Regulatory Elements and Gene Expression in
Primates and Diverse Human Cell-types
by
Nathan C. Sheffield
Graduate Program in Computational Biology and Bioinformatics
Duke University

Date: ________________________
Approved:

________________________
Gregory E. Crawford, Supervisor

________________________
Terrence S. Furey

________________________
Gregory A. Wray

________________________
Alexander J. Hartemink

________________________
Joseph E. Lucas

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Graduate Program in
Computational Biology and Bioinformatics
in the Graduate School of Duke University
2013
ABSTRACT

Regulatory Elements and Gene Expression in Primates and Diverse Human Cell-types

by

Nathan C. Sheffield

Graduate Program in Computational Biology and Bioinformatics
Duke University

Date: __________________________

Approved:

______________________________
Gregory E. Crawford, Supervisor

______________________________
Terrence S. Furey

______________________________
Gregory A. Wray

______________________________
Alexander J. Hartemink

______________________________
Joseph E. Lucas

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate Program in Computational Biology and Bioinformatics in the Graduate School of Duke University 2013
Abstract

After finishing a human genome reference sequence in 2002, the genomics community has turned to the task of interpreting it. A primary focus is to identify and characterize not only protein-coding genes, but all functional elements in the genome. The effort has identified millions of regulatory elements across species and in hundreds of human cell-types. Nearly all identified regulatory elements are found in non-coding DNA, hypothesizing a function for previously unannotated sequence. The ability to identify regulatory DNA genome-wide provides a new opportunity to understand gene regulation and to ask fundamental questions in diverse areas of biology.

One such area is the aim to understand the molecular basis for phenotypic differences between humans and other primates. These phenotypic differences are partially driven by mutations in non-coding regulatory DNA that alter gene expression. This hypothesis has been supported by differential gene expression analyses in general, but we have not yet identified specific regulatory variants responsible for differences in transcription and phenotype. I have worked to identify regulatory differences in the same cell-type isolated from human, chimpanzee, and macaque. Most regulatory elements were conserved among all three species, as expected based on their central role in regulating transcription. However, several hundred regulatory elements were gained or lost on the lineages leading to modern human and chimpanzee. Species-specific regulatory elements are enriched near differentially expressed genes, are positively correlated with increased transcription, show evidence of branch-specific positive selection, and overlap with active chromatin marks.
Species-specific sequence differences in transcription factor motifs found within this regulatory DNA are linked with species-specific changes in chromatin accessibility. Together, these indicate that species-specific regulatory elements contribute to transcriptional and phenotypic differences among primate species.

Another fundamental function of regulatory elements is to define different cell-types in multicellular organisms. Regulatory elements recruit transcription factors that modulate gene expression distinctly across cell-types. In a study of 112 human cell-types, I classified regulatory elements into clusters based on regulatory signal tissue specificity. I then used these to uncover distinct associations between regulatory elements and promoters, CpG-islands, conserved elements, and transcription factor motif enrichment. Motif analysis identified known and novel transcription factor binding motifs in cell-type-specific and ubiquitous regulatory elements. I also developed a classifier that accurately predicts cell-type lineage based on only 43 regulatory elements and evaluated the tissue of origin for cancer cell-types. By correlating regulatory signal and gene expression, I predicted target genes for more than 500k regulatory elements. Finally, I developed a web resource to enable researchers to explore these regulatory patterns and better understand how expression is modulated within and across human cell-types.

Regulation of gene expression is fundamental to life. This dissertation uses identified regulatory DNA to better understand regulatory systems. In the context of either evolutionary or developmental biology, understanding how differences in regulatory DNA contribute to phenotype will be central to completely understanding human biology.
Contents

Abstract iv
List of Tables x
List of Figures xi
List of Abbreviations and Symbols xiii
Acknowledgements xiv

1 Introduction 1

1.1 Background 4

1.1.1 Chromatin structure 4

1.1.2 Challenges to studying regulatory elements 5

1.1.3 Open chromatin and regulatory element assays 7

1.2 Identifying and characterizing regulatory elements 10

1.2.1 Open chromatin defines regulatory elements 11

1.2.2 Regulatory elements are located in noncoding regions 12

1.2.3 More than 30% of the genome may be regulatory 13

1.2.4 Most regulatory elements are cell-type-specific 13

1.2.5 Transcription factor binding affects chromatin structure 14

1.2.6 Chromatin structure affects gene expression 14

1.2.7 Regulatory elements can be classified by function 15

1.3 Motivation 17
2 Regulatory Elements Across Primates

2.1 Introduction .................................................. 20

2.2 Results ......................................................... 22
   2.2.1 DNase-seq identifies species-specific DHSs ....................... 22
   2.2.2 Species-specific DHSs are cell-type-specific ...................... 30
   2.2.3 Species-specific DHSs are associated with biological function . 34
   2.2.4 Species-specific DHSs are near species-specific genes .......... 35
   2.2.5 Species-specific DHSs show evolutionary selection and constraint 38
   2.2.6 DHS gains near DHS losses may indicate binding site turn-over . 41
   2.2.7 DHSs deleted in the human or chimp lineage ...................... 43

2.3 Discussion ..................................................... 46

2.4 Methods ....................................................... 50
   2.4.1 Cell-types and tissue culture .................................... 50
   2.4.2 DNaseI HS library preparation .................................. 51
   2.4.3 DNase-seq aligned to native genomes and lifted to hg19 ......... 51
   2.4.4 Identifying species-specific hypersensitive sites ................. 52
   2.4.5 Skin biopsy location and Fibroblast heterogeneity ............... 54
   2.4.6 Identifying DNaseI hypersensitive sites in all species (Common sites) 56
   2.4.7 Why are there more species-specific increases than decreases? .. 57
   2.4.8 Testing for selection ........................................... 58
   2.4.9 Testing for evolutionary constraint .............................. 60
   2.4.10 Expression analysis using digital gene expression (DGE-seq) ... 60
   2.4.11 Cross-species correlation of genome-wide expression .......... 61
   2.4.12 Identification of differentially expressed genes between species 62
   2.4.13 Correlation of species-specific DHSs with expression .......... 62
### 2.4.14 Variation in motifs in species-specific DHSs

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 Clusters of Regulatory Elements in Humans</td>
<td>65</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>65</td>
</tr>
<tr>
<td>3.2 Results</td>
<td>66</td>
</tr>
<tr>
<td>3.2.1 Clusters of regulatory elements drive cell-type specific expression</td>
<td>66</td>
</tr>
<tr>
<td>3.2.2 Expansion to 40 human cell-types</td>
<td>71</td>
</tr>
<tr>
<td>3.3 Discussion</td>
<td>72</td>
</tr>
<tr>
<td>3.4 Methods</td>
<td>72</td>
</tr>
<tr>
<td>4 Regulatory Elements in Humans Across Tissues</td>
<td>74</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>74</td>
</tr>
<tr>
<td>4.2 Results</td>
<td>77</td>
</tr>
<tr>
<td>4.2.1 DNaseI hypersensitive sites cluster cell-types by biological similarity</td>
<td>77</td>
</tr>
<tr>
<td>4.2.2 SOM clusters capture variation in CpG-island, promoter, and conservation</td>
<td>79</td>
</tr>
<tr>
<td>4.2.3 A logistic classifier predicts cell-type lineage with few DHS inputs</td>
<td>82</td>
</tr>
<tr>
<td>4.2.4 DHS clusters are enriched for known and novel TF motifs</td>
<td>85</td>
</tr>
<tr>
<td>4.2.5 Motifs in similar clusters reveal subtle motif differences</td>
<td>88</td>
</tr>
<tr>
<td>4.2.6 Motif discovery results are consistent with experimental ChIP data</td>
<td>90</td>
</tr>
<tr>
<td>4.2.7 Global TF trends suggest AP-1 is a chromatin-accessibility factor</td>
<td>90</td>
</tr>
<tr>
<td>4.2.8 Chromatin and expression correlation predicts long-range interaction</td>
<td>93</td>
</tr>
<tr>
<td>4.2.9 A web resource for exploring results</td>
<td>100</td>
</tr>
<tr>
<td>4.2.10 Clusters are enriched for biologically relevant pathways</td>
<td>101</td>
</tr>
<tr>
<td>4.3 Discussion</td>
<td>101</td>
</tr>
<tr>
<td>4.4 Methods</td>
<td>103</td>
</tr>
<tr>
<td>4.4.1 DNase-seq protocols</td>
<td>103</td>
</tr>
<tr>
<td>4.4.2 Data normalization and processing</td>
<td>104</td>
</tr>
</tbody>
</table>
4.4.3 Microarray protocol and normalization ....................... 104
4.4.4 Classifying regulatory elements with a self-organizing map .... 105
4.4.5 CpG-island, promoter, and conserved element overlap ........ 106
4.4.6 Tissue lineage identity classifier ................................. 107
4.4.7 Motif analysis .................................................... 107
4.4.8 General motif summary results ................................. 108
4.4.9 Discovered motifs in shared clusters ......................... 108
4.4.10 Validating motif discovery with ChIP results ................. 109
4.4.11 Mapping regulatory elements to target genes they regulate ... 110
4.4.12 Comparing DHS-gene correlation to experimental 3C evidence 110
4.4.13 POUF51/NANOG and ES-specific connections .............. 114
4.4.14 Beta-globin locus and olfactory receptor genes ............ 114
4.4.15 Pathway analysis .............................................. 115

5 Conclusion ......................................................... 116

5.1 The future of cross-species DHS comparisons .................. 117
5.2 The future of cross-cell-type comparison in humans ............ 118
5.3 Toward synthetic biology ........................................ 120

6 Appendix ......................................................... 122

References ......................................................... 125

Biography ......................................................... 149
List of Tables

2.1 Samples list ............................................ 21
2.2 Concordance between DNase-ChIP and DNase-seq ................. 24
2.3 ENCODE samples list .................................... 25
2.4 Fibroblast overlaps ..................................... 26
2.5 Lymphoblast overlaps ................................... 34
2.6 LCL overlap with histone modifications .......................... 34
2.7 Selection .................................................... 38
2.8 Fibroblast biopsy location overlaps ............................. 56
3.1 COREs and gene expression, PolIII, and CTCF binding ............ 67
List of Figures

