Organization and evolution of transcription factor occupancy in the human genome

Jeffrey David Vierstra

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

John A. Stamatoyannopoulos, M.D., Chair

Philip Bradley, Ph.D.

Alejandro Wolff-Yadlin, Ph.D.

Program Authorized to Offer Degree:
Department of Genome Sciences

University of Washington

Abstract

Organization and evolution of transcription factor occupancy in the human genome

Jeffrey David Vierstra

Chair of the Supervisory Committee:
Assistant Professor John A. Stamatoyannopoulos, M.D.
Department of Genome Sciences

Cis-regulatory DNA encodes the circuitry that enables cell development and differentiation. *Cis*-regulatory DNA is densely populated by recogntition sequences for transcription factors and the cooperative binding TFs to these sequences determines cell-fate and function by the precise transcriptional regulation of their cognate genes. As such, a mechanistic understanding of gene regulation hinges on our ability to quantify transcription factor occupancy. To map transcription factor occupancy with in the human genome, I took part in the development of digital genomic footprinting – a technique leveraging the endonuclease DNase I that enables the unbiased and simultaneous detection of transcription factor occupancy genome-wide. We applied digital genomic footprinting to 41 diverse cell- and tissue-types to comprehensively map the human *cis*-regulatory lexicon. We show that this small genomic compartment contains an expansive repertoire of conserved recognition sequences for DNA-binding proteins and that nuclease patterns within these sequences mirror nucleotide-level evolutionary conservation and track the crystallographic topography of protein-DNA interfaces. We also show that both genetic and epigenetic variants affecting chromatin states are concentrated within footprints. Finally, we describe a large collection of novel regulatory factor recognition motifs that are highly conserved in both sequence and function, and exhibit cell-selective occupancy patterns that closely parallel major regulators of development, differentiation and pluripotency. These results provide for the first time an exhaustive map of TF occupancy within the human genome.

The architecture of individual *cis*-regulatory sites is critical for their function. While digital genomic footprinting provides rich information about the occupancy of TFs within individual *cis*-regulatory elements, it is currently not possible to resolve the genome-wide relationship of

transcription factors (TFs) and nucleosomes. To address this deficiency, I developed an extension to digital genomic footprinting that couples the detection of individual TF footprints to nucleosome occupancy. We find that TF occupancy is the major determinant of the positioning of *cis*-regulatory proximal nucleosomes, and that the positioning and occupancy of promoter-associated nucloeosomes is related to transcriptional start sites selection and output. The approach we describe provides a new view on the structure of *cis*-regulatory chromatin.

In the second part of this thesis, I used a comparative genomics approach to study the evolution of *cis*-regulatory DNA and protein occupancy. To do this, I mapped DNase I hypersensitive sites (DHSs) in 45 mouse cell types and primary tissues, and systematically compared these with human DHS maps from orthologous cell and tissue compartments. While I uncovered a small set of core regulatory sequences that encode a developmental program, the vast majority of *cis*-regulatory DNA is rapidly evolving independently in mouse and human. Overall, I find that the activity of *cis*-regulatory DNA is directly linked to the the composition of TF recognition sequences within and that the aggregate recognition sequence space for each transcription factor within accessible regulatory DNA of orthologous mouse and human cell types has been strictly conserved. These results demonstrate the remarkable plasticity of the mammalian *cis*-regulatory program and that TF occupancy is driven by an evolutionary inflexible *trans*-environment rather than conservation of individual regulatory elements.

Taken together, this thesis provides a framework to understand the organization and evolution of global TF occupancy within the mammalian genome.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Over the course of my dissertation I have been surrounded by tremendous group of individuals who collectively have shaped me as the fledging scientist that I am today.

First and foremost, I would like to thank my advisor John Stamatoyannopoulos for providing access to the resources and vision that guaranteed my success. Your persisent focus on the big picture helped shaped my own scientic vision.

Over the course of the last four years I had the pleasure of working with some truely amazing colleagues within the laboratory. The bioinformatics team has provided essential tools, knowledge and grit in completing projects. Thanks to Bob Thurman, Richard Sandstrom, Eric Haugen, Alex Reynolds, Audra Johnson for their computational expertise, attentivness and patience. I would especially like to thank Eric Rynes with whom I have worked closely – your accomplishments in science and music are truely inspirational. It is also necesary to acknowledge Shane Neph, whose hard-work, dry humor, friendship and development of a software package has made this thesis possible. Rich Humbert has been a great friend for exploring my curiosities outside of science.

This thesis is largely a product of the vast experimental data generated over many years and countless hours of optimization. I would like to thank in particular Theresa Canfield, Morgan Diegel, Shinny Vong, Jun Neri, Dan Bates, Scott Hansen, Tanya Kutyavin, Pete Sabo and Raj Kaul for creating one the most largest and high-quality genomics datasets in the world.

Countless hours of discussion, debate and collaboration played an integral part in shaping this thesis. In particular I would like to thank Matt Maurano his clarity of thought and scientific focus. I would be remiss to not mention Sam John as his mentorship was critical to my success. Above all, thanks to Hao Wang for experimental expertise and companionship – I am indebted by your generosity.

Finally, I would like to thank my wife Evelyn Salinas for her constant support and companionship.

The work presented in this thesis has been generously supported by the Genome Scienes Genomic Training Grant, the National Institute of Health ENCODE Project, and a National

Science Foundation Graduate Research Fellowship.

DEDICATION

For Evelyn Salinas

Chapter 1

INTRODUCTION

The completion of the human genome sequencing project (Lander et al., 2001) has given rise to large efforts to understand how nearly 3.3 billion bases encode development, differentiation and function of thousands of distinct cell and tissues types (ENCODE Project Consortium et al., 2007). Of the 3.3 Gb, only roughly 1% corresponds to translated sequence (i.e., the exome) that gives rise to proteins and the precise control of the expression and activity of these proteins determines the function of a cell. Recently, it has been revealed that the remaining 99% of the genome contains millions of *cis*-regulatory switches that encode a complex *cis*-regulatory circuitry (ENCODE Project Consortium et al., 2012; Thurman et al., 2012). *Cis*-regulatory DNA is densely populated by recogntion sequences for sequence-specific transcription factors whose occupancy control the transcription of the genes they regulate.

This thesis focuses on the the organization and evolution of transcription factor occupancy within the mammalian genome that gives rise to determisitic chromatin states that control cell identity and function. Part I describes the extensive mapping of transcription factor occupancy and *cis*-regulatory architecture with in the human genome using DNase I mapping. Part II describes evolutionary mechanisms of mammalian *cis*-regulatory DNA.

1.1 Experimental detection of *cis*-regulatory DNA

Chromatin accessibility is a hallmark of active eukaryotic *cis*-regulatory DNA. Operationally, regions of the accessible regions of the genome are defined by their exquistite sensistivity to cleavage action by various biological and chemical nucleases. Over 30 years ago, it was discovered that chromatin acessiblity marks all classes of *cis*-regulatory DNA (Wu, 1980) and since then, it has becoming fundemental in identifying *cis*-regulatory elements and dechipering *cis*-regulatory logic.

Accesible chromatin can be detected using a wide-variety of enzymatic (i.e., nucleases) and chemical agents each with unique properties and preferences. Of these, deoxyribonuclease I (DNase I) has proven to be the most robust for the *in vivo* detection of active *cis*-regulatory DNA. Regions hypersensitive to DNase I cleavage action are called DNase I hypersensitive

sites or DHSs.

1.1.1 Mapping nuclease hypersensitive sites

The basic method underlying the detection accessible regultory DNA is the exposure of chromatin to limiting amounts of a cleavage agent. As accessible chromatin is hypersensitive to cleavage action, a limiting amount of cleavage agent is critical, as the goal of this technique is to recover the cleavage products which is not possible if a digestion is driven to completion. The following section outlines some to the current and historical techniques used to detect DNase I hypersensitive sites.

Targeted detection

End-labeling. Traditionally, the mapping of hypersensitive sites has relied on indirect end labeling approaches. Following digestion, DNA derived from nuclei exposed to a cleavage agent is electropheretically separated and probed with a radiolabled probe targeting a region of interest.

Quantivative chromatin profiling. This method leverages the quantitative PCR (qPCR) to measure the relative amount of cleavage within the boundaries of a short amplicon (Dorschner et al., 2004). PCR primers densely tilled across a region of interest creating overlaping amplicons between 75 and 200 bp in length. The primer pairs are used to amplify DNA from digested and undigested nuclei. Due to increased cleavage activity, PCR amplicons overlaping accessible chromatin will have a a relative increase in the amplification cycles (C_t)[1]. As such, the ΔC_t is a measure of relative accessibility.

Genome-scale detection

Genome-scale detection of accessible chromatin requires the purification of DNase I cleavages. Over the past decade, a number of methodologies have been developed which have evolved with the rapid transformation of microarray and sequencing technology and currently culminate in leveraging recent advances in massively parallel sequencing.

DNase I fragment clone pools. A method devloped by Sabo et al. (2004) that captured the 5' DNase I cleavage sites within chromatin and mapped them to the genome in an unbiased manner. Briefly, DNase I nicks are labelled with a small DNA adapter labelled with biotin

[1] C_t is the number of cycles for the PCR detection strategy to reach a threshold. ΔC_t is a common approach to compare the relative quantities of DNA between independent experiments.

containing a recognition sequence for the Type IIS restriction endonuclease MmeI. Type IIS restriction enzyzmes are characterized by cleavage sites that are offset from the recognition sequence. Leveraging this property, cleavage of purified adapter-containing fragments, yields 20 bp fragments that can be oligomerized, cloned, and sequenced as a contiguous read using Sanger sequencing technology[2].

Tiling microarrays. Microarray technologies heralded in the era of genome-wide detection of *cis*-regulatory elements. Contemporaneously, Sabo et al. (2006) and Crawford et al. (2006) developed strategies to enrich for DNase I cleavages and hybrize them to a densely tiled microarray covering large portions of the non-coding genome. The relative accessibility was determined by mixing a labeled DNase I experiments performed on native chromatin vs. naked DNA. While both strategies employ array hybridization as the final detection modality, important differences in how DNase I cleavages are purified distiguish these two methodologies.

To create a library compatible for hybridization on a microarray, DNase I digested chromatin is reduced to small fragments with one or both ends generated by DNase I cleavage. To create these fragments, Crawford et al. (2006) utilized the Type IIS restriction endonuclease strategy employed by the oligomerizated clone pool detection method, such that DNase I cleavages were labeled with a biotinylated oligo containing a MmeI recognition sequence. In contrast, Sabo et al. (2006) utlized a novel strategy that specifically selected for 'two-hit' DNase I fragments directly from the digested chromatin. The key difference between these two strategies is the effective signal-to-noise ratio (SNR). While DNase I cleavage action is intensly focused within 1-3% of genome, the majority of total cleavages occur outside of DHSs. Although the strategy employed by Crawford et al. (2006) enriches for DNase I cleavages, the 'two-hit' enrichment strategy significantly enriches for clevages within DHS by selecting DNA fragments that are derived from two cleavages within close proximity (<500 bp).

Massively parallel sequencing. Sequencing of individual DNase I fragments is a natural extension to microarray technology and has enabled extensive regulatory mapping of many genomes (Boyle et al., 2008; Thurman et al., 2012)

1.1.2 Fine-scale mapping of individual hypersensitive sites

Within DNase I hypersensitive sites, DNase I cleavage is not uniform; rather punctuated binding by sequence-specific transcription factors occludes bound DNA from from cleavage, leav-

[2]Oligmerized DNase I clone pooles are perhaps the first sequencing based strategy to detect functional non-coding DNA elements in any genome.

ing 'footprints' that demarcate transcription factor occupancy. Numerous methods have been developed to detect the binding of transcription factors within chromatin, some of which are briefly described below.

In vitro footprinting. The discovery of DNase I footprinting over 30 years ago (Galas and Schmitz, 1978; Galas, 2001) revolutionized the the analysis of *cis*-regulatory sequences. The original footprinting technique was performed *in vitro* utilizing a ^{32}P–labeled DNA fragment, such that DNA fragment was then bound by a sequence specific transcription factor and exposed to DNase I. Separation of the resulting ^{32}P–labelled small fragments on a sequencing gel revealed the locations of 'footprints'.

Genomic footprinting. Genomic footprinting is a modified the *in vitro* method of (Galas and Schmitz, 1978) for detection of 'footprints' within large complex genomes. The techinque utilizes the electrophoretic transfer of DNA fragments from an acrylamide sequencing gel to a nylon membrane, which are detected by hybidization with a radiolabelled restriction fragment targeting a specific region.

Primer extension. This technique uses radiolabelled primers that anneal to denatured DNA fragments generated by nuclease digestion which are extended to the site of cleavage. Labelled and extended fragments are visualized directly on a denaturing sequencing gel.

Ligation-mediated PCR. LM-PCR is a hybrid approach that combines locus specific primer extension with an exponential amplification strategy allowing for site specific footprinting in large complex genomes (Dai et al., 2000). Here, single strand breaks such as those created by nuclease cleavages are converted to blunt-end duplex DNA fragments by primer extension. Following the primer extension, a linker oligonucleotide is ligated to the blunt ends, enabling the amplification of the fragments using a locus-specific primer and a sequence unique to the ligated linker. The products are end labeled with a third locus-specific primer and are visualized on a sequencing gel.

Digital genomic footprinting. Hesselberth et al. (2009) developed this method as an extension to DNase-seq. High quality DNase I sequencing libraries are deeply sequenced (>500 million sequencing tags) to reveal footprints genome-wide. Digital genomic footprinting was first demonstrated in *Sacchromyces cerevisae*, and has been used to derive TF occupancy maps in mammals (Boyle et al., 2011; Neph et al., 2012b), plants (Alessandra Sullivan, personal communication) and bacteria (Hui Li, personal communication). The application of digital genomic footprinting to human genome is the subject of the chapters found within Part II.

1.2 General features of accessible chromatin

1.2.1 Morphology of accessible chromatin

In the most general sense, accessible chromatin is characterized by a depletion of nucleosome occupancy. Within accesible DNA the per-nucleotide cleavage is non-uniform due to the protection of individual bases by occupancy of transcription factors. Most importantly, at individual regions the observed heterogenous per-nucleotide cleavage patterns are extremely reproducible between experiments performed under a similar set of conditions.

1.2.2 Proteins associated with accesible chromatin

The binding of TFs to their cognate recognition elements cooperatively evicts nucleosomes and establish an active chromatin element. DNase I hypersensitivity is quantitative marker of TF occupancy which is discussed further in Chapters 2 and 3. Many types of proteins are associated with *cis*-regulatory DNA which interpret the regulatory program encoded in the DNA sequence.

Histones. While the nucleosome comprised by a histone octamer is a general repressive element that occuldes access of transcription factors to their cognate binding sites, many nucleosome containing variant or post-translationally modified histones are associated within active or poised chromatin. For example, methylation of the fourth lysine reside on histon H3 (H3K4me3) is associated actively transcribed promoters (Santos-Rosa et al., 2002) and the variant histone H2A.Z is enriched at the -1, +1 and +2 positions (Schones et al., 2008). Although it is clear that modified and variant histones mark with active *cis*-regulatory DNA, their biological role is not well-understood.

Sequence-specific transcription factors. Transcription factors contain DNA binding domains that recognize specific physical structures that determined by the sequence motifs. The human genome has over 1,000 transcription factors (Vaquerizas et al., 2009) that differentially bind within accssible chromatin to enact *cis*-regulatory programs. The binding preferences for many transcription factors is unknown.

Chromatin remodellers. Sequence-specific transcription factors have been shown to tether chromatin remodelling complexes that deposit and position variant and post-transcriptionally modified histones (Ernst et al., 2001; Huang et al., 2011). As a result, general transcriptional regulators can recognize and occupy the remodelled chromatin. For example, a major subunit of

the RNA polyermase II comlex specifically recognizes and tighly binds H3K4 trimethylation (Vermeulen et al., 2007). Thus, evidence suggests that interactions between sequence specific transcription factors and chromatin remodellers provide machanism in which general transcription machinery can bind to specific regions in the genome.

1.2.3 Localization of *cis*-regulatory DNA within the genome

The vast majority of human *cis*-regulatory DNA resides distal to the transcription start sites (TSSs). Recent comprehensive mapping of DHSs within diverse human tissues demonstrated that >80% of *cis*-regulatory DNA is found further than 50 kb from annotated genes (Thurman et al., 2012). New technologies have recently demonstrated that these distally located regulatory elements are highly interconnected in a complex 3-dimensional structure in which distal elements interact with the 5' of genes (de Wit and de Laat, 2012) to promote the stable formation of transcriptional complexes (Weintraub, 1988).

1.3 Biology of accessible chromatin

Sequence specific transcription factors interpret the *cis*-regulatory sequences encoded within the genome to enact *cis*-regulatory programs that determine developmental trajectories, cell-fate descision and responses to enviromental stimuli. Binding by transcription factors to their cognate *cis*-regulatory sequences in the place of a canonical nucleosome triggers the remodelling of local chromatin structures resulting in accessible chromatin . The landscape accessible chromatin determines in *cis*-regulatory potential, thus the pattern of accessibilty within individual cells encodes a cellular identity (Stergachis et al., 2013b).

1.3.1 Constitutive, inducible, developmental, and tissue-specific

Broadly, *cis*-regulatory DNA can be described by its activity patterns both within the context of an individual cell-type or many cell types. Overall, each human cell type contains roughly 150,000-250,000 DHSs, collectively encompassing 1-3% of the nucleotide content of the 3.3 Gb haploid genome.

Constituitive. These regulatory regions are accessible in nearly all cell-types and conditions and typically mark the promoters and distal *cis*- elements of structural and housekeeping genes in a poised state. Overall, constituitively accessible *cis*-regulatory DNA comprises a small minority of the overall regulatory landscape in metazoans (<5%) (Thurman et al., 2012).

Tissue-specific. The *cis*-regulatory landscape is largely comprised by *cis*-regulatory DNA active in one or few cell- and tissue-types. The tremendous diversity of distal regulatory DNA provides a mechanism by which the transcription of individual genes can be regulated in different cellular contexts.

Inducible. Inducible regulatory DNA is almost exclusively activated by extracellular stimuli and is associated with changes in transcriptional output. One of the most well characterized indicible regulatory elements in the human genome is the enhancesome that regulates expression of the interferon-beta gene in cells that are infected with viruses. In this system, viral infection induces the cooperative binding of the transcription factors NF-$\kappa\beta$, IRF3 and the c-Jun complex to their target sites encoded within the *INFB* enhancesome core and recruit chromatin remodelling and transcription machinery that enable transcription of effector genes (Thanos and Maniatis, 1995). In many cases, although not all, inducible regulatory sites are associated with transcription of their linked genes.

Primed. *Cis*-regulatory priming allows for the rapid activation, as well as deactivation, of gene regulation. In constrast to inducible sites in which accessibility is tightly linked to the transcriptional activity, many *cis*-elements exist in the protentiated or primed state, such that they are accessible regardless of current regulatory activity. This class of *cis*-regulatory element is epitomized by sites which are bound by steroid hormone receptors such as glucocorticoid receptor (GR). Nuclear receptors bind their cognate ligand in the cytsol, which causes conformational changes leading to homo- and/or heterodimeization and translocation into the nucleus, where they opportunistically bind chromatin nearly exlusively into pre-accessible chromatin (John et al., 2011). Indeed, GR binding within the nucleus can be detected within minutes of exposure to glucocortoid hormones (John et al., 2011). It is now appreciated that the accessiblity of primed *cis*-regulatory DNA is maintained by other sequence-specific transcription factors; for example, primed regulatory sites that are GR targets are potentiated by the sequence-specfific transcription factor complex AP-1 (Biddie et al., 2011).

Developmental. This class of *cis*-regulatory element is associated with the developmental program such that they are induced during cellular differentiation and maturation.

1.4 Mechanisms for the establishment, maintenance and propagation of accessible chromatin

The precise manner in which *cis*-regulatory DNA is activated, maintained and propagated is largely unknown. Activation of regulatory DNA is generally to occur via the cooperative binding of transcription factors that can outcompete a nucleosome at individual chromatin templates. Not unlike allosteric ligand induced cooperativity found in hemaglobin, transcription factors do not need to interact direcly with eachother to induce cooperative eviction of nucleosomes (Mirny, 2010). Interestingly, DNA sequences found with the accessible chromatin lansdcape of a cell reflect the transcription factors active within that cell type and indicate that the *trans*-environment shapes the cis-regulatory landscape such that modulation of transcription factor concentrations can shape chromatin accessibility.

Classic experiments have demonstrated that regulatory DNA can be propagated in the absence of protein synthesis and that accessbility can be maintained for over 20 cellular generations (Groudine and Weintraub, 1982). These observations suggest the existence of a mechnism that can maintain the epigenetic state of individual chromatin templates and it has been prososed that positive feedback between transcription factor occupancy and the maintence of a permissive binding enviroment via chromatin remodellers (Voss and Hager, 2014). Indeed, transcription factors interact tightly with chromatin remodellers providing a plausible mechanism for the propagation of active *cis*-regulatory elements through cell division. As the experiments carried out by Groudine and Weintraub demonstrated, the accessibility of chromatin can be stably maintained for many generations without any associated function. Recent experiments comparing the accessible chromatin landscape between cells derived from distinct embryological orgins has suggested that in addition to an immediate regulatory capacity active *cis*-regulatory DNA may serve as a developmental memory or clock serving to inform current and future regulatory trajectories (Stergachis et al., 2013b). Indeed, the embrylogical origins of diverse cells can be recapitulated solely via their acessible chromatin landscapes. Thus, accessible chromatin most likely has both direct and indirect effects on genome regulation.

1.5 Aims of thesis

In this thesis I aim to provide insight into the organization and function of mammalian *cis*-regulatory elements. In part I, I develop a framework to investigate the fine-scale structure of *cis*-regulatory DNA and provide mechnasitic insight into how the collective binding of tran-

scription factors organizes chromatin and drive transcriptional regulation. Chapter 1 describes how chromatin accessibility reflects the cumulative occupancy of transcription factors within the genome. Chapter 2, 3, and 4 describe the per-nucleotide occupancy patterns of transcription factors and adjacent nucleosomes at base-pair resolution in diverse human cell types. Finally, in Part II I use a comparative genomics approach to understand the constraint on individual *cis*-regulatory elements over evolutionary timescales.

Part I

QUANTIFYING PROTEIN OCCUPANCY WITHIN THE HUMAN GENOME

Chapter 2

TRANSCRIPTION FACTOR DRIVERS OF CHROMATIN ACCESSBILITY

This chapter has been adapted with minor changes from a section in: Thurman, R. E. et al. The accessible chromatin landscape of the human genome. *Nature.* 489, 75—82 (2012).

2.1 Results

DNase I hypersensitive sites result from cooperative binding of transcriptional factors in place of a canonical nucleosome (Felsenfeld et al., 1996; Gross and Garrard, 1988). To quantify the relationship between chromatin accessibility and the occupancy of regulatory factors, we compared sequencing-depth-normalized DNase I sensitivity in the ENCODE common cell line K562 to normalized chromatin immunoprecipitation and high-throughput sequencing (ChIP-seq) signals from all 42 transcription factors mapped by ENCODE ChIP-seq (ENCODE Project Consortium et al., 2012) in this cell type (Figure 2.1a–b). Simple summation of the ChIP-seq signals markedly parallels quantitative DNase I sensitivity at individual DHSs (Figure 2.1a–b) and across the genome ($r = 0.79$, Figure 2.2a). For example, the β-globin locus control region contains a major enhancer element at hypersensitive site 2 (HS2), which appears to be occupied by dozens of transcription factors (Figure 2.1b). Such highly overlapping binding patterns have been interpreted to signify weak interactions with lower-affinity recognition sequences potentiated by an accessible DNA template (Biggin, 2011). However, HS2 is a compact element with a functional core spanning ≈110 bp that contains 5–8 sites of transcription factor–DNA interaction *in vivo* depending on the cell type (Forsberg et al., 2000; Reddy et al., 1994; Talbot and Grosveld, 1991). The fact that the cumulative ChIP-seq signal closely parallels the degree of nuclease sensitivity at HS2 and elsewhere is thus most readily explained by interactions between DNA-bound factors and other interacting factors that collectively potentiate the accessible chromatin state (Figure 2.2b). Given the relatively limited number of factors studied, it may seem surprising that such a close correlation should be evident. However, most of the factors selected for ENCODE ChIP-seq studies have well-described or even fundamental roles in transcriptional regulation, and many were identified originally based on their high affinity for DNA. Alternatively, as originally proposed by (Weisbrod and Weintraub, 1979), a limited number of factors may be involved in establishment and maintenance of chromatin remodelling, whereas others may interact nonspecifically with the remodelled state. We also found that the recognition sequences for a small number of factors were consistently linked with elevated chromatin accessibility across all classes of sites and all cell types (not shown), indicating that regulators acting through these sequences are key drivers of the accessibility landscape.

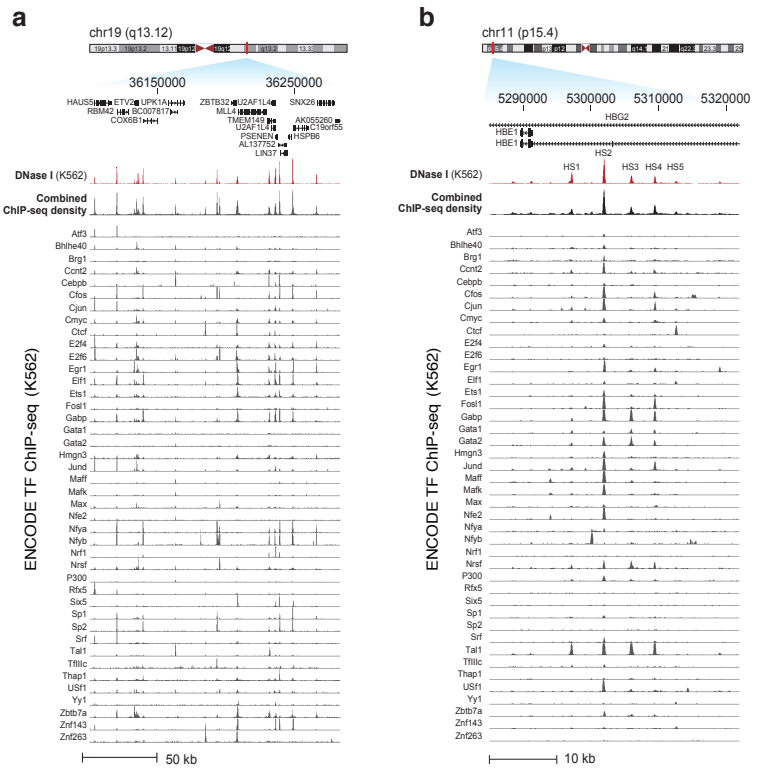

Figure 2.1: **DNase I accessbility mirrors TF occupancy**. (a) DNase□I tag density is shown in red for a 175-kb region of chromosome 19. Below, normalized ChIP-seq tag density for 45 ENCODE ChIP-seq experiments from K562 cells, with a cumulative sum of the individual tag density tracks shown immediately below the K562 DNase I data. (b) Same as (a) for a genomic segment emcompassing the β-globin locus control region.

2.2 Methods

2.2.1 ChIP-seq signal processing

Raw sequencing tags (BAM format) from ChIP-seq experiments in K652 cells were downloaded from the ENCODE DCC. Sequencing tags from replicate experiments were merged and mapped to hg19 with BWA using default settings. Tag densities were calculated in 150-bp sliding windows every 20 bp over the entire genome and normalized to 10 million reads. Aggregate transcription factor occupancy was computed by summation of the normalized ChIP-seq densities for individual factors (n = 42). The pair-wise Pearson correlation was computed between

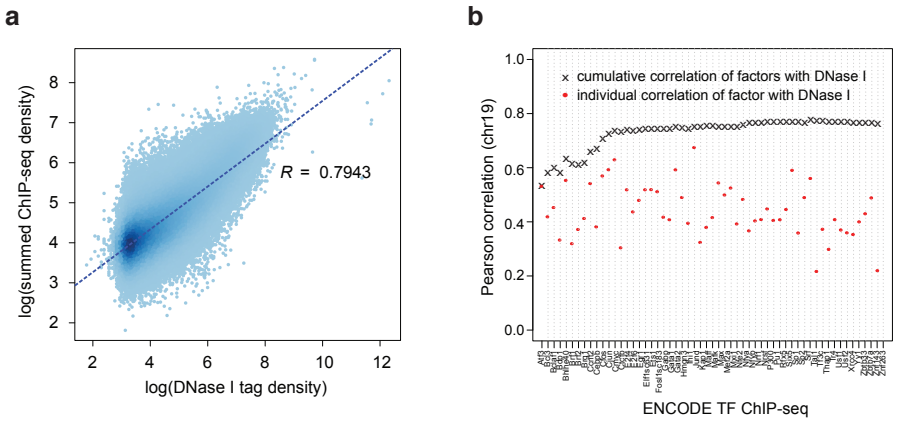

Figure 2.2: **Quantitative relationship of DNase I accessbility and TF ChIP-seq signal**. (a) Genome-wide correlation ($r = 0.7943$) between ChIP-seq and DNase I tag densities (log_{10}) in K562 cells. (b) Additive correlation (y-axis) of ChIP-seq with DNaseI across chromosome 19 with increasing numbers of TFs. TFs are ordered alphabetically (x-axis). Correlation values for individual factors are shown in red.

DNaseI accessibility and transcription factor occupancy in DNase I peaks using normalized DNaseI and the aggregate ChIP-seq density at DHS peaks. Cumulative Pearson correlations of DNase I density and ChIP-seq densities were iteratively calculated for the entire chromosome 19 by the sequential addition of transcription factor ChIP-seq densities in the order specified in Figure 2.2.

Chapter 3

MAPPING TRANSCRIPTION FACTOR FOOTPRINTS GENOME-WIDE

This chapter has been adapted with minor changes from: *Neph, S., *Vierstra J., *Stergachis A.B., *Reynolds A. et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature 489, 83–90 (2012). Asterisk denotes equal contribution to manuscript.

Abstract

Regulatory factor binding to genomic DNA protects the underlying sequence from cleavage by DNase I, leaving nucleotide-resolution 'footprints'. Using genomic DNase I footprinting across 41 diverse cell and tissue types, we detected 45 million transcription factor occupancy events within regulatory regions, representing differential binding to 8.4 million distinct short sequence elements. Here we showthat this small genomic sequence compartment, roughly twice the size of the exome, encodes an expansive repertoire of conserved recognition sequences for DNA-binding proteins that nearly doubles the size of the human *cis*-regulatory lexicon. We find that genetic variants affecting allelic chromatin states are concentrated infootprints, and that these elements are preferentially sheltered from DNA methylation. High-resolution DNase I cleavage patterns mirror nucleotide-level evolutionary conservation and track the crystallographic topography of protein-DNA interfaces, indicating that transcription factor structure has been evolutionarily imprinted on the human genome sequence. We identify a stereotyped 50-base-pair footprint that precisely defines the site of transcript origination within thousands of human promoters. Finally, we describe a large collection of novel regulatory factor recognition motifs that are highly conserved in both sequence and function, and exhibit cell-selective occupancy patterns that closely parallel major regulators of development, differentiation and pluripotency.

3.1 Introduction

Sequence-specific transcription factors interpret the signals encoded within regulatory DNA. The discovery of DNase I footprinting over 30 years ago (Galas and Schmitz, 1978) revolutionized the analysis of *cis*-regulatory sequences in diverse organisms, and directly enabled the discovery of the first human sequence-specific transcription factors (Dynan and Tjian, 1983). Binding of transcription factors to regulatory DNA regions in place of canonical nucleosomes triggers chromatin remodelling, resulting in nuclease hypersensitivity (Gross and Garrard, 1988). Within DNase I hypersensitive sites (DHSs), DNase I cleavage is not uniform; rather, punctuated binding by sequence-specific regulatory factors occludes bound DNA from cleavage, leaving footprints that demarcate transcription factor occupancy at nucleotide resolution (Galas and Schmitz, 1978; Hesselberth et al., 2009) (Figure 3.1). DNase I footprinting has been applied widely to study the dynamics of transcription factor occupancy and ooperativity within regulatory DNA regions of individual genes (Thanos and Maniatis, 1995), and to identify cell- and lineage-selective transcriptional regulators (Tsai et al., 1989).

3.2 Results

3.2.1 Regulatory DNA is populated with DNase I footprints

To map DNase I footprints comprehensively within regulatory DNA, we adapted digital genomic footprinting (Hesselberth et al., 2009) to human cells. The ability to resolve DNase I footprints sensitively and precisely is critically dependent on the local density of mapped DNase I cleavages (Figure 3.2a–d), and efficient footprinting of a large genome such as human requires substantial concentration of DNase I cleavages within the small fraction (1-3%) of the genome contained in DNase I-hypersensitive regions. We selected highly enriched DNase I cleavage libraries from 41 diverse cell types in which 53–81% of DNase I cleavage sites localized to DNase I-hypersensitive regions (Thurman et al., 2012) (Table 3.1), representing nearly tenfold higher signal-to-noise ratio than previous results from yeast (Hesselberth et al., 2009), and two- to fivefold greater enrichment than achieved using end-capture of single DNase I cleavages (Boyle et al., 2008; Sabo et al., 2004). We then performed deep sequencing of these libraries, and obtained 14.9 billion Illumina sequence reads, 11.2 billion of which mapped to unique locations in the human genome (Table 3.1). We achieved an average sequencing depth of >273 million DNase I cleavages per cell type that enabled extensive and accurate discrimination of DNase I

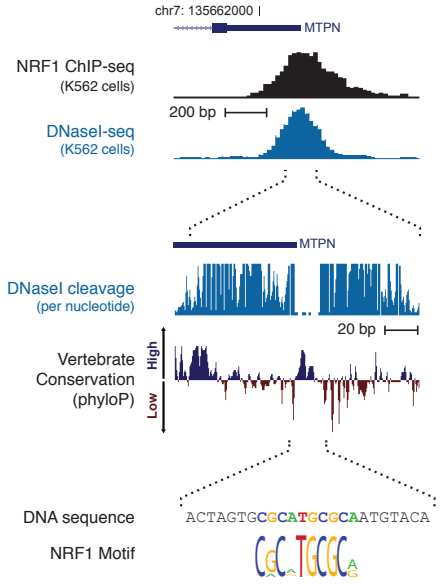

Figure 3.1: **Digital genomic footprinting in human cells**. DNase I footprinting of K562 cells identifies the individual nucleotides within the *MTPN* promoter that are bound by NRF1.

footprints.

To detect DNase I footprints systematically, we implemented a detection algorithm based on the original description of quantitative DNase I footprinting (Galas and Schmitz, 1978) (Methods). We identified an average of >1.1 million high-confidence (false discovery rate (FDR) of 1%) footprints per cell type (range 434,000 to 2.3 million; Table 3.1), and collectively 45,096,726 6–40 base pair (bp) footprint events across all cell types. We resolved cell-selective footprint patterns to reveal 8.4 million distinct elements with a footprint, each occupied in one or more cell type. At least one footprint was found in >75% of DHSs (Figire 3.2c–d and Table 3.2), with detection strongly dependent on the number of mapped DNase I cleavages within each DHS. 99.8% of DHSs with >250 mapped DNase I cleavages contained at least one footprint, indicating that DHSs are not simply open or nucleosome-free chromatin features, but are constitutively populated with DNase I footprints. Modelling DNase I cleavage patterns using empirically derived intrinsic DNA cleavage propensities for DNase I showed that only a miniscule fraction (0.24%) of discovered 1% FDR footprints from cell and tissue samples could be caused

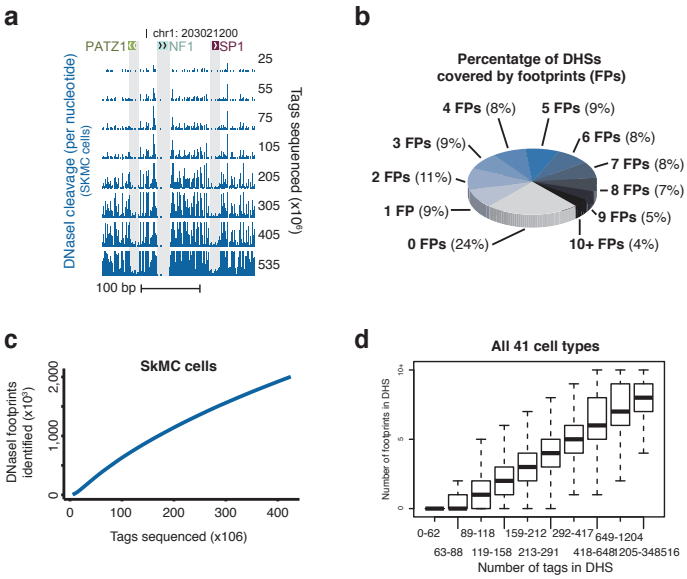

Figure 3.2: **Identification and distribution of DNase I footprints**. (a) As more DNase I cleavages are sequenced from SKMC cells, individual DNase I footprints are easier to distinguish. (b) The number of DNase I footprints identified in SKMC cells at varying DNase I cleavage tag sequencing depths. (c) The majority (76%) of SKMC DHSs contain at least one footprint. (d) The number of footprints detected in each DHSs varies with sequencing coverage. DHSs from all 41 cell types were broken into deciles determined by sequence tag coverages. The box plot shows that distribution of the number of footprints within DHSs for each decile within each of the 41 cells types assayed.

by inherent DNase I sequence specificity (Methods).

DNase I footprints were distributed throughout the genome, including intergenic regions (45.7%), introns (37.7%), upstream of transcriptional start sites (TSSs, 8.9%), and in 5' and 3' untranslated regions (UTRs, 1.4% and 1.3%, respectively; Figure 3.3a–b). DNase I footprints were enriched in promoters (3.6-fold; $P < 2.2 \times 10^{-16}$; Binomial test) and 59 UTRs (2.4-fold; $P < 2.2 \times 10^{-16}$; Binomial test), commensurate with high DNase I cleavage densities observed in these regions. We found that 2.0% of footprints localized within exons, raising the possibility that occupancy by DNA binding proteins could further restrict sequence diversity within coding DNA, thus superimposing an unexpected layer of constraint on codon usage.

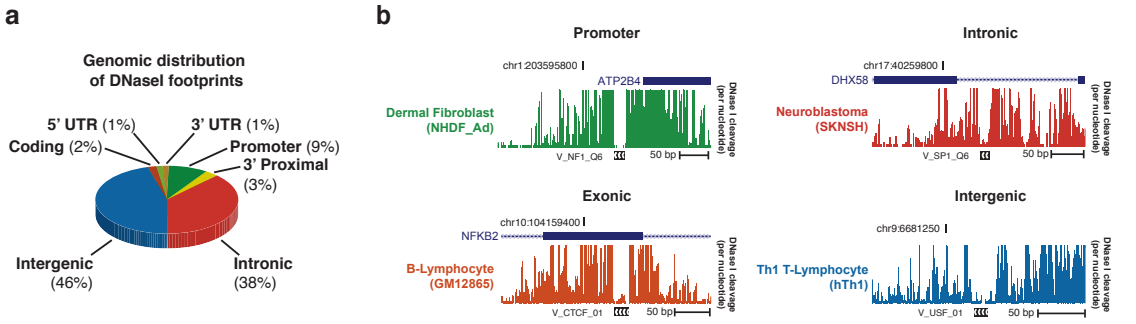

Figure 3.3: **Genomic distribution of DNase I footprints**. (a) The genomic distribution of footprints found in 41 cell types in relation to annotated genomic features. (b) Examples of DNase I footprints associated with different genomic features.

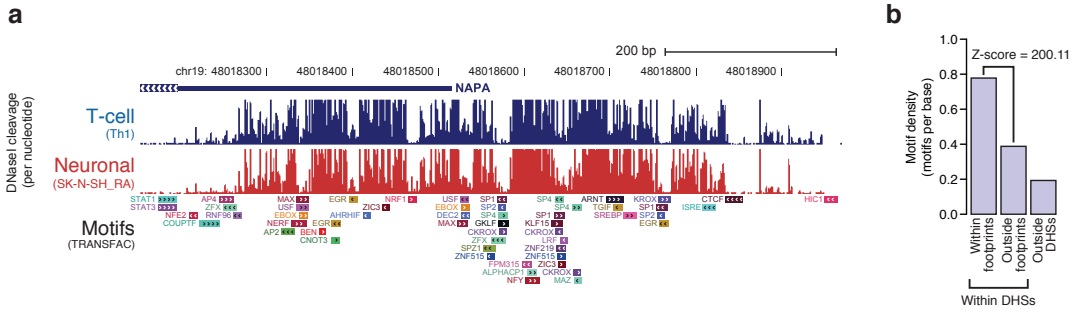

Figure 3.4: **DNase I foorprints are populated with the recognition sequences for sequence-specific transcription factors**. (a) Example locus harboring eight clearly defined DNase I footprints in T-helper type I (T_H1) and SKnSH (stimulated with retinoic acid) with TRANSFAC database motif instances indicated below. (b) The density of TRANSFAC motifs in DNase I footprints, DHSs (not in footprints) and non-DNase I hypersensitive regions. Motifs are significantly enriched in footprints (Z-score = 204.22; Genome Structure Correction).

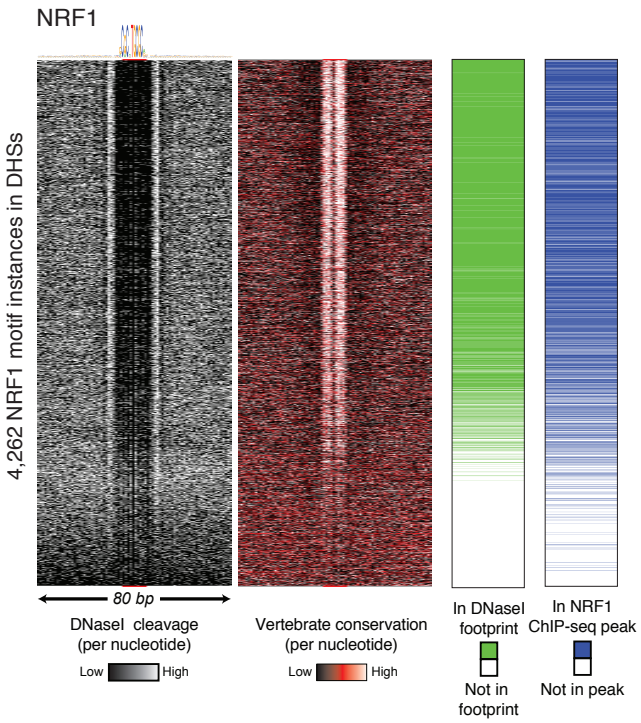

Figure 3.5: **Per-nucleotide cleavage patterns reflect TF occupancy**. (a) Heat maps showing per-nucleotide DNase I cleavage (left) and vertebrate conservation by phyloP (right) for 4,262 NRF1 motif within K562 DHSs ranked by the local density of DNase I cleavages. Green ticks indicate the presence of DNase I footprints over motif instances. Blue ticks indicate the presence of ChIP-seq peaks over the motif instances.

3.2.2 Footprints are quantitative markers of factor occupancy

We next examined the correspondence between DNase I footprints and known regulatory factor recognition sequences within DNase I hypersensitive chromatin. Comprehensive scans of DNase I hypersensitive regions for high-confidence matches to all recognized transcription factor motifs in the TRANSFAC (Matys et al., 2006) and JASPAR (Bryne et al., 2008) databases revealed a striking enrichment of motifs within footprints ($P < 0$, $Z = 204.22$ for TRANSFAC; $Z = 169.88$ for JASPAR; Figure 3.4a–b).

To quantify the occupancy at transcription factor recognition sequences within DHSs genome-wide, we computed for each instance a footprint occupancy score (FOS) relating the density of DNase I cleavages within the core recognition motif to cleavages in the immediately flank-

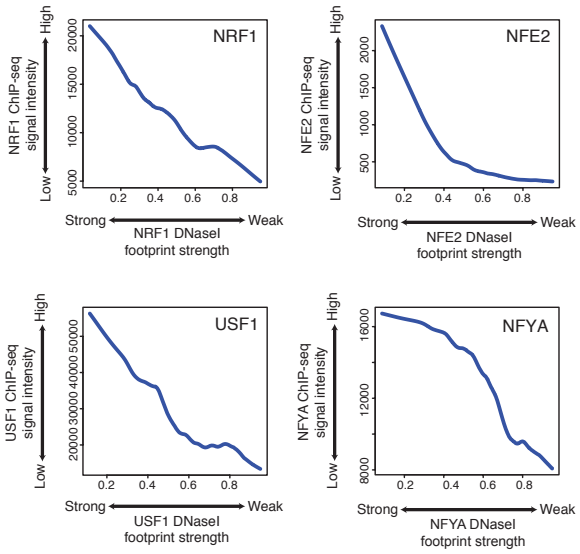

Figure 3.6: **DNase I footprints mirror orthogonal measures of TF occupancy**. Lowess regression of NRF1, USF1, NFE2, and NFYA ChIP-seq signal intensities versus DNase I footprinting occupancy (footprint occupancy score) at DNase I footprints containing their cognate recognition sequences.

ing regions (see Methods). The FOS can be used to rank motif instances by the 'depth' of the footprint at that position, and is expected to provide a quantitative measure of factor occupancy (Galas and Schmitz, 1978). To examine this relationship for a well-studied sequence-specific regulator (NRF1; Chan et al., 1993), we plotted DNase I cleavage patterns surrounding all 4,262 NRF1 motifs contained within DHSs and ranked these by FOS. Whereas only a subset of these motif instances (2,351) coincided with high-confidence footprints, the vast majority of NRF1 motif instances in DNase I footprints (89%) overlapped reproducible sites of NRF1 occupancy identified by chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) (Figure 3.5). In parallel, we analysed nucleotide-level evolutionary conservation patterns around NRF1-binding sites, revealing that FOS closely parallels phylogenetic conservation within the core motif region, indicating strong selection on factor occupancy (Figure 3.5). We observed a nearly monotonic relationship between FOS and ChIP-seq signal intensities at NRF1-binding sites within DNase I footprints of K562 cells (3.7a). Similarly strong correlations between footprint occupancy and either ChIP-seq signal or phylogenetic conservation were

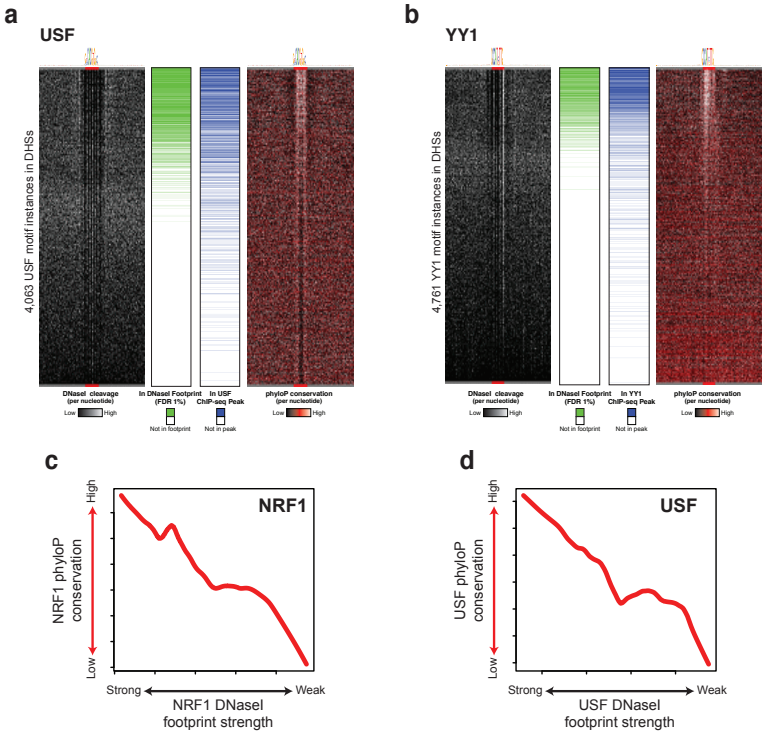

Figure 3.7: **Association of footprint, occupancy, and sequence conservation**. (a–b) Heat maps showing per-nucleotide DNase I cleavage (left) and vertebrate conservation as in Figure 3.5 for USF and YY1 motifs within K562 DHSs ranked by tag density. (c–d) Lowess regression of NRF1 and USF1 maximum phyloP score versus DNase I footprint occupancy score at K562 DNase I footprints marked by NRF1 and USF1 motifs.

evident for diverse factors (Figure 3.6 and Figure 3.7a–d). We found that footprint occupancy and nucleotide-level conservation correlated for 80% of all transcription factor motifs in the TRANSFAC database, of which 50% were statistically significant ($P < 0.05$; Methods). This relationship between footprint occupancy and conservation is most readily explained by evolutionary selection on factor occupancy, with higher conservation of higher affinity binding sites. Taken together, these results indicate that footprint occupancy provides a quantitative measure of sequence-specific regulatory factor occupancy that closely parallels evolutionary constraint and ChIP-seq signal intensity.

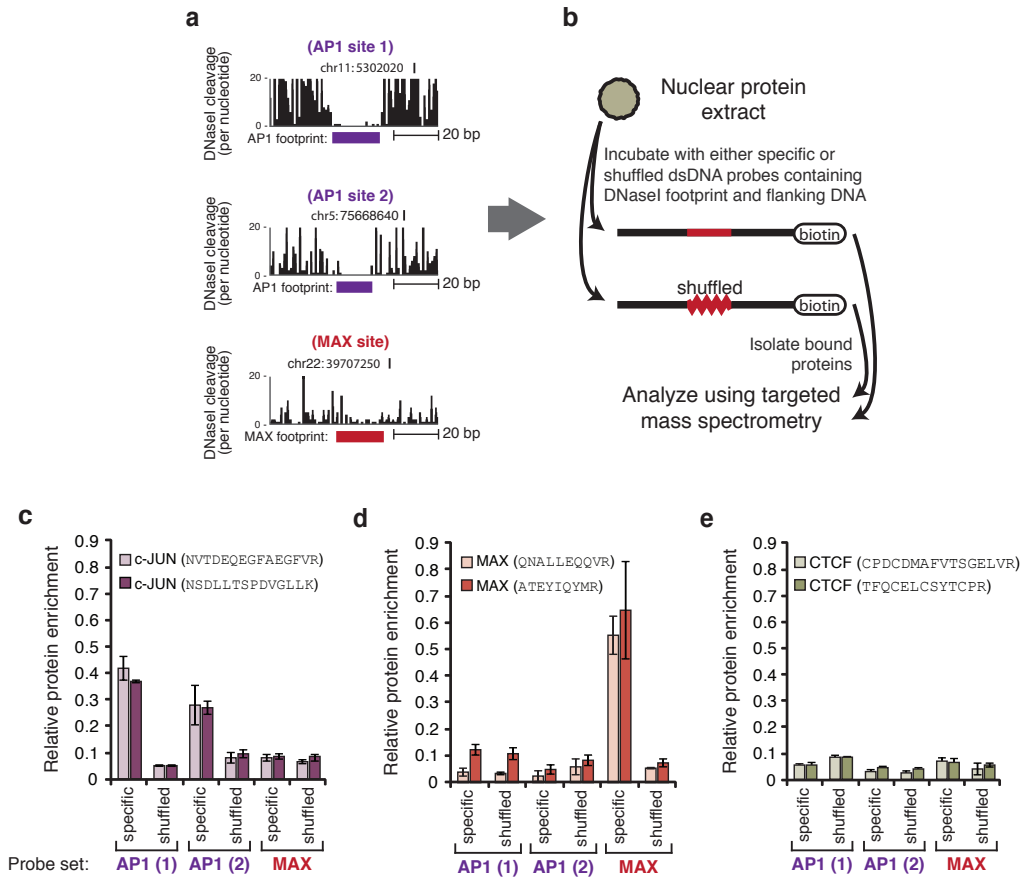

Figure 3.8: **Validation of footprints as potential sites of protein occupancy *in vitro***. (a), Three genomic loci of varying footprint strength targeted using DNA interacting protein precipitation (DIPP). (b) Schematic overview of the DIPP protocol. (c–d), Targeted mass spectrometry measurements of the proteins enriched using the different probe sets. The AP1 protein c-Jun was enriched specifically using the AP1 probes (c) and MAX was enriched specifically using the MAX probe (d). (e) As a negative control for DIPP, we tested for CTCF binding to the six probes. CTCF did not appear to be enriched in any of the pulldowns.

