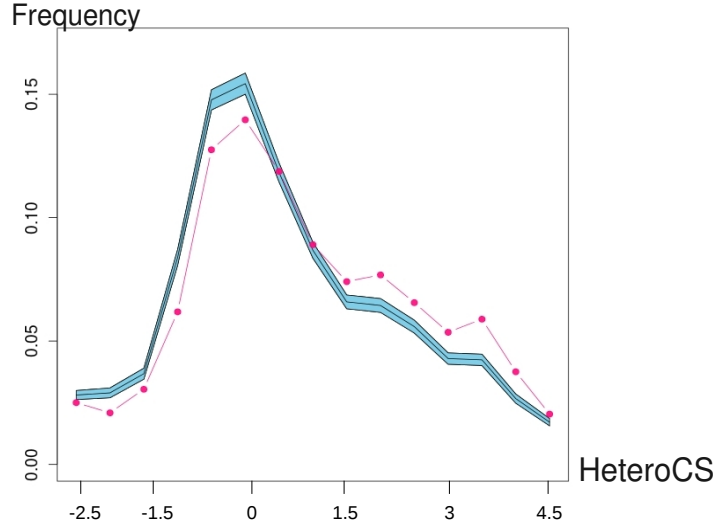


Figure 3.7: *TF pair co-localization score distribution compared to the random, expected distribution for a significantly co-localized TF pair, GABPA and SP1, as an example.* Shaded region of blue depicts the 95% confidential interval of the expected distribution for the co-localization scores between these two TFs, whereas the red dotted line shows the observed score distribution in each co-localization score bin. X-axis corresponds to the heterotypic co-localization score between these two TFs for each score bin.

The phenomenon of homotypic binding site clustering also stands out in TF co-localization pair calling. Even though our pair calling threshold is stringent, there are 15 out of 40 TFs whose BSs are significantly co-localized with their own type of BSs (37.5%), which is a more than 6-fold enrichment over the average proportion of the called pairs between TFs (around 5%). Even when comparing with the probability of calling significant pairs within each TF co-localization network, there is still an enrichment of over threefold. This re-enforces the fact that the co-localization of homotypic BSs is a wide-spread phenomenon that contributes significantly to TF spatial binding patterns.

TF1	TF2	z score (top 20%)	z score (10%)	z score (5%)
ELK1_f1	EGR1_f2	6.41	5.15	4.01
ELK1_f1	ETS1_si	8.37	7.83	5.67
ELK1_f1	GABPA_f1	9.65	9.5	7.48
MAX_f1	SP1_f1	8.11	6.8	4.31
MAX_f1	MAZ_f1	8.83	7.39	5.84
NFKB1_f1	STAT3_si	7.74	7.88	5.56
NFKB1_f1	NFIC_f2	8.19	6.89	5.54
NFKB1_f1	CEBPB_f1	8.83	7.29	4.08
STAT3_si	NFKB1_f1	6.88	6.09	4.53
STAT3_si	NFIC_f2	9.61	7.68	5.86
STAT3_si	ATF2_f1	9.85	7.28	6.81
STAT3_si	CEBPB_f1	10.4	8.49	6.94
STAT3_si	MEF2A_f1	8.29	6.08	4.56
STAT3_si	IKZF1	6.21	6.85	5.73
STAT3_si	NFATC1	6.67	6.38	5.05
NFIC_f2	NFKB1_f1	6.48	5.75	4.22
NFIC_f2	STAT3_si	8.54	7.63	6.07
NFIC_f2	ATF2_f1	7.62	6.87	4.64
NFIC_f2	CEBPB_f1	11.4	9.77	6.59
NFIC_f2	MEF2A_f1	4.01	5.11	4.98
NFIC_f2	NFATC1	6.46	6.28	5.32
E2F4_do	SP1_f1	10.7	8.34	6.87
E2F4_do	ETS1_si	8.1	7.69	5.15
E2F4_do	GABPA_f1	8.4	6.69	5.79
ATF2_f1	STAT3_si	10.8	9.99	7.84
ATF2_f1	NFIC_f2	10.6	8.74	6.74
ATF2_f1	CEBPB_f1	13	10.3	7.6
ATF2_f1	MEF2A_f1	6.94	6.45	5.16
ATF2_f1	MEF2C_f1	7.2	5.99	3.82
ATF2_f1	IKZF1	6.66	6.3	4.82
CEBPB_f1	NFKB1_f1	6.76	6.25	5.14
CEBPB_f1	STAT3_si	11.3	8.78	9.64
CEBPB_f1	NFIC_f2	10.9	9.83	6.88
CEBPB_f1	ATF2_f1	10	8.02	7.17
CEBPB_f1	MEF2A_f1	10.7	8.94	6.81
CEBPB_f1	MEF2C_f1	8.64	7.1	5.44
CEBPB_f1	IKZF1	5.82	5.09	4.54
CEBPB_f1	NFATC1	6.95	5.21	6.35
EGR1_f2	ELK1_f1	6.98	4.92	4.51
EGR1_f2	SP1_f1	10.9	8.9	6.38
ELF1_f1	ETS1_si	7.94	5.51	4.47
ELF1_f1	GABPA_f1	6.03	4.36	3.97
ELF1_f1	SP1_f1	9.98	8.56	6.51
ETS1_si	ELK1_f1	6.88	5.96	6.39
ETS1_si	SP1_f1	8.49	6.76	6.15
ETS1_si	E2F4_do	7.1	5.82	5.79
ETS1_si	ELF1_f1	6.15	4.4	4.54
ETS1_si	GABPA_f1	8.39	7.7	7.27
GABPA_f1	ELK1_f1	10.9	9.19	7.55
GABPA_f1	SP1_f1	12.2	10.5	7.09
GABPA_f1	E2F4_do	7	6.34	4.69
GABPA_f1	ELF1_f1	8.67	7.01	5.04

Figure 3.8: TF spatial co-localization pairs identified genome wide that are significantly enriched in high heterotypic clustering score groups. The columns of z-scores show the deviation of the observed frequency of BSs falling into top score groups (20%,10% and 5%, respectively) from the expected distribution.

TF1	TF2	z score (top 20%)	z score (10%)	z score (5%)
GABPA_f1	ETS1_si	9.45	8.82	8.05
MEF2A_f1	STAT3_si	11	11.3	7.6
MEF2A_f1	NFIC_f2	11	9.14	7.81
MEF2A_f1	ATF2_f1	9.71	9.43	6.88
MEF2A_f1	CEBPB_f1	12.4	11.8	10.4
MEF2A_f1	MEF2C_f1	10.5	9.68	9.31
MEF2A_f1	NFATC1	7.29	6.86	5.89
MEF2C_f1	ATF2_f1	9.92	8.6	6.32
MEF2C_f1	CEBPB_f1	11	9.69	7.86
MEF2C_f1	MEF2A_f1	11.9	10.6	8.58
MEF2C_f1	NFATC1	7.98	6.25	5.63
IKZF1	STAT3_si	7.52	6.46	5.19
IKZF1	ATF2_f1	6.45	5.35	4.95
IKZF1	CEBPB_f1	7.79	8.23	7.34
NFATC1	STAT3_si	10.5	9.89	5.73
NFATC1	NFIC_f2	8.76	9.12	6.38
NFATC1	CEBPB_f1	10.2	10.4	8.79
NFATC1	MEF2A_f1	7.13	6.74	6.12
NFATC1	MEF2C_f1	7.23	6.89	5.83
SP1_f1	MAX_f1	6.54	5.15	4.8
SP1_f1	E2F4_do	6.58	5.47	5.02
SP1_f1	EGR1_f2	9.31	5.52	4.05
SP1_f1	ELF1_f1	8.38	5.74	3.84
SP1_f1	ETS1_si	7.53	6.61	4.44
SP1_f1	GABPA_f1	8.42	5.99	5.64
SP1_f1	MAZ_f1	12.5	10.5	8.07
MAZ_f1	MAX_f1	7.27	5.83	6.26
MAZ_f1	SP1_f1	11.4	9.09	6.84

Figure 3.9: TF co-localization pairs identified genome wide: continued.

TF1 name	TF2 name	z score (top 20%)	z score (top 10%)
ELK1_f1	EGR1_f2	3.83	3.38
ELK1_f1	ETS1_si	5.2	4.07
ELK1_f1	GABPA_f1	7.98	6.65
ELK1_f1	SP1_f1	5.76	5.25
ELK1_f1	MAZ_f1	3.09	4.46
NFKB1_f1	STAT3_si	6.35	5.81
NFKB1_f1	NFIC_f2	7.29	5.45
NFKB1_f1	CEBPB_f1	6.95	3.24
NRF1_f1	GABPA_f1	8.53	5.03
STAT3_si	NFKB1_f1	5.15	5.6
STAT3_si	NFIC_f2	7.46	6.7
STAT3_si	ATF2_f1	7.23	5.23
STAT3_si	CEBPB_f1	5.9	4.53
STAT3_si	MEF2A_f1	9.41	6.28
STAT3_si	MEF2C_f1	6.1	3.49
STAT3_si	IKZF1	3.82	3.99
STAT3_si	NFATC1	5.7	4.46
NFIC_f2	NFKB1_f1	7.72	5.95
NFIC_f2	STAT3_si	7.2	5.63
NFIC_f2	ATF2_f1	8.38	5.07
NFIC_f2	CEBPB_f1	8.85	5.96
NFIC_f2	MEF2A_f1	5.44	6.17
NFIC_f2	MEF2C_f1	3.82	3.45
NFIC_f2	NFATC1	4.03	3.9
JUND_f1	NFATC1	7.51	4.71
JUND_f1	BATF_si	5.93	3.25
JUND_f1	MEF2A_f1	9.73	9.26
JUND_f1	MEF2C_f1	7.43	6.7
E2F4_do	EGR1_f2	4.32	3.35
E2F4_do	ELF1_f1	3.76	4.31
E2F4_do	ETS1_si	6.24	5.68
E2F4_do	GABPA_f1	7.14	6.54
E2F4_do	SP1_f1	7.09	6.62
ATF2_f1	STAT3_si	7.45	8.58
ATF2_f1	NFIC_f2	6.9	5.83
ATF2_f1	CEBPB_f1	7.78	5.82
ATF2_f1	MEF2A_f1	4.97	5.77
ATF2_f1	IKZF1	3.06	2.98
ATF2_f1	NFATC1	7.58	5.34
BATF_si	JUND_f1	6.84	3.33
BATF_si	CEBPB_f1	8.46	5.67
CEBPB_f1	NFKB1_f1	9.08	7.07
CEBPB_f1	STAT3_si	8.02	7.05
CEBPB_f1	NFIC_f2	10.9	9.61
CEBPB_f1	ATF2_f1	9.36	8.56
CEBPB_f1	BATF_si	7.57	7.71
CEBPB_f1	MEF2A_f1	9.9	8.07
CEBPB_f1	MEF2C_f1	6.78	5.65
CEBPB_f1	IKZF1	3.67	2.76
CEBPB_f1	NFATC1	6.18	3.41
EGR1_f2	ELK1_f1	2.82	2.92
EGR1_f2	E2F4_do	3.42	2.8

Figure 3.10: TF spatial co-localization pairs identified within A1 subcompartment.

TF1 name	TF2 name	z score (top 20%)	z score (top 10%)
EGR1_f2	SP1_f1	4.56	2.81
ELF1_f1	E2F4_do	2.64	3
ETS1_si	ELK1_f1	4.67	4.14
ETS1_si	E2F4_do	5.92	5.02
ETS1_si	GABPA_f1	4.95	6.06
ETS1_si	SP1_f1	7.07	6.09
GABPA_f1	ELK1_f1	6.06	5.87
GABPA_f1	NRF1_f1	6.15	4.22
GABPA_f1	E2F4_do	4.47	4.54
GABPA_f1	ETS1_si	5.56	5.94
GABPA_f1	ZNF143	3.42	3.09
GABPA_f1	SP1_f1	7.43	6.1
GABPA_f1	MAZ_f1	6.21	5.41
MEF2A_f1	STAT3_si	7.62	9.36
MEF2A_f1	NFIC_f2	7.58	6.72
MEF2A_f1	JUND_f1	4.94	3.81
MEF2A_f1	ATF2_f1	7.3	5.99
MEF2A_f1	CEBPB_f1	7.22	5.27
MEF2A_f1	MEF2C_f1	6.65	4.73
MEF2A_f1	IKZF1	5.02	5.4
MEF2A_f1	NFATC1	3.99	4.37
MEF2C_f1	STAT3_si	7.16	5.73
MEF2C_f1	NFIC_f2	7.25	7.34
MEF2C_f1	JUND_f1	4.36	3.65
MEF2C_f1	CEBPB_f1	7.16	5.57
MEF2C_f1	MEF2A_f1	9.05	6.99
MEF2C_f1	IKZF1	3.14	4.64
MEF2C_f1	NFATC1	4.98	4.31
IKZF1	STAT3_si	6.89	6.25
IKZF1	ATF2_f1	6.24	6.32
IKZF1	CEBPB_f1	6.38	8.32
IKZF1	MEF2A_f1	7.72	6.95
IKZF1	MEF2C_f1	7.28	7.35
IKZF1	NFATC1	5.77	7.94
NFATC1	STAT3_si	7.9	9.54
NFATC1	NFIC_f2	8.99	9.41
NFATC1	JUND_f1	3.96	4.02
NFATC1	ATF2_f1	9.68	8
NFATC1	CEBPB_f1	7.01	5.51
NFATC1	MEF2A_f1	6.39	7.24
NFATC1	MEF2C_f1	7.03	6.51
NFATC1	IKZF1	4.01	4.7
NFATC1	POU2	4.31	4.09
POU2	NFATC1	6.8	7.42
ZNF143	GABPA_f1	5.67	4.48
SP1_f1	ELK1_f1	3	2.89
SP1_f1	E2F4_do	4.29	3.45
SP1_f1	EGR1_f2	4.27	2.71
SP1_f1	ETS1_si	5.01	3.36
SP1_f1	GABPA_f1	3.63	3.74
SP1_f1	MAZ_f1	8.89	6
MAZ_f1	ELK1_f1	2.89	2.73

Figure 3.11: TF co-localization pairs identified within A1: continued.

TF1 name	TF2 name	Z-score (top 20%)	z-score (top 10%)
ELK1_f1	GABPA_f1	4.85	4.16
NFKB1_f1	STAT3_si	4.39	4.61
NFKB1_f1	NFIC_f2	5.62	3.45
NFKB1_f1	CEBPB_f1	7.19	5.14
NFYA_f1	E2F4_do	5.03	5.45
NRF1_f1	GABPA_f1	6.3	6.48
STAT3_si	NFKB1_f1	5.34	4.69
STAT3_si	NFIC_f2	6.21	4.2
STAT3_si	JUND_f1	4.73	3.56
STAT3_si	ATF2_f1	4.57	3.81
STAT3_si	CEBPB_f1	7.55	5.44
STAT3_si	IKZF1	4.8	5.21
STAT3_si	NFATC1	3.48	3.55
NFIC_f2	NFKB1_f1	3.35	3.24
NFIC_f2	STAT3_si	7.18	5.62
NFIC_f2	ATF2_f1	4.09	3.17
NFIC_f2	CEBPB_f1	8.94	6.11
NFIC_f2	NFATC1	3.47	3.87
JUND_f1	STAT3_si	3.46	3.01
JUND_f1	ATF2_f1	2.99	3.17
JUND_f1	BATF_si	4.17	2.84
JUND_f1	MEF2A_f1	5.11	4.51
JUND_f1	MEF2C_f1	8.49	4.28
E2F4_do	NFYA_f1	4.55	5.15
ATF2_f1	STAT3_si	7.25	5.04
ATF2_f1	NFIC_f2	4.31	4.1
ATF2_f1	JUND_f1	5.4	3.2
ATF2_f1	CEBPB_f1	7.16	5.91
ATF2_f1	MEF2A_f1	4.72	3.46
ATF2_f1	MEF2C_f1	8.36	4.71
ATF2_f1	IKZF1	5.66	5.3
ATF2_f1	NFATC1	3.35	3
BATF_si	JUND_f1	3.59	2.81
CEBPB_f1	NFKB1_f1	4.01	2.98
CEBPB_f1	STAT3_si	5.98	4.79
CEBPB_f1	NFIC_f2	7.55	7.46
CEBPB_f1	ATF2_f1	7.55	4.09
CEBPB_f1	MEF2A_f1	5.52	4.19
CEBPB_f1	MEF2C_f1	5.47	2.89
CEBPB_f1	IKZF1	4.8	3.88
CEBPB_f1	NFATC1	3.47	3.02
ETS1_si	GABPA_f1	4.7	4.03
GABPA_f1	ELK1_f1	6.23	5.53
GABPA_f1	NRF1_f1	8.44	6.69
GABPA_f1	ETS1_si	5.43	4.22
GABPA_f1	ZNF143	4.1	4.28
MEF2A_f1	JUND_f1	5.66	3.96
MEF2A_f1	ATF2_f1	5.41	2.93
MEF2A_f1	CEBPB_f1	6.58	8.16
MEF2A_f1	MEF2C_f1	8.48	8.08
MEF2C_f1	JUND_f1	3.14	2.97
MEF2C_f1	ATF2_f1	5.63	3.7

Figure 3.12: TF spatial co-localization pairs identified within A2 subcompartment.

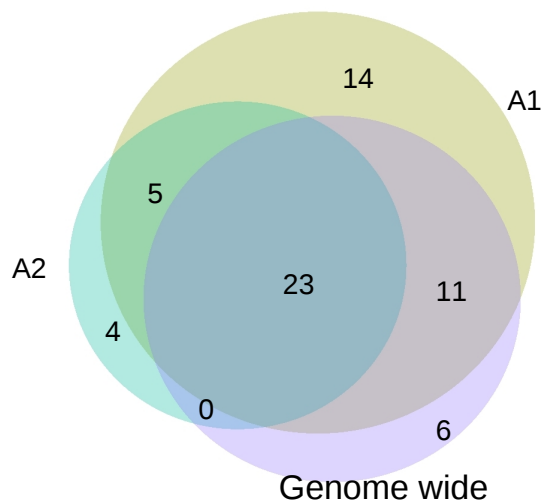


Figure 3.13: The overlap between called significant TF co-localization pairs obtained from genome-wide analysis and within A1 and A2 subcompartments. There are 23 pairs of TFs appearing in all three sets of results.

Molecule A	Molecule B	Interaction detection method	Interaction AC in EBI-IntAct database
MEF2A	MEF2C	tandem affinity purification	EBI-11322070
ATF2	NFATC1	tandem affinity purification	EBI-10711208
CEBPB	BATF	peptide array	EBI-10890773
CEBPB	ATF2	peptide array	EBI-10890828
NFkB1	STAT3	two hybrid	EBI-3940875
GABPA	SP1	pull down	EBI-7786331
JUND	ATF2	peptide array	EBI-10891948
JUND	BATF	peptide array	EBI-10891973
JUND	NFATC1	tandem affinity purification	EBI-11319967
JUND	MEF2A	tandem affinity purification	EBI-11319891

Figure 3.14: Previously identified TF physical interaction pairs that also appear to be significantly spatial co-localized according to Hi-C contact map.



### 3.3.3 TFCN1 and TFCN2 have markedly different binding preference in A1 and A2 chromosome subcompartments

To further investigate the two distinct TF co-localisation networks, we explored their relationship with chromosome compartmentalization as inferred by Hidden Markov Model in Rao *et al*, 2014 [1]. Due to naturally low (and hence insufficient here) ChIP-seq data for the B compartment, we focused our attention on the A compartment, which is enriched in actively transcribed genes. As was revealed by Rao *et al*, 2014, the A compartment identified using lower resolution maps can be further partitioned into two different subcompartments, A1 and A2, based on the inter-chromosome contact map.

First, we compared BS occupancy (as was defined by the proportion of putative BS occupied by ChIP-seq peaks in DHS, as in Section 2.3.3) for each TF respectively between A1 and A2 subcompartments. We found TFs within TFCN1 tend to show significantly higher occupancy in A1, while TFs from TFCN2 have higher occupancy in A2 (See Figure 3.15). Particularly, in TFCN2, JunD had more than a twofold increase in occupancy within A2 compared to A1 and IKZF1 also has over one-fold increase. Other TFs that show more than a 70% occupancy increase in A2 compared to A1 including STAT3, BATF and NFIC, key factors in immune responses and NfκB signalling pathway, are all within TFCN2. In contrast, TFs with higher occupancy in A1 are mostly non-cell type specific TFs within TFCN1. In addition, we note that four out of five TFs belonging to ETS family grouped into TFCN1 (GABPA, ELK1, ETS1 and ELF1) are associated with higher occupancy in A1 compared to A2 (except EGR1).

To further confirm the observed binding occupancy differences in different subcompartments, we performed ChIP-seq SignalValue comparisons between paired BSs similar to those that we did for homotypic BSs analysis, but here the BSs only differ in their A1/A2 subcompartment identity, while chromatin state, different types of histone marks and the homotypic clustering score level (low, mid or high, as previously defined) are all required to be the same. TFs showing significant differences in ChIP-

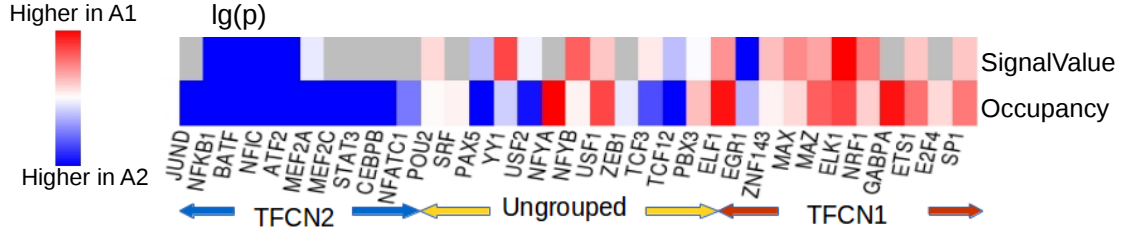


Figure 3.15: *BS occupancy for individual TFs in relation to chromosome subcompartment.* Colours indicate the  $-\log_{10}(p)$  (G-test with William correction) of BS occupancy differences (bottom row) and ChIP-seq SignalValue differences (top row) between A1 and A2 chromosome subcompartment for individual TFs.

seq SignalValue within each TF co-localization network is depicted in Figure 3.15. TFs with significantly higher SignalValues in either subcompartment are consistent with previous binding occupancy analysis (Figure 3.15).

### 3.3.4 TFCN1 and TFCN2 exist independent of TF occupancy differences within A1/A2 subcompartments

Since there were clear differences in TF binding between the A1 and A2 subcompartments, we explored further whether this was sufficient to account for the presence of two distinct TF spatial networks. The co-localisation of these groups of TFs may not be a consequence of direct TF-to-TF association, but instead it could be an indirect effect; the TFs in each network may tend to co-occur due to their subcompartment preference. To determine whether the TFs in each group continued to co-localise within each subcompartment, we further analysed the co-localization score distribution for each TF within each subcompartment independently. We used control sets for co-localization scores generated by randomly shuffling ChIP-seq identified BSs within each subcompartment for each chromosome separately, instead of shuffling all BSs. Surprisingly, the two co-localization networks reoccurred almost in the same manner in both of the analysis done independently in the A1 and A2 subcompartment (Figure 3.16 and Figure 3.17). The same cluster-defining metric was performed as before and

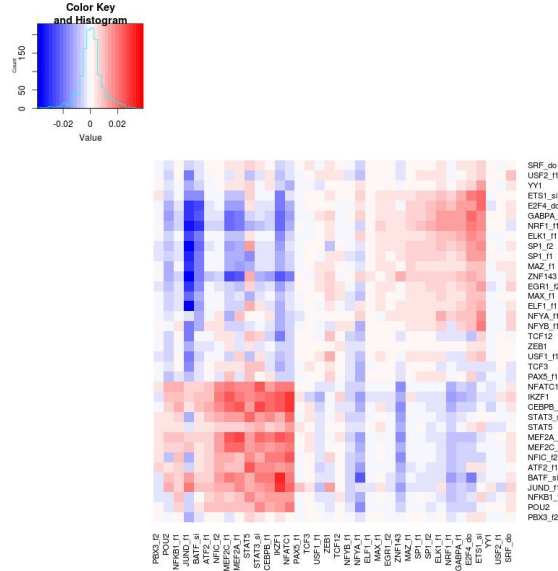


Figure 3.16: TF co-localization enrichment map within A1 subcompartment given distance measure using Equation3.3 and average linkage hierarchycal clustering.

not a single TF swapped clusters in either the A1 or A2 subcompartment analysis, as compared to the original analysis in Figure 3.1 and Table 3.3. Though, a few additional TFs in the A2 subcompartment did not fall into either cluster, probably due to relatively small sample size of BSs in A2. This suggests that TFCN1 and 2 is not a simple reflection of TF binding occupancy differences between the A1/A2 subcompartments; instead, its presence is robust both in a genome-wide scale and within each subcompartment.

