Deciphering Transcriptional Regulatory Circuits: Transcription Factor Binding and Regulatory Variants Identification

by

Ningxin Ouyang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2022

Doctoral Committee:

Associate Professor Alan Boyle, Chair
Assistant Professor Carlos Aguilar
Professor Jun Li
Associate Professor Ryan Mills
Professor Maureen Sartor

Ningxin Ouyang
nouyang@umich.edu
ORCID iD: 0000-0002-1182-8861
©Ningxin Ouyang 2022

To my loving family

ACKNOWLEDGEMENTS

The work presented in this dissertation could not have been accomplished without the support of so many people. First and foremost, I would like to thank my advisor, Dr. Alan Boyle, for the continuous support, guidance and patience during my PhD study. I have benefited greatly from your wealth of knowledge, your trust in our abilities as scientist and your enthusiasm for science and life. Thank you for establishing an open and positive lab climate. Such a friendly, inclusive, engaging, and fun lab culture has made the last 6 years an enjoyable journey. It was a great experience to grow with the lab and work with so many wonderful people.

I would like to express my sincere gratitude to all members of the Boyle lab. Thank you all for making the Boyle lab a great place to work. From the various lab outings, events, parties to constructive discussions, the moments spent in the lab with all of you will be invaluable memories I would cherish for life. Thank you, Adam Diehl, for your technical support and thoughtful, detailed feedback for my projects and writings. Thank you, Jessica Switzenberg, for your invaluable help in manuscripts writing. Thank you everyone in our dry lab family for our enjoyable time spent together in the lab. Thank you, Shengcheng Dong, Sam Zhao, Bradley Crone, Christopher Castro.

I would like to express my deepest appreciation to my committee: Dr. Carlos Aguilar, Dr. Jun Li, Dr. Ryan Mills, and Dr. Maureen Sartor for their insightful comments, immense support and guidance over the years.

I would like to thank Dr. Margit Burmeister for being a wonderful graduate advisor. I would also like to thank the Bioinformatics administrative specialist Julia Eussen, for all the support and assistance throughout the years.

Finally, I must express my very profound gratitude to my parents for their unfailing love and support throughout my years of study. Thank you, mon, for taking the 14-hour international flight every year just to spend time with me.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

ABSTRACT

Transcription factors can bind cis-regulatory DNA elements to achieve their regulatory properties. Identification of transcription factor binding sites remains a crucial goal in deciphering transcriptional regulatory circuits. The vast majority of genetic variants identified from whole genome sequencing studies and leading disease-causing SNPs implicated in genome-wide association studies (GWAS) lie well outside of protein coding regions. The functional effect of variants within non-coding sequence is often through creation or disruption of individual transcription factor binding that alters downstream gene regulatory activity. In addition, transcription factors can act cooperatively to regulate transcription in a context-specific manner. Some binding complexes, such as CTCF together with cohesin proteins, can also mediate 3D chromatin interactions that also have downstream gene regulatory control. My dissertation is focused on deciphering these interactions driving gene regulatory circuits. In this dissertation, I develop an improved footprinting algorithm to map transcription factor binding sites genome-wide, study regulatory variants associated with transcription factor binding affinity, and explore transcription factor cooperativity and their role in 3D chromatin interaction.

In Chapter 2, I will introduce the TRACE algorithm, a multi-threaded computational footprinting method to predict transcription factor binding sites, using chromatin accessibility data (DNase-seq or ATAC-seq) and sequence information. In the development of the method, I Implemented a multivariate hidden Markov model

(HMM) in an unsupervised training manner for identifying and labeling DNase footprints. TRACE exhibited the best overall performance among all existing footprinting methods after a comprehensive evaluation.

In Chapter 3, I investigated the association between genetic variants and transcription factor binding activity to identify footprint QTLs (fpQTLs) at a base pair resolution, contributing to a better knowledge of the mechanism behind the linkage between genotypic variation and gene regulation as well as disease phenotypes. Overall, detection of fpQTLs provides additional information for a more complete characterization of the landscape of human regulatory variation and its direct effect on gene expression. In Chapter 4, I employed an artificial neural network called Self-Organizing Maps (SOMs) to identify "clusters" of transcription factors and define co-binding patterns. I specifically examined the transcription factor enrichment at chromatin loop anchors and studied how they might modulate downstream looping effects.

Together, the studies in this dissertation provide improved transcription factor binding site prediction, deliver improved functional interpretation of noncoding variation, and expand our knowledge on transcription factor cooperativity and their effect on 3D organization of chromatin.

CHAPTER I

Introduction

Binding of transcription factors (TFs) to specific DNA sequences is elementary for transcriptional regulation. Accurate identification of transcription factor binding sites (TFBSs) is critical for understanding gene expression and whole regulatory networks. Whole genome sequencing studies have revealed that the majority of variants lies outside of protein coding regions. The functional effect of noncoding single nucleotide variants (SNVs) is often through altering TFBSs. Accurate understanding of variants' impact on transcription factor binding can lead to a better interpretation of how noncoding variants affect gene function and lead to disease phenotypes. Gene regulatory control also relies on chromatin interactions driven through 3-dimensional conformation. Co-binding of particular TFs with another, including CTCF, the cohesin complex, and a host of accessory TFs, often work cooperatively to mediate the formation of 3D chromatin structure and elicit specific regulatory outcomes. Elucidating these co-binding patterns is central to understanding gene regulatory mechanisms.

In this dissertation, I evaluated existing footprinting methods and developed new algorithm for an improved genome-wide transcription factor mapping (Chapter 2). Then I extended the usage of this algorithm in characterizing and mapping functional

variation. I performed association test on TF binding affinity and genetic variation to identify footprint QTLs (fpQTLs) (Chapter 3). Finally, I explored TF co-binding patterns and investigated their potential impact on CTCF looping regulation and 3D chromatin contacts (Chapter 4).

In this introductory chapter, I will discuss a variety of experimental or computational appraches in TFBSs identification and their features and limitations. I will also describe some statistical analyses in genome-wide regulatory variants mapping and computational tools in regulatory variants characteriazation. Finally, I will discuss TF cooperative binding activity and CTCF-mediated chromatin interctions.

1.1 Identification of transcription factor binding sites

My dissertation has a focus on transcription factors and their binding pattern and mechanism. The first key element in my study is genome-wide TFBSs mapping. TFs are DNA binding proteins that can recognize their binding sequence and are essential in gene expression regulation. TFBSs are building blocks for regulatory sequences such as cis regulatory elements. Thus, identification of TFBSs is essential in understanding gene expression and regulatory networks. In this chapter I will discuss some commonly used TFBSs mapping techniques and tools, many of which utilize data from high-throughput functional genomics assays (Figure 1.1). The standard approaches that were widely used include Position Weight Matrices (PWMs) and Chromatin immunoprecipitation followed by sequencing (ChIP-seq). TFBSs can also be detected by DNase footprint by investigating chromatin accessibility patterns, as TFs often leave protected regions (footprints) from DNaseI digestion. However, these existing experimental or computational tools in TFBSs identification have different drawbacks that limit their application and accuracy, which will be discussed

in this chapter. As a result, improved TFBSs prediction methods are needed to map TFs with better quality and higher availability.

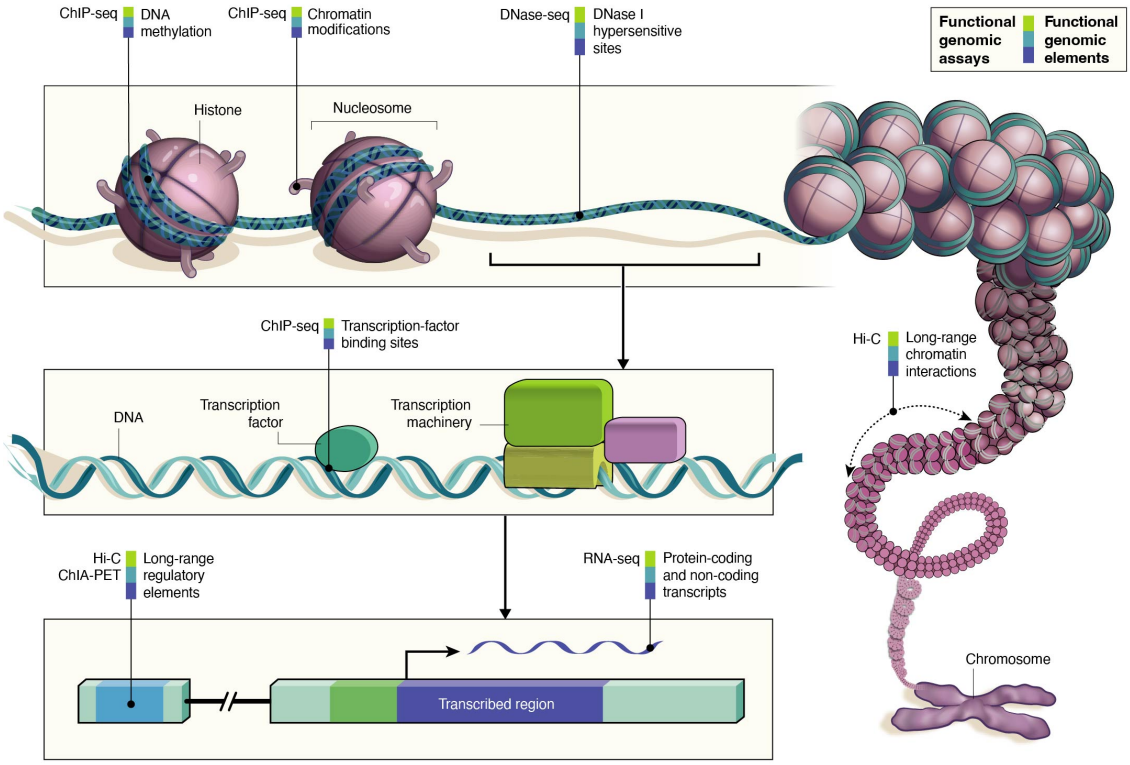

Figure 1.1: ENCODE data types. Figure adapted from Ecker et al. 2012 [1].

1.1.1 Chromatin Immunoprecipitation sequencing (ChIP-seq)

Chromatin immunoprecipitation followed by sequencing (ChIP–seq) is one of the early applications of next-generation sequencing. ChIP–seq measures proteins binding to DNA, and can be used to identify transcription factor binding, profile nucleosome positioning and histone modifications, and detect methylated DNA regions in a genome-wide manner [2]. ChIP–seq workflow includes crosslinking protein to DNA and then lysing cells and fragment (around 200-300 bp) with sonication. DNA fragments bound with specific proteins are enriched and unbound chromatin and proteins

are discarded. Purified DNA fragments can then be identified with next-generation sequencing techniques. Protein-specific antibody is required to immunoprecipitate the DNA–protein complex.

ChIP-seq has a low resolution (200-500 bp), but it can identify specific TFBSs with downstream analysis. However, quality (specificity & sensitivity) of antibody against the protein of interest limits its data quality. Success of ChIP-seq assays depends on the existence of a good antibody against the TFs of interest and the availability of large numbers of cells. Dependency on antibodies is one of the major limitations of ChIP-related technologies. Notably, available antibodies for TFs are very limited and only a very small fraction of TFs have ChIP-seq data available in ENCODE (Figure 1.2). There are only about 160 validated antibodies, that is only 9% of predicted TFs. Previous study also showed that 20–35% of the commercially produced antibodies tested were unsatisfactory [3, 4]. ENCODE has put in enormous effort in generating sequencing data and has produced 2443 TF ChIP-seq assays over the three phases [5]. However, TF ChIP-seq data is still in shortage, considering the total number of TFs and cell lines. In fact, ChIP-seq assay is labor intensive and relatively cost inefficient, so it is not feasible to map all TFs in all cell lines or tissues even if the corresponding antibody is available.

1.1.2 Position Weight Matrices (PWMs)

A position weight matrix (PWM) describes the DNA binding preferences of a TF. It is a probabilistic model that denotes the fraction of nucleotide occurrences (A, C, G or T) at each location of the motif. Most of PWMs from early studies were identified using systematic evolution of ligands by exponential enrichment (SELEX) data. More recent PWMs were generated from ChIP-seq data. Several public databases provide PWMs for a large amount of TFs from different sources. JASPAR is one

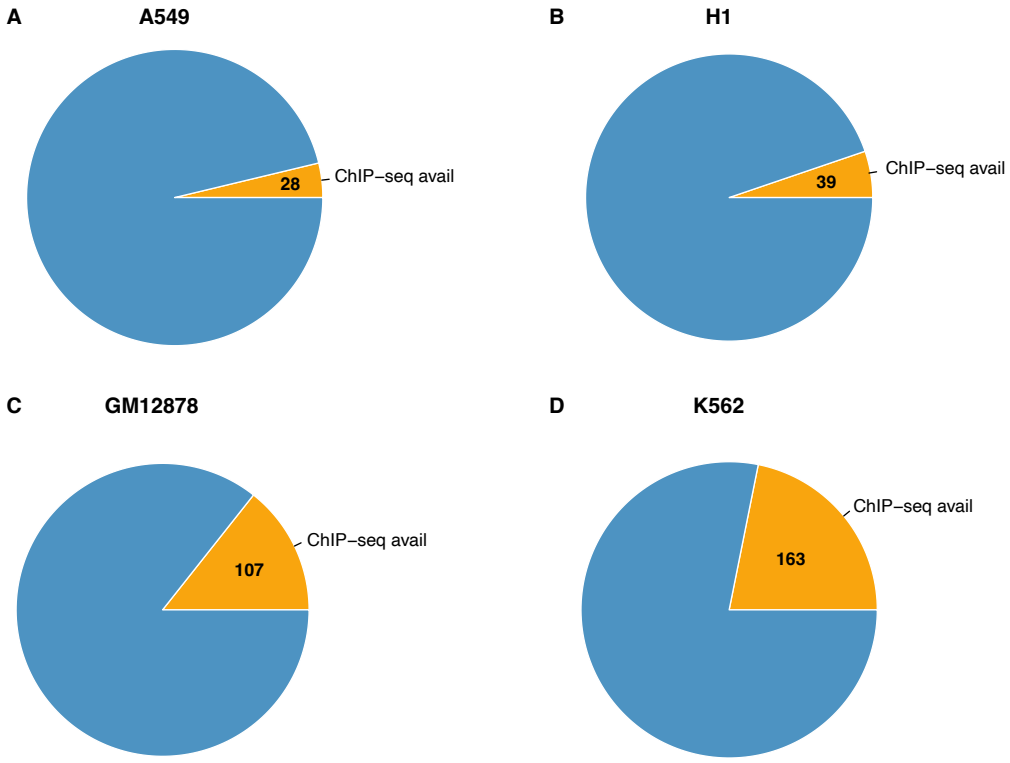

Figure 1.2: Numbers of TFs with ChIP-seq data available in (A) A549, (B) H1, (C) K562 and (D) GM12878 among all motifs in JASPAR CORE database (non-redundant). Each pie plot has a total number of 746 TFs, orange portion represents number of TFs has ChIP-seq data in that cell line in ENCODE.

of these databases and it gathers a non-redundant collection of PWMs from around 800 TFs along with cluster information of closely related TFs.

PWM-based methods to identify TFBSs have high resolution but also have significant drawbacks. One of its limitations is that binding preference might not be sufficient to measure binding affinity. The presence of a sequence motif for a TF does not necessarily imply there is actual protein binding in a particular cell line. Consequently, sequence-based methods can detect a large number of putative TFBSs, but only a small fraction of them are functional sites or active binding sites, which can potentially result in a high false positive rate (FPR). For instance, CTCF

is one of the most studied and abundant TFs, but it still has a FPR of 61.6%. A more extreme example is GATA2, which has a 96.2% FPR. Together, it shows the difficulty of using PWMs alone for TFBS mapping.

1.1.3 DNase I footprinting assays

The traditional DNase I footprinting assay is another experimental method that can be used to detect TFBSs. It is a classical method to investigate in vivo protection of DNA by protein binding. DNase I footprinting assays can be used to identify TFBS since a critical feature of footprinting is that the affinity of TFs for their binding sites is greater than the affinity of DNase I for the same sequences, leading to the protection of TF-occupied DNA from nuclease attack. In open chromatin regions, TFs will protect the DNA it is bound to from DNase I cut, leaving nucleotide-resolution footprints.

DNase I can cut the DNA at open chromatin regions, but TFs will protect the sequence that they bind, resulting in a reduced number of cuts at those regions. The limited DNase I digestion at the protein protected regions will result in a series of nested fragments that can be resolved by running a gel. In the gel, the missing bands are TF bound regions (footprints).

1.1.4 DNase I hypersensitive sites sequencing (DNase-seq) and computational footprinting methods

DNase I hypersensitive sites sequencing (DNase-seq) is chromatin accessibility assay that measures the absence of nucleosomes [6]. It maps open chromatin regions that are more accessible for protein interaction, which are shown to be enriched for regulatory elements including enhancers, promoters, silencers, insulators and locus control regions. DNase-seq identifies DNase I hypersensitive sites (DHSs) where DNase I can cut at higher frequency, allowing for genome-wide footprinting and

nucleotide-level identification of TFBSs by searching footprint-like regions with low numbers of DNase I cuts surrounded by regions with high numbers of cuts. Since the development of the first DNase-seq data based computational footprinting method in *Saccharomyces cerevisiae* [7], several chromatin accessibility data based computational methods have been developed to detect footprints genome-wide (Table 1.1) [8, 9, 10, 11, 12, 13, 14, 15, 16]. ATAC-seq can also be used in footprinting as it identifies open chromatin regions where Tn5 transposase can insert [17].

Computational footprinting algorithms are categorized into *de novo* methods (the Boyle method, DNase2TF, HINT, PIQ and Wellington) and motif-centric methods (DeFCoM, BinDNase, CENTIPEDE, FLR). *De novo* methods are TF-agnostic, which detect footprints by investigating chromatin accessibility pattern across input regions. These generic footprints have the desired DNase digestion pattern, but not necessarily match with any sequence motif. On the other hand, motif-centric methods are TF-specific, which assess the TF-binding probability at each pre-generated candidate binding site for TFs of interest.

Binding sequence preference is another key factor that can be considered in footprints identification, especially when predicting binding sites for specific TF with a known motif.

1.1.5 Limitations of current TFBSs prediction methods

Current widely used TFBSs prediction includes ChIP-seq, motif sequence scan, and footprinting. ChIP-seq provides enrichment peaks for DNA binding proteins but has a low resolution and is labor intensive. More recent ChIP based assays such as ChIP-exo [18] has higher resolution and improved cost efficiency, but still have some of the same disadvantages as ChIP-seq, including labor intensive and limited antibody availability. PWM-based methods can identify binding sites at

Table 1.1: Computational footprinting methods

	Input data	Algorithm	Year published	Language
Boyle method	DNase-seq	HMM	2011	C++
Neph method (FOS)	DNase-seq	Sliding window	2012	C++
HINT	DNase-seq, and/or ChIP-seq	HMM	2014/2016	python
DNase2TF	DNase-seq, mappability	Sliding window	2014	R
pyDNase (Wellington)	DNase-seq	Sliding window	2013	python
Footprint mixture (FLR)	DNase-seq	Mixture model	2014	R
CENTIPEDE	DNase-seq, PWM bit-score and/or sequence conservation and/or distance to the nearest TSS	Bayesian mixture model	2011/2015	R
BinDNase	DNase-seq, PWM score	Logistic regression	2015	R
DEFCOM	DNase-seq, PWM score	SVM	2017	python

high resolution but usually have high false positive rates.

De novo footprinting methods only detect general footprints instead of binding sites for specific TFs, so they require additional sequencing scan or motif database query steps to label binding sites of TF of interest. Motif-centric methods are TF-specific, but require pre-identified candidate binding sites for TFs of interest. Another limitation of existing footprint methods is that many of them include supervised

training steps, which means they still need ChIP-seq data to generate training sets. Their applications heavily rely on availability and quality of ChIP-seq data.

Collectively, the limitations of experimental assays, PWM-based methods, and existing computational footprinting algorithms call for a much-needed improvement in computational footprinting algorithm with increased prediction accuracy and expanded application. New algorithms that overcome these drawbacks can provide much improved TFBSs profiles and contribute to better understanding of transcription regulation.

1.2 Effect of genetic variation on TF binding

Regulatory variants can modulate gene expression in a context-specific manner and are often linked to disease phenotypes. However, the molecular mechanisms underlying their regulating effect are still poorly understood. Although genome-wide association studies have linked genetic variants to many human phenotypes such as gene expression, chromatin accessibility, protein level, and histone modifications, and functional annotation of the genome has been improved with machine learning and deep learning techniques, our understanding of biological mechanisms underlying causal variants are still limited. Multiple studies have linked functional regulatory variants with TF motif sequence, so that identification of the effect of genetic variation on TF binding is key to understanding and interpreting downstream consequences of gene expression variations. Genome Wide Association Studies (GWAS) has provided rich resource to understanding human phenotypes by identifying putative causal genetic variation statistically correlated with traits or diseases. 85% leading single nucleotide polymorphisms (SNPs) identified in GWAS studies lie outside of protein coding sequences [19]. Therefore, understanding the role of noncoding

variation in altering TF binding is important in studying the impact of variants on gene expression and understand gene regulation. Identifying how variation affects TF binding activity will allow us to generate more informed hypotheses on the function of regulatory genetic variants in the genome.

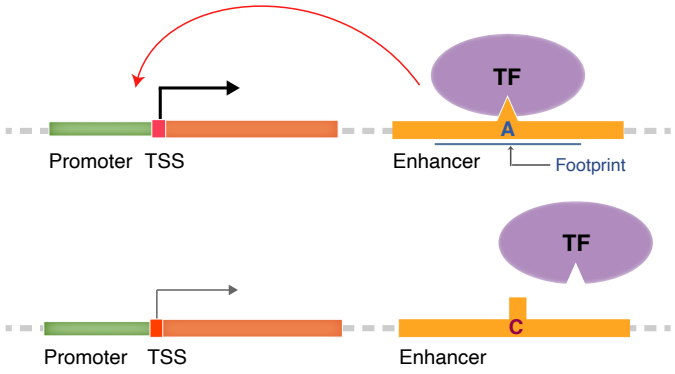

Figure 1.3: Simplified schematic that a genetic variant disrupts TF binding at footprint, and therefore decreases expression of gene that is regulated by that enhancer.

1.2.1 Non-coding genetic variation

Development of high throughput sequencing technologies has allowed whole genome analysis for variant detection. GWAS studies have discovered a large number of genetic variants linked to diseases phenotypes, however a substantial portion of these associations are still unexplained, encouraging further studies on underlying biological mechanisms. Whole genome sequencing studies from projects including the 1000 Genomes Project have revealed that 95% of genomic variation falls into non-coding sequences [20, 21], and 85% of leading GWAS SNPs that were potentially disease causing also lie outside of protein coding regions [19]. In addition, the genetic variations in cis-regulatory sequence have been linked to altered gene regulation and

human disease. The abundance of variants in non-coding regions and their association with diseases raises the importance of understanding how non-coding variants affect gene function and disease phenotypes.

The functional effect of noncoding SNVs can occur through the disruption or creation of transcription factor binding sites. Genetic variation can lead to alteration of TF binding sequence and/or change in DNase sensitivity, and therefore affect the likelihood of footprints occurrence and gene expression (Figure 1.3). In fact, transcription factor binding sites make up only 8% of the genome but contain 31% of variants identified by GWAS. Previous studies have reported that genetic variations in transcription factor binding sites that alter likelihood of transcription factor binding are also associated with human diseases. For example, a previous study reported variants associated with Type 2 diabetes, located in open chromatin region, significantly increased reporter gene expression, consistent with increased transcription factor binding [22]. A systematic study on genetic variation and its impact on TF binding, as well as potential downstream effects including gene expression and disease phenotypes can contribute to a better understanding and interpretation of genetic variations consequences.

1.2.2 Quantitative trait loci (QTLs) mapping

Association mapping of quantitatively measurable molecular traits has merged as a powerful approach in study of genetic variation. Quantitative trait locus (QTL) approach can be applied to any quantitatively measurable cellular traits with defined locus in the genome. QTL mapping of chromatin traits identifies genetic variants that regulate chromatin both proximally and distally. Genome-wide mapping of expression quantitative trait loci (eQTLs) has become an essential method to study the impact of variants on gene expression and understand gene regulation [23, 24]. The

Genotype-Tissue Expression (GTEx) provides the largest eQTL datasets of human tissues, characterizing genetic associations in 838 individuals over 49 human tissues. eQTL is the most extensively conducted QTL analysis where genetic variants associated with gene expression levels from RNA-seq are mapped on the whole genome. However, only a small fraction of previously localized disease-associated variants are also eQTL SNPs, although they are enriched in gene regulatory elements [25, 26]. In addition, the regulatory mechanism behind variants' effect on gene regulation is not clearly explained. Gene expression change might be correlated with casual variants through different mechanisms, instead of the eQTL directly regulate transcription. One putative mechanism of their effect is that some SNPs can create or disrupt TFBSs and alter TF binding affinity, thereby they can affect regulatory networks and change gene expression level. This also emphasizes the importance of investigating the relationship between gene variation and TF binding activity.

Besides gene expression level, chromatin accessibility is another molecular trait that has been worked on intensively and been linked to genetic variants. dsQTLs were first identified at which DNase-seq read depth correlates significantly with genotypes at nearby variants in 70 HapMap Yoruba lymphoblastoid cell lines (LCLs). They were also shown to be enriched in TFBSs and a large portion were found to be associated with differential expression levels of nearby genes. They were then further generalized as chromatin accessibility QTLs (caQTLs) and can be generated using ATAC-seq data. caQTLs were also reported to be enriched in TFBSs, suggesting the mechanism behind caQTLs might also include alteration in TF binding activity. Genome-wide association test between TF binding affinity and genetic variants and investigation on these linkages can further extend our knowledge on their mechanism and potential provide new functional annotations.