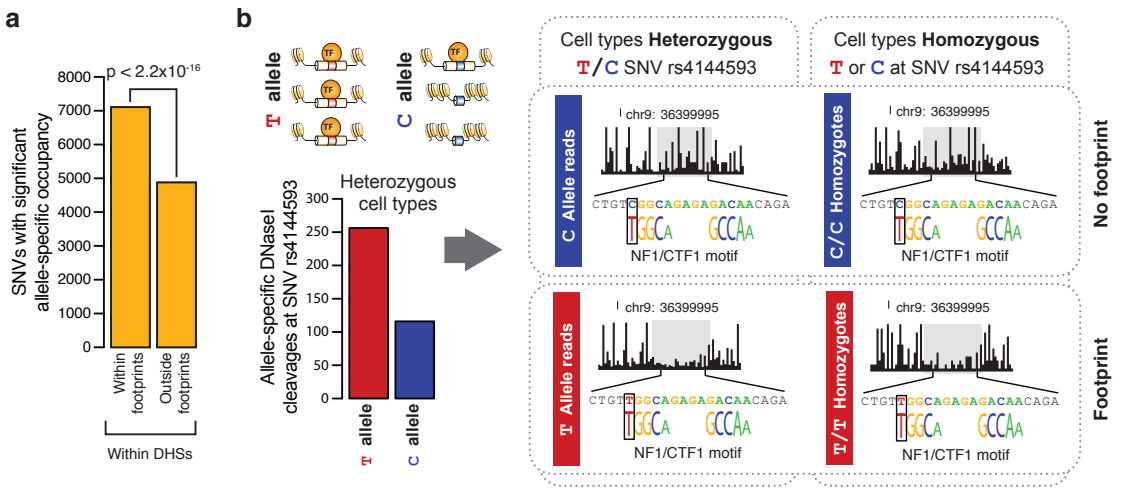

Figure 3.9: **Genetic variation within DNase I footprints modulates TF occupancy**. (a) Heterozygous SNVs associated with allele-specific occupancy are significantly enriched inside footprints compared to the rest of the DHS ($P < 2.2 \times 10^{-16}$, Fisher's exact test). (b) Schematic and plots showing the effect of T/C SNV rs4144593 on protein occupancy and chromatin accessibility. The y axis of the bar graph shows the number of DNase I cleavage events containing either the T or C allele. Middle plots show T or C allele-specific DNase I cleavage profiles from ten cell lines heterozygous for the T/C alleles at rs4144593. Right plots show DNase I cleavage profiles from 18 cell lines homozygous for the C allele at rs4144593 and one cell line homozygous for the T allele at rs4144593. Cleavage plots are cut off at 60% cleavage height.

To validate the potential for selective binding of footprints by factors predicted on the basis of motif-to-footprint matching, we developed an approach to quantify specific occupancy in the context of a complex transcription factor milieu using targeted mass spectrometry (DNA interacting protein precipitation or DIPP; Methods). Using DIPP, we affirmed specific binding by several different classes of transcription factor (Figure 3.8a–e). Together with the analysis of ChIP-seq data described above, these results indicate that the localization of transcription factor recognition motifs within DNase I footprints can accurately illuminate the genomic protein occupancy landscape.

3.2.3 Footprints harbour functional SNVs and lack methylation

The potential for single nucleotide variants (SNVs) within a transcription factor recognition sequence to abrogate binding of its cognate factor is well known (Rockman and Wray, 2002).

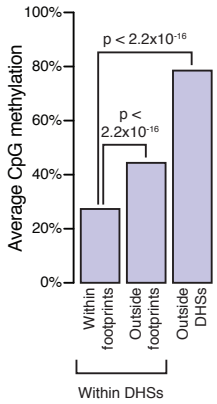

Figure 3.10: **Epigenetic variation within DNase I footprints**. The average CpG methylation within IMR90 DNase I footprints, IMR90 DHSs (but not in footprints) and non-hypersensitive genomic regions in IMR90 cells. CpG methylation is significantly depleted in DNase I footprints ($P < 2.2 \times 10^{-16}$, Mann-Whitney U-test).

The depth of sequencing performed in the context of our footprinting experiments provided hundreds- to thousands-fold coverage of most DHSs, enabling precise quantification of allelic imbalance within DHSs harbouring heterozygous variants. We scanned all DHSs for heterozygous SNVs identified by the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2010) and measured, for each DHS containing a single heterozygous variant, the proportion of reads from each allele. We identified likely functional variants conferring significant allelic imbalance in chromatin accessibility and analysed their distribution relative to DNase I footprints. This analysis revealed significant enrichment ($P < 2.2 \times 10^{-16}$; Fisher's exact test) of such variants within DNase I footprints (Figure 3.9a). For example, rs4144593 is a common T-to-C (T/C) variant that lies within a DHS on chromosome 9. This variant falls on a high-information position within a footprint containing an NF1/CTF1 motif and substantially disrupts footprinting of this motif, resulting in allelic imbal- ance in chromatin accessibility (Figure 3.9b).

Protein–DNA interactions are also sensitive to cytosine methylation (Lister et al., 2009; Tate and Bird, 1993). Comparing DNase I footprints and whole-genome bisulphite sequencing methylation data from pulmonary fibroblasts (IMR90), we found that CpG dinucleotides contained within DNase I footprints were significantly less methylated than CpGs in non-footprinted regions of the same DHS ($P < 2.2 \times 10^{-16}$; Mann–Whitney U-test; Figure 3.10).

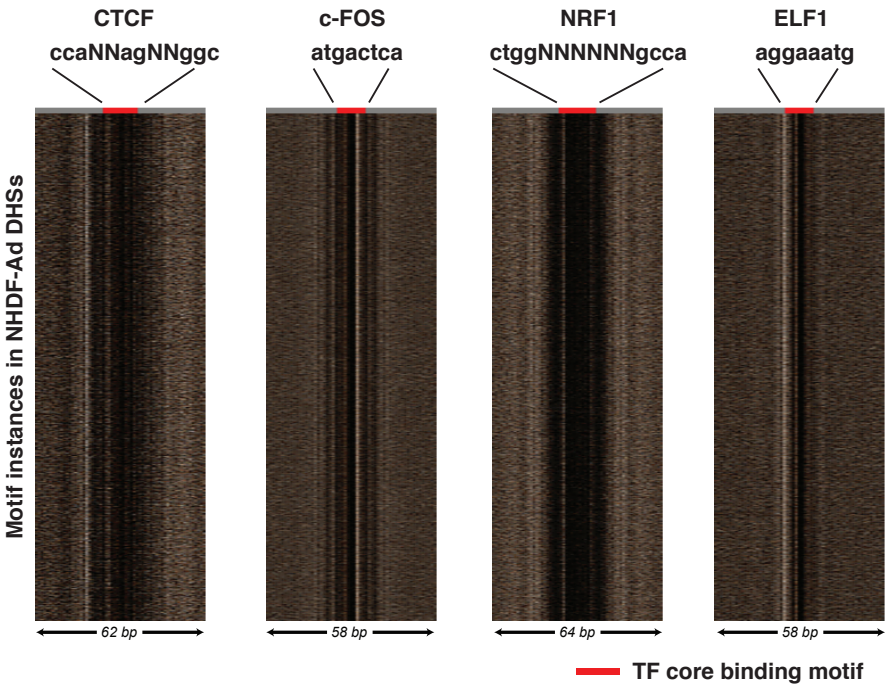

Figure 3.11: **Stereotyped cleavage patterns for different TFs**. The per-nucleotide DNase I cleavage patterns at motif instances of 4 different transcription factors in adult dermal fibroblasts (NHDF-Ad). The different motif instances (rows) are randomly ordered.

Footprints therefore seem to be selectively sheltered from DNA methylation, indicating a widespread connection between regulatory factor occupancy and nucleotide-level patterning of epigenetic modifications.

3.2.4 Transcription factor structure is imprinted on the genome

We observed surprisingly heterogeneous base-to-base variation in DNase I cleavage rates within the footprinted recognition sequences of different regulatory factors. And yet, the per site cleavage profiles for individual factors were highly stereotyped, with nearly identical ocal cleavage patterns at thousands of genomic locations (Figure 3.11). This raised the possibility that DNase I cleavage patterns may provide information concerning the morphology of the DNA–protein interface. We obtained the available DNA–protein co-crystal structures for human transcrip-

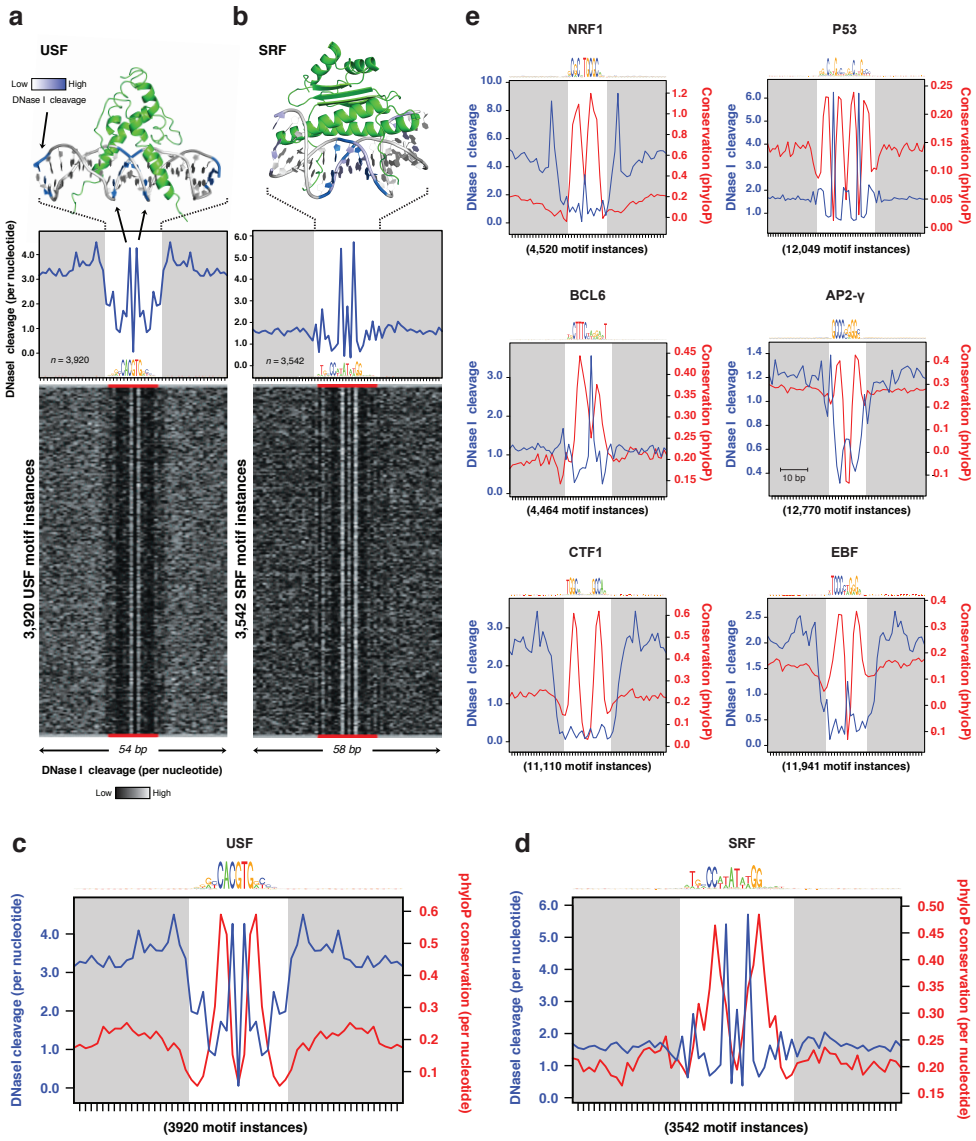

Figure 3.12: **Footprint structure parallels transcription factor structure and is imprinted on the human genome.**. (a–b) The co-crystal structure of upstream stimulatory factor (USF1) and serum response factor (SRF) bound to their DNA ligands is juxtaposed above the average nucleotide-level DNase I cleavage pattern (blue) at motif instances in DNase I footprints. Nucleotides that are sensitive to cleavage by DNase I are coloured blue on the co-crystal structure. The motif logo generated from DNase I footprints is displayed below the DNase I cleavage pattern. Below is a randomly ordered heat map showing the per-nucleotide DNase I cleavage for each motif instance of USF in DNase I footprints. (c–e) The per-base DNase I hypersensitivity (blue) and vertebrate phylogenetic conservation (red) for all DNase I footprints in dermal fibroblasts matching three well-annotated transcription factor motifs. The white box indicates width of consensus motif. The number of motif occurrences within DNase I footprints is indicated below each graph.

tion factors, and mapped aggregate DNase I cleavage patterns at individual nucleotide positions onto the DNA backbone of the co-crystal model. Figure 3.12a–b show two examples: USF1 (Ferré-D'Amaré et al., 1994) and SRF (Pellegrini et al., 1995). For both factors, DNase I cleavage patterns clearly parallel the topology of the protein–DNA interface, including a marked depression in DNase I cleavage at nucleotides involved in protein–DNA contact, and increased cleavage at exposed nucleotides such as those within the central pocket of the leucine zipper. These data show that nucleotide-level aggregate DNase I cleavage patterns reflect fundamental features of the protein–DNA interaction interface at unprecedented resolution.

We next asked how these patterns related to evolutionary conservation. Plotting nucleotide-level aggregate DNase I cleavage in parallel with per-nucleotide vertebrate conservation calculated by phyloP (Pollard et al., 2010) revealed striking antiparallel patterning of cleavage versus conservation across nearly all motifs examined (six representative examples are shown in Figure 3.12c–e). Notably, conservation is not limited to only DNA contacting protein residues, but exhibits graded changes that mirror DNase I accessibility across the entirety of the protein–DNA interface (Figure 3.12c–d). Taken together, these results imply that regulatory DNA sequences have evolved to fit the continuous morphology of the transcription factor–DNA binding interface.

3.2.5 A 50-bp footprint localizes transcription initiation

Transcription initiation requires the binding of multi-protein complexes that position RNA polymerase II (Buratowski et al., 1989; Kim et al., 1997, 2005; Pugh and Tjian, 1991). Using a modified footprint detection algorithm designed to detect larger features (see Methods), we scanned the regions upstream from GENCODE TSSs and identified highly stereotyped 80-bp chromatin structure comprising a prominent 50-bp central DNase I footprint, flanked symmetrically by 15-bp regions of uniformly elevated levels of DNase I cleavage (Figure 3.13a). Alignment of per-nucleotide DNase I cleavage profiles from 5,041 prominent footprints mapped in different K562 promoters highlights the homogeneous, nearly invariant nature of the structure (Figure 3.13b).

Plotting evolutionary conservation in parallel with DNase I cleavage revealed two distinct peaks in evolutionary conservation within the central footprint (Figure 3.13c) compatible with binding sites for paired canonical sequence-specific transcription factors. The density of capped analysis of gene expression (CAGE) tags (Figure 3.13d; green line) and ends of expressed sequenced tags (ESTs) (Figure 3.13d; orange line) relative to the central 50-bp footprint revealed

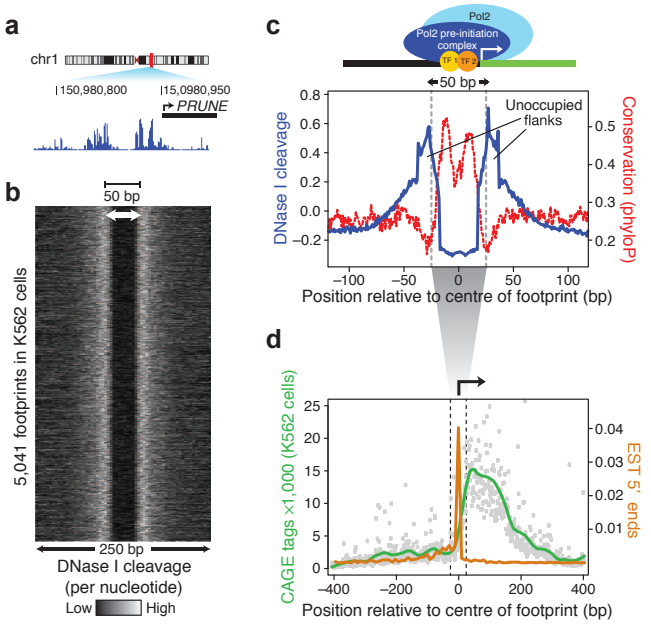

Figure 3.13: **A highly stereotyped chromatin structural motif marks sites of transcription initiation in human promoters**. (a) A 35–55-bp footprint is the predominant feature of many promoter DHSs and is in tight spatial coordination with the transcription start site. (b) Heat map of the per-nucleotide DNase I cleavage pattern at 5,041 instances of this stereotypical footprint in K562 cells. (c) Aggregate per-base DNase I cleavage profile (blue line) and mean per-nucleotide conservation score (phyloP) surrounding instances of this stereotypical footprint in K562 cells (red dashed line). (d) Aggregate strand corrected CAGE sequencing data (green line) and the average nearest 59 end of a spliced EST (orange line) surrounding instances of this stereotypical footprint in K562 cells.

that, at the vast majority of promoters, RNA transcript initiation localized precisely within the stereotyped footprint. It is notable that the location of this footprint is often offset, typically from many GENCODE-annotated TSSs. This probably derives from the incomplete nature of many of the transcript ends used to define TSSs (Affymetrix ENCODE Transcriptome Project and Cold Spring Harbor Laboratory ENCODE Transcriptome Project, 2009).

These data together define a new high-resolution chromatin structural signature of transcription initiation and the interaction of the pre-initiation complex with the core promoter. Indeed, chromatin occupancy of TATA-binding protein (TBP), a critical component of the pre-initiation complex, is maximal precisely over the centre of the 50-bp footprint region (Figure

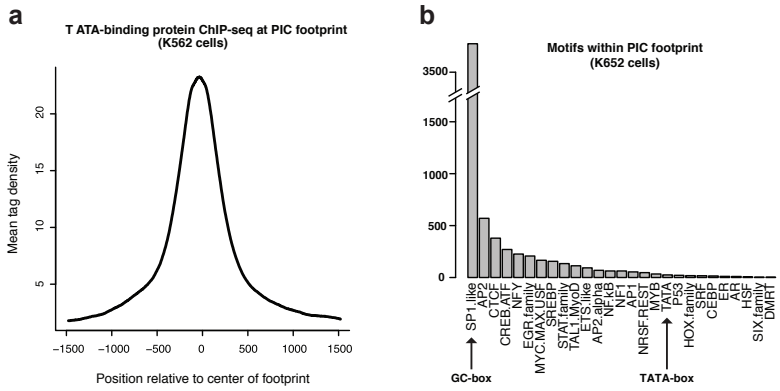

Figure 3.14: **General transcriptional activators occupy the PIC footprint**. (a) Mean ChIP-seq tag density for TATA-binding protein centered on the TSS-linked footprint in K562 cells. (b) Motifs associated with general transcription factors are found within the footprint. TRANSFAC motifs reduced by similarity and non-overlapping instances of each motif group were enumerated inside of the PIC footprint.

3.14a). Sequence analysis of the two conservation peaks within the 50-bp footprint identified motifs for GC-box-binding proteins such as SP1 and, less frequently, other general transcription factors (though with the notable absence of TATA motifs) (Figure 3.14b), indicating that TBP (and potentially other pre-initiation complex components) interacts preferentially with general transcriptional factors bound to GC-box-like features in the central footprinted region. The results are therefore consistent with a model in which a limited number of sequence-specific factors function both to prime the chromatin template for recruitment of RNA polymerase II and to guide transcriptional positioning.

3.2.6 Distinguishing indirect transcription factor occupancy

Many transcriptional regulators are posited to interact indirectly with the DNA sequence of some target sites through mechanisms such as tethering (Biddie et al., 2011). Approaches such as ChIP-seq detect chromatin occupancy, but cannot by themselves distinguish sites of direct DNA binding from non-canonical indirect binding. We therefore asked whether DNase I footprint data could illuminate ChIP-seq-derived occupancy profiles by differentiating directly bound factors from indirect binding events. We first partitioned ChIP-seq peaks from each of 38 ENCODE transcription factors (ENCODE Project Consortium et al., 2012) mapped in K562

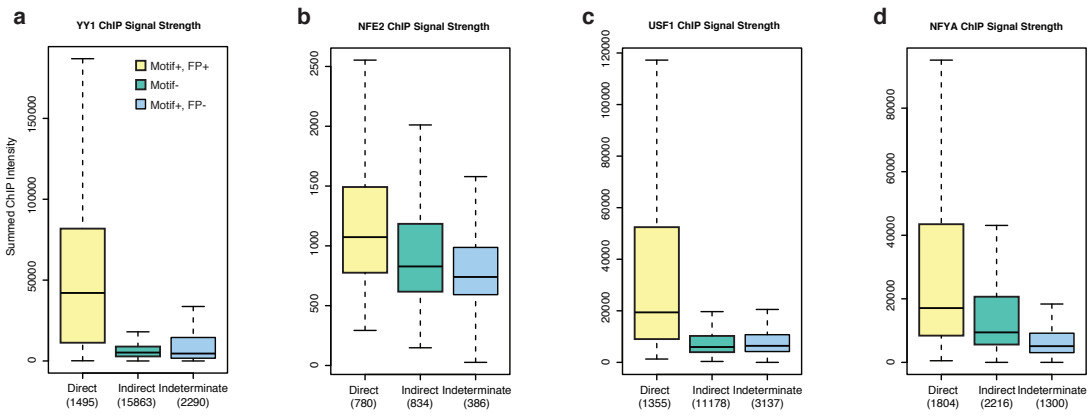

Figure 3.15: **Occupancy of transcription factors differs by mode of interaction with chromatin**. (a–d) ChIP-seq peaks of the factors YY1 (a), NFE2 (b), USF1 (c), and NFYA(d) were partitioned into three classes: direct (footprinted motif; yellow), indirect (no motif; green) and indeterminate (motif with no footprint; blue). The sum of the raw sequencing tags is displayed for each instance of a ChIP-seq peak in each partition. The number of ChIP-seq peaks contributing to each partition is displayed below.

cells into three categories of predicted sites: ChIP-seq peaks containing a compatible footprinted motif (directly bound sites); ChIP-seq peaks lacking a compatible motif or footprint (indirectly bound sites); and ChIP-seq peaks overlying a compatible motif lacking a footprint (indeterminate sites). Predicted indirect sites showed significantly reduced ChIP-seq signal compared with predicted directly bound sites (Figure 3.15a–d), consistent with lack of direct crosslink- ing to DNA (and therefore reduced ChIP efficiency). Indeterminate sites exhibited low ChIP-seq signal and were therefore excluded from further analysis (Figure 3.15a–d).

The fraction of ChIP-seq peaks predicted to represent direct versus indirect binding varied widely between different factors, ranging from nearly complete direct sequence-specific binding (for example, CTCF), to nearly complete indirect binding (for example, TBP; Figure 3.16a). In many cases factors that preferentially engage in direct DNA binding at distal sites show predominantly indirect occupancy in promoter regions and vice versa (Figure 3.16b–c).

Next, we analysed the frequency with which indirectly bound sites of one transcription factor coincided with directly bound sites of a second factor, indicative of protein–protein interactions (for example, tethering). This analysis recovered many known protein–protein interactions, such as CTCF–YY1 and TAL1–GATA1 (Wadman, 1997), as well as many novel associa-

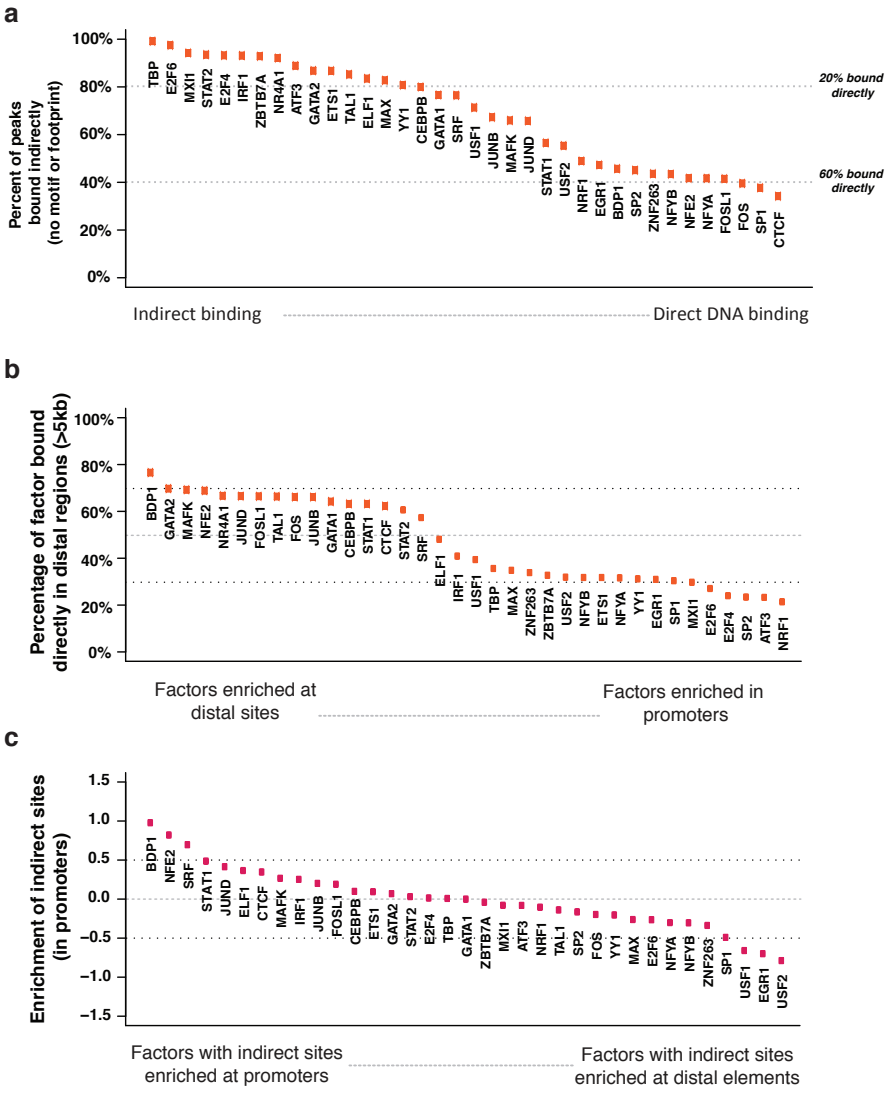

Figure 3.16: **Distribution of indirect binding by transcription factor**. (a) Transcription factors are ordered by the percentages of total peaks bound indirectly (bottom). ChIP-seq peaks are ordered by intensity and binned into groups of 500 peaks (x-axis). The fraction of ChIP-seq peaks containing a discovered motif (y-axis) is plotted. Red and green lines represent the known binding motif, except for TATA-binding protein, for which a TATA-box was not identified. The dotted horizontal line on the bottom plot represents 20% and 60% direct binding (80% and 40% indirect, respectively). (b) The percentage of K562 ChIP-seq peaks bound directly in distal regions was computed for each factor. We define distal as sites greater than 5 kb from any GENCODE level 1 and 2 annotated promoter. (b) Enrichment of indirect ChIP-seq peaks found in promoters. The enrichment is defined as the log_2 ration between the fraction of indirect sites in promoters and distal regions.

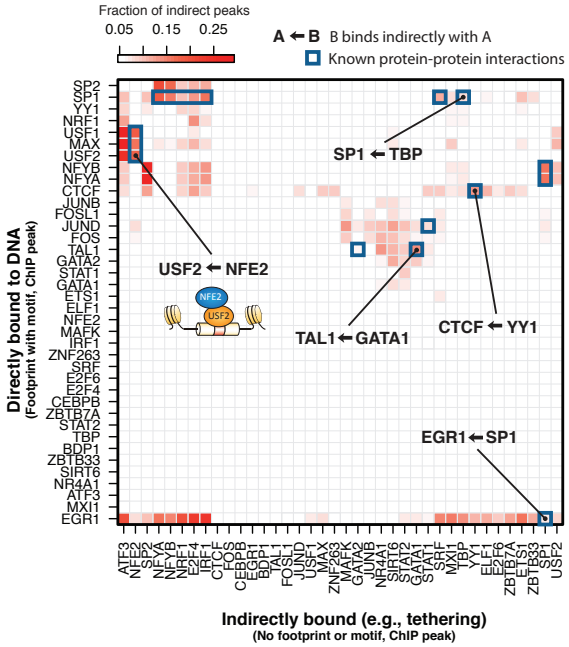

Figure 3.17: **Distinguishing direct and indirect binding of transcription factors**. Heat map of the enrichment of pairs of transcription factors in a direct–indirect association. Direct peaks are defined by ChIP occupancy accompanied by a footprint overlapping a compatible motif. Indirect peaks do not have a compatible motif. The colour of each cell is determined by the fraction of indirect peaks that co-localize with the direct peaks of another factor.

tions (Figure 3.17). We observed enrichment for NFE2 indirect interactions at promoter-bound USF2 sites, compatible with their known interaction (Zhou et al., 2010). At distal sites, we observed the opposite, with NFE2 predominantly directly bound accompanied by USF2 indirect peaks (Figure 3.16b–c), indicating the possibility of a reciprocal or looping mechanism. Notably, directly bound promoter-predominant transcription factors were enriched for co-localization with indirect peaks compared to distal regions (Figure 3.18a–b). These results suggest that combining DNase I footprinting with ChIP-seq has the potential to expose a previously unappreciated landscape of complex transcription factor occupancy modes.

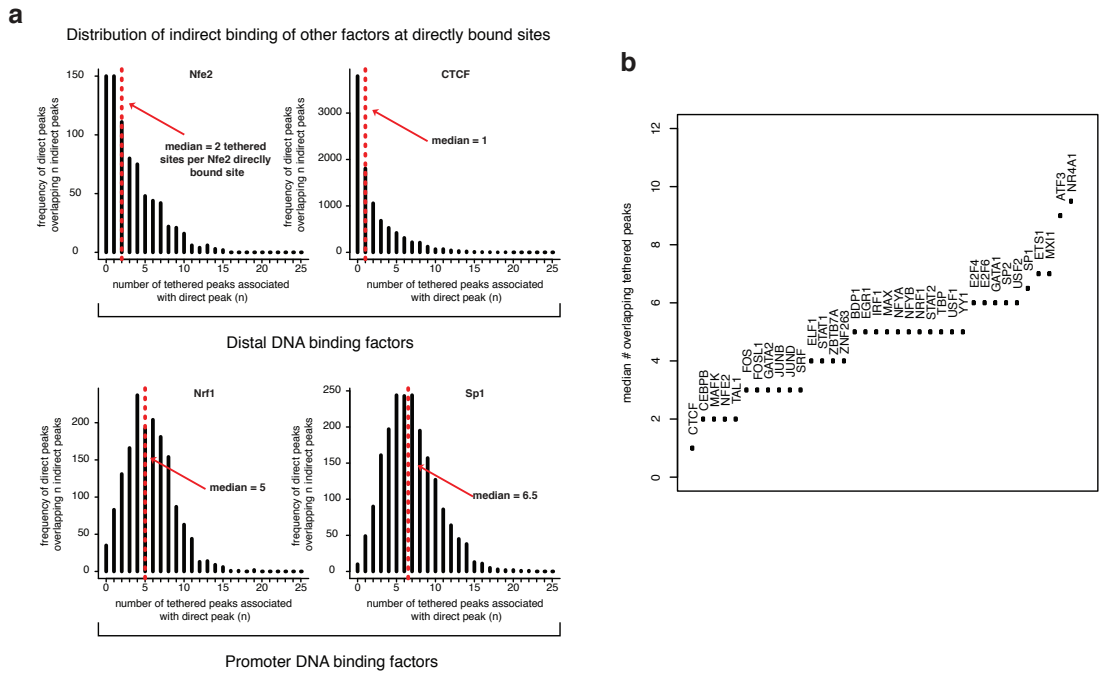

Figure 3.18: **Directly bound promoter elements mediate indirect transcription factor interactions**. (a) The number of overlapping indirect ChIP-seq peaks of other factors was computed for each directly bound ChIP-seq peak and represented as a histogram. On average, directly bound NFE2 ChIP-seq peaks overlap two indirect peaks of other factors, while Sp1 overlaps on average 6.5 indirect peaks. (b) The median value of overlapping indirect peaks at directly bound sites was computed for many factors.

3.2.7 Footprints encode an expansive *cis*-regulatory lexicon

Since the discovery of the first sequence-specific transcription factor (Gilbert and Müller-Hill, 1966), considerable effort has been devoted to identifying the cognate recognition sequences of DNA-binding proteins (Badis et al., 2009; Mukherjee et al., 2004). Despite these efforts, high-quality motifs are available for only a minority of the 1,400 human transcription factors with predicted sequence-specific DNA binding domains (Vaquerizas et al., 2009).

We reasoned that the genomic sequence compartment defined by DNase I footprints in a given cell type ideally should contain much, if not all, of the factor recognition sequence information relevant for that cell type. Consequently, applying *de novo* motif discovery to the

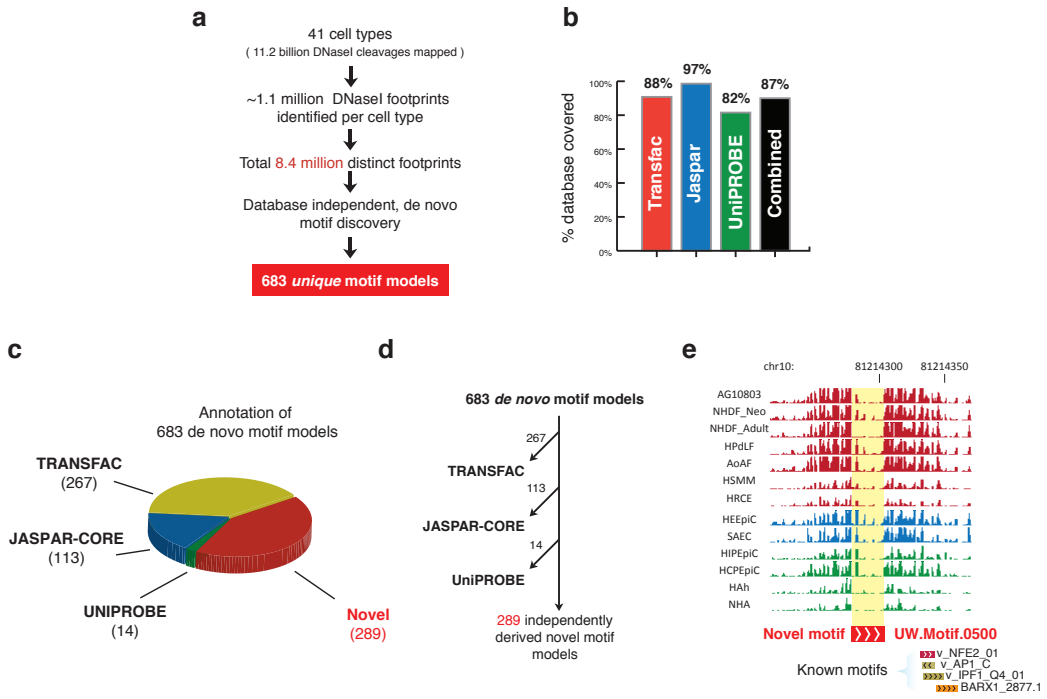

Figure 3.19: ***De novo* motif discovery expands the human regulatory lexicon**. (a) Overview of *de novo* motif discovery using DNase I footprints. (b) Annotation of the 683 *de novo*-derived motif models using previously identified transcription factor motifs. A total of 394 of these *de novo*-derived motifs match a motif annotated within the TRANSFAC, JASPAR or UniPROBE databases. (c) Pie chart annotating the partition of *de novo* motifs into known and novel motifs. (d) Diagram of the depletion scheme used to identify novel motifs. 683 motifs were filtered in successive order using TOMTOM with TRANSFAC, JASPAR-CORE and UniPROBE. The numbers on the arrows display the number of *de novo* motifs matched to the corresponding database. (e) Example of a DNase I footprint found in multiple cell types that is annotated solely by one of the novel *de novo*-derived motifs.

footprint compartments gleaned from multiple cell types should greatly expand our current knowledge of biologically active transcription factor binding motifs.

We performed unbiased *de novo* motif discovery within the footprints identified in each of the 41 cell types that yielded 683 unique motif models (Figure 3.19a and Methods). We compared these models with the universe of experimentally grounded motif models in the TRANSFAC (Matys et al., 2006), JASPAR (Bryne et al., 2008) and UniPROBE (Newburger and Bulyk, 2009) databases. Owing to the redundancy of motif models contained within these databases, we first

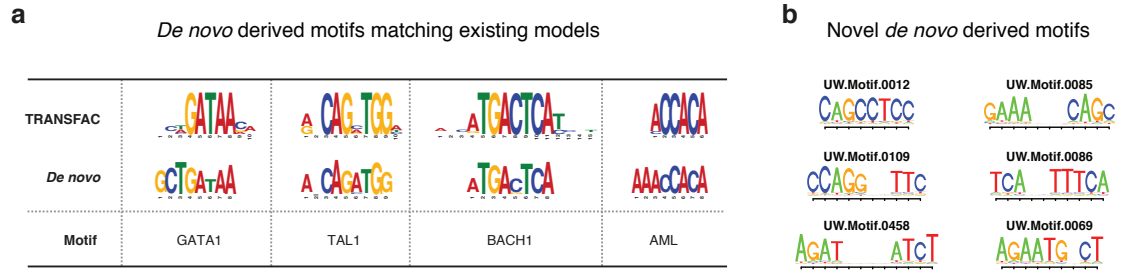

Figure 3.20: **Example of motif models derived from DNase I footprints**. (a) Example consensus logos of *de novo* derived motifs that match TRANSFAC models. (b) Example consensus logos of novel *de novo* derived motifs using DNaseI footprints.

collapsed all duplicate models (Methods). A total of 394 of the 683 (58%) *de novo* motifs matched distinct experimentally grounded motif models, accounting collectively for 90% of all unique entries across the three databases (Figure 3.19b and Figure 3.20a–b). The wholesale *de novo* derivation of the vast majority of known regulatory factor recognition sequences from the small genomic compartment defined by DNase I footprints highlights the marked concentration of regulatory information encoded within this sequence space.

Notably, 289 of the footprint-derived motifs were absent from major databases (Figure 3.19c–d and Figure 3.20b). These novel motifs populate millions of DNase I footprints (Figure 3.19e), and show features of in vivo occupancy and evolutionary constraint similar to motifs for known regulators, including marked anti-correlation with nucleotide-level vertebrate conservation (Figure 3.12c–e).

To test whether novel motifs were functionally conserved in an evolutionarily distant mammal, we analysed DNase I cleavage patterns around human novel motifs mapped within DHSs assayed in primary mouse liver tissue (Figure 3.21). This analysis demonstrated that many novel motifs show nearly identical DNase I footprint patterns in both human cells and mouse liver, indicating that these novel motifs correspond to evolutionarily conserved transcriptional regulators that are functional in both mouse and human.

Given the conservation of protein occupancy in a distant mammal, we assessed whether the novel motifs are under selection in human populations by analysing nucleotide diversity across

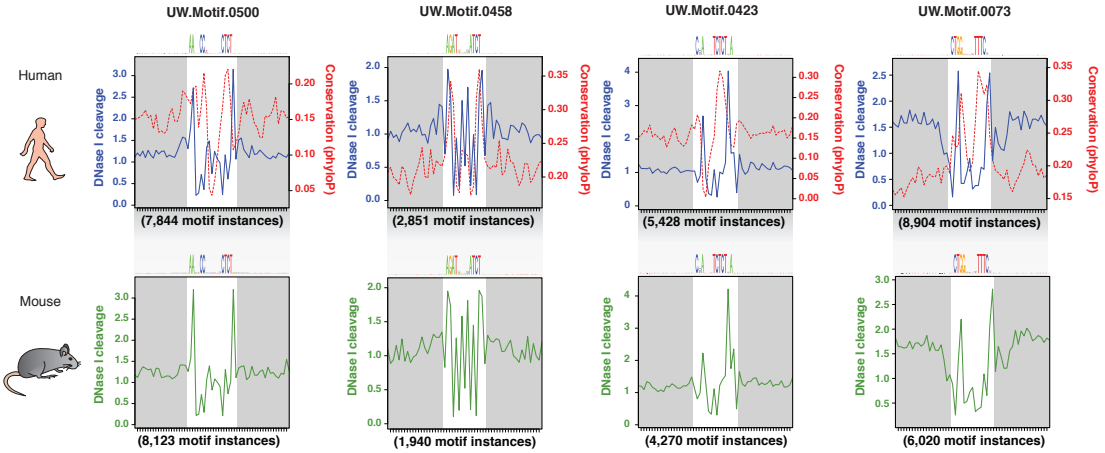

Figure 3.21: **Effects of data quality and TF occupancy on digital genomic footprinting**. (a) Density histograms of the fragment lengths in DNase I libraries from He et al. and Vierstra *et al.* Arrows indicate the means and the red arrows highly the 10.4 bp periodicity. (b) Signal portion of tags (SPOT) scores from ENCODE digital genomic footprinting datasets.

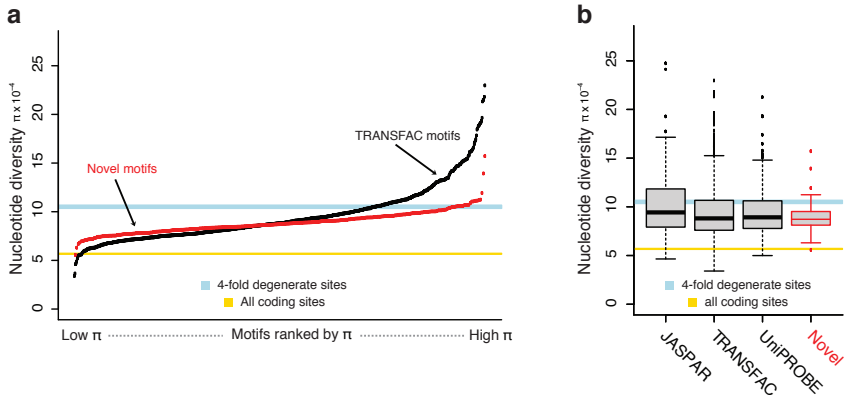

Figure 3.22: **Novel motifs are under significant selection within human populations**. (a) The average human nucleotide diversity (π, y-axis) across all motif instances within DNase I footprints is plotted for each of the motif models in the TRANSFAC database (black, ordered by mean π) and for each of the novel *de novo*-derived motif models (red, ordered by mean π). (b) Box-and-whisker plot comparing the average nucleotide diversity at instances of the 289 novel *de novo*-derived motif models to instances of motifs present in databases of known specificities (x axis). The box defines the 25% and 75% percentiles and the whiskers display 1.5 times the inner quartile range of the distribution of π values in each respective database. The blue bar indicates the average nucleotide diversity (π) at fourfold degenerate coding sites (width is equal to 95% confidence interval); gold bar indicates π at all coding sites (width is equal to 95% confidence interval).

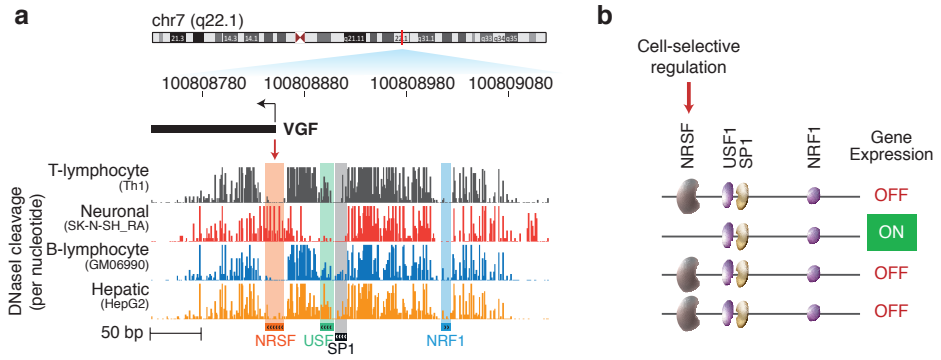

Figure 3.23: **Comparative DNase I footprints reveals *cis*-regulatory logic**. Comparative footprinting of the nerve growth factor gene (VGF) promoter in multiple cell types reveals both conserved (NRF1, USF1 and SP1) and cell-selective (NRSF) DNase I footprints.

all motif instances found within accessible chromatin. Using high-quality genomic sequence data from 53 unrelated individuals (Drmanac et al., 2010) (Table 3.4), we calculated the average nucleotide diversity (Nei and Li, 1979) for each individual motif space (Figure 3.22a). Reduced diversity levels are indicative of functional constraint, through the elimination of deleterious alleles from the population by natural selection. We found that novel motifs are collectively under strong purifying selection in human populations. On average, the new motifs are more constrained than most motifs found in the major databases (Figure 3.22a–b), even after exclusion of motifs containing highly mutable CpG dinucleotides, which underlie the marked increase in nucleotide diversity seen with a subset of known motifs (Figure 3.22a, right). Collectively, these results demonstrate that DNase I footprints encode an expansive *cis*-regulatory lexicon encompassing both known transcription factor recognition sequences and novel motifs that are functionally conserved in mouse and bear strong signatures of ongoing selection in humans.

3.2.8 Novel motif occupancy parallels regulators of cell fate

Cell-selective gene regulation is mediated by the differential occupancy of transcriptional regulatory factors at their cognate *cis*-acting elements. For example, the nerve growth factor gene

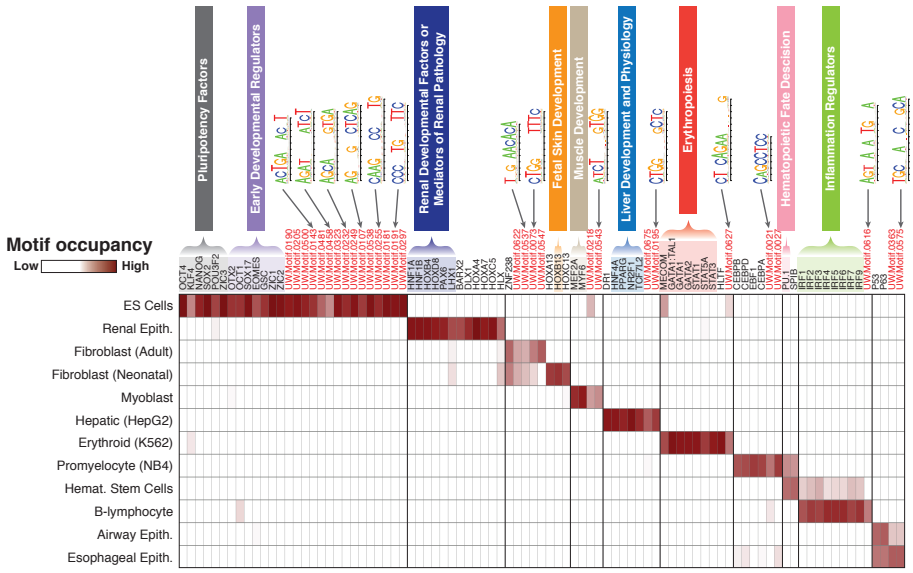

Figure 3.24: **Multi-lineage DNase I footprinting reveals cell-selective gene regulators**. heat map of footprint occupancy computed across 12 cell types (columns) for 89 motifs (rows), including well-characterized cell/tissue-selective regulators, and novel *de novo*-derived motifs (red text). The motif models for some of these novel *de novo*-derived motifs are indicated next to the heat map. ES, embryonic stem.

VGF is selectively expressed only within neuronal cells (Figure 3.23), presumably due to the repressive action of the transcriptional regulator NRSF (also called REST) at the VGF promoter in non-neuronal cell types (Schoenherr and Anderson, 1995). Although VGF is expressed only in neuronal cells, its promoter is DNase I- hypersensitive in most cell types (not shown). Examination of nucleotide-level cleavage patterns within the VGF promoter exposes its fundamental *cis*-regulatory logic, coordinated by the transcriptional regulators NRSF, SP1, USF1 and NRF1. Whereas the NRSF motif is tightly occupied in non-neuronal cells, in neuronal cells, NRSF repression is relieved, and recognition sites for the positive regulators USF1 and SP1 become highly occupied, resulting in VGF expression. These data collectively illustrate the power of genomic footprinting to resolve differential occupancy of multiple regulatory factors in parallel at nucleotide resolution.

We next extended this paradigm using genome-wide DNase I footprints across 12 functionally distinct cell types to identify both known and novel factors showing highly cell-specific

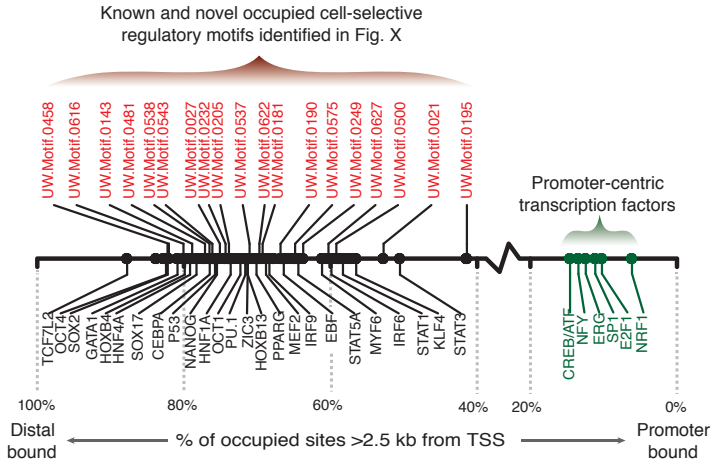

Figure 3.25: **Novel motifs are enriched distally to promoters**. The proportion of motif instances in DNase I footprints within distal regulatory regions for known (black) and novel (red) cell-type-specific regulators in Figure 3.24 is indicated. Also noted are these values for a small set of known promoter-proximal regulators (green). ES, embryonic stem.

occupancy patterns. To calculate the footprint occupancy of a motif, we enumerated for each motif and cell type the number of motif instances encompassed within DNase I footprints and normalized this by the total number of DNase I footprints in that cell type. Figure 3.24 shows a heat-map representation of cell-selective occupancy at motifs for 60 known transcriptional regulators and for 29 novel motifs. This approach appropriately identified a number of known cell-selective transcrip- tional regulators including: (1) the pluripotency factors OCT4 (also called POU5F1), SOX2, KLF4 and NANOG in human embryonic stem cells (Takahashi et al., 2007); (2) the myogenic factors MEF2A and MYF6 in skeletal myocytes (Yun and Wold, 1996); and (3) the erythrogenic regulators GATA1, STAT1 and STAT5A in erythroid cells (Halupa et al., 2005; Pevny et al., 1991; Socolovsky et al., 2001) (Figure 3.24).

Many of the footprint-derived novel motifs displayed markedly cell- selective occupancy patterns highly similar with the aforementioned well-established regulators. This suggests that many novel motifs correspond to recognition sequences for important but uncharacterized regulators of fundamental biological processes. Notably, both known and novel motifs with high cell-selective occupancy predominantly localized to distal regulatory regions (Figure 3.25, further highlighting the role of distal regulation in developmental and cell-selective processes

(Grosveld et al., 1987; Treisman and Maniatis, 1985).