1.1 Transcriptional regulation. ............................................. 2
1.2 Open chromatin .......................................................... 5
1.3 Genomic distribution of regulatory elements. ..................... 11
1.4 DHS saturation. .......................................................... 12
2.1 Comparison of DHSs and DGE-seq data across species .......... 23
2.2 Identification of species-specific differences in DHSs .......... 24
2.3 Representative examples of chimpanzee DHS gain and chimpanzee DHS loss 27
2.4 Comparison of species-specific DHS to independently derived cells ... 27
2.5 DHS Gains and Common overlap reproducibility ................ 28
2.6 Differential edgeR analysis of original human and chimpanzee samples to independently analyzed cell samples ............ 29
2.7 Comparison of human DHS gains and losses to DNase-seq data from other human cell-types ................................. 30
2.8 Summary comparison to other human cell-types ................. 31
2.9 Summary comparison to other human cell-types ................ 31
2.10 Comparison of chimpanzee DHS to other human cell-types ....... 32
2.11 Species-specific DHSs are associated with functional chromatin marks 33
2.12 Species-specific DHSs are correlated with expression and selection 36
2.13 Correlation between gene expression and DHS for fibroblasts ....... 37
2.14 Correlation between gene expression and DHS for lymphoblasts .... 38
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.15</td>
<td>DHS and Phastcons overlaps</td>
<td>40</td>
</tr>
<tr>
<td>2.16</td>
<td>Regulatory element shuffling p-values</td>
<td>41</td>
</tr>
<tr>
<td>2.17</td>
<td>Regulatory element shuffling examples</td>
<td>42</td>
</tr>
<tr>
<td>2.18</td>
<td>Functional mutations associated with DHS gains and losses</td>
<td>44</td>
</tr>
<tr>
<td>2.19</td>
<td>DNase-seq signal in cCONDEL regions</td>
<td>45</td>
</tr>
<tr>
<td>3.1</td>
<td>Clusters of Open Regulatory Elements (COREs)</td>
<td>68</td>
</tr>
<tr>
<td>3.2</td>
<td>Clusters of open regulatory elements and other signals</td>
<td>69</td>
</tr>
<tr>
<td>3.3</td>
<td>An ES-cell-specific CORE</td>
<td>70</td>
</tr>
<tr>
<td>3.4</td>
<td>An NHEK-cell-specific CORE</td>
<td>70</td>
</tr>
<tr>
<td>3.5</td>
<td>COREs from 40 cell-types</td>
<td>71</td>
</tr>
<tr>
<td>4.1</td>
<td>Batch correction results</td>
<td>77</td>
</tr>
<tr>
<td>4.2</td>
<td>SOM clustering of DHS profiles</td>
<td>78</td>
</tr>
<tr>
<td>4.3</td>
<td>Distribution of conservation, promoters, and CpG-islands</td>
<td>80</td>
</tr>
<tr>
<td>4.4</td>
<td>Detailed SOM results, and association of conservation, CpG islands, promoters, and tissue specificity</td>
<td>81</td>
</tr>
<tr>
<td>4.5</td>
<td>Tissue and sex classifiers based on DNaseI data</td>
<td>83</td>
</tr>
<tr>
<td>4.6</td>
<td>De novo motif discovery results</td>
<td>86</td>
</tr>
<tr>
<td>4.7</td>
<td>Complete motif enrichment</td>
<td>87</td>
</tr>
<tr>
<td>4.8</td>
<td>Variations in IRF-like motifs in hematopoietic clusters</td>
<td>89</td>
</tr>
<tr>
<td>4.9</td>
<td>Motif specificity in SOM clusters</td>
<td>91</td>
</tr>
<tr>
<td>4.10</td>
<td>Correlation between DHS and expression</td>
<td>93</td>
</tr>
<tr>
<td>4.11</td>
<td>Connections comparison and pathway enrichment</td>
<td>94</td>
</tr>
</tbody>
</table>
## List of Abbreviations and Symbols

### Abbreviations

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA</td>
<td>DeoxyriboNucleic Acid</td>
</tr>
<tr>
<td>bp</td>
<td>base pair (a single DNA nucleotide)</td>
</tr>
<tr>
<td>kb</td>
<td>kilobase (1,000 base pairs)</td>
</tr>
<tr>
<td>Mb</td>
<td>megabase (1,000,000 base pairs)</td>
</tr>
<tr>
<td>DHS</td>
<td>DNaseI Hypersensitive Site</td>
</tr>
<tr>
<td>ChIP</td>
<td>Chromatin ImmunoPrecipitation</td>
</tr>
<tr>
<td>ENCODE</td>
<td>ENCYclopedia Of DNA Elements</td>
</tr>
<tr>
<td>DGE</td>
<td>Digital Gene Expression</td>
</tr>
<tr>
<td>seq</td>
<td>sequence</td>
</tr>
<tr>
<td>CORE</td>
<td>Cluster of Open Regulatory Elements</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>SOM</td>
<td>Self-Organizing Map</td>
</tr>
<tr>
<td>TF</td>
<td>Transcription Factor</td>
</tr>
<tr>
<td>CpG</td>
<td>Cytosine-phosphate-Guanine</td>
</tr>
<tr>
<td>LCR</td>
<td>Locus Control Region</td>
</tr>
<tr>
<td>ICR</td>
<td>Imprinted Control Region</td>
</tr>
<tr>
<td>CNV</td>
<td>Copy Number Variation</td>
</tr>
<tr>
<td>nm</td>
<td>nanometer</td>
</tr>
</tbody>
</table>
Acknowledgements

Foremost, I thank my advisors, Terry Furey and Greg Crawford, for their guidance and patience through several years of sometimes exhilarating but sometimes disappointing results. I also thank Boris Lenhard, Greg Wray, Alex Hartemink, and Joseph Lucas, who have each spent hours teaching me and helping me improve myself and my project. I thank the other scientists who are co-authors with me on various publications; this dissertation is based on very collaborative projects and my effort is only meaningful in the context of the effort of others. Among these, I particularly wish to thank Yoichiro Shibata for his expertise and collaboration on the primate experiments. Finally, completing this thesis would not have been possible without the love and support of my family, especially of my wife Carolyn.

I am indebted to several granting institutions for funding my research; most notable among these is the National Science Foundation for the Graduate Research Fellowship that provided most of my support. I am also grateful to both the National Science Foundation and the Norwegian Research Council for the Nordic Research Opportunity, which enabled me to visit Norway to collaborate and gain valuable experience and skills with Boris Lenhard. I thank the Duke Graduate School for a teaching mini-grant, which helped improve my scientific writing skills and gave me an opportunity to teach others. The Duke Primate Genomics initiative generously provided summer fellowship funds that helped cover both experimental costs and computational equipment and improved my project. The NIH, through the ENCODE project, funded most of the expense for the data that I
analyzed. I appreciate conference travel funds from several sources, including the National Science Foundation, the Duke Graduate School, and an NIH training grant to the Duke University Graduate Program in Computational Biology and Bioinformatics. Finally, I acknowledge the leadership, administrators, and instructors of the Computational Biology and Bioinformatics Graduate Program for creating the system to teach and train me.
Introduction

The sequencing of the human genome has both facilitated progress and uncovered new challenges. The sequence itself is a trove of data benefiting diverse biological disciplines. Insights enabled by the human genome project are fueling the drive toward personalized medicine and impacting the diagnosis and treatment of human disease (Green and Guyer, 2011). Though our knowledge has increased considerably, there are many unanswered questions. Foremost is the task of converting so much data into meaningful information. With the genome sequence in hand, how do we make sense of what it says? Assigning even a basic function for the majority of the genome has still not been completed. Despite monumental progress, a primary goal remains to determine what the genome actually encodes. This is not a new goal. Attempts to discover the functional meaning of the whole human genome sequence dates to before the sequencing project began. About 40 years ago, with preliminary evidence, Ohno suggested that only 6% of the human genome
consists of genes and promoters (Ohno, 1972). He was not far off—more recent estimates put that number near 2–3% (Alexander et al., 2010; Lander et al., 2001). Since Ohno’s proposal, the community has debated the function of the remaining 94%+ of the genome: is it junk DNA (Alexander et al., 2010)? Simply decoding the complete sequence has not answered the question, partly because non-coding DNA lacks the information-rich genetic code that has made identifying protein-coding sequences possible. However, the community has now amassed experimental evidence that much noncoding, intergenic DNA is important for transcriptional regulation. These sequences regulate transcription by controlling when a gene is expressed with respect to, for example, cell-type, developmental

![Transcriptional regulation flowchart](image)

**Figure 1.1**: Transcriptional regulation flowchart showing how regulatory elements convert input information into phenotype. A variety of internal and external factors alter chromatin structure and transcription factor binding. These regulatory inputs control the availability of regulatory elements like enhancers, silencers, or promoters. Regulatory elements drive gene expression, which ultimately control phenotype (along with other regulatory influences like RNA decay).
stage, or environment. It is still possible that some of the genome does not have a direct cellular function, such as particular repetitive sequences that have not yet been thoroughly explored (Lander et al., 2001), but it is clear that much of the genome is regulatory. This has led to a multifaceted effort to identify and characterize the regulatory sequences in the human genome.

Cataloging regulatory elements is vital for a complete understanding of human biology. Like all multicellular organisms, humans are composed of a diverse set of cell-types with widely divergent phenotypes that interact in complex ways. The diversity of cellular phenotype is possible because of differential gene regulation, despite cells having identical genomes. Each cell-type activates different genes via appropriate regulatory elements that encode instructions dictating a cell’s response to both external and internal stimuli. In an information flowchart, regulatory elements could be considered decision-making entities that transfer input information into the organism’s response (Fig. 1.1). A difference in phenotype or response to stimulus is often driven by differences in gene expression, which are in turn governed by regulatory elements. Thus, regulatory element activity can be viewed as a genome-based signal that drives differences in phenotype. In order to piece together how humans work, we must understand the differences among cell-types, including decoding their specific regulatory elements.

Thanks to recent biotechnology advances, we can now examine regulatory sequences in depth genome-wide. Several methods have recently been applied to genome-wide regulatory element study (Schones and Zhao, 2008), including the DNaseI hypersensitivity assay (DNase-seq) (Boyle et al., 2008; Song and Crawford, 2010), formaldehyde-assisted isolation of regulatory elements (FAIRE) (Giresi et al., 2007), and chromatin immunoprecipitation (ChIP) (Johnson et al., 2007). Each of these experiments has advantages and disadvantages. The ENCODE consortium (Myers et al., 2011) has recently generated massive amounts of data from each of these experiments. To complement this data, it has been and will continue to be necessary to develop computational algorithms and tools to aid in its interpretation.
Such data and analysis have the potential to help us understand how chromatin structure contributes to transcription factor binding, gene expression, and ultimately phenotypic differences. This chapter will provide an overview of the ENCODE effort to define regulatory elements based on these experiments, summarize the general results, and discuss implications of the millions of regulatory elements distributed throughout the genome.

1.1 Background

1.1.1 Chromatin structure

Chromatin structure has long been known to affect tissue-specific transcriptional regulation (Elgin, 1996; Wu and Gilbert, 1981). Fig. 1.2 illustrates the familiar organization of DNA in the genome. Briefly, the double helix wraps around histone protein octamers to form nucleosomes that then, with additional scaffold proteins, form higher-order 30 nm fibers. Transcriptionally silenced regions are generally packaged into tightly-packed heterochromatin. Actively transcribed genes typically remain in the more loosely-packed euchromatin, where DNA is more accessible to the transcriptional machinery. Within euchromatin, some stretches of DNA are less associated with histones. These “unwrapped” regions are referred to as open, accessible, or nucleosome-depleted regions. Nucleosome-depleted regions can interact with DNA-binding proteins, which can then regulate nearby chromatin structure and gene expression.

