Computational Methods to Dissect Tissue-Specific Landscapes of Transcription Factor and DNA Interactions

by

Ricardo D'Oliveira Albanus

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2020

Doctoral Committee:

Professor Stephen C. J. Parker, Chair
Professor Margit Burmeister
Professor Gerald Higgins
Professor Laura J. Scott
Professor Patricia J. Wittkopp

Just look at your father
And you'll see how you took after him.
Me, I'm just another
Like my brothers
Of my mother's genes.

All they really want for you
Are the things they didn't do.
All they ever wanted, a little clue.
All they ever wanted, the truth.

Kate Bush
Never for Ever

What's he building in there?
What the hell is he building in there?
He has subscriptions to those magazines

He never waves when he goes by
He's hiding something from the rest of us
He's all to himself; I think I know why...

And what's that tune he's always whistling?

Tom Waits
Mule Variations

Ricardo D'Oliveira Albanus

albanus@umich.edu

ORCID iD: 0000-0003-3651-0136

To my parents, Ricardo and Celia,

and my sister, Silvia.

ACKNOWLEDGEMENTS

Pursuing a PhD in the United States and was by no means an easy task. Fortunately, I found in Michigan much more than I ever anticipated. I am honored to have a chance to thank all of those who supported me during this time.

First and foremost, I thank Dr. Steve Parker for being a role-model of what the scientific pursuit represents – hard-work, collegiality, unabashed criticism and rigor, and a never-exhausting enthusiasm for new discoveries. I cannot emphasize enough how much I learned from him in these past years. Besides being an incredibly gifted scientist and mentor, Steve is a very positive and optimistic person. He established a work culture that will serve as my yardstick for years to come. This nourishing work culture, of course, would not be possible without the carefully selected roster of current and past Parker Lab members: I thank Venkat Elangovan, John Hensley, Anya Kiseleva, Yoshi Kyono, Nandini Manickam, Peter Orchard, Vivek Rai, and Arushi Varshney for many years of support, motivation, friendship, and scientific insights.

My dissertation committee members – Drs. Margit Burmeister, Gerry Higgins, Laura Scott, and Patricia Wittkopp – were invaluable mentors. They have given me helpful feedback and assisted me in keeping focused during my dissertation work. I eagerly anticipated our committee meetings and always left feeling positive afterwards. I am specifically grateful to Trisha for being the reason why I initially chose the University of Michigan and, later, for recognizing my mentorship needs better than I did and stewarding me towards a more successful path.

The DCMB environment was nothing short of helpful and conductive to my academic success. I owe it all to the faculty and staff here. Primarily, I would like to thank Margit Burmeister, Julia Eussen, and Dan Burns for helping me navigate a foreign academic environment. Their advice was extremely helpful. Next, I'd like to thank Maureen Sartor and Laura Scott for their BIOINF 545 course, which was by far the most important during my graduate studies. It was my honor to serve later as a graduate student instructor for this very same course (then, under the stewardship of Drs. Maureen Sartor, Steve Parker, and Alex Tsoi). I also thank Dr. Brian Athey for inspiring me with his engaging and energetic leadership of our department.

I am also honored to acknowledge my previous mentors for instilling in me the love for science. I am indebted to Dr. Jorge Quillfeldt for introducing me to the ideas of Stuart Kauffman, which became a major landmark in my understanding of biology and guided everything that came afterwards. I am grateful to Dr. Rodrigo Dalmolin first for turning my eye to bioinformatics and then for his mentorship and friendship. I thank Prof. José Cláudio Fonseca Moreira for being a gifted teacher, mentor, and friend that always encouraged my scientific exploration.

I chose the two excerpts in the epigraph to represent the *yin-yang* of excitement and alienation that PhD life can bring. I would easily become the aforementioned Tom Waits' character if it were not for all my friends, old and new. I specially want to thank Mariel Barbachan, Sushma Chaluvadi, Shashank Jariwala, and Alexandr Kalinin for their friendship and support. Along the way, many others came and went – we shared laughs, stories, and adventures that helped make these few last years some of the best in my life. I thank my sister, Silvia D'Oliveira Albanus, for a wonderful nine months in her company here. I also want to thank the Bohn family for all their love and support from the very beginning. Rosaura Lembert has been a fundamental cornerstone to my mental well-being for over a decade. Finally, Patrícia Pereira came unexpectedly in Christmas, MI, and very quickly became one of the most important

people in my life; I look forward to our next years together.

Lastly, I thank my parents, Ricardo Falkenberg Albanus and Celia Porto D'Oliveira, for raising me in a caring environment that allowed me freedom to pursuit my curiosity for all subjects from a very early age. I would have not gotten this far if it were not for their unconditional love and support.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

LIST OF FIGURES ix

ABSTRACT xi

CHAPTERS

I. Introduction 1

1.1 The organization of the human genome 1
1.2 Regulatory domains control gene expression 2
1.3 Transcription factors are the effectors of gene regulation . . . 5
1.4 Computational methods to predict TF binding 6
1.5 Nucleosomes positioning has active and passive roles in gene regulation 8
1.6 Pioneer transcription factors define cellular identity 10
1.7 Complex diseases and the flow of biological information . . . 11
1.8 Information theory and biological information 13
1.9 Thesis outline 14

II. Information Theoretical Properties of Transcription Factor and Chromatin Interactions 16

2.1 Abstract 16
2.2 Introduction 16
2.3 Results 17
2.3.1 Chromatin information reflects TF-chromatin interaction patterns 17
2.3.2 Footprint-free prediction of TF binding and chromatin information 19
2.3.3 Chromatin information varies across TFs 20

2.3.4 Chromatin information is associated with TF-DNA residence times . . . 21
2.3.5 High chromatin information TFs associate with nucleosome phasing . . . 23
2.3.6 Chromatin information asymmetry at TF motifs . . . 24
2.3.7 Chromatin information patterns are tissue-specific and associate with genetic control of gene expression 25
2.3.8 High chromatin information TF motifs are associated with chromatin accessibility . . . 27
2.4 Discussion . . . 28
2.5 Limitations . . . 28
2.6 Supplementary Results . . . 29
2.7 Methods . . . 31
2.8 Acknowledgements . . . 47
2.9 Appendix: Additional Figures . . . 49

III. Analyses of Thymic Precursors Chromatin Accessibility to Elucidate Thymocyte Development . . . 70

3.1 Foreword . . . 70
3.2 Abstract . . . 70
3.3 Results . . . 71
3.3.1 Introduction . . . 71
3.3.2 Chromatin accessibility varies across thymocyte development . . . 73
3.3.3 TF binding identification by ATAC-seq footprinting 76
3.3.4 Changes in global TF binding during thymocyte development . . . 78
3.4 Discussion . . . 84
3.5 Methods . . . 86
3.6 Acknowledgements and Publication . . . 90
3.7 Appendix: Additional Figures . . . 91

IV. Implications and Future Directions . . . 103

4.1 *In vivo* TF-chromatin interaction signatures are dynamic and reflect biophysical and regulatory properties of TFs . . . 103
4.2 TF-chromatin interaction patterns identify candidate pioneer TFs . . . 105
4.3 TF binding prediction methods are affected by TF-chromatin interactions . . . 106
4.4 Chromatin information as a novel metric of ATAC-seq quality control . . . 107
4.5 Concluding remarks . . . 108

BIBLIOGRAPHY 111

LIST OF FIGURES

Figure

1.1	Overview of nucleosome organization	3
1.2	Schematic of an enhancer and promoter interaction at a chromatin loop regulating gene expression	5
1.3	Footprints overview	8
1.4	Pioneer TFs help establish the chromatin accessibility landscape in the cell	10
1.5	A simplified vision of the biological layers that link genotype to phenotype	12
1.6	Schematic of information content calculation	14
2.1	Information content of TF-chromatin interactions	18
2.2	Chromatin information informs residence times and TF-nucleosome interactions	22
2.3	The chromatin information landscape of human tissues	26
2.4	ATAC-seq datasets signal-to-noise comparisons	49
2.5	GM12878 V-plots	50
2.6	TF binding prediction methods comparisons across datasets	51
2.7	HINT-ATAC performance using narrow or broad peak calls	52
2.8	BMO and PIQ comparisons	53
2.9	TF binding prediction methods comparisons across sequencing depths	54
2.10	CENTIPEDE and ssCENTIPEDE perform similarly across datasets	55
2.11	BMO and ChIP-seq f-VICEs are correlated	55
2.12	Selection of additional ATAC-seq samples using ubiquitous and conserved CTCF-cohesin binding sites	56
2.13	Normalization of f-VICE	57
2.14	f-VICE distributions across samples	58
2.15	f-VICE correlation with FRAP recovery times in multiple datasets	59
2.16	GM12878 CTCF/cohesin^{+} and CTCF/cohesin^{-} regions	60
2.17	Sonicated GM12878 ATAC-seq data	61
2.18	Chromatin information clusters in GM12878	62
2.19	f-VICE correlation with nucleosome phasing	63
2.20	Motifs with information asymmetry in GM12878	64
2.21	Protein domain enrichments	65

2.22 Enrichment of high and low f-VICE motifs in *cis*-eQTLs 66
2.23 DNA 6-mers f-VICE analyses 67
2.24 f-VICE allelic imbalance analyses 68
2.25 f-VICE and PWM score AUC-PR 69
3.1 Analysis of stage-specific and ubiquitous ATAC-seq clusters 75
3.2 ChIP-Enrich results for the ATAC-seq clusters 77
3.3 ATAC-seq signal and CENTIPEDE footprint calls around the functionally validated *E4p Cd4* gene enhancer 79
3.4 ATAC-seq signal and footprint calls within functionally validated enhancers for the *Cd8* gene 80
3.5 ATAC-seq signal and footprint calls within the functionally validated *TCE1 Gata3* enhancer 82
3.6 Footprint occupancies across samples and clusters 83
3.7 Individual footprint enrichments in each of the samples 84
3.8 Isolation of staged thymocytes 91
3.9 ATAC-seq profiles of thymocytes 92
3.10 Correlation of the ATAC-seq signal between replicates 93
3.11 Open chromatin regions defined by ATAC-seq peaks 94
3.12 Additional information on k-means clustering 95
3.13 Footprint enrichment results from GAT 96
3.14 CENTIPEDE footprint calls within functionally validated enhancers for the *Cd8* gene 97
3.15 ATAC-seq signal and CENTIPEDE footprint calls around functionally validated β E enhancer for the *Trb* gene 98
3.16 Footprint occupancies across samples and clusters 99
3.17 HOMER motif enrichment analysis 100
3.18 CENTIPEDE footprint calls within functionally validated regulatory elements for the *Cd4* gene 101
4.1 Future directions 110

ABSTRACT

The intricately ordered structure of the human genome is a product of dynamic interactions between DNA and proteins such as nucleosomes and transcription factors (TFs), which allow cells to respond to environmental changes while maintaining robustness of genetic programs. Changes in the non-coding genome can affect gene regulation and lead to increased disease predisposition, but the underlying mechanisms are not fully understood. Therefore, understanding how the genome is organized and regulated is a central question in biomedical research. My thesis aims to develop and apply novel computational methods to understand general biological mechanisms of genome regulation, with a focus on TF-DNA interactions.

In the initial part of this thesis, I develop computational methods to quantify TF-DNA interaction patterns by applying information theory to high-throughput molecular profiles of chromatin accessibility data (using the assay for transposase-accessible chromatin followed by high-throughput sequencing, ATAC-seq) to measure a property which we name chromatin information. To circumvent the requirement of high-throughput molecular profiles of TF binding (chromatin immunoprecipitation followed by sequencing, ChIP-seq) to obtain chromatin information measurements, I develop BMO, a novel algorithm to predict TF binding from chromatin accessibility data that outperforms current state-of-the-art methods. Using BMO in combination with the information theoretical approach developed here, I quantify the chromatin information patterns of hundreds of TF motifs across different human tissues and

cell lines. Only a subset of TFs (10-20%) have high chromatin information, and are therefore associated with organized chromatin. By integrating multiple layers of molecular profiles, I find that high chromatin information TFs have longer TF-DNA residence times, associate with nucleosome phasing, and are enriched to overlap regions associated with the genetic control of gene expression. I then use genetic data to find evidence that high chromatin information TFs associate with increased chromatin accessibility and may therefore act as pioneer TFs.

In the last part of this thesis, I apply TF binding prediction algorithms to characterize the regulatory landscape associated with thymocyte development. The results from these analyses support that thymocyte development is a highly dynamic process and help prioritize novel candidate TFs and regulatory elements for future experimental validation.

This work represents a novel fusion of two research domains – information theory and genomics – which allowed to capture properties of TF-chromatin interactions, with important implications for gene regulation, cell state dynamics, and understanding the pathological mechanisms associated with non-coding disease-associated genetic variants.

CHAPTER I

Introduction

1.1 The organization of the human genome

Every cell in an individual human has a nearly identical genetic code, but its differential interpretation leads to diverse tissues and organs. This common 3.2 billion bases of genetic information would stretch out to approximately two meters, but must fit within a few micrometers inside the cell nucleus, requiring the genome to be compacted by about six orders of magnitude [1]. This dense packaging is facilitated by a constellation of proteins that simultaneously compact the genetic material and allow the relevant subset of genes and regulatory circuits to be accessible in a cell-specific manner. The compact form of DNA and proteins is referred to as chromatin.

Chromatin can be broadly categorized as euchromatin or heterochromatin [2]. Euchromatin is the less condensed form of chromatin and is associated with active regions of the genome, where genes are expressed. Heterochromatin, on the other hand, usually corresponds to the repressed sections of the genome. As cells differentiate and respond to their environment, different regions become accessible or repressed [3, 4]. Chromatin organization is, therefore, a highly dynamic process and understanding its regulation is a central question in biology.

The fundamental unit of chromatin is the nucleosome, an octamer of histone protein cores. Each nucleosome is wrapped by approximately 147 base-pairs (bp) of DNA,

and is linked to its neighbors by a stretch of linker DNA forming a "beads-on-a-string" structure (Figure 1.1) [2]. Each of the histone protein cores can be subject to a myriad of post-translational modifications at different amino-acid residues (*e.g.* methylation, acetylation) affecting nucleosome behavior [5]. These modifications modulate nucleosome biophysical parameters (*e.g.* solubility and mobility [6]) and can lead to more or less accessible nucleosome arrangements as well as allow binding of specialized proteins called chromatin remodelers. Chromatin remodeler further affect chromatin organization by adding or removing other post-translational modifications [5]. The higher-order organization patterns of chromatin include topologically associated domains (TADs), which correspond to regions of the genome that may or may not be in close linear proximity, but interact in three-dimensional space through chromatin loops [7]. TADs separate functionally distinct regions of the genome [8].

1.2 Regulatory domains control gene expression

The genomic regions that regulate gene expression are called regulatory elements. These regions contain DNA sequences that facilitate recruiting of the transcription machinery, composed of transcription factors (TFs) and RNA polymerase. Regulatory elements can be broadly characterized as promoters and enhancers based on proximity to their target genes [9]. Promoters are located immediately upstream of the gene transcription start sites (TSS) and are the most well-studied class of regulatory elements. Enhancers are located distally to TSS regions, typically tens of thousands of base-pairs [10, 11], but up to 1 Mbp in the same chromosome [12]. The only well-described instance of enhancers acting in different chromosomes (in *trans*) pertains to the process of determining which copy of the olfactory receptors is expressed in a given olfactory neuron [13]. Enhancers are responsible for driving more complex spatio-temporal activation patterns in gene expression [14], and are postulated to act as the effectors of developmental and environmental signals [15, 16]. The

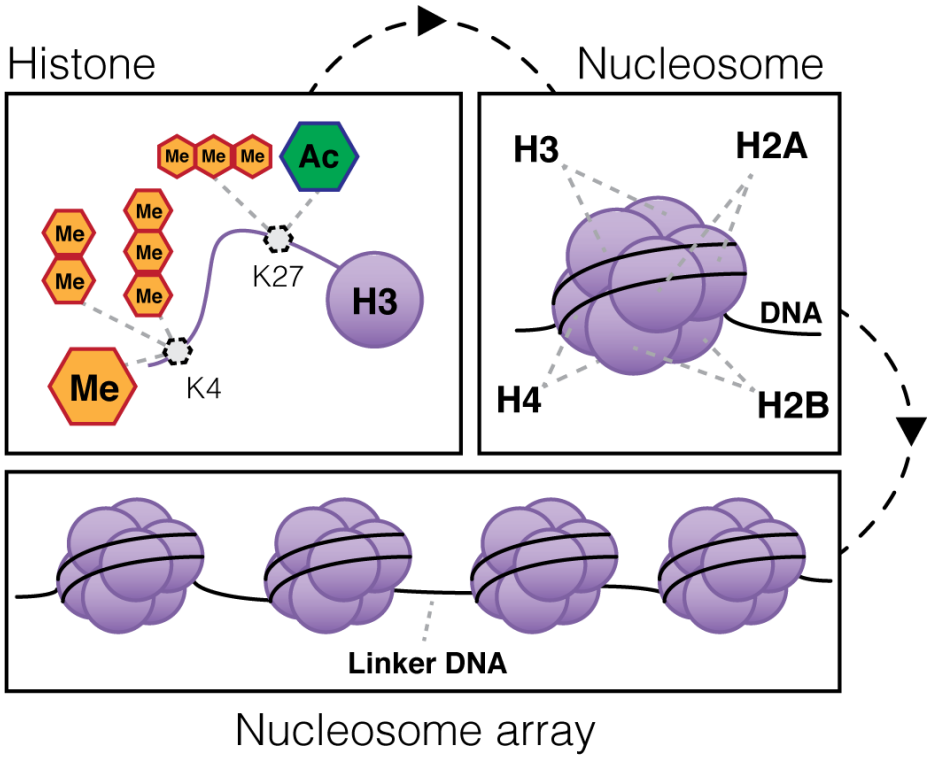


Figure 1.1: Overview of nucleosome organization. Nucleosomes are composed of core histone subunits (upper panels) and form nucleosome arrays with genomic DNA (lower panel). Histones can be post-translationally modified to change their biological properties. The upper-left panel shows a schematic of the most well-studied post-translational modifications in the H3 tail lysines. Me, methylation (mono-di-tri); Ac, acetylation.

most accepted mechanism for enhancer function is through 3-dimensional proximity to its target gene promoter within TADs (Figure 1.2) [17]. The uncertainty associated with their target genes requires more complex experimental designs to measure enhancer activity [18, 19]. Therefore, enhancers are harder to study than promoters. In contrast to the well-characterized genetic code, which links DNA triplets to amino acid residues during gene translation, the regulatory grammars encoded by non-coding DNA elements are much more complex and still remain far from understood.

Only in the last decade has the systematic dissection of the enhancer repertoire in a given cell type became possible. This was in large part due to advances in high-throughput molecular profiling techniques, such as chromatin immunoprecipitation

followed by sequencing (ChIP-seq) of histone modifications, DNase I hypersensitive site sequencing (DNase-seq) [20], the assay for transposase-accessible chromatin using sequencing (ATAC-seq) [21], and massively-parallel reporter assays [22] (for an in depth review of these and other techniques, refer to ref. [23]). In addition, broad community efforts to characterize epigenomic profiles across tissues and cell lines, such as the ENCODE [24] and Roadmap Epigenomics [25] projects, have provided a wealth of datasets serving as a reference for other studies.

The availability of multiple histone modifications ChIP-seq datasets across cell types enabled the application of statistical techniques to segment the genome into tissue-specific sub-classes of regulatory elements based on their molecular profiles. ChromHMM is a tool that uses hidden Markov models to define chromatin states, which are recurring combinations of specific histone modifications [26]. Repressed regions of the genome are rich in marks including, but not limited to histone 3 lysine 27 trimethylation (H3K27me3) and histone 3 lysine 9 bi/tri-methylation (H3K9me2 and H3K9me3). Promoters are characterized by histone 3 lysine 4 trimethylation (H3K4me3) and enhancers by histone 3 lysine 4 mono-methylation (H3K4me1). In addition, the presence of H3K27ac can distinguish active from poised promoters and enhancers (Figures 1.1 and 1.2) [26–28].

More recently, the development of high-throughput 3-dimensional interaction mapping techniques (*e.g.* Hi-C [29] and promoter-capture Hi-C [30]) enabled the determination of candidate target genes for enhancers. In addition, analyses of large cohorts with dense genotyping data combined with gene expression profiles allowed the assignment of target genes to enhancers based on genetic modulation of gene expression (*cis*-expression quantitative trait loci - *cis*-eQTLs) [31–34]. These studies have shown that the interaction landscape between enhancer and promoters are highly dynamic across tissues, developmental stages, and environmental perturbations [16, 35–37]. Importantly, disruptions in enhancer activity or interaction patterns can lead to

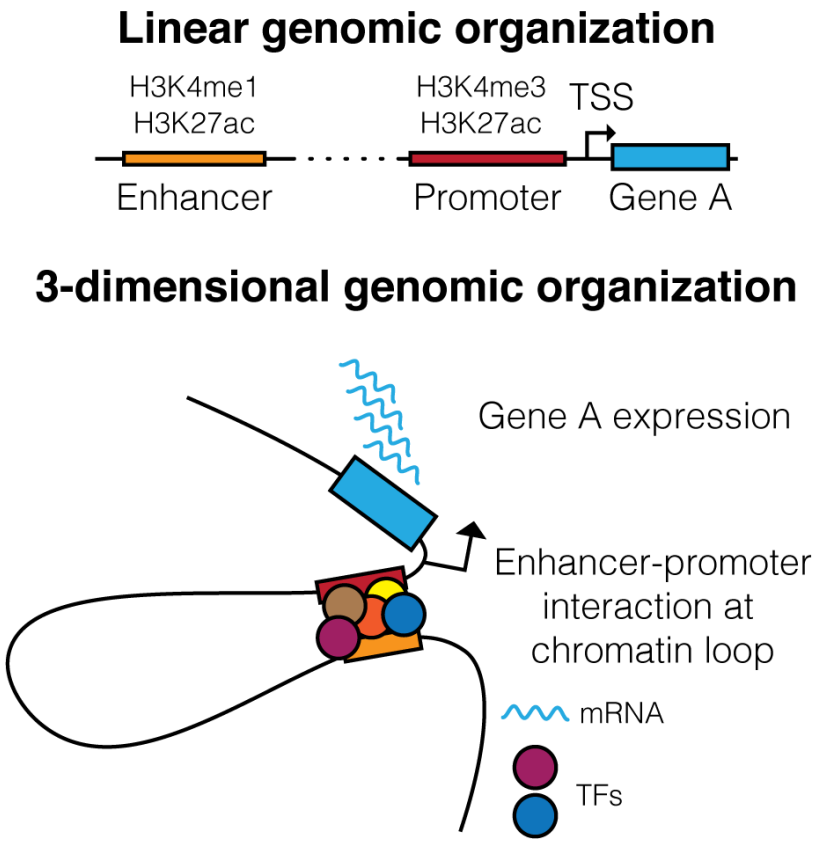


Figure 1.2: Schematic of an enhancer and promoter interaction at a chromatin loop regulating gene expression.

pathological conditions [38–40]. Therefore, understanding the biological principles determining regulatory element activity and interaction patterns is a promising area of research.

1.3 Transcription factors are the effectors of gene regulation

TFs are a class of proteins that regulate gene activity by recognizing and binding specific DNA sequences at regulatory elements, called TF motifs. TFs can act by recruiting the cell's transcription machinery to their target genes or by recruiting chromatin remodelers. These chromatin remodelers will then reshape the local chromatin architecture and modulate binding of other TFs. TFs can be broadly categorized as activators and repressors based on their effects on the target gene ex-

pression. Activators can work by directly recruiting the transcriptional machinery to the gene promoter [2] or by recruiting chromatin remodelers that induce a more accessible chromatin configuration for other TFs [41]. Repressors can act by competing with activators for the same motif [42] or by recruiting repressive chromatin remodelers, such as the polycomb complex [4]. The most widely used method for determining genome-wide TF binding is with ChIP-seq assays. This approach, however, is limited to one TF at a time. Due to sequencing costs, biological material and antibody specificity requirements, the application of TF ChIP-seq is limited to cases where there is some *a priori* knowledge of the TF(s).

TF motifs are commonly represented as position weight matrices (PWMs). PWMs encode the probabilities of observing any given base at every position of the TF binding site [43]. Currently, the most common approach to characterize TF PWMs is to perform a ChIP-seq experiment for the TF of interest and statistically determine overrepresented DNA sequences that inform the TF binding preferences [43]. Other approaches include *in vitro* assays, such as the systematic evolution of ligands by exponential enrichment (SELEX) [44, 45]. These *in vitro* approaches measure pure TF-DNA binding affinities, but may not be biologically accurate because they do not account for the myriad of factors modulating TF activity in the cellular environment. There currently are numerous TF motif databases available that integrate data from these different experimental sources [24, 46, 47].

1.4 Computational methods to predict TF binding

Due to the resource limitations imposed by TF ChIP-seq experiments, one attractive area of research is *in silico* TF binding prediction. The availability of comprehensive motif libraries and cell lines with large number of TF ChIP-seq experiments allowed the development and benchmarking of computational approaches to predict TF binding using either DNA sequence alone or DNA sequence combined with molec-

ular profile from the sample of interest. These approaches provide a first-pass identification of putative TFs relevant to the biological phenomena being studied. These candidate TFs can then be functionally validated using more specific assays [33].

The most naïve approach to predict TF binding is using TF motifs to determine all the putative binding sites for the TFs of interest using sequence similarity [48]. The major limitation of this approach is that most motif matches are located in inaccessible chromatin and are unlikely to be bound [49]. In addition, members of the same TF family can recognize very similar motifs and many TFs bind DNA indirectly through co-binding partners, which make motif-TF assignments unreliable for some TFs. One of the focus of the field is to identify redundancy across databases [50] and correctly assign motifs to TFs or TF families [46] to mitigate these issues.

To overcome the limitations of using sequence alone to predict TF binding, it is necessary to include functional genomic data and generate tissue-specific predictions. The most widely adopted methods to predict TF binding use chromatin accessibility data (DNase-seq or ATAC-seq) to find small localized regions of protected DNA in otherwise open and accessible regions that are thought to be due to TF binding, which are called TF footprints [51–54]. TF footprints are characterized by a stereotypical pattern of a low accessibility region flanked by high accessibility (Figure 1.3). The footprint location can then be intersected with motif matches in order to determine candidate TFs. Other methods, such as CENTIPEDE [49], do not directly search for footprints and instead model the chromatin accessibility shape around the TF motif to predict TF binding.

There is evidence suggesting that only a small fraction (10-20%) of TFs associate with footprints [55]. Therefore, the application of footprinting-based methods may be restricted. It is hypothesized that TF residence time, which corresponds to the duration of TF binding on DNA (described in detail in the next section), affects TF footprints [52]. However, the residence times of most TFs are currently unknown.

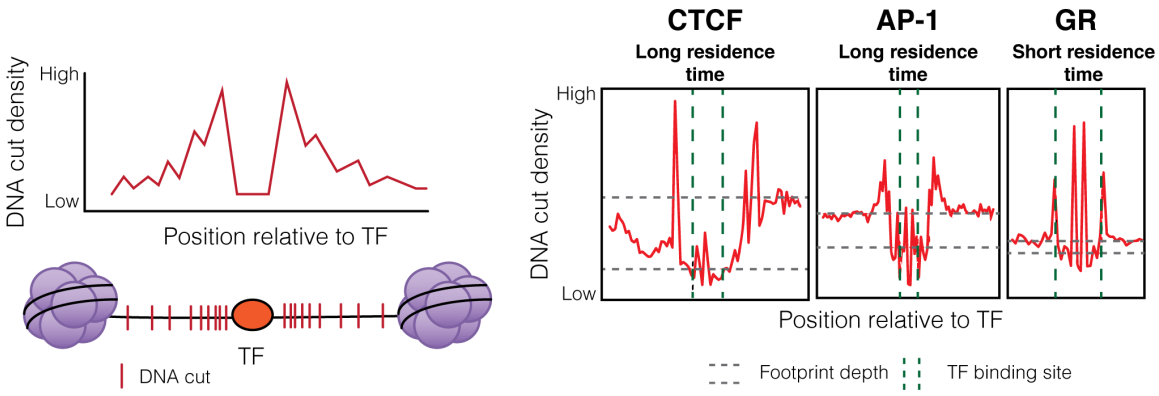


Figure 1.3: Footprints overview. Left: schematic of a TF footprint. Right: Aggregate DNase cut profiles across TF binding sites (based on ChIP-seq data) showing footprints for three TFs (adapted from [52]). Note that the cut density fluctuations at the TF binding site reflect assay-specific sequence bias.

Other features, such as the level of chromatin accessibility around the TF motif [56, 57] and the presence of nearby co-occurring motifs [58], have been shown to positively correlate with TF binding. These features can provide footprint-independent measurements to predict TF binding, but they have not been as extensively characterized in this context.

1.5 Nucleosomes positioning has active and passive roles in gene regulation

Nucleosome positioning is an essential property of chromatin architecture. A significant portion of the genome is characterized by highly-ordered (phased) nucleosome arrays. These regions are enriched to overlap active regulatory regions such as promoters and enhancers [59]. The CCCTC-binding factor (CTCF) is the TF with the most well-characterized ability to rearrange nucleosomes upon binding [60]. Sequencing of micrococcal nuclease sensitive sites (MNase-seq) experiments, which inform nucleosome positioning, show that CTCF binding sites are flanked by >10 evenly-spaced nucleosomes [59, 61]. Other regions with known well-positioned nucleosomes

include TSS [62]. Studies in yeast have shown that the positioning of the TSS-flanking nucleosomes correlates with transcriptional activity [63].

The regulatory mechanisms associated with nucleosome positioning are still not completely understood. One explanation is that the nucleosome competes with TFs [64] or acts as a barrier for TF binding [59]. However, recent studies have shown that TF-nucleosome interactions are more intricate. One such study systematically characterized TF-nucleosome interactions *in vitro* and demonstrated that TFs have widely different nucleosome binding preferences [65], ranging from those that are not affected by the nucleosome, those that bind at specific regions of nucleosome dyad, and those that are inhibited by the nucleosome. The nucleosome-binding TFs are postulated to act as anchors of nucleosome positioning. Another study demonstrated that the TF SOX2 can only bind its motif if it is either located in the center of the nucleosome dyad, where it encounters less steric hindrance, or if its binding partner, OCT4, is bound nearby and evicted the nucleosome. The same study showed that OCT4, on the other hand, is not affected by nucleosome positioning and therefore likely act as a facilitator of SOX2 (and possibly other TFs) [54]. Together, these studies demonstrate that nucleosomes simultaneously affect and are affected by TF binding.

An increasing body of research indicates that the dynamics of TF-chromatin interactions are a key component of the regulatory element activity. One important biophysical property of TF-chromatin interactions is TF residence time, which corresponds to the duration of TF binding on DNA [66]. It has been shown that differences in TF residence time can modulate the target gene expression and induce competition between TFs [67–69]. Importantly, there is evidence indicating that TF residence times can be biologically modulated [68, 70], with implications for gene regulation. However, the methodology to experimentally determine TF residence times is technically demanding, resulting in data available only for a few dozen TFs. Therefore,

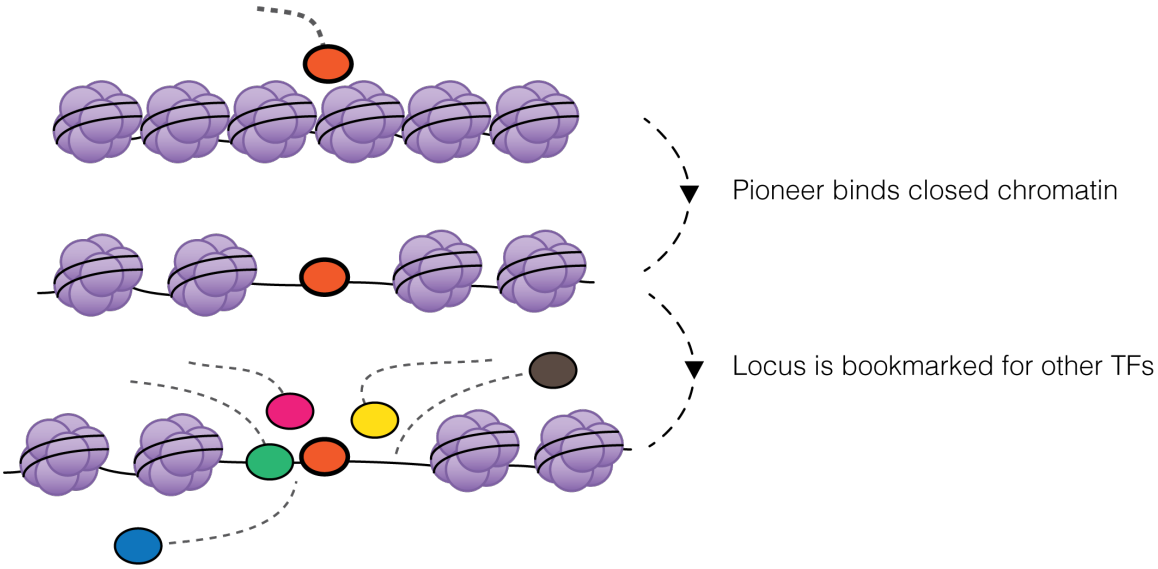


Figure 1.4: Pioneer TFs help establish the chromatin accessibility landscape in the cell. In this schematic, the orange TF behaves as a pioneer and the other TFs behave migrants [51].

developing new methods to investigate TF-chromatin dynamics is a potentially rewarding area to gain insights into the organization of the genome.

1.6 Pioneer transcription factors define cellular identity

A special class of TFs, called pioneers, has been found to be able to bind DNA in closed chromatin [71]. It is believed that pioneers help establish cell identity by bookmarking cell-specific regions of the genome where other non-pioneer TFs (migrants [51]) exert their regulatory activity. Known pioneers include FOXA2, GATA, NF-Y, SOX2, and OCT4, which are all associated with cellular differentiation [71, 72].

The characteristics that differentiate pioneers from non-pioneers are still not understood, but seem to result mainly from nucleosome affinity [73, 74]. A recent study determined the *in vitro* nucleosome affinities for multiple TFs in order to quantify their pioneering capabilities [74]. Rather than finding a binary separation between pioneers and non-pioneers, this study found a continuum of nucleosome affinity asso-

ciated with different TF families. One interpretation of this result is that some TFs have intrinsic pioneer activity, while others act as pioneers by combinatorial effects of TFs with distinct TF-chromatin interaction patterns. This is supported by a previous study that determined that MYC co-binds with OCT4 in nucleosome-dense regions [73] and with the previously mentioned SOX2 and OCT4 study [75]. Similarly, another study demonstrated that the glucocorticoid receptor can have "pioneer-like" properties depending on its oligomerization state [76]. Therefore, future studies aiming to characterize pioneer TFs will need to take into account the biological context modulating TF activity.

1.7 Complex diseases and the flow of biological information

One of the critical questions currently in biomedical research is how genetic variation encodes disease predisposition. Over 90% of the genetic signals identified by genome-wide association studies (GWAS) of type 2 diabetes (T2D) occur in non-protein-coding regions of the human genome [77]. This observation represents a unifying theme across common diseases [78]. Therefore, determining the mechanisms of disease predisposition driven by these non-coding genetic variants is central to biomedical research [79]. In order to understand the complex interactions between genetics and disease, it is necessary to understand the many layers by which the genotype propagates into phenotype.

The proposed mechanism by which non-coding genetic variation influences disease predisposition is through disruption of TF binding sites at key regulatory regions [78, 81]. Disrupting the DNA sequence at regulatory elements affects TF binding and chromatin organization at these regions. This, in turn, affects gene expression levels [40]. The changes in gene expression are reflected in protein levels, which lead to disrupted protein and metabolomic networks and, ultimately, to phenotypes such as disease status (Figure 1.5) [80].

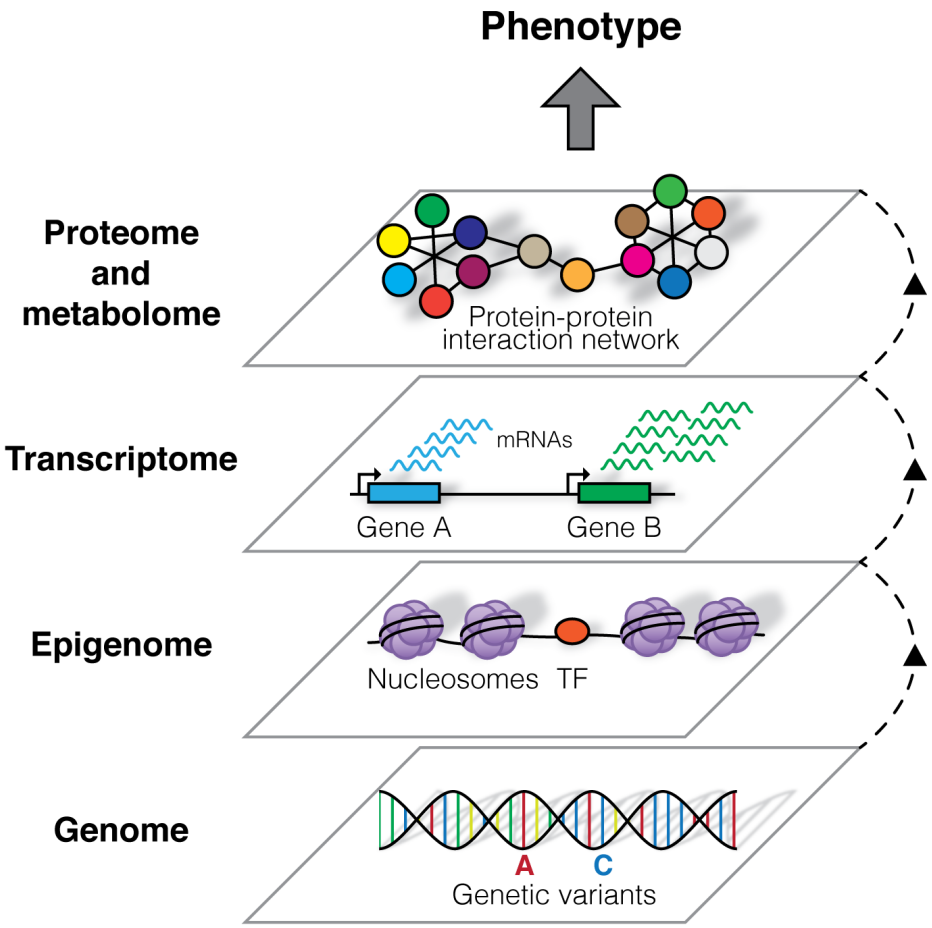


Figure 1.5: A simplified vision of the biological layers that link genotype to phenotype. Adapted from [80]. In reality, there are bidirectional interactions between layers as well as interactions that skip one or more layers.

A landmark study demonstrated that one of the non-coding genetic variants associated with obesity disrupts a binding site for the TF ARID5B at the FTO locus. This disruption leads to increased expression of the IRX3 and IRX5 genes during adipocyte development. These changes in gene expression, in turn, affect lipid storage and mitochondrial function, with affects body weight [82]. Since then, other studies (including from our group during the progress of this dissertation) have demonstrated the role of disease-associated genetic variants in altering TF binding (either by disrupting or creating TF binding sites) [33, 83–85]. These findings underscore that chromatin organization is the first layer affected by non-coding genetic variation. Therefore, studying the epigenome can provide important clues about how disease predisposition is encoded in the genome.

1.8 Information theory and biological information

The publication of *A Mathematical Theory of Communication* by Claude Shannon [86] established the theoretical framework to mathematically quantify information. Shannon's Information Theory principles are based on measuring the amount of entropy [randomness, $S(X)$] in a given signal (message). This can be done if one knows *a priori* the range of the signal (alphabet) by calculating the probability of the observed message given the alphabet (Equation 1.1). As a corollary, one can calculate the information content [$I(X)$], the amount of information in the signal (Equation 1.2, Figure 1.6).

$$S(X) = -\sum_i P(x_i) log P(x_i) \tag{1.1}$$

$$I(X) = 1 - S(X) \tag{1.2}$$

Shannon's Information has been applied to various levels of biological organiza-

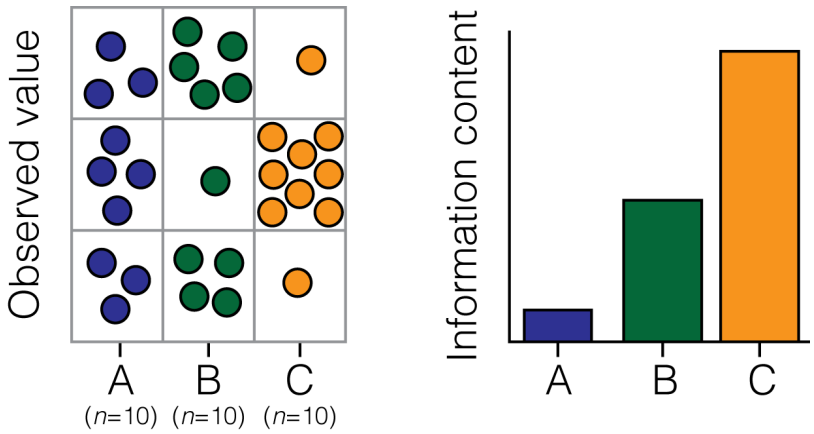


Figure 1.6: Schematic of information content calculation. Left: observed values for three groups (A-C) across three possible signal bins (*i.e.* alphabet length = 3). Right: Information content of groups A-C.

tion - DNA sequences [87], protein-protein interaction networks [88], and ecological communities [89]. However, few studies used an information theoretical approach to epigenomic data [90–92]. These studies found that the information content patterns in DNA methylation could be used to predict higher-order chromatin organization properties, such as TAD boundaries, and to prioritize genes affecting a phenotype of interest. It is, therefore, reasonable to expect that applying Information Theory to other epigenomic modalities will yield important insights in the understanding of genome organization. Specifically, we reason that chromatin accessibility is an ideal candidate for this type of analysis because it simultaneously indicates with high-resolution the location of regions with accessible chromatin and their underlying chromatin architecture.