3.2.9 Discussion

We describe an expansive map of regulatory factor occupancy at millions of precisely demarcated sequence elements across the human genome revealed by genomic DNase I footprinting applied to a wide spectrum of cell types. These elements collectively define a highly information-rich genomic sequence compartment that encodes the recognition landscape of hundreds of DNA-binding proteins. This compartment has been extensively shaped by evolutionary forces to match closely the physical properties of its cognate interacting proteins. Mining footprint sequences for recognition motifs has nearly doubled the human *cis*-regulatory lexicon, exposing a previously hidden trove of elements with evolutionary, structural and functional profiles that parallel the collections of experimentally derived genomic regulators brought to light during the past 30 years. Because the ability to resolve footprints is dependent on sequencing depth, and the sequencing level of DNase I cleavage events in most DHSs is not saturating (even in cell types with >500 million mapped unique DNase I cleavages), the present study, although extensive in many respects, represents only an initial foray into this biologically rich space. Identification of the cognate DNA-binding proteins for novel recognition sequences presents a significant challenge, although one that can be addressed with confidence using emerging technolo- gies and our extensive experimental data demonstrating both occu- pancy in vivo and strong evolutionary signatures of function. On a broader level, the approach that we describe here can, in principle, be applied to derive the *cis*-regulatory lexicon of any organism. We anticipate that the extensive new resources we describe, particularly in combination with other ENCODE data, will help to advance many aspects of human gene regulation research.

3.3 Methods

3.3.1 Digital genomic footprinting

DNase I digestion and high-throughput sequencing were performed on intact human nuclei from various cell types, following published methods (Hesselberth et al., 2009; Sabo et al., 2006). Briefly, roughly 10 million cells were grown in appropriate culture media and nuclei were extracted using NP-40 in an isotonic buffer. The NP-40 detergent (also known as IGEPAL-CA630) was removed and the nuclei were incubated for 3 minutes at 37°C with limiting concentrations of the DNA endonuclease, DNase I (Sigma) supplemented with Ca^{2+} and Mg^{2+}. The digestion was stopped with EDTA and the samples were treated with proteinase K. The small 'double-hit' fragments (<500 bp) were recovered by sucrose ultra-centrifugation, end-repaired and ligated with adapters compatible with the Illumina sequencing platform.

The following human cell types were subjected to DNase I digestion at the 36mer or 27mer level: AG10803, AoAF, CD20+, CD34+ Mobilized, fBrain, fHeart, fLung, GM06990, GM12865, HAEpiC, HA-h, HCF, HCM, HCPEpiC, HEEpiC, HepG2*, H7-hESC, HFF, HIPEpiC, HMF, HMVEC-dBl-Ad, HMVEC-dBl-Neo, HMVEC-dLy-Neo, HMVEC-LLy, HPAF, HPdLF, HPF, HRCEpiC, HSMM, Th1, HVMF, IMR90, K562, NB4, NH-A, NHDF-Ad, NHDF-neo, NHLF, SAEC, SKMC, and SK-N-SH RA.

Tags were aligned to the reference genome, build GRCh37/hg19 (specified by ENCODE `http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/encodeDCC/referenceSequences`) using bowtie (Langmead et al., 2009), version 0.12.7 with parameters: `--mm -n 3 -v 3 -k 2`, and `--phred33-quals` for Illumina HiSeq sequencer runs or `--phred64-quals` for Illumina GAII sequencer runs.

3.3.2 Identification of DNase I footprints

For each cell type, we computed the DNase I cleavage per nucleotide by assigning to each base of the human genome an integer score equal to the number of uniquely mappable sequence tags with 5' ends mapping to that position. To identify DNase I footprints comprehensively across the genome, we used an improved and conceptually simplified approach versus that applied previously to the yeast genome (Hesselberth et al., 2009). We focused on high cleavage density regions, hotspot regions as identified by the hotspot algorithm (John et al., 2011), within each cell type. We scanned the genome for 6-40 nt stretches of successive nucleotides with low

DNase I cleavage rates relative to the immediately flanking regions, the signature of localized protection from DNase I cleavage (Galas and Schmitz, 1978; Hesselberth et al., 2009). We then filtered findings to those occurring within the hotspot regions.

A priori, footprints comprise three components: a central area of direct factor engagement, and an immediately flanking component to each side. Upon factor engagement, local DNA architecture is distorted, frequently resulting in enhanced cleavage rates for flanking nucleotides outside of the factor recognition sequence. Greater disparity between the central and flanking components is indicative of higher factor occupancy.

To quantify this, we applied a simple Footprint Occupancy Statistic (FOS) such that

$$FOS = \frac{C+1}{L} + \frac{C+1}{R} \tag{3.1}$$

where C represented the average number of tags in the central component, L represented the average number of tags in the left flanking component, R represented the average number of tags in the right flanking component, and a smaller FOS value indicated greater average contrast levels between the central component and its flanking regions.

We sought to optimize the statistic across a range of central component (6-40 nt) and flanking component (3-10 nt) sizes. The output of the algorithm was the set of footprints with optimal FOS scores, subject to the criteria that L and R were greater than C, and all central components were disjoint and non-adjoining. When two or more potential footprints (those with L and R greater than C) had overlapping or abutting central components, we selected the one with the lowest FOS (or, in rare cases of identical scores, the 5'-most footprint relative to the forward strand). We then rescanned the entire local region to identify additional footprints. A local region was defined as the smallest genomic segment to contain all potential footprints of shared bases (by transitivity). No newly identified footprint consisted of a central component that overlapped or abutted the central component of any previously selected footprint. The rescan process was iterated until no new footprint was identified within the local region.

Human genomic positions uniquely mappable using 36 nt (and 27 nt as appropriate) sequence reads were computed using the same algorithm previously applied to yeast (Hesselberth et al., 2009). Any computed footprint whose central component consisted of non-uniquely mappable bases (thus having no mapped cleavage events by definition) that covered at least 20% of its length was discarded. Typically, fewer than 1% of unthresholded footprints were discarded during this process.

Due to the large number of tests for footprints performed over the genome, it was necessary

to control for the expected number of false positives that arose due to chance through multiple testing (Benjamini and Hochberg, 1995). We applied a false discovery rate (FDR) measure, defined as the expected value of the fraction of truly null features called significant divided by the total number of features called significant. To estimate FDR, we first generated a null set of pseudo-cleavages. For each hotspot in one cell type, we randomly reassigned the same number of tags found within the region to uniquely mappable positions within the same genomic interval. Analogous with experimental data, each base received an in silico cleavage score equal to the number of tags with 5' ends mapped to that base. We then considered the identical footprint positions under the randomized scenario that were derived as output for the non-thresholded experimental data, thus encompassing the same number of footprint calls for FDR calculation purposes. We computed the maximum FOS threshold at which the number of footprints in the null set divided by the number of footprints in the observed set was less than or equal to 1%. The 1% FDR estimates were computed separately for all 41 cell types, covering a wide range of total tag levels and number of hotspot regions, to produce an average FOS threshold of 0.95 with a standard deviation of 0.02. We applied a final FOS threshold of 0.95 to footprints across all cell types. The central components of these FDR thresholded footprints, henceforth footprints, made up the final output of the procedure.

To combine footprints across cell-types, we computed the multiset union of all footprints across all cell types. For each element of the union, we collected all significantly overlapping footprints, which were defined as those footprints with 65% or more of their bases in common with the element. A footprint's genomic coordinates were redefined to the minimum and maximum coordinates from its overlap set, which always included the footprint itself. All redefined footprints from the union then passed through a subsumption and uniqueness filter: when a footprint was genomically contained within another, the filter discarded the smaller of the two or selected just one footprint if identical. Footprints passing through the filter comprised the final set of 8.4 million combined footprints across all cell types. Unlike footprints from any single cell type, the combined set included overlapping footprints.

3.3.3 Footprinting vs. tag levels

Random subsamples (sampling without replacement) of the 543 million uniquely mappable DNase I-seq tags from SKMC were generated. Increasing sample sizes utilized tags generated from smaller samples in addition to new tags generated from the randomized process. Footprints were called at each subsampled tag level (Figure 3.2a).

3.3.4 FDR 1% DNase I hypersensitive sites

We counted the number of footprints falling within every DNase I hypersensitive sites (DHS, defined as 150 nt in length)2 and grouped peaks by their number of footprints. Any peak containing more than 10 footprints was grouped with peaks containing exactly 10 footprints. The analysis was performed in every cell type separately, and then results were combined. We also decile-partitioned the DHSs by the number of sequencing tags mapped to them. For each partition, we drew a box plot to indicate the distribution of the number of footprints falling within the DHSs (Figure 3.2b). We also determined the average number of footprints falling in DHSs (Table 3.2).

3.3.5 Genomic annotation of footprints

We counted and summarized the number of combined footprints (8.4 million) falling into common genomic element categories (defined by at least 1 nt of overlap), such as those overlapping introns, coding elements, and intergenic regions. We utilized annotations from Gencode, version 7 (Harrow et al., 2012). Promoter regions were defined as within $\pm$2.5 kb from a transcriptional start site (TSS). Regions within $\pm$2.5 kb of transcriptional end sites were categorized as 3'-proximal. Other feature categories, such as Coding, 5'-UTR, 3'-UTR, and Introns were derived directly from Gencode annotations using transcriptional and coding start and stop site information, as well as exon boundary coordinates. When a footprint satisfied more than one category condition (for example, when a footprint was found near more than one annotated transcript), we assigned it to only a single category. The order of category assignment in such cases was: coding, 5'-UTR, 3'-UTR, promoter, 3'-proximal, intronic, and intergenic.

3.3.6 Putative motif binding sites and footprints

We determined the significance of overlap between footprints and predicted motifs within hotspot regions utilizing the Genome Structure Correction (GSC) test (ENCODE Project Consortium et al., 2012). Merged genomic hotspot regions across all 41 cell types made up the domain. The multiset union of all footprints, part of the domain by definition, as well as motif predictions within the domain (FIMO $P < 10^{-5}$ using TRANSFAC) were used as inputs to GSC. Program parameters were: `-n 10000 -s 0.1 -r 0.1 -t m`. Significance was reported as a Z-score (empirical p-value was 0).

3.3.7 Average Motif Density Per-nucleotide

We determined the average per-nucleotide number of overlapping motif instances over segments of a genome-wide partition. We separately merged the hotspot regions and footprint regions across the 41 cell types. Using genome-wide FIMO scan predictions over TRANSFAC ($P < 10^{-5}$), we counted the number of motif scan bases contained within the merged footprint partition and divided by the total number of bases within the partition. Similarly, we found the average over the genomic complement between merged hotspots and merged footprints. Finally, we found a genome-wide average outside of hotspots and divided by the number of nucleotides with known base labels (A,C,G,T), thereby ignoring large centromeric and telemeric regions.

3.3.8 DNase I cleavages vs. ChIP-seq density

Motif models from TRANSFAC (Wingender et al., 1996), version 2011.1, JASPAR Core (Bryne et al., 2008), and UniPROBE (Newburger and Bulyk, 2009) were used in conjunction with the FIMO motif scanning software (Grant et al., 2011), version 4.6.1 using a $P < 10^{-5}$ threshold, to find all motif instances within DNase I hotspots of the K562 cell line. We buffered ($\pm$35 nt) a discovered motif instance and counted at each base position the number of uniquely mapping DNase I sequencing tags with 5' ends mapping to the position. We sorted buffered motif instances by their total counts, and then normalized each instance's counts to a mean value of 0 and variance 1. A heatmap, with 1 row per motif instance, was generated using matrix2png (Pavlidis and Noble, 2003), version 1.2.1. A phyloP (Pollard et al., 2010) evolutionary conservation score heatmap over the same ordered motif instances and bases was generated using the same processing techniques. Motif instances that overlapped footprints by at least 3 nt were annotated. Uniformly processed hg19 K562 ChIP-seq peaks generated from experiments as part of the ENCODE Consortium were downloaded from the UCSC Table Browser (Rosenbloom et al., 2011). Motif instances overlapping ChIP-seq peaks by at least 1 nt were also annotated.

3.3.9 Footprint Strength vs. ChIP-seq

For a given ChIP-seq factor, we collected footprints that overlapped putative binding sites within hotspot regions by at least 3 nt. We calculated the summed ChIP-seq signal density over each region, after buffering by $\pm$50 nt from footprint centroid. Footprints were ordered by their FOS values, and signal data were plotted using lowess curve fitting with a span of 25%.

ChIP-seq data (raw tag counts) included those from first replicates only. Average tag count numbers replaced cases where multiple measurements over the same genomic coordinates existed in the ChIP-seq data.

3.3.10 Footprint strength vs. evolutionary conservation

We additionally calculated the maximum phyloP evolutionary conservation score over the same set of footprints. The maximum score was derived over the core footprint region (no buffering). As before, footprints were ordered by their FOS values, and signal data were plotted using lowess curve fitting with a span of 25% (Figure 3.7c–d).

3.3.11 DNA Interacting Protein Precipitation (DIPP) Experiments

Nuclei were isolated using a standard protocol previously described (Dorschner et al., 2004; Sabo et al., 2006). Briefly, K562 cells were grown in RPMI (GIBCO) supplemented with 10% Fetal Bovine Serum (PAA), sodium pyruvate (GIBCO), L-glutamine (GIBCO), penicillin and streptomycin (GIBCO), and washed once with Dulbecco's PBS (GIBCO). Nuclear extraction was performed by resuspending cells at 2.5×10^6 cells/mL in 0.05% NP-40 (Roche) in Buffer A (15mM Tris pH 8.0, 15mM NaCl, 60mM KCl, 1mM EDTA pH 8.0, 0.5mM EGTA pH 8.0, 0.5mM Spermidine). After an 8 minute incubation on ice, nuclei were pelleted at 400 *g* for 7 minutes and washed once with Buffer A. Nuclei were then transferred to a 37°C water bath and resuspended at 1.25×10^7 nuclei/mL in Extraction Buffer (10mM Tris pH 8.0, 600mM NaCl, 1.5mM EDTA pH 8.0, 0.5mM spermidine). After 3 minutes at 37°C the sample was transferred to ice and rocked at 4°C for 2 hours. The soluble and insoluble fractions were separated by centrifugation at 3,220*g* for 15 minutes. The soluble fraction was then dialyzed for 2 hours at 4°C using a 3,500 Da molecular weight cut off (MWCO) cartridge (Pierce) against 500mL Dialysis Buffer (15mM Tris pH 7.5, 15mM NaCl, 60mM KCl, 5μM $ZnCl_2$, 6mM $MgCl_2$, 1 mM DTT, 0.5mM Spermidine, 40% Glycerol). The dialysis buffer was refreshed after 1 hour of dialysis. Dialyzed protein samples were quantified using a BCA assay (Pierce), flash frozen using liquid nitrogen and stored at -80°C until use.

Three genomic loci were targeted that demonstrated varying footprinting strengths. These footprints included (in hg19 coordinates) a MAX footprint (`chr22:39707228-39707245`) and two AP1 footprints – AP1 site 1 footprint (`chr11:5301978-5302005`) and AP1 site 2 footprint (`chr5:75668604-75668626`). For each of these sites, a 70-85 base pair region of DNA centered

on the DNase I footprint was selected. The selected DNA regions, in hg19 coordinates, were; `chr22:39707201-39707270` for the MAX site; `chr11:5301945-5302029` for the AP1 site 1; and `chr5:75668577-75668646` for the AP1 site 2. DNA oligos were ordered for the forward and reverse strand for each of these sites, with the forward strand oligo containing a 5' biotin modification (Integrated DNA Technologies). For each of these sites, we also shuffled the footprinting sequence and ordered DNA oligos that contained this shuffled footprinting sequence along with the same flanking sequence as for the oligos above (Integrated DNA Technologies). The sequences of each of the probes can be found in Table 3.3.

Generation of dsDNA bound beads for DNA Interacting Protein Precipitation (DIPP) For each probe set, 500 picomoles of the forward strand biotinylated DNA oligo was mixed with 1 nanomoles of the reverse strand DNA oligo in Annealing Buffer (20 mM Tris pH 8.0, 100 mM KCl, 10mM $MgCl_2$). The reaction was denatured at 90°C for 5 minutes, slowly cooled to 65°C over 10 minutes, held at 65°C for 5 minutes and then cooled to 25°C. For each reaction, 100μl of Dynabeads MyOne Streptavidin T1 beads (Invitrogen) were washed twice with 0.75 mL of Bead Buffer (20 mM Tris pH 8.0, 2 M NaCl, 0.5 mM EDTA, 0.03% NP-40) and resuspended in 0.8mL Bead Buffer similar to how previously described (Mittler et al., 2009). Annealed dsDNA probes were then added to the beads and rocked at room temperature for 1 hour. Beads were then washed twice with 0.8 mL Bead Buffer to remove unbound oligos. 1 mL of Blocking Buffer (20 mM Hepes pH 7.9, 300 mM KCl, 50μg/mL bovine serum albumin (BSA), 50μg/mL glycogen, 5 mg/mL polyvinylpyrrolidone (PVP), 2.5 mM DTT, 0.02% NP-40) was added to each bead reaction and incubated at room temperature for 2 hours. Beads were then washed twice with 0.75 mL of Binding Buffer (20 mM Tris-HCl pH 7.3, 5 μM $ZnCl_2$, 100mM KCl, 0.2 mM EDTA pH 8.0, 10 mM potassium glutamate, 2 mM DTT, 0.04% NP-40, 10% glycerol).

The protein extract was pre-cleared prior to oligo binding using the following method. 60μl of fresh Dynabeads MyOne Streptavidin T1 beads (Invitrogen) were washed twice with 0.3 mL of Bead Buffer and once with 0.3 mL of Binding Buffer and then added to 80 μg of 600mM soluble K562 nuclear protein extract and 80 μg of poly [d(I-C)] (Roche) in a 400 μl total reaction volume with Binding Buffer. This reaction was incubated at 4°C for 1.5 hours, the beads were removed and the buffered protein extract was cleared by centrifugation at 10,000*g* for 8 minutes at 4°C.

To each of the washed dsDNA bound bead reactions, 200μl of the pre-cleared buffered protein extract was added. This was incubated at 4°C for 2 hours then washed 3 times with 1 mL Binding Buffer, twice with 0.5 mL 50mM Ammonium Bicarbonate pH 7.8 and resuspended

in 100μl 0.1% PPS Silent Surfactant (Protein Discovery) in 50mM Ammonium Bicarbonate pH 7.8. Bead bound proteins were boiled at 95°C for 5 minutes, reduced with 5 mM DTT at 60°C for 30 minutes and alkylated with 15 mM iodoacetic acid (IAA) at 25°C for 30 minutes in the dark. Proteins were then digested with 2μg Trypsin (Promega) at 37°C for 1.5 hours while shaking. The supernatant, which now contains digested peptides, was then transferred to a new tube, the pH was adjusted to <3.0 by 5 μl of 5 M HCl and incubated at 25°C for 20 minutes and then cleared by centrifugation at 20,817*g* for 10 minutes. The digested samples were desalted using an Oasis MCX cartridge 30mg/60μm (Waters) as previously described (Stergachis et al., 2011). Peptide samples were then resuspended in 30μl 0.1% formic acid in H_2O. These peptide samples were stored at -20°C until injected on the mass spectrometer.

Proteotypic peptides for c-Jun, MAX and CTCF were identified as previously described (Stergachis et al., 2011). These peptides were; `CPDCDMAFVTSGELVR` and `TFQCELCSYTCPR` for CTCF; `NSDLLTSPDVGLLK` and `NVTDEQEGFAEGFVR` for c-Jun; and `QNALLEQQVR` and `ATEYIQYMR` for MAX. For each doubly charged monoisotopic precursor, we monitored singly charged monoisotopic y3 to yn-1 product ions. All cysteines were monitored as carbamidomethyl cysteines. Ions were isolated in both Q1 and Q3 using 0.7 FWHM resolution. Peptide fragmentation was performed at 1.5mTorr in Q2 using calculated peptide specific collision energies (MacLean et al., 2010a). Data was acquired using a scan width of 0.002 m/z and a dwell time of 40ms.

Peptide samples were analyzed with a TSQ-Vantage triple-quadrupole instrument (Thermo) using a nanoACQUITY UPLC (Waters). A 5μl aliquot of each sample was separated on a 20cm long 75μm I.D. packed column (Polymicro Technologies) using Jupiter 4u Proteo 90A reverse-phase beads (Phenomenex) and chromatography conditions as previously described (Mittler et al., 2009). The injection order for each sample was randomized, and each sample was measured in three separate replicate injections.

Targeted measurements were imported into Skyline for analysis (MacLean et al., 2010b). Chromatographic peak intensities from all monitored product ions of a given peptide were integrated and summed to give a final peptide peak height. For each peptide, peak heights from different samples and replicate runs were normalized such that the injection with the highest intensity was given a value of 1. Final peptide data were generated by taking the average normalized value of a peptide across replicates of a sample.

3.3.12 Allelic imbalance in footprints

A set of known autosomal single nucleotide variants (SNVs) was downloaded from the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2010). To avoid positions subject to mapping bias, SNVs were filtered to exclude any two within a read length (up to 36 nt) of one another. Allele counts used the same DNase I- seq alignments from which the cut-counts were derived. For each cell type, reads overlapping each SNV were queried from the alignment in BAM format using the SAMtools (Li et al., 2009). Reads supporting a base call were counted only if they were mapped with no more than one mismatch excluding the SNV position being counted. If more than one read from a library was mapped at the same chromosome offset and strand, a single read was sampled at random to avoid overcounting from possible PCR duplicates. In order to call an individual heterozygous at a SNV conservatively, both alleles observed by 1000 Genomes had to be supported by at least four distinct reads. To call homozygotes conservatively, one of the known alleles had to be supported by at least 10 reads, and there had to be no reads supporting the other known allele, but a single read supporting another base was tolerated as a sequencing error where total read depth exceeded 50.

In the vicinity of each SNV (36 nt), DNase I cut-counts from individuals homozygous for the same allele were added together, using the same genomic cut-count tracks used for calling footprints. In heterozygous individuals, reads overlapping the SNV were queried from the alignment BAM files but not subjected to the mismatch and duplicate filters used to obtain unbiased counts. The cut position represented by each read was reported as the aligned genomic position of the first base of the read, so cut-counts from reads aligning to the negative genomic strand may be offset by 1 nt, relative to the convention normally used for genomic cut counts. For each allele, the phased cut-counts for that allele from all heterozygous individuals were then added together.

At each SNV, the reads supporting each allele from all individuals heterozygous at the SNV were added together. Heterozygous sites were divided into two sets, those within the merged FDR 1% footprints across all cell types and those outside. A read-depth distribution was derived from each set, and the intersection was determined to generate a read-depth-matched random sample as large as possible. At each particular read depth, all sites from the set with fewer instances of that depth were included, and a random sample without replacement was taken from the set with more instances. Finally, we counted sites in each set showing allelic imbalance with two-sided binomial test $P < 0.01$. The difference between these counts was

tested for significance with a one-sided Fisher's exact test.

3.3.13 CpG methylation calculation within footprints, DHSs, and non-DHSs

IMR90 methylation calls (Lister et al., 2009) were filtered to CpGs covered by at least 40 reads. Methylation at each CpG is defined as the count of reads showing methylation (protection from bisulfite conversion) divided by the total read depth. We generated three sets of genomic coordinates with this signal: IMR90 FDR 1% footprints, IMR90 DNase I peaks (subtracting overlapping footprint bases), and locations of CpGs in the GRCh37/hg19 genome reference sequence, removing elements that overlap IMR90 DNase I hotspots. For each contiguous region in these datasets, we took the mean methylation of all overlapping CpGs that passed the 40-read coverage threshold. Regions with no such overlap were ignored. To compute p-values, vectors of mean methylation values were compared using a two-sided Mann-Whitney test.

3.3.14 Rendering of DNA-protein complexes

Crystallography data showing DNA-protein complexes for selected factors were obtained from the Protein Data Bank (Bernstein et al., 1977; Ferré-D'Amaré et al., 1994; Párraga et al., 1998) and rendered with MacPyMOL (version 1.3) (http://www.pymol.org). Nucleotide residues were colored from white to blue, indicating increasing relative DNase I cleavage propensity as aggregated across all motif instances.

3.3.15 Visualization of DNase I cleavage profiles by motif occurrence

Motif models (from TRANSFAC, JASPAR Core, and UniPROBE) were used in conjunction with the FIMO motif scanning software, version 4.6.1 using a p-value < 1^{-5} threshold, to find all motif instances within DNase I hotspots of each cell type. The left and right coordinates of each motif instance were padded by 35 nt. Using the bedmap tool from the BEDOPS suite (Neph et al., 2012a), version 1.2, the per-nucleotide DNase I cleavage values from deeply sequenced DNase I-seq libraries were recovered for each motif occurrence. A similar approach was used for phyloP vertebrate conservation. Aggregate plots were made by averaging over all strand-oriented motif occurrences the number of DNase I cleavages and per-base conservation scores. Palindromic motif occurrences were left in the dataset, reasoning that a transcription factor may bind to either orientation of the genomic region and binding events on either strand result in conformal changes to DNA that result in strand-specific cleavage patterns. Sequence logos

were generated by assessing the information content of the oriented genomic sequences from all motif occurrences (Crooks et al., 2004).

3.3.16 Analysis of PIC footprint

The cleavage profiles ±500 nt of all GENCODE V7 (level 1 and 2; manual curation) (Harrow et al., 2012) transcription start sites were used as regions to search for a 35-55 base-pair footprint following the method outline above with modifications. To amplify the signal in regions of low tag density and to remove noise in the data, the DNase I cutcounts were squared (x^2). The FOS score was then calculated for every segment 35-55 base-pairs in width using a fixed flank width of 10 base-pairs (left and right). The scored segments were ranked in ascending order (low FOS to high FOS) and the top non-overlapping segments were collected until no segments remained. Finally, a FOS threshold was selected (0.75, uniformly across 41 cell types) and these putative footprints were used in the subsequent analysis.

Graphical profiles were generated by enumerating the per-nucleotide DNase I cleavages and phyloP conservation in a 250 base-pair window centered on the footprint. CAGE tags from the nuclear poly-A fraction (replicate 1) generated by RIKEN was downloaded from the UCSC Browser and the 5' stranded oriented ends were summed per-base. The footprint was stranded oriented to the nearest GENCODE V7 TSS (Djebali et al., 2012). We enumerated the per-base CAGE tags in an 800 base-pair window centered on the footprint. To evaluate the spatial relationship of transcription we calculated the distance to the nearest spliced EST curated from GenBank (Pruitt et al., 2009) (Figure 3.13d).

3.3.17 Determining direct and indirect transcription factor binding

Uniformly processed hg19 K562 ChIP-seq peaks generated from experiments as part of the ENCODE Consortium were downloaded from the UCSC Genome Browser. Peaks overlapping DNase I hypersensitive hotspot regions2 by at least 20% were stratified into three categories: direct peaks, indirect peaks and indeterminate peaks. Direct peaks contained an appropriate motif instance (FIMO scan software13, version 4.6.1, using a $P < 10^{-5}$ threshold and motifs from TRANSFAC, version 2011.1) that overlapped a DNase I footprint by at least 1 nt. Indirect peaks did not contain a cognate motif and indeterminate peaks were ambiguous (contained a motif which did not overlap a footprint). To identify enriched direct/indirect binding pairs, we counted the number of overlapping occurrences of all possible direct/indirect combinations.

We normalized each ChIP-seq peak-pair count by the total number of indirect peaks for the indirectly bound factor, in order to reduce the effect of noise (due to incomplete motif models, insufficient DNase I coverage, and/or non-specific antibodies).

3.3.18 *De novo* motif discovery

We created different footprint subsets for each cell type for the purpose of *de novo* motif discovery. A proximal subset was defined as all footprints within 2000 nt of the canonical transcriptional start site of genes (Pruitt et al., 2009), a non-proximal set was defined as all footprints not in the proximal subset, a distal set was defined as all footprints more than 10,000 nt from any transcriptional start site, and cell-type-specific footprints were those footprints found within cell-type-specific DHSs. Cell-type-specific DHSs and constituent footprints were those found in only a single cell type.

We developed an exhaustive motif discovery procedure for inputs consisting of millions of genomic regions. To accomplish the exhaustive search, several simple heuristic filtering and clustering techniques were employed, along with a compute cluster. *De novo* motif discovery was performed separately for every cell type and on every footprint subset. For each subset, we symmetrically padded the central components of footprints by 4 nt and extracted genomic sequence information to create target regions for *de novo* discovery. We counted the number of target regions within which each subsequence pattern occurred, separately considering every 8 nt permutation over the 4-letter DNA nucleotide alphabet, with up to 8 intervening IUPAC `N` degenerate symbols. For background estimates, nucleotide labels within every target region were randomly shuffled, thereby maintaining local nucleotide label compositions. The number of regions within which each pattern existed was determined after each of 1000 shuffling operations in order to establish sample mean and variance values for expectation. These estimates for patterns further served as conservative estimates for longer patterns in the background case. For example, the estimates for `cgttacc` also served as estimates for the `acgNttacc` pattern. A Z-score was computed for each observed subsequence pattern by subtracting the mean background frequency estimate from the observed frequency and then dividing by the estimated standard deviation. Patterns with Z-score of at least 14 were listed in descending Z-score order and then further filtered and clustered to remove redundant motifs. Initially, the highest Z-score pattern was added to an output list, and each subsequent pattern was compared to every entry in the list. If a similar entry was found, the pattern was discarded; otherwise, the pattern was added to the bottom of the output list. Pattern similarities were determined by sequen-

tially comparing characters. When two patterns were the same length and their N placeholders aligned, they were considered similar if they had one character difference; otherwise, they were declared similar if they had up to two character differences. The reverse character sequence of every pattern then underwent the same filtering. The re-tuned motif list underwent a similar second stage filter that included all alignment possibilities and reverse complement combinations. Sequence patterns were converted to positional weight matrices (PWMs) by scanning all target sequences and normalizing over the nucleotide alphabet. Only exact matches to a subsequence pattern, ignoring all N placeholders, were considered during PWM construction, which underwent further filtering. The PWM corresponding to the highest Z-score pattern was added to an output list and a comparison list. PWMs for subsequent patterns, still in descending Z-score order, were compared to every entry in the comparison list and then added to the bottom of that list. If no similar entry was found, the PWM was also added to the output list. During comparisons, Pearson correlation coefficients were calculated over all alignment possibilities and reverse complement combinations. We converted PWMs into 1-dimensional vector representations. Vectors were temporarily padded using samples from the genome-wide background nucleotide frequency distribution and renormalized for various alignments as needed. If a correlation value of at least 0.75 was found, two PWMs were considered similar. We reverted PWMs to their subsequence pattern forms and rescanned target regions, allowing up to one nucleotide mismatch from the pattern's subsequence representation. PWM filtering comparisons were performed as before, and PWM outputs from this stage formed the output.

The *de novo* discovery results for all footprint subsets and cell types were combined, clustered, and filtered further into a final set of 683 motifs. The PWM representations were converted to their subsequence pattern forms and combined in descending Z-score order. The first pattern was added to the output list. Each subsequent pattern was compared to every entry of the output list. If no similar entry was found, the pattern was added to the bottom of the list. Pattern comparisons included all alignment possibilities and reverse complement combinations. For a given alignment, the patterns were compared sequentially, character by character. They were considered similar if they had one character difference; otherwise, they were declared similar if they had up to two character differences.

For the final stage of clustering, we determined the proportion of instances of one pattern that genomically overlapped instances from another pattern. All pairwise combinations between patterns were considered. We scanned twice for every patterns instances. The first scan included only those instances that do not deviate from their motif pattern. The second

included all instances that have up to one mismatch. Scanning occurred over all padded footprints, merged across all cell types. If the proportion of overlapping instances between two patterns was 0.1 or more in the first case and 0.33 in the second case, in either motif comparison direction, we discarded the pattern of lower Z-score. We considered all cases with any amount of overlap (at least 1 nt). For example, if two patterns instances overlapped at one part of the genome by 5 nt, and two more instances overlapped in another part of the genome by 2 nt, we conservatively counted both cases toward the proportion of overlaps (in contrast to the potential requirement of counting overlapping proportions at fixed offsets between instances). All patterns passing through this step made up the set of final motif models.

3.3.19 Motif matching

We compared *de novo* motifs to motifs available as part of various databases, including TRANSFAC, version 2011.1, JASPAR Core, and UniPROBE using the TOMTOM software (Gupta et al., 2007), version 4.6.1. We filtered TRANSFAC and JASPAR Core for motifs annotated to the human genome, and mouse motifs in UniPROBE. TOMTOM parameters were set to their default values. When partitioning the *de novo* motifs, assigning each to a single category, the order of match assignment preference was to TRANSFAC, JASPAR Core, UniPROBE, and then to the novel motif category (Figure 3.15).

3.3.20 Mouse scans of novel human motifs

Novel *de novo* motifs (those with no motif match to entries of the TRANSFAC, JASPAR Core, and UniPROBE databases) were scanned across DNase I hotspot regions of the mouse genome (build NCBI37/mm9) using FIMO at $P < 10^{-5}$. Average cleavage profiles were generated and compared to analogous profiles of the human genome.

3.3.21 Nucleotide diversity in DNase I footprints

To quantify the nature of selection operating on regulatory DNA, we surveyed nucleotide diversity (π) in footprint calls. Population genetics analyses were performed on 53 unrelated, publicly available human genomes (Table 3.4) released by Complete Genomics, version 1.1032 (Drmanac et al., 2010). Relatedness was determined both by pedigree and with KING (Manichaikul et al., 2010). Two Maasai individuals in the public dataset (NA21732 and NA21737) were not reported

as related, but were found with KING to be either siblings or parent-child. NA21737 was removed from the analysis.

We defined four-fold degenerate sites using NCBI-called reading frames and the NimbleGenSeqCapEZ Exome version 2.0 definition, downloaded from the NimbleGen website (http://www.nimblegen.com/products/seqcap/ez/v2/). Repeats were defined by RepeatMasker, downloaded from the UCSC Genome Browser, version 29Jan2009/open-3-2-7 (Smit et al., 1996–2010). Exome and repeats were removed from all footprints prior to analysis.

3.3.22 π calculation

π for a single variant is $2pq$, where p = major allele frequency and q = minor allele frequency. π was calculated for each cell type by summing π for all variants and dividing by total number of bases considered. Variant sites were filtered by coverage (>20% of individuals must have calls). Additionally, Complete Genomics makes partial calls at some sites (i.e., one allele is `A` and the other is `N`). These were counted as fully missing.

3.3.23 Cell type predominance: motifs within footprints

We scanned hotspot regions for motifs in each cell type using the FIMO software tool with a $P < 10^{-5}$ and defaults for other parameters. Scans included motif templates from TRANSFAC, JASPAR Core, UniPROBE, and novel *de novo* (those with no match to motifs in the aforementioned databases). We filtered predicted motifs to those that overlapped footprints by at least 1 nt. For each cell type, we counted the number of discovered motif instances for a motif template and normalized to the total number of bases within footprints.

3.3.24 Proximal vs. distal regulators

For every motif template, we quantified the number of gene-distal and gene-proximal instances overlapping footprints by at least 1 nt, with proximal defined as within 2500 nt of the TSSs of genes in the reference sequence (NCBI RefSeq; Pruitt et al., 2009). The number of motifs found within a partition was scaled by the number of bases covered by footprints in that partition. Finally, we rescaled the partition values to proportions that summed to one.

3.3.25 Data availability

DNase I-seq production data for Digital Genomic Footprinting (DGF) are available through the NCBI's Gene Expression Omnibus (GEO) data repository (accessions GSE26328 and GSE18927), and also through the UCSC Genome Browser (see http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeUwDgf).

Additional files are available via the EBI ftp server which contains an organized file structure with the ENCODE data. Analysis datasets are located at: ftp://ftp-private.ebi.ac.uk (Login: encode-box-01 Password: enc*deDOWN) in the directories in byDataType.

Acknowledgements

This work was supported by National Institutes of Health (NIH) grants HG004592 (J.A.S.) and RC2HG005654 (J.A.S. and M.G.). J.V. is supported by a National Science Foundation Graduate Research Fellowship under grant no. DGE-071824. Additional support was provided in part by the University of Washington Proteomics Resource (UWPR95794). We thank F. Urnov for critical reading of the manuscript and many discussions, and S. Thomas for insights.

Table 3.1: **Mapping and footprint statistics for 41 cell lines used in this study.**

Cell type	Total sequencing tags	Uniquely mapping sequencing tags	SPOT score	Hotspots (Z > 2)	Hotspot bases (Z > 2)	Hotspot regions (FDR 1%)	Hotspot bases (FDR 1%)	Footprints	Footprint bases
AG10803	369,891,377	284,236,136	0.72	251,730	127,453,885	167,583	67,986,704	1,106,404	16,098,123
AoAF	393,665,970	331,043,131	0.68	247,711	157,461,278	149,702	68,131,539	1,566,170	22,727,963
CD20+	300,895,833	240,594,387	0.57	176,008	143,218,339	74,681	55,800,743	603,190	8,277,635
CD34+ Mobilized	287,605,852	221,098,234	0.7	206,033	119,621,324	125,354	59,782,641	902,386	13,494,265
fBrain	247,674,296	202,264,605	0.72	282,491	145,389,927	166,439	72,393,628	1,022,782	16,300,790
fHeart	328,340,544	264,719,957	0.55	292,700	200,606,802	149,848	77,501,676	954,914	14,276,245
fLung	318,544,188	268,295,068	0.67	338,476	206,554,970	184,428	89,331,237	1,181,235	18,884,673
GM06990	238,189,351	137,532,640	0.62	194,407	124,993,150	71,679	45,274,471	434,561	6,588,543
GM12865	375,878,018	263,505,515	0.56	227,299	132,981,145	118,973	60,969,626	811,374	11,983,394
HAEpiC	347,415,753	281,238,046	0.76	294,231	153,728,366	189,376	78,262,576	1,506,475	20,870,168
HA-h	463,552,126	367,030,503	0.53	316,350	215,720,624	183,402	96,933,344	966,188	13,539,911
HCF	375,266,490	289,961,235	0.7	251,268	136,811,533	158,618	69,285,351	1,057,743	15,531,256
HCM	348,691,463	274,790,101	0.73	259,190	148,481,494	176,297	80,809,789	1,130,292	16,902,956
HCPEpiC	305,511,987	253,925,492	0.76	304,490	166,336,755	194,361	80,354,004	1,296,454	18,684,469
HEEpiC	472,469,677	342,975,637	0.58	326,246	170,658,884	195,205	76,649,605	1,263,648	18,866,093
HepG2	248,320,583	168,883,956	0.57	199,174	115,287,739	73,091	35,775,390	448,678	6,938,557
H7-hESC	401,363,495	302,050,785	0.61	491,178	236,940,779	248,021	99,574,058	1,279,454	18,940,427
HFF	344,017,295	262,521,646	0.59	271,655	158,223,732	173,197	85,614,520	590,904	8,338,681
HIPEpiC	333,832,037	254,744,863	0.58	331,341	179,188,377	209,537	82,901,328	1,089,936	16,510,277
HMF	384,696,734	311,000,443	0.75	266,757	137,564,020	175,343	71,776,020	1,434,330	19,950,879
HMVEC-dBl-Ad	292,823,995	239,063,258	0.72	194,045	124,284,025	142,138	75,710,345	1,085,741	14,936,244

Continued on next page

Table 3.1 – *Continued from previous page*

Cell type	Total sequencing tags	Uniquely mapping sequencing tags	SPOT score	Hotspots (Z > 2)	Hotspot bases (Z > 2)	Hotspot regions (FDR 1%)	Hotspot bases (FDR 1%)	Footprints	Footprint bases
HMVEC-dBl-Neo	368,114,784	293,473,622	0.57	198,849	142,843,889	145,392	87,577,362	1,061,860	15,410,057
HMVEC-dLy-Neo	338,875,033	270,345,138	0.58	196,251	122,494,916	132,021	67,831,113	989,626	14,367,015
HMVEC-LLy	417,841,726	313,021,953	0.62	176,634	115,933,614	122,591	65,162,445	872,721	12,565,891
HPAF	320,990,135	255,470,482	0.7	256,698	140,840,570	169,984	81,414,365	1,090,215	15,983,810
HPdLF	371,416,176	304,268,872	0.67	266,670	168,672,350	156,380	73,248,071	1,404,872	20,203,519
HPF	368,365,528	296,713,698	0.66	235,885	138,436,787	138,004	63,605,792	1,175,289	17,027,101
HRCEpiC	284,056,343	236,736,388	0.6	307,274	155,486,022	178,791	71,906,908	1,187,325	17,124,566
HSMM	467,134,471	367,269,086	0.66	331,104	177,231,683	215,419	94,115,944	1,668,243	23,986,641
Th1	232,708,777	171,609,858	0.64	154,717	117,670,939	63,672	43,580,258	498,505	7,448,335
HVMF	341,994,992	279,802,866	0.63	265,941	159,102,198	154,706	72,203,801	1,263,833	18,402,231
IMR90	309,171,904	242,507,116	0.53	286,260	141,276,695	184,888	79,251,166	970,277	14,355,207
K562	268,452,588	179,970,820	0.56	256,735	157,203,075	125,859	64,943,646	498,683	7,161,934
NB4	404,801,445	323,812,091	0.56	236,509	141,522,055	119,640	62,875,330	1,049,300	15,418,984
NH-A	307,812,903	231,589,045	0.57	280,019	148,952,898	176,271	80,278,529	977,923	14,329,589
NHDF-Ad	300,516,213	235,650,107	0.81	296,898	151,532,500	212,841	93,354,642	1,429,399	20,950,088
NHDF-neo	482,603,639	373,361,757	0.7	275,166	153,433,828	172,878	76,999,986	1,532,853	22,147,781
NHLF	454,391,713	357,163,548	0.71	294,352	166,314,415	190,888	85,453,334	1,567,106	22,751,625
SAEC	296,719,796	243,838,476	0.58	291,390	159,165,382	184,542	72,503,786	1,256,188	18,742,067
SKMC	632,856,867	543,886,965	0.81	311,537	158,366,070	193,202	74,105,557	2,370,723	31,607,291
SK-N-SH_RA	217,691,024	164,615,431	0.7	160,880	91,155,430	70,493	37,635,241	498,926	7,609,202

Table 3.2: **Summary of footprints within DHSs**.

Cell type	Total FPs	Total DHS peaks	FPs in DHS peaks	DHS peaks with FP	Mean FP per DHS peak
AG10803	1,106,404	181,473	677,479	139,806	4.84585
AoAF	1,566,170	165,258	820,187	148,612	5.51898
CD20	603,190	104,139	303,432	72,752	4.17077
CD34+ Moblized	902,386	147,098	560,210	117,862	4.7531
fBrain	1,022,782	182,501	636,950	140,256	4.54134
fHeart	954,914	173,135	562,780	129,032	4.36155
fLung	1,181,235	205,880	681,428	160,948	4.23384
GM06990	434,561	92,709	195,168	49,295	3.95918
GM12865	811,374	143,716	487,801	104,614	4.66287
HAEpiC	1,506,475	205,033	913,983	172,375	5.30229
HAh	966,188	200,014	506,977	134,600	3.76655
HCF	1,057,743	174,667	647,025	135,144	4.78767
HCM	1,130,292	193,375	696,405	146,587	4.7508
HCPEpiC	1,296,454	210,380	826,565	167,674	4.9296
HEEpiC	1,263,648	209,838	834,743	173,806	4.80273
HepG2	448,678	90,775	228,280	54,600	4.18095
hESCT0	1,279,454	266,618	808,678	189,181	4.27463
HFF	590,904	192,282	384,995	106,555	3.61311
HIPEpiC	1,089,936	225,744	731,881	164,569	4.44726
HMF	1,434,330	190,512	874,301	162,132	5.39253
HMVEC_dBlAd	1,085,741	162,593	644,136	123,503	5.21555
HMVEC_dBlNeo	1,061,860	168,436	633,452	124,918	5.07094
HMVEC_dLyNeo	989,626	153,107	603,547	120,801	4.99621
HMVEC_LLy	872,721	144,886	550,573	111,126	4.95449
HPAF	1,090,215	188,071	684,069	140,068	4.88383
HPdLF	1,404,872	171,349	785,700	147,294	5.33423
HPF	1,175,289	154,397	683,890	131,805	5.18865
HRCE	1,187,325	192,147	723,271	146,937	4.92232
HSMM	1,668,243	228,282	937,370	184,856	5.07081
hTH1	498,505	84,201	220,748	53,494	4.12659
HVMF	1,263,833	170,340	688,248	137,947	4.98922
IMR90	970,277	199,752	646,563	139,353	4.63975
K562	498,683	142,986	305,128	72,048	4.23507
NB4	1,049,300	143,838	588,282	117,445	5.009

Continued on next page

Table 3.2 – *Continued from previous page*

Cell type	Total FPs	Total DHS peaks	FPs in DHS peaks	DHS peaks with FP	Mean FP per DHS peak
NHA	977,923	191,510	601,546	130,914	4.59497
NHDF_Ad	1,429,399	230,696	891,028	179,529	4.96314
NHDF_Neo	1,532,853	187,962	840,887	160,662	5.23389
NHLF	1,567,106	206,254	896,218	173,139	5.17629
SAEC	1,256,188	198,442	791,686	160,216	4.94137
SkMC	2,370,723	205,493	1,230,494	198,952	6.18488
SK-N-SH_RA	498,926	89,968	259,755	61,111	4.25054

Table 3.3: **Sequence oligos used for DIPP.**

Probe ID	IDT sequence
MAX Specific F	/5BiosG/CTGGAGACTTGCAGGGTGGACAGACACACGTGGGGAAGGTTCCCGCTGCACACAACTCAACTCTGACCTG
MAX Specific R	CAGGTCAGAGTTGAGTTGTGTGCAGCGGGAACCTTCCCCACGTGTGTCTGTCCACCCTGCAAGTCTCCAG
MAX Shuffled F	/5BiosG/CTGGAGACTTGCAGGGTGGACAGACACAGTTCAGAGCGGCCGTGGCTGCACACAACTCAACTCTGACCTG
MAX Shuffled R	CAGGTCAGAGTTGAGTTGTGTGCAGCCACGGCCGCTCTGAACTGTGTCTGTCCACCCTGCAAGTCTCCAG
AP1 Site 1 Specific F	/5BiosG/CATCTGGGCACACACCCTAAGCCTCAGCATGACTCATCATGACTCAGCATTGCTGTGCTTGAGCCAGAAG
AP1 Site 1 Specific R	CTTCTGGCTCAAGCACAGCAATGCTGAGTCATGATGAGTCATGCTGAGGCTTAGGGTGTGTGCCCAGATG
AP1 Site 1 Shuffled F	/5BiosG/CATCTGGGCACACACCCTAAGCCTTGCAACCGGACAACAAGCGCCTTATTTTCTGTGCTTGAGCCAGAAG
AP1 Site 1 Shuffled R	CTTCTGGCTCAAGCACAGAAAATAAGGCGCTTGTTGTCCGGTTGCAAGGCTTAGGGTGTGTGCCCAGATG
AP1 Site 2 Specific F	/5BiosG/TGGGATTATCAGGCTGGAGTTCTCTGTCATTAGGATGACTCATCATTTTTCTATCTCTGCTTCCATTGCT
AP1 Site 2 Specific R	AGCAATGGAAGCAGAGATAGAAAAATGATGAGTCATCCTAATGACAGAGAACTCCAGCCTGATAATCCCA
AP1 Site 2 Shuffled F	/5BiosG/TGGGATTATCAGGCTGGAGTTCTCTGTAGTATATTCCTCATATTCTTGAGCTATCTCTGCTTCCATTGCT
AP1 Site 2 Shuffled R	AGCAATGGAAGCAGAGATAGCTCAAGAATATGAGGAATATACTACAGAGAACTCCAGCCTGATAATCCCA

Table 3.4: **Complete genomics sequence IDs**. Genomes used for computing nucleotide diversity (π).

Assemblies	**Population**
GS20845-1100-37-ASM	Gujarati
GS20846-1100-37-ASM	
GS20847-1100-37-ASM	
GS20850-1100-37-ASM	
HG00731-1100-37-ASM	Hispanic
HG00732-1100-37-ASM	
GS19735-1100-37-ASM	
GS19648-1100-37-ASM	
GS19649-1100-37-ASM	
GS19669-1100-37-ASM	
GS19670-1100-37-ASM	
GS19238-1100-37-ASM	African
GS19020-1100-37-ASM	
GS19025-1100-37-ASM	
GS19026-1100-37-ASM	
GS21732-1100-37-ASM	
GS21733-1100-37-ASM	
GS21767-1100-37-ASM	
GS18501-1100-37-ASM	
GS18502-1100-37-ASM	
GS18504-1100-37-ASM	
GS18505-1100-37-ASM	
GS18508-1100-37-ASM	
GS18517-1100-37-ASM	
GS19219-1100-37-ASM	
GS19700-1100-37-ASM	African-American
GS19701-1100-37-ASM	
GS19703-1100-37-ASM	
GS19704-1100-37-ASM	
GS19834-1100-37-ASM	
GS12890-1100-37-ASM	European
GS12890-1100-37-ASM	
GS12891-1100-37-ASM	
GS12892-1100-37-ASM	

Continued on next page

Table 3.4 – *Continued from previous page*

Assemblies	**Population**
GS20502-1100-37-ASM	
GS20509-1100-37-ASM	
GS20510-1100-37-ASM	
GS20511-1100-37-ASM	
GS06985-1100-37-ASM	
GS06994-1100-37-ASM	
GS07357-1100-37-ASM	
GS10851-1100-37-ASM	
GS12004-1100-37-ASM	
GS18940-1100-37-ASM	Asian
GS18942-1100-37-ASM	
GS18947-1100-37-ASM	
GS18956-1100-37-ASM	
GS18526-1100-37-ASM	
GS18537-1100-37-ASM	
GS18555-1100-37-ASM	
GS18558-1100-37-ASM	

Chapter 4

COUPLING TRANSCRIPTION FACTOR OCCUPANCY TO NUCLEOSOME ARCHITECTURE WITH DNASE-FLASH

This chapter has been adapted with minor changes from: Vierstra, J., Wang, H., John, S., Sandstrom, R. and Stamatoyannopoulos, J. A. Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH. *Nat. Meth.* 11, 66–72 (2014).

Abstract

It is currently not possible to resolve the genome-wide relationship of transcription factors (TFs) and nucleosomes at the level of individual chromatin templates despite rapidly increasing data on TF and nucleosome occupancy in the human genome. Here we describe DNase I-released fragment length analysis of hypersensitivity (DNase-FLASH) an approach that directly couples mapping of TF occupancy within regulatory DNA, via quantification of DNA microfragments released from individual TF recognition sites, to the surrounding nucleosome architecture, via analysis of larger DNA fragments, in a single assay. DNase-FLASH enables coupling of individual TF footprints to nucleosome occupancy, identifying TFs that precisely demarcate the regulatory DNA-nucleosome interface.

4.1 Introduction

In vivo, genomic DNA is contacted chiefly by nucleosomes and sequence-specific transcription factors (TFs), which bind regulatory DNA regions in a mutually exclusive fashion (Gross and Garrard, 1988; Workman et al., 1988). A detailed description of the chromatin architecture of regulatory DNA regions has remained elusive, and hinges on understanding the relationship between the occupancy of transcription factor recognition sites, both individually and in combination, and the precise positioning of surrounding nucleosomes on the same chromatin molecule. Current approaches have the potential to measure occupancy of transcription factors or nucleosomes separately, but not concurrently. While, genomic DNase I footprinting (Hesselberth et al., 2009; Neph et al., 2012b) provides nucleotide resolution of transcription factor footprints and enables quantification of their occupancy, it does not provide information on co-occupancy of factors nor their connection with surrounding chromatin features. MNase has long been used to map nucleosome architectures, yet even under ideal conditions, confidently mapping nucleosome positioning within the human genome by MNase-seq requires exceptionally deep sequencing coverage totaling billion of reads (Gaffney et al., 2012; Valouev et al., 2011). MNase chromatin immunoprecipitation-based approaches require only modest levels of sequencing, however, are limited to a single nucleosome variant per experiment. Most importantly, while MNase is effective for mapping nucleosome positioning, its marked propensity for cleavage within inter-nucleosomal linker regions (Axel, 1975) renders it ineffective as a TF footprinting agent.