1.2.3 Computational tools to predict regulatory variants

Computational non-coding annotation tools have been developed to address the prioritization of genetic variants and are available to predict regulatory variants in noncoding regions with different scoring schemes. They utilize many functional genomics assays and association studies, such as ChIP-seq and DNase-seq, as well as eQTLs, DNase footprints and TF motif sequences, along with other annotations, such as sequence conservation and 3D chromatin interactions from assays like Hi-C and ChIA-PET. Machine learning techniques and, more recently, deep learning methods have been widely used for functional variations prediction, which can be grouped into three categories: disease risk predictors, fitness consequence predictors and regulatory function predictors.

Tools like RegulomeDB utilize ChIP-seq, DNase-seq, and eQTL annotations as well as scores from another non-coding SNP annotation method DeepSEA, in a random forest machine learning model to prioritize non-coding variation [27, 28]. Deep learning-based method DeepSEA utilizes a convolutional neural network to predict functional SNPs in single nucleotide resolution [29]. It predicts variant effects on regulatory function based on 919 predictors from functional genomics features including TF binding, open chromatin, and histone mark profiles, across various cell lines. These computational non-coding annotation tools refine the functional regulatory variants from candidate variation and generate quantitative score predictions. These methods can be utilized to assess the QTL associated variants to further study their functional properties.

1.3 TF co-binding patterns and 3D chromatin structure

The importance of 3D chromatin structure in gene regulation has been demonstrated in many recent studies. DNA binding sites of the insulator protein CTCF are present in many chromatin loop boundaries, and variations in CTCF occupancy are associated with looping divergence, but their contribution to 3D chromatin structural evolution remains unknown. Study of TF cooperativity, especially CTCF and cohesin proteins related complex which mediate 3D chromatin interaction, is key to extended understanding gene regulatory networks.

1.3.1 TF cooperativity

Binding of TFs to specific DNA sequences are elementary for transcription regulation. TFs can act cooperatively to regulate gene expression under varying conditions. They often recruit co-regulators and epigenetic modifiers to affect the 3D chromatin structure. Determining TF co-binding patterns is central to understanding gene regulatory mechanisms. Co-binding of particular TFs with another, including CTCF, the cohesin complex, and a host of accessory TFs, may work cooperatively to mediate the formation of 3D chromatin structure and elicit specific regulatory outcomes.

Recent studies have reported hundreds of co-binding TF pairs [30], and also showed that TF co-localization is prevalent and, in many cases, co-binding TFs form DNA-mediated complexes instead of direct interaction. TF co-binding may result from either TF binding at neighboring sites along the DNA strand, or through TF protein-protein interactions. Most TF co-binding studies utilize ChIP-seq data as their main protein binding information source; however, this method cannot discern between the two modes of co-occupancy because ChIP measures enrichment of protein interacted DNA but not necessarily DNA that were directly bound by that

protein. To better understand TF cooperativity, a comprehensive map of TFs co-localization due to directed DNA binding or protein-protein interactions and ability to explore and interpret these complex relationships is needed.

1.3.2 CTCF-mediated chromatin interactions

Our previous study demonstrated that looping variation may produce differential expression by refining altered enhancer–promoter interactions, and raised questions about the necessity of CTCF variability in chromatin looping dynamics [31]. CTCF is known to bind at insulators to mediate enhancer-blocking activity [32], and also plays an important role in 3D genome structure formation as chromatin looping mediator. Clustered factors might regulate CTCF binding and promote CTCF loops. The loop anchors bound with more factors clustered with CTCF were also shown to have a greater capacity to mediate stronger loops [33]. In addition, CTCF-mediated chromatin loops were reported to be involved in enhancer-promoter loops [34, 35] and CTCF binding might facilitate enhancer-promoter loop formation.

A few TFs have been shown to co-localize with CTCF and can regulate binding, participate in CTCF looping, and help modulate downstream looping effects. Previously reported CTCF co-localization proteins including BHLHE40, BPTF, CHD8, PARP1, SIN3A, TAF3, YY1 [36, 37, 33, 38, 39, 40, 41]. The cohesin complex proteins consisting of SMC1, SMC3, RAD21 and SA1/2 subunits are some of the most studied CTCF co-binding factors [42, 43]. A more systematic investigation on TF cooperativity can help identify additional TFs that might participate in CTCF-mediated chromatin loop stabilization and can lead to better understanding of how CTCF related protein complex mediate chromatin interaction.

1.3.3 3D structure measurements

Several high-throughput sequencing-based assays have been developed to measure the 3-dimensional interactions in the genome, including 3D conformation capture that detects the interaction between two loci (3C), 3C-based technologies that capture interactions between one locus and the rest of the genome (4C), between multiple loci (5C) and genome-wide (Hi-C), and chromatin-interaction analysis by paired-end-tag sequencing (ChIA-PET) [44, 45, 46, 47]. Analysis of data generated from these assays can identify loci pairs with higher interaction frequencies than expected by random chance. They detect 3D chromatin structures including topologically associated domains (TADs) exhibiting more interactions than outside the domain and TF-mediated chromatin loops involved in gene regulation. Loop interactions are at the edges of TADs and are often mediated by CTCF and cohesin complex. ChIA-PET features an immunoprecipitation step to enrich for chromatin complex with a specific protein. It maps long-range chromatin interactions bound by specific proteins genome-wide at high resolution. However, 3D conformation assays usually require very deep sequencing and can contain experimental noise. In addition, direct comparison between observations from different 3D conformation assays can be very challenging due to lack of benchmarks.

Several analysis tools have been developed to process ChIA-PET data, including Mango, a bias-correcting ChIA-PET analysis pipeline [48]. It can detect significant chromatin loop anchors, enabling protein-mediated functional interactions study and downstream functional annotation.

1.4 Conclusion

TFBSs are building blocks for cis-regulatory elements. Precise genome-wide identification of TFBSs is essential in understanding gene regulatory network and can provide rich resource for downstream regulatory network analysis. In this dissertation, I developed a new computational footprinting algorithm TRACE, which incorporates DNase-seq or ATAC-seq data and PWMs within a multivariate Hidden Markov Model (HMM) to detect footprint, with a better prediction performance and improved applicability (Chapter 2). TRACE's ability to predict individual-specific and tissue-specific footprint enables genome-wide test on impact of regulatory variants on TF binding activity. I performed association test on TF binding affinity and genetic variation to identify footprint QTLs (fpQTLs), providing powerful information for functional interpretation of human noncoding variation (Chapter 3). Finally, I investigated TF cooperativity mechanism and explored how TF co-binding patterns and CTCF-related molecular complexes interact to determine regulatory effects and impact chromatin conformation (Chapter 4).

CHAPTER II

TRACE: Transcription Factor Footprinting Using Chromatin Accessibility Data and DNA Sequence

2.1 Abstract

Transcription is tightly regulated by cis-regulatory DNA elements where transcription factors can bind. Thus, identification of transcription factor binding sites (TFBSs) is key to understanding gene expression and whole regulatory networks within a cell. The standard approaches used for TFBS prediction, such as position weight matrices (PWMs) and chromatin immunoprecipitation followed by sequencing (ChIP-seq), are widely used, but have their drawbacks including high false positive rates and limited antibody availability, respectively. Several computational footprinting algorithms have been developed to detect TFBSs by investigating chromatin accessibility patterns, however these also have limitations. We have developed a footprinting method to predict Transcription factor footpRints in Active Chromatin Elements (TRACE) to improve the prediction of TFBS footprints. TRACE incorporates DNase-seq data and PWMs within a multivariate Hidden Markov Model (HMM) to detect footprint-like regions with matching motifs. TRACE is an unsupervised method that accurately annotates binding sites for specific TFs automatically with no requirement for pre-generated candidate binding sites or ChIP-seq training data. Compared to published footprinting algorithms, TRACE has the best over-

all performance with the distinct advantage of targeting multiple motifs in a single model.

2.2 Introduction

Identification of cis-regulatory elements where transcription factors (TFs) bind remains a key goal in deciphering transcriptional regulatory circuits. Standard approaches to identify sets of active transcription factor binding sites (TFBSs) include the use of position weight matrices (PWMs) [49] and ChIP-seq [50]. While these methods have been successful, both suffer from drawbacks that limit their usefulness. PWMs are able to identify high-resolution binding sites but are prone to extremely high false positive rates in the genome. On the other hand, while ChIP-seq binding measurements are highly specific and have a significantly reduced false positive rate, the resolution is comparatively low, labor intensive, and depends on suitable antibodies that are only available for a limited number of TFs. Newer experimental techniques for identification of DNA-bound protein binding sites, such as ChIP-exo [18] and CUT&RUN [51], have the advantage of high resolution and cost efficiency, but still share the same labor intensive and limited antibody availability disadvantages as ChIP-seq. To complement these approaches, another experimental method has been developed using data from high-throughput sequencing after DNase I digestion (DNase-seq) [52]. DNase-seq identifies stretches of open regions of chromatin where DNase I cuts at a higher frequency. Within these regions, TFBSs can be identified at nucleotide resolution by searching for footprint-like regions with low numbers of DNase I cuts embedded in high-cut peaks.

Hesselberth et al. (2009) first proposed a DNase-seq signal based computational method to detect footprints at base pair resolution in *Saccharomyces cerevisiae*. Since

then, several computational footprinting algorithms have been developed to detect TFBSs by investigating chromatin accessibility patterns, which can be categorized as *de novo* (the Boyle method, DNase2TF, HINT, PIQ and Wellington) and motif-centric (DeFCoM, BinDNase, CENTIPEDE, FLR) [8, 9, 10, 11, 12, 13, 14, 15, 16]. *De novo* methods detect footprints across input regions based on their DNase digestion pattern. However, most of these methods were not designed to distinguish between binding sites for specific TFs, and cannot automatically label TF-specific binding sites of interest. In contrast, motif-centric methods can predict TF-specific sites, but require pre-generated candidate binding sites for TFs and assess their probability of being TF-bound (active binding sites). This limits their performance as these methods are unable to detect additional regions of candidate binding sites. Moreover, some of these methods are supervised, requiring ChIP-seq data to generate positive and negative training sets, and can only be applied to TFs with high-quality antibodies. This is a constraint as only a minority of TFs have ChIP-seq data available [53].

In addition to DNase digestion patterns, more detailed modeling of sequence preference information has been used in TFBSs identification. Hoffman and Birney (2010) [54] have previously proposed a Hidden Markov Model (HMM)-based method, termed Sunflower, to predict TFBSs based solely on sequence data. Instead of scanning for motif sequences directly, this model takes into consideration the competition between multiple TFs to provide a binding profile for all factors included in the model. While Sunflower still suffers from sequence-only method limitations for identifying TFBSs, it has a greater ability to distinguish the specific TF that binds at each predicted site.

We have developed an unsupervised footprinting method, TRACE, based on a

HMM framework [55, 56] and inspired by the success of Sunflower and other existing footprinting methods. TRACE predicts footprints and label binding sites for a set of desired TFs by integrating both DNase-seq data and PWMs. Our method is not dependent on pre-generated candidate binding sites or available ChIP-seq data, making it more flexible and broadly applicable compared to previous methods.

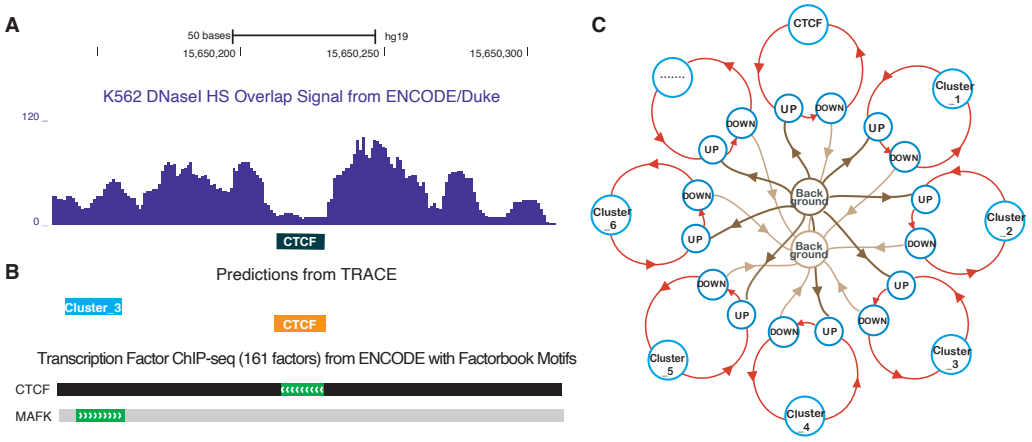

Figure 2.1: Computational footprinting can detect TFBSs at nucleotide resolution. (A) An example of digestion pattern at footprints: DNase I base overlap signal centered at CTCF motif sites (black box). (B) Predicted binding sites from TRACE using our 10-motif CTCF model match corresponding region of transcription factor binding obtained by ChIP-seq experiments with DNA binding motifs by the ENCODE Factorbook repository. (MAFK is a member of cluster 3 motifs.) (C) Simplified example schematic of a 7-motif CTCF model. Circles represent different hidden states including multiple motifs, lines with arrows represent transitions between different states. For simplicity, TOP states are not shown in the model structure.

2.3 Results

2.3.1 The TRACE Model

TRACE is an HMM-based unsupervised method with the number of hidden states dependent on the numbers and lengths of included PWMs (Figure 2.1). The basic structure of our model includes two Background states (the start and end of each open chromatin region delineated by DNase I cut sites), a target TF state (Figure 2.1C,

CTCF), a generic footprint state (Figure 2.1C, fp), and a series of bait motif states (Figure 2.1C, motif_1-motif_6). Each of the non-background states is surrounded by a set of UP, TOP and DOWN states (upslope, summit, and downslope of small peaks surrounding each footprint). Target TF states and bait motif states contain a number of discrete chains of states representing binding sites for each motif included in the model. The generic footprint state represents the regions that have a footprint-like digestion pattern, but do not match any PWMs in the model. TRACE includes a series of bait motifs representing commonly co-occurring motifs that significantly increase the performance of the model. For example, the 7-motif CTCF model in Fig. 1C includes a CTCF binding site state chain, 6 additional bait motifs (motif_1, motif_2, ..., motif_6), and generic footprints whose sequences do not match any of the included motifs. For each of these motifs, our model can distinguish its TF-bound states from unbound states based on the distinct DNase-seq digestion patterns of the motif sites (Figure 2.2).

TRACE takes PWMs and DNase-seq signals as inputs and models the emission distribution as a multivariate normal distribution using cut count signal and its derivative, and PWM scores at each genomic position. Each binding site (footprint) is expected to be in a region of low sequence density surrounded by a peak of density to either side with a high PWM score (Figure 2.1A, 2.1B).

2.3.2 TRACE outperforms existing methods

To evaluate the performance of TRACE relative to published computational footprinting methods, we tested 9 methods (DeFCoM, BinDNase, CENTIPEDE, FLR, DNase2TF, HINT, PIQ, Wellington, and a PWM-only comparison) on 99 TFs. For a fair comparison across all methods, *de novo* methods were applied to DNase-seq peaks containing the same sets of motif sites that were assessed by motif-centric

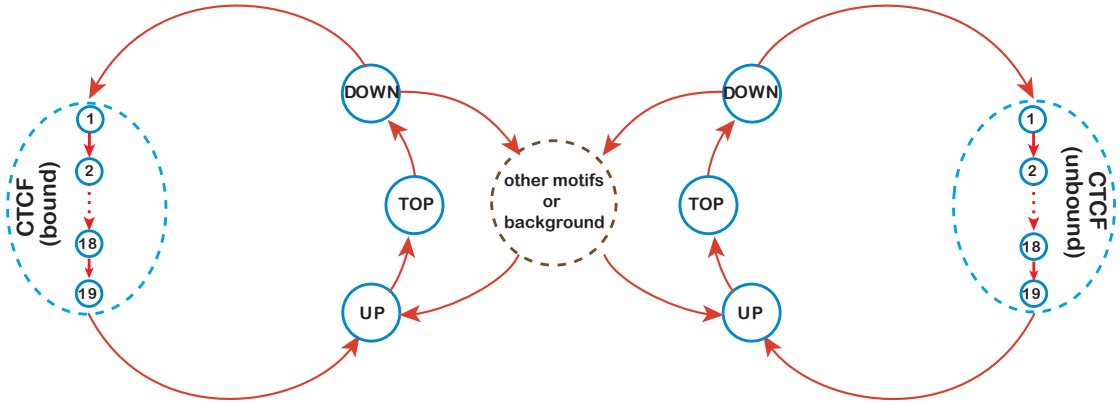

Figure 2.2: Detailed schematic of bound and unbound CTCF state in CTCF model. Circles represent different hidden states including binding sites and peaks, lines with arrows represent transitions between different states. For simplicity, all other motifs, generic footprint states and background are represented by a dashed line circle.

methods. Receiver operating characteristic curve area under the curve (ROC AUC), and Precision-Recall (PR) AUC of predictions of each TF were computed for each method based on the P-values or scores provided, and ranked across all methods (Figure 2.3A).

Previous studies evaluating computational footprinting methods focus on ROC AUC as a measurement of performance. Although this is a decent classification performance assessment, the number can be inflated by false positive predictions. For example, the ROC AUC statistic might imply a relatively favorable classification if the method tends to call most samples as positive hits when the data is highly unbalanced, as is the case for many of TFs tested. In addition, partial ROC AUC (ROC pAUC) were computed at a 5% false positive rate (FPR) cutoff. PR AUC was also included in the evaluations as it provides a better assessment of false positives. Compared with other footprinting methods, TRACE has the best overall performance based on average rank in both ROC pAUC and PR AUC across the 99 tested TFs (Figure 2.3C, 2.3D). It ranked first overall for 25.5% of TFs and in the

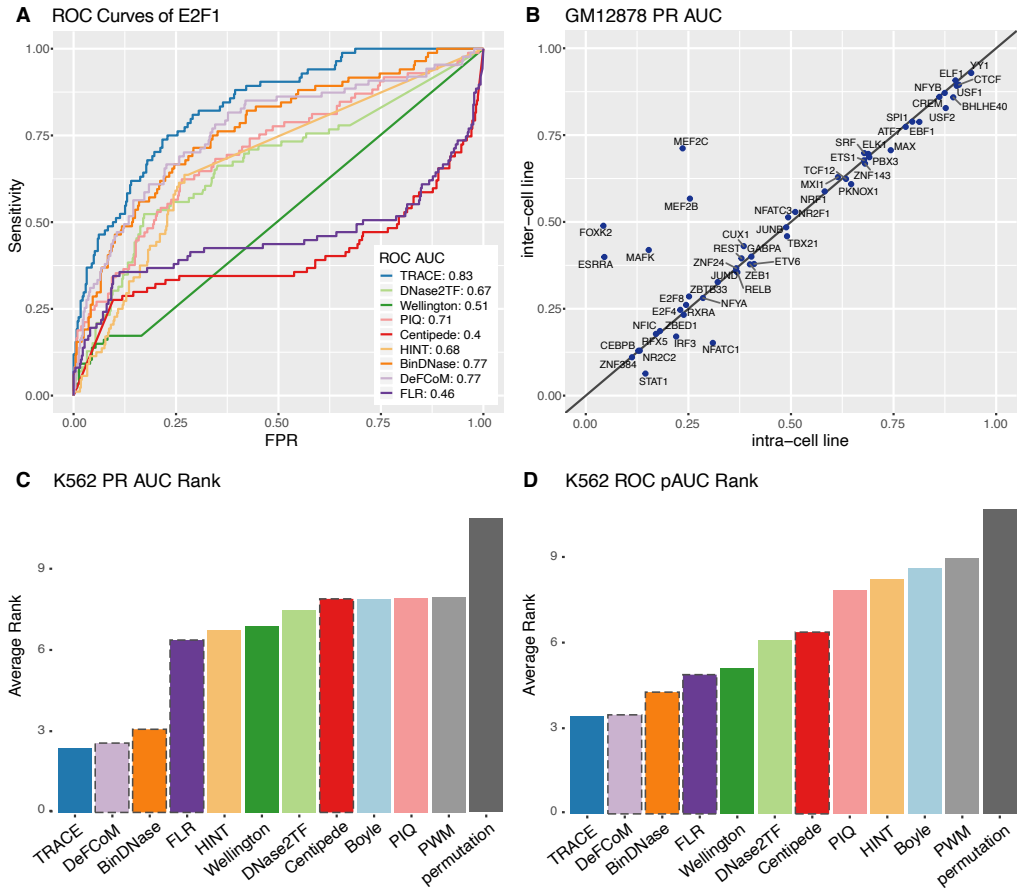

Figure 2.3: TRACE's performances are stable across-cell line and it outperforms other computational methods. (A) Example ROC curves of E2F1 for all methods evaluated. (B) Cross-cell line comparison of binding sites prediction in GM12878. Each point represents a TF tested, x-axis and y-axis are PR AUCs of applying TRACE using models trained from GM12878 and models trained from K562, respectively. Points above the diagonal line indicate TFs for which inter-cell line model performed better. (C, D) Average rank of PR AUC and ROC pAUC of existing methods across all TFs tested. The Bars with a dashed outline represent motif-centric methods.

top 5 for 96.9% of TFs. Compared to other unsupervised methods, TRACE ranked first for 87.7% of TFs. TRACE also outperformed supervised approaches including DeFCoM and BinDNase. TRACE can predict TF footprints with performance equal to or better than the best published methods without the requirement of positive and negative training datasets.

2.3.3 Bait motifs improve footprinting prediction accuracy

TRACE provides identification of binding sites for any desired TFs at nucleotide resolution. By incorporating DNase-seq data and PWM information, it can detect footprints with an anticipated DNase digestion pattern and matching motifs (Figure 2.1B). One important feature of our model is that states for different motifs are independent of each other, enabling its ability to distinctly label binding sites for multiple TFs. In addition, adding extra motifs to the model for a specific TF can potentially increase the accuracy of identifying TF-specific binding sites. These additional motifs as baits, discouraging the prediction of weakly matching sites and introducing competition, thus decreasing false positive rates [54]. However, including PWMs with similar sequence preference does not provide useful information and could decrease our model's ability to distinguish between binding sites of different motifs. To avoid this, only root motifs from each motif cluster in the JASPAR CORE vertebrates clustering were used [57] and the cluster that contains the TF of interest was excluded. Each root motif encompasses all of the position-specific scoring matrices (PSSMs) of a cluster generated by the RSAT matrix-clustering tool [58]. In a N-motif model, the root motifs from N-1 clusters with the greatest number of occurrences were selected. These N-1 motifs provide additional information, making the model more sensitive to identifying binding sites for the TF of interest.

Overall, the addition of bait motifs to the model yielded significant improvements over our original method, which had a similar HMM structure but did not include motif information (an option provided in TRACE) [8]. Using a 10-motif model (the TF of interest plus 9 extra motifs), the average PR AUC from TRACE increased by 0.20 (63.1%) over our original method and ROC pAUC improved by 20%.

By comparing models containing different numbers of extra motifs, we found that

additional TFs can increase the quality of TFBS identification in most cases. However, this was at the expense of considerably increased computational time. We determined that an optimal trade-off between performance increase and computational time was the 10-motif model, which is used in the remainder of this study.

2.3.4 TRACE can be applied accurately across cell lines

Cross cell-line validation was performed using models trained from K562 DNase-seq data and subsequently applied to GM12878 to test their performance compared to models trained on GM12878. Due to less available validation data in GM12878, this comparison utilized 52 TFs. The results indicated that TRACE can provide accurate predictions in one cell line using a model trained from another cell line, and that intra-cell line and inter-cell line predictions have comparable overall performance (Figure 2.3B). This suggests that the data processing steps can successfully capture the signature information of DNase digestion and diminish between-dataset variance to a degree sufficient for effective prediction across cell lines. It also indicates that the DNase digestion pattern of binding sites is preserved for most TFs across cell types. Some exceptions were observed however, for example ESRRA had significantly better performance in the inter-cell line test compared to the intra-cell line test. This TF has far fewer active binding sites in GM12878 (7.6% prevalence) than K562 (31.3% prevalence), and TRACE may not be able to learn an accurate model from the GM12878 data. This suggests that the model should be trained using high quality and most representative of the true genome-wide binding datasets, and the trained model can be applied across all cell types of interest.

TRACE's cross-cell line application allows for fast and large-scale TFBSs prediction using existing models without repetitive model training, which is the most time-consuming step. It also shows TRACE's advantage over supervised methods'

limited usage as only a very small fraction of TFs have ChIP-seq data available (Figure 1.2). To further showcase this flexibility, we have generated models for 526 JASPAR motifs and made them available through our GitHub site.

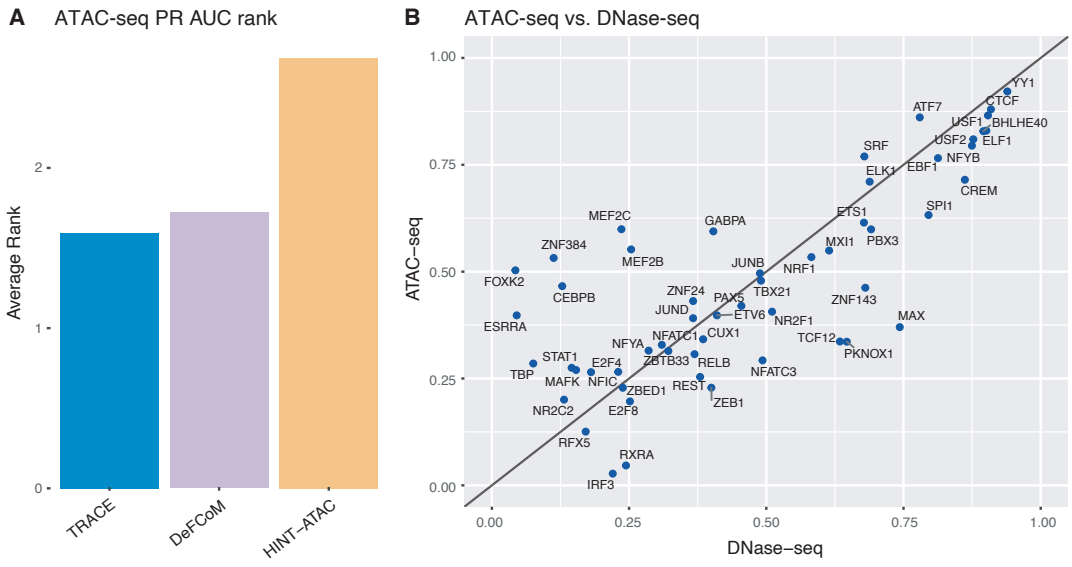

Figure 2.4: TRACE can perform well on ATAC-seq data. (A) Average rank of PR AUC across all TFs tested using ATAC-seq data for TRACE, DeFCom, and HINT-ATAC. (B) DNase-seq and ATAC-seq based TRACE performance comparison on PR AUC.