1.9 Thesis outline

In this work, I developed and applied novel computational methods to analyze chromatin accessibility data and provide new insights into genome organization. This dissertation represents the fusion of two research domains - Information Theory and genomics. In Chapter II, I describe the work that constitutes the bulk of this thesis.

First, I describe an information theoretical approach to detect TF-chromatin interaction signatures using chromatin accessibility data. I then develop BMO, a new method for predicting TF binding using chromatin accessibility without relying on footprints. BMO outperforms current state-of-the-art computational approaches that rely on TF footprints. I then combine these two methodologies to systematically characterize TF-chromatin interaction patterns across multiple human tissues. Finally, I integrate these results with new and existing epigenomic molecular profiles to dissect the biological properties associated with TF-chromatin interactions, describing how they reflect biophysical and regulatory properties of TF-nucleosome interactions. In Chapter III, I describe a collaborative effort to characterize the epigenomic changes during thymocyte differentiation using high-quality chromatin accessibility molecular profiles. In Chapter IV, I lay out the future applications of my work in advancing our understanding of genome regulation and complex diseases.

CHAPTER II

Information Theoretical Properties of Transcription Factor and Chromatin Interactions

2.1 Abstract

Interactions between transcription factors (TFs) and chromatin are fundamental to genome organization and regulation and, ultimately, cell state. Here, we use information theory to measure signatures of TF-chromatin interactions encoded in the patterns of the accessible genome, which we term chromatin information enrichment (CIE). We calculate CIE for hundreds of TF motifs across human tissues and identify two classes: low and high CIE. The 10-20% of TF motifs with high CIE associate with higher protein-DNA residence time, including different binding sites subclasses of the same TF, increased nucleosome phasing, specific protein domains, and the genetic control of both chromatin accessibility and gene expression. These results show that variations in the information content of chromatin architecture reflects functional biological variation, with implications for cell state dynamics and memory.

2.2 Introduction

Chromatin is the association between DNA, RNA, and diverse nuclear proteins, including nucleosomes. It enables the ∼2-meter human genome to be packaged inside

the nucleus while allowing active genes and their corresponding regulatory elements to remain accessible [93]. Nucleosome positioning is an essential aspect of chromatin architecture and has been shown to have both passive and active roles in transcription factor (TF) binding [65, 67, 94]. Understanding TF-chromatin interactions is therefore critical to dissect the regulatory circuits leading to differences in transcriptional activity across diverse species, tissue, stimulatory, and genetic contexts. Information theory provides a powerful framework to quantify ordered patterns in data [86] and has been successfully used to characterize genome-wide DNA methylation patterns [90]. Here, we hypothesized that chromatin local architecture encodes rich signatures of TF interactions and developed information-theoretical tools to measure these patterns in human tissues.

2.3 Results

2.3.1 Chromatin information reflects TF-chromatin interaction patterns

We first aimed to quantify patterns of how chromatin accessibility informs TF-chromatin interactions. We reasoned that TF binding creates a localized impact on chromatin architecture, which may result in TF-specific signatures. To measure chromatin architecture, we focused on the assay for transposase-accessible chromatin using sequencing (ATAC-seq) [21], that can simultaneously quantify both TF and nucleosome signatures, which are reflected in the ATAC-seq fragment length patterns. This chromatin architecture can be visualized using V-plots [95], which show the aggregate ATAC-seq fragment midpoint distribution around TF binding sites and can result in a stereotyped "V" pattern of points for bound TFs that associate with nucleosome phasing (evenly positioned nucleosomes around TF binding sites; Figure 2.1A, upper plot). The extent of organization of data in the V-plot can be quantified using Shannon's entropy [86]. We calculated the information content of the ATAC-

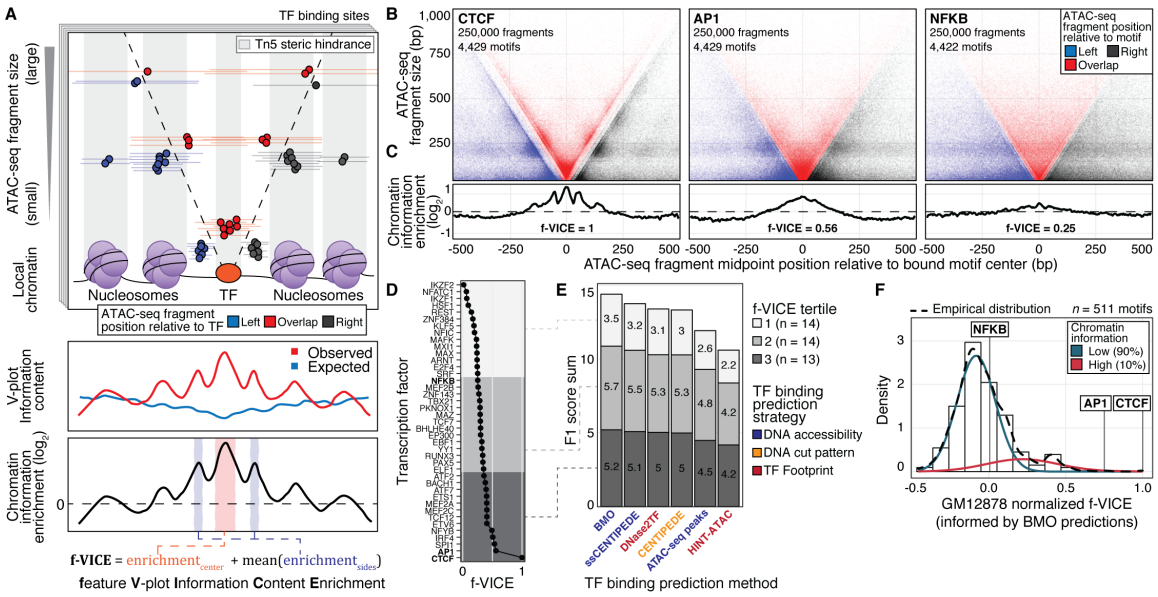


Figure 2.1: Information content of TF-chromatin interactions. (A) Upper: TF binding impacts the chromatin architecture and the observed ATAC-seq fragment distribution around TF binding sites. Middle and bottom: calculation of CIE and f-VICE. (B-C) V-plots and CIEs of CTCF, AP-1, and NFKB (GM12878 ATAC-seq data generated in this study). V-plots were downsampled to highlight differences in chromatin architecture (but not for f-VICE calculation). (D) f-VICEs calculated for TFs with GM12878 ChIP-seq data. (E) F1 score sum of TF binding prediction algorithms. Numbers inside bars represent the F1 score sum for TFs in that tertile. F1 scores reflect precision and recall at the cutoff threshold used to define predicted bound motif instances. (F) Normalized GM12878 BMO-informed f-VICE distribution.

seq fragment size distribution around TF binding sites as a way to quantify V-plot organization (Figure 2.1A, middle plot). To adjust for potential bias arising from non-uniform ATAC-seq coverage across the V-plot, we devised a metric called chromatin information enrichment (CIE) (Figure 2.1A, middle and lower plots). We summarized CIE into a single value, named feature V-Plot Information Content Enrichment (f-VICE), representing the CIE at landmark TF and nucleosomal positions across the V-plot, which are expected to have high CIE when the nucleosomes are phased around the TF binding site (Figure 2.1A, lower plot). Therefore f-VICE quantifies the degree of chromatin architecture organization around a TF. Higher f-VICE values indicate organized local chromatin around the TF binding site.

We initially focused on the GM12878 lymphoblastoid cell line, for which there is high-quality, deeply-sequenced ATAC-seq data [21] and 41 TF chromatin immunoprecipitation followed by sequencing (ChIP-seq) experiments that pass our inclusion criteria [24]. To increase our ability to detect TF-chromatin interactions, we generated an independent GM12878 ATAC-seq dataset with higher signal-to-noise ratio (Figure 2.4). Using these datasets, we created V-plots and calculated f-VICEs centered on bound motif instances for 41 TFs. The ATAC-seq fragment pattern was most ordered around CTCF, a known chromatin organizer [60], where we detected clusters of fragments distributed periodically in a "V" pattern indicating nucleosome phasing (Figure 2.1B-C, 2.5). Accordingly, CTCF f-VICE was highest among the tested TFs (Fig 1D). Other TFs, exemplified by AP-1 and NFKB, had diverse f-VICEs (Figures 2.1B-D, 2.5). These patterns were consistent across ATAC-seq libraries, indicating the robustness of the f-VICE metric (Figure 2.5). These results indicate extensive differences in the TF-chromatin interactions, which are captured in the CIE patterns.

2.3.2 Footprint-free prediction of TF binding and chromatin information

One alternative to determine f-VICEs for TFs without ChIP-seq data is to rely on binding predictions using chromatin accessibility data. This motivated us to first evaluate the performance of current TF binding prediction algorithms. Most algorithms search for footprints, which are regions of low chromatin accessibility embedded within larger accessible regions, thought to be caused by cleavage protection from bound TFs [51, 52, 54]. However, a recent report indicated that ∼80% of TFs do not have footprints [55]. Hence, we developed BMO, an unsupervised method to predict TF binding without relying on footprints. BMO classify motifs based on two separate features: 1) motif accessibility (number of ATAC-seq fragments) and 2) the number of additional co-occurring motifs in the motif vicinity. Both of these signals were shown to correlate with TF occupancy [56–58, 96]. The hypothesis underlying

BMO predictions is that genomic regions with many motif instances accessible and within proximity of each other will act as attractors to TF molecules diffusing in the nucleus, therefore increasing the likelihood of TF binding. One way to visualize this concept is to imagine the TF molecules as "Brownian bees" and the TF binding sites as flowers - the higher the number of open (accessible) flowers in the flower bed, the more likely the bees will interact with them (hence, bee-model of TF binding - BMO). We benchmarked BMO and other methods [49, 51, 52, 54] using TF ChIP-seq data from GM12878 and HepG2. To compare across methods, we used F1 scores, which account for the precision and recall at the thresholds used to separate between the predicted bound and unbound classes. Overall, the footprint-agnostic methods (BMO, CENTIPEDE, and a custom implementation of CENTIPEDE, called ssCENTIPEDE) outperformed footprint-based methods on most tested TFs, particularly on those with lower f-VICEs (Figures 2.1E, figs. 2.6 to 2.10; Supplementary Results). These findings indicate that TF binding is more accurately predicted using a simple chromatin accessibility model tuned to each TF motif and that footprinting-based methods are more sensitive to the local TF-chromatin architecture.

2.3.3 Chromatin information varies across TFs

Having determined that footprint-based methods are less accurate for predicting TF binding, we proceeded with BMO predictions to estimate f-VICEs for TFs without ChIP-seq data. BMO-predicted f-VICEs were significantly correlated with f-VICEs calculated from TF ChIP-seq data (Pearson's $\rho \geqslant 0.72$, $p \leqslant 1e-10$; Figure 2.11). We therefore concluded that BMO can be used to estimate f-VICEs without ChIP-seq data and performed BMO TF binding predictions to calculate f-VICEs for 540 non-redundant TF motifs. We used high-quality ATAC-seq datasets from four additional human tissues (pancreatic islets [84], pancreatic islet sorted alpha and beta cells [97], and CD4+ cells [98]), selected applying a strategy that uses the highly

stereotyped chromatin architecture in ubiquitous and conserved CTCF/cohesin binding sites to infer sample quality (Figure 2.12). We normalized f-VICEs within each sample using linear regression models to control for differences in bound motif predictions and overall chromatin accessibility (Figure 2.13). The resulting normalized f-VICE value represents how much the chromatin information deviates from the expected chromatin information given the motif accessibility and number of predicted bound instances, with positive values indicating more organized local chromatin. The majority of the 540 TF motifs followed an approximately normal f-VICE distribution, but we observed an upper tail with higher f-VICEs resulting from potentially from a mixture of two separate f-VICE distribution. This motivated us to fit two Gaussian distributions to the data in order to classify between low or high f-VICE motifs. The median percentage of motifs associated with high f-VICEs across datasets was 14% (Figures 2.1G, 2.14), which is comparable to the percentage of motifs associated with DNase footprint protection across datasets (median=19%) from another study [55] and supports our conclusion that footprint-based algorithms will not perform well on most TFs. Together, these results reinforce the use of BMO for accurately calculating f-VICE and indicate that a minority of TFs associate with high CIE.

2.3.4 Chromatin information is associated with TF-DNA residence times

TF residence time, which corresponds to the duration of DNA binding for a TF, is an important biophysical measurement that can influence TF activity [67, 68]. Based on the high f-VICEs for CTCF and AP-1 and low f-VICE for NFKB (Figure 2.1C-D), which agree with the known residence times for these TFs (Table 2.1), we hypothesized that CIE correlates with residence time. We correlated BMO-informed f-VICEs with previously measured fluorescence recovery after photobleaching (FRAP) data from mammalian cell lines (Table 2.1), which provide an upper bound of TF residence time [99, 100]. Using a robust linear regression to protect against outlier

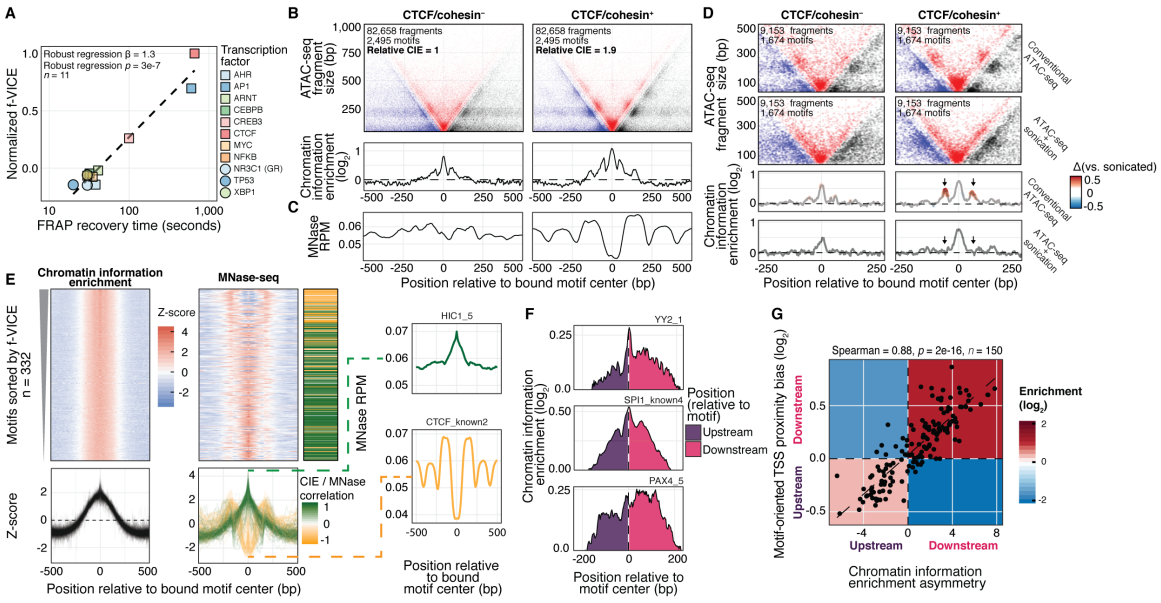


Figure 2.2: Chromatin information informs residence times and TF-nucleosome interactions. (A) Correlation of FRAP recovery times and GM12878 f-VICEs. Dashed line, linear model fit. (B) V-plots and CIEs of CTCF/cohesin$^+$ and CTCF/cohesin$^-$ motifs. (C) GM19238 MNase-seq reads per million mapped reads at the same motifs. (D) CTCF/cohesin$^+$ and CTCF/cohesin$^-$ motifs in the sonicated and conventional GM12878 ATAC-seq data. Colors, differences relative to sonicated. (E) Left: CIE and MNase-seq profiles (k-means cluster three). Middle: Heatmap of MNase and CIE Z-score correlations. Right: Example motifs with positive and negative CIE/MNase correlation. (F) Top 3 motifs with CIE asymmetry Z-scores in GM12878. (G) Scatter plot of motif-oriented TSS position bias and CIE asymmetry in TSS-proximal motifs. Enrichments calculated by permuting the signs of observed values (n=10,000).

influence, we found that f-VICE significantly correlated with FRAP recovery times in all samples ($\beta \geqslant 0.7$, Bonferroni adjusted $p \leqslant 0.001$; Figures 2.2A, 2.15). This suggests that TFs associated with high CIE have longer residence times.

CTCF and cohesin co-bind at CTCF motifs to regulate chromatin loop maintenance [101]. A recent study found that cohesin has a residence time 10- to 20-fold higher than CTCF [100]. We reasoned this difference reflected in the local chromatin architecture and calculated the CIE of GM12878 CTCF binding sites with and without the presence of cohesin (CTCF/cohesin$^+$ and CTCF/cohesin$^-$), controlling for potential confounding biases from motif strength, ATAC-seq and ChIP-

seq signal intensities (Figure 2.16A). CTCF/cohesin$^+$ had 1.9-fold higher CIE compared to CTCF/cohesin (Figures 2.2B, 2.16B), indicating these distinct CTCF occupancy classes have different CIE signatures. We next compared the nucleosome positioning signals inferred from lymphoblastoid cell line micrococcal nuclease sequencing (MNase-seq) profiles . Only the CTCF/cohesin$^+$ class had phased nucleosomes around the binding site (Figures 2.2C, 2.16C), consistent with longer residence times associating with nucleosome phasing. To experimentally validate these results, we generated chromatin accessibility data using a modified ATAC-seq protocol with an additional sonication step to disrupt the fragment size information (Figure 2.17). There were no detectable nucleosome phasing patterns in the motif-flanking CIE of the sonicated sample (Figure 2.2D; see vertical arrows). These results show that CIE signatures of the two classes of CTCF binding result from differences in TF-chromatin interactions instead of differences in chromatin accessibility.

2.3.5 High chromatin information TFs associate with nucleosome phasing

To systematically characterize the association between CIE and nucleosome positioning, we compared GM12878 CIE patterns across TF motifs to lymphoblastoid MNase-seq profiles. First, we used k-means clustering to divide motifs into broad CIE shape categories based on their Z-scores. We found three clusters representing a continuum of CIE at the motif region (Figure 2.18A). Clusters one and two had "through-shaped" CIE shapes, with lower CIE at the motif compared to motif-flanking regions, while cluster three had a "peak-shape", with the highest CIE at the motif region (Figure 2.2E, 2.18A-B) and encompassed >95% of the high f-VICE motifs (Figure 2.18C-D). Notably, we observed two distinctly anti-correlated MNase signal patterns for the motifs in cluster three, corresponding to one group of motifs with high MNase signal at the motif center and another with high MNase signal at the motif flanking regions, but not at the motif (Figure 2.2E). This result is consistent with

TFs binding at the center of the nucleosome dyad or between phased nucleosomes [65] and further suggest that a subset of the TFs that cannot evict nucleosomes encounter less steric hindrance and bind at the dyad center, comparably with what has been observed in SOX2 [75]. CIE and MNase signals were anti-correlated at high f-VICE motifs (Fig 2.2E; yellow-green heatmap), indicating that the highest CIE TFs associate with nucleosome phasing. We quantified nucleosome phasing and found that it was significantly correlated with f-VICE in clusters two and three (Spearman's $\rho \geqslant 0.42$, $p \leqslant$1e-7 Figure 2.19). These results suggest that TF-chromatin interaction patterns are driven by TF residence time, resulting in distinct CIE signatures.

2.3.6 Chromatin information asymmetry at TF motifs

Previous reports suggested that a subset of TFs directionally bind DNA, with potential effects on gene regulation [51, 102, 103]. To investigate this further, we extended our information content analyses to quantify CIE asymmetry. Of the 540 motifs tested, 150 had significantly asymmetric CIE (Bonferroni corrected p<0.05; Figures 2.2F, 2.20A). The direction of CIE asymmetry was correlated with the direction of the nearest TSS relative to each motif instance (Spearman's ρ=0.66, p=2e-16; Figure 2.20B). To determine if this result was an artifact of TSS proximity, we calculated CIE asymmetry separately for TSS-proximal ($\leqslant$ 1 kb) and TSS-distal ($\geqslant$ 10 kb) motif instances. The TSS-distal and TSS-proximal CIE asymmetry directions agreed more than expected by chance (111/150, binomial test p=4e-9; Figure 2.20C-D), suggesting that CIE asymmetry is intrinsic to the TF motif. The magnitude of asymmetry was higher in TSS-proximal motifs (Figure 2.20D), suggesting that TSS proximity amplifies TF CIE asymmetry. Accordingly, the correlation between nearest TSS direction and CIE asymmetry was stronger at TSS-proximal motifs (Spearman's $\rho = 0.88$, p=2e-16; Figure 2.2G). These results support that directional binding is a property of TF-chromatin interactions.

2.3.7 Chromatin information patterns are tissue-specific and associate with genetic control of gene expression

We next aimed to investigate cross-tissue differences in CIEs. We performed an unsupervised hierarchical clustering of motif f-VICEs and found that it recapitulated the expected tissue grouping (Figure 2.3A). The motifs driving the clustering patterns included known tissue-specific transcriptional regulators (Figure 2.21), consistent with CIE reflecting TF activity. A recent study demonstrated that NF-KB (p65) residence time is determined by its DNA-binding domain (DBD) [104], which motivated us to ask if DBDs are associated with CIE. We assigned high-confidence DBDs and protein domains to motifs and designed a permutation-based rank test to calculate DBD f-VICE enrichments. We observed both common and tissue-specific f-VICE enrichments, including IRF and ETS in blood-related samples, PAX in islet-related samples, and HMG/SOX and FOX domains in HepG2 (FDR $< 10\%$; Figures 2.3B, 2.22). Our findings show the landscape of TF-chromatin interactions varies across tissues and reflects properties of TF biology.

The prevalence of tissue-specific differences in CIEs led us to examine the role of high f-VICE TFs in regulating gene expression. We calculated the enrichment of the motifs categorized as high or low f-VICE in GM12878 (Figure 2.1F) to overlap lymphoblastoid *cis*-expression quantitative trait loci (*cis*-eQTLs) datasets [32, 34], which represent gene expression genetic control regions in these cells. High f-VICE motifs had 15-30% higher (median=24%) fold-enrichment in *cis*-eQTLs compared to low f-VICE motifs (Figures 2.3C, 2.23A), but no differences in eQTL effect sizes (Figure 2.23B). These results indicate that high f-VICE TFs are more likely to mediate genetic effects on gene expression, but not their magnitude.

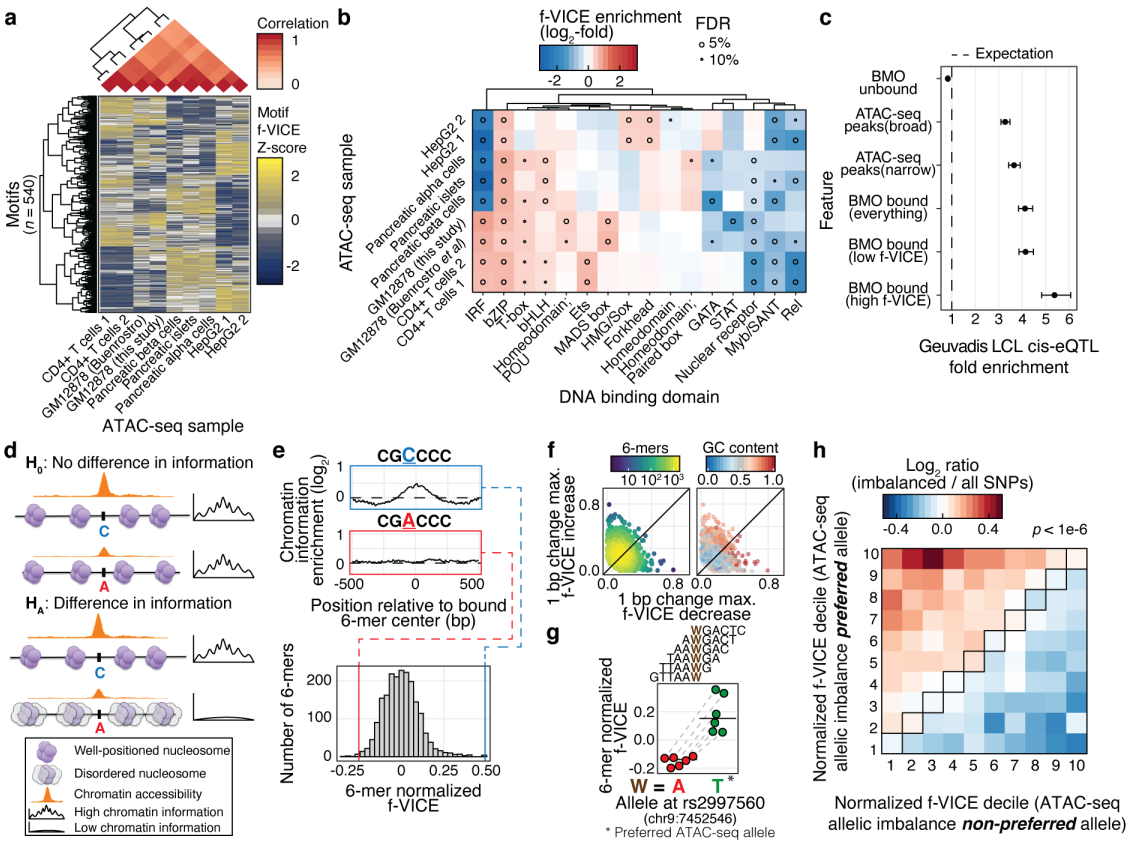


Figure 2.3: The chromatin information landscape of human tissues. (A) Hierarchical clustering of f-VICE Z-scores. (B) f-VICE enrichments across DBDs. (C) LCL *cis*-eQTLs enrichments. Error bars, effect size SD. (D) Hypothesis schematic. (E) Upper: Two 6-mers with Hamming distance of 1. Lower: pancreatic islets 6-mer normalized f-VICE distribution. (F) Range of f-VICE differences associated with 1-bp difference in 6-mer sequence. (G) Predicted f-VICE change associated with rs2997560, which has ATAC-seq allelic imbalance in pancreatic islets (T is the preferred allele). Horizontal bars, median. (H) Log_2 ratio of f-VICE decile changes associated with the preferred and non-preferred alleles of imbalanced SNPs versus all tested SNPs in pancreatic islets.

2.3.8 High chromatin information TF motifs are associated with chromatin accessibility

Given the highly ordered chromatin (Figure 2.1), high predicted residence times (Figures 2.2A, B, 2.15), and nucleosome phasing properties (Figure 2.2E, 2.19) associated with high f-VICE TFs, we hypothesized that their regulatory effects (Figure 2.3C) could result from acting as or recruiting pioneer factors (TFs that induce chromatin accessibility) [51, 71]. If true, we would expect increased CIE for single nucleotide polymorphism (SNP) alleles with increased chromatin accessibility (*i.e.* with ATAC-seq allelic imbalance; Figure 2.3D). Because we do not have sufficient coverage at a single locus to detect changes in chromatin information. we first performed a motif-agnostic approach to calculate the f-VICEs associated with every DNA 6-mer in the human genome, using linear regression models to control for differences in chromatin accessibility and number of BMO predicted bound 6-mer instances. This strategy allows the interrogation of genetic variants by determining the DNA 6-mers formed by each allele and their corresponding f-VICEs. DNA 6-mers have a distribution of f-VICEs (Figures 2.3E, 2.23A) and GC-pure 6-mers had the highest f-VICEs (Figure 2.23B), which is consistent with GC-rich sequences driving enhancer activity [105] and suggest that high GC-content regions represent anchors of nuclear architecture. Notably, a single base-pair change can lead to large differences in 6-mer f-VICEs (Figures 2.3E-F, 2.23C-E), suggesting that genetic variation impacts CIE. We separated the DNA 6-mers into f-VICE deciles and found that they had different biological properties (Figure 2.24B-C). Using the f-VICEs obtained from the predicted bound 6-mers, we predicted the f-VICEs changes associated with either allele at SNPs with significant ATAC-seq allelic imbalance (binomial test $p<0.05$) in GM12878 and pancreatic islets. The preferred ATAC-seq alleles were significantly biased to form higher f-VICE 6-mers compared to the less favored allele (permutation test $p<3e-4$; Figure 2.3G-H, 2.25). These findings support a model where TFs with distinct properties

(pioneers) bookmark regions of the genome to allow binding of other TFs (migrants) [51, 71]. Notably, TF motifs that are predictive of binding without any chromatin accessibility data (based solely on the motif match score) have significantly higher f-VICEs in GM12878 and HepG2 (robust linear regression $p \leqslant 0.001$; Figure 2.25). This suggests that high f-VICE TFs, like CTCF, are more likely to bind any strong motif, while the remaining TFs require motifs located in accessible regions.

2.4 Discussion

In this study, we develop and use for the first time, entropy-based algorithms to analyze chromatin accessibility data and dissect TF-chromatin interaction patterns across human tissues. TF-chromatin interactions are captured in the information content patterns of chromatin accessibility and reflect functional properties of TFs, such as TF-DNA residence times, nucleosome phasing, and protein DNA binding domains. We find that a subset of TFs (10-20%) have high chromatin information and are more highly associated with the genetic control of both chromatin accessibility and gene expression, therefore defining cell state. In addition, we show that footprinting-based algorithms to predict TF binding are sensitive to the TF-chromatin information landscape we describe. We develop and cross-validate a novel tool for predicting TF binding based on chromatin accessibility that outperforms footprinting-based methods. Collectively, our results show a dynamic landscape of TF-chromatin interactions, with implications for gene regulation and cell state memory.

2.5 Limitations

One of the major limitations from our methodology is the reliance on TF motifs to calculate chromatin information patterns. This approach does not allow us to determine whether the observed results reflect the putative TFs binding to the motifs

or if these are due to other proteins recruited to the motif. One example is the CTCF and cohesin interactions dissected in this study, which was only possible using ChIP-seq data to resolve which binding sites were occupied by one or both proteins. Therefore, additional data will be necessary to address similar scenarios on a case-by-case basis. Similarly, we did not address issues of motif proximity, which could potentially affect our interpretation of results from TFs that frequently bind in close proximity.

2.6 Supplementary Results

BMO builds on previous reports that the degree of chromatin accessibility around a motif [56, 57, 96] and the presence of co-occurring motifs [58] positively correlates with TF binding, and uses TF-specific negative binomials of these two signals to estimate the likelihood of a bound instance. We benchmarked the performance of BMO and other unsupervised TF binding prediction algorithms using ATAC-seq datasets from the GM12878 and HepG2 [106] cell lines and their corresponding TF ChIP-seq data (n=41 and n=59, respectively. We compared BMO to three footprinting-based algorithms (HINT-ATAC [54], DNase2TF [52], PIQ [51]), to CENTIPEDE, which learns informative DNA cut patterns indicating TF binding [49], and to a baseline classifier that labels TF motifs within ATAC-seq peaks as bound. To evaluate methods, we calculated the area under the precision-recall curve (AUC-PR), which informs the performance of the classifier in ranking bound and unbound motif instances, and the F1 score, which measures the performance of the threshold used to call bound motif instances.

BMO outperformed all methods in our high-signal GM12878 dataset (Figure 2.1E), whereas BMO and CENTIPEDE had similarly high performance in lower-signal datasets (Figures 2.4, 2.6). DNase2TF had lower performance in the lower-signal datasets (Figure 2.6). PIQ cannot use custom TF motif scans and therefore

required separate benchmarking, which revealed lower performance compared to BMO (Figure 2.8). These results were consistent across ATAC-seq replicates and cell lines, including downsampled data representing shallower sequencing depths (Figure 2.9). Of note, the AUC-PR of footprinting-based methods was lower overall due to their inability to classify motifs occurring outside ATAC-seq peaks, which we reason contain true TF binding sites and negatively affect PR-AUCs. While their F1 scores indicate that this effect is less pronounced when taking into account the thresholds to call bound motif instances, their performance was still consistently lower than non footprinting-based methods (Figure 2.6). Overall, the two footprinting-agnostic methods (BMO and CENTIPEDE) outperformed footprint-based methods on most tested TFs, indicating that TF binding is more accurately predicted using a simple chromatin accessibility model tuned to each TF motif.

We next sought to determine if the CENTIPEDE approach relied on spatial DNA cut patterns, or if the overall accessibility in the region was sufficient for high performance. We devised an alternative implementation of CENTIPEDE that ignores the DNA cut positions (signal-sum CENTIPEDE; ssCENTIPEDE) and masks any footprint-like patterns. This ssCENTIPEDE approach performed almost identically to CENTIPEDE (Figures 2.10, 2.1E), again indicating that footprint patterns in chromatin profiles are not necessary for high prediction performance. One corollary expectation from this conclusion is that footprint-based algorithms should perform comparatively worse when predicting binding for TFs with a low impact on local chromatin. To test this, we compared performance across f-VICE tertiles representing low (tertile one), intermediate (tertile two), and high (tertile three) f-VICEs. Notably, BMO and (ss)CENTIPEDE had relatively higher performance on lower f-VICE tertiles one and two (Figures 2.1E, 2.6). Our findings indicate that footprinting-based methods are more sensitive to the local TF-chromatin architecture.

2.7 Methods

GM12878 cell culture. We cultured GM12878 cells following the ENCODE GM12878 cell culture protocol (`www.encodeproject.org/documents/1bb75b62-ac29-4368-9855-68d410e1963a`), with added plasmocin (Invivogen, San Diego, CA; 50 ug/mL) to the growth media to prevent mycoplasma contamination.

GM12878 ATAC-seq data generation. We conducted ATAC-seq as described in [107] using a home-made Tn5 that we synthesized as described in [108]. For each replicate we incubated 250,000 cells with 12.5 μL of 1:1 mix of Tn5 enzyme that carry 5-methylC-MEDS-A oligos and MEDS-B oligos at 37° C for 30 minutes in a 50 μL reaction. We column-purified the tagmented DNA using the Zymo DNA Clean & Concentrator-5 kit (Zymo Research, Irvine, CA) and constructed Illumina sequencing library using the Kitzman lab custom indexing primers. We PCR-amplified a total of 11 cycles until amplification curve reached its mid-log phase ($\frac{1}{3}$ to $\frac{1}{2}$ of max signal), and then purified the PCR products using SPRI beads prepared as in [109] and eluted in 22 μL of TTE8 buffer. Sequencing was performed on an Illumina HiSeq 4000 platform at the University of Michigan Sequencing Core and a total of ∼33 million paired-end 52 bp reads were generated.

Sonicated GM12878 ATAC-seq data generation. For each replicate we incubated 250,000 cells with three different concentrations of enzyme (0.2X, 1X, and 5X; 1X corresponds to 2.5 μL of Tn5 that carry 5-methylC-MEDS-A oligos) at 37° C for 30 minutes in a 50 μL reaction. We column-purified the tagmented DNA using the Zymo DNA Clean & Concentrator-5 kit (Zymo Research, Irvine, CA), and sonicated to ∼350 bp using the Covaris M220 sonicator (peak incident power - 50W; duty factor - 20%; cycles per burst - 200; treatment time - 60 sec). We constructed Illumina sequencing library using the ACCEL-NGS Methyl-seq DNA Library kit (Swift Bio-

sciences #DL-ILMMS-12; revision 160106) with the following modifications to the manufacturer's protocol: 1) We skipped "Ligation" and "Post-ligation SPRI" steps (pg. 10-11), as the 5' end of the fragments had already been tagged during the transposition step. Accordingly, we eluted DNA with 20 muL of TTE8 (10 mM Tris-HCl, 0.1 mM EDTA, 0.05% Tween-20, pH 8) for Post-Extension SPRI step (pg. 10), instead of 15 μL, to adjust for the difference in volume before proceeding to the "Indexing PCR" step (pg. 11); 2) We used 2:1 beads:sample ratio for "Post-Extension SPRI" step (pg. 10) and 1.8:1 beads:sample ratio for "Post-PCR SPRI" step (pg. 12); and 3) For indexing PCR, we used the Kitzman lab custom primers (barcode plate #5) to prime the P5 end and the "IndexD7XX" primers (Swift Biosciences #DI-ILMMS-48) to prime the P7 end. We PCR-amplified a total of 14 cycles for 0.2X, 1X samples and 16 cycles for 5X samples until amplification curve reached its mid-log phase ($\frac{1}{3}$ to $\frac{1}{2}$ of max signal), and then purified the PCR products using SPRI beads prepared as in [109] and eluted in 22 μL of TTE8 buffer. Sequencing was performed on an Illumina HiSeq 2500 platform at the University of Michigan Sequencing Core and a total of $\sim$33 million paired-end 126 bp reads were generated.

ATAC-seq data processing. Reads were trimmed for barcodes and aligned to the hg19 reference human genome using BWA mem (v. 0.7.15) [110] similarly to our previous study [33], with additional parameters -I 200,200,5000 to avoid larger ATAC-seq fragments being discarded. We removed duplicate alignments using Picard (broadinstitute.github.io/picard) and retained properly paired and uniquely mapped alignments with high mapping quality using samtools view (v. 1.3.1) [111] with flags -f 3 -F 4 -F 8 -F 256 -F 1024 -F 2048 -q 30. We called broad and narrow peaks using MACS2 (v. 2.1.1.20160309) [112] with flags -g hs –nomodel –shift -100 –extsize 200 -B [–broad] –keep-dup all and kept peaks that did not intersect blacklisted regions (`sites.google.com/site/anshulkundaje/projects/blacklists`), using bedtools

(v2.26.0) [113], and that reached 5% FDR. All data was processed uniformly using Snakemake [114].

Motif processing. We used the PWM scans from [33]. Briefly, we used biallelic SNPs and short indels from the 1,000 Genomes project (release v5) [115] to generate comprehensive scans with FIMO [48], using the background nucleotide frequencies from hg19 and a p<1e-4. We only kept motif instances that intersected mappable regions and did not intersect blacklisted regions (`sites.google.com/site/anshulkundaje/projects/blacklists`). In order to reduce motif redundancy, we performed PWM clustering in our motif database using the matrix-clustering tool from RSAT [50], with parameters -lth cor 0.7 -lth Ncor 0.7. For each of the 540 clusters obtained, we retained the motif with the highest total PWM information content for downstream analyses.

V-plots, chromatin information enrichments, and f-VICEs. V-plots [95] were generated by creating a matrix of aggregated fragments from the selected set of genomic features (motifs), removing all instances that overlap within ±500 bp of each other. We used the script measure_signal (using flags -r 500) , which is part of a suite of tools to analyze ATAC-seq data we developed for this study (`github.com/ParkerLab/atactk`). Each cell in the V-plot matrix outputted by measure_signal correspond to the number of fragment midpoints at a position relative to the feature center (x-axis) and fragment size (y-axis). We binned the matrix in the x-axis using a sliding window of 10 bp width, with 2-bp overlap, and summed all the values within the window corresponding to a given fragment size. For each x-axis bin, we calculated the normalized information content (I) of the corresponding fragment size distribution (y-axis) using the equation 2.1, where H(x) is the maximum-likelihood Shannon's entropy function implemented of the entropy R package [116], and H_{max} is the Shannon's entropy of the information length (*i.e.* the range of fragment size

distribution of the entire V-plot). To calculate the expected normalized information content, we randomly permuted the position and size labels in the fragments and repeated the steps outlined above.

$$I(x) = 1 - \frac{H(x)}{H_{max}} \tag{2.1}$$

Chromatin information enrichment was calculated as the log_2 of the observed normalized information content divided by the expected normalized information content. For Figure 2.1C, V-plots were downsampled to equal number of ATAC-seq fragments and motifs between TFs by selecting the top n motifs ranked by the number of ATAC-seq fragments within ±500 bp, and then further downsampling to 250,000 fragments, where n represents the smallest number of bound motifs among the plotted TFs. These downsampled V-plots were only used for visualization purposes and not used for the f-VICE calculations described below.

To obtain the feature V-plot information content enrichment (f-VICE) for each motif, we summed of the average chromatin information enrichment in the V-plot regions corresponding to the center (-25 to 25 bp) and proximal (-70 to -50 and 50 to 70 bp) chromatin information enrichment peaks referent to the CTCF V-plot, which correspond to small fragments spanning the TF binding site and to those positioned between the TF and the first pair of proximal nucleosomes, respectively (Figure 1A). This value was then normalized across all motifs using the residuals of the linear model f-VICE $\sim log_{10}(m) + log_{10}(f)$, where m corresponds to the number of predicted bound motif instances for each motif and f corresponds to the total number of ATAC-seq fragments at the predicted bound motif instances for each motif (Figure 2.13). Negative values indicate that the feature has less chromatin information information than expected based on its accessibility. The residuals for each sample were divided by the corresponding CTCF value in that sample to normalize it to 1. The linear model normalization was not performed in the ChIP-seq f-VICEs reported in Figures

2.1D and 2.11 due to lack of data points to accurately fit the linear model.

Additional ATAC-seq samples selection. In addition to the GM12878 ATAC-seq dataset generated for this study, we analyzed an additional eight publicly available datasets corresponding to pancreatic islets [84, 97], CD4+ cells [98], GM12878 [21], and HepG2 [106]. With the exception of HepG2, these datasets were selected from a survey of all the public ATAC-seq datasets available until the end of 2017. We selected for our analyses datasets with at least 20 million high-quality autosomal reads and transcription start site (TSS) enrichment $\geqslant 6$. In addition, we only retained samples with the stereotypical chromatin information enrichment indicative of nucleosome phasing at ubiquitous and conserved CTCF-cohesin binding sites. These regions provide a reference V-plot, with expected high accessibility and periodical chromatin information enrichment patterns in any high-quality sample. The ubiquitous and conserved CTCF-cohesin sites were defined as CTCF motifs overlapping ENCODE CTCF and Rad21 ChIP-seq peaks in at least in at least 54/59 (CTCF) and 2 (Rad21) different human tissues, located in bi-directionally mappable regions between human and mouse using bnMapper [117] that also corresponded to CTCF motif matches in the mm9 reference genome. To quantify samples, we defined our high-quality GM12878 dataset as reference and calculated the chromatin information enrichment correlation $\leqslant 200$ bp from the motif center. Samples with correlation $<$ 0.8 (Spearman) were discarded (Figure 2.12). Finally, we only retained tissues/cells that had at least two samples that passed our stringent selection criteria.