Other hierarchical clustering metrics give similar results within A1, but some of them perform poorly within A2 (for instance, the average linkage hierarchical clustering with euclidean distance, see Figure 3.21) compared to other methods (Wards, Figure3.20 or using the distance measure from Equation3.3, Figure3.17).

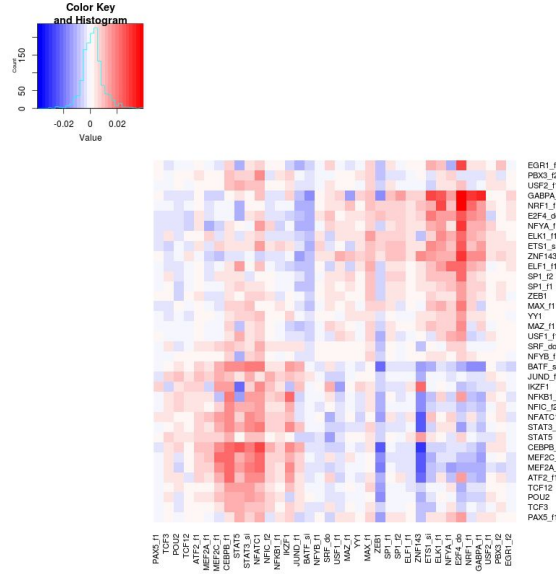


Figure 3.17: TF co-localization enrichment map within A2 subcompartment given distance measure using Equation3.3.

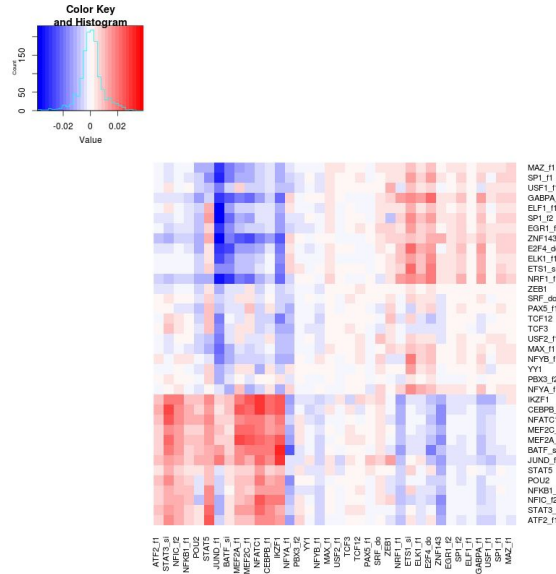


Figure 3.18: TF co-localization enrichment map within A1 subcompartment (Wards approach).





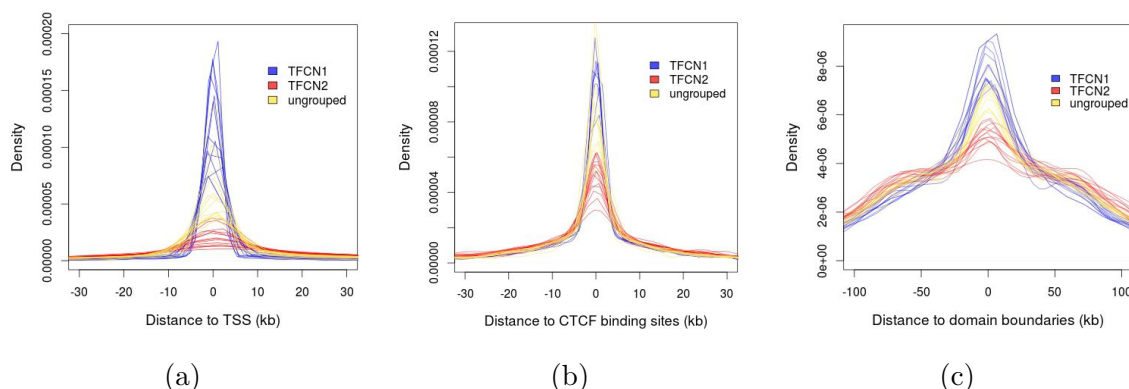


Figure 3.22: The distribution of TF BSs in respect to transcription start site (TSS), CTCF binding sites and chromosome domain boundary (shown are genomic distances in kb).

### 3.3.6 Revisit the relationship between homotypic BS co-localization and BS occupancy with regards to TFCN1 and TFCN2 respectively

Homotypic BS co-localization has been shown to increase TF binding site occupancy in either the enhancer or promoter state, for BSs with both strong and weak DNA binding motifs, though the effect is more prominent in enhancers and weak promoters, for weak BSs rather than strong ones. In light of the TF co-localization networks identified in GM12878 cells, we want to revisit the question related to the role of BS spatial co-localization in determining TF binding occupancy, both homotypic and between different types of TFs.

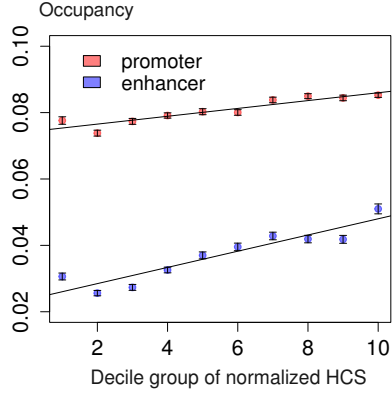
We first re-examined the homotypic case, by grouping TFs according to TF co-localization networks, to see if the previously observed trends hold for different TFCNs. We found strong correlations between BS occupancy and homotypic BS co-localization similarly within both TF co-localization networks, confirming the fact that the impact of homotypic BS co-localization on BS occupancy preserves regardless of TF co-localization network (Figure 3.23(a,b)). Interestingly, we noticed that in

TFCN1, BS occupancy is much higher in the promoter state in comparison to the enhancer state, while in TFCN2, the opposite trend is observed. It is consistent with the finding that BSs of TFCN2 are less centred around TSS, which is mainly associated with the promoter state, compared to those in TFCN1. In addition, the trend that weak promoter without H3k36Me3 get better correlation compared to strong, highly active promoter also holds within each TF co-localization network (Figure 3.23(c,d)).

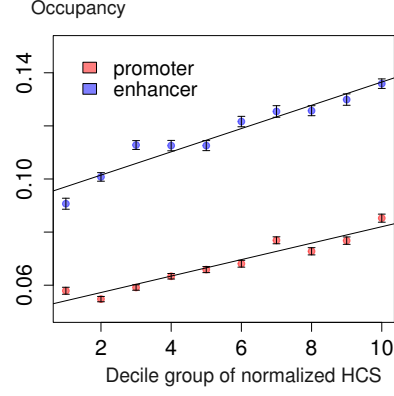
### **3.3.7 Intra-TFCN heterotypic BS co-localization significantly enhances TF binding, while the role of inter-TFCN BS co-localization varies between enhancer and promoter regions**

BSs of TFs within the same TF co-localization network tend to have higher than expected levels of spatial co-localization, similar to the spatial clustering of BSs within the same type of TF. Thus, we hope to see if the co-localization of BSs within the same TFCN may also influence BS occupancy as well. Assuming the effects of TFs within a specific TFCN are simply additive, we defined an integrated measure of spatial heterotypic co-localization for a group of TFs collectively, around each genome loci  $i$  containing putative BSs according to Equation 3.4. Here only the interactions between different TFs within the same co-localization network have been taken into account, while the homotypic co-localization was not involved, in order to remove the known effects of homotypic BS clustering. As shown in Figure 3.24 (a) and (b), for both TF co-localization networks, BSs associated with the enhancer state showed a strong positive correlation with intra-TFCN heterotypic co-localization (adjusted Pearson's  $R^2$  of 0.71 and 0.77 for TFCN1 and 2 respectively). However, for BSs falling into regions of the promoter state, the effect of intra-TFCN heterotypic co-localization seems to be more pronounced in TFCN2 compared to TFCN1 in terms of both correlation coefficient and the magnitude of BS occupancy increase (adjusted  $R^2$  of 0.75 and 59% of increase for TFCN2; while adjusted  $R^2$  of 0.30 and less than 5% of increase for TFCN1).

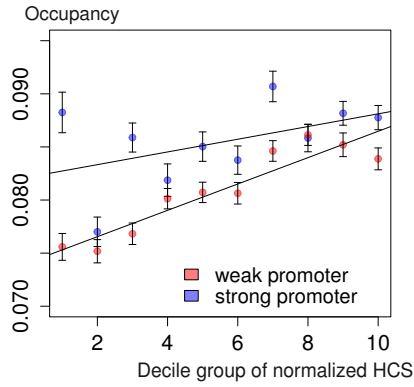




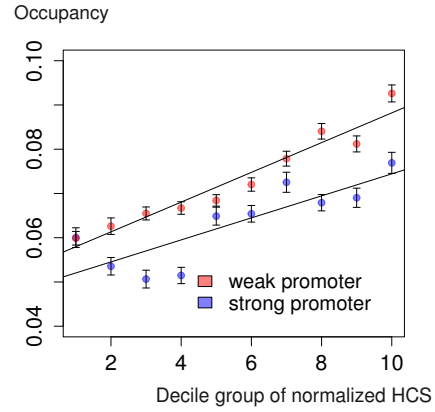
(a)



(b)

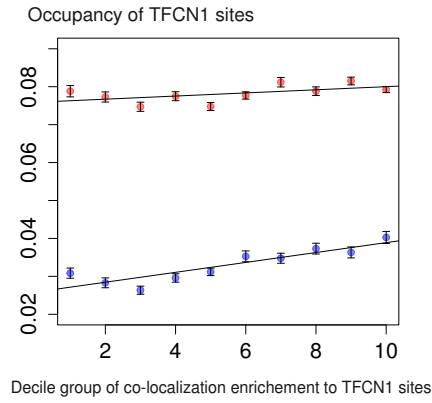


(c)

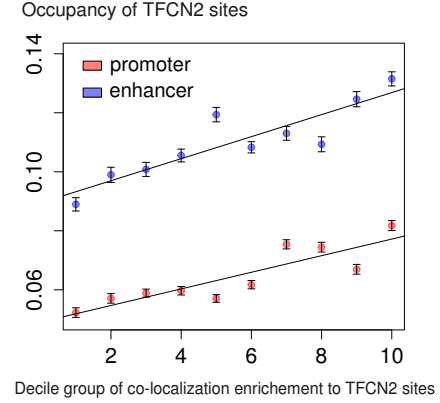


(d)

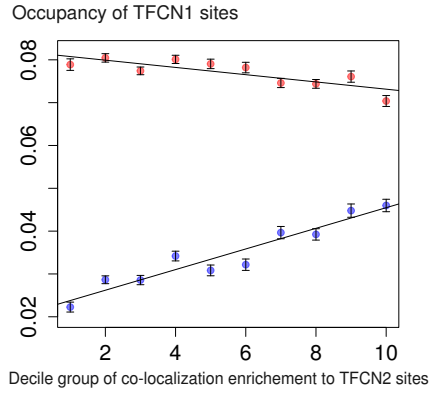
Figure 3.23: The relationship between BS occupancy and homotypic BS co-localization within either TFCN1 (a) or TFCN2 (b). Adjusted Pearson's  $R^2$  of 0.87 and 0.84 for the enhancer and the promoter state within TFCN1; adjusted  $R^2$  of 0.93 and 0.91 for the enhancer and the promoter state within TFCN2. For the comparison between strong and weak promoter within TFCN1 (c) or TFCN2 (d), weak promoter consistently shows better correlation between between BS occupancy and homotypic BS co-localization (adjusted Pearson's  $R^2$  of 0.85 and 0.93 for TFCN1 and 2 respectively), compared to the strong ones (adjusted  $R^2$  of 0.15 and 0.63 for TFCN1 and 2).



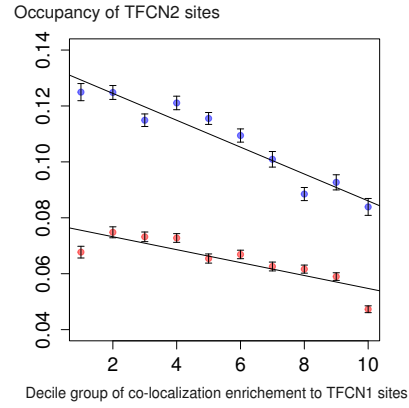
(a)



(b)



(c)



(d)

Figure 3.24: TF BS occupancy in relation to integrated heterotypic co-localization within the same TFCN (a:TFCN1 and b:TFCN2) or between the two TFCNs(c and d). Figure c shows the occupancy of TFCN1 in relation to the co-localization level with regards to TFCN2 sites, while d describes the relationship the other way around (occupancy of TFCN2 regarding co-localization with TFCN1 sites).

The effect of intra-TFCN heterotypic co-localization is in line with homotypic clustering, which enhances TF binding in either the enhancer or the promoter state. We wish to further explore if this is also true when considering BS co-localization levels between different TFCNs, namely, the degree to which BSs from TFCN1 is co-localized with putative BSs within TFCN2 and vice versa. Surprisingly, for TFCN2,

in both the promoter and the enhancer states, occupancy drops significantly when associated with high level of TFCN1 BS co-localization (adjusted  $R^2$  of 0.69 (38%) and 0.87 (33%) of decrease for the promoter state and the enhancer state, respectively, Figure 3.24(d)). In contrast, when examining BS occupancy of TFCN1 in relation to TFCN2 BS co-localization, BSs in the promoter and the enhancer state show the opposite trend. In the promoter state, decreased occupancy is observed ( $R^2$  of 0.63, Figure 3.24(c)), while in the enhancer state, correlation becomes positive instead (adjusted  $R^2$  of 0.66 ). We will discuss more about this unexpected diverged trend in the discussion section of this chapter.

The above analysis integrated all TF BSs within a certain co-localization network together to derive BS occupancy, however, it is unclear if similar trend can be observed for individual TFs. Therefore, we next focused our attention on each TF within either TFCN1 or TFCN2 and compared BS occupancy between BSs associated with high and low level (the top 1/3 and the bottom 1/3) of intra and inter-TFCN heterotypic co-localization. The results are shown in Figure 3.25, where the effect of homotypic BS co-localization is also included for comparison. 67% of TFs within TFCN1 (indicated by yellow bars) and 91% TFs within TFCN2 (indicated by blue bars) show a significant occupancy increase in response to high levels of intra-TFCN co-localization ( $p < 0.01$ , Chi-square test with Yates' correction). The magnitude of occupancy increase is comparable or sometimes even slightly higher than the occupancy gain due to homotypic co-localization for TFs within TFCN2, while within TFCN1, the magnitude of increase is mostly lower than the effect of homotypic co-localization. When further grouping BSs according to their associated chromosome states (Figure 3.26), in the enhancer state, similar conclusion as described above could be reached; whereas in the promoter state, no significant occupancy gain is observed for most of TFs within TFCN1, but nearly all TFs within TFCN2 display significant occupancy increase in response to high level of intra-TFCN BS heterotypic co-localization.

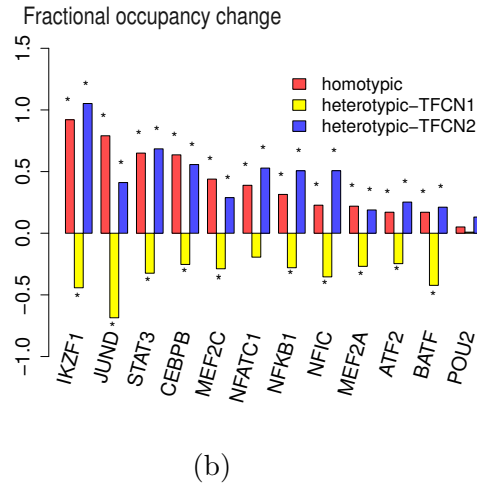
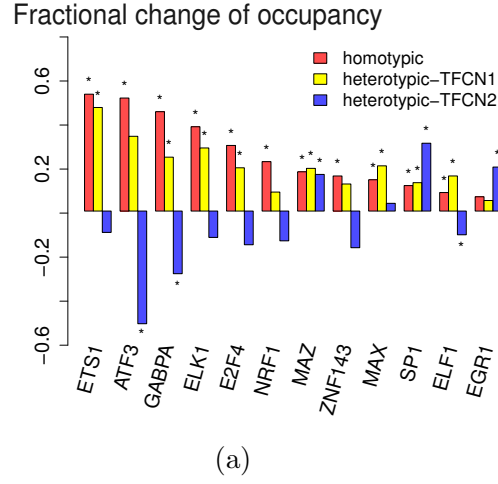


Figure 3.25: *The influence of homotypic and heterotypic BS spatial clustering on TF binding.* For each TF, the percent changes in BS occupancy associated with high level of homotypic or heterotypic BS co-localization when compared to the low are depicted; (a) for TFs within TFCN1 while (b) for TFCN2. The presence of the star above each bar indicates statistical significance (Chi-square test with Yates' correction,  $p < 0.01$ ).

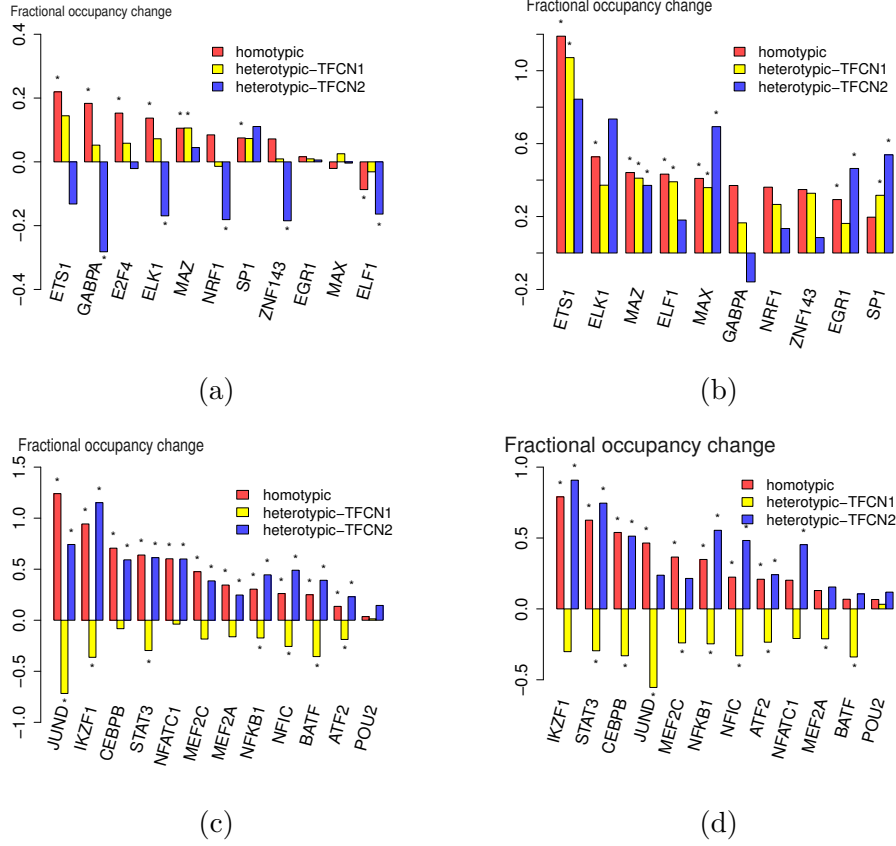


Figure 3.26: The influence of homotypic and heterotypic BS co-localization on TF binding similar to 3.25. Here, BSs are further grouped according to either the promoter (a: TFs within TFCN1 and c: TFs within TFCN2) or the enhancer state (b: TFs within TFCN1 and d: TFs within TFCN2).

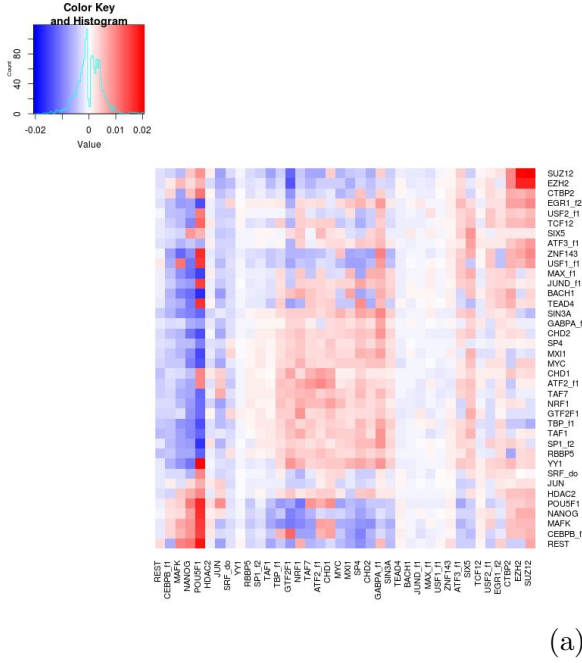
When investigating the effect of inter-TFCN heterotypic co-localization, unexpectedly, for either the enhancer or the promoter state, most of TFs within TFCN2 witness significant occupancy decrease associated with high level of spatial co-localization to TFCN1 (Figure 3.26(c and d)). It confirms the finding in Figure 3.24, and further demonstrating that this is a shared property among most of TFs within a specific co-localization network. Interestingly, for TFCN1, TFs mostly have occupancy increase in response to co-localization with TFCN2 in the enhancer state, while show decreased occupancy in the promoter state instead (Figure 3.26(a and b)), consistent with what was shown in Figure 3.24. We note that within each co-localization

network, the magnitude of occupancy change may vary between different TFs. For instance, MAX, EGR and SP1 in TFCN1 show significant high level (more than 40%) of occupancy increase in the enhancer state in response to co-localization with TFCN2 BSs, and they are exactly the ones with negligible level of occupancy decrease or even slight increase (SP1) in the promoter state. In contrast, GABPA is the one with the highest level of occupancy decrease in the promoter state and it still shows slight decrease of occupancy even in the enhancer state, which is an exception.

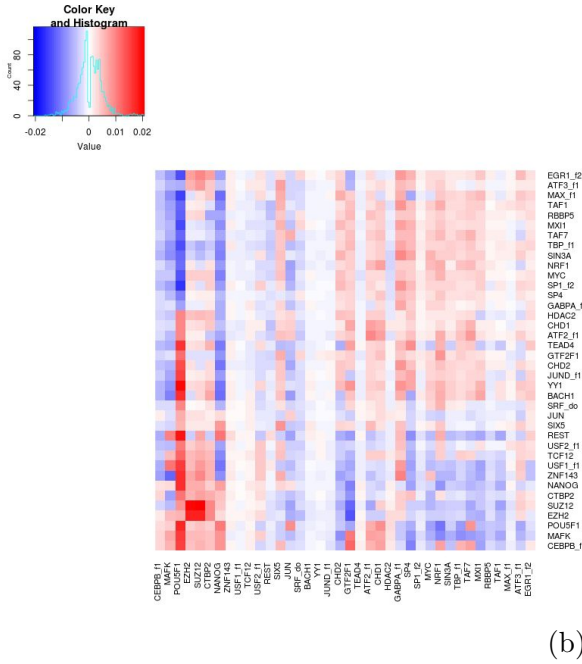
### 3.3.8 TF binding sites co-localization in human embryonic stem cells

Human lymphoblastoid cell line GM12878 has been used as a model cell line in all the above analysis, with regard to TF co-localization network and occupancy. To investigate if TFs co-localize similarly in different cell types, and to show the applicability of our computational approach to Hi-C maps gained from different sources, we studied another cell line – human h1 embryo stem cell (h1-ESC), which has sufficient number of ENCODE ChIP-seq profiles for different TFs [67], though the contact map was with a much lower resolution of 40kb. Similar to TF co-localization analysis performed in GM12878 described above, we investigated TF co-localization pattern in h1-ESC for 38 TFs with available and sufficient ChIP-seq profiles using the Hi-C contact map reported by Dixon *et al*, 2012 [120]. Given there are only 3 TFs in TFCN2 with presence in h1-ESC with available ChIP-seq data, it is not possible to draw an analogue of TF co-localization networks in ESC to the previously defined ones in GM12878. We also note that due to the lack of deeply sequenced Hi-C library and less optimized Hi-C protocol compared to the one presented in Rao *et al*, 2014 [1], the TF co-localization analysis results described below might suffer from certain drawbacks due to the quality of the contact map (see discussion section for details).