Thus, one avenue for chromatin structure to affect transcriptional regulation is through open chromatin being bound by sequence-specific transcription factors. Scientists have been studying regulatory elements using experiments that locate open chromatin since the 1970s (Weisbrod and Weintraub, 1979). However, despite the realized importance and interest in studying regulatory DNA, initial studies of the human genome sequence instead focused on protein-coding genes. This is partly due to difficulties with noncoding DNA: Identifying and assigning functions to regulatory elements is complicated by several issues.
FIGURE 1.2: Open chromatin and the organization of DNA in a cell. DNA is tightly wound around histone proteins to form nucleosomes. Nucleosomes are further wrapped around scaffold proteins to form the 30 nm fiber. In unwrapped regions where histones are displaced, the DNA is open for recognition by transcription factors. Open chromatin regions are also susceptible to DNaseI, which cleaves phosphodiester bonds in the DNA backbone to sever DNA molecules.

1.1.2 Challenges to studying regulatory elements

What makes studying regulatory elements so difficult? First is the sheer number of elements; there are far more than genes. Our initial estimate of 30,000 to 40,000 genes in the human genome (Lander et al., 2001) has more recently been reduced to 20,000–25,000 (Claverie, 2005). In contrast, the number of proposed regulatory elements currently stands in the millions and continues to rise (Lindblad-Toh et al., 2011; Thurman et al., 2012). This complexity has given rise to the term regulome, symbolizing the growing set of regulatory components. The fragmentation and volume of regulatory sequence (perhaps one-third of the genome vs. 2% being coding regions) makes it difficult to study (Thurman et al., 2012).

Second, they are difficult to find. Unlike protein-coding genes, which have a rigid genetic code, regulatory elements seem to lack such a rigid code and are thus more difficult to identify computationally. Where regulatory elements do follow patterns, such as
conforming to transcription factor binding motifs, they are more elastic, which has led to more sequence variation in regulatory elements than in protein-coding genes. Sequence conservation yields some clues (Sandelin et al., 2004), but active functional elements may not be under detectable evolutionary constraint (Birney et al., 2007; McGaughey et al., 2008). For this reason, we must currently rely mainly on experimental methods to identify regulatory elements.

Third, after identifying regulatory elements, further difficulties arise in determining their targets (Heintzman and Ren, 2009). Regulatory elements can act both directly and indirectly to modulate transcriptional levels. For direct (or cis-) regulation, target genes are not obvious because an element does not necessarily regulate the single nearest gene; instead, they can act at large distances (Amano et al., 2009), skip genes (Visser et al., 2012), or affect multiple genes (Spilianakis et al., 2005). Elements can also act indirectly (in trans-) to affect genes en masse by regulating a transcription factor with many targets. These layers form the robust system required to thrive in a changing environment, but they are not easy to dissect and explain.

Fourth, regulatory elements have a variety of functions (Maston et al., 2006). For example, there are distinctions between promoters, enhancers (Bulger and Groudine, 2011), enhancer blockers (Hou et al., 2008), insulators (West et al., 2002), LCRs (Dean, 2006), and Polycomb-bound silencers (Simon and Kingston, 2009). One step further, within each category there may be additional subdivisions, such as the distinction between TATA-box vs. CpG-rich promoters (Carninci et al., 2006) or enhancer sub-classes (Bulger and Groudine, 2011). Each performs a different but necessary purpose, but the sequence alone is currently unable to classify the elements by type. Computational research continues to identify subtle sequence patterns (Lee et al., 2011), but our understanding of sequence remains a key limitation.

Finally, perhaps the most important challenge to studying transcriptional regulation is that the regulome is dynamic. The human regulome varies in many dimensions, such as
age, environment, developmental time, cell-cycle stage, and tissue type. In contrast, the genome is essentially constant and identical in each cell and tissue type. To identify every regulatory element in the human genome is a currently infeasible task; it would require interrogating every cell-type under every possible developmental stage in any environment and against all genetic backgrounds. Despite these challenges, considerable recent progress has been made toward identifying and characterizing regulatory elements. Several of these difficulties have only started to become tractable in the past few years due in large part to technological improvements in sequencing and computation, which have driven new discovery in almost every biological field. In the study of transcriptional regulation, the major genome-wide findings have been primarily driven by results from chromatin accessibility and transcription factor experiments that assay regulatory elements.

1.1.3 Open chromatin and regulatory element assays

Three common experimental techniques used to assay chromatin structure and identify regulatory elements are DNaseI hypersensitivity, formaldehyde assisted identification of regulatory elements (FAIRE), and chromatin immunoprecipitation (ChIP). These assays each have strengths and weaknesses.

**DNaseI hypersensitivity** Scientists have long used DNaseI hypersensitivity assays to distinguish between open regions of DNA and those protected from digestion by nucleases (Wu, 1980). Deoxyribonucleases (DNases) are enzymes that cleave phosphodiester bonds in DNA (Fig. 1.2). There are several types of DNases, including restriction enzymes, which are sequence-specific DNases. In contrast, DNaseI cleaves non-specifically (without sequence preference). DNaseI is a fairly large enzyme, which limits its ability to penetrate tight spaces and restricts it to cleaving DNA that is easily accessible. In a DNaseI hypersensitivity experiment, DNA is treated with a small concentration of DNaseI. The cuts can then be located within the genome and quantified to annotate open chromatin. In
the original experiments, this was done on individual loci and cuts were mapped using the electrophoretic separation of radiolabelled digested fragments on polyacrylamide gels. Now, we use high-throughput sequencing to interrogate digestion genome-wide with greater resolution (Boyle et al., 2008). As a result, after sequencing short DNA molecules corresponding to DNaseI cuts and aligning these to a reference sequence, we get a profile of open regions in the genome. Because transcription factors also tend to bind in such open areas, DNA that is accessible to DNaseI primarily corresponds to regulatory elements. Since many (probably most) changes in the accessibility of DNA are associated with regulatory processes (Felsenfeld and Groudine, 2003), DNaseI assays have been the gold standard in locating regulatory elements. One particular advantage of DNaseI experiments is that they can detect all types of active elements, even without prior knowledge of function (Gross and Garrard, 1988); however, they do not identify what specific factors bind there. They simply distinguish between open and closed DNA.

*Formaldehyde-assisted isolation of regulatory elements (FAIRE)*  
FAIRE is a more recent technique that highlights similar open chromatin regions (Giresi et al., 2007). It is a relatively simple experiment that involves only a few steps: first, a formaldehyde step to fix protein-DNA interactions, followed by sonication to fragment the genome, and finally a phenol-chloroform extraction to separate bound from unbound DNA. Unbound DNA is then sequenced and aligned to identify nucleosome-depleted regions. The results overlap considerably, but not completely, with DNaseI regions (Song et al., 2011). The advantages of FAIRE are that it is highly reproducible and that the samples require relatively minimal preprocessing, reducing potential artifacts and enabling the experiment to be done on a variety of sample types (Giresi et al., 2007). However, the final signal is more diffuse and lacks additional information used for fine-resolution DNaseI footprint mapping (Boyle et al., 2011; Pique-Regi et al., 2011).
Chromatin immunoprecipitation (ChIP)  Like DNase assays, ChIP dates from several decades ago (Solomon et al., 1988). In ChIP, like FAIRE, a lysate is cross-linked to fix protein-DNA reactions, and sonicated or digested to shear the genome. Then, antibodies are used to pull down a particular protein of interest, cross-links are reversed, and DNA is sequenced. Two common uses of ChIP are 1) to find where specific factors bind, and 2) to identify histone tail modifications. In the first case, ChIP differs from DNase and FAIRE assays in that it targets a specific factor. This is both an advantage and a disadvantage; the factor bound is revealed, but it requires specific antibodies and individual experiments for each factor of interest. Thus, we can only use ChIP on factors that are known a priori. In the second case, ChIP uses antibodies against histone proteins (rather than transcription factors), with different antibodies targeting histones with different chain modifications. This type of ChIP experiment does not identify individual transcription factors, but nucleosomes with particular modifications. This usually results in a much more diffuse signal covering multiple nucleosomes; it does not identify specific TFs but it does give a clue as to the function of the region because some modifications have been associated with certain types of regulatory functions (Heintzman and Ren, 2009).

There have been several major advances in ChIP technology recently; most notable is ChIP-seq, the combination with sequencing to look at genome-wide TF binding[9]. ChIP has also been recently modified to require fewer cells (Dahl and Collas, 2008), include methylation status (Statham et al., 2012), and provide better resolution with exonucleases (Rhee and Pugh, 2011). One of the key remaining limitations with ChIP is the availability of high quality antibodies (Egelhofer et al., 2011).

Other similar assays  MNase-seq is similar to DNase-seq, but replaces DNaseI with MNase (micrococcal nuclease) (Schones and Zhao, 2008). The principles of the experiment are the same, but MNase is a smaller molecule than DNaseI, enabling it to digest smaller, less accessible areas. It is able to cleave the linker region between nucleosomes. It also has
exonuclease activity, so it digests back from the initial cuts until it reaches a bound protein protecting the DNA. Rather than sequencing where the cuts are, this method sequences the DNA left intact, indicating the location where something (a nucleosome or transcription factor) is bound. This can highlight exact nucleosome positioning, but it requires deeper sequencing because it sequences the inverse regions: locations of nucleosomes, rather than open regions (less than 5% of a genome is open in any given DNase experiment (Song et al., 2011)). For this reason, it has primarily been used in smaller genomes, like yeast, but a limited amount of data from human is also available (Schones and Zhao, 2008). Other similar techniques include Sono-seq (Auerbach et al., 2009) and Nome-seq (You et al., 2011) which also identify open chromatin; Sono-seq uses sonication instead of DNaseI to fragment DNA, while Nome-seq uses a methyltransferase to mark regions of open chromatin with GC methylation.

One key limitation of all of the above methods is that the regions they identify are only potentially functional. Open chromatin, and even evidence for a bound transcription factor, implies but does not demonstrate regulatory potential (Maston et al., 2006). It is possible for regions to be open or even bound but still lack regulatory effect. Reporter assays are commonly used to validate function, but currently these are done on an individual-site basis, limiting the number of sites that can be reasonably tested.

1.2 Identifying and characterizing regulatory elements

The most common experiment among those mentioned above has been ChIP. Several groups in the ENCODE Consortium have published the results of hundreds of ChIP experiments for different cell-types and transcription factors (Myers et al., 2011; Rosenbloom et al., 2012). Recent reviews have summarized these results (Park, 2009; Collas, 2010). My work has focused primarily on identifying regulatory elements using DNaseI hypersensitivity and FAIRE. This work has both confirmed hypotheses formed on smaller samples and
FIGURE 1.3: The distribution of the genome as a whole compared to the distribution of regulatory elements with respect to known transcribed regions. Dark bars reflect the genomic distribution while the light bars reflect the distribution of regulatory elements in four categories: exon, promoter, intron, and intergenic regions. Regulatory elements are located near (promoters) or within (introns, exons) known genes at greater frequency than would be expected given the genomic distribution.

has revealed new findings regarding how transcriptional regulation works at the genomic level (Thurman et al., 2012; Song et al., 2011).

