Tracking Transcription Factors on the Genome by their DNase-seq Footprints

by

Galip Gürkan Yardımcı

Graduate Program in Computational Biology and Bioinformatics
Duke University

Date:_______________________
Approved:

___________________________
Gregory E. Crawford, Co-Supervisor

___________________________
Uwe Ohler, Co-Supervisor

___________________________
Raluca Gordan

___________________________
Alexander Hartemink

___________________________
L. Ryan Baugh

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate Program in Computational Biology and Bioinformatics in the Graduate School of Duke University 2014
ABSTRACT

Tracking Transcription Factors on the Genome by their DNase-seq Footprints

by

Galip Gürkan Yardımcı

Graduate Program in Computational Biology and Bioinformatics
Duke University

Date:_______________________
Approved:

___________________________
Gregory E. Crawford, Co-Supervisor

___________________________
Uwe Ohler, Co-Supervisor

___________________________
Raluca Gordan

___________________________
Alexander Hartemink

___________________________
L. Ryan Baugh

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate Program in Computational Biology and Bioinformatics in the Graduate School of Duke University 2014
Abstract

Transcription factors control numerous vital processes in the cell through their ability to control gene expression. Dysfunctional regulation by transcription factors lead to disorders and disease. Transcription factors regulate gene expression by binding to DNA sequences (motifs) on the genome and altering chromatin. DNase-seq footprinting is a well-established assay for identification of DNA sequences that bind to transcription factors. We developed computational techniques to analyze footprints and predict transcription factor binding. These transcription factor specific predictive models are able to correct for DNase sequence bias and characterize variation in DNA binding sequence. We found that DNase-seq footprints are able to identify cell-type or condition specific transcription factor activity and may offer information about the type of the interaction between DNA and transcription factor. Our DNase-seq footprint model is able to accurately discover high confidence transcription factor binding sites and discover alternative interactions between transcription factors and DNA. DNase-seq footprints can be used with ChIP-seq data to discover true binding sites and better understand transcription regulation.
Dedication

To my mother
Contents

Abstract ........................................................................................................................................iv

List of Tables ................................................................................................................................x

List of Figures ..............................................................................................................................xi

Acknowledgements .................................................................................................................... xiii

1 Background ............................................................................................................................... 1

1.1 Genes and Chromatin ........................................................................................................... 1

1.2 Transcription factors ............................................................................................................ 3

1.3 Methods of detecting transcription factor binding ............................................................... 4

1.3.1 Chromatin Immunoprecipitation ................................................................................... 5

1.3.2 Computational Methods ............................................................................................... 7

1.3.3 DNase-seq and Footprinting ......................................................................................... 8

1.3.4 Additional genomic assays to interrogate transcription control .................................. 12

1.4 Goals and contributions ....................................................................................................... 13

1.5 Outline ................................................................................................................................ 14

2 Data preparation ....................................................................................................................... 16

2.1 DNase-seq ........................................................................................................................... 16

2.2 Detection of DNase Hypersensitive Sites .......................................................................... 17

2.3 Identification of transcription factor binding sites using ChIP-seq .................................... 18

2.4 Scanning the genome for candidate binding sites .............................................................. 19

2.5 RNA-seq ............................................................................................................................ 20

3 Aggregated DNase-seq footprints identify cell-type or condition specific transcription factor binding ........................................................................................................... 21

3.1 Introduction ......................................................................................................................... 21
3.2 Aggregate DNase-seq profiles and footprints ........................................ 23
3.3 Cell type specific DNase-seq footprints ............................................. 26
3.4 Condition specific DNase-seq footprints ............................................ 28
  3.4.1 Androgen Receptor ................................................................. 28
  3.4.2 Androgen activated DNase-seq footprint ........................................ 30
  3.4.3 Multiple AR footprints may correspond to different modes of binding ...... 32
3.5 Methods .............................................................................................. 37
  3.5.1 DNase-seq aggregate profiles ......................................................... 37
  3.5.2 Clustering of DNase-seq profiles ..................................................... 38
3.6 Discussion ............................................................................................. 38

4 Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection .......................................................................................... 41
  4.1 Introduction .......................................................................................... 41
  4.2 Results .................................................................................................. 44
    4.2.1 ChIP-seq sites display a heterogeneous mixture of DNase digestion profiles. 45
    4.2.2 Sequence bias of DNase-seq is protocol specific ................................ 45
    4.2.3 DNase footprints are an aggregation of a footprint and background profiles 46
    4.2.4 DNase footprints predict direct binding events with high sensitivity ....... 48
    4.2.5 ChIP-seq peaks without a DNase footprint suggests weak or indirect binding .................................................................................................................... 51
    4.2.6 DNase footprints detect microarchitecture changes in chromatin landscape 54
    4.2.7 Unsupervised EM training can identify de novo DNase bias .................. 56
    4.2.8 DNase footprints can identify variants of the main sequence motif ......... 58
4.3 Methods .......................................................................................................................... 58
  4.3.1 DNase-seq assay and sequence bias of DNase.......................................................... 58
  4.3.2 Modeling of DNase-seq footprints ......................................................................... 60
  4.3.3 Discovery of Co-binding elements ........................................................................... 64
4.4 Discussion ....................................................................................................................... 65

5 CTCF binding is necessary and sufficient for maintaining open chromatin .......... 81
  5.1 Introduction .................................................................................................................. 81
  5.2 Results .......................................................................................................................... 82
    5.2.1 CTCF knockdown changes genome-wide chromatin architecture ...................... 82
    5.2.2 Genome-wide CTCF binding is weakened .............................................................. 86
    5.2.3 Chromatin changes are caused by direct and indirect effects of CTCF binding loss .......................................................................................................................................................................................... 88
  5.3 Methods ......................................................................................................................... 92
    5.3.1 Detection of differential DHS sites ......................................................................... 92
  5.4 Discussion ....................................................................................................................... 93

6 Development of Broad ChIP-seq peak caller ................................................................. 95
  6.1 Introduction .................................................................................................................... 95
  6.2 Results ........................................................................................................................... 96
    6.2.1 Successful shift size estimation depends on protocol specific shearing .............. 96
  6.3 Kernel density estimates can detect broad regions of enrichment............................. 99
    6.3.1 Broad peaks are highly reproducible between replicates ...................................... 101
  6.4 Discussion ....................................................................................................................... 103

7 Conclusions ....................................................................................................................... 105
  7.1 Summary ....................................................................................................................... 105
  7.2 Discussion ....................................................................................................................... 108
7.3 Future directions........................................................................................................... 110

References ........................................................................................................................... 112

Biography ............................................................................................................................ 122
List of Tables

Table 4.1 : auROC and Sensitivity at 1% FPR values for D-s and FLR for each transcription factor.................................................................78

Table 4.2 : ENCODE antibody names used for ChIP experiments transcription factors and the IDs of PWMs used to look for sequence motif matches in these peaks........79

Table 4.3 : P-values for Kolmogorov-Smirnoff test (one-sided) for comparing the distributions of FLR and D-s metrics assigned to sequence motifs from two different cell types..........................................................................................................................80

Table 6.2 : Number of peak pairs that are under the IDR threshold for broad peak caller compared with peak calls from MACS and SPP.......................................................102
List of Figures

Figure 1.1: ChIP-seq protocol ......................................................................................... 6
Figure 1.2: DNase-seq protocol ......................................................................................... 9
Figure 3.1: Aggregate DNase-seq plots ............................................................................ 25
Figure 3.2: Aggregate plots of DNase-reads around motifs for factors with high regression coefficients ............................................................................................................. 27
Figure 3.3: Base pair resolution around AR motif matches reveals a unique pattern of protection by the AR .......................................................................................................................... 30
Figure 3.4: AR binding displays three distinct modes of AR-DNA interaction that are specific to ligand activated AR ............................................................................................................. 34
Figure 3.5: DNaseI protection around AR motifs ................................................................ 35
Figure 4.1: Scenarios relevant to identifying DNase footprints ......................................... 68
Figure 4.2: Aggregate DNase plots identify distinct TF binding profiles .......................... 69
Figure 4.3: DNase-seq displays cleavage bias that is protocol specific .............................. 70
Figure 4.4: Workflow of binary classification scheme ...................................................... 71
Figure 4.5: auROC and sensitivity values for D-s and FLR values in for classification of all sequence motifs detected across the genome ............................................................. 72
Figure 4.6: Comparison of Footprint Likelihood Ratio (FLR) to general DNase sensitivity (D-s) score ........................................................................................................................................ 72
Figure 4.7: Aggregate DNase-seq plots for E2F1 sequence motif matches in E2F1 ChIP-seq peaks from MCF7 cell line ........................................................................................................... 73
Figure 4.8: Receiver operating characteristic (ROC) curves for 3 transcription factors using FLR as a predictor for all genome wide sequence motif matches ........................................ 73
Figure 4.9: Real and bias simulated aggregate DNase-seq plots ...................................... 74
Figure 4.10: Footprint scores indicate mode of transcription factor interaction ................. 75
Figure 4.11: Cell type specific footprints in shared DHS sites ........................................... 76
Figure 4.12: EM footprint components distinguish background bias and footprints, as well as alternate motif usage.

Figure 5.1: CTCF knockdown experiments.

Figure 5.2: Changes in DHSs.

Figure 5.3: Changes in CTCF footprints.

Figure 5.4: Drivers of chromatin modifications caused by knockdown.

Figure 5.5: Indexes of genes that are differentially regulated.

Figure 6.1: Cross correlation plot for ChIP-seq data for BLJ.

Figure 6.2: Representative ChIP-seq pileup and density estimates.
Acknowledgements

I would like to thank my advisors, Dr. Uwe Ohler and Dr. Greg Crawford, for giving me an opportunity to work with them when the chance presented itself. I thoroughly enjoyed working with them at professional and personal levels. I wish them all the best for the rest of their careers.

I would like to extend my thanks to my dissertation committee members - Dr. Alex Hartemink, Dr. Raluca Gordan and Dr. Ryan Baugh - for their advice and feedback. I would like to especially thank Dr. Raluca Gordan for being there to help every time I needed it.

I would like thank all the great people I worked with at Duke University, especially all members of Ohler and Crawford labs. Without exception, they were all fantastic collaborators and great team members. From Ohler lab I would like to specifically like to thank Bill Majoros and Iulian Pruteanu for sharing their considerable experience in statistical models. Anirudh Natarajan from Ohler lab was also a great colleague to work with and I wish him well. From Crawford labs, I would like to thank Alok Tewari and John Mayfield. From day one they kindly invited me to collaborate with them in their interesting projects and it was a great experience to work with them.

I would like to thank my friends in Duke Computational Biology and Bioinformatics community for choosing to study here with me. It was a big adaption for me to get used to living in Durham and their company has made it fun enough to be here. Michael Mayhew, Diana Fusco, Lingling Zheng, Chetan Rupakheti and especially Jing Hu, you have made my life in North Carolina better.
Last but obviously not least, I would like to thank my parents and family. Their constant and unconditional support and dedication to me has made it possible to achieve the things I have, especially this dissertation.
1 Background

1.1 Genes and Chromatin

Organisms store the information necessary for their existence in their genomes. The term ‘Genome’ was coined by combining the words gene and chromosome. Traditionally genes consist of discrete DNA sequences that describe different proteins. Chromosomes are pieces of coiled DNA that store numerous genes and regulatory DNA sequences.

Every cell of a multicellular organism carries at least one copy of its complete genome. Cells of multicellular organism fulfill many different functions; this is possible thanks to their ability to transform into different cell types through production (synthesis) of different sets of proteins. It was shown that a frog egg can successfully still differentiate into a number of different cells after having its genome replaced by that of a fully differentiated frog cell’s genome. This classical experiment showed that cells achieve differentiation without altering the contents of their genome.

Cells are able to differentiate and fulfill different tasks thanks to mechanisms that control the production of specific proteins cells need to differentiate. First step in protein synthesis is copying a DNA sequence into RNA; this process is called transcription. Transcription is followed by processing of RNA sequence that contains genetic information and synthesis of protein from the RNA molecule (translation). Cells control
production of each protein at different layers of this process; however regulation of transcription is paramount.

RNA polymerase II (PolII) enzyme is responsible for transcription of DNA to RNA in eukaryotic cells. Typically PolII first assembles preiniation complex (PIC) at the start of a gene, then moves along the gene, copying DNA into RNA. However PolII does not assemble PIC at every gene due to factors that regulate PIC assembly. Such regulation is achieved through interplay between two mechanisms: chromatin and transcription factors.

Chromatin refers to the complex structure that assists cells in storing DNA. Human genome from start to end is approximately as long as a meter; chromatin allows compaction of human genome to fit a copy into each of the millions of cells that make up the human body. Basic building block of chromatin is the nucleosome, approximately 146 base pairs of DNA wrapped roughly twice around a complex made of histone proteins. Each histone has a short polypeptide tail that can be chemically modified to alter the strength of interaction between histone and DNA.

In addition to its duties in compacting DNA, chromatin also controls transcription because DNA wrapped tightly around nucleosomes cannot interact with PolII. However, nucleosomes are dynamic structures and their interaction with DNA can be controlled. Proteins called transcription factors can bind to DNA sequences that normally wrap around nucleosomes, either displacing nucleosomes or weakening their
interaction with DNA. Thus, binding of transcription factor, makes it possible for PolIII to assemble PIC and transcribe genes.

### 1.2 Transcription factors

Due to their ability to modify chromatin and control transcription, transcription factors play a central role in regulation of gene transcription. For transcription factors to control transcription of specific genes, they must bind to specific regions in the genome. This specificity is achieved thanks to different transcription factors having an affinity for certain DNA sequences, as determined by their three dimensional structures [1]. These DNA sequences are short (6-20bp in eukaryotes) and degenerate which can be represented by a consensus sequence or a set of consensus sequences. This set is called the sequence motif.

The most widely used model that represents the range of DNA sequences that TFs have an affinity to is called position weight matrix (PWM) [2]. A PWM is built from relative frequencies of DNA bases at each position that interacts with transcription factor, representing the contribution of DNA base to binding. PWMs are generally built by assembling a number of sequences that are known to bind to transcription factor (identified by experimental methods such as SELEX [3]), followed by calculating the relative frequency of each DNA base at every position along the length of the DNA sequence. Alternatively, entries in a PWM can also be formulated in log likelihoods ratio: relative frequencies from foreground (known binding sequences) divided by
relative frequencies from background (typically estimated from whole genome or specific regions of genome). In this formulation, likelihood of being bound for a DNA sequence can be calculated by adding log likelihood values accordingly. This value is typically called the PWM score.

Since DNA sequences that bind to a specific transcription factor are short and degenerate, there are thousands of sequences across the genome that resemble the sequence motif modeled by PWM (i.e. high PWM scores). Most of these DNA sequences are known to be not functional. Querying the genome in this manner also gives a static picture of DNA-TF interactions even though lot of variation is known to occur between different cell types that depend on cell type specific state of chromatin [4]. As a result, experimental methods are utilized to accurately characterize transcription factor binding sites

1.3 Methods of detecting transcription factor binding

There are a variety of experimental methods that characterize whether a specific DNA sequence binds to a transcription factor (or generally a protein) or not, such as DNA affinity chromatography [5] and gel shift assays [6]. These methods are both low throughput, therefore not suited to characterize a large number of DNA sequences. Another drawback of such methods is that they are performed in vitro (in test tube). As a result, they do not tell us whether this TF-DNA interaction occurs in a living cell.
1.3.1 Chromatin Immunoprecipitation

The most common assay for detecting transcription factor binding in vivo (in living cells) is chromatin immunoprecipitation (ChIP), which utilizes antibodies that specifically target transcription factor/proteins [7, 8]. In this assay, proteins in cells are cross-linked to chromatin using a chemical such as formaldehyde, and whole chromatin is fragmented into manageable sizes by sonication. DNA sequences that are bound by the transcription factor are isolated with the help of the antibody that is specific to transcription factor of interest. After removal of crosslinks, these DNA fragments are characterized by either using microarrays (ChIP-chip) or sequencing (ChIP-seq) as shown in Figure 1.

ChIP based assay result in an enrichment of DNA fragments in the vicinity of the binding site. Resulting enrichment sites are typically broader (100+ bp) compared to the real binding site (6-20bp) [9]. Regions that have a statistically significant enrichment of DNA fragments (called ‘peaks’) are identified using computational approaches (peak callers).

Due to sequencing protocol, ChIP-seq experiments result in two peaks (upstream and downstream) that surround the putative binding site. Distance between pairs of peaks is typically referred a ‘shift size’ [9]. Certain peak callers try to account for this property: SPP models the shift size by calculating cross-correlation values [10]; MACS
first estimates shift size then shifts each DNA fragment by this value before looking for peaks [11].