2.3.5 TRACE calls accurate footprints using ATAC-seq data

ATAC-seq provides chromatin accessibility information [17] and has been proposed to be useful in footprinting analyses. TRACE was tested using ATAC-seq and OMNI-ATAC-seq data to evaluate the performance of our model compared to other models designed to work with this particular data type. The results were compared with HINT-ATAC [59] and DeFCoM, as their original publications included ATAC-seq-based evaluation, and showed similar improvement in performance as in the case of DNase data.

Overall, TRACE maintains the best performance among these three methods, as

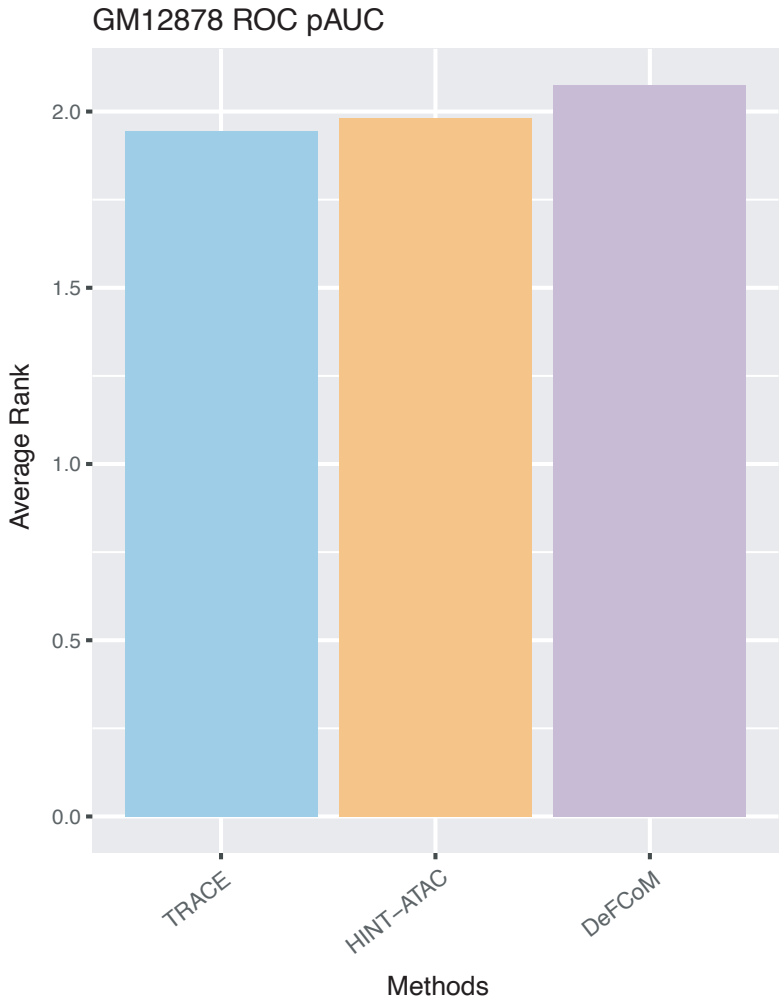

Figure 2.5: Average rank of ROC pAUC across all TFs tested using ATAC-seq data for TRACE, DeFCoM and HINT-ATAC.

it ranked first for both PR AUC and ROC pAUC (Figure 2.4A, 2.5). Prediction accuracy for TRACE was compared using DNase-seq and ATAC-seq data for each TF in GM12878 (Figure 2.4B). This analysis showed that ATAC-seq data provides comparable TFBS identification potential as DNase-seq, but that TRACE works slightly better comparing PR AUCs using DNase-seq data (60% of TFs). TFs that showed significant lower PR AUC using DNase-seq were caused by training data imbalances

from GM12878 DNase-seq peaks. For example, training sets from ATAC-seq data for FOXK2, ZNF384, CEBPB and TBP all have at least 100% increase of prevalence compared to DNase-seq training sets. To determine that the performance difference between these two datasets was not due to the deeper sequencing depth of DNase-seq, TRACE was performed on a DNase-seq dataset that had comparable and/or fewer reads than ATAC-seq. This had minimal effect on TRACE's performance and similar results were obtained (Figure 2.6). We further downsampled our datasets and found that footprinting performance would drop significantly if number of reads was below 50 million.

2.3.6 DNase footprinting has stable performance despite variable levels of data imbalance

It has been noted that not all TFs have accurately predicted active binding sites by computational footprinting, regardless of the algorithm applied. Our evaluation of existing footprinting methods indicates that all methods share similar performance trends across all TFs (Figure 2.7A left panel). This pattern also exists when assessing candidate binding sites by PWM scores alone (Figure 2.7A right panel). The footprinting performance gain against PWMs is only marginal for some TFs, and using PWM scores alone can even outperform all footprinting methods for 2 TFs among the 99 TFs tested here (Figure 2.7B).

The poor performance from footprinting might be partially due to the imbalance of positive (P) and negative (N) examples in data sets, as evaluation statistics of prediction for each TF were shown to be associated with its prevalence (fraction of positive samples, P/(P+N), see Methods) (Figure 2.7A). Data imbalance affects the quality of model training and, if the data distribution is too skewed, training quality will likely be diminished. Some poor performing models were associated with too

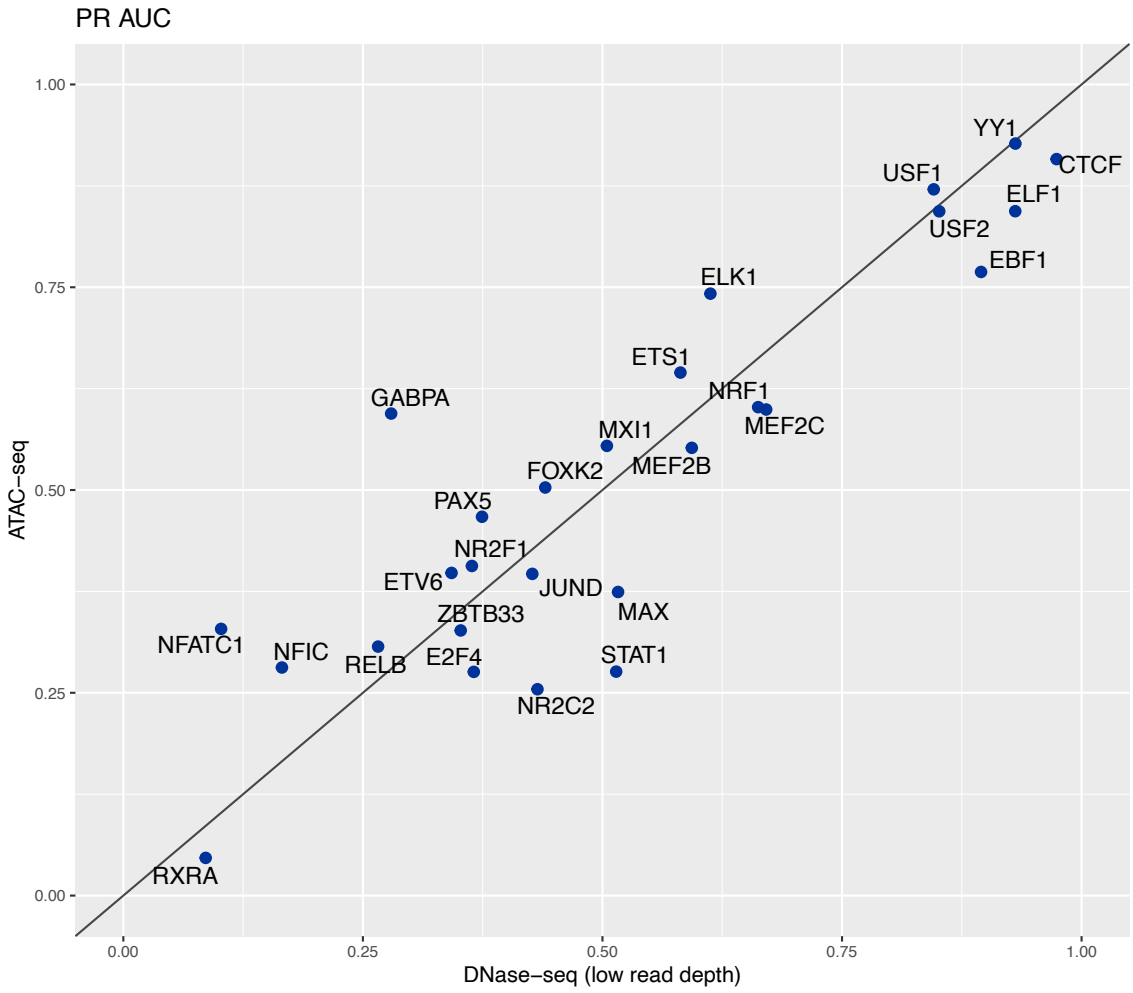

Figure 2.6: TRACE performance comparison on DNase-seq and ATAC-seq with comparable level of read depth. DNase-seq data used in this analysis has fewer reads than the one in Figure 2.4B.

few positive examples, due to their inability to distinguish active and inactive states in model training. However, this only accounts for a small subset and cannot explain the general trend of poor performance in TFBSs predictions. Comparing final models for each TF did not reveal significant correlation between prediction accuracies and statistics from different models.

To further explore how computational footprinting may be limited by data imbalance, the best footprinting performance for each TF was compared with a matched-

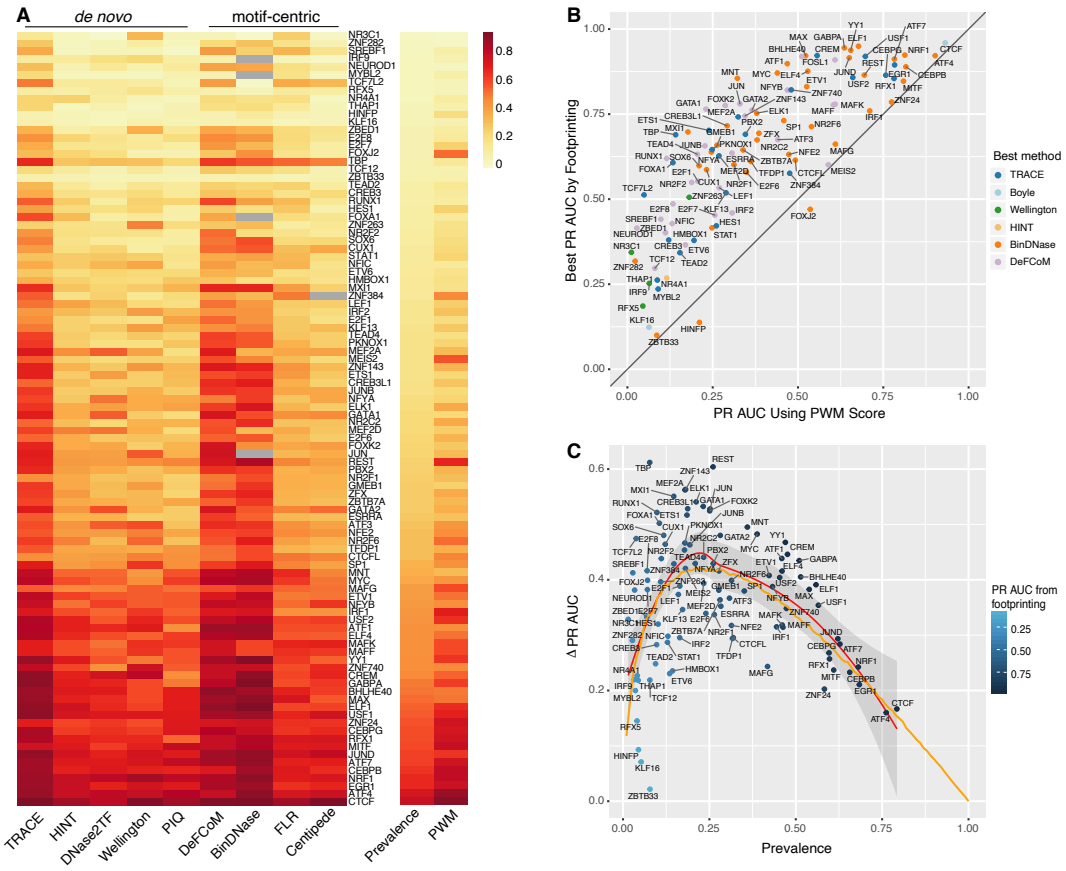

Figure 2.7: Computational footprinting methods share similar performance patterns. (A) Heatmap of PR AUC of all TFs tested from existing methods sorted by prevalence. (B) Comparison between the best PR AUC among all footprinting methods (y-axis) and PR AUC from using PWM score alone (x-axis) for every TF tested. (C) Performance improvement of footprinting methods over permutation for each TF colored by its best PR AUC from footprinting. Orange line is from a simulation test using positive instances drawn from N(10, 8), and negative instances from N(0, 7) to demonstrate expected PR AUC trend as binding prevalence changes.

imbalance permutation test of labeled sites (Figure 2.7C, 2.8, 2.9, 2.11). To complement this, simulations were performed with different levels of classification skill and varying imbalance to estimate how PR AUC and ROC AUC values reflect the classifier performance (Figure 2.8). As imbalance changes within a classification skill, we can expect the PR AUC will change correspondingly, but ROC AUC and ROC

pAUC will stay the same [60]. However, ROC curve often provides an overly optimistic assessment caused by true negatives used in false positive rate calculation, especially when there is a large skew in the data distribution [61].

Instead of comparing AUCs across TFs directly, their performance improvement over random labels (baseline) was measured. To examine the general performance gain using computational footprinting, max PR AUCs or ROC AUCs were collected from all existing methods, including TRACE, and then AUCs were subtracted from the corresponding permutation test. This number was used as a measurement of footprinting performance advantage over randomly predicted labels. The regression line of PR AUC increase against baseline has a skewed bell-like shape, consistent with the shape of simulated performance generated from a steady model skill (Figure 2.7C, 2.8, 2.11). This suggests that the performance of footprinting is roughly at a stable level and not associated with data imbalance. A higher evaluation statistic does not necessarily mean a better classification quality for that TF in some cases. Although prevalence may affect evaluation statistic values, no evidence was found that the true classification quality is determined by this data imbalance. Instead, there tends to be a stable level of footprinting classification performance increase compared to random across all TFs.

2.4 Discussion

Incorporating DNase-seq data and PWM information enables TRACE to detect footprints with the desired DNase digestion pattern and matching motifs. By including multiple motifs in the same model, our method provides a better overall TFBS prediction than other existing computational footprinting methods. Since different motifs are treated as separate states in our model, TRACE also has the potential of

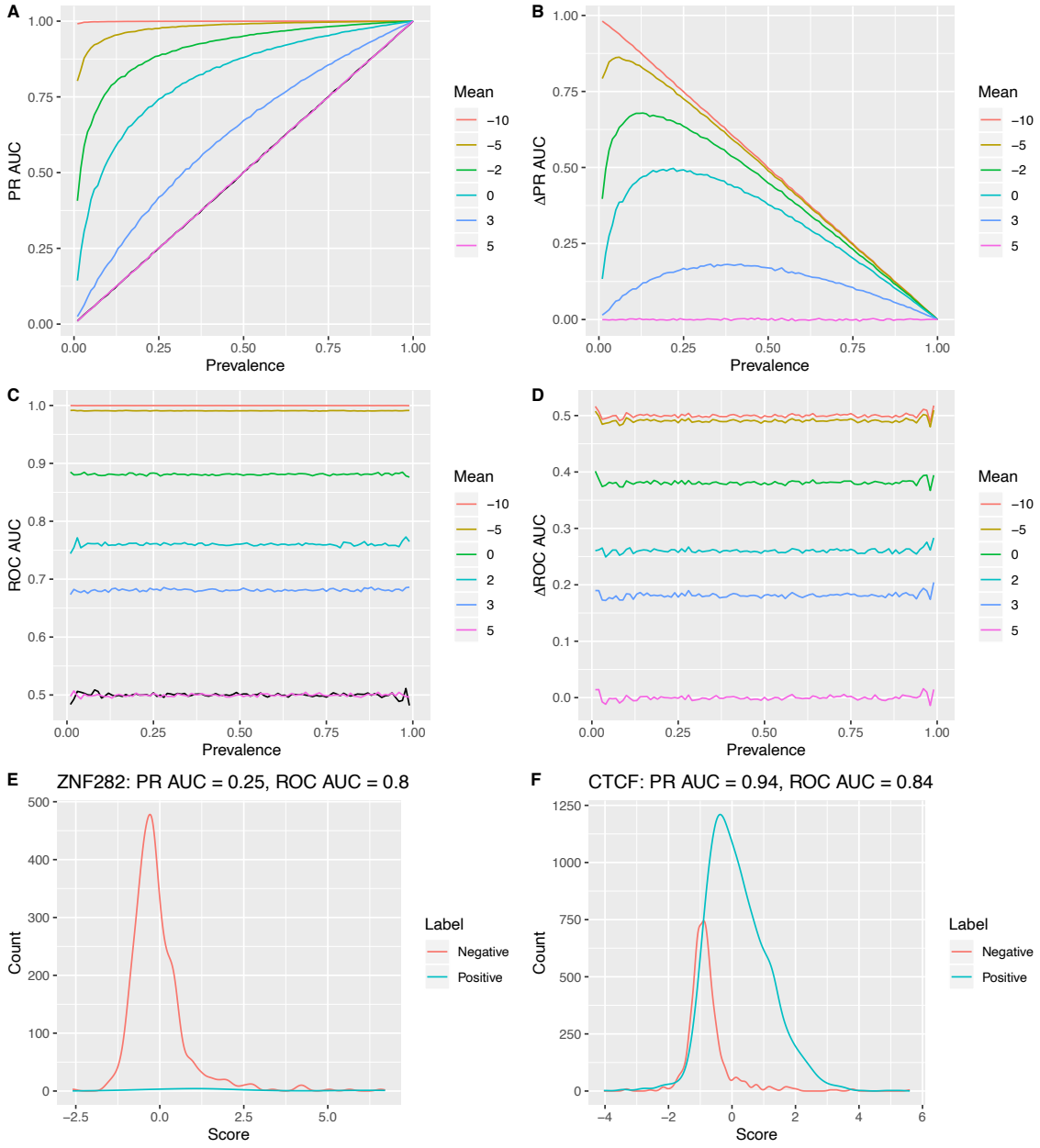

Figure 2.8: (A) PR AUC and (C) ROC AUC from simulation test with different prevalence. Scores for positive examples for all simulation were drawn from the same distribution N(5, 3). Each line represents a different negative example distribution with their own mean value, varying from 5 to -10. Black lines represent AUCs from random labels. (B) PR AUC increase and (D) ROC AUC increase from simulation test with different prevalence. (E, F) Score distributions for ZNF282 and NR2C2 as examples of TFs with different level of data imbalance.

targeting multiple TFs in a single model. Our method annotates binding sites for the desired TFs across input regions automatically, without requiring pre-generated

candidate binding sites or additional motif matching steps. In addition, as an unsupervised algorithm, its application is not limited to TFs with available ChIP-seq data.

Although computational footprinting has demonstrated the ability to predict TFBSs at an approximately consistent level, variation in evaluation statistics is still observed across TFs. A previous study showed that not all TFs will leave clear footprint-like nuclease cleavage patterns, and their protection of DNA from cleavage is correlated with residence time [62]. For some TFs, this can result in footprinting methods being unable to detect a consistent footprint-like DNase digestion pattern, and might fail to correctly label its binding sites. However, there is only limited residence time data available for a small number of TFs, and no comprehensive examination on residence time's impact on footprinting quality has been completed. Although residence time is known to be associated with enzymatic digestion patterns, it is also correlated with the number of active binding sites. GR, AP-1 and CTCF were tested by Sung et al. (2014) as TFs or TF subunits with short, intermediate and long residence time, respectively. For those TFs included in our test (NR3C1 as GR group, JUN, JUNB, JUND as AP-1 group, and CTCF), we observed that TFs with longer residency time tend to have a greater prevalence and a better PR AUC from footprinting (Figure 2.12). However, neither ROC AUC nor ROC pAUC of these TFs were correlated with residence time. This indicates the possibility that the association between residence time and footprinting ability might be caused by the correlation between performance evaluation statistics and TFBS prevalence. The observed performance disparity may only reflect the changes in fraction of active binding sites among all putative motif sites.

Our evaluation on all footprinting methods indicates that there might be a limited

classification accuracy gain that computational footprinting achieves, as the best performance for different TFs all centered at a certain level of classification quality. Our analysis suggests that evaluation statistics of classification from footprinting may be largely influenced by TFBS prevalence, and comparing them directly across TFs may be misleading. Computational footprinting in general might have a maximum potential for how well it can detect TFBSs and only very limited improvement can be achieved beyond this point.

2.5 Methods

2.5.1 Data and software

DNase-seq data in BAM and BED formats and ChIP-seq data in BED format were retrieved from the ENCODE download portal (Supplemental Table S1). ATAC-seq data for GM12878 cells using the standard ATAC-seq protocol were obtained from GSE47753 [17]. Omni-ATAC-seq data were obtained from the Sequencing Read Archive (SRA) with the BioProject accession PRJNA380283 [63]. 129 PWMs and cluster information (Supplemental Table S2) were downloaded from the JASPAR database [57]. Motif sites were identified using FIMO (MEME v5.0.3) with default parameters [64]. Evaluation statistics were generated using the Python package scikit-learn [65].

2.5.2 Data processing

After bias correction based on model and bias values reported in He et al. (2014) [66], we first counted the number of DNase-seq reads at each location using the 5' end of the reads, which is the DNase I digestion site. These cut counts were then normalized by non-zero mean of the surrounding 10k bp window (within data set normalization) as well as the percentile and standard deviation from the entire region

(between data set normalization) (Supplemental Methods). Normalized signals were then smoothed using the local regression method R [67] function LOESS [68] and their derivatives were calculated using the savitzky-golay filter in the Python package Scipy [69]. The first derivatives represent the slope of the processed signal curve and their signs indicate the increase or decrease data changes. UP, TOP and DOWN states in the peak have positive, zero and negative slopes, respectively.

2.5.3 ATAC-seq pipeline

ATAC-seq data for GM12878 was obtained from GSE47753, and Omni-ATAC-seq data was obtained from the Sequencing Read Archive (SRA) with the BioProject accession PRJNA380283. These data were processed following the Kundaje lab's ATAC-seq pipeline. (https://github.com/kundajelab/atac_dnase_pipelines)

2.5.4 Model details

Our model was built based on the idea of a generalized HMM, in which each motif consists of n states, each representing one position in its PWM (n is the length of PWM.). Each state in a motif can only transition to the next state in that motif, and the last state in this motif will transition to the state of the small peak. Thus, each motif can still be considered as an individual large state, but its parameters at each base pair can be captured separately. There are also footprint states representing generalized footprints which do not match any motif in included in the model. (Figure 2.1B) shows this in a simplified structure of TRACE model.

Two Background states represent starting and ending positions for each region of interest. Small peaks that surround footprints are divided into UP, TOP, and DOWN states with a one-direction connection. The DOWN state will either transit to a footprint state or the end of an open chromatin region state (Background state).

Only the last state in the motif states, or start of the region (Background state), can transit to an UP state.

To better predict transcription factor (TF) functional binding sites, we included two sets of states for each motif to represent active and inactive binding sites. For each TF, TRACE will differentiate and predict its functional binding sites, and those regions with a matching motif but are not necessarily bound by that TF.

2.5.5 Bait motif selection

In addition to the TF of interest states, our TRACE model also includes other motifs which serve as bait motifs. Adding bait motifs can potentially reduce the false positive labels of regions with footprint-like digestion patterns, but a weak sequence match with the TF of interest. These footprint-like regions might have a higher sequence preference for the bait motifs. This binding competition can increase the accuracy of identifying TF binding sites.

To include useful information from the bait motifs in the model, these motifs should not have similar binding preference, otherwise they will only contain repetitive sequence information and be treated as the same states by TRACE. We obtained hierarchical clustering information of position frequency matrices (PFMs) from the JASPAR database using the RSAT matrix-clustering tool to ensure all motifs included in the model are different. The motifs at the root of each tree encompass all the position-specific scoring matrices (PSSMs) of a cluster. These root alignments are the only PWMs that should be added to the TRACE model as baits. The root motif from the cluster that contains the TF of interest should also be excluded from the model. We scanned each root motif across the genome and ranked their numbers of occurrences to determine which motifs to be added in the model. For a N-motif model for a certain TF, the bait motifs will be (N-1) the most abundant root motifs

from the clusters that do not contain the TF of interest.

2.5.6 Evaluation

To assess the performance of TRACE and existing computational footprinting tools, we evaluated DeFCoM, BinDNase, CENTIPEDE, FLR, PWM score only, DNase2TF, HINT, PIQ and Wellington based on scores or P-values provided by each method. Candidate binding sites (motif sites) that overlapped with DNase-seq peaks confirmed by ChIP-seq were used as the positive set, and those not in ChIP-seq peaks but still overlapping DNase-seq peaks made up the negative set. Prevalence was calculated as number of active binding sites (positive set) divided by total number of motif sites (positive set and negative set).

To provide a fair comparison across all methods, we applied *de novo* methods to DNase-seq peaks (with 100bp flanking regions to each side) containing the same sets of motif sites that were included in motif-centric methods tests. For *de novo* methods, only the predictions overlapping with motif sites of tested TFs were included in our evaluation; candidate binding sites that were missing from their predictions were also included with an assigned minimum score. For motif-centric methods and PWM only evaluations, only candidate binding sites provided are assessed, thus all predictions were included in the evaluation. Annotations and corresponding scores or P-values were used to calculate the ROC AUC, ROC pAUC at a 5% FPR cutoff and PR AUC values for all TFs.