BMO transcription factor binding prediction. BMO builds on previous reports that the degree of chromatin accessibility around a motif [56, 57, 96] and the presence of co-occurring motifs [58] positively correlate with TF occupancy. BMO fits per-motif negative binomial (NB) on these two signals to estimate the likelihood of a TF motif instance being bound. BMO performs three steps: 1) calculate the back-

ground ATAC-seq fragment NB distribution, 2) calculate the co-occurring motifs NB distribution, and 3) combine the p values from the two distributions.

Using all genomic matches for a given motif PWM, we calculated the number of ATAC-seq fragments overlapping a region ±100 bp from every motif instance. We ignored fragments that integrated within the motif coordinates in order to mitigate ATAC-seq bias, as the nucleotide sequence in the motif regions is relatively constant across features and is more subject to assay-specific bias compared to the motif-flanking regions. The background ATAC-seq fragments NB distribution was fitted on 10,000 randomly selected motif instances of the same PWM occurring outside ATAC-seq peaks. We repeated this step 100 times and used the average mean and overdispersion as the NB parameters. This sampling approach is 1-2 orders of magnitude faster compared to using all motif instances outside ATAC-seq peaks, yielding identical results on a representative subset of our data that accounted for the number of motif matches per PWM. We then used this NB distribution to calculate the ATAC-seq signal p value of every instance of that motif PWM.

The co-occurring motifs NB was obtained by determining the number of additional instances of the same motif PWM within ±100 bp of every motif instance. We used this distribution to to calculate the p value of the number of co-occurring motifs per motif instance.

The two p values from the ATAC-seq fragment and co-occurring motif negative binomials were combined by summing their Z scores [118] using the sumz function of the R package metap. This yielded a single p value representing chromatin accessibility and number of co-occurring motifs. A given motif instance will have more significant p values if it is located in accessible chromatin and/or have many instances of the same motif nearby. Multiple testing correction was performed using the Benjamini-Yekutieli method [119] and motif instances were considered bound where the adjusted p value < 0.05. Fitted NB distributions were obtained using the R packages MASS

[120] and fitdistrplus [121].

CENTIPEDE For each PWM scan result, we generated a strand-specific (relative to the motif orientation) base-pair resolution matrix encoding the number of Tn5 transposase integration events in a region ±100 bp from each motif occurrence using make_cut_matrix with parameters -d -r 100. This matrix and the motif PWM score were used as input for CENTIPEDE (v. 1.2), and a motif occurrence was considered bound if the outputted posterior probability was higher than 0.99. To calculate AUC-PRs, we used the posteriors outputted by the software as scoring metric. We developed make_cut_matrix as part of atactk (`github.com/ParkerLab/atactk`).

Signal-sum CENTIPEDE (ssCENTIPEDE). To run ssCENTIPEDE. we performed CENTIPEDE predictions using the total number of DNA cuts in the vicinity of each motif instance as input instead of the positions of the DNA cuts. This strategy informs motif accessibility while omitting positional patterns that can be used by CENTIPEDE as a signature of TF binding. We ran CENTIPEDE similarly as described above, with the only difference being that we summed across the rows of the input Tn5 cuts matrix to generate a one-column matrix containing the total number of Tn5 cuts in the motif vicinity. This ensures that the positional information (i.e. where the transposition events occur relative to the motif) is omitted from CENTIPEDE.

DNase2TF. In order to run DNase2TF (v. 1.0) on ATAC-seq data, we offset all the cut points calculated using paired_end_bam2split.r by 4 bp before using them as input to the software, which was run with default parameters. We intersected the called footprints with each motif file and considered bound those motif instances that intersected a footprint with FDR < 0.05.

HINT. We performed footprinting analyses with HINT-ATAC (RGT v. 1.1.1) using as input the broad ATAC-seq peaks and filtered BAM file from each sample. In their methods, the authors used MACS2 narrow peaks, but we found that they had lower performance compared to broad peaks (Figure 2.7), so we used the latter for the analyses. We intersected the HINT output file with each motif file and considered bound every motif instance that intersected a footprint.

PIQ. We performed PWM scans using the pwmmatch.exact.r script included with PIQ (v. 1.3). BAM files were processed with bam2rdata.r due to an error in the code of pairedbam2rdata.r which prevented any of our BAM files from being processed. Footprinting was performed using the *pertf.r* script. Because PIQ performs its own PWM scans, we compared PIQ to BMO only on PWM matches that were shared between PIQ and BMO (using bedtools intersect).

Dataset downsampling. In order to compare the TF binding prediction methods across multiple sequencing depths, we uniformly downsampled BAM files using the -s flags of samtools view (v1.9), which downsamples files while maintaining read pairs intact (this behavior is not present in version 1.3). These downsampled files were used as input for peak calling and all other steps required prior to running each TF binding prediction method.

TF binding evaluation. We defined as true positives for a given TF all motif matches that fully intersected a ChIP-seq ENCODE conservative irreproducible discovery rate (IDR) narrow peak in the respective sample. We only analyzed TFs that had motifs in our database and at least 1,000 bound motif instances. For TFs with multiple motifs, we selected the motif with the highest total PWM information. For TFs with multiple ChIP-seq experiments, we selected the one with the highest number of bound motifs. To evaluate methods, we calculated the area under the

precision-recall curve (AUC-PR), which informs the performance of the classifier in ranking bound and unbound motif instances, and the F1 score, which measures the performance of the threshold used to call bound motif instances. We did not use areas under the receiver-operator characteristic curve (AUC-ROC) given the highly skewed class imbalance between bound and unbound motifs, which makes AUC-ROCs an unreliable metric to evaluate TF binding predictions [122, 123]. AUC-PRs were calculated using packages ROCR (v. 1.0-7) and PRROC (v. 1.3) in R [124, 125]. To rank predictions, we used the $-log_{10}$ adjusted p values for BMO, the number of reported tags from HINT, the posteriors calculated by CENTIPEDE, the $-log_{10}$p values calculated by DNase2TF, the purity score outputted by PIQ, and MACS2 $-log_{10}$ p values for motifs in peaks. F1 scores were calculated using the equation 2.2 at the following thresholds for each method: BMO adjusted p value$<$0.05, CENTIPEDE posterior$\geqslant$0.99, any motif instance overlapping a HINT predicted footprint, any motif instance overlapping a DNase2TF predicted footprint with FDR value $<$ 0.05, and any motif instance called bound by PIQ.

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2.2}$$

Mixture models for f-VICE distributions. High and low f-VICE distributions were calculated using the R package mixtools (v. 1.1.0) [126] using as input the normalized f-VICEs for each ATAC-seq sample, after removing low signal motifs, where the number of predicted bound instances for the motif was contained in the lowest decile of that sample. We used as cutoff a posterior probability of 0.5 to split between the high and low f-VICE distributions.

FRAP/f-VICE robust regression and CTCF-Cohesin regions comparisons. To measure the correlation between FRAP recovery times and f-VICE, we performed a literature search of reported FRAP recovery times, which are referenced in Table

2.1. Robust linear regressions of f-VICE and FRAP recovery times were performed with the rlm function of the R package MASS (v. 7.3-50) [120]. Te model used was f-VICE $\sim \log_{10}$(FRAP recovery time), without scaling or centering of variables. For each TF with FRAP recovery times, we used the f-VICE from the motif with highest total PWM information content in our database. f-VICEs for these motifs were normalized using the same linear regression model described earlier, but including all the motifs in our database (n=1,850). For each sample, we required that the gene corresponding to each TF had RNA-seq TPM $\geqslant 1$ in a related tissue in GTEx (except for pancreatic islets, where we used the RNA-seq data from [84]).

CTCF-Cohesin regions in GM12878 were obtained by selecting CTCF motifs that intersected conservative IDR GM12878 CTCF ChIP-seq peaks (ENCODE accessions ENCFF096AKZ, ENCFF710VEH, and ENCFF963PJY) and the merged GM1287 RAD21 optimal IDR peaks (ENCODE accessions ENCFF753RGL and ENCFF002CPK). CTCF regions without cohesin were obtained similarly as above, but removing CTCF motifs that intersected any of the GM12878 RAD21 ChIP-seq peaks. All operations were performed with bedtools (v. 2.26.0). The choice of optimal IDR peaks for RAD21 aimed to increase the number of RAD21 peaks are included in the CTCF-cohesin$^+$ regions, therefore increasing stringency of the comparisons. We performed a quantile-based downsampling approach to make the CTCF/cohesin$^+$ and CTCF/cohesin$^-$ regions comparable regarding ChIP-seq signal, ATAC-seq signal, and FIMO motif scores. This was done by selecting all CTCF motifs encompassing the CTCF/cohesin$^+$ and CTCF/cohesin$^-$ regions and, for each feature (ATAC-seq fragments, ChIP-seq signal, or motif scores), calculating quantiles (n=20). The, for every quantile, we counted the number of motifs belonging to the CTCF/cohesin$^+$ and CTCF/cohesin$^-$ regions and randomly downsampled the group with more motifs instances to have the same number of motifs as the other in that quantile. This ensured that both regions had the same number of motifs and comparable distributions of ATAC and ChIP signals

and motif scores (as an example of this normalization, refer to Figure 2.16A). Pseudocode:

```
for feature in ATAC, ChIP, PWM:
    split feature in 20 quantiles
    for quantile in 1..20:
        set1 = CTCF/cohesin+ ∈ feature_quantile
        set2 = CTCF/cohesin- ∈ feature_quantile
        smallest_set = smallest(set1, set2)
        largest_set = largest(set1, set2)
        n = size(smallest_set)
        randomly select n items from largest_set
```

For the main figures, we used CTCF and RAD21 experiments ENCFF963PJY and ENCFF753CPK, respectively (the same comparisons using the other CTCF/RAD21 datasets are presented in Figure 2.17). The quantity labeled as relative chromatin information enrichment in Figure 2B corresponds to the sum of positive chromatin information enrichment (above dashed line) in each V-plot, divided by the CTCF/cohesin value for normalization.

Clustering. Chromatin information enrichment Z-score clusters were obtained using the R k-means implementation, using parameters k=3 and 1,000 random starts (Figure 2.18E). Cross-tissue clustering and dendrograms were calculated using the euclidean distances of the pairwise Spearman correlation of f-VICEs across samples. Normalized f-VICE values were converted to motif-wise Z scores before clustering.

MNase-seq data processing. Paired-end MNase unmapped reads from the lymphoblastoid cell line GM19238 were obtained from SRA, under accession SRR452483 [59]. Reads were mapped to the hg19 reference using BWA and processed in an identical fashion to the ATAC-seq data, with an additional step to retain only sequenced fragments of length 147±2 bp, therefore enriching for mononucleosomal fragments. The MNase aggregate signal plots were generated using ngsplot (`github.com/`

`shenlab-sinai/ngsplot`). For each motif plot, we used for input the BED files corresponding to the regions that were used to generate the corresponding V-plot. Motif MNase Z-scores for the clustering analyses were calculated using the MNase reads per million mapped reads (RPM) signal tracks outputted by ngsplot and the equation 2.3. MNase/V-plot correlations were calculated using positions $\leqslant$ 150 bp from the motif center.

$$Z(x) = \frac{x - mean(X)}{sd(X)} \tag{2.3}$$

V-plot asymmetry. V-plot asymmetry was calculated as the $\log_2$ ratio between the positive information content enrichment in the left and right of the motif center. To estimate significance, we used a permutation test where each fragment midpoint had a 50% chance of changing its direction relative to the motif while keeping the same distance (*i.e.* multiply its x-axis value by -1). We calculated the asymmetry of the permuted V-plots (n = 100,000) to generate a null distribution of asymmetry. Because the null was normally distributed based on Kolmogorov-Smirnof and Shapiro normality tests, we were able to estimate p values beyond the number of permutations by calculating the observed asymmetry Z-score relative to the null distribution. To calculate the nearest TSS directionality bias, we counted the number of active protein-coding TSS (GENCODE V19) (determined with the presence of LCL Cap analysis gene expression (CAGE) tag clusters, described in the next session) on either side of the motif and calculated the $\log_2$ ratio of the two. For the proximal and distal motif V-plots, we restricted our analyses to motifs occurring $\leqslant$ 1kb or $\geqslant$ 10kb from the nearest CAGE-supported TSS of any type (e.g. lincRNAs, pseudogenes; GENCODE V19). Enrichments of the plots in Figure 2F were calculated by randomly permuting the signal of the points in the x- and y-axis (n=10,000 permutations).

CAGE tag cluster identification. We downloaded CAGE data (fastq files) for 154 LCL samples [127] and mapped to hg19 using STAR (version 2.5.4b; default parameters) [128] and pruned the mapped reads to high quality reads (using samtools view v. 1.3.1; options -F 4 -q 255). We used the paralu method [129] to identify clusters of CAGE start sites (CAGE tag clusters). We called TCs in each individual sample using raw tag counts, requiring at least 2 tags at each included start site and allowing single base-pair tag clusters ('singletons') if supported by >2 tags. We then merged the tag clusters on each strand across samples. For each resulting segment, we calculated the number of LCL samples in which TCs overlapped the segment. We included the segment in the consensus TCs set if it was supported by independent TCs in at least 10 individual LCL samples, resulting in n=10> tag clusters. We then filtered out regions blacklisted by the ENCODE consortium due to poor mappability using bedtools (v. 2.26.0) to obtain the final set of LCL tag cluster regions.

DNA binding domain enrichments. DNA binding domains (DBD) enrichments were performed using a f-VICE rank sum permutation test. We assigned DBDs to the non-redundant motifs that mapped between our database and the one reported in [46], which has manually curated DBD-motif assignments. In order to map motifs between databases, we used tomtom [130] and selected motif matches with p-value < 0.05 after a conservative Bonferroni adjustment using all comparisons as denominator (i.e. number of motifs in our database times the number of motifs in the queried database), which yielded high-confidence DBD assignments for 402 of 540 motifs. We used the f-VICE rank from each motif to calculate the f-VICE rank sum the DBD and compared the observed value to a null distribution of 100,000 rank sums obtained from randomly permuting gene labels. This approach ensures that all the DBD retain their sizes during each permutation. We retained DBDs with at least 5 motifs and calculated the f-VICE enrichments for each DBD using the $\log_2$ of observed f-VICE rank sum

divided by the median of the null. FDR was calculated separately per sample, using the empirical p-value from the 100,000 permutations. We simultaneously performed a similar analysis using InterPro protein domains (v. 72) [131] (Figure 2.22). In order to assign domains to motifs, we first mapped our motifs to CIS-BP database (Build 1.02) [132], which has high-confidence motif-gene assignments, and retained genes that mapped to a single motif using the same approach described above. Each gene was then linked to a motif f-VICE score ($n = 475$) and we only retained domains with at least 5 genes after motif-gene mapping. Permutation and enrichments were calculated identically as described above.

***cis*-eQTL enrichments.** Feature enrichments in eQTLs were calculated using GREGOR [133] and QTL tools fenrich [134]. We used the lymphoblastoid cell line (LCL) eQTLs sets from Geuvadis [32] and GTEX [34] (FDR<5%). GREGOR background estimations were performed using SNPs with LD 0.99 for eQTL, with a maximum distance of 1 Mb from the variants of interest. Variants used as input for GREGOR were pruned to have maximum linkage disequilibrium r^2 of 0.8 with any other variant. For fenrich, we used the most significant SNP per gene as input.

ATAC-seq allelic imbalance analyses. To determine SNP allelic bias in ATAC-seq data, we used the publicly available data from Buenrostro *et al*, the Parker lab GM12878 sample discussed here, or the ABCU196 islet sample introduced in [84]. For GM12878 data, adapters were trimmed using cta (v. 0.1.2), and reads mapped to hg19 using bwa mem (default options except for the -M flag). Bam files were filtered to high-quality autosomal read pairs using samtools view (-f 3 -F 4 -F 8 -F 256 -F 2048 -q 30; v. 1.3.1). WASP (v. 0.2.1, commit 5a52185; python version 2.7) [135] was used to diminish reference bias; for remapping the reads as part of the WASP pipeline, we used the same mapping and filtering parameters described above for the initial mapping and filtering. Duplicates were

removed using WASP's rmdup_pe.py script. We used the phased GM12878 VCF file downloaded from `ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv3.3.1/GRCh37/HG001_GRCh37_GIAB_highconf_CG-IllFB-IllGATKHC-Ion-10X-SOLID_CHROM1-X_v.3.3.1_highconf_phased.vcf.gz` . To avoid double-counting alleles, overlapping read pairs were clipped using bamUtil clipOverlap (v. 1.0.14; `genome.sph.umich.edu/wiki/BamUtil:_clipOverlap`). For the Buenrostro *et al* data, the bam files from the samples were then merged to create a single GM12878 bam file using samtools merge (v. 1.3.1). For each heterozygous autosomal SNP, we then counted the number of reads containing each allele, using only bases with base quality of at least 20. We used a two-tailed binomial test that accounted for reference allele bias to evaluate the significance of the allelic bias at each SNP (as described in [84]; when calculating the expected fracRef, SNPs in the top 25th percentile of read coverage were downsampled to the 50th percentile coverage and SNPs with coverage less than 10 were excluded). When performing the binomial test we downsampled the coverage at each SNP such that each SNP had coverage = 20 (to reduce coverage-related biases). The islet ATAC-seq data was processed and tested as described in [33], except that we also downsampled coverage at each SNP to 20 when performing the binomial test.. We did not test SNPs in regions blacklisted by the ENCODE Consortium because of poor mappability (wgEncodeDacMapabilityConsensusExcludable.bed and wgEncodeDukeMapabilityRegionsExcludable.bed). We retained for downstream analyses all loci with nominally significant binomial p values ($p < 0.05$) and at least 2 reads (10%) mapped to any allele.

6-mer f-VICE calculations. We generated a list of all possible DNA 6-mers and scanned the reference human genome (hg19) to obtain the coordinates for all their corresponding matches. Similarly to motifs, we only retained 6-mer matches that were in mappable regions and did not intersect blacklisted regions. We used BMO to

determine the subset of 6-mer instances that was predicted bound. We calculated the normalized f-VICE for each 6-mer using the same approach as in the motifs, including using linear regression to control for chromatin accessibility and number of predicted bound 6-mer instances. This yielded a table with normalized f-VICEs for every 6-mer in each analyzed sample. Given that only a small fraction of 6-mers instances deviate from the reference sequence in any given sample, we reasoned that the effects from loci that did not match the reference genome would be diluted by all the other predicted bound 6-mer instances that matched the reference genome. Therefore, we did not perform haplotype-aware 6-mer scans. For each locus with significant allelic imbalance, we determined the six 6-mers formed by each allele and obtained their corresponding normalized f-VICEs values from the corresponding sample f-VICE table. Because of the sparsity of the ATAC-seq coverage in any individual loci with ATAC-seq allelic imbalance, it was not possible to directly calculate the chromatin information changes associated with the preferred and non-preferred alleles.

We used GAT [136] to calculate enrichments across cell-type specific ChromHMM states (obtained from [84]). We divided the 6-mers into f-VICE deciles and determined genomic regions populated exclusively by predicted bound 6-mers belonging to each decile (*i.e.* no overlap between deciles; bedtools subtract v. 2.26.0). These regions were used as input for GAT, and the workspace was the union of all regions assigned to any decile (using bedtools merge).

Data availability. Code and scripts used for the analyses performed in this study are publicly available at `http://github.com/ParkerLab/chromatin_information`. BMO and atactk are publically available at `http://github.com/ParkerLab/BMO` and `http://github.com/ParkerLab/atactk`.

2.8 Acknowledgements

I'd like to thank several people who contributed to this project. I thank Dr. Yasuhiro Kyono and Dr. Jacob Kitzman for generating the GM12878 ATAC-seq data. I am grateful to John Hensley for all his valuable and timely help with software development, and I thank Arushi Varshney and Peter Orchard for helping with the eQTL, GREGOR, and allelic imbalance analyses. I thank Prof. Steve Parker conceptualizing and organizing this project and for his insight and support throughout this work. I specifically contributed towards the computational analyses and manuscript preparation.

Table 2.1: FRAP recovery times from literature

Factor	**Organism**	**Motif**	**FRAP recovery (s)**	**Reference**
AHR	*H. sapiens*	AHR_1	38	[137]
AP1	*H. sapiens*	MA0476.1	600	[70]
ARNT	*H. sapiens*	ARNT_2	41	[137]
CEBP	*H. sapiens*	CEBPB_known5	32	[137]
CREB	*H. sapiens*	CREB3_1	100	[138]
CTCF	*H. sapiens*	CTCF_known2	660	[139]
FOXA1	*M. musculus*	FOXA_known4	300	[140]
MYC	*H. sapiens*	MYC_known13	37	[137]
NFKB	*H. sapiens*	NFKB_known5	30	[141]
NR3C1	*C. aethiops*	NR3C1_known18	30	[142]
NR3C2	*H. sapiens*	NR3C2_1	30	[143]
TP53	*H. sapiens*	TP53_4	20	[144]
XBP	*H. sapiens*	XBP1_2	30	[137]

2.9 Appendix: Additional Figures

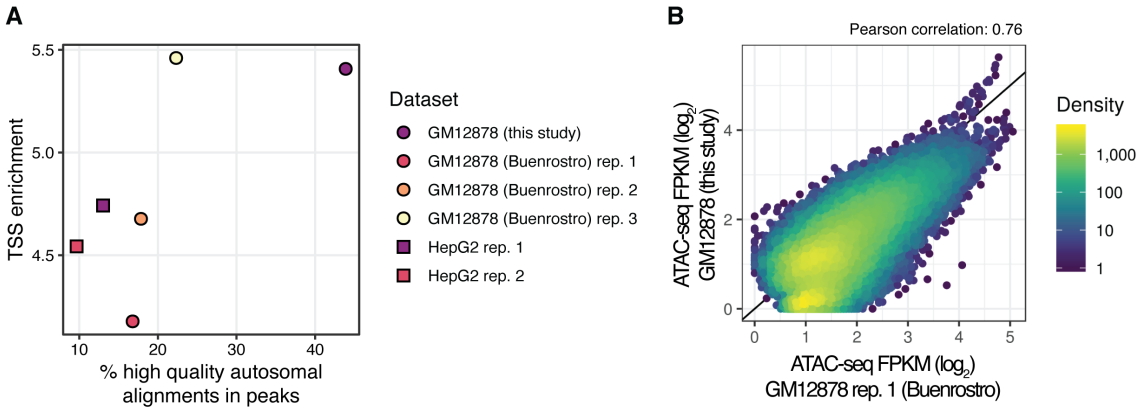


Figure 2.4: ATAC-seq datasets signal-to-noise comparisons. (a) Scatter plots of the percent high quality autosomal alignments (%HQAA) in ATAC-seq peaks distribution and TSS enrichments of GM12878 and HepG2 datasets, obtained using Ataqv (`github.com/ParkerLab/ataqv`). (b) Scatter plot of the ATAC-seq signal in the union of the MACS2 broad peaks called in the two GM12878 datasets. Each point corresponds to one ATAC-seq peak and the position correspond to its coverage (fragments per kilobase per million, FPKM) in each dataset. Solid diagonal line, identity (x=y).

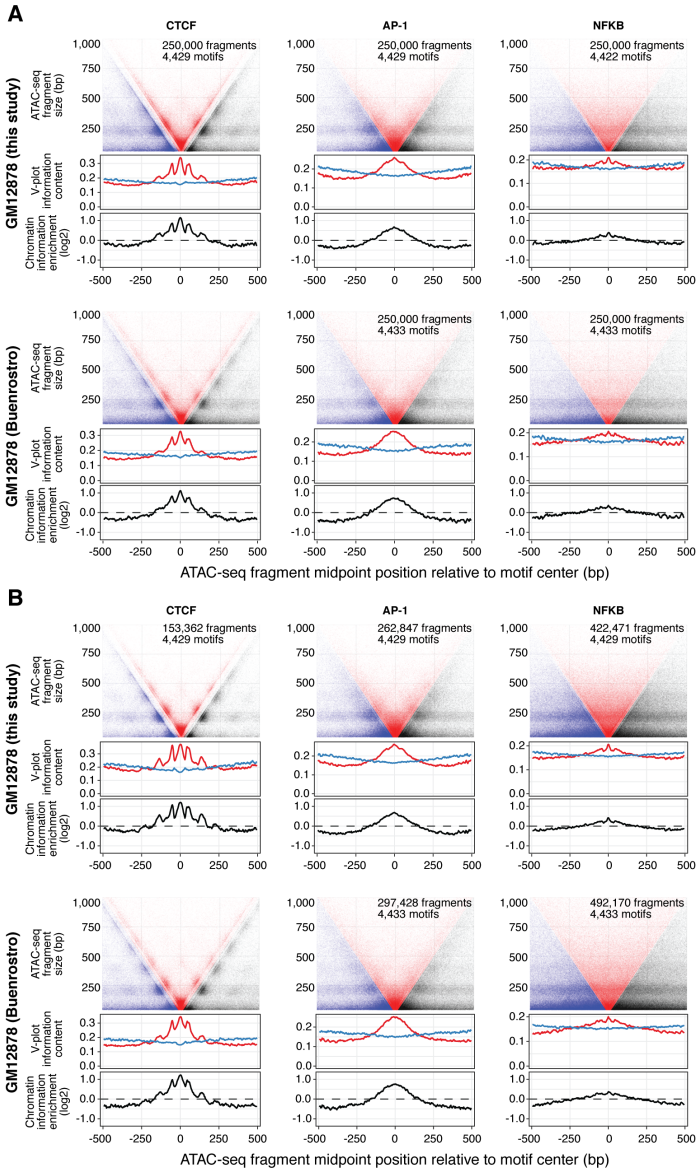


Figure 2.5: GM12878 V-plots. (a) V-plots for the same TFs as in Figure 1B across GM12878 datasets. Upper: ATAC-seq fragment distribution. Middle: observed (red) and expected (blue) information content tracks, used to calculate chromatin information enrichments (bottom). V-plots were downsampled to equal number of ATAC-seq fragments and motifs between TFs by selecting the top n motifs, ranked by number of ATAC-seq fragments, and then further downsampling to 250,000 fragments. n represents the smallest number of bound motifs among the plotted TFs per sample. (b) Similar to (A), but randomly downsampling to exactly n motifs (without ranking by signal or further downsampling the number of ATAC-seq fragments). This was performed to demonstrate that the differences in chromatin architecture are intrinsic to the TF and evident regardless of downsampling method.

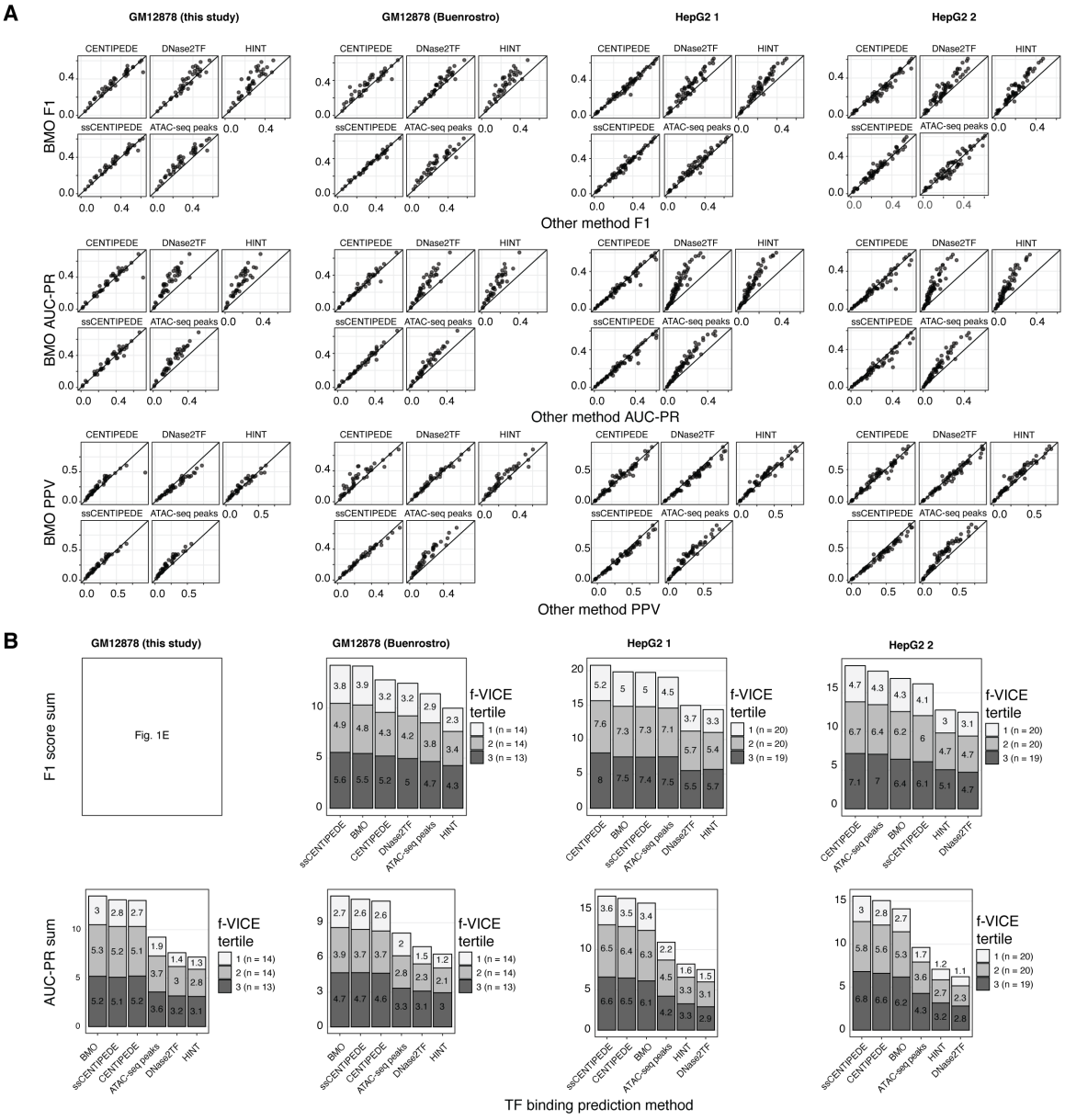


Figure 2.6: TF binding prediction methods comparisons across datasets. (a) F1, positive predictive value (PPV), and AUC-PR scatter plots of BMO versus other TF binding prediction methods across multiple ATAC-seq datasets. Each point corresponds to a TF with ChIP-seq data. (b) Total F1-score and AUC-PR across datasets, separated into f-VICE tertiles. Solid diagonal line, identity (x=y).

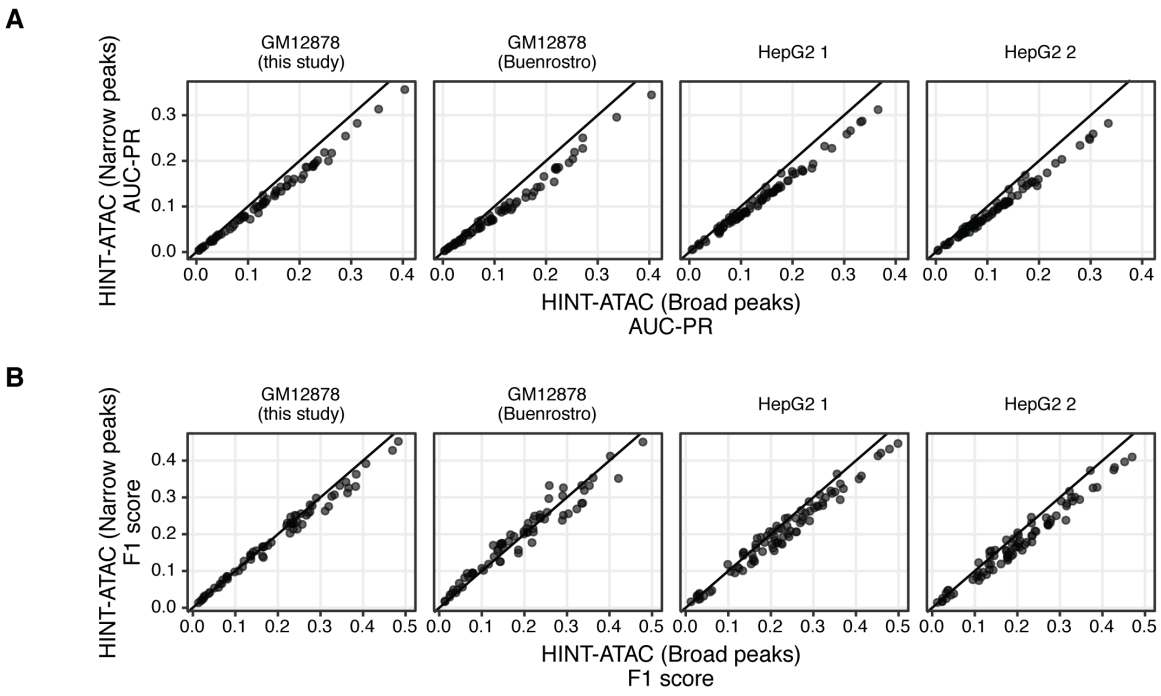


Figure 2.7: HINT-ATAC performance using narrow or broad peak calls. Scatter plots of AUC-PRs (a) and F1 scores (b) across datasets. Each point corresponds to a TF with ChIP-seq data. Solid diagonal line, identity (x=y).

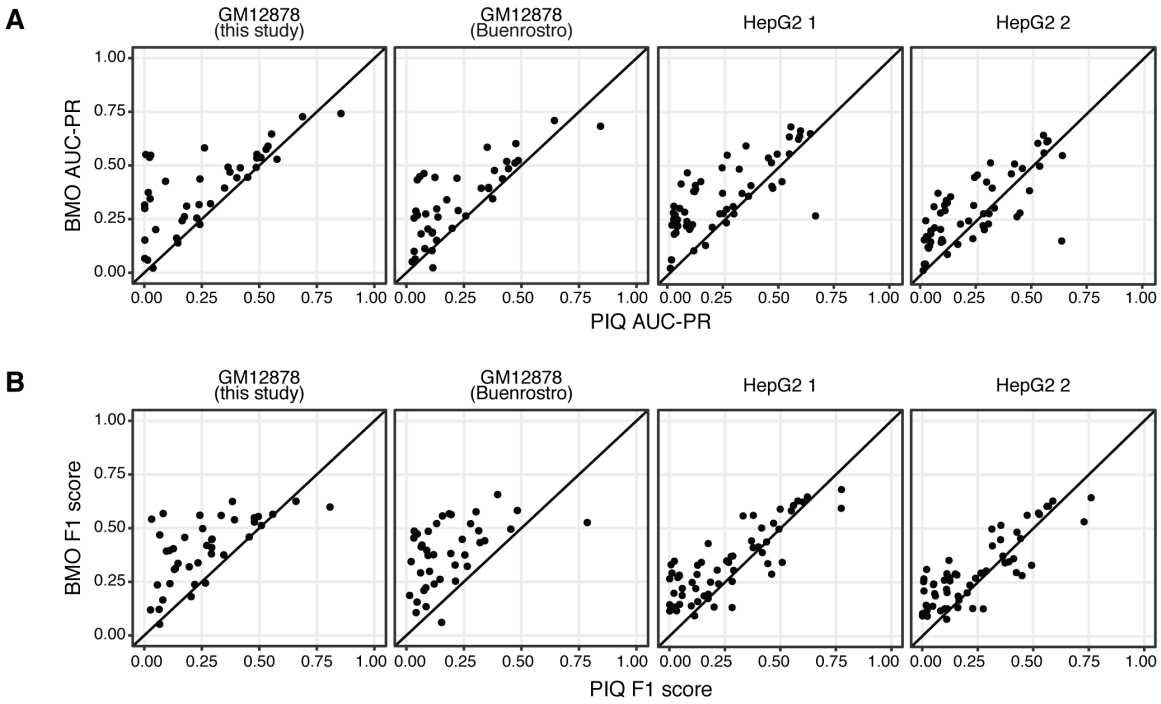


Figure 2.8: BMO and PIQ comparisons. Scatter plots of AUC-PR (a) and F1 scores (b) across datasets comparing BMO and PIQ. Each point corresponds to a TF with ChIP-seq data. Solid diagonal line, identity (x=y).

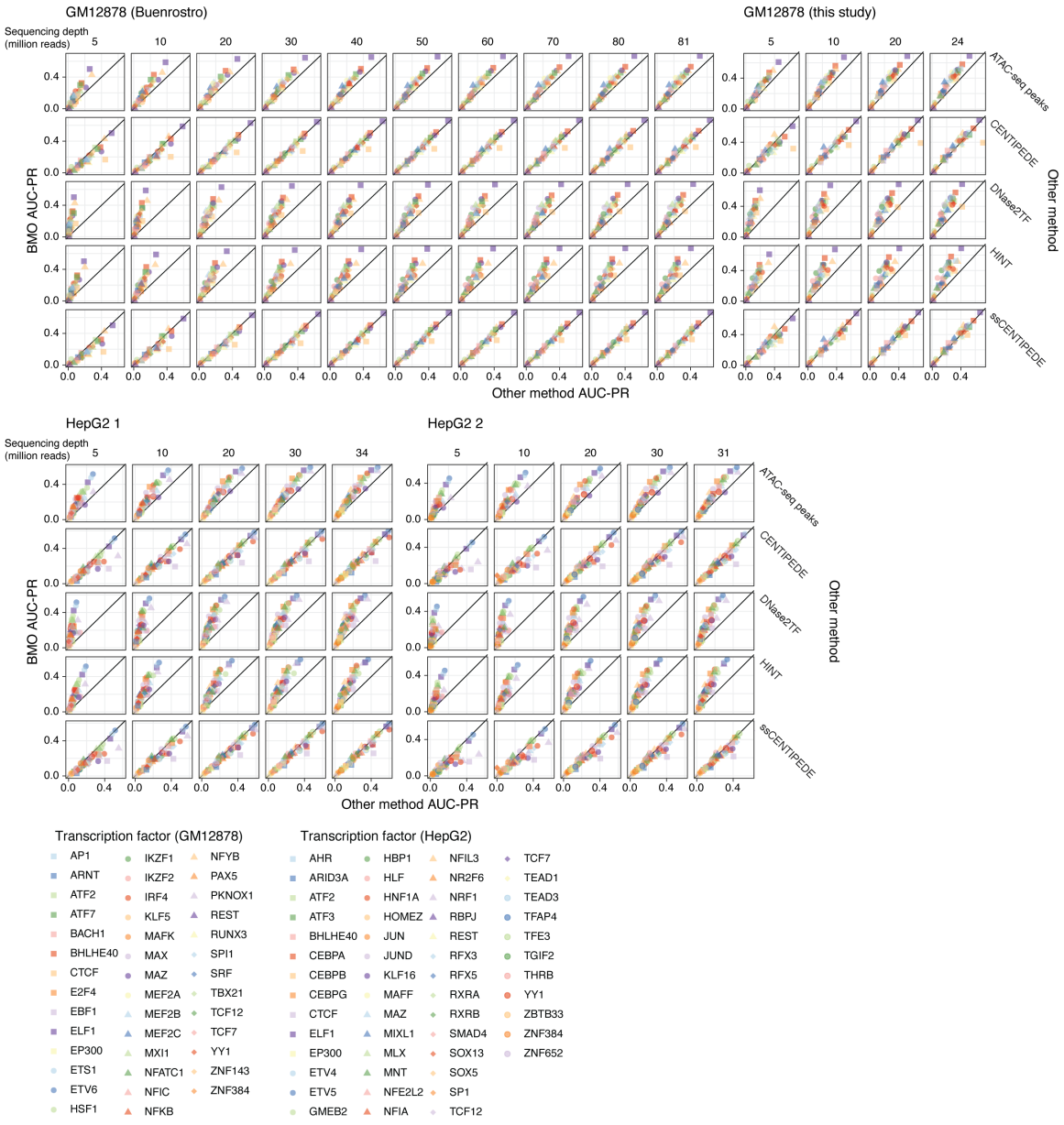


Figure 2.9: TF binding prediction methods comparisons across sequencing depths. AUC-PR scatter plots of BMO versus other methods across different sequencing depths, shown in millions of reads in the top of each facet column. Each point corresponds to one TF with ChIP-seq data. Solid diagonal lines, identity (x=y).

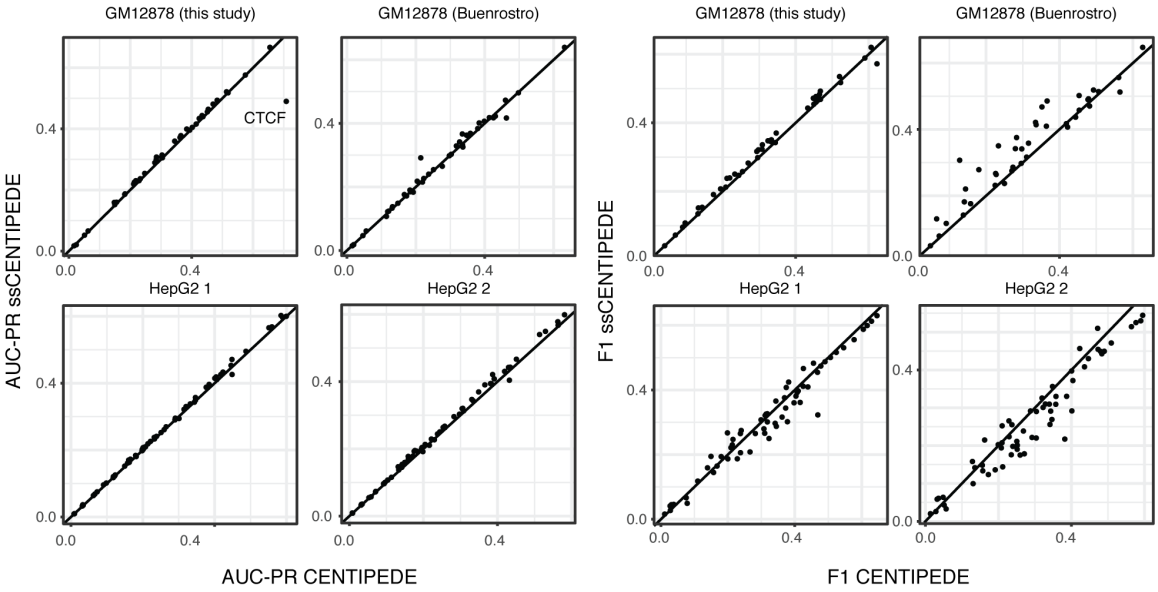


Figure 2.10: CENTIPEDE and ssCENTIPEDE perform similarly across datasets. Scatter plots of ssCENTIPEDE and CENTIPEDE AUC-PRs and F1 scores across multiple datasets. Each point corresponds to one TF with ChIP-seq data. Solid diagonal line, identity (x=y).