Figure 1.1: ChIP-seq protocol for proteins and histone tail modifications. Alternative sequencing methods are shown in detail. Adapted from Park et al. [9]
Although ChIP-seq is most widely used protocol for identifying in vivo transcription factor binding sites, it has certain drawbacks. First requirement is a high quality antibody specific to transcription factor of interest, which may not be always available. Binding of a single transcription factor can be queried per experiment, requiring hundreds of experiments to exhaustively characterize binding sites for all active transcription factors in a given cell type. Furthermore, due to the nature of the protocol, resolution of peaks is low; thus one cannot easily determine which exact DNA sequence binds to transcription factor [9]. Recent studies showed that certain regions of chromatin interact with large number of transcription factors in a nonspecific manner and bodies of genes undergoing high levels of transcription exhibit an enrichment of unexpected transcription factors [12, 13]. Such regions are thought to be artifacts associated with this protocol. However ChIP-seq still remains the gold standard in the field to characterize transcription factor binding sites [9].

1.3.2 Computational Methods

Experiments that improved our understanding of transcription factor-DNA interactions have also spurned the development of computational methods that predict regions of transcription factor binding [14-18]. Some of these methods purely depend on DNA sequence whereas others utilize information from other genomic assays that measure chromatin state and histone tail modifications.
When the binding motif is known, either described as a consensus sequence or a PWM, the simplest way of searching for candidate binding sites is accomplished by scanning the genome for DNA sequences that are significantly similar to PWM model/sequence motif. However it is well known that such PWM scanning approaches lead to many false positives and fails to identify cell type specific binding sites.

More successful approaches tend to leverage additional information, such as conservation of blocks of DNA sequences across different species. Although these methods are more accurate because conserved sequences are likely to be functional, these methods also give a static picture of binding [18]. Methods that use information chromatin based assays are more accurate and better suited to discover cell type specific binding sites [14, 15, 17]. These methods typically look for candidate binding sites (discovered by PWM scans) that fall in regions of open chromatin, which can be detected by ChIP experiments that target certain histone modifications. It has been reported that when different assays that detect open chromatin are used as classifier features, most informative feature is DNase-seq [16].

1.3.3 DNase-seq and Footprinting

1.3.3.1 DNase-seq

Deoxyribonuclease (DNase) enzyme is an endonuclease that cleaves phosphodiester bonds in the DNA backbone. Unlike restriction enzymes, DNase cleaves DNA irrespective of DNA sequence, albeit with a small degree of sequence bias [19].
Regions of chromatin that are not tightly wrapped around nucleosomes are more likely to be cleaved by DNase; these regions are called DNase Hypersensitive Sites (DHS).

DHSs have been shown to mark regulatory regions such as promoters, enhancers, insulators, silencers and locus control regions [20, 21].

Figure 1.2: DNase-seq protocol
Recent advent of high throughput DNase-seq protocol has made it possible to characterize regulatory elements of complete genome of a cell type by a single experiment. DNase-seq starts by treating chromatin with DNase. DNA oligomers resulting from DNase cleavage are ligated to linkers and cleaved by MmeI. Resulting 20bp long DNA fragments are sequenced after amplification. These fragments are then aligned against the genome, each fragment corresponding to a DNase cleavage event. This protocol is shown in Figure 2.

After DNase-seq fragments are aligned against genome, peak calling algorithms are used to detect regions significantly enriched for DNase cleavage events [22]. These peaks define the boundaries of DHSs which are enriched for regulatory elements of the genome. DHSs have also been shown to be enriched for transcription factor binding sites and used as a predictor of transcription factor binding. However, DNase-seq experiments can be used to detect high resolution transcription factor binding events by applying DNase footprinting approach [23, 24].

1.3.3.2 DNase footprinting

An old and well established assay, DNase footprinting assay is used to discover protein interaction sites on a DNA molecule [25]. Whereas open chromatin or a DNA molecule in test tube is susceptible to DNase cleavage, the DNA sequence that binds to a protein (or a transcription factor) is protected from DNase cleavage by the physical interaction between protein and DNA. DNase footprint assay operates on this principle;
detection of absence of DNase cleavage on a DNA molecule allows detection of the site of protein-DNA interaction.

A number of computational approaches that use DNase-seq footprinting approaches have been designed to identify transcription factor binding sites. These methods range from simple DNase-seq fragment count ratios [26] to Hidden Markov models [24], Bayesian networks [27] and generative models[28]. However, their aim is same: detection of DNase-seq signal depleted regions surrounded by relative enrichment of DNase-seq signal. In searching for such regions, these methods make different assumptions about footprints and can be broken into two groups.

Early and traditional footprinting algorithms look for regions of depleted DNase-seq signal of variable size (typically ranging from 5-20bp), surrounded by two peaks, a peak trough peak shape. There are simple approaches that calculate the ratio of DNase-seq fragments in the trough to peaks. Hidden Markov models [24] and Bayesian network [27] based models that similarly look for a variable size peak trough shape have been implemented. One common property of these algorithms is the variable footprint size and subsequent identification of transcription factor that caused the footprint. These methods also assume all footprints have similar peak trough peak shapes.

Some footprinting algorithms leverage more information than a simple peak trough peak shape. CENTIPEDE algorithm models each transcription factor with a unique footprint model that is weighted against a uniform background model [15]. This
method also factors in other variables to predict transcription factor binding, such as DNase accessibility surrounding the footprint, PWM score, distance to transcription start sites. As such, it is not a pure footprinting algorithm. Wellington algorithm, instead of looking for depleted DNase-seq signal, identifies footprints by identifying regions where DNase-seq reads preferentially map against one strand vs. another [29]. These algorithms aim to predict transcription factor binding rather than strictly identifying traditional footprints.

There is an open debate in the field concerning how accurately DNase-seq footprints detect transcription factor binding sites. Furthermore, effect of DNase sequence bias of DNase on footprint discovery has been brought into question [30, 31]. Regardless, DNase-seq footprinting remains a strong alternative to ChIP based assays for its ability to characterize many transcription factor binding sites with a single experiment.

1.3.4 Additional genomic assays to interrogate transcription control

There are numerous other high throughput assays that enhance our understanding of transcription regulation and role of transcription factors. Three assays of interest for understanding transcription factor binding and chromatin conformation are: Protein Binding Microarrays (PBM) [32], Chromosome Conformation Capture (3C) [33] and Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) assays [34].
Protein binding microarrays measure the affinity of a transcription factor for all combinations of DNA sequences of length k. Utilizing De Bruijn sequence; a microarray is designed in such a way every DNA of length occurs once in one spot. After microarrays are treated with transcription factor, intensity values are used to measure the strength of interaction between DNA sequence in each cell and transcription factor. These PBMs are called universal PBMs. Custom PBMs that only contain in-vivo DNA sequences of interest have been also been designed [35].

3C and ChIA-pet assays aim to characterize three dimensional interactions between regions of chromatin. 3C assay works by first establishing crosslinking between interacting DNA fragments, followed end to end ligation of these fragments. After ligated fragments are sequenced, interacting regions can be determined by mapping them back to genome. ChIA-PET assay adds an immunoprecipitation step to 3C protocol. Therefore it is used to detect three dimensional chromatin interactions caused by a protein-DNA interaction.

1.4 Goals and contributions

The main goal of my work is to design and characterize DNase-seq footprinting approach for its ability to detect TF-DNA interactions and its potential in getting a better understanding of TF-DNA interactions. By using DNase-seq in conjunction with ChIP-seq data when it’s available, I profiled transcription factor binding activity under different conditions and in different cell types. Furthermore, I built DNase-seq
footprinting models that aim to purely model DNase-seq footprints that can correct for assay specific biases. With the development of this algorithm, we aimed to robustly assess DNase-seq footprinting’s ability to detect genome-wide transcription factor binding and other possible applications. These possible applications include identification of variation in consensus binding sequence and characterization of different TF binding modes (direct vs. indirect). Coupled with DNase-seq footprinting approaches, I also focused on building a custom peak caller for ChIP-seq experiments for which shift size cannot be estimated.

1.5 Outline

Chapter 1 describes data parsing and preparation steps I use to analyze various kinds of genomic assays that are available.

Chapter 2 describes use of aggregated DNase-seq footprints and their ability to detect changes in condition or cell type specific transcription factor binding. I describe ways to detect variation in the footprint signal that possibly corresponds to different ways a transcription factor may interact with DNA. Cell type specific footprint study was published in Natarajan et al. [36] and condition specific footprint study with footprint characterization was published in Tewari et al [37].

Chapter 3 describes a novel DNase-seq footprint model that detects individual DNase-seq footprints and assesses the accuracy of model thoroughly. I used this model
to analyze variation in sequence motifs and to characterize possible different modes of binding for the transcription factor. This work is under revision as Yardımcı et al.

Chapter 4 describes my work in investigation of changes in chromatin that result from knockdown of transcription factor CTCF. My footprinting model was used to validate loss of CTCF binding that result from CTCF. Global changes in chromatin were characterized by using RNA-seq and DNase-seq experiments. Manuscript for this work is in preparation.

Chapter 5 describes a custom ChIP-seq peak caller I designed to handle ChIP-seq experiments with no obvious shift size. I applied this peak caller to analysis of transcription factor landscape of two transcription factors in Arabidopsis thaliana. Manuscript for this work is in preparation.

Chapter 6 provides overall conclusions of my work and possible future directions.
2 Data preparation

This chapter describes experimental and computational protocols for datasets that used for modeling and analysis.

2.1 DNase-seq

DNase-seq experiments were performed on multiple different cell types as described in [21]. Shortly, nuclei were extracted and digested with optimal concentrations of DNaseI. Following digestion, resulting DNA fragments were blunt ended and ligated to a biotinylated linker. DNA fragments that were attached to the linker were isolated. Following digestion with MmeI, they were captured using streptavidin-conjugated magnetic beads. A second linker was ligated to the MmeI-digested ends, and DNA fragments were amplified and subsequently purified via gel electrophoresis. Resulting libraries were sequenced and aligned to the human reference genome (NCBI Build 37) using BWA [38]. Alignments were filtered to remove alignments against problematic repetitive regions such as alpha satellites and PCR artifacts characterized by many sequences mapped to small genomic locations as described in [20].

Resulting aligned fragments indicate a DNaseI cleavage event between the 5’ most DNA base pair it maps against and the base upstream. Therefore a DNase-seq experiment yields a vector of DNaseI cleavage counts across the genome. Depending on
the model and the analyses we perform, count data was directly used in most cases. For peak discovery, we transformed count data as explained in the following section.

2.2 Detection of DNase Hypersensitive Sites

To identify regions that are enriched for DNase-seq fragments - the so called DNase hypersensitive sites (DHS) – we use F-seq algorithm [22]. F-seq transforms discrete DNaseI cleavage count data into continuous probability density by kernel density estimation. DNaseI cleavage counts are assumed to independent and identically distributed samples and are transformed using a Gaussian kernel function with mean 0 and variance 1.

For all DHSs used, default parameters were used (such as bandwidth parameter =600). As estimation of density using all samples is computationally expensive and exceeds available precision of computing systems, F-seq uses shifting windows that depend on the bandwidth parameter, typically in the range of few thousand kilobases for human genome.

To discover regions that are significantly enriched for density signal (DHSs), F-seq estimates the background signal using uniform distribution of DNase-seq fragments. Performing this sampling process a large number of times yields a normally distributed kernel density estimates. Threshold for calling a DHS is s standard deviations above the mean of this normal distribution.
2.3 **Identification of transcription factor binding sites using ChIP-seq**

For numerous analyses we performed, ChIP-seq experiments were used as the gold standard method to detect transcription factor binding. Recently a wealth of ChIP-seq experiments for numerous factors in different cell types has been performed under the ENCODE consortium by various groups [39, 40]. ChIP-seq experiments are processed by peak-calling algorithms which identify regions with a significant enrichment of ChIP-seq fragments (peaks). These peaks are used to define the region surrounding transcription factor binding sites. ChIP-seq peaks reported different groups are identified by different peak calling algorithms, introducing an extra level of variability to results are available online. Typically ChIP-seq experiment for a single TF is at least performed twice for two biological replicates and these two replicates are analyzed separately by peak calling algorithms.

To rule out variation resulting from use of different peak calling algorithms and combine multiple biological replicates in a principled manner, ENCODE consortium has published guidelines and a pipeline to standardize peak calling procedures [41]. Resulting ChIP-seq peaks called by this pipeline are called ENCODE uniform peaks. To avoid variation in peak calls due to methods and different control assays, we also adopted and used ENCODE uniform pipeline set of blacklist filtered ChIP-seq peaks called by SPP peak caller [10]. These peak calls are available at [http://www.broadinstitute.org/~anshul/projects/encode/rawdata/peaks_spp/mar2012/di](http://www.broadinstitute.org/~anshul/projects/encode/rawdata/peaks_spp/mar2012/di)
For the purposes of ranking ChIP-seq peaks in terms of strength of ChIP-seq signal, we used signal enrichment value reported by the SPP peak caller.

### 2.4 Scanning the genome for candidate binding sites

To identify candidate binding sites of transcription factors, we scan the human genome (hg19 assembly) using position weight matrices (PWMs). PWMs model the occurrence frequency of each base pair independently within a transcription factor binding site (TFBS), with a typical size between 6 to 20 nucleotides for a eukaryotic transcription factor. To eliminate zero entries in PWMs, we add pseudo counts (0.0005) to each entry in the PWM. We slide a window of the size of each PWM across the genome and measure the similarity of each window to the TF binding motif using PWM log-likelihood scoring, calculating the log likelihood ratio of that nucleotide being generated by the PWM as opposed to a background frequency model. In this scoring scheme, we call windows that are assigned a positive score as “sequence motif matches”. For the background frequency model, we use a first order Markov model that is estimated from a 500bp window centered at the currently considered window. Compared to a fixed zero-order background model for whole human genome, the local background model allows us to more realistically account for nucleotide fluctuations across genomic domains, and accounts for higher-order nucleotide dependencies such as CpG islands [42].
2.5 RNA-seq

To characterize gene expression, we performed RNA-seq experiments. Total RNA was isolated using TRIzol Reagent (Invitrogen) and further purified with on-column DNase digestion using Qiagen RNeasy (Qiagen) kit according to manufacturer’s protocol. Two micrograms of total RNA were used for standard TruSeq library preparation with polyA selection (performed by the Duke Genome Sequencing and Analysis Core Resource) for mRNA sequencing using 2 x 50 bp paired-end reads on an Illumina HiSeq 2000.

RNA-seq reads were aligned to the mouse genome (NCBI37/mm9) using Tophat v2.0.8b [43, 44], allowing for up to 2 mismatches with UCSC gene transcriptome-guided mapping. Normalized Fragments Per Kilobase of transcript per Million mapped reads (FPKM) expression values were obtained for reference transcripts via Cufflinks, and pairwise differential gene expression tests were carried out with Cuffdiff v2.0.2 [43].
3 Aggregated DNase-seq footprints identify cell-type or condition specific transcription factor binding

3.1 Introduction

Transcription factors act as one of the central actors of transcription regulation by binding to different regions of the genome. TF binding changes the state of chromatin and controls transcription of genes [45]. By regulating gene expression, TFs enable various complex processes such as cell cycle progression [46], developmental temporal processes and cellular differentiation [47, 48]. Some of these processes require transcription factors to control gene expression in a cell-type specific manner. Understanding the mechanisms through which transcription factors achieve cell-type specific expression is a key question in biology.

To detect and characterize transcription factor binding activity, different high throughput methods are used, such as chromatin immune-precipitation (ChIP) followed by microarray or sequencing (ChIP-chip and ChIP-seq) [7, 8]. ChIP based methods allow identification of large regions (typically a few hundred bps to 1kb) which are enriched for transcription factor binding events. However they require a high quality antibody for TF of interest for successful IP and each individual TF has to be assayed independently for every cell type. Furthermore ChIP based assays tend to detect a subset of regions of enrichment which may not contain the regulatory element TFs bind to, due to indirect binding events caused by chromatin looping [49] or artifacts [12, 13].
An alternative approach is to identify regulatory sequences TFs bind to (typically called transcription factor binding sites or TFBS in short), which are typically 6 to 20bp long and exhibit a consensus DNA sequence. Such consensus sequences can be characterized by SELEX [3] (Systemic evolution of ligands by exponential enrichment) or PBM [32] (Protein binding microarray) experiments and modeled using position weight matrices [2] (PWM). PWM models predict many TFBSs across the genome but it is known that most of these candidate TFBSs are not utilized. Furthermore binding site predictions purely dependent on DNA sequence of the genome do not yield any information about cell type specific utilization of binding sites.

Chromatin accessibility, measured by DNase-seq, has been shown to be strong indicator for transcription factor binding [16]. Total transcription factor landscape of multiple TFs, detected by multiple ChIP-seq assays, has been shown to reconstitute accessible chromatin landscape reliably [50]. However, a more direct indicator of TF binding is the DNase-seq footprint, which results from direct physical interaction between TF and DNA. Physical interaction caused by TF binding prevents DNase from cleaving DNA at binding site which results in a depleted DNase-seq signal at binding sites [23, 24]. By aggregating DNase-seq signal from DNase-seq experiment around a number of candidate binding sites, peak trough peak like footprint shape may become apparent that suggests genome-wide TF binding in the system/cell type experiment was conducted on.
We used aggregate DNase-seq plots to characterize cell type or condition specific TF binding activity in genome-wide fashion. Aggregate DNase-seq profiles robustly identify binding activity for transcription factors as predicted by experimental setup or computational analyses. Furthermore we identify different sub DNase-seq profiles that make up main aggregate profile for a single transcription factors, suggesting a single TF may interact with DNA in alternative modes, thus leaving footprints with different profiles.