Although DNase I has traditionally been applied to localize TF-contacted DNA elements, its ability to cleave accessible inter-nucleosomal linkers has been exploited to probe nucleosome structure both *in vitro* (Lutter, 1979) and *in vivo* (Staynov, 2008). In addition, DNase I has played a unique role in understanding nucleosome-DNA contacts due to an inherent propensity for nuclease action within the minor groove of DNA on the nucleosome surface, resulting in cleavage products with a characteristic 10.4 bp period (Noll, 1974).

Here we report an approach for high-resolution exposition of regulatory DNA architecture that exploits the full range DNase I-chromatin engagement. We show that systematic analysis of DNase I-released DNA fragments over a wide range of sizes enables recognition and high-resolution analysis of three key features of regulatory DNA biology, including the quantitative occupancy of individual TFs; TF-nucleosome interactions; and nucleosome architecture surrounding regulatory DNA. This unifying approach, which we term DNase I-released fragment

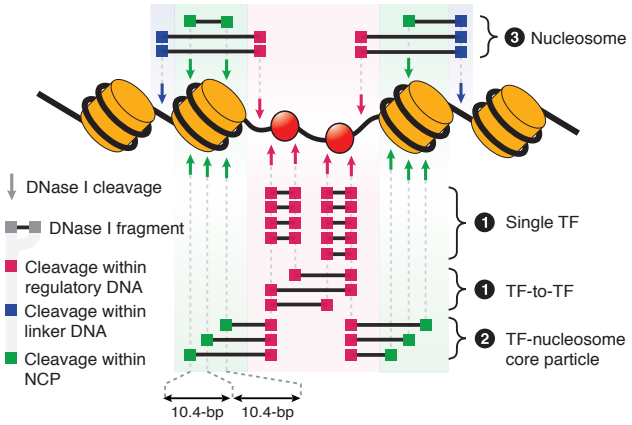

Figure 4.1: **Model of DNase I cleavage action within chromatin**. Arrows indicate cleavage site, canonical DNase I hypersensitive sites are shaded in red, adjacent nucleosomes in green and inter-nucleosomal linker regions in blue. Cleavages within the nucleosome core particle are expected to occur at a 10.4 bp periodicity.

length analysis of hypersensitivity (DNase-FLASH), enables for the first time direct measurement of the influence of transcription factor occupancy on nucleosome positioning at a single locus *in vivo*. Although applied to the human genome, the principles are generic and thus should be readily extensible to any eukaryotic organism.

4.2 Results

4.2.1 Mapping nucleosomes and transcription factor footprints

We previously observed that DNase I digestion of nuclear chromatin releases small (<500-bp) 'double-cut' DNA fragments from DNase I hypersensitive sites (DHSs) (Hesselberth et al., 2009). In theory, these fragments could encompass several primary events, represented by fragments derived from: (1) two cleavages within the core TF binding region of canonical DNase I hypersensitive sites; (2) one cleavage within the core TF binding region and a second on an adjacent nucleosome; (3) one cleavage within the core TF binding region and a second within an inter-nucleosomal linker region; and (4) two cleavages from internucleosomal linker regions or within the nucleosome core particle itself (Figure 4.1). We speculated that TF occupancy-

DNase I library	Total fragments mapped ($Q > 1$)	Properly mapped (no chrM)	Percent uniquely mapped	Total cleavages detected
>200 bp	119,623,788	88,763,949	75.0%	179,527,898
200–400 bp	209,192,086	162,097,644	77.5%	324,195,288
Combined	328,815,874	251,861,593	76.6%	503,723,186

Table 4.1: **Summary of sequencing statistics corresponding to the two size fractions gel purifed**.

associated fragments (class 1) could be directly separated from nucleosomal fragments (classes 3 and 4) simply on the basis of size. To test this, we prepared DNase I double-cut fragments from chromatin of primary human gingival fibroblasts and biochemically separated (over sucrose gradients) two distinct size fractions (Figure 4.2 and Figure 4.3), the first containing fragments <200-bp, and the second fragments ranging from 200- to 400-bp in size. We prepared Illumina sequencing libraries from each size fraction and performed paired-end sequencing, yielding between 89.7 and 162 million uniquely mappable fragments derived from >500 million DNase I cleavage events (Table 4.1). The observed distribution of fragment lengths was bimodal, with peaks corresponding to small fragments (mean ≈60-bp) and large fragments (mean ≈165-bp) (Figure 4.4a). Moreover, we observed that the distribution contained secondary peaks in regular intervals. To test for periodicity and ascertain the period, we performed Fourier analysis of the fragment length distribution and visualized signal intensity vs. frequency on a power spectrum (Figure 4.4b). This analysis revealed a dominant frequency of precisely 10.4 bp, consistent with nuclease cleavage within the exposed minor groove of DNA wrapping the nucleosome surface (Klug and Lutter, 1981; Lutter, 1979).

On the basis of fragment length, we partitioned the data into two classes: fragments <125 bp and 126- to 185 bp (Figure 4.5). For each size range, we calculated the fragment density over the entire genome in non-overlapping 5 bp intervals and developed a simple algorithm to map nucleosomes by identifying peaks enriched in nucleosome-length fragments within broader zones of increased DNase I cleavage (DNase I 'hotspots'; John et al., 2011). Using this approach, we conservatively identified 541,737 nucleosomes within 226,985 DNase I hotspots. In total, we detected at least one nucleosome associated with 91.3% of DHSs, while the vast majority (71.3%) of DHSs contained two or more.

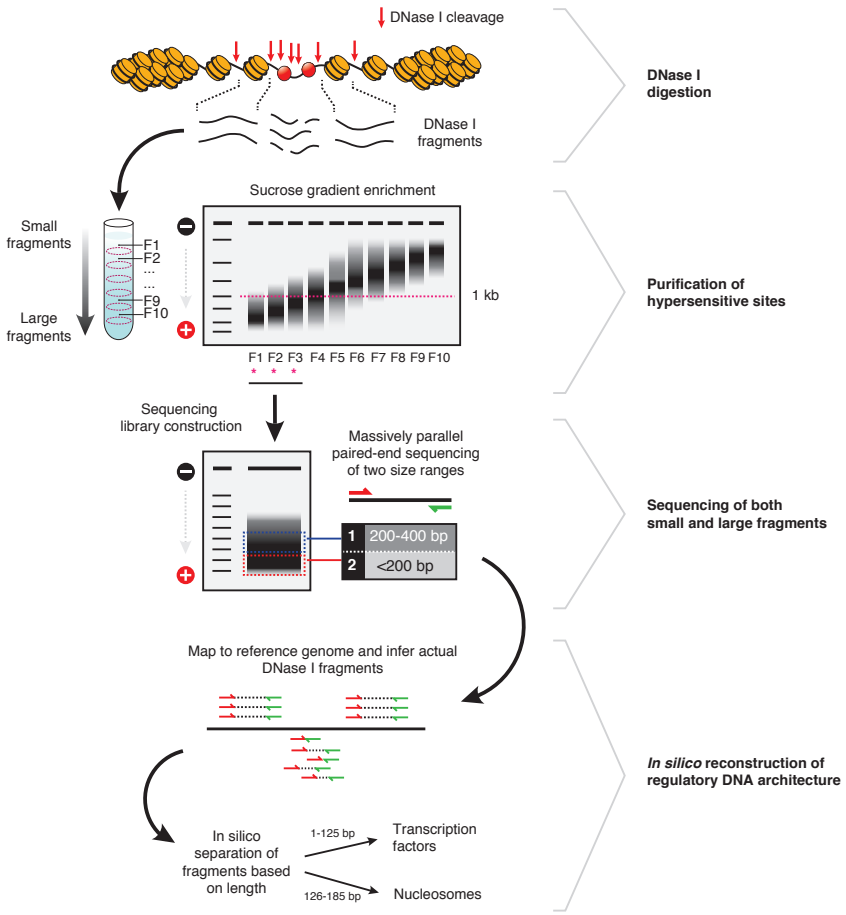

Figure 4.2: **Outline of the DNase-FLASH method**. A standard DNase I digestion of intact human nuclei (John et al., 2011; Neph et al., 2012b; Thurman et al., 2012) preferentially releases small fragments pertaining to sites of transcription factor binding and nucleosome occupancy. DNase I hypersensitive sites (DHSs) are purified via ultracentrifugation of DNA layered on top of a 9% sucrose cushion. Small fragments remained near the top of tube, while larger fragments migrated to the bottom. The separation was monitored on a 1.5% agarose gel and fragments <1 kb (red dashed line) were used to construct a DNase I library. Canonical DHS fragments are small (<100 bp) while nucleosomal fragments are larger (>125 bp), therefore two size ranges were selected from the same DNase I library via agarose gel extraction and sequenced independently in paired-end mode. The paired sequencing reads were mapped to the genome and used to reconstruct DNase I fragments in silico. Stratification of the inferred fragments on the basis on length revealed sites of transcription factor or nucleosome occupancy.

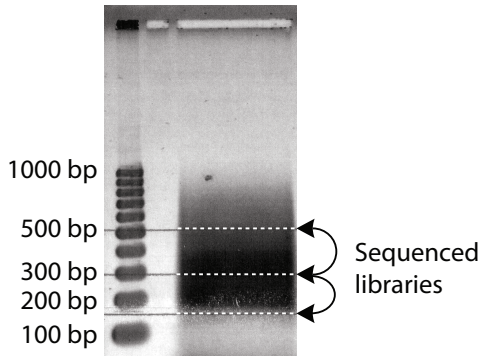

Figure 4.3: **Size selection of DNase I fragments**. DNase I fragments were purified and ligated with adapters compatible with the Illumina sequencing platform. The ligated fragments were amplified with 8 cycles of PCR and loaded onto a 2% agarose gel and run for 100V for 1 hr in Tris-acetate/EDTA (TAE) buffer. DNA was purified from two independent regions of the gel and sequenced. Dotted lines indicate the excised gel slices.

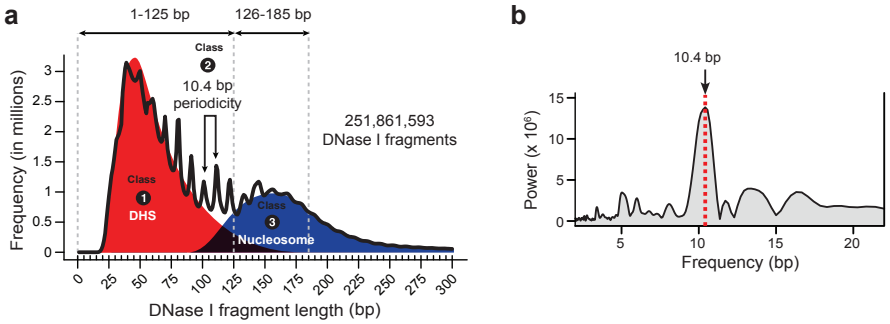

Figure 4.4: **DNase I fragment size spectrum**. (a) A histogram of inferred DNase I fragment lengths from the >250 million fragments sequenced. The shaded curves highlight the regions on the bimodal distribution that correspond to fragments derived from cleavages in a DHS (orange) or around nucleosomes (blue). Grey dashed lines demarcate the size ranges used for the stratification of fragments. (b) The periodicity of the histogram in a was analyzed by Fourier transformation and plotted as a power spectrum (x-axis, frequency in base-pairs; y-axis, magnitude of signal).

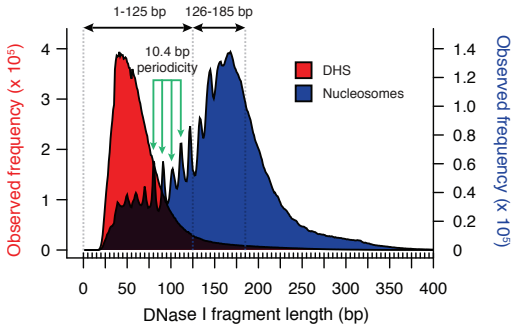

Figure 4.5: **Distribution of the lengths of DNase I fragments overlapping DHS or nucleosomes**. The two curves correspond to fragments overlapping the top 10% most accessible DHS (red) or nucleosomes (blue). Grey dashed lines demarcate the size ranges used for the stratification of fragments. Green arrows highlight the 10.4 bp periodicity.

We compared the length of fragments that overlapped either canonical DHS or nucleosomal peaks identified by our algorithm and observed a strong separation between the two distributions, with peak maxima occurring at ≈45 bp for DHS and ≈165 bp for nucleosomes (Figure 4.5). Importantly, the 10.4 bp periodic interval was only present in the distribution corresponding to nucleosome-size fragments. These separations were clearly evident at the level of individual elements, highlighted by a compact signal over core DHS regions that was well separated from flanking broader ≈150 bp peaks consistent with 5'- and 3'-flanking nucleosomes (Figure 4.6a–c). We also plotted per nucleotide cleavage, revealing clear DNase I footprints within the sub-125 bp fraction (Figure 4.6a–c). Additionally, while fragments released from within the nucleosome occurred in stereotyped 10.4 bp periods, the profile of DNase I cleavages within the nucleosome core revealed no detectable cleavage preferences (Figure 4.7).

Because nucleosomes immediately flanking many regulatory regions are enriched in H2A.Z (Jin et al., 2009), we performed native chromatin immunoprecipitation (ChIP) for this variant on MNase-digested chromatin. In promoter-proximal regions, DHSs are stereotypically associated with H3K4me3-modified nucleosomes13, and therefore we also performed native ChIP for this modification. Superimposing both the H2A.Z and H3K4me3 signals on the DNase I cleavage densities revealed that the former tightly tracked the latter (Figure 4.6a–c), and directly

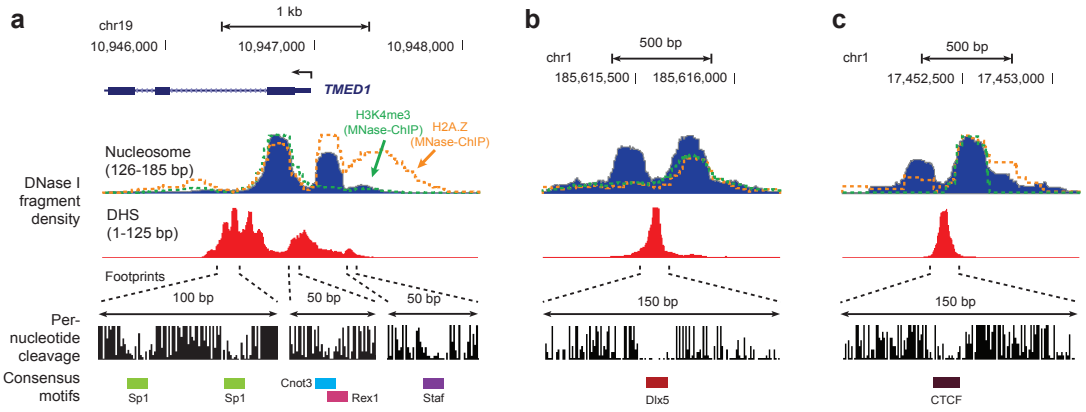

Figure 4.6: **DNase-FLASH data at exemplar loci**. DNase I hypersensitive sites in both TSS proximal (a) and distal configurations (b,c). (a) Nucleosome positioning (blue) DHS (red) and consensus motifs for 5 TFs are shown. Dotted lines correspond to the density of the histone modifications H3K4me3 (green) and H2A.Z (orange) measured by native ChIP-seq of MNase digestion chromatin from the same cell type. (b–c) The *cis*-regulatory architecture surrounding distal binding sites of (b) DLX5 and (c) CTCF.

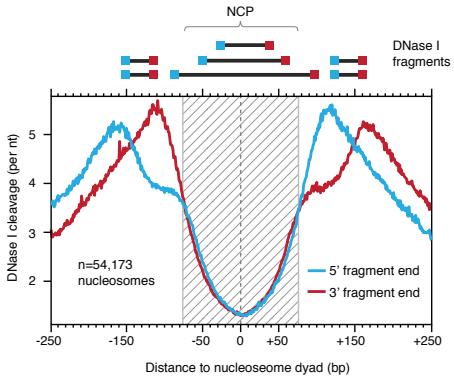

Figure 4.7: **DNase I cleavage on the nucleosome core**. Per nucleotide DNase I cleavage profile around the 10% ($n = 54,173$) most accessible nucleosomes. Blue, 5′ ends of DNase I fragments. Red, 3′ ends of DNase I fragments. Hatched region highlights the ≈147 bp nucleosome core particle. Top, localization of theoretical DNase I fragments released from digestion of the nucleosome. Within the nucleosome core the 5′ and 3′ cleavages are offset by 3 bp.

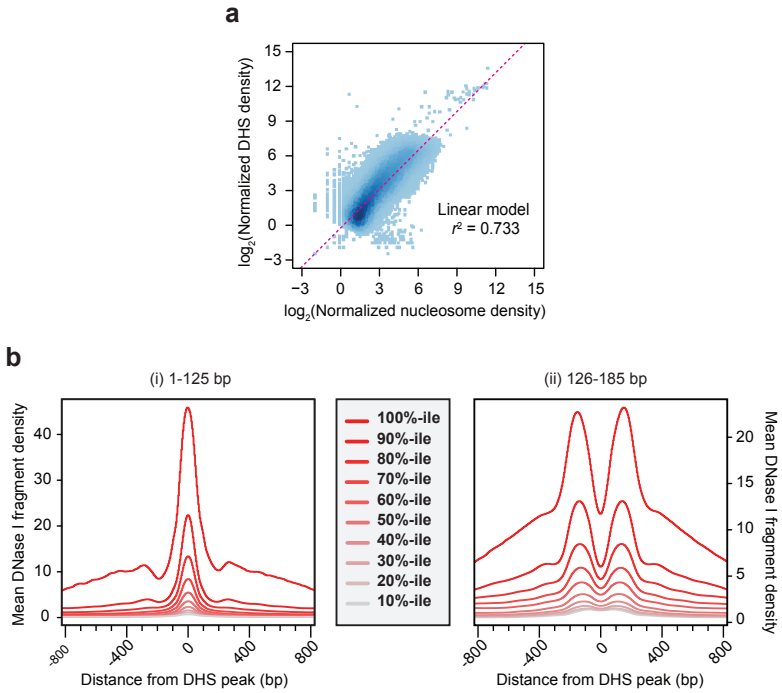

Figure 4.8: **DHS signal parellels stoichiometry of TF/nucleosome occupancy**. DNase I fragments were partitioned into two classes: fragments ≤125 bp and 126-185 bp. For each size range, we calculated the fragment density over the entire genome in non-overlapping 5 bp intervals and normalized the values to the total fragments in each class. (a) Scatterplot of small (y-axis) vs. large fragment (x-axis) reveals a strong correlation (linear model; $r^2 = 0.73$, $P < 2.2 \times 10^{-16}$). (b-c) Each of 197,777 DNase I hypersensitive peaks (150 bp window) were separated into deciles on the basis of maximum small fragment density within the window. For each decile, an aggregate signal profile was computed for small fragments (≤125 bp) (i) or large fragments (126-185 bp).

confirms the nucleosomal identity of the DHS-flanking peaks.

4.2.2 Nucleosomes bookend TF-occupied regulatory DNA

We next asked how nucleosome occupancy flanking core DHS regions was related to TF occupancy within these regions. We computed the maximum density of fragments mapping to the core TF-occupied region of each DHS vs. the density of fragments emanating from flanking nucleosome-occupied regions (Figure 4.8a). We found a strong linear correlation between these (linear model, $r^2 = 0.73$; $P < 2.2 \times 10^{-16}$) indicating that positionally stable occupancy of

flanking nucleosomes is directly proportional to transcription factor binding within regulatory DNA. Partitioning each DHS into ten bins of equal size (deciles) based on accessibility and computing average density profiles of both small (<125 bp) and large (126- to 185 bp) fragment size classes revealed a marked inverse relationship between the density profiles corresponding to DHS and nucleosomes (Figure 4.8b). The inverse relationship of the two size fraction density profiles is compatible with a general mutual exclusivity of transcription factor and nucleosome occupancy at individual chromatin molecules (Figure 4.6a–c). In addition, we also found the average large fragment (126- to 185 bp) signal to be half that of the corresponding DHS signal (<125 bp) for each decile, indicating that for each DHS identified, we reliably detect two adjacent flanking nucleosomes (Figure 4.8b).

4.2.3 DNase I microfragments define high-occupancy binding sites

We next investigated the abundance of small DNA fragments released by DNase I from core hypersensitive regions. In conventional footprinting assays, transcription factor occupancy is frequently associated with enhanced DNase I cleavage of immediately flanking nucleotides, presumably due to distortion of the minor groove (Stamatoyannopoulos et al., 1995). We therefore hypothesized that small DNase I-released fragments would preferentially result from such sites, and thus have a regular relationship with TF occupancy, with progressively smaller fragments more closely approximating the TF recognition site and thus signifying progressively higher average occupancy. To analyze this relationship, we studied several TFs with well-defined recognition sequences that are known from prior studies to exhibit wide ranges of affinity and average occupancy of cognate recognition sites within DHSs: the polyfunctional genomic regulator CTCF; the nuclear respiratory factor NRF1; and the CCAAT-box binding factor NF-Y. For each TF, we sorted their putative binding sites within accessible chromatin by the total DNase I cleavage density ±25 bp of the recognition sequence (collectively, 32,426 CTCF sites, 4,716 NRF1 sites, 5,092 NF-Y sites) (Figure 4.9a and Figure 4.10a,e). We then partitioned the recognition sites for each TF into deciles, and plotted in parallel the mean length of DNA fragments overlapping the recognition sites within each decile (Figure 4.9b and Figure 4.10b,f). This analysis revealed a clear relationship between the size of released DNA fragments and transcription factor occupancy, with the most highly occupied sites releasing smaller fragments that directly overlapped the recognition sequence (Figure 4.9a–b, Figure 4.10a–b and Figure 4.10e–f).

To analyze more closely the landscape of small DNase I fragments immediately related to TF recognition sequences, we computed, for each CTCF recognition sequence in DHSs, the

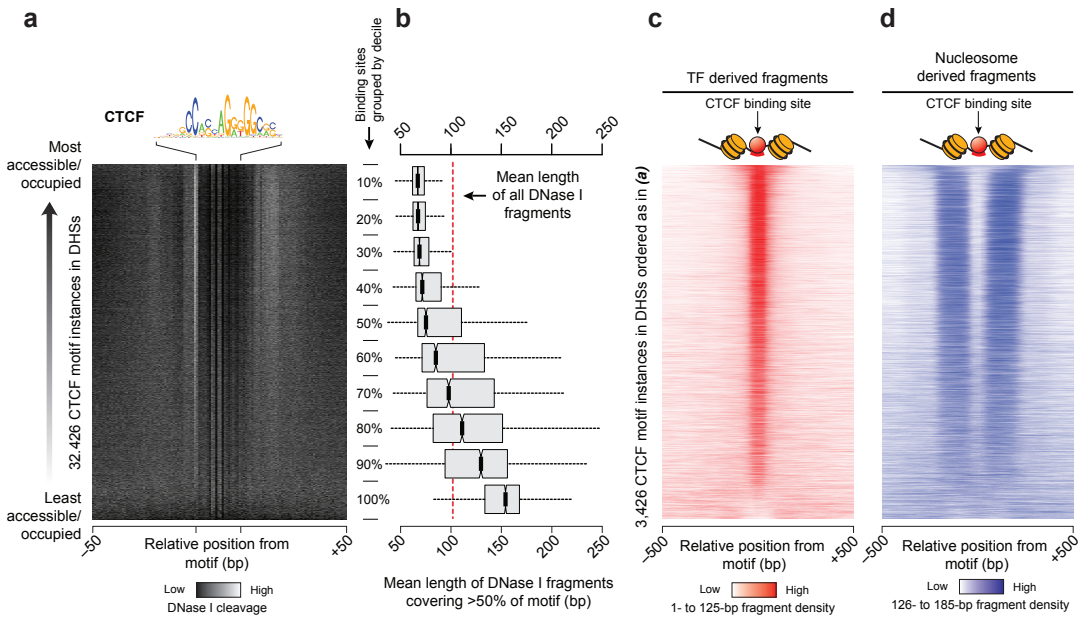

Figure 4.9: **DNase I fragment length parallels transcription factor occupancy**. (a) Heatmap (bottom) of the per-nucleotide cleavages surrounding ±50-bp of 32,426 predicted CTCF binding sites in accessible chromatin. Each row of the heatmap corresponds to one predicted CTCF binding site and columns correspond to each nucleotide ±50 bp surrounding the TF recognition sequence. Rows are sorted by decreasing cleavage density within the ±25 bp window surrounding the motif instance. (b) Distribution of the mean fragment lengths overlapping each CTCF motif. Each box corresponds to 10% of the predicted CTCF binding sites ordered as in (a). The dashed red line represents the mean fragment length of all fragments detected irrespective of binding site motif. (c-d) Heatmaps of the density of (c) TF derived fragments (≤125 bp) and (d) nucleosome derived fragments (126- to 185 bp) ±500 bp surrounding predicted CTCF binding sites ordered as in (a).

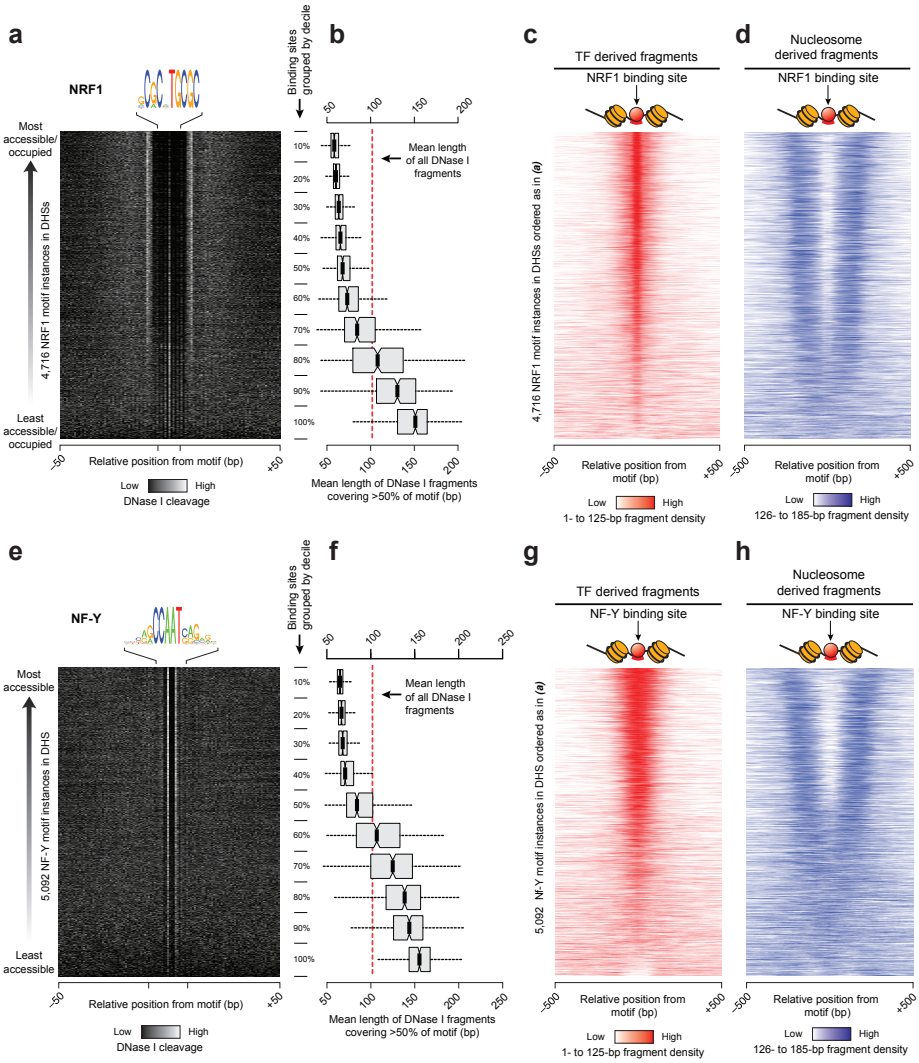

Figure 4.10: **DNase I fragment length parallels transcription factor occupancy at NRF1 and NF-Y bindings sites**. (a) Heatmap (bottom) of the per-nucleotide cleavages surrounding ±50-bp of 4,716 predicted NRF1 binding sites in accessible chromatin. Each row of the heatmap corresponds to one predicted NRF1 binding site and columns correspond to each nucleotide ±50 bp surrounding the TF recognition sequence. Rows are sorted by decreasing cleavage density within the ±25 bp window surrounding the motif instance. (b) Distribution of the mean fragment lengths overlapping each NRF1 motif. Each box corresponds to 10% of the predicted NRF1 binding sites ordered as in (a). The dashed red line represents the mean fragment length of all fragments detected irrespective of binding site motif. (c-d) Heatmaps of the density of (c) TF derived fragments (≤125 bp) and (d) nucleosome derived fragments (126- to 185 bp) ±500 bp surrounding predicted CTCF binding sites ordered as in (a). (e–f) Same as (a–d) for the transcription factor NF-Y.

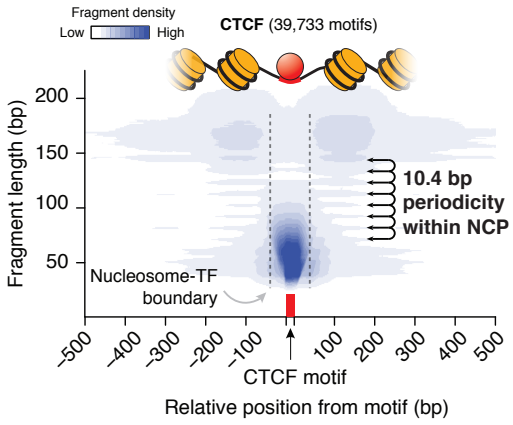

Figure 4.11: **Transcription factors position nucleosomes and organize chromatin structure in regulatory regions**. The density of DNase I fragments (white, low; blue, high) is mapped by fragment length (y-axis), relative to the position of the predicted binding site (x-axis). The red box corresponds to the consensus CTCF motif. Black arrows in the profile for CTCF highlight the ≈10.4 bp periodicity of the fragment lengths within the nucleosome core particle.

spatial density of small and large DNaseI-released fragments within a 500 bp region centered on the recognition sequence (Figure 4.9c), and ranked the sites as in Figure 4.9a. This analysis showed that the highest density of small fragments occurred over the region spanned by the CTCF recognition sequence. A parallel analysis of nucleosomal-size fragments (126- to 185 bp) revealed depletion of such fragments covering the binding site, but increased density in the flanking sequences (Figure 4.9d). A similar fragment analysis for both NRF1 (Figure 4.10c–d) and NF-Y (Figure 4.10h–h) confirmed the distinct localization of TF- and nucleosome-derived fragments.

We next calculated the density of DNase I fragments lengths covering each nucleotide position around CTCF recognition motifs in DHSs (Figure 4.11). This showed that the majority of very small DNase I fragments precisely spanned the CTCF motif, thus localizing the site of protein-DNA interaction and directly confirming our finding that short DNase I-released fragments correlate with transcription factor occupancy. We also found that larger fragments (126- to 185 bp) accumulated adjacent to the CTCF binding sites, suggesting a strong effect of CTCF occupancy on proximal nucleosome positioning (Fu et al., 2008). Furthermore, clear 10.4

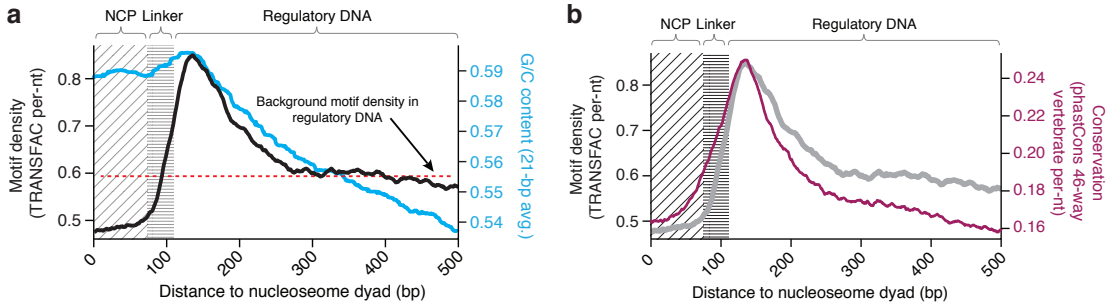

Figure 4.12: **Nucleotide conservation surrounding nucleosomes flanking DHSs**. (a) Density of predicted transcription factor binding sites (TRANSFAC, $P < 10^{-5}$) around the 10% most accessible nucleosomes reveals that TF recognition sequences are enriched at nucleosome boundaries. The dashed red line marks average motif density in DHS genome-wide. Blue line indicates the average G+C content of the genomic sequence underlying these regions measured in 21 bp intervals. (a) Nucleotide conservation surrounding the 10% most accessible nucleosomes. Grey, density of predicted transcription factor binding sites (TRANSFAC, $P < 10^{-5}$). Magenta, the average (mean) phastCons (46-way vertebrate) score. The hatched region highlights the 73 bp of DNA wrapped around the nucleosome core particle relative to the dyad axis; horizontal lines demarcate the 35 bp of linker DNA.

bp laddering of DNA fragment lengths was observed within sites of nucleosome occupancy, indicating CTCF-mediated translational positioning constraints.

4.2.4 TFs demarcate the regulatory DNA-nucleosome interface

DHSs are densely populated by transcription factor recognition sequences and footprints (Neph et al., 2012b). To investigate how these features transition to flanking sequences occupied by well-positioned nucleosomes, we selected the strongest 10% (≈54,000) peri-DHS nucleosomal regions ranked by fragment density (126- to 185-bp) and computed per-base motif density as a function of distance from the predicted nucleosome dyad (Figure 4.12a). This revealed marked selective depletion of TF recognition sequences coinciding with the nucleosome core particle, increased density within the linker region, and maximal density ≈120-bp from the nucleosome dyad that was also reflected by nucleotide-level conservation (Figure 4.12b). In addition, we found that the positions of maximum motif density coincided with sequence dinucleotide patterns associated with strong nucleosome positioning (Valouev et al., 2011) (Figure 4.13a), indicating that TF binding shifts the nucleosome from occupying a more energetically favorable

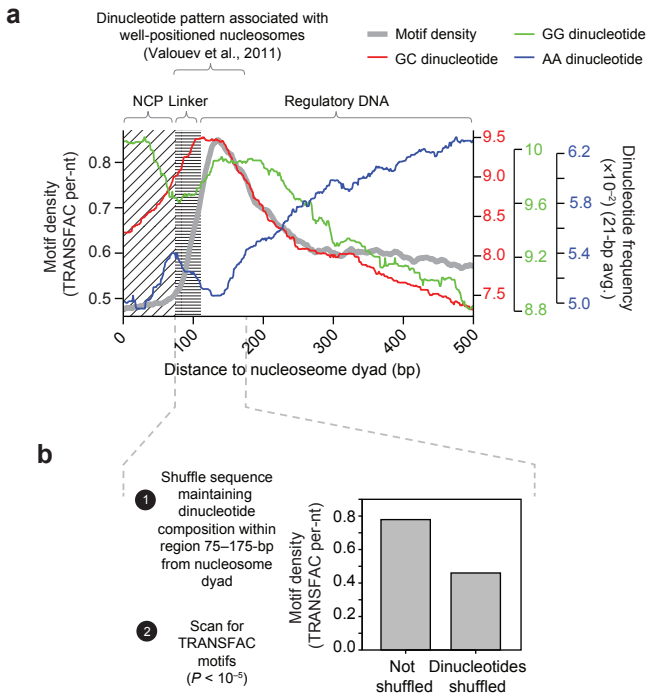

Figure 4.13: **Sequence composition surrounding nucleosomes flanking DHSs**. (a) Sequence features surrounding the 10% most accessible nucleosomes. Grey, density of predicted transcription factor binding sites (TRANSFAC, P < 10–5). Colored curves show the dinucletide frequency for GC dinucleotides (green) and AA dinucleotides (blue). The hatched region highlights the 73 bp of DNA wrapped around the nucleosome core particle relative to the dyad axis; horizontal lines demarcate the 35 bp of linker DNA. (b) Analysis of TRANSFAC motif density controlling for sequence dinucleotide composition.

location. Notably, controlling for sequence composition by permutation of dinucleotides could not account for the increase in motif density (Figure 4.13b). Taken together, these results are compatible with transcription factor occupancy as a major determinant of nucleosome positioning around human regulatory DNA.

To quantify the impact of TF occupancy on nucleosome positioning, we first computed the distribution of DNase I-released nucleosomal fragments surrounding highly occupied (top 10%) TF recognition sequences in DHSs. The nucleosome profile surrounding high occupancy CTCF and AP-1 sites contained narrow punctate peaks at 125 bp, 375 bp, and 575 bp relative to the center of the binding sites (Figure 4.14a–b), consistent with phased organization of nucleo-

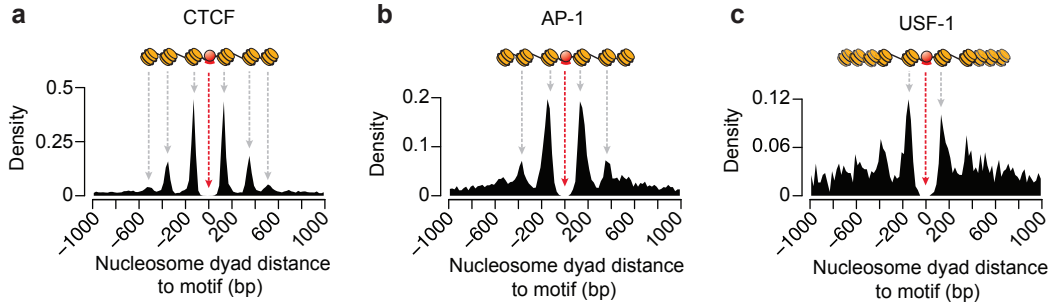

Figure 4.14: **Aggregate nucleosome organization surrounding TF binding sites**. (a–c) The nucleosome profiles surrounding the top 10% most highly occupied transcription factor binding sites of (a) CTCF, (b) AP-1 and (c) USF-1.

somes (Cuddapah et al., 2009; Fu et al., 2008). By contrast, the nucleosomal profile surrounding high occupancy USF-1 binding sites was markedly heterogeneous, with the notable depletion of the 375 bp and 575 bp peaks (Figure 4.14c).

We next expanded our analysis to all recognition sites for these regulators in DHSs by plotting nucleosome organization around each TF recognition sequence ranked by overall DNase I accessibility (Figure 4.15a–c). These data collectively illustrate that structurally and functionally diverse transcription factors such as CTCF and AP-1 impart strong nucleosome positioning in an occupancy dependent fashion. These data (Figure 4.15a–c) are compatible with a statistical distribution model of nucleosome positioning (Kornberg and Stryer, 1988), in which a barrier or boundary (e.g., a transcription factor) exerts an effect on nucleosome positioning that is maximal on proximal nucleosomes and diminishes with distance from the boundary. By contrast, transcription factors such as USF-1 are poor effectors of nucleosome positioning (Figure 4.15c), confirming that not all TF barriers impart equivalent influences (Kundaje et al., 2012), and indicate a hierarchy of TF-mediated positioning effects.

We next queried whether specific TFs were responsible for limiting the encroachment of immediately flanking nucleosomes on the core regulatory DNA region. We aligned all +1 nucleosomes in unidirectional promoters and calculated the density of transcription factor recognition sequences within a 2 kb window (Figure 4.16a–d). Consistent with the above findings (Figure 4.12a), TF recognition elements for diverse factors (CREB/ATF, NRF1, YY1 and NF-

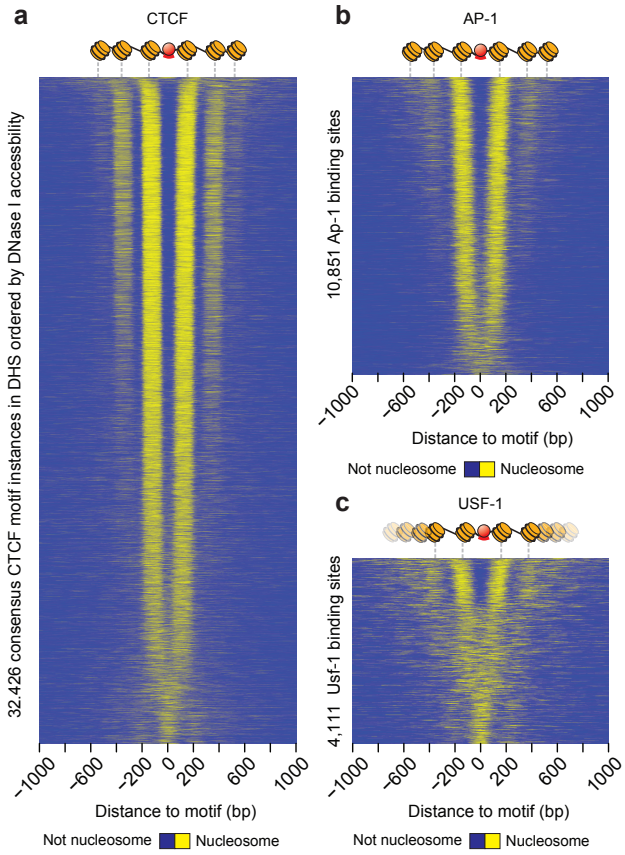

Figure 4.15: **TF occupancy effects on nucleosome organization**. (a–c) Heatmaps visualizing the positioning of nucleosomes ±1 kb surrounding the predicted binding sites for (a) CTCF, (b) AP-1 and (c) USF-1. Rows indicate the location of nucleosomes identified by the density of large fragments and are ordered by decreasing DNase I cleavage with ±25-bp surround the TF recognition sequence. Yellow indicates that a nucleosome is present at that position, while blue indicates absence of a nucleosome.

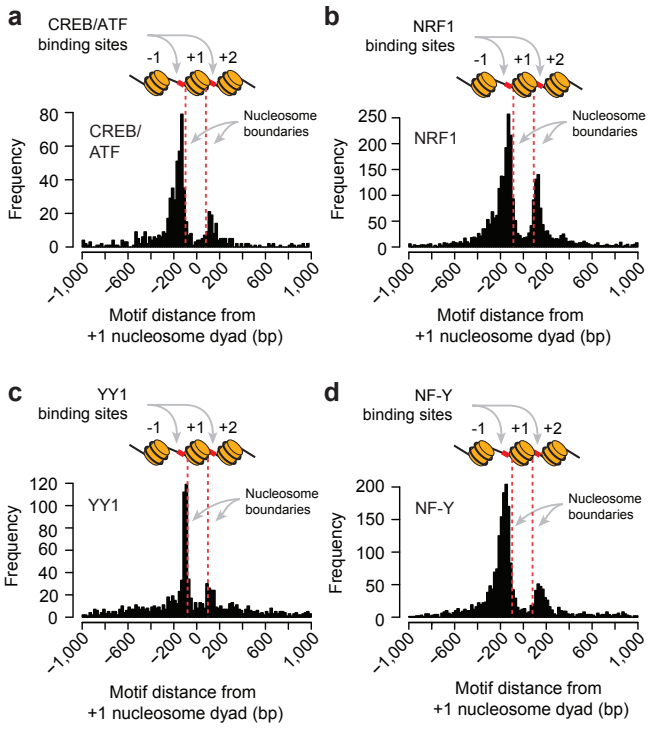

Figure 4.16: **Boundary TFs demarcate the regulatory DNA-nucleosome interface**. (a-d) A histogram of distances between +1 nucleosomes at unidirectional promoters and the predicted motif instances for CREB/ATF, YY1, NRF1, and NF-Y. Dotted red lines denote the boundaries of the +1 nucleosomes.

Y), were found immediately adjacent to the +1 nucleosome (Figure 4.16a-d). Notably, we also observed enrichment of motifs in the linker DNA between the +1 and +2 nucleosomes, signifying that nucleosome encroachment on regulatory DNA is limited by the occupancy of specific flanking TF binding elements.

4.2.5 Nucleosome organization and TSS selection

The lack of specific sequence motifs recognized by the RNA polymerase II preinitiation complex in promoters suggests an instrumental role for promoter nucleosome architecture in specifying transcriptional start sites (TSSs) (Rhee and Pugh, 2012; Yamashita et al., 2011). Figure 4.6a and Figure 4.17a–c illustrate usages of +1 nucleosomes that are consistent with TSSs of annotated gene models. To gain insight into the mechanism of TSS selection, we first scanned 18,454 DHSs

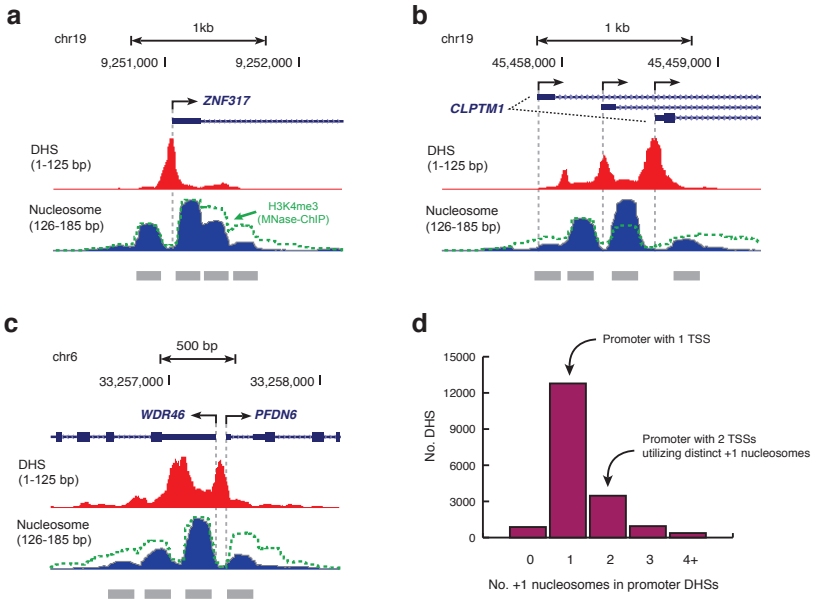

Figure 4.17: **Nucleosome positioning within promoter DHSs**. (a–c) Exemplary promoters demonstrating the relationship of the +1 nucleosome and the annotated transcription start site.

overlapping unidirectional promoters for the presence of a flanking +1 nucleosome. In total, we identified 24,234 nucleosomes associated with at least one TSS. While the vast majority of accessible promoters contained only one +1 nucleosome, a small subset contained two or more, a characteristic feature of alternative transcription start sites (Figure 4.17d).

To measure the broader effect of promoter architecture on transcriptional start site selection, we separated +1 nucleosomes (126- to 185 bp) associated with unidirectional transcription into quartiles on the basis of fragment density. For each quartile we computed aggregate average fragment densities of DHSs and nucleosomes (Figure 4.18), which revealed distinct chromatin structures for highly and lowly occupied +1 nucleosomes. We then compared the location of annotated TSSs relative to the +1 nucleosomes in these four occupancy quartiles (Figure 4.18a). We found that TSSs within promoters containing highly occupied +1 nucleosomes were narrowly focused at the +1 nucleosome boundary, while TSSs with promoters containing weakly occupied +1 nucleosomes were broadly dispersed (Figure 4.18a–b). These results confirm that the position of TSSs are tightly associated with +1 nucleosome occupancy (Mavrich et al., 2008;

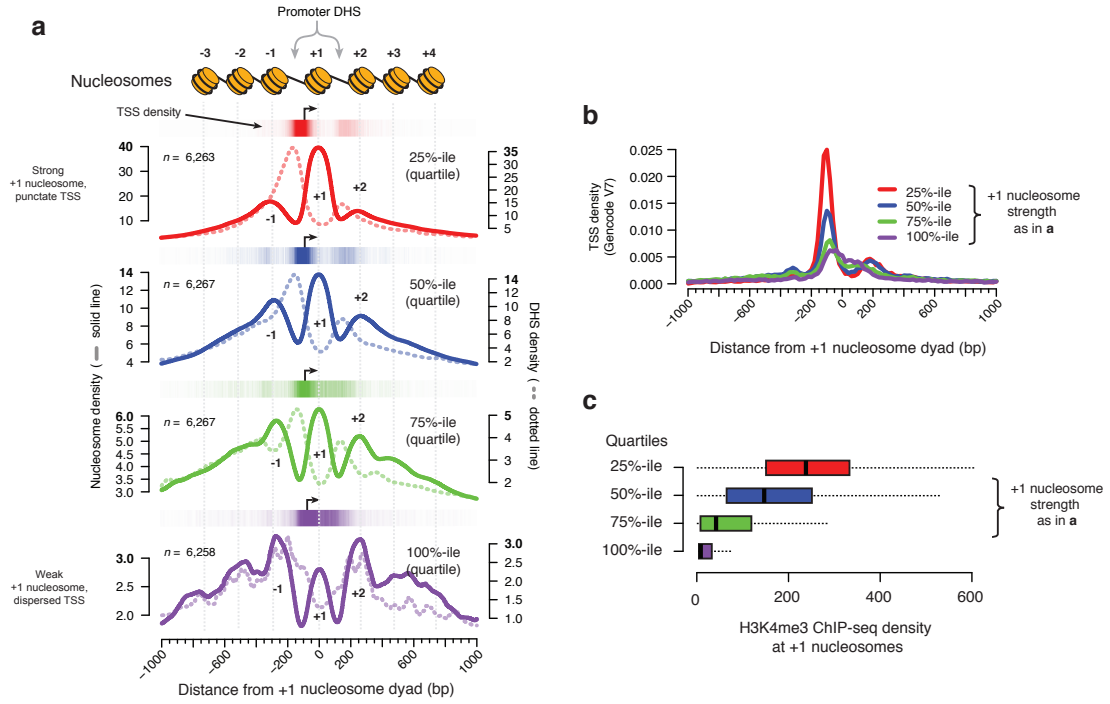

Figure 4.18: **Nucleosome organization and transcription start site selection**. (a) Aggregate DNase I fragment density profiles ±1 kb from +1 nucleosomes grouped into quartiles on the basis of signal intensity (red, 25% most accessible; purple, 25% least accessible). Dashed lines correspond to ≤125 bp fragments (DHSs), while solid lines correspond to 126- to 185 bp fragments (nucleosomes). Colored ticks above the DNase I profiles show the density of annotated TSSs relative to the +1 nucleosome. (b) TSS density. (c) H3K4me3 enrichment around the +1 nucleosomes in each quartile from (a) is correlated with DNase I accessibility (colors as in a, box represents the 25%- and 75%-ile and the dashed lines indicate 1.5x the interquartile range).

Schones et al., 2008; Zhang et al., 2009). Taken together with the strong association of TF recognition sequences at the +1 nucleosome boundary, these data suggest that the positioning of the +1 nucleosome by TFs may play a predominant role in TSS localization by recruitment of transcriptional machinery and remodelers that reinforce the promoter chromatin structure.

Trimethylation of lysine 4 on histone H3 (H3K4me3) is triggered by the RNA polymerase complex, marking an active promoter state and strong correlated with transcriptional activity (Barski et al., 2007; Bernstein et al., 2002; Thurman et al., 2012). To probe the connection between nucleosomal organization, TSS selection, and transcriptional activity, we measured the enrichment of H3K4me3 at each of the +1 nucleosome accessibility quartiles (Figure 4.18c). We observed a strong association between these two independent measurements such that highly occupied +1 nucleosomes were associated with increased H3K4 trimethylation.