In h1-ESC, a small group of TFs seem to have distinct co-localization patterns compared to the rest (Figure 3.27), which comprises several key pluripotency factors. For instance, we observed the spatial co-localization between pluripotency factors



(a)



(b)

Figure 3.27: *TF co-localization enrichment map in human ES cells*. Colours show the enrichment (red) or depletion (blue) of spatial co-localization between pairs of TFs. Average linkage hierarchical clustering based on the distance measure of Equation 3.3 is shown in (a), while the results of Ward clustering based on euclidean distance is depicted in (b). Note that the value of four pixels shown here corresponding to the co-localization between E2H2 and Suz12 (E2H2-SUZ12, SUZ12-E2H2, E2H2-E2H2, SUZ12-SUZ12) are extremely high compared to others. Thus, to make the color map have a balanced color scale (just for the sake of visualization), those 4 pairs have been divided by 6, though they are still the maximum score even after minimizing by 6 folds.

POU5F1 and NANOG together with MAFK based on either of the clustering metrics shown in Figure 3.27. It is consistent with the report that NANOG and POU5F1 interact with each other and also other components of multiple repression complexes [208, 209] to regulate self-renewal and modulate ES cell fate. MAFK is a b-ZIP protein from MAF family that is extensively involved in the cell pluripotency network as well as cell differentiation processes [210]. Other TFs co-localising with these pluripotency factors are: CTBP2, EZH2, SUZ12 and CEBPB. EZH2 and SUZ12 are key components of polycomb repressive complex 2 (PRC2), while CTBP2 is a transcriptional co-repressor and chromosome remodeler [211]. CTBP2 helps to recruit histone modification enzymes, particularly for gene repression [211, 212]. Their associated histone remodelers involve: histone deacetylase 1 (HDAC1), HDAC2 and G9a, which can remove the activating mark H3K9Ac and add methyl groups to lysine 9 instead. CTBP2 also interacts with lysine specific demethylase 1 (LSD1), which removes the activating mark H3K4me on gene promoters and enhancers [213]. Therefore, the co-localization between CTBP2 and PRC2 components observed here is consistent with their repressive functions in gene regulation. The above spatial co-localization pattern is also in line with the genomic co-binding of several pluripotency factors to inactive chromatin with polycomb marks in stem cells. For instance, POU5F1 and NANOG have been shown to bind to bivalent promoters with both polycomb marks and H3K4Me3 in ES cells [214].

Another interesting thing to note is that in our co-localization color map, the two PRC2 components EZH2 and SUZ12 show extremely strong co-localization between themselves. Their co-localization scores are more than 6 fold higher than the maximum score between any other pairs. The very strong spatial co-localization of Polycomb complex components coincides with recent reports from either promoter-capture Hi-C [187] or fluorescence microscopy images [215].

To further validate our co-localization analysis, we compared the higher than expected co-localization partners of POU5f1, a crucial TF in stem cell self-renewal and differentiation, to its physically interacting partners identified via Mass Spectrometry [201, 216]. From Figure 3.27, there are 12 TFs with consistently higher than expected co-localization scores for POU5F1. Interestingly, 6 of them coincide



with the results gained from Mass Spectrometry analysis, namely, CTBP2, ZNF143, HDAC2, CHD1, ATF2 and NANOG, while only 1 TF (SP1) identified by [201] appear to have lower than expected co-localization score. This again suggests that TF spatial co-localization is closely linked to functional TF-TF interaction, not only in lymphoblastoid cells, but also in ESC. Therefore, the TF co-localization patterns we observed can provide insights for identifying further TF interacting partners, or at least, TF pairs with high level of spatial co-localization are more likely to show physically interaction than others.

### 3.4 Discussion

We presented for the first time a Hi-C map based analysis to probe the spatial TF binding site co-localization in human GM12878 lymphoblastoid cells. We found two distinct TF spatial co-localization networks, one of which is enriched in lymphocyte specific TFs. Our analysis provide robust confirmation of TF co-localization networks across different scales, both genome-wide and within each chromosome subcompartment.

Our TF co-localization analysis also provides functional insights into TF interaction and co-function in gene regulation. Using population Hi-C contact map, we identified 40 pairs of significantly co-localized TFs according to the genome wide chromosome contact map in GM12878, which were enriched in previously reported, physically interacting TF pairs. In h1-ESC, even given the limited resolution and data quality of the Hi-C contact map, the co-localized TFs identified from our study showed very good consistency with the available Mass Spectrometry analysis of partner TFs [201, 216].

TF BSs in TFCN1 are more centred around TSS, CTCF and chromosome domain boundaries, while BSs in TFCN2 have much wider distributions. This suggests that TFs in TFCN2 could be more likely to be involved in gene regulation via distal enhancers, while those in TFCN1 are more likely to bind directly to gene promoters.

This hypothesis has been validated later by the comparison of BS occupancy between the enhancer and the promoter states. TFs in TFCN2 show higher occupancy in the enhancer state; in contrast, TFCN1 has higher occupancy in the promoter state. Considering that TFCN2 are more enriched in cell-type specific functions, these results are consistent with the observation that distal enhancer-promoter interaction plays a critical role in tissue-specific gene regulation [144, 176]. It could be possible that those tissue specific TFs may establish their function through other looping mechanisms rather than via CTCF, while constitutive TF binding (TFCN1) mostly co-occur with CTCF mediated looping, thereby centring around chromosome domain boundaries as well, which are enriched in CTCF binding sites [107].

### **3.4.1 Cell-type specific TF co-localization network and its relationship with chromosome subcompartments**

Our results revealed a close relationship between transcription factor co-localization networks and chromosome A1/A2 subcompartment organisation in GM12878 in terms of TF binding site occupancy. Nonetheless, each TF co-localization network was still robustly preserved within each subcompartment. We observed a substantial bias in TF occupancy for TFCN1 and TFCN2 towards A1 and A2 chromosome subcompartment, suggesting the potential regulatory network differences between A1 and A2. It partly answers the question why given similar epigenetic marks and chromosome accessibility, A1 and A2 show drastically different chromosome contact patterns. This can possibly be a consequence of TF binding and gene regulatory network differences, where A2 is more enriched in TFCN2, which in turn is associated with tissue specific transcription regulation. Note another interesting point that A2 subcompartment harbours more long genes than A1 [1]. Long genes are more likely to associate with cell-type specific gene regulation [217].

A/B compartment segmentation has been observed across different tissues, organisms and different cells [218, 219, 140], however, chromosome subcompartments could possibly only exist in a subset of cell types, especially considering the enrichment

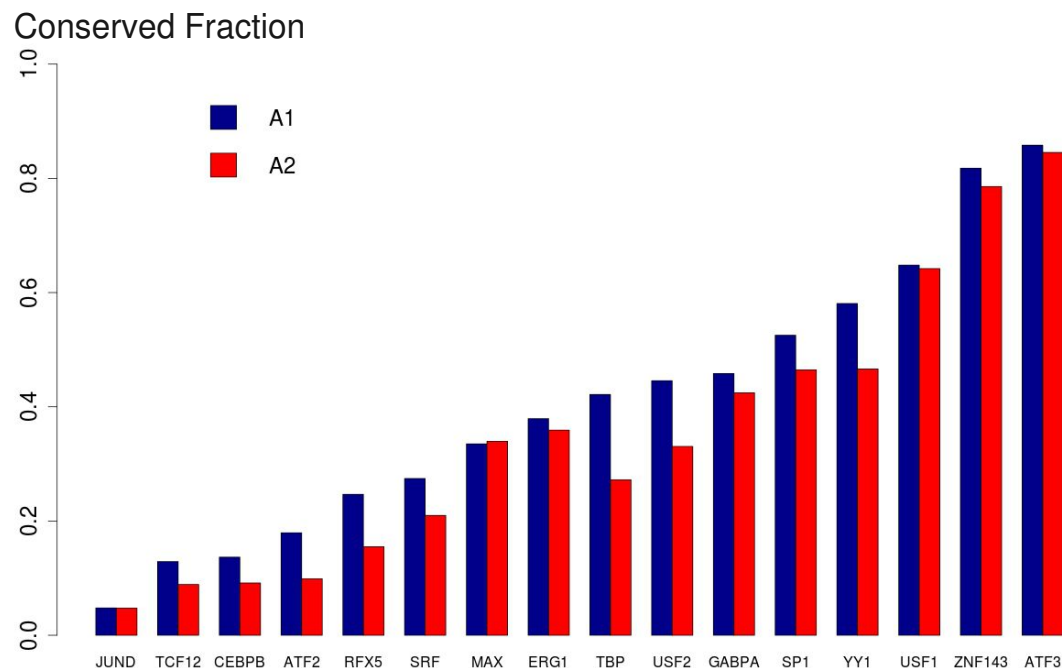


Figure 3.28: The fraction of conserved ChIP-seq peaks between GM12878 and hESC for each TF within A1/A2 chromosome subcompartments. Shown are TFs with ChIP-seq profiles available in both cell lines.

of TFCN2 in A2, which is associated with lymphocyte specific TF binding. Another point worth mentioning is that when investigating BSs conservation between GM12878 and h1-ESC for each TF, BSs in A2 subcompartment showed less conservation compared to those in A1 for nearly all TFs in common, as shown in Figure 3.28. Therefore, it would be interesting to investigate further if A2 subcompartment is a cell-type, or at least a cell lineage-specific subcompartment when more Hi-C contact map become available for other blood cell lines or cell lines of related lineages.

In Figure 3.28, we also noticed that the 3 TFs (JUND, CEBPB and ATF2) that belong to TFCN2 and also have available ChIP-seq in ESC all show a low fraction of common ChIP-seq peaks in GM12878 and ESC cells, whereas the TFs in TFCN1 tend to show more consistency across these two cell types. Taking together, it suggested that TFCN2 may have cell lineage-specific roles which may be facilitated by the co-

localisation in A2 subcompartments, consistent with the observation that TFCN2 were enriched in lymphocyte specific pathways.

### **3.4.2 TF binding in relation to heterotypic BS co-localization in two co-localization networks**

The complicated relationship between inter-TFCN heterotypic co-localization and BS occupancy is an interesting observation, though there is no straight forward explanation to this. It is far beyond the scope that can be explained by the simple TF target search mechanism with facilitated diffusion. Since TFCN1 and TFCN2 behave differently, the effect of heterotypic BS co-localization cannot be captured by a single common principle. For the promoter state, BS occupancy drops in response to high level of spatial co-localization with another TFCN. It can be viewed as a reflection of spatial segregation between TFCN1 and TFCN2 in terms of BS occupancy. This rule also holds with regards to the enhancer state in TFCN2, which shows decreased occupancy in relation to TFCN1 co-localization level; but the trend is reversed for the enhancer state in TFCN1, where higher BS occupancy is observed for higher level of co-localization regarding to TFCN2. Taking into consideration that TFCN2 has higher occupancy in the enhancer state and is a cell-type specific co-localization network, these observations emphasizes the biological function for the binding of TFs within TFCN2 to enhancers. It might be interesting to investigate further if certain TFs within TFCN2 can help to open up genome regions or bring in active epigenetic marks that enhances the binding of TFs , especially in enhancers, from both co-localization networks.

Interestingly, there are some special TFs within TFCN2 that can function as pioneer factors which help to maintain the baseline chromatin accessibility, for instance, JUND, a component of AP1 transcription factor complex [150], and *CEBP $\beta$*  [220]. AP1 transcription factor complex has been shown to prime chromatin and select regions in the genome for the binding of other TFs, for instance, nuclear receptors including glucocorticoid receptor (GR). The recruitment of GR is dependent on AP1

binding and ablating AP1 binding significantly attenuates chromatin accessibility [150]. Another example is *CEBP $\beta$* , which is also involved in steroid response. Disruption of *CEBP $\beta$*  binding led to attenuation of pre-programmed chromatin accessibility and thus attenuated the binding of other associated TFs [220]. Their results further suggested that selective targeting of certain TFs is mediated by cell-specific priming proteins, including TFs in the CEBP family and possibly other chromatin remodellers. This evidence together may help us understand why in the enhancer state, BS occupancy in both co-localization networks all positively correlate with spatial co-localization to TFCN2, which includes AP1 components and *CEBP $\beta$* . Priming proteins that can establish the pre-programmed chromatin accessibility may be more important for the function of enhancers rather than promoters, as chromatin accessibility in enhancers is known to show a lot more variation across cell-types compared to promoters [221].

### 3.4.3 The influence of Hi-C protocol and sequencing depth on Hi-C contact map quality and data analysis

The TF spatial co-localization analysis was done on two different cell lines: human GM12878 lymphoblastoid cells and h1-ESC. However, during the data analysis process, I found several problems associated with the Hi-C contact map in h1-ES cells that can potentially affect our down-stream data analysis and make the results less convincing. The Hi-C contact map of GM12878 was generated using the optimized protocol for spatial proximity ligations within intact nuclei, which best preserved the chromosome organisation in live cells. However, the one for h1-ESC used the old protocol, which ruptured the nuclei before restriction, marking of DNA ends and ligation. Even though with cross-linked chromosome, disturbance of nuclear structure might also result in a tremendous loss of certain chromosome association patterns, especially long-range contacts, as can be seen from the failure to detect loops near TADs boundaries [120], which is apparent with the improved protocol [1]. This further indicated that only formaldehyde cross-linking itself may not capture all types of chromosome proximity information equally efficiently, proximity ligation within

intact nuclei is a crucial step in improving contact map quality. Interestingly, later on, other studies were even able to use proximity ligation without any cross-linking in intact nuclei to capture sufficient Hi-C contact reads [111], though more efforts were required to preserve the native chromosome organisation as much as possible, for example, with embedded nuclei in agar plugs before restriction and ligation.

The map for GM12878 with kilo-base resolution also used super deep sequencing for the Hi-C library, which put maximum efforts in capturing proximity contacts; while in ESC, the lack of sequencing depth may result in loss of contact information.

The above disadvantages associated with h1-ESC Hi-C data had the following unfavourable influences on our data analysis: 1) the lack of sequencing depth and the Hi-C protocol defects made it difficult to detect sufficient long-range contacts, e.g. contacts between genome regions with sequence separation longer than 1Mb. Although I have tried to expand the window size up to 300kb for contact read counting between distal genome regions, total reads counts were still below 20 on average, which is insufficient for our analysis. If so, TF co-localization contributed by distal chromosome looping could be largely underestimated. 2) In our initial quality control step, we need to remove bins of genomic loci associated with insufficient mappable reads (here we chose 1/3 of the genome-wide median as a threshold) to avoid suspicious score amplification in the normalization of Hi-C contacts. In GM12878 cells, this step only removed around 8% of genomic regions, while in h1-ESC, it resulted in the loss of 31.5% of genomic regions. Even using a much relaxed threshold of 1/5 of the genome-wide median, 25% of sequences still needed to be discarded, giving a large fraction of the genome with very few mappable contact reads.

Therefore, in order to perform convincing TF co-localization analysis based on population Hi-C contact map, we suggest 1) it is necessary to use the optimized protocol for cell lysis, restriction digestion, biotin marking of DNA ends and proximity ligation within intact nuclei, even though with cross-linked product; 2) deeply sequenced Hi-C libraries are preferred when the library has been determined to be of high quality. The sequencing depth that results in the contact map resolution of 40kb may be insufficient for high-confidence TF co-localization analysis using our procedure.

## Chapter 4

# Proper scaling of Position Weight Matrices to enable TF binding strength comparison across different TFs

### 4.1 introduction

TF binding preferences to DNA have been studied extensively *in vitro*, *in vivo* and using computational methods. *In vitro* methods such as protein binding microarray(PBM) [50], high-throughput SELEX measurements [49] and DNase I-seq [21] have provided fundamental insight into the specificity of TF binding. I have reviewed the approaches to determine TF binding specificity and the bioinformatic representation of TF binding motifs using Position Weight Matrices (PWM) in section 1.1.3 and 1.1.4. We use motif in this chapter in reference to the PWM motif for a specific TF. Berg *et al.* [57] first showed that the score obtained by the PWM model is proportional to the binding energy between this TF and the DNA. In most cases the actual binding energy between the protein and DNA is not known, and the proportionality is

scaled with a factor commonly termed  $\lambda$ . Berg *et al.* originally introduced  $\lambda$  to relate the population of base-pair choices to binding free energy [56], as an analogy to the inverse temperature factor in statistical physics to describe the energy distribution and also to serve as a factor to tune the number of potential binding sites in order to satisfy the constraints on overall energy distribution.

There are some experimental techniques available to infer TF binding strength to different DNA sequences. For instance, the PBM approach [50] allows the estimation of the relative binding strength of a protein to “naked” DNA *in vitro*, but the data availability is restricted to a limited number of TFs due to high cost of the technology. In addition, PBMs are not suitable for TFs with longer motifs, as their accuracy will decrease with the increasing length of the DNA probe [50]. Therefore, in most of cases, computational approaches are preferred when estimating BS strength, especially the ones based on PWM method. [25, 81].

However, to date, there is no available software or easily implemented algorithms to computationally determine TF binding energy to specific DNA sequences, especially when it comes to binding site strength comparison between different types of TFs at large scale. This is problematic when scanning the genome with a library of PWMs, as scoring functions treat each PWM independently, and the absolute score associated with a “good match” to the PWM of one transcription factor might be associated with a mismatch for another factor. A more sophisticated application of binding site strength estimation is, for example, modelling the relationship between enhancer occupancy and gene expression [25, 222], when the affinity of TFs to specific BSs are important parameters in binding kinetic models.

In the majority of bioinformatic studies, the scaling factor  $\lambda$  is unknown and PWM scores are used at face value as measure of affinity. For example, in our own work [94] we used the PWM score without scaling to compare binding site strength across different TFs in *E. coli*, which might lead to a bias due to the absolute differences between the highest and lowest PWM scores across all TFs of interest. One approach is to scale the PWM score by a p-value for each specific score threshold [223]. This method provides a good way to define putative binding sites by choosing a proper



statistical threshold, but it is difficult to correlate these p-values with binding energy estimation, as is required *e.g.* for quantitative studies of enhancer activity [25, 222]. Other work has tried to assess the range of  $\lambda$  on the basis of fitting calculated affinity landscapes to ChIP-seq profiles [224, 225]. However, ChIP data is intrinsically noisy and the height of a ChIP peak is not an reliable representation of the real binding affinity, undermining the stability and accuracy of  $\lambda$  obtained from these methods. In Roeder *et.al* [224], the estimated  $\lambda$  for the same TF in different conditions diverged greatly in nearly one third of TFs they studied. Furthermore, there is a wide band of possible  $\lambda$  values that nearly equally optimize the correlation. Aforementioned fitting methods are further reliant on chromatin accessibility data acquired under the same experimental conditions, which is often not available for specific conditions and TFs.

We propose a simple approximation to estimate the scaling parameter  $\lambda$  based on existing PWM matrices, the average maximum mismatch energy tolerance estimated by high-throughput binding energy measurements [63] and the distribution of PWM scores of a certain TF across the genome of a specific organism. This method is independent of genome-wide binding profiles and accessibility data. Furthermore, in the cases where there are potentially inconsistent PWMs for a particular TF (*e.g.* derived on the basis of individual binding sites vs. derived from high-throughput efforts), we provide a method to convert the known  $\lambda$  for one PWM matrix of the same TF into another suitable value for a new PWM matrix. This method is based on a computational model of the facilitated diffusion of TFs on the DNA that our group established earlier [156]. We calculate sequence-specific residence times of TFs at the DNA, which is correlated with affinity. We can therefore derive  $\lambda$  for different PWMs of the same TF on the basis of the consistency of simulated residence time. These two strategies (a) calculating  $\lambda$  to scale PWM scores based on the mismatch energy theory using a simple equation and (b) converting the scaling parameter  $\lambda$  between different PWMs of the same TF on the basis of simulated residence time of facilitated diffusion provide simple but useful estimations of binding energy across different TFs using properly scaled PWM scores.

## 4.2 Methods

### 4.2.1 PWM matrices for TFs for yeast, fly and vertebrates

Position frequency matrices (PFM) used to construct PWM matrices were downloaded from the JASPAR database (JASPAR-CORE-2014 non-redundant PFM) [167]. Additional sources of PFM such as those contained in the BioConductor package *PW-MEnrich.Dmelanogaster.background* [226] were used as a source of different matrices for the same TFs. PFMs constructed with less than 30 reference sequences of validated binding sites were removed, as we deemed those insufficient descriptions of binding preference. Given that typical TF binding sites span at least six base pairs, we removed any motifs less than 6 base pairs in length.

A bioinformatics approach was used to derive PWM scores [227] as follows:

$$S_j = \sum_{k=1}^L \log_2 \frac{v_{j,k}}{f_{j+k}} \quad (4.1)$$

where  $j$  is the DNA position for the PWM score calculation,  $L$  is the length of the motif and  $k$  represents  $k^{th}$  nucleotide in the PWM motif. In addition, if there is a specific nucleotide in position  $(j + k)$  on the DNA,  $f_{j+k}$  is the frequency of this nucleotide in the whole genome of a specific organism. Nucleotide frequency used for this study in each organism were as follows: *D. melanogaster*: 0.28 for A and T, 0.22 for G and C; *S. cerevisiae*: 0.31 for A and T, 0.19 for C and G; vertebrate including human and mouse: 0.29 for A and T, 0.21 for C and G. Please note that the choice of background frequencies can be critical, and that adjustments to local extrema may be necessary. We used a pseudo-count  $\mu$  to adjust the frequency of nucleotides and obtain  $v_{j,k}$  to avoid zero frequency as follows [60]

$$v_{j,k} = \frac{n_{j,k} + f_{j+k} \cdot \mu}{\sum_x n_{x,k} + \mu} \quad (4.2)$$

where  $\mu$  is chosen to be 1 [60] and we also show that the choice of the pseudo-count  $\mu$

does not have significant influence on our results (Figure 4.12);  $n_{x,k}$  is the frequency of certain nucleotide  $x$  in a specific position  $k$  of the motif.

### 4.2.2 Simple equation to calculate $\lambda$

$\lambda$  is the scaling factor that allows for direct comparison of different PWMs in terms of binding energy to DNA. The binding energy of a TF to the DNA at a specific position can be expressed as:

$$E_j = E_0 \cdot e^{-S_j/\lambda} \quad (4.3)$$

where  $E$  is the binding energy,  $S_j$  is the PWM score and  $E_0$  is a scaling parameter. This is useful in a variety of contexts, such as comparing the binding strength of different TFs. In addition, the expected amount of time that the TF is bound to a particular sequence can be estimated as:

$$\tau_j = \tau_0(\lambda) \cdot e^{-S_j/\lambda} \quad (4.4)$$

where  $S_j$  is the PWM score at position  $j$  in the genome,  $\tau_0$  is the average residence time calculated as in [156]. This equation is widely used in simulations of TF binding kinetics [228].

Given the utility of the  $\lambda$  for estimating binding strength and occupancy time, it is very important to have a simple strategy for estimating it. We derive our equation based on the following core assumptions: 1. The top 0.1% of the highest scoring matches of the PWM to intergenic regions are considered to be possible TF binding sites, as suggested by [229]. Their genome-wide study of different eukaryotic TFs revealed an average of 1 binding site in every 1-5 thousand base pairs of intergenic sequence. This top 0.1% score threshold has also been similarly adopted in other studies [81]. In addition, if varying the top PWM score thresholds in Equation 6 from top 0.01% to top  $1 \cdot 10^{-4}$  and  $1 \cdot 10^{-5}$ , the rank of calculated  $\lambda$  still shows good correlation in each group of organism (Figure 4.11). 2. The maximum mismatch en-

ergy between the consensus binding motif and specific DNA sequences is proportional to the information content of the PWM matrix of the TF. Note that the mismatch energy we refer to in the text is derived from information theory, with the unit of bits, which can also be described as “mismatch bits. The information content ( $If$ ) of the PWM matrix is defined below,

$$If = \sum_{k=1}^L \sum_{i \in A, T, C, G} \frac{n_{i,k}}{\sum_x n_{x,k}} \log_2 \frac{v_{i,k}}{f_i} \quad (4.5)$$

where  $k$  is the  $k^{th}$  nucleotide in the PWM motif,  $\frac{n_{i,k}}{\sum_x n_{x,k}}$  is the frequency of nucleotide  $i$  in position  $k$ ,  $f_i$  is the background nucleotide frequency.

Based on the mismatch energy theory for estimating TF binding strength [57], the mismatch energy at a particular binding site  $j$  of TF species  $i$  in the genome can be expressed as:

$$E_{mismatch,i,j} = \Delta S_{i,j} / \lambda_i = (S_{max,i} - S_{i,j}) / \lambda_i$$

where  $S_{i,j}$  stands for the PWM score at position  $j$ ,  $S_{max,i}$  is for the maximum PWM score of TF species  $si$  and  $\lambda_i$  is the scaling parameter we want to estimate.