3.2 Aggregate DNase-seq profiles and footprints

Protein-DNA interactions are known to prevent DNaseI cleavage, early DNase footprinting assays are based on this principle. In a high throughput DNase-seq experiment, same principle results in relatively small number of DNase-seq reads mapping to regulatory sequences (such as TFBSs) that are bound by TFs or other DNA binding molecules, such as PolII. This observation may not hold at every putative regulatory sequence for different reasons, such as low DNase-seq coverage or possible lack of TF binding. However a clear footprint shape (peak-trough-peak) becomes obvious when DNase-seq signal is aggregated over multiple regulatory elements, suggesting that majority of these regulatory elements are bound by a TF.

To build DNase-seq aggregate plots, first we identify a set of high confidence candidate binding sites for a specific TF (c.f. section 2.4). Typically such sets are generated by scanning the genome using PWM for the TF of interest and filtering the set
of predicted sites. For ubiquitously active transcription factors such as CTCF, we observe a clear footprint in aggregate plots from many different cell types, showing aggregate plot’s utility in detecting TF binding activity. Even though aggregate plots of active TFs mostly share similar peak-trough-peak shapes, we detect some variation in the overall shape of the footprints (Figure 1), possibly due to fact that different structural families of TFs bind DNA in different conformations [1]. DNaseI also cleavages in a strand specific manner, with footprint shape apparent on both 5’ and 3’ strand DNase-seq fragments (Figure 1A, B) [15]. For simplicity, DNase-seq signal from two strands are combined in our analyses.
Figure 3.1: Aggregate DNase-seq plots around CTCF motif matches, with DNase-seq reads mapping against forward and reverse strands separately (A) and together (B). These two profiles are not necessarily symmetric. REST/NRSF aggregate plot from 50,000 highest scoring candidate bindings sites (C) and aggregate plots using only 10,000 candidate sites (D) or known TFBSs (E). Footprint becomes obvious at higher scoring candidate binding sites and known binding sites.

Footprints in aggregate plots become more defined if candidate binding sites are filtered in a more stringent manner or only true TF binding sites are used. As an example, an aggregate plot built from 10000 (Figure 1D) highest scoring PWM matches
as opposed to 50000 (Figure 1C) shows a cleaner footprint. Similarly, using only PWM matches that are known to be bound by TF of interest (i.e. using PWM matches that map in ChIP-seq peaks) also makes the footprint more obvious (Figure 1E). These observations further establish that footprints in DNase-seq aggregate plots accurately detect genome-wide TF binding over a collection of candidate binding sites.

3.3 Cell type specific DNase-seq footprints

To understand regulation of cell type specific gene expression, we designed a logistic regression model. Using DNase-seq and exon arrays experiments for 19 different cell types, we identified genes that are expressed at significantly high or low levels for certain cell types. DHSs were scanned for binding sites of many transcription factors using PWM models. Logistic regression classifier was trained to classify cell-type expressed genes using PWM scores as independent variables. TFs that had high regression coefficients were considered regulators of cell type specific gene expression [36].

Since ChIP-seq experiments were not available for majority of TFs we identified as drivers of cell type specific gene expression, we instead used DNase-seq aggregate plots to validate binding. In order to build these aggregate plots, we picked the top 10,000 candidate binding site of each TF in the genome. For this 10,000 candidate binding sites, we generated aggregate DNase-seq plots using a 100bp window around the binding site for each cell type. We expected to see footprints in DNase-seq aggregate
profiles for TFs in the cell line in which it predicts gene expression. As a control, aggregate plots were examined as to whether they showed a footprints in other cell lines where they did not have a higher regression coefficient.

![Figure 3.2](image)

**Figure 3.2**: Aggregate plots of DNase-reads around motifs for factors with high regression coefficients. (Red lines) The cell line in which the TF is identified as a regulator. (A) CRX shows a footprint in medulloblastoma but not in the other cell lines shown. (B) REST shows a footprint in other cell lines but not in medulloblastoma, where it is not expressed. (C, D) EGR2 and SPIB show footprints in the GM12878 cell line

For several TFs, we observed indicative footprints in the region of the binding site. For example, CRX is predictive of up-regulated genes in the medulloblastoma cell line and it has a signal depleted region at the motif (Figure 2A). Importantly, in other cell lines such as GM12878, LnCAP and MCF7, it does not have a signal depleted region at the binding site. This is suggestive that the CRX factor is actually bound to DHS
present in the medulloblastoma cell line. Interestingly, there are other factors such as OTX2 that are highly expressed in the medulloblastoma cell line with a similar PWM to CRX. Further, OTX2 is known to be important for transcriptional regulation in medulloblastomas [51]. This highlights a caveat in predicting expression from motifs. Specifically, while we do identify biologically relevant motifs, we can only narrow down the set of potential regulators to the family of factors that can bind to that specific motif.

We identified REST as a regulator in the medulloblastoma cell line. Since, it is not expressed in this cell line, we would expect to not see a footprint in that cell line. This is precisely what we observe with a footprint visible in the other cell lines shown (Figure 2B). Additional footprinting evidence is detected for EGR2 and SPIB in the GM12878 cell line (Figure 2C, D). However, we note that footprints are not always as clear. For instance, SPIB motif also seems to have a small footprint in the LnCAP cell line. This could be due to expression of other factors with a similar motif in this cell line. Overall, DNase-seq aggregate plots prove useful in detecting overall binding activity of TFs and we observe a footprint or lack of a footprint in cell lines where we expect binding or no binding, respectively.

3.4 Condition specific DNase-seq footprints

3.4.1 Androgen Receptor

To characterize condition specific transcription factor binding, DNase-seq aggregate plots were generated and compared between different conditions. For this
purpose, we analyzed DNase-seq experiments conducted on LnCAP cell line before and after activation of androgen receptor (AR). Androgen receptor is a ligand activated nuclear receptor, which acts as a transcription factor. AR locates to nucleus after its activation by androgen, where it can bind chromatin and regulate gene expression. We activated AR using synthetic androgen R1881 (12 hours), performed DNase-seq experiments before and after stimulation and characterized the footprint AR leaves after activation.
Figure 3.3: Base pair resolution around AR motif matches reveals a unique pattern of protection by the AR. (a) Aggregate plot of DNase-seq signal around AR motif matches within poised DHS sites that also bind the AR. The pattern of DNaseI cuts within the motif closely follows the known structure of the AR dimer as well as the information content of the AR DNA recognition motif determined de novo from ChIP-seq sequences that overlap DHS sites. Aggregate DNase-seq signal centered around CTCF motif (b) and NRSF (c) matches genome-wide displaying a structurally different footprint from that of the AR. (d) Aggregate plot of DNase-seq signal around AR motif matches within DHS sites unique to LNCaP-induced cells that also bind the AR. (e) Aggregate plot of DNase-seq signal around the centers of 10,000 randomly sampled DHS sites shared between LNCaP and LNCaP-induced cells. Note that overall the aggregate signal is higher in LNCaP as compared to LNCaP-induced cells within all DHS sites.

3.4.2 Androgen activated DNase-seq footprint

Investigation of aggregate DNase-seq plots revealed an overall increase in DNase signal around AR motifs (Figure 3A) compared to other transcription factor motifs such
as CTCF and NRSF (Figures 3B and 3C). Furthermore we observe symmetrical depletion of DNase-seq signal around AR motifs in DHS sites that closely matches the information content of the AR binding motif dimer (Figure 3A, red line), suggesting that we depletion of DNase-seq signal precisely at base pairs make contacts with AR molecule. For AR binding sites that map to DHS sites before and after androgen induction, we observed a similar pattern of protection despite lower overall DNase-seq signal intensity (Figure 3A, blue line). AR binding sites that become accessible only after androgen induction only exhibited the footprint after androgen treatment (Figure 3D, blue line). Importantly, the overall enrichment of DNase-seq signal in LNCaP-induced cells is specific to DHS regions that bind to AR and have an AR motif, as opposed to all DHS sites (Figure 3E). The observed evidence of AR motif protection prior to androgen induction (Figure 3A) may represent binding of an alternate factor that is displaced upon AR activation, such as has been reported for specific loci by GATA2[52][52][52][52][52][52]. From the compendium of cell lines that have been processed for DNase-seq through the ENCODE project, we identified H1 embryonic stem cells and D721 medulloblastoma cells as having relatively low expression levels of the AR. DNase-seq signal around AR motifs within DHS sites in these two cell lines resemble that of LNCaP cells prior to hormone treatment (Figure 5A), suggesting that such a protection pattern in non-AR activated cell lines could result from alternative transcription factor binding to DNA at these regions.
3.4.3 Multiple AR footprints may correspond to different modes of binding

To further investigate the AR footprint, we performed k-means clustering to search for discrete DNase-seq patterns around AR motif matches. DNase-seq signal was represented by a vector of DNaseI cuts spanning 15 bp around the center of the AR motif. We identified three reproducible clusters, each of which represented part of the observed composite footprint (Figure 4A). These clusters were much less frequently detected across repeated iterations of clustering in uninduced LNCaP cells. To quantify the degree to which these three patterns were present in LNCaP-induced cells compared to untreated cells, we examined the correlation between cluster centers obtained by performing k-means clustering 100 times for induced and uninduced LNCaP DNase-seq data. Specifically, the correlation of each cluster center to the cluster centers from all previous iterations was computed. Correlations tightly distributed around 1.0 represent highly reproducible clusters across different runs, suggesting that the three patterns are robust and consistently observed at AR motifs. Correlations loosely distributed about values less than 1.0 indicate that the three DNase-seq patterns at AR motif matches are less reproducible. We found this correlation distribution to be significantly higher (Mann-Whitney p-value < 2.2e-16) for LNCaP-induced cells (Figure 4B), with the most robust clustering associated with AR binding (Figure 5B) (Mann-Whitney p-value < 0.001 between each column of correlations). Increasing the value of k consistently identified the same three general patterns in LNCaP-induced DNase-seq data within the
AR motif, with multiple clusters aggregating to each general pattern (Figure 5E). Using correlation analysis to analyze clusters from different values of k revealed that k=3 is the most appropriate value (Figure 5F), supporting that three distinct patterns of DNaseI cleavage exist within AR motifs. Overall, the three distinct patterns of DNaseI protection appeared to be a robust phenomenon more often detected in LNCaP-induced DNase-seq data, suggesting that AR activation stabilizes specific chromatin structure around AR motifs.

AR binding has been associated with enrichment of palindromic full-site AR motifs (such as depicted in Figure 3A) as well as half-site motifs [53, 54]. The directional footprinting in clusters 1 and 2 is indicative of only half of the full canonical AR motif being protected from DNaseI cleavage, whereas Cluster 3 is consistent with full-site protection. Our ability to detect this indicates that specific half-site usage is consistent across the entire population of cells, and does not fluctuate randomly. The spike in the center of Cluster 3 corresponds to the degenerate bases in the middle of the AR motif, indicating reduced DNA protection between AR proteins within the dimer. A recent report examining the dynamics of AR dimerization showed in an exogenous system that the AR binding enhancer element of TMPRSS2 requires an AR dimer [55]. Consistently, we observed DNase-seq digestion pattern similar to that shown in Cluster 3 within this enhancer element (Figure 5D).
Figure 3.4: AR binding displays three distinct modes of AR-DNA interaction that are specific to ligand activated AR. (a) K-means clustering of LNCaP-induced DNase-seq signal into three consistent clusters within AR binding sites. (b) K-means clustering (k=3) was repeated 100 times on both LNCaP and LNCaP-induced DNase-seq data around all DHS sites with a full-site canonical AR motif. Shown is the distribution of correlations between cluster centers for each run. Asterisk denotes the statistically significant difference between distributions (Mann-Whitney p-value < 2.2e-16) (c) Motif analysis of the entire 25 bp span up and downstream from AR motif matches for each cluster. MEME motifs identified within this interval (E < 0.1, E-value shown below logo) that significantly match a known motif (E < 0.05, by TomTom) are marked with an asterisk. The name of the most significant match according to TomTom is indicated next to the logo, as is the percentage of regions that contain the enriched motif. For matches resembling Forkhead box family factors (FOX), we note that these motifs are very similar to each other. DNase-seq signal is shown as the aggregate signal from all cluster members with the dotted lines marking the location of the AR motif within the plot.
Figure 3.5: DNaseI protection around AR motifs. (a) Aggregate plot of LNCaP (blue) and H1 embryonic stem cell (purple) and D721 medulloblastoma cell (green) DNase-seq signal around high scoring AR PWM matches within DHS sites in each cell line. (b) Correlation between cluster centers for 100 runs for sub-samples of 1000 regions containing AR motifs. The three clusters observed are most reproducible in LNCaP induced DHS sites that also have an AR binding site. AR activation to the nucleus also creates similar clusters, but to a statistically significant lesser degree. Also shown are the correlations across cluster runs for LNCaP DNase-seq data at the same AR binding and non-binding regions as analyzed for LNCaP induced data. Asterisks indicate significantly different correlation distributions (Mann-Whitney p-value < 0.05). (c) AR ChIP-seq peak score for the Massie and Yu AR binding peaks that contain a cluster 1, 2, or 3 AR PWM match. There is no statistical difference between the AR binding strength in each cluster. (d) Example of cluster 3 AR motif
within the upstream AR binding enhancer of TMPRSS2. This enhancer was predicted to require a full AR dimer to function, consistent with a cluster 3 DNase-seq pattern. Y-axis represents number of DNaseI cuts at each base pair, and axes are fixed at 31 cuts for both LNCaP and LNCaP-induced rows. (e) Cluster centers from k-means clustering with k=6. Note that for clusters 2 and 5, the variation in signal detected falls outside the AR motif itself. (f) Correlation between cluster centers for 100 iteration of k-means clustering with various values of k for DNase-seq signal around AR motif matches. The strongest and most reproducible clusters were found for k=3.

While we posited that full-site protection might reflect a stronger AR-DNA association, AR ChIP-seq peak scores were evenly distributed between the three clusters suggesting similar binding strength (Figure 5C). We next explored if each cluster exhibited different co-factor motif enrichment by de novo motif analysis of the 25 base pairs upstream and downstream of the motif clusters using MEME [56](Figure 4C). Within these intervals, we detected more significant enrichment of Forkhead box family (FOX) motifs in the highly protected portions (dips) of clusters 1 and 2. A motif consistent with NFIC was enriched only upstream of cluster 3. These analyses suggest that the two well-defined dips observed around the composite footprint (Figure 3A) correspond to FOX factor-mediated DNA protection, which is seen to a more noticeable degree in LNCaP-induced DNase-seq data. Overall, our footprinting analysis revealed three different stable modes of DNase-seq protection with AR binding that represent two phenomenon: full or half-site protection at full-site DNA motifs.
3.5 Methods

3.5.1 DNase-seq aggregate profiles

To visualize DNase-seq profiles that are associated with specific transcription factors, we scan the genome for candidate binding sites of a TF using PWMs and aggregate the DNase-seq fragment counts in a window that surround the candidate binding site. Typically the window is composed of 100 bp long regions that are upstream and downstream of the binding site, resulting in a $200 + W$ bp long window where $W$ is the size of the binding site determined by the number of columns in PWM.

Specifically we scan the genome for candidate binding sites (c.f. Section 2.4) and pick highest scoring $N$ sites, where $N$ is typically in the order of tens of thousands. For each candidate binding site, counts of DNase-seq fragments at each position in $200 + W$ bp window are stored in vector $V$. For a set of candidate binding sites, the entirety of count data is stored in matrix $M$ where $i$th row is the vector that corresponds to $i$th candidate binding site. Aggregate DNase-seq profile is obtained by summing over each column of $M$.

For a subset of transcription factors that do not bind to palindromic sequence motifs, aggregate DNase-seq profiles for fragments that map to Watson and Crick strand are different [15]. For the purposes of the analyses conducted here, strand difference did not have an impact on our models and strand specific count vectors have been combined into a single vector.
3.5.2 Clustering of DNase-seq profiles

As name suggests, DNase-seq aggregate profiles are an aggregation of individual DNase-seq profiles around numerous candidate binding sites and not all candidate binding sites have footprint like profiles for different reasons (inaccessible chromatin, lack of transcription factor binding, etc...). To separate footprint profile from noise and discover if main footprint profile is a combination of different sub profiles, we used an unsupervised clustering approach.