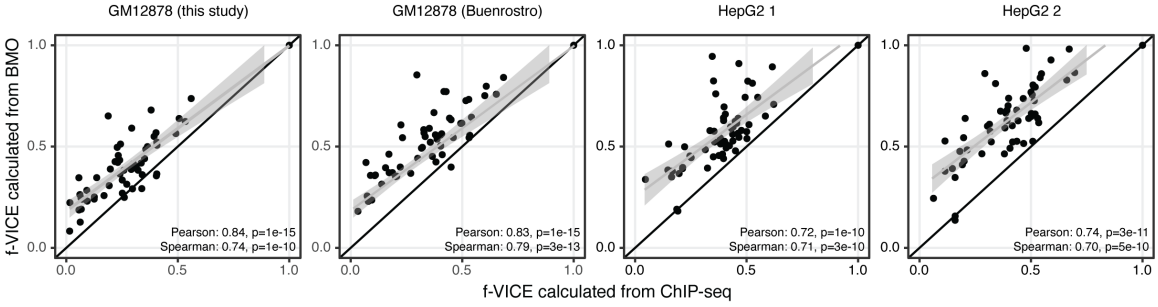
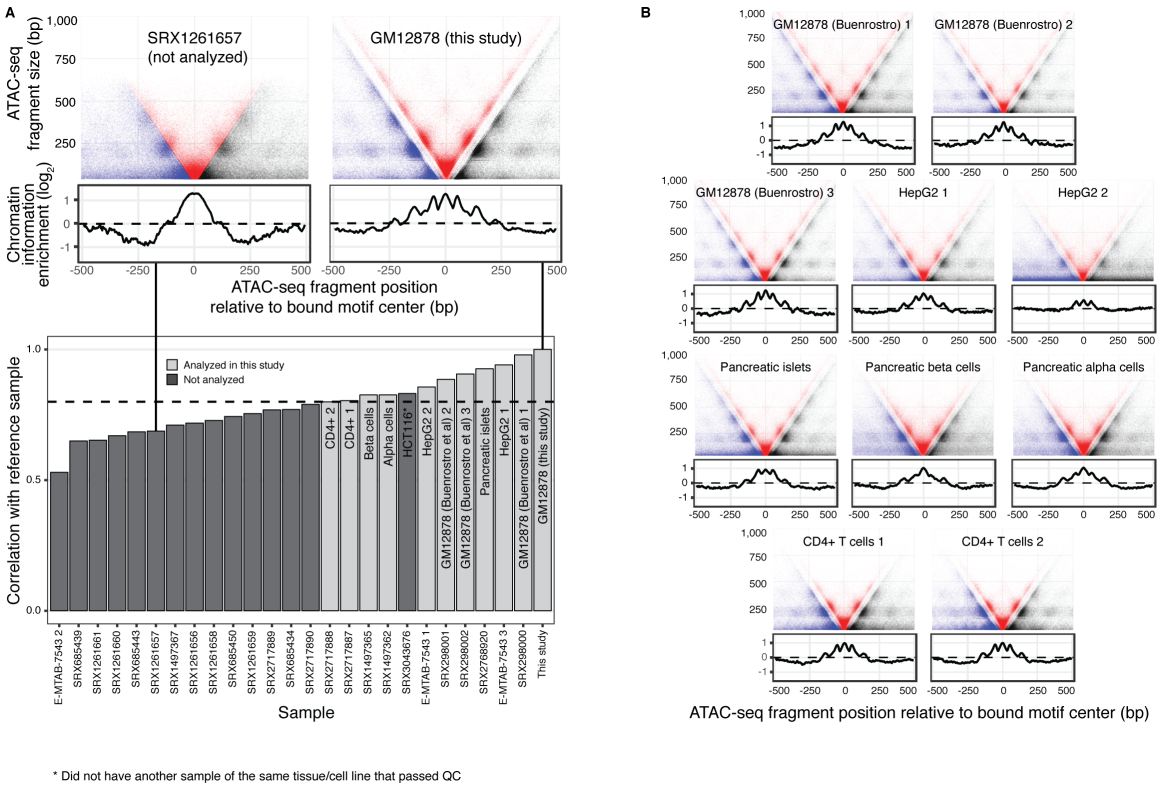


Figure 2.11: BMO and ChIP-seq f-VICEs are correlated. Correlation of f-VICEs calculated from BMO predictions and from the respective ChIP-seq data across ATAC-seq datasets. Note that BMO f-VICEs are consistently higher than ChIP-seq f-VICEs. This is due to a higher number of predicted bound motif instances in BMO, which motivated us to normalize f-VICEs using the linear regression approach described in the methods (f-VICEs are not normalized using regression in this figure owing to low n). Solid diagonal line, identity (x=y). Grey lines and shaded areas, linear model $y \sim x$ fit and 95% confidence intervals.



* Did not have another sample of the same tissue/cell line that passed QC

Figure 2.12: Selection of additional ATAC-seq samples using ubiquitous and conserved CTCF-cohesin binding sites. (a) Upper: examples of V-plots for the reference ubiquitous and conserved CTCF-cohesin binding sites indicating a high-quality and low-quality sample (the latter shown for exemplification purposes and not included in this study). Lower: chromatin information enrichment correlation between the CTCF-cohesin binding sites across multiple experiments to a reference sample. (b) V-plots of the same regions in the other samples selected for this study. Y-axes labels are the same as the upper plot in panel A.

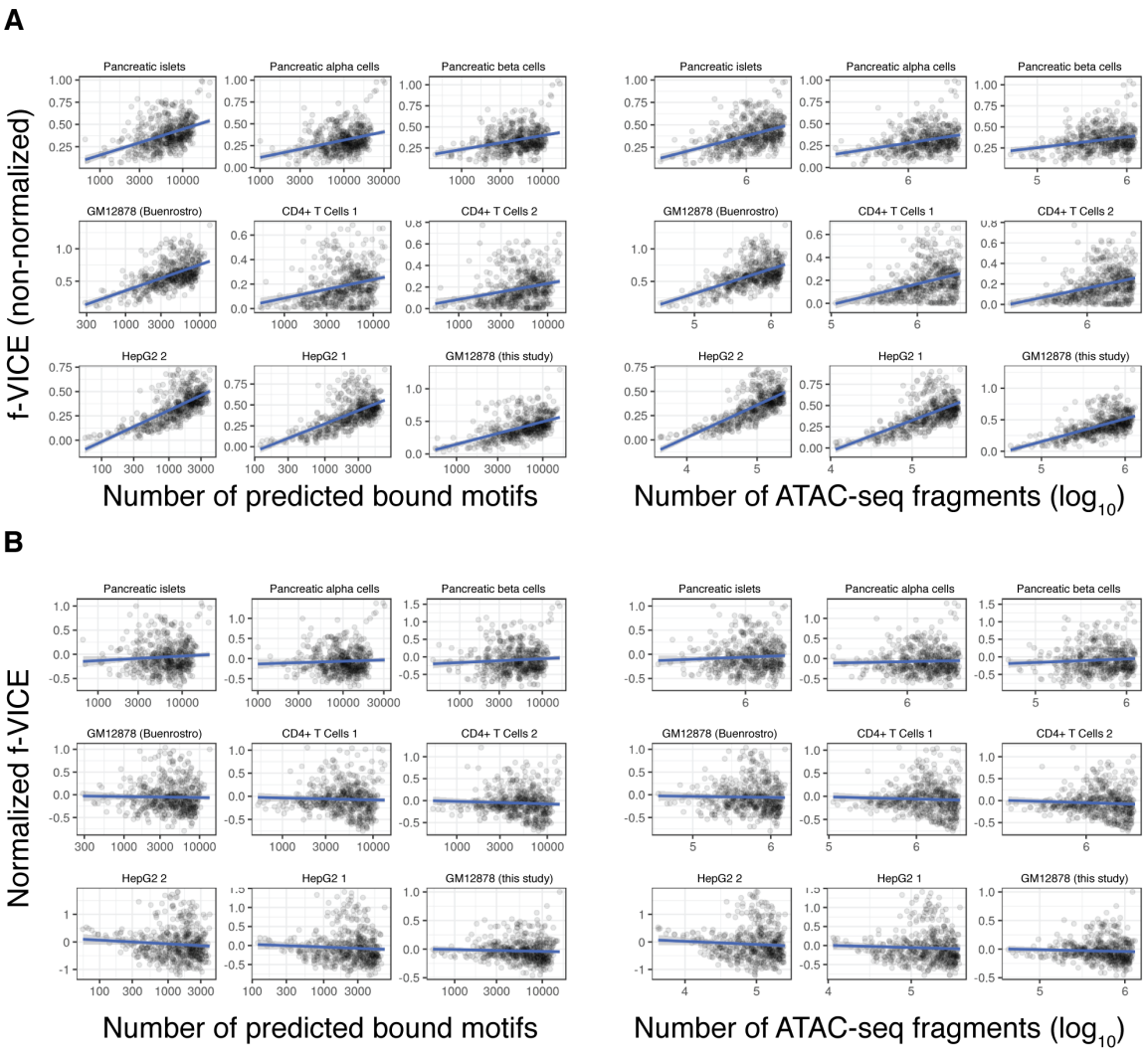


Figure 2.13: Normalization of f-VICE. (a) Scatter plots of f-VICE as a function of number of predicted bound motifs or ATAC-seq signal. (b) Same data after normalization using a linear regression model that accounts for both variables (described in the Methods section). Each point corresponds to a motif (n=540).

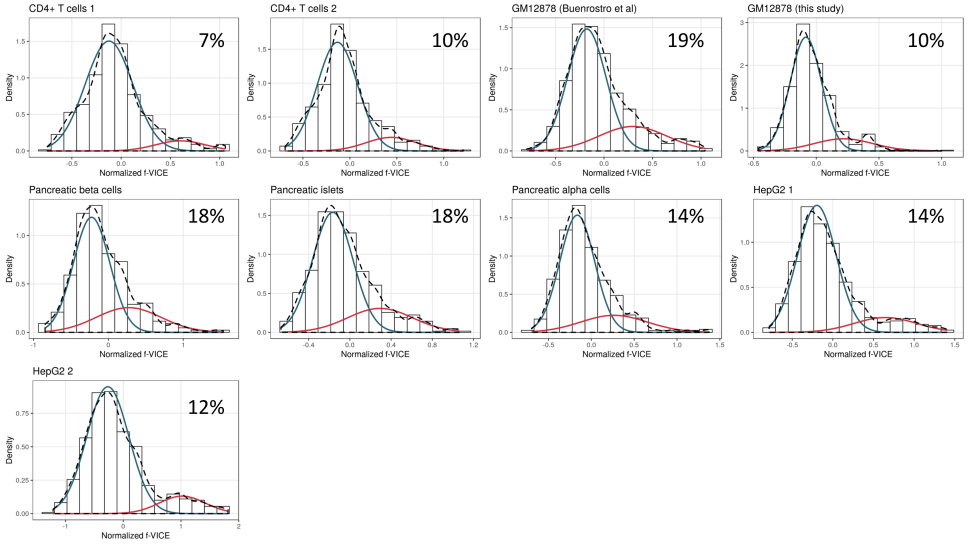


Figure 2.14: f-VICE distributions across samples. Histograms and density plots of the empirical (dashed) and high/low f-VICE distributions Gaussian fits (red and blue, respectively) across all the datasets surveyed in this work. Percentages in the upper right corner of plots represent the high f-VICE distribution.

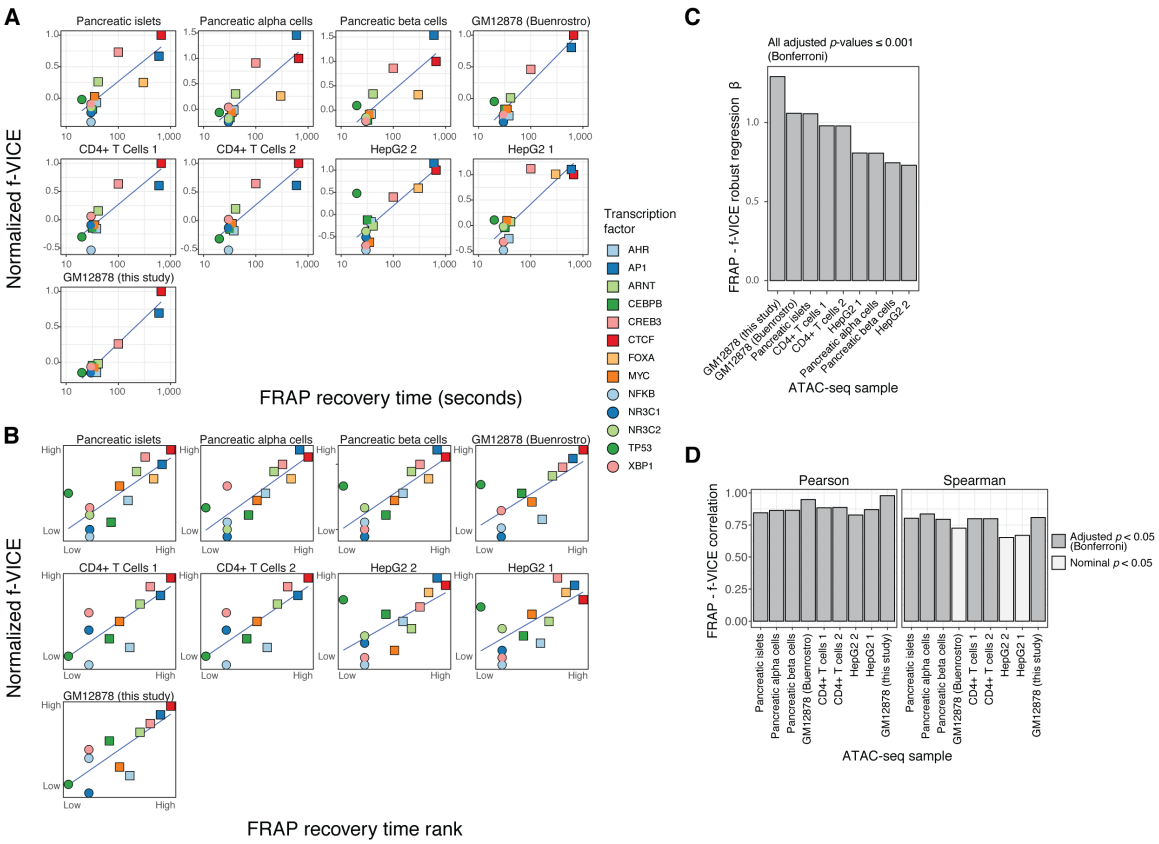


Figure 2.15: f-VICE correlation with FRAP recovery times in multiple datasets. (a) Scatter plots of mammalian FRAP recovery times and f-VICEs across multiple datasets. (b) Similar to (a), but showing f-VICE and FRAP ranks (similar to a Spearman correlation). Solid blue lines in (a) and (b), linear model $y \sim x$. (c) Robust linear regression betas for the plots shown in (a). Robust linear regression model f-VICE $\sim log_{10}$(FRAP) (d) Pearson and Spearman correlations of the plots shown in (a). All correlations were significant at p<0.05.

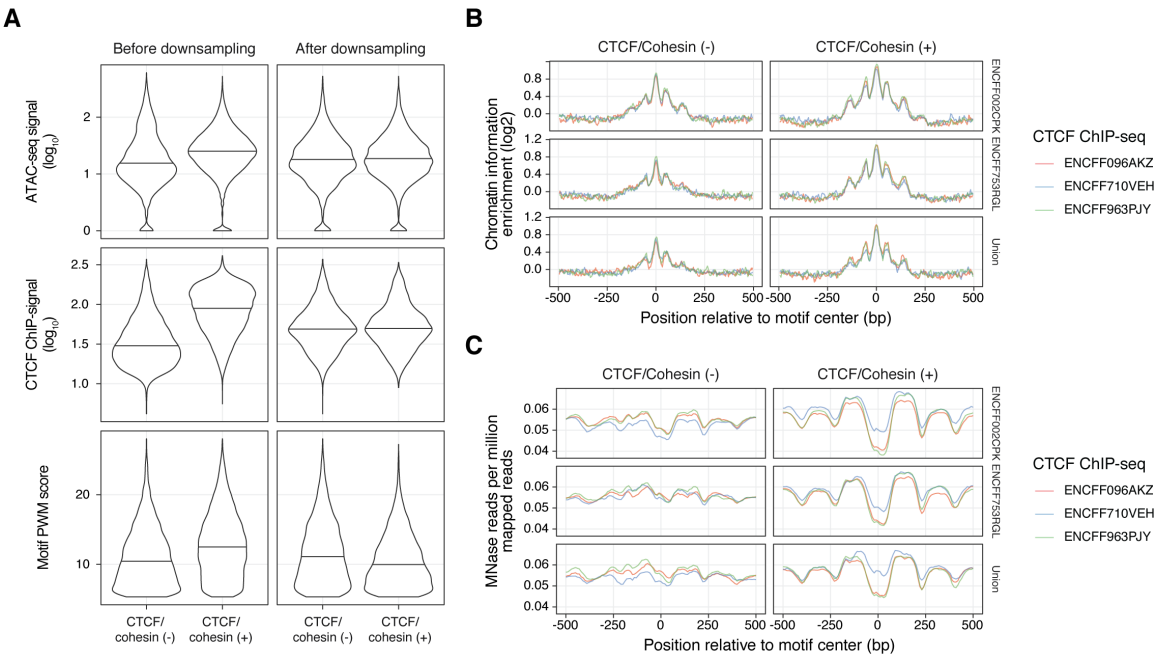


Figure 2.16: GM12878 CTCF/cohesin^{+} and CTCF/cohesin^{-} regions. (a) Example distributions of ATAC-seq signal, ChIP-seq signal, and motif PWM match score before and after quantile-based downsampling of the CTCF/cohesin^{+} and CTCF/cohesin^{-}. Datasets: ENCFF963PJY (CTCF) and ENCFF002CPK (Rad21). Other CTCF/Rad21 ChIP-seq dataset combinations not shown. Horizontal lines, median. (b) Chromatin information tracks of CTCF/cohesin^{+} and CTCF/cohesin^{-} using genomic regions obtained from different GM12878 CTCF and RAD21 ChIP-seq datasets combinations. The facet labeled "Union" corresponds to the union of the two GM12878 RAD21 ChIP-seq datasets. (c) Corresponding MNase signal at the regions shown in (b).

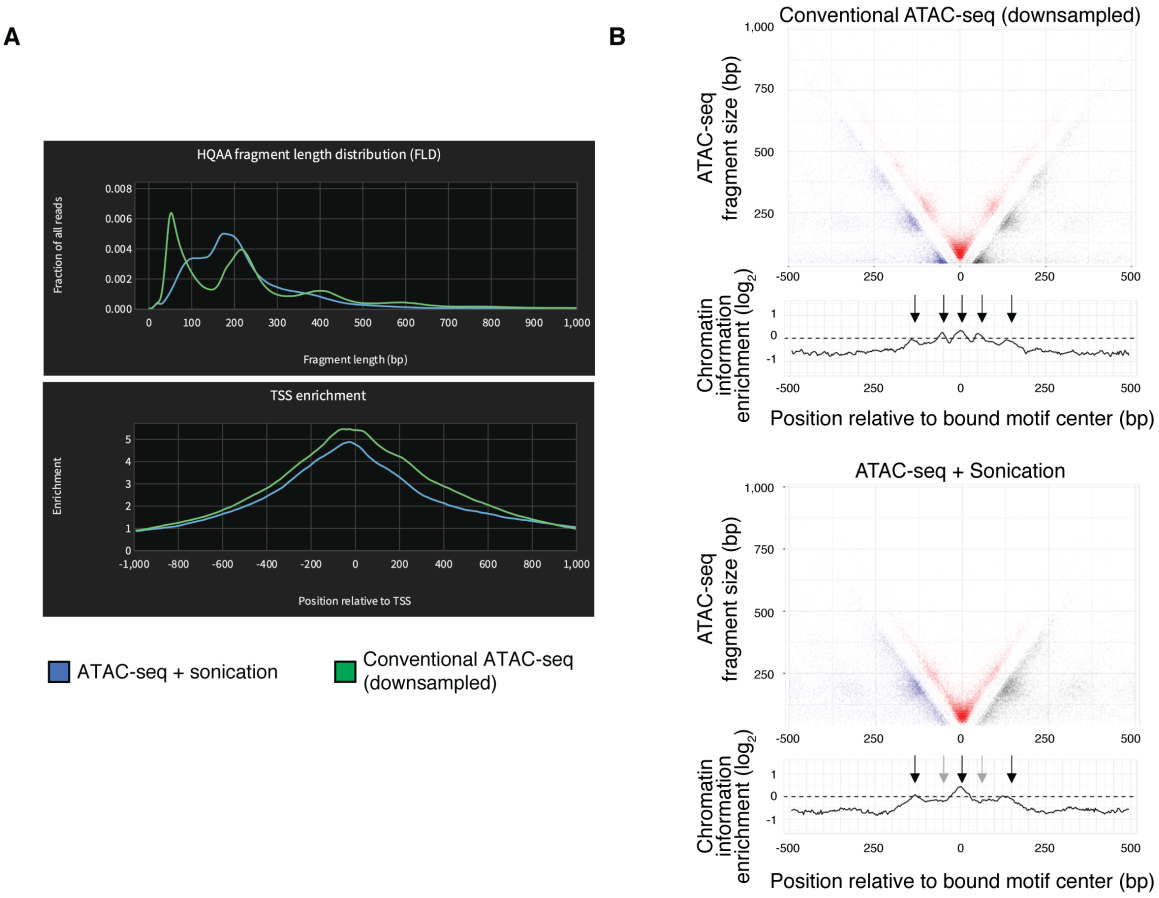


Figure 2.17: Sonicated GM12878 ATAC-seq data. (a) Ataqv (`github.com/ParkerLab/ataqv`) screenshot showing the fragment size distribution and TSS enrichments of the conventional and sonicated GM12878 ATAC-seq datasets generated in this study. HQAA = high-quality autosomal alignments. (b) V-plots of the reference conserved CTCF-cohesin regions in the two datasets. "Conventional ATAC-seq" refers to the sample labeled as "GM12878 (this study)" in other figures. However, this dataset was downsampled to the same depth as the sonicated dataset (3.45 million reads) for the analyses presented in this figure and Figure 2a in order to make datasets directly comparable. Black arrows, CIE peaks in both samples. Gray arrows, CIE peaks not in the sonicated sample.

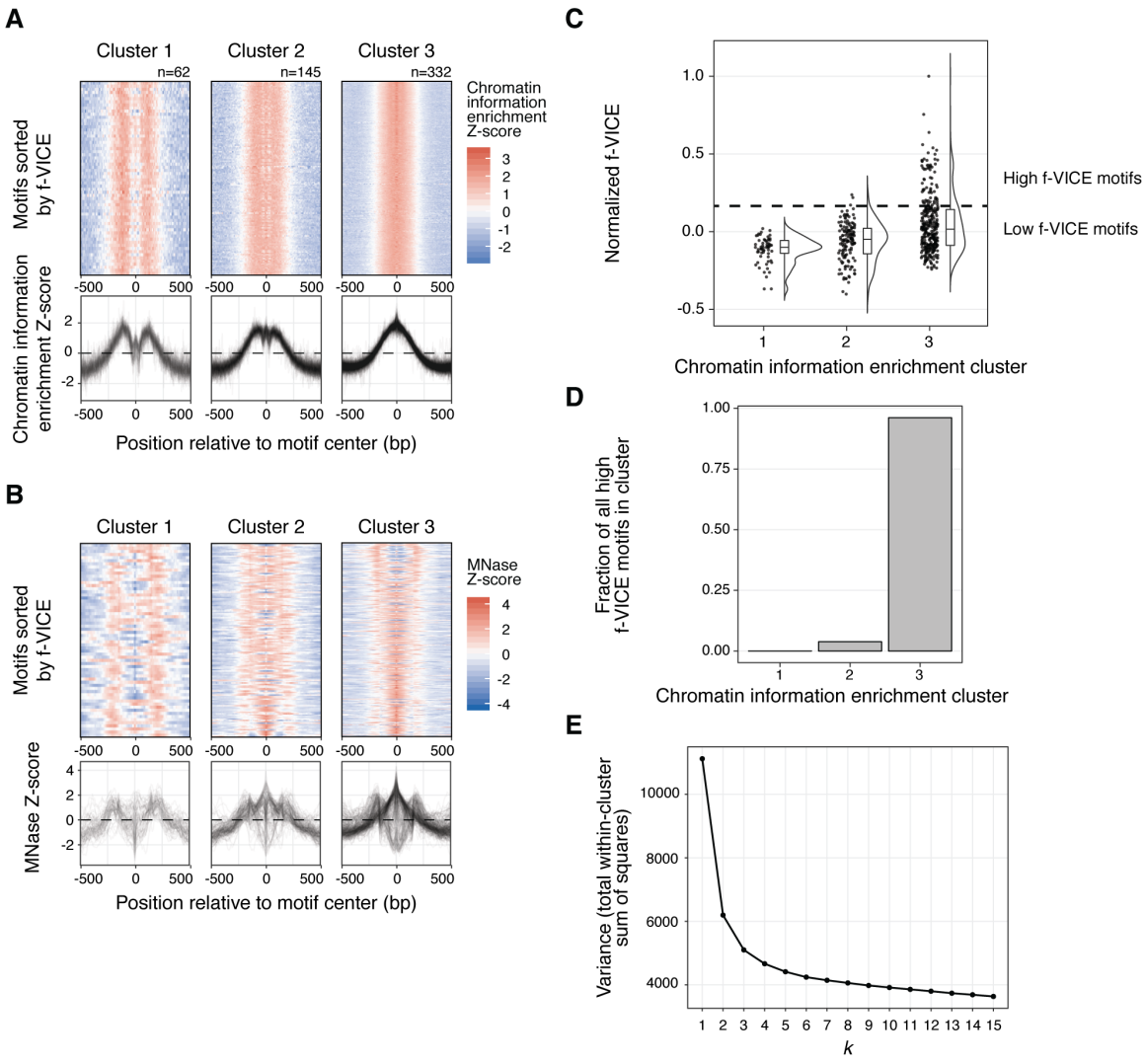


Figure 2.18: Chromatin information clusters in GM12878. (a) Chromatin information enrichment Z-score clusters and (b) their corresponding MNase Z-scores. All clusters are sorted by f-VICE on the y-axis. (c) Distribution of f-VICEs across chromatin information enrichment clusters. (d) Fraction of total high f-VICE TFs per chromatin information cluster (based on the mixture model distributions). (e) Elbow plot showing within-cluster variance for different k values in the chromatin information enrichment k-means clustering in panel (a).

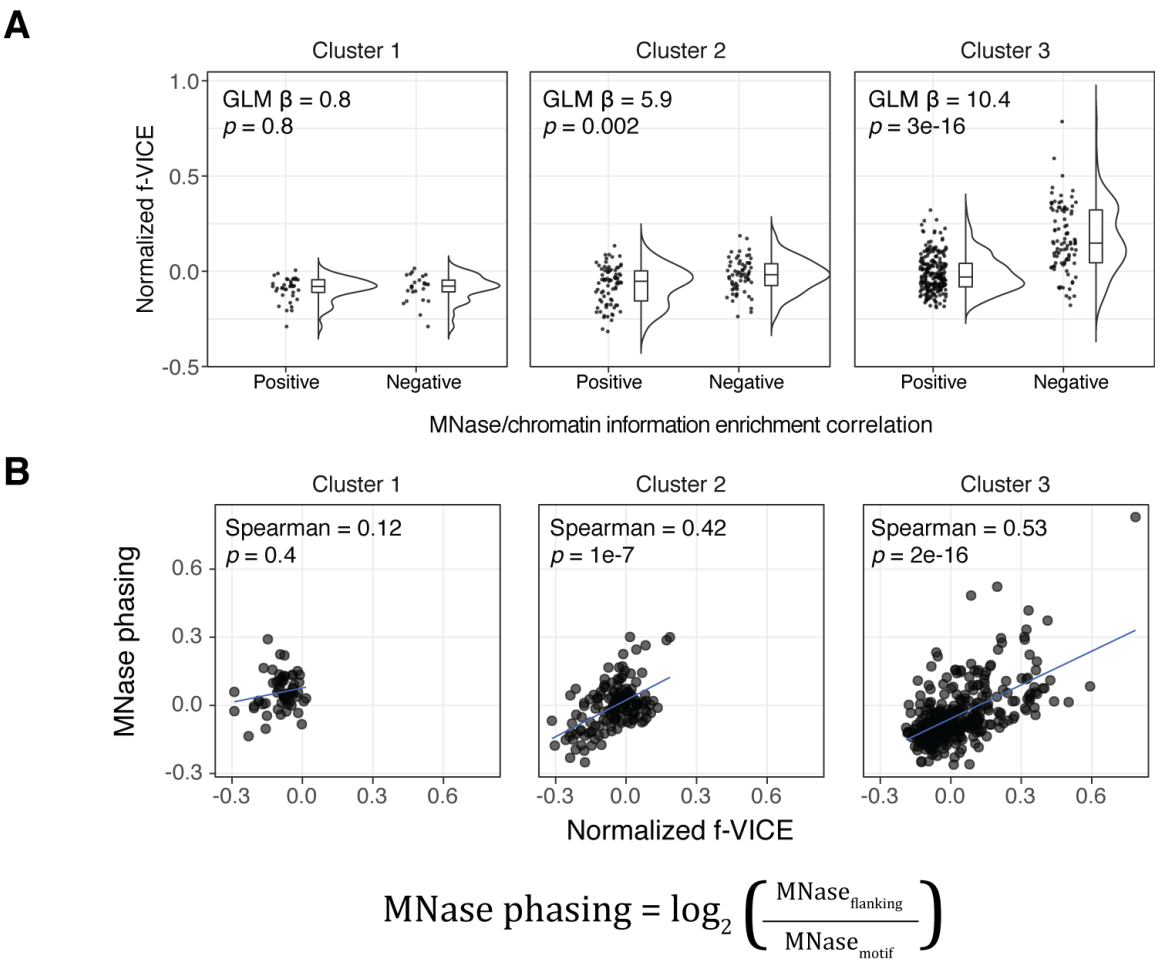


Figure 2.19: f-VICE correlation with nucleosome phasing. In this plot, we show two independent approaches for correlating f-VICE to nucleosome phasing. (a) f-VICE distributions for motifs with positive and negative correlation between chromatin information and MNase Z-scores ($\leq$150 bp from motif center) across the different k-means clusters, labeled 1-3 in the header. GLM = generalized linear model. (b) Correlation between f-VICE and the $\log_2$ ratio of the MNase signal at the motif vicinity ($\pm$125-150 bp from motif center) divided by the MNase signal at the motif ($\pm$25 bp from motif center). Positives ratios indicate nucleosome phasing.

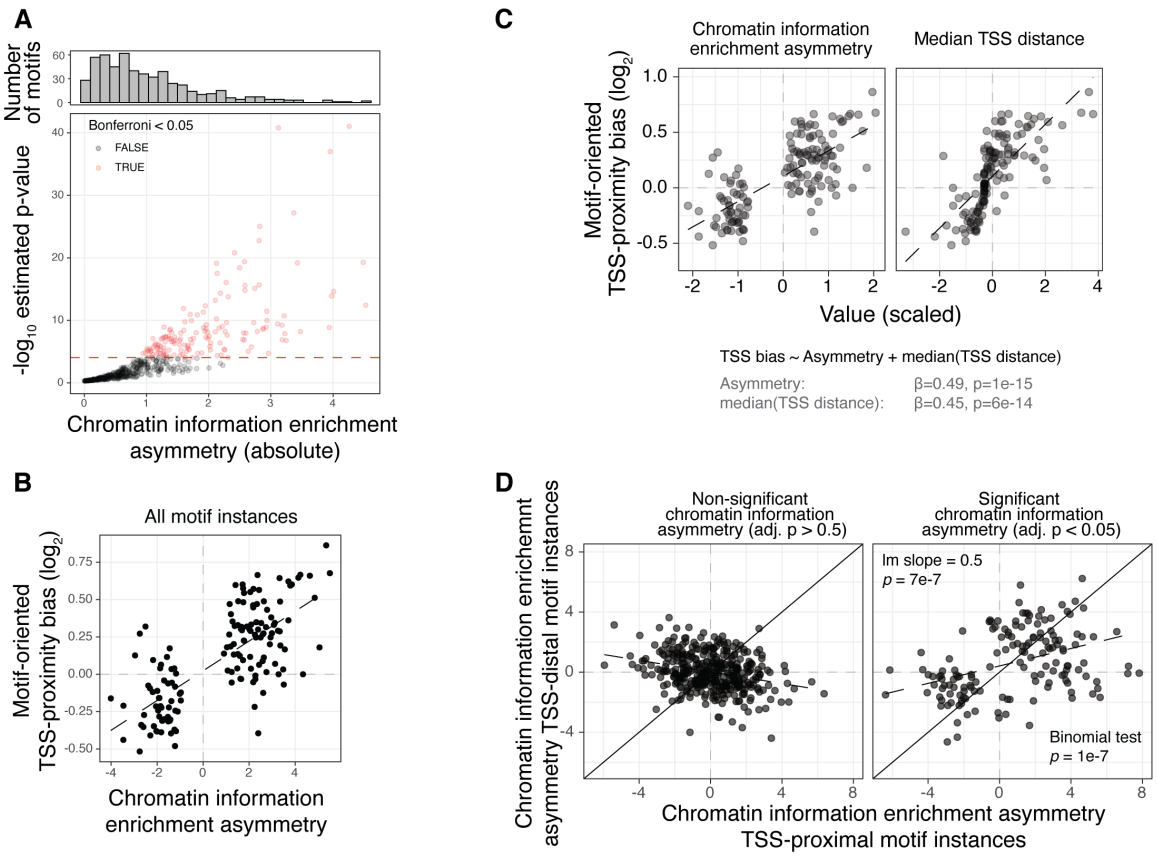


Figure 2.20: Motifs with information asymmetry in GM12878. (a) Chromatin information asymmetry distribution in GM12878. Red dashed line represent the Bonferroni p-value cutoff threshold. (b) Relationship between chromatin information asymmetry and motif-oriented TSS proximity bias based on all motif instances. (c) Motif-oriented TSS proximity bias chromatin as a function of information asymmetry and median nearest TSS distance. We performed a regression analysis of nearest TSS direction bias and chromatin information enrichment asymmetry, controlling for TSS distance (Methods). Chromatin information enrichment asymmetry remained significant when controlling for TSS distance. (d) Concordance of chromatin information asymmetry direction between TSS-distal and TSS-proximal motif instances. Solid diagonal line, identity (x=y). Dashed black lines, linear model (lm) fit in the data.

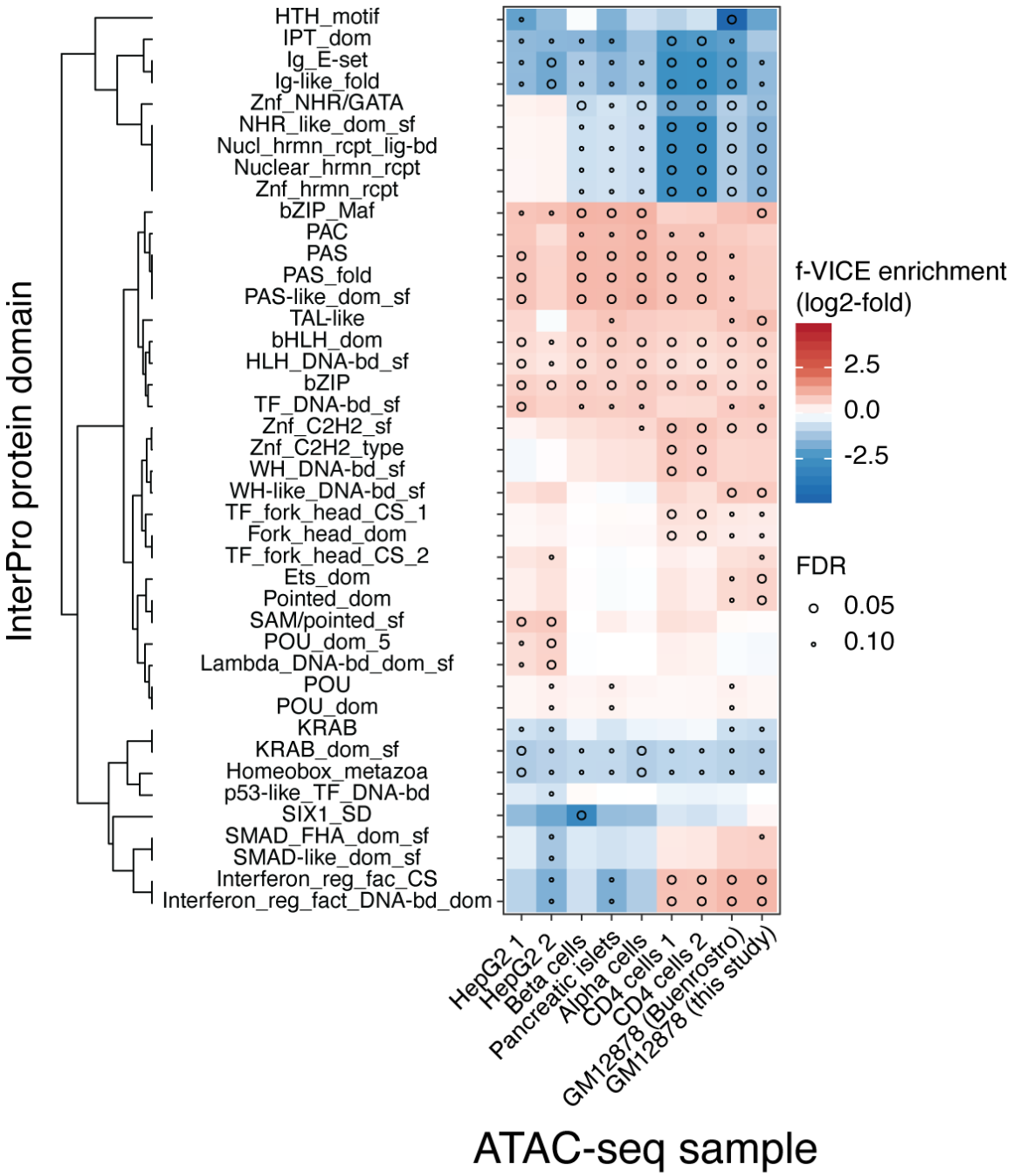


Figure 2.21: Protein domain enrichments. InterPro protein domains f-VICE enrichments across samples.