The lower boundary of potential binding sites is approximated by the top 0.1% of PWM scores following the same reason as mentioned above and corresponds to the maximum mismatch energy tolerance level as follows

$$E_{maxMismatch,i} = \frac{S_{max,i} - S_{top0.1\%,i}}{\lambda_i}$$

where  $E_{maxMismatch,i}$  stands for maximum mismatch energy tolerance for TF species  $i$ , thus,  $\lambda_i$  can be calculated using:

$$\lambda_i = \frac{S_{max,i} - S_{top0.1\%,i}}{E_{maxMismatch,i}} \quad (4.6)$$

Because different transcription factors have different DNA binding domains, the maximum mismatch energy range can vary from one TF to another. Since there is only

data available for 4 individual TFs using microfluidic platform-based binding energy measurements [63], we estimated the maximum mismatch energy for other TFs by using the available data as the average rate and assuming that the mismatch energy tolerance is proportional to the information content of the PWM matrix as follows:

$$E_{maxMismatch,i} = \langle E_{maxMismatch} \rangle \times \frac{If_i}{\langle If \rangle} \quad (4.7)$$

where  $If_i$  represents the information content of a specific PWM matrix,  $\langle If \rangle$  stands for the average information content corresponding to the average maximum mismatch energy measured by [63], which is 13.2 bits.

We chose an average mismatch energy tolerance of 6 bits based on the study by [63]. They showed by mechanical trapping of molecular interactions a significant decline in binding energy by at most 2 to 3 nucleotide mismatches, and each mismatch nucleotide contributes 2 bits in mismatch energy. Even if more mutations are introduced, the binding energy does not drop further since it has already reached the background non-specific binding energy level.

This experiment was applied only to TFs belonging to the bHLH family. In the absence of more comprehensive data, we must assume that all TFs share this value; although if more general TF in vitro binding energy measurement results become available, we suggest adjusting the specific top score threshold and corresponding average mismatch energy bits accordingly. Another report featuring TFs from different families including: p53, Max, Glucocorticoid Receptor [230] also provides additional support for 6 bits as average mismatch energy tolerance level since TFs from different families in their study have similar binding kinetics.

In order to control for PWM motif length, in the analysis of  $\lambda$  value comparison across different species and TF families, each  $\lambda$  value was transformed into a Z-score. Specifically, PWM motifs were grouped by motif length, with each group having more than 50 PWM motifs (The groups were: 7-8bp, 9-10bp, 11-12bp, 13-15bp,  $\geq 16$  bp), and the  $\lambda$  values were normalized by the mean and standard deviation within each of these groups (Figure 4.8 lists the mean and standard deviation value for each group,

Figure 4.9 depicts the distribution of  $\lambda$  at different motif length with color coded points that represent different species).

### 4.2.3 Estimating $\lambda$ of a new PWM matrix for the same TF based on the residence time landscape of the facilitated diffusion model

Sometimes there may be more than one PWM available for a specific TF. For instance, different TF motif databases (such as JASPAR [167], SwissRegulon [166], FlyFactor-Survey [231], and HOCOMOCO [165]) may have different versions of PWM motifs for the same TF. In order to directly compare the TF binding energy when using two alternative versions of a PWM, it is important to have a way of scaling the results by  $\lambda$ .  $\lambda$  can be adjusted using the formalism introduced in the previous sections. As a compute-efficient alternative, we developed a more optimal strategy for estimating  $\lambda$ , which does not require the assumption that the PWM information content influences the energy mismatch tolerance. Instead, we base our strategy on the estimation of the sequence specific residence time of a particular TF, which is a biological meaningful quantity and can be correlated with *in vitro* sequence-dependent sliding measurement of TFs [81]. For the same TF, the sequence-specific residence time distribution calculated by Equation 4 should be as consistent as possible, even when using slightly different PWMs, if an appropriate  $\lambda$  is chosen for scaling. Based on this, given a known  $\lambda$  for one PWM, we are able to find another suitable  $\lambda$  for the new PWM.

Note that the stronger the PWM score, the more likely it is that the sequence is bound by a TF and that the TF's residence time is a biologically meaningful quantity, but there is a much greater number of weak and medium strength binding sites than strong sites in the genome. Therefore, if we scored each potential binding site equally, the background of weak and medium-strength binding sites would have a greater affect on the estimated  $\lambda$  than the strong binding sites. Therefore we compare residence times across different quantiles on a logarithmic binding strength scale, so that the strongest binding sites have the most influence on our  $\lambda$  estimates. Specifically, in the

following analysis, we take the  $-\log_{10}$  of the cumulative distribution of PWM scores and select all binding sites with values greater than 3.0 (recall that this corresponds to the 0.1% percent of binding sites, which were chosen as the lower boundary of weak binding sites). We divide these top-scoring binding sites into bins every 0.1 log-quantile and calculate the average residence time for each of these bins.

Our strategy identifies the  $\lambda$  that would produce the most similar residence times for each of these log-quantiles. Assuming that for the first PWM, we already have an estimate of  $\lambda$  by either binding profile fitting or other methods, we can use Equation 4.4 to calculate the residence time for each binding strength log-quantile, as described above. In the following analysis of this paper, since there are very few well-characterized  $\lambda$  values from profile fitting, for proof-of-principle, we borrow the values obtained from Equation 6 as pre-calculated  $\lambda$ . Note that  $\tau_0$  is calculated via the strategy described in [156] from all intergenetic regions in the genome, which has a different value for each unique PWM.

Now for the second PWM, we can vary  $\lambda$  between the potential values of 0.1 and 3, which was shown to be a possible  $\lambda$  range [224], and calculate the corresponding residence times at each log-quantile level. We can now compare the reference residence times from the first PWM with the residence times for the second PWM across each binding site strength level, and for each value of  $\lambda$ . The  $\lambda$  that minimizes the mean square error between two sets of calculated residence times is chosen as the suitable  $\lambda$  value for the second PWM matrix. Since outliers can have a big influence on the mean square error, we calculated the sum of the absolute differences for the natural logarithm of residence times between the two PWM matrices for these quartile bins (Equation 8) to make a comparison with the method that uses mean square error.

$$\sum_q |\ln \tau_{q,\lambda} - \ln \tau_{q,ref}| \quad (4.8)$$

where  $q$  represents each quantile in the quantile series,  $\tau_{q,\lambda}$  is the residence time in a specific quantile of a particular  $\lambda$ ,  $\tau_{q,ref}$  is the residence time in the same quantile of the known  $\lambda$  of the reference PWM matrix. The  $\lambda$  derived by these two methods show good consistency with adjusted  $R^2$  of 0.9644 ( $p=6.3 \cdot 10^{-9}$ ). Thus, there should

not be significant bias using either of these two methods.

The R scripts for both converting  $\lambda$  between two PWM matrices and estimating  $\lambda$  using Equation 6 are provided in the following link:

[https://github.com/XyMa/PWM\\_scale](https://github.com/XyMa/PWM_scale).

## 4.3 Results

### 4.3.1 Estimating scaling parameter $\lambda$ for binding site affinity across different species and TF families based on Equation 6

The  $\lambda$  parameter is the important link between PWM score, the estimated binding energy and simulated TF residence time. Estimating TF binding site affinity by comparing PWM scores at face value can lead to large biases, especially when it comes to the comparisons between multiple types of TFs. Several properties of the PWM matrix itself can influence PWM scores. For example, the information content of PWM matrices can be positively correlated to the maximum possible PWM score (Figure 4.1 with an  $R^2$  value of 0.597). Thus, the absolute value of PWM scores cannot be compared directly across different TFs as an indicator of binding site strength. Therefore, proper scaling of PWM score is needed in order to compare binding site affinity across different types of TFs. Based on the methods proposed by Berg *et al.* [57], the TF binding energy for a specific binding site can be computed by Equation 4.3 using the estimated  $\lambda$ .

$\lambda$  calculated by this method are all within the range suggested by [224], which are listed in Table 4.1 for different organisms. The values for vertebrate species refer to all available vertebrate TFs obtained from the non-redundant PFM JASPAR database. The upper and the lower bound of  $\lambda$  across all organisms are all in the range of 0.25 to 2.83. This indicates that all eukaryotic TFs, no matter which organisms



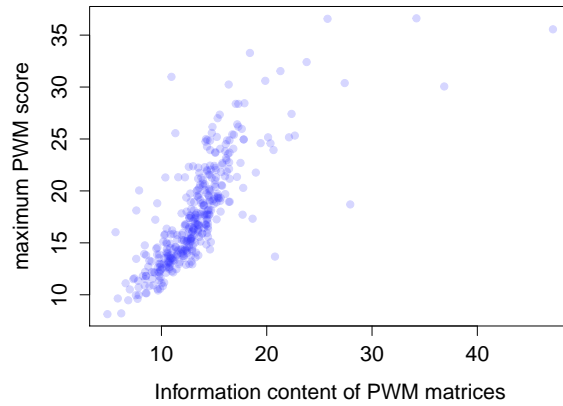


Figure 4.1: *The relationship between maximum PWM score and information content of PWM matrices.* Individual dots represents each PWM matrix generated from the non-redundant PFM JASPAR-CORE database [167] after the filtering procedures specified in the Methods section. There is a strong positive correlation between the information content of the PWM and the maximum possible PWM score that could be generated by that PWM, with an adjusted  $R^2$  value of 0.597. .

they belong to, all share energetically similar DNA binding mechanisms, because  $\lambda$  can be interpreted as a metric for the chemical property of stickiness between the TF molecule and DNA. To demonstrate the biological application of  $\lambda$ , Figure 4.2 shows an example of the *D. melanogaster Even-skipped stripe 1* enhancer with the comparison between PWM score and the affinity estimation using  $\lambda$  scaling. The usefulness of  $\lambda$  estimates becomes apparent when comparing the first two binding sites indicated by blue arrows in this locus; the second binding site has a higher PWM score, but its binding strength is lower than the first one once the  $\lambda$  scaling factor is taken into account. Similar situations also appear in the overlapping binding site of Bicoid and Kruppel indicated by the third arrow. Thus, only comparing the raw value of PWM score [94] may lead to false interpretations of binding site importance. Although there is no specific experimental evidence for the relative importance of binding sites for this specific enhancer, this example serves to demonstrate how a different interpretation of the contribution of individual binding sites can lead to

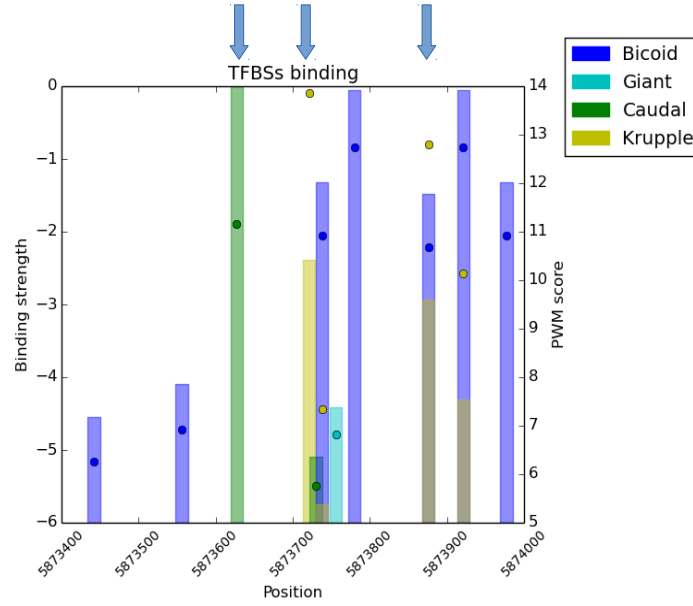


Figure 4.2: A comparison between PWM score and binding site strength in the *D. melanogaster even-skipped stripe 1* enhancer. The *even-skipped stripe 1* enhancer on chromosome 2R is dense with binding sites. We compare the raw PWM scores (circles) and the  $\lambda$ -scaled binding strength (height of the bars) for each of these binding sites, colour-coded by the type of TF. Based on raw PWM scores, one might assume that the Caudal site indicated by the first blue arrow would have a lower binding strength than the Kruppel site indicated by the second blue arrow; however, once the binding strength is scaled by  $\lambda$  using Equation 4.5, it becomes evident that the opposite is the more likely scenario. The third arrow points to a location where a Kruppel and a Bicoid binding site overlap. Here, the  $\lambda$  adjusted binding strength estimates would suggest that Bicoid binding site is stronger, while a raw PWM score would suggest the opposite. These results illustrate how using raw PWM scores may result in biased interpretation of the relative binding strength of TFs.

alternative testable hypotheses.

Next, we calculated  $\lambda$  for each TF in *S. cerevisiae*, *D. melanogaster* and available vertebrate TFs in JASPAR [167]. Figure 4.3A to 4.3C show the overall  $\lambda$  distribution in each group of organisms. After controlling for motif length, there is a significant difference between vertebrate and *S. cerevisiae* motifs (Welch t-test p-value=0.008)

Table 4.1: Maximum, minimum and the average values of  $\lambda$  in 3 groups of organisms.

	<i>S. cerevisiae</i>	<i>D. melanogaster</i>	Vertebrates
maximum	2.83	2.72	2.82
minimum	0.26	0.35	0.25
mean	1.25	1.40	1.73

(Figure 2D) and between *D. melanogaster* and vertebrate motifs (p-value=0.043) , but no significant difference is detected between *S. cerevisiae* and *D. melanogaster*. Furthermore, we grouped  $\lambda$  values, normalized by PWM motif length, according to different TF families in JASPAR [167] (Figure 4.4). The distribution of raw  $\lambda$  values across different TF-families are depicted in Figure 4.5. The basic leucine-zipper family and helix-loop-helix family are two families with the highest average z-score of  $\lambda$ , compared with other groups with Welch t-test p-values equal to  $8.9 \cdot 10^{-4}$  and  $3.7 \cdot 10^{-5}$  respectively. TF families that belong to the same superfamily show similar  $\lambda$  distribution. For example,  $\beta$ - $\beta$ - $\alpha$  zinc-finger family and zinc-finger nuclear receptor both belong to the zinc-finger TF super family, and there is no significant difference is detected between these two (Welch t-test p value=0.35), while both of them are significantly lower than the aforementioned two families (p-value= 0.012 and  $5.0 \cdot 10^{-5}$ ). In addition, homeobox and forkhead TF families, both of which belong to the helix-turn-helix(HTH) TF super family, show no difference in  $\lambda$  z-score distribution (p value=0.27), but appear to have lower average  $\lambda$  compared with leucine-zipper, helix-loop-helix family and zinc-finger super family (Welch t-test p-value equals to  $5.2 \cdot 10^{-6}$ ,  $1.6 \cdot 10^{-7}$  and  $2.2 \cdot 10^{-4}$ , respectively).

Since  $\lambda$  is the denominator to the PWM score differences between one binding site and the consensus sequence in Equation 4.3, a larger  $\lambda$  indicates lower mismatch energy when  $\Delta S_j$  is the same. Thus, with the same possible mismatch energy range, if  $\lambda$  is larger, the PWM score can have a greater range from the consensus sequence to the potentially weakest binding site, which indicates the binding motif for the TF family has higher flexibility as suggested by [232]. This is consistent with the finding that the TFs in the zinc-finger super-family including the nuclear receptor and  $\beta$ - $\beta$ - $\alpha$  zinc-finger families are less constrained to a particular motif than HTH super family. Additionally, cross species comparison of  $\lambda$  indicates that from yeast to vertebrate,

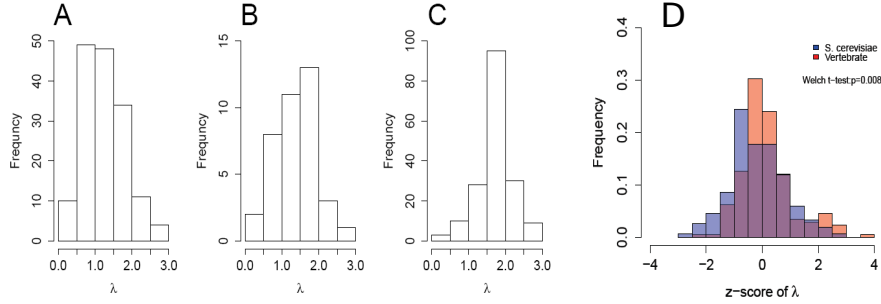


Figure 4.3:  $\lambda$  distributions across different organisms. The histograms depict the  $\lambda$  values estimated from Equation 4.6 for the JASPAR non-redundant core motifs in *S. cerevisiae* (A), *D. melanogaster* (B) and available vertebrates (C) [167]. Figure D depicts the comparison between z-score distribution of  $\lambda$  for vertebrate and yeast TFs after controlling for motif length.

more flexible TF motifs are used, which is consistent with the result from [233] that organisms which appeared more recently in evolution tend to use more TFs with motifs of higher flexibility.

#### 4.3.2 Comparison of $\lambda$ values estimated with Equation 4.6 to $\lambda$ values derived from fitting ChIP-seq data

We compared our estimated  $\lambda$  values with those estimated from ChIP-seq experiments by Zabet *et al* [225] (See Table 4.2). Equation 6 provides a close approximation of all five values estimated in this paper (adjusted  $R^2=0.64$ , p-value=0.061). We also compare our results with the  $\lambda$  values reported by Roeder *et.al* 2007 [224] for 11 yeast TF motifs from TRANSFAC [234] (See Table 4.2). For each of the 11 TFs, Roeder and colleagues fit  $\lambda$  values to ChIP-seq data from cells grown in different growth mediums, leading to a range of potential  $\lambda$  values for each TF. However, for each specific cell growth condition, only the most optimal value of  $\lambda$  was selected for each TF, even in circumstances in which there is a plateau in the parameter space with many possible  $\lambda$  values fitting the data nearly equivalently. The  $\lambda$  value ranges from their study and

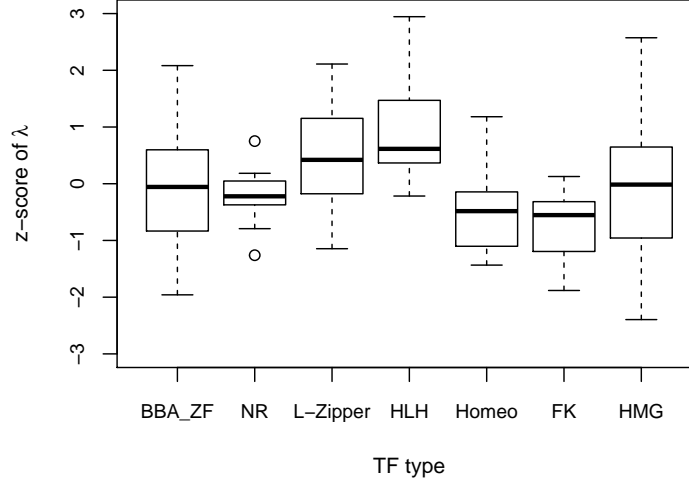


Figure 4.4:  $\lambda$  z-score distribution comparison across major TF families. BBA-ZF represents the  $\lambda$  distribution for  $\beta$ - $\beta$ - $\alpha$  zinc-finger family; NR is zinc-finger nuclear receptor family; L-zipper stands for the basic leucine-zipper family; HLH is helix-loop-helix family; Homeo is homeobox family; FK is fork-head family and HMG is high mobility group family. For each group,  $\lambda$  was calculated by Equation 6 and z-score is obtained by normalizing  $\lambda$  in each PWM motif length group.

the estimated results from Equation 4.6 using default parameters are listed in table 4.2. For 8 out of 11 motifs (6 out of 8 TFs), our results are within their estimated range or very close (absolute differences within 0.25), however, another 3 motifs for 2 TFs show poor correlation. It is possible that in some specific cases the assumed default parameters in Equation 4.6 could deviate from the real binding properties of these TFs, which can potentially lead to bias in the estimation of  $\lambda$ . Alternatively, these  $\lambda$  values may lie within the parameter plateau region, and might be a suitable fit for the experimental data.

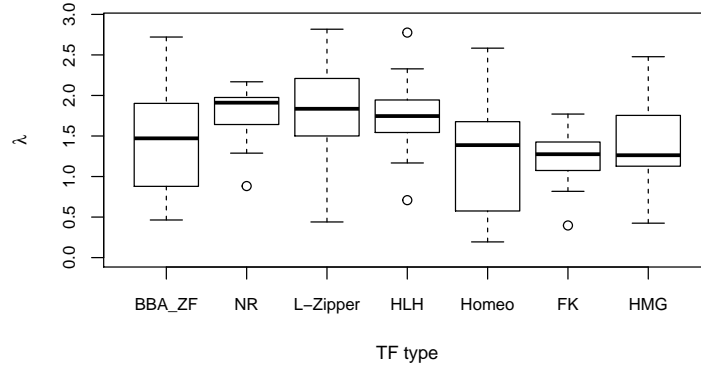

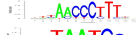

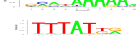



Figure 4.5: *Estimated  $\lambda$  distribution across major TF families.* BBA-ZF represents the  $\lambda$  distribution for  $\beta$ - $\beta$ - $\alpha$  zinc-finger family; NR is zinc-finger nuclear receptor family; L-zipper stands for the basic leucine-zipper family; HLH is helix-loop-helix family; Homeo is homeobox family; FK is fork-head family and HMG is high mobility group family. For each group,  $\lambda$  was calculated by Equation 4.6.

### 4.3.3 Converting $\lambda$ between different PWM matrices of the same TF

In many cases there are two PWMs available for the same TF, and one of these PWMs might already have a reliable estimate of  $\lambda$ , from any number of experimental or computational approaches [225]. In such circumstances, we provide a strategy to estimate the unknown  $\lambda$  associated with the alternative PWM matrix. It would be possible to calculate the unknown  $\lambda$  from Equation 6, but this does not incorporate the additional data available (i.e. the known  $\lambda$ ). Our alternative strategy not only incorporates this data, but also loosens the assumption in Equation 4.6 that the maximum mismatch energy for DNA binding is proportional to information content.

The procedure to compute a suitable  $\lambda$  is based on the concept of sequence-specific residence time (Equation 4.4), as illustrated in Figure 4.6. Initially, a well-characterized  $\lambda$  is computed or measured for the first PWM of a particular TF, and then we use this value to derive a  $\lambda$  that is appropriate for the second PWM of the same TF. As part

TF name	Estimated $\lambda$ from Zabet et.al, 2014	Estimated $\lambda$ from Equation 6	Motif logo
Gt	1	1.17	
Kr	2	1.93	
Bcd	1.5	1.44	
Hb	1	0.70	
Cad	1.5	1.06	







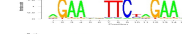
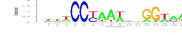



PFM name from TRANSFAC	Estimated $\lambda$ range from Roider et.al, 2007	Estimated $\lambda$ from Equation 6	Motif Logo
GAL4_01	0.25-1.45	1.17	
GAL4_C	0.25-1.30	1.10	
GCN4_01	0.50-0.60	0.52	
GCN4_C	0.50	0.64	
HSF_04	0.80-0.90	1.05	
HAP1_B	0.75	1.00	
MCM1_02	1.45-1.70	1.32	
MIG_1	0.90	1.10	
ABF1_01	0.60-0.65	1.37	
ABF1_C	0.45-0.50	1.01	
RAP1_C	0.15-0.60	1.22	

Table 4.2: Comparison of  $\lambda$  values estimated with Equation 6 to  $\lambda$  values derived from fitting ChIP-seq data of Zabet *et al.* 2014 and Roider *et al.* 2007.

of the calculation of the  $\lambda$  for the second PWM, Figure 4.6C shows a heatmap of the estimated residence times for a TF named lame duck (lmd) in a particular binding strength quantile, at different values of  $\lambda$  (ranging from 0.1 to 3.0 as suggested by both [224] and the range of estimated  $\lambda$  using Equation 6 across different organisms). Both PWMs for the TF come from FlyFactorSurvey database [231], but they are derived from different reports with motif logos shown in Figure 4.6B. Blank regions in the heatmap indicate the choice of  $\lambda$  would generate a residence time outside the range of pre-calculated possible residence times using the first PWM and the existing  $\lambda$  value, implying that these  $\lambda$  values for the second PWM are unsuitable. As shown in the heatmap, blank regions often appear in very low values of  $\lambda$ , while if  $\lambda$  is too large, the possible residence time range from weak to strong binding sites is often

restricted, which means high affinity sites cannot be effectively distinguished from low affinity sites.  $\lambda$  values with residence times all within the reference range can be further selected, as specified in Methods. Figure 4.6D-F compares the residence time values between two different PWMs, at different values of  $\lambda$  for the second PWM. We see that the  $\lambda$  in Figure 4.6D and 4.6F would not allow for consistent residence times between the two PWMs, but Figure 4.6E does provide consistent results. Therefore, the  $\lambda$  adopted in Figure 4.6E is picked up as the suitable value for the second PWM matrix. More examples of residence time heatmaps for converting  $\lambda$  between different PWMs are shown in Figure 4.13.