For a set of candidate binding sites, we gather DNase-seq count data around each candidate binding site in matrix M. We clustered M using k-means algorithm. We used a smaller window size (+/-25nt) around binding sites to avoid discovery of larger profiles that may not be relevant to distribution of count data in the immediate vicinity of binding sites. To separate footprint profile from background profile, we used two clusters (k=2). For discovery of sub profiles that make up the main profiles, we generally found that k=3 is sufficient. The distance metric of choice for dealing with DNase-seq count data is Pearson correlation, since coverage and depth of sequencing varies across the genome and regions around different candidate binding sites.

3.6 Discussion

Early DNaseI footprinting assay approaches date back 1978 [25] and have been used to detect in vitro protein DNA interactions in a low throughput manner. Recent advent of high throughput DNase based assays on in vivo chromatin, using sequencing
based technologies [20, 57], has allowed us to characterize DNase accessible regions (DNase hypersensitive sites) which overlap with different kinds of regulatory elements. DHSs themselves have been shown to be enriched for transcription factor binding sites [15, 17]. However, DNase-seq footprinting yields more convincing evidence for direct TF binding at higher resolution through detection of signal depletion caused by the protein-DNA interaction at the candidate binding site. DNase-seq footprinting approaches have been used in numerous studies to detect and profile TF binding events in a genome-wide fashion [20, 23, 26, 29].

We used DNase-seq aggregate plots to detect the global genome-wide “footprint” active TFs leave on DNase-seq signal landscape centered on candidate TFBSs. Detection of footprint in a specific cell type yields strong evidence for binding activity of TFs that bind to the TFBS of interest. Footprints can be particularly useful when different assays that detect TF binding, such as ChIP-seq, are missing. For certain TF structural families that tend to bind similar motifs, identity of TF family member that causes the footprint may be hard to infer. For such cases we suggest coupling of DNase-seq data with gene expression data to pinpoint the most likely family member based on its expression levels [26]. DNase-seq aggregate plots can also be used to characterize TF binding activity after activation of TF; such as in the cause of AR. Aggregate plots show a better defined footprint signal after AR activation by androgen.
Even though we see a typical depleted DNase-seq signal at binding site surrounded by two peaks (shoulders), the exact shape and size of shoulders vary amongst different TFs. Furthermore, some TFs leave extended footprints or sub profiles that make up the main footprint profile. Such complexity suggests DNase-seq footprints may be sensitive to variation in the TF-DNA interaction itself. Modeling the variation in footprint profiles in base pair resolution can reveal if it corresponds to a functional variation in binding levels or DNA motif TFs utilize. Base pair resolution footprint models further have the potential to identify individual DNase-seq footprints across the genome and predict each binding site.
4 Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection

4.1 Introduction

The identification and characterization of eukaryotic transcriptional regulatory regions to understand complex gene expression patterns has long been hampered by large genome sizes and regulatory elements that can act over large distances. The recent integration of high-throughput experimental and computational approaches has made it possible to characterize functionally important regions via the identification of chromatin states using a variety of histone modifications [58, 59].

A complementary and unbiased approach to identify candidate regulatory regions has been the mapping of accessible chromatin by DNase-seq, which identifies DNase hypersensitive (DHS) sites across the genome. DHS sites typically span a few hundred bases and have been shown to overlap all types of known regulatory elements, including promoters, enhancers, insulators, locus control regions, active histone modifications and the majority of transcription factor binding sites [20, 40, 50, 57, 60]. The reduction of the entire genome to a set of small putatively functional DHS sites has enabled computational approaches that decode condition-specific expression patterns, for instance by building models that can distinguish the expression of genes based on the sequence features in nearby regulatory regions [36, 61, 62].

A wealth of genome-wide association evidence supports the notion that these non-coding regions of the genome are relevant to disease risk [63-66]. Elucidating the
mechanism of these diseases will require formal understanding of how non-coding sequence variation impacts gene expression levels. To accomplish this goal, it is necessary to know the exact location of where regulatory interactions occur at base-pair resolution, such as binding of transcription factors to specific DNA elements. For decades, DNase footprinting – the identification of regions of local DNA protection from DNase cleavage - has been the method of choice to identify transcription factor binding sites. However, this method is laborious and only focuses on a single small region of the genome [25]. With sufficient sequencing coverage now available, DNase-seq has the potential to pinpoint DNase footprints all across the genome by locating relative depletion of DNase-seq signal at a candidate binding site compared to the flanking DNA. Depending on the precise interaction of a TF with its cognate binding site, footprints for different factors differ from each other and display distinct cleavage profiles [15, 36]. This observation raises the possibility of learning TF-specific footprint models through intersection of ChIP and DNase data from the same cellular state, and using this data to predict binding sites for these TFs in other cell types and conditions where only DNase-seq data has been collected.

This application of DNase-seq has spurred a number of studies for predicting functional transcription factor binding sites (TFBS) from three directions, which make different assumptions but are often not clearly distinguished. The simplest method assesses only the number of DNase-seq reads that surround a candidate binding site
While robust, this method does not reflect the footprint shape and specific DNase-seq cleavage profiles around the binding site. A second strategy uses a variety of approaches to model and identify footprints [23, 24, 26, 29]. These de novo approaches generally assume that all transcription factor binding events lead to similar generic footprint shapes, and do not generally require knowledge of the sequence specificity for each TF. A third method integrates DNase signal strength and shape with conservation and other sequence features, all within a relatively large window size (200 bp), to identify footprints via a hierarchical mixture model [15]. While this method utilizes DNase footprint shape, its predictions are also driven by other features such as general chromatin accessibility across the entire window and sequence conservation.

One factor that complicates the use of DNase to reliably detect footprints is its sequence cleavage bias, which was originally described over 30 years ago [19] and has been recently revisited in the context of genome-wide studies [30, 31, 67]. Here, we describe a method that solely models DNase footprint shape in a high resolution, transcription factor-specific manner. Focusing on a small window (50bp) surrounding candidate binding sites, this method simultaneously accounts for DNase sequence cleavage bias to distinguish true footprints from background effects. Together, this allows us to identify high confidence transcription factor binding sites at high resolution. Further, we demonstrate advantages of footprinting in addition to chromatin
immunoprecipitation on untangling direct and indirect binding and pinpointing transcription factor binding sites at single-nucleotide resolution.

4.2 Results

At least four different scenarios arise when searching for bona fide DNase footprints at candidate TF binding sites, which we define as DNA sequences that match the sequence preferences of a specific TF (i.e. a sequence motif match). We provide examples of these scenarios using NRSF ChIP-seq and DNase-seq data from the GM12878 lymphoblastoid cell line [8, 68]. First, true positives are sequence motif matches that overlap both a DNase footprint and a ChIP-seq peak for a TF associated with the sequence motif. These are highly likely to represent direct binding sites (Figure 1A). Second, true negatives are sequence motif matches without a DNase-seq footprint that do not map in a ChIP-seq peak (Figure 1B). Third, ChIP may not have the resolution to tell apart which one of two sequence motif matches is indeed bound, but this may be resolved by the presence of a footprint (Figure 1C). Fourth, sequence motif matches that overlap ChIP-seq peaks but do not exhibit a DNase-seq footprint (Figure 1D) may represent weak or indirect binding of TFs, long range chromatin looping [49], or simply artifacts due to false positive ChIP-seq peak calls [12, 13]. Together, these scenarios illustrate the challenges of identifying footprints and the motivation behind our modeling approach.
4.2.1 ChIP-seq sites display a heterogeneous mixture of DNase digestion profiles.

To show that DNase-seq generates unique digestion profiles that are TF-specific, we generated aggregate DNase-seq profiles by summing DNase signal centered on sequence motif matches within ChIP-seq peaks. This is exemplified by three well-studied TFs, including CTCF (Figure 2A), STAF (Figure 2B), and NRF1 (Figure 2C). Unexpectedly, we observed that a simple unsupervised k-means clustering of these aggregate DNase-seq profiles leads to two distinct aggregate DNase-seq profiles: one with an obvious peak/trough/peak footprint shape, and a second more flattened signal that retains a distinct shape in the absence of a footprint (Figure 2D). This indicates that while each TF has its own distinct DNase footprint shape visible in aggregate plots, aggregate plots are a mixture of footprint digestion profiles and a background digestion profiles.

4.2.2 Sequence bias of DNase-seq is protocol specific

We hypothesized that background profile observed in the k-means clustering reflected recently reported intrinsic cleavage biases of DNase I [30, 67], which may affect a portion of the DNase-seq aggregate profiles [31]. To accurately separate footprint profile from background bias profile, we measured the intrinsic sequence cleavage bias of DNaseI by performing DNase-seq (Song and Crawford, 2010) on deproteinized genomic DNA from K562 and MCF7 cells, and calculated the relative frequency of DNase cleavage for all possible DNA 6-mers. We normalized these relative frequencies
of cleavage by dividing them by the relative frequency of each 6-mer across the genome. This ratio of relative frequencies for each 6-mer, which we call cleavage propensity, represents any positive or negative bias DNase I has compared to what would be expected from the prevalence of each 6-mer across the genome. These cleavage propensities ranged over two orders of magnitude and were highly reproducible between genomic DNA isolated from the two cell lines (0.99 Spearman correlation), ruling out possible cell type specific biases such as different DNA methylation patterns (Fig 3A). We also estimated cleavage propensities on a published DNase-seq data set generated from deproteinized IMR90 cell genomic DNA [67] that used an alternative DNase-seq protocol [23, 69]. We observed that the 6-mer cleavage propensities between DNase-seq protocols were far less correlated (Figure 3B, 0.75 Spearman correlation), suggesting that the two protocols each display significant differences in sequence bias [30]. Therefore, accurately modeling footprints using different DNase-seq protocols will likely require using protocol-specific deproteinized DNase-seq datasets. For this study, we limited the modeling and analysis to data that use the single cut, high molecular weight DNase-seq protocol [21].

4.2.3 DNase footprints are an aggregation of a footprint and background profiles

The observations described thus far motivated us to develop TF specific footprint models to assess the predictive accuracy of DNase-seq footprinting alone that also takes into account DNase cleavage bias. Specifically this method would not include additional
chromatin accessibility or genomic features that may indicate the general presence of a regulatory region (Figure 4). Following the footprint modeling strategy used in earlier approaches such as CENTIPEDE [15], our models sought to reflect the relative propensity of DNase-I cleavage at each position around sequence motif matches using factor-specific multinomial distributions. In addition, to accurately quantify the extent of DNase sequence bias, we incorporated a separate non-uniform background model that accounts for variability in signal profiles in the absence of functional footprints at candidate binding sites. This intrinsic sequence bias background model is estimated from the DNase-seq profile that would result from the selected DNA sequences alone, i.e. the relative cleavage propensity of each 6-mer surrounding the sequence motif matches (Figure 3A).

To train these models, we used expectation-maximization (EM), which is an unsupervised training algorithm for mixture modeling. As Figure 1A and 1D illustrate, ChIP-seq data does not distinguish between true and false positives, or in other words, direct from indirect binding. We therefore used ChIP-seq data to initialize the true positive motif set, but allow the model to reassign motifs to the background should they display evidence of being derived from cleavage bias. This approach ensures that bound sequence motif matches that follow the background DNase profile were excluded from training of the footprint model, leading to a cleaner footprint model.
Importantly, to avoid conflating the DNase-seq signal in the larger genomic region with the one at an individual site, we use a relatively small window (+/- 25bp) around the motif match. This is a more targeted approach compared to previous approaches that opted for a 200 bp or larger window [15, 17]. We expect the smaller window to better exclude influences from neighboring footprints from the same or different TFs.

4.2.4 DNase footprints predict direct binding events with high sensitivity

To assess the performance of using footprint models alone to predict TF binding at candidate binding sites, we set up a binary classification scheme using ChIP-seq data as the gold standard (Figure 4). To discover candidate binding sites (sequence motif matches), we scanned the human genome (hg19) using 21 PWMs from the JASPAR and TRANSFAC databases (Table 2). These PWMs were largely chosen based on their availability of ChIP-seq data from more than one cell type. Following model training with EM algorithm, each sequence motif match in the test set is assigned a footprint score based on the log-odds ratio of footprint and background components. We call this log-odds ratio the Footprint Likelihood Ratio (FLR). In the case of multiple sequence motif matches within a single ChIP-seq peak (Figure 1C), we assigned the highest FLR value to each of the sequence motif matches.

We next compared the performance of FLR footprinting scores against the simple D-s score of overall DNase hypersensitivity [17], which is defined as the number of
DNase-seq reads within a 200bp window around the sequence motif match. Therefore, the D-s score predicts binding without modeling the footprint. We again used the highest D-s score in cases of multiple sequence motif matches within a single ChIP-seq peak. D-s has been reported to perform as well as CENTIPEDE [17], which predicts whether a sequence motif match is bound by jointly modeling the DNase footprint, overall DNase hypersensitivity with a 200 bp window, and other static genomic features such as conservation. We did not compare performance of FLR to CENTIPEDE since CENTIPEDE models multiple features over a larger window and our objective is to quantify the predictive power of DNase-seq footprinting alone. We could not also compare performance against another recent footprint metric [26] because of the unavailability of parts of software implementation. Following previous comparisons with Cuellar-Partida et al., we evaluated the predictive performance via the area under the ROC curve (auROC) and the sensitivity at 1% false positive rate (FPR) threshold. AuROC assesses how well the classifier separates positives from negatives across different thresholds, whereas the sensitivity at 1% FPR assesses the classifiers’ ability to identify true positives at a relatively low level of false positives.

For most factors, the majority of unbound sequence motif matches map outside DNase hypersensitive sites. By definition, D-s score separates sequence motif matches within DHS sites from those that fall outside of DHS sites. Thus, as anticipated, D-s outperformed our pure footprint model in terms of auROC and sensitivity when we
consider all genome wide sequence motif matches (Figure 5). Due to the generally low coverage of DNase-seq signal outside DHS sites, FLR cannot perform well since this model requires DNase-seq coverage at multiple positions in a four times narrower window (50 bp vs. 200bp) without aggregating the signal. Limiting the classification of sequence motif matches to those occurring only within DHS sites, the two approaches perform comparably, with D-s score still outperforming FLR in auROC (Figure 6A), but FLR outperforming D-s in terms of sensitivity at 1% FPR (Figure 6B). Thus, while D-s is reliable in generally distinguishing bound from unbound sequence motif matches, there is a subset of bound sequence motif matches exhibiting strong DNase footprints that lead to higher sensitivity at 1% FPR using the FLR.

Predictive performance of FLR and D-s score varied between the 21 transcription factors, and between different transcriptional factor families (Table 1). For instance, some of the best scoring FLR footprints belonged to zinc finger family members (e.g., CTCF, NRSF, NRF1), which are typically characterized by unique sequence preferences corresponding to high information content PWMs. In contrast, other factors that are part of the basic helix-loop-helix transcription factor family (e.g., Max, C-Myc) tend to bind variations of a core motif such as the E box. Therefore, learning specific footprints for members of such families is more difficult since scanning the genome with a PWM that corresponds to one family member inevitably introduces matches to other family members, confounding the model. Another challenge for high resolution motif based
footprint models is that the sequence motif for certain transcription factors may not be
correct. For E2F1, a TF that binds to a motif very similar to E2F4, scanning the genome
with 8-mers from PBM experiments yields an aggregate DNase-seq plot with much
deeper trough compared to the PWM scan (Figure 7). This demonstrates that 8-mers
from PBM experiments are better suited to discover candidate binding sites for some
TFs, and further illustrates that accurate footprinting can be dependent on the class of
transcription factor and the accuracy of the known sequence motif.

4.2.5 ChIP-seq peaks without a DNase footprint suggests weak or
indirect binding

The success of footprinting methods will clearly depend on the amount of data
available to detect local DNase patterns. This trend is illustrated in a tripartite structure
of the auROC curves (Figure 8). After a quick rise in true positives for high FLR scores,
the performance quickly flattens out; indicating that for a considerable subset of
candidate sites there is not sufficient data to make a clear call as to whether a footprint is
detected. As evidenced by the sudden rise of true positive rate at high false positive
rates, the FLR mixture models identify a third subset of ChIP-seq peaks that do not
exhibit a DNase footprint, even though they map within a DHS site.

To ensure sequence bias was not a driving factor for this distribution of scores,
we ranked sequence motif matches according to FLR and estimated the DNase-seq
signal that would result purely from DNase sequence bias using cleavage propensity
values. Estimated signal was similar for upper and lower quintiles of FLR ranked
sequence motif matches, showing that the footprint profile is not caused by sequence bias itself (Figure 9).

We investigated possible reasons for the lack of footprints at sequence motif matches that are ChIP-seq positive. For this analysis we adopted a ChIP-seq peak-centric rather than motif-centric approach and assigned the maximum FLR score found within each peak. ChIP-seq peaks without a sequence motif match were excluded from this analysis. For all transcription factors we analyzed, there is a positive correlation (ranging from 0.07 to 0.54, Spearman) between FLR and ChIP-seq enrichment value. Sorting ChIP-seq peaks according to FLR shows a consistent decrease of the median ChIP-seq enrichment for all factors (Figure 10A). This trend suggests that direct interactions between the TF and candidate binding site is reflected by higher FLR scores within stronger ChIP-seq signals. Absence of the footprint may therefore suggest indirect binding events or weaker interactions between the TF and the sequence motif match.