Permutation tests were performed by shuffling labels from footprinting prediction results. Multiple simulation tests were also included based on different levels of positive and negative samples separation and different positive example fractions. Scores for positive and negative groups were randomly drawn from the normal distribution of different means and standard deviations.

2.6 Software availability

TRACE is an open source software; the source code, trained models, and predictions are available on GitHub at https://github.com/Boyle-Lab/TRACE.

2.7 Publication

The study in this chapter has been published in Genome Research [70]: Ouyang, N., & Boyle, A. P. (2020). TRACE: transcription factor footprinting using chromatin accessibility data and DNA sequence.

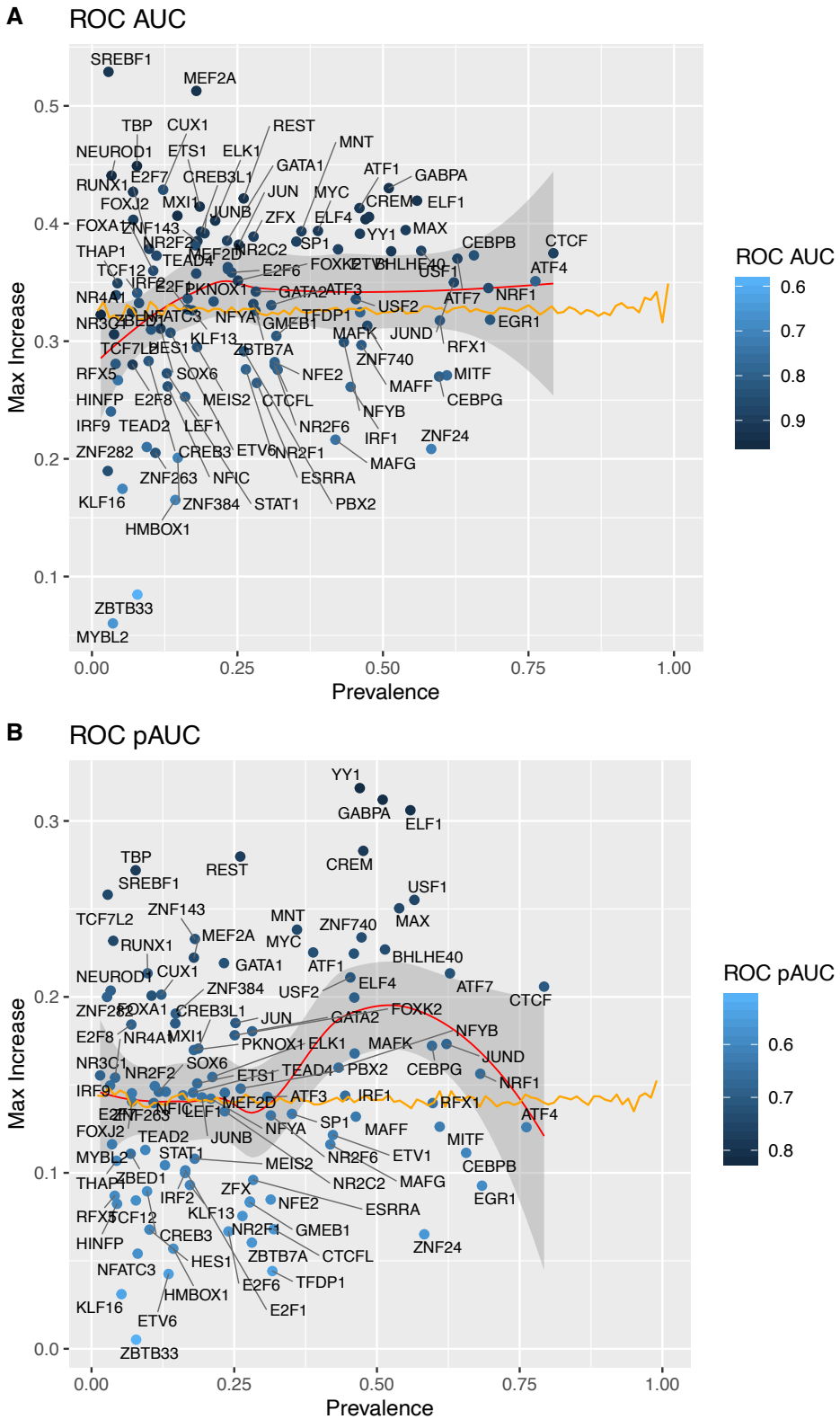

Figure 2.9: Increase of footprinting methods' best (A) ROC AUC and (B) ROC pAUC over permutation are not correlated with prevalence. Orange line is from simulation test using positive set from N(10, 8), negative set from N(0, 7).

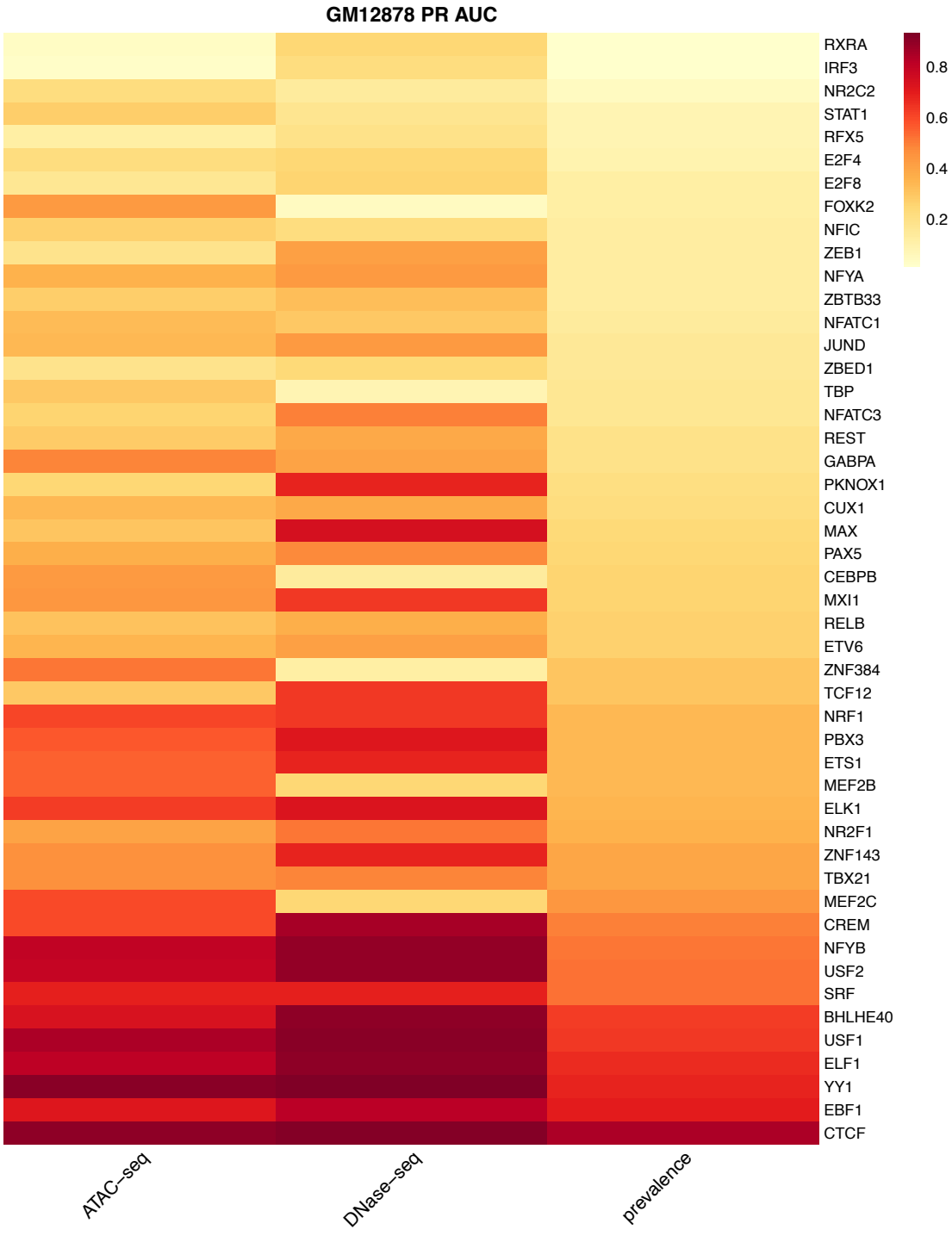

Figure 2.10: Heatmap of PR AUC of all TFs tested using DNase-seq and ATAC-seq data, sorted by prevalence.

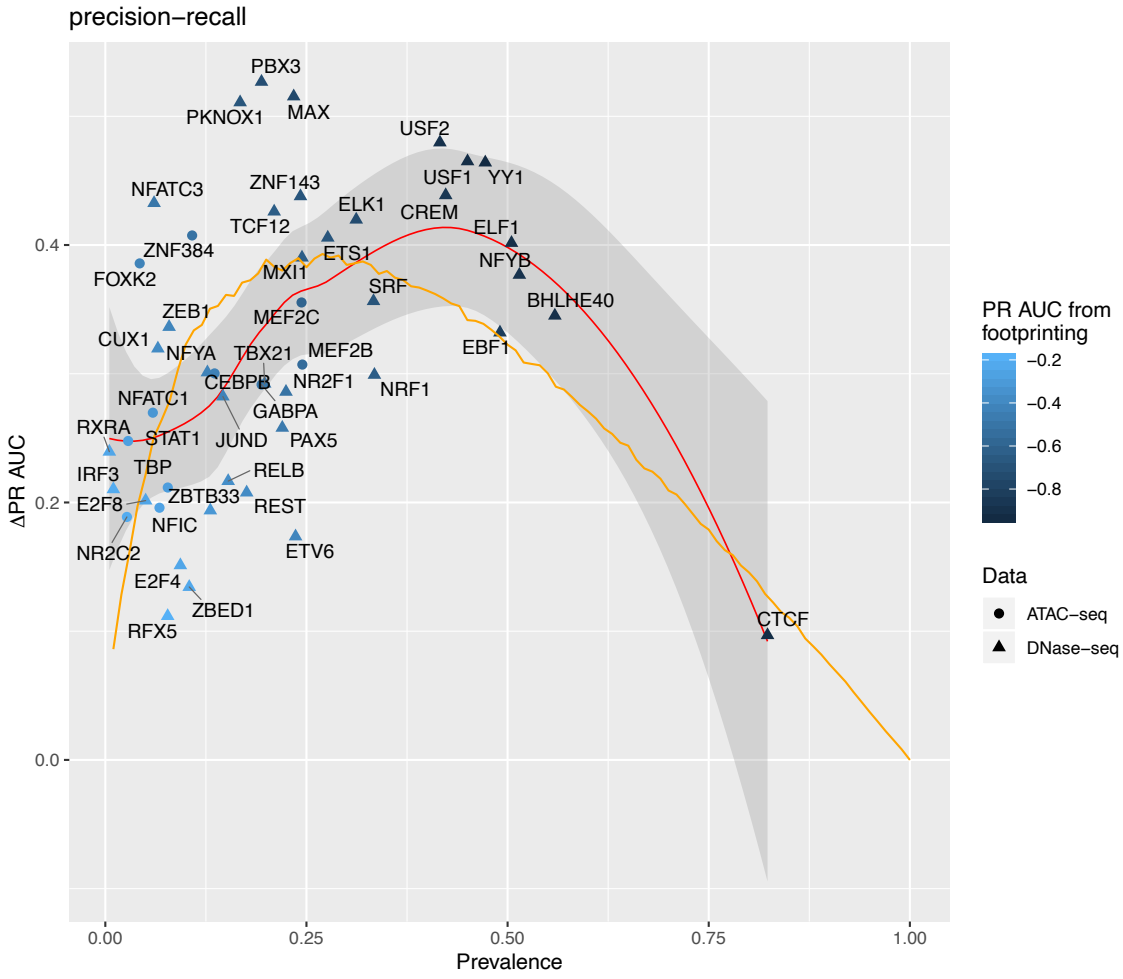

Figure 2.11: Performance improvement of TRACE model over permutation for each TF in GM12878, colored by its best PR AUC from DNase-seq or ATAC-seq data. Orange line is from simulation test using positive instances drawn from N(12, 9), and negative instances from N(0, 9) to demonstrate expected PR AUC trend as binding prevalence changes.

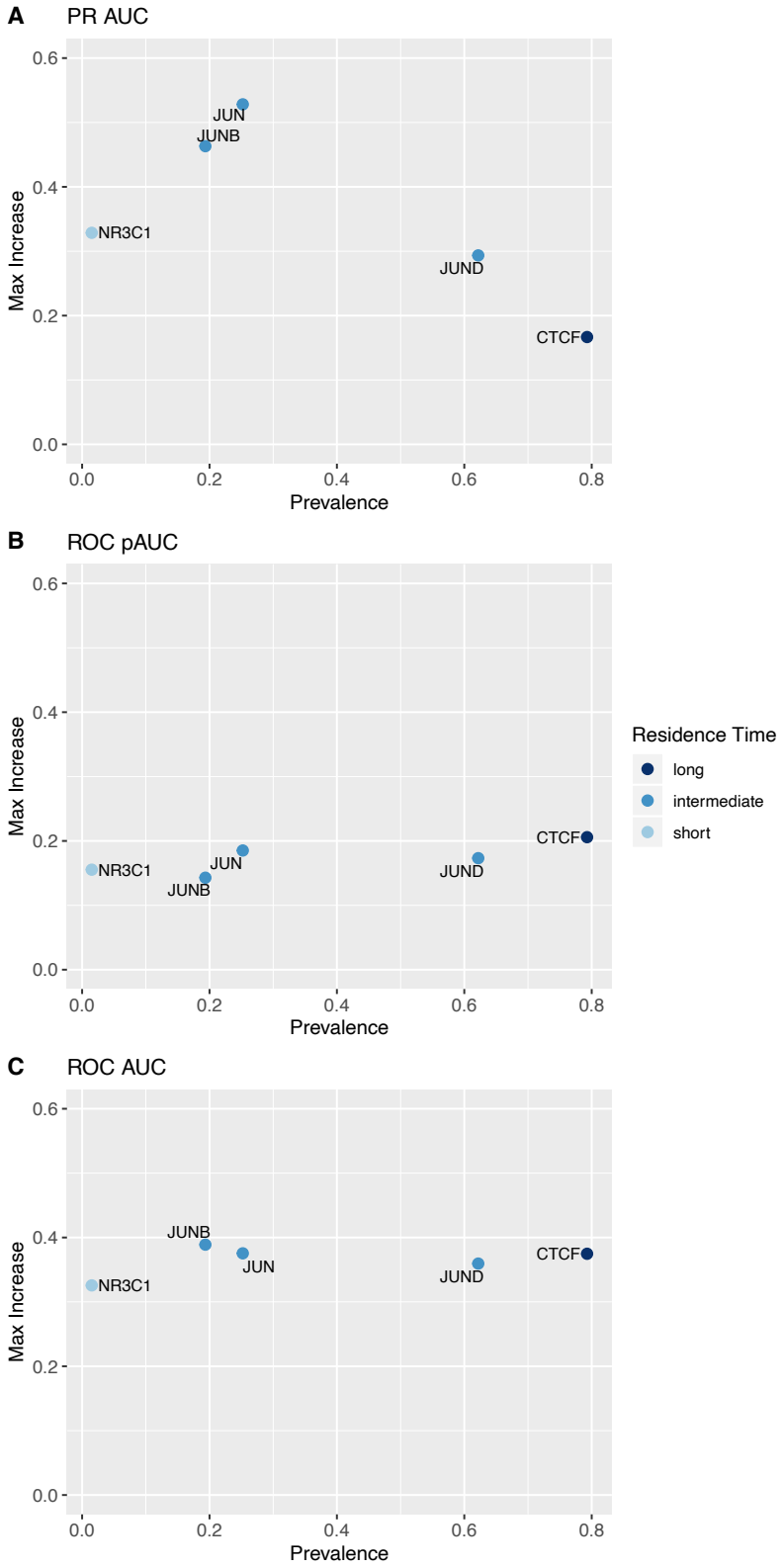

Figure 2.12: Comparison of footprinting performance increase on TFs with (A) short, (B) intermediate and (C) long residence time.

CHAPTER III

Characterizing Regulatory Variants by Fine-mapping Footprint QTLs

3.1 Abstract

Association fine-mapping of molecular traits has emerged as an essential method for understanding the function of genetic variation. Sequencing-based assays, including RNA-seq, DNaseI-seq and ChIP-seq data, have been widely used to measure different cellular traits and enabled genome-wide mapping of quantitative trait loci (QTLs). The disruption of cis-regulatory sequence, often occurring through variation within transcription factor binding motifs, has been strongly associated with gene dysregulation and human disease. We recently developed a computational method, TRACE, for transcription factor binding sites identification. TRACE provides powerful information for association mapping of footprint and regulatory variants utilizing DNase footprinting for precise genome-wide prediction of individual-specific transcription factor binding sites. Using binding activities predicted from TRACE and genome-wide genotypes of Yoruba lymphoblastoid cell lines (LCLs), we mapped footprint-variants significant associations, termed footprint QTLs (fpQTLs). Detection of fpQTLs provides a rich resource for the continued study of regulatory variants, leading to improved functional interpretation of noncoding variation.

3.2 Introduction

Transcription factors (TFs) can recognize and bind to DNA sequence, including cis-regulatory elements, to perform their regulatory activaty. Transcription factor binding sites (TFBSs) make up only 8% of the genome but contain 31% of genome-wide association studies (GWAS) identified genetic variants, suggesting they play a largely underappreciated role in human disease [71]. Binding of TFs to specific DNA sequences is fundamental for transcription regulation and the effect of regulatory variants on these regions is more directly interpretable. The functional effect of noncoding single nucleotide variants (SNVs) is often observed as a strengthening or weakening of individual transcription factor binding activity, thus identification of variation effects on transcription factor binding is key to understanding and interpreting the downstream consequences on gene expression and disease phenotypes.

Association mapping of quantitatively measurable molecular traits has emerged as a powerful approach for studying genetic variation. Genome-wide mapping of expression quantitative trait loci (eQTLs) has become an essential method to examine the impact of variants on gene expression and regulation [72, 24]. However, the underlying regulatory mechanism is not immediately evident due to linkage with causal variants rather than having a direct effect on transcription. One putative mechanism is that variants can change the likelihood of transcription factor binding and thereby affect regulatory networks. Previous study demonstrated that mutations in transcription factor binding sites alter transcription factor binding affinity, and are associated with human diseases, including cancer and type 2 diabetes [22].

Binding of TFs to DNA often exhibits regions that are protected from DNaseI digestion, known as footprints. DNase I hypersensitive site sequencing (DNase-seq)

measures open chromatin regions where DNase I cuts at higher frequencies, allowing for genome-wide footprinting. We recently developed a computational footprinting method, TRACE that predicts footprints and labels binding sites for desired TFs [70]. Our model integrates both DNA accessibility (DNase-seq or ATAC-seq) and genome sequence information (PWMs to predict footprints, and binding scores can reflect binding activity changes caused by variation in genotype and chromatin accessibility by utilizing DNase-seq or ATAC-seq and DNA sequence from different individuals. Here We leverage DNase-seq, RNA-seq, and genotypes for the HapMap Yoruba lymphoblastoid cell lines (LCLs) to enable genome-wide identification of genetic variant-footprint associations in a base pair resolution and examine their impact on phenotypic variation [20, 73, 21, 74]. We refer to loci with significant associations in TF binding between genotypes and inter-individual variation as "footprint QTLs", or fpQTLs. This genome-wide mapping of fpQTLs provides additional information for the functional interpretation of human noncoding variation.

3.3 Results

3.3.1 Genome-wide identification of fpQTLs

Genetic variation can alter TF binding sequences and therefore affect the likelihood of TF binding. As an example, rs1338681 is a SNP within a CTCF binding motif (Figure 3.1C), where the alternative allele disrupts the motif and leads to reduced binding affinity at the footprint (Figure 3.1D). Genome-wide mapping of footprint-variant association can be accomplished by utilizing TFBSs prediction datasets. TRACE employs both chromatin accessibility data and sequence information to assess binding activity at footprints, allowing for the investigation of the impact of variants on TF binding activity. We generated individual digestion signal profile and PWM scores for 57 HapMap Yoruba lymphoblastoid cell lines (LCLs),

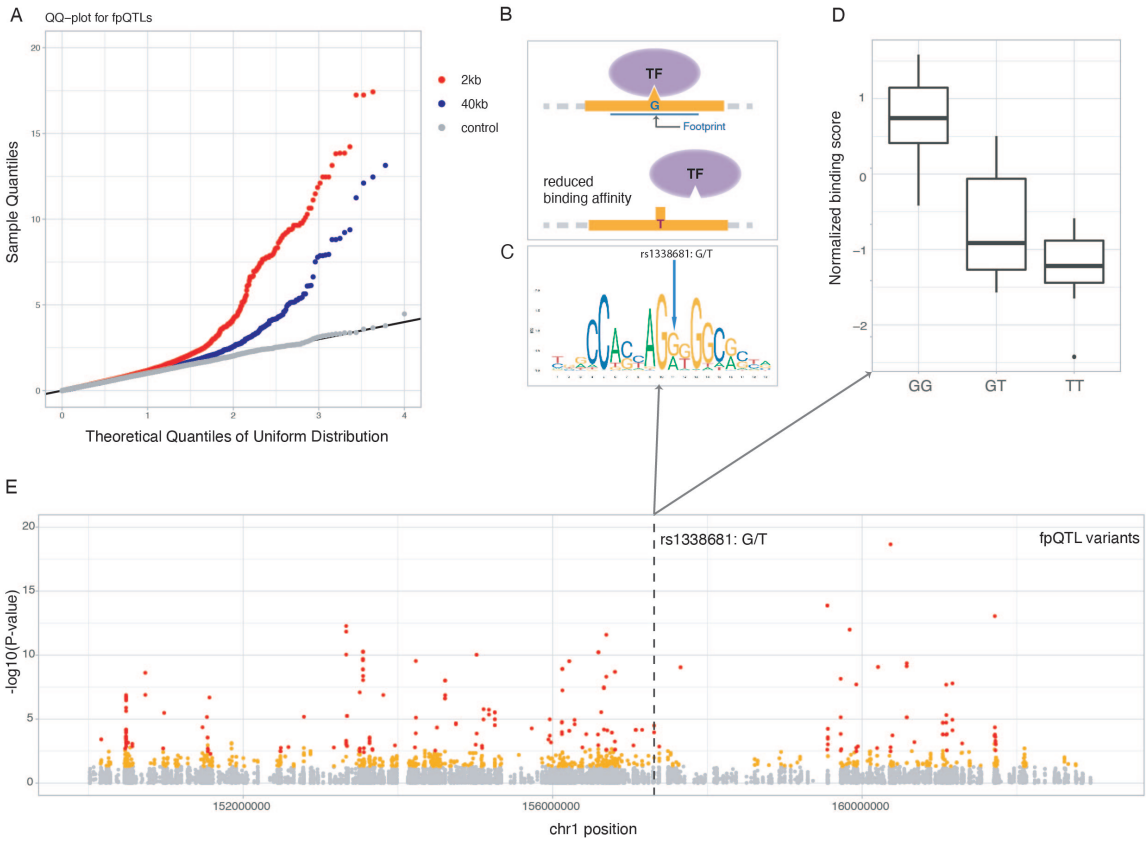

Figure 3.1: Genome-wide detection of fpQTLs and an example fpQTL SNP. (A) QQ-plots for all tests of association between footprint scores and variants within 2kb (red) and 40kb (blue) regions surrounding the target footprint. Permutation controls (gray) confirmed that observed p-values are uniform under the null (B, C, D) Example of a fpQTL rs1338681: (B) The T allele reduced TF binding affinity through (C) disrupting the CTCF binding motif, and (D)is associated with reduced binding score. (E) -log10 p values of associations of variants-footprint tested, significant fpQTLs at 10% FDR are in red color.

using available DNase-seq data and genome-wide genotypes, following the same data processing pipeline as TRACE with some adjustments that account for genetic variation. These processed signals were subsequently used to predict individual-specific TFBSs with a modified version of TRACE. Individual-specific marginal posterior probability for each region being bound by a certain TF was also assigned to the corresponding predicted footprint.

For each TFBS within a sample, the predicted score from TRACE was used as a quantitative trait to estimate the sample-specific TF binding activity at that

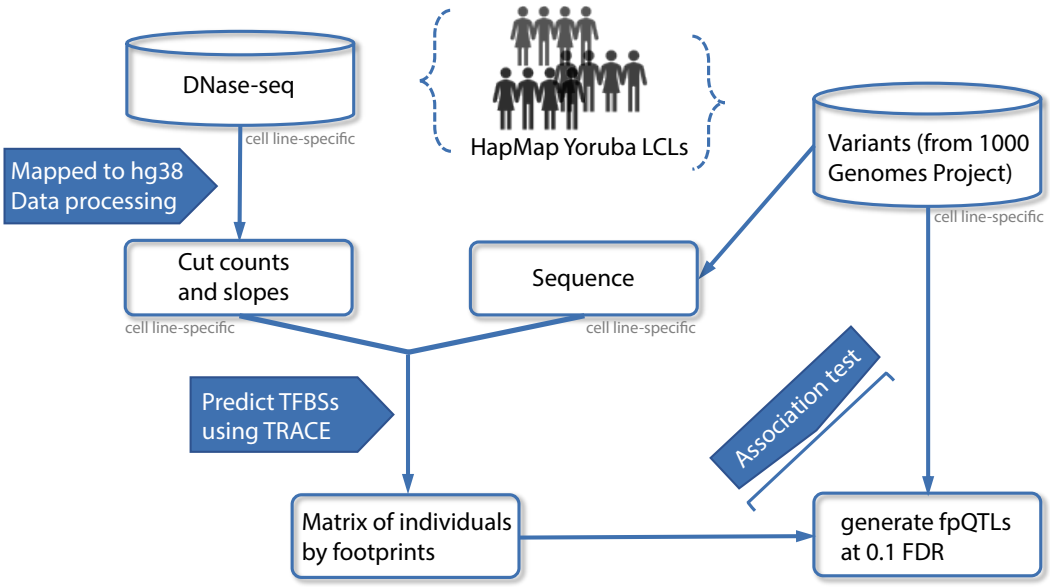

Figure 3.2: Workflow of fpQTLs identification.

footprint. Given genome-wide TFBS prediction and genotype, an association test was completed between binding scores at each footprint and and all SNP genotypes in a cis candidate window of 2kb and 40kb centered at the target binding site (Figure 3.2). Significant footprint-genetic variant pairs were identified where TF binding affinity was significantly correlated with genome variants at 10% FDR, thus referred to as footprint QTLs (fpQTLs). In general, fpQTLs within in a 2kb window showed a higher significance level than fpQTLs in a 40kb window (Figure 3.1A). In the following analyses, only include fpQTLs identified in a 2kb window are included.