In order to evaluate the consistency of  $\lambda$  estimation between the above method and using Equation 4.6, we use the examples of 20 *D.melanogaster* TFs with more than 1 version of PWMs available from different experiments. These PWMs are obtained from the *BioConductor* R package *PWMErich.Dmelanogaster.background* [226] and their labels are listed in Figure 4.7. Since there are only few  $\lambda$  available from binding profile fitting, just for the purpose of illustration, the reference values of  $\lambda$  were pre-calculated from Equation 4.6 instead. New  $\lambda$  values for PWMs obtained from other experiments are computed using both methods and they show good consistency with each other with adjusted  $R^2$  equals to 0.88 (Figure 4.14). Converting  $\lambda$  between these two PWMs in the opposite direction also show similar results (data not shown). It indicates that both methods provide consistent estimates of  $\lambda$ , even though they have different core assumptions.

## 4.4 Discussion

Estimating TF binding site strength based on PWM is essential to the modelling of TF-DNA interactions, however a proper scaling parameter is needed when using the PWM score to derive estimations of TF binding energy. Here, we provide two different methods for estimating the scaling parameter  $\lambda$  based on different situations. A simple estimation requiring the minimum information can be achieved via using Equation 4.6. It only needs a PWM matrix and the genomic sequences of a certain organism



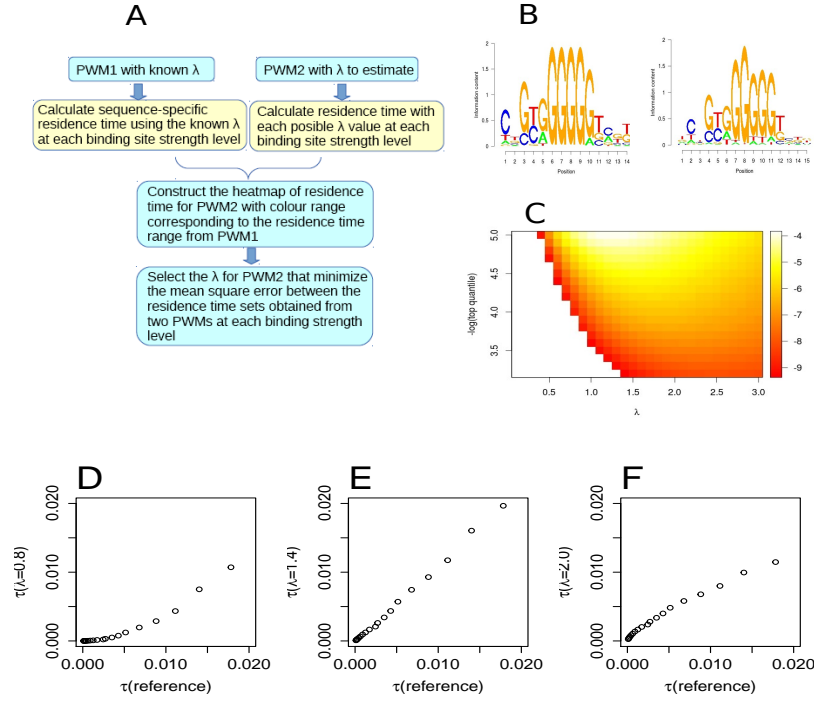


Figure 4.6: *Conversion of  $\lambda$  between two PWM matrices for the lmd transcription factor.* The flow chart shows the procedure to obtain an optimised  $\lambda$ , given two different PWMs and one known and one unknown  $\lambda$  (A). Subfigure B illustrates the two alternate PWMs for *lmd* which are available in the FLYFACTORSURVEY database [231]. Equation 6 suggests that PWM1 has a  $\lambda$  of 1.6, and we are trying to find a  $\lambda$  for PWM2. Subfigure C is a heatmap of the residence time distribution of PWM2 for different values of  $\lambda$  and different binding site strength level. Each column of the heatmap represents a specific  $\lambda$  value and each row represents a specific binding site strength level measured by the  $-\log_{10}$  of the corresponding top quantiles from low affinity to high affinity sites. Blank regions in the heatmap indicates  $\lambda$  values which lead to residence time out of the reference scale that is an indication of unsuitable  $\lambda$  values. D, E and F show the correlation of residence time between PWM1 and PWM2 using specific  $\lambda$  values of 0.8, 1.4 and 2.0, respectively. The curve in subfigure E has the lowest mean square error, and so we assign PWM2 to have a  $\lambda = 1.4$ .

as inputs, which is easy to implement compared to other complicated methods using fitting to ChIP-seq profiles [224, 225]. The second method converts a  $\lambda$  specific to one PWM into  $\lambda$  for a different PWM of the same TF. This metric follows the

PWM1	PWM2
CG11085_Cell_FBgn0030408	CG11085_SOLEXA_FBgn0030408
Dfd_Cell_FBgn0000439	Dfd_SOLEXA_FBgn0000439
nub	nub_SOLEXA_5_FBgn0085424
Unc4_Cell_FBgn0024184	Unc4_SOLEXA_FBgn0024184
Eve_Cell_FBgn0000606	Eve_SOLEXA_FBgn0000606
hb	hb_SOLEXA_5_FBgn0001180
AbdB_Cell_FBgn0000015	AbdB_SOLEXA_FBgn0000015
Bap_Cell_FBgn0004862	Bap_SOLEXA_FBgn0004862
BH2_Cell_FBgn0004854	BH2_SOLEXA_FBgn0004854
gl_SANGER_5_FBgn0004618	gl_SOLEXA_5_FBgn0004618
her_SANGER_10_FBgn0001185	her_SOLEXA_10_FBgn0001185
Hgtx_Cell_FBgn0040318	Hgtx_SOLEXA_FBgn0040318
ken_SANGER_10_FBgn0011236	ken_SOLEXA_5_FBgn0011236
klu_SANGER_10_FBgn0013469	klu_SOLEXA_5_FBgn0013469
lmd_SANGER_5_FBgn0039039	lmd_SOLEXA_5_FBgn0039039
Lag1_Cell_FBgn0040918	Lag1_SOLEXA_FBgn0040918
Lbl_Cell_FBgn0008651	Lbl_SOLEXA_FBgn0008651
Six4_Cell_FBgn0027364	Six4_SOLEXA_FBgn0027364
ab_SANGER_10_FBgn0259750	ab_SOLEXA_5_FBgn0259750
en_FlyReg_FBgn0000577	en_SOLEXA_2_FBgn0000577

Figure 4.7: The list of labels of PWMs used in evaluating the consistency between the two  $\lambda$  estimation methods. Those PWMs were downloaded using PWMEnrich.Dmelanogaster.background package in R.

motif_length_bin	lambda_mean_in_each_length_bin	standard_deviation_of_lambda_in_each_length_bin
7-8	1.0402	0.4960802
9-10	1.394569	0.5560687
11-12	1.873732	0.4381651
13-15	2.01405	0.2984683
>=16	2.032448	0.2435366

Figure 4.8: The mean and standard deviation of  $\lambda$  in each group of certain motif length.

logic of sequence-specific residence time from the facilitated diffusion model of TF target search [156]. This method is particularly useful when converting a previously estimated  $\lambda$  into the new one for a more up-to-date or otherwise alternative PWM matrix.

These two methods are consistent with one another (Figure 4.10) and also with previ-

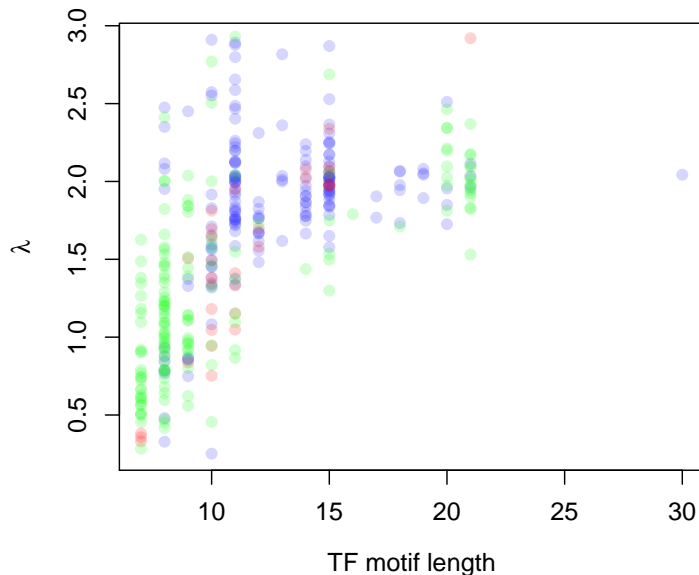


Figure 4.9: *Estimated  $\lambda$  distribution in relation to PWM motif length.* Each color coded point represents a specific  $\lambda$  value of a TF estimated by Equation 6 for *S. cerevisiae* (green), *D. melanogaster* (red), and vertebrate (blue). There is a positive correlation between estimated  $\lambda$  value and TF motif length with adjusted  $R^2$  equals 0.33.

ously established methods. For instance, Equation 4.6 can provide very similar results compared with the estimated  $\lambda$  from ChIP-seq data fitting [224, 225]. Although our estimates of  $\lambda$  are mostly consistent with those estimated by Zabet [225] and Roeder [224], it is not possible to compare our  $\lambda$  estimates to experimentally derived values at scale currently, simply because this type of data is unavailable for most of TFs. Having more experimentally derived estimates would enable us to adjust currently fixed parameters in our equation for different TF families, such as the top-scoring threshold, which was assumed to be a constant in our current equation, but could vary across different TF families. The consistent value range of  $\lambda$  in different organisms calculated by this method provides additional support for the applicability of this simple equation. Furthermore, the estimated distribution of  $\lambda$  values for different TF families

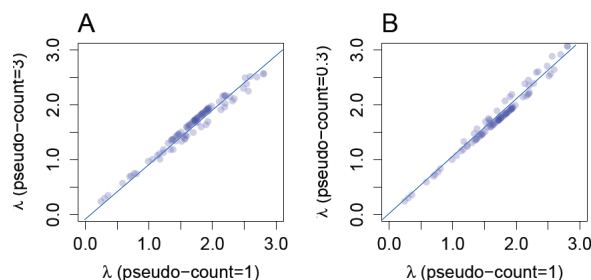


Figure 4.10: Comparison of  $\lambda$  values calculated by using different pseudo-count values in PWM matrices. Subfigure A shows the comparison between the  $\lambda$  values obtained by using PWM matrices with pseudocounts of 1 and 3 (the adjusted  $R^2$  is 0.973), while subfigure B compares pseudocounts of 1 and 0.3 (the adjusted  $R^2$  is 0.978). Each dot represents a TF from 100 randomly chosen vertebrate TFs in JASPAR database [167].

make sense in the light of motif choice for each TF families [235]. For example, TFs in the zinc-finger TF super-family including nuclear receptor zinc-finger and  $\beta$ - $\beta$ - $\alpha$  zinc-finger families have more flexible binding motifs, which are able to suit a wider range of possible binding sites compared to the helix-turn-helix super-family, which own a more restricted motif consensus sequence [236]. In contrast, some TF families belong to the same super-family and also share similar binding domain properties can have strong similarity in  $\lambda$  distribution, *e.g.* homeobox family and forkhead family (they both belong to the helix-turn-helix super-family). The two TF families that show the highest average z-score of  $\lambda$  values (namely, basic leucine-zipper and helix-loop-helix families) tend to form homodimers and heterodimers, though some TFs in other TF families also tend to dimerise *e.g.* some members in homeobox family. If PWM motifs for either monomers or dimers are available, the corresponding  $\lambda$  scores can be roughly estimated following the same procedure using Equation 6, or we can further use the second method mentioned before to convert  $\lambda$  values between different PWM matrices by the keeping residence time consistent. However, our method only

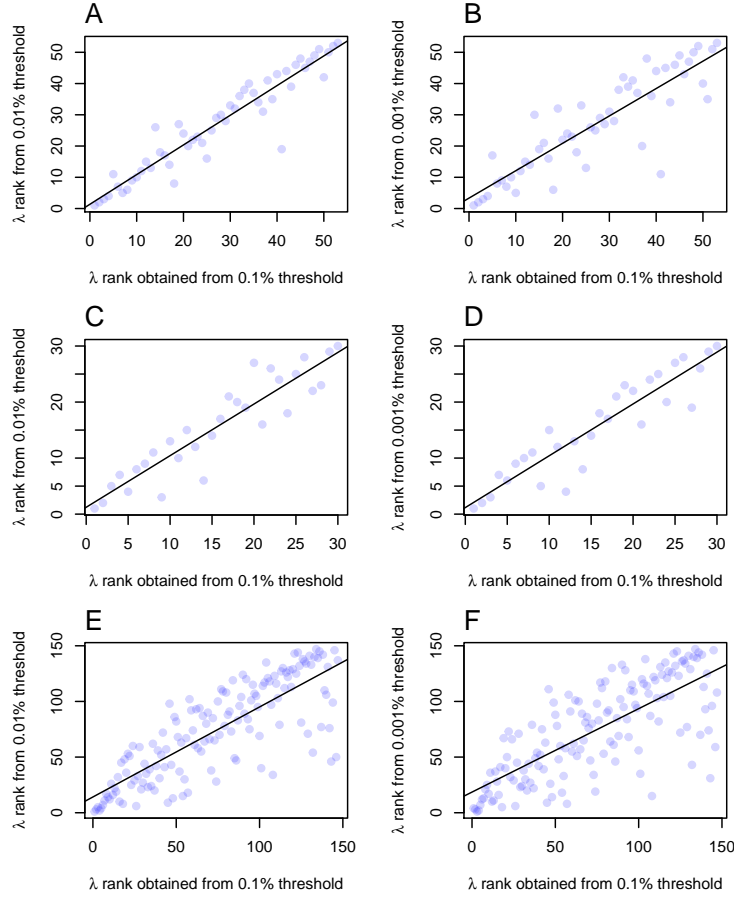


Figure 4.11: *Correlation of  $\lambda$  rank obtained by using different top score threshold in Equation 4.6.* We compare the  $\lambda$  rank for different TFs in each group of organisms (A, B for *S. cerevisiae*, C and D for *D. melanogaster*, E and F for vertebrate PWM motifs) for by adopting different top score threshold of top 0.01% or 0.001% instead of the default value of 0.1% in Equation 6. The adjusted  $R^2$  for the  $\lambda$  rank correlation between 0.1% and 0.01% thresholds for *S. cerevisiae*, *D. melanogaster*, and vertebrate motifs are 0.94, 0.89 and 0.80, respectively, with p-values all less than  $10^{-8}$ . As for the  $\lambda$  rank correlation between 0.1% and 0.001% thresholds, the adjusted  $R^2$  are 0.87, 0.92 and 0.74, respectively (p-values all less than  $10^{-6}$ ).

considers TF-DNA interaction, ignoring the effects of TF-TF interactions that could stabilize TF binding.

There are some points that should be noted when using equation 4.6: first, it cannot

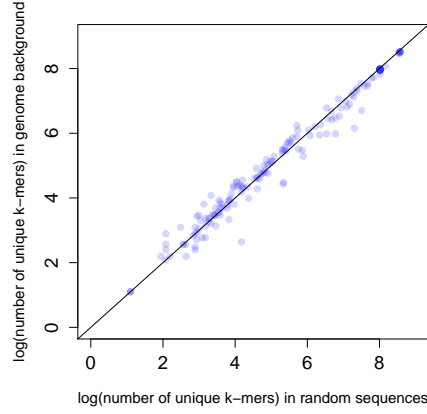


Figure 4.12: *Comparison of unique k-mer number passing 0.1% top PWM score threshold in genomic background versus that in random sequences.* For each TF PWM motif, we calculated the logarithm of the number of unique k-mers that passes the threshold in both genomic background and random sequences that have the same GC content and they show good correlation with adjusted  $R^2$  equals 0.98, p-value  $< 10^{-16}$ .

be applied to very short TF motifs that are less than 6 base pairs in length. Since this method depends on calculating the difference between the top 0.1% of PWM scores and the maximum score, if the motif is only 5 base pairs in length, the number of possible choices for sequence combination of 5 base pairs is only 1024, then the top 0.1% of PWM scores is the top score. However, most eukaryotic motifs are more than 6 base pairs long. Eukaryotic TFs on average cover 15 bp of DNA with a core motif length of 8-15 bp [25]. Thus, this limitation should not be a problem in the majority of cases. However, if a higher threshold *e.g.* top  $1 \cdot 10^{-5}$  is applied with certain adjustment for average mismatch bits in the denominator, it requires the PWM motif to be at least 10bp long, which will limit the applicability of this method. That is why we use top 0.1% as our default threshold choice.

Though we set the default cut-off threshold as the top 0.1% PWM score, but varying this threshold up to the top 0.001% does not significantly influence the rank of  $\lambda$  (Figure 4.11). Note in Equation 4.6, the average mismatch energy bit score in the

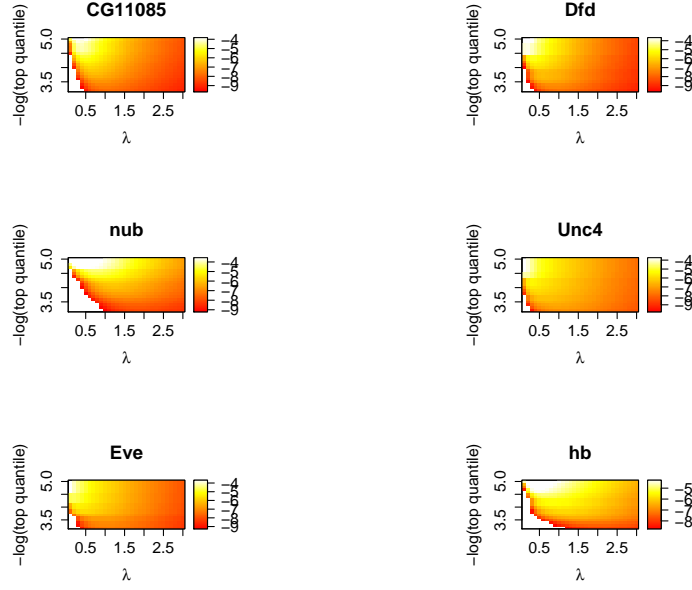


Figure 4.13: *Heatmaps for  $\lambda$  conversion between different PWMs.* These are additional examples of heatmaps of sequence-specific residence time that are used for  $\lambda$  conversion between different PWM matrices of the same TF. Alternative versions of PWM matrices are from *BioConductor* R package of *PWMEnrich.Dmelanogaster.background* [226]. Each column of the heatmaps represents a specific  $\lambda$  value and each row represents a specific binding site strength level.

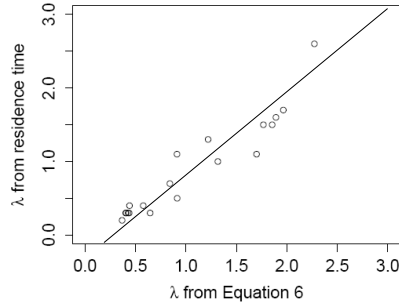


Figure 4.14: *Consistency of  $\lambda$  estimation between two methods.* This figure shows the correlation between  $\lambda$  values obtained from Equation 6 and from  $\lambda$  conversion using the heatmap of sequence-specific residence time. The adjusted  $R^2$  is 0.88, p-value =  $5.9 \cdot 10^{-5}$ .

denominator is the one corresponding to the certain top score threshold, which means if a new top score threshold is adopted, the average mismatch energy bit score should be updated accordingly, but given very limited binding energy measurement data, it is difficult to select specific values for each corresponding binding site strength level. Thus, we simply compared the rank correlation of  $\lambda$ , which is not affected by the linear scaling factor of average mismatch energy bits. Although we estimate  $\lambda$  by top scoring genomic sequences, it does not substantially affect this analysis if this is done on random sequences with the same GC content, since given the size of the genome, local binding site patterns will not have much influence on the general distribution of binding site strength. Figure 4.12 shows the number of unique k-mers passing the 0.1% top score threshold in genomic sequences correlates well with that in random sequences of the same GC content.

Another core assumption in the first method is that the mismatch energy tolerance range measured in bits is proportional to the information content of a specific PWM. We reason that if the information content is a good indication of how specific a TF is, the energy drop measured in bits between strong and weak binding sites ( $S_{max,i} - S_{top0.1\%,i}$ )/ $\lambda_i$  should have some relationship with the binding specificity of a particular TF. The more specific a TF is, the more significant the energy drop can be. Given the scarce biophysical data for the binding energy range in relation to binding specificity, we simply assume this relationship to be linear. This enables us to deal with the bias from the differences in information content of most PWMs; however, it might not be true for PWMs with extremely high information content. For instance, a yeast transcription factor named IXR1 has an information content of 47 bits according to the PFM from JASPAR [167], which is substantially larger than the average information content of 13.2 bits we used in our analysis. The binding energy of IXR1 is probably overestimated in the above case, which may lead to a lower  $\lambda$ , but these cases are very rare and only 7 PWMs in our analysis (less than 1.5%) are associated with information content greater than 20. Further, we note that the experiment by Maerkl *et al.*, 2007 [63] was applied only to TFs belonging to a specific TF family—the bHLH family. In the absence of any alternative data, we assume that this value is scaled by the information content of the PWM. If more in-vitro binding



energy measurements should become available in the future, we suggest to adjust the specific top score threshold and their corresponding average mismatch energy bits accordingly to different TF families if possible.

There are two limitations of this method, which may lead to some biases between different organisms and different TF families. One limitation is related to the calculation of mismatch energy tolerance in different groups of TF families. We apply a single cut-off threshold of top 0.1% PWM score for weak binding sites suggested by [229], but it could be possible for different TF families, different thresholds should be used due to their structural variations in DNA binding domains. However, it is difficult to choose specific thresholds for each TF family based on the availability of data currently. Further, from the definition of information content of the PWM matrix, it sums up information content gain from each nucleotide [228], which implies longer motifs including more flanking base-pairs may have slightly higher information content compared to the shorter ones just with core motifs, which is an artefact of computation. However, there is no satisfactory way to deal with this. One possible solution could be using the information content per nucleotide instead of the total information content, but this may be problematic as the information content contributed by flanking sequences constitutes only a tiny fraction compared to the core motifs. Therefore, if we divide total information content by the length of the motif, the dilution of information content could lead to even larger biases. Hence, in our analysis of comparing  $\lambda$  value distribution across different organisms and TF families, we choose to use a strategy that controls for motif length by normalizing it to the mean in each motif length bin. Another potential solution is trying to define a core motif from one PWM matrix, but this requires having sufficient knowledge about the TF motif of interest. Additionally,  $\lambda$  will not be a reliable measure of biochemical stickiness of the TF to the DNA if the PWM itself is not an accurate representation of TF binding. A PWM assumes that each nucleotide position independently contributes to TF binding affinity, which may not be the case [185, 237]. For example, a study by Storm *et. al* 2007 [64] used both a single nucleotide model and a di-nucleotide model to fit the binding energy measurements in [63]. Although in their study, they reached the conclusion that the di-nucleotide model provides a better fit

to the experimental data, the single nucleotide model could also perform well when non-specific binding energy was taken into account. In addition, the composition of the position frequency matrix of the PWM may contain biases due to the difficulties of attaining an unbiased validated binding site set.

Also, it should be pointed out that residence time in this paper refers to an estimate based on the biophysical model proposed in [81, 156]. However, other papers report inconsistent scales of residence time according to different experimental approaches. For example, the residence time estimations obtained by Competition-ChIP methods [6] do not share the same order of magnitude compared to the residence times measured by FRAP or single molecular tracking [230, 85, 84], which can probably be an artifact of experimental methods or alternatively, the range of residence time truly varies greatly across different TFs [28]. Because the experimentally determined values are not comparable to each other, we simply adopt bioinformatics-based approaches to compute residence time. Since our method converts  $\lambda$  between different PWM matrices of the same TF under the concept of residence time, it avoids fitting inconsistent experimental observations and potential variations in DNA-binding kinetics for different TFs.