To further explore whether high FLR scores indicate direct interactions, whereas low scores are indicative of indirect binding, we compared footprint enriched and depleted ChIP-seq peaks to ChIP-seq data from 83 other DNA-binding proteins. We characterized significant associations with footprint-enriched and depleted peaks as direct associations and indirect associations, respectively. For example, NRSF ChIP-seq peaks (Figure 10B) enriched with strong NRSF footprints were associated with binding
to CoRest (K-S test, $P < 1.4\times10^{-12}$) and Znf143 ($P < 5\times10^{-71}$; Figure 10C) occupied sites. CoRest is a known co-regulator of NRSF [70]. For NRSF ChIP-seq peaks that are depleted for NRSF footprints, a significant association was detected for PolII ($P < 1.4\times10^{-20}$) and TBP associated factor 1 Taf1 ($P < 2.9\times10^{-15}$; Figure 10D). We performed a similar analysis for CTCF ChIP-seq peaks that are enriched for strong CTCF footprints, and a significant direct association was detected for YY1 ($P < 7\times10^{-171}$) and ZNF143 ($P < 7\times10^{-117}$). Both YY1 and ZNF143 have been previously documented to bind together with CTCF [71, 72]. CTCF ChIP-seq peaks that do not contain strong CTCF footprints are also enriched for PolII ($P < 1.4\times10^{-7}$) and Taf1 ($P < 1.5\times10^{-7}$). For CTCF and NRSF, the indirect association with PolII and Taf1 may indicate indirect binding or chromatin looping with the basal promoter machinery that result in a TF-DNA interaction detected by ChIP-seq, but does not leave a footprint. Interestingly, many of these binding sites with the weakest FLR footprint scores have a modest increase in ChIP-seq intensities relative to FLR footprints with stronger scores (Figure 10), indicating that this indirect interaction with the basal machinery is stable. Another possible explanation for this observation are possible ChIP-seq data artifacts; recent studies have reported that regions in the yeast genome going through active transcription are enriched for unrelated transcription factors, indicating that some ChIP enrichments are due to nonfunctional binding [12, 13].

Footprint enriched associations are factor-specific as expected; however we detected significant (threshold $P < 5\times10^{-2}$, Bonferroni correction) indirect associations with
P300 for seven factors and Ikzf1 for five factors. P300 is a well-known transcription co-activator often marking sites with enhancer activity by acting as an acetyl-transferase to increase histone acetylation levels, promote relaxation of local chromatin structure, and recruit transcriptional machinery [73, 74] whereas Ikzf1 is associated with chromatin remodeling proteins [75]. Indirect association with P300 suggests a similar mechanism where active transcription may introduce artifacts into ChIP-seq data whereas Ikzf1 association indicates we may not be able to detect footprints in regions of active chromatin remodeling.

4.2.6 DNase footprints detect microarchitecture changes in chromatin landscape

The above analyses support that ChIP-seq data likely represents a combination of both direct and indirect binding, which cannot be distinguished using ChIP-seq data alone. While overall DNase hypersensitivity (e.g., measured by D-s) and DNase footprints (e.g., measured by FLR) present themselves as equally good predictors of TF binding events across the genome (Figure 6), there is a tradeoff between the two approaches. Specifically, while D-s is a good indicator of detecting either any binding event (either direct or indirect) around the sequence motif match, FLR has the potential to distinguish between indirect and direct binding events. This becomes important in situations where differential TF binding event does not change the overall chromatin accessibility, but does change chromatin microarchitecture.
To explore this scenario, we used DNase-seq data from different cell types for which it is known that specific transcription factors are differentially expressed. We first compared DNase-seq data from the medulloblastoma cell line D721 and the lymphoblastoid cell line GM12878. In medulloblastoma cells, the NRSF repressor is not expressed, which results in expression of neuronal genes [76]. In GM12878 cells, the presence of NRSF actively represses these genes. We identified NRSF sequence motif matches that mapped within DHS sites that are present in both cell lines. Within these regions, we detected that the NRSF footprint was often only detectable in GM12878 cells at individual loci (Figure 11A), and at all potential NRSF binding sites (Figure 11B). We computed the overall distribution of FLR and D-s scores for NRSF sequence motif matches that map in DHS sites that are present in both cell types (Figure 11C). When comparing NRSF motif matches between GM12878 and D721 cell types, we only observed a significantly different distribution of FLR scores (K-S test, p < 3.15e-12) while D-s score differences were not significant (p < 0.4). This indicates that there is a class of regulatory elements (i.e. DHS sites present in multiple cell types) where the D-s statistic is not optimal for accurately identifying differential TF binding sites.

To test whether this advantage of FLR was specific to NRSF, we also compared DHS sites from skin fibroblasts that were transformed into induced pluripotent stem cells (iPSCs) by overexpression of Oct4, Sox2, Klf4 and c-Myc Yamanaka factors [77]. When we compared the DHS sites that were present in both the iPSC lines and the
parental skin fibroblast cell lines, the difference between the two cell types was more significant for FLR footprint scores for each of the four Yamanaka factors than D-s scores (Figure 11D and Table 3).

The ability of FLR to distinguish a change in binding events in shared DHS sites therefore appears to be a distinct advantage of FLR over a more general statistic that only captures overall DNase hypersensitivity in a large window around the sequence motif match. In cases where the disappearance of a footprint does not abolish a DHS site, our footprint model can be used to detect subtle but functionally significant changes in the chromatin microarchitecture landscape.

4.2.7 Unsupervised EM training can identify de novo DNase bias

To illustrate the robustness and flexibility of the TF-specific mixture modeling approach, we evaluated how well it would allow us to separate bound from unbound sites, even in the absence of the intrinsic sequence bias detected using the deproteinized naked DNA DNase-seq data, using unsupervised EM training of mixture components. Rather than using two mixture components for a footprint and background, we often observed that the EM algorithm frequently split the foreground footprint signal into two separate footprints, one with a stronger DNase-seq signal upstream of the motif and another with stronger DNase-seq signal downstream of the motif. In these cases, we combined the two separate de novo foreground footprint components into one
according to their mixing weights, and learned a distinct de novo background model from the third component.

We next compared both de novo foreground footprint and background model components to the previously described intrinsic sequence bias background model. For most of the TFs analyzed, the de novo background model had a higher Pearson correlation to intrinsic sequence bias background model than foreground footprint model (Figure 12A). The foreground footprint model largely correlated negatively or very weakly with the sequence bias background, indicating that EM was able to identify a distinct, non-background subset of footprints that is clearly not due to sequence bias (Figure 12A, upper left quadrant). However, some factors did not allow for a clear separation between the foreground footprint and de novo background model, with respect the intrinsic sequence bias (Figure 12A, upper right quadrant). This may be due to these TFs either leaving weak footprints, or that binding sites for these TFs strongly resemble the cutting bias preferences of DNaseI. Furthermore, using de novo background instead of intrinsic sequence bias background yields a slightly better performance for the classification pipeline on the average, suggesting that de novo background is a better estimation of the true background signal than using the intrinsic sequence bias background (Table 4).
4.2.8 DNase footprints can identify variants of the main sequence motif

While the two foreground footprint components were symmetric for most TFs, we found that this was not the case for CTCF and ZNF143. For these factors, we detected some foreground components that exhibited an extended footprint profile beyond the core motif (Figure 12B, C). These extended footprints suggest additional interactions within a subset of the TF binding sites. Building extended PWMs on the sets of sequence motifs that make up each footprint component, we found a variant of the main CTCF motif (Figure 12B) that has been reported by previous footprint studies [24], ChIP-seq [78] and ChIP-exo experiments [79]. We detected a similar extended motif for ZNF143 that has not been previously characterized (Figure 12C). As these factors are members of the zinc finger family, it is not surprising to find extended variants of the main motif, which can be bound by additional domains that do not make contact with the core sequence motif. This finding demonstrates the usefulness of unsupervised mixture models in which footprint model components reflect different modes of binding and can detect ways that certain TFs interacts with either the core or extended consensus motif.

4.3 Methods

4.3.1 DNase-seq assay and sequence bias of DNase

To quantify the DNase I sequence-dependent cleavage bias, we performed DNase-seq experiments using deproteinized DNA from K562 and MCF7 cell lines.
Deproteinized genomic DNA was isolated from ~20 million K562 or MCF7 cells (cultured in RPMI-1640 or DMEM media + 10% FBS, respectively) using a DNeasy Blood and Tissue kit (Qiagen). Isolated total DNA was then subjected to two phenol-chloroform extractions followed by ethanol precipitation to ensure removal of bound protein. DNase-seq experiments were performed essentially as previously described [21] with the following modifications: DNase I digestions were performed in 120μL solution with ~50μg of deproteinized DNA. DNase digestions were stopped by addition of 50mM EDTA and 15 min incubation at 75°C. Digestions performed with 1.2U and 2.4U total DNase I were selected based on .8% agarose gel sizing to be similar to typical DNase-seq assay digestions and pooled for the two cell lines separately. Libraries were constructed from pooled digests as described and sequenced on the Illumina HiSeq2000 platform with 50 bp single-end reads at the Duke Sequencing and Analysis Core Resource. This protocol generates essentially fixed-length 20 bp insert fragments due to MmeI cleavage of DNA 20 bp from the MmeI recognition site included in the first ligated linker. Therefore, the first 20 bp of Reads were aligned against hg19 assembly using Burrows-Wheeler aligner (BWA), allowing up to 2 mismatches and multi-reads mapping up to 4 locations. Aligned reads were further processed through a pipeline to remove potential PCR artifacts to match standard DNase-seq analysis previously applied to all Duke generated DNase-seq libraries and as described previously in [20, 21]. This resulted in approximately 140 million mapped reads per cell line.
After sequencing tags were aligned, we counted the number of DNA 6-mers centered at each DNase cleavage site (between the 3rd and 4th bp) and calculated the relative frequency of each 6-mer. These relative frequencies are normalized by the relative frequency of each 6-mer in the human genome and genomic relative frequencies are corrected for regions that are mappable using our DNase-seq protocol. This ratio is called cleavage propensity of the 6-mer. As there is no chromatin structure or transcription factor binding events in such deproteinized cell lines, any deviation from one indicates an increased or decreased propensity of DNase cleavage for a 6-mer.

We use cleavage propensity values to generate the background signal profile around a set of motif matches that would solely result from DNase sequence bias. More specifically, DNase-seq signal at each base pair is estimated as cleavage propensity of the 6-mer centered at the current position (i.e. corresponding to the 4th base of the 6-mer). For a set of sequence motif matches, we first obtain the DNA sequence surrounding each sequence motif match. We estimate the aggregate sequence bias background signal around that set of motif matches by iterating over each base pair and each sequence motif match.

4.3.2 Modeling of DNase-seq footprints

DNase-I footprints reflect the propensity of a DNase cleavage at each position at and around the transcription factor binding site; with low probability of DNase cleavage at the binding site due to protection of transcription factor and a high probability of
cleavage surrounding the binding site. To capture different positional propensity of cleavage due to transcription factor-DNA interaction, we modeled the relative probability of a DNase cleavage at each position using a TF-specific multinomial distribution. For training and testing our models, we used ENCODE Duke DNase-seq data from GM12878 cell type [39].

Any given candidate TFBS identified by a PWM match may be bound or unbound in the particular condition for which we have data. At bound locations, we assume to observe a footprint; unbound locations should not resemble the footprint and are therefore represented by a background model, which is also modeled by a multinomial distribution.

While many motif matches in ChIP-seq peaks are enriched for footprints, some resemble the background component. DNase-seq signal around remaining motif matches that are not in ChIP-seq peaks follow the opposite trend, being enriched for background component. Since motif matches are made of a mixture of footprint and background, we used a multinomial mixture model to learn both models simultaneously. Even with ChIP data as a starting point, a mixture model is able to correctly assign motif matches in ChIP-seq peaks that do not have a footprint-like signal to the background, while estimating the parameters of footprint model, allowing us to arrive at a cleaner footprint model. Representing the DNase-seq signal in an M base pair wide window around a motif match as sample X, then
\[ \Pr(X) = \lambda_{fp} \Pr(X|\theta_{fp}) + (1 - \lambda_{fp}) \Pr(X|\theta_{bg}) \]

For a two-component mixture model, with \( \lambda \) representing the mixing weights, and \( \theta_{fp} \) and \( \theta_{bg} \) are the parameters of footprint and background multinomials. We estimate mixing weights and parameters of background and footprint multinomials using the Expectation Maximization algorithm. To initialize \( \theta_{fp} \), we calculate maximum likelihood estimates from aggregate DNase-seq counts around motif matches in ChIP-seq peaks for multinomial model. For \( \theta_{bg} \), we either set the parameters as the maximum likelihood estimates from the 6-mer DNase sequence bias cleavage frequencies, or allow it to be estimated de novo.

We used a two component fixture to learn \( \theta_{fp} \) while \( \theta_{bg} \) was fixed at maximum likelihood estimate from deproteinized naked DNA DNase-seq profile. To learn the de novo background profile, we used a three component mixture because EM has a tendency to fit two symmetric footprints in a two component mixture. In the three component mixture, two components that make up the footprint were combined together to learn the final footprint model, based on their mixture weights. Background component is chosen automatically as the component with most mass at the sequence motif since footprint models have less mass at sequence motif.

For assessing the accuracy of footprint model as a predictive model of transcription factor binding, we use a four-fold cross validation scheme. After footprint
and background parameters are trained, we calculate a footprint log-likelihood ratio (FLR) for each motif match in our test:

\[
FLR = \log \frac{P(X|\theta_{fp})}{P(X|\theta_{bg})}
\]

FLR measures how similar DNase-seq signal around each motif match to footprint model. As the resolution of ChIP-seq does not allow resolving closely spaced motif matches, we use the highest FLR value of all the motif matches (both stringent and relaxed sets) within a ChIP-seq peak. For this task, we used ENCODE uniform pipeline set of blacklist filtered ChIP-seq peaks called by SPP peak caller [10, 41] available at (http://www.broadinstitute.org/~anshul/projects/encode/rawdata/peaks_spp/mar2012/distinct/idrOptimalBlackListFilt/). For ChIP-seq peak enrichment values, we used the signal enrichment value reported by the peak caller.

To allow for a fair comparison, we follow the same evaluation for the D-s score. Following [17], we assess accuracy using two metrics: the area under the ROC curve (auROC) and the sensitivity at 1% false positive rate. We built footprint models for 21 different TFs for which ChIP data and PWMs were already available for Tier 1 ENCODE cell types (Table 2). For comparison of FLR and D-s score distributions from different cell types, we used subsamples with equal number of DNase-seq reads for a fair comparison.
4.3.3 Discovery of Co-binding elements

For each transcription factor with a footprint model, potential co-binding factors were identified by querying for overlapping ChIP-seq peaks of other DNA binding proteins, such as RNA Polymerase II and P300. A set of ENCODE uniform pipeline processed ChIP-seq peaks were utilized for this analysis. Two ChIP-seq peaks are overlapping if the intersection of two ChIP-seq peaks is at least 1bp long. For each TF, we assigned a footprint score to each ChIP-seq peak of that TF, as defined by the maximum of FLR value of all sequence motif matches under the peak.

For each TF in the set of 21 TFs that have footprint models, ChIP-seq peaks that contain a stringent sequence motif match are sorted according FLR and overlapping ChIP-seq peaks of all other DNA binding proteins (i.e. we call this set of proteins associated binders) are identified. To discover if overlapping ChIP-seq peaks of other factors are enriched at peaks with large FLR values (peaks with a DNase-seq footprint) or small FLR values (peaks without a DNase-seq footprint), we used two-sample Kolmogorov-Smirnoff (K-S) test. Specifically, we calculate the significance of the hypothesis that the distribution of FLR scores of ChIP-seq peaks is significantly different from the distribution of FLR scores of ChIP-seq peaks that only overlap with peaks of another associated binder. One-sided tests are performed to discover binders that associate with footprint enriched peaks vs. footprint depleted peaks. Binders that
associate with footprint enriched and depleted peaks are called direct and indirect associations respectively.

4.4 Discussion

Since the advent of next generation sequencing assays, DNase hypersensitivity and DNase footprinting via DNase-seq has been widely used as predictors of transcription factor binding locations in various organisms and cell types. A variety of approaches for sequence motif-based or de novo discovery of DNase footprints have been developed with differences in their definition of a DNase footprint [15, 24, 26, 29]. However, since some of these methods incorporate data that is independent of the footprint, the extent of how well footprinting alone predicts TF-DNA interactions has been unclear. Furthermore, sequence bias in DNase-seq data has recently been documented in multiple studies [30, 67], and has even called into question the overall ability to identify footprints in DNase data [31].