4.3 Discussion

Activation of regulatory DNA by the binding of sequence-specific transcription factors in place of the canonical nucleosome is fundamental to eukaryotic gene regulation. Despite rapidly increasing data on TF and nucleosome occupancy in the human genome, TF-nucleosome interactions have been extremely difficult to study *in vivo* due to the lack of techniques to simultaneously measure their occupancy in a single experiment. Therefore, the ability to couple TF occupancy to nucleosome structure represents a fundamental advance over current and widely used methodologies used to map the structure of chromatin and regulatory DNA. DNase-FLASH represents a significant departure from MNase-based strategies (Henikoff et al., 2011), in which the detection of nucleosomes occurs irrespective of transcription factor binding. Accordingly, the TF-centric approach inherent to DNase-FLASH enables assessment of the direct contribution and influence of transcription factor occupancy on nucleosome positioning at individual loci.

DNase-FLASH is also a powerful method for connecting the fine-scale architecture of regulatory DNA to function. In this study, we mapped the chromatin structure surrounding human promoters, confirming the association of nucleosome positioning with transcription start site selection and activity (Mavrich et al., 2008; Rhee and Pugh, 2012; Schones et al., 2008; Zhang et al., 2009) and find strong evidence that +1 nucleosomes are positioned by the binding of sequence-specific transcription factors. Similarily, the approach we describe should be particularly useful for the analysis of chromatin dynamics, such as the actuation of regulatory elements

in response to developmental or conditional environmental cues.

4.4 Methods

4.4.1 Nuclei isolation and DNase I digestion

Adult gingival fibroblasts (AG09319) cells obtained from Coriell were grown to 90% confluency in MEM supplemented with 15% FBS (PAA) and 1× non-essential amino acids (Gibco) and detached with 0.5% Trypsin-EDTA (Gibco). The cells were washed once with growth media, pelleted by centrifugation (500*g* at 25°C), followed by a wash with cold 1× Dulbecco's PBS (Gibco). The following steps were performed on ice or at 4°C. 10 million cells were pelleted (500*g*) and resuspended in 2 mL of buffer A (15 mM Tris-HCl, 15 mM NaCl, 60 mM KCl, 0.5 mM spermidine, 1 mM EDTA, 0.5 mM EGTA, pH 8.0). The nuclei were extracted by the addition 2 mL of 2× extraction buffer (Buffer A supplemented with 0.5% IGEPAL-630CA), mixed by gentle inversion, and incubated on ice for ten minutes. Nuclei were immediately pelleted by centrifugation (500*g* for 5 minutes) and washed twice with cold buffer A. The nuclei pellet was incubated in at 37°C for 1 minute and then resuspended in 1 mL of prewarmed DNase I digestion buffer (15 mM Tris-HCl, 90 mM NaCl, 60 mM KCl, 6 mM $CaCl_2$, 0.5 mM spermidine, 1 mM EDTA, 0.5 mM EGTA, pH 8.0) and incubated at 37°C for 3 minutes. The digestion was stopped with the addition of 1 mL of stop buffer (50 mM Tris-HCl, 100 mM NaCl, 0.1% SDS, 100 mM EDTA, pH 8.0), 50 units of Proteinase K (Fermentas) and incubated at 55°C for 1 hour. Following proteinase digestion, RNA was degraded by the addition of RNaseA (30 minutes at 37°C). 10 mul of the digest was loaded onto a 1% agarose gel post-stained with SYBR Green I (Invitrogen) and visualized with a Typhoon 8600 scanner (GE Healthcare) to monitor digestion conditions.

4.4.2 Purification of DNase I fragments and sequencing

The digested DNA was purified by phenol-chloroform extraction and concentrated to 250 ul using a 10 kDa MWCO Centricon column (Milipore). The purified DNA was loaded on top of a 9% sucrose cushion (20 mM Tris-HCl, 1M NaCl, 5 mM EDTA) and spun for 24 hours at 25,000 rpm at 20°C. 250 μl fractions were extracted from the gradient using a Biomek NX liquid handling robot (Beckman). 10 μl of each fraction was loaded onto a 1.5% agarose gel and visualized by SYBR Green I (Invitrogen) post-staining using a Typhoon 8600 scanner (GE Healthcare). Fractions containing fragments <1 kb were pooled and purified over a MiniElute column (Qiagen). The small fragments were end-repaired, A-tailed and ligated with adapters

compatible with the Illumina sequencing platform. Fragments containing adapters were enriched with 8 cycles of PCR. The enrichment PCR reaction was loaded onto a 2% agarose gel and two bands were excised (100-300 bp and 300-500 bp) and purified using the QIAquick gel extraction kit (Qiagen). Each DNase I library was end-sequenced independently on two "lanes" in pair-end mode (2×36 bp) with the Illumina HiSeq2000 (Illumina).

4.4.3 In silico separation of fragments by length

Adapters were trimmed from reads (for fragments <36 bp) and mapped to the UCSC human genome version 37 (hg19) using bwa (Li and Durbin, 2009) version 0.5.8 (r1442), with default parameters. The mapped reads were assembled into paired reads with the following parameters: `-a 750`. The pair reads were filtered for mapping quality ($Q > 0$) and orientation. Reads were discarded that mapped to the mitochondrial genome. To separate DHS from nucleosomes, the mapped fragments were separated on the basis of their length into two bins: (1) 1-125 bp and (2) 126-185 bp. Smoothed densities were generated by counting the number of fragments in each respective length class in non-overlapping 5 bp windows across the genome. Per-nucleotide cut-counts were generated by summing the quantity of tags whose 5' end mapped to a particular genome coordinate.

4.4.4 Hotspots and DHS peaks

To identify regions hypersensitive to DNase I cleavage, we applied the hotspot algorithm2 (available for download at: `http://www.uwencode.org/proj/hotspot`). Hotspot is a program for identifying regions of local enrichment of short-read sequence tags mapped to the genome using a binomial distribution model. Regions flagged by the algorithm are called "hotspots." Local high-density peaks (DHS) are detected using a greedy-hill climbing algorithm that finds local maxima within "hotspots". DHS peaks are defined as 150 bp windows surrounding a local maximum.

4.4.5 Transcription factor binding site predictions

The TRANSFAC human database (version April 2011) (Wingender et al., 1996) was scanned genome-wide using FIMO (Grant et al., 2011) (MEME suite version 3.6) using the following parameters: `--verbosity 1 --output-pthresh 1e-5 --text`. Motifs instances were retained with a p-value threshold of $< 10^{-}5$. Motif instances passing the p-value threshold were then

overlapped with DNase I hotspots and used for all analysis unless otherwise specified.

4.4.6 Nucleosome detection

The 126–185 bp fragment densities were smoothed by wavelet transformation using the software package wavelet (Neph et al., in prep; available for download at: http://faculty.washington.edu/dbp/WMTSA/NEPH/wavelets.html) with the following parameters: `--boundary reflected --level 4`. An ad-hoc strategy was used for peak finding based on the second-derivative of the smoothed data. We applied a local-hill finding algorithm to refine the peak detection. Each nucleosome was defined as ± 75 bp from the peak (total 150 bp).

4.4.7 Annotation of transcription start sites

GENCODE5 (version 7) was downloaded from the UCSC Genome Browser and used for all transcript annotations.

4.4.8 Identification of +1 nucleosomes

Identification of +1 nucleosomes followed a step-wise approach. First, we curated all DHS peaks overlapping an annotated TSS on the basis of stranded-ness. For simplicity, bidirectional promoters we removed from further analysis. We labeled a nucleosome as +1 if it overlapped a TSS or was the nearest nucleosome 3' to the TSS. We then aligned all promoters by the +1 nucleosome(s) corrected for transcript directionality.

4.4.9 Native chromatin-immunoprecipitation of modified histones

AG09319 cells were propagated to 70-80% confluence, and detached from the flasks by 0.05% Trypsin-EDTA (Invitrogen). The cells were resuspended in RSB buffer (10 mM Tris-HCl pH 7.5, 10 mM NaCl, 3 mM $MgCl_2$, 0.5 mM spermidine) and the cell membranes were lysed on ice for 10 min with addition of NP-40 (Roche) to a final concentration of 0.02%. The nuclei were collected by centrifugation at 300*g* for 5 min at 4°C and washed once with RSB buffer. Aliquots of 2×10^7 nuclei were resuspended in 200 μl MN buffer (10 mM Tris-HCl pH 7.5, 10 mM NaCl, 3 mM $MgCl_2$, 1 mM $CaCl_2$) supplemented with protease inhibitor cocktail (Roche) and digested by 17 units of micrococcal nuclease (Worthington, NJ) for 10 min at 37°C. The digestion was stopped by adding 80 μl MNase stop buffer (500 mM NaCl, 50 mM EDTA pH8.0).

The supernatants (S1) were collected by centrifugation at 5000 rpm for 3 min at 4°C. The insoluble pellets were resuspended in 200 μl MNase buffer supplemented with 80 μl MNase stop buffer and sonicated by sonication (Bioruptor, Diagenode Inc., NJ) for 10 sec (setting 'H') at 4°C. The supernatants (S2) were collected by centrifugation at top speed for 3 minutes at 4°C, combined with S1, and diluted 5-fold with Dilution buffer (10 mM Tris-HCl pH 7.5, 150 mM NaCl, 1.5 mM EDTA, pH 8.0, 0.5 mM spermidine). For chromatin immunoprecipitation, the antibodies (Active motif #39113 for H2A.Z, Cell Signaling #9751 for tri-methyl-histone H3 lysine 4) were conjugated to Dynabeads M-280 sheep anti-rabbit IgG (Life Technologies) in 1 ml 1$\times$ PBS solution for at least 6 hours at 4°C, and incubated with micrococcal nuclease-digested, diluted chromatin at 4°C overnight. The complexes were washed 4 times with elution buffer (50 mM Tris-HCl, pH 7.5, 10 mM EDTA, 5 mM sodium butyrate, 150 mM NaCl), twice with TE buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA), and then washed in incubation buffer (10 mM Tris-HCl pH 8.0, 0.3 M NaCl, 5 mM EDTA pH 8.0, 0.5% SDS) briefly. The supernatants were recovered from the beads and treated with proteinase K at 55°C for 4 hours. The DNA was purified by phenol-chloroform extraction and ethanol precipitation. The small fragments were end-repaired, A-tailed and ligated with adapters compatible with the Illumina sequencing platform and sequenced in pair-end mode (2 $\times$ 36 bp) using the Illumina HiSeq2000 (Illumina). Sequencing reads were mapped using the same methods as the DNase I tags. Fragment densities were computed by counting the number of mapped fragments mapping to non-overlapping 5 bp windows covering the entire genome.

Acknowledgements

J.V. is supported by a US National Science Foundation Graduate Research Fellowship under grant DGE-071824. This work was supported by US National Institutes of Health NHGRI grants U54HG004592 and U54HG007010 to J.A.S.

Chapter 5

DNASE I FOOTPRINTS REFLECT TRANSCRIPTION FACTOR OCCUPANCY, NOT SEQUENCE BIAS

This chapter has been adapted with minor changes from: Vierstra, J. and Stamatoyannopoulos, J. A. DNase I footprints reflect transcription factor occupancy, not sequence bias. *Nat. Meth.* In review (2014).

Abstract

This work was produced as a technical comment to He, H. H. et al. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat. Meth.* 11, 73–78 (2014). The article by He et al. challenged the utility to digital genomic footprinting as a reliable method to identify and quantify transcription factor occupancy within the human genome. This chapter outlines the techinal and interpretive errors found within this article and puts the authors primary conclusions into question.

5.1 Introduction

Genomic DNase I footprinting has emerged as a powerful approach for global analysis of the in vivo occupancy patterns of sequence-specific DNA binding proteins (Hesselberth et al., 2009; Mercer et al., 2011; Neph et al., 2012b; Stergachis et al., 2013a). In a recent report in *Nature Methods*, He et al. claim to have: (a) produced a "refined" protocol for "obtaining high-quality, reproducible DNase-seq data"; (b) analyzed "the use of DNase-seq footprinting data to discover transcription factor (TF) binding sites at nucleotide resolution"; (c) found footprinting data to be "uninformative" concerning the occupancy of many or even most TFs; and (d) discovered an "intrinsic bias in transcription factor footprint identification" owing to DNase I sequence preferences.

Here we scrutinize these claims, and highlight a cascade of interpretive and technical errors that collectively vitiate the major claims of He et al. As we show below: (i) the "refined" DNase-seq procedure of He et al. in fact depletes precisely the DNA fragment population deriving from the TF-bound core regions of DHSs, resulting in strikingly low data quality compared with public DNase footprinting data; (ii) the "intrinsic bias" He et al. claim to have discovered is in fact an artifact resulting from failure to distinguish occupied from non-occupied TF recognition sites; (iii) despite claims to the contrary, He et al. do not identify or analyze any actual DNase I footprints; and (iv) the analyses of He et al. run counter to evolutionary selection measures.

5.2 Results

Refinement typically refers to a process whereby a crude substance is cleansed of its impurities. Within the core of DNase I hypersensitive sites (DHSs), TF-occupied DNA is released by the enzyme into a population of small (<100bp) DNA fragments by opposing single-stranded nicks (Vierstra et al., 2014). In their Figure 3a, He et al. plot the size distribution of DNase I-released DNA fragments obtained from their experiments. Juxtaposition of this figure with its exact counterpart found in Figure S13 from the simultaneously published paper from Vierstra et al. shows that the data of He et al. are almost completely lacking the vital population of small DNase I-released fragments deriving from the core TF binding regions of DHSs (Figure 5.1a). He et al. observe that their "fragment size distribution was dominated by a periodicity of 10.4 bp consistent with one turn of the DNA helix". However, it has long been known that this periodic cleavage pattern only emerges when a DNA molecule subjected to nuclease attack is directly apposed to a surface such as glass (Rhodes and Klug, 1980) or wound around a nucle-

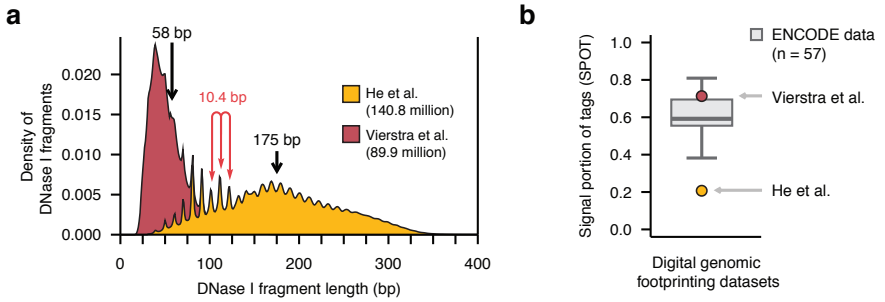

Figure 5.1: **Effects of data quality and TF occupancy on digital genomic footprinting**. (a) Density histograms of the fragment lengths in DNase I libraries from He et al. and Vierstra *et al.* Arrows indicate the means and the red arrows highly the 10.4 bp periodicity. (b) Signal portion of tags (SPOT) scores from ENCODE digital genomic footprinting datasets.

osome (Lutter, 1978). It therefore appears that the protocol reported by He et al. has indeed refined their DNase-seq data - by expunging precisely the fragment population deriving from the core non-nucleosomal, TF-occupied regions of DHSs.

Given the above, it is not surprising that the data of He et al. fall well below the quality standard of published genomic DNase I footprinting data. The hallmark of high-quality DNase-seq data is high signal-to-noise ratio, which can be computed simply as the proportion of mapped cleavages localizing within statistically significant peaks in tag density (DNase 'hotspots'). This basic metric – the 'signal' proportion of tags – is routinely computed for all DNase-seq data sets produced by major consortia such as the Roadmap Epigenomics Project Bernstein et al. (2010) and the ENCODE Project ENCODE Project Consortium et al. (2012); Thurman et al. (2012). Figure 5.1b compares the signal proportion of tags computed for He et al. with those of 57 ENCODE DNase I footprinting data sets (Neph et al., 2012b). In our experience, data such as those of He et al. – in which 80% of mapped DNase I cleavages fall outside of DNase hotspots – generally fail to yield accurately discernable DNase I footprints with any frequency, and are thus routinely rejected at an early stage in the quality control process (Thurman et al., 2012).

Recently, Lazarovici et al. (2013) extensively described the intrinsic sequence recognition and

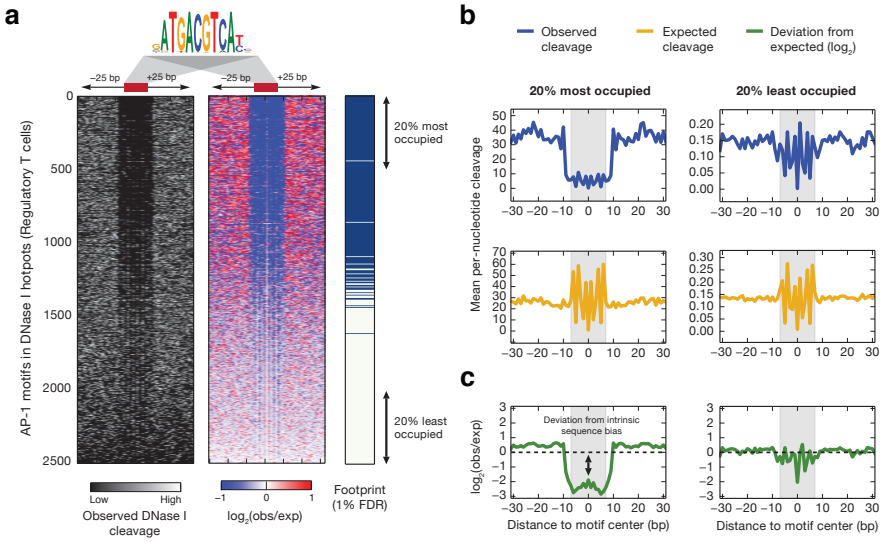

Figure 5.2: **Observed and expected DNase I cleavage profiles at AP-1 recognition elements**. (a) Heatmaps of per-nucleotide DNase I cleavages and discovered footprints surrounding JunD (AP-1) recognition sequences in regulatory T cells sorted by total cleavage density. Left, observed cleavages normalized by row. Right, the ratio of the observed cleavages to expected cleavages computed by reassigning tags to a hexamer model DNase I cleavage bias. Blue ticks indicate that the recognition sequence has an associated DNase I footprint. (b) Aggregate profile of mean per-nucleotide DNase I cleavages at the 20% most (left column) and 20% least (right column) accessible recognition sequences. Top row, observed cleavages. Middle, expected cleavages computed using the hexamer model. (c) the log_2 ratio of observed to expected.

cleavage preferences of DNase I through analysis of over 300 million nuclease cleavage events on naked DNA (Lazarovici et al., 2013). However, He et al. elected to sidestep these data and instead derive DNase I cleavage preferences from their in nucleo data which, as noted above, derive chiefly from nucleosome-associated DNA. They claim that their model is equivalent to that of Lazarovici et al. (their Figure 5c). However, this figure is misleading because the axes are discordant, and the details of how this calculation was performed are not available. Re-extracting the data of He et al. and comparing with those of Lazarovici et al. reveals that for nearly 40% of hexamers recognized by DNase I, the corresponding cleavage rates differ by more than 2-fold between the naked DNA- and in nucleo-derived models (data not shown).

Based on their model, He et al. claim that DNase I "footprints" in fact reflect intrinsic nu-

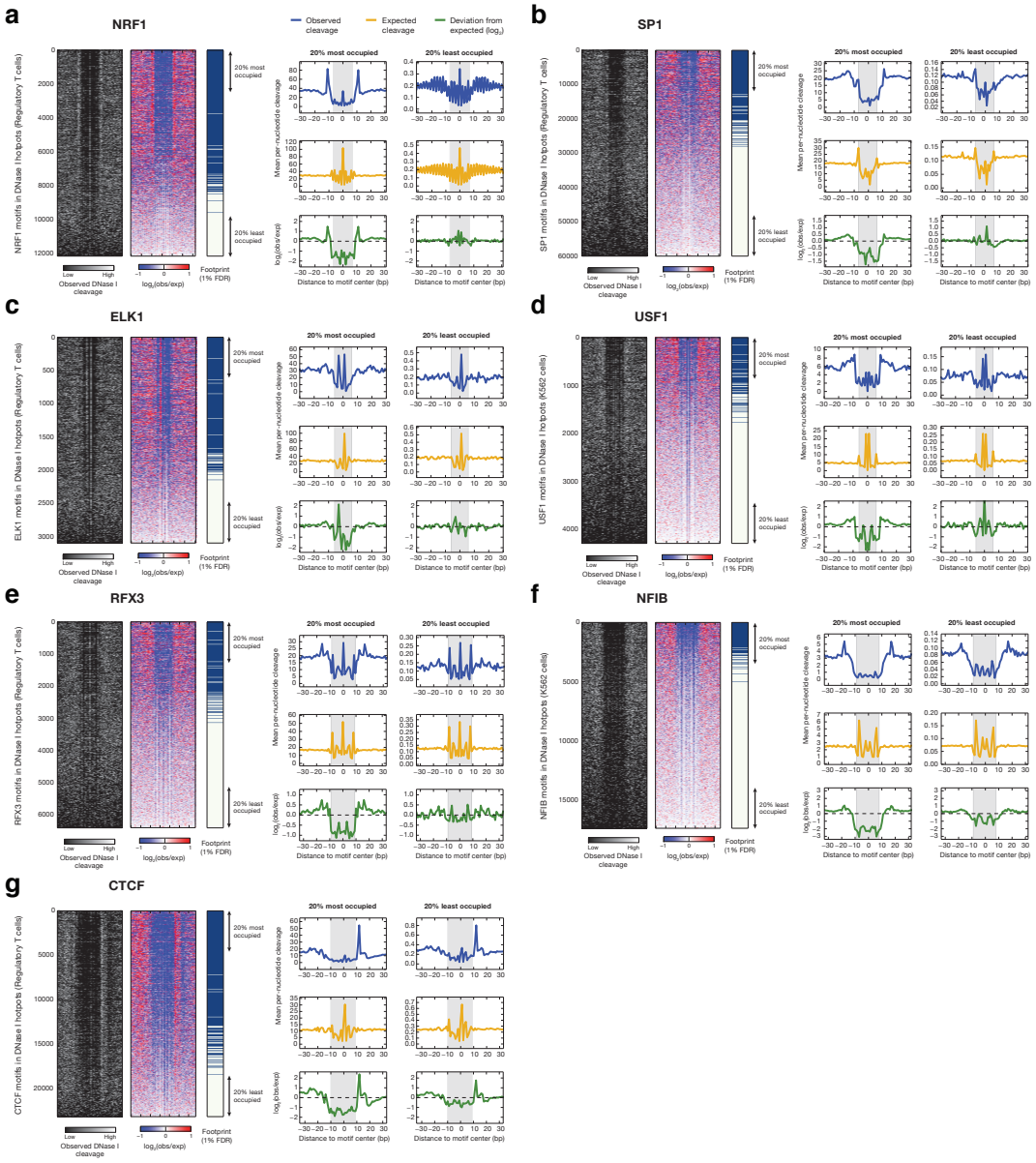

Figure 5.3: **Effects of protein occupancy and sequence context on DNase I cleavage profiles**. (a) Heatmaps of per-nucleotide DNase I cleavages and discovered footprints surrounding NRF1 recognition sequences in regualtory T cells. Left, observed cleavages. Right, the ratio of the observed cleavages to expected cleavages computed by reassigning tags to a hexamer model DNase I cleavage bias. Blue ticks indicate that the recognition sequence has an associated DNase I footprint. Line plots show the aggregate profile of mean per-nucleotide DNase I cleavages at the 20% most (left column) and 20% least (right column) accessible NRF1 recognition sequences. Top row, observed cleavages. Middle, expected cleavages computed using the hexamer model. Bottom, the log_2 ratio of observed to expected. (b-g) The same as (a) for the recognition sequences for SP1, ELK1, USF1, RFX3, NFIB, and CTCF within accessible chromatin.

cleotide cleavage preferences. Before assessing the specific merits of this claim, it is necessary to clarify the salient terminology. Since their discovery in 1978 by Galas and Schmitz (see Galas, 2001, for personal account), DNase I footprints have been uniformly understood to represent specific short polynucleotide segments wherein the cleavage pattern resulting from nuclease attack has been attenuated by the occupancy of a DNA-binding protein. By contrast, aggregated DNase I cleavage plots (which were first described by Hesselberth et al., 2009) represent averaged per-nucleotide DNase I cleavage across hundreds to thousands of instances of a given TF recognition sequence within DHSs genome-wide. Notably, only a fraction of such instances are in fact occupied *in vivo*, with the exact proportion varying with the TF and the cellular context. As such, aggregated DNase I cleavage plots incorporate a mixture of both occupied and unoccupied templates. The result is that for a highly occupying factor, the aggregated cleavage plot represents mainly the cleavage pattern of occupied templates, but for a lowly occupying factor the plot represents mainly the intrinsic nuclease cleavages.

The claim of He et al. that the observed DNase I cleavage pattern associated with most TF recognition sequences does not differ substantially from the intrinsic cleavage preferences predicted by their model is based on the observation (their Figure 6b) that Pearson's correlation values computed on observed vs. modeled cleavage data are typically high. It is, of course, well known that high correlation may be observed between two fixed data series that show parallel directional trends, even though the absolute differences between the series may be large. In the present case, this problem is confounded further by the co-mingling of occupied and unoccupied templates.

For example, He et al. claim that the "footprint" (read: aggregated DNase I cleavage pattern) of JunD (part of the canonical AP-1 complex) has a Pearson correlation of 0.8 with the modeled DNase I cleavage pattern. However, this calculation takes no account of recognition sequence occupancy. Figure 5.2 shows the true relationship between measured and predicted per-nucleotide DNase I cleavage patterns at individual JunD motifs within DHSs, and whether a canonical TF footprint is detected over the motif. Figure 5.2b shows measured (blue) and predicted (yellow) aggregate averaged DNase I cleavage plots for the most (top 20%) and least (bottom 20%) occupied sites, and the absolute differences between the measured and predicted values are shown in Figure 5.2c. Figure 5.3 shows analogous data for additional TFs reported by He et al. to show high correlation between measured and modeled DNase I cleavage patterns; similar results can be obtained for most TFs. Figure 5.4a that the DNase I cleavage patterns of occupied vs. unoccupied templates differ substantially.

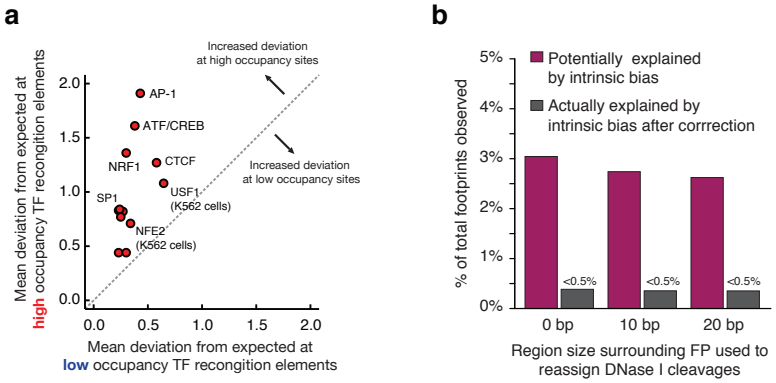

Figure 5.4: **TF occupancy is a hallmark of DNase I footprints**. (a) Comparison of observed mean per-nucleotide deviations from expected cleavage patterns at low (x-axis) and high (y-axis) occupancy TF recognition elements. Mean per-nucleotide deviation was computed from a region spanning ±5-bp of the recognition sequence. (b) Proportion of estimated false-positive footprints due to sequence bias alone. Purple, K562 footprints that reach the 1% FDR threshold after cleavage reassignment using a hexamer model. Grey, K562 footprints that score as well or better than observed footprints after cleavage reassignment.

He et al. claim to have revealed an "intrinsic bias in transcription factor footprint identification." This claim is particularly surprising given that He et al. never performed TF footprint detection. The claim also directly contradicts a prior result on this topic (Neph et al., 2012b). To quantify the degree to which intrinsic DNase I cleavage preferences might compromise detection of DNase I footprints, we reassigned observed DNA cleavages according to the intrinsic sequence preference of DNase I, and then re-computed the footprint discovery score for each site (Figure 5.4b). From this analysis it is obvious that intrinsic sequence preference has a negligible effect on footprint discovery, possibly accounting for well under 1% of detected footprints. Intuitively this makes sense, since footprints are typically 6–20 bp in extent, and generation of false-positive footprints would require a series of 1 bp-shifted mutually overlapping hexamers, each with similarly low intrinsic cleavage preferences.

Another notable absence from He et al. is any consideration of nucleotide-level evolutionary conservation, which has been shown to run antiparallel with average per nucleotide DNase I cleavage for a given TF, and to correlate strongly with TF occupancy at footprints (Neph et al., 2012b) (Figure 5.5 and 5.6). The relationship between DNase I footprints and per-nucleotide conservation is perhaps most obvious in the recent report from Stergachis et al. (2013a) de-

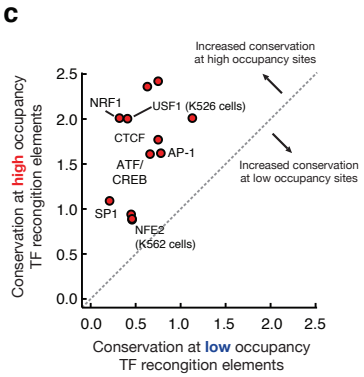

Figure 5.5: **High occupancy TF recognition elements are associated with increased sequence conservation**. (a) Mean per-nucleotide sequence conservation (phyloP 100-way alignment) at the 20% most (red) and 20% least (blue) accessible AP-1 recognition sequences. (b–h) The same as (a) for the recognition sequences of NRF1, SP1, ELK1, USF1, RFX3, NFIB, and CTCF within accessible chromatin.

scribing dense footprinting of the human exome, and the many major impacts of coding TF occupancy on base-level evolutionary events. It remains to be explained how these findings could be forthcoming if the claims of He et al. were valid.

In summary, we have shown that the results of He et al. derive from a combination of (i) poor data quality; (ii) confusion of DNase I footprints on the genome with aggregated DNase I cleavage plots; and, most importantly, (iii) failure to consider TF occupancy - which is the core objective of DNase I footprinting.

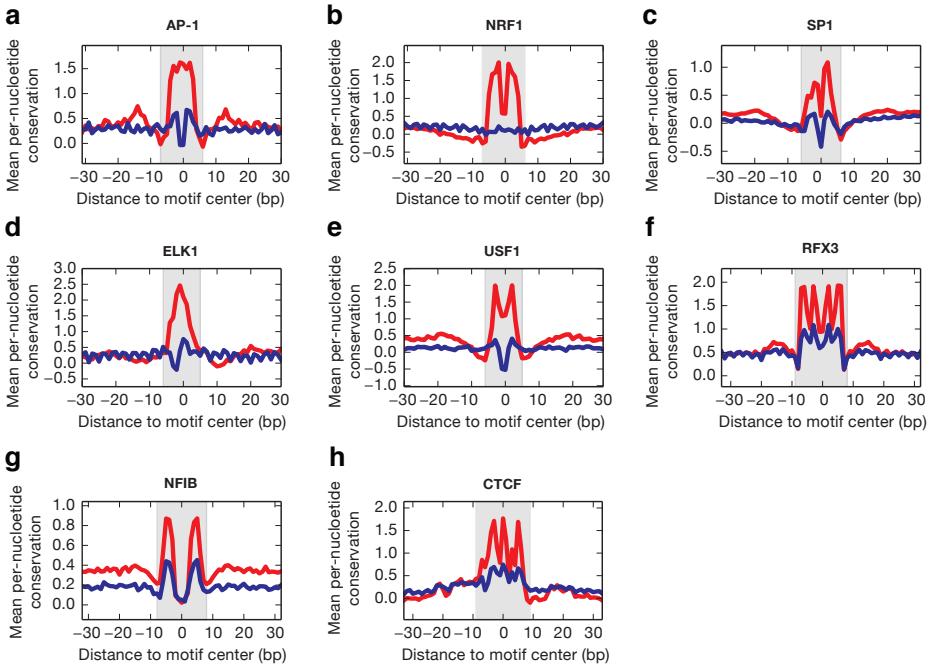

Figure 5.6: **Evolutionary selection on high occupancy binding sites**. Comparison of the per-nucleotide conservation at low (x-axis) and high (y-axis) occupancy TF recognition elements.

5.3 Methods

5.3.1 Evaluation of data quality

To identify DNase I hypersensitive regions, we applied the hotspot algorithm (John et al., 2011) (available at http://www.uwencode.org/proj/hotspot). Hotspot is a program for identifying regions of local enrichment of short-read sequence tags mapped to the genome using a binomial distribution model. Regions flagged by the algorithm are called 'hotspots'. Both the He et al. and Vierstra et al. DNase I datasets were evaluated in the context of 57 published ENCODE datasets using the signal portion of tags (SPOT) metric. SPOT scores reflect the amount total cleavages observed that occurred within hotspots, such that low SPOT scores indicate increased rates of background cleavages. SPOT scores were calculated from 5 million sequencing tags.

5.3.2 Transcription factor binding site predictions

TF recognition sequence models were obtain from Jolma et al. (2013) and scanned genome-wide using FIMO (MEME suite 3.6) (Grant et al., 2011) using the following parameters: "`--verbosity 1 --ouptut-pthresh 1e5 --text`". TF recognition sequences were then overlapped with DNase I FDR 1% hotspots and used for all analyses.

5.3.3 Modeling intrinsic cleavage specificity

DNase I sequence bias was estimated from digestion of naked DNA derived from human fetal lung fibroblast cells (IMR90) (Lazarovici et al., 2013). For each nucleotide j within a genomic window $[i, l]$ the normalized expected cleavage rate is:

$$p_j = \frac{a_j}{\sum_{k=i}^{l} a_k} \tag{5.1}$$

We define a_k as the relative cleavage bias of the 6-mer spanning the positions $[k-3, k+2]$ as described in Lazarovici et al. We redistributed the total observed cleavages

$$N_{i,l} = \sum_{k=i}^{l} n_k \tag{5.2}$$

where n_k is the observed cleavages at base k to determine the expected cleavages:

$$n_j^{'} = N_{i,l} \times p_j. \tag{5.3}$$

The per-nucleotide deviation (d) from intrinsic sequence specificity was defined as

$$d = log_2(\frac{n_j}{n_j^{'}}) \tag{5.4}$$

The sequence bias normalization was computed separately for each strand and then recombined for visualization purposes.

5.3.4 Conservation analysis

The per-nucleotide phyloP (Siepel et al., 2005) 100-way conservation track was downloaded from the UCSC Genome Browser (Meyer et al., 2013).

Part II

EVOLUTION OF *CIS*-REGULATORY DNA

Chapter 6

MOUSE REGULATORY DNA LANDSCAPES REVEAL GLOBAL PRINCIPLES OF *CIS*-REGULATORY EVOLUTION

This chapter has been adapted with minor changes from: Vierstra, J. Rynes, E., Sandstrom R., Zhang, M. et al. Mouse regulatory DNA landscapes reveal global principles of *cis*-regulatory evolution. *Science.* In press (2014).

Abstract

To study the evolutionary dynamics of regulatory DNA, we mapped >1.3 million DNase I hypersensitive sites (DHSs) in 45 mouse cell types and primary tissues, and systematically compared these with human DHS maps from orthologous cell and tissue compartments. Since their last common ancestor, the mouse and human genomes have undergone extensive *cis*-regulatory rewiring that combines branch-specific evolutionary innovation and loss with widespread functional repurposing (cell type switching) of DHSs mediated by acquisition of novel recognition sites for cell fate regulators. Strikingly, in spite of pervasive evolutionary remodeling of the locations and content of individual *cis*-regulatory regions, we find that the aggregate recognition sequence space for each transcription factor within accessible regulatory DNA of orthologous mouse and human cell types has been strictly conserved. Our findings provide new insights into the evolutionary forces shaping mammalian regulatory DNA landscapes.

6.1 Introduction

The laboratory mouse Mus musculus is the major model organism for mammalian biology and has provided extensive insights into human developmental and disease processes (Hardouin and Nagy, 2000), At 2.7 Gb, the mouse genome is of comparable size, structure, and sequence composition with the 3.3 Gb human genome (Lander et al., 2001; Mouse Genome Sequencing Consortium et al., 2002), and >80% of mouse genes have clear human orthologs (Guénet, 2005; Hardouin and Nagy, 2000). The mouse has also played a major role in analysis of the regulation of human genes via human-to-mouse transgenic experiments, which have collectively demonstrated that the mouse is capable of recapitulating salient features of human gene regulation, often with striking precision, even in the case of human genes that lack mouse orthologs (Peterson et al., 1993; Wilson et al., 2008). By contrast, comparative analyses of regulatory regions governing individual gene systems (Dermitzakis and Clark, 2002), as well as the occupancy patterns of several transcription factors (Odom et al., 2007; Schmidt et al., 2010; Stefflova et al., 2013), have highlighted the potential for *cis*-regulatory divergence. However, broader efforts to identify and quantify the major forces shaping the evolution of the mammalian *cis*-regulatory landscape have been hampered by the lack of expansive and highly detailed regulatory DNA maps that can be directly compared between mouse and human.

DNase I hypersensitive sites (DHSs) are highly sensitive markers of all major classes of *cis*-regulatory elements, and systematic genome-wide mapping of human DHSs in diverse cell and tissue contexts has provided fundamental insights into basic principles of transcriptional regulation (Thurman et al., 2012), development and differentiation (Stergachis et al., 2013b), and the genetics of common diseases and traits (Maurano et al., 2012). To gain a comprehensive perspective on the evolutionary dynamics of mammalian regulatory DNA and its implications for understanding human transcriptional regulatory programs, we undertook comprehensive mapping of DHSs in diverse mouse cell and tissue types, and systematically compared the resulting maps DHSs mapped in orthologous human cells and tissues, as well as a broader catalogue encompassing over 200 additional cell types and states.

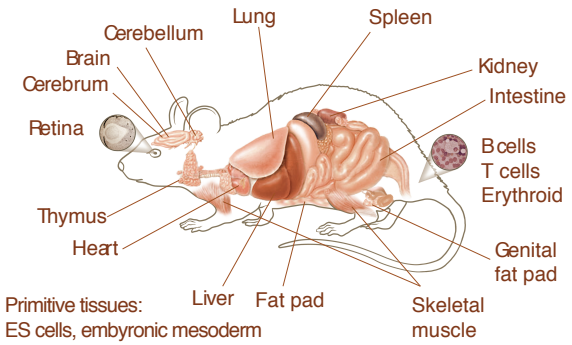

Figure 6.1: **Overview of mouse tissues used in this study**. (a) The accessible landscape of the mouse was derived from 45 tissues and cell types.

6.2 Results

6.2.1 The regulatory DNA landscape of the mouse genome

To define the mouse regulatory DNA landscape expansively, we mapped DHSs in 45 mouse cell and tissue types including adult primary tissues (n=19); purified adult and primitive primary cells (n=10); primary embryonic tissues (n=4); embryonic stem cell (ESC) lines (n=4); and model immortalized primary (n=3) and malignant cell lines (n=5) (Figure 6.1, figure 6.2 and table 6.1). We identified between 74,386 and 218,597 DHSs per cell type at a false discovery rate (FDR) threshold of 1% (John et al., 2011), and collectively delineated 1,334,703 distinct DHSs, each of which was detected in one or more mouse cell/tissue types. The genomic distribution of DHSs relative to annotated genes and transcripts was similar to that observed in the human genome (Thurman et al., 2012) (Figure 6.3a). On average, 13.5% of DHSs coincided with promoters, with the remaining 86.5% distributed across the intronic and intergenic compartments in roughly equal proportions. The vast majority of intergenic distal elements were located within 250 kb of the nearest annotated transcriptional start site (TSS) (Figure 6.3b). However, compared with the human genome, average intergenic DHS-to-TSS distances in the mouse genome were significantly compressed (Fig. S1D, median 48.7 kb vs. 91.6 kb for human), consistent with genetic loss in the murine lineage (Mouse Genome Sequencing Consortium et al., 2002). Notably, this compaction of mouse DHSs around TSSs is greater than expected from the ratio of genetic

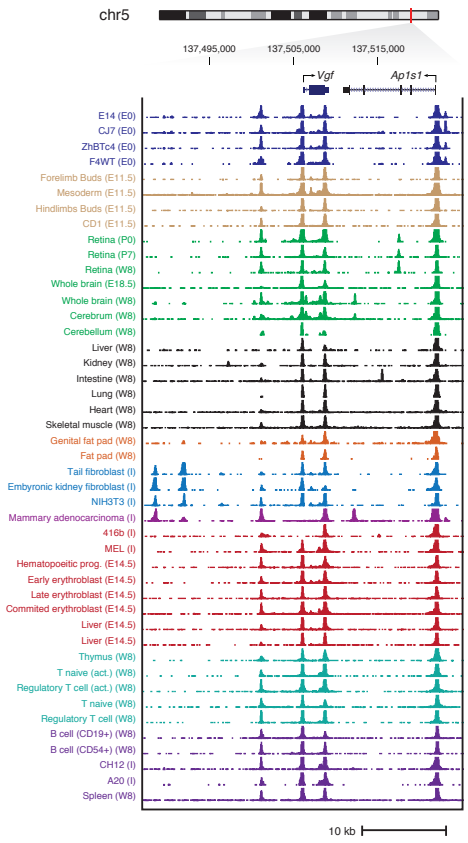

Figure 6.2: **Comprhensive mapping of the accessible regulatory landscape of the mouse genome**. Exemplar DNase I cleavage profile within the 35 kb surrounding the Vgf locus in 45 diverse mouse cell and tissue types.

material in mouse vs. human (2.7 Gb vs. 3.3 Gb) indicating differential rates of genome remodeling within mouse vs. human DHS-rich regions (Figure 6.3c). In fact, comparing the overall DHS landscape between mouse and human yielded revealed a marked difference in both size and density at distally positioned elements (Figure 6.4a–b).

6.2.2 Conservation and divergence of regulatory DNA

To gain insight into the evolution of mammalian regulatory DNA, we comprehensively integrated the mouse DHS maps with human maps generated using the same methods derived

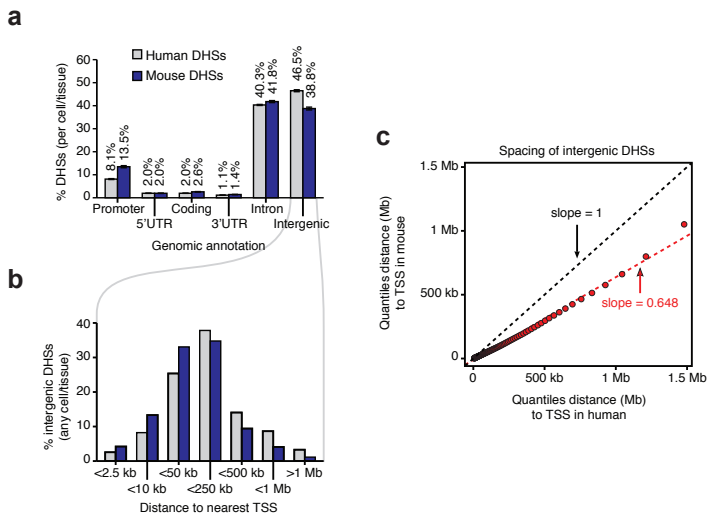

Figure 6.3: **General characteristics of the mouse *cis*-regulatory landscape**. (a) Distribution of mouse and human DHSs with respect to genome annotations (per cell- or tissue-type). Promoters are defined as 1 kb upstream of an annotated TSS. Bars indicate the mean and errors bars indicate the standard error of the mean. (b) Distribution of distances between all intergenic DHSs identified to the nearest annotated TSS in both human and mouse. (c) Q-Q plot of distances from intergenic DHSs to the nearest annotated TSS reveals a linear compression in regulatory DNA spacing. For clarity, the bottom and top 1% of the data points are not shown. Dashed black line, expected relationship between two identical distributions. Red line, linear regression of mouse vs. human distance quantiles.

from 232 cell/tissues types from the ENCODE Project (n=103) (Thurman et al., 2012) and the Roadmap Epigenomics Project (n=126) (Bernstein et al., 2010). These human maps collectively encompass 3 million distinct DHSs from primary cells, adult and fetal tissues, immortalized and malignant lines, and ESCs (Table 6.2). To identify DHSs shared between mouse and human, we projected the genomic sequence underlying all mouse and human DHSs to the other species using high-quality pairwise alignments and a conservative reciprocal mapping and filtering strategy (Figure 6.5a–b and figure 6.6). Using this strategy, collectively 59.5% of mouse DHS regions (range 52.5–78.8% per cell type) could be aligned with high confidence to the human genome, of which 35.6% (38.6–60% per cell type) coincided with a human DHS (Figure 6.5a and table 6.3). The remaining 23.9% (13–22.7% per cell type) of mouse DHSs aligning to the human genome outside of the currently defined human DHS compartment may correspond either with yet-to-be defined human DHSs, or with human lineage-specific extinction of an ancestral

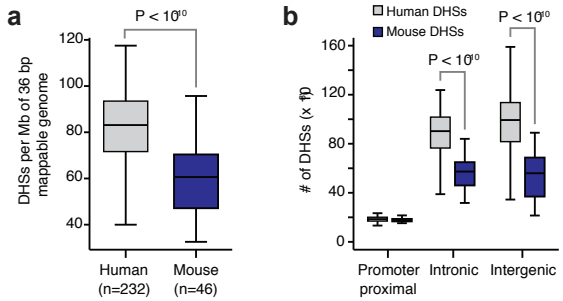

Figure 6.4: **Expansion of the human *cis*-regulatory landscape**. (a) Comparison of densities of the human and mouse DHS landscapes reveals a relative increase in human regulatory DNA ($P < 10^{-10}$, Wilcoxon rank-sum test). Densities were normalized by the size of the 36 bp alignable genome (see Methods). (b) Genomic distribution of DHSs in mouse and human by annotation. Mouse and human have nearly equivalent amounts of promoter proximal DHSs, while human tissues have significant increase in the quantity of distal regulatory elements ($P < 10^{-10}$, Wilcoxon rank-sum test).

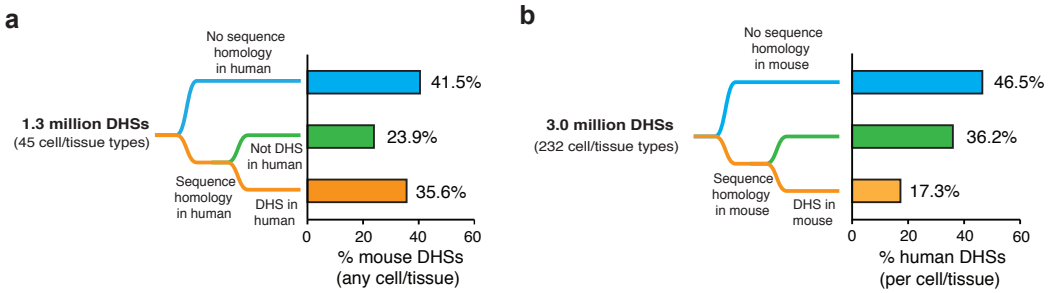

Figure 6.5: **Sequence and functional conservation of the mouse and human DHS landscape**. (a) Proportions of the mouse regulatory DNA landscape with sequence homology and functional conservation with human. (b) Same as (a) but for human DHSs.

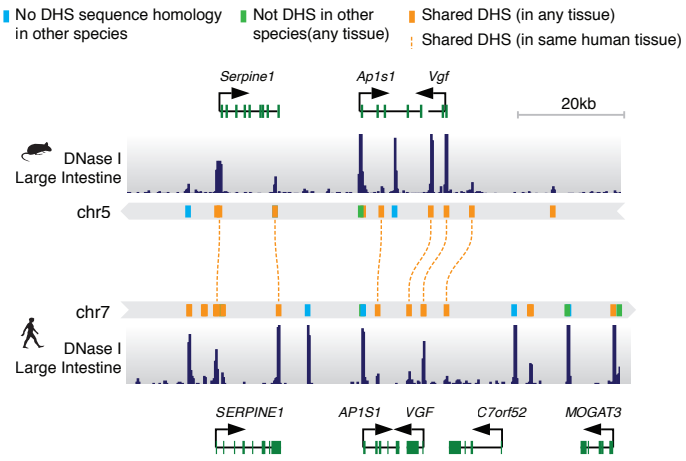

Figure 6.6: **Conservation of the *cis*-regulatory elements surrounding Vgf/VGF**. Example of the conservation of the *cis*-regulatory elements surrounding within the Vgf/VGF locus in mouse and human intestine

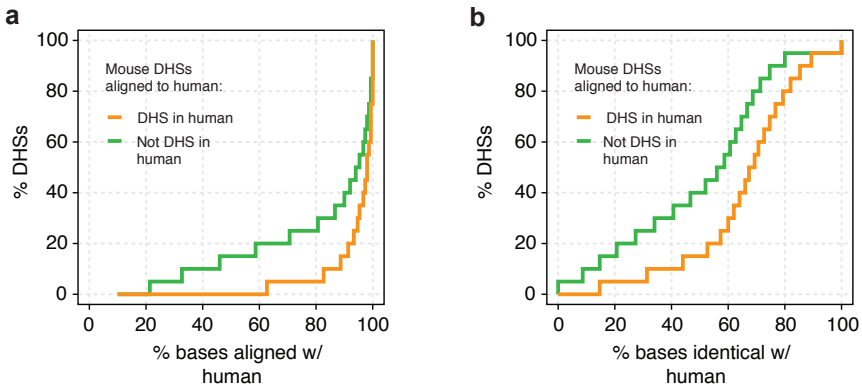

Figure 6.7: **Sequence alignment of mouse DHSs to the human genome**. (a) Cumulative fraction of mouse DHSs as a function of the proportion of bases that align to the human using the mm9 to hg19 "over" chain downloaded from the UCSC Genome Browser. The dashed grey line indicates the minimum alignment parameter used for cross-mapping DHSs (see Methods). (b) Cumulative fraction of mouse DHSs as a function of the proportion of aligned nucleotides identical to human.

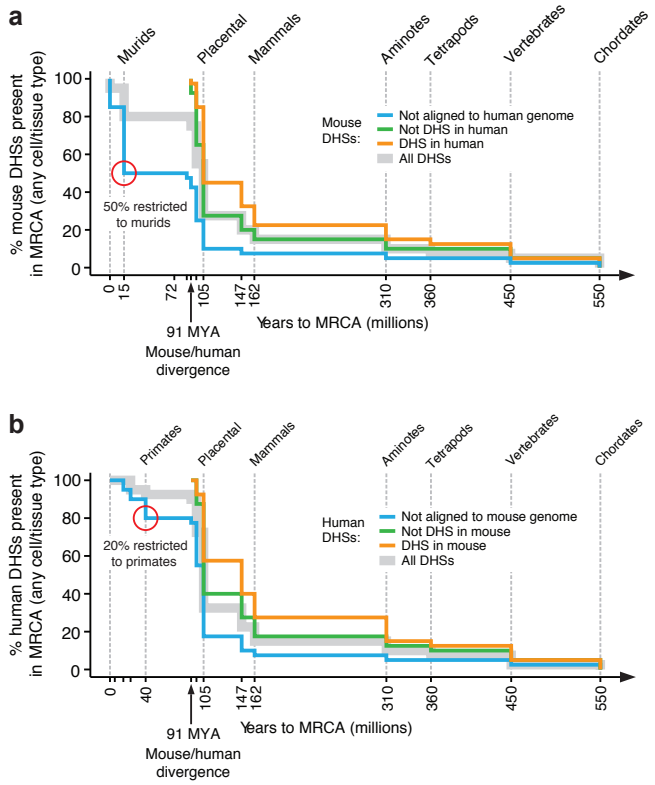

Figure 6.8: **Evolutionary conservation of mouse DHSs reveals pervasive turnover of individual *cis*-regulatory elements**. (a) Proportion of mouse DHSs with conserved sequence across the chordate phylogeny with respect to DHSs conservation with mouse (b) Same as (a) for human DHSs.

element. In support of the latter, mouse DHSs aligning outside of human DHSs show an excess of nucleotide sequence divergence evidenced by fewer alignable or identical nucleotides than mouse DHSs aligning to human DHSs (Figure 6.7a–b). A lower proportion of human DHSs align with a mouse DHS (17.3%, figure 6.5b and table 6.4); however, this is largely a reflection of the >2-fold greater number DHSs delineated in human vs. mouse. Given the breadth of mouse and human tissues analyzed, the above values suggest upper and lower limits of regulatory DNA conservation between mouse and human.