Although in many cases PWMs are not optimal representations of binding motifs, they have become widely adopted to identify potential TF binding sites. However, it is important to remember that the face value of a PWM score is not directly correlated to the binding energy, but rather depends on the scaling parameter of  $\lambda$ . Previously, researchers either assumed that  $\lambda$  adopts similar values across all PWMs or estimated it through compute-intensive binding profile fitting procedures [224, 225]. There are several alternative ways of identifying potential sites based on PWM score that estimate how likely these individual binding sites can occur in the genome as a representation for binding site strength [223]. Other studies combine more local information *e.g.* DNA sequence conservation and epigenetic marks with PWMs to identify potential binding sites and this has been shown to obtain higher confidence and better performance in mammalian genomes [238]. These methods are useful in defining potential binding sites, but their results could be difficult to interpret in terms of TF binding energy which is essential in modelling TF binding dynamics and

enhancer activities [25]. Here we provide two simple strategies for estimating  $\lambda$ , which will let us more clearly link PWM scores with the energetics of TF binding.

# Chapter 5

## Conclusion, discussions and future directions

### 5.1 Conclusion

Transcription factor (TF) binding has shown to be influenced by DNA binding sequence motifs and epigenetic modifications; however, the relationship between genome spatial organisation and TF binding is not very well studied. Using the high resolution Hi-C contact map of human GM12878 cell line of 5kb [1], we systematically investigated the genome-wide spatial co-localization of TF binding sites, both within the same type and between different types of TFs.

We discovered two distinct co-localization groups of transcription factors. The first is formed from the interactions between more general, constitutive TFs and is enriched within the A1 chromosome sub-compartment. The second, formed from the interactions between cell-type specific TFs, is enriched within the A2 sub-compartment and is associated with significantly fewer conserved histone marks, when comparing lymphoblastoid and stem cell data. We called 40 pairs of significantly co-localized TFs according to the genome wide chromosome contact map, which were enriched

in previously reported, physically interacting TF pairs, thus validating our approach and linking transcription factor spatial distribution with function.

We also investigated TF binding site occupancy at a genome-wide scale in relation to the effect of spatial co-localisation of homotypic and heterotypic binding sites on TF binding dynamics and occupancy. The positive relationship between homotypic binding site co-localization and TF binding is confirmed by both BS occupancy analysis and ChIP-seq SignalValue comparison between carefully matched BSs. Particularly, in the case of homotypic BS clustering, enhancers and weak promoters witness stronger correlations and a larger magnitude of BS occupancy increase when they are found in spatial proximity. Moreover, BSs with weak DNA binding sequences show better correlation compared to those with strong binding sequences. Meanwhile, we built a computational simulation model (updated-FastGRiP) based on the facilitated diffusion mechanism to illustrate that homotypic BS clustering induced an occupancy boost. We observed a substantial occupancy gain when multiple BSs are close together in 3D, which is very sensitive to the distances between homotypic BSs located on different DNA strands and can also be strongly influenced by the number of homotypic BSs involved.

The relationship between binding site occupancy and heterotypic binding sites co-localization is complicated by the presence of the two TF co-localization networks and further, different chromatin states. It is clear that intra-TFCN BSs co-localization has similar effects as homotypic BS co-localization, which serves to enhance TF binding. However, co-localization of TFs associated with another network mostly has the opposing effect, indicating spatial networks segregation, with the exception that within the enhancer state, TFCN1 corresponds positively to TFCN2 BSs co-localization. Different TFs also exhibit different trends to certain degrees. The above complex landscape cannot be simply explained by biophysical models of TF diffusion and DNA association dynamics, but instead, points towards complicated TF-TF, TF-DNA interactions. To gain further insight, research into specific TFs are required, possibly using a combination of high-throughput *in vitro* protein-DNA binding assays [239, 51](on longer fragments of DNA instead of currently used short fragments) and *in vivo* knock-down of potential key pioneer TFs.

Meanwhile, I have also developed approaches to computationally estimate TF-DNA binding strength and enabled horizontal binding strength comparison across multiple TFs, which is a more basic research topic associated with TF-DNA interactions, but hard to solve efficiently using currently available software. Scoring DNA sequences against Position Weight Matrices (PWMs) is a widely adopted method to identify putative transcription factor BSs. While common bioinformatics tools produce scores that can reflect the binding strength between a specific transcription factor and the DNA, these scores are not directly comparable between different TFs. I developed two different ways to find the scaling parameter  $\lambda$  that allow us to infer binding energy from a simple PWM score. The first approach uses a PWM and background genomic sequence as input to estimate  $\lambda$  for a specific TF, which we applied to show that  $\lambda$  distributions for different TF families correspond with their DNA binding properties. The second method is able to reliably convert  $\lambda$  between different PWMs of the same TF, so that we can directly compare PWMs that were generated by different approaches or from different data bases. These two approaches provide consistent and computationally efficient ways to scale PWM scores and estimate TF binding sites strength.

## **5.2 Discussions and future directions: Hypothesis for chromosome domain and subcompartment formations**

Local chromosome folding into TADs can be reasonably well explained by various self-interaction models, including loop extrusion [112]. However, the global arrangement of chromosomes at a larger scale, i.e. from Mb inter-TADs contacts to chromosome territories and interfaces, are beyond the range that can be recapitulated directly by specific, small-scale interaction mechanisms only involving structural proteins like CTCF and cohesin. Recent study from Kubo *et al* observed that upon acute loss of CTCF, chromatin organization is almost preserved in mESC, except for Lamin associ-

ated domains [131]. Also, the existence of chromosome subcompartments such as A1 and A2 cannot be explained simply by the segregation of open and closed chromatin. Instead, TFs and other chromosome remodelling proteins can potentially provide a much larger pool for dynamic interactions, i.e. bring distal regulatory elements and promoters together [186][110], to generate more diverse chromosome contacts patterns. The insight of our TF binding profile analysis coincides with ideas of chromosome folding mechanisms provided by biophysical simulations taking into account distinct groups of chromosome bridging proteins. Given DNA bridging molecules are multivalent, studies have shown that bridging induced attraction can result in the segregation of chromosome domains and compartment-like structures [141]. Though no assumption of interactions between individual bridging proteins were specified, this study assumed the existence of different groups of protein complexes that can form stable bridges between distal chromosome segments, which can possibly involve physical interactions between TFs, chromosome re-modellers and mediator proteins in a cell-type specific manner.

We further showed that significant TF spatial co-localization pairs derived from Hi-C contact maps are enriched in known physical interactions between TFs. We hypothesize that physically interacting TFs might help to bring distal regulatory elements together, so that the co-regulated genes utilizing similar sets of physically interacting TFs co-localize together in nucleus, similar to the concept of transcription factories [186, 110]. In contrast to CTCF or cohesin, TF mediated chromosome contacts could be more cell type specific, as the abundance of different TFs vary greatly in different tissues. This might give rise to the distinct patterns of inter-TAD contacts across different cell types [122]. In addition, the spatially co-localized TFs we identified can also be viewed as potential candidate for TF-TF interaction, or at least, co-regulators for gene expression in specific cell types.

There is a long standing discussion about if it is the genome organisation that determines TF binding and histone marks, or if the binding of TF proteins that shapes genome organisation and histone marks. Benveniste *et al* demonstrated a good level of prediction of epigenetic marks from TF-binding profiles, which supported the view that binding of TF proteins to DNA determines the epigenetic state of chromosomes.

Our finding of robust TF co-localization network genome wide and also within each chromosome compartment or subcompartment suggests the existence of a TF spatial regulatory network across multiple scales, which might help to determine the genome structure. Furthermore, in GM12878, the opposite trend in TF occupancy differences between A1 and A2 chromosome subcompartment associated with TFCN1 and TFCN2 indicates a potential link between TF binding and chromosome subcompartment segregation. Even though our analysis does not provide a direct causal relationship between TF binding and the origin of chromosome compartment and subcompartment, it points towards the complex protein-protein, protein-DNA interaction shaping the genome organisation collaboratively.



## 5.3 List of Publications

Xiaoyan Ma, Daphne Ezer, Carmen Navarro and Boris Adryan. Reliable scaling of position weight matrices for binding strength comparisons between transcription factors. *BMC Bioinformatics*, <https://doi.org/10.1186/s12859-015-0666-1>, Aug 2015.

Xiaoyan Ma, Daphne Ezer, Boris Adryan and Tim J. Stevens. Hi-C contact maps reveal two distinct chromatin interaction networks of human transcription factors. *in prep.*

Justin Malin, Daphne Ezer, Xiaoyan Ma, Steve Mount, Hiren Karathia, Seung Gu Park, Boris Adryan, and Sridhar Hannenhalli. Crowdsourcing: Spatial clustering of low-affinity binding sites amplifies in vivo transcription factor occupancy. *BioRxiv*, Aug 2015

# Bibliography

- [1] Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.
- [2] M Carey. The enhanceosome and transcriptional synergy. *Cell*, 92(1):5–8, jan 1998.
- [3] D L Cook, A N Gerber, and S J Tapscott. Modeling stochastic gene expression: implications for haploinsufficiency. *Proc Natl Acad Sci U S A*, 95(26):15641–15646, dec 1998.
- [4] Bahram Houchmandzadeh, Eric Wieschaus, and Stanislas Leibler. Establishment of developmental precision and proportions in the early drosophila embryo. *Nature*, 415(6873):798–802, feb 2002.
- [5] Mark B Gerstein, Anshul Kundaje, Manoj Hariharan, Stephen G Landt, Koon-Kiu Yan, Chao Cheng, Xinmeng Jasmine Mu, Ekta Khurana, Joel Rozowsky, Roger Alexander, Renqiang Min, Pedro Alves, Alexej Abyzov, Nick Addleman, Nitin Bhardwaj, Alan P Boyle, Philip Cayting, Alexandra Charos, David Z Chen, Yong Cheng, Declan Clarke, Catharine Eastman, Ghia Euskirchen, Seth Fritze, Yao Fu, Jason Gertz, Fabian Grubert, Arif Harmanci, Preti Jain, Maya Kasowski, Phil Lacroute, Jing Leng, Jin Lian, Hannah Monahan, Henriette O’Geen, Zhengqing Ouyang, E Christopher Partridge, Dorrelyn Patacsil, Florencia Pauli, Debasish Raha, Lucia Ramirez, Timothy E Reddy, Brian Reed,

- Minyi Shi, Teri Slifer, Jing Wang, Linfeng Wu, Xinqiong Yang, Kevin Y Yip, Gili Zilberman-Schapira, Serafim Batzoglou, Arend Sidow, Peggy J Farnham, Richard M Myers, Sherman M Weissman, and Michael Snyder. Architecture of the human regulatory network derived from encode data. *Nature*, 489(7414):91–100, sep 2012.
- [6] Colin R Lickwar, Florian Mueller, Sean E Hanlon, James G McNally, and Jason D Lieb. Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature*, 484(7393):251–255, 2012.
- [7] Mike Levine. Transcriptional enhancers in animal development and evolution. *Curr Biol*, 20(17):R754–63, sep 2010.
- [8] Harold D Kim and Erin K O’Shea. A quantitative model of transcription factor-activated gene expression. *Nat Struct Mol Biol*, 15(11):1192–1198, nov 2008.
- [9] Shane Neph, Jeff Vierstra, Andrew B Stergachis, Alex P Reynolds, Eric Haugen, Benjamin Vernot, Robert E Thurman, Sam John, Richard Sandstrom, Audra K Johnson, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90, 2012.
- [10] Eva Nogales, Robert K Louder, and Yuan He. Structural insights into the eukaryotic transcription initiation machinery. *Annu Rev Biophys*, 46:59–83, may 2017.
- [11] Merle Hantsche and Patrick Cramer. Conserved rna polymerase ii initiation complex structure. *Curr Opin Struct Biol*, 47:17–22, apr 2017.
- [12] Benjamin P Berman, Yutaka Nibu, Barret D Pfeiffer, Pavel Tomancak, Susan E Celniker, Michael Levine, Gerald M Rubin, and Michael B Eisen. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the drosophila genome. *Proc Natl Acad Sci U S A*, 99(2):757–762, jan 2002.

- [13] Bart Deplancke. Experimental advances in the characterization of metazoan gene regulatory networks. *Brief Funct Genomic Proteomic*, 8(1):12–27, jan 2009.
- [14] Matthew T Weirauch, Ally Yang, Mihai Albu, Atina G Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S Najafabadi, Samuel A Lambert, Ishminder Mann, Kate Cook, Hong Zheng, Alejandra Goity, Harm van Bakel, Jean-Claude Lozano, Mary Galli, Mathew G Lewsey, Eryong Huang, Tuhin Mukherjee, Xiaoting Chen, John S Reece-Hoyes, Sridhar Govindarajan, Gad Shaulsky, Albertha J M Walhout, Francois-Yves Bouget, Gunnar Ratsch, Luis F Larrondo, Joseph R Ecker, and Timothy R Hughes. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6):1431–1443, sep 2014.
- [15] Alexander E Kel, Ellen Gößling, Ingmar Reuter, Evgeny Cheremushkin, Olga V Kel-Margoulis, and Edgar Wingender. Matchtm: a tool for searching transcription factor binding sites in dna sequences. *Nucleic acids research*, 31(13):3576–3579, 2003.
- [16] Meghana M Kulkarni and David N Arnosti. Information display by transcriptional enhancers. *Development*, 130(26):6569–6575, 2003.
- [17] Srinivas Veerla, Markus Ringnr, and Mattias Hglund. Genome-wide transcription factor binding site/promoter databases for the analysis of gene sets and co-occurrence of transcription factor binding motifs. *BMC Genomics*, 11:145, mar 2010.
- [18] Holger Klein and Martin Vingron. Using transcription factor binding site co-occurrence to predict regulatory regions. *Genome Inform*, 18:109–118, 2007.
- [19] Tommy Kaplan, Xiao-Yong Li, Peter J Sabo, Sean Thomas, John A Stamatoyannopoulos, Mark D Biggin, and Michael B Eisen. Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early drosophila development. *PLoS Genet*, 7(2):e1001290, feb 2011.

- [20] Olivier Crauk and Nathalie Dostatni. Bicoid determines sharp and precise target gene expression in the drosophila embryo. *Curr Biol*, 15(21):1888–1898, nov 2005.
- [21] Shane Neph, Jeff Vierstra, Andrew B Stergachis, Alex P Reynolds, Eric Haugen, Benjamin Vernot, Robert E Thurman, Sam John, Richard Sandstrom, Audra K Johnson, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90, 2012.
- [22] Nicolae Radu Zabet and Boris Adryan. Computational models for large-scale simulations of facilitated diffusion. *Molecular BioSystems*, 8(11):2815–2827, 2012.
- [23] Xin He, Md Abul Hassan Samee, Charles Blatti, and Saurabh Sinha. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. *PLoS Comput Biol*, 6(9), sep 2010.
- [24] Md Abul Hassan Samee, Bomyi Lim, Nria Samper, Hang Lu, Christine A Rushlow, Gerardo Jimnez, Stanislav Y Shvartsman, and Saurabh Sinha. A systematic ensemble approach to thermodynamic modeling of gene expression from sequence data. *Cell Syst*, 1(6):396–407, dec 2015.
- [25] Ah-Ram Kim, Carlos Martinez, John Ionides, Alexandre F Ramos, Michael Z Ludwig, Nobuo Ogawa, David H Sharp, and John Reinitz. Rearrangements of 2.5 kilobases of noncoding DNA from the drosophila even-skipped locus define predictive rules of genomic cis-regulatory logic. *PLoS Genetics*, 9(2):e1003243, 2013.
- [26] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, jun 2007.
- [27] M J Solomon, P L Larsen, and A Varshavsky. Mapping protein-dna interactions in vivo with formaldehyde: evidence that histone h4 is retained on a highly transcribed gene. *Cell*, 53(6):937–947, jun 1988.

- [28] Myong-Hee Sung, Michael J Guertin, Songjoon Baek, and Gordon L Hager. DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Molecular Cell*, 56(2):275–285, 2014.
- [29] Terrence S Furey. Chip-seq and beyond: new and improved methodologies to detect and characterize protein-dna interactions. *Nat Rev Genet*, 13(12):840–852, dec 2012.
- [30] Y Blat and N Kleckner. Cohesins bind to preferential sites along yeast chromosome iii, with differential regulation along arms versus the centric region. *Cell*, 98(2):249–259, jul 1999.
- [31] Luis G Acevedo, A Leonardo Iniguez, Heather L Holster, Xinmin Zhang, Roland Green, and Peggy J Farnham. Genome-scale chip-chip analysis using 10,000 human cells. *BioTechniques*, 43(6):791–797, dec 2007.
- [32] Gordon Robertson, Martin Hirst, Matthew Bainbridge, Misha Bilenky, Yongjun Zhao, Thomas Zeng, Ghia Euskirchen, Bridget Bernier, Richard Varhol, Allen Delaney, Nina Thiessen, Obi L Griffith, Ann He, Marco Marra, Michael Snyder, and Steven Jones. Genome-wide profiles of stat1 dna association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*, 4(8):651–657, aug 2007.
- [33] Anton Valouev, David S Johnson, Andreas Sundquist, Catherine Medina, Elizabeth Anton, Serafim Batzoglou, Richard M Myers, and Arend Sidow. Genome-wide analysis of transcription factor binding sites based on chip-seq data. *Nat Methods*, 5(9):829–834, sep 2008.
- [34] Stephen G Landt, Georgi K Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E Bernstein, Peter Bickel, James B Brown, Philip Cayting, Yiwen Chen, Gilberto DeSalvo, Charles Epstein, Katherine I Fisher-Aylor, Ghia Euskirchen, Mark Gerstein, Jason Gertz, Alexander J Hartemink, Michael M Hoffman, Vishwanath R Iyer, Youngsook L Jung, Subhradip Karmakar, Manolis Kellis, Peter V Kharchenko, Qunhua Li,

- Tao Liu, X Shirley Liu, Lijia Ma, Aleksandar Milosavljevic, Richard M Myers, Peter J Park, Michael J Pazin, Marc D Perry, Debasish Raha, Timothy E Reddy, Joel Rozowsky, Noam Shores, Arend Sidow, Matthew Slattery, John A Stamatoyannopoulos, Michael Y Tolstorukov, Kevin P White, Simon Xi, Peggy J Farnham, Jason D Lieb, Barbara J Wold, and Michael Snyder. Chip-seq guidelines and practices of the encode and modencode consortia. *Genome Res*, 22(9):1813–1831, sep 2012.
- [35] Peter J Park. Chip-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, 10(10):669–680, oct 2009.
- [36] Peter V Kharchenko, Michael Y Tolstorukov, and Peter J Park. Design and analysis of chip-seq experiments for dna-binding proteins. *Nat Biotechnol*, 26(12):1351–1359, dec 2008.
- [37] Ruiqiang Li, Yingrui Li, Karsten Kristiansen, and Jun Wang. Soap: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–714, mar 2008.
- [38] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25, mar 2009.
- [39] Joel Rozowsky, Ghia Euskirchen, Raymond K Auerbach, Zhengdong D Zhang, Theodore Gibson, Robert Bjornson, Nicholas Carriero, Michael Snyder, and Mark B Gerstein. Peakseq enables systematic scoring of chip-seq experiments relative to controls. *Nat Biotechnol*, 27(1):66–75, jan 2009.
- [40] Yong Zhang, Tao Liu, Clifford A Meyer, Jrme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. Model-based analysis of chip-seq (macs). *Genome Biol*, 9(9):R137, sep 2008.
- [41] David A Nix, Samir J Courdy, and Kenneth M Boucher. Empirical methods for controlling false positives and estimating confidence in chip-seq peaks. *BMC Bioinformatics*, 9:523, dec 2008.

- [42] Timothy L Bailey, Nadya Williams, Chris Misleh, and Wilfred W Li. Meme: discovering and analyzing dna and protein sequence motifs. *Nucleic Acids Res*, 34(Web Server issue):W369–73, jul 2006.
- [43] Gregory E Crawford, Ingeborg E Holt, James Whittle, Bryn D Webb, Denise Tai, Sean Davis, Elliott H Margulies, YiDong Chen, John A Bernat, David Ginsburg, Daixing Zhou, Shujun Luo, Thomas J Vasicek, Mark J Daly, Tyra G Wolfsberg, and Francis S Collins. Genome-wide mapping of dnase hypersensitive sites using massively parallel signature sequencing (mpß). *Genome Res*, 16(1):123–131, jan 2006.
- [44] Housheng Hansen He, Clifford A Meyer, Sheng'en Shawn Hu, Mei-Wei Chen, Chongzhi Zang, Yin Liu, Prakash K Rao, Teng Fei, Han Xu, Henry Long, X Shirley Liu, and Myles Brown. Refined dnase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat Methods*, 11(1):73–78, jan 2014.
- [45] Eduardo G Gusmao, Manuel Allhoff, Martin Zenke, and Ivan G Costa. Analysis of computational footprinting methods for dnase sequencing experiments. *Nat Methods*, 13(4):303–309, apr 2016.
- [46] S Small, A Blair, and M Levine. Regulation of even-skipped stripe 2 in the drosophila embryo. *EMBO J*, 11(11):4047–4057, nov 1992.
- [47] Melissa M Harrison, Xiao-Yong Li, Tommy Kaplan, Michael R Botchan, and Michael B Eisen. Zelda binding in the early drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition. *PLoS genetics*, 7(10):e1002266, 2011.
- [48] Arttu Jolma, Teemu Kivioja, Jarkko Toivonen, Lu Cheng, Gonghong Wei, Martin Enge, Mikko Taipale, Juan M Vaquerizas, Jian Yan, Mikko J Sillanpää, et al. Multiplexed massively parallel selex for characterization of human transcription factor binding specificities. *Genome research*, 20(6):861–873, 2010.
- [49] Emmanuelle Roulet, Stéphane Busso, Anamaria A Camargo, Andrew JG Simpson, Nicolas Mermoud, and Philipp Bucher. High-throughput selex–sage method



for quantitative modeling of transcription-factor binding sites. *Nature biotechnology*, 20(8):831–835, 2002.

- [50] Sonali Mukherjee, Michael F Berger, Ghil Jona, Xun S Wang, Dale Muzzey, Michael Snyder, Richard A Young, and Martha L Bulyk. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nature Genetics*, 36(12):1331–1339, 2004.
- [51] Arttu Jolma, Jian Yan, Thomas Whittington, Jarkko Toivonen, Kazuhiro R Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, Kimmo Palin, Juan M Vaquerizas, Renaud Vincentelli, Nicholas M Luscombe, Timothy R Hughes, Patrick Lemaire, Esko Ukkonen, Teemu Kivioja, and Jussi Taipale. Dna-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339, jan 2013.
- [52] Korneel Hens, Jean-Daniel Feuz, and Bart Deplancke. A high-throughput gateway-compatible yeast one-hybrid screen to detect protein-dna interactions. *Methods Mol Biol*, 786:335–355, 2012.
- [53] Carine Gubelmann, Sebastian M Waszak, Alina Isakova, Wiebke Holcombe, Korneel Hens, Antonina Iagovitina, Jean-Daniel Feuz, Sunil K Raghav, Jovan Simicevic, and Bart Deplancke. A yeast one-hybrid and microfluidics-based pipeline to map mammalian gene regulatory networks. *Mol Syst Biol*, 9:682, aug 2013.
- [54] Elodie Portales-Casamar, David Arenillas, Jonathan Lim, Magdalena I Swanson, Steven Jiang, Anthony McCallum, Stefan Kirov, and Wyeth W Wasserman. The pazar database of gene regulatory information coupled to the orca toolkit for the study of regulatory sequences. *Nucleic Acids Res*, 37(Database issue):D54–60, jan 2009.
- [55] Steven M Gallo, Dave T Gerrard, David Miner, Michael Simich, Benjamin Des Soye, Casey M Bergman, and Marc S Halfon. Redfly v3.0: toward a comprehensive database of transcriptional regulatory elements in drosophila. *Nucleic Acids Res*, 39(Database issue):D118–23, jan 2011.

- [56] Otto G Berg and Peter H von Hippel. Selection of DNA binding sites by regulatory proteins statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol*, 193:723–750, 1987.
- [57] Otto G Berg and Peter H von Hippel. Selection of DNA binding sites by regulatory proteins. *Trends in Biochemical Sciences*, 13(6):207–211, 1988.
- [58] G D Stormo. Dna binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, jan 2000.
- [59] O G Berg and P H von Hippel. Selection of dna binding sites by regulatory proteins. *Trends Biochem Sci*, 13(6):207–211, jun 1988.
- [60] Wyeth W Wasserman and Albin Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4):276–287, 2004.
- [61] Sridhar Hannenhalli. Eukaryotic transcription factor binding sites—modeling and integrative search methods. *Bioinformatics*, 24(11):1325–1331, jun 2008.
- [62] Pei-Chen Peng and Saurabh Sinha. Quantitative modeling of gene expression using dna shape features of binding sites. *Nucleic Acids Res*, 44(13):e120, jul 2016.
- [63] Sebastian J Maerkl and Stephen R Quake. A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315(5809):233–237, 2007.
- [64] Gary D Stormo and Yue Zhao. Putting numbers on the network connections. *BioEssays*, 29(8):717–721, 2007.
- [65] Justin Crocker, Ella Preger-Ben Noon, and David L Stern. The soft touch: Low-affinity transcription factor binding sites in development and evolution. *Curr Top Dev Biol*, 117:455–469, jan 2016.
- [66] Jan Grau, Ivo Grosse, Stefan Posch, and Jens Keilwagen. Motif clustering with implications for transcription factor interactions. aug 2015.