1.2.1 Open chromatin defines regulatory elements

The working hypothesis of the past several decades has been that open chromatin identifies regulatory regions. Genome-wide results now confirm this finding: Thurman et al. (2012) showed that the DNaseI-seq profile recapitulates the sum of TF ChIP-seq signals. In K562 cells, with ChIP-seq results available for more than 42 factors, the correlation between the cumulative ChIP results and DNaseI-seq is very high ($\approx 0.8$). Almost 95% of known ChIP-seq peaks are in regions identified as open by DNaseI assays. This result shows that open chromatin is a reasonable proxy for generic TF binding and highlights the utility of open chromatin assays.
Figure 1.4: The cumulative number of bases contained in DNaseI-hypersensitive sites according to annotations by Thurman et al. (2012). This plot shows how the total number of accessible bases in the genome (y-axis) grows as additional samples (x-axis) are considered. Starting with a single sample on the left, the cumulative number of accessible bases is calculated by taking the union of the DNase sites from each successive sample. Two different threshold levels are shown, peaks (hypersensitive) and hotspots (more generally sensitive). Ultimately, greater than 1.2 billion nucleotides are annotated as within a DNase hotspot, with over 500 million nucleotides locating within a DNase peak.

1.2.2 Regulatory elements are located in noncoding regions

The distribution of DNA in the genome relative to known transcribed regions is shown in Fig. 1.3, and compared to the distribution of regulatory elements. The vast majority of regulatory elements are noncoding, with about 5% identifying known promoter elements, and the other noncoding elements almost evenly split between intergenic and intronic sequences (Thurman et al., 2012). There are some rare exonic open chromatin regions, which could regulate splicing (Majewski and Ott, 2002), or they may overlap intronic elements. The clear emphasis on distal regulation (most elements are intergenic or intronic) supports the growing realization that distal elements are a key source of phenotypic complexity (Bulger and Groudine, 2011).
1.2.3 More than 30% of the genome may be regulatory

Using a sample of 126 cell-types, Thurman et al. (2012) reported that nearly one-third of the genome shows sensitivity to DNaseI digestion with 15% of the genome being DNaseI hypersensitive. These percentages differ depending on the threshold used to define hypersensitivity, but they provide an idea about the overall proportion of the genome that is sensitive to DNaseI. These numbers should be considered minimums, as they are restricted to the cell-types, environments, and developmental stages assayed. Fig. 1.4 shows how the number and percentage approaches saturation with increasing cell-types. This trend is not simply a result of false-positive DNase-seq signal, which is estimated to be less than 0.5% (Thurman et al., 2012). The numbers will grow as new cell-types and contexts are assayed. Without testing all cell-types and contexts, the exact percentage of regulatory DNA in the human genome cannot be determined, but these results demonstrate that it may be quite high. Ultimately, these regions will have to be tested to establish function.

1.2.4 Most regulatory elements are cell-type-specific

With open chromatin data from a variety of cell-types, it is possible to classify regulatory elements as cell-type-specific, shared among multiple cell-types, or ubiquitous. This cell-type-specificity classification necessarily depends on the cell-types used. Many elements currently considered cell-type-specific may turn out to be present in several related but not yet assayed cell-types. Nevertheless, highly cell-type-specific sites are likely to retain that feature, though it may be adjusted from “cell-type-specific” to “narrow-cell-lineage-specific.”

Gaulton et al. (2010) identified thousands of pancreatic-islet-specific FAIRE sites. Song et al. (2011) used both FAIRE and DNaseI to show similar results in seven other cell-types. These results have illustrated that each cell-type is likely to have a unique set of specific open chromatin regions that guide specific cell fates and functions, in addition to
those shared by other cell-types. Both of these studies made two additional observations in relation to cell-type-specificity: First, the cell-type-specific DNase sites were associated with cell-type-specific gene expression; and second, cell-type-specific regulatory elements tended to cluster with respect to genomic location.

Most recently, Thurman et al. (2012) showed that the majority of DNase sites are found in relatively few cell-types, with the distribution depending on the genomic context of the DNase site. For example, promoter DNase sites were more likely to be ubiquitous rather than cell-type-specific.

1.2.5 Transcription factor binding affects chromatin structure

One way a cell can affect cell-type-specific expression is by creating a cell-type-specific chromatin conformation. It is now becoming clear that a large class of regulatory elements is involved in establishing cell-type-specific chromatin structure. For example, CTCF sites have many roles affecting chromatin accessibility (Phillips and Corces, 2009); CTCF is the canonical insulator, but it also can create both active and repressive loops (Hou et al., 2010). Recent studies have shown that other factors also work by altering chromatin structure. Biddie et al. (2011) showed that DNaseI-hypersensitive sites, specifically with AP1 binding sites, predefine binding sites for GR receptor binding in glucocorticoid cells. Along the same lines, Shibata et al. (2012) associated AP1 motifs with DNaseI signal differences across primates. Chromatin looping is also likely to be regulated by other factors, such as mediator, p300, and cohesin, which work together to establish cell-type-specific chromatin structure (Kagey et al., 2010). These results collectively illustrate the important interaction between transcription factors and chromatin structure.

1.2.6 Chromatin structure affects gene expression

A connection between transcription factor binding and chromatin structure represents one way a cell could regulate gene expression. Recently, Degner et al. (2012) drew the
connection between chromatin and expression in a study that used matched DNaseI-seq, genotype, and expression data from a common cell-type in 70 individuals. They identified about 9,000 sites where DNaseI signal correlated with genotype, and further showed that, in many cases, the DNaseI difference also correlates with gene expression. In a cross-species comparison, Shibata et al. (2012) showed that DNaseI differences among primates are closely tied to expression differences. These results illustrate how this data may be able to inform computational models to predict gene expression, which has been an area of interest in other organisms (Beer and Tavazoie, 2004; Yuan et al., 2007; Segal et al., 2008). Along these lines, Natarajan et al. (2012) built a model to predict gene expression from motif analysis and DNaseI signal. They showed that performance improves when distal elements are included in the model, highlighting the relevance of distal regulatory sites. These results are beginning to unravel the complex interactions between chromatin structure and gene expression.

1.2.7 Regulatory elements can be classified by function

One of the most important ways to annotate a regulatory element after discovery is to determine what factors bind there. The primary experimental tool for such an annotation is ChIP. In the ENCODE project, data for hundreds of ChIP experiments have already been made available on the UCSC genome browser (Rosenbloom et al., 2012). These data enable us to identify the frequency and importance of individual factors (Pérez-Lluch et al., 2011; Reddy et al., 2009; Wontakal et al., 2011; Wu et al., 2012), as well as explore cooperativity among factors (Karczewski et al., 2011). DNaseI footprinting has also been used to propose what factors bind an element (Boyle et al., 2011; Pique-Regi et al., 2011). These analyses require not only experimental data, but appropriate computational algorithms to extract meaningful signal.

Regulatory elements can also be divided into functional classes (i.e., silencers, promoters, enhancers). Some classes, like promoters, are relatively straightforward to positionally
and functionally define: they lie just upstream of genes and operate on the adjacent gene. Others, like distal enhancers, seem to be anywhere and act on anything. In general, all functional classes share the one major property that they tend to be found in regions of open chromatin. However, they vary in other properties, such as TF binding and histone marks (Kouzarides, 2007). For example, insulators are often bound by CTCF, whereas enhancers are commonly characterized by certain histone marks (H3K4me1 and H3K27ac) or cofactors (p300) (Heintzman and Ren, 2009). Quantifying histone marks therefore gives us a relatively straightforward way to provide an initial annotation of regulatory elements genome-wide.

The ENCODE Consortium and others have been performing ChIP-seq to identify multiple histone modifications in several human cell-types. Heintzman and Ren (2009) showed that enhancer histone marks are more often cell-type-specific, in contrast to promoter and insulator marks that are fairly consistent across cell-types. As histone mark data have become available from more cell-types, it has become possible to extrapolate to elements of unknown class and improve current genome annotation. On the basis of this type of data, computational researchers are designing algorithms to characterize regulatory elements. For example, Ernst and Kellis (2010) designed a Hidden Markov Model to assign categories to genomic elements on the basis of their histone modification signatures. Lee et al. (2011) trained a Support Vector Machine to predict mammalian enhancers from genomic sequence and validated the predictions with ChIP and DNaseI data. Classifying regulatory elements by function will be a vital step in understanding transcriptional regulation.

A complementary way to characterize regulatory elements is by cell-type specificity, rather than by experimental signature. In Chapter 4, I will show how a self-organizing map (a machine learning method) can classify regulatory elements based on their open chromatin signal pattern across all cell-types. Regulatory elements cluster into a limited number of similar patterns. These kinds of classifications will help us derive more meaningful annotations of regulatory elements. By grouping elements by patterns, we may be able
to leverage information across open chromatin sites to better annotate how they act. For instance, we have discovered that many groups of DNase sites with similar patterns of accessibility across cell-types share common transcription factor motifs suggesting similarly bound factors.

1.3 Motivation

In earlier decades, our understanding of regulatory DNA was restricted by our ability to identify it, which required painstaking locus-by-locus experiments. In the past few years, though, these assays (DNase-seq, FAIRE-seq, and ChIP-seq) have made it possible to query regulatory DNA genome-wide. This new technology has led to important discoveries in gene regulation with relevance across biology. This dissertation focuses on two studies of gene regulation: First, in the evolutionary context of comparative genomics, and second, in the context of human cell-type diversity. In each case, I have studied how identifying genome-wide regulatory DNA can inform our models of gene regulation.

The study of regulatory elements in primates is a fundamental step toward answering the question of what makes us human. Humans and chimpanzees are very similar in DNA, particularly in protein-coding DNA, where they differ by only 1 nucleotide in 1,000 (99.9% identity). However, despite this genotypic similarity, it is clear that there are many important phenotypic differences between humans and other primates, including physical traits, such as size, hair, brain size, and gastrointestinal system, and also behavioral differences, including diet, cognitive ability, and social structure. Some of these phenotypic differences have important medical implications, including susceptibility to diabetes, heart disease, and pathogens. What is the genetic basis for such differences? To address this question, Chapter 2 will describe how regulatory element activity differs across primate species. Years before the human genome sequence was known, it was hypothesized that the key differences between humans and other primates would be in regulatory DNA, rather
than in protein-coding DNA (King and Wilson 1975). Now, with the complete human and chimpanzee sequences available, we have begun to show that this is actually the case. Because humans and chimpanzees share such similar protein-coding DNA, it is likely that important differences can be found in regulatory DNA. To better understand the regulatory differences that underlie human uniqueness, I developed algorithms to analyze differential regulatory element activity across primates. I also went on to hypothesize that specific sequence variants can confer the regulatory differences we see. By connecting sequence to regulatory element activity, and then on to gene expression, we come one step closer to understanding how genotype connects to phenotype in the context of human uniqueness.