3.3.2 Proximal and distal fpQTLs

Since different alleles within a motif can directly lead to the creation or disruption in TF binding sequence, we hypothesized that fpQTL SNPs lying inside the target footprints could have a greater effect on TF binding compared to those outside of

footprint. Similarly, the distance of a gene variant relative to motif sites can impact its effect size on binding activity. fpQTLs generated previously were separated into three groups: 5.15% of SNPs were locating inside associated footprints, 29.2% were outside of the target footprint but fell within a +/-100 bp window centered at the footprint, and others were outside the +/- 100bp window but were within a +/-1 kb window (Figure 3.3A). Each footprint can be linked to multiple SNPs, however, 12.5% of fpQTL loci were significantly linked to SNPs that fell within the target footprint, and 55.4% had associated SNPs lying in a +/-100 bp window centered at the footprint.

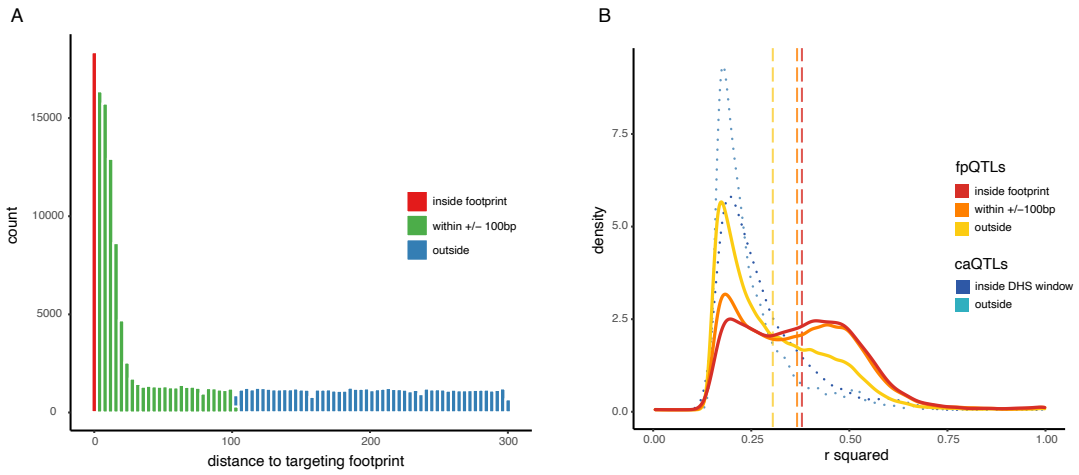

Figure 3.3: Properties of Proximal and distal fpQTLs (A) Distribution of fpQTLs by variant' distance to target footprint. (B) Distribution of effect sizes for proximal and distal fpQTLs (solid curves) and caQTLs (dotted curve) showed larger effect size for more proximal QTLs. Dashed lines are average effect size for the three positional categories of fpQTLs.

To test if distance factors into the impact of genetic variants on footprint, R squared from the linear model were used as effect size measurements. The distribution of effect size in three positional categories indicated that fpQTLs inside the

binding sites tend to have a larger effect on binding activity, and more distal fpQTLs tend to have smaller effect size (Figure 3.3B, solid curve). This is consistent with our expectations that formation or disruption of a binding sequence can exert the greatest effect on TF binding. The same trend was observed with chromatin accessibility QTLs (caQTLs) as proximal SNPs have a more extreme effect on chromatin accessibility than more distant ones (Figure 3.3B dotted curves). [73].

3.3.3 Functional significance of fpQTLs

The high number of detected fpQTLs provide a rich resource to generalize genomic properties of variants, including their enrichment with other cellular traits or diseases. fpQTLs share a large overlap with caQTLs that were also identified in LCLs using the same DNase-seq dataset. Utilizing GWAS catalog data, many fpQTL SNPs were found to overlap with GWAS-associated SNPs for various disease/traits including Type 2 diabetes and cancer such as prostate cancer, consistent with our prior knowledge, as well as blood protein levels, mean corpuscular volume and serum metabolite levels. Fisher's exact test also showed significant enrichment in some traits or diseases such as IgM levels and systemic lupus erythematosus (Figure 3.4C).

Several computational tools have been developed to prioritize genetic variants and predict their functional consequences based on a variety of genomic features. To study genomic properties of fpQTLs, fpQTL SNPs were queried against RegulomeDB [75, 76] and ExPecto [77] annotations, and functional scores were obtained for each SNP. Compared to randomly selected non-fpQTL variants, fpQTL SNPs showed higher functional significance in both scoring systems (Figure 3.4A, 3.4B).

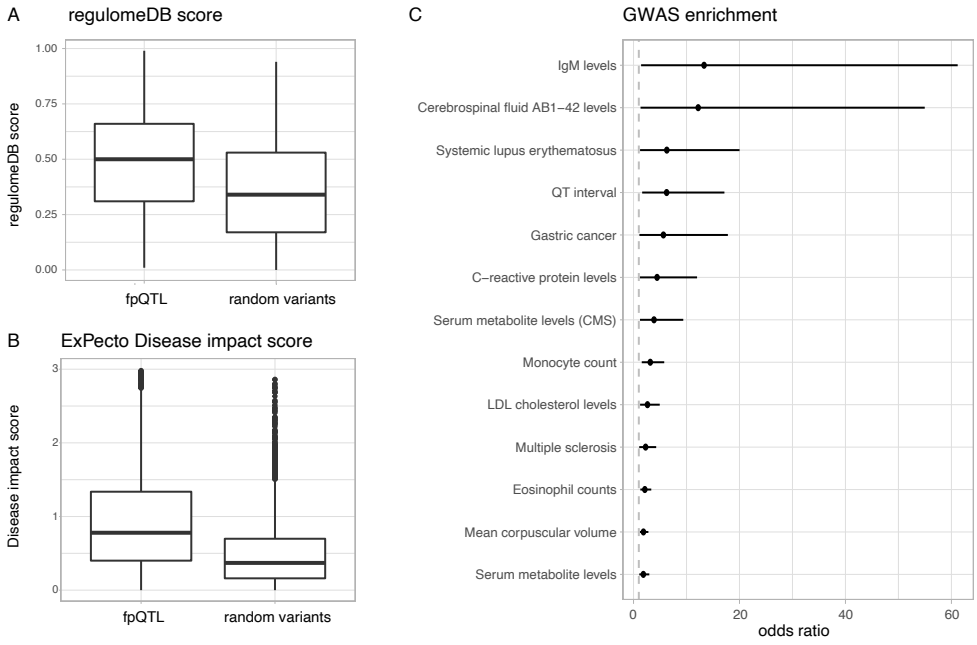

Figure 3.4: Functional significance of fpQTLs (A) RegulomeDB scores of detected fpQTL SNPs and random selected variants. (B) ExPecto disease impact score of detected fpQTL SNPs and random selected variants. (C) GWAS catalogue diseases or traits enrichment in fpQTLs from Fisher exact test (P value < 0.02 and odds ratio < 1). Odds ratio and its 95% confidence interval are shown as dots and horizontal lines.

3.3.4 Cis-regulation of gene expression by fpQTLs

Genetic variations that alter the likelihood of transcription factor binding can thereby affect regulatory networks, thus we hypothesized that a substantial portion of fpQTLs will also contribute to variation in the expression level of nearby genes. To examine the effect of creation or disruption of transcription factor binding as a potential mechanism involved in the linkage between genotypic and phenotypic variation, tissue specific eQTLs from Epstein-Barr virus (EBV) transformed lymphocytes were retrieved from the Genotype-Tissue Expression (GTEx) project. We found that 12.8% of fpQTL SNPs are also significantly associated with variation in the expression levels of nearby genes.

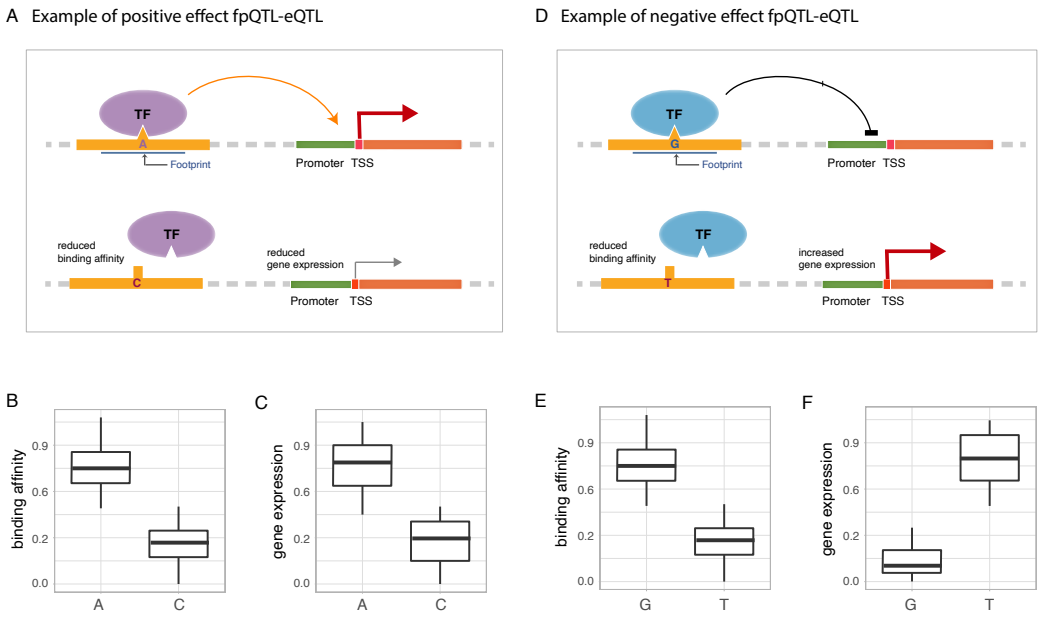

Figure 3.5: Simplified schematic of (A, B, C) positive and (D, E, F) negative impact fpQTL-eQTL. (A) A positively directed genetic variant disrupts TF binding at footprint and decreases expression of the gene regulated by that regulatory element. the A/C allele is associated with (B) level of TF binding and (C) gene expression. (D) A negatively directed genetic variant disrupts TF binding at footprint and induced gene expression. the G/T allele is associated with (E) level of TF binding and (F) gene expression.

We observed that the joint fpQTLs-eQTLs SNPs are enriched in a closer window surrounding the TSS, 9.2% are within a +/- 1kb window around the TSS, 78.6% are outside +/- 1kb but within a +/- 100kb window. SNPs closer to the TSS tended to be more significant than distal ones (Figure 3.6). However, unlike dsQTLs-eQTLs pairs, where usually increased chromatin accessibility is associated with increased gene expression levels, the direction of the SNP's impact on TF binding activity and gene expression is highly diverse due to the diversity of activating or repressing functionality of different TFs. To study the correlation between the impact of SNPs on TF occupancy and gene expression levels, fpQTLs and eQTLs regression slopes were

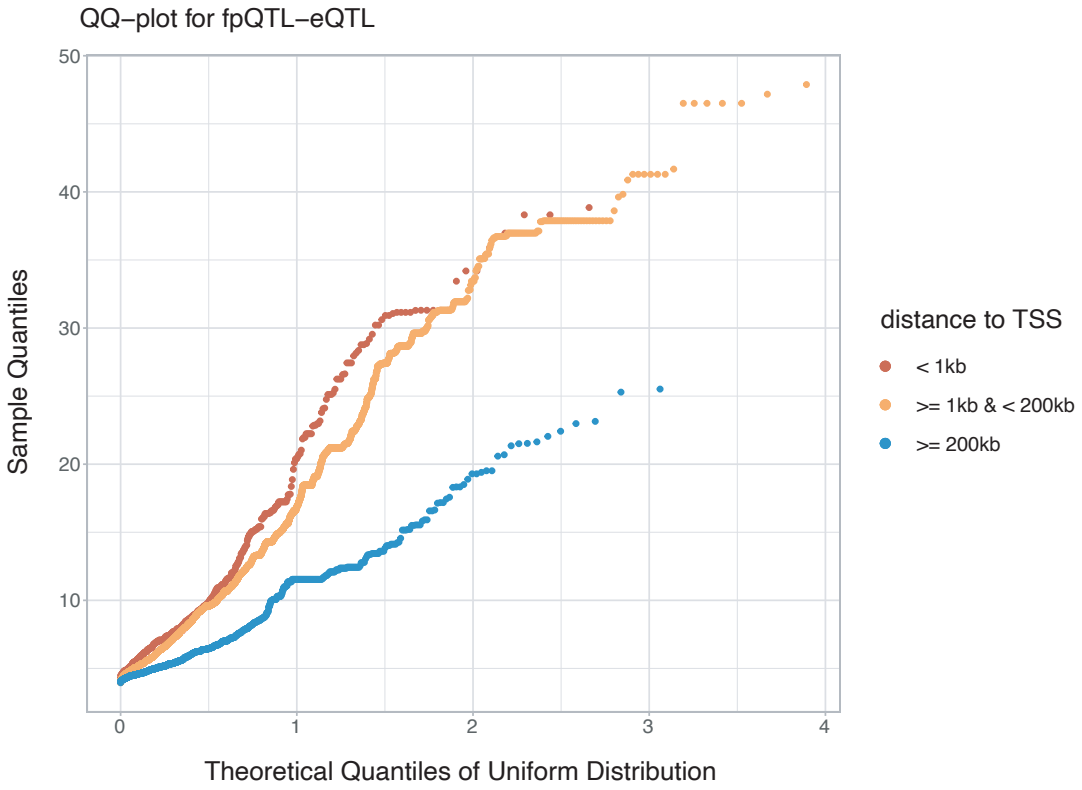

Figure 3.6: QQ-plots for fpQTL-eQTL, grouped by distances to TSS. The dots don't start off following the y=x line because only the fpQTL-eQTLs with significant p values were plotted.

calculated from linear models of footprint scores and gene expression respectively, and the sign of the slope indicates either an up-regulating or down-regulating effect on binding affinity or gene expression. Congruent signs for both slopes suggest the SNP associated with TF binding disruption is also associated with reduced expression of a nearby gene (Figure 3.5A, 3.5B, 3.5C). Conflicting directions of the SNPs' effect on TF binding and gene expression indicate that reduced TF binding activity is linked to increased gene expression (Figure 3.5D, 3.5E, 3.5F). To determine whether a specific TF has directional impact on gene expression, as defined as the

changes in binding activity tending towards have a consistently positive or negative effect on gene expression across all detected fpQTLs, we performed a binomial test on each TF separately to characterize direction-specific TFs with significant association between regulating directionalities of TF binding activity and transcription level. There were 43 TFs with Bonferroni-corrected p-values smaller than 0.05 and a fold change greater than 2 (Figure 3.7A). Among these TFs, nuclear transcription factor Y, beta subunit (NFYB) is a most significant negative effect TF but has a moderate fold change of positive loci / negative loci. This suggests that, although NFYB loci are more enriched for having a negative regulatory effect, some still exhibited up-regulation on gene expression. In fact, the NF-Y complex is known as an activator protein that can bind promoter regions and regulate transcription through heterodimers or heterotrimers formation. However, the NFYB subunit of the NF-Y complex has been previously reported to be a repressor in multi-omics analysis [78, 79]. Another significant directional TF is eomesodermin (EOMES) that act as a transcriptional activator and is involved in differentiation of CD8+ T cells [80, 81, 82], consistent with its high positive log fold change value.

3.4 Discussion

Association studies using sequencing-based assays such as RNA-seq, RIBO-seq, ChIP-seq and DNaseI-seq have revealed plentiful QTLs involved in gene expression [74] and transcriptome [83, 84, 85], ribosome occupancy and protein abundance [86], histone modification [87], DNA methylation [88] and chromatin accessibility [73]. Here we define footprint QTLs (fpQTLs) as genetic variations that are significantly associated with footprint binding affinity, providing additional information on the regulatory functions of genetic variation. Both proximal and distal fpQTLs were

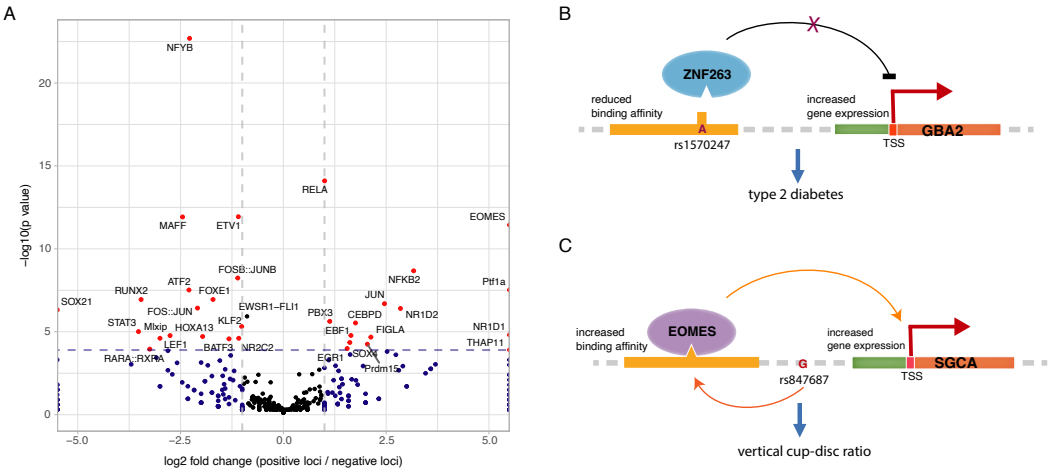

Figure 3.7: Properties of fpQTL-eQTL (A) Volcano plot for TFs on their gene regulation directionality. (B) rs1570247 disrupts ZNF263 binding and is correlated with increased expression of GBA2 gene. This SNP is also linked to type 2 diabetes. (C) rs847687 is associated with increased binding of EOMES and increased expression of the SGCA gene. This SNP is linked to vertical cup-disc ratio trait.

identified using an association test. Although a substantial portion of fpQTL SNPs enriched closer to the targeting footprint with a greater effect size, a number of distant regulatory relationships also exist.

fpQTLs from this study provide a rich source of information for examination of SNP genomic properties and evaluatation of the regulatory potential of untested variants to achieve a better understanding of regulatory mechanisms. Among fpQTL SNPs, many also overlapped with eQTL SNPs and potential disease-causing SNPs identified by GWAS. For example, rs1570247, a causal SNP correlated with type 2 diabetes, disrupts the binding of the previously reported transcriptional repressor ZNF263 [89], and is associated with increased expression of the Glucosylceramidase Beta 2 (GBA2) gene (Figure 3.7B). This provides a potential underlying mechanism for the linkage between genetic variants, gene expression and disease phenotype, as

the alternative allele can alter ZNF263 binding. Subsequent disruption of ANF263 binding can affect GBA2 gene expression along with other co-factors, and cause the disease phenotype. Another example is rs847687, which increases the binding of the activating transcription factor EOMES that was discussed in the previous section, and is associated with increased expression of the Sarcoglycan Alpha (SGCA) gene (Figure 3.7C). This has also been identified as a disease risk locus and was linked to vertical cup-disc ratio through the same gene by a GWAS study [90], postulating that EOMES binding might be involved in the cup-disc ratio phenotype via transcriptional regulation.

3.5 Methods

3.5.1 Data and software

DNase-seq data for 57 HapMap Yoruba lymphoblastoid cell lines (LCLs) in fastq formats were retrieved from http://eqtl.uchicago.edu/dsQTL_data/. Reads were remapped to hg38 using the variant-aware aligner VG toolkit [91]. Corresponding variant file in VCF format were downloaded from 1000 Genomes. 780 PWMs and cluster information were downloaded from the JASPAR database [57].

3.5.2 Modified TRACE workflow

Signal processing of DNase-seq data followed the TRACE pipeline, which includes cutting bias correction, normalization, and local regression smoothing. The original program was modified to accept sequence information with variants, and the motif score feature can reflect different alleles. To generate individual-specific TFBSs, a 10-motif was trained for each TF and the same model was applied on all individual datasets.

3.5.3 fpQTL association testing

For each predicted footprint and its individual-specific binding score from TRACE, we tested for association of the binding score with the genotypes of all SNPs where the minor allele frequency was greater than 5% within a cis-candidate region (2 kb and 40 kb cis-candidate windows centered on the target footprint). For each footprint and each SNP falling within the candidate window, standard linear regression between genotype and binding score was performed in R, and a p-value was generated by testing the alternative hypothesis that the slope in the linear regression model is not 0.10% FDR, which was estimated using the "qvalue" R package, was used as a threshold to select significant footprint-variant pairs.

3.5.4 caQTLs generation

To generate caQTLs, DNase-seq reads that were mapped to the hg38 in the previous section were processed following the same pipeline described in Degner et al. (2012) [73]. Reads starting within the 5 bp window centered at SNPs were discarded. Raw DNase sensitivity was calculated by counting the number of reads falling in each non-overlapping 100bp window, normalized by total number of mapped reads and mappability. Additional normalization steps for hypersensitivity phenotypes include GC-content correction, mean-center and variances scale, and quantile-normalization to standard normal distribution. Unidentified confounders were removed with PCA, thus 4 PCs were removed. Subsequent association testing was conducted to generate dsQTLs at 10% FDR.

3.5.5 Functional significance analysis for fpQTLs

Genomic annotations and function scores were collected from RegulomeDB [75, 76] and ExPecto [77].All identified fpQTL SNPs and randomly selected background non-

fpQTL SNPs were queried, and lists of RegulomeDB scores and ExPecto disease significant scores were obtained. GWAS catalogue data [92] built in Dec 2021 were retrieved from the GWAS Catalogue website. Fisher's exact test was performed to assess the enrichment of disease and other traits. Odds ratio and its 95% confidence interval were calculated for each disease/trait.

3.5.6 Impact direction of the fpQTLs-eQTLs SNPs

The impact direction of the fpQTLs-eQTLs SNPs is defined as the consistency of the signs of the slopes of the variant-footprint and variant-expression linear regression model. It can be interpreted as the up or down-regulation effect of the increased or reduced TF binding resulting from the alternative allele (ALT) relative to the reference allele (REF) on gene expression (i.e., if it has a negative impact direction, the disruption of TF binding will lead to increased expression of a nearby gene). Log fold-change for each TF was calculated by taking the log-ratio of the number of fpQTLs-eQTLs SNPs associated with that specific TF that have a positive impact direction to the number of SNPs with negative effects.

CHAPTER IV

Integration of TFBSs with 3D Chromatin Structure to Understand CTCF Looping Regulation

4.1 Abstract

Genetic regulation relies on transcription factor binding as well as spatial folding of chromosomes and chromatin looping. 3D chromatin structure plays a pivotal role in gene expression by bringing distal regulatory elements into spatial proximity. Transcription factors can co-bind with one other and recruit other regulatory proteins through direct or indirect physical interaction. CTCF and cohesin complex are known to work cooperatively to mediate the formation of 3D chromatin structure and achieve specific regulatory outcomes. Illumination of TF co-localization is central to understanding gene regulatory mechanisms. Here I explore how TF co-binding patterns and CTCF-related molecular complexes interact to determine regulatory effects and impact chromatin conformation.

4.2 Introduction

Transcription factors (TFs) act cooperatively to regulate gene expression under varying conditions across cell types. Determining TF co-binding patterns is essential in understanding architecture of the gene regulatory network in a cell. One way of mapping this network is through ChIP-seq assays that provide genome-wide binding

profiles of a large number of transcription-related factors and analysis of these data has revealed complex co-binding patterns [93]. This combinatorial binding of TFs exhibits distinct genomic properties and drives regulatory function in a context-specific fashion. However, in depth systematic investigation on TFs co-localization can be challenging due to the high dimensionality of the large data sets. Furthermore, ChIP-seq data based TF cooperativity studies cannot discern between the two modes of co-occupancy: protein-protein interaction or direct DNA binding. An overall understanding of the mechanisms and consequences behind cooperative TF binding is still needed.

The insulator protein CTCF is known to serve as a chromatin looping mediator in forming 3D genome structure. CTCF can mediate chromosomal contacts and plays a critical role in genome organization [94]. These changes in chromatin interaction through 3D conformation of the genome allows regulatory elements where TFs bind to achieve their gene regulation properties across long distances. 3D chromatin interactions can connect regulatory elements to target genes and regulate gene expression. CTCF binding sites are present in many chromatin loop boundaries, and variations in CTCF occupancy are associated with chromatin looping dynamics. In addition, previous studies have showed that depletion of CTCF binding can lead to disruption of these chromatin interactions [95, 96].

A few TFs have been shown to co-localize with CTCF and can regulate binding, participate in CTCF looping, and help modulate downstream looping effects. However, no comprehensive analysis of molecular complexes around CTCF has been performed. The most widely explored co-factors that co-localize with CTCF are the cohesin complex proteins, consisting of SMC1, SMC3, RAD21 and SA1/2 subunits. Cohesin has its established role in chromatid cohesion and CTCF has been shown

to be required for cohesin subunit Scc3/SA1 recruitment to Chromatin [42, 43]. Cohesin is essential in CTCF-mediated chromatin loops stabilization and is critical for CTCF function genome-wide [97]. To deepen our understanding of how CTCF mediates higher-order chromatin organization, the factors that are involved in CTCF complex and CTCF mediated loops need to be further explored.