To trace the evolutionary origins and dynamics of individual regulatory regions, we aligned all mouse and human DHS sequences to >30 vertebrate genomes spanning roughly 550 million years of evolutionary distance (Figure 6.8a–b). These results indicate that the vast majority of

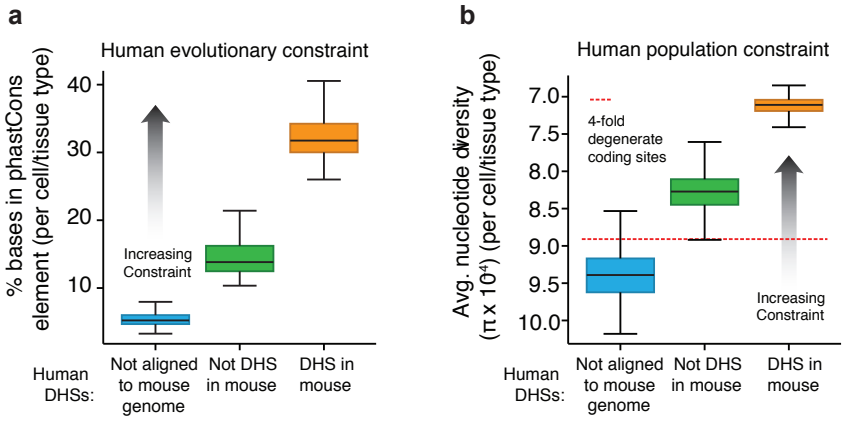

Figure 6.9: **Sequence constraint within regulatory DNA**. (a) Percentage of human DHSs containing a conserved phastCons element. Boxes indicate 25%- and 75%-iles for all human tissues. Whiskers denote 1.5 times the interquartile range. (b) Nucleotide diversity (π) with human DHSs estimated from the complete genomes of 53 unrelated individuals. Boxes indicate the 25%- and 75%-iles for all human tissues. Whiskers denote 1.5 times the interquartile range. Red dashed line indicates π at four-fold degenerate coding sites.

mouse and human regulatory DNA arose concomitant with the radiation of placental mammals. However, despite the deep sequence conservation of many DHSs, turnover of individual regulatory regions within different branches of the evolutionary tree appears frequently (Shibata et al., 2012). For example, of the 80% of mouse DHS sequences that predate the divergence of humans from a common ancestor, only 58.5% are detectable in human, and comparison of mouse DHSs aligning to a human DHS or to a non-DHS region yields nearly identical evolutionary profiles (Figure 6.8a–b). In general, the proportion of DHSs that encompass evolutionarily conserved sequence elements defined by phastCons (Pollard et al., 2010) increases with alignability and conservation of DNase I hypersensitivity (Figure 6.9a). Unexpectedly, however, 40% of mouse-human shared DHSs lack any phastCons conserved elements.

The aforementioned trends are also reflected in patterns of human variation. Analysis of nucleotide diversity (π) within DHSs revealed graded constraint depending on the extent of sequence and DHS conservation (Figure 6.9b). Notably, mean π within human-specific DHSs approximated that of four-fold synonymous sites within coding regions, compatible with relaxed (but not absent) nucleotide-level constraint on this compartment. While nucleotide-level constraint is often invoked as a proxy for function, we found that despite decreased constraint

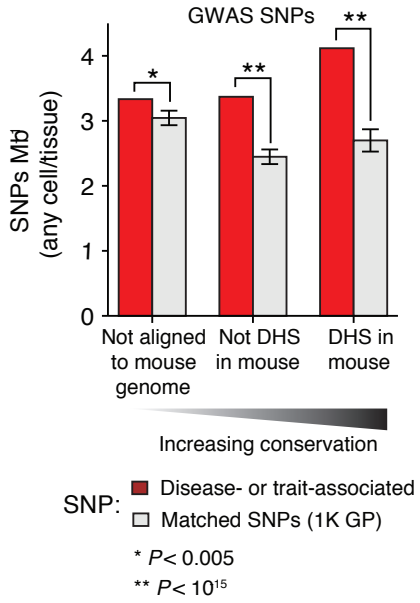

Figure 6.10: **Enrichment of genetic variation within DHSs**. Disease- and trait-associated variants are enriched within both conserved and species-specific regulatory elements. The observed variant density is compared against 1,000 sets of randomly matched SNPs (see Methods).

(both evolutionary and recent), human-specific DHSs are significantly enriched (vs. all DHSs) in disease- and trait-associated variants identified by genome-wide association studies (Figure 6.10, Permutation test, $P_{null} < 0.005$).

Collectively, the above results indicate that while mouse-human shared DHSs are collectively under selection over evolutionary timescales and within human populations, the sequence information with the *cis*-regulatory compartment is rapidly evolving in both mice and humans.

6.2.3 Preferential conservation of regulatory elements

Compared with the entire mouse DHS compartment, the overall density of shared DHSs was higher in gene-proximal regions such as promoters, exons and UTRs (Figure 6.11a). However, the relative proportion of shared DHSs (to all DHSs) increased markedly with distance from the TSS (Figure 6.11b and figure 6.12). From 0 to 50kb upstream of the TSS, the proportion of DHSs that are shared with human (avg. 27%) is lower than the average for intergenic regions

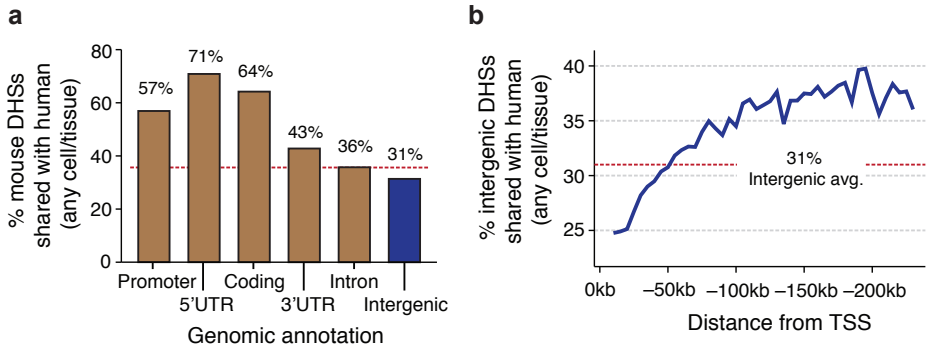

Figure 6.11: **Rapid evolution of proximal DHSs**. (a) Gene proximal DHSs are more likely to be conserved than distal DHSs. Dashed red line indicates the average conservation of DHSs. (b) The rate of intergenic DHS conservation vs. distance to nearest TSS indicates a rapidly evolving *cis*-regulatory domain.

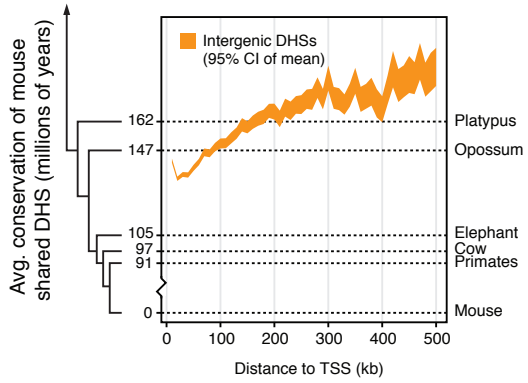

Figure 6.12: **Deep phylogenetic conservation of far distal elements**. The average conservation of intergenic shared DHSs increases with distance to the nearest TSS. The tree on the y-axis shows part of the mammalian phylogeny. Orange shading indicates the 95% percent confidence interval on the mean estimated by bootstrap analysis (500 replicates).

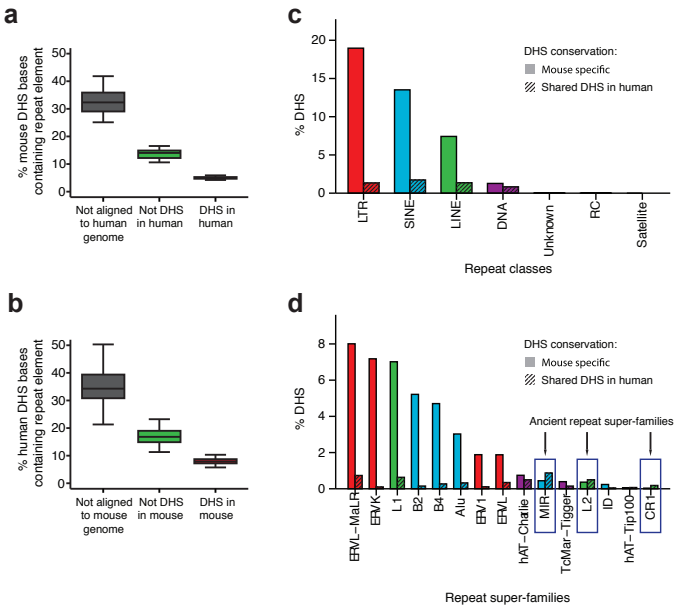

Figure 6.13: **Repeat associated innovation of *cis*regulatory DNA**. (a–b) Proportion of the mouse (a) and human (b) DHS landscapes that overlap a repetitive element identified by RepeatMasker. (c) Mouse specific regulatory DNA is enriched for all classes repetitive elements. (d) LTR endogenous retroelements, L1 and SINE elements constitute the majority of mouse specific regulatory DNA that has arisen via repetitive elements. Enrichment of ancient repeat-families (blue boxes; SINE/MIR, LINE/L2 and LINE/CR1) within conserved DHSs indicate that repetitive elements contributed to innovations in mammalian regulatory DNA long before the mouse and human divergence ≈91 MYA.

(avg. 31%, figure 6.11a), while in far distal regions this proportion increases substantially to a plateau of 38%. Whereas several cases of highly conserved regulatory elements functioning over hundreds of kilobases have come to light (Lettice et al., 2003; Montavon et al., 2011), the above data suggest that such elements comprise a genomic compartment that may be functionally distinct from a more rapidly evolving gene-proximal region, and less buffered against evolutionary alteration.

6.2.4 Innovation of species-specific regulatory DNA

Genesis of novel regulatory DNA sequences appears to have played a substantial role in shaping the DHS landscape in both mouse and human (Fig. 1B and fig. S2A). Over 50% of the mouse

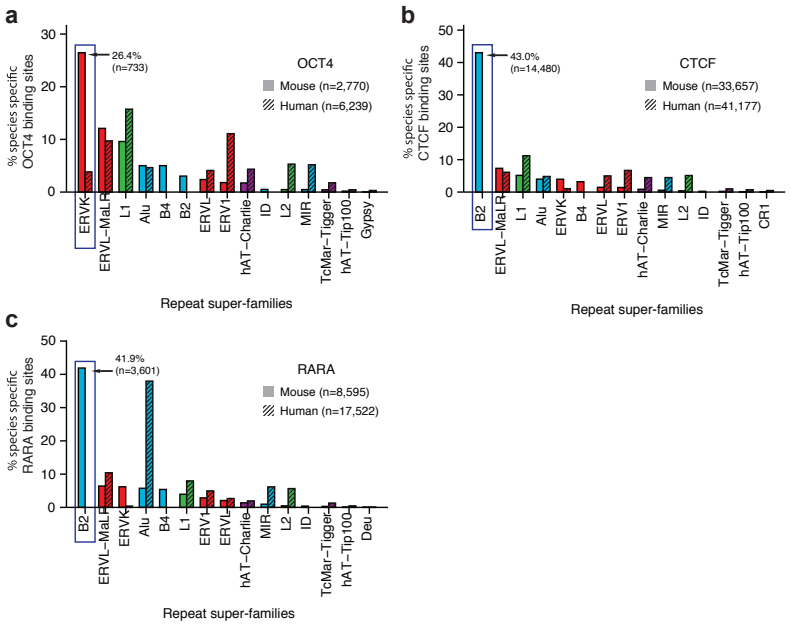

Figure 6.14: **Stereotypical expansion of TF binding sites on specific repetitive elements**. Examples of transcription factors for which lineage specific expansion of their putative binding elements has occurred by specific repeat super-families. (a) Many of the recognition sequences for the canonical pluripotency master-regulator OCT4 has arisen on LTR/ERVL retrotransposable elements. (b) Enrichment of mouse specific CTCF recognition sequences within rodent specific SINE/B2 elements confirms and extends previous reports (Bourque et al., 2008). (c) Recognition sequences for retinoic acid receptor (RARα) have also expanded on SINE/B2 elements.

and human genomes comprise repetitive DNA (Lander et al., 2001; Mouse Genome Sequencing Consortium et al., 2002), which is proportionately reflected in their respective DHS compartments (Figure 6.13a–b). Species-specific DHSs were enriched (relative to all DHSs) for nearly all classes and super-families of repetitive elements (Figure 6.13c), and 5–10% of shared DHSs overlap ancient repeats that predate the mouse and human divergence (Figure 6.13d). These features are compatible with an important role for transposons in the evolution and of the mammalian regulatory genomes (Okada et al., 2010).

Transposable elements have recently been implicated in the rapid expansion of a TF recognition element into novel genomic contexts, impacting transcriptional programs (Bourque et al., 2008; Jacques et al., 2013; Ward et al., 2013). To explore the generality of this phenomenom,

we estimated the total proportion of TF recognition sequences residing within species specific DHSs that arose from transposon expansion during mouse and human evolution. We found substantial asymmetries in these proportions (Figure 6.14). For example, the recognition elements for the pluripotency factor OCT4 have been greatly expanded in the murine lineage on a LTR/ERVL element, accounting for >25% of mouse-specific OCT4 sites vs. <5% in humans with a similar class of retroelement (Figure 6.14a). By contrast, expansion of recognition elements for the polyfunctional genomic regulator CTCF and the retinoic receptor-alpha (RARα) have been driven chiefly by SINE elements in both mouse and human (Figure 6.14b–c). These results suggest that expansion of the binding elements for transcription factors by repetitive elements is a general feature shaping both common and shared mammalian *cis*-regulatory landscapes.

6.2.5 A conserved core mammalian regulon

DHS patterns encode cellular fate and identity in a manner that reflects both current and future regulatory potential and informs developmental trajectory (Stergachis et al., 2013b). To visualize their cell- and tissue-selective activity patterns, we clustered shared DHS by their magnitude (total normalized DNase I cleavages) measured in each of the 45 mouse cell- and tissue-types (Figure 6.15a and Methods). The vast majority of shared DHSs (78.8%) evinced tissue-selective accessibility, and were readily organized into distinct cohorts of tissue-selective DHSs. A minority (21.2%) exhibited high accessibility across multiple tissue types, but only a very small fraction (<5%) showed constitutive activity (Figure 6.15b). Tissue-selective shared DHSs were enriched in distal regions (Figure 6.16) and reflected both tissue-level organization as well as anatomic or functional compartments within tissues. For example, 91,951 shared DHSs were specifically active in the brain; these in turn comprised four sub-clusters corresponding to distinct anatomical and developmental partitions (Figure 6.15a, green box). Similarly, blood DHSs were sub-compartmentalized into major hematopoietic lineages, comprising DHSs restricted to T cells, B cells, myeloid cells and erythroid cells (Figure 6.15a, red boxes). Across all compartments, cell/tissue-selective shared DHSs were preferentially localized around genes critical for the development and maintenance of their respective cell or tissue type (Figure 6.17).

We hypothesized that tissue-selective shared DHSs should encompass elements critical for basic mammalian regulatory processes such as development and differentiation, and that this would be reflected in their TF recognition sequence content. To analyze the contribution of individual TFs to specific shared DHS activity clusters, we computed the number of DHSs

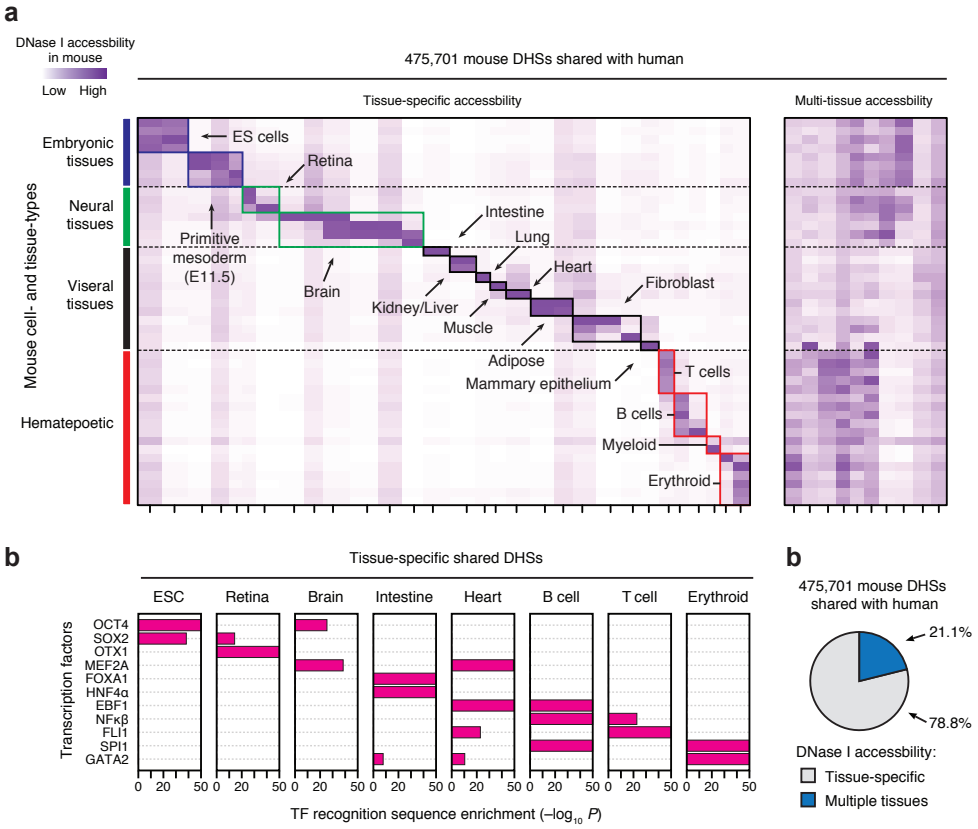

Figure 6.15: **Cell and tissue lineage encoding within shared regulatory elements**. (a) k-means clustering of DHSs by accessibility at each of the 475,701 shared DHS in mouse reveals tissue/cell restricted regulatory DNA. Columns correspond clusters of mouse DHSs that are also accessible in human and rows correspond to the 45 mouse cell/tissue types. Colors (axes and boxes) distinguish tissue groupings. Left, tissue-selective clusters. Right, clusters containing DHSs active in multiple tissues. (b) Proportion of shared DHSs that are tissue-selective or active in multiple tissues. (c) Enrichment of TF recognition sequences within tissue-selective DHSs computed using the cumulative hypergeometric distribution.

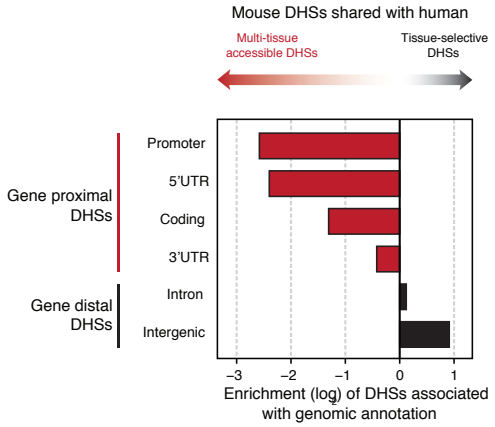

Figure 6.16: **Tissue-selective DHSs are predominantly distal to genes**. Enrichment (log_2) of tissue-selective shared DHSs categorized by genomic annotation.

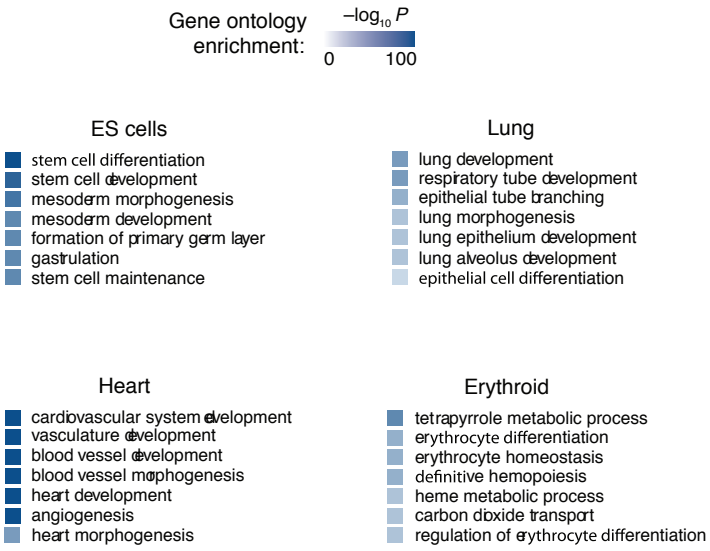

Figure 6.17: **Shared DHSs localize to genes important in development and differentiation**. Gene ontology terms associated with tissue-selective DHSs in mouse discovered using GREAT (63).

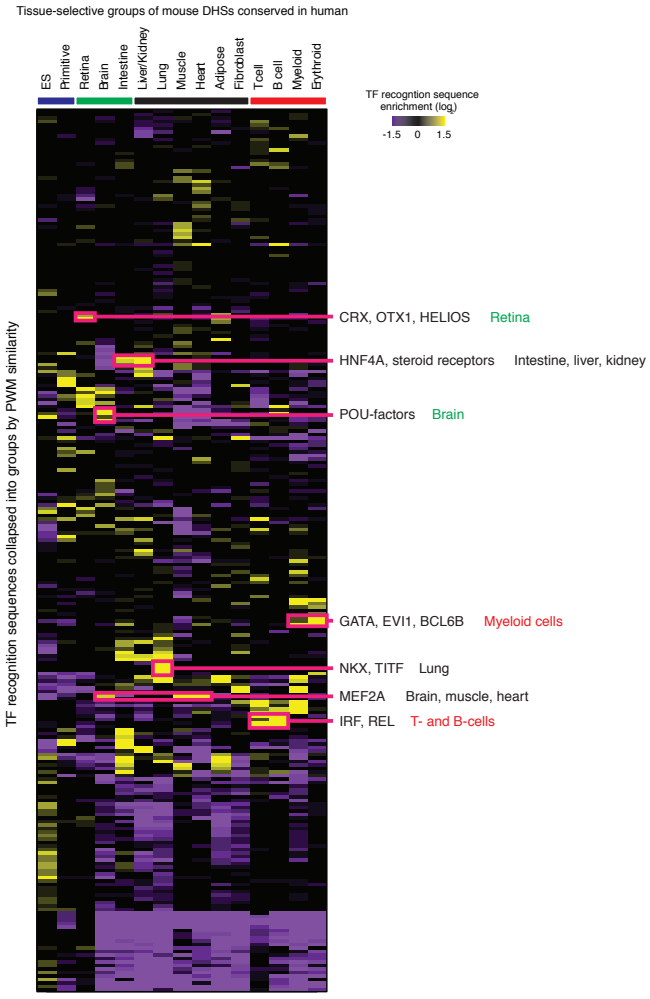

Figure 6.18: **Motif content within tissue-selective shared DHSs in mouse**. Individual position weight matrices were clustered to form groups of transcription factors with similar recognition sequences (see Methods). Each cell indicates the enrichment of individual PWM instances comprising the group (rows) in the tissue-selective mouse DHSs conserved in human (columns).

within each cluster that contained a recognition sequence for each TF, and compared these values to the overall distribution of TF recognition sequences for that TF within all shared DHSs. This analysis revealed a pronounced enrichment for nearly all known lineage- or cell identity-specifying regulators within tissue-selective DHSs, which were further organized combinatorially into their respective functional compartments (Figure 6.15c and figure 6.18). For example, recognition sequences for the pluripotency regulators Oct4, Sox2, Klf4, and Nanog were collectively concentrated within ES-selective shared DHS landscapes, consistent with their coordinated expression in ES cells. while Klf4 recognition sites were also enriched within intestine- and erythroid-specific DHSs, consistent with the known role of Krüppel-like factors in intestinal epitheliogenesis (Shields et al., 1996) and in erythropoiesis (Xiong et al., 2013). Analogously, the recognition elements for the cardiac regulators Mef2a, Ebf1, Fli1 and Gata4 (Edmondson et al., 1994; Garel et al., 1999; Schachterle et al., 2012) are enriched within heart-selective mouse-human shared DHSs, while Ebf1 and Fli1 elements are also enriched, respectively, in B and T cell shared DHSs, consistent with important functions in defining their respective cell fates (Anderson et al., 1999; Nechanitzky et al., 2013). Many members of TF families, such as the GATA factors recognize similar sequence elements, yet perform different regulatory roles in different tissues (Ko and Engel, 1993; Merika and Orkin, 1993). Indeed, the tissue-selective enrichments we observed were still highly significant even after TF recognition sequences are systematically grouped by similarity (Figure 6.18).

Together, the above results indicate that mouse-human shared DHSs densely encode regulatory information fundamental to diverse cell and tissue specification programs, and thus collectively define a core mammalian regulon.

6.2.6 Lineage repurposing of regulatory DNA

Since most shared DHSs showed strong cell/tissue-selectivity in mouse, we next asked to what degree these selectivity patterns were preserved in human. Computing the Jaccard similarity index over all possible combinations of mouse and human cell types revealed surprisingly limited similarity between mouse and human in the tissue-selective usage of shared DHSs (Figure 6.19a and figure 6.20a–b), even when accounting for variability in DNase I signal density and peak identification (Figure 6.21). Unsupervised hierarchical clustering loosely grouped shared DHSs in cells or tissues derived from the same progenitor or developmental lineage (Figure 6.19b).

The weak correspondence between orthologous tissues suggested that a substantial fraction of shared DHSs had undergone functional 'repurposing' via alteration of tissue activity from one

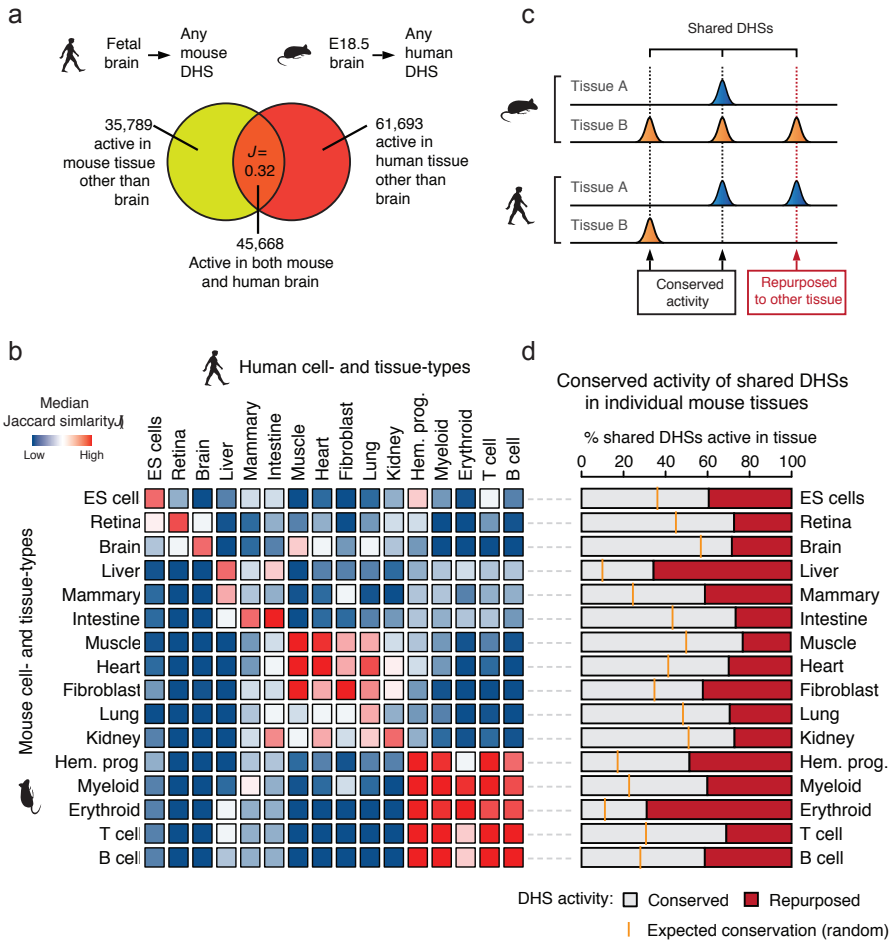

Figure 6.19: **Conservation and repurposing of regulatory DNA accessibility**. (a) Usage of the shared DHS landscape in human fetal brain and mouse embryonic brain indicate limited conservation in tissue-level accessibility patterns. (b) Pairwise comparison (median Jaccard distance) of shared DHS landscape usage between all mouse (rows) and human (columns) tissues largely mirrors their conserved morphological and embryological origins. (c) The activity patterns of individual shared DHSs during mouse and human evolution may have been conserved (activity in at least one similar tissue) or repurposed to another tissue. (d) The conservation of *cis*-regulatory DNA accessibility in mouse in individual tissue types indicates substantial repurposing of *cis*-regulatory DNA. Orange ticks indicate the expected overlap of randomly selected DHSs..

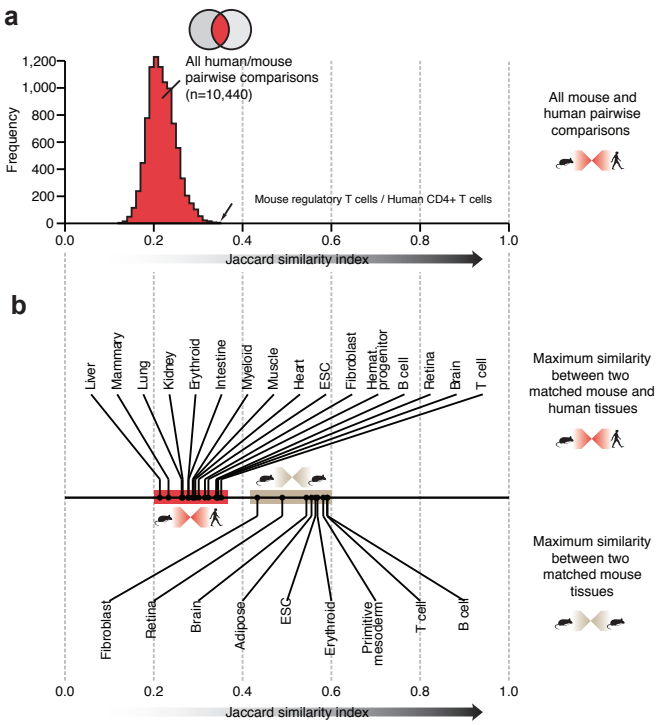

Figure 6.20: **Conservation of accessibility within shared DHSs**. (a) Histogram of all pairwise Jaccard similarity indices comparing the shared accessibility of DHSs common to both mouse and human tissues. (b) The maximum shared DHS landscape usage between matched mouse and human tissues (top, red) is markedly lower than a similar comparison between comparing mouse tissues with similar embryological origins (bottom, brown).

cell type or lineage in mouse to a different one in the human during the 91M years since divergence from a common ancestor (Figure 6.19c). Indeed, analysis of well-matched mouse and human tissue pairs revealed substantial repurposing ranging from 22.9%-69% of shared DHSs, depending on the tissue, significantly more than expected (Figure 6.19d). For example, of the 77,060 shared DHSs active in mouse muscle, 59,658 (77.4%) were also DHSs in human muscle, while the remaining 17,402 (22.6%) were DHSs in a different human tissue ((Figure 6.19d, 7th from top). Overall 35.7% of shared DHSs (12.7% of mouse DHSs overall) have undergone repurposing (Figure 6.22a), with the majority of the effect manifested at distal elements (Figure 6.22b). Flexible repurposing of regulatory DNA from one tissue context to another thus emerges as an

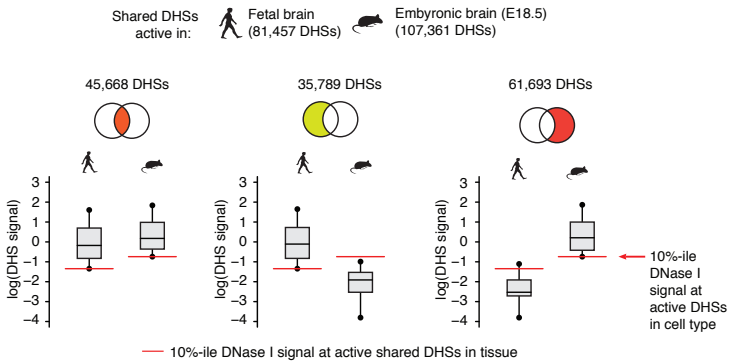

Figure 6.21: **Identification of conserved and repurposed DHSs is robust to DNase I cleavage intensity**. Box plots show the normalized DNase I cleavage intensity within mouse and human shared DHSs active in fetal and embryonic brain. Left, shared DHSs active in both mouse and human brain. Middle, shared DHSs active in human brain, but not active in mouse embryonic brain. Right, shared DHS active in mouse brain, but not active in human fetal brain. Boxes indicate the inner quartile range and the whiskers denote the 10%- and 90%-iles. Red line indicates the DNase I cleavage intensity at the weakest 10% shared DHSs in each tissues.

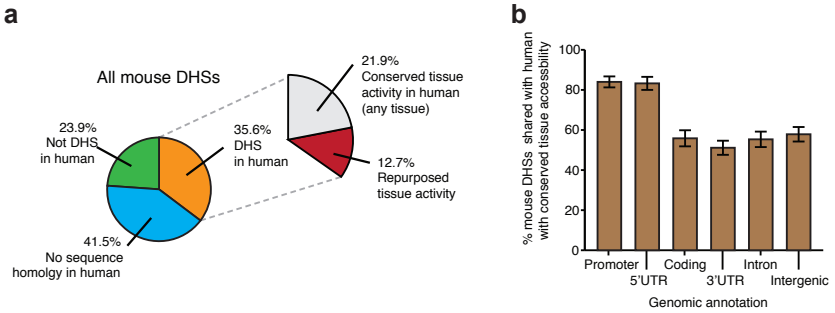

Figure 6.22: **Genomic distribution of shared DHSs with conserved tissue accessibility**. The proportion of DHSs with conserved tissue accessibility within each genomic compartment. Promoters are defined as 1 kb upstream of an annotated TSS. Bar indicate mean for all mouse tissues with a matched human tissue. Errors bars indicate the standard error of the mean.

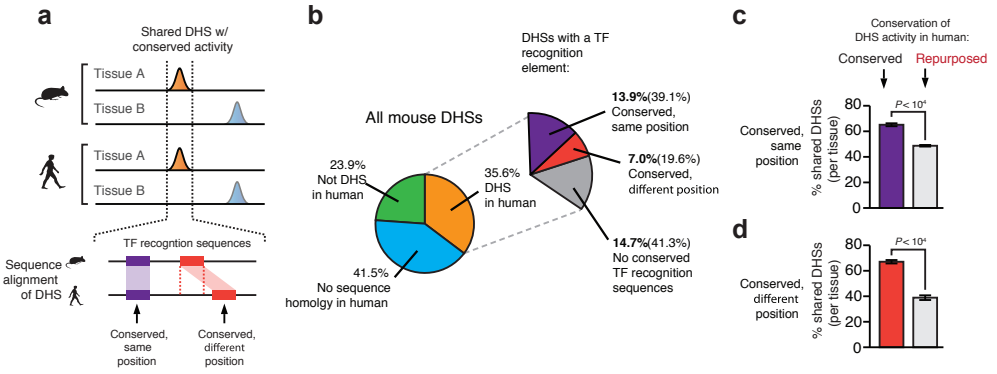

Figure 6.23: **Conservation of transcription factor recognition sequences**. (a) Model depicting the relationship of TF recognition sequences and conservation of accessibility at shared DHSs. DHSs with similar tissues activity patterns contain positionally and/or operationally conserved TF recognition sequences. (b) Overall conservation of TF recognition sequences within shared DHSs. (c–d) Positional and operational conservation of TF recognition sequences is linked to functional conservation.

important evolutionary mechanism shaping the mammalian *cis*-regulatory landscape.

6.2.7 Turnover of TF recognition elements within repurposed DHSs

Regulatory DNA densely encodes recognition sequences for transcriptional regulators (Neph et al., 2012b). We thus next examined the conservation of individual TF recognition elements within the shared DHS compartment, distinguishing between elements that were positionally conserved vs. those that were operationally conserved — i.e., were present but located at a different position within the DHS and thus arose independently (Figure 6.23a). Overall, within the shared DHS compartment, we found 39% of TF recognition sequences were positionally conserved, with 19.8% operationally conserved (Figure 6.23b). Both positional and operational conservation were concentrated within DHSs that maintained their tissue activity profile (vs. repurposed DHSs) (Figure 6.23c–d and figure 6.24a). Surprisingly, however, 41.2% of shared DHSs lacked any positionally or operationally conserved TF recognition elements (Figure 6.23b) and figure 6.24b–c). Notably, the TF motif densities did not differ significantly between shared DHSs with positionally, operationally, or non-conserved TFs (Figure 6.24d).

We next investigated the relationship between conservation of TF recognition sites and the

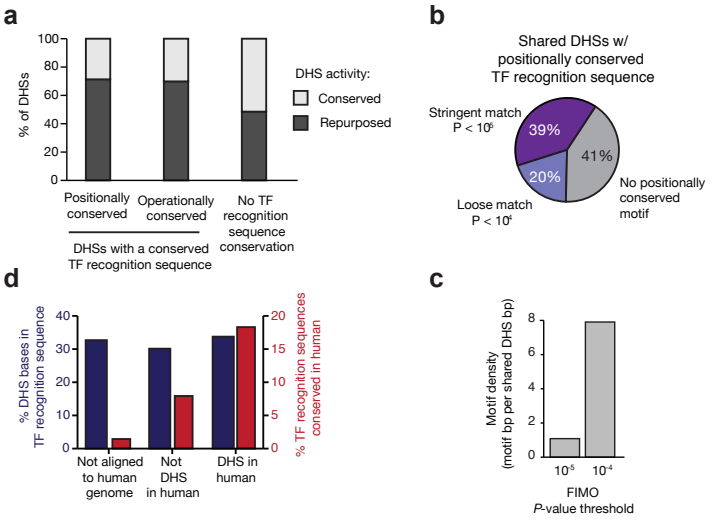

Figure 6.24: **Conservation of transcription factor recognition sequences within DHS**. (a) Proportion of mouse DHSs containing a positionally conserved, operationally conserved or no conserved TF recognition element that have conserved or repurposed DHS activity in human. (b) Effect of motif thresholds on the identification of conserved motifs. Left, proportion of mouse DHSs shared with human that contain a positionally conserved transcription factor recognition sequence at stringent (purple) and relaxed (lavender) thresholds. (c) The number of putative TF recognition sequences increases drastically with reduced P-value thresholds. (d) The density (blue, % of DHS covered by TF recognition sequences) of TF recognition sequences within mouse DHSs is uniform across with respect to functional conservation, while the proportion of TF recognition sequence base-pairs conserved (red, % of TF recognition sequence bases conserved) is increased in mouse DHSs conserved in human.

maintenance of tissue accessibility patterns. Reasoning that known TF regulators of cell fate or lineage would play an outsized role in repurposing, we hypothesized that recognition sequences for such TFs would be preferentially maintained (or gained) in DHSs with conserved tissue activity spectra, but preferentially lost at repurposed DHSs (Figure 6.25a). We found this to be the case across the spectrum of lineage-regulating TFs. For example, recognition sequences for the retinal master regulator OTX1 were >4-fold depleted within mouse retinal DHSs that had undergone repurposing in human compared with orthologous DHSs that had conserved retinal activity (Figure 6.25b). Analogously, recognition sites for the intestinal master regulator HNF1β were selectively depleted in repurposed intestinal DHSs, and those of the major erythroid regulator GATA1 were selectively depleted in repurposed erythroid DHSs (Figure 6.25b). These

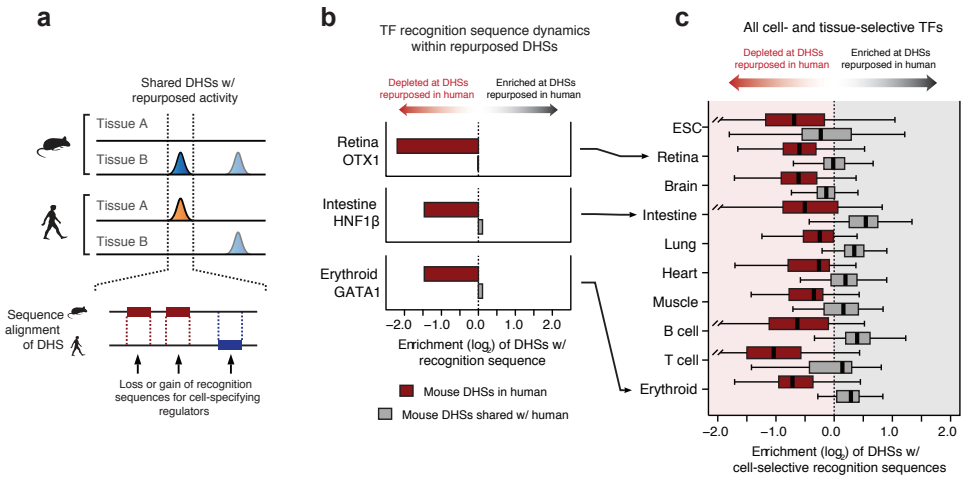

Figure 6.25: **Evolutionary dynamics of transcription factor recognition sequences**. (a) Model depicting the relationship of TF recognition sequence turnover and the repurposing of DHSs (b) Recognition sequences for cell-selective transcription factors are preferentially lost at mouse DHSs that are repurposed in human, while maintained in or gained in human. Representative examples of individual TF regulators in retina, intestine and erythroid tissues. (c) Same as (d) for recognition sequences of all cell-selective TF regulators within mouse DHSs repurposed in human.

findings are generalized in Figure 6.25c, which shows results for all cell/tissue-selective recognition sequences identified in Figure 6.18. Across the spectrum of tissues, we consistently observed depletion of cell-selective TF recognition sequence within repurposed DHSs. These results thus link the conservation and repurposing of DHSs to preservation vs. turnover of specific TF recognition sequences. They also suggest an incremental process wherein the composition of TFs within a single DHS is remodeled over evolutionary time via sequential small mutations (Payne and Wagner, 2014) that could ultimately affect function and phenotype (Prud'homme et al., 2006; Williams et al., 2008). The presence of a substantial population of shared DHSs without preserved TF recognition sites yet preserved DHS activity (and even preserved tissue-selectivity patterns) highlights the plasticity of individual *cis*-regulatory templates, and indicates that the same higher-level regulatory outcome can be encoded by a large number of different combinations of instructive TF recognition events.

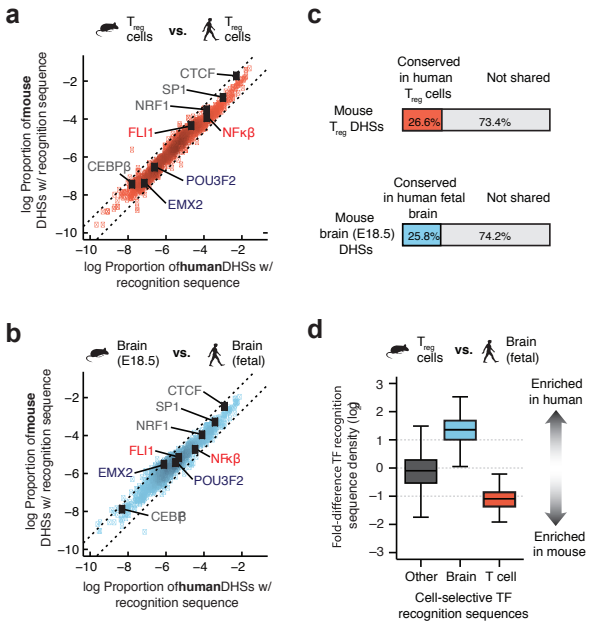

Figure 6.26: **Conservation of *cis*-regulatory content in brain and regulatory T cells**. (a) Density of individual TF recognition sequences in both human (x-axis) and mouse (y-axis) regulatory T cells. Dotted black lines demarcate a 2-fold difference in density between mouse and human. (b) Same as (a) for human and mouse brain. (c) Proportion of mouse DHSs that are conserved in a matched human tissue. Top, mouse regulatory T cells DHSs that are conserved in human regulatory T cells. Bottom, mouse embryonic brain DHSs that are conserved in human fetal brain. (d) A comparison of TF recognition sequence density in human fetal brain to mouse regulatory T cells reveals cell-selective regulators

6.2.8 Conservation of global TF recognition landscapes

We next asked how the marked plasticity of TF recognition elements within the evolving *cis*-regulatory landscape was reflected in global patterns in the types and quantities of such elements. To investigate this, we computed the global density of recognition sequences for each of 744 TFs within all mouse and human DHSs (separately, and irrespective of conservation status) from each cell/tissue type. This analysis revealed striking conservation of the proportion of the regulatory DNA landscape of each cell type devoted to recognition sites of each TF. Figure 6.26a–b show examples for mouse vs. human regulatory T cell DHSs, and mouse brain vs. human fetal brain; in each case, a linear relationship is observed indicating that the proportion of the DHS space devoted to recognition sequences of each of 744 TFs has been strictly

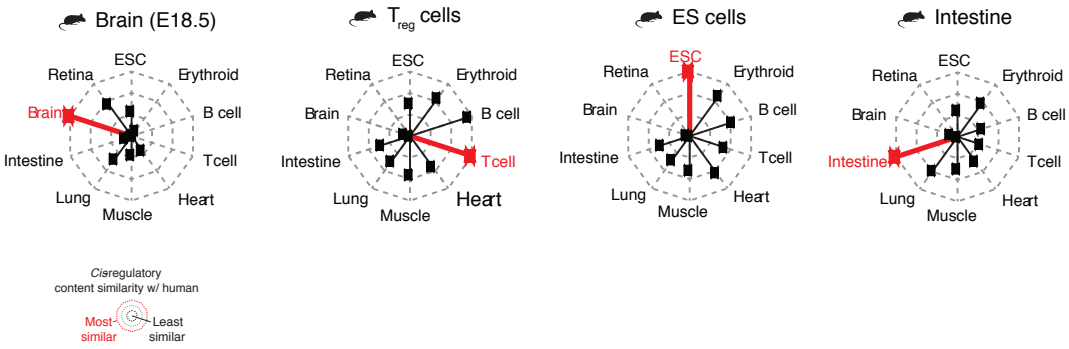

Figure 6.27: **Conservation of *cis*-regulatory content dominates over the conservation of individual regulatory elements** . Radar plots showing the median similarity of the *cis*regulatory content between mouse and human tissues. Similarity is defined as the pairwise Euclidean distance between the density of TF recognition sequences in mouse and human tissues.

conserved (Figure 6.26a–b). These findings stand in marked contrast to the weak conservation (25%) of individual mouse regulatory T cell and brain DHSs in human (Figure 6.26c). TF recognition sequence content varied between cell/tissue types, with effector TFs selectively enriched within their cognate cell type (Figure 6.26d), and TF recognition sequence density was consistently most similar between orthologous cell/tissue pairs vs. non-orthologous cells/tissues (Figure 6.27). It has been proposed that in large genomes such as mouse and human, maximization of the occupancy of any given TF demands an excess of its recognition sites in order to ensure high occupancy of sites with critical regulatory roles across a range of TF concentrations (Lin and Riggs, 1975). Consistent with this model, the majority of DHSs in both the mouse and human genome show relaxed sequence constraint over evolutionary distances (Figure 6.9a) and within current human populations (Figure 6.9b). Collectively, these results show that in spite of marked plasticity within the regulatory DNA compartment, the aggregate recognition space for each TF has been strikingly conserved during the mouse-human evolutionary interval. It is particularly notable that this finding obtains across a wide spectrum of TFs that encompasses diverse functional roles and biophysical mechanisms of DNA recognition.

6.2.9 Perspective

Taken together, the results reported herein have important implications for understanding the major mechanisms and forces governing the evolution of mammalian regulatory DNA. The pervasive repurposing and wholesale turnover of individual regulatory elements with concomitant conservation of global TF recognition site content indicates that the combination of a highly conserved *trans*-regulatory environment and large genome (under weakened selection) has shaped the mammalian *cis*-landscape by facilitating both the de novo creation and the *cis*-migration of operational TF binding elements. We speculate that high *cis*-regulatory plasticity may be a key facilitator of mammalian evolution by increasing the potential for innovation of novel functions in the context of an evolutionarily inflexible *trans*-regulatory environment.

6.3 Methods

6.3.1 Nuclei isolation from solid mouse tissues

Solid mouse tissues were minced in 2 mm square pieces and resuspended in 3 mL of homogenization buffer (20 mM tricine, 25 mM D-sucrose, 15 mM NaCl, 60 mM KCl, 2 mM $MgCl_2$, 0.5 mM spermidine, pH 7.8) per gram of tissue. The nuclei were released by 5-10 strokes in a Dounce homogenizer with a loose-fitting type-A pestle and the resulting homogenate was filtered through a 100μm filter. For some samples, the homogenate was cryopreserved before DNaseI treatment (10% DMSO was added for a controled freeze to -80°C; stored in liquid nitrogen). Prior to DNase I treatment an additional buffer exchange was performed (for both fresh and frozen samples) by first adding 15 mL of sucrose buffer (10 mM Tris-HCl, 250 mM D-sucrose, 1 mM $MgCl_2$, pH 7.5), collecting nuclei by centrifugation (600*g* for 10 minutes at 4°C), and then resuspending nuclei pellet in 10 mL of fresh sucrose buffer. The nuclei were filtered through a 20µm filter and collected by centrifugation (600g for 10 minutes at 4°C). The nuclei pellet was washed once in 10 mL of buffer A (15 mM Tris-HCl, 15 mM NaCl, 60 mM KCl, 1 mM EDTA, 0.5 mM EGTA, 0.5 mM spermidine) and resuspended at concentration of 2×10^6 per mL.

6.3.2 Nuclei isolation of mouse cultured and primary cells

Cells (primary or cultured) were washed once with Dulbecco's PBS (without $MgCl_2$ or $CaCl_2$). Nuclei were extracted by resuspending cells in buffer A supplemented with detergent (IGEPAL-CA630) (Sigma) and incubating for 10 minutes on ice. Following incubation, the nuclei were collected by centrifugation (600*g*) and resuspended in buffer A at a concentration of 2×10^6 nuclei per mL. Optimal detergent concentrations for nuclei extraction were empirically derived for each cell type (commonly ranging from 0.010–0.10%).

6.3.3 DNase I digestion of mouse nuclei

Fresh nuclei were incubated for 3 minutes at 37°C with limiting concentrations of the DNA endonuclease deoxyribonuclease I (DNase I) (Sigma) in buffer A supplemented with Ca^{2+}. The digestion was stopped with stop buffer (50 mM Tris-HCl, 100 mM NaCl, 0.1% SDS, 100 mM EDTA, 1 mM spermidine, 0.5 mM spermine, pH 8.0) and the samples were treated with proteinase K and RNase A. The small 'double-hit' fragments (<750 bp) were recovered by sucrose

ultracentrifugation, end-repaired and ligated with adapters compatible with the Illumina sequencing platform. A detailed protocol describing genome-wide mapping of DNase I hypersensitivity can found in (John et al., 2013).