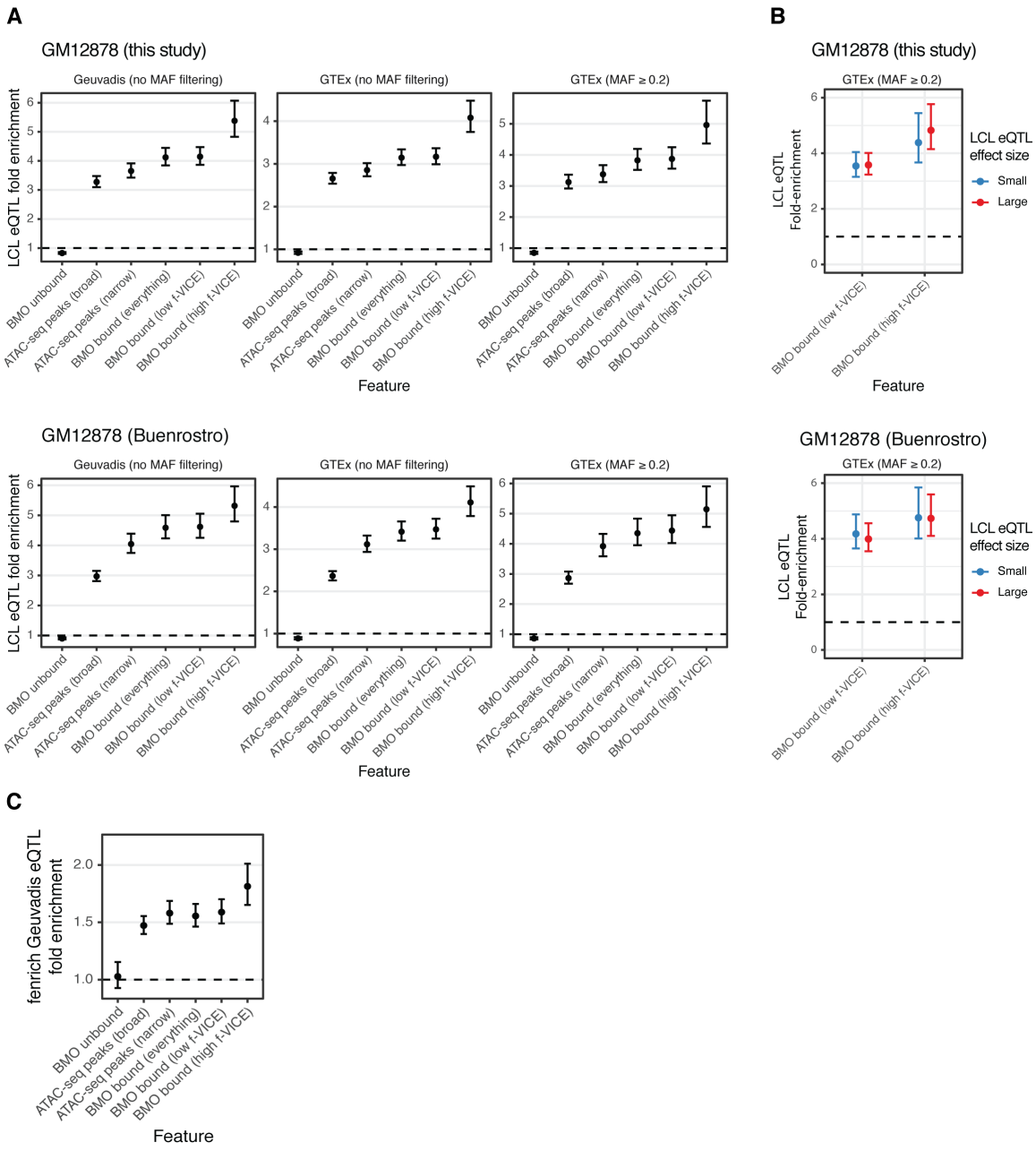


Figure 2.22: Enrichment of high and low f-VICE motifs in *cis*-eQTLs. (a) eQTL enrichments of different features across GM12878 ATAC-seq and lymphoblastoid cell lines (LCL) eQTL datasets. Enrichments are shown for GTEx with and without minor allele frequency (MAF) filtering to demonstrate that the observed results are not due to disproportionate representation of low MAF variants in any feature. (b) Enrichments of high and low f-VICE BMO predictions on high and low effect size GTEx eQTLs (above and below the median, respectively) across the two GM12878 datasets. (c) Geuvadis LCL eQTL enrichment calculated using QTL tools fenrich in our GM12878 dataset. Error bars in all plots represent the standard deviation of the effect size.

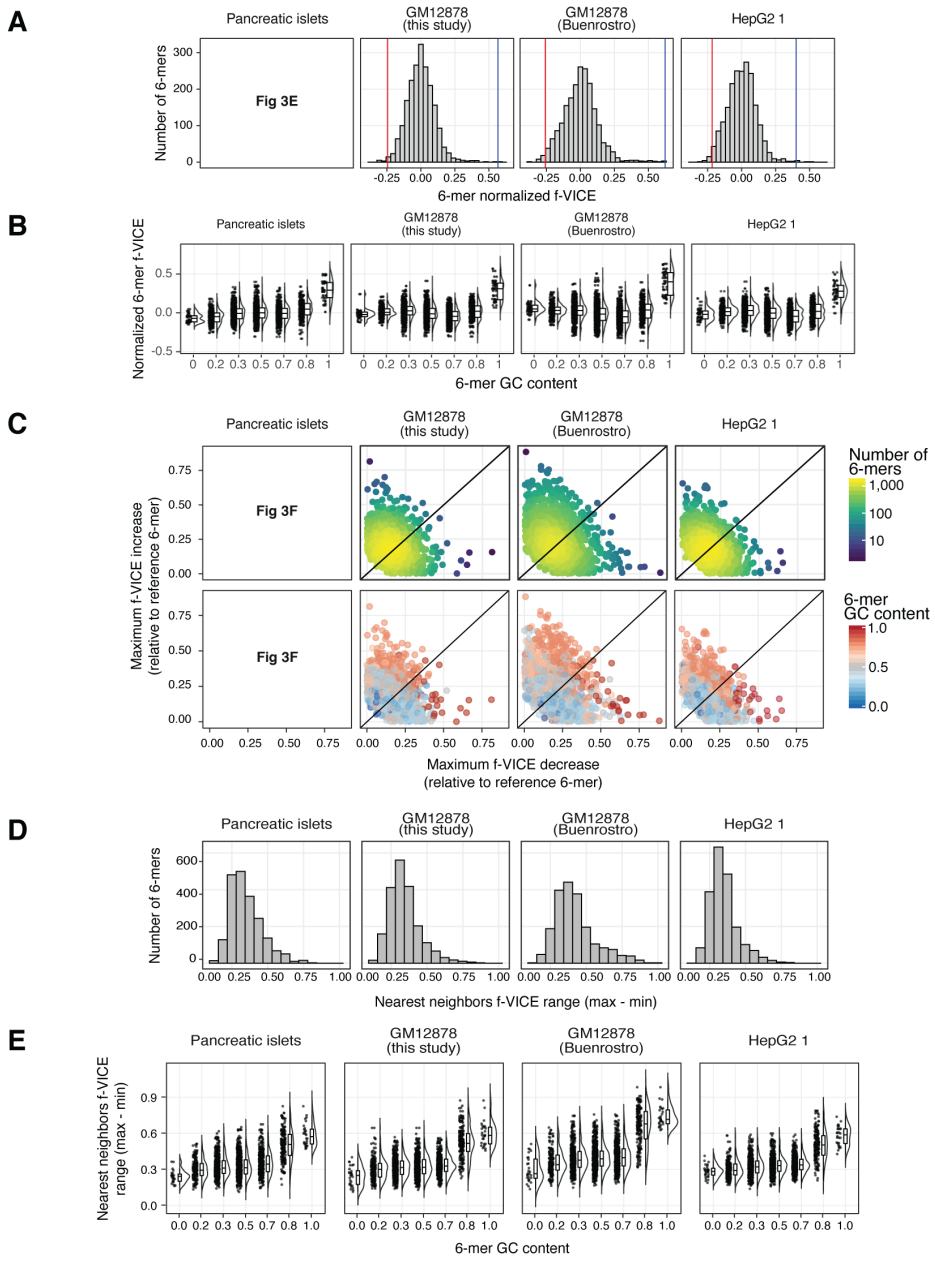


Figure 2.23: DNA 6-mers f-VICE analyses. (a) DNA 6-mers normalized f-VICE distributions across datasets. Horizontal lines represent the normalized f-IVCEs of the two 6-mers shown in Figure 3e (CGCCCC in blue and CGACCC in red). (b) 6-mer f-VICEs as a function of GC content. (c) Scatter plot of f-VICE differences for all 6-mers relative to 1bp neighbors in sequence space (i.e. 6-mers with a Hamming distance of 1). (d) Distribution of the f-VICE range of each 6-mer relative to its 1bp neighbors in sequence space. (e) Distribution of f-VICE range as a function of GC content. Note that high GC content 6-mers are more likely to have immediate neighbors in sequence space with lower f-VICEs.

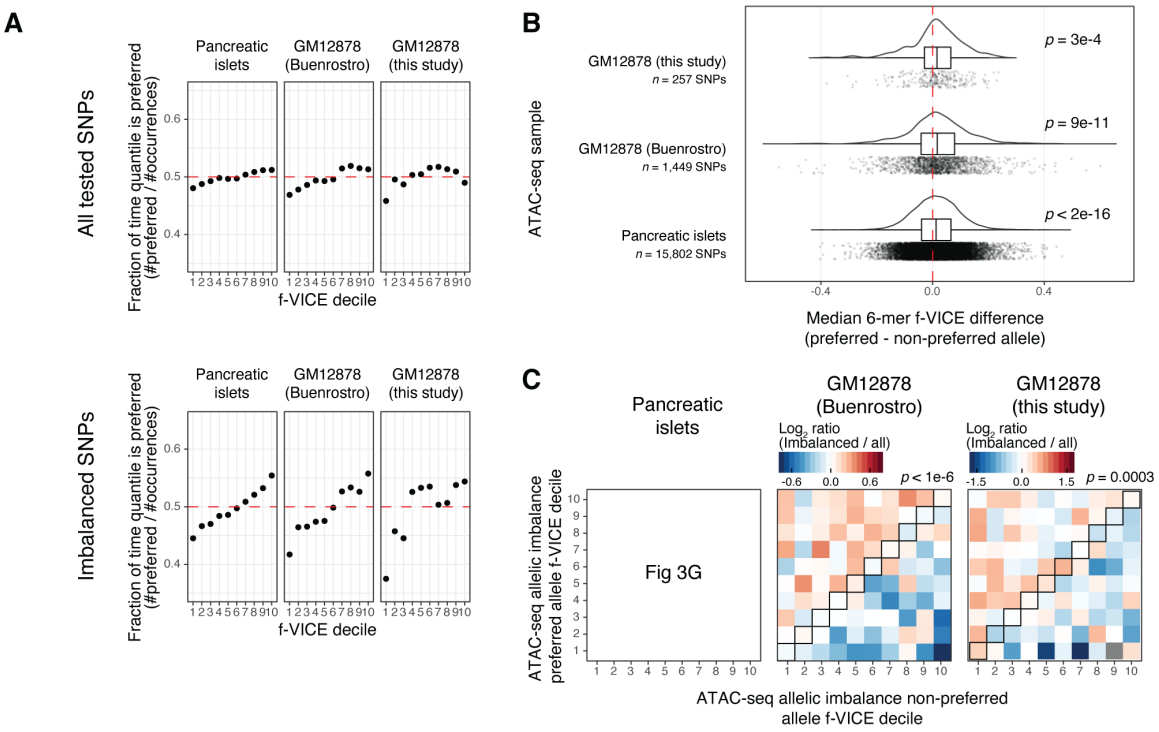


Figure 2.24: f-VICE allelic imbalance analyses. (a) Proportion of time the preferred ATAC-seq allele forms a 6-mer belonging to each f-VICE decile in all tested SNPs (upper) and all SNPs with significant allelic imbalance (lower). (b) Distribution of 6-mer f-VICE difference between the preferred and non-preferred alleles at loci with significant ATAC-seq imbalance. Each point corresponds to a DNA 6-mer overlapping a locus with allelic imbalance. Red dashed line corresponds to the expectation. P-values obtained from binomial tests. (c) f-VICE decile transition matrices. Each square corresponds to the ratio of imbalanced versus all tested SNPs. P-values obtained from permutation tests.

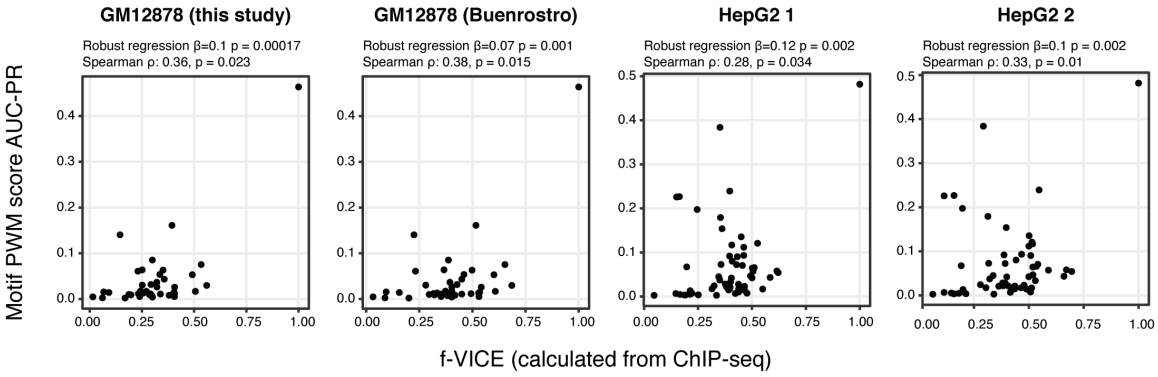


Figure 2.25: f-VICE and PWM score AUC-PR. Scatter plots of f-VICE and FIMO position weight matrix (PWM) score AUC-PR relative to ChIP-seq data. Robust linear regressions calculated using formula AUC-PR $\sim$ f-VICE.

CHAPTER III

Analyses of Thymic Precursors Chromatin Accessibility to Elucidate Thymocyte Development

3.1 Foreword

Even though this chapter is positioned later in the dissertation, the work described here was performed in parallel with the work from the previous chapter. Note that we used CENTIPEDE for the TF binding prediction. This was due to BMO being still under development at the time the manuscript was submitted for peer revision.

3.2 Abstract

In vertebrates, multiple transcription factors (TFs) bind to gene regulatory elements (promoters, enhancers, and silencers) to execute developmental expression changes. ChIP experiments are often used to identify where TFs bind to regulatory elements in the genome, but the requirement of TF-specific antibodies hampers analyses of tens of TFs at multiple loci. Here we tested whether TF binding predictions using ATAC-seq can be used to infer the identity of TFs that bind to functionally validated enhancers of the *Cd4*, *Cd8*, and *Gata3* genes in thymocytes. We performed ATAC-seq at four distinct stages of development in mouse thymus, probing the chromatin accessibility landscape in double negative (DN), double positive (DP), CD4

single positive (SP4) and CD8 SP (SP8) thymocytes. Integration of chromatin accessibility with TF motifs genome-wide allowed us to infer stage-specific occupied TF binding sites within known and potentially novel regulatory elements. Our results provide genome-wide stage-specific T cell open chromatin profiles, and allow the identification of candidate TFs that drive thymocyte differentiation at each developmental stage.

3.3 Results

3.3.1 Introduction

T cells develop in the thymus, where biologically distinct events driven by the interplay of multiple transcription factors (TFs) acting in coordination take place at each thymocyte stage. After migration of thymic seeding progenitors from the bone marrow and their occupation of supportive niches in the thymic medulla, early thymic progenitors (ETP) develop through immature double negative (DN; CD4-CD8-) cells to the double positive (DP; CD4+CD8+) stage, and then mature into either CD4 single-positive (SP4) helper or CD8 SP (SP8) killer T cells. While ETP retain multilineage differentiation capacity, they gradually lose the potential to become non-T lineage cells and become increasingly restricted to a T lineage fate [145–148]. During the DN stages, committed, developing T cells undergo immune system-specific DNA recombination, and must successfully recombine a Trb gene allele (encoding the T cell β receptor, TCRβ) to then pass the β selection checkpoint (when the formation of a pre-TCR complex is assessed)[149, 150] in order to survive. At the next DP stage (where both CD4 and CD8 are expressed on the cell surface), the TCRα receptor rearranges, and only cells expressing a functional cell surface TCR complex (TCRα plus TCRβ) that is able to bind with appropriate affinity to the major histocompatibility complex (MHC) survive positive selection [151]. DP cells recognizing MHC

class I can then mature into SP8 T cells, while DP cells recognizing MHC class II mature into SP4 T cells. Finally, negative selection eliminates by apoptosis cells that bind to self-peptides presented by the MHC, and only cells that do not exhibit high affinity to self-peptides survive [151].

Although T cell developmental stage-specific gene expression profiling has been previously described [152, 153], the mechanisms that regulate those spatial and temporal expression patterns are far less well understood for all but a handful of genes. DNA-binding TFs play a central role governing gene expression in each cell, often eliciting transcriptional responses through specialized regulatory elements, including promoters, enhancers, and silencers. A widely accepted model for gene expression is that multiple transcription factors bind to an enhancer, assemble an enhanceosome, and then recruit co-activators and chromatin-remodeling proteins to the promoter [154, 155]. Given the limitations of ChIP-seq to detect a single TF per assay, an alternative approach for detecting TF binding is using open chromatin assays, such ATAC-seq [33, 84]. The genome is highly compact except within transcribed genes and regulatory elements, where chromatin is open and sensitive to cleavage by DNaseI [156–158] or transposition by Tn5 transposase [21]. The binding of TFs to DNA affects DNase/transposase cleavage in the vicinity of the bound site, allowing for TF occupancy to be predicted from the chromatin accessibility data [33, 84, 159]. Thus DNase/ATAC footprinting can be used to identify TF binding motif sequences within regulatory elements.

To generate genome-wide profiles of stage-specific chromatin accessibility and TF binding during thymocyte development, we performed ATAC-seq at four different stages of adult thymocyte development: DN, DP, SP4 and SP8 stages. The open chromatin regions identified by ATAC-seq highlighted both known, biologically validated regulatory elements, as well as many novel potential regulatory elements. Furthermore, footprinting analysis [33, 84] of those open chromatin regions revealed the

high-resolution landscape of predicted TF-bound motifs within those sequences. Our ATAC-seq data enabled the discovery of both stage-independent and stage-specific domains of open chromatin, and the TF footprinting data revealed 10-20 novel protein bound sequences within the previously validated enhancers of the *Cd4*, *Cd8*, *Trb* and *Gata3* genes. Furthermore, enrichment analyses of TF binding in stage-specific open chromatin allowed the identification of TF motifs potentially driving each stage of thymocyte development. These data demonstrate that stage-specific changes in open chromatin are highly dynamic as thymocytes develop and provide deep insight into how the stage-specific binding of multiple TFs orchestrate transcriptional regulatory networks.

3.3.2 Chromatin accessibility varies across thymocyte development

T cell developmental stage-specific genome-wide mapping of accessible chromatin. To gain insight into developing T cell stage-specific chromatin opening, DN, DP, SP4 and SP8 cells were isolated from adult thymi by flow cytometry (Figure 3.8). 50,000-100,000 cells were processed for ATAC-library preparation as described [21]. The ATAC-seq reads were then mapped to mouse reference genome mm10 using BWA [110] and peaks were called using MACS2 [112]. ATAC-seq signals depicted in the IGB browser [160] were reproducible in thymocytes recovered from 4 individual animals (Figure 3.9), and all peaks were highly correlated across biological and technical replicates (median Spearman correlations: DN = 0.89, DP = 0.87, SP4 = 0.88, SP8 = 0.90; Figure 3.10). ATAC-seq signals at the DP stage (which comprises approximately 85% of total thymocytes, Figure 3.8) reflected profiles that were similar to DNase-seq peaks of total adult thymocytes [161] (Figure 3.9), as anticipated. On a global scale, DP ATAC-seq peak signals were highly correlated with DNase-seq peak signals of total thymocytes (median Spearman correlation = 0.70 to 0.79 Figure 3.10). Based on these results, we concluded that ATAC-seq provides a biologically reliable strategy to attain

deeper insights into T cell stage-specific chromatin accessibility and transcription factor binding.

We identified 150,139 (DN), 107,110 (DP), 115,074 (SP4) and 104,411 (SP8) genomic open chromatin peaks at 5% FDR (Figure 3.11). These open chromatin domains correspond to 1.63% (DN), 1.22% (DP), 1.32% (DP) and 1.26% (SP4) of the mouse genome. 73,177 peaks were present at all four stages of thymocyte development, while the others were stage-specific. 20% (DN), 27% (DP), 24% (SP4) and 26% (SP8) of the ATAC peaks overlapped with promoter regions (defined as 200 bp upstream of a gene transcriptional start site). 10% (DN), 9% (DP), 8% (SP4) and 9% (SP8) of the ATAC peaks overlapped with an exon, but not with a promoter. 73% (DN), 63% (DP), 68% (SP4) and 65% (SP8) of the ATAC peaks overlapped with neither an exon nor a promoter (Figure 3.11).

We next sought to quantify the full spectrum (from specific to ubiquitous) of patterns of chromatin accessibility across the analyzed thymocyte developmental stages in an unbiased manner. We performed k-means clustering using the ATAC-seq signal. This analysis yielded 6 clusters of accessible regions: four that were specific for each stage (DN, DP, SP4, SP8), one that was ubiquitous, and one that was a combination of DN and ubiquitous (Fig. 3.1a, Figure 3.12). The ubiquitous cluster covered more genomic territory than any of the stage-specific clusters, while the DN-specific cluster covered more territory than the other stage-specific clusters (Fig. 3.1b), which is consistent with the previous conclusion that in general differentiated cells maintain a more compact chromatin architecture than their immature counterparts [162].

We next measured the distance of each peak in the four clusters to the nearest TSS and found that the ubiquitous cluster was significantly closer to TSS than the other clusters ($p< 10^{-3}$, pairwise Kolmogorov-Smirnov tests with Bonferroni correction), suggestive of it being more associated with promoters and housekeeping genes than cell-identity features (Fig. 3.1c). Supporting this hypothesis, we found that SP4- and

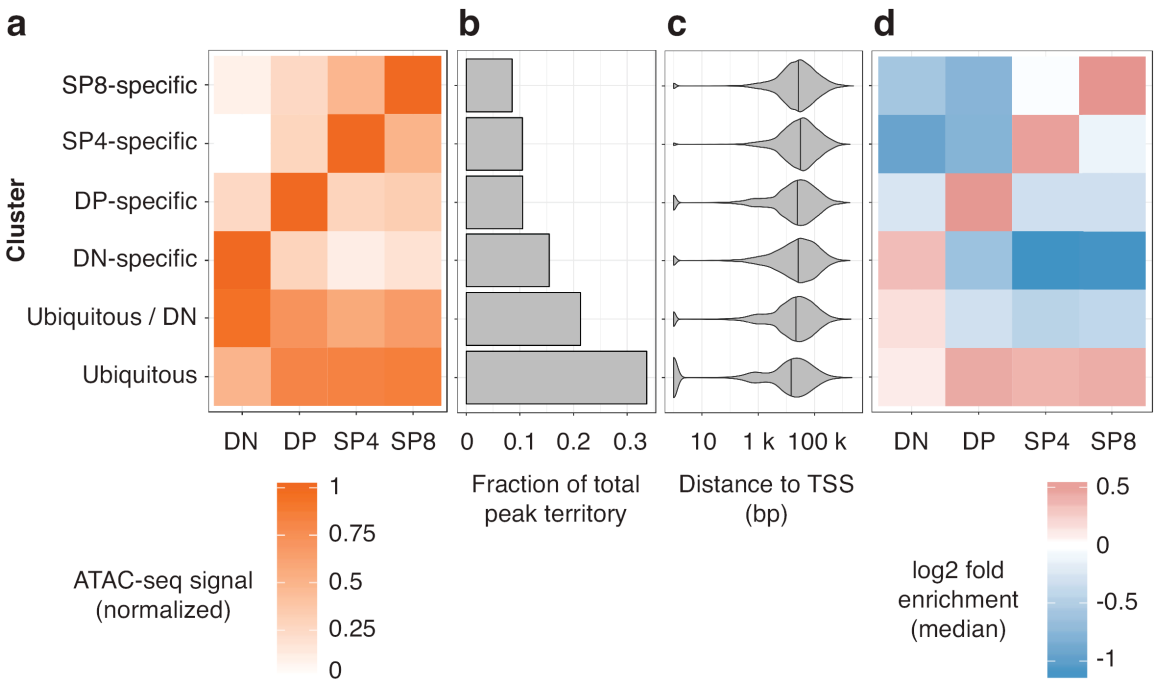


Figure 3.1: Analysis of stage-specific and ubiquitous ATAC-seq clusters. (**a**) k-means clustering results for the ATAC-seq signal in the four samples. Colors indicate the mean ATAC-seq signal for each sample in the respective cluster (i.e. the cluster centers). (**b**) Fraction of total peaks territory covered by each of the clusters. (**c**) TSS distance distribution (in log10 scale) for each of the clusters. Vertical bars in the violin plots correspond to the median of the dataset. (**d**) Median GAT footprint enrichment of all motifs for each dataset in the k-means clusters. GAT footprint enrichment heatmaps for each motif are shown in Figure 3.13

SP8-specific clusters were the most enriched for T cell related GO terms using ChIP-Enrich [163] (Figure 3.2). The DP-specific cluster also had high enrichment for terms related to T cell differentiation, but to a lesser extent. Conversely, the DN-specific and ubiquitous clusters were strongly enriched for non-specific developmental terms, suggesting that these might regulate more general functions. These results form a comprehensive map of developmental dynamics in the open chromatin landscape across thymocyte maturation.

3.3.3 TF binding identification by ATAC-seq footprinting

In order to achieve greater insights into genomic DNA sequences that are bound by TFs, we performed TF footprinting predictions using CENTIPEDE [49]. To validate the performance of CENTIPEDE footprint calls in our thymocyte data, we first compared our results with GATA3 ChIP-seq data in DN and DP thymocytes (GSE20898) [164] and CTCF ChIP-seq in total thymocytes (ENCODE, ENCSR000CDZ) [161]. We used the Genomic Annotation Tester (GAT) tool [136] to statistically evaluate the overlap between footprint calls and ChIP-seq bound motifs, while controlling for genome and feature sizes, as well as mapability issues (see Methods for details). Tests on both datasets showed significant overlap between ChIP-seq and footprint data ($p < 10^{-3}$), indicating that the footprint predictions recapitulate actual protein binding events detected by ChIP-seq for the corresponding TF. These data demonstrate the effectiveness of the footprint calls from these deeply sequenced ATAC-seq data in order to generate a high confidence catalogue of putative TF-bound sequences in a thymocyte stage-specific manner.

We next focused on TF binding motifs for T cell activators and repressors that were predicted by Jojic et al. [165] from stage-specific gene expression profiling. We predicted the binding for the Jojic factors and their families. As expected, we found that the footprints called in each sample were enriched both in its own specific cluster

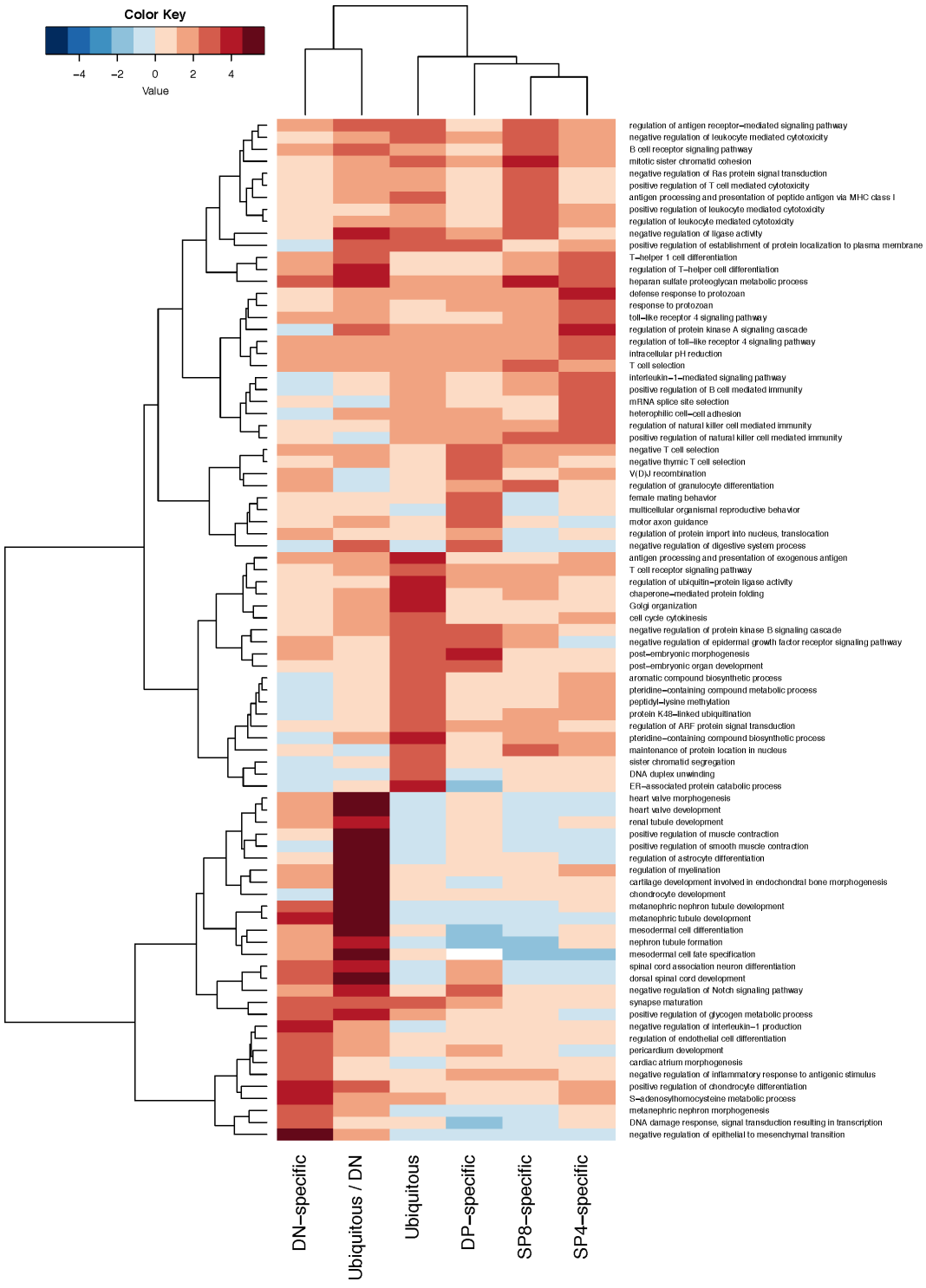


Figure 3.2: ChIP-Enrich results for the ATAC-seq clusters. GO term enrichment values for the top 15 terms called in each of the clusters and their corresponding enrichments for all clusters.

and the ubiquitous cluster, but depleted in the other sample-specific clusters using GAT, indicating that CENTIPEDE did not detect bound TF binding sites in regions that were not active in the sample being analyzed (Fig. 3.1d and Figure 3.13).

We next focused on footprint calls within functionally validated enhancers. The classic definition of enhancer requires that it must be functionally validated by tests for both sufficiency and necessity in regulating its specific target gene expression, but to date only few T cell enhancers have been tested for both *in vivo*. The ATAC-seq data identified open chromatin regions within functionally validated regulatory elements for the *Cd4* (Fig. 3.3 and Figure 3.18), *Cd8* (Fig. 3.4 and Figure 3.14), *Trb* (Figure 3.15) and *Gata3* (Fig. 3.5) genes. The fact that the ATAC footprints recapitulated TF binding to the previously characterized motifs within these regulatory elements underscore the robust nature of the footprint approach employed in this study. Furthermore, our footprint data unveiled 8-20 novel sequences that were predicted to be bound within each of these regulatory elements (Figures 3.3, 3.4, 3.5, 3.18, 3.14, and 3.15). Based on these data, we propose that TFs bind to these sequences to assemble an active structural element that initiates and/or maintains the activity of each of these regulatory modules.

3.3.4 Changes in global TF binding during thymocyte development

We next sought to identify higher-resolution differences in predicted TF binding across samples by measuring the pattern of chromatin accessibility anchored on footprint motifs. We found striking differences for the footprint motifs across the samples and clusters in which they were active (Fig. 3.6 and Figure 2.14). CTCF had strong detectable binding patterns only in the ubiquitous cluster, and a similar pattern was observed for EGR3. TCF7 (aka TCF-1) had significant binding in all clusters. TCF4, on the other hand, was detected more strongly in the DP and DN clusters, was mostly absent in the common clusters, and almost undetectable in SP4

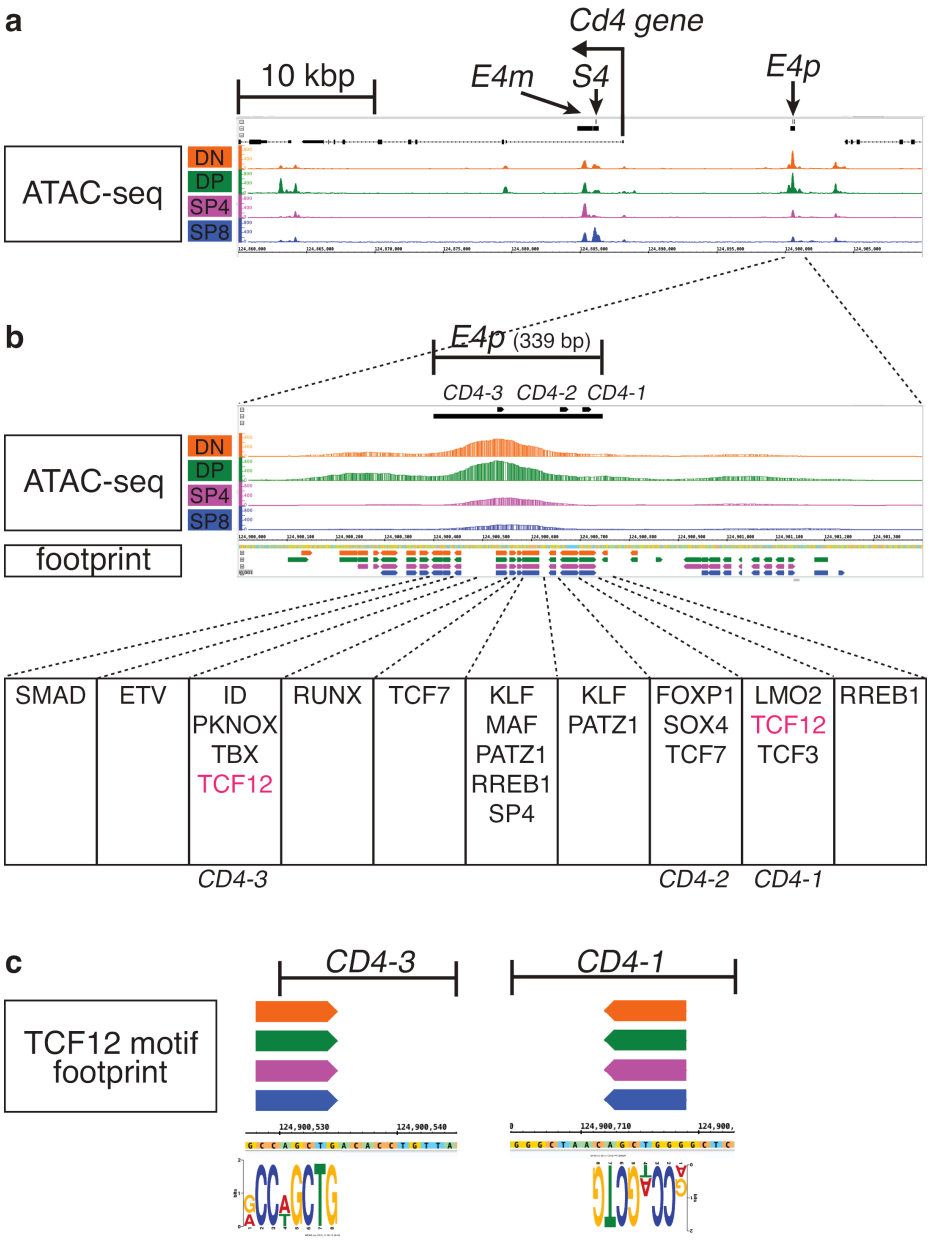


Figure 3.3: ATAC-seq signal and CENTIPEDE footprint calls around the functionally validated *E4p Cd4* gene enhancer. (**a**) ATAC-seq signals are shown on the IGB browser within around 50 kbp of the *Cd4* locus; mm10, chr6:124,860,001124,910,000. The positions of the *E4p*, *E4m* enhancers and S4 silencer [166–168] are shown at the top. (**b**) ATAC signal and footprint calls around *E4p* are depicted. *CD4-1*, *CD4-2* and *CD4-3* sequences were first identified by DNaseI footprinting in the SL3B T cell line40. (**c**) TCF12 (aka HEB) motif footprints in the *CD4-1* and *CD4-3* sequences. Footprint calls within S4 and *E4m* are shown in Figure 3.18.

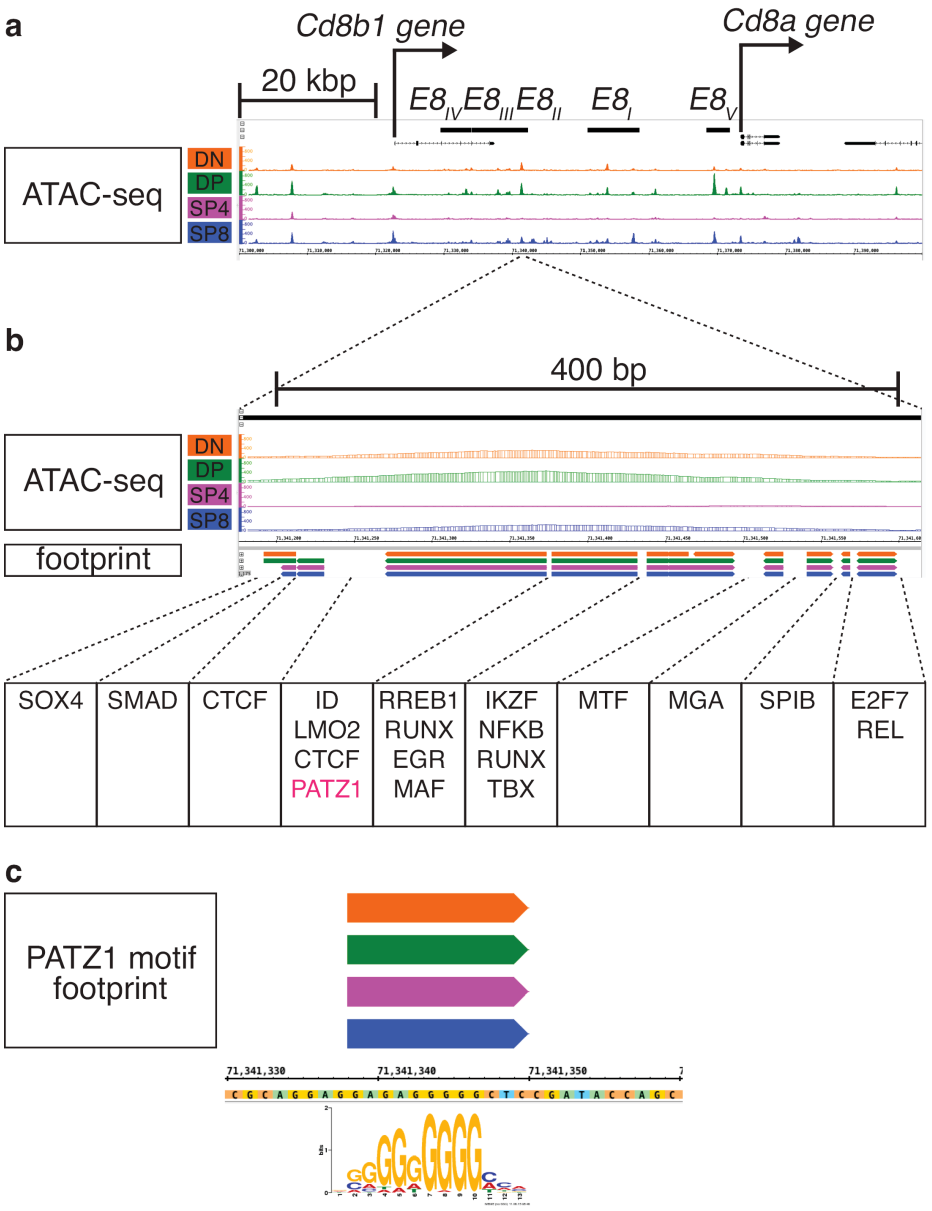


Figure 3.4: ATAC-seq signal and footprint calls within functionally validated enhancers for the *Cd8* gene. (a) ATAC-seq signals are shown on the IGB browser within around 100 kbp of the *Cd8* locus; mm10, chr6:71,300,001-71,400,000. The positions of the $E8_I$ -$E8_V$ enhancers [169–171] are depicted at the top (b) ATAC signals and footprint calls at an ATAC peak identified in $E8_{II}$ are shown. (c) A PATZ1 motif footprint within $E8_{II}$ is shown. Footprint calls within $E8_I$ and $E8_V$ are shown in Figure 3.14.

and SP8 thymocytes, even though it was one of the most significantly enriched motifs in these two stages ($p = 0.001$). RUNX patterns were visible in the common and DN clusters, but not in the more differentiated stages. Although GATA footprints were enriched in all clusters, we could not detect strong binding patterns, which is suggestive that it may have weaker interactions with DNA29. Interestingly, we did not find any SP4- or SP8-specific occupancy patterns, even though some motifs, such as TCF3 and ID4, had higher enrichment values in the SP4- and SP8-specific clusters than in the ubiquitous cluster. These different patterns between stages for TCF3 and ID4 suggest that the availability (expression or protein levels) of these TFs changes or that different TFs recognize these motifs at each stage.

We finally asked which footprints were enriched in each of the stage-specific open chromatin clusters defined in Fig. 3.1a. Each cluster showed enrichment of different TF motif footprints (Fig. 3.7). Of note, we independently performed motif enrichment in the ATAC-seq clusters using HOMER (Figure 3.17), but this approach did not capture the nuanced enrichments we detected with the footprinting approach, as HOMER only takes the motif occurrences and not the ATAC-seq signal into account. These data support the concept that many TFs bind to specific transcriptional regulatory elements at each developmental stage to achieve stage-specific gene expression patterns, and that the binding of these individual factors is reflected in the dynamic changes in transcriptional networks that must accompany thymocyte developmental progression from one stage to the next.