Using transcription factor binding information as raw data, here I employed an artificial neural network called Self-Organizing Maps (SOMs) to identify "clusters" of TFs and define co-binding patterns. These were further characterized by their correlation with chromatin states and histone modifications to improve our understanding of the mechanisms behind the downstream regulatory outcomes. I also studied TF enrichment at chromatin loop anchors to investigate key factors in mediating the 3D organization of chromatin and 3D-cooperation between chromatin loops and TF co-binding complexes.

4.3 Methods

4.3.1 SOM training and plotting

At each DNase I hypersensitive site (DHS), the binding state of all TFs at that region was encoded into binary states (bound / unbound) determined by either TFBSs predicted from TRACE or ChIP-seq peak of same group of TFs overlapping. These binary binding information vectors ware used as input into an empty SOM which then went through unsupervised training. After training, SOMs can detect clusters of TFs and generate nodes of different TF co-binding patterns. These co-localization patterns were represented by hexagons in SOM map which can be color coded by values representing different biological signals including binding activity of a specific TF and regulatory element enrichment (Figure 4.1A). Each hexagon denotes a generated patterns and is supported by multiple regions that were bound

by TFs involved in that pattern (Figure 4.1B). To plot enhancer enrichment, we first intersected all regions with enhancer locations, and calculated the mean value of each binary overlapping information vectors of regions labeled with each pattern. Then the signal can be used to color each hexagon to represent the fragment of enhancer regions.

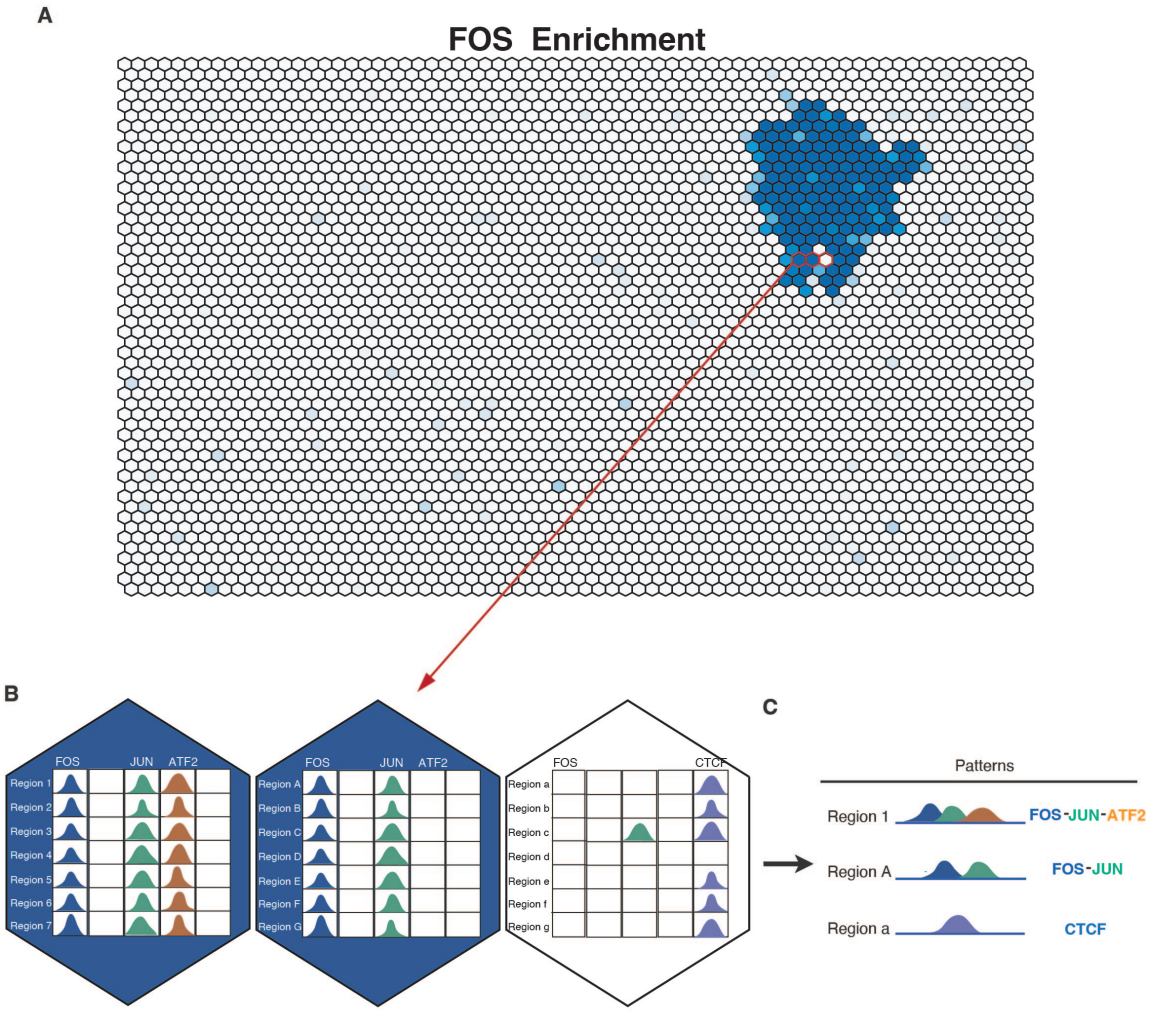

Figure 4.1: SOM map and co-binding patterns (A) SOM map colored by FOS binding values (darker blue represents higher FOS binding). (B) Each hexagon represents a co-localization pattern. (C) Overlapping TF binding were assembled into different co-binding patterns.

4.3.2 ChIA-PET data analysis

ChIA-PET data for GM12878 and K562 were retrieved from the ENCODE download portal in fastq format. After quality check, paired reads were aligned using BWA mem with default parameters. Unmapped and secondary reads, as well as reads with quality scores less than 30 were filtered out using Samtools view. Processed data from all biological and technical replicates were concatenated. MANGO pipeline was then utilized to further analyze the data to generate significant loop interactions.

4.3.3 Epigenetic properties at loop anchors

ChIP-seq data for GM12878, and K562 cells in bigWig format were downloaded from the ENCODE download portal. The rtracklayer R package [98] was used to retrieve histone modification signals at each location in 20-kb windows surrounding the loop anchors. For each histone mark and cell line, we also compared the average signals in left and right 10-kb portions of each window region. If the right half has higher signal than left, we will flip the signal track direction. Mean signal at each location was calculated and then normalized by mean signals at randomly selected 20-kb windows across the genome. This process was applied to promoter and enhancer enriched loops anchors separately and the processed score vectors were plotted as line graphs for each cell type using ggplot2.

4.3.4 TF enrichment at loop anchors and cis-regulatory elements

Core 15-state model chromHMM chromatin states annotations were downloaded from Roadmap Epigenomics Consortium [99, 100]. Their annotations were used to label cis-regulatory element at chromatin loop anchors by intersecting significant loop anchors generated from previous section and regulatory element annotations. To study the enrichment of TF at certain group of loop anchors (group X) such as

enhancer loop anchors and promoter loop anchors, we first generated a background group. We counted total number of regions in group X and randomly selected the same number of regions from the pool of all loop anchors as background group. We next calculated the number of regions in group X and background group having a specific TF binding respectively. By comparing the counts from a certain group and background, we can show which TF tend to be enriched or depleted in anchor group X.

4.4 Results

4.4.1 TF co-binding patterns from Self-Organizing Map

Existence of TF co-localization may result from either direct TF binding at neighboring sites along the DNA strand, or through protein-protein interactions. ChIP-seq data is widely used as main TF binding information source in most co-occupancy studies; however, using ChIP-seq data alone cannot differentiate the two modes of TF co-binding. In contrast, TRACE can predict footprints of TFs that bind directly to DNA by accessing chromatin accessibility pattern and sequence binding preference [70]. By utilizing TFBSs data, more TFs can be included to help distinguish between DNA-directed co-binding and co-localization due to protein-protein interactions. This allows us to work on a larger number of TFs, considering the limited availability of ChIP-seq data for TFs, and focus on the combinatorial effects of neighboring TFBSs on regulatory outcomes. Binding information from footprinting can generate combinatorial binding for more than 400 TFs, the high dimensionality of the data makes it very challenging to study TF co-binding patterns. Here, an artificial neural network called Self-Organizing Maps (SOMs) was employed to identify "clusters" of TFs and define co-binding patterns [101].

Binding sites for 425 TFs in K562 were generated using TRACE. K562 was used

in this study because it has the most TF ChIP-seq data available. At each genomic region, whether the binding sites of TFs overlapping the region was encoded into binary states. The binary binding information were used in SOM training in order to cluster all overlapping binding into sets of similar co-binding patterns. A SOM map comprised of a series of "neurons" was generated by SOM algorithm based on mutual overlap of TF binding. Each neuron contains a common co-binding pattern and was assigned to a node in the map grid (Figure 4.1A).

The SOM captured a large variety of co-localization patterns and detected some previously reported TF cooperativities. For example, JUN-FOS-ATF co-binding pattern were identified from SOM. JUN and FOS are activating protein 1 (AP-1) transcription factors. ATF often form complexes with AP-1 proteins and bind to TPA-responsive elements (TREs)-like sequences together [102]. CTCF-ZNF143 co-binding were also detected, which is consistent with previous studies. ZNF143 is known to be co-localized with CTCF, and knocking down ZNF143 can lead to loss of chromatin interaction at individual loci [103, 104].

4.4.2 co-binding patterns from ChIP-seq peaks and footprints

To distinguish between TF co-localization due to DNA-directed co-binding and protein-protein interactions, we applied SOM on both TFBSs data (footprint SOM) and ChIP-seq data (ChIP-seq SOM) and found low level of agreement between TF co-binding patterns identified from footprints and ChIP-seq peaks (Figure 4.2). For example, for genomic regions with CTCF binding that were determined by ChIP-seq peaks, only 35.7% of them were labeled with CTCF involved patterns from footprint, and for footprint labeled CTCF-bound regions, 72.6% of them were also interacted with CTCF from ChIP-seq data (Figure 4.2A, 4.2B).

The difference in footprint SOM and ChIP-seq SOM labeled patterns can be

caused by TF co-localization due to direct DNA binding or protein-protein interactions. Our footprinting method, TRACE investigates chromatin accessibility pattern and motif sequence to detect TFBSs, its predicted binding sites reflect there is direct TF binding at DNA. Therefore, a footprint-based SOM will only cluster together TFs binding at neighboring sites on the DNA. ChIP-seq enriches for DNA regions interacted with specific proteins but not necessarily by direct binding of that protein, so co-localization detected by ChIP-seq SOM might be formed by protein-protein interactions. For example, cohesin proteins like RAD21 lacks a direct DNA-binding domain, instead they function through interacting with CTCF. CTCF-cohesin protein patterns can be identified by ChIP-seq SOM but not footprint SOM. The footprint SOM can help filter out protein-protein interactions and allows study on cooperativity of TFs that are all directly binding at DNA.

With the binding mechanism difference from direct DNA-protein interaction and protein-protein interaction, a low level of consistency between co-binding modules identified by the footprint SOM and ChIP-seq SOM were observed, along with low overlapping rate between ChIP-seq peaks and binding motif. Previous studies have shown that most ChIP-seq peaks lack the TF's sequence motif [105], suggesting in many ChIP-seq peaks there are not direct TF interaction with DNA. Motif information is one of the features used in our TFBSs prediction, so the low agreement between ChIP-seq peaks and motif sequence is also consistent with lack of predicted binding sites in ChIP-seq peaks. In fact, our comparison also showed that very small fraction of ChIP-seq peaks for specific TFs contain binding sites predictions from TRACE (Figure 4.2C). It is consistent with our observation that the ChIP-seq identified co-localization has a more extreme small portion that have similar co-binding patterns by footprint than fraction of footprint patterns that were consistence with

ChIP-seq patterns.

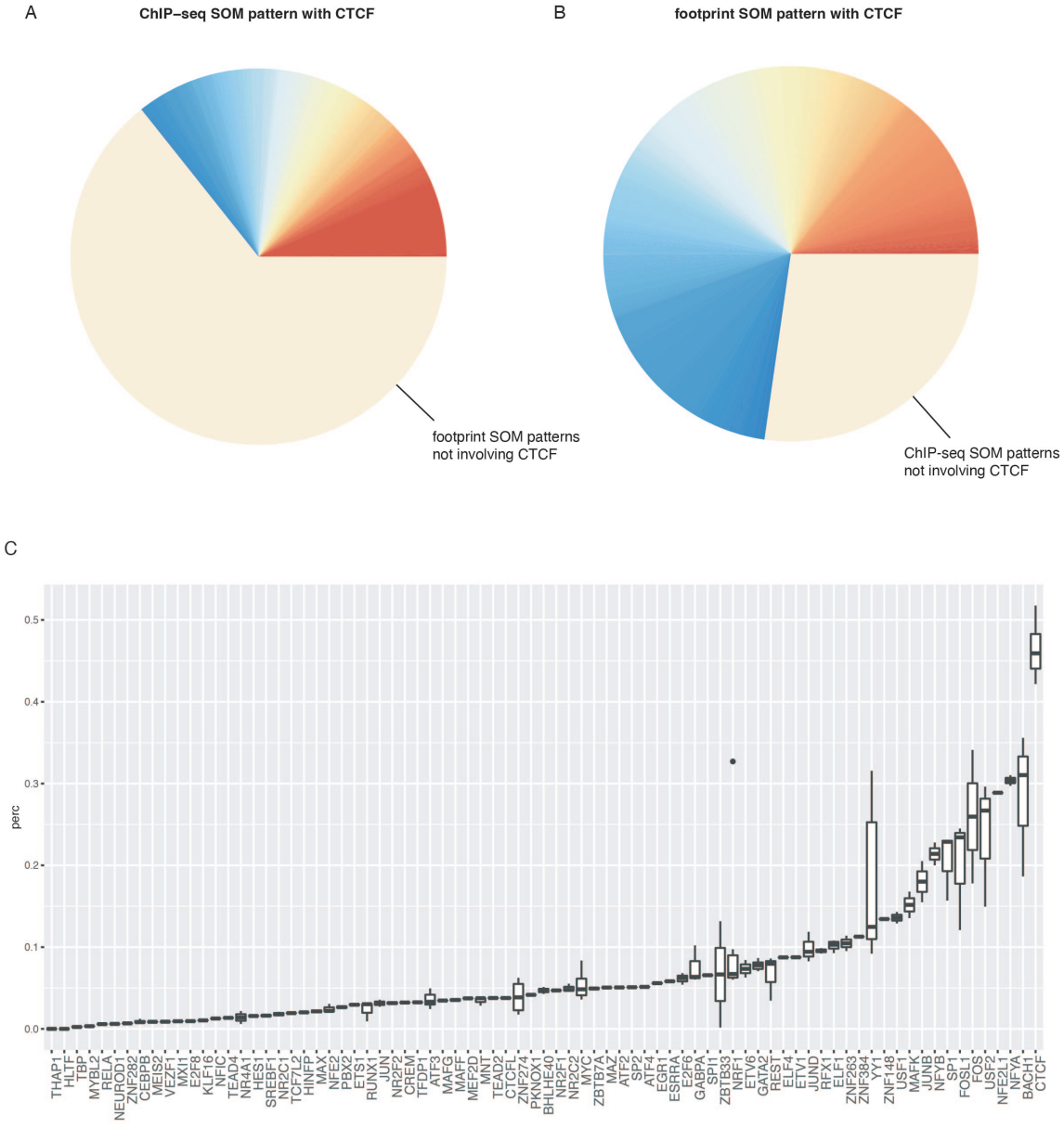

Figure 4.2: Low consistency between TF binding from ChIP-seq and footprints (A) For all regions labeled as ChIP-seq SOM generated CTCF involved patterns, the pie plot shows the distribution of their pattern annotated by footprint SOM (B) For all regions labeled as footprint SOM generated CTCF involved patterns, the pie plot shows the distribution of their pattern annotated by ChIP-seq SOM. (C) Fraction of ENCODE ChIP-seq peaks for a TF with binding sites predictions from TRACE.

4.4.3 TF enrichment at loop anchors

We next studied TF binding activities in chromatin interaction. To identify chromatin loops in the human genomes, we utilized publicly available Paired-End Tag sequencing (ChIA-PET) data. Loop anchors were generated using ChIA-PET data analysis pipeline Mango [48] and then labeled with TF binding. Here, difference in co-localization resulting from DNA-directed or protein-protein interaction is not the main focus. Instead,we want to include proteins that might lack a sequence motif in our analysis, so ChIP-seq peaks were used to label TF binding. Among all TFs we tested, CTCF, REST, RAD21, SMC3 and ZNF143 were most prevalent and strongly enriched in loop anchor regions (Figure 4.3A). CTCF and cohesion complex (RAD21, SMC3) are known to mediate 3D chromatin structure formation. Previous studies have shown that anchor regions of chromatin loops are very strongly enriched for CTCF, cohesion complex protein RAD21 and SMC3, and ZNF143 [106]. ZNF143 is known to cooperate with cohesin complex to help establish the CTCF involved conserved chromatin loops. Besides these 5 TFs, interestingly, the other TFs all exhibited slightly enrichment in loop anchors. One possible reason is that the background regions used in comparison were selected from open chromatin regions (DNase-seq peaks) but not necessarily have evidence of TF binding. In contrast, loop anchors generated from ChIA-PET data should be bound by at least one protein that were enriched in the assay and presumably are more likely to be interacted with TFs. As a result, loop anchor regions will show a higher enrichment for TFs binding than background, and the fitted line of plotted dots can be used as a base line instead of the diagonal line to determine strong enrichment (Figure 4.3A). We then conducted a same comparison but remove all high-occupancy target (HOT) regions since those regions are expected to have one or a few proteins binding. By

only considering non-HOT regions, the overall abundance of TF were reduced, and CTCF, REST, RAD21, SMC3 and ZNF143 showed even more extreme enrichment.

Some other TFs enriched in loop anchor regions includes Myc-associated zinc finger protein (MAZ) and myc-associated factor X (MAX). Binding sites of MAZ are often found at adjacent regions to CTCF, and just like CTCF, it also interacts with cohesin subunit with a suggested role in 3D chromatin structure formation [107, 108]. MAX serves as a cofactor for DNA binding and often form homodimer or heterodimer [109]. MAX–MAZ is known to form protein complex in multiple cell lines and previous study suggested that they can modify chromatin structure and alter gene expression with the involvement of chromatin-remodelling gene CHD2 [110].

4.4.4 Enhancer and promoter loop anchors

Our previous study demonstrated that looping variation may produce differential expression by refining altered enhancer–promoter interactions and raised questions about the necessity of CTCF variability in chromatin looping dynamics. Here we followed up on this finding by examining the TF binding complex at enhancer and promoter loop anchors.

To study TF enrichment in cis-regulatory elements, we first defined enhancer and promoter loop anchors by intersecting loop anchors with chromatin states annotations from chromHMM. The same 5 TFs (CTCF, REST, RAD21, SMC3 and ZNF143) were still enriched in enhancer regions when comparing to randomly selected background group from open chromatin regions (Figure 4.3B). However, when we used the randomly selected regions from pools of all generated loop anchors as background group, these 5 TFs showed similar level of abundance in enhancer regions with background (Figure 4.3C). Unexpectedly, all other tested TFs still displayed a

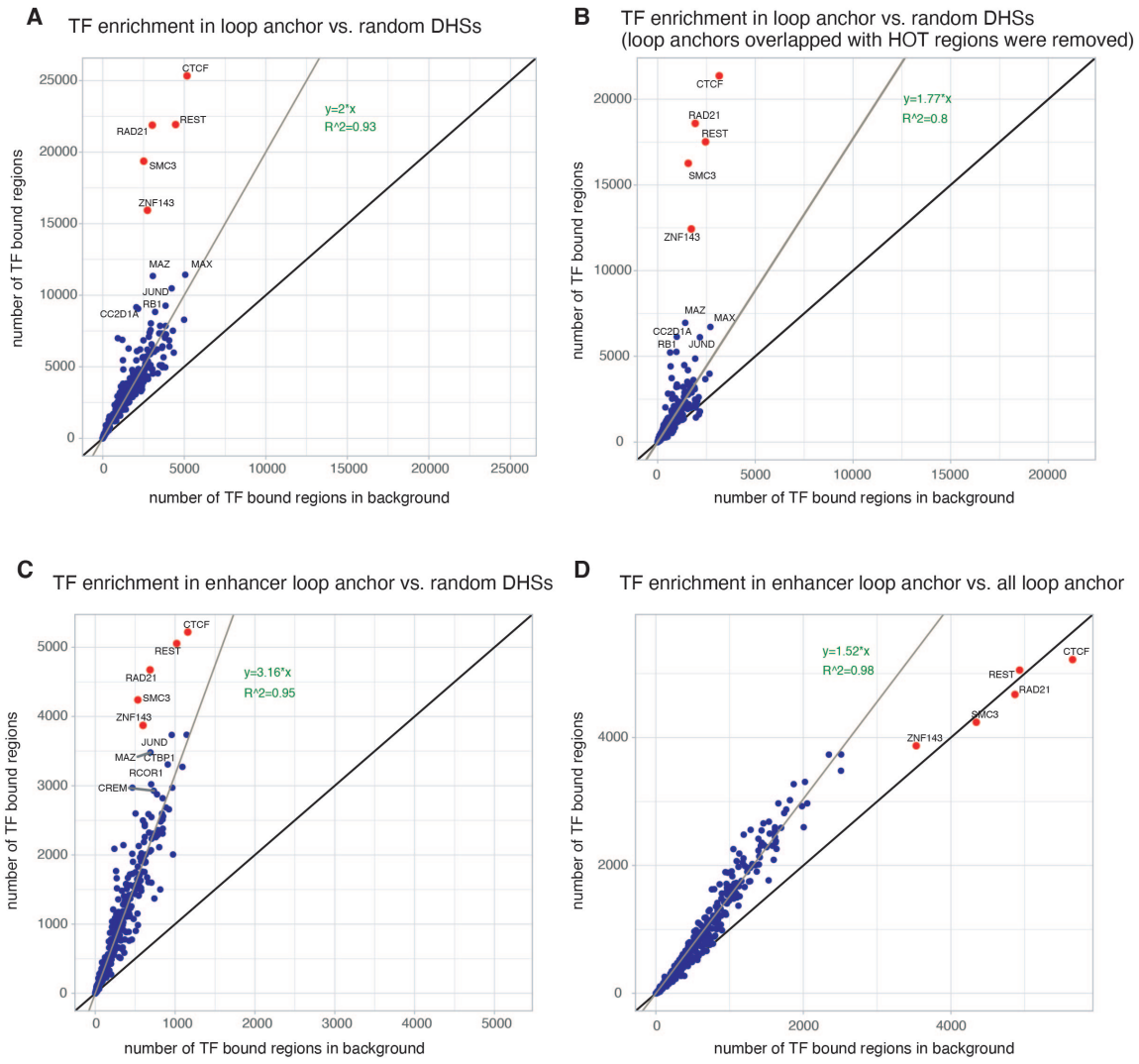

Figure 4.3: Enrichment of transcription factors. (A) TF enrichment at loop anchors. (B) TF enrichment at loop anchors after removing HOT regions. (C) TF enrichment at enhancer loop anchors over random DHSs as background. (D) TF enrichment at enhancer loop anchors over random loop anchors as background.

slight enrichment in active regulatory element loop anchors. One possible explanation is that the CTCF is a known chromatin interaction mediator, and CTCF, together with TFs that are clustered with it, exist at most loop anchors no matter which regulatory element they were annotated, so when we compare occurrence of the TFs at enhancer and promoter anchors versus random anchors, they tend to show similar level of abundance instead of an enrichment at specific group of loop

anchors. But it does not affect proteins that are not essential in chromatin loop formation, and as a result, the shift of the majority of TFs towards enrichment in enhancer and promoter loop anchors still exist. However, no particular TF showed extreme enrichment in enhancer loop anchors when comparing to all background loop anchors if the common shift shared by all tested TFs were considered as a base line.

4.4.5 TF co-binding patterns at loop anchors

We next studied the TF co-localization at loop anchors by training SOM on loop anchors labeled by TF binding. Our goal was to define common binding patterns at loop anchors and enhancer-promoter loop anchors. Common co-binding patterns at anchor regions that we discovered using SOM include different combinations of CTCF, ZNF143, RAD21, and SMC3, as expected. REST, MAZ, MAX, E2F6 are some of other most occurred TFs in co-binding patterns. For the regions grouped by co-binding patterns that are more enriched with enhancer regions, we tested the abundance of some TFs that showed enrichment at loop anchors in previous sections. Among all enhancer enriched co-binding patterns, a few of them contain CTCF and REST (Figure 4.4B, 4.4F). MAX and MAZ were exist in many of enhancer related patterns, and most of these patterns do not consist of CTCF or REST (Figure 4.4C, 4.4D). JUND is another TF appearing at most enhancers (Figure 4.4E), which is consistence with a previous study that reported JUND is major enhancer molecule and can induce Bcl6 expression [111].

To further investigate binding proteins at regulatory element loop anchors, we also employed a SOM on enhancer loop anchors and promoter loop anchors. A variety of binding patterns were revealed, however, number of genomic regions supporting these binding patterns was too small due to limited data availability and quality,

so that these may be false positives and are not true co-localizations at regulatory loop anchors. This approach can be reconducted when chromatin interaction and regulatory element data with better quality become available.