6.3.4 DNase I fragment sequence alignment and normalization

Mouse (NCBI37) and human (GRCh37) 36 bp sequence reads were mapped using bowtie (Langmead et al., 2009), version 0.12.7 with parameters: `--mm -v 3 -k 2`, with `-phred33-quals` for Illumina HiSeq sequencer runs or `--phred64-quals` for Illumina GAII sequencer runs. Only uniquely mapping reads with up to 2 mismatches were retained; this was accomplished by additionally filtering the `-k 2` results, when present, for actual uniqueness within the potential 2-3 mismatch alignments (any remaining 3-mismatch-only alignments were discarded). Signal tracks were generated using BEDOPS (Neph et al., 2012a), summing reads within a window size of $\pm$75bp in 20 bp steps and subsequently normalizing to the total number of reads per dataset and then scaling to one million reads. The NCBI and Mouse Genome Sequencing Consortium build 37 (MGSCv37/mm9) including chromosome Y was used as the reference assembly for all sequence alignments.

6.3.5 DHS identification and master list creation

We identified DNase I hypersensitive regions of chromatin accessibility (hotspots) and more highly accessible DNaseI hypersensitive sites (DHSs, or peaks) within the hotspots, using the hotspot algorithm (John et al., 2011). Briefly, the hotspot algorithm is a scan statistic that uses the binomial distribution to gauge enrichment of tags based on a local background model estimated around every tag. General-sized regions of enrichment are identified as hotspots, and then 150 bp peaks within hotspots are called by looking for local maxima in the tag density profile (sliding window tag count in 150 bp windows, stepping every 20 bp). Further stringencies are applied to the local maximum detections to prevent overcalling of spurious peaks. Hotspot also includes a FDR (false discovery rate) estimation procedure for thresholding hotspots and peaks, based on a simulation approach. Random reads are generated at the same sequencing depth as the target sample, hotspots are called on the simulated data, and the random and observed hotspots are compared via their Z-scores (based on the binomial model) to estimate the FDR.

The DHSs called on individual cell-types were consolidated into a master list of unique, non-

overlapping DHS positions by first merging the FDR 1% peaks across all cell-types. Then, for each resulting interval of merged sites, the DHS with the highest Z-score was selected for the master list. Any DHSs overlapping the peaks selected for the master list were then discarded. The remaining DHSs were then merged and the process repeated until each original DHS was either in the master list or discarded.

Due to the variability in sequencing depth within the mouse DNase I experiments, we down-sampled each mouse dataset to 25 million sequencing tags (random sampling without replacement) and used the sampled tags to detect hotspots and DHSs. Data availability

All mouse sequence data generated for this study can be accessed with GEO accession numbers found within table S1. Processed data such as hotspots and peaks are released as part of the Mouse ENCODE Consortium and available for download at the data portal website (http://www.mouse-encode.org).

6.3.6 Human DNase I data

Human DNase I data was generated as part of the ENCODE Project (ENCODE Project Consortium et al., 2012) or the Roadmap Epigenomics Mapping Consortium (Bernstein et al., 2010). Raw sequence tag data can be accessed through the GEO accession numbers in table S2. Processed data such as hotspots and DHS peaks can be accessed at http://www.encode-roadmap.org/.

6.3.7 Genomic annotations

Genome annotations used for all analysis correspond to the GENCODE version 10 (Harrow et al., 2012) (human) or ENSEMBL 65 (Flicek et al., 2013) (mouse). Promoters are defined as 1 kb upstream of a TSS.

6.3.8 Alignment of DHSs between mouse and human

Pair-wise genome alignments ("over" chain files) between mouse (mm9) and human (hg19) were downloaded from the UCSC Genome Browser (Meyer et al., 2013). Using these chain files, DHSs were mapped between species using the software "bnMapper" (bx-python software package; available at http://bitbucket.org/james_taylor/bx-python) (Denas et al. 2013, submitted) using the following parameters: `-fBED12 --gap 20 --threshold 0.1`. The mapped blocks were then intersected with a DHS peak list from the query species using the software

BEDOPS (Neph et al., 2012a), requiring only 1 base-pair of overlap. We applied this mapping strategy for each target-query pair (i.e., human → mouse and mouse → human) and then retained DHSs in each list that were in strict reciprocal relationships.

6.3.9 Alignment of mouse DHS sequence to the vertebrate phylogeny

Pairwise genome alignments ("over" chain files) were downloaded from the UCSC Genome Browser between mouse (mm9) or human (hg19) to the following genomes: panda (ailMel1), lizard (anoCar2), cow (bosTau7), lancelet (braFlo1), marmoset (calJac3), dog (canFam2), guinea pig (cavPor3), zebrafish (danRer7), tenrec (echTel1), horse (equCab2), cat (felCat4), fugu (fr2), chicken (galGal3), stickleback (gasAcu1), human (hg19), elephant (loxAfr3), turkey (melGal1), opossum (monDom5), platypus (ornAna1), rabbit (oryCun2), medaka (oryLat2), sheep (oviAri1), chimp (panTro3), lamprey (petMar1), orangutan (ponAbe2), macaque (rheMac2), rat (rn5), pig (susScr2), tetradon (tetNig2), and frog (xenTro3). Using the sequence alignment strategy described above (see Cross-alignment of DHSs between mouse and human), each mouse DHS was to each of these genomes.

To estimate the proportion of regulatory elements conserved throughout vertebrate evolution, we overlayed the sequence alignments on a recent proposed vertebrate phylogenies obtained from (Bininda-Emonds et al., 2007; Hedges, 2002; Murphy et al., 2007). We considered a DHS to be conserved if it aligned successfully to one or more species within a branch of the phylogeny that share a MRCA with mouse.

6.3.10 Evolutionary sequence constraint

The phastCons (Siepel et al., 2005) element track corresponding to a 46-way multiple alignment of vertebrate species was downloaded from the UCSC Genome Browser (Meyer et al., 2013). To assess conservation, we computed the fraction of phastCons elements that overlapped mouse DHSs grouped by conservation status with human.

6.3.11 Nucleotide diversity

Human nucleotide diversity measurements π were calculated using whole genome sequences from 53 unrelated, publicly available human genomes released by Complete Genomics (51) (version 1.1034) as previously described (Thurman et al., 2012; Vernot et al., 2012). π for a heterozygous site is $2pq$, where p is the major allele frequency and q the minor allele frequency. π was

calculated for each human tissue by summing 2pq all variants and dividing by the total number of nucleotides considered. Repetitive elements identified by RepeatMasker (Jurka et al., 2005; Smit et al., 1996–2010) (see Analysis of repeat content within DHS) were removed from all π calculations.

6.3.12 Functional variation within human DHSs

A catalog of single nucleotide polymorphisms (SNPs) identified by genome-wide association studies was downloaded from NHGRI GWAS Catalog on December 3, 2013 from: http://www.genome.gov/admin/gwascatalog.txt. SNPs within coding regions were removed yielding a total of 6,571 total loci. We randomly sampled the same number of SNPs from the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2010) matched by distance to TSS, intergenic or intronic, and allele frequency as in (Maurano et al., 2012). We repeated this sampling process 1,000 times to estimate the parameters of a normal distribution (μ, σ). These parameters were used to calculate the upper-tail p-value of the observed overlap of GWAS SNPs.

6.3.13 Analysis of repeat content within DHSs

We scanned both hg19 and mm9 for repeats using the RepeatMasker program (Smit et al., 1996–2010) (http://repeatmasker.org) using the default parameters except for the specification of the species (e.g., human or mouse). The RepeatMasker database version 2012-04-18 was obtained from RepBase (Jurka et al., 2005).

6.3.14 Tissue-selective mouse DHSs conserved in human

We calculated the maximum density (tags per 150 bp, tiled at 20 bp genome-wide and normalized for sequencing depth) of DNase I cleavage within each DHS master peak across all 45 mouse tissues/experiments. To limit the effects of outliers, we normalized each value to 90% of the maximum (across all experiments) at each DHS peak. We then performed k-means clustering on a matrix with the columns comprising mouse DNase I experiments and the rows corresponding to individual DHSs contained within the master DHS peak list. The clustering was performed using an efficient implementation of the k-means++ algorithm (Arthur and Vassilvitskii, 2007) (GraphLab API, http://graphlab.org/toolkits/clustering/) setting $k = 45$.

6.3.15 TF recognition sequence predictions

Transcription factor binding sites identified by scanning the entire genome for consensus sequences using the FIMO (Grant et al., 2011) tool from the MEME Suite (Bailey et al., 2009) (version 4.6). A 5th order Markov model was generated from 36 bp mappable genome sequence and used as the background model. Motif models were curated from TRANSFAC (Matys et al., 2006) (version 11), JASPAR (Bryne et al., 2008), and a SELEX-derived set from (Jolma et al., 2013). Putative binding sites with a FIMO $P < 10^{-4}$ were retained.

6.3.16 Grouping TF recognition sequence models by similarity

Motif models used for the genome-wide scans were compared pairwise using the software TOMTOM (Gupta et al., 2007) tool from the MEME Suite (Bailey et al., 2009) (version 4.6) with the following parameters: `-dist kullback -query-pseudo 0.1 -target-pseudo 0.1 -text -min-overlap 0 -thresh 1`. The same 5th order Markov model background model as the FIMO genome-wide scans was used. The resulting pairwise comparisons were hierarchically clustered using Pearson correlation as the distance metric and complete linkage. Clusters were selected by cutting the tree at a height of 0.1.

6.3.17 TF recognition sequence enrichment within tissue-selective DHSs

We computed the number of DHSs containing an instance of each motif model. Using the cumulative hypergeometric distribution, a p-value for the number of observed DHSs containing a TF recognition sequence within a particular cluster of DHSs with respect the overall prevalence of the recognition sequence within all mouse DHSs conserved in human. The p-values were thresholded using the Bonferroni correction method.

6.3.18 Gene ontology analysis of tissue-selective DHSs

DHSs within lineage-specific clusters were supplied as input to GREAT (McLean et al., 2010). The analysis was run in the "basal plus extension" configuration, such that proximal regions were defined as 5 kb upstream and 1 kb downstream and distal regions were limited to 1 Mb.

6.3.19 Analysis of conserved DHS landscape usage mouse and human tissues

To compare the usage of the conserved DHS landscape in between two tissues we computed the Jaccard index. The Jaccard index is defined as

$$\frac{A \cap B}{A \cup B} \tag{6.1}$$

where A and B are the number DHSs active in a mouse or human tissue. To compare multiple independent samples of similar tissues (as in fig. 3C) we used the median Jaccard index of all pairwise combinations.

6.3.20 Comparison of tissue activity of mouse DHSs conserved in human

For each shared DHS peak, tissue activity was defined as whether the DHS peak was identified via DNase I assay within that tissue (binarization of DHS signal into "accessible" or "inaccessible"). We associated tissues/cell types/experiments into tissue categories and took the union of all accessible DHSs within the datasets comprising the group. As a control, we compared the conservation of tissue activity against human DHSs with shuffled tissue activity profiles by sampling DHS from all shared human DHS without replacement keeping the number sampled constant with the number of active DHSs in each tissue.

6.3.21 TF recognition sequence conservation

To identify positionally conserved transcription factor recognition sequences, predicted transcription factor binding sites identified with FIMO (see TF recognition sequence predictions) within all mouse DHSs were aligned to the human genome, using the "over chain" pairwise alignment downloaded from UCSC Genome Browser (Meyer et al., 2013). We then obtained the human coordinates for the aligned mouse binding sites and overlapped them with predicted transcription factor binding sites in human. Importantly, both mouse and human genomes were scanned using the same motif models. A relaxed threshold (FIMO $P < 10^{-4}$) was used for human genome scans. A motif was labeled as conserved if the same motif was identified in mouse and human and the human motif matched the exact coordinates as the aligned mouse motif.

6.3.22 Positional and functional conservation of TF recognition sequences

Functionally conserved TF recognition sequences was identified by first filtering all DHSs with a one or more positionally conserved TF recognition sequence (see above). We then searched for shared DHSs that contained independent instances in mouse and human of a motif that corresponded to the recognition sequence for the same transcription factor.

6.3.23 TF recognition sequence turnover at repurposed DHSs

To assess the dynamics of TF recognition sequence evolution within shared DHSs with respect to conservation of tissue-specific DNase I accessibility we partitioned DHSs active in each mouse tissue, for both mouse and human, into two groups: conserved and repurposed. For each of these groups, we then computed the proportion of DHSs containing an instance of a motif (FIMO $P < 10^{-5}$). The assumption of this analysis is that the TF recognition sequence density should either be maintained (neutral loss or gain) or increase (net gain) when comparing mouse DHS with conserved accessibility in human vs. DHSs that have been repurposed human. For human, however, we expect to see a relative reduction in TF recognition sequences in a comparison of DHSs with conserved vs. repurposed accessibility. Tissue-selective TF recognition sequences were defined by the motifs enriched ($P < 10^{-5}$) within the tissue-specific clusters (Figure 6.15c).

6.3.24 Comparison of *cis*-regulatory sequence content

We examined the *cis*-regulatory sequence content between mouse and human tissues by computing the proportion of all DHSs in a tissue (regardless of DHSs sequence or functional conservation) containing at least one instance of each TF recognition sequence (see TF recognition sequence predictions; FIMO $P < 10^{-5}$). Brain- and T cell-selective TF recognition sequences were defined by the motifs enriched $P < 10^{-5}$) within the tissue-specific clusters (identified in figure 6.15c).

To assess the similarity of each mouse tissue to all human tissues we computed the Euclidean distance pairwise between the proportions of DHSs contain each TF recognition sequences. When comparing a single tissue to multiple tissues, as in figure 6.27, we used the median Euclidean distance between all possible pairwise combinations.

Acknowledgments

This work was supported by NIH grants U54HG007010 to J.A.S. and 1RC2HG005654 to J.A.S. and M.G. Additional support was provided by NIH grants R37DK44746 to M.G. and M.A.B., and 2R01HD04399709 to L.S. J.V. is supported by a National Science Foundation Graduate Research Fellowship under grant no. DGE-071824. J.V. and J.A.S. designed the experiments and analysis. E.R., R.S. and R.E.T. aided in data analysis and management. All other authors participated in data generation and sample collection. J.V. and J.A.S. wrote the manuscript

with help from E.R. We would like to thank H. Wang and E.K. Salinas for help with figures. All sequence data generated in this study can be accessed with GEO accession numbers found within tables 6.1 and 6.2.

Table 6.1: **Mouse cell- and tissue-types**. Overview of strain, developmental age and DHS mapping statistics for each mouse tissue used within this study.

Cell- or tissue-type	Strain*	Age#	Tissue	GEO Accession	SPOT	Tags	Tags (-chrM)	DHS peaks (FDR 1%)	Downsampled DHS peaks (FDR 1%)
E14	129/Ola	E0	ES	GSM1014154	0.457	33,828,611	32,646,012	175,237	165,014
CJ7	129S1/SVimJ	E0	ES	GSM1014187	0.427	35,432,673	35,123,245	161,628	146,807
ZhBTc4	129/Ola	E0	ES	GSM1014169	0.558	30,861,073	29,216,912	164,855	158,561
ESC (F4WT)	WW6	E0	ES	GSM1014159	0.427	24,221,226	23,283,969	167,073	167,055
416b	B6D2F1/J	I	Myeloid	GSM1014163	0.538	42,933,235	40,389,545	141,483	129,350
MEL	Unknown	I	Erythroid	GSM1014191	0.502	28,778,970	28,632,367	129,572	125,720
Erythroblast (CD117+; CD71-; Ter119-)	CD-1	E14.5	Hem. Prog.	GSM1014155	0.435	26,102,618	26,038,902	147,180	146,269
Erythroblast (CD117+; CD71+; Ter119-)	CD-1	E14.5	Erythroid	GSM1014158	0.491	26,514,730	26,250,932	116,843	115,265
Erythroblast (CD117-; CD71+; Ter119+)	CD-1	E14.5	Erythroid	GSM1014156	0.324	39,265,301	39,135,425	74,521	68,271
Erythroblast (CD117+ ;CD71+; Ter119+)	CD-1	E14.5	Erythroid	GSM1014157	0.406	51,563,832	51,115,669	109,017	95,539
Na•ve T cell (activated)	C57BL/6*	W8	T cell	GSM1014149	0.557	28,386,529	28,157,831	85,736	83,850
Regualtory T cell (activated)	C57BL/6*	W8	T cell	GSM1014200	0.543	32,305,988	32,264,326	100,369	97,162

Continued on next page

Table 6.1 – *Continued from previous page*

Cell- or tissue-type	Strain*	Age#	Tissue	GEO Accession	SPOT	Tags	Tags (-chrM)	DHS peaks (FDR 1%)	Downsampled DHS peaks (FDR 1%)
Na•ve T cell (resting)	C57BL/6*	W8	T cell	GSM1014192	0.519	33,637,621	33,535,765	100,535	93,602
Regulatory T cell (resting)	C57BL/6*	W8	T cell	GSM1014148	0.560	28,564,791	28,422,561	93,807	91,491
B cell (CD19+)	C57BL/6*	W8	B cell	GSM1014190	0.425	28,983,865	28,854,216	94,291	92,120
B cell (CD53+)	C57BL/6J	W8	B cell	GSM1014170	0.491	29,461,779	29,172,586	92,539	88,889
Liver	129*	E14.5	Erythroid	GSM1014162	0.367	29,791,189	29,512,936	102,545	97,849
Liver	C57BL/6J	E14.5	Erythroid	GSM1014183	0.351	44,019,764	43,575,379	112,444	97,047
Liver	C57BL/6J	W8	Liver	GSM1014195	0.423	19,812,754	19,748,170	104,995	104,979
A20 (lymphoma)	BALB/cAnN	I	B cell	GSM1014167	0.446	29,879,112	29,050,553	123,283	120,322
Mammary adenocarcinoma	RIII	I	Mammary	GSM1014196	0.686	32,345,915	32,111,316	91,073	84,282
Whole brain	C57BL/6J	E18.5	Brain	GSM1014197	0.483	25,257,085	25,249,691	167,715	166,503
Whole brain	C57BL/6J	W8	Brain	GSM1014151	0.689	36,089,477	36,079,464	224,978	204,809
Cerebrum	C57BL/6J	W8	Brain	GSM1014168	0.437	43,877,302	43,834,852	242,095	201,693
Cerebellum	C57BL/6J	W8	Brain	GSM1014164	0.397	21,727,916	21,690,965	106,514	106,499
Lung	C57BL/6J	W8	Lung	GSM1014194	0.423	21,041,166	20,978,206	163,431	163,463
Kidney	C57BL/6J	W8	Kidney	GSM1014193	0.451	46,472,928	46,357,946	184,503	160,029
Fat pad (mammary)	C57BL/6J	W8	Adipose	GSM1014165	0.355	19,628,835	19,566,613	139,661	139,658
Fibroblast (tail)	C57BL/6J	I	Fibroblast	GSM1014199	0.580	36,767,603	30,042,276	163,505	156,571
Fibroblast (NIH3T3)	NIH/Swiss	I	Fibroblast	GSM1014177	0.516	30,124,872	28,204,510	131,639	129,066

Continued on next page

Table 6.1 – *Continued from previous page*

Cell- or tissue-type	Strain*	Age#	Tissue	GEO Accession	SPOT	Tags	Tags (-chrM)	DHS peaks (FDR 1%)	Downsampled DHS peaks (FDR 1%)
Fibroblast (embryonic kidney)	Spretus.BL6-Xist	I	Fibroblast	GSM1014171	0.463	25,485,737	22,217,276	137,147	137,138
Heart	C57BL/6J	W8	Heart	GSM1014166	0.433	28,947,564	28,077,411	146,214	141,807
Embyro (headless)	CD-1	E11.5	Prim. mesoderm	GSM1014172	0.402	29,943,418	29,909,171	149,434	138,370
Embryo (forelimb buds)	CD-1	E11.5	Prim. mesoderm	GSM1014174	0.458	26,816,011	26,734,717	150,619	147,330
Fat pad (genital)	C57BL/6J	W8	Adipose	GSM1014173	0.439	25,555,735	24,690,742	169,155	169,143
Embryo (hindlimb buds)	CD-1	E11.5	Prim. mesoderm	GSM1014179	0.389	30,266,505	30,132,401	157,008	144,567
Embryo (mesoderm)	CD-1	E11.5	Prim. mesoderm	GSM1014178	0.456	103,121,762	102,805,749	228,571	148,193
Intestine (large)	C57BL/6J	W8	Intestine	GSM1014186	0.398	30,356,956	29,995,670	131,435	129,237
Muscle (skeletal)	C57BL/6J	W8	Muscle	GSM1014189	0.389	102,881,281	97,647,043	193,949	126,944
Retina	C57BL/6J	W8	Retina	GSM1014175	0.323	28,638,758	28,252,986	100,570	97,088
Retina	C57BL/6J	D7	Retina	GSM1014198	0.418	27,024,790	26,973,855	109,073	108,605
Retina	C57BL/6J	D0	Retina	GSM1014188	0.484	36,128,496	35,959,145	134,712	123,526
Spleen	C57BL/6J	W8	B cell	GSM1014182	0.610	25,037,085	24,916,157	86,664	86,661
Thymus	C57BL/6J	W8	T cell	GSM1014185	0.495	26,399,759	26,286,891	105,449	103,977
CH12 (lymphoma)	B10.H-2aH-4bp/Wts	I	B cell	GSM1014153	0.536	68,983,868	68,760,538	171,305	141,262

* indicates that the substrain is not known

E = days post-conception, D = days post-natal, W = weeks post-natal, I = immortalized or malignant

Table 6.2: **Human cell- and tissue-types**. Overview of developmental age and DHS mapping statistics for each human tissue used within this study.

Cell- or tissue-type	Source	Age#	Tissue	GEO Accession	SPOT	Tags	Tags (-chrM)	DHS peaks (FDR 1%)
A549	ENCODE		Lung	GSM736580	0.438	33,328,713	20,075,666	118,965
AG04449	ENCODE		Fibroblast	GSM736562	0.462	29,035,304	22,616,899	166,158
AG04450	ENCODE		Fibroblast	GSM736514	0.464	27,024,153	22,671,606	148,086
AG09309	ENCODE		Fibroblast	GSM736551	0.695	38,396,845	26,118,731	201,320
AG09319	ENCODE		Fibroblast	GSM736531	0.670	28,332,355	19,707,391	141,216
AG10803	ENCODE		Fibroblast	GSM736598	0.747	33,534,037	25,715,256	171,180
AoAF	ENCODE		Blood vessel	GSM736583	0.716	38,226,561	31,210,444	173,907
BE2_C	ENCODE		Brain	GSM736508	0.614	44,063,888	42,391,862	175,969
BJ	ENCODE		Fibroblast	GSM736518	0.749	42,763,886	24,605,012	162,086
CACO2	ENCODE		Intestine	GSM736500	0.707	27,117,636	25,576,071	122,479
CD14	ENCODE		Myeloid	GSM736513	0.429	67,698,560	67,035,406	175,178
CD19	REMC	y34	B cell	GSM701493	0.483	24,575,305	23,887,915	84,515
CD20	ENCODE		B cell	GSM1024765	0.572	36,983,818	36,413,583	104,139
CD34	REMC	y33	Hem. prog.	GSM530657	0.769	22,001,770	21,463,361	139,457
CD34	ENCODE		Hem. prog.	GSM1024770	0.691	49,756,223	48,561,124	164,050
CD4	REMC	y37	T cell	GSM701539	0.625	28,031,950	27,199,471	93,360
CD4pos_N	ENCODE		T cell	GSM1024789	0.356	24,083,134	22,482,769	82,384
CMK	ENCODE		Myeloid	GSM736607	0.572	31,319,078	21,490,670	133,032
fAdrenal	REMC	d96	Adrenal	GSM530653	0.440	27,428,606	27,094,195	188,072
fAdrenal	REMC	d85	Adrenal	GSM665799	0.352	33,450,853	32,872,463	181,935
fAdrenal	REMC	d108	Adrenal	GSM817167	0.313	28,128,420	27,626,714	136,447
fAdrenal	REMC	d101	Adrenal	GSM1027311	0.291	44,006,828	43,589,652	180,997

Continued on next page

Table 6.2 – *Continued from previous page*

Cell- or tissue-type	Source	Age#	Tissue	GEO Accession	SPOT	Tags	Tags (-chrM)	DHS peaks (FDR 1%)
fAdrenal	REMC	d113	Adrenal	GSM878658	0.358	36,200,139	34,885,120	199,720
fBrain	REMC	d122	Brain	GSM530651	0.720	25,521,213	25,472,137	182,501
fBrain	REMC	d117	Brain	GSM595920	0.590	25,078,460	25,016,845	195,888
fBrain	REMC	d85	Brain	GSM595923	0.405	23,665,122	23,608,044	184,688
fBrain	REMC	d96	Brain	GSM595928	0.555	20,987,395	20,887,542	177,090
fBrain	REMC	d112	Brain	GSM665804	0.386	34,750,897	34,671,585	191,014
fBrain	REMC	d142	Brain	GSM665819	0.434	34,909,221	34,462,349	167,734
fBrain	REMC	d101	Brain	GSM878650	0.416	30,829,149	30,786,898	216,130
fBrain	REMC	d104	Brain	GSM878651	0.584	35,036,073	34,982,802	191,232
fBrain	REMC	d109	Brain	GSM878652	0.436	24,953,459	24,905,787	178,475
fBrain	REMC	d105	Brain	GSM1027328	0.490	35,021,221	34,978,828	204,455
fHeart	REMC	d96	Heart	GSM530654	0.580	25,623,855	25,253,759	173,135
fHeart	REMC	d101	Heart	GSM530661	0.540	32,812,462	26,756,542	209,039
fHeart	REMC	d117	Heart	GSM665809	0.574	20,497,225	20,237,506	157,105
fHeart	REMC	d103	Heart	GSM665814	0.560	24,283,747	23,937,451	172,946
fHeart	REMC	d147	Heart	GSM665824	0.412	34,249,454	33,857,207	187,545
fHeart	REMC	d110	Heart	GSM665830	0.649	36,948,524	36,362,253	189,064
fHeart	REMC	d105	Heart	GSM774203	0.600	57,235,297	56,289,045	220,074
fHeart	REMC	d120	Heart	GSM878630	0.596	39,316,934	38,082,008	217,926
fHeart	REMC	d91	Heart	GSM817220	0.585	34,557,454	33,668,945	202,827
fIntestine_Lg	REMC	d103	Intestine	GSM665815	0.343	25,940,579	25,753,257	148,014
fIntestine_Lg	REMC	d105	Intestine	GSM665818	0.367	26,525,135	26,385,840	165,904
fIntestine_Lg	REMC	d110	Intestine	GSM665826	0.391	36,983,856	36,509,010	174,469
fIntestine_Lg	REMC	d107	Intestine	GSM701490	0.357	21,461,793	21,194,682	165,770

Continued on next page

Table 6.2 – *Continued from previous page*

Cell- or tissue-type	Source	Age#	Tissue	GEO Accession	SPOT	Tags	Tags (-chrM)	DHS peaks (FDR 1%)
fIntestine_Lg	REMC	d108	Intestine	GSM701495	0.439	27,932,860	27,773,196	84,894
fIntestine_Lg	REMC	d115	Intestine	GSM774213	0.391	22,437,785	22,188,540	147,422
fIntestine_Lg	REMC	d113	Intestine	GSM774214	0.296	53,885,997	53,430,571	174,813
fIntestine_Lg	REMC	d91	Intestine	GSM774220	0.439	21,201,047	21,106,664	154,797
fIntestine_Lg	REMC	d120	Intestine	GSM701531	0.349	29,832,797	29,469,112	166,626
fIntestine_Lg	REMC	d98	Intestine	GSM774228	0.453	81,556,735	80,494,968	286,072
fIntestine_Sm	REMC	d110	Intestine	GSM665825	0.412	36,648,756	36,283,133	182,941
fIntestine_Sm	REMC	d115	Intestine	GSM701487	0.406	19,998,953	19,734,999	136,731
fIntestine_Sm	REMC	d105	Intestine	GSM665835	0.406	40,799,365	40,095,822	193,223
fIntestine_Sm	REMC	d87	Intestine	GSM817161	0.428	29,166,499	28,923,298	172,118
fIntestine_Sm	REMC	d91	Intestine	GSM774205	0.476	22,479,708	22,244,548	173,146
fIntestine_Sm	REMC	d107	Intestine	GSM774210	0.401	23,494,853	23,287,680	162,640
fIntestine_Sm	REMC	d108	Intestine	GSM701496	0.402	24,154,098	23,959,814	149,240
fIntestine_Sm	REMC	d120	Intestine	GSM774225	0.343	63,431,414	63,044,896	189,437
fIntestine_Sm	REMC	d98	Intestine	GSM774229	0.417	35,105,452	34,694,378	183,531
fKidney	REMC	d122	Kidney	GSM530655	0.450	26,390,341	26,243,980	168,058
fKidney	REMC	d121	Kidney	GSM878666	0.380	24,583,989	24,482,089	185,281
fKidney	REMC	d105	Kidney	GSM1027329	0.452	23,680,869	23,598,291	207,896
fKidney	REMC	d85	Kidney	GSM1027342	0.585	35,548,444	35,467,952	240,229
fKidney_L	REMC	d147	Kidney	GSM665822	0.398	29,828,784	29,783,443	174,558
fKidney_L	REMC	d110	Kidney	GSM665829	0.356	35,168,628	35,102,532	192,878
fKidney_L	REMC	d115	Kidney	GSM665834	0.620	26,541,650	26,356,072	218,057
fKidney_L	REMC	d98	Kidney	GSM1027336	0.431	49,114,616	48,912,663	238,901
fKidney_R	REMC	d117	Kidney	GSM665810	0.360	35,268,222	35,173,250	194,782

Continued on next page

Table 6.2 – *Continued from previous page*

Cell- or tissue-type	Source	Age#	Tissue	GEO Accession	SPOT	Tags	Tags (-chrM)	DHS peaks (FDR 1%)
fKidney_R	REMC	d107	Kidney	GSM817163	0.437	24,749,718	24,724,172	177,763
fKidney_R	REMC	d87	Kidney	GSM1027346	0.519	27,437,943	27,358,757	204,630
fKidney_renal_cortex	REMC	d108	Kidney	GSM701502	0.540	25,114,272	25,032,574	202,065
fKidney_renal_cortex	REMC	d113	Kidney	GSM701511	0.583	29,566,179	29,481,922	241,258
fKidney_renal_cortex	REMC	d120	Kidney	GSM701532	0.440	29,888,464	29,710,664	204,864
fKidney_renal_cortex	REMC	d97	Kidney	GSM878629	0.536	19,863,251	19,749,761	180,904
fKidney_renal_cortex	REMC	d96	Kidney	GSM1027316	0.519	27,413,186	27,173,032	199,449
fKidney_renal_cortex	REMC	d89	Kidney	GSM878667	0.491	34,431,625	34,140,860	227,044
fKidney_renal_pelvis	REMC	d91	Kidney	GSM774222	0.426	38,430,223	38,227,807	198,240
fKidney_renal_pelvis	REMC	d127	Kidney	GSM817177	0.302	31,071,466	30,965,110	175,309
fKidney_renal_pelvis	REMC	d103	Kidney	GSM878662	0.438	51,210,201	50,856,330	247,500
fLung	REMC	d122	Lung	GSM530656	0.470	25,021,977	24,742,635	157,966
fLung	REMC	d101	Lung	GSM530662	0.670	24,455,569	24,229,966	220,166
fLung	REMC	d103	Lung	GSM595916	0.400	41,845,529	41,604,307	196,107
fLung	REMC	d67	Lung	GSM595921	0.416	20,778,962	20,718,549	166,407
fLung	REMC	d85	Lung	GSM595924	0.662	24,489,273	24,450,572	205,880
fLung	REMC	d96	Lung	GSM595927	0.554	17,950,380	17,899,596	158,920
fLung	REMC	d112	Lung	GSM665805	0.471	36,058,035	35,938,309	199,936
fLung	REMC	d82	Lung	GSM665806	0.401	37,705,535	37,364,344	175,444
fLung_L	REMC	d110	Lung	GSM665828	0.496	38,447,839	38,319,871	186,870
fLung_L	REMC	d113	Lung	GSM701512	0.505	39,104,651	38,962,497	187,442
fLung_L	REMC	d108	Lung	GSM701524	0.581	21,555,838	21,499,339	195,089
fLung_L	REMC	d115	Lung	GSM774237	0.663	89,759,962	89,203,598	233,911
fLung_L	REMC	d87	Lung	GSM1027345	0.668	40,500,507	40,386,730	189,797

Continued on next page

Table 6.2 – *Continued from previous page*

Cell- or tissue-type	Source	Age#	Tissue	GEO Accession	SPOT	Tags	Tags (-chrM)	DHS peaks (FDR 1%)
fLung_R	REMC	d117	Lung	GSM665807	0.410	28,258,863	28,164,047	166,666
fLung_R	REMC	d91	Lung	GSM774206	0.563	38,209,949	38,084,356	194,569
fLung_R	REMC	d107	Lung	GSM817164	0.552	30,722,966	30,611,703	181,390
fLung_R	REMC	d98	Lung	GSM774227	0.485	71,593,015	71,347,894	218,885
fLung_R	REMC	d105	Lung	GSM774231	0.546	53,568,139	53,408,427	203,664
fMuscle_arm	REMC	d115	Muscle	GSM701506	0.437	29,500,649	29,372,970	196,225
fMuscle_arm	REMC	d91	Muscle	GSM774223	0.478	90,984,367	90,581,737	324,206
fMuscle_back	REMC	d98	Muscle	GSM701536	0.517	29,018,101	28,803,299	216,313
fMuscle_back	REMC	d105	Muscle	GSM774235	0.478	89,523,543	89,154,405	243,666
fMuscle_back	REMC	d85	Muscle	GSM817217	0.5	31,601,635	31,512,300	207,352
fMuscle_back	REMC	d104	Muscle	GSM878639	0.472	71,300,684	66,777,257	248,101
fMuscle_leg	REMC	d127	Muscle	GSM774242	0.556	26,120,290	25,999,443	191,850
fMuscle_leg	REMC	d96	Muscle	GSM878626	0.509	41,739,601	41,562,780	213,588
fMuscle_leg	REMC	d97	Muscle	GSM817213	0.471	25,731,993	25,477,370	206,751
fMuscle_leg	REMC	d101	Muscle	GSM878631	0.488	35,011,994	34,889,731	207,372
fMuscle_leg	REMC	d113	Muscle	GSM878653	0.588	37,504,372	37,317,264	222,306
fMuscle_trunk	REMC	d120	Muscle	GSM701533	0.521	30,627,558	30,400,821	204,765
fMuscle_trunk	REMC	d121	Muscle	GSM878664	0.368	40,333,900	40,149,841	185,668
fOvary	REMC	pool	Ovary	GSM1027306	0.303	28,055,073	27,948,213	135,870
fPlacenta	REMC	d108	Placenta	GSM774215	0.32	74,216,981	73,680,118	233,243
fPlacenta	REMC	d91	Placenta	GSM774219	0.38	28,484,130	28,356,431	183,711
fPlacenta	REMC	d85	Placenta	GSM817219	0.523	32,115,085	31,947,808	198,726
fPlacenta	REMC	d113	Placenta	GSM878659	0.43	28,287,203	28,136,595	184,044
fPlacenta	REMC	d105	Placenta	GSM1027332	0.566	41,258,148	40,911,915	213,832

Continued on next page

Table 6.2 – *Continued from previous page*

Cell- or tissue-type	Source	Age#	Tissue	GEO Accession	SPOT	Tags	Tags (-chrM)	DHS peaks (FDR 1%)
fRetina	REMC	d125	Retina	submitted	0.7	43,144,177	43,011,675	184,852
fRetina	REMC	d87	Retina	submitted	0.49	35,442,562	34,942,536	209,105
fRetina	REMC	d103	Retina	submitted	0.41	46,845,714	46,713,047	343,803
fSpinal_cord	REMC	d105	Spinal cord	GSM817189	0.324	41,246,058	41,088,815	227,684
fSpinal_cord	REMC	d96	Spinal cord	GSM1027308	0.307	32,060,944	31,943,861	170,412
fSpinal_cord	REMC	d113	Spinal cord	GSM878661	0.405	39,555,509	39,344,297	205,017
fSpinal_cord	REMC	d89	Spinal cord	GSM878663	0.371	35,237,430	34,879,741	191,578
fSpinal_cord	REMC	d87	Spinal cord	GSM1027339	0.367	22,834,430	22,683,314	167,471
fSpleen	REMC	d112	B cell	GSM701509	0.324	28,603,001	28,216,178	172,219
fStomach	REMC	d147	Stomach	GSM774202	0.325	29,750,284	29,366,538	157,567
fStomach	REMC	d107	Stomach	GSM774212	0.357	25,534,879	24,955,355	164,391
fStomach	REMC	d91	Stomach	GSM701528	0.269	21,448,212	21,054,457	131,796
fStomach	REMC	d98	Stomach	GSM701538	0.427	30,676,457	30,139,257	202,942
fStomach	REMC	d105	Stomach	GSM774232	0.346	28,486,554	28,351,708	168,103
fStomach	REMC	d127	Stomach	GSM817173	0.277	30,624,637	30,374,867	146,317
fStomach	REMC	d101	Stomach	GSM817199	0.286	35,346,480	34,995,534	163,730
fStomach	REMC	d96	Stomach	GSM1027318	0.301	57,000,342	56,662,485	185,823
fStomach	REMC	d108	Stomach	GSM878660	0.368	30,410,923	30,126,473	177,326
fStomach	REMC	d121	Stomach	GSM878665	0.387	30,430,524	30,060,424	180,573
fThymus	REMC	d147	T cell	GSM665823	0.315	34,790,401	34,559,542	130,756
fThymus	REMC	d105	T cell	GSM774204	0.343	28,903,878	28,696,303	116,826
fThymus	REMC	d98	T cell	GSM774230	0.341	77,445,578	76,852,917	150,360
fThymus	REMC	d127	T cell	GSM817172	0.293	84,341,630	83,920,582	161,208
fThymus	REMC	d104	T cell	GSM1027313	0.336	41,442,960	41,201,419	129,323

Continued on next page

Table 6.2 – *Continued from previous page*

Cell- or tissue-type	Source	Age#	Tissue	GEO Accession	SPOT	Tags	Tags (-chrM)	DHS peaks (FDR 1%)
fThymus	REMC	d108	T cell	GSM878656	0.381	30,036,394	29,784,612	113,568
fThymus	REMC	d113	T cell	GSM878657	0.396	28,428,579	28,166,498	120,968
GM04503D	ENCODE		Fibroblast	GSM1024777	0.646	42,011,708	34,097,581	200,043
GM04504A	ENCODE		Fibroblast	GSM1024775	0.751	35,761,055	31,272,576	190,772
GM06990	ENCODE		B cell	GSM736558	0.546	22,440,189	19,729,256	92,709
GM12864	ENCODE		B cell	GSM736525	0.475	29,163,780	22,094,200	137,656
GM12865	ENCODE		B cell	GSM736512	0.525	37,660,121	34,216,112	143,716
GM12878	ENCODE		B cell	GSM736496	0.5	26,277,477	22,759,410	117,684
H7_hESC_T14	ENCODE		ES	GSM736638	0.372	33,507,208	26,763,822	140,102
H7_hESC_T2	ENCODE		ES	GSM736638	0.286	24,392,212	24,354,266	156,697
H7_hESC_T5	ENCODE		ES	GSM736638	0.343	44,134,819	42,890,856	211,653
HAc	ENCODE		Brain	GSM736586	0.422	45,892,152	40,763,545	180,083
HAEpiC	ENCODE		Placenta	GSM736631	0.764	31,211,767	29,842,070	205,033
HAh	ENCODE		Brain	GSM736594	0.485	32,927,492	27,040,382	200,014
HAsp	ENCODE		Spinal cord	GSM736537	0.425	39,414,103	34,517,745	194,537
HBMEC	ENCODE		Blood vessel	GSM736509	0.543	45,157,036	35,967,071	199,815
HBVP	ENCODE		Blood vessel	GSM1024750	0.359	47,747,677	45,100,866	209,369
HBVSMC	ENCODE		Blood vessel	GSM1024768	0.355	26,910,879	25,087,761	160,148
HCFaa	ENCODE		Heart	GSM736494	0.518	32,427,458	29,403,643	184,440
HCF	ENCODE		Heart	GSM736568	0.688	35,153,888	27,752,878	174,667
HCM	ENCODE		Heart	GSM736516	0.721	38,668,400	30,918,209	193,375
HConF	ENCODE		Fibroblast	GSM736547	0.506	36,022,255	32,701,475	153,668
HCPEpiC	ENCODE		Brain	GSM736569	0.742	26,199,536	24,538,293	210,380
HCT116	ENCODE		Intestine	GSM736600	0.454	39,022,377	26,710,185	114,060

Continued on next page

Table 6.2 – *Continued from previous page*

Cell- or tissue-type	Source	Age#	Tissue	GEO Accession	SPOT	Tags	Tags (-chrM)	DHS peaks (FDR 1%)
HEEpiC	ENCODE		Esophagael	GSM736585	0.569	39,750,750	31,034,571	209,838
Hela	ENCODE		Cervical	GSM736564	0.579	36,519,330	26,208,222	123,470
HEPG2	ENCODE		Liver	GSM736637	0.57	22,126,494	18,183,774	90,775
HESC	ENCODE		ES	GSM736582	0.358	24,431,583	20,890,242	150,729
hESCT0	ENCODE		ES	GSM736638	0.635	33,752,751	27,784,900	266,618
HFF	ENCODE		Fibroblast	GSM736602	0.545	46,526,167	37,514,733	192,282
HFF_MyC	ENCODE		Fibroblast	GSM736524	0.484	39,972,176	29,812,887	209,807
HGF	ENCODE		Fibroblast	GSM736579	0.483	30,751,179	30,148,924	145,887
HIPEpiC	ENCODE		Iris	GSM736589	0.56	32,033,391	26,191,760	225,744
HL60	ENCODE		Myeloid	GSM736626	0.589	33,148,093	32,066,521	161,716
HMEC	ENCODE		Mammary	GSM736634	0.425	43,782,139	31,914,218	140,574
HMF	ENCODE		Fibroblast	GSM736628	0.798	33,118,548	27,188,138	179,452
HMVEC_dAd	ENCODE		Blood vessel	GSM1024745	0.377	28,492,574	25,367,629	125,234
HMVEC_dBlAd	ENCODE		Blood vessel	GSM736609	0.726	46,050,351	42,643,211	162,593
HMVEC_dBlNeo	ENCODE		Blood vessel	GSM736571	0.529	54,866,074	49,657,528	168,436
HMVEC_dLyAd	ENCODE		Blood vessel	GSM736599	0.575	48,098,604	33,370,795	127,713
HMVEC_dLyNeo	ENCODE		Blood vessel	GSM736577	0.578	44,427,289	41,368,217	153,107
HMVEC_dNeo	ENCODE		Blood vessel	GSM736611	0.586	41,076,964	29,813,741	141,037
HMVEC_LBl	ENCODE		Blood vessel	GSM736542	0.485	50,131,824	47,911,926	169,983
HMVEC_LLy	ENCODE		Blood vessel	GSM736507	0.605	52,381,779	41,515,731	144,886
HNPCEpiC	ENCODE		Blood vessel	GSM736621	0.605	30,449,370	20,884,189	212,433
HPAEC	ENCODE		Blood vessel	GSM1024763	0.298	33,475,651	32,804,748	123,918
HPAF	ENCODE		Blood vessel	GSM736555	0.716	54,924,185	47,644,186	188,071
HPdLF	ENCODE		Fibroblast	GSM736632	0.686	25,814,740	23,004,769	171,349

Continued on next page

Table 6.2 – *Continued from previous page*

Cell- or tissue-type	Source	Age#	Tissue	GEO Accession	SPOT	Tags	Tags (-chrM)	DHS peaks (FDR 1%)
HPF	ENCODE		Lung	GSM736574	0.672	30,693,222	27,902,228	154,397
HRCE	ENCODE		Kidney	GSM736549	0.657	26,420,059	24,358,945	192,147
HRE	ENCODE		Kidney	GSM736527	0.534	25,850,984	22,900,558	187,131
HRGEC	ENCODE		Kidney	GSM736499	0.424	20,987,760	19,623,147	141,865
HRPEpiC	ENCODE		Retina	GSM736630	0.741	38,479,119	27,376,806	227,086
HSMM_D	ENCODE		Muscle	GSM736530	0.498	34,007,549	28,658,027	221,749
HSMM	ENCODE		Muscle	GSM736560	0.637	44,844,225	37,479,300	228,282
hTH17	ENCODE		T cell	GSM1024790	0.268	18,274,006	17,865,182	77,514
hTH1	ENCODE		T cell	GSM1024753	0.333	20,742,528	20,459,985	82,245
hTH2	ENCODE		T cell	GSM1024739	0.689	43,122,486	42,711,501	128,977
hTR	ENCODE		T cell	GSM1024744	0.504	40,015,882	37,326,516	125,848
HUVEC	ENCODE		Blood vessel	GSM736575	0.401	23,090,713	20,957,203	119,094
HVMF	ENCODE		Connective	GSM736534	0.591	22,052,455	19,811,379	170,340
Jurkat	ENCODE		T cell	GSM736501	0.497	67,808,791	62,773,068	159,613
K562	ENCODE		Erythroid	GSM736629	0.542	35,811,565	22,416,086	142,986
LHCN_M2_D4	ENCODE		Muscle	GSM1024787	0.727	64,634,390	37,759,025	218,246
LHCN_M2	ENCODE		Muscle	GSM1024787	0.709	37,903,192	30,643,938	192,347
LNCap	ENCODE		Prostate	GSM736565	0.62	49,624,105	28,747,933	183,224
M059J	ENCODE		Brain	GSM1024773	0.69	61,673,188	56,087,660	220,835
MCF7	ENCODE		Mammary	GSM736581	0.683	44,891,456	32,444,710	207,878
MCF7	ENCODE		Mammary	GSM736581	0.437	29,211,804	22,787,939	126,717
MCF7_ER	ENCODE		Mammary	GSM736581	0.652	63,497,188	45,615,312	220,065
NB4	ENCODE		Myeloid	GSM736604	0.531	36,021,761	32,684,492	143,838
NHA	ENCODE		Brain	GSM736544	0.562	40,038,155	34,637,438	191,510

Continued on next page

Table 6.2 – *Continued from previous page*

Cell- or tissue-type	Source	Age#	Tissue	GEO Accession	SPOT	Tags	Tags (-chrM)	DHS peaks (FDR 1%)
NHBE_RA	ENCODE		Broncheal	GSM1024781	0.344	41,688,982	29,967,781	149,972
NHDF_Ad	ENCODE		Fibroblast	GSM736567	0.805	44,829,222	44,170,209	230,696
NHDF_Neo	ENCODE		Fibroblast	GSM736498	0.698	38,935,890	33,799,898	187,962
NHEK	ENCODE		Skin	GSM736545	0.357	27,995,767	24,307,712	145,203
NHLF	ENCODE		Lung	GSM736612	0.706	41,241,537	33,520,502	206,254
NT2_D1	ENCODE		ES	GSM1024751	0.351	36,885,055	33,804,957	184,238
PANC1	ENCODE		Pancreas	GSM736517	0.418	26,460,252	22,683,675	116,642
PrEC	ENCODE		Prostate	GSM1024742	0.323	35,291,903	32,675,776	167,623
RPMI_7951	ENCODE		Skin	GSM1024779	0.682	38,648,543	34,251,134	167,310
RPTEC	ENCODE		Kidney	GSM736543	0.487	35,166,095	22,996,745	169,261
SAEC	ENCODE		Broncheal	GSM736608	0.62	24,221,968	22,257,395	198,442
SKMC	ENCODE		Muscle	GSM736593	0.801	28,517,820	28,117,588	205,493
SK_N_MC	ENCODE		Brain	GSM736522	0.353	26,552,446	24,428,899	146,328
SKNSH	ENCODE		Brain	GSM736559	0.622	19,593,615	18,614,810	89,968
T_47D	ENCODE		Mammary	GSM1024762	0.584	55,382,458	41,173,122	153,537
WERI_Rb1	ENCODE		Retina	GSM736495	0.546	35,473,957	33,340,703	191,374
WI_38	ENCODE		Lung	GSM736613	0.7	26,765,352	20,405,387	166,381
WI_38_TAM	ENCODE		Lung	GSM736613	0.622	28,554,163	24,833,322	205,334

y = years post-natal, d = days post-conception, blank = primary, cultured or malignant cells

Table 6.3: **Conservation of mouse DHSs in human by cell- and tissue-type**. Conservation statistics for DHSs identified within individual mouse tissues

		Not aligned		Not DHS in human		DHS in human		
Cell- or tissue-type	**Age***	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**Total peaks#**
E14	E0	80,583	45.5%	26,273	14.8%	70,120	39.6%	176,976
CJ7	E0	75,270	47.4%	22,106	13.9%	61,331	38.6%	158,707
ZhBTc4	E0	77,919	45.8%	25,009	14.7%	67,164	39.5%	170,092
ESC (F4WT)	E0	82,365	46.1%	26,526	14.8%	69,807	39.1%	178,698
416b	I	58,933	42.7%	22,653	16.4%	56,363	40.9%	137,949
MEL	I	60,654	45.3%	21,290	15.9%	52,055	38.8%	133,999
Erythroblast (CD117+; CD71-; Ter119-)	E14.5	61,617	38.3%	24,796	15.4%	74,311	46.2%	160,724
Erythroblast (CD117+; CD71+; Ter119-)	E14.5	52,228	41.9%	18,410	14.8%	53,914	43.3%	124,552
Erythroblast (CD117-; CD71+; Ter119+)	E14.5	29,920	40.2%	10,006	13.5%	34,460	46.3%	74,386
Erythroblast (CD117+; CD71+; Ter119+)	E14.5	42,430	40.7%	14,691	14.1%	47,232	45.3%	104,353
Na?ve T cell (activated)	W8	34,496	37.2%	12,749	13.7%	45,558	49.1%	92,803
Regualtory T cell (activated)	W8	41,758	39.4%	15,106	14.2%	49,249	46.4%	106,113

Continued on next page

Table 6.3 – *Continued from previous page*

		Not aligned		Not DHS in human		DHS in human		
Cell- or tissue-type	**Age***	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**Total peaks#**
Na?ve T cell (resting)	W8	40,697	39.7%	15,604	15.2%	46,262	45.1%	102,563
Regulatory T cell (resting)	W8	38,518	38.2%	13,688	13.6%	48,598	48.2%	100,804
B cell (CD19+)	W8	38,256	37.9%	14,083	13.9%	48,643	48.2%	100,982
B cell (CD53+)	W8	38,280	39.2%	13,590	13.9%	45,769	46.9%	97,639
Liver	E14.5	41,251	38.3%	15,036	13.9%	51,537	47.8%	107,824
Liver	E14.5	41,073	38.2%	14,842	13.8%	51,734	48.1%	107,649
Liver	W8	42,346	37.9%	20,548	18.4%	48,890	43.7%	111,784
A20 (lymphoma)	I	54,396	42.6%	20,896	16.4%	52,336	41.0%	127,628
Mammary adenocarcinoma	I	29,491	32.4%	11,840	13.0%	49,635	54.6%	90,966
Whole brain	E18.5	39,091	21.9%	31,667	17.8%	107,361	60.3%	178,119
Whole brain	W8	62,899	28.8%	43,684	20.0%	112,014	51.2%	218,597
Cerebrum	W8	60,346	28.2%	48,480	22.7%	105,164	49.1%	213,990
Cerebellum	W8	32,608	28.5%	18,803	16.4%	62,965	55.1%	114,376
Lung	W8	59,285	33.4%	32,120	18.1%	85,947	48.5%	177,352
Kidney	W8	59,542	34.8%	30,698	17.9%	80,980	47.3%	171,220
Fat pad (mammary)	W8	55,297	36.5%	28,524	18.8%	67,563	44.6%	151,384
Fibroblast (tail)	I	52,603	31.2%	29,090	17.3%	86,887	51.5%	168,580
Fibroblast (NIH3T3)	I	47,016	33.8%	21,600	15.5%	70,671	50.7%	139,287