My study of regulatory elements also led in another direction: to explore how different cell-types—within a single species—can have an identical genome and yet behave so differently. Chapters 3 and 4 shift from a cross-species comparison to a within-species comparison looking at human cell-type diversity. Humans are made up of many cell-types, each with unique morphology and function. These diverse cell-types communicate and cooperate, each contributing a vital and unique task to form a complete multicellular organism. Because these cell-types have the same genome, the diversity in a single human is due not to different genetics, but to different regulatory programs employed by each cell-type. While each cell has the same initial set of instructions, it employs a different program to effect a different outcome. To really understand how individual cell-types perform their unique functions, we must understand the regulatory signals that program them. Such an understanding is interesting as a fundamental biological question of how multicellular organisms work, and also has broader application in human medicine. For example, cancers can subvert normal regulatory processes to enable out-of-control cell proliferation. To better understand tissue regulatory specificity, I identified tissue-specific patterns in regulatory activity and then further analyzed these elements for other biological patterns. My analysis shows that tissue specific regulatory element activity can predict tissue identity for unknown tissues and identify the regulators that confer tissue specificity.
Although we still have much to learn, the past 40 years have seen phenomenal advances in our understanding of how the human genome works. These advancements have been driven by the combined effort of both experimental and computational research. Already, there are clear benefits and insights derived from our study of transcriptional regulation, as well as other levels of gene regulation, including mRNA splicing, dispersion, and decay. As we improve our annotations of regulatory elements, our goal to convert this data into information will be realized, ultimately to find the function (if present) of what was once called junk DNA.
Regulatory Elements Across Primates

Understanding the molecular basis of phenotypic differences between humans and other primates has been a priority in medicine, behavior, and evolution research (Robertson, 2010; Olson and Varki, 2003; Sholtis and Noonan, 2010). The genetic basis for these differences can now be explored genome-wide due in part to the rising number of completely sequenced primate genomes. However, finding genotype-phenotype connections is difficult since the vast majority of sequence changes do not contribute to phenotypic differences across species. It was hypothesized over 40 years ago that phenotypic differences between humans and our closest primate relatives are shaped largely by changes in non-coding regulatory elements. Variation in gene regulation have been indirectly confirmed by studying gene expression differences across matched cell or tissue types isolated from different primates (Cáceres et al., 2003; Khaitovich et al., 2005; Gilad et al., 2006; Enard et al., 2002; Blekhman et al., 2002).
Table 2.1: Samples used for DNase-seq and DGE-seq analyses. Basic sequence count statistics are provided with each sample. Dermal fibroblasts are primary. * Standard location of skin punch biopsy for chimpanzees at Yerkes Regional Primate Research Center is the upper arm.

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Species</th>
<th>Age (yr)</th>
<th>Sex</th>
<th>Coriell No.</th>
<th>Cell-type</th>
<th>Site of biopsy</th>
<th>DNase-seq reads</th>
<th>DGE-seq reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>HL1</td>
<td>H. sapiens</td>
<td>12</td>
<td>M</td>
<td>AG16408</td>
<td>Lymphoblastoid</td>
<td>peripheral vein</td>
<td>29,332,310</td>
<td>10,950,474</td>
</tr>
<tr>
<td>HF1</td>
<td></td>
<td></td>
<td></td>
<td>AG16409</td>
<td>Skin Fibroblast</td>
<td>unspecified</td>
<td>37,624,821</td>
<td>8,876,138</td>
</tr>
<tr>
<td>HL2</td>
<td>H. sapiens</td>
<td>36</td>
<td>M</td>
<td>GM02184</td>
<td>Lymphoblastoid</td>
<td>peripheral vein</td>
<td>25,272,133</td>
<td>5,701,620</td>
</tr>
<tr>
<td>HF2</td>
<td></td>
<td></td>
<td></td>
<td>GM02185</td>
<td>Skin Fibroblast</td>
<td>unspecified</td>
<td>25,272,133</td>
<td>5,701,620</td>
</tr>
<tr>
<td>HL3</td>
<td>H. sapiens</td>
<td>48</td>
<td>F</td>
<td>GM05920</td>
<td>Lymphoblastoid</td>
<td>peripheral vein</td>
<td>28,414,314</td>
<td>6,729,690</td>
</tr>
<tr>
<td>HF3</td>
<td></td>
<td></td>
<td></td>
<td>GM05879</td>
<td>Skin Fibroblast</td>
<td>upper arm</td>
<td>17,998,960</td>
<td>4,870,902</td>
</tr>
<tr>
<td>CL1</td>
<td>P. troglodytes</td>
<td>9</td>
<td>M</td>
<td>S003648</td>
<td>Lymphoblastoid</td>
<td>peripheral vein</td>
<td>29,343,610</td>
<td>4,223,955</td>
</tr>
<tr>
<td>CF1</td>
<td></td>
<td></td>
<td></td>
<td>S003649</td>
<td>Skin Fibroblast</td>
<td>unspecified*</td>
<td>34,974,357</td>
<td>9,711,114</td>
</tr>
<tr>
<td>CL2</td>
<td>P. troglodytes</td>
<td>14</td>
<td>M</td>
<td>S003623</td>
<td>Lymphoblastoid</td>
<td>peripheral vein</td>
<td>41,226,270</td>
<td>5,788,846</td>
</tr>
<tr>
<td>CF2</td>
<td></td>
<td></td>
<td></td>
<td>S003624</td>
<td>Skin Fibroblast</td>
<td>unspecified*</td>
<td>59,866,736</td>
<td>10,349,784</td>
</tr>
<tr>
<td>CL3</td>
<td>P. troglodytes</td>
<td>16</td>
<td>M</td>
<td>S007602</td>
<td>Lymphoblastoid</td>
<td>peripheral vein</td>
<td>20,329,252</td>
<td>4,727,760</td>
</tr>
<tr>
<td>CF3</td>
<td></td>
<td></td>
<td></td>
<td>S007603</td>
<td>Dermal Fibroblast</td>
<td>unspecified*</td>
<td>24,840,265</td>
<td>7,420,880</td>
</tr>
<tr>
<td>MF1</td>
<td>M. mulatta</td>
<td>5</td>
<td>F</td>
<td>AG06252</td>
<td>Skin Fibroblast</td>
<td>upper arm</td>
<td>29,330,217</td>
<td>3,626,144</td>
</tr>
<tr>
<td>MF2</td>
<td>M. mulatta</td>
<td>1</td>
<td>M</td>
<td>AG08305</td>
<td>Skin Fibroblast</td>
<td>upper arm</td>
<td>31,937,760</td>
<td>5,142,449</td>
</tr>
<tr>
<td>MF3</td>
<td>M. mulatta</td>
<td>1</td>
<td>M</td>
<td>AG08308</td>
<td>Skin Fibroblast</td>
<td>upper arm</td>
<td>20,324,578</td>
<td>7,004,730</td>
</tr>
</tbody>
</table>

2010; Somel et al., 2009; Xu et al., 2010; Babbitt et al., 2010), but these studies have failed to pinpoint the regulatory elements responsible for these changes (Carroll, 2008). Genome-wide scans of non-coding DNA sequences under branch-specific positive selection have identified putative regulatory elements that have undergone functional changes (Pollard et al., 2006a; Prabhakar et al., 2006; Haygood et al., 2007). These studies identified hundreds of regulatory regions with evidence of accelerated sequence substitution during human origins, possibly reflecting adaptive changes in gene regulation. Scans for selection do not, however, provide information about the functional or trait consequences of these evolutionary changes.

Understanding the relationship between mutation, natural selection, and variation in gene regulation is an important goal in evolutionary genomics. Heritable differences in gene expression must have a genetic basis, but exactly what sequence variants have led to these differences are largely unknown. In this study, we used changes in chromatin configuration to better understand this genotype-phenotype relationship. We identified evolutionary conserved and altered regulatory element activity by performing genome-wide
DNase-seq (Boyle et al., 2008; Song and Crawford, 2010) in primary skin fibroblasts and lymphoblastoid cell lines (LCLs) isolated from three human and three chimpanzee individuals (Fig. 2.1a and Tab. 2.1). Each DNase-seq experiment identifies nucleosome-depleted DNaseI hypersensitive sites (DHSs) that mark all types of regulatory elements, including promoters, enhancers, silencers, insulators, and locus control regions. The comprehensiveness of this assay is supported by ChIP experiments for active histone marks, p300, CTCF, and other transcription factors (Birney et al., 2007; Heintzman et al., 2007; Xi et al., 2007). In addition to human and chimpanzee, we performed DNase-seq on fibroblasts from three Rhesus macaque individuals to polarize human-chimpanzee chromatin differences and to distinguish between gains and losses of regulatory elements on the human and chimpanzee branches (EBV-derived lymphoblastoid cells are not available for this species). We also performed Digital Gene Expression sequence (DGE-seq) experiments using the same cell cultures to simultaneously compare levels of mRNA abundance (Gilad et al., 2006; Blekhman et al., 2008). Analyses of these data provide insights into the relationship between evolutionary changes in regulatory elements, their tissue-specific activity, and the resulting functional consequences in gene expression.

2.2 Results

2.2.1 DNase-seq identifies species-specific DHSs

To directly compare DNase-seq data generated from human and non-human primate fibroblast and lymphoblastoid cell line (LCL) samples, we mapped all data to the human genome (build hg19). Non-human DNase-seq sequences were first aligned to their native primate genome and then converted to human coordinates using liftOver (Karolchik et al., 2012) (Fig. 2.1a). We limited analyses to high confidence orthologous regions of the human, chimp, and macaque genomes to eliminate potential artifacts due to mis-aligned, missing sequence, or CNVs (Materials and Methods). Comparisons across individuals within a
Figure 2.1: (a) Analysis pipeline. DNase-sequences from each species were aligned to the native genome and lifted over to the human genome for analysis. Regions are filtered at various steps of the analysis to remove alignment and orthology artifacts (Materials and Methods). Correlation plots of DNase-seq signals (b) and DGE-seq signals (c) expression data show that both chromatin and expression data from human (Hu), chimpanzee (Ch), and macaque (Ma) are more highly correlated between biological replicates from the same tissue within a single species. Additionally, the same cell-type from different species is more similar than different cell-types from the same species.

Species and against tiling array DNase-chip (Crawford et al., 2006; Shibata and Crawford, 2009) data generated from the same material supported data accuracy and reproducibility (Materials and Methods and Tab. 2.2). DNase-seq signals from individuals within a species were more highly correlated than signals from different species (Fig. 2.1b). Human and chimpanzee DNase-seq signals from fibroblasts were better correlated than human and macaque signals as expected since human and chimp share a more recent common ancestor. Chromatin structure differed more in cell-types (fibroblasts vs. LCLs) from the same species than in the same cell-type across different species (Fig. 2.1b). For example, human and chimpanzee fibroblast DNase-seq signals are more similar than human fibroblast and human LCL DNase-seq signals. The same correlation patterns were also found in gene expression data generated from the same samples (Fig. 2.1c).