These footprinting data identified potential stage-specific regulators. Out of the 34 SOX family TF motifs tested, 10 and 9, respectively, were within the top 20 enrichment scores for DN and DP, but not in in SP4 and SP8, suggesting that a SOX family TF(s) is important for DN- and DP-specific gene expression. Of 3 PBX family TF motifs tested, 2 were within the top 20 fold-enrichment scores for the DP-specific cluster. These data suggest a role for PBX family TFs in DP-specific gene

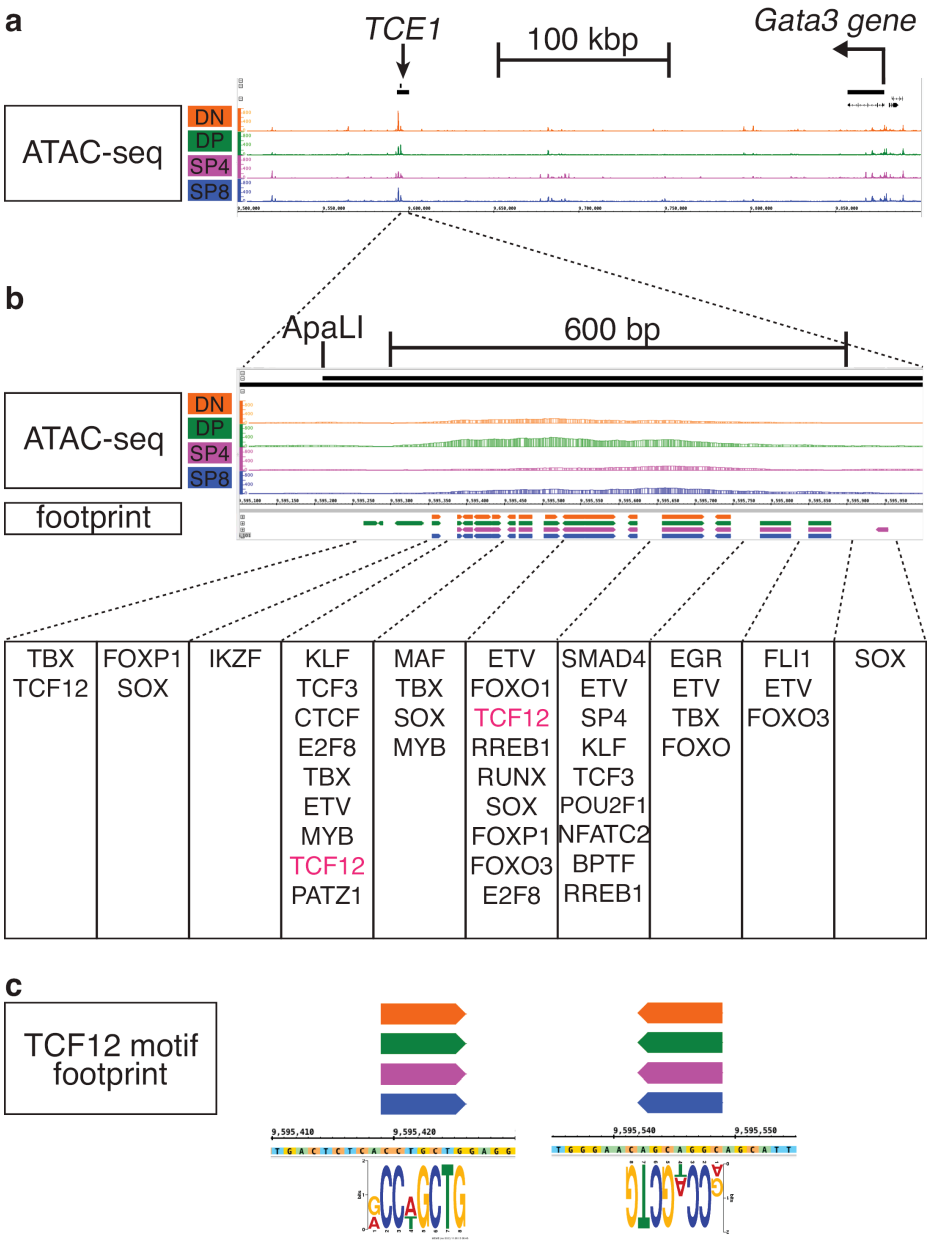


Figure 3.5: ATAC-seq signal and footprint calls within the functionally validated *TCE1 Gata3* enhancer. (a) ATAC-seq signals are shown on the IGB browser around 400 kbp of the *Gata3* gene; mm10, chr2: 9,500,0019,900,000. (b) ATAC peak and TF footprint calls at an ATAC peak found in TCE1 core [172, 173]. (c) TCF12 (aka HEB) motif footprints.

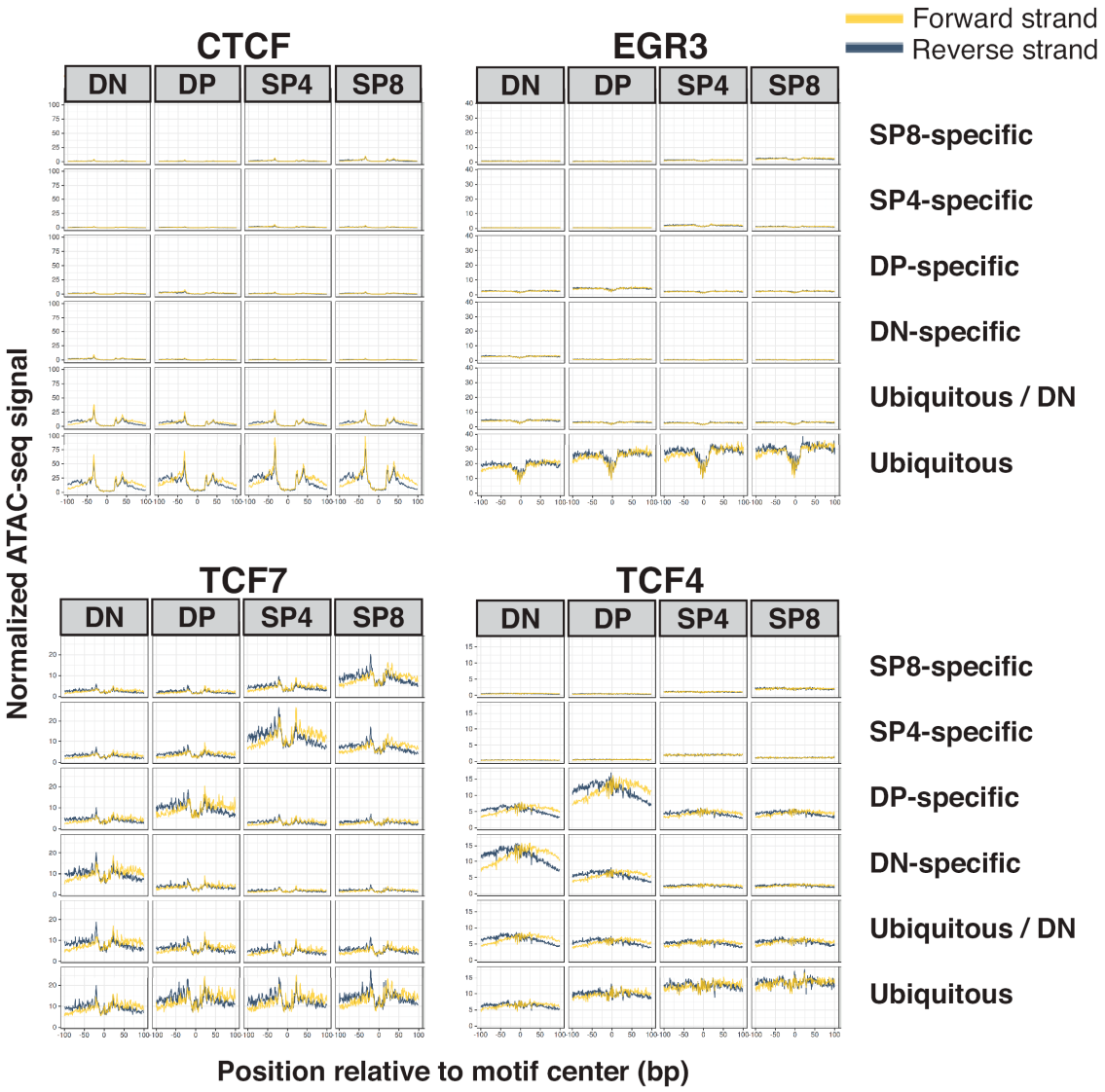


Figure 3.6: Footprint occupancies across samples and clusters. Normalized occupancy signals (see Methods) at ±100 bp of motif center for CTCF, EGR3, TCF7 and TCF4. Horizontal facets correspond to the ATAC-seq samples, and vertical facets correspond to the k-means clusters.

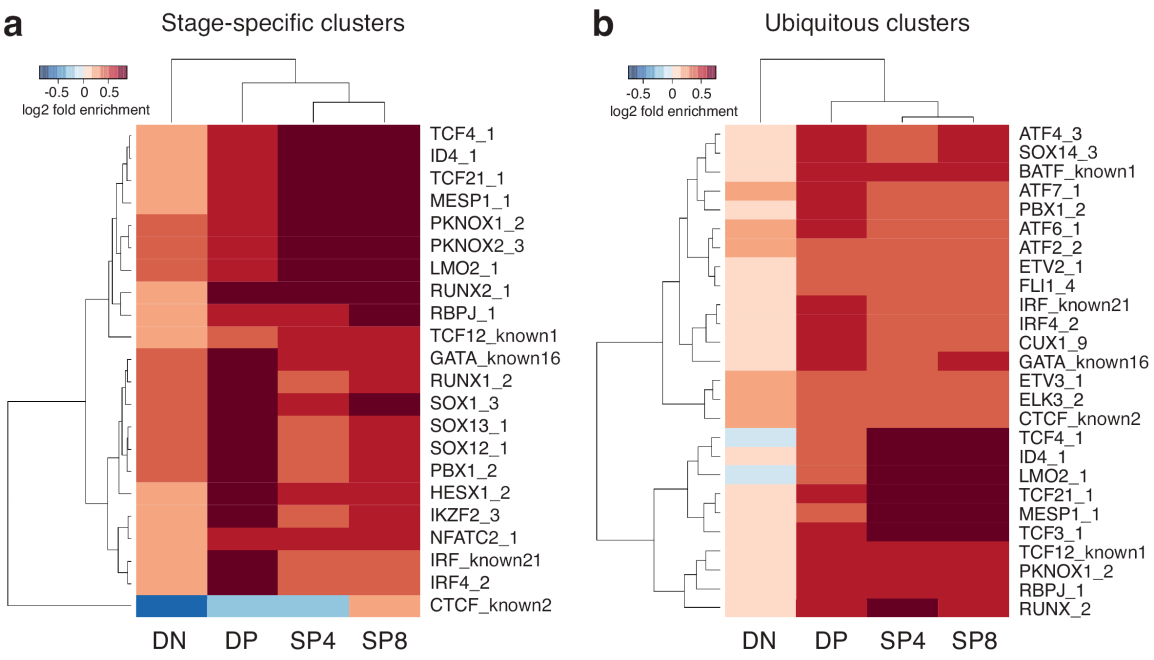


Figure 3.7: Individual footprint enrichments in each of the samples. (a) GAT enrichment scores for the top factors that maximize the variance between the stage-specific clusters of each of the ATAC-seq samples. (b) Similar to (a), but comparing the enrichments for each ATAC-seq sample in the ubiquitous cluster. Darker red colors correspond to stronger enrichments, and blue colors correspond to depletions.

expression. Out of 4 PKNOX family TF motifs tested, 3 were within the top 20 enrichment scores for the SP4 and SP8 clusters, indicating that PKNOX family TFs contribute to SP4- and SP8-specific gene transcription. Finally, out of 7 MAF family TF motifs tested, 1 and 3 were within the top 20 fold-enrichment scores for the SP4 and SP8 clusters, respectively, supporting the hypothesis that MAF family TFs are important for SP8-specific gene expression.

3.4 Discussion

We performed ATAC-seq experiments and footprint analyses at four major stages of thymocyte development in order to compile a catalogue of stage-specific accessible chromatin sequences as well as to identify specific sequences bound by TFs. We

identified ubiquitous and stage-specific open chromatin regions, recapitulating the identity of functionally validated regulatory elements, as well as revealing novel regulatory loci. The ATAC-seq footprinting data for predicted $\alpha\beta$T cell activators and repressors highlighted TF-bound motifs within those regulatory regions, as well as bound motifs that were enriched in each of the thymocyte stage-specific accessible chromatin clusters, providing an in-depth view into the inner regulatory workings of thymocyte development. We identified between 8 and 20 novel sequences that were predicted to be bound by proteins within previously identified regulatory elements for the *Cd4*, *Cd8*, *Trb* and *Gata3* genes, which supports the idea that an approximately 8-20 TFs bind to an enhancer in order to form a TF complex/enhanceosome that is capable of supporting the initiation and/or activation of enhancer activity. Thus one future goal is to investigate the ability of individually bound sequences to contribute to enhancer activity, which can be tested by *in vivo* ablation or mutation of specific TF motifs. The genome-wide footprinting approach detailed here is an alternative to ChIP experiments, but the two are complementary. It is well known and has been documented that several different proteins can bind to a given sequence motif (*e.g.* all six vertebrate GATA factors bind with reasonably high affinity to the AGATAA sequence motif, so identification of a given *cis* element in the absence of data regarding the tissue specificity of a given family of factors may only be marginally informative). ChIP experiments, in contrast, can capture indirect binding by virtue of protein-protein interactions that occur in larger complexes formed with a specific DNA binding protein [174, 175], potentially complicating assignment of which factor is genuinely bound to DNA at any given site.

The thymocyte stage-specific open chromatin regions identified here by ATAC-seq followed by k-means clustering approach highlighted the positions for thousands of potentially novel developmental stage-specific regulatory elements. The ATAC peaks provided evidence for the previously predicted closed- or open-chromatin status in

both the *Cd4* and *Cd8* loci during thymocyte development [176]. Furthermore, the identification of two major ATAC peaks within the 7.1 kbp that originally defined the Gata3 enhancer, TCE1 [172, 173] suggests that one or both of these two open chromatin domains (of approximately 600 bp and 500 bp) play a major role in the enhancer activity of TCE1. In agreement with this hypothesis, one of these ATAC-seq peaks aligns perfectly with a 1.2 kbp "core" sequence that exerts similar reporter gene activation in thymocytes of transgenic mice that is roughly equivalent to the whole 7.1 kbp TCE1 sequence [173].

Enrichment of footprints in the stage-specific open chromatin clusters highlighted TF families binding to the motif as potential stage-specific regulators. These data provide an additional layer of information to the $\alpha\beta$ T cell factors [165] predicted from lineage specific gene expression profiles. The most immediate future plans following these identifications are to investigate whether or not each bound sequence is necessary for any specific enhancer/silencer activity, which can be tested by *in vivo* genomic DNA mutation of the TF motif. In summary, the genome-wide view of open chromatin presented here as well as the identification of the sequence motifs bound by TFs at four different stages of thymocyte development is a useful point from which to begin to assemble precise models for transcriptional regulation of T cell stage-specific gene expression.

3.5 Methods

ATAC-seq. ATAC libraries were prepared as described previously [21]. In brief, 50,000 to 100,000 DN, DP, CD4 SP and CD8 SP thymocytes were isolated by flow cytometry (Figure 3.8). Cells were processed for ATAC reaction, and then the ATAC libraries were PCR amplified with barcoded primers. The ATAC-seq libraries were paired end 75 bp sequenced on a HiSeq 4000 at the UM Sequencing Core. Raw reads were trimmed for barcodes and aligned to the mm10 reference genome using BWA

[110]; duplicates were removed with Picard and then filtered for high quality (mapq $\geqslant$ 30), properly paired alignments and uniquely mapped as described in our previous study [33].

Peak calling. In order to account for sequencing depth differences between each library, we down-sampled reads (keeping read pairs intact) to the median depth of all libraries after the pruning steps described above. This ensures that sequencing depth would not confound the analysis. After this step, we combined all replicates from each stage into a single BAM file to increase sequencing depth (ranging from 120 to 134 million reads per stage) and called peaks using MACS2 [112] with options -nomodel -shift -50 -extsize 100 -B -keep-dup all. For testing the reproducibility between samples, we generated a set of regions that were called (narrow) peaks in at least one of the merged samples, retrieved the number of fragments mapping to these regions in each replicate and calculated the pairwise Pearson correlations between all replicates of the same stage.

k-means clustering and functional enrichments. To perform k-means clustering, we generated a set of genomic regions that were called peaks in at least one of the samples (master peaks list) by using bedtools merge in the combined MACS2 output for all samples. For each sample, we calculated the FPKM in each of the master peaks regions, and normalized the signal by dividing the values by the TSS enrichment of the sample, which accounts for the signal-to-noise ratio, and then applied robust IQR scaling:

$$X_{scaled} = \frac{(x_i - median(X))}{IQR(X)} \tag{3.1}$$

where IQR is the distance between the 1st and 3rd quartiles, to make the values comparable across samples. This signal was then row-wise normalized by the

maximum of every sample:

$$Y_{normalized} = \frac{y_i}{max(Y)} \tag{3.2}$$

Using this matrix of genomic coordinates per samples, we ran the k-means implementation available in R 3.3.1 for k = 1, 2, ..., 15 k values and determined that $k = 6$ was suitable for our analyses. Increasing k to higher values only marginally decreased variance and yielded repetitive clusters patterns, with 1,000 random starts for robustness. We analyzed the within cluster variances for all (Figure 2.8). In order to perform functional annotation of the clusters, we used the ChIP-Enrich R package [163], which allow us to directly compare the enrichment scores and p values for the same GO terms across samples.

PWM scans and ATAC-seq footprints. In the current study we focused on TF binding motifs for T cell activators and repressors that were predicted by Jojic *et al.* [165] from stage-specific gene expression profiling (171 $\alpha\beta$T cell factors, Supplementary Table 10 in ref. [165]). Position weight matrix (PWMs) for each motif was obtained from ENCODE [177], JASPAR [178] and TF pairs identified by Jolma *et al.* [45]. Total 417 binding motifs for 67 out of 171 Jojic $\alpha\beta$ T cell factors were derived from these databases. We scanned the mm10 genome for the PWMs for the 417 motifs using FIMO [43] with the G-C content back- ground frequency for mm10 (41.7%), and used the default 10^{-4} P value threshold, also filtering for motif occurrences intersecting regions with known mapability issues (blacklisted regions). CENTIPEDE [49] was used to call footprints from the ATAC-seq data as we have done previously [33, 84]. Briefly, for each PWM scan result we generated a strand-specific (relative to the motif orientation) single base pair resolution matrix encoding the number of Tn5 transposase integration events in a region ±100 bp from each motif occurrence. A motif occurrence was considered bound if the CENTIPEDE posterior probability was

higher than 0.99 and its coordinates were entirely contained by an ATAC-seq peak. To generate the motif occupancy plots for each factor, we aggregated the signal used as input for CENTIPEDE for all the predicted bound motifs, as well as an equal number of motifs with posteriors less than or equal to 0.5 and not intersecting ATAC-seq peaks in that sample. The normalized signal plotted was obtained by dividing the bound signal by the unbound.

Overlap of ATAC-seq footprint and ChIP-seq. In order to test the correspondence between foot- print calls and ChIP-seq data for GATA and CTCF, we used GAT [136] with the workspace set as all the GATA or CTCF motif matches in the mm10 genome, the respective ChIP-seq peaks as the segments, and the respective CENTIPEDE footprint calls as the annotation. By limiting the workspace only to the specific motifs, the data stringently delimit the space for genomic interval overlap testing. The footprint enrichments in the ATAC-seq clusters were performed separately for each sample and for each motif. We used as workspace all the motif occurrences within the master peaks regions (see k-means clustering above) for the individual motif being analyzed. As annotations, we used the cluster designations from the k-means analysis. The segments were all the footprints for that motif in that sample. Additionally, we used the option -n to 1,000 in order to increase statistical robustness. This resulted in a table with the GAT results for every motif in each cluster and in each sample.

Data Availability. ATAC-seq and footprint data have been deposited in GEO database [179] and are accessible through accession number GSE107076.

3.6 Acknowledgements and Publication

The results described in this chapter are published in [180]. This project resulted from a collaborative effort between the Parker and Engel groups. I thank all co-authors and specifically co-first author Tomonori Hosoya for generating the ATAC-seq data and his help contextualizing the computational results. I specifically contributed towards the computational analyses and manuscript preparation. Finally, I thank Professors Parker and Engel for supervising all aspects of this project.

3.7 Appendix: Additional Figures

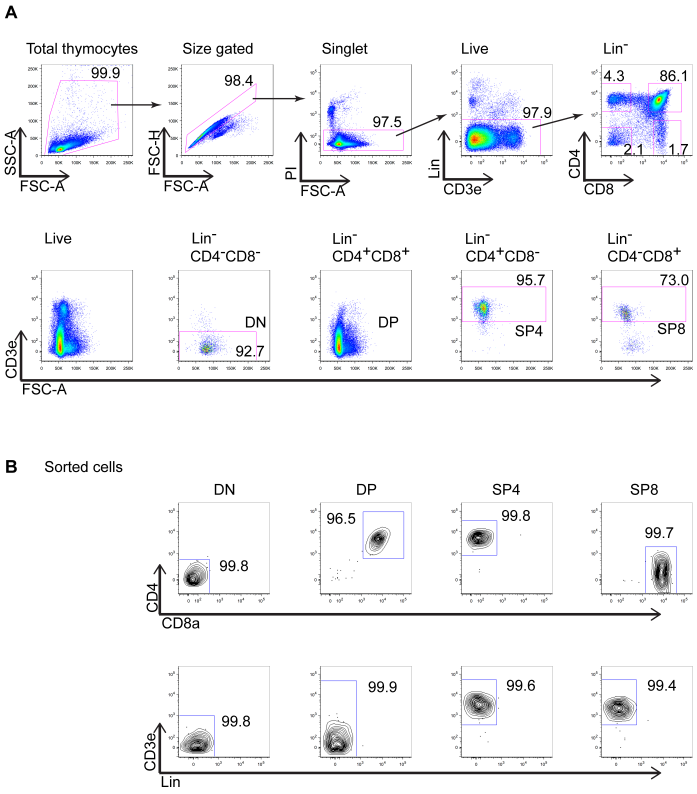


Figure 3.8: Isolation of staged thymocytes. (a) Individual stages of thymocytes were isolated using a FACSAria III (BD). Representative dot plots from 1 mouse (out of 4 animals examined in two different experiments) are shown. Area Scaling was set with total thymocytes and doublets were gated out using FSC-A vs. FSC-H. PI was used to discriminate dead from live cells. The gates (blue) are shown for DN (Lin-CD4-CD8-CD3-), DP (Lin-CD4+CD8+), SP4 (Lin- CD4+CD8-CD3+) and SP8 (Lin-CD4-CD8+CD3+) cells. The numbers near the boxed areas indicate the mean percentage of cells in each gate. The lineage cocktail used was a mixture of e450- conjugated antibodies recognizing TER119 (TER119), B220 (RA2-6B2), CD19 (1D3), Mac1 (M1/70), Gr1 (RB6-8C5), CD11c (N418), NK1.1 (PK136) and gdTCR (GL3). Cells were also stained with PECy-CD4 (RM4-5), APC-CD8 (53-6.7) and PE-CD3e (145-2C11) purchased from eBiosciences or BioLegend. (b) A small fraction of the sorted cells were reanalyzed to check for purity.

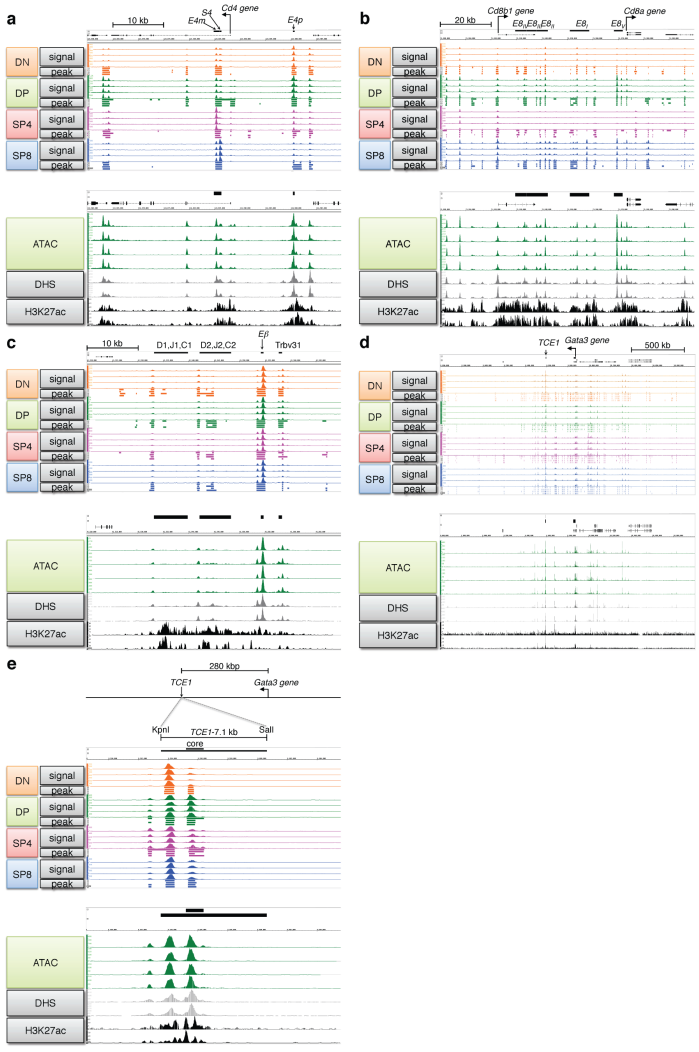


Figure 3.9: ATAC-seq profiles of thymocytes. ATAC-seq signals (MACS2 bedgraph converted to bigwig format) and peaks (MACS2 broad peak calling) surrounding the Cd4 (a), Cd8 (b) and Trb (c, encoding TCRβ), Gata3 (d) loci and TCE1 enhancer for Gata3 gene (e). Data are on the IGB browser around 50 kbp of the Cd4 locus (a, mm10, chr6:124,860,001- 124,910,000), around 100 kbp of the Cd8 locus (b, chr6:71,300,001-71,400,000), around 50 kbp of Trb gene beta enhancer (c, Eβ, chr6:41,520,001-41,570,000), within +/- 1.2 Mbp of the Gata3 gene (d, chr2: 8,600,001-11,000,000) and around TCE1 enhancer for Gata3 gene (e). (Top) ATAC- seq peaks were generated in quadruplicate in order to analyze chromatin accessibility in DN (orange), DP (green), SP4 (pink) and SP8 (blue) stage thymocytes. (Bottom) ATAC-seq peaks in DP stage (green), which compose approximately 85% of total thymocytes, were compared with DNase-seq (DHS, middle, ENCODE, ENCSR000COB, isogenic replicate 1 and 2) and H3K27ac ChIP-seq (bottom, ENCODE, ENCSR000CCH, isogenic replicate 1 and 2) in thymocytes.

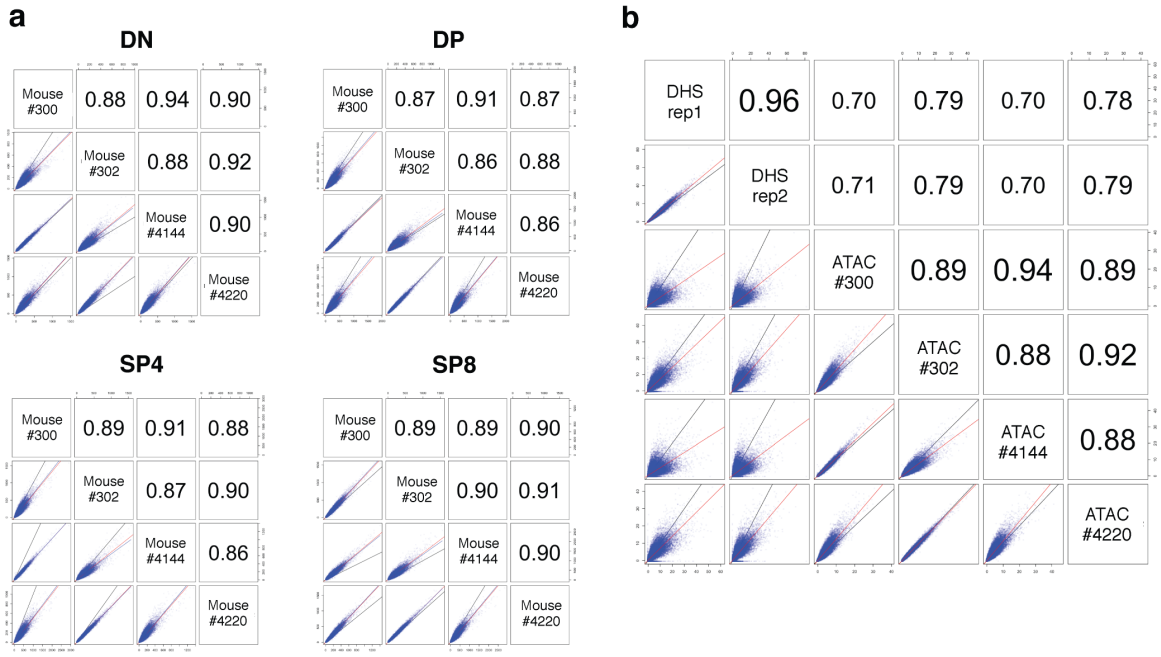


Figure 3.10: Correlation of the ATAC-seq signal between replicates. (a) Correlation between ATAC-seq libraries. (b) Correlation between DP ATAC-seq libraries and adult thymocytes DNase-seq data from total adult thymocytes (ENCODE,ENCSR000COB). Bottom facets: each data point corresponds to an ATAC-seq peak that was called in at least one sample (see Methods). The values plotted are the number of fragments in each peak in the corresponding samples (labelled on the diagonal), and the red and black lines correspond, respectively to the linear model fit from the two datasets and the identity (i.e. x = y). Upper facets: Spearman correlation values for the comparisons.

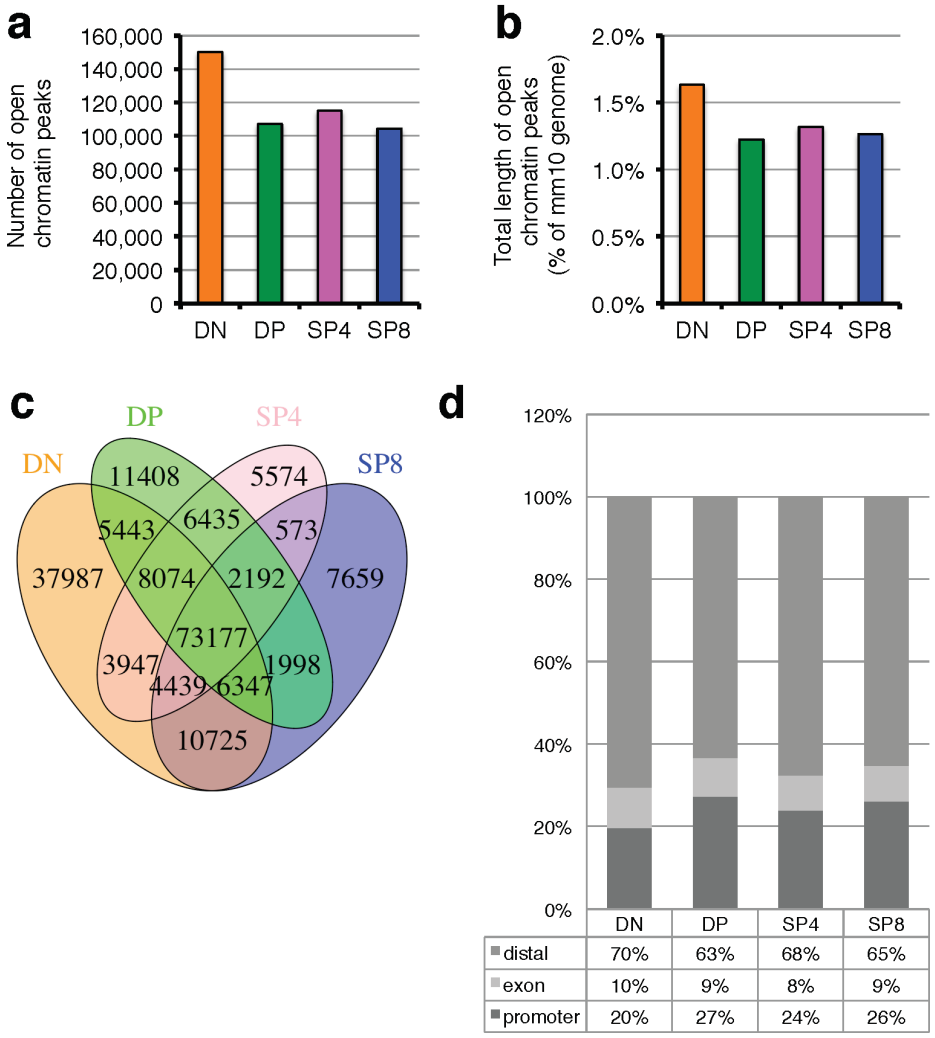


Figure 3.11: Open chromatin regions defined by ATAC-seq peaks. (a) Number of ATAC-seq peaks in DN (orange), DP (green), SP4 (pink) and SP8 (blue) stage thymocytes. (b) Total length of ATAC-seq peaks assigned by MACS2 (narrow peak calling) is shown as a fraction of the whole mouse genome length. (c) The overlap among ATAC-seq peak calls at the four stages examined in this study is shown as a Venn diagram. (d) ATAC-seq peaks that overlap within 200 bp 5' to a gene were characterized as promoters (left). The peaks that overlap with exons, but are not with the promoters, are shown as exons (middle). The peaks that overlap with neither promoters nor exons are characterized as distal (right). The gene annotations were downloaded from the UCSC Table Browser. Annotations for RefSeq genes and UCSC genes are combined.

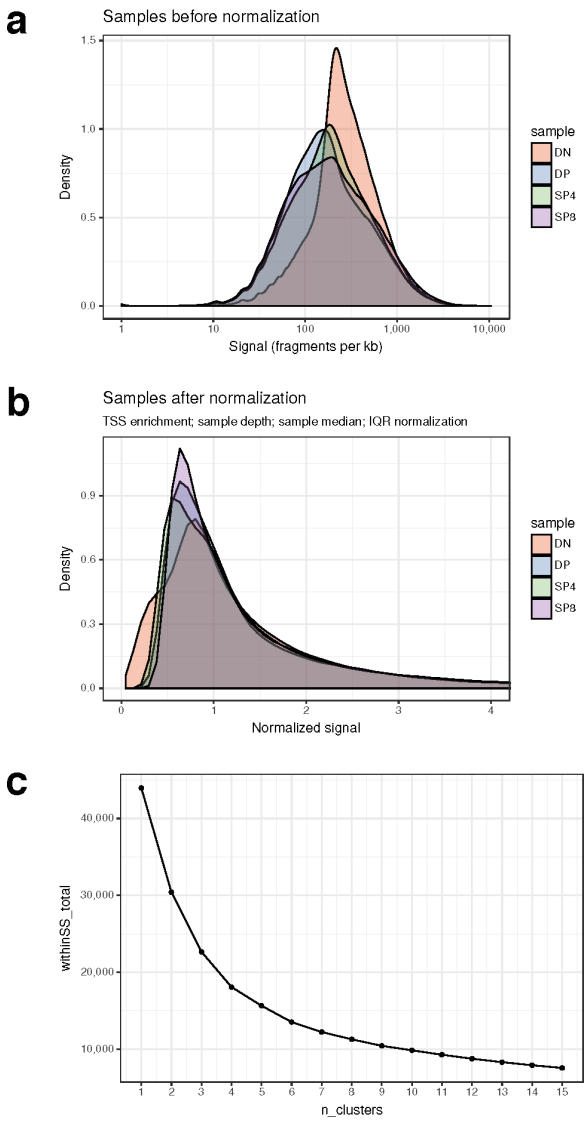


Figure 3.12: Additional information on k-means clustering. Distribution of the ATAC-seq signal in the master peaks before (a) and after (b) normalization. (c) Elbow plot showing variances within clusters for each value of k.

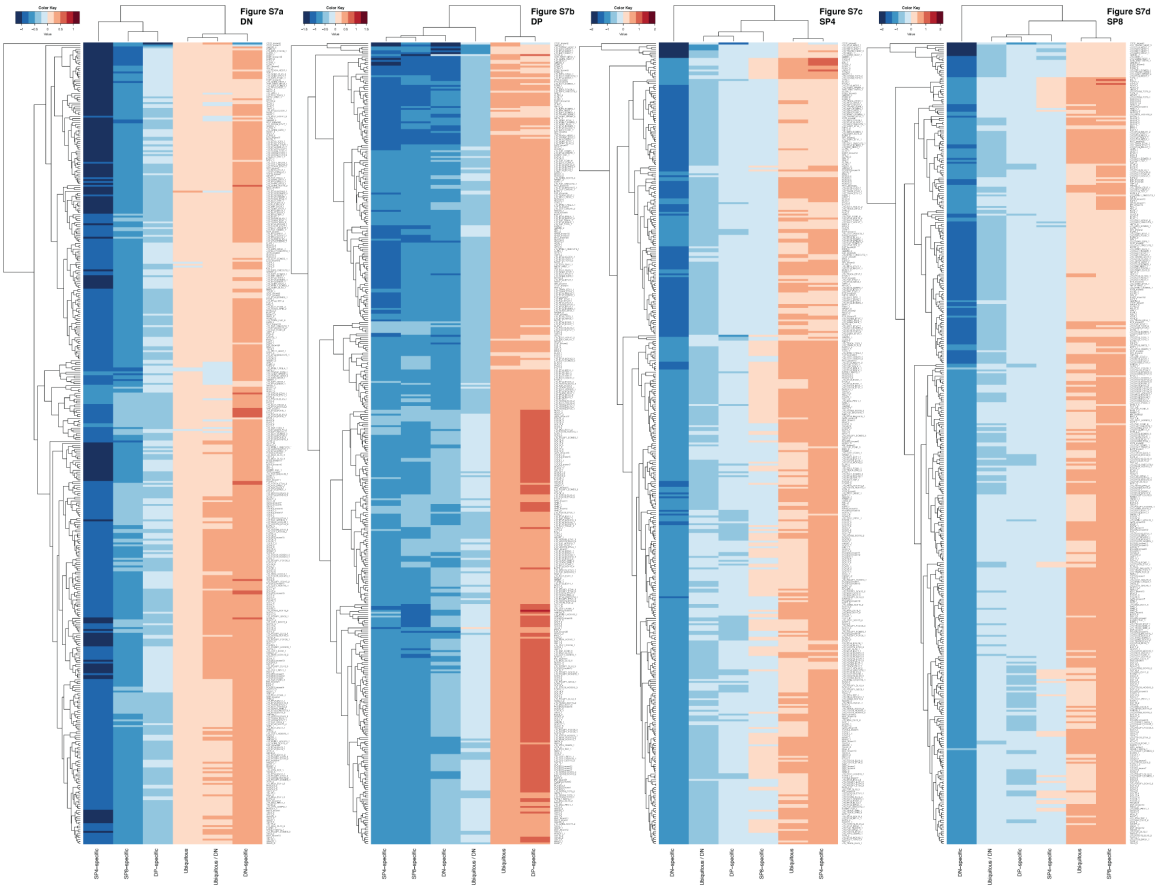


Figure 3.13: Footprint enrichment results from GAT. Heatmaps showing GAT enrichments of each of the motifs for the individual k-means clusters in (a) DN, (b) DP, (c) SP4, and (d) SP8 samples.

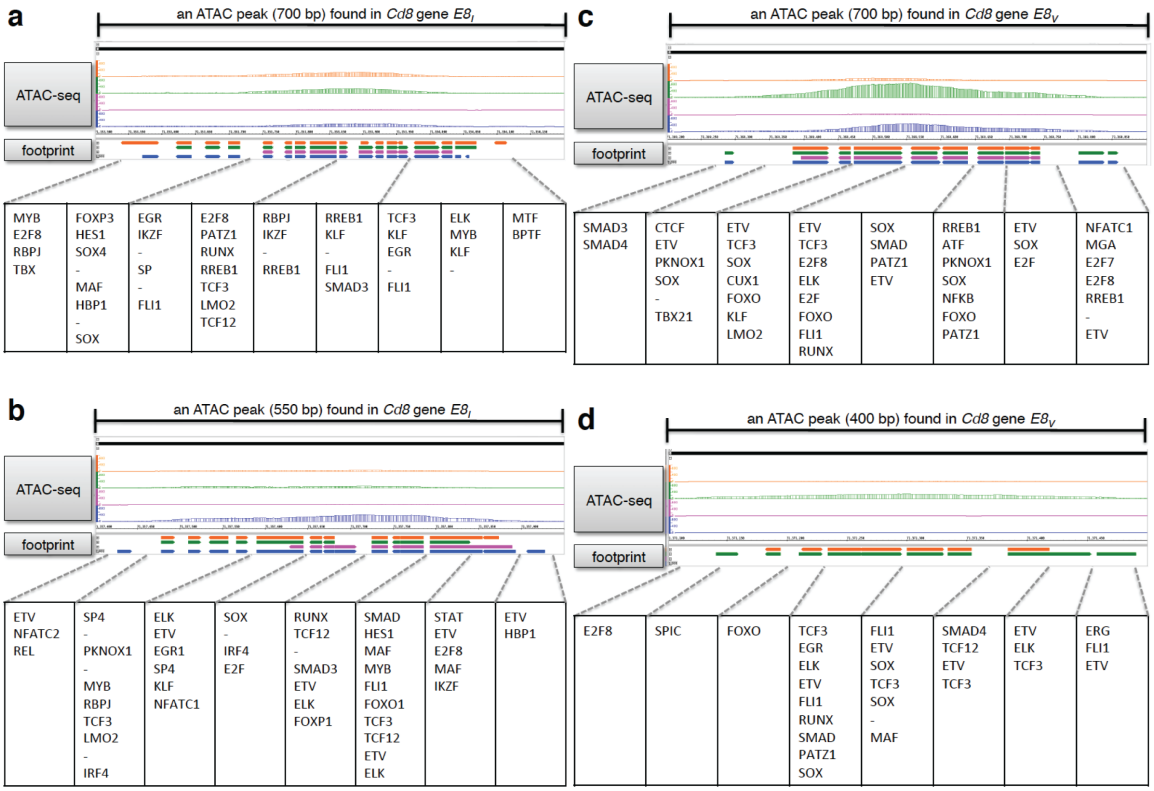


Figure 3.14: CENTIPEDE footprint calls within functionally validated enhancers for the *Cd8* gene. Related to Fig. 3.4. ATAC signals and footprint calls within peaks found in E8I silencer (a,b) and E8V enhancer (c,d). Transgenic reporter analysis showed that expression of the Cd8 genes is regulated by multiple enhancers (E8I, E8II, E8III, E8IV and E8V) [33, 154, 155]. Removal of either E8I/E8II or E8V from the mouse genome results in reduced CD8 expression [84, 156]. In the present analysis, open chromatin regions were identified in the E8 enhancers and exons, but also revealed the presence of possibly novel and previously undetected regulatory elements (Fig. 3.4). The footprint data recapitulated TF binding to IKAROS motifs 15, RUNT motifs [157] and PATZ1 [158] motifs within the known Cd8 enhancers. We conclude that experimental information obtained at the Cd4 and Cd8 loci by ATAC demonstrate agreement between protein binding to previously well characterized enhancers (and one silencer) and our footprint predictions, supporting the hypothesis that TF footprints revealed by ATAC-seq provide a reliable complement to experimental ChIP-qPCR and ChIP-seq data. This highlights the use of ATAC-seq to generate highly informative predictions of TF binding to regulatory elements prior to experimental validation.