- [67] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.
- [68] Zhaohui S Qin, Lee Ann McCue, William Thompson, Linda Mayerhofer, Charles E Lawrence, and Jun S Liu. Identification of co-regulated genes through bayesian clustering of predicted regulatory binding sites. *Nat Biotechnol*, 21(4):435–439, apr 2003.
- [69] Alexander P Lifanov, Vsevolod J Makeev, Anna G Nazina, and Dmitri A Papatsenko. Homotypic regulatory clusters in drosophila. *Genome Res*, 13(4):579–588, apr 2003.
- [70] Valer Gotea, Axel Visel, John M Westlund, Marcelo A Nobrega, Len A Pennacchio, and Ivan Ovcharenko. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res*, 20(5):565–577, may 2010.
- [71] J L Johnson and A McLachlan. Novel clustering of sp1 transcription factor binding sites at the transcription initiation site of the human muscle phosphofructokinase p1 promoter. *Nucleic Acids Res*, 22(23):5085–5092, nov 1994.
- [72] Xin He, Thyago S P C Duque, and Saurabh Sinha. Evolutionary origins of transcription factor binding site clusters. *Mol Biol Evol*, 29(3):1059–1070, mar 2012.
- [73] Chaolin Zhang, Zhenyu Xuan, Stefanie Otto, John R Hover, Sean R McCorkle, Gail Mandel, and Michael Q Zhang. A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. *Nucleic Acids Res*, 34(8):2238–2246, may 2006.
- [74] CA Brackley, ME Cates, and D Marenduzzo. Facilitated diffusion on mobile dna: configurational traps and sequence heterogeneity. *Physical review letters*, 109(16):168103, 2012.

- [75] Garth R Ilsley, Jasmin Fisher, Rolf Apweiler, Angela H De Pace, and Nicholas M Luscombe. Cellular resolution models for even skipped regulation in the entire drosophila embryo. *elife*, 2:e00522, aug 2013.
- [76] Johan Elf, Gene-Wei Li, and X Sunney Xie. Probing transcription factor dynamics at the single-molecule level in a living cell. *Science*, 316(5828):1191–1194, 2007.
- [77] Nicolae Radu Zabet and Boris Adryan. Estimating binding properties of transcription factors from genome-wide binding profiles. *Nucleic Acids Res*, 43(1):84–94, jan 2015.
- [78] Arthur D. Riggs, Suzanne Bourgeois, and Melvin Cohn. The lac repressor-operator interaction. *J Mol Biol*, 53(3):401–417, nov 1970.
- [79] R B Winter, O G Berg, and P H von Hippel. Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. the escherichia coli lac repressor-operator interaction: kinetic measurements and conclusions. *Biochemistry*, 20(24):6961–6977, nov 1981.
- [80] Otto G Berg, Robert B Winter, and Peter H Von Hippel. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. models and theory. *Biochemistry*, 20(24):6929–6948, 1981.
- [81] Jason S Leith, Anahita Tafvizi, Fang Huang, William E Uspal, Patrick S Doyle, Alan R Fersht, Leonid A Mirny, and Antoine M van Oijen. Sequence-dependent sliding kinetics of p53. *Proceedings of the National Academy of Sciences*, 109(41):16552–16557, 2012.
- [82] J G Kim, Y Takeda, B W Matthews, and W F Anderson. Kinetic studies on cro repressor-operator dna interaction. *J Mol Biol*, 196(1):149–158, jul 1987.
- [83] N Shimamoto. One-dimensional diffusion of proteins along dna. its biological and chemical significance revealed by single-molecule measurements. *J Biol Chem*, 274(22):15293–15296, may 1999.

- [84] Jiji Chen, Zhengjian Zhang, Li Li, Bi-Chang Chen, Andrey Revyakin, Bassam Hajj, Wesley Legant, Maxime Dahan, Timothée Lionnet, Eric Betzig, et al. Single-molecule dynamics of enhanceosome assembly in embryonic stem cells. *Cell*, 156(6):1274–1285, 2014.
- [85] Florian Mueller, Timothy J Stasevich, Davide Mazza, and James G McNally. Quantifying transcription factor kinetics: At work or at play? *Critical reviews in biochemistry and molecular biology*, 48(5):492–514, 2013.
- [86] Stephen E Halford and John F Marko. How do site-specific dna-binding proteins find their targets? *Nucleic Acids Res*, 32(10):3040–3052, jun 2004.
- [87] Ana-Maria Florescu and Marc Joyeux. Description of nonspecific dna-protein interaction and facilitated diffusion with a dynamical model. *J Chem Phys*, 130(1):015103, jan 2009.
- [88] Ulrich Gerland, J David Moroz, and Terence Hwa. Physical constraints and functional characteristics of transcription factor–dna interaction. *Proceedings of the National Academy of Sciences*, 99(19):12015–12020, 2002.
- [89] C Loverdo, Olivier Benichou, Raphael Voituriez, A Biebricher, I Bonnet, and P Desbiolles. Quantifying hopping and jumping in facilitated diffusion of dna-binding proteins. *Physical review letters*, 102(18):188101, 2009.
- [90] Nicolae Radu Zabet and Boris Adryan. A comprehensive computational model to simulate transcription factor binding in prokaryotes. In *Information Processing in Cells and Tissues*, volume 7223 of *Lecture Notes in Computer Science*, pages 35–37. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
- [91] CA Brackley, ME Cates, and D Marenduzzo. Intracellular facilitated diffusion: searchers, crowders, and blockers. *Physical Review Letters*, 111(10):108101, 2013.
- [92] Chris A Brackley, Mike E Cates, and Davide Marenduzzo. Effect of dna conformation on facilitated diffusion. *Biochem Soc Trans*, 41(2):582–588, apr 2013.

- [93] G. Foffano, D. Marenduzzo, and E. Orlandini. Facilitated diffusion on confined dna. *Phys. Rev. E*, 85(2), feb 2012.
- [94] Daphne Ezer, Nicolae Radu Zabet, and Boris Adryan. Physical constraints determine the logic of bacterial promoter architectures. *Nucleic Acids Research*, page gku078, 2014.
- [95] Eilon Sharon, David van Dijk, Yael Kalma, Leeat Keren, Ohad Manor, Zohar Yakhini, and Eran Segal. Probing the effect of promoters on noise in gene expression using thousands of designed sequences. *Genome Res*, 24(10):1698–1706, oct 2014.
- [96] Gyrgy Vmosi, Sndor Damjanovich, Jnos Szllosi, and Gyrgy Vereb. Measurement of molecular mobility with fluorescence correlation spectroscopy. *Curr Protoc Cytom*, Chapter 2:Unit2.15, oct 2009.
- [97] Michael Carnell, Alex Macmillan, and Renee Whan. Fluorescence recovery after photobleaching (frap): acquisition, analysis, and applications. *Methods Mol Biol*, 1232:255–271, 2015.
- [98] Tatsuya Morisaki and James G McNally. Photoswitching-free frap analysis with a genetically encoded fluorescent tag. *PLoS ONE*, 9(9):e107730, sep 2014.
- [99] Davide Mazza, Alice Abernathy, Nicole Golob, Tatsuya Morisaki, and James G McNally. A benchmark for chromatin binding measurements in live cells. *Nucleic Acids Res*, 40(15):e119, aug 2012.
- [100] Agata Pernu and Jrg Langowski. Imaging fos-jun transcription factor mobility and interaction in live cells by single plane illumination-fluorescence cross correlation spectroscopy. *PLoS ONE*, 10(4):e0123070, apr 2015.
- [101] Noriyuki Sugo, Masatoshi Morimatsu, Yoshiyuki Arai, Yoshinori Kousoku, Aya Ohkuni, Taishin Nomura, Toshio Yanagida, and Nobuhiko Yamamoto. Single-molecule imaging reveals dynamics of creb transcription factor bound to its target sequence. *Sci Rep*, 5:10662, jun 2015.

- [102] Tatsuya Morisaki, Waltraud G Mller, Nicole Golob, Davide Mazza, and James G McNally. Single-molecule analysis of transcription factor binding at transcription sites in live cells. *Nat Commun*, 5:4456, jul 2014.
- [103] Peter Hinow, Carl E Rogers, Christopher E Barbieri, Jennifer A Pietenpol, Anne K Kenworthy, and Emmanuele DiBenedetto. The dna binding activity of p53 displays reaction-diffusion kinetics. *Biophys J*, 91(1):330–342, jul 2006.
- [104] Sylvain D Ethier, Hisashi Miura, and Jose Dostie. Discovering genome regulation with 3c and 3c-related technologies. *Biochim Biophys Acta*, 1819(5):401–410, may 2012.
- [105] Zhihu Zhao, Gholamreza Tavoosidana, Mikael Sjlander, Anita Gndr, Piero Mariano, Sha Wang, Chandrasekhar Kanduri, Magda Lezcano, Kuljeet Singh Sandhu, Umashankar Singh, Vinod Pant, Vijay Tiwari, Sreenivasulu Kurukuti, and Rolf Ohlsson. Circular chromosome conformation capture (4c) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet*, 38(11):1341–1347, nov 2006.
- [106] Jose Dostie, Todd A Richmond, Ramy A Arnaout, Rebecca R Selzer, William L Lee, Tracey A Honan, Eric D Rubio, Anton Krumm, Justin Lamb, Chad Nusbaum, Roland D Green, and Job Dekker. Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome Res*, 16(10):1299–1309, oct 2006.
- [107] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragozy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293, 2009.
- [108] Zhijun Duan, Mirela Andronescu, Kevin Schutz, Sean McIlwain, Yoo Jung Kim, Choli Lee, Jay Shendure, Stanley Fields, C Anthony Blau, and William S Noble. A three-dimensional model of the yeast genome. *Nature*, 465(7296):363–367, may 2010.

- [109] Ming Hu, Ke Deng, Siddarth Selvaraj, Zhaohui Qin, Bing Ren, and Jun S Liu. Hicnorm: removing biases in hi-c data via poisson regression. *Bioinformatics*, 28(23):3131–3133, dec 2012.
- [110] Eitan Yaffe and Amos Tanay. Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet*, 43(11):1059–1065, oct 2011.
- [111] Ilya M Flyamer, Johanna Gassler, Maxim Imakaev, Hugo B Brando, Sergey V Ulianov, Nezar Abdennur, Sergey V Razin, Leonid A Mirny, and Kiku Tachibana-Konwalski. Single-nucleus hi-c reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature*, 544(7648):110–114, apr 2017.
- [112] Adrian L Sanborn, Suhas S P Rao, Su-Chen Huang, Neva C Durand, Miriam H Huntley, Andrew I Jewett, Ivan D Bochkov, Dharmaraj Chinnappan, Ashok Cutkosky, Jian Li, Kristopher P Geeting, Andreas Gnirke, Alexandre Melnikov, Doug McKenna, Elena K Stamenova, Eric S Lander, and Erez Lieberman Aiden. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci U S A*, 112(47):E6456–65, nov 2015.
- [113] Robert A Beagrie, Antonio Scialdone, Markus Schueler, Dorothee C A Kraemer, Mita Chotalia, Sheila Q Xie, Mariano Barbieri, Ins de Santiago, Liron-Mark Lavitas, Miguel R Branco, James Fraser, Jose Dostie, Laurence Game, Niall Dillon, Paul A W Edwards, Mario Nicodemi, and Ana Pombo. Complex multi-enhancer contacts captured by genome architecture mapping. *Nature*, 543(7646):519–524, mar 2017.
- [114] Michael E G Sauria, Jennifer E Phillips-Cremins, Victor G Corces, and James Taylor. Hifive: a tool suite for easy and efficient hic and 5c data analysis. *Genome Biol*, 16:237, oct 2015.
- [115] Charalampos Lazaris, Stephen Kelly, Panagiotis Ntziachristos, Iannis Aifantis, and Aristotelis Tsirigos. Hic-bench: comprehensive and reproducible hi-c



- data analysis designed for parameter exploration and benchmarking. *BMC Genomics*, 18(1):22, jan 2017.
- [116] Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R Lajoie, Job Dekker, and Leonid A Mirny. Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nat Methods*, 9(10):999–1003, oct 2012.
  - [117] P. A. Knight and D. Ruiz. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis*, 33(3):1029–1047, jul 2013.
  - [118] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.*, 21(2):343–348, may 1967.
  - [119] Job Dekker, Marc A Marti-Renom, and Leonid A Mirny. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet*, 14(6):390–403, jun 2013.
  - [120] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, apr 2012.
  - [121] Emily M Smith, Bryan R Lajoie, Gaurav Jain, and Job Dekker. Invariant tad boundaries constrain cell-type-specific looping interactions between promoters and distal elements around the cfr locus. *Am J Hum Genet*, 98(1):185–201, jan 2016.
  - [122] James Fraser, Carmelo Ferrai, Andrea M Chiariello, Markus Schueler, Tiago Rito, Giovanni Laudanno, Mariano Barbieri, Benjamin L Moore, Dorothee C A Kraemer, Stuart Aitken, Sheila Q Xie, Kelly J Morris, Masayoshi Itoh, Hideya Kawaji, Ines Jaeger, Yoshihide Hayashizaki, Piero Carninci, Alistair R R Forrest, FANTOM Consortium, Colin A Semple, Jose Dostie, Ana Pombo, and Mario Nicodemi. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol Syst Biol*, 11(12):852, dec 2015.

- [123] Elphge P Nora, Bryan R Lajoie, Edda G Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L van Berkum, Johannes Meisig, John Sedat, Joost Gribnau, Emmanuel Barillot, Nils Blthgen, Job Dekker, and Edith Heard. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, 485(7398):381–385, apr 2012.
- [124] Daro G Lupiez, Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hlya Kayserili, John M Opitz, Renata Laxova, Fernando Santos-Simarro, Brigitte Gilbert-Dussardier, Lars Wittler, Marina Borschiwer, Stefan A Haas, Marco Osterwalder, Martin Franke, Bernd Timmermann, Jochen Hecht, Malte Spielmann, Axel Visel, and Stefan Mundlos. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025, may 2015.
- [125] Orsolya Symmons, Veli Vural Uslu, Taro Tsujimura, Sandra Ruf, Sonya Nasari, Wibke Schwarzer, Laurence Ettwiller, and Franois Spitz. Functional and topological characteristics of mammalian regulatory domains. *Genome Res*, 24(3):390–400, mar 2014.
- [126] Geoffrey Fudenberg and Leonid A Mirny. Higher-order chromatin structure: bridging physics and biology. *Curr Opin Genet Dev*, 22(2):115–124, apr 2012.
- [127] Kim Nasmyth and Christian H Haering. Cohesin: its roles and mechanisms. *Annu Rev Genet*, 43:525–558, 2009.
- [128] Elzo de Wit, Erica S M Vos, Sjoerd J B Holwerda, Christian Valdes-Quezada, Marjon J A M Verstegen, Hans Teunissen, Erik Splinter, Patrick J Wijchers, Peter H L Krijger, and Wouter de Laat. Ctf binding polarity determines chromatin looping. *Mol Cell*, 60(4):676–684, nov 2015.
- [129] Armelle Lengronne, Yuki Katou, Saori Mori, Shihori Yokobayashi, Gavin P Kelly, Takehiko Itoh, Yoshinori Watanabe, Katsuhiko Shirahige, and Frank Uhlmann. Cohesin relocation from sites of chromosomal loading to places of convergent transcription. *Nature*, 430(6999):573–578, jul 2004.

- [130] HyeonJun Kim and Joseph J Loparo. Multistep assembly of dna condensation clusters by smc. *Nat Commun*, 7:10200, jan 2016.
- [131] Naoki Kubo, Haruhiko Ishii, David Gorkin, Franz Meitinger, Xiong Xiong, Rongxin Fang, Tristin Liu, Zhen Ye, Bin Li, Jesse Dixon, Arshad Desai, Huimin Zhao, and Bing Ren. Preservation of chromatin organization after acute loss of ctfc in mouse embryonic stem cells. *bioRxiv*, 2017.
- [132] Mariano Barbieri, Mita Chotalia, James Fraser, Liron-Mark Lavitas, Jose Dostie, Ana Pombo, and Mario Nicodemi. Complexity of chromatin folding is captured by the strings and binders switch model. *Proc Natl Acad Sci U S A*, 109(40):16173–16178, oct 2012.
- [133] Catherine Naughton, Nicolaos Avlonitis, Samuel Corless, James G Prendergast, Ioulia K Mati, Paul P Eijk, Scott L Cockcroft, Mark Bradley, Bauke Ylstra, and Nick Gilbert. Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. *Nat Struct Mol Biol*, 20(3):387–395, mar 2013.
- [134] Samuel Corless and Nick Gilbert. Effects of dna supercoiling on chromatin architecture. *Biophys Rev*, 8(3):245–258, jul 2016.
- [135] K Havas, A Flaus, M Phelan, R Kingston, P A Wade, D M Lilley, and T Owen-Hughes. Generation of superhelical torsion by atp-dependent chromatin remodeling activities. *Cell*, 103(7):1133–1142, dec 2000.
- [136] Alistair N Boettiger, Bogdan Bintu, Jeffrey R Moffitt, Siyuan Wang, Brian J Beliveau, Geoffrey Fudenberg, Maxim Imakaev, Leonid A Mirny, Chao-ting Wu, and Xiaowei Zhuang. Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature*, 529(7586):418–422, jan 2016.
- [137] Fabrizio Benedetti, Julien Dorier, and Andrzej Stasiak. Effects of supercoiling on enhancer-promoter contacts. *Nucleic Acids Res*, 42(16):10425–10432, aug 2014.
- [138] A Vologodskii and N R Cozzarelli. Effect of supercoiling on the juxtaposition and relative orientation of dna sites. *Biophys J*, 70(6):2548–2556, jun 1996.

- [139] Julio Mateos-Langerak, Sandra Goetze, Heinrich Leonhardt, Thomas Cremer, Roel van Driel, and Christian Lanctt. Nuclear architecture: Is it important for genome function and can we prove it? *J Cell Biochem*, 102(5):1067–1075, dec 2007.
- [140] Tim J Stevens, David Lando, Srinjan Basu, Liam P Atkinson, Yang Cao, Steven F Lee, Martin Leeb, Kai J Wohlfahrt, Wayne Boucher, Aoife O’Shaughnessy-Kirwan, Julie Cramard, Andre J Faure, Meryem Ralser, Enrique Blanco, Lluís Morey, Miriam Sans, Matthieu G S Palayret, Ben Lehner, Luciano Di Croce, Anton Wutz, Brian Hendrich, Dave Klenerman, and Ernest D Laue. 3d structures of individual mammalian genomes studied by single-cell hi-c. *Nature*, 544(7648):59–64, apr 2017.
- [141] Chris A Brackley, James Johnson, Steven Kelly, Peter R Cook, and Davide Marenduzzo. Simulated binding of transcription factors to active and inactive regions folds human chromosomes into loops, rosettes and topological domains. *Nucleic Acids Res*, 44(8):3503–3512, may 2016.
- [142] Glenn A Maston, Sara K Evans, and Michael R Green. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet*, 7:29–59, 2006.
- [143] Aaron Arvey, Phaedra Agius, William Stafford Noble, and Christina Leslie. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res*, 22(9):1723–1734, sep 2012.
- [144] Robert E Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T Maurano, Eric Haugen, Nathan C Sheffield, Andrew B Stergachis, Hao Wang, Benjamin Vernot, Kavita Garg, Sam John, Richard Sandstrom, Daniel Bates, Lisa Boatman, Theresa K Canfield, Morgan Diegel, Douglas Dunn, Abigail K Ebersol, Tristan Frum, Erika Giste, Audra K Johnson, Ericka M Johnson, Tanya Kuttyavin, Bryan Lajoie, Bum-Kyu Lee, Kristen Lee, Darin London, Dimitra Lotakis, Shane Neph, Fidencio Neri, Eric D Nguyen, Hongzhu Qu, Alex P Reynolds, Vaughn Roach, Alexias Safi, Minerva E Sanchez, Amartya Sanyal,

Anthony Shafer, Jeremy M Simon, Lingyun Song, Shinny Vong, Molly Weaver, Yongqi Yan, Zhancheng Zhang, Zhuzhu Zhang, Boris Lenhard, Muneesh Tewari, Michael O Dorschner, R Scott Hansen, Patrick A Navas, George Stamatoyannopoulos, Vishwanath R Iyer, Jason D Lieb, Shamil R Sunyaev, Joshua M Akey, Peter J Sabo, Rajinder Kaul, Terrence S Furey, Job Dekker, Gregory E Crawford, and John A Stamatoyannopoulos. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, sep 2012.

- [145] Eduardo G Gusmao, Christoph Dieterich, Martin Zenke, and Ivan G Costa. Detection of active transcription factor binding sites with the combination of dnase hypersensitivity and histone modifications. *Bioinformatics*, 30(22):3143–3151, nov 2014.
- [146] Eugen Fazius, Vladimir Shelest, and Ekaterina Shelest. Sitar: a novel tool for transcription factor binding site prediction. *Bioinformatics*, 27(20):2806–2811, oct 2011.
- [147] Dmitry Y Oshchepkov and Victor G Levitsky. In silico prediction of transcriptional factor-binding sites. *Methods Mol Biol*, 760:251–267, 2011.
- [148] Tom Whittington, Andrew C Perkins, and Timothy L Bailey. High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. *Nucleic Acids Res*, 37(1):14–25, jan 2009.
- [149] Alan P Boyle, Lingyun Song, Bum-Kyu Lee, Darin London, Damian Keefe, Ewan Birney, Vishwanath R Iyer, Gregory E Crawford, and Terrence S Furey. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res*, 21(3):456–464, mar 2011.
- [150] Simon C Biddie, Sam John, Pete J Sabo, Robert E Thurman, Thomas A Johnson, R Louis Schiltz, Tina B Miranda, Myong-Hee Sung, Saskia Trump, Stafford L Lightman, Charles Vinson, John A Stamatoyannopoulos, and Gordon L Hager. Transcription factor ap1 potentiates chromatin accessibility and glucocorticoid receptor binding. *Mol Cell*, 43(1):145–155, jul 2011.