We identified genomic regions exhibiting significant differences in DNase-seq signal between species (Robinson and Oshlack, 2010) (Materials and Methods). Data from macaque samples were used to classify regions as DHS gains or DHS losses on the human...
Table 2.2: DNase-seq data is validated by DNase-chip. DNase-chip libraries from the 15 samples were hybridized to 1% ENCODE DNA arrays of the corresponding species type. Probes consisted of segments of DNA sequences matching reference sequence builds hg18 (human), panTro2 (chimpanzee), and rheMac2 (macaque), respectively. The array intensities were compiled and significant DHS sites were called using ChIPotle (P<0.000001 peak cutoff). We intersected the DNase-chip sites from each of the 3 individuals for each species and determined the amount that overlap DNase-seq data. The top 100,000 F-seq called sites were used from each DNase-seq sample for this comparison.

<table>
<thead>
<tr>
<th>Sample</th>
<th>% Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>HF1</td>
<td>71.3</td>
</tr>
<tr>
<td>HF2</td>
<td>71.6</td>
</tr>
<tr>
<td>HF3</td>
<td>73.3</td>
</tr>
<tr>
<td>CF1</td>
<td>79.0</td>
</tr>
<tr>
<td>CF2</td>
<td>87.1</td>
</tr>
<tr>
<td>CF3</td>
<td>77.1</td>
</tr>
<tr>
<td>MF1</td>
<td>87.8</td>
</tr>
<tr>
<td>MF2</td>
<td>87.9</td>
</tr>
<tr>
<td>MF3</td>
<td>82.4</td>
</tr>
</tbody>
</table>

Figure 2.2: Species-specific DHSs were identified by edgeR (Materials and Methods). Boxplots show the distribution of number of reads per sample in 300 bp windows. For human DHS gains (a), the 3 human samples are all significantly more open than the other 2 species. Likewise, human DHS losses (b) show lower signal in human compared to both chimpanzee and macaque. A representative sampling of distributions from all DHS is shown in (c), as well as Common DHSs (d) found in all three species that are matched for signal intensity compared to human DHS gains and human DHS losses. (e) Distribution of species-specific DHS Gains and DHS Losses relative to promoters, introns, 3 UTR, and intergenic regions. (f) Representative screen shots of human-specific DHS Gains and Losses compared to a Common region.

or chimpanzee branch (Materials and Methods). More specifically, we defined a human DHS gain as a region with significantly more DNase-seq signal in human than in either chimpanzee or macaque (Fig. 2.2a), and a human DHS loss as a region with significantly less DNase-seq signal in human than in either chimpanzee or macaque (Fig. 2.2b). In essence, these data identify regulatory regions that originated or disappeared in fibroblasts

24
Table 2.3: Human cell types analyzed by DNase-seq for the ENCODE project by our group. All data is available on the UCSC genome browser (http://genome.ucsc.edu/)

<table>
<thead>
<tr>
<th>UCSC Name</th>
<th>Additional Description</th>
<th>Sex</th>
<th>Cell Name</th>
<th>Cell Provider</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chorion</td>
<td></td>
<td>F</td>
<td>Chorion</td>
<td>Amy Murtha</td>
</tr>
<tr>
<td>Medullo</td>
<td>Medulloblastoma (D721)</td>
<td>M</td>
<td>D721</td>
<td>Hai Yan and Matt Wortham</td>
</tr>
<tr>
<td>ProgFib</td>
<td>Fibroblast from progeria patient (FB0167P)</td>
<td>M</td>
<td>Fibro_FB0167P</td>
<td>Progeria Foundation</td>
</tr>
<tr>
<td>FibroP</td>
<td>Fibroblast from parkinson’s patients M(B1), F(B2), F(B3)</td>
<td>F</td>
<td>Fibro_park</td>
<td>Paul Tesar</td>
</tr>
<tr>
<td>GM12878</td>
<td>Lymphoblastoid</td>
<td>F</td>
<td>GM12878</td>
<td>Coriell</td>
</tr>
<tr>
<td>GM12891</td>
<td>Lymphoblastoid</td>
<td>M</td>
<td>GM12891</td>
<td>Coriell</td>
</tr>
<tr>
<td>GM12892</td>
<td>Lymphoblastoid</td>
<td>F</td>
<td>GM12892</td>
<td>Coriell</td>
</tr>
<tr>
<td>GM18507</td>
<td>Lymphoblastoid</td>
<td>M</td>
<td>GM18507</td>
<td>Coriell</td>
</tr>
<tr>
<td>GM19238</td>
<td>Lymphoblastoid</td>
<td>F</td>
<td>GM19238</td>
<td>Coriell</td>
</tr>
<tr>
<td>GM19239</td>
<td>Lymphoblastoid</td>
<td>M</td>
<td>GM19239</td>
<td>Coriell</td>
</tr>
<tr>
<td>GM19240</td>
<td>Lymphoblastoid</td>
<td>F</td>
<td>GM19240</td>
<td>Coriell</td>
</tr>
<tr>
<td>H1-MESC</td>
<td>H1 Embryonic Stem Cells</td>
<td>M</td>
<td>H1-ES</td>
<td>Cellular Dynamics</td>
</tr>
<tr>
<td>H9-MESC</td>
<td>H9 Embryonic Stem Cells</td>
<td>F</td>
<td>H9-ES</td>
<td>Paul Tesar</td>
</tr>
<tr>
<td>HeLa-S3</td>
<td>cervical carcinoma</td>
<td>F</td>
<td>HelaS3</td>
<td>ATCC CCL-2.2</td>
</tr>
<tr>
<td>HeLa-S3</td>
<td>cervical carcinoma induced with interferon alpha (IFNa4h)</td>
<td>F</td>
<td>HelaS3_IFNA</td>
<td>ATCC CCL-2.2</td>
</tr>
<tr>
<td>HepG2</td>
<td>hepatocellular carcinoma</td>
<td>M</td>
<td>HepG2</td>
<td>ATCC HB-8065</td>
</tr>
<tr>
<td>HUVEC</td>
<td>human umbilical vein endothelial cell</td>
<td>M</td>
<td>HUVEC</td>
<td>Lonza CC-2517 (Lot# 7F3239)</td>
</tr>
<tr>
<td>K562</td>
<td>chronic myeloid leukemia</td>
<td>F</td>
<td>K562</td>
<td>ATCC CCL-243</td>
</tr>
<tr>
<td>MCF-7</td>
<td>epithelial breast adenocarcinoma</td>
<td>F</td>
<td>MCF7</td>
<td>Yijun Ruan</td>
</tr>
<tr>
<td>Melano</td>
<td>Melanocyte, cat 2200, NHM22=lot#1002, NHM23=lot#1014</td>
<td>M</td>
<td>Melanocyte</td>
<td>William Kaufmann and Patrick Tompkins</td>
</tr>
<tr>
<td>HSMM</td>
<td>CM33j(B1), CM01-201-001(B2), CM01-154-002(B3), CM01-135-003(B4)</td>
<td>F(B1), F(B2), M(B3), F(B4)</td>
<td>Myoblast</td>
<td>Melanie Ehrlich</td>
</tr>
<tr>
<td>Myometr</td>
<td>myometrial</td>
<td>F</td>
<td>Myometrial</td>
<td>Amy Murtha</td>
</tr>
<tr>
<td>HSMMMtube</td>
<td>CM33j(B1), CM01-201-001(B2), CM01-154-002(B3), CM01-135-003(B4)</td>
<td>F(B1), F(B2), M(B3), F(B4)</td>
<td>Myotube</td>
<td>Melanie Ehrlich</td>
</tr>
<tr>
<td>NHEK</td>
<td>Normal Human Epidermal Keratinocyte</td>
<td>F</td>
<td>NHEK</td>
<td>Lonza CC-2501 (Lot# 4F11553)</td>
</tr>
<tr>
<td>PanIslets</td>
<td>VGP024(B1), VKJ113(B3)</td>
<td>M(B1), F(B2), M(B3)</td>
<td>Pancr_islets</td>
<td>Francis Collins and Jorge Ferrer</td>
</tr>
<tr>
<td>AoSMC</td>
<td>Smooth muscle, Serum Free (SM_SFM)</td>
<td>M</td>
<td>SM_SFM</td>
<td>Lonza CC-2571 (Lot# 7F4356)</td>
</tr>
</tbody>
</table>

during human origins. Chimpanzee DHS gains and DHS losses were similarly defined (Fig. 2.3). For approximately 90% of gains, a corresponding DHS was completely absent in all three individuals from each of the other species (Fig. 2.2f). For the remaining sites, DHSs were annotated in multiple species, but a consistently higher DNase-seq signal was present in one species compared to the others (data not shown). We found that the majority of the human DHS gains (72–79%) and chimpanzee DHS losses (73–74%), and
Table 2.4: Percent overlap of human- and chimpanzee- DHS gains/losses/common detected in fibroblasts with DHS sites identified in ENCODE human cell types. FibroP, Fibrobl, and ProgFib cells were independently derived fibroblasts samples.