Continued on next page

Table 6.3 – *Continued from previous page*

		Not aligned		Not DHS in human		DHS in human		
Cell- or tissue-type	**Age***	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**Total peaks#**
Fibroblast (embryonic kidney)	I	47,897	32.4%	26,985	18.2%	73,010	49.4%	147,892
Heart	W8	45,619	29.5%	24,139	15.6%	85,013	54.9%	154,771
Embyro (headless)	E11.5	40,219	26.4%	21,259	13.9%	90,981	59.7%	152,459
Embryo (forelimb buds)	E11.5	45,168	28.1%	24,222	15.1%	91,401	56.8%	160,791
Fat pad (genital)	W8	67,057	37.1%	35,318	19.6%	78,197	43.3%	180,572
Embryo (hindlimb buds)	E11.5	44,267	28.1%	24,695	15.7%	88,798	56.3%	157,760
Embryo (mesoderm)	E11.5	43,181	26.6%	24,325	15.0%	94,953	58.4%	162,459
Intestine (large)	W8	53,391	38.6%	23,019	16.6%	62,061	44.8%	138,471
Muscle (skeletal)	W8	41,230	29.5%	20,894	15.0%	77,526	55.5%	139,650
Retina	W8	30,844	29.0%	14,902	14.0%	60,469	56.9%	106,215
Retina	D7	33,666	28.3%	15,969	13.4%	69,356	58.3%	118,991
Retina	D0	36,126	26.9%	19,810	14.7%	78,419	58.4%	134,355
Spleen	W8	36,278	38.1%	13,352	14.0%	45,499	47.8%	95,129
Thymus	W8	44,140	38.9%	16,902	14.9%	52,478	46.2%	113,520
CH12 (lymphoma)	I	63,397	42.1%	26,139	17.3%	61,138	40.6%	150,674

* E = days post-conception, D = days post-natal, W = weeks post-natal, I = immortalized or malignant

Total peaks from master peaks list (see Methods)

Table 6.4: **Conservation of human DHSs in mouse by cell- and tissue-type**. Conservation statistics for DHSs identified within individual human tissues

		Not aligned		Not DHS in mouse		DHS in mouse		
Cell- or tissue-type	**Age***	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**Total peaks#**
A549		56,327	38.8%	36,445	25.1%	52,537	36.2%	145,309
AG04449		73,878	36.6%	58,713	29.1%	69,302	34.3%	201,893
AG04450		65,008	35.8%	49,633	27.3%	67,180	36.9%	181,821
AG09309		89,264	36.9%	70,819	29.3%	81,928	33.9%	242,011
AG09319		58,682	33.7%	47,490	27.3%	67,822	39.0%	173,994
AG10803		73,912	35.6%	58,341	28.1%	75,129	36.2%	207,382
AoAF		75,624	35.7%	58,858	27.8%	77,085	36.4%	211,567
BE2_C		76,347	36.9%	58,379	28.2%	72,261	34.9%	206,987
BJ		70,919	35.8%	58,337	29.4%	69,043	34.8%	198,299
CACO2		52,579	35.6%	41,067	27.8%	54,211	36.7%	147,857
CD14		86,352	43.0%	53,618	26.7%	61,053	30.4%	201,023
CD19	y34	37,781	35.9%	23,176	22.0%	44,230	42.0%	105,187
CD20		44,868	36.0%	27,592	22.1%	52,157	41.9%	124,617
CD34	y33	64,748	38.5%	41,169	24.5%	62,410	37.1%	168,327
CD34		75,036	38.6%	47,615	24.5%	71,881	37.0%	194,532
CD4	y37	40,534	35.3%	24,588	21.4%	49,685	43.3%	114,807
CD4pos_N		35,699	36.3%	22,677	23.0%	40,083	40.7%	98,459
CMK		66,174	41.8%	38,850	24.5%	53,261	33.6%	158,285
fAdrenal	d96	81,661	35.9%	64,442	28.3%	81,421	35.8%	227,524
fAdrenal	d85	77,024	34.2%	63,716	28.3%	84,713	37.6%	225,453
fAdrenal	d108	58,885	35.9%	43,321	26.4%	61,722	37.7%	163,928
fAdrenal	d101	76,831	35.4%	60,523	27.9%	79,806	36.7%	217,160

Continued on next page

Table 6.4 – *Continued from previous page*

		Not aligned		Not DHS in mouse		DHS in mouse		
Cell- or tissue-type	**Age***	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**Total peaks#**
fAdrenal	d113	86,410	36.5%	69,264	29.2%	81,146	34.3%	236,820
fBrain	d122	65,968	30.6%	64,895	30.1%	84,531	39.2%	215,394
fBrain	d117	70,308	30.1%	71,318	30.6%	91,777	39.3%	233,403
fBrain	d85	63,723	28.1%	67,364	29.7%	95,625	42.2%	226,712
fBrain	d96	64,562	30.4%	62,892	29.6%	84,916	40.0%	212,370
fBrain	d112	68,107	29.1%	75,146	32.1%	90,532	38.7%	233,785
fBrain	d142	57,884	28.4%	63,803	31.3%	82,264	40.3%	203,951
fBrain	d101	74,876	29.0%	79,855	30.9%	103,390	40.1%	258,121
fBrain	d104	68,753	30.4%	66,989	29.6%	90,316	40.0%	226,058
fBrain	d109	61,976	29.2%	63,351	29.8%	87,251	41.0%	212,578
fBrain	d105	73,427	30.1%	73,343	30.1%	97,025	39.8%	243,795
fHeart	d96	70,589	33.0%	59,591	27.9%	83,492	39.1%	213,672
fHeart	d101	87,103	33.9%	75,635	29.4%	94,348	36.7%	257,086
fHeart	d117	63,750	32.2%	56,484	28.6%	77,564	39.2%	197,798
fHeart	d103	71,673	33.1%	65,427	30.2%	79,361	36.7%	216,461
fHeart	d147	77,034	33.2%	73,569	31.7%	81,337	35.1%	231,940
fHeart	d110	78,676	33.5%	67,543	28.8%	88,482	37.7%	234,701
fHeart	d105	90,590	34.2%	75,571	28.5%	98,821	37.3%	264,982
fHeart	d120	88,246	33.4%	77,863	29.5%	98,246	37.2%	264,355
fHeart	d91	83,176	33.7%	70,734	28.7%	92,972	37.7%	246,882
fIntestine_Lg	d103	61,072	32.2%	48,760	25.7%	79,643	42.0%	189,475
fIntestine_Lg	d105	75,869	36.6%	54,478	26.3%	76,969	37.1%	207,316
fIntestine_Lg	d110	79,221	36.2%	59,820	27.3%	79,759	36.5%	218,800
fIntestine_Lg	d107	74,910	36.3%	56,451	27.3%	75,225	36.4%	206,586

Continued on next page

Table 6.4 – *Continued from previous page*

		Not aligned		Not DHS in mouse		DHS in mouse		
Cell- or tissue-type	**Age***	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**Total peaks#**
fIntestine_Lg	d108	37,563	34.6%	27,122	25.0%	43,890	40.4%	108,575
fIntestine_Lg	d115	62,644	34.0%	46,380	25.2%	75,358	40.9%	184,382
fIntestine_Lg	d113	79,272	37.3%	57,025	26.8%	76,378	35.9%	212,675
fIntestine_Lg	d91	67,186	35.1%	48,357	25.3%	75,831	39.6%	191,374
fIntestine_Lg	d120	76,130	37.0%	55,063	26.8%	74,358	36.2%	205,551
fIntestine_Lg	d98	134,881	39.7%	98,046	28.8%	107,213	31.5%	340,140
fIntestine_Sm	d110	83,624	36.4%	63,576	27.6%	82,777	36.0%	229,977
fIntestine_Sm	d115	60,130	35.3%	41,929	24.6%	68,262	40.1%	170,321
fIntestine_Sm	d105	88,488	37.0%	66,555	27.9%	83,810	35.1%	238,853
fIntestine_Sm	d87	74,051	34.9%	54,681	25.8%	83,166	39.2%	211,898
fIntestine_Sm	d91	76,506	35.9%	55,677	26.1%	81,038	38.0%	213,221
fIntestine_Sm	d107	69,329	34.4%	51,749	25.7%	80,171	39.8%	201,249
fIntestine_Sm	d108	69,186	36.9%	49,896	26.6%	68,172	36.4%	187,254
fIntestine_Sm	d120	84,929	36.7%	61,618	26.6%	84,730	36.6%	231,277
fIntestine_Sm	d98	82,490	36.4%	59,834	26.4%	84,068	37.1%	226,392
fKidney	d122	67,815	31.8%	55,748	26.1%	89,780	42.1%	213,343
fKidney	d121	72,854	31.3%	65,120	28.0%	94,674	40.7%	232,648
fKidney	d105	84,763	32.6%	74,115	28.5%	100,989	38.9%	259,867
fKidney	d85	103,037	34.9%	83,422	28.3%	108,643	36.8%	295,102
fKidney_L	d147	69,030	30.8%	65,513	29.2%	89,463	39.9%	224,006
fKidney_L	d110	78,574	31.9%	74,341	30.2%	93,040	37.8%	245,955
fKidney_L	d115	91,843	33.5%	75,945	27.7%	106,496	38.8%	274,284
fKidney_L	d98	99,602	33.8%	86,395	29.3%	108,716	36.9%	294,713
fKidney_R	d117	80,948	32.3%	73,424	29.3%	96,164	38.4%	250,536

Continued on next page

Table 6.4 – *Continued from previous page*

		Not aligned		Not DHS in mouse		DHS in mouse		
Cell- or tissue-type	**Age***	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**Total peaks#**
fKidney_R	d107	69,926	31.4%	60,148	27.0%	92,751	41.6%	222,825
fKidney_R	d87	84,389	33.0%	70,543	27.6%	100,749	39.4%	255,681
fKidney_renal_cortex	d108	82,569	32.3%	72,489	28.3%	100,743	39.4%	255,801
fKidney_renal_cortex	d113	102,935	34.1%	87,225	28.9%	111,276	36.9%	301,436
fKidney_renal_cortex	d120	84,310	32.5%	73,978	28.5%	100,858	38.9%	259,146
fKidney_renal_cortex	d97	71,790	31.6%	60,590	26.7%	94,514	41.7%	226,894
fKidney_renal_cortex	d96	80,546	32.3%	68,368	27.4%	100,344	40.3%	249,258
fKidney_renal_cortex	d89	94,585	33.5%	81,256	28.8%	106,134	37.6%	281,975
fKidney_renal_pelvis	d91	82,257	33.3%	66,083	26.8%	98,685	39.9%	247,025
fKidney_renal_pelvis	d127	69,085	31.2%	61,380	27.7%	91,090	41.1%	221,555
fKidney_renal_pelvis	d103	103,589	34.0%	89,644	29.5%	111,127	36.5%	304,360
fLung	d122	63,460	32.0%	53,397	26.9%	81,302	41.0%	198,159
fLung	d101	94,196	34.5%	81,299	29.8%	97,612	35.7%	273,107
fLung	d103	82,352	33.5%	75,419	30.7%	88,087	35.8%	245,858
fLung	d67	68,319	32.5%	57,978	27.6%	83,720	39.9%	210,017
fLung	d85	85,272	33.2%	76,009	29.6%	95,765	37.3%	257,046
fLung	d96	62,810	31.4%	56,107	28.1%	80,912	40.5%	199,829
fLung	d112	83,034	33.2%	76,799	30.7%	90,608	36.2%	250,441
fLung	d82	71,530	32.3%	67,830	30.6%	82,387	37.2%	221,747
fLung_L	d110	77,121	32.6%	70,644	29.9%	88,642	37.5%	236,407
fLung_L	d113	77,822	32.7%	68,552	28.8%	91,683	38.5%	238,057
fLung_L	d108	79,252	32.6%	71,016	29.2%	92,935	38.2%	243,203
fLung_L	d115	97,920	34.5%	81,141	28.6%	104,725	36.9%	283,786
fLung_L	d87	76,916	33.1%	67,133	28.9%	88,630	38.1%	232,679

Continued on next page

Table 6.4 – *Continued from previous page*

		Not aligned		Not DHS in mouse		DHS in mouse		
Cell- or tissue-type	**Age***	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**Total peaks#**
fLung_R	d117	66,759	31.5%	61,657	29.1%	83,796	39.5%	212,212
fLung_R	d91	78,225	32.6%	68,454	28.5%	93,602	39.0%	240,281
fLung_R	d107	71,058	31.6%	64,334	28.6%	89,454	39.8%	224,846
fLung_R	d98	91,338	33.8%	79,063	29.2%	100,088	37.0%	270,489
fLung_R	d105	82,470	32.8%	72,146	28.7%	96,787	38.5%	251,403
fMuscle_arm	d115	76,837	31.4%	75,098	30.6%	93,131	38.0%	245,066
fMuscle_arm	d91	140,585	36.8%	120,026	31.4%	121,621	31.8%	382,232
fMuscle_back	d98	85,784	31.9%	79,664	29.6%	103,328	38.4%	268,776
fMuscle_back	d105	97,212	32.8%	88,002	29.7%	110,872	37.4%	296,086
fMuscle_back	d85	80,229	31.6%	75,674	29.8%	97,798	38.5%	253,701
fMuscle_back	d104	103,912	34.6%	90,807	30.3%	105,186	35.1%	299,905
fMuscle_leg	d127	73,330	31.0%	67,296	28.4%	95,923	40.6%	236,549
fMuscle_leg	d96	82,734	31.8%	78,408	30.1%	99,001	38.1%	260,143
fMuscle_leg	d97	79,719	31.4%	76,368	30.1%	97,614	38.5%	253,701
fMuscle_leg	d101	80,365	31.6%	74,766	29.4%	98,995	39.0%	254,126
fMuscle_leg	d113	87,071	32.2%	81,261	30.0%	102,386	37.8%	270,718
fMuscle_trunk	d120	81,120	31.9%	74,126	29.2%	98,762	38.9%	254,008
fMuscle_trunk	d121	70,422	30.8%	65,285	28.6%	92,650	40.6%	228,357
fOvary	pool	54,725	32.8%	40,692	24.4%	71,443	42.8%	166,860
fPlacenta	d108	116,138	42.4%	73,834	27.0%	83,665	30.6%	273,637
fPlacenta	d91	90,994	41.8%	55,110	25.3%	71,338	32.8%	217,442
fPlacenta	d85	101,630	44.2%	60,218	26.2%	68,195	29.6%	230,043
fPlacenta	d113	91,694	42.6%	55,904	26.0%	67,825	31.5%	215,423
fPlacenta	d105	109,086	44.4%	65,882	26.8%	70,684	28.8%	245,652

Continued on next page

Table 6.4 – *Continued from previous page*

		Not aligned		Not DHS in mouse		DHS in mouse		
Cell- or tissue-type	**Age***	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**Total peaks#**
fRetina	d125	72,230	32.2%	62,616	27.9%	89,552	39.9%	224,398
fRetina	d87	78,004	30.6%	73,024	28.6%	104,094	40.8%	255,122
fRetina	d103	127,649	32.1%	140,437	35.4%	129,095	32.5%	397,181
fSpinal_cord	d105	77,694	27.9%	91,064	32.7%	109,827	39.4%	278,585
fSpinal_cord	d96	56,804	26.8%	60,443	28.6%	94,399	44.6%	211,646
fSpinal_cord	d113	72,777	29.0%	71,650	28.5%	106,887	42.5%	251,314
fSpinal_cord	d89	63,565	27.1%	69,648	29.7%	101,447	43.2%	234,660
fSpinal_cord	d87	53,882	26.2%	60,496	29.4%	91,642	44.5%	206,020
fSpleen	d112	75,576	35.2%	57,468	26.7%	81,904	38.1%	214,948
fStomach	d147	61,626	31.3%	49,890	25.3%	85,604	43.4%	197,120
fStomach	d107	65,459	31.6%	52,623	25.4%	89,089	43.0%	207,171
fStomach	d91	52,436	31.0%	41,716	24.7%	74,922	44.3%	169,074
fStomach	d98	83,927	32.7%	68,202	26.6%	104,575	40.7%	256,704
fStomach	d105	65,861	31.4%	54,465	26.0%	89,306	42.6%	209,632
fStomach	d127	59,019	32.1%	44,425	24.2%	80,389	43.7%	183,833
fStomach	d101	64,719	31.5%	53,793	26.2%	87,116	42.4%	205,628
fStomach	d96	74,012	31.9%	61,046	26.3%	96,665	41.7%	231,723
fStomach	d108	69,215	31.2%	58,683	26.4%	94,240	42.4%	222,138
fStomach	d121	73,677	32.7%	58,625	26.0%	93,026	41.3%	225,328
fThymus	d147	60,164	37.3%	40,987	25.4%	60,059	37.3%	161,210
fThymus	d105	48,894	34.2%	33,197	23.2%	60,748	42.5%	142,839
fThymus	d98	68,133	37.9%	45,099	25.1%	66,619	37.0%	179,851
fThymus	d127	73,567	38.2%	49,173	25.5%	69,990	36.3%	192,730
fThymus	d104	57,055	36.7%	37,830	24.3%	60,581	39.0%	155,466

Continued on next page

Table 6.4 – *Continued from previous page*

		Not aligned		Not DHS in mouse		DHS in mouse		
Cell- or tissue-type	**Age***	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**Total peaks#**
fThymus	d108	48,765	35.6%	31,881	23.3%	56,443	41.2%	137,089
fThymus	d113	51,506	35.3%	34,246	23.5%	60,161	41.2%	145,913
GM04503D		89,876	37.4%	70,603	29.4%	79,709	33.2%	240,188
GM04504A		84,495	37.1%	65,604	28.8%	77,926	34.2%	228,025
GM06990		43,530	38.0%	27,508	24.0%	43,604	38.0%	114,642
GM12864		67,850	41.4%	40,417	24.6%	55,753	34.0%	164,020
GM12865		69,664	41.0%	42,297	24.9%	57,913	34.1%	169,874
GM12878		55,743	39.3%	34,566	24.3%	51,686	36.4%	141,995
H7_hESC_T14		59,597	34.5%	44,000	25.4%	69,296	40.1%	172,893
H7_hESC_T2		60,813	30.8%	52,290	26.5%	84,526	42.8%	197,629
H7_hESC_T5		101,067	40.4%	70,850	28.3%	78,116	31.2%	250,033
HAc		77,240	35.4%	61,257	28.0%	79,958	36.6%	218,455
HAEpiC		97,121	39.8%	69,829	28.6%	76,895	31.5%	243,845
HAh		86,423	35.2%	70,941	28.9%	88,082	35.9%	245,446
HAsp		88,285	37.4%	75,303	31.9%	72,765	30.8%	236,353
HBMEC		92,189	38.3%	69,819	29.0%	78,658	32.7%	240,666
HBVP		96,603	38.5%	74,378	29.6%	80,001	31.9%	250,982
HBVSMC		67,366	34.3%	54,369	27.7%	74,805	38.1%	196,540
HCFaa		86,530	39.2%	64,883	29.4%	69,583	31.5%	220,996
HCF		76,528	35.7%	57,601	26.9%	80,296	37.4%	214,425
HCM		87,385	37.2%	64,247	27.4%	83,160	35.4%	234,792
HConF		64,326	34.1%	49,361	26.2%	74,723	39.7%	188,410
HCPEpiC		97,182	38.4%	72,293	28.5%	83,826	33.1%	253,301
HCT116		53,125	38.8%	31,591	23.1%	52,187	38.1%	136,903

Continued on next page

Table 6.4 – *Continued from previous page*

		Not aligned		Not DHS in mouse		DHS in mouse		
Cell- or tissue-type	**Age***	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**Total peaks#**
HEEpiC		101,853	41.6%	68,683	28.1%	74,232	30.3%	244,768
Hela		59,123	39.5%	39,411	26.3%	51,264	34.2%	149,798
HEPG2		39,242	35.0%	26,199	23.4%	46,656	41.6%	112,097
HESC		71,678	39.3%	45,463	24.9%	65,350	35.8%	182,491
hESCT0		136,414	44.1%	85,660	27.7%	86,969	28.1%	309,043
HFF		85,055	36.4%	65,221	27.9%	83,648	35.8%	233,924
HFF_MyC		98,289	38.6%	70,575	27.7%	85,532	33.6%	254,396
HGF		62,573	35.1%	48,782	27.4%	66,760	37.5%	178,115
HIPEpiC		107,400	39.9%	78,797	29.3%	82,905	30.8%	269,102
HL60		81,343	43.2%	47,679	25.3%	59,430	31.5%	188,452
HMEC		63,060	37.1%	46,803	27.5%	60,317	35.4%	170,180
HMF		80,142	37.3%	59,590	27.7%	75,340	35.0%	215,072
HMVEC_dAd		51,193	33.7%	37,593	24.7%	63,319	41.6%	152,105
HMVEC_dBlAd		71,178	36.5%	51,785	26.6%	71,870	36.9%	194,833
HMVEC_dBlNeo		74,224	36.6%	54,039	26.6%	74,553	36.8%	202,816
HMVEC_dLyAd		53,129	34.4%	39,295	25.5%	61,810	40.1%	154,234
HMVEC_dLyNeo		65,563	35.4%	48,152	26.0%	71,372	38.6%	185,087
HMVEC_dNeo		59,515	34.7%	43,238	25.2%	68,778	40.1%	171,531
HMVEC_LBl		75,549	37.0%	55,622	27.2%	73,266	35.8%	204,437
HMVEC_LLy		61,126	35.1%	44,974	25.8%	67,968	39.0%	174,068
HNPCEpiC		98,777	38.8%	74,578	29.3%	80,926	31.8%	254,281
HPAEC		52,042	34.6%	38,140	25.4%	60,126	40.0%	150,308
HPAF		84,471	37.1%	61,520	27.1%	81,433	35.8%	227,424
HPdLF		74,324	35.8%	59,192	28.5%	73,981	35.7%	207,497

Continued on next page

Table 6.4 – *Continued from previous page*

		Not aligned		Not DHS in mouse		DHS in mouse		
Cell- or tissue-type	**Age***	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**Total peaks#**
HPF		65,833	34.7%	50,672	26.7%	73,258	38.6%	189,763
HRCE		88,432	38.3%	65,038	28.1%	77,596	33.6%	231,066
HRE		85,205	37.6%	63,436	28.0%	78,219	34.5%	226,860
HRGEC		63,575	37.0%	47,061	27.4%	61,400	35.7%	172,036
HRPEpiC		100,268	37.4%	80,359	30.0%	87,440	32.6%	268,067
HSMM_D		101,244	37.5%	80,135	29.7%	88,246	32.7%	269,625
HSMM		104,311	38.4%	83,219	30.7%	83,901	30.9%	271,431
hTH17		35,035	37.7%	22,567	24.3%	35,419	38.1%	93,021
hTH1		35,919	36.6%	22,049	22.5%	40,238	41.0%	98,206
hTH2		59,737	40.0%	36,764	24.6%	52,780	35.4%	149,281
hTR		57,337	38.3%	34,943	23.3%	57,424	38.4%	149,704
HUVEC		50,074	33.7%	39,793	26.8%	58,633	39.5%	148,500
HVMF		78,271	37.4%	58,697	28.1%	72,042	34.5%	209,010
Jurkat		80,694	43.8%	47,148	25.6%	56,320	30.6%	184,162
K562		75,406	44.5%	41,067	24.3%	52,826	31.2%	169,299
LHCN_M2_D4		96,994	38.0%	72,396	28.3%	86,072	33.7%	255,462
LHCN_M2		87,242	38.2%	64,367	28.2%	76,491	33.5%	228,100
LNCap		86,111	41.2%	61,243	29.3%	61,736	29.5%	209,090
M059J		104,568	40.4%	73,394	28.3%	81,008	31.3%	258,970
MCF7		103,258	42.7%	65,227	27.0%	73,364	30.3%	241,849
MCF7		62,412	41.2%	40,681	26.8%	48,493	32.0%	151,586
MCF7_ER		110,407	43.3%	69,194	27.2%	75,090	29.5%	254,691
NB4		70,437	41.2%	41,264	24.1%	59,308	34.7%	171,009
NHA		85,719	36.9%	68,675	29.5%	78,051	33.6%	232,445

Continued on next page

Table 6.4 – *Continued from previous page*

		Not aligned		Not DHS in mouse		DHS in mouse		
Cell- or tissue-type	**Age***	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**# peaks**	**% peaks**	**Total peaks#**
NHBE_RA		68,025	37.4%	46,201	25.4%	67,562	37.2%	181,788
NHDF_Ad		102,754	37.7%	80,642	29.6%	88,807	32.6%	272,203
NHDF_Neo		80,333	35.4%	66,749	29.4%	80,125	35.3%	227,207
NHEK		67,911	39.1%	47,560	27.4%	58,321	33.6%	173,792
NHLF		92,846	37.4%	74,196	29.9%	81,323	32.7%	248,365
NT2_D1		88,796	41.0%	55,529	25.6%	72,255	33.4%	216,580
PANC1		55,251	38.7%	36,159	25.3%	51,479	36.0%	142,889
PrEC		81,066	40.9%	55,146	27.9%	61,781	31.2%	197,993
RPMI_7951		74,892	37.3%	53,533	26.7%	72,143	36.0%	200,568
RPTEC		79,909	39.2%	56,747	27.8%	67,167	33.0%	203,823
SAEC		96,014	41.2%	66,279	28.5%	70,540	30.3%	232,833
SKMC		91,154	36.8%	73,780	29.8%	82,738	33.4%	247,672
SK_N_MC		63,753	37.0%	52,338	30.3%	56,359	32.7%	172,450
SKNSH		31,456	28.0%	28,969	25.8%	52,062	46.3%	112,487
T_47D		71,740	40.0%	45,070	25.1%	62,636	34.9%	179,446
WERI_Rb1		88,488	40.7%	63,636	29.3%	65,350	30.0%	217,474
WI_38		75,170	37.4%	56,069	27.9%	69,727	34.7%	200,966
WI_38_TAM		97,277	39.9%	71,270	29.2%	75,185	30.8%	243,732

* y = years post-natal, d = days post-conception, blank = primary, cultured or malignant cells

Total peaks from master peaks list (see Methods)

REFERENCES

1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D. et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, October 2010.

Affymetrix ENCODE Transcriptome Project and Cold Spring Harbor Laboratory ENCODE Transcriptome Project. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature*, 457(7232):1028–1032, February 2009.

Anderson, M.K., Hernandez-Hoyos, G., Diamond, R.A. and Rothenberg, E.V. Precise developmental regulation of Ets family transcription factors during specification and commitment to the T cell lineage. *Development (Cambridge, England)*, 126(14):3131–3148, June 1999.

Arthur, D. and Vassilvitskii, S. *k-means++: the advantages of careful seeding*. Society for Industrial and Applied Mathematics, January 2007.

Axel, R. Cleavage of DNA in nuclei and chromatin with staphylococcal nuclease. *Biochemistry*, 14(13):2921–2925, July 1975.

Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R. et al. Diversity and complexity in DNA recognition by transcription factors. *Science*, 324(5935): 1720–1723, June 2009.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E. et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, 37(Web Server issue): W202–8, July 2009.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E. et al. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, May 2007.

Benjamini, Y. and Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B. Statistical Methodology.*, 57(1):298–300, 1995.

Bernstein, B.E., Humphrey, E.L., Erlich, R.L., Schneider, R., Bouman, P. et al. Methylation of histone H3 Lys 4 in coding regions of active genes. *Proceedings of the National Academy of Sciences of the United States of America*, 99(13):8695–8700, June 2002.

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A. et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology*, 28(10): 1045–1048, October 2010.

Bernstein, F., Koetzle, T. and Williams, G. The protein data bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, 1977.

Biddie, S.C., John, S., Sabo, P.J., Thurman, R.E., Johnson, T.A. et al. Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. *Molecular cell*, 43(1):145–155, July 2011.

Biggin, M.D. Animal transcription networks as highly connected, quantitative con-

tinua. *Developmental cell*, 21(4):611–626, October 2011.

Bininda-Emonds, O.R.P., Cardillo, M., Jones, K.E., MacPhee, R.D.E., Beck, R.M.D. et al. The delayed rise of present-day mammals. *Nature*, 446(7135):507–512, March 2007.

Bourque, G., Leong, B., Vega, V.B., Chen, X., Lee, Y.L. et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Research*, 18 (11):1752–1762, November 2008.

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H. et al. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322, January 2008.

Boyle, A.P., Song, L., Lee, B.K., London, D., Keefe, D. et al. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Research*, 21(3):456–464, March 2011.

Bryne, J.C., Valen, E., Tang, M.H.E., Marstrand, T., Winther, O. et al. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Research*, 36(Database issue):D102–6, January 2008.

Buratowski, S., Hahn, S., Guarente, L. and Sharp, P.A. Five intermediate complexes in transcription initiation by RNA polymerase II. *Cell*, 56(4):549–561, February 1989.

Chan, J.Y., Han, X.L. and Kan, Y.W. Cloning of Nrf1, an NF-E2-related transcription factor, by genetic selection in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 90(23): 11371–11375, December 1993.

Crawford, G.E., Holt, I.E., Whittle, J., Webb, B.D., Tai, D. et al. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Research*, 16(1):123–131, January 2006.

Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. WebLogo: a sequence logo generator. *Genome Research*, 14(6):1188–1190, June 2004.

Cuddapah, S., Jothi, R., Schones, D.E., Roh, T.Y., Cui, K. et al. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Research*, 19 (1):24–32, January 2009.

Dai, S.M., Chen, H.H., Chang, C., Riggs, A.D. and Flanagan, S.D. Ligation-mediated PCR for quantitative in vivo footprinting. *Nature biotechnology*, 18(10):1108–1111, October 2000.

de Wit, E. and de Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes & Development*, 26(1):11–24, January 2012.

Dermitzakis, E.T. and Clark, A.G. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Molecular and cellular biology*, 19(7):1114–1121, July 2002.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T. et al. Landscape of transcription in human cells. *Nature*, 489(7414): 101–108, September 2012.

Dorschner, M.O., Hawrylycz, M., Humbert, R., Wallace, J.C., Shafer, A. et al. High-throughput localization of functional elements by quantitative chromatin profiling. *Nature methods*, 1(3):219–225, December 2004.

Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L. et al. Human genome sequencing using unchained base

reads on self-assembling DNA nanoarrays. *Science*, 327(5961):78–81, January 2010.

Dynan, W.S. and Tjian, R. The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. *Cell*, 35(1):79–87, November 1983.

Edmondson, D.G., Lyons, G.E., Martin, J.F. and Olson, E.N. Mef2 gene expression marks the cardiac and skeletal muscle lineages during mouse embryogenesis. *Development (Cambridge, England)*, 120(5):1251–1263, May 1994.

ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R. et al. Identification and analysis of functional elements in 1human genome by the ENCODE pilot project. *Nature*, 447(7146): 799–816, June 2007.

ENCODE Project Consortium, Bernstein, B.E., Birney, E., Dunham, I., Green, E.D. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489 (7414):57–74, September 2012.

Ernst, P., Wang, J., Huang, M., Goodman, R.H. and Korsmeyer, S.J. MLL and CREB Bind Cooperatively to the Nuclear Coactivator CREB-Binding Protein. *Molecular and cellular biology*, 21(7):2249–2258, April 2001.

Felsenfeld, G., Boyes, J., Chung, J., Clark, D. and Studitsky, V. Chromatin structure and gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 93(18):9384–9388, September 1996.

Ferré-D'Amaré, A.R., Pognonec, P., Roeder, R.G. and Burley, S.K. Structure and function of the b/HLH/Z domain of USF. *EMBO Journal*, 13(1):180–189, January 1994.

Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K. et al. Ensembl 2013. *Nucleic Acids Research*, 41(Database issue):D48–55, January 2013.

Forsberg, E.C., Downs, K.M. and Bresnick, E.H. Direct interaction of NF-E2 with hypersensitive site 2 of the beta-globin locus control region in living cells. *Blood*, 96(1): 334–339, July 2000.

Fu, Y., Sinha, M., Peterson, C.L. and Weng, Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genetics*, 4(7):e1000138, 2008.

Gaffney, D.J., McVicker, G., Pai, A.A., Fondufe-Mittendorf, Y.N., Lewellen, N. et al. Controls of nucleosome positioning in the human genome. *PLoS Genetics*, 8(11): e1003036, 2012.

Galas, D.J. and Schmitz, A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Research*, 5(9):3157–3170, September 1978.

Galas, D.J. The invention of footprinting. *Trends in Biochemical Sciences*, 26(11):690–693, January 2001.

Garel, S., Marín, F., Grosschedl, R. and Charnay, P. Ebf1 controls early cell differentiation in the embryonic striatum. *Development (Cambridge, England)*, 126(23):5285–5294, December 1999.

Gilbert, W. and Müller-Hill, B. Isolation of the lac repressor. *Proceedings of the National Academy of Sciences of the United States of America*, 56(6):1891–1898, December 1966.

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

Gross, D.S. and Garrard, W.T. Nuclease hypersensitive sites in chromatin. *Annual review of biochemistry*, 57:159–197, 1988.

Grosveld, F., van Assendelft, G.B., Greaves, D.R. and Kollias, G. Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell*, 51(6): 975–985, December 1987.

Groudine, M. and Weintraub, H. Propagation of globin DNAase I-hypersensitive sites in absence of factors required for induction: a possible mechanism for determination. *Cell*, 30(1):131–139, August 1982.

Guénet, J.L. The mouse genome. *Genome Research*, 15(12):1729–1740, December 2005.

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

Halupa, A., Bailey, M.L., Huang, K., Iscove, N.N., Levy, D.E. et al. A novel role for STAT1 in regulating murine erythropoiesis: deletion of STAT1 results in overall reduction of erythroid progenitors and alters their distribution. *Blood*, 105(2):552–561, January 2005.

Hardouin, S.N. and Nagy, A. Mouse models for human disease. *Clinical genetics*, 57(4): 237–244, April 2000.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research*, 22(9):1760–1774, September 2012.

He, H.H., Meyer, C.A., Hu, S.S., Chen, M.W., Zang, C. et al. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nature methods*, 11(1):73–78, January 2014.

Hedges, S.B. The origin and evolution of model organisms. *Nature reviews Genetics*, 3 (11):838–849, November 2002.

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

Hesselberth, J.R., Chen, X., Zhang, Z., Sabo, P.J., Sandstrom, R. et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature methods*, 6(4): 283–289, April 2009.

Huang, G., Zhao, X., Wang, L., Elf, S., Xu, H. et al. The ability of MLL to bind RUNX1 and methylate H3K4 at PU.1 regulatory regions is impaired by MDS/AML-associated RUNX1/AML1 mutations. *Blood*, 118(25): 6544–6552, December 2011.

Jacques, P.É., Jeyakani, J. and Bourque, G. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genetics*, 9(5):e1003504, May 2013.

Jin, C., Zang, C., Wei, G., Cui, K., Peng, W. et al. H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. *Nature Genetics*, 41(8): 941–945, August 2009.

John, S., Sabo, P.J., Thurman, R.E., Sung, M.H., Biddie, S.C. et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Genetics*, 43(3): 264–268, March 2011.

John, S., Sabo, P.J., Canfield, T.K., Lee, K., Vong, S. et al. Genome-scale mapping of DNase I hypersensitivity. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, Chapter 27:Unit 21.27, July 2013.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R. et al. DNA-binding specificities of human transcription factors. *Cell*, 152 (1-2):327–339, January 2013.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research*, 110 (1-4):462–467, 2005.

Kim, T.K., Lagrange, T., Wang, Y.H., Griffith, J.D., Reinberg, D. et al. Trajectory of DNA in the RNA polymerase II transcription preinitiation complex. *Proceedings of the National Academy of Sciences of the United States of America*, 94(23):12268–12273, November 1997.

Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M.A. et al. A high-resolution map of active promoters in the human genome. *Nature*, 436(7052):876–880, August 2005.

Klug, A. and Lutter, L.C. The helical periodicity of DNA on the nucleosome. *Nucleic Acids Research*, 9(17):4267–4283, September 1981.

Ko, L.J. and Engel, J.D. DNA-binding specificities of the GATA transcription factor family. *Molecular and cellular biology*, 13(7): 4011–4022, July 1993.

Kornberg, R.D. and Stryer, L. Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. *Nucleic Acids Research*, 16(14A):6677–6690, July 1988.

Kundaje, A., Kyriazopoulou-Panagiotopoulou, S., Libbrecht, M., Smith, C.L., Raha, D. et al. Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Research*, 22(9):1735–1747, September 2012.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C. et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3): R25, 2009.

Lazarovici, A., Zhou, T., Shafer, A., Dantas Machado, A.C., Riley, T.R. et al. Probing DNA shape and methylation state on a genomic scale with DNase I. *Proceedings of the National Academy of Sciences of the United States of America*, 110(16):6376–6381, April 2013.

Lettice, L.A., Heaney, S.J.H., Purdie, L.A., Li, L., de Beer, P. et al. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human molecular genetics*, 12 (14):1725–1735, July 2003.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

Lin, S. and Riggs, A.D. The general affinity of lac repressor for E. coli DNA: implications for gene regulation in procaryotes and eucaryotes. *Cell*, 4(2):107–111, February 1975.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G. et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322, November 2009.

Lutter, L.C. Kinetic analysis of deoxyribonuclease I cleavages in the nucleosome core: evidence for a DNA superhelix. *Journal of Molecular Biology*, 124(2):391–420, September 1978.

Lutter, L.C. Precise location of DNase I cutting sites in the nucleosome core determined by high resolution gel electrophoresis. *Nucleic Acids Research*, 6(1):41, January 1979.

MacLean, B., Tomazela, D.M., Abbatiello, S.E., Zhang, S., Whiteaker, J.R. et al. Effect of collision energy optimization on the measurement of peptides by selected reaction monitoring (SRM) mass spectrometry. *Analytical Chemistry*, 82(24):10116–10124, December 2010a.

MacLean, B., Tomazela, D.M., Shulman, N., Chambers, M., Finney, G.L. et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*, 26(7):966–968, April 2010b.

Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, November 2010.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S. et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, 34(Database issue):D108–10, January 2006.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337 (6099):1190–1195, September 2012.

Mavrich, T.N., Ioshikhes, I.P., Venters, B.J., Jiang, C., Tomsho, L.P. et al. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Research*, 18(7):1073–1083, July 2008.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T. et al. GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology*, 28 (5):495–501, May 2010.

Mercer, T.R., Neph, S., Dinger, M.E., Crawford, J., Smith, M.A. et al. The human mitochondrial transcriptome. *Cell*, 146(4): 645–658, August 2011.

Merika, M. and Orkin, S.H. DNA-binding specificity of GATA family transcription factors. *Molecular and cellular biology*, 13(7): 3999–4010, July 1993.

Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M. et al. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Research*, 41 (Database issue):D64–9, January 2013.

Mirny, L.A. Nucleosome-mediated cooperativity between transcription factors. *Proceedings of the National Academy of Sciences of the United States of America*, 107(52):22534–22539, December 2010.

Mittler, G., Butter, F. and Mann, M. A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. *Genome Research*, 19(2):284–293, February 2009.

Montavon, T., Soshnikova, N., Mascrez, B., Joye, E., Thevenet, L. et al. A regulatory archipelago controls Hox genes transcription in digits. *Cell*, 147(5):1132–1145, November 2011.

Mouse Genome Sequencing Consortium, Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J. et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, December 2002.

Mukherjee, S., Berger, M.F., Jona, G., Wang, X.S., Muzzey, D. et al. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nature Genetics*, 36(12):1331–1339, December 2004.

Murphy, W.J., Pringle, T.H., Crider, T.A., Springer, M.S. and Miller, W. Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Research*, 17 (4):413–421, April 2007.

Nechanitzky, R., Akbas, D., Scherer, S., György, I., Hoyler, T. et al. Transcription factor EBF1 is essential for the maintenance of B cell identity and prevention of alternative fates in committed cells. *Nature immunology*, 14(8):867–875, August 2013.

Nei, M. and Li, W.H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, 76(10):5269–5273, October 1979.

Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E. et al. BEDOPS: high-performance genomic feature operations. *Bioinformatics*, 28(14):1919–1920, July 2012a.

Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E. et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90, September 2012b.

Newburger, D.E. and Bulyk, M.L. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Research*, 37 (Database issue):D77–82, January 2009.

Noll, M. Internal structure of the chromatin subunit. *Nucleic Acids Research*, 1(11): 1573–1578, November 1974.

Odom, D.T., Dowell, R.D., Jacobsen, E.S., Gordon, W., Danford, T.W. et al. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature Genetics*, 39(6):730–732, June 2007.

Okada, N., Sasaki, T., Shimogori, T. and Nishihara, H. Emergence of mammals by emergency: exaptation. *Genes to cells : devoted to molecular & cellular mechanisms*, 15(8): 801–812, August 2010.

Párraga, A., Bellsolell, L., Ferré-D'Amaré, A.R. and Burley, S.K. Co-crystal structure of sterol regulatory element binding protein 1a at 2.3 A resolution. *Structure*, 6(5):661–672, May 1998.

Pavlidis, P. and Noble, W.S. Matrix2png: a utility for visualizing matrix data. *Bioinformatics*, 19(2):295–296, January 2003.

Payne, J.L. and Wagner, A. The robustness and evolvability of transcription factor binding sites. *Science*, 343(6173):875–877, February 2014.

Pellegrini, L., Tan, S. and Richmond, T.J. Structure of serum response factor core bound to DNA. *Nature*, 376(6540):490–498, August 1995.

Peterson, K.R., Zitnik, G., Huxley, C., Lowrey, C.H., Gnirke, A. et al. Use of yeast artificial chromosomes (YACs) for studying control of gene expression: correct regulation of the genes of a human beta-globin locus YAC following transfer to mouse erythroleukemia cell lines. *Proceedings of the National Academy of Sciences of the United States of America*, 90(23):11207–11211, December 1993.

Pevny, L., Simon, M.C., Robertson, E., Klein, W.H., Tsai, S.F. et al. Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. *Nature*, 349(6306):257–260, January 1991.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. and Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, January 2010.

Prud'homme, B., Gompel, N., Rokas, A., Kassner, V.A., Williams, T.M. et al. Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature*, 440(7087):1050–1053, April 2006.

Pruitt, K.D., Tatusova, T., Klimke, W. and Maglott, D.R. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Research*, 37(Database issue): D32–6, January 2009.

Pugh, B.F. and Tjian, R. Transcription from a TATA-less promoter requires a multisubunit TFIID complex. *Genes & Development*, 5 (11):1935–1945, November 1991.

Reddy, P.M., Stamatoyannopoulos, G., Papayannopoulou, T. and Shen, C.K. Genomic footprinting and sequencing of human beta-globin locus. Tissue specificity and cell line artifact. *Journal of Biological Chemistry*, 269 (11):8287–8295, March 1994.

Rhee, H.S. and Pugh, B.F. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature*, 483(7389): 295–301, March 2012.

Rhodes, D. and Klug, A. Helical periodicity of DNA determined by enzyme digestion. *Nature*, 286(5773):573–578, August 1980.

Rockman, M.V. and Wray, G.A. Abundant raw material for cis-regulatory evolution in humans. *Molecular biology and evolution*, 19 (11):1991–2004, November 2002.

Rosenbloom, K.R., Dreszer, T.R., Long, J.C., Malladi, V.S., Sloan, C.A. et al. ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic Acids Research*, November 2011.

Sabo, P.J., Hawrylycz, M., Wallace, J.C., Humbert, R., Yu, M. et al. Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proceedings of the National Academy of Sciences of the United States of America*, 101(48):16837–16842, November 2004.

Sabo, P., Kuehn, M., Thurman, R., Johnson, B., Johnson, E. et al. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nature methods*, 3(7):511–518, 2006.

Santos-Rosa, H., Schneider, R., Bannister, A.J., Sherriff, J., Bernstein, B.E. et al. Active genes are tri-methylated at K4 of histone H3. *Nature*, 419(6905):407–411, September 2002.

Schachterle, W., Rojas, A., Xu, S.M. and Black, B.L. ETS-dependent regulation of a distal Gata4 cardiac enhancer. *Biophysical Journal*, 361(2):439–449, January 2012.

Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D. et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, 328(5981):1036–1040, May 2010.

Schoenherr, C.J. and Anderson, D.J. The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. *Science*, 267(5202):1360–1363, March 1995.

Schones, D.E., Cui, K., Cuddapah, S., Roh, T.Y., Barski, A. et al. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887–898, March 2008.

Shibata, Y., Sheffield, N.C., Fedrigo, O., Babbitt, C.C., Wortham, M. et al. Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. *PLoS Genetics*, 8(6):e1002789, June 2012.

Shields, J.M., Christy, R.J. and Yang, V.W. Identification and characterization of a gene encoding a gut-enriched Krüppel-like factor expressed during growth arrest. *Journal of Biological Chemistry*, 271(33):20009–20017, August 1996.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8): 1034–1050, August 2005.

Smit, A., Hubley, R. and Green, P. RepeatMasker Open-3.0 Software Package. 1996–2010. <http://www.repeatmasker.org>.

Socolovsky, M., Nam, H., Fleming, M.D., Haase, V.H., Brugnara, C. et al. Ineffective erythropoiesis in Stat5a(-/-)5b(-/-) mice due to decreased survival of early erythroblasts. *Blood*, 98(12):3261–3273, December 2001.

Stamatoyannopoulos, J.A., Goodwin, A., Joyce, T. and Lowrey, C.H. NF-E2 and GATA binding motifs are required for the formation of DNase I hypersensitive site 4 of the human beta-globin locus control region. *EMBO Journal*, 14(1):106–116, January 1995.

Staynov, D.Z. DNase I footprinting of the nucleosome in whole nuclei. *Biochemical and biophysical research communications*, 372(1): 226–229, July 2008.

Stefflova, K., Thybert, D., Wilson, M.D., Streeter, I. and Aleksic, J. Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell*, 154(3):530–540, August 2013.

Stergachis, A.B., MacLean, B., Lee, K., Stamatoyannopoulos, J.A. and MacCoss, M.J. Rapid empirical discovery of optimal peptides for targeted proteomics. *Nature methods*, 8(12):1041–1043, 2011.

Stergachis, A.B., Haugen, E., Shafer, A., Fu, W., Vernot, B. et al. Exonic transcription factor binding directs codon choice and affects protein evolution. *Science*, 342(6164): 1367–1372, December 2013a.

Stergachis, A.B., Neph, S., Reynolds, A., Humbert, R., Miller, B. et al. Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell*, 154 (4):888–903, August 2013b.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T. et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, 131(5): 861–872, November 2007.

Talbot, D. and Grosveld, F. The 5'HS2 of the globin locus control region enhances transcription through the interaction of a multimeric complex binding at two functionally distinct NF-E2 binding sites. *EMBO Journal*, 10(6):1391–1398, June 1991.

Tate, P.H. and Bird, A.P. Effects of DNA methylation on DNA-binding proteins and gene expression. *Current opinion in genetics & development*, 3(2):226–231, April 1993.

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

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T. et al. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, September 2012.

Treisman, R. and Maniatis, T. Simian virus 40 enhancer increases number of RNA polymerase II molecules on linked DNA. *Nature*, 315(6014):73–75, May 1985.

Tsai, S.F., Martin, D.I., Zon, L.I., D'Andrea, A.D., Wong, G.G. et al. Cloning of cDNA for the major DNA-binding protein of the erythroid lineage through expression in mammalian cells. *Nature*, 339(6224):446–451, June 1989.

Valouev, A., Johnson, S.M., Boyd, S.D., Smith, C.L., Fire, A.Z. et al. Determinants of

nucleosome organization in primary human cells. *Nature*, 474(7352):516–520, June 2011.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A. and Luscombe, N.M. A census of human transcription factors: function, expression and evolution. *Nature reviews Genetics*, 10(4):252–263, April 2009.

Vermeulen, M., Mulder, K.W., Denissov, S., Pijnappel, W.W.M.P., van Schaik, F.M.A. et al. Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell*, 131(1):58–69, October 2007.

Vernot, B., Stergachis, A.B., Maurano, M.T., Vierstra, J., Neph, S. et al. Personal and population genomics of human regulatory variation. *Genome Research*, 22(9):1689–1697, September 2012.

Vierstra, J., Wang, H., John, S., Sandstrom, R. and Stamatoyannopoulos, J.A. Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH. *Nature methods*, 11(1):66–72, January 2014.

Voss, T.C. and Hager, G.L. Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nature reviews Genetics*, 15(2):69–81, February 2014.

Wadman, I.A. The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. *EMBO Journal*, 16(11): 3145–3157, June 1997.

Ward, M.C., Wilson, M.D., Barbosa-Morais, N.L., Schmidt, D., Stark, R. et al. Latent regulatory potential of human-specific repetitive elements. *Molecular cell*, 49(2):262–272, January 2013.

Weintraub, H. Formation of stable transcription complexes as assayed by analysis of individual templates. *Proceedings of the National Academy of Sciences of the United States of America*, 85(16):5819–5823, August 1988.

Weisbrod, S. and Weintraub, H. Isolation of a subclass of nuclear proteins responsible for conferring a DNase I-sensitive structure on globin chromatin. *Proceedings of the National Academy of Sciences of the United States of America*, 76(2):630–634, February 1979.

Williams, T.M., Selegue, J.E., Werner, T., Gompel, N., Kopp, A. et al. The regulation and evolution of a genetic switch controlling sexually dimorphic traits in Drosophila. *Cell*, 134(4):610–623, August 2008.

Wilson, M.D., Barbosa-Morais, N.L., Schmidt, D., Conboy, C.M., Vanes, L. et al. Species-specific transcription in mice carrying human chromosome 21. *Science*, 322 (5900):434–438, October 2008.

Wingender, E., Dietze, P., Karas, H. and Knüppel, R. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Research*, 24(1):238–241, January 1996.

Workman, J.L., Abmayr, S.M., Cromlish, W.A. and Roeder, R.G. Transcriptional regulation by the immediate early protein of pseudorabies virus during in vitro nucleosome assembly. *Cell*, 55(2):211–219, October 1988.

Wu, C. The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. *Nature*, 286(5776):854–860, August 1980.

Xiong, Q., Zhang, Z., Chang, K.H., Qu, H., Wang, H. et al. Comprehensive characterization of erythroid-specific enhancers in the genomic regions of human Kruppel-like factors. *BMC genomics*, 14(1):587, August 2013.

Yamashita, R., Sathira, N.P., Kanai, A., Tanimoto, K., Arauchi, T. et al. Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. *Genome Research*, 21(5):775–789, May 2011.

Yun, K. and Wold, B. Skeletal muscle determination and differentiation: story of a core regulatory network and its context. *Current opinion in cell biology*, 8(6):877–889, December 1996.

Zhang, Y., Moqtaderi, Z., Rattner, B.P., Euskirchen, G., Snyder, M. et al. Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. *Nature structural & molecular biology*, 16 (8):847–852, August 2009.

Zhou, Z., Li, X., Deng, C., Ney, P.A., Huang, S. et al. USF and NF-E2 cooperate to regulate the recruitment and activity of RNA polymerase II in the beta-globin gene locus. *Journal of Biological Chemistry*, 285(21): 15894–15905, May 2010.

VITA

Jeffrey Vierstra received his B.S in Genetics from the University of Wisconsin. While not busy digesting human nuclei with DNase I, you can find him in woods and mountains of Washington.