In this study, we have modeled and thoroughly assessed the predictive performance of DNase-seq footprints for 21 transcription factors using models that learn distinct and factor specific footprint shapes with signal depletion centered on sequence motif matches, and include a background model to account for DNase sequence bias. We have shown that sequence motif centered footprints are not always present when the transcription factors are crosslinked to DNA, indicating weaker or indirect TF/DNA binding. Aggregate DNase-seq profiles at motif matches are therefore a mixture of
background and footprint profiles, which likely explains the imperfect predictive
performance of footprinting. Finally, we have shown that the overall footprint shape is
generally not dependent on the bias of DNase enzymatic cleavage and is more likely
caused by direct TF binding. Following the findings of [17], raw DNase accessibility is a
comparable predictor for genome-wide TF binding and is able to generally discern
bound motifs from unbound ones. However, we show a class of regulatory elements
where overall DNase accessibility does not accurately reflect changes in binding to
specific factors, indicating that single-nucleotide resolution footprint models are needed
to accurately identify these types of binding events.

Although ChIP-seq is a well-established method to query protein-DNA
interaction in a genome-wide manner, the extent to which ChIP-seq peaks result from
direct binding or are confounded by indirect binding or noise is unknown [12, 13, 49,
80]. Here, characterizing ChIP-seq peaks based on the presence/absence of a footprint,
we were able to resolve functional differences within the ChIP-seq data, such as indirect
binding events in the vicinity of sequence match that may result from active
transcription around the candidate binding site [12, 13]. Another possible explanation
for ChIP peaks without footprints may be the spatial closeness of two ChIP-seq peaks
due to 3 dimensional conformation of the genome [49]. In the current state, footprint
models may not yet be advanced enough to replace specific ChIP data, but given the
complexity of binding landscape and architecture of the genome, we propose that
DNase-seq footprints may be used in combination with ChIP-seq to distinguish high confidence direct binding events at high resolution.

Finally, for two of the transcription factors evaluated here, the mixture modeling detected more than one distinct footprint profile, coinciding with differences in sequence motifs and suggesting the effect of different binding modes. While we used traditional PWMs that represent all binding sites of a TF in unison, motif variations and differential binding modes may exist, backed by in vitro PBM experiments [32] and ChIP-seq experiments [78, 79]. Further we show that predictive performance of footprinting is higher for transcription factors with highly specific sequence motifs and we show a case where PWM model fails to accurately pinpoint candidate binding sites associated with the factor, confounding our high resolution footprint model. These findings suggest joint learning of DNA sequence and DNase-seq footprint is a promising next step to more accurately detect and characterize TF-DNA interactions in vivo.
Figure 4.1: Scenarios relevant to identifying DNase footprints. On the right, representative examples of DNase-seq data from GM12878 cell type and ChIP-seq data for NRSF from ENCODE [39]. The location of sequence motif match for the transcription factor NRSF is indicated with a yellow box. On the left, a schematic representation of TF-DNA interaction is shown and whether a footprint is detected or not detected at the motif match. A) A DNase footprint centered at the motif maps within a ChIP-seq peak indicating a direct binding event. B) A motif that maps within a DHS site, but has no appreciable ChIP-seq signal, nor footprint, indicating no interaction between TF and sequence motif match. C) Multiple sequence motif matches within a DHS site may only have a single footprint, showing that TF may be more likely to interact with one of the motif matches. D) ChIP-seq peak with a sequence motif match that does not have a footprint suggests a possible indirect binding event.
Figure 4.2: Aggregate DNase plots identify distinct TF binding profiles. Aggregate DNase-seq signal was calculated for motifs that map within ChIP-seq peaks for A) CTCF, B) STAF (ZNF143), and C) NRF1. Note that each transcription factor displays variation of general footprint shapes, indicating that footprint detection requires a TF specific approach. D) Top panel shows aggregate DNase-seq signal centered on REST motif matches that map within REST ChIP-seq peaks. K-means clustering of the REST aggregate plot (top) identifies two types of DNase aggregate profiles (bottom). Cluster 1 identifies subset of REST binding sites that does not display depletion of DNase signal, while Cluster 2 represents REST binding sites with depletion of DNase-seq signal.
Figure 4.3: DNase-seq displays cleavage bias that is protocol specific. A) Scatter plot of cleavage propensities of all possible DNA 6-mers (log10 scale) for deproteinized genomic DNA from MCF7 and K562 cell lines using the single hit high molecular weight DNase-seq protocol [21]. B) Scatter plot comparing cleavage propensities of 6mers from deproteinized genomic DNA from K562 using the single hit DNase-seq protocol vs. deproteinized genomic DNA from IMR90 cell line using an independent two hit small molecular weight DNase-seq protocol [69]. The inset box represents maximum and minimum cleavage propensity values for single hit DNase-seq protocol performed on K562 cell line. Spearman correlation is indicated in each plot.
**Data Collection**
Scan the genome for sequence motif matches using PWMs
Build sets of sequence motif matches by thresholding
Collect DNase-seq data around sequence motif matches

**Building Training and Test Sets**
Intersect sequence motif matches with coordinates of ChIP-seq peaks and DHSs
Build sets of motif matches inside and outside ChIP-seq peaks and DHSs

**Model Training**
Initialize footprint and background components
Train mixture model
Learn footprint and background components

**Validation**
Calculate FLR metric
Perform 4-fold cross validation
Calculate auROC and sensitivity metrics

Figure 4.4: Workflow of binary classification scheme.
Figure 4.5: auROC (A) and sensitivity (B) values for D-s and FLR values in for classification of all sequence motifs detected across the genome.

Figure 4.6: Comparison of Footprint Likelihood Ratio (FLR) to general DNase sensitivity (D-s) score. Motif matches for 21 transcription factors that map within DHS sites were compared to ChIP-seq data to calculate A) auROC and B) sensitivity at 1% false positive rate for FLR and D-s scores. Each transcription factor is indicated as a blue (FLR) or red (D-s) circle, dashed lines represent the means.
Figure 4.7: Aggregate DNase-seq plots for E2F1 sequence motif matches in E2F1 ChIP-seq peaks from MCF7 cell line, obtained by either PBM scan (left) and PWM scan (right). Aggregate plot obtained using PBM scan show a substantially cleaner footprint shape compared to PWM scan, showing that PBM scan accurately pinpoints exact locations of TF-DNA interaction.

Figure 4.8: Receiver operating characteristic (ROC) curves for 3 transcription factors using FLR as a predictor for all genome wide sequence motif matches. A rapid increase in true positive rate at high (>0.9) false positive rates is a common trend for many factors, indicating that a portion of sequence motifs that are in ChIP-seq peaks get low FLR scores.
Figure 4.9: A) Aggregate DNase-seq signal around top 25% and bottom 25% of sequence motifs for ZNF 143 - sorted by FLR score - are shown on left side. Top 25% profile has a footprint shape as expected, but no footprint shape in the lower 25% set. On right side we see the DNA sequence estimated signal for the same two quartiles, and neither show a footprint shape profile. This indicates that sequence bias is not a main contributor to footprint shape. B) Same plot for the RFX transcription factor.
Figure 4.10: Footprint scores indicate mode of transcription factor interaction.
A) Median ChIP-seq intensity scores of ChIP-seq peaks of 5 factors sorted by FLR footprint scores and divided into ten bins. The highest FLR scores are in the first bin. Note footprint score correlates with ChIP-seq signal, with the exception of the weakest footprinting scores where they are inversely correlated. B) Boxplots of NRSF ChIP-seq intensity scores across footprint scores. C) A heat-map showing overlapping ChIP-seq peaks for the top and bottom 10% highest and lowest footprint scores. CoRest and Znf143 binding is enriched for the strongest NRSF footprints (left) and are depleted in the weakest NRSF footprints (right). D) Conversely, Taf1 and Pol2 binding is depleted for the strongest NRSF footprints (left), and enriched for the weakest NRSF footprints (right).
Figure 4.11: Cell type specific footprints in shared DHS sites. A) Representative example of DNase-seq data from GM12878 and Medulloblastoma (D721) cell lines. This DHS site is present in both cell types, but a clear footprint for NRSF is only detected in GM12878 at the sequence motif match (blue shaded box). B) Aggregate DNase-seq signal around NRSF motifs in GM12878 (left) and medulloblastoma (right) cell lines indicate that NRSF does not leave a footprint in the medulloblastoma cell line. C) Boxplots showing distribution of FLR and D-s scores in GM12878 and Medulloblastoma cell lines for the NRSF motif in DHS sites that are present in both cell types. Distribution of FLR scores displays a difference between GM12878 and Medulloblastoma, whereas D-s scores display no difference. D) Similar boxplots showing distributions of FLR and D-s to identify differential footprint scores between skin fibroblasts and iPSc cells for OCT4, Sox2, C-Myc, and KLF4 Yamanaka factors. FLR scores were more sensitive to changes in transcription factor binding between two cell types, reflected by smaller p-values indicated in each box and Table 3.
Figure 4.12: EM footprint components distinguish background bias and footprints, as well as alternate motif usage. A) Correlation of intrinsic DNase-seq sequence bias profile (generated from deproteinized naked DNA DNase-seq) compared to the de novo foreground footprint component (X-axis) and de novo background component (Y-axis) of multinomial mixture model. For 19 transcription factors the de novo background component learned by mixture model correlates more with intrinsic sequence bias model. The majority of de novo foreground footprint models correlate negatively with intrinsic sequence bias model. B) Combined footprint model for CTCF against the de novo background component in the upper panel and the two footprint components (C1 and C2) that make up the footprint in the lower two panels, with the sequence logo associated with each component for CTCF. Vertical lines delimit the PWM we used for this factor. An additional motif associated with the depletion in second footprint component can be seen upstream of the main motif. C) Similarly for ZNF143, extended motif corresponds to a bigger footprint for the second component.
Table 4.1: auROC and Sensitivity at 1% FPR values for D-s and FLR for each transcription factor

<table>
<thead>
<tr>
<th>TF</th>
<th>Neg</th>
<th>Pos</th>
<th>AUC</th>
<th>Sensitivity at 1% FPR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Whole Genome</td>
<td>Whole Genome</td>
</tr>
<tr>
<td>E2F4</td>
<td>49689</td>
<td>468</td>
<td>0.99</td>
<td>0.89</td>
</tr>
<tr>
<td>NFY-A</td>
<td>48591</td>
<td>1435</td>
<td>0.99</td>
<td>0.88</td>
</tr>
<tr>
<td>NFY-B</td>
<td>46423</td>
<td>3603</td>
<td>0.98</td>
<td>0.9</td>
</tr>
<tr>
<td>NRF1</td>
<td>46423</td>
<td>3718</td>
<td>0.99</td>
<td>0.93</td>
</tr>
<tr>
<td>RFX5</td>
<td>50119</td>
<td>189</td>
<td>0.98</td>
<td>0.9</td>
</tr>
<tr>
<td>ZNF143</td>
<td>49330</td>
<td>870</td>
<td>0.99</td>
<td>0.93</td>
</tr>
<tr>
<td>BHLHE40</td>
<td>47763</td>
<td>2349</td>
<td>0.97</td>
<td>0.85</td>
</tr>
<tr>
<td>USF1</td>
<td>49840</td>
<td>415</td>
<td>0.97</td>
<td>0.77</td>
</tr>
<tr>
<td>USF2</td>
<td>49870</td>
<td>276</td>
<td>0.96</td>
<td>0.79</td>
</tr>
<tr>
<td>YY1</td>
<td>48686</td>
<td>2095</td>
<td>0.99</td>
<td>0.94</td>
</tr>
<tr>
<td>CTCF</td>
<td>29715</td>
<td>20430</td>
<td>0.95</td>
<td>0.76</td>
</tr>
<tr>
<td>MEF2A</td>
<td>49112</td>
<td>1177</td>
<td>0.97</td>
<td>0.56</td>
</tr>
<tr>
<td>NRSF</td>
<td>47807</td>
<td>2434</td>
<td>0.81</td>
<td>0.64</td>
</tr>
<tr>
<td>PAX5</td>
<td>49756</td>
<td>599</td>
<td>0.94</td>
<td>0.81</td>
</tr>
<tr>
<td>PU1</td>
<td>49235</td>
<td>1269</td>
<td>0.94</td>
<td>0.81</td>
</tr>
<tr>
<td>SP1</td>
<td>50044</td>
<td>322</td>
<td>1</td>
<td>0.91</td>
</tr>
<tr>
<td>SRF</td>
<td>49143</td>
<td>866</td>
<td>0.94</td>
<td>0.83</td>
</tr>
<tr>
<td>TCF3</td>
<td>50100</td>
<td>251</td>
<td>0.99</td>
<td>0.82</td>
</tr>
<tr>
<td>ZEB1</td>
<td>50249</td>
<td>40</td>
<td>1</td>
<td>0.84</td>
</tr>
<tr>
<td>C-MYC</td>
<td>49966</td>
<td>332</td>
<td>0.99</td>
<td>0.85</td>
</tr>
<tr>
<td>MAX</td>
<td>49639</td>
<td>683</td>
<td>0.99</td>
<td>0.88</td>
</tr>
</tbody>
</table>
Table 4.2: ENCODE antibody names used for ChIP experiments transcription factors and the IDs of PWMs used to look for sequence motif matches in these peaks. TRANSFAC and JASPAR databases were used to find corresponding PWMs.

<table>
<thead>
<tr>
<th>Factor Name</th>
<th>Matrix ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>E2F4</td>
<td>M00738</td>
</tr>
<tr>
<td>NFYA</td>
<td>MA0060.2</td>
</tr>
<tr>
<td>NFYB</td>
<td>MA0060.2</td>
</tr>
<tr>
<td>NRF1</td>
<td>M00652</td>
</tr>
<tr>
<td>RFX5</td>
<td>M00280</td>
</tr>
<tr>
<td>ZNF143</td>
<td>M00262</td>
</tr>
<tr>
<td>BHLHE40</td>
<td>M00985</td>
</tr>
<tr>
<td>USF1</td>
<td>MA0093.2</td>
</tr>
<tr>
<td>USF2</td>
<td>M00726</td>
</tr>
<tr>
<td>YY1</td>
<td>M00793</td>
</tr>
<tr>
<td>CTCF</td>
<td>MA0139.1</td>
</tr>
<tr>
<td>MEF2A</td>
<td>MA0052.2</td>
</tr>
<tr>
<td>REST/NRSF</td>
<td>MA0138.2</td>
</tr>
<tr>
<td>Pax5</td>
<td>MA0014.2</td>
</tr>
<tr>
<td>PU1</td>
<td>MA0080.3</td>
</tr>
<tr>
<td>SP1</td>
<td>MA0079.3</td>
</tr>
<tr>
<td>SRF</td>
<td>MA0083.2</td>
</tr>
<tr>
<td>TCF3</td>
<td>MA0091.1</td>
</tr>
<tr>
<td>ZEB1</td>
<td>MA0103.2</td>
</tr>
<tr>
<td>C-Myc</td>
<td>MA0059.1</td>
</tr>
<tr>
<td>MAX</td>
<td>MA0058.2</td>
</tr>
<tr>
<td>OCT4::Sox2</td>
<td>MA0142.1</td>
</tr>
<tr>
<td>KLF4</td>
<td>MA0039.2</td>
</tr>
</tbody>
</table>
Table 4.3: P-values for Kolmogorov-Smirnov test (one-sided) for comparing the distributions of FLR and D-s metrics assigned to sequence motifs from two different cell types. The expression of transcription factors are induced or abolished in one cell type vs. the other, FLR metric is consistently more sensitive to these changes (more significant differences as measured by smaller p-values).

<table>
<thead>
<tr>
<th>Factor, Cell Types Compared</th>
<th>FLR(P-value)</th>
<th>D-s(P-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NRSE, GM12878 vs. Medulo</td>
<td>1.18e-10</td>
<td>0.4</td>
</tr>
<tr>
<td>OCT4::SOX2, iPS vs. Skin</td>
<td>2.8e-12</td>
<td>2.5e-10</td>
</tr>
<tr>
<td>KLF4, iPS vs. Skin</td>
<td>1.05e-7</td>
<td>0.009</td>
</tr>
<tr>
<td>C-myc, iPS vs. Skin</td>
<td>0.0015</td>
<td>0.096</td>
</tr>
</tbody>
</table>
5 CTCF binding is necessary and sufficient for maintaining open chromatin.

5.1 Introduction

CTCF is a zinc-finger family transcription factor that plays numerous roles in a cell, including transcriptional repression [81-84] and activation [85], insulator activity [86, 87], and regulation of three-dimensional conformation of chromatin [88, 89]. The importance of CTCF is reflected by it being ubiquitously expressed across cell types [90], it binding invariantly to the same regions across cell types [91], and it being conserved from Drosophila to human [90, 92].