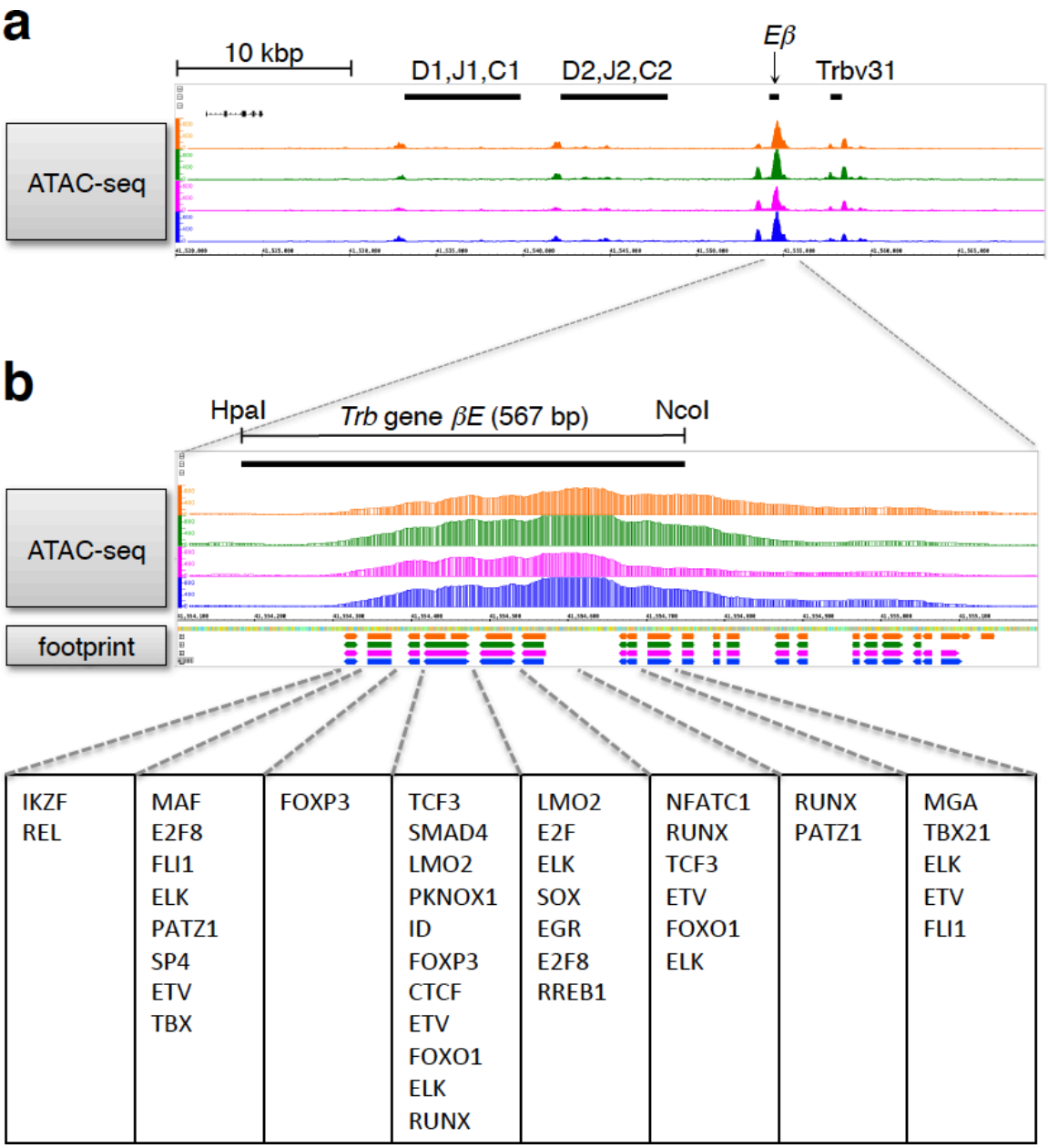


Figure 3.15: ATAC-seq signal and CENTIPEDE footprint calls around functionally validated β E enhancer for the Trb gene. (a) ATAC-seq signals are shown on the IGB browser within around 50 kbp of the Trb locus (encoding TCRβ); mm10, chr6: 41,520,001-41,570,000. Eβ enhancer, DJC and Trbv31 regions are shown at the top. (b) ATAC signals and footprint calls around Eβ are shown. Deletion of Eβ enhancer from mouse genome blocks $\alpha\beta$T cell development [110, 159]

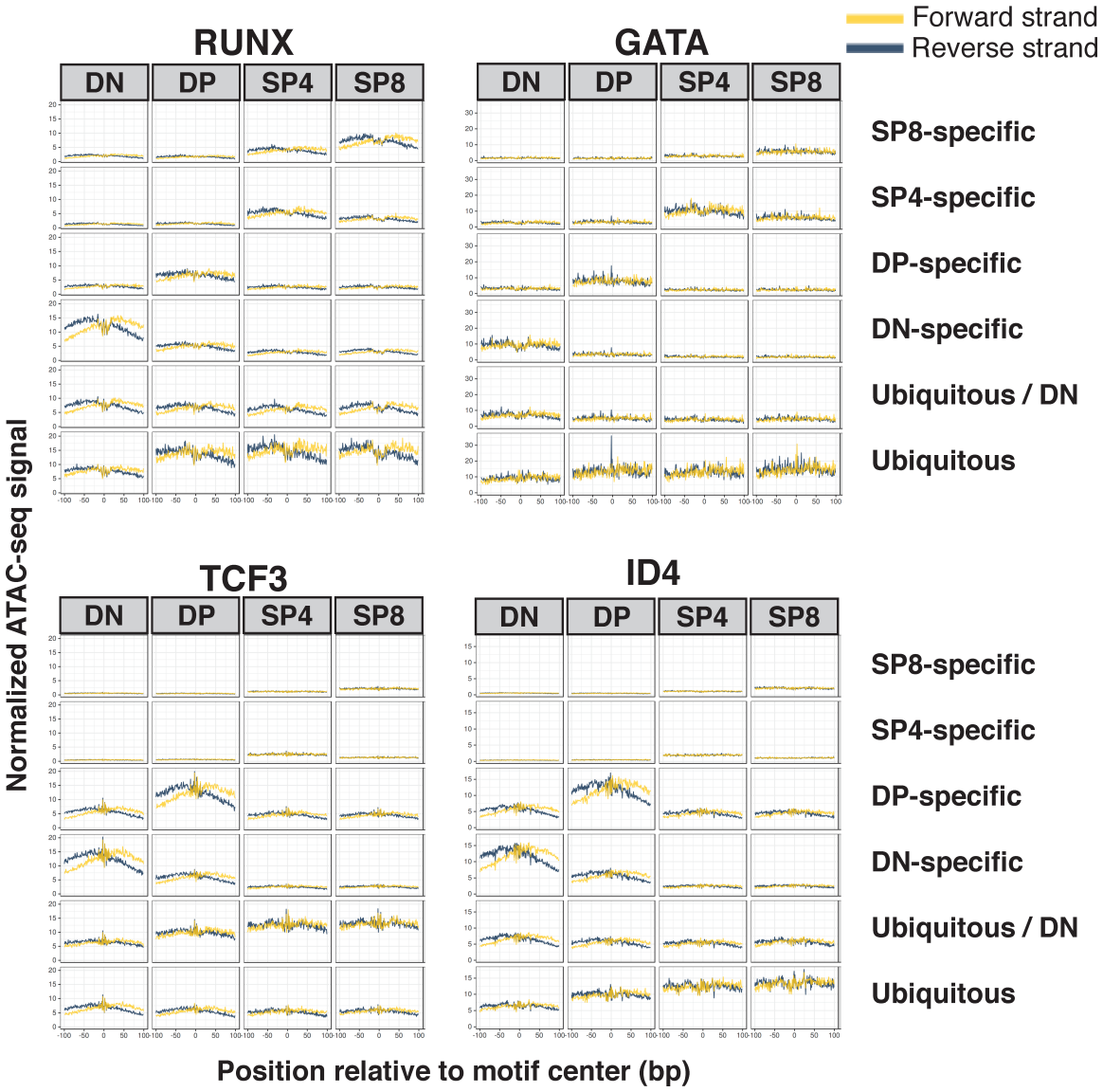


Figure 3.16: Footprint occupancies across samples and clusters. Normalized occupancy signals (see Methods) at ±100 bp of motif center for RUNX1, GATA, TCF3, and ID4. Horizontal facets correspond to the ATAC-seq samples, and vertical facets correspond to the k-means clusters.

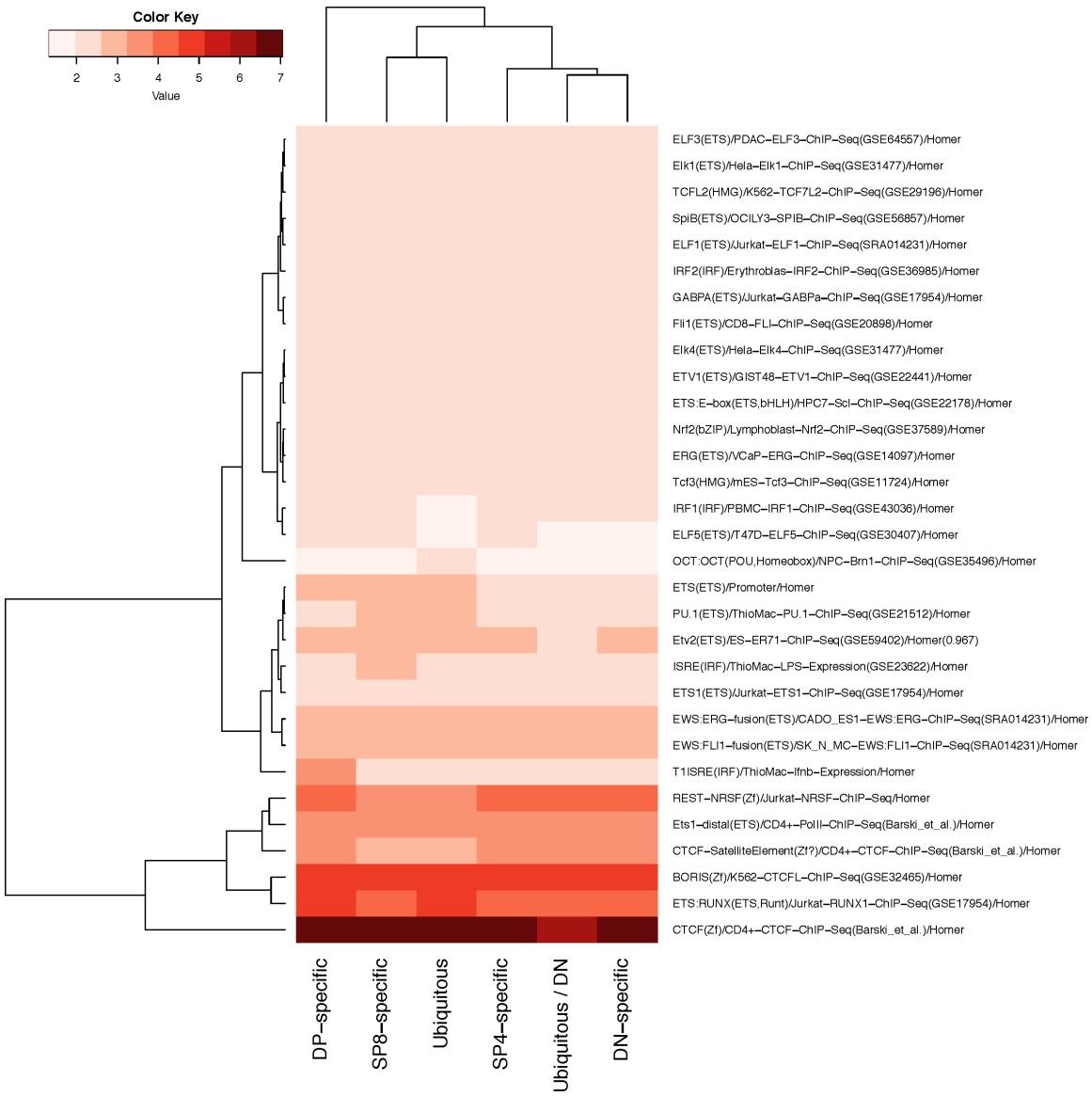


Figure 3.17: HOMER motif enrichment analysis. HOMER known motifs that were called as significant in each of the clusters are plotted with their enrichment values in the color scale.

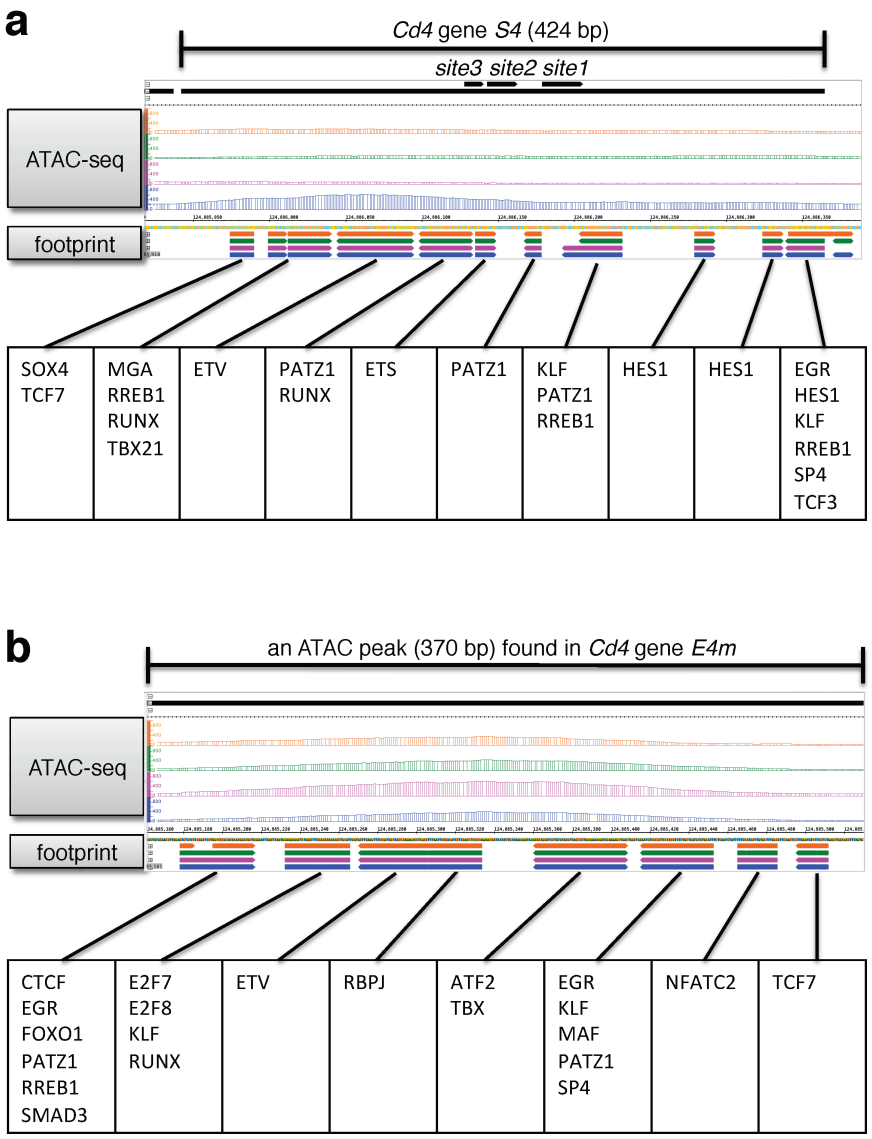


Figure 3.18: CENTIPEDE footprint calls within functionally validated regulatory elements for the *Cd4* gene. Related to Figure 3.3. ATAC signals and footprint calls around S4 silencer (a) and E4m enhancer (b). Transgenic reporter analyses have shown that Cd4 transcription is regulated by at least two proximal enhancers (E4p and E4m) and a silencer (S4) [146–148]. In mice ablated for E4p only, CD4 expression was reduced in pre-selection DP thymocytes 5 but was completely abrogated when both E4p and E4m were deleted [150]; . Removal of S4 from the mouse genome resulted in ectopic Cd4 expression in DN and SP8 cells, which are both CD4-negative in wild type mice [151–153]. We identified ATAC-seq open chromatin regions in the Cd4 locus E4p, E4m and S4 (Fig. 3.3). The ATAC-peak at E4p belongs to DP-specific cluster shown in Fig. 3.1a, while the ATAC-peak found in E4m belongs to SP4-specific cluster, in keeping with its developmental function [149, 150]. The open chromatin at S4 belongs to SP8-specific cluster, as would be expected from its Cd4 gene silencing function. Minor open chromatin regions were also found approximately 3 kbp 5' to

E4p, as well proximal to the Cd4 1st and 2nd exons. Our digital footprint data in mouse primary thymocytes identified approximately 10 sequences bound by protein within E4p, S4 and E4m, respectively. Based on known and novel protein binding found here, we propose that approximately 10 proteins shown in bind to/around E4p, S4 and E4m in order to contribute to activity of the enhancers/silencer.

CHAPTER IV

Implications and Future Directions

This dissertation is based on the overarching principle that the interactions between TFs and chromatin are fundamental to understand genome regulation and the genetic causes of disease predisposition. The body of work described here emerges from analyses of several modalities of high-throughput molecular profiles with a focus on applying an information theoretical perspective to genomic data. My work aimed to improve our understanding of genome organization and regulation by developing novel computational methods to quantify the interactions between transcription factors and chromatin. During the course of this dissertation, several themes emerged with implications for future studies of genome organization.

4.1 *In vivo* TF-chromatin interaction signatures are dynamic and reflect biophysical and regulatory properties of TFs

Previous studies have determined interaction patterns between TFs and nucleosomes *in vitro* using protein binding microarrays, affinity purification, and optical tweezers [65, 75, 94]. However, these studies were performed using purified nucleosomes and TFs, therefore not taking into account for the complex set of biological factors regulating TF-chromatin interactions in their native cellular environment. Using the information theoretical approach developed during this dissertation, we were

able to determine TF-chromatin interaction patterns by measuring the organization of the chromatin accessibility signals around TF binding sites. This metric, which we called chromatin information, quantifies the underlying local chromatin architecture (nucleosome positioning) from the information content patterns in ATAC-seq fragments. This allowed the determination of *in vivo* TF-chromatin interaction patterns for hundreds of TF motifs, thus enabling the dissection of tissue-specific patterns that cannot be captured in the aforementioned *in vitro* studies.

We present evidence that the TF-chromatin interaction patterns are associated with biophysical aspects of TF biology, such as TF-DNA residence times and specific DNA binding domains. By measuring the local chromatin organization associated with TF binding, we provide for the first time a method to estimate TF-DNA residence times for hundreds of TFs simultaneously using sequencing data. This represents an important technological advance. We hope these tools will be useful for researchers interested in understanding TF biophysics.

Our work shows that TF-chromatin interaction patterns vary widely between TFs, and even between subclasses of the same TFs (as exemplified by the CTCF/cohesin results described in Chapter II). We found that the majority of TFs (up to 90%) do not associate with organized chromatin. We therefore hypothesize that these low chromatin information TFs have less active roles in chromatin organization. Importantly, we find differences in TF-chromatin interaction patterns across tissues and cell types, indicating that different subsets of TFs drive these chromatin information patterns. Interestingly, some of the TF families associated with the highest chromatin information in each tissue are associated with development and include known tissue-specific TF families (Figure 2.1). This suggests that the chromatin information metric we developed here can be used to nominate candidate tissue-specific regulators of chromatin organization. Supporting this, our results show that the subset of TFs associated with organized chromatin are more enriched to overlap regions associated

with the genetic control of gene expression (*cis*-eQTLs), consistent with these high chromatin information TFs having a prominent regulatory role in the cell types in which they are expressed.

4.2 TF-chromatin interaction patterns identify candidate pioneer TFs

Given the enrichment of high chromatin information TF motifs overlapping *cis*-eQTLs compared to other TFs, we hypothesized that high chromatin information TF could act by establishing the conditions for other TFs regulate gene expression. One mechanism for this would be if high chromatin information TFs acted as pioneer TFs.

Pioneer TFs are a special class of TFs that are postulated to bind closed chromatin and induce chromatin accessibility to enable binding of other TFs [71]. While some studies published during the course of this work provided clues about what differentiates pioneers from non-pioneer TFs [73, 74], determining which TFs act as pioneers in any given tissue remains an open question. During this work, we sought to tackle this question by using a genetics-based approach to estimate chromatin information at heterozygous loci. Our results showed that genetic variants associated with higher chromatin accessibility were more likely to form binding sites for TFs associated with organized chromatin. This result is consistent with high chromatin information TFs acting as pioneer TFs. Supporting this, motifs associated with known pioneer TFs such as OCT, SOX, KLF, and FOXA2, are among the highest in chromatin information. While it will be necessary to experimentally validate this claim, our chromatin information framework provides an approach to determine candidate pioneer TFs in a given biological sample. Importantly, this methodology does not rely on time-course or *in vitro* TF-nucleosome dynamics experiments, which demand more resources. An exciting future direction from this work is to determine how the repertoire of these

putative pioneer TFs vary as a function of development or in response to experimental perturbations. Addressing this question has the potential to increase our understanding of genome regulation.

4.3 TF binding prediction methods are affected by TF-chromatin interactions

Determining genome-wide TF binding sites is critical to understand gene regulatory networks. TF ChIP-seq experiments, however, are expensive and *a priori* knowledge of the TF(s) to be profiled. Therefore, a viable alternative is to predict TF binding using chromatin accessibility data. In Chapter III of this dissertation, we demonstrated the use of CENTIPEDE [49], a TF binding prediction method, to characterize the dynamics of TF binding during thymocyte development. By finding known and potentially novel TFs involved in thymocyte development, we present a proof-of-concept that TF binding prediction algorithms can be used to dissect the complex dynamics associated with biological processes.

The most widely used methods to predict TF binding from chromatin accessibility data rely on the presumed local protection the TF confers to DNA cleavage (TF footprints). Previous studies, however, suggested that not all TFs associate with footprints [52, 55]. As part of this dissertation, we evaluated TF footprint-based binding prediction algorithms and determined that they are sensitive to the local chromatin architecture associated with the TF-chromatin interaction. Footprinting-based methods had high prediction power of TFs associated with highly organized chromatin, such as CTCF, but were less accurate for TFs that did not associate with organized chromatin. BMO, the method we developed in this study, is less sensitive to local TF-chromatin interaction patterns. BMO outperformed footprinting-based algorithms in the majority of TFs. Using BMO, we were able to calculate the TF-chromatin in-

teraction patterns for hundreds of TFs. Our results show that the majority TFs do not associate with organized chromatin. Most TFs, therefore, are not suitable for footprint-based algorithms. This represents an important advance in the field of genomics, as not only do we demonstrate that the current framework of TF binding prediction is limited, but we also provide an alternative to circumvent these limitations. We hope that the use of BMO or other non footprint-based algorithms will become the standard for TF binding predictions. An important future direction for this work is to develop a statistical framework that allow to use BMO across multiple replicates. This will be advantageous for studies analyzing cohorts where large numbers of chromatin accessibility profiles are available.

4.4 Chromatin information as a novel metric of ATAC-seq quality control

One of the most important steps in analyzing any high-throughput molecular profile is assessing data quality, which requires well-established quality control (QC) metrics. Because of its relatively young age at the start of this work, ATAC-seq had a limited set of QC tools available. In particular, we felt the need for a tool that could compare multiple samples simultaneously and provide metrics to quantify relevant differences between samples that could affect downstream comparisons (*e.g.* fragment length distribution and signal-to-noise ratio). To address this issue, our group developed ataqv (`https://github.com/ParkerLab/ataqv`), a QC tool for ATAC-seq data. Ataqv which was used extensively during this work. The metrics analyzed by ataqv were sufficient for broadly categorizing samples as higher or lower quality. However, ataqv did not assess the finer information content patterns indicative of TF-chromatin interactions. As part of this work, we leveraged the highly ordered chromatin architecture around constitutive and evolutionarily conserved CTCF-cohesin binding sites to

provide a known quantity (a "standard candle", in astronomical jargon) of chromatin organization to be used as a QC metric. By generating a V-plot of the ATAC-seq signal around these ubiquitous CTCF-cohesin binding sites, it is possible to obtain important information regarding sample quality from a single plot. This approach is complementary to the other QC metrics provided by ataqv, as it provides an intuitive way to evaluate multiple samples (see Figure 2.5 as an example). The CTCF-cohesin V-plot visually informs metrics such as signal-to-noise ratio and ATAC-seq fragment size distribution. In addition, it can also be analyzed quantitatively by calculating the correlation of the chromatin information between the sample of interest and a reference high-quality sample. We expect that including the CTCF-cohesin V-plots as part of the ataqv package will help other researchers more easily assess the quality of their ATAC-seq experiments.

4.5 Concluding remarks

The work performed during this dissertation highlights the necessity of interdisciplinary approaches to analyze biological data. By applying techniques from the information theory and signal processing fields to high-throughput molecular profiles, we obtained a unique perspective on the organization and regulation of the human genome. We demonstrate that applying information theoretical principles to chromatin accessibility data allows for a powerful readout of different aspects related to genome organization and TF biology. These aspects include nucleosome positioning and TF-DNA residence times. The methodology developed here provides the groundwork for future studies that aim to characterize TFs in their native biological context, as well as measure the effect of biological perturbations on the TF-chromatin interaction landscape.

During the course of this dissertation, my work supported studies which aimed to determine the molecular mechanisms underlying T2D GWAS variants in human skele-

tal muscle [33] and pancreatic islets [84]. These studies advanced our understanding of T2D by not only nominating the causal variants and identifying the respective effector TFs in these tissues, but also proposing a novel mechanism for T2D etiology in pancreatic islets involving independent genetic variants potentially acting in confluence to disrupt binding sites of RFX6. RFX6 is a key pancreatic islet TF. In addition, the work developed here helped support other studies characterizing the understanding T2D pathophysiology [181–183].

The expected decrease in sequencing costs will enable larger chromatin accessibility datasets across large genetically diverse cohorts, which will allow dissection of the effects of genetic variation on modulating chromatin organization patterns. By integrating these chromatin accessibility profiles with existing 3-dimensional chromatin organization reference datasets, it will be possible to determine how genetic perturbations affect the chromatin information patterns locally and within higher-order chromatin domains (Figure 4.1A). We hypothesize that some genetic variants can act by disrupting chromatin organization, therefore increasing or decreasing the entropy levels at the chromatin domain level (chromatin information quantitative trait loci – ciQTLs). This could provide a novel mechanism by which genetic variation affects disease predisposition.

Another exciting future direction is the use of single-nuclei resolution molecular profiles (snATAC-seq and snRNA-seq) as a platform for genomic information theory tool development. This will enable the dissection of genome organization across the individual cell types that form bulk tissue samples, which will not only increase our understanding of the underlying tissue biology, but also allow the identification of relevant cell types associated with disease predisposition. Single-nuclei assays are ideal for the inference of tissue-specific co-accessibility [36] and can potentially be used for co-expression [184] networks. By applying these methodologies to groups of nuclei representing the different cell types in each sample, it will be possible to

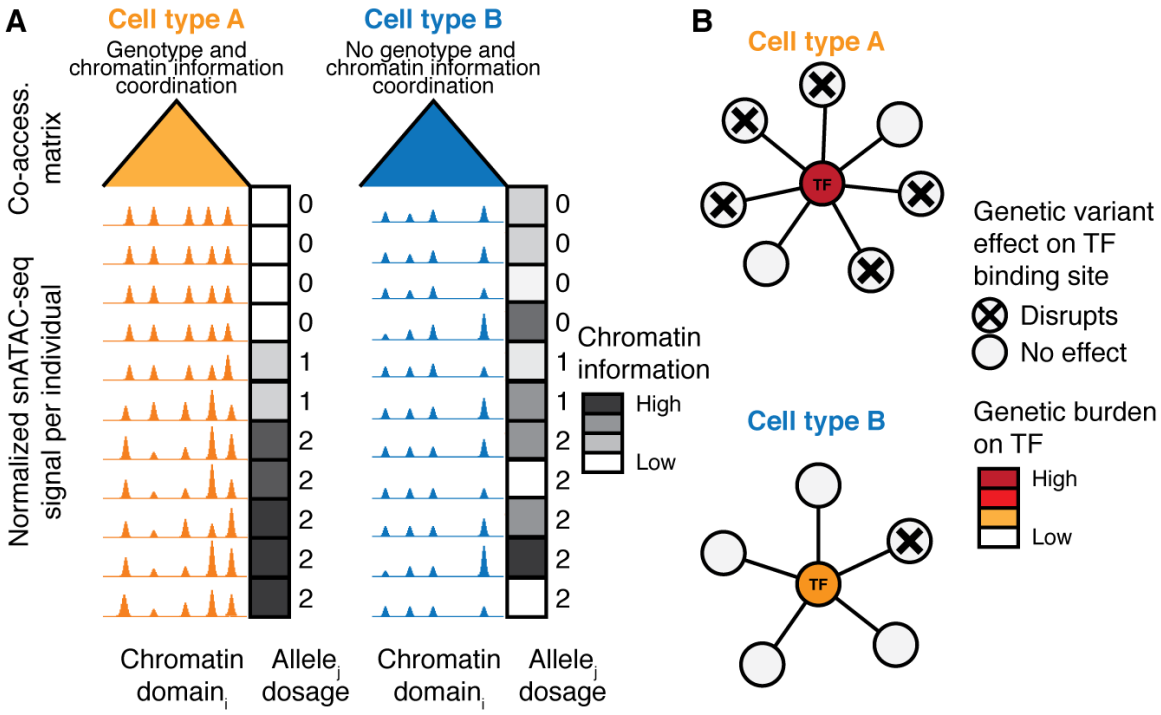


Figure 4.1: Future directions. (A) Chromatin information quantitative trait loci (ciQTLs) analyses across cell types. (B) Calculating the impact of disease-associated genetic variants on gene regulatory networks inferred from co-expression and co-accessibility data.

decrease the noise associated with analyzing bulk samples. The use of graph theory to analyze these networks will provide powerful tools to dissect the molecular pathways associated with disease predisposition in each cell type (Figure 4.1B). Together, these technological and analytical advances will help improve our understanding of the human genome, with implications for biomedical research.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Piovesan, A., Pelleri, M. C., Antonaros, F., Strippoli, P., Caracausi, M., and Vitale, L. (December, 2019) On the length, weight and GC content of the human genome. *BMC Research Notes,* **12**(1), 106.

[2] Cooper, G. M. (2000) The cell: a molecular approach, ASM Press [u.a.], Washington, DC 2. ed edition OCLC: 247541326.

[3] Chen, T. and Dent, S. Y. R. (February, 2014) Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nature Reviews. Genetics,* **15**(2), 93–106.

[4] Allshire, R. C. and Madhani, H. D. (2018) Ten principles of heterochromatin formation and function. *Nature Reviews. Molecular Cell Biology,* **19**(4), 229–244.

[5] Tessarz, P. and Kouzarides, T. (November, 2014) Histone core modifications regulating nucleosome structure and dynamics. *Nature Reviews. Molecular Cell Biology,* **15**(11), 703–708.

[6] Gibson, B. A., Doolittle, L. K., Schneider, M. W., Jensen, L. E., Gamarra, N., Henry, L., Gerlich, D. W., Redding, S., and Rosen, M. K. (October, 2019) Organization of Chromatin by Intrinsic and Regulated Phase Separation. *Cell,* **179**(2), 470–484.e21.

[7] Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blthgen, N., Dekker, J., and Heard, E. (May, 2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature,* **485**(7398), 381–385.

[8] Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., and Aiden, E. L. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell,* **159**(7), 1665–1680.

[9] Haberle, V. and Stark, A. (October, 2018) Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology,* **19**(10), 621–637.

[10] Veyrieras, J.-B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M., and Pritchard, J. K. (October, 2008) High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation. *PLoS Genetics,* **4**(10), e1000214.

[11] Kumasaka, N., Knights, A. J., and Gaffney, D. J. (January, 2019) High-resolution genetic mapping of putative causal interactions between regions of open chromatin. *Nature Genetics,* **51**(1), 128–137.

[12] Lettice, L. A., Heaney, S. J. H., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E., and de Graaff, E. (July, 2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics,* **12**(14), 1725–1735.

[13] Monahan, K., Horta, A., and Lomvardas, S. (January, 2019) LHX2- and LDB1-mediated trans interactions regulate olfactory receptor choice. *Nature,* **565**(7740), 448–453.

[14] Long, H. K., Prescott, S. L., and Wysocka, J. (2016) Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell,* **167**(5), 1170–1187.

[15] Simeonov, D. R., Gowen, B. G., Boontanrart, M., Roth, T. L., Gagnon, J. D., Mumbach, M. R., Satpathy, A. T., Lee, Y., Bray, N. L., Chan, A. Y., Lituiev, D. S., Nguyen, M. L., Gate, R. E., Subramaniam, M., Li, Z., Woo, J. M., Mitros, T., Ray, G. J., Curie, G. L., Naddaf, N., Chu, J. S., Ma, H., Boyer, E., Van Gool, F., Huang, H., Liu, R., Tobin, V. R., Schumann, K., Daly, M. J., Farh, K. K., Ansel, K. M., Ye, C. J., Greenleaf, W. J., Anderson, M. S., Bluestone, J. A., Chang, H. Y., Corn, J. E., and Marson, A. (2017) Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature,* **549**(7670), 111–115.

[16] Calderon, D., Nguyen, M. L. T., Mezger, A., Kathiria, A., Mller, F., Nguyen, V., Lescano, N., Wu, B., Trombetta, J., Ribado, J. V., Knowles, D. A., Gao, Z., Blaeschke, F., Parent, A. V., Burt, T. D., Anderson, M. S., Criswell, L. A., Greenleaf, W. J., Marson, A., and Pritchard, J. K. (October, 2019) Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nature Genetics,* **51**(10), 1494–1505.

[17] Furlong, E. E. M. and Levine, M. (2018) Developmental enhancers and chromosome topology. *Science (New York, N.Y.),* **361**(6409), 1341–1345.

[18] Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K. D., Plajzer-Frick, I., Akiyama, J., De Val, S., Afzal, V., Black, B. L., Couronne, O., Eisen, M. B., Visel, A., and Rubin, E. M. (November, 2006) In vivo enhancer analysis of human conserved non-coding sequences. *Nature,* **444**(7118), 499–502.

[19] Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L. A. (January, 2007) VISTA Enhancer Browser–a database of tissue-specific human enhancers. *Nucleic Acids Research,* **35**(Database), D88–D92.

[20] Crawford, G. E., Holt, I. E., Whittle, J., Webb, B. D., Tai, D., Davis, S., Margulies, E. H., Chen, Y., Bernat, J. A., Ginsburg, D., Zhou, D., Luo, S., Vasicek, T. J., Daly, M. J., Wolfsberg, T. G., and Collins, F. S. (January, 2006) Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Research,* **16**(1), 123–131.

[21] Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods,* **10**(12), 1213–1218.

[22] Arnold, C. D., Gerlach, D., Stelzer, C., Bory, . M., Rath, M., and Stark, A. (March, 2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science (New York, N.Y.),* **339**(6123), 1074–1077.

[23] Meyer, C. A. and Liu, X. S. (November, 2014) Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews Genetics,* **15**(11), 709–721.

[24] The ENCODE Project Consortium (September, 2012) An integrated encyclopedia of DNA elements in the human genome. *Nature,* **489**(7414), 57–74.

[25] Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., Farnham, P. J., Hirst, M., Lander, E. S., Mikkelsen, T. S., and Thomson, J. A. (October, 2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology,* **28**(10), 1045–1048.

[26] Ernst, J. and Kellis, M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods,* **9**(3), 215–216.

[27] Parker, S. C. J., Stitzel, M. L., Taylor, D. L., Orozco, J. M., Erdos, M. R., Akiyama, J. A., van Bueren, K. L., Chines, P. S., Narisu, N., Black, B. L., Visel, A., Pennacchio, L. A., Collins, F. S., Becker, J., Benjamin, B., Blakesley, R., Bouffard, G., Brooks, S., Coleman, H., Dekhtyar, M., Gregory, M., Guan, X., Gupta, J., Han, J., Hargrove, A., Ho, S.-l., Johnson, T., Legaspi, R., Lovett, S., Maduro, Q., Masiello, C., Maskeri, B., McDowell, J., Montemayor, C., Mullikin, J., Park, M., Riebow, N., Schandler, K., Schmidt, B., Sison, C., Stantripop, M., Thomas, J., Thomas, P., Vemulapalli, M., and Young, A. (2013) Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences,* **110**(44), 17921–17926.

[28] Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., and Kellis, M. (February, 2015) Integrative analysis of 111 reference human epigenomes. *Nature,* **518**(7539), 317–330.

[29] Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (October, 2009) Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science,* **326**(5950), 289–293.

[30] Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S. W., Andrews, S., Grey, W., Ewels, P. A., Herman, B., Happe, S., Higgs, A., LeProust, E., Follows, G. A., Fraser, P., Luscombe, N. M., and Osborne, C. S. (June, 2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics,* **47**(6), 598–606.

[31] Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., Ingle, C. E., Dunning, M., Flicek, P., Koller, D., Montgomery, S., Tavar, S., Deloukas, P., and Dermitzakis, E. T. (October, 2007) Population genomics of human gene expression. *Nature Genetics,* **39**(10), 1217–1224.

[32] Lappalainen, T., Sammeth, M., Friedlnder, M. R., t Hoen, P. A. C., Monlong, J., Rivas, M. A., Gonzlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlf, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., Tikhonov, A., Sultan, M., Bertier, G., MacArthur, D. G., Lek, M., Lizano, E., Buermans, H. P. J., Padioleau, I., Schwarzmayr, T., Karlberg, O., Ongen, H., Kilpinen, H., Beltran, S., Gut, M., Kahlem, K., Amstislavskiy, V., Stegle, O., Pirinen, M., Montgomery, S. B.,

Donnelly, P., McCarthy, M. I., Flicek, P., Strom, T. M., The Geuvadis Consortium, Lehrach, H., Schreiber, S., Sudbrak, R., Carracedo, ., Antonarakis, S. E., Hsler, R., Syvnen, A.-C., van Ommen, G.-J., Brazma, A., Meitinger, T., Rosenstiel, P., Guig, R., Gut, I. G., Estivill, X., and Dermitzakis, E. T. (September, 2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature,* **501**(7468), 506–511.

[33] Scott, L. J., Erdos, M. R., Huyghe, J. R., Welch, R. P., Beck, A. T., Wolford, B. N., Chines, P. S., Didion, J. P., Narisu, N., Stringham, H. M., Taylor, D. L., Jackson, A. U., Vadlamudi, S., Bonnycastle, L. L., Kinnunen, L., Saramies, J., Sundvall, J., Albanus, R. D. O., Kiseleva, A., Hensley, J., Crawford, G. E., Jiang, H., Wen, X., Watanabe, R. M., Lakka, T. A., Mohlke, K. L., Laakso, M., Tuomilehto, J., Koistinen, H. A., Boehnke, M., Collins, F. S., and Parker, S. C. J. (2016) The genetic regulatory signature of type 2 diabetes in human skeletal muscle. *Nature Communications,* **7**.

[34] GTEx Consortium (October, 2017) Genetic effects on gene expression across human tissues. *Nature,* **550**(7675), 204–213.

[35] Rubin, A. J., Barajas, B. C., Furlan-Magaril, M., Lopez-Pajares, V., Mumbach, M. R., Howard, I., Kim, D. S., Boxer, L. D., Cairns, J., Spivakov, M., Wingett, S. W., Shi, M., Zhao, Z., Greenleaf, W. J., Kundaje, A., Snyder, M., Chang, H. Y., Fraser, P., and Khavari, P. A. (2017) Lineage-specific dynamic and pre-established enhancer-promoter contacts cooperate in terminal differentiation. *Nature Genetics,* **49**(10), 1522–1528.

[36] Pliner, H. A., Packer, J. S., McFaline-Figueroa, J. L., Cusanovich, D. A., Daza, R. M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., Adey, A. C., Steemers, F. J., Shendure, J., and Trapnell, C. (2018) Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Molecular Cell,* **71**(5), 858–871.e8.

[37] Ulirsch, J. C., Lareau, C. A., Bao, E. L., Ludwig, L. S., Guo, M. H., Benner, C., Satpathy, A. T., Kartha, V. K., Salem, R. M., Hirschhorn, J. N., Finucane, H. K., Aryee, M. J., Buenrostro, J. D., and Sankaran, V. G. (April, 2019) Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nature Genetics,* **51**(4), 683–693.

[38] Lupiez, D., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., Spielmann, M., Visel, A., and Mundlos, S. (May, 2015) Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell,* **161**(5), 1012–1025.

[39] Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., and Bergmann, S. (2016) Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nature Methods,* **13**(4), 366–370.