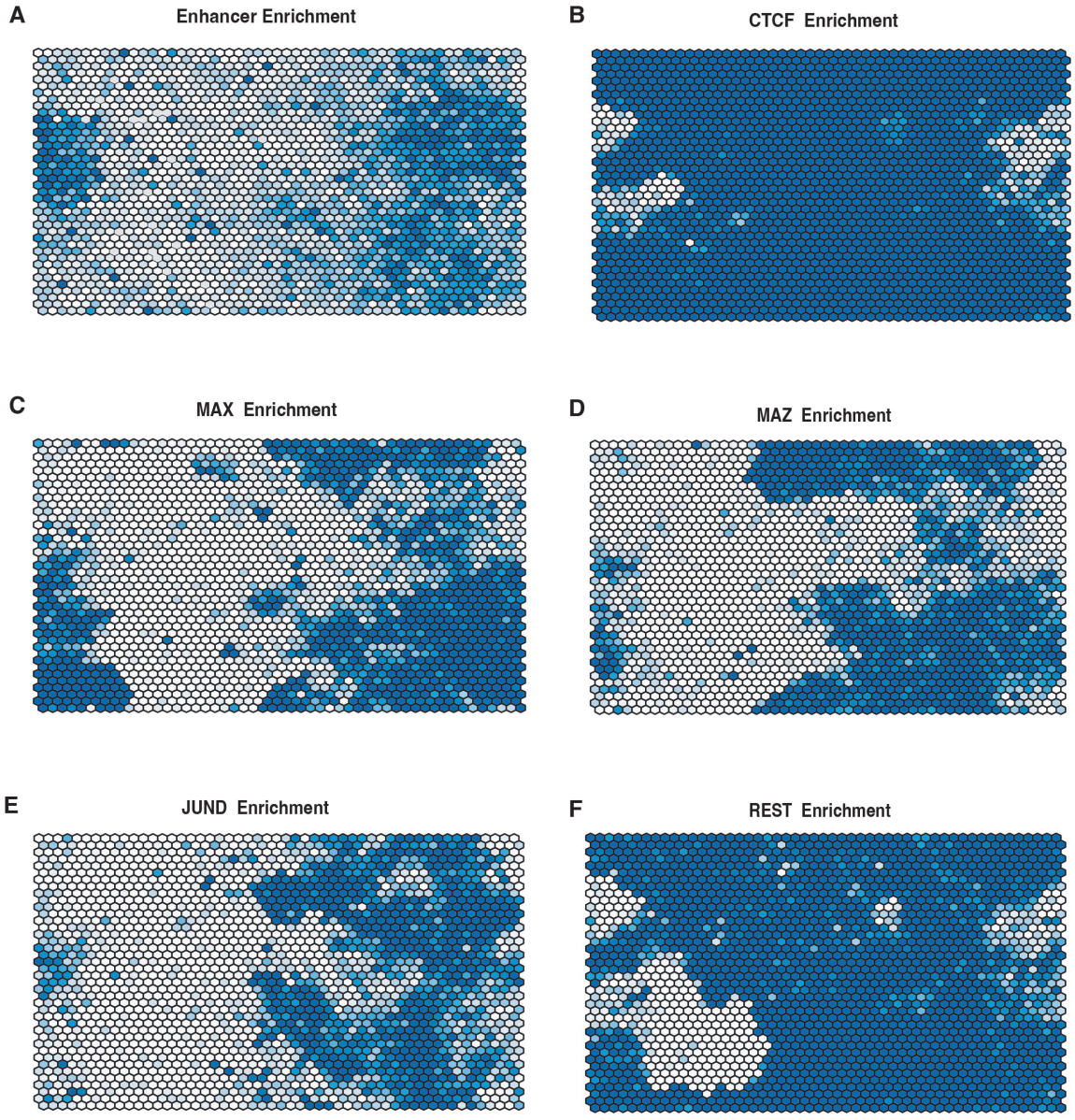

Figure 4.4: Loop anchor SOM maps, shaded by (A) enhancer enrichment. (B) CTCF binding value. (C) MAX binding value. (D) MAZ binding value. (E) JUND binding value. (F) REST binding value.

4.5 Discussion

Our analysis in TF enrichment at loop anchors raised questions about the function and binding mechanism of REST as it is highly enriched in loop anchors and showed highly preference in co-binding with CTCF. 63.8% CTCF bound regions were also bound by REST and 73.1% regions bound by REST also interacted with CTCF. Furthermore, 61.3% REST bound regions also interacted with cohesin protein RAD21 or SMC3. This suggests that REST might play an important role in chromatin interaction and mediate chromatin looping together with cohesin proteins, it might also form special co-localization complex with CTCF to exert specific regulatory effect. Repressor element-1 silencing transcription factor (REST) is a known transcriptional repressor that can recognize neuron-restrictive silencer elements (NRSEs) and can recruit chromatin-modifying enzymes to regulate gene expression [112]. However, how REST regulates gene expression has not yet been fully understood. To further exam regulatory properties of REST, follow-up study on REST binding needs to be conducted to better understand its binding mechanism and regulatory function.

Another two TFs that we studied were USF1 and USF2. USF1 and USF2 often form homo- and heterodimers, and preferentially bind to TSS proximal regions. They have been reported to bind at domains with high levels of active histone marks and low levels of repressive histone marks [113]. We observed that 45.5% of loop anchors with USF1/USF2 binding are TSS or TSS flanking regions (Figure 4.5A), but for the rest loop anchors not bound by USFs, only 18.3% are proximal to TSS. We previously found that USF1/USF2 related pattern's binding specificity can be fully explained by transposable elements (TEs), however, in general, TE-derived and native chromatin loops presented similar activating histone marks patterns [31]. Here we compared

chromatin modification enrichment at USF1/USF2 bound chromatin loops that were TE derived with those were not TE derived. High level signal of H3K27ac, H3K4me1, H3K4me3 and H3K9ac were enriched at USFs bound anchors, along with low level of signal of H3K27me3 (Figure 4.5B), consistent with previous studies. However, the signal was very noisy and there were minor peaks occurring repeatedly surrounding the major peak at region summit. The chromatin marks signal became extremely noisy when we plotted TE derived and not TE derived loop anchors separately. We can still detect high level of active histone marks and observed similar magnitude between two groups, but there were no strong patterns (Figure 4.5C, 4.5D). As a result, no clear difference can be detected between these two groups of USFs bound loop anchors. This could be caused by small number of USF1/USF2 binding sites, as the aggregate signal can be easily impacted by some extreme samples. We also examined the histone marks at each region individually and found a certain amount of variety of signal patterns across USFs bound loop anchors. Given the current data availability and quality, the conclusion on epigenetic properties at TE or non-TE derived loop anchors with USF1/USF2 binding cannot be confidently drawn.

The unexpected TF enrichment at active regulatory elements also raised questions about potential TF binding patterns associated with enhancer and promoter loop anchors. Most TFs were shown to be more prevalent than random in promoter and enhancer loop anchors except CTCF, cohesin complex proteins RAD21 and SMC3 and ZNF143 which were previously reported to be enriched in chromatin interaction regions and can mediates CTCF-bound loops. Although we proposed some possible explanations that might lead to a false enrichment, whether the shift of majority TFs in enrichment plots has biological meaning needs further exploration.

In addition, our analysis on TF enrichment could also be limited by chromHMM

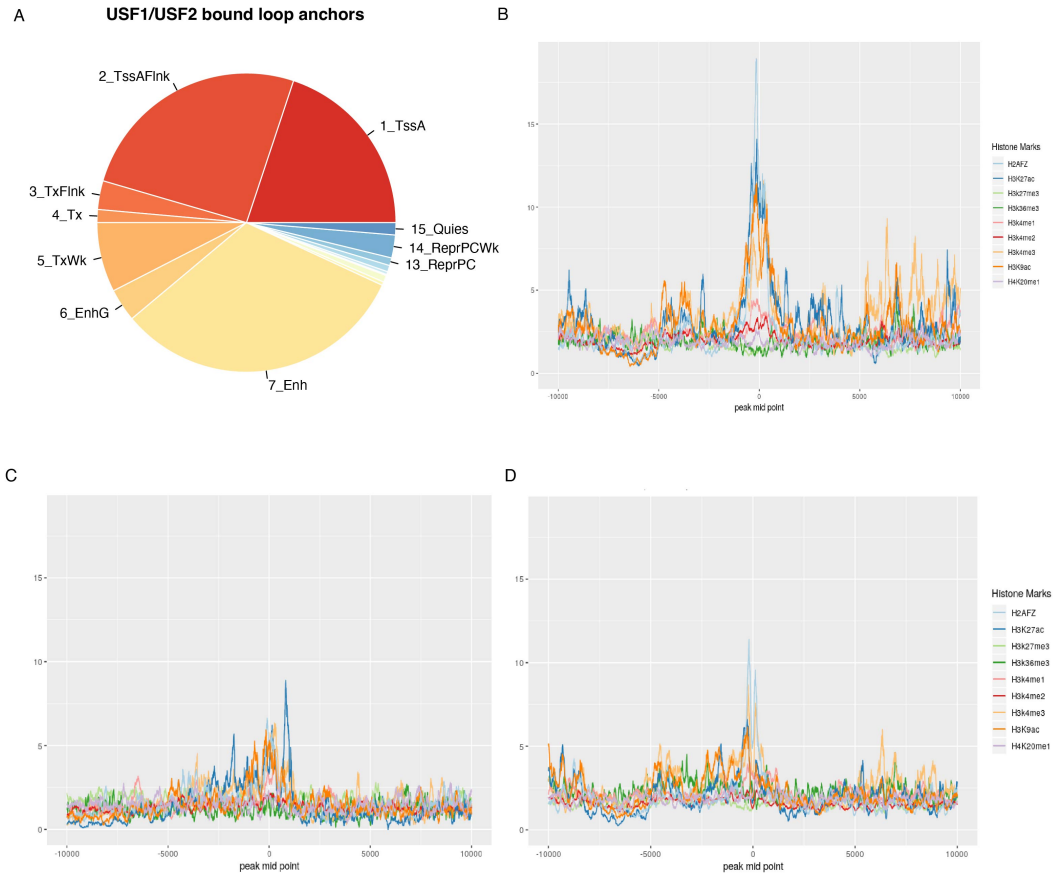

Figure 4.5: Comparison of chromatin state and histone modifications at USF1/USF2 loop anchors. (A) chromatin states annotations at all loop anchors bound by USF1/USF2. TssA: Active TSS, TssAFlnk: Flanking Active TSS, TxFlnk: Transcription at gene 5' and 3', Tx: Strong transcription, TxWk: Weak transcription, EnhG: Genic enhancers, Enh: Enhancers, ZNF/Rpts: ZNF genes & repeats, Het: Heterochromatin, TssBiv: Bivalent/Poised TSS, BivFlnk: Flanking Bivalent TSS/Enh, EnhBiv: Bivalent Enhancer, ReprPC: Repressed PolyComb, ReprPCWk: Weak Repressed PolyComb, Quies: Quiescent/Low. (B, C, D) Average histone modification signals centered at anchor summits in loop anchors that are (B) bound by USF1 and are (C) TE-derived or (D) not TE-derived .

annotation accuracy. Although chromHMM is a widely used chromatin segmentation tool and is able to label regulatory elements with certain accuracy, it often performs poorly when annotating promoters and enhancers [114] as their annotations can result in many regions containing a noisy mixture of adjacent promoter and enhancer states. chromHMM has many limitations, such as loss of information caused by transforming read counts into binary values, the proper choice of binarization cutoff,

unrealistic data distribution assumptions (independent Bernoulli distributions), and limitation of data input used [115]. Therefore, the group of enhancer and promoter loop anchors we annotated using chromHMM might not be correctly labeled with the corresponding regulatory element. Furthermore, some publicly available chromHMM annotations also utilized CTCF binding information as input feature and have CTCF binding region or insulator has a separated states in their prediction (these data were not used in this chapter). As a result, the regions with strong CTCF binding will most likely be labeled as insulators instead of enhancers or promoters so that we will not observe CTCF and its co-factors enriched at enhancer or promoter loop anchors if utilizing these chromHMM annotations to determine regulatory elements. The analysis of TF binding at active regulatory elements related chromatin interactions can be reconducted when more accurate genome-wide regulatory element annotations are available.

CHAPTER V

Conclusions and Future Directions

The main focus of this dissertation was to decipher the gene regulatory circuit driving by transcription factor binding and cooperativity, 3D chromatin looping, and regulatory variants. Transcription factor can recognize and bind to specific DNA sequence to regulate gene expression in a cell type specific manner. They may work cooperatively to mediate the formation of 3D chromatin structure to exert specific gene regulatory effect. Transcription factor binding sites are key building blocks for gene regulatory network. They are enriched with regulatory variants which can lead to gene dysregulation and diseases through altering TF binding affinity.

The work presented in this dissertation provided improved genome-wide prediction of transcription factor binding sites. Individual-specific and tissue-specific footprinting allowed association fine-mapping of footprint quantitative trait loci that can contribute to a better understanding and interpretation of genetic variations consequences. Precise mapping of a large amount of TFs also enabled a systematic study on TF cooperativity.

5.1 Improved genome-wide transcription factor binding sites prediction

In chapter 2, I developed a new computational method TRACE to predict TFBSs through DNase footprinting. By incorporating DNase-seq or ATAC-seq data and

PWM information, our model is able to detect footprints with the desired DNase digestion pattern and matching motifs. Although several DNase-seq signal based computational algorithms have been developed to detect footprints by investigating chromatin accessibility patterns [8, 9, 10, 11, 12, 13, 14, 15, 16], they have various restraints that limit their prediction accuracy and applicability. TRACE is a Hidden Markov Model (HMM)-based unsupervised footprinting method. Its basic structure includes different hidden states such as TFBSs, small peaks which flank footprints, and background which are start and end of each input regions. A unique feature of TRACE model is that it includes "bait" motifs of additional TFs as separate hidden state that helps the model discriminate true binding for the TF of interest. Also, as an unsupervised model, its application is not limited to TFs with available ChIP-seq data. In this chapter, I tested TRACE on 100 TFs and generated binding scores for each binding sites predicted. A comprehensive evaluation on existing footprinting methods was conducted. Compared to other supervised and unsupervised footprinting methods, TRACE has the best overall performance as it has the best ranking in both receiver operating characteristic curve area under the curve (ROC AUC) and Precision-Recall (PR) AUC across 100 tested TFs. I also demonstrated that TRACE can be applied accurately across cell lines, so that a model trained in one cell line can be applied to all other cell lines with a comparable accuracy, which significantly increased its applicability and reduced computational cost.

The current TRACE model is TF-specific, which means one separate model needs to be trained for each TF. Considering its capability of targeting multiple motifs in a single model, we can further extend the model and develop a general model that contains all cluster root motifs, so that one single model can predict binding sites for all TFs. It should be noted that adding motifs with similar sequence binding

preference to the model can lead to collision of those motif states, so that motifs included in the model need to be carefully selected.

Another future direction is genome-wide footprinting in all cell lines with available DNase-seq or ATAC-seq data. We recently improved memory usage of TRACE by truncating the input data and reduced its computational time by GPU computing. It makes TFBSs prediction in a huge amount of cell lines feasible, allowing global and nucleotide-precision analyses of cell-context-dependent gene regulatory mechanisms and providing rich resources for various downstream analysis.

5.2 Functional interpretation of regulatory variants

Genetic variation in regulatory elements has been linked to various diseases and phenotypic traits. Identification of variation effects on transcription factor binding is key to understanding and interpreting downstream consequences of changes on gene expression. In chapter 3, I performed association test on genetic variants and TF binding activity predicted from modified version of TRACE, and generated abundant footprint quantitative trait loci (fpQTLs). Many of the fpQTLs SNPs are located inside footprints and exert its regulatory effect by disruption or creation of TF bind sequence. More distal fpQTLs also exist, but generally exhibit a smaller effect size. fpQTLs SNPs are found enriched with higher functional score predicted from computational variant function prediction tools. And a substantial fraction of them are also associated with human diseases or traits such as type 2 diabetes. Regulatory variants altering TF binding affinity may also contribute to variation in expression levels of nearby genes. Many fpQTLs we detected also overlap eQTLs and these SNPs can impact gene expression by alter TF binding activity. We observed positive impact and negative impact TFs by testing the regulating direction of fpQTL-eQTL

SNPs, their functional direction is mostly consistent with the activating or repressing nature of the binding TF.

The next step is to perform gene set enrichment test on the fpQTLs associated genes and study the enrichment of pathways, biological processes, or molecular functions for genes associated with altered binding activity of each TF. Other cellular traits associated QTLs such as pQTLs, meQTLs can also be included in the analysis to extend the study on fpQTLs' property and functional mechanisms.

5.3 TF co-binding patterns and their contribution to CTCF-mediated chromatin interactions and molecular complexes

Co-binding of TFs and regulatory proteins, such as CTCF and cohesin complex, may work cooperatively to mediate the formation of 3D chromatin structure and exert specific regulatory effect. In chapter 4, I employed an artificial neural network called Self-Organizing Maps (SOMs) to explore TF co-binding patterns and studied how CTCF-related molecular complexes interact to impact chromatin conformation, in order to improve our understanding of TF cooperativity and the mechanisms that determine chromatin looping patterns and downstream regulatory outcomes. Footprint and ChIP-seq based SOMs detected a variety of TF co-binding patterns, these two groups showed low level of agreement, which is as expected since the co-localization of TFs can be due to direct DNA binding or protein-protein interactions. Footprint SOM exclusively measures TF co-binding through direct DNA binding but ChIP-seq does not discern between these two cases. Using chromatin loop anchors generated from ChIA-PET data, we found 5 TFs significant enriched in loop anchor regions, including CTCF, RAD21, SMC3, ZNF143, and REST. CTCF, cohesion complex proteins (RAD21, SMC3) and ZNF143 are known to mediate 3D chromatin structure formation collaboratively [106, 42, 43]. But this observation raised ques-

tions about binding mechanisms of REST. REST is a known transcription silencing factor, but how it exerts transcription regulation effect and its effect on chromatin loop is not fully understood. Follow-up study on REST binding mechanism should be conducted in order to better understand its role in gene regulation and 3D chromatin structure formation. Additional experimental validations of TF co-binding will also be beneficial. Moreover, the analysis was performed on K562 due to the largest TF ChIP-seq availability in K562. We can further extend the study on more cell lines to validate the observations when more data are available in other cell lines. We also tested TF enrichment and co-binding pattern in loop anchors overlapping active regulatory element. However, limited by data quality and quantity, there was no significant results produced. When better quality data become available, the analysis can be reperformed.

5.4 Concluding remarks

In this dissertation, I developed an HMM-based computational footprinting method TRACE to precisely predict transcription factor binding sites genome-wide. Compared to existing methods, TRACE showed better prediction performance and extended applicability. TRACE can also provide individual-specific or tissue-specific footprints and binding scores, which allows association test on genetic variation and transcription factor binding affinity across population. Using this method, I identified fpQTLs and their associated regulatory variants. Detection of fpQTLs provides insights in landscape of human regulatory variation and its direct effect on gene expression. In addition, I explored transcription factor cooperativity utilizing footprint prediction from TRACE.

As we are expanding the usage of TRACE and generating footprints from all cell

lines and tissues with available sequencing data, we hope to provide powerful resource to context-dependent or independent studies on TF binding, co-binding mechanism, functional variants, and more.

Current transcription factor binding sites prediction and their binding measurement focus on chromatin accessibility pattern and sequence binding preference, but there are many other biological features that might also impact transcription factor binding but systematic investigation on their effect is lacking. For example, DNA methylation is an epigenetic mark that can changes DNA information content without varying DNA sequence. Abnormal methylation in transcription factor binding sites has been linked to gene dysregulation but we only have limited knowledge as to its impact on transcription factor binding. With the adaptability of including additional features and capability of measuring individual-specific and tissue-specific TF binding activity, TRACE can be adapted to assess the impact of methylation on TF binding alteration. With further extension of TRACE model utilizing more biological information, it can help examine the combinatorial effect on TF binding activity and downstream phenotype consequences.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] J. R. Ecker, W. A. Bickmore, I. Barroso, J. K. Pritchard, Y. Gilad, and E. Segal, "Genomics: {ENCODE} explained," *Nature*, vol. 489, pp. 52–55, sep 2012.

[2] T. S. Furey, "ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions," *Nature reviews. Genetics*, vol. 13, pp. 840–852, dec 2012.

[3] P. J. Park, "ChIP–seq: advantages and challenges of a maturing technology," *Nature Reviews Genetics 2009 10:10*, vol. 10, pp. 669–680, sep 2009.

[4] J. Wang, J. Zhuang, S. Iyer, X. Y. Lin, T. W. Whitfield, M. C. Greven, B. G. Pierce, X. Dong, A. Kundaje, Y. Cheng, O. J. Rando, E. Birney, R. M. Myers, W. S. Noble, M. Snyder, and Z. Weng, "Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors," *Genome Research*, vol. 22, pp. 1798–1812, sep 2012.

[5] The ENCODE Project Consortium, "Perspectives on ENCODE," *Nature 2020 583:7818*, vol. 583, pp. 693–698, jul 2020.

[6] A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford, "High-Resolution Mapping and Characterization of Open Chromatin across the Genome," *Cell*, vol. 132, pp. 311–322, jan 2008.

[7] J. R. Hesselberth, X. Chen, Z. Zhang, P. J. Sabo, R. Sandstrom, A. P. Reynolds, R. E. Thurman, S. Neph, M. S. Kuehn, W. S. Noble, S. Fields, and J. A. Stamatoyannopoulos, "Global mapping of protein-DNA interactions in vivo by digital genomic footprinting.," *Nature methods*, vol. 6, pp. 283–9, apr 2009.

[8] A. P. Boyle, L. Song, B.-K. Lee, D. London, D. Keefe, E. Birney, V. R. Iyer, G. E. Crawford, and T. S. Furey, "High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells," *Genome Research*, vol. 21, pp. 456–464, mar 2011.

[9] E. G. Gusmao, M. Allhoff, M. Zenke, and I. G. Costa, "Analysis of computational footprinting methods for DNase sequencing experiments," *Nature Methods*, vol. 13, pp. 303–309, apr 2016.

[10] J. Kähärä and H. Lähdesmäki, "BinDNase: a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data," *Bioinformatics*, vol. 31, pp. 2852–2859, sep 2015.

[11] S. Neph, J. Vierstra, A. B. Stergachis, A. P. Reynolds, E. Haugen, B. Vernot, R. E. Thurman, S. John, R. Sandstrom, A. K. Johnson, M. T. Maurano, R. Humbert, E. Rynes, H. Wang, S. Vong, K. Lee, D. Bates, M. Diegel, V. Roach, D. Dunn, J. Neri, A. Schafer, R. S. Hansen, T. Kutyavin, E. Giste, M. Weaver, T. Canfield, P. Sabo, M. Zhang, G. Balasundaram, R. Byron, M. J. MacCoss, J. M. Akey, M. A. Bender, M. Groudine, R. Kaul, and J. A. Stamatoyannopoulos, "An expansive human regulatory lexicon encoded in transcription factor footprints," *Nature 2012 489:7414*, vol. 489, pp. 83–90, sep 2012.

[12] J. Piper, M. C. Elze, P. Cauchy, P. N. Cockerill, C. Bonifer, and S. Ott, "Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data," *Nucleic Acids Research*, vol. 41, pp. e201–e201, nov 2013.

[13] R. Pique-Regi, J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad, and J. K. Pritchard, "Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data," *Genome Research*, vol. 21, pp. 447–455, mar 2011.

[14] R. I. Sherwood, T. Hashimoto, C. W. O'Donnell, S. Lewis, A. A. Barkal, J. P. van Hoff, V. Karun, T. Jaakkola, and D. K. Gifford, "Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape," *Nature Biotechnology*, vol. 32, pp. 171–178, feb 2014.

[15] M.-H. Sung, M. J. Guertin, S. Baek, and G. L. Hager, "DNase Footprint Signatures Are Dictated by Factor Dynamics and DNA Sequence," *Molecular Cell*, vol. 56, pp. 275–285, oct 2014.

[16] G. G. Yardımcı, C. L. Frank, G. E. Crawford, and U. Ohler, "Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection.," *Nucleic acids research*, vol. 42, pp. 11865–78, oct 2014.

[17] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf, "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position.," *Nature methods*, vol. 10, pp. 1213–8, dec 2013.

[18] H. S. Rhee and B. F. Pugh, "ChiP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy," *Current Protocols in Molecular Biology*, vol. 0 21, no. SUPPL.100, 2012.

[19] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio, "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits," *Proceedings of the National Academy of Sciences*, vol. 106, pp. 9362–9367, jun 2009.

[20] T. . G. P. 1000 Genomes Project Consortium, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean, "A map of human genome variation from population-scale sequencing.," *Nature*, vol. 467, pp. 1061–73, oct 2010.

[21] International HapMap Consortium, "A second generation human haplotype map of over 3.1 million SNPs," *Nature 2007 449:7164*, vol. 449, pp. 851–861, oct 2007.

[22] K. J. Gaulton, T. Nammo, L. Pasquali, J. M. Simon, P. G. Giresi, M. P. Fogarty, T. M. Panhuis, P. Mieczkowski, A. Secchi, D. Bosco, T. Berney, E. Montanya, K. L. Mohlke, J. D. Lieb, and J. Ferrer, "A map of open chromatin in human pancreatic islets," *Nature Genetics 2010 42:3*, vol. 42, pp. 255–259, jan 2010.

[23] V. G. Cheung, R. S. Spielman, K. G. Ewens, T. M. Weber, M. Morley, and J. T. Burdick, "Mapping determinants of human gene expression by regional and genome-wide association.," *Nature*, vol. 437, pp. 1365–9, oct 2005.

[24] J. B. Veyrieras, S. Kudaravalli, S. Y. Kim, E. T. Dermitzakis, Y. Gilad, M. Stephens, and J. K. Pritchard, "High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation," *PLOS Genetics*, vol. 4, p. e1000214, oct 2008.

[25] K. K. H. Farh, A. Marson, J. Zhu, M. Kleinewietfeld, W. J. Housley, S. Beik, N. Shoresh, H. Whitton, R. J. Ryan, A. A. Shishkin, M. Hatan, M. J. Carrasco-Alfonso, D. Mayer, C. J. Luckey, N. A. Patsopoulos, P. L. De Jager, V. K. Kuchroo, C. B. Epstein, M. J. Daly, D. A. Hafler, and B. E. Bernstein, "Genetic and epigenetic fine mapping of causal autoimmune disease variants," *Nature*, vol. 518, pp. 337–343, feb 2015.