- [151] Siming Li, Christopher M Armstrong, Nicolas Bertin, Hui Ge, Stuart Milstein, Mike Boxem, Pierre-Olivier Vidalain, Jing-Dong J Han, Alban Chesneau, Tong Hao, Debra S Goldberg, Ning Li, Monica Martinez, Jean-Francois Rual, Philippe Lamesch, Lai Xu, Muneesh Tewari, Sharyl L Wong, Lan V Zhang, Gabriel F Berriz, Laurent Jacotot, Philippe Vaglio, Jrme Reboul, Tomoko Hirozane-Kishikawa, Qianru Li, Harrison W Gabel, Ahmed Elewa, Bridget Baumgartner, Debra J Rose, Haiyuan Yu, Stephanie Bosak, Reynaldo Sequerra, Andrew Fraser, Susan E Mango, William M Saxton, Susan Strome, Sander Van Den Heuvel, Fabio Piano, Jean Vandenhaute, Claude Sardet, Mark Gerstein, Lynn Doucette-Stamm, Kristin C Gunsalus, J Wade Harper, Michael E Cusick, Frederick P Roth, David E Hill, and Marc Vidal. A map of the interactome network of the metazoan *c. elegans*. *Science*, 303(5657):540–543, jan 2004.
- [152] Nevan J Krogan, Gerard Cagney, Haiyuan Yu, Gouqing Zhong, Xinghua Guo, Alexandr Ignatchenko, Joyce Li, Shuye Pu, Nira Datta, Aaron P Tikuisis, Thanuja Punna, Jos M Peregrn-Alvarez, Michael Shales, Xin Zhang, Michael Davey, Mark D Robinson, Alberto Paccanaro, James E Bray, Anthony Sheung, Bryan Beattie, Dawn P Richards, Veronica Canadien, Atanas Lalev, Frank Mena, Peter Wong, Andrei Starostine, Myra M Canete, James Vlasblom, Samuel Wu, Chris Orsi, Sean R Collins, Shamanta Chandran, Robin Haw, Jennifer J Rilstone, Kiran Gandi, Natalie J Thompson, Gabe Musso, Peter St Onge, Shaun Ghanny, Mandy H Y Lam, Gareth Butland, Amin M Altaf-Ul, Shigehiko Kanaya, Ali Shilatifard, Erin O’Shea, Jonathan S Weissman, C James Ingles, Timothy R Hughes, John Parkinson, Mark Gerstein, Shoshana J Wodak, Andrew Emili, and Jack F Greenblatt. Global landscape of protein complexes in the yeast *saccharomyces cerevisiae*. *Nature*, 440(7084):637–643, mar 2006.
- [153] Johanna Weindl, Zaher Dawy, Pavol Hanus, Juergen Zech, and Jakob C Mueller. Modeling promoter search by *e. coli* rna polymerase: one-dimensional diffusion in a sequence-dependent energy landscape. *J Theor Biol*, 259(3):628–634, aug 2009.
- [154] Petter Hammar, Prune Leroy, Anel Mahmutovic, Erik G Marklund, Otto G Berg, and Johan Elf. The lac repressor displays facilitated diffusion in living

- p>cells.
- Science*
- , 336(6088):1595–1598, 2012.
- [155] Leonid Mirny, Michael Slutsky, Zeba Wunderlich, Anahita Tafvizi, Jason Leith, and Andrej Kosmrlj. How a protein searches for its site on dna: the mechanism of facilitated diffusion. *J. Phys. A: Math. Theor.*, 42(43):434013, oct 2009.
  - [156] Nicolae Radu Zabet and Boris Adryan. A comprehensive computational model of facilitated diffusion in prokaryotes. *Bioinformatics*, 28(11):1517–1524, 2012.
  - [157] Justin Malin, Daphne Ezer, Xiaoyan Ma, Steve Mount, Hiren Karathia, Seung Gu Park, Boris Adryan, and Sridhar Hannenhalli. Crowdsourcing: Spatial clustering of low-affinity binding sites amplifies in vivo transcription factor occupancy. *BioRxiv*, aug 2015.
  - [158] Marieke Simonis, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo de Wit, Bas van Steensel, and Wouter de Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4c). *Nat Genet*, 38(11):1348–1354, nov 2006.
  - [159] Borbala Mifsud, Filipe Tavares-Cadete, Alice N Young, Robert Sugar, Stefan Schoenfelder, Lauren Ferreira, Steven W Wingett, Simon Andrews, William Grey, Philip A Ewels, Bram Herman, Scott Happe, Andy Higgs, Emily LeProust, George A Follows, Peter Fraser, Nicholas M Luscombe, and Cameron S Osborne. Mapping long-range promoter contacts in human cells with high-resolution capture hi-c. *Nat Genet*, 47(6):598–606, jun 2015.
  - [160] Ralph Stadhouders, Petros Kolovos, Rutger Brouwer, Jessica Zuin, Anita van den Heuvel, Christel Kockx, Robert-Jan Palstra, Kerstin S Wendt, Frank Grosveld, Wilfred van Ijcken, and Eric Soler. Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. *Nat Protoc*, 8(3):509–524, mar 2013.
  - [161] Jonathan Cairns, Paula Freire-Pritchett, Steven W. Wingett, Csilla Várnai, Andrew Dimond, Vincent Plagnol, Daniel Zerbino, Stefan Schoenfelder, Biola-Maria Javierre, Cameron Osborne, Peter Fraser, and Mikhail Spivakov.

Chicago: robust detection of dna looping interactions in capture hi-c data. *Genome Biology*, 17(1):127, 2016.

- [162] H. S Carslaw and J. C Jaeger. *Conduction of Heat in Solids*. Clarendon Press, 1959.
- [163] Thayaparan Paramanathan, Daniel Reeves, Larry J Friedman, Jane Kondev, and Jeff Gelles. A general mechanism for competitor-induced dissociation of molecular complexes. *Nature communications*, 5, 2014.
- [164] Zeba Wunderlich and Leonid A Mirny. Spatial effects on the speed and reliability of protein–dna search. *Nucleic acids research*, 36(11):3570–3578, 2008.
- [165] Ivan V Kulakovskiy, Yulia A Medvedeva, Ulf Schaefer, Artem S Kasianov, Ilya E Vorontsov, Vladimir B Bajic, and Vsevolod J Makeev. HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic acids research*, 41(D1):D195–D202, 2013.
- [166] Mikhail Pachkov, Piotr J Balwiercz, Phil Arnold, Evgeniy Ozonov, and Erik van Nimwegen. SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic acids research*, 41(D1):D214–D220, 2013.
- [167] Anthony Mathelier, Xiaobei Zhao, Allen W Zhang, François Parcy, Rebecca Worsley-Hunt, David J Arenillas, Sorana Buchman, Chih-yu Chen, Alice Chou, Hans Ienasescu, et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research*, page gkt997, 2013.
- [168] Charles E Grant, Timothy L Bailey, and William Stafford Noble. Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.
- [169] Matthew D Simon, Stefan F Pinter, Rui Fang, Kavitha Sarma, Michael Rutenberg-Schoenberg, Sarah K Bowman, Barry A Kesner, Verena K Maier, Robert E Kingston, and Jeannie T Lee. High-resolution xist binding maps reveal two-step spreading during x-chromosome inactivation. *Nature*, 504(7480):465–469, 2013.



- [170] Erik Splinter, Elzo de Wit, Elphège P Nora, Petra Klous, Harmen JG van de Werken, Yun Zhu, Lucas JT Kaaij, Wilfred van IJcken, Joost Gribnau, Edith Heard, et al. The inactive x chromosome adopts a unique three-dimensional conformation that is dependent on xist rna. *Genes & development*, 25(13):1371–1383, 2011.
- [171] Michael M Hoffman, Jason Ernst, Steven P Wilder, Anshul Kundaje, Robert S Harris, Max Libbrecht, Belinda Giardine, Paul M Ellenbogen, Jeffrey A Bilmes, Ewan Birney, et al. Integrative annotation of chromatin elements from encode data. *Nucleic acids research*, page gks1284, 2012.
- [172] Paulina Kolasinska-Zwierz, Thomas Down, Isabel Latorre, Tao Liu, X Shirley Liu, and Julie Ahringer. Differential chromatin marking of introns and expressed exons by h3k36me3. *Nature genetics*, 41(3):376–381, 2009.
- [173] Alessandro Vezzoli, Nicolas Bonadies, Mark D Allen, Stefan MV Freund, Clara M Santiveri, Brynn T Kvinlaug, Brian JP Huntly, Berthold Göttgens, and Mark Bycroft. Molecular basis of histone h3k36me3 recognition by the pwwp domain of brpf1. *Nature structural & molecular biology*, 17(5):617–619, 2010.
- [174] Jaroslav Jelinek, Shoudan Liang, Yue Lu, Rong He, Louis S Ramagli, Elizabeth J Shpall, Marcos RH Estecio, and Jean-Pierre J Issa. Conserved dna methylation patterns in healthy blood cells and extensive changes in leukemia measured by a new quantitative technique. *Epigenetics*, 7(12):1368–1378, 2012.
- [175] Joseph D Fleming, Giulio Pavesi, Paolo Benatti, Carol Imbriano, Roberto Mantovani, and Kevin Struhl. Nf-y coassociates with fos at promoters, enhancers, repetitive elements, and inactive chromatin regions, and is stereo-positioned with growth-controlling transcription factors. *Genome research*, 23(8):1195–1209, 2013.
- [176] Kyle L MacQuarrie, Abraham P Fong, Randall H Morse, and Stephen J Tapscott. Genome-wide transcription factor binding: beyond direct target regulation. *Trends Genet*, 27(4):141–148, apr 2011.

- [177] Emma K Farley, Katrina M Olson, Wei Zhang, Alexander J Brandt, Daniel S Rokhsar, and Michael S Levine. Suboptimization of developmental enhancers. *Science*, 350(6258):325–328, oct 2015.
- [178] Andrea I Ramos and Scott Barolo. Low-affinity transcription factor binding sites shape morphogen responses and enhancer evolution. *Philos Trans R Soc Lond, B, Biol Sci*, 368(1632):20130018, dec 2013.
- [179] B Hentsch, A Mouzaki, I Pfeuffer, D Rungger, and E Serfling. The weak, fine-tuned binding of ubiquitous transcription factors to the il-2 enhancer contributes to its t cell-restricted activity. *Nucleic Acids Res*, 20(11):2657–2665, jun 1992.
- [180] William W Fisher, Jingyi Jessica Li, Ann S Hammonds, James B Brown, Barret D Pfeiffer, Richard Weiszmann, Stewart MacArthur, Sean Thomas, John A Stamatoyannopoulos, Michael B Eisen, Peter J Bickel, Mark D Biggin, and Susan E Celniker. Dna regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in drosophila. *Proc Natl Acad Sci U S A*, 109(52):21330–21335, dec 2012.
- [181] Xu Luo and MICHIELE Sawadogo. Antiproliferative properties of the usf family of helix-loop-helix transcription factors. *Proceedings of the National Academy of Sciences*, 93(3):1308–1313, 1996.
- [182] T Kevin Howcroft, Charles Murphy, Jocelyn D Weissman, Sam J Huber, Michèle Sawadogo, and Dinah S Singer. Upstream stimulatory factor regulates major histocompatibility complex class i gene expression: the u2 $\delta$ e4 splice variant abrogates e-box activity. *Molecular and cellular biology*, 19(7):4788–4797, 1999.
- [183] Raoul Kopelman. Rate processes on fractals: Theory, simulations, and experiments. *Journal of Statistical Physics*, 42(1):185–200, Jan 1986.
- [184] Aurélien Bancaud, Sébastien Huet, Nathalie Daigle, Julien Mozziconacci, Joël Beaudouin, and Jan Ellenberg. Molecular crowding affects diffusion and binding

of nuclear proteins in heterochromatin and reveals the fractal organization of chromatin. *The EMBO journal*, 28(24):3785–3798, 2009.

- [185] Arttu Jolma, Jian Yan, Thomas Whittington, Jarkko Toivonen, Kazuhiro R Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, et al. DNA-binding specificities of human transcription factors. *Cell*, 152(1):327–339, 2013.
- [186] Stefan Schoenfelder, Tom Sexton, Lyubomira Chakalova, Nathan F Cope, Alice Horton, Simon Andrews, Sreenivasulu Kurukuti, Jennifer A Mitchell, David Umlauf, Daniela S Dimitrova, et al. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nature genetics*, 42(1):53–61, 2010.
- [187] Stefan Schoenfelder, Robert Sugar, Andrew Dimond, Biola-Maria Javierre, Harry Armstrong, Borbala Mifsud, Emilia Dimitrova, Louise Matheson, Filipe Tavares-Cadete, Mayra Furlan-Magaril, Anne Segonds-Pichon, Wiktor Jurkowski, Steven W Wingett, Kristina Tabbada, Simon Andrews, Bram Herman, Emily LeProust, Cameron S Osborne, Haruhiko Koseki, Peter Fraser, Nicholas M Luscombe, and Sarah Elderkin. Polycomb repressive complex prc1 spatially constrains the mouse embryonic stem cell genome. *Nat Genet*, 47(10):1179–1186, oct 2015.
- [188] O Puig, F Caspary, G Rigaut, B Rutz, E Bouveret, E Bragado-Nilsson, M Wilm, and B Sraphin. The tandem affinity purification (tap) method: a general procedure of protein complex purification. *Methods*, 24(3):218–229, jul 2001.
- [189] Kanwardeep S Kaleka, Amber N Petersen, Matthew A Florence, and Nashaat Z Gerges. Pull-down of calmodulin-binding proteins. *J Vis Exp*, (59), jan 2012.
- [190] Anna Brckner, Ccile Polge, Nicolas Lentze, Daniel Auerbach, and Uwe Schlatter. Yeast two-hybrid, a powerful tool for systems biology. *Int J Mol Sci*, 10(6):2763–2788, jun 2009.
- [191] Ulf Reimer, Ulrich Reineke, and Jens Schneider-Mergener. Peptide arrays: from macro to micro. *Curr Opin Biotechnol*, 13(4):315–320, aug 2002.

- [192] Ling Ren, Edith Chang, Khadijah Makky, Arthur L Haas, Barbara Kaboord, and M Walid Qoronfleh. Glutathione s-transferase pull-down assays using dehydrated immobilized glutathione resin. *Anal Biochem*, 322(2):164–169, nov 2003.
- [193] Melissa J Fullwood, Mei Hui Liu, You Fu Pan, Jun Liu, Han Xu, Yusoff Bin Mohamed, Yuriy L Orlov, Stoyan Velkov, Andrea Ho, Poh Huay Mei, Elaine G Y Chew, Phillips Yao Hui Huang, Willem-Jan Welboren, Yuyuan Han, Hong Sain Ooi, Pramila N Ariyaratne, Vinsensius B Vega, Yanquan Luo, Peck Yean Tan, Pei Ye Choy, K D Senali Abayratna Wansa, Bing Zhao, Kar Sian Lim, Shi Chi Leow, Jit Sin Yow, Roy Joseph, Haixia Li, Kartiki V Desai, Jane S Thomsen, Yew Kok Lee, R Krishna Murthy Karuturi, Thoreau Herve, Guillaume Bourque, Hendrik G Stunnenberg, Xiaoan Ruan, Valere Cacheux-Rataboul, Wing-Kin Sung, Edison T Liu, Chia-Lin Wei, Edwin Cheung, and Yijun Ruan. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, 462(7269):58–64, nov 2009.
- [194] Elzo de Wit and Wouter de Laat. A decade of 3c technologies: insights into nuclear organization. *Genes Dev*, 26(1):11–24, jan 2012.
- [195] Chris A Brackley, Jill M Brown, Dominic Waithe, Christian Babbs, James Davies, Jim R Hughes, Veronica J Buckle, and Davide Marenduzzo. Predicting the three-dimensional folding of cis-regulatory regions in mammalian genomes using bioinformatic data and polymer models. *Genome Biol*, 17:59, mar 2016.
- [196] Natalia Tapia, Caitlin MacCarthy, Daniel Esch, Adele Gabriele Marthaler, Ulf Tiemann, Marcos J Arazo-Bravo, Ralf Jauch, Vlad Cojocaru, and Hans R Schler. Dissecting the role of distinct oct4-sox2 heterodimer configurations in pluripotency. *Sci Rep*, 5:13533, aug 2015.
- [197] Peter Langfelder, Bin Zhang, and Steve Horvath. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. *Bioinformatics*, 24(5):719, 2008.

- [198] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.
- [199] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.
- [200] Peggy J Farnham. Insights from genomic profiling of transcription factors. *Nat Rev Genet*, 10(9):605–616, sep 2009.
- [201] Debbie L C van den Berg, Tim Snoek, Nick P Mullin, Adam Yates, Karel Bezstarosti, Jeroen Demmers, Ian Chambers, and Raymond A Poot. An oct4-centered protein interaction network in embryonic stem cells. *Cell Stem Cell*, 6(4):369–381, apr 2010.
- [202] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29, may 2000.
- [203] Gene Ontology Consortium. Gene ontology consortium: going forward. *Nucleic Acids Res*, 43(Database issue):D1049–56, jan 2015.
- [204] Aaron W Reinke, Gevorg Grigoryan, and Amy E Keating. Identification of bzip interaction partners of viral proteins hbx, meq, bzlf1, and k-bzip using coiled-coil arrays. *Biochemistry*, 49(9):1985–1997, mar 2010.
- [205] F Galvagni, S Capo, and S Oliviero. Sp1 and sp3 physically interact and co-operate with gabp for the activation of the utrophin promoter. *J Mol Biol*, 306(5):985–996, mar 2001.
- [206] Xu Li, Wenqi Wang, Jiadong Wang, Anna Malovannaya, Yuanxin Xi, Wei Li, Rudy Guerra, David H Hawke, Jun Qin, and Junjie Chen. Proteomic analyses reveal distinct chromatin-associated and soluble transcription factor complexes. *Mol Syst Biol*, 11(1):775, jan 2015.

- [207] Jian Wang, Keke Huo, Lixin Ma, LiuJun Tang, Dong Li, Xiaobi Huang, Yanzhi Yuan, Chunhua Li, Wei Wang, Wei Guan, Hui Chen, Chaozhi Jin, Junchen Wei, Wanqiao Zhang, Yongsheng Yang, Qiongming Liu, Ying Zhou, Cuili Zhang, Zhihao Wu, Wangxiang Xu, Ying Zhang, Tao Liu, Donghui Yu, Yaping Zhang, Liang Chen, Dewu Zhu, Xing Zhong, Lixin Kang, Xiang Gan, Xiaolan Yu, Qi Ma, Jing Yan, Li Zhou, Zhongyang Liu, Yunping Zhu, Tao Zhou, Fuchu He, and Xiaoming Yang. Toward an understanding of the protein interaction network of the human liver. *Mol Syst Biol*, 7:536, oct 2011.
- [208] Yui-Han Loh, Qiang Wu, Joon-Lin Chew, Vinsensius B Vega, Weiwei Zhang, Xi Chen, Guillaume Bourque, Joshy George, Bernard Leong, Jun Liu, Kee-Yew Wong, Ken W Sung, Charlie W H Lee, Xiao-Dong Zhao, Kuo-Ping Chiu, Leonard Lipovich, Vladimir A Kuznetsov, Paul Robson, Lawrence W Stanton, Chia-Lin Wei, Yijun Ruan, Bing Lim, and Huck-Hui Ng. The oct4 and nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet*, 38(4):431–440, apr 2006.
- [209] Jiancong Liang, Ma Wan, Yi Zhang, Peili Gu, Huawei Xin, Sung Yun Jung, Jun Qin, Jiemin Wong, Austin J Cooney, Dan Liu, and Zhou Songyang. Nanog and oct4 associate with unique transcriptional repression complexes in embryonic stem cells. *Nat Cell Biol*, 10(6):731–739, jun 2008.
- [210] Meenakshi B Kannan, Vera Solovieva, and Volker Blank. The small maf transcription factors maff, mafg and mafk: current knowledge and perspectives. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1823(10):1841–1846, 2012.
- [211] J Turner and M Crossley. Cloning and characterization of mctbp2, a co-repressor that associates with basic krppel-like factor and other mammalian transcriptional regulators. *EMBO J*, 17(17):5129–5140, sep 1998.
- [212] G Chinnadurai. Ctbp, an unconventional transcriptional corepressor in development and oncogenesis. *Mol Cell*, 9(2):213–224, feb 2002.

- [213] Yujiang Shi, Fei Lan, Caitlin Matson, Peter Mulligan, Johnathan R Whetstine, Philip A Cole, Robert A Casero, and Yang Shi. Histone demethylation mediated by the nuclear amine oxidase homolog lsd1. *Cell*, 119(7):941–953, dec 2004.
- [214] Christina Galonska, Michael J Ziller, Rahul Karnik, and Alexander Meissner. Ground state conditions induce rapid reorganization of core pluripotency factor binding before global epigenetic reprogramming. *Cell Stem Cell*, 17(4):462–470, oct 2015.
- [215] Ajazul H Wani, Alistair N Boettiger, Patrick Schorderet, Ayla Ergun, Christine Mnger, Ruslan I Sadreyev, Xiaowei Zhuang, Robert E Kingston, and Nicole J Francis. Chromatin topology is coupled to polycomb group protein subnuclear organization. *Nat Commun*, 7:10291, jan 2016.
- [216] Mercedes Pardo, Benjamin Lang, Lu Yu, Haydn Prosser, Allan Bradley, M Madan Babu, and Jyoti Choudhary. An expanded oct4 interaction network: implications for stem cell biology, development, and disease. *Cell Stem Cell*, 6(4):382–395, apr 2010.
- [217] Ken Sugino, Chris M Hempel, Benjamin W Okaty, Hannah A Arnson, Saori Kato, Vardhan S Dani, and Sacha B Nelson. Cell-type-specific repression by methyl-cpg-binding protein 2 is biased toward long genes. *J Neurosci*, 34(38):12877–12883, sep 2014.
- [218] Jon-Matthew Belton, Rachel Patton McCord, Johan Harmen Gibcus, Natalia Naumova, Ye Zhan, and Job Dekker. Hi-c: a comprehensive technique to capture the conformation of genomes. *Methods*, 58(3):268–276, nov 2012.
- [219] Job Dekker and Leonid Mirny. The 3d genome as moderator of chromosomal communication. *Cell*, 164(6):1110–1121, mar 2016.
- [220] Lars Grontved, Sam John, Songjoon Baek, Ying Liu, John R Buckley, Charles Vinson, Greti Aguilera, and Gordon L Hager. C/ebp maintains chromatin accessibility in liver and facilitates glucocorticoid receptor recruitment to steroid response elements. *EMBO J*, 32(11):1568–1583, may 2013.

- [221] Axel Visel, Matthew J Blow, Zirong Li, Tao Zhang, Jennifer A Akiyama, Amy Holt, Ingrid Plajzer-Frick, Malak Shoukry, Crystal Wright, Feng Chen, Veena Afzal, Bing Ren, Edward M Rubin, and Len A Pennacchio. Chip-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854–858, feb 2009.
- [222] Luca Giorgetti, Trevor Siggers, Guido Tiana, Greta Caprara, Samuele Notarbartolo, Teresa Corona, Manolis Pasparakis, Paolo Milani, Martha L Bulyk, and Gioacchino Natoli. Noncooperative interactions between transcription factors and clustered dna binding sites enable graded transcriptional responses to environmental inputs. *Molecular Cell*, 37(3):418–428, 2010.
- [223] Hélène Touzet, Jean-Stéphane Varré, et al. Efficient and accurate p-value computation for position weight matrices. *Algorithms Mol Biol*, 2(1510.1186):1748–7188, 2007.
- [224] Helge G Roider, Aditi Kanhere, Thomas Manke, and Martin Vingron. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, 23(2):134–141, 2007.
- [225] Nicolae Radu Zabet and Boris Adryan. Estimating binding properties of transcription factors from genome-wide binding profiles. *Nucleic Acids Research*, page gku1269, 2014.
- [226] R Stojnic and D Diez. PWMEnrich: PWM enrichment analysis. R package version 4.2.0., 2014.
- [227] Gary D Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, 2000.
- [228] Gary D Stormo and Yue Zhao. Determining the specificity of protein–DNA interactions. *Nature Reviews Genetics*, 11(11):751–760, 2010.
- [229] Zeba Wunderlich and Leonid A Mirny. Different gene regulation strategies revealed by analysis of binding motifs. *Trends in Genetics*, 25(10):434–440, 2009.



- [230] Florian Mueller, Paul Wach, and James G McNally. Evidence for a common mode of transcription factor interaction with chromatin as revealed by improved quantitative fluorescence recovery after photobleaching. *Biophysical journal*, 94(8):3323–3339, 2008.
- [231] Lihua Julie Zhu, Ryan G Christensen, Majid Kazemian, Christopher J Hull, Metewo Selase Enuameh, Matthew D Basciotta, Jessie A Brasefield, Cong Zhu, Yuna Asriyan, David S Lapointe, et al. FlyFactorSurvey: a database of drosophila transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Research*, 39(suppl 1):D111–D117, 2011.
- [232] Carl O Pabo and Robert T Sauer. Transcription factors: structural families and principles of DNA recognition. *Annual Review of Biochemistry*, 61(1):1053–1095, 1992.
- [233] Shalev Itzkovitz, Tsvi Tlusty, and Uri Alon. Coding limits on the number of transcription factors. *BMC Genomics*, 7(1):239, 2006.
- [234] Volker Matys, Olga V Kel-Margoulis, Ellen Fricke, Ines Liebich, Sigrid Land, A Barre-Dirrie, Ingmar Reuter, D Chekmenev, Mathias Krull, Klaus Hornischer, et al. TRANSFAC and its module transcompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, pages D108–D110, 2006.
- [235] Nicholas M Luscombe, Susan E Austin, Helen M Berman, and Janet M Thornton. An overview of the structures of protein-DNA complexes. *Genome Biol*, 1(1):1–37, 2000.
- [236] C O Pabo and R T Sauer. Transcription factors: structural families and principles of dna recognition. *Annu Rev Biochem*, 61:1053–1095, 1992.
- [237] Martha L Bulyk, Philip LF Johnson, and George M Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Research*, 30(5):1255–1261, 2002.

- [238] Roger Pique-Regi, Jacob F Degner, Athma A Pai, Daniel J Gaffney, Yoav Gilad, and Jonathan K Pritchard. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome research*, 21(3):447–455, 2011.
- [239] Sonali Mukherjee, Michael F Berger, Ghil Jona, Xun S Wang, Dale Muzzey, Michael Snyder, Richard A Young, and Martha L Bulyk. Rapid analysis of the dna-binding specificities of transcription factors with dna microarrays. *Nat Genet*, 36(12):1331–1339, dec 2004.