[40] GTEx Consortium, Gamazon, E. R., Segr, A. V., van de Bunt, M., Wen, X., Xi, H. S., Hormozdiari, F., Ongen, H., Konkashbaev, A., Derks, E. M., Aguet, F., Quan, J., Nicolae, D. L., Eskin, E., Kellis, M., Getz, G., McCarthy, M. I., Dermitzakis, E. T., Cox, N. J., and Ardlie, K. G. (July, 2018) Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nature Genetics,* **50**(7), 956–967.

[41] Ortega, E., Rengachari, S., Ibrahim, Z., Hoghoughi, N., Gaucher, J., Holehouse, A. S., Khochbin, S., and Panne, D. (2018) Transcription factor dimerization activates the p300 acetyltransferase. *Nature,* **562**(7728), 538–544.

[42] Ohkuma, Y., Horikoshi, M., Roeder, R. G., and Desplan, C. (March, 1990) Engrailed, a homeodomain protein, can repress in vitro transcription by competition with the TATA box-binding protein transcription factor IID. *Proceedings of the National Academy of Sciences of the United States of America,* **87**(6), 2289–2293.

[43] Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (July, 2015) The MEME Suite. *Nucleic Acids Research,* **43**(W1), W39–W49.

[44] Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J. M., Vincentelli, R., Luscombe, N. M., Hughes, T. R., Lemaire, P., Ukkonen, E., Kivioja, T., and Taipale, J. (2013) DNA-binding specificities of human transcription factors. *Cell,* **152**(1-2), 327–339.

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

[46] Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., and Weirauch, M. T. (2018) The Human Transcription Factors. *Cell,* **172**(4), 650–665.

[47] Fornes, O., Castro-Mondragon, J. A., Khan, A., vanderLee, R., Zhang, X., Richmond, P. A., Modi, B. P., Correard, S., Gheorghe, M., Baranai, D., Santana-Garcia, W., Tan, G., Chneby, J., Ballester, B., Parcy, F., Sandelin, A., Lenhard, B., Wasserman, W. W., and Mathelier, A. (November, 2019) JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research,* p. gkz1001.

[48] Grant, C. E., Bailey, T. L., and Noble, W. S. (April, 2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics (Oxford, England),* **27**(7), 1017–1018.

[49] Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., and Pritchard, J. K. (2011) Accurate inference of transcription factor binding from

DNA sequence and chromatin accessibility data. *Genome Research,* **21**(3), 447–455.

[50] Castro-Mondragon, J. A., Jaeger, S., Thieffry, D., Thomas-Chollier, M., and Van Helden, J. (2017) RSAT matrix-clustering: Dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Research,* **45**(13), 1–13.

[51] Sherwood, R. I., Hashimoto, T., O'Donnell, C. W., Lewis, S., Barkal, A. A., Van Hoff, J. P., Karun, V., Jaakkola, T., and Gifford, D. K. (2014) Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature Biotechnology,* **32**(2), 171–178.

[52] Sung, M. H., Guertin, M. J., Baek, S., and Hager, G. L. (2014) DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Molecular Cell,* **56**(2), 275–285.

[53] Gusmao, E. G., Allhoff, M., Zenke, M., and Costa, I. G. (2016) Analysis of computational footprinting methods for DNase sequencing experiments. *Nature Methods,* **13**(4), 303–309.

[54] Li, Z., Schulz, M. H., Look, T., Begemann, M., Zenke, M., and Costa, I. G. (February, 2019) Identification of transcription factor binding sites using ATAC-seq. *Genome Biology,* **20**(1), 45.

[55] Baek, S., Goldstein, I., and Hager, G. L. (2017) Bivariate Genomic Footprinting Detects Changes in Transcription Factor Activity. *Cell Reports,* **19**(8), 1710–1722.

[56] He, H. H., Meyer, C. A., Hu, S. S., Chen, M. W., Zang, C., Liu, Y., Rao, P. K., Fei, T., Xu, H., Long, H., Liu, X. S., and Brown, M. (2014) Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nature Methods,* **11**(1), 73–78.

[57] Yardimci, G. G., Frank, C. L., Crawford, G. E., and Ohler, U. (October, 2014) Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. *Nucleic Acids Research,* **42**(19), 11865–11878.

[58] Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., Pierce, B. G., Dong, X., Kundaje, A., Cheng, Y., Rando, O. J., Birney, E., Myers, R. M., Noble, W. S., Snyder, M., and Weng, Z. (September, 2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Research,* **22**(9), 1798–1812.

[59] Gaffney, D. J., McVicker, G., Pai, A. A., Fondufe-Mittendorf, Y. N., Lewellen, N., Michelini, K., Widom, J., Gilad, Y., and Pritchard, J. K. (November, 2012) Controls of Nucleosome Positioning in the Human Genome. *PLOS Genetics,* **8**(11), e1003036.

[60] Fu, Y., Sinha, M., Peterson, C. L., and Weng, Z. (July, 2008) The Insulator Binding Protein CTCF Positions 20 Nucleosomes around Its Binding Sites across the Human Genome. *PLoS Genetics,* **4**(7), e1000138.

[61] Clarkson, C. T., Deeks, E. A., Samarista, R., Mamayusupova, H., Zhurkin, V. B., and Teif, V. B. (December, 2019) CTCF-dependent chromatin boundaries formed by asymmetric nucleosome arrays with decreased linker length. *Nucleic Acids Research,* **47**(21), 11181–11196.

[62] Valouev, A., Johnson, S. M., Boyd, S. D., Smith, C. L., Fire, A. Z., and Sidow, A. (June, 2011) Determinants of nucleosome organization in primary human cells. *Nature,* **474**(7352), 516–520.

[63] Kubik, S., Bruzzone, M. J., Jacquet, P., Falcone, J. L., Rougemont, J., and Shore, D. (2015) Nucleosome Stability Distinguishes Two Different Promoter Types at All Protein-Coding Genes in Yeast. *Molecular Cell,* **60**(3), 422–434.

[64] Luo, Y., North, J. A., Rose, S. D., and Poirier, M. G. (March, 2014) Nucleosomes accelerate transcription factor dissociation. *Nucleic Acids Research,* **42**(5), 3017–3027.

[65] Zhu, F., Farnung, L., Kaasinen, E., Sahu, B., Yin, Y., Wei, B., Dodonova, S. O., Nitta, K. R., Morgunova, E., Taipale, M., Cramer, P., and Taipale, J. (October, 2018) The interaction landscape between transcription factors and the nucleosome. *Nature,* **562**(7725), 76–81.

[66] Mueller, F., Stasevich, T. J., Mazza, D., and McNally, J. G. (2013) Quantifying transcription factor kinetics: At work or at play?. *Critical Reviews in Biochemistry and Molecular Biology,* **48**(5), 492–514.

[67] Lickwar, C. R., Mueller, F., Hanlon, S. E., McNally, J. G., and Lieb, J. D. (2012) Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature,* **484**(7393), 251–255.

[68] Loffreda, A., Jacchetti, E., Antunes, S., Rainone, P., Daniele, T., Morisaki, T., Bianchi, M. E., Tacchetti, C., and Mazza, D. (2017) Live-cell p53 single-molecule binding is modulated by C-terminal acetylation and correlates with transcriptional activity. *Nature Communications,* **8**(1).

[69] Claus, K., Popp, A. P., Schulze, L., Hettich, J., Reisser, M., EscoterTorres, L., Uhlenhaut, N. H., and Gebhardt, J. (November, 2017) DNA residence time is a regulatory factor of transcription repression. *Nucleic Acids Research,* **45**(19), 11121–11130.

[70] Malnou, C. E., Brockly, F., Favard, C., Moquet-Torcy, G., Piechaczyk, M., and Jariel-Encontre, I. (2010) Heterodimerization with different jun proteins controls c-Fos intranuclear dynamics and distribution. *Journal of Biological Chemistry,* **285**(9), 6552–6562.

[71] Zaret, K. S. and Carroll, J. S. (November, 2011) Pioneer transcription factors: establishing competence for gene expression. *Genes & Development,* **25**(21), 2227–2241.

[72] Magnani, L., Eeckhoute, J., and Lupien, M. (November, 2011) Pioneer factors: directing transcriptional regulators within the chromatin environment. *Trends in Genetics,* **27**(11), 465–474.

[73] Soufi, A., Garcia, M., Jaroszewicz, A., Osman, N., Pellegrini, M., and Zaret, K. (April, 2015) Pioneer Transcription Factors Target Partial DNA Motifs on Nucleosomes to Initiate Reprogramming. *Cell,* **161**(3), 555–568.

[74] Fernandez Garcia, M., Moore, C. D., Schulz, K. N., Alberto, O., Donague, G., Harrison, M. M., Zhu, H., and Zaret, K. S. (June, 2019) Structural Features of Transcription Factors Associating with Nucleosome Binding. *Molecular Cell,*.

[75] Li, S., Zheng, E. B., Zhao, L., and Liu, S. (September, 2019) Nonreciprocal and Conditional Cooperativity Directs the Pioneer Activity of Pluripotency Transcription Factors. *Cell Reports,* **28**(10), 2689–2703.e4.

[76] Paakinaho, V., Johnson, T. A., Presman, D. M., and Hager, G. L. (August, 2019) Glucocorticoid receptor quaternary structure drives chromatin occupancy and transcriptional outcome. *Genome Research,* **29**(8), 1223–1234.

[77] Fuchsberger, C., Flannick, J., Teslovich, T. M., Mahajan, A., Agarwala, V., Gaulton, K. J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D. J., Rivas, M. A., Perry, J. R. B., Sim, X., Blackwell, T. W., Robertson, N. R., Rayner, N. W., Cingolani, P., Locke, A. E., Tajes, J. F., Highland, H. M., Dupuis, J., Chines, P. S., Lindgren, C. M., Hartl, C., Jackson, A. U., Chen, H., Huyghe, J. R., Van De Bunt, M., Pearson, R. D., Kumar, A., Mller-Nurasyid, M., Grarup, N., Stringham, H. M., Gamazon, E. R., Lee, J., Chen, Y., Scott, R. A., Below, J. E., Chen, P., Huang, J., Go, M. J., Stitzel, M. L., Pasko, D., Parker, S. C. J., Varga, T. V., Green, T., Beer, N. L., Day-Williams, A. G., Ferreira, T., Fingerlin, T., Horikoshi, M., Hu, C., Huh, I., Ikram, M. K., Kim, B. J., Kim, Y., Kim, Y. J., Kwon, M. S., Lee, J., Lee, S., Lin, K. H., Maxwell, T. J., Nagai, Y., Wang, X., Welch, R. P., Yoon, J., Zhang, W., Barzilai, N., Voight, B. F., Han, B. G., Jenkinson, C. P., Kuulasmaa, T., Kuusisto, J., Manning, A., Ng, M. C. Y., Palmer, N. D., Balkau, B., Stankov, A., Abboud, H. E., Boeing, H., Giedraitis, V., Prabhakaran, D., Gottesman, O., Scott, J., Carey, J., Kwan, P., Grant, G., Smith, J. D., Neale, B. M., Purcell, S., Butterworth, A. S., Howson, J. M. M., Lee, H. M., Lu, Y., Kwak, S. H., Zhao, W., Danesh, J., Lam, V. K. L., Park, K. S., Saleheen, D., So, W. Y., Tam, C. H. T., Afzal, U., Aguilar, D., Arya, R., Aung, T., Chan, E., Navarro, C., Cheng, C. Y., Palli, D., Correa, A., Curran, J. E., Rybin, D., Farook, V. S., Fowler, S. P., Freedman, B. I., Griswold, M., Hale, D. E., Hicks, P. J., Khor, C. C., Kumar, S., Lehne, B., Thuillier, D., Lim, W. Y., Liu, J., Van Der Schouw, Y. T., Loh, M., Musani, S. K., Puppala, S., Scott, W. R., Yengo, L., Tan, S. T., Taylor,

H. A., Thameem, F., Wilson, G., Wong, T. Y., Njolstad, P. R., Levy, J. C., Mangino, M., Bonnycastle, L. L., Schwarzmayr, T., Fadista, J., Surdulescu, G. L., Herder, C., Groves, C. J., Wieland, T., Bork-Jensen, (2016) The genetic architecture of type 2 diabetes. *Nature,* **536**(7614), 41–47.

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

[79] Gallagher, M. D. and Chen-Plotkin, A. S. (2018) The Post-GWAS Era: From Association to Function. *American Journal of Human Genetics,* **102**(5), 717–730.

[80] Civelek, M. and Lusis, A. J. (January, 2014) Systems genetics approaches to understand complex traits. *Nature Reviews Genetics,* **15**(1), 34–48.

[81] Quang, D. X., Erdos, M. R., Parker, S. C. J., and Collins, F. S. (December, 2015) Motif signatures in stretch enhancers are enriched for disease-associated genetic variants. *Epigenetics & Chromatin,* **8**(1), 23.

[82] Claussnitzer, M., Dankel, S. N., Kim, K.-H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I. S., Beaudry, J. L., Puviindran, V., Abdennur, N. A., Liu, J., Svensson, P.-A., Hsu, Y.-H., Drucker, D. J., Mellgren, G., Hui, C.-C., Hauner, H., and Kellis, M. (2015) FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *New England Journal of Medicine,* **373**(10), 895–907.

[83] Alasoo, K., Rodrigues, J., Mukhopadhyay, S., Knights, A. J., Mann, A. L., Kundu, K., HIPSCI Consortium, Hale, C., Dougan, G., and Gaffney, D. J. (March, 2018) Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nature Genetics,* **50**(3), 424–431.

[84] Varshney, A., Scott, L. J., Welch, R. P., Erdos, M. R., Chines, P. S., Narisu, N., Albanus, R. D., Orchard, P., Wolford, B. N., Kursawe, R., Vadlamudi, S., Cannon, M. E., Didion, J. P., Hensley, J., Kirilusha, A., Bonnycastle, L. L., Taylor, D. L., Watanabe, R., Mohlke, K. L., Boehnke, M., Collins, F. S., Parker, S. C. J., and Stitzel, M. L. (2017) Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. *Proceedings of the National Academy of Sciences,* **114**(9), 2301–2306.

[85] Gupta, R. M., Hadaya, J., Trehan, A., Zekavat, S. M., Roselli, C., Klarin, D., Emdin, C. A., Hilvering, C. R. E., Bianchi, V., Mueller, C., Khera, A. V., Ryan,

R. J. H., Engreitz, J. M., Issner, R., Shoresh, N., Epstein, C. B., de Laat, W., Brown, J. D., Schnabel, R. B., Bernstein, B. E., and Kathiresan, S. (2017) A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression. *Cell,* **170**(3), 522–533.e15.

[86] Shannon, C. E. (July, 1948) A Mathematical Theory of Communication. *Bell System Technical Journal,* **27**(3), 379–423.

[87] Vinga, S. (May, 2014) Information theory applications for biological sequence analysis. *Briefings in Bioinformatics,* **15**(3), 376–389.

[88] Mc Mahon, S. S., Sim, A., Filippi, S., Johnson, R., Liepe, J., Smith, D., and Stumpf, M. P. (November, 2014) Information theory and signal transduction systems: From molecular information processing to network inference. *Seminars in Cell & Developmental Biology,* **35**, 98–108.

[89] Sherwin, W., Chao, A., Jost, L., and Smouse, P. (December, 2017) Information Theory Broadens the Spectrum of Molecular Ecology and Evolution. *Trends in Ecology & Evolution,* **32**(12), 948–963.

[90] Jenkinson, G., Pujadas, E., Goutsias, J., and Feinberg, A. P. (2017) Potential energy landscapes identify the information-theoretic nature of the epigenome. *Nature Genetics,* **49**(5), 719–729.

[91] Jenkinson, G., Abante, J., Feinberg, A. P., and Goutsias, J. (December, 2018) An information-theoretic approach to the modeling and analysis of whole-genome bisulfite sequencing data. *BMC Bioinformatics,* **19**(1), 87.

[92] Jenkinson, G., Abante, J., Koldobskiy, M. A., Feinberg, A. P., and Goutsias, J. (December, 2019) Ranking genomic features using an information-theoretic measure of epigenetic discordance. *BMC Bioinformatics,* **20**(1), 175.

[93] Segal, E. and Widom, J. (August, 2009) What controls nucleosome positions?. *Trends in Genetics,* **25**(8), 335–343.

[94] Rudnizky, S., Khamis, H., Malik, O., Melamed, P., and Kaplan, A. (May, 2019) The base pair-scale diffusion of nucleosomes modulates binding of transcription factors. *Proceedings of the National Academy of Sciences,* p. 201815424.

[95] Henikoff, J. G., Belsky, J. A., Krassovsky, K., MacAlpine, D. M., and Henikoff, S. (November, 2011) Epigenome characterization at single base-pair resolution. *Proceedings of the National Academy of Sciences,* **108**(45), 18318–18323.

[96] Cuellar-Partida, G., Buske, F. A., McLeay, R. C., Whitington, T., Noble, W. S., and Bailey, T. L. (2012) Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics,* **28**(1), 56–62.

[97] Ackermann, A. M., Wang, Z., Schug, J., Naji, A., and Kaestner, K. H. (March, 2016) Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. *Molecular Metabolism,* **5**(3), 233–244.

[98] Corces, M. R., Trevino, A. E., Hamilton, E. G., Greenside, P. G., Sinnott-Armstrong, N. A., Vesuna, S., Satpathy, A. T., Rubin, A. J., Montine, K. S., Wu, B., Kathiria, A., Cho, S. W., Mumbach, M. R., Carter, A. C., Kasowski, M., Orloff, L. A., Risca, V. I., Kundaje, A., Khavari, P. A., Montine, T. J., Greenleaf, W. J., and Chang, H. Y. (October, 2017) An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nature Methods,* **14**(10), 959–962.

[99] Mueller, F., Mazza, D., Stasevich, T. J., and McNally, J. G. (2010) FRAP and kinetic modeling in the analysis of nuclear protein dynamics: What do we really know?. *Current Opinion in Cell Biology,* **22**(3), 403–411.

[100] Hansen, A. S., Pustova, I., Cattoglio, C., Tjian, R., and Darzacq, X. (2017) CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *eLife,* **6**, 1–33.

[101] Hnisz, D., Day, D. S., and Young, R. A. (2016) Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. *Cell,* **167**(5), 1188–1200.

[102] Kundaje, A., Kyriazopoulou-Panagiotopoulou, S., Libbrecht, M., Smith, C. L., Raha, D., Winters, E. E., Johnson, S. M., Snyder, M., Batzoglou, S., and Sidow, A. (September, 2012) Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Research,* **22**(9), 1735–1747.

[103] Grossman, S. R., Engreitz, J., Ray, J. P., Nguyen, T. H., Hacohen, N., and Lander, E. S. (2018) Positional specificity of different transcription factor classes within enhancers. *Proceedings of the National Academy of Sciences,* p. 201804663.

[104] Callegari, A., Sieben, C., Benke, A., Suter, D. M., Fierz, B., Mazza, D., and Manley, S. (January, 2019) Single-molecule dynamics and genome-wide transcriptomics reveal that NF-kB (p65)-DNA binding times can be decoupled from transcriptional activation. *PLOS Genetics,* **15**(1), e1007891.

[105] Yanez-Cuna, J. O., Arnold, C. D., Stampfel, G., Bory, u. M., Gerlach, D., Rath, M., and Stark, A. (July, 2014) Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. *Genome Research,* **24**(7), 1147–1156.

[106] Chesi, A., Wagley, Y., Johnson, M. E., Manduchi, E., Su, C., Lu, S., Leonard, M. E., Hodge, K. M., Pippin, J. A., Hankenson, K. D., Wells, A. D., and Grant, S. F. A. (March, 2019) Genome-scale Capture C promoter interactions implicate

effector genes at GWAS loci for bone mineral density. *Nature Communications,* **10**(1), 1260.

[107] Buenrostro, J. D., Wu, B., Chang, H. Y., and Greenleaf, W. J. (2015) ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology,* **2015**(January), 21.29.1–21.29.9.

[108] Picelli, S., Bjrklund, A. K., Reinius, B., Sagasser, S., Winberg, G., and Sandberg, R. (December, 2014) Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Research,* **24**(12), 2033–2040.

[109] Rohland, N. and Reich, D. (May, 2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research,* **22**(5), 939–946.

[110] Li, H. and Durbin, R. (July, 2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* **25**(14), 1754–1760.

[111] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (August, 2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England),* **25**(16), 2078–2079.

[112] Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nussbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008) Model-based Analysis of ChIP-Seq (MACS). *Genome Biology,* **9**(9), R137.

[113] Quinlan, A. R. and Hall, I. M. (March, 2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England),* **26**(6), 841–842.

[114] Kster, J. and Rahmann, S. (October, 2012) Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics (Oxford, England),* **28**(19), 2520–2522.

[115] 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., and Abecasis, G. R. (October, 2015) A global reference for human genetic variation. *Nature,* **526**(7571), 68–74.

[116] Hausser, J. and Strimmer, K. (2009) Entropy Inference and the James-Stein Estimator, with Application to Nonlinear Gene Association Networks. *Journal of Machine Learning Research,* **10**, 1469–1484.

[117] Denas, O., Sandstrom, R., Cheng, Y., Beal, K., Herrero, J., Hardison, R. C., and Taylor, J. (February, 2015) Genome-wide comparative analysis reveals human-mouse regulatory landscape and evolution. *BMC Genomics,* **16**(1), 87.

[118] Liptak, T. (1958) On the combination of independent tests. *Magyar Tud Akad Mat Kutato Int Kozl,* **3**, 171–197.

[119] Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of statistics,* **29**(4), 1165–1188.

[120] Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S, Statistics and Computing, Statistics,Computing Venables,W.N.:Statistics w.S-PLUSSpringer-Verlag, New York 4 edition.

[121] Delignette-Muller, M. L. and Dutang, C. (2015) fitdistrplus: An R Package for Fitting Distributions. *Journal of Statistical Software,* **64**(4), 1–34.

[122] Saito, T. and Rehmsmeier, M. (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE,* **10**(3), 1–21.

[123] Davis, J. and Goadrich, M. (2006) The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning - ICML '06,* pp. 233–240.

[124] Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (October, 2005) ROCR: visualizing classifier performance in R. *Bioinformatics,* **21**(20), 3940–3941.

[125] Grau, J., Grosse, I., and Keilwagen, J. (August, 2015) PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics,* **31**(15), 2595–2597.

[126] Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. S. (October, 2009) mixtools: An R Package for Analyzing Mixture Models. *Journal of Statistical Software,* **32**(1), 1–29.

[127] Garieri, M., Delaneau, O., Santoni, F., Fish, R. J., Mull, D., Carninci, P., Dermitzakis, E. T., Antonarakis, S. E., and Fort, A. (November, 2017) The effect of genetic variation on promoter usage and enhancer activity. *Nature Communications,* **8**(1), 1–9.

[128] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (January, 2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics,* **29**(1), 15–21.

[129] Frith, M. C., Valen, E., Krogh, A., Hayashizaki, Y., Carninci, P., and Sandelin, A. (January, 2008) A code for transcription initiation in mammalian genomes. *Genome Research,* **18**(1), 1–12.

[130] Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., and Noble, W. S. (February, 2007) Quantifying similarity between motifs. *Genome Biology,* **8**(2), R24.

[131] Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., Brown, S. D., Chang, H.-Y., El-Gebali, S., Fraser, M. I., Gough, J., Haft, D. R., Huang, H., Letunic, I., Lopez, R., Luciani, A., Madeira, F., Marchler-Bauer,

A., Mi, H., Natale, D. A., Necci, M., Nuka, G., Orengo, C., Pandurangan, A. P., Paysan-Lafosse, T., Pesseat, S., Potter, S. C., Qureshi, M. A., Rawlings, N. D., Redaschi, N., Richardson, L. J., Rivoire, C., Salazar, G. A., Sangrador-Vegas, A., Sigrist, C. J. A., Sillitoe, I., Sutton, G. G., Thanki, N., Thomas, P. D., Tosatto, S. C. E., Yong, S.-Y., and Finn, R. D. (2018) InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research,*.

[132] Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M. G., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J. S., Govindarajan, S., Shaulsky, G., Walhout, A. J. M., Bouget, F.-Y., Ratsch, G., Larrondo, L. F., Ecker, J. R., and Hughes, T. R. (September, 2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell,* **158**(6), 1431–1443.

[133] Schmidt, E. M., Zhang, J., Zhou, W., Chen, J., Mohlke, K. L., Chen, Y. E., and Willer, C. J. (August, 2015) GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics,* **31**(16), 2601–2606.

[134] Delaneau, O., Ongen, H., Brown, A. A., Fort, A., Panousis, N. I., and Dermitzakis, E. T. (May, 2017) A complete tool set for molecular QTL discovery and analysis. *Nature Communications,* **8**, 15452.

[135] van de Geijn, B., McVicker, G., Gilad, Y., and Pritchard, J. K. (November, 2015) WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nature methods,* **12**(11), 1061–1063.

[136] Heger, A., Webber, C., Goodson, M., Ponting, C. P., and Lunter, G. (August, 2013) GAT: a simulation framework for testing the association of genomic intervals. *Bioinformatics,* **29**(16), 2046–2048.

[137] Phair, R. D., Scaffidi, P., Elbi, C., Vecerova, J., Dey, A., Ozato, K., Brown, D. T., Hager, G., Bustin, M., and Misteli, T. (2004) Global Nature of Dynamic Protein-Chromatin Interactions In Vivo: Three-Dimensional Genome Scanning and Dynamic Interaction Networks of Chromatin Proteins. *Molecular and Cellular Biology,* **24**(14), 6393–6402.

[138] Mayr, B. M., Guzman, E., and Montminy, M. (2005) Glutamine rich and basic region/leucine zipper (bZIP) domains stabilize cAMP-response element-binding protein (CREB) binding to chromatin. *Journal of Biological Chemistry,* **280**(15), 15103–15110.

[139] Nakahashi, H., Kwon, K. R. K., Resch, W., Vian, L., Dose, M., Stavreva, D., Hakim, O., Pruett, N., Nelson, S., Yamane, A., Qian, J., Dubois, W., Welsh, S., Phair, R. D., Pugh, B. F., Lobanenkov, V., Hager, G. L., and Casellas, R.

(2013) A Genome-wide Map of CTCF Multivalency Redefines the CTCF Code. *Cell Reports,* **3**(5), 1678–1689.

[140] Sekiya, T., Muthurajan, U. M., Luger, K., Tulin, A. V., and Zaret, K. S. (2009) Nucleosome-binding affinity as a primary determinant of the nuclear mobility of the pioneer transcription factor FoxA. *Genes and Development,* **23**(7), 804–809.

[141] Bosisio, D., Marazzi, I., Agresti, A., Shimizu, N., Bianchi, M. E., and Natoli, G. (2006) A hyper-dynamic equilibrium between promoter-bound and nucleoplasmic dimers controls NF-B-dependent gene activity. *EMBO Journal,* **25**(4), 798–810.

[142] Groeneweg, F. L., Van Royen, M. E., Fenz, S., Keizer, V. I. P., Geverts, B., Prins, J., De Kloet, E. R., Houtsmuller, A. B., Schmidt, T. S., and Schaaf, M. J. M. (2014) Quantitation of glucocorticoid receptor DNA-binding dynamics by single-molecule microscopy and FRAP. *PLoS ONE,* **9**(3), 1–12.

[143] Tirard, M., Almeida, O. F. X., Hutzler, P., Melchior, F., and Michaelidis, T. M. (2007) Sumoylation and proteasomal activity determine the transactivation properties of the mineralocorticoid receptor. *Molecular and Cellular Endocrinology,* **268**(1-2), 20–29.

[144] Hinow, P., Rogers, C. E., Barbieri, C. E., Pietenpol, J. A., Kenworthy, A. K., and DiBenedetto, E. (2006) The DNA binding activity of p53 displays reaction-diffusion kinetics. *Biophysical Journal,* **91**(1), 330–342.

[145] Scripture-Adams, D. D., Damle, S. S., Li, L., Elihu, K. J., Qin, S., Arias, A. M., Butler, R. R., Champhekar, A., Zhang, J. A., and Rothenberg, E. V. (October, 2014) GATA-3 Dose-Dependent Checkpoints in Early T Cell Commitment. *The Journal of Immunology,* **193**(7), 3470–3491.

[146] Allman, D., Sambandam, A., Kim, S., Miller, J. P., Pagan, A., Well, D., Meraz, A., and Bhandoola, A. (February, 2003) Thymopoiesis independent of common lymphoid progenitors. *Nature Immunology,* **4**(2), 168–174.

[147] Bell, J. J. and Bhandoola, A. (April, 2008) The earliest thymic progenitors for T cells possess myeloid lineage potential. *Nature,* **452**(7188), 764–767.

[148] Van de Walle, I., Dolens, A.-C., Durinck, K., De Mulder, K., Van Loocke, W., Damle, S., Waegemans, E., De Medts, J., Velghe, I., De Smedt, M., Vandekerckhove, B., Kerre, T., Plum, J., Leclercq, G., Rothenberg, E. V., Van Vlierberghe, P., Speleman, F., and Taghon, T. (September, 2016) GATA3 induces human T-cell commitment by restraining Notch activity and repressing NK-cell fate. *Nature Communications,* **7**(1), 11171.

[149] Raulet, D. H., Garman, R. D., Saito, H., and Tonegawa, S. (March, 1985) Developmental regulation of T-cell receptor gene expression. *Nature,* **314**(6006), 103–107.

[150] Pardoll, D. M., Fowlkes, B. J., Bluestone, J. A., Kruisbeek, A., Maloy, W. L., Coligan, J. E., and Schwartz, R. H. (March, 1987) Differential expression of two distinct T-cell receptors during thymocyte development. *Nature,* **326**(6108), 79–81.

[151] Zerrahn, J., Held, W., and Raulet, D. H. (March, 1997) The MHC Reactivity of the T Cell Repertoire Prior to Positive and Negative Selection. *Cell,* **88**(5), 627–636.

[152] Heng, T. S. P., Painter, M. W., Elpek, K., Lukacs-Kornek, V., Mauermann, N., Turley, S. J., Koller, D., Kim, F. S., Wagers, A. J., Asinovski, N., Davis, S., Fassett, M., Feuerer, M., Gray, D. H. D., Haxhinasto, S., Hill, J. A., Hyatt, G., Laplace, C., Leatherbee, K., Mathis, D., Benoist, C., Jianu, R., Laidlaw, D. H., Best, J. A., Knell, J., Goldrath, A. W., Jarjoura, J., Sun, J. C., Zhu, Y., Lanier, L. L., Ergun, A., Li, Z., Collins, J. J., Shinton, S. A., Hardy, R. R., Friedline, R., Sylvia, K., and Kang, J. (October, 2008) The Immunological Genome Project: networks of gene expression in immune cells. *Nature Immunology,* **9**(10), 1091–1094.

[153] Mingueneau, M., Kreslavsky, T., Gray, D., Heng, T., Cruse, R., Ericson, J., Bendall, S., Spitzer, M. H., Nolan, G. P., Kobayashi, K., von Boehmer, H., Mathis, D., and Benoist, C. (June, 2013) The transcriptional landscape of T cell differentiation. *Nature Immunology,* **14**(6), 619–632.

[154] Thanos, D. and Maniatis, T. (December, 1995) Virus induction of human IFN gene expression requires the assembly of an enhanceosome. *Cell,* **83**(7), 1091–1100.

[155] Panne, D., Maniatis, T., and Harrison, S. C. (June, 2007) An Atomic Model of the Interferon- Enhanceosome. *Cell,* **129**(6), 1111–1123.

[156] Weintraub, H. and Groudine, M. (September, 1976) Chromosomal subunits in active genes have an altered conformation. *Science,* **193**(4256), 848–856.

[157] Stalder, J., Groudine, M., Dodgson, J. B., Engel, J. D., and Weintraub, H. (April, 1980) Hb switching in chickens. *Cell,* **19**(4), 973–980.

[158] Stalder, J. (June, 1980) Tissue-specific DNA cleavages in the globin chromatin domain introduced by DNAase I. *Cell,* **20**(2), 451–460.

[159] Wall, L., DeBoer, E., and Grosveld, F. (September, 1988) The human beta-globin gene 3' enhancer contains multiple binding sites for an erythroid-specific protein.. *Genes & Development,* **2**(9), 1089–1100.

[160] Freese, N. H., Norris, D. C., and Loraine, A. E. (July, 2016) Integrated genome browser: visual analytics platform for genomics. *Bioinformatics,* **32**(14), 2089–2095.

[161] Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B. D., Shen, Y., Pervouchine, D. D., Djebali, S., Thurman, R. E., Kaul, R., Rynes, E., Kirilusha, A., Marinov, G. K., Williams, B. A., Trout, D., Amrhein, H., Fisher-Aylor, K., Antoshechkin, I., DeSalvo, G., See, L.-H., Fastuca, M., Drenkow, J., Zaleski, C., Dobin, A., Prieto, P., Lagarde, J., Bussotti, G., Tanzer, A., Denas, O., Li, K., Bender, M. A., Zhang, M., Byron, R., Groudine, M. T., McCleary, D., Pham, L., Ye, Z., Kuan, S., Edsall, L., Wu, Y.-C., Rasmussen, M. D., Bansal, M. S., Kellis, M., Keller, C. A., Morrissey, C. S., Mishra, T., Jain, D., Dogan, N., Harris, R. S., Cayting, P., Kawli, T., Boyle, A. P., Euskirchen, G., Kundaje, A., Lin, S., Lin, Y., Jansen, C., Malladi, V. S., Cline, M. S., Erickson, D. T., Kirkup, V. M., Learned, K., Sloan, C. A., Rosenbloom, K. R., Lacerda de Sousa, B., Beal, K., Pignatelli, M., Flicek, P., Lian, J., Kahveci, T., Lee, D., James Kent, W., Ramalho Santos, M., Herrero, J., Notredame, C., Johnson, A., Vong, S., Lee, K., Bates, D., Neri, F., Diegel, M., Canfield, T., Sabo, P. J., Wilken, M. S., Reh, T. A., Giste, E., Shafer, A., Kutyavin, T., Haugen, E., Dunn, D., Reynolds, A. P., Neph, S., Humbert, R., Scott Hansen, R., De Bruijn, M., Selleri, L., Rudensky, A., Josefowicz, S., Samstein, R., Eichler, E. E., Orkin, S. H., Levasseur, D., Papayannopoulou, T., Chang, K.-H., Skoultchi, A., Gosh, S., Disteche, C., Treuting, P., Wang, Y., Weiss, M. J., Blobel, G. A., Cao, X., Zhong, S., Wang, T., Good, P. J., Lowdon, R. F., Adams, L. B., Zhou, X.-Q., Pazin, M. J., Feingold, E. A., Wold, B., Taylor, J., Mortazavi, A., Weissman, S. M., Stamatoyannopoulos, J. A., Snyder, M. P., Guigo, R., Gingeras, T. R., Gilbert, D. M., Hardison, R. C., Beer, M. A., and Ren, B. (November, 2014) A comparative encyclopedia of DNA elements in the mouse genome. *Nature,* **515**(7527), 355–364.

[162] Gaspar-Maia, A., Alajem, A., Meshorer, E., and Ramalho-Santos, M. (January, 2011) Open chromatin in pluripotency and reprogramming. *Nature Reviews Molecular Cell Biology,* **12**(1), 36–47.

[163] Welch, R. P., Lee, C., Imbriano, P. M., Patil, S., Weymouth, T. E., Smith, R. A., Scott, L. J., and Sartor, M. A. (2014) ChIP-Enrich: Gene set enrichment testing for ChIP-seq data. *Nucleic Acids Research,* **42**(13), 1–13.

[164] Wei, G., Abraham, B. J., Yagi, R., Jothi, R., Cui, K., Sharma, S., Narlikar, L., Northrup, D. L., Tang, Q., Paul, W. E., Zhu, J., and Zhao, K. (August, 2011) Genome-wide Analyses of Transcription Factor GATA3-Mediated Gene Regulation in Distinct T Cell Types. *Immunity,* **35**(2), 299–311.

[165] Jojic, V., Shay, T., Sylvia, K., Zuk, O., Sun, X., Kang, J., Regev, A., and Koller, D. (June, 2013) Identification of transcriptional regulators in the mouse immune system. *Nature Immunology,* **14**(6), 633–643.

[166] Sawada, S. and Littman, D. R. (November, 1991) Identification and characterization of a T-cell-specific enhancer adjacent to the murine CD4 gene.. *Molecular and Cellular Biology,* **11**(11), 5506–5515.

[167] Siu, G., Wurster, A., Duncan, D., Soliman, T., and Hedrick, S. (August, 1994) A transcriptional silencer controls the developmental expression of the CD4 gene.. *The EMBO Journal,* **13**(15), 3570–3579.

[168] Wurster, A. L., Siu, G., Leiden, J. M., and Hedrick, S. M. (October, 1994) Elf-1 binds to a critical element in a second CD4 enhancer.. *Molecular and Cellular Biology,* **14**(10), 6452–6463.

[169] Hostert, A., Tolaini, M., Roderick, K., Harker, N., Norton, T., and Kioussis, D. (October, 1997) A Region in the CD8 Gene Locus That Directs Expression to the Mature CD8 T Cell Subset in Transgenic Mice. *Immunity,* **7**(4), 525–536.

[170] Hostert, A., Tolaini, M., Festenstein, R., McNeill, L., Malissen, B., Williams, O., Zamoyska, R., and Kioussis, D. (May, 1997) A CD8 genomic fragment that directs subset-specific expression of CD8 in transgenic mice. *Journal of Immunology (Baltimore, Md.: 1950),* **158**(9), 4270–4281.

[171] Ellmeier, W., Sunshine, M. J., Losos, K., Hatam, F., and Littman, D. R. (October, 1997) An Enhancer That Directs Lineage-Specific Expression of CD8 in Positively Selected Thymocytes and Mature T Cells. *Immunity,* **7**(4), 537–547.

[172] Hosoya-Ohmura, S., Lin, Y.-H., Herrmann, M., Kuroha, T., Rao, A., Moriguchi, T., Lim, K.-C., Hosoya, T., and Engel, J. D. (May, 2011) An NK and T Cell Enhancer Lies 280 Kilobase Pairs 3' to the Gata3 Structural Gene. *Molecular and Cellular Biology,* **31**(9), 1894–1904.

[173] Ohmura, S., Mizuno, S., Oishi, H., Ku, C.-J., Hermann, M., Hosoya, T., Takahashi, S., and Engel, J. D. (January, 2016) Lineage-affiliated transcription factors bind the Gata3 Tce1 enhancer to mediate lineage-specific programs. *Journal of Clinical Investigation,* **126**(3), 865–878.

[174] Porcher, C., Liao, E. C., Fujiwara, Y., Zon, L. I., and Orkin, S. H. (October, 1999) Specification of hematopoietic and vascular development by the bHLH transcription factor SCL without direct DNA binding. *Development (Cambridge, England),* **126**(20), 4603–4615.

[175] Wright, C. W. and Duckett, C. S. (January, 2009) The Aryl Hydrocarbon Nuclear Translocator Alters CD30-Mediated NF- B-Dependent Transcription. *Science,* **323**(5911), 251–255.

[176] Kioussis, D. and Ellmeier, W. (December, 2002) Chromatin and CD4, CD8A and CD8B gene expression during thymic differentiation. *Nature Reviews Immunology,* **2**(12), 909–919.

[177] Kheradpour, P. and Kellis, M. (March, 2014) Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Research,* **42**(5), 2976–2987.

[178] Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C.-y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., Zhang, A. W., Parcy, F., Lenhard, B., Sandelin, A., and Wasserman, W. W. (January, 2016) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research,* **44**(D1), D110–D115.

[179] Edgar, R. (January, 2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research,* **30**(1), 207–210.

[180] Hosoya, T., Albanus, R. D., Hensley, J., Myers, G., Kyono, Y., Kitzman, J., Parker, S. C. J., and Engel, J. D. (April, 2018) Global dynamics of stage-specific transcription factor binding during thymocyte development. *Scientific Reports,* **8**(1), 5605.

[181] Taylor, D. L., Knowles, D. A., Scott, L. J., Ramirez, A. H., Casale, F. P., Wolford, B. N., Guan, L., Varshney, A., Albanus, R. D., Parker, S. C. J., Narisu, N., Chines, P. S., Erdos, M. R., Welch, R. P., Kinnunen, L., Saramies, J., Sundvall, J., Lakka, T. A., Laakso, M., Tuomilehto, J., Koistinen, H. A., Stegle, O., Boehnke, M., Birney, E., and Collins, F. S. (April, 2018) Interactions between genetic variation and cellular environment in skeletal muscle gene expression. *PLOS ONE,* **13**(4), e0195788.

[182] Kycia, I., Wolford, B. N., Huyghe, J. R., Fuchsberger, C., Vadlamudi, S., Kursawe, R., Welch, R. P., Albanus, R. d., Uyar, A., Khetan, S., Lawlor, N., Bolisetty, M., Mathur, A., Kuusisto, J., Laakso, M., Ucar, D., Mohlke, K. L., Boehnke, M., Collins, F. S., Parker, S. C. J., and Stitzel, M. L. (April, 2018) A Common Type 2 Diabetes Risk Variant Potentiates Activity of an Evolutionarily Conserved Islet Stretch Enhancer and Increases C2CD4A and C2CD4B Expression. *The American Journal of Human Genetics,* **102**(4), 620–635.

[183] Varshney, A., Kyono, Y., Elangovan, V. R., Wang, C., Erdos, M. R., Narisu, N., Albanus, R. D., Orchard, P., Stitzel, M. L., Collins, F. S., Kitzman, J. O., and Parker, S. C. J., Integrating Enhancer RNA signatures with diverse omics data identifies characteristics of transcription initiation in pancreatic islets. preprint, Genomics (October, 2019).

[184] Castro, M. A. A., de Santiago, I., Campbell, T. M., Vaughn, C., Hickey, T. E., Ross, E., Tilley, W. D., Markowetz, F., Ponder, B. A. J., and Meyer, K. B. (January, 2016) Regulators of genetic risk of breast cancer identified by integrative network analysis. *Nature Genetics,* **48**(1), 12–21.