<table>
<thead>
<tr>
<th>Cell Type</th>
<th>% Overlap with human DHS gains (intersects)</th>
<th>% Overlap with human DHS losses (intersects)</th>
<th>% Overlap with common DHS (intersects)</th>
<th>% Overlap with chimpanzee DHS gains (intersects)</th>
<th>% Overlap with chimpanzee DHS losses (intersects)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chorion</td>
<td>195 (23.3%)</td>
<td>56 (19.58%)</td>
<td>943 (74.90%)</td>
<td>89 (13.17%)</td>
<td>82 (38.86%)</td>
</tr>
<tr>
<td>D721</td>
<td>57 (6.82%)</td>
<td>28 (9.79%)</td>
<td>833 (66.16%)</td>
<td>48 (7.10%)</td>
<td>36 (17.06%)</td>
</tr>
<tr>
<td>FB0167P</td>
<td>612 (73.21%)</td>
<td>76 (26.57%)</td>
<td>1254 (99.29%)</td>
<td>108 (15.98%)</td>
<td>155 (73.46%)</td>
</tr>
<tr>
<td>FB8470</td>
<td>598 (71.53%)</td>
<td>32 (11.19%)</td>
<td>1178 (93.57%)</td>
<td>56 (8.28%)</td>
<td>155 (73.46%)</td>
</tr>
<tr>
<td>Fibroblasts_park</td>
<td>661 (79.07%)</td>
<td>64 (22.38%)</td>
<td>1283 (99.60%)</td>
<td>108 (15.98%)</td>
<td>155 (73.46%)</td>
</tr>
<tr>
<td>GM12878</td>
<td>102 (12.20%)</td>
<td>22 (7.69%)</td>
<td>944 (74.98%)</td>
<td>49 (7.25%)</td>
<td>41 (19.43%)</td>
</tr>
<tr>
<td>GM12891</td>
<td>84 (10.05%)</td>
<td>37 (12.94%)</td>
<td>911 (72.36%)</td>
<td>69 (10.21%)</td>
<td>42 (19.91%)</td>
</tr>
<tr>
<td>GM12892</td>
<td>92 (11.00%)</td>
<td>45 (15.73%)</td>
<td>902 (71.64%)</td>
<td>84 (12.43%)</td>
<td>43 (20.38%)</td>
</tr>
<tr>
<td>GM18507</td>
<td>66 (7.89%)</td>
<td>22 (7.69%)</td>
<td>894 (71.01%)</td>
<td>35 (5.18%)</td>
<td>34 (16.11%)</td>
</tr>
<tr>
<td>GM19238</td>
<td>77 (9.21%)</td>
<td>32 (11.19%)</td>
<td>923 (73.31%)</td>
<td>57 (8.4%)</td>
<td>38 (18.01%)</td>
</tr>
<tr>
<td>GM19239</td>
<td>75 (8.97%)</td>
<td>33 (11.54%)</td>
<td>912 (72.44%)</td>
<td>33 (5.7%)</td>
<td>44 (22.83%)</td>
</tr>
<tr>
<td>GM19240</td>
<td>74 (8.85%)</td>
<td>33 (11.54%)</td>
<td>628 (73.71%)</td>
<td>52 (7.69%)</td>
<td>38 (18.01%)</td>
</tr>
<tr>
<td>H1_ES</td>
<td>66 (7.89%)</td>
<td>40 (13.99%)</td>
<td>870 (69.10%)</td>
<td>65 (9.62%)</td>
<td>46 (21.80%)</td>
</tr>
<tr>
<td>H9_ES</td>
<td>74 (8.85%)</td>
<td>46 (16.08%)</td>
<td>928 (73.71%)</td>
<td>51 (7.54%)</td>
<td>42 (19.91%)</td>
</tr>
<tr>
<td>HelaS3</td>
<td>195 (23.33%)</td>
<td>58 (20.28%)</td>
<td>1055 (83.80%)</td>
<td>90 (13.31%)</td>
<td>84 (39.81%)</td>
</tr>
<tr>
<td>HelaS3_IFNA</td>
<td>229 (27.39%)</td>
<td>57 (19.93%)</td>
<td>1063 (84.43%)</td>
<td>101 (14.94%)</td>
<td>85 (40.28%)</td>
</tr>
<tr>
<td>HepG2</td>
<td>82 (9.81%)</td>
<td>45 (15.73%)</td>
<td>968 (76.89%)</td>
<td>69 (10.21%)</td>
<td>61 (32.91%)</td>
</tr>
<tr>
<td>HUVEC</td>
<td>212 (25.36%)</td>
<td>32 (11.19%)</td>
<td>1079 (85.70%)</td>
<td>72 (10.65%)</td>
<td>79 (37.44%)</td>
</tr>
<tr>
<td>K562</td>
<td>95 (11.36%)</td>
<td>44 (15.38%)</td>
<td>922 (73.23%)</td>
<td>60 (8.88%)</td>
<td>49 (23.22%)</td>
</tr>
<tr>
<td>MCF7</td>
<td>109 (13.04%)</td>
<td>37 (12.94%)</td>
<td>969 (76.97%)</td>
<td>63 (9.32%)</td>
<td>62 (29.38%)</td>
</tr>
<tr>
<td>Melanocyte</td>
<td>379 (45.33%)</td>
<td>28 (9.79%)</td>
<td>1178 (93.57%)</td>
<td>66 (9.76%)</td>
<td>100 (47.39%)</td>
</tr>
<tr>
<td>Myoblast</td>
<td>367 (43.90%)</td>
<td>54 (18.88%)</td>
<td>1166 (92.61%)</td>
<td>105 (15.53%)</td>
<td>110 (52.13%)</td>
</tr>
<tr>
<td>Myometrial</td>
<td>360 (46.99%)</td>
<td>68 (23.78%)</td>
<td>1226 (97.38%)</td>
<td>114 (16.80%)</td>
<td>142 (67.30%)</td>
</tr>
<tr>
<td>Myotube</td>
<td>353 (42.22%)</td>
<td>61 (21.33%)</td>
<td>1123 (89.20%)</td>
<td>118 (17.46%)</td>
<td>111 (52.61%)</td>
</tr>
<tr>
<td>NHEK</td>
<td>385 (46.05%)</td>
<td>56 (19.58%)</td>
<td>1126 (89.44%)</td>
<td>76 (11.24%)</td>
<td>116 (54.98%)</td>
</tr>
<tr>
<td>Pancreatic islets</td>
<td>165 (19.74%)</td>
<td>44 (15.38%)</td>
<td>1013 (80.46%)</td>
<td>85 (12.57%)</td>
<td>66 (31.28%)</td>
</tr>
<tr>
<td>SM_SFm</td>
<td>437 (52.27%)</td>
<td>108 (37.76%)</td>
<td>1246 (98.97%)</td>
<td>156 (23.08%)</td>
<td>120 (56.87%)</td>
</tr>
</tbody>
</table>

a minority of the human DHS losses (11–27%) and chimpanzee DHS gains (8–17%), overlapped a DHS found in one or more of three independently derived human fibroblasts (Fig. 2.4a, Tab. 2.3). We also found similar trends comparing six independently derived LCLs analyzed by our group (Fig. 2.4b, Tab. 2.4), and 20 independently derived human fibroblast samples analyzed by another ENCODE group (Fig. 2.5) (Sabo et al., 2004, 2006). These provide evidence that the identified DHS gains and losses represent significant and
reproducible functional changes between species. Fibroblasts have been shown to have specific expression profiles associated with different biopsy locations (Chang et al., 2002; Rinn et al., 2006). DHS gains and losses are not enriched around these genes (Materials and Methods).

Identified DHS gains (Fig. 2.2a) and losses (Fig. 2.2b) deviated in sequence read depth from the general chromatin spectrum (Fig. 2.2c). To more directly compare DHS gains and losses with sites that do not change between species, we also identified a set of DHS regions
Species-specific gains and common DHS sites are highly reproducible in independently isolated matched cell types. (a) Human DHS gains/losses, Chimp DHS gains/losses, and Common DHS sites detected in fibroblast cells were compared to independently derived ENCODE Fibroblast DNase-seq data generated by the University of Washington ENCODE group. (b) LCL human DHS gains/losses/common compared to 4 independently derived ENCODE lymphoblastoid samples also identified by the University of Washington ENCODE group.

with similar DNase-seq signal intensity across all three species, which we call Common DHS regions (Fig. 2.2d, Materials and Methods). Using a false discovery rate (FDR) of 1%, we detected 836 human DHS gains, 286 human DHS losses, 676 chimpanzee DHS gains, 211 chimpanzee DHS losses, and 1259 Common regions. The higher number of DHS gains compared to DHS losses could be due to purifying selection, or more simply...
may be related to the asymmetry in their detection criteria (see Materials and Methods for a more complete discussion). True species-specific DHS gains and losses could not be identified in LCLs due to the lack of macaque EBV-derived LCL samples. However, we identified 103 DHSs with higher DNase-seq signals in human (LCL human DHS gain), 181 DHSs with lower signals in human (LCL human DHS loss), and 1583 DHSs with similar signals in both (LCL common DHS).

Similar numbers of gains and losses were found when comparing chimpanzee DNase-seq data to data from an independent set of human fibroblasts and LCLs at the same FDR (Fig. 2.6). Furthermore, only 66 differential open chromatin sites were detected when comparing human fibroblast data to additional independently derived human fibroblasts. Likewise, only 1 differential DHS was detected when comparing human LCLs to additional independently derived human LCLs. This is less than 1% of all differential open chromatin sites when comparing human vs. chimpanzee, indicating a low false positive rate (Fig. 2.6).
2.2.2 Species-specific DHSs are cell-type-specific

As part of the ENCyclopedia Of DNA Elements (ENCODE) project (Myers et al., 2011), we have generated DNase-seq data from 27 diverse human cell-types (Song et al., 2011) (Tab. 2.3, Crawford unpublished ENCODE data). We determined the overlap of our identified DHS gains and losses in fibroblasts with DHSs in these other human cell-types. Seven hundred and sixty-seven (92%) fibroblast human DHS gains were found in at least one of three other independently derived human skin fibroblast ENCODE cell lines from normal (Fibrobl) and diseased individuals (Parkinson’s: FibroP; Progeria: ProgFib).
Figure 2.8: Box plot showing DNase intensity of human DHS gains/losses and common regions across 27 human cell types shown in Fig. 2.7 heatmaps.

Figure 2.9: Boxplot of the binary comparison of human DHS gain/loss, chimpanzee DHS gain/loss, and Common regions to the DHS peak calls from the 27 other human cell types.
FIGURE 2.10: Comparison of chimpanzee DHS gains and DHS losses to DNase-seq data from other human cell types. These regions were compared to DNase-seq data generated from 27 other human cell types (Tab. 2.3). Heatmap signal intensities are of maximum DNase-seq parzen scores in log scale, where red indicates a higher DNase-seq score and blue indicates lower DNase-seq scores. (a) Chimpanzee DHS sites were identified as differentially open (chimpanzee DHS gain) in chimpanzee fibroblasts compared to human/macaque fibroblasts. (b) Chimpanzee DHS sites identified as differentially closed (Chimpanzee DHS loss) compared to human and macaque fibroblasts. (c) DNase-seq signal values for Common regions representing DHS sites in all three species. Note that more than 50% of Common regions are also DHS sites in other human tissues. (d, e, f) DNase-seq signal values for same regions as (a, b, c), but DNase-seq data is from orthologous region from human and macaque fibroblasts. (g, h, i) DNase-seq values for same regions as (a, b, c), but from human and chimpanzee LCLs. (bottom) Box plot shows intensity values shown in heatmaps.

supporting the reproducibility of these data (Fig. 2.7a, Fig. 2.6a, Fig. 2.9). Additionally, human DHS gains showed a high level of overlap with some, but not all, non-fibroblast human cell-types (Fig. 2.7a, Tab. 2.4, Fig. 2.8, Fig. 2.9). This suggests that DHS gains are largely cell-type specific. Few human DHS losses were identified as a DHS in any of the other human cell-types (Fig. 2.7b and Fig. 2.8, Fig. 2.9). In contrast, Common DHSs were detected in most other human cell-types (Fig. 2.7c, Fig. 2.8, Fig. 2.9) suggesting DHSs active among all three primates have more general roles in regulating transcription. Similar trends were seen when comparing LCL human DHS gain/loss/common regions (Tab. 2.5). This suggests Common DHSs mark DHSs present in most or all non-human
FIGURE 2.11: DHSs detected only in human, but not chimpanzee, LCLs are more enriched for ChIP signals in matched human cell-types. LCL histone modification and CTCF ChIP-seq data were previously generated from the GM12878 cell line (Ernst et al., 2011). Fisher’s exact test P value significance levels indicated by asterisks or NS (not significant) are provide for LCL human DHS gains compared to LCL human DHS loss and LCL common DHS regions (Tab. 2.6).

primate cell-types, as can be seen for chimp lymphoblast DHSs (Fig. 2.10i). Expected chimp and macaque DNase signal intensity are detected in orthologous regions (Fig. 2.7d, Fig. 2.8f). Similar to previous analysis of cell-type specific DHSs (Song et al., 2011), we found species-specific gains and losses of DHSs depleted in promoter regions relative to Common DHSs and enriched in distal intergenic regions and within introns (Fig. 2.2e).