CTCF-DNA interactions have been thoroughly characterized by ChIP-seq [78, 93], ChIP-exo [79] assays and DNase-seq footprinting approaches [24], which indicate alternative CTCF-DNA interaction models that correspond to variants of consensus binding motifs. ChIA-PET [89] and Hi-C [88] studies have shown enrichment of CTCF binding at boundaries of long range interaction topological domains. These findings indicate that CTCF is implicated in organizing the three dimensional conformation of the genome and regulating gene expression by controlling promoter-enhancer interactions [94]. However, it is exact mechanism underlying this process are unknown.

We characterized genome-wide CTCF-DNA interactions and the effects of CTCF binding on chromatin state by knocking down CTCF in human MCF7 cells. Changes in genome-wide chromatin accessibility and transcription control were studied using DNase-seq and RNA-seq assays. We find 1000’s of regulatory regions that become
significantly less accessible as a direct result of reduced CTCF levels in the cell, indicating that CTCF is required for maintenance of these DHS sites. The degree of reduced accessibility is highly correlated with similarity to the canonical CTCF binding motif. However, DNase footprinting analysis identifies a subset of DHS sites remain open even though they do not bind CTCF, indicating that other factors are involved in keeping these sites accessible. We also detect 1000’s of DHS sites that become more accessible in response to CTCF knockdown, and these newly accessible sites appear to function as enhancers. Knocking down CTCF results in 100’s of changes in gene expression, and these differentially expressed genes cluster together and around differential DHS sites. Together, these results indicate that CTCF plays different roles in maintaining chromatin accessibility and gene expression across the genome.

5.2 Results

5.2.1 CTCF knockdown changes genome-wide chromatin architecture

To understand role of CTCF in chromatin organization, we knocked down CTCF expression in MCF7 cell line. We transfected MCF7 cells with either a lentiviral vector that contains a CTCF targeting shRNA (MCF7-CTCFKD), or a scrambled shRNA construct (MCF7-SCR). By western blot, MCF7-CTCFKD cell line displayed approximately 10% of the normal CTCF protein levels compared to MCF7-SCR (Figure1A).
Figure 5.1: CTCF was knocked down using a shRNA that targets CTCF coding sequence. Western blots show significantly lowered levels of CTCF expression in MCF7\_CTCFKD cell line whereas B-actin expression is unaffected (A). CTCF knockdown results in changes in chromatin accessibility. In this representative locus, DHSs are abolished or weakened after CTCF knockdown (B).

To characterize the global impact of CTCF knockdown on chromatin architecture, we generated DNase-seq and RNA-seq for two independent biological replicates of the knocked down and control cell lines. As an additional control, we performed DNase-seq and RNA-seq on MCF7 cells that were transduced with the RNAi construct, but expanded until CTCF expression recovered to normal levels. Comparison of DNase-seq data reveals a large of number DHS sites that are lost after CTCF knockdown (Figure 1B). When CTCF is reexpressed after continued passaging, many of the DHS sites that become less accessible return to the open chromatin state (Figure 1B). ChIP-seq analysis reveals that DHS sites that disappear often overlap CTCF binding sites. In addition, some DHS sites become more accessible after CTCF knockdown, and these sites do not overlap CTCF binding sites (Figure 1C).
Characterization of DHSs reveals that overall accessibility of chromatin is reduced after CTCF knockdown (Figure 2A). This trend is also apparent for DHS sites that are present in both cell types (Figure 2B). DHS sites that are only present in MCF7-SCR are generally weaker DHS sites whereas DHS sites that are only present in MCF7-CTCFKD have the lowest DNase enrichment scores. This indicates that after knocking down CTCF results, strong DHS sites disappear, while new DHS sites that open up have relatively weaker signal.
Figure 5.2: Boxplots showing distributions of parzen scores of DHSs from MCF7-SCR (blue) and MCF7-CTCFKO (red) cell lines (A). On left boxplots of distributions of parzen scores of DHSs that are shared between MCF7-SCR and MCF7-CTCFKO cell lines, color coded the same way. On the right DHSs that are specific to each cell type (B). Distribution of log fold change values for each DHS tile (C) and for only tiles that are significantly different between two cell types (D).

We next used DESeq [95] to identify quantitative differences in DHS sites. To control for variable sizes of DHS sites, we took the union set of DHS sites from MCF7-SCR and MCF7-CTCFKO, and divided up these DHS sites into overlapping 300bp windows [96]. After extracting the DNase read counts for each biological replicate, we
used DESeq to identify DHS sites that were significantly different (P.val < 10e-2). Out of 267664 total DHS sites, we identified 49991 DHS sites where DNase-seq read counts are statistically different. This totals 19% of the DHS sites detected within these two cell types. This trend can be visualized by looking at the distribution of log. fold change (FC) values of each DHS sites, which shows that while the majority of sites have log FC values close zero, we detect a skewing towards negative values supporting that overall chromatin accessibility is reduced after CTCF knockdown (Figure 2C). Analysis of only the significantly different DHS sites also supports that there are more DHS sites that disappear in MCF-CTCFKD, compared to the smaller number of new DHS sites that open up in MCF-CTCFKD cells (Figure 2D). This indicates that loss of CTCF in the cell results in both loss of some DHS sites and the gain of other DHS sites.

5.2.2 Genome-wide CTCF binding is weakened

We used DNase-seq footprinting to investigate changes in CTCF binding activity across the genome. High confidence CTCF candidate binding sites were identified by scanning the genome using the CTCF PWM from JASPAR database. Reducing CTCF in the cell to approximately 10% of normal levels generates aggregate DNase-seq plots with a well-defined footprint, indicating that the residual CTCF protein still binds to many CTCF motifs across the genome (Figure 3A).
Figure 5.3: Aggregate plots of all high confidence (high PWM score) CTCF binding sites from MCF7-SCR (blue) and MCF7- CTCFKD (red) cell types, side by side (A). Aggregate plots of high confidence CTCF binding sites that map in a DHS in both cell types both only have a footprint in MCF7-SCR cell type (B). Distribution of PWM scores of candidate sites that have a footprint in both cell types and map in a DHS both cell types (blue); no footprint but map in a DHS site in both cell types (purple) and not in a DHS in MCF7_CTCFKD cell type (red) in panel C.

Aggregate DNase-seq plots suggest that even though there are detectable levels of CTCF binding after knockdown, DNase accessibility around CTCF binding sites is either reduced or lost. Out of 19922 CTCF binding sites that map in a DHS in MCF7_SCR cell line, 45% of such sites are no longer in a DHS after knockdown, suggesting loss of CTCF binding causes loss of DNase hypersensitivity around the binding site. In order to better characterize CTCF binding at high resolution, we used DNase-seq footprint models were utilized to quantify the strength of CTCF-motif interaction.
Using footprint models enables characterization of TF-DNA interaction at higher resolution, particularly when overall chromatin accessibility does not change. We used our previously described footprint models assign a footprint likelihood ratio (FLR) to each candidate binding site. A positive FLR indicates positive binding evidence and negative FLR suggests there is no TF-DNA interaction. We characterized CTCF binding at 19922 CTCF candidate binding sites that map within DHSs before knockdown. 11,337 of these sites map in significantly weakened DHS tiles after knockdown of CTCF, while 6,932 of candidate binding sites retain a DNase footprint (positive FLR value) coupled with there being no significant change in chromatin accessibility. Interestingly we discovered 1,631 candidate CTCF binding sites where there is no significant change in chromatin accessibility but the footprint is substantially weakened (Figure 3B). Boxplots of PWM scores of these three categories (Figure 3C) reveal that highest scoring motifs are still bound with remaining CTCF after knockdown, whereas weaker scoring motifs lose their footprints but retain their open chromatin status. The lowest scoring motifs lose both footprints and chromatin accessibility.

5.2.3 Chromatin changes are caused by direct and indirect effects of CTCF binding loss

Analysis of high confidence CTCF candidate binding sites showed that large portion of differential DHS sites overlap with CTCF motifs. Distribution of CTCF PWM scores in differential DHS tiles (Figure 4A) show an enrichment of CTCF motifs in DHS
sites that decrease in DNase hypersensitivity after knockdown. Conversely, we detect a depletion of CTCF motif matches in tiles that significant increase in DNase hypersensitivity following knockdown. This is consistent with findings in previous section, as loss of CTCF binding results in loss of DNase hypersensitivity surrounding the binding site.

Figure 5.4: Distribution of CTCF PWM scores in DHS tiles that close after knockdown (A) and open after knockdown (B) shown in histograms. PWM scores lower than zero indicates there is no CTCF motif in a tile. Histogram of log fold change values of gene expression values as measured by RNA-seq experiments. Gene regulation is affected in both directions as we see a bimodal distribution of
upregulated and downregulated genes (B). Changes in gene expression with respect to changes in DNase accessibility. Boxplots show distribution of gene expression fold change values of genes that contain DHS sites that opens (blue) or closes (red). Differential DHS sites are broken into three categories based on whether they map to a promoter, intron, or exon. The gene expression value was determined for the nearest gene.

Other than blocking the spread of heterochromatin [93], CTCF is known regulates transcription by causing DNA looping and regulating enhancer-promoter interactions [97]. To analyze the effects of CTCF knockdown on transcription, we performed RNA-seq experiments for MCF7-SCR and MCF7-CTCFKD cell lines. Using CuffLinks [43], 1611 genes were found be differentially expressed after CTCF knockdown; 771 genes are upregulated whereas 840 genes are downregulated (Figure 4B). Coupling transcriptional changes with changes in chromatin and comparing the effects of knockdown, we observed that DHS sites that open up overlap with gene bodies whose expression levels increase. Expectedly we also observed the reverse trend as DHS site loss overlaps with genes that are downregulated following knockdown (Figure 4C).
Figure 5.5: Indexes of genes that are differentially regulated on 22 of the human chromosomes in red. Differentially expressed genes cluster together on each chromosome.

Further investigation of locations of differentially expressed genes reveals significant clustering. Figure 5 presents every gene on 22 chromosomes as blue dots where differentially expressed genes are shown as red crosses, which tend to cluster together. Distances between boundaries of differentially expressed gene pairs are significantly lower than distances between pairs of randomly selected genes (Kolmogorov Smirnov Test, \( P < 7 \times 10^{-3} \), one sided). Therefore, knockdown of CTCF may disrupt formation of correct topological domains and effect regulation of genes that group together into the same domain [88]. Further Hi-C experiments are needed to validate the effects CTCF knockdown on three-dimensional organization of the genome.
5.3 **Methods**

5.3.1 Detection of differential DHS sites

To discover significant changes in chromatin accessibility as measured by DNase-seq, we adopted approaches that are designed to discover differential gene expression measured by RNA-seq such as edgeR and DESeq [95, 98]. These methods are designed to model discrete count data and calculate statistically significant differences in count values that cover genes across the genome. Since DNA-seq data results in similar discrete counts that quantify chromatin accessibility of loci, these methods are applicable for our purposes.

To identify changes in chromatin accessibility - or differential DHSs - , we compare DNase-seq read counts from two different cell types or two different conditions from the cell type. As we query intervals that are not static and predefined like gene bodies, we instead use the union of sets of DHSs identified from two different conditions. Union operation allows us to limit to search space to genomic regions which are chromatin accessible in at least one cell line. The set of unions are broken into overlapping (50bp) windows that are 300bps long. This tiling process is necessary to run the analysis in comparable windows and avoid false discoveries that may result from two DHSs that overlap minimally but may be differentially accessible in different parts of the union DHS [96].
For each tile we define, we count the number of DNase-seq fragments that fall into each tile from every DNase-seq replicate we have for each condition. This data is normalized by DESeq package; each tile is assigned a fold change value and p-value that measures the significance of change in accessibility. Each tile that is assigned a p-val < 0.01 is considered to have a significant change in accessibility.

5.4 Discussion

CTCF is a well-studied transcription factor that has been shown to play an active role in determining chromatin architecture and three-dimensional conformation of the genome [88, 89, 97]. Conserved from Drosophila to human, CTCF binding sites are mostly invariant between different human cell lines [90, 91]. The role of CTCF in determining three-dimensional interactions between different genomic regions has been characterized recently by multiple studies. In order to assess the effects of this important and versatile factor on genome-wide chromatin architecture, we used DNase-seq to study changes in open chromatin that result from the reduction of CTCF in the cell.

By knocking down CTCF expression to 10% of baseline levels, we discovered that 19% of chromatin is significantly altered, as measured by DNase-seq experiments. We validated that CTCF binding is abolished at a large number of binding sites by performing DNase footprinting. However, we observed that residual CTCF protein present after knockdown was enough to bind to the strongest CTCF motifs, showing that the main determinant of CTCF binding is its DNA sequence specificity.
Characterization of changes in chromatin states revealed that most DNA sites that close (loose chromatin accessibility) contain a CTCF motif. This suggests that CTCF binding itself strongly contributes to opening of chromatin and is responsible for the formation of a large number of DHSs.

Analysis of gene expression by RNA-seq also revealed that loss of CTCF significantly altered the expression of more than 1600 genes. We observed that these differentially expressed genes significantly cluster together, and that changes in chromatin accessibility also correlate with changes in gene expression. These findings suggest that CTCF binding loss changes the three-dimensional organization of the genome, which in turn affects enhancer-promoter interactions and transcriptional regulation [88, 93, 94].

Our results show that CTCF has large direct and indirect effects on genome-wide accessible chromatin, through its binding to DNA and its effects on transcriptional regulation. These findings are consistent with the large number of roles and large number of binding sites attributed to CTCF. Our experiments are the first in characterizing CTCF’s effects on maintaining accessible chromatin and further highlight its importance in maintaining correct chromatin architecture and regulation. Further chromosome conformation capture experiments are needed to understand the intermediate mechanisms by which CTCF regulates chromatin architecture and transcriptional regulation.
6 Development of Broad ChIP-seq peak caller

6.1 Introduction

Next generation sequencing based assays that generate short DNA sequences are used to identify specific regions of genome; such as transcription factor binding sites (ChIP-chip, ChIP-seq) [7, 8] and regions of open chromatin (DNase-seq, FAIRE-seq) [20, 57, 99, 100]. Due to noise and artifacts, these DNA sequences do not exclusively map to regions of interest but are distributed across the genome. To accurately identify regions of interest, computational tools are used to identify regions with a statistically significant enrichment of DNA fragments [10, 11, 22, 101, 102]. These tools are called peak callers.

With the advent of next generation sequencing based assays, a large number of peak callers have been designed. Most of these methods use sliding window approach coupled with a statistical model for calculating statistical significance. Additionally, individual assays exhibit specific properties due to differences in experimental protocol, such as twin peaks that surround each binding site in a ChIP-seq experiment [10]. The distance between twin peaks is called shift size and most ChIP-seq peak callers estimate this distance and factor it into their enrichment models.

Currently numerous NGS assay peak callers are available, especially for ChIP-seq protocol [10, 11, 102]. However, comparisons of peak caller performance have not yielded a clearly best performing tool, and performance of peak callers tend to vary
across factors [103]. ENCODE consortium has generated guidelines to standardize peak calling approaches and reduce variability [41]. SPP and MACS [10, 11] are the two most widely used peak callers due to their intuitive model design and performance. Both of these methods model shift size and factor it into their peak calling model heavily. They are known to be not suited to work for histone tail ChIP-seq experiments, which tend to have broad regions of enrichment [41].

Here we describe an alternative peak caller for ChIP-seq experiments where shift size is difficult to estimate due to ChIP-seq protocol we use. Using ChIP-seq experiments for Bluejay (BLJ) and Jackdaw (JKD) proteins in Arabidopsis thaliana root cells; we estimated regions of enrichment by transforming ChIP-seq signal using kernel density estimation. Resulting peaks were analyzed for reproducibility using IDR pipeline and our peak caller tool yielded better reproducibility for these two proteins compared to MACS and SPP. This suggests that peak calling method we present here is able to capture binding sites more accurately compared to peak callers that depend on shift size.

6.2 Results

6.2.1 Successful shift size estimation depends on protocol specific shearing

In a typical ChIP-seq protocol, DNA fragments that are cross-linked to a transcription factor or a protein complex are sonicated to yield short (100-500bp) DNA fragments. These fragments are sequenced from their 5’ ends. This particular property of the protocol leads to detection of two peaks that surround the true binding site. We
observe one peak upstream in sense direction for fragments that align to Watson strand and another peak upstream in antisense direction for fragments that align to Crick strand. This bimodal distribution property is generally factored into by peak callers to detect regions of true enrichment [9].

Due to stability limitations of certain transcription factors of BLJ and JKD (members of C2H2 zinc finger family), sonication step may not be performed for extended durations. Due to a briefer sonication step, resulting fragments are longer than usual (1000+ bp) DNA fragments in a typical ChIP-seq protocol. To manage fragment length, these longer fragments are shortened into 50bp during library preparation. This difference in ChIP-seq protocol causes DNA fragments to be uniformly distributed around the true binding in a wide window (>1kb) as opposed to twin peak distribution seen in normal ChIP-seq protocols. This difference poses challenges for most ChIP-seq peak callers since they make assumptions that depend on enrichment in a tight window around the binding site and detection of twin peak distribution.
Running MACS peak caller for BLJ and JKD results in failure to regions of enrichment because MACS uses a relatively tight sliding window (typically ~250bp) that cannot detect broad regions of enrichment in our data. Lack of twin peak property in our ChIP-seq data can be observed in cross-correlation plots (obtained by running SPP peak caller). Only 50bp shift-size results in local maxima for different cross correlations plots, which corresponds to length of sequencing fragments (Figure 1). High values of...
cross correlation at shift-size value equal to fragment length is well documented as phantom peaks and do not correspond to shift size detected at twin peak sites. This presents the difficulties in using a regular ChIP-seq peak caller to detect peaks for BLJ and JKD ChIP-seq datasets.