[26] M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, and J. A. Stamatoyannopoulos, "Systematic localization of common disease-associated variation in regulatory DNA," *Science (New York, N.Y.)*, vol. 337, pp. 1190–1195, sep 2012.

[27] A. P. Boyle, E. L. Hong, M. Hariharan, Y. Cheng, M. A. Schaub, M. Kasowski, K. J. Karczewski, J. Park, B. C. Hitz, S. Weng, J. M. Cherry, and M. Snyder, "Annotation of functional variation in personal genomes using RegulomeDB," *Genome research*, vol. 22, pp. 1790–1797, sep 2012.

[28] S. Dong and A. P. Boyle, "Predicting functional variants in enhancer and promoter elements using RegulomeDB," *Human Mutation*, vol. 40, pp. 1292–1298, sep 2019.

[29] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning-based sequence model.," *Nature methods*, vol. 12, pp. 931–4, oct 2015.

[30] A. Jolma, Y. Yin, K. R. Nitta, K. Dave, A. Popov, M. Taipale, M. Enge, T. Kivioja, E. Morgunova, and J. Taipale, "DNA-dependent formation of transcription factor pairs alters their binding specificity," *Nature 2015 527:7578*, vol. 527, pp. 384–388, nov 2015.

[31] A. G. Diehl, N. Ouyang, and A. P. Boyle, "Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes," *Nature Communications*, vol. 11, dec 2020.

[32] A. T. Hark, C. J. Schoenherr, D. J. Katz, R. S. Ingram, J. M. Levorse, and S. M. Tilghman, "CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus," *Nature*, vol. 405, pp. 486–489, may 2000.

[33] G. Hu, X. Dong, S. Gong, Y. Song, A. P. Hutchins, and H. Yao, "Systematic screening of CTCF binding partners identifies that BHLHE40 regulates CTCF genome-wide distribution and long-range chromatin interactions," *Nucleic acids research*, vol. 48, pp. 9606–9620, sep 2020.

[34] G. Ren, W. Jin, K. Cui, J. Rodrigez, G. Hu, Z. Zhang, D. R. Larson, and K. Zhao, "CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression," *Molecular cell*, vol. 67, pp. 1049–1058.e6, sep 2017.

[35] A. Sanyal, B. R. Lajoie, G. Jain, and J. Dekker, "The long-range interaction landscape of gene promoters," *Nature*, vol. 489, pp. 109–113, sep 2012.

[36] M. E. Donohoe, L. F. Zhang, N. Xu, Y. Shi, and J. T. Lee, "Identification of a Ctcf cofactor, Yy1, for the X chromosome binary switch," *Molecular cell*, vol. 25, pp. 43–56, jan 2007.

[37] T. Guastafierro, B. Cecchinelli, M. Zampieri, A. Reale, G. Riggio, O. Sthandier, G. Zupi, L. Calabrese, and P. Caiafa, "CCCTC-binding factor activates PARP-1 affecting DNA methylation machinery," *The Journal of biological chemistry*, vol. 283, pp. 21873–21880, aug 2008.

[38] K. Ishihara, M. Oshimura, and M. Nakao, "CTCF-dependent chromatin insulator is linked to epigenetic remodeling," *Molecular cell*, vol. 23, pp. 733–742, sep 2006.

[39] Z. Liu, D. R. Scannell, M. B. Eisen, and R. Tjian, "Control of embryonic stem cell lineage commitment by core promoter factor, TAF3," *Cell*, vol. 146, pp. 720–731, sep 2011.

[40] M. Lutz, L. J. Burke, G. Barreto, F. Goeman, H. Greb, R. Arnold, H. Schultheiß, A. Brehm, T. Kouzarides, V. Lobanenkov, and R. Renkawitz, "Transcriptional repression by the insulator protein CTCF involves histone deacetylases," *Nucleic acids research*, vol. 28, pp. 1707–1713, apr 2000.

[41] Z. Qiu, C. Song, N. Malakouti, D. Murray, A. Hariz, M. Zimmerman, D. Gygax, A. Alhazmi, and J. W. Landry, "Functional interactions between NURF and Ctcf regulate gene expression," *Molecular and cellular biology*, vol. 35, pp. 224–237, jan 2015.

[42] E. D. Rubio, D. J. Reiss, P. L. Welcsh, C. M. Disteche, G. N. Filippova, N. S. Baliga, R. Aebersold, J. A. Ranish, and A. Krumm, "CTCF physically links cohesin to chromatin," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, pp. 8309–8314, jun 2008.

[43] W. Stedman, H. Kang, S. Lin, J. L. Kissil, M. S. Bartolomei, and P. M. Lieberman, "Cohesins localize with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators," *The EMBO journal*, vol. 27, pp. 654–666, feb 2008.

[44] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, "Capturing chromosome conformation," *Science*, vol. 295, pp. 1306–1311, feb 2002.

[45] M. J. Fullwood, M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, E. G. Chew, P. Y. H. Huang, W. J. Welboren, Y. Han, H. S. Ooi, P. N. Ariyaratne, V. B. Vega, Y. Luo, P. Y. Tan, P. Y. Choy, K. D. A. Wansa, B. Zhao, K. S. Lim, S. C. Leow, J. S. Yow, R. Joseph, H. Li, K. V. Desai, J. S. Thomsen, Y. K. Lee, R. K. M. Karuturi, T. Herve, G. Bourque, H. G. Stunnenberg, X. Ruan, V. Cacheux-Rataboul, W. K. Sung, E. T. Liu, C. L. Wei, E. Cheung, and Y. Ruan, "An Oestrogen Receptor α-bound Human Chromatin Interactome," *Nature*, vol. 462, p. 58, nov 2009.

[46] E. Lieberman-Aiden, N. L. Van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, "Comprehensive mapping of long range interactions reveals folding principles of the human genome," *Science (New York, N.Y.)*, vol. 326, p. 289, oct 2009.

[47] S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden, "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping," *Cell*, vol. 159, pp. 1665–1680, dec 2014.

[48] D. H. Phanstiel, A. P. Boyle, N. Heidari, and M. P. Snyder, "Mango: a bias-correcting ChIA-PET analysis pipeline," *Bioinformatics*, vol. 31, p. 3092, oct 2015.

[49] G. D. Stormo, T. D. Schneider, L. Gold, and A. Ehrenfeucht, "Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli.," *Nucleic Acids Research*, vol. 10, p. 2997, may 1982.

[50] A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao, "High-Resolution Profiling of Histone Methylations in the Human Genome," *Cell*, vol. 129, pp. 823–837, may 2007.

[51] P. J. Skene and S. Henikoff, "An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites," *eLife*, vol. 6, jan 2017.

[52] A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford, "High-Resolution Mapping and Characterization of Open Chromatin across the Genome," *Cell*, vol. 132, pp. 311–322, jan 2008.

[53] J. Wang, J. Zhuang, S. Iyer, X. Lin, T. W. Whitfield, M. C. Greven, B. G. Pierce, X. Dong, A. Kundaje, Y. Cheng, O. J. Rando, E. Birney, R. M. Myers, W. S. Noble, M. Snyder, and Z. Weng, "Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors.," *Genome research*, vol. 22, pp. 1798–812, sep 2012.

[54] M. M. Hoffman and E. Birney, "An effective model for natural selection in promoters.," *Genome research*, vol. 20, pp. 685–92, may 2010.

[55] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis*. Cambridge: Cambridge University Press, 1998.

[56] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[57] A. Khan, O. Fornes, A. Stigliani, M. Gheorghe, J. A. Castro-Mondragon, R. van der Lee, A. Bessy, J. Chèneby, S. R. Kulkarni, G. Tan, D. Baranasic, D. J. Arenillas, A. Sandelin, K. Vandepoele, B. Lenhard, B. Ballester, W. W. Wasserman, F. Parcy, and A. Mathelier, "JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework," *Nucleic Acids Research*, vol. 46, pp. D260–D266, jan 2018.

[58] J. A. Castro-Mondragon, S. Jaeger, D. Thieffry, M. Thomas-Chollier, and J. van Helden, "RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections," *Nucleic Acids Research*, vol. 45, pp. e119–e119, jul 2017.

[59] Z. Li, M. H. Schulz, T. Look, M. Begemann, M. Zenke, and I. G. Costa, "Identification of transcription factor binding sites using ATAC-seq," *Genome Biology*, vol. 20, p. 45, dec 2019.

[60] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, mar 2015.

[61] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *ACM International Conference Proceeding Series*, vol. 148, pp. 233–240, 2006.

[62] M. H. Sung, M. J. Guertin, S. Baek, and G. L. Hager, "DNase footprint signatures are dictated by factor dynamics and DNA sequence," *Molecular cell*, vol. 56, no. 2, pp. 275–285, 2014.

[63] M. R. Corces, A. E. Trevino, E. G. Hamilton, P. G. Greenside, N. A. Sinnott-Armstrong, S. Vesuna, A. T. Satpathy, A. J. Rubin, K. S. Montine, B. Wu, A. Kathiria, S. W. Cho, M. R. Mumbach, A. C. Carter, M. Kasowski, L. A. Orloff, V. I. Risca, A. Kundaje, P. A. Khavari, T. J. Montine, W. J. Greenleaf, and H. Y. Chang, "An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues," *Nature Methods*, vol. 14, pp. 959–962, oct 2017.

[64] C. E. Grant, T. L. Bailey, and W. S. Noble, "FIMO: scanning for occurrences of a given motif.," *Bioinformatics (Oxford, England)*, vol. 27, pp. 1017–8, apr 2011.

[65] F. Pedregosa FABIANPEDREGOSA, V. Michel, O. Grisel OLIVIERGRISEL, M. Blondel, P. Prettenhofer, R. Weiss, J. Vanderplas, D. Cournapeau, F. Pedregosa, G. Varoquaux, A. Gramfort, B. Thirion, O. Grisel, V. Dubourg, A. Passos, M. Brucher, M. Perrot andÉdouardand, A. Duchesnay, and F. Duchesnay EDOUARDDUCHESNAY, "Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot," tech. rep., 2011.

[66] H. H. He, C. A. Meyer, S. S. Hu, M.-W. Chen, C. Zang, Y. Liu, P. K. Rao, T. Fei, H. Xu, H. Long, X. S. Liu, and M. Brown, "Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification," *Nature Methods*, vol. 11, pp. 73–78, jan 2014.

[67] R Core Team, "R: A Language and Environment for Statistical Computing," *R Foundation for Statistical Computing*, pp. https://www.R–project.org, 2018.

[68] W. S. Cleveland, E. Grosse, and W. M. Shyu, "Local Regression Models," in *Statistical Models in S* (J.M. Chambers and T.J. Hastie, ed.), ch. 8, pp. 309–376, New York: Wadsworth & Brooks/Cole, oct 1992.

[69] E. Jones, T. Oliphant, and P. Peterson, "{SciPy}: Open source scientific tools for {Python}," 2014.

[70] N. Ouyang and A. P. Boyle, "TRACE: transcription factor footprinting using chromatin accessibility data and DNA sequence," *Genome Research*, jul 2020.

[71] The ENCODE Project Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature 2012 489:7414*, vol. 489, pp. 57–74, sep 2012.

[72] V. G. Cheung, R. S. Spielman, K. G. Ewens, T. M. Weber, M. Morley, and J. T. Burdick, "Mapping determinants of human gene expression by regional and genome-wide association," *Nature 2005 437:7063*, vol. 437, pp. 1365–1369, oct 2005.

[73] J. F. Degner, A. A. Pai, R. Pique-Regi, J. B. Veyrieras, D. J. Gaffney, J. K. Pickrell, S. De Leon, K. Michelini, N. Lewellen, G. E. Crawford, M. Stephens, Y. Gilad, and J. K. Pritchard, "DNase-I sensitivity QTLs are a major determinant of human expression variation," *Nature*, vol. 482, pp. 390–394, feb 2012.

[74] J. K. Pickrell, J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J.-B. Veyrieras, M. Stephens, Y. Gilad, and J. K. Pritchard, "Understanding mechanisms underlying human gene expression variation with RNA sequencing," *Nature*, vol. 464, p. 768, apr 2010.

[75] A. P. Boyle, E. L. Hong, M. Hariharan, Y. Cheng, M. A. Schaub, M. Kasowski, K. J. Karczewski, J. Park, B. C. Hitz, S. Weng, J. M. Cherry, and M. Snyder, "Annotation of functional variation in personal genomes using RegulomeDB," *Genome Research*, vol. 22, pp. 1790–1797, sep 2012.

[76] S. Dong and A. P. Boyle, "Prioritization of regulatory variants with tissue-specific function in the non-coding regions of human genome," *Nucleic acids research*, oct 2021.

[77] J. Zhou, C. L. Theesfeld, K. Yao, K. M. Chen, A. K. Wong, and O. G. Troyanskaya, "Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk," *Nature Genetics 2018 50:8*, vol. 50, pp. 1171–1179, jul 2018.

[78] R. G. Tharyan, A. Annibal, I. Schiffer, R. Laboy, I. Atanassov, A. L. Weber, B. Gerisch, and A. Antebi, "NFYB-1 regulates mitochondrial function and longevity via lysosomal prosaposin," *Nature Metabolism 2020 2:5*, vol. 2, pp. 387–396, may 2020.

[79] H. Zhao, D. Wu, F. Kong, K. Lin, H. Zhang, and G. Li, "The Arabidopsis thaliana nuclear factor Y transcription factors," *Frontiers in Plant Science*, vol. 7, p. 2045, jan 2017.

[80] I. Atreya, C. C. Schimanski, C. Becker, S. Wirtz, H. Dornhoff, E. Schnürer, M. R. Berger, P. R. Galle, W. Herr, and M. F. Neurath, "The T-box transcription factor eomesodermin controls CD8 T cell activity and lymph node metastasis in human colorectal cancer," *Gut*, vol. 56, p. 1572, nov 2007.

[81] L. Baala, S. Briault, H. C. Etchevers, F. Laumonnier, A. Natiq, J. Amiel, N. Boddaert, C. Picard, A. Sbiti, A. Asermouh, T. Attié-Bitach, F. Encha-Razavi, A. Munnich, A. Sefiani, and S. Lyonnet, "Homozygous silencing of T-box transcription factor EOMES leads to microcephaly with polymicrogyria and corpus callosum agenesis," *Nature Genetics 2007 39:4*, vol. 39, pp. 454–456, mar 2007.

[82] K. Shimizu, Y. Sato, M. Kawamura, H. Nakazato, T. Watanabe, O. Ohara, and S. ichiro Fujii, "Eomes transcription factor is required for the development and differentiation of invariant NKT cells," *Communications Biology 2019 2:1*, vol. 2, pp. 1–13, apr 2019.

[83] A. Battle, S. Mostafavi, X. Zhu, J. B. Potash, M. M. Weissman, C. McCormick, C. D. Haudenschild, K. B. Beckman, J. Shi, R. Mei, A. E. Urban, S. B. Montgomery, D. F. Levinson, and D. Koller, "Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals," *Genome research*, vol. 24, pp. 14–24, jan 2014.

[84] T. Lappalainen, M. Sammeth, M. R. Friedländer, P. A. 'T Hoen, J. Monlong, M. A. Rivas, M. Gonzàlez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. Van Iterson, J. Almlöf, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. Macarthur, M. Lek, E. Lizano, H. P. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom, H. Lehrach, S. Schreiber, R. Sudbrak, Á. Carracedo, S. E. Antonarakis, R. Häsler, A. C. Syvänen, G. J. Van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigó, I. G. Gut, X. Estivill, and E. T. Dermitzakis, "Transcriptome and genome sequencing uncovers functional variation in humans," *Nature*, vol. 501, no. 7468, pp. 506–511, 2013.

[85] S. B. Montgomery, M. Sammeth, M. Gutierrez-Arcelus, R. P. Lach, C. Ingle, J. Nisbett, R. Guigo, and E. T. Dermitzakis, "Transcriptome genetics using second generation sequencing in a Caucasian population," *Nature*, vol. 464, pp. 773–777, apr 2010.

[86] A. Battle, Z. Khan, S. H. Wang, A. Mitrano, M. J. Ford, J. K. Pritchard, and Y. Gilad, "Genomic variation. Impact of regulatory variation from RNA to protein.," *Science (New York, N.Y.)*, vol. 347, pp. 664–7, feb 2015.

[87] G. McVicker, B. Van De Geijn, J. F. Degner, C. E. Cain, N. E. Banovich, A. Raj, N. Lewellen, M. Myrthil, Y. Gilad, and J. K. Pritchard, "Identification of genetic variants that affect histone modifications in human cells," *Science (New York, N.Y.)*, vol. 342, no. 6159, pp. 747–749, 2013.

[88] N. E. Banovich, X. Lan, G. McVicker, B. van de Geijn, J. F. Degner, J. D. Blischak, J. Roux, J. K. Pritchard, and Y. Gilad, "Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels," *PLoS genetics*, vol. 10, sep 2014.

[89] R. J. Weiss, P. N. Spahn, A. G. Toledo, A. W. Chiang, B. P. Kellman, J. Li, C. Benner, C. K. Glass, P. L. Gordts, N. E. Lewis, and J. D. Esko, "ZNF263 is a transcriptional regulator of heparin and heparan sulfate biosynthesis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 117, pp. 9311–9317, apr 2020.

[90] J. E. Craig, X. Han, A. Qassim, M. Hassall, J. N. Cooke Bailey, T. G. Kinzy, A. P. Khawaja, J. An, H. Marshall, P. Gharahkhani, R. P. Igo, S. L. Graham, P. R. Healey, J. S. Ong, T. Zhou, O. Siggs, M. H. Law, E. Souzeau, B. Ridge, P. G. Hysi, K. P. Burdon, R. A. Mills, J. Landers, J. B. Ruddle, A. Agar, A. Galanopoulos, A. J. White, C. E. Willoughby, N. H. Andrew, S. Best, A. L. Vincent, I. Goldberg, G. Radford-Smith, N. G. Martin, G. W. Montgomery, V. Vitart, R. Hoehn, R. Wojciechowski, J. B. Jonas, T. Aung, L. R. Pasquale, A. J. Cree, S. Sivaprasad, N. A. Vallabh, A. C. Viswanathan, F. Pasutto, J. L. Haines, C. C. Klaver, C. M. van Duijn, R. J. Casson, P. J. Foster, P. T. Khaw, C. J. Hammond, D. A. Mackey, P. Mitchell, A. J. Lotery, J. L. Wiggs, A. W. Hewitt, and S. MacGregor, "Multitrait analysis of glaucoma identifies new risk loci and enables polygenic prediction of disease susceptibility and progression," *Nature genetics*, vol. 52, p. 160, feb 2020.

[91] E. Garrison, J. Sirén, A. M. Novak, G. Hickey, J. M. Eizenga, E. T. Dawson, W. Jones, S. Garg, C. Markello, M. F. Lin, B. Paten, and R. Durbin, "Variation graph toolkit improves read mapping by representing genetic variation in the reference," oct 2018.

[92] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, and H. Parkinson, "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations," *Nucleic acids research*, vol. 42, jan 2014.

[93] M. B. Gerstein, A. Kundaje, M. Hariharan, S. G. Landt, K. K. Yan, C. Cheng, X. J. Mu, E. Khurana, J. Rozowsky, R. Alexander, R. Min, P. Alves, A. Abyzov, N. Addleman, N. Bhardwaj, A. P. Boyle, P. Cayting, A. Charos, D. Z. Chen, Y. Cheng, D. Clarke, C. Eastman, G. Euskirchen, S. Frietze, Y. Fu, J. Gertz, F. Grubert, A. Harmanci, P. Jain, M. Kasowski, P. Lacroute, J. Leng, J. Lian, H. Monahan, H. Oĝgeen, Z. Ouyang, E. C. Partridge, D. Patacsil, F. Pauli, D. Raha, L. Ramirez, T. E. Reddy, B. Reed, M. Shi, T. Slifer, J. Wang, L. Wu, X. Yang, K. Y. Yip, G. Zilberman-Schapira, S. Batzoglou, A. Sidow, P. J. Farnham, R. M. Myers, S. M. Weissman, and M. Snyder, "Architecture of the human regulatory network derived from ENCODE data," *Nature*, vol. 489, pp. 91–100, sep 2012.

[94] C. T. Ong and V. G. Corces, "CTCF: an architectural protein bridging genome topology and function," *Nature reviews. Genetics*, vol. 15, no. 4, pp. 234–246, 2014.

[95] E. de Wit, E. S. Vos, S. J. Holwerda, C. Valdes-Quezada, M. J. Verstegen, H. Teunissen, E. Splinter, P. J. Wijchers, P. H. Krijger, and W. de Laat, "CTCF Binding Polarity Determines Chromatin Looping," *Molecular cell*, vol. 60, pp. 676–684, nov 2015.

[96] Y. Guo, Q. Xu, D. Canzio, J. Shou, J. Li, D. U. Gorkin, I. Jung, H. Wu, Y. Zhai, Y. Tang, Y. Lu, Y. Wu, Z. Jia, W. Li, M. Q. Zhang, B. Ren, A. R. Krainer, T. Maniatis, and Q. Wu, "CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function," *Cell*, vol. 162, p. 900, aug 2015.

[97] M. Merkenschlager and D. T. Odom, "CTCF and Cohesin: Linking Gene Regulatory Elements with Their Targets," *Cell*, vol. 152, pp. 1285–1297, mar 2013.

[98] M. Lawrence, R. Gentleman, and V. Carey, "rtracklayer: an R package for interfacing with genome browsers," *Bioinformatics*, vol. 25, pp. 1841–1842, jul 2009.

[99] J. Ernst and M. Kellis, "ChromHMM: automating chromatin-state discovery and characterization," *Nature Methods 2012 9:3*, vol. 9, pp. 215–216, feb 2012.

[100] Roadmap Epigenomics Consortium, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y. C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K. H. Farh, S. Feizi, R. Karlic, A. R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L. H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, and M. Kellis, "Integrative analysis of 111 reference human epigenomes," *Nature*, vol. 518, pp. 317–329, feb 2015.

[101] T. Kohonen, "Self-Organizing Maps," vol. 30, 2001.

[102] J. Hess, P. Angel, and M. Schorpp-Kistner, "AP-1 subunits: quarrel and harmony among siblings," *Journal of Cell Science*, vol. 117, pp. 5965–5973, dec 2004.

[103] S. D. Bailey, X. Zhang, K. Desai, M. Aid, O. Corradin, R. Cowper-Sallari, B. Akhtar-Zaidi, P. C. Scacheri, B. Haibe-Kains, and M. Lupien, "ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters," *Nature Communications 2015 6:1*, vol. 6, pp. 1–10, feb 2015.

[104] Q. Zhou, M. Yu, R. Tirado-Magallanes, B. Li, L. Kong, M. Guo, Z. H. Tan, S. Lee, L. Chai, A. Numata, T. Benoukraf, M. J. Fullwood, M. Osato, B. Ren, and D. G. Tenen, "ZNF143 mediates CTCF-bound promoter–enhancer loops required for murine hematopoietic stem and progenitor cell function," *Nature Communications 2021 12:1*, vol. 12, pp. 1–12, jan 2021.

[105] M. Karimzadeh and M. M. Hoffman, "Virtual ChIP-seq: predicting transcription factor binding by learning from the transcriptome," *bioRxiv*, p. 168419, mar 2019.

[106] N. Heidari, D. H. Phanstiel, C. He, F. Grubert, F. Jahanbani, M. Kasowski, M. Q. Zhang, and M. P. Snyder, "Genome-wide map of regulatory interactions in the human genome," *Genome Research*, vol. 24, pp. 1905–1917, dec 2014.

[107] G. Maity, I. Haque, A. Ghosh, G. Dhar, V. Gupta, S. Sarkar, I. Azeem, D. McGregor, A. Choudhary, D. R. Campbell, S. Kambhampati, S. K. Banerjee, and S. Banerjee, "The MAZ transcription factor is a downstream target of the oncoprotein Cyr61/CCN1 and promotes pancreatic cancer cell invasion via CRAF-ERK signaling," *The Journal of biological chemistry*, vol. 293, pp. 4334–4349, mar 2018.

[108] T. Xiao, X. Li, and G. Felsenfeld, "The Myc-associated zinc finger protein (MAZ) works together with CTCF to control cohesin positioning and genome organization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 118, p. 2021, feb 2021.

[109] C. Grandori, S. M. Cowley, L. P. James, and R. N. Eisenman, "The Myc/Max/Mad Network and the Transcriptional Control of Cell Behavior," *http://dx.doi.org/10.1146/annurev.cellbio.16.1.653*, vol. 16, pp. 653–699, nov 2003.

[110] K. Zhang, N. Li, R. I. Ainsworth, and W. Wang, "Systematic identification of protein combinations mediating chromatin looping," *Nature Communications 2016 7:1*, vol. 7, pp. 1–11, jul 2016.

[111] E. Arguni, M. Arima, N. Tsuruoka, A. Sakamoto, M. Hatano, and T. Tokuhisa, "JunD/AP-1 and STAT3 are the major enhancer molecules for high Bcl6 expression in germinal center B cells," *International Immunology*, vol. 18, pp. 1079–1089, jul 2006.

[112] L. Ooi and I. C. Wood, "Chromatin crosstalk in development and disease: lessons from REST," *Nature reviews. Genetics*, vol. 8, pp. 544–554, jul 2007.

[113] A. Rada-Iglesias, A. Ameur, P. Kapranov, S. Enroth, J. Komorowski, T. R. Gingeras, and C. Wadelius, "Whole-genome maps of USF1 and USF2 binding and histone H3 acetylation reveal new aspects of promoter structure and candidate genes for common human disorders," *Genome Research*, vol. 18, pp. 380–392, mar 2008.

[114] B. Zacher, M. Michel, B. Schwalb, P. Cramer, A. Tresch, and J. Gagneur, "Accurate Promoter and Enhancer Identification in 127 ENCODE and Roadmap Epigenomics Cell Types and Tissues by GenoSTAN," *PloS one*, vol. 12, jan 2017.

[115] A. Mammana and H. R. Chung, "Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome," *Genome biology*, vol. 16, jul 2015.