### 6.3 Kernel density estimates can detect broad regions of enrichment

In order detect the broad peaks (regions of enrichment that are longer than 1kb) that are prevalent in our data due to differences in experimental protocol; we designed a custom peak caller that uses kernel density estimates to smoothen the ChIP-seq signal. We generated control signal by sonication of DNA without immunoprecipitation and smoothed the control data in the same manner. After normalization, control signal is subtracted from ChIP signal and regions of enrichment are discovered by a sliding window approach. We call our algorithm broad peak caller.
Figure 6.2: Upper panel shows ChIP-seq read pileup on Chr1 (50000-60000). Lower panels show kernel density estimates for two bandwidth values (600 and 1200) for ChIP (blue), control (red) and their difference (purple). Green lines denote peaks boundaries and black line denotes the threshold.

To calculate kernel density estimates, we used F-seq peak caller (cf. Section 2.2) software. Even though F-seq could have been used to call peaks in our data, it is not designed to use a control experiment to estimate background and therefore was not suited to work with ChIP-seq data in its current state. The degree of smoothening in kernel density estimation depends on bandwidth parameter. In Figure 2, the effects of bandwidth parameter on smoothening and resulting peaks are visible. As we increase the bandwidth, neighboring peaks are joined together. For BLJ and JKD datasets, bandwidth parameter was set to 1000 empirically after experimenting with different values. This value was also consistent with the fact that ChIP DNA fragments were approximately 1000bp long. After density estimates for ChIP experiment and control experiment are obtained, we normalize the densities for each chromosome by the total
number of ChIP-seq tags to counter the effects of differing sequencing depths for each experiment. Control density is subtracted from ChIP signal to factor out sequencing biases and other chromatin effects.

We use this differential density to look for regions of ChIP signal enrichment. We define a peak as a region where differential density is higher than a threshold value for at least consecutive 100 base-pairs. Regions of enrichment smaller than 100bps are filtered out as they are unlikely to result from 1000+ bp ChIP DNA fragments. Threshold for enrichment can be varied to obtain regions of different degrees of enrichment. Resulting peaks above the threshold are assigned a fold enrichment score, defined as the ratio of aggregate density for ChIP and control. For p-value calculations, we adopted a similar approach to MACS peak caller. ChIP-seq fragment counts were modeled by Poisson distribution. Background fragment rate from control experiment was estimated as the maximum of average read count from four windows of different sizes: width of original peak, 1kb, 5kb and 10kb [11].

6.3.1 Broad peaks are highly reproducible between replicates

To test reproducibility of experiments, ChIP-seq assays are typically run with multiple biological replicates to guarantee reproducibility of results and characterize variation in the system under study. A typical problem in this framework is to find a principled approach to combining analysis results (especially calling regions of enrichments or peaks) from different replicates [41].
Table 6.1: Number of peak pairs that are under the IDR threshold for broad peak caller compared with peak calls from MACS and SPP. Peak calls in the table are for BLJ

<table>
<thead>
<tr>
<th>IDR threshold</th>
<th>#Peaks (Broad peaks)</th>
<th>#Peaks (MACS)</th>
<th>#Peaks (SPP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>58</td>
<td>152</td>
<td>0</td>
</tr>
<tr>
<td>0.02</td>
<td>262</td>
<td>238</td>
<td>30</td>
</tr>
<tr>
<td>0.03</td>
<td>558</td>
<td>330</td>
<td>84</td>
</tr>
<tr>
<td>0.04</td>
<td>971</td>
<td>423</td>
<td>120</td>
</tr>
<tr>
<td>0.05</td>
<td>1490</td>
<td>520</td>
<td>154</td>
</tr>
<tr>
<td>0.06</td>
<td>1988</td>
<td>626</td>
<td>197</td>
</tr>
<tr>
<td>0.07</td>
<td>2470</td>
<td>745</td>
<td>244</td>
</tr>
<tr>
<td>0.08</td>
<td>2963</td>
<td>879</td>
<td>301</td>
</tr>
<tr>
<td>0.09</td>
<td>3441</td>
<td>1035</td>
<td>376</td>
</tr>
<tr>
<td>0.1</td>
<td>3920</td>
<td>1204</td>
<td>448</td>
</tr>
</tbody>
</table>

For this purpose, we adopted irreproducible discovery rate (IDR) framework that quantitatively assesses reproducibility of each peak across multiple experimental replicates. In short, IDR framework compares two ranked lists of discoveries (peaks) and creates a curve which assesses consistency of findings. These ranked lists should contain high confidence and low confidence peaks for the model to fit bivariate rank distributions. In turn, this allows separation of signal from noise based on consistency of ranks.

We applied IDR framework to our broad ChIP-seq peaks of BLJ and JKD. For both transcription factors, two biological replicates of ChIP-seq were available for IDR analysis. As IDR framework requires both high confidence and low confidence peak calls to compare ranks, we used a liberal threshold (75\textsuperscript{th} percentile of genome-wide
differential density signal). This threshold results in identification of 40000+ peaks for both transcription factors. For SPP and MACS peaks, we similarly adopted loose thresholds to identify 50000+ peaks.

To run the IDR framework, peaks from two replicates have to be paired. The criteria for pairing two peaks is whether they overlap or not for at least 1bp. For peak ranking, we used fold enrichment values for broad peaks as they rank high confidence observations more robustly compared to p-values. As authors of IDR pipeline suggest, we used p-values for MACS and fold enrichment values for SPP.

Broad peak caller resulted in highest number of reproducible peaks between two replicates. As seen in table 1, at IDR threshold of 0.01, MACS performs best followed by broad peak caller and SPP in that order. For all other thresholds, broad peak caller consistently outperforms SPP and MACS. SPP particularly results in very poor reproducibility, since its peak calling algorithm heavy depends on cross-correlation values and correct estimation of shift size. As table 1 presents, broad peak caller’s ability to detect more reproducible peaks is due its ability to detect enrichments independent of shift size and twin peak structure.

6.4 Discussion

Accurate peak calling methods for next generating sequencing based assays require good understanding of processes that generate DNA fragments [9, 41]. Numerous peak calling algorithms have been devised for ChIP-seq data [10, 11, 102] but
their performance may vary from type of pulled down protein or the experimental protocol [103]. Peak caller performance for transcription factor ChIP-seq data is generally considered satisfactory. However, ChIP-seq experiments that result in broad enriched regions (such as histone modification ChIP-seq data) are challenging for existing peak callers [41].

We present a peak calling algorithm that is suited to detect broad regions of enrichment. Broad regions of enrichment in our datasets result from a specific fragment size controlling step in experimental protocol we adopted. Compared to widely used peak callers MACS and SPP, our algorithm performs better in terms of detecting reproducible peaks for BLJ and JKD. SPP and MACS’ poor performance result the assumptions associated with these algorithms that heavily depend on prevalence of twin peaks of enrichment that surround true binding sites.

As there are no ChIP experiments for BLJ and JKD that are generated using traditional ChIP-seq protocols, a direct comparison was not available. As new experiments become available, further comparisons are necessary to assess the ability of different kinds of peak callers for different ChIP-seq protocols.
7 Conclusions

7.1 Summary

The advent of high throughput next generation sequencing (NGS)-based assays has transformed the field of genomics by dramatically increasing our ability to characterize different biological processes in a genome-wide fashion. Recently, the ENCODE consortium and other research groups have performed large numbers of various NGS assays to characterize gene transcription and regulatory processes in multiple human cell types [39, 40]. As the amount of such data increases at a staggering rate, our main challenge is to analyze these datasets to understand the underlying biological processes, by using accurate models that can account for experimental biases and artifacts.

I focused on the DNase-seq assay, with the goal of understanding transcriptional regulation in human cells. Building on the well-established DNase footprinting principles, I specifically studied mechanisms of transcription factor binding and characterized TF-DNA binding events using information on DNA sequence and the state of chromatin.

DNase-seq footprints can be analyzed at different levels. First, aggregated over multiple candidate binding sites, the genome-wide footprint of a transcription factor reveals the overall activity of that factor in the assayed cell type, as shown in Chapter 3. Second, a comparison of aggregated DNase-seq profiles across different cell types
enables the discovery of cell type specific transcription factor binding, as illustrated in Chapter 3. Complementary to these analyses, in Chapter 4, I described a method for building transcription factor-specific footprints models that enable discovery of individual candidate binding sites distributed across the genome. For multiple transcription factors, footprints are enriched at high confidence transcription factor binding sites (as measured by ChIP-seq) and allow discovery of true binding sites at a much higher resolution (6-20bp) compared to ChIP-seq experiments (100 or more bp).

Aggregate DNase-seq profiles have distinct shapes specific to each factor. However, these profiles can be broken into sub-profiles that possibly correspond to alternative mechanisms of TF-DNA interactions. For example, for the androgen receptor protein, sub-profiles that exhibit symmetric half footprints were identified. For TFs that belong to the zinc finger structural family, extended footprints that correspond to an additional binding motif were discovered. These observations are consistent with proposed TF-DNA interaction mechanisms and have been validated either experimentally or computationally [20, 78, 79]. These findings show that DNase-seq footprints can detect alternative binding modes for transcription factors.

Unsupervised investigation of aggregated DNase-seq profiles showed that aggregate profiles contain a sub-profile that does not resemble a footprint but has a distinct shape. I showed that these sub-profiles are likely caused by DNase sequence bias, which was measured by performing DNase-seq experiments on genomic DNA.
Contrary to recent reports that claim either strong or no influence of sequence bias on footprint discovery [26, 31], I found that the influence of sequence bias varies according to the strength of transcription factor binding. High confidence binding sites (as measured by ChIP-seq) correspond to clear footprints, whereas low confidence binding sites exhibit a strong bias and no footprint. ChIP-seq verified binding sites without a footprint also tend to localize in regions with a high level of transcription. Recent reports of ChIP-seq artifacts in such regions [12, 13] explain the absence of footprints.

By incorporating sequence bias into the footprint models I built, the effect of bias on footprint discovery was reduced. Sequence bias profiles also differ from one TF to another. For a few TFs, the bias profile resembles the actual footprint profile. Therefore, it is harder to separate the footprint from the bias-driven profile for certain transcription factors, in which case modeling of bias becomes a necessity for accurate discovery of footprints.

I have also used footprints to characterize genome-wide changes in chromatin architecture after knockdown of CTCF expression, as illustrated in Chapter 5. By analyzing CTCF footprints and DHS sites, I found that CTCF binding is mainly driven by sequence specificity. Binding sites that closely resemble the CTCF sequence motif do not lose CTCF binding whereas weak sites result in either loss of just the footprint or overall chromatin accessibility.
Coupled with DNase-seq footprinting approaches, I analyzed binding sites of Jackdaw and Bluejay transcription factors in Arabidopsis root cells using ChIP-seq data. To accurately identify peaks of these factors, which showed broad enrichment due a difference in our ChIP-seq protocol, I designed a computational tool for identifying broad peaks, illustrated in Chapter 6. The new tool performed better than well-known peak callers such as MACS and SPP. This work further highlighted the importance of understanding the implications of the experimental protocol on the resulting data.

7.2 Discussion

Through their ability to regulate gene transcription, TFs play an important role in controlling many vital biological processes, ranging from cell cycle progression and maintenance of metabolic balance, to cellular differentiation [46-48]. Malfunction of transcription factors has been implicated in many diseases and disorders [104, 105]. Therefore, understanding TF binding mechanisms and discovery of TF binding sites are paramount to understanding transcriptional regulation and its role in complex biological processes.

Recently, ChIP-seq experiments and DNase-seq footprinting approaches have been widely used to map transcription factor binding sites in high throughput and on a genome-wide scale [23, 24, 39]. Even though they are useful for this purpose, each assay can be confounded by biases, artifacts and other factors related to the chromatin environment [12, 13, 31, 49]. By building the first footprint model that incorporates
DNase sequence bias, I showed that, at least in the case of transcription factors with known sequences motifs, DNase-seq footprinting can be a robust replacement for ChIP-seq for identifying high confidence binding sites. This raises the possibility of using a single DNase-seq assay to identify the genomic binding landscapes of multiple transcription factors simultaneously, instead of performing multiple ChIP-seq experiments.

DNase-seq footprints show enrichment in high quality ChIP-seq peaks and are depleted in regions of chromatin associated with ChIP artifacts, highlighting their ability to identify high confidence binding sites where TFs directly interact with their DNA motifs. Therefore, by footprinting we can identify possible artifacts and indirect or nonspecific TF-DNA interactions in ChIP-seq experiments, which can result in a better characterization of TF binding landscapes. Furthermore, alternative DNA motifs discovered by footprinting models show that footprinting offers a deeper characterization of TF-DNA interactions, as opposed to a binary picture of binding versus not binding.

However, there are challenges in using footprinting approaches in isolation to identify transcription factor binding sites. Indirect binding cannot be detected by footprinting. Some factors can interact with DNA in a non-sequence specific manner and certain families of TFs bind to similar sets of sequence motifs. For the former case, ChIP-
seq assays are necessary. For the latter, protein binding microarrays and analysis of flanking DNA sequence offer promise in determining transcription factor identity [106].

The advent of NGS-based assays has significantly increased our ability to study chromatin organization and gene transcription. After the accumulation of a large number of experimental data sets, various groups in the field of genomics have recently begun to characterize assay-specific artifacts and biases [12, 13, 30, 31, 67]. As we continue to use these assays to understand underlying biological processes, such as transcription factor binding, it is necessary to compare and contrast different methods and assays to better understand what exactly is measured by each assay, and to decipher the complexity of the underlying biological processes. Joint analysis [58, 59] of various assays offers greater promise in achieving a more holistic understanding of the complex process of transcription regulation.

### 7.3 Future directions

By building TF-specific base-pair resolution footprint models, I showed that these footprint models can capture variation in the sequence motif that corresponds to alternative footprints. Therefore, joint modeling of DNA sequence and DNase-seq footprint offers great promise in detecting variation in TF DNA binding motifs more accurately. Such a joint model may also be used in de novo motif discovery for ChIP-seq datasets, as I have shown that footprints aid in pinpointing the true binding site (6-20bp) in larger ChIP-seq peaks (100+ bp). By adopting a Hidden Markov Model framework,
such as micro-MUMMIE [107], DNase-seq fragment counts and DNA sequence may be modeled jointly. Using ChIP-seq peaks for training regions, a joint model can be used to discover DNA motif that overlap with footprints. This model can be extended to multiple footprint models of similar or different transcription factors. As a result, we can potentially discover co-binding of transcription factors and their joint footprints.

Characterization of ChIP-seq peaks of transcription factors using footprints revealed that not every ChIP-seq peak has a footprint. As my analyses and recent studies suggest, a portion of ChIP-seq peaks may result from artifacts and three dimensional chromatin interactions. To better understand the ChIP-seq landscape, Bayesian network models that describe ChIP-seq signal conditioned on footprints, chromatin state and three dimensional interactions can be a promising strategy. Such models will help annotate TF binding events across the genome and will add deeper insight into the binary view of transcription factor binding. Furthermore, such models will be able to learn dependencies between different factors that influence TF-DNA binding, and will offer insights into the mechanisms of transcriptional regulation of gene expression.
References


Biography

Galip Gürkan Yardımcı was born in the great city of Istanbul in 1982 to Nilgün Yardımcı and Gürol Yardımcı. A lover of chemistry, biology and computers; he chose to pursue computer science B.S. degree in İsik University in Istanbul (2002). Not satisfied with his education and desiring to study biology and genetics, he chose to pursue higher education and got his M.S. degree in Bioengineering at Sabancı University in Istanbul (2005). He received TUBITAK M.S. degree scholarship from Scientific and Technological Research Council of Turkey. Still not satisfied, he chose to challenge himself and pursue a PhD degree in USA where all cutting edge science is done. He was awarded Chancellor’s Scholarship at admission at Duke University and HHMI VIP scholarship during his studies at Computational Biology and Bioinformatics program at Duke University.

Publications

“Modified Association Rule Mining Approach for the MHC-Peptide Binding Problem”, G. Gürkan Yardımcı, Alper Küçükural, Yucel Saygın and Ugur Sezerman, LNCS, 2006

“Predicting Cell-Type Specific Gene Expression Using DNaseI Hypersensitive Site Data” Anirudh Natarajan, G. Gürkan Yardımcı, Nathan C. Sheffield, Gregory E. Crawford and Uwe Ohler. Genome Research 2012