We also compared chimpanzee DHS gains and losses to DNase-seq results from a diverse set of 27 human cell-types. We found that chimpanzee DHS gains did not largely overlap with DHSs from any of the 27 human cell-types (5–23%, Fig. 2.10a, Tab. 2.4) while chimpanzee DHS losses were more likely to overlap human DHSs, especially those from human fibroblasts (73%, Fig. 2.106b, Tab. 2.4). Thus, comparisons to diverse cell-types indicate that Common DHSs have been selectively maintained through millions of years of primate evolution suggesting a role in housekeeping function. In contrast, more recently evolved DHSs unique to humans and chimpanzees are likely functional in a small fraction of cell-types with related functions.
Table 2.5: Percent overlap of human-DHS gains, losses, and common detected in lymphoblasts with DHS sites identified in ENCODE human cell types. Note high degree of overlap with independently derived LCLs (GM).

<table>
<thead>
<tr>
<th>Culture</th>
<th>Human gains</th>
<th>Human losses</th>
<th>Common</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM12878</td>
<td>96.12%</td>
<td>16.02%</td>
<td>99.94%</td>
</tr>
<tr>
<td>GM19239</td>
<td>96.12%</td>
<td>14.92%</td>
<td>99.87%</td>
</tr>
<tr>
<td>GM19240</td>
<td>93.20%</td>
<td>14.36%</td>
<td>99.87%</td>
</tr>
<tr>
<td>GM18507</td>
<td>92.33%</td>
<td>12.15%</td>
<td>99.43%</td>
</tr>
<tr>
<td>GM19238</td>
<td>92.33%</td>
<td>19.34%</td>
<td>99.87%</td>
</tr>
<tr>
<td>GM12891</td>
<td>85.44%</td>
<td>20.99%</td>
<td>99.43%</td>
</tr>
<tr>
<td>Myotube</td>
<td>37.86%</td>
<td>20.99%</td>
<td>77.95%</td>
</tr>
<tr>
<td>Chorion</td>
<td>36.89%</td>
<td>23.76%</td>
<td>65.70%</td>
</tr>
<tr>
<td>HelaS3</td>
<td>35.92%</td>
<td>21.35%</td>
<td>77.38%</td>
</tr>
<tr>
<td>HelaS3_JFNA</td>
<td>35.92%</td>
<td>17.68%</td>
<td>77.20%</td>
</tr>
<tr>
<td>Myometrial</td>
<td>35.92%</td>
<td>18.23%</td>
<td>76.55%</td>
</tr>
<tr>
<td>MCF7</td>
<td>33.98%</td>
<td>13.81%</td>
<td>74.48%</td>
</tr>
<tr>
<td>Fibroblasts_park</td>
<td>33.01%</td>
<td>16.57%</td>
<td>80.48%</td>
</tr>
<tr>
<td>HUVEC</td>
<td>33.01%</td>
<td>21.55%</td>
<td>79.91%</td>
</tr>
<tr>
<td>Myoblast</td>
<td>33.01%</td>
<td>23.20%</td>
<td>78.33%</td>
</tr>
<tr>
<td>Pancreatic_islets</td>
<td>33.01%</td>
<td>20.44%</td>
<td>78.02%</td>
</tr>
<tr>
<td>FB0167P</td>
<td>32.04%</td>
<td>16.57%</td>
<td>77.57%</td>
</tr>
<tr>
<td>H1_ES</td>
<td>32.04%</td>
<td>24.86%</td>
<td>72.46%</td>
</tr>
<tr>
<td>NHEK</td>
<td>32.04%</td>
<td>15.47%</td>
<td>77.26%</td>
</tr>
<tr>
<td>H9_ES</td>
<td>31.07%</td>
<td>17.68%</td>
<td>74.48%</td>
</tr>
<tr>
<td>D721</td>
<td>30.10%</td>
<td>12.71%</td>
<td>69.36%</td>
</tr>
<tr>
<td>K562</td>
<td>29.13%</td>
<td>19.89%</td>
<td>78.27%</td>
</tr>
<tr>
<td>Melanocyte</td>
<td>29.13%</td>
<td>12.15%</td>
<td>77.83%</td>
</tr>
<tr>
<td>HepG2</td>
<td>28.16%</td>
<td>18.23%</td>
<td>75.99%</td>
</tr>
<tr>
<td>FB8470</td>
<td>27.18%</td>
<td>10.50%</td>
<td>72.52%</td>
</tr>
<tr>
<td>SM_SFM</td>
<td>23.30%</td>
<td>11.60%</td>
<td>75.05%</td>
</tr>
</tbody>
</table>

Table 2.6: Fisher’s exact test for histone modification and CTCF ChIP-seq peaks that intersect LCL Human DHS gains/losses/commons.

<table>
<thead>
<tr>
<th>K4me1</th>
<th>K4me2</th>
<th>K4me3</th>
<th>K9me3</th>
<th>K27me3</th>
<th>K27ac</th>
<th>CTCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human DHS gain vs loss</td>
<td>$&lt; 10^{-16}$</td>
<td>$&lt; 10^{-14}$</td>
<td>$&lt; 10^{-12}$</td>
<td>0.0005996</td>
<td>0.03788</td>
<td>$&lt; 10^{-12}$</td>
</tr>
<tr>
<td>Human DHS gain vs common</td>
<td>$&lt; 10^{-16}$</td>
<td>0.2836</td>
<td>0.5729</td>
<td>0.04043</td>
<td>0.0001605</td>
<td>0.00914</td>
</tr>
<tr>
<td>Human DHS loss vs common</td>
<td>$0.02426$</td>
<td>$&lt; 10^{-16}$</td>
<td>$&lt; 10^{-16}$</td>
<td>0.008462</td>
<td>$&lt; 10^{-16}$</td>
<td>$&lt; 10^{-16}$</td>
</tr>
</tbody>
</table>

2.2.3 *Species-specific DHSs are associated with biological function*

Species-specific DHSs were compared to cell-type matched human ChIP-seq data for multiple active histone marks and transcription factor binding sites. We found that human-only DHSs were better associated with these marks compared to chimpanzee-only DHSs (Fig. 2.11). This enrichment was highest for H3K4me1, H3K4me2, H3K4me3, and H3K27ac, consistent with chromatin marks predictive of enhancers (Heintzman et al., 2007;
Ernst et al., 2011) (Fig. 2.11, Tab. 2.6). H3K4 methylation signals were detected in a higher percentage of LCL human DHS gains compared to Common DHSs, while CTCF, a known insulator protein, is enriched in LCL Common DHSs (Fig. 2.11, Tab. 2.6). The combination of adjacent chromatin marks and their location relative to genes (Fig. 2.2e) provides further evidence that species-specific regulatory elements are functional. These data suggest most regulatory elements gained or lost after the human-chimpanzee divergence are preferentially associated with enhancers, while Common regions are preferentially associated with promoters and insulators.

2.2.4 Species-specific DHSs are near species-specific genes

We expect species-specific DHSs that contribute to phenotypic differences would be located near genes differentially expressed across species. To test this, we measured the proximity of fibroblast DHS gains and losses to genes with variable expression (Fig. 2.1a). From matched fibroblast expression data, we used edgeR (Robinson and Oshlack, 2010) analysis to identify 1047 human upregulated genes, 881 human downregulated genes, 785 chimpanzee upregulated genes and 788 chimpanzee downregulated genes. Human DHS gains were significantly enriched (permutation test, \( P = 0.00039 \)) near genes with increased expression in human and depleted (\( P = 0.008 \)) near genes with decreased expression in human (Fig. 2.12a–6b). Similarly, human DHS losses were enriched (\( P = 0.008 \)) near genes downregulated in humans and depleted (\( P = 0.002 \)) near genes upregulated in humans (Fig. 2.12b). The same relationships between DNase-seq signal and expression held true for chimpanzee (Fig. 2.12b). Analogously, we found that significantly upregulated genes were more likely to be near chromatin gains and downregulated genes near chromatin losses in each species compared to genes similarly expressed in both species (Fig. 2.13, Materials and Methods). These results support a direct role for species-specific DHS differences in species-specific gene regulation. The direction of these correlations indicate that DHS gains and losses are more commonly associated with enhancers than repressors. The LCL DNase-
Figure 2.12: (a) 58 human DHS gains (yellow arrow) overlapped human upregulated genes (HumanExpUp), significant enrichment compared to 100,000 random permutations ($P = 0.00039$). Only 17 human DHS gains (blue arrow) overlapped human downregulated genes (HumanExpDown), lower than random expectation ($P = 0.008$) (b) Comparison of DHS gains and losses with expression gains and losses. Yellow represents DHS and expression matches that occur more often than random permutations, while blue represent less often. P value indicated in each box. (c) Percentage of regions that display evidence of positive selection on the human (purple) or chimpanzee (brown) branch. Both human-specific DHS gains and losses show more evidence of positive selection on the human branch, while chimpanzee-specific DHS gains and losses show more evidence of positive selection on the chimpanzee branch (*$P < 0.03$, **$P < 0.002$). Common sites show an equivalent amount of selection on both branches. (d) Percentage overlap of DHS Gains (both human and chimpanzee combined), DHS Losses, and Common DHSs compared to evolutionarily constrained regions from GSC (Materials and Methods). Regions were divided into three compartments: promoter, intron, and intergenic regions. The black dot is the null expectation of finding a constrained region and error bars are one standard deviation.
FIGURE 2.13: Species-specific upregulated/downregulated gene expression levels are correlated with species-specific DHS gains/losses, respectively. (a) 48 Human upregulated genes (HumanExpUp) intersect genes that were located closest to human DHS gains (yellow arrow), which is higher than random permutations. 6 Human upregulated genes (HumanExpUp) overlap with genes located closest to human DHS losses (blue arrow), which is less than random permutations. (b) Comparison of upregulated genes (expression gains) and downregulated genes (expression losses) to genes located nearest to DHS gains and losses. P values were derived from 1000 random permutations (Materials and Methods).

seq and expression data from human and chimp show a similar trend (Fig. 2.14). Many species-specific expression differences were not readily explained by the presence of a nearby species-specific DHS. For example, though statistically, genes upregulated in human were enriched near human DHS gains, this was true for only 58 of 1182 higher expressed genes (Fig. 2.12a). This may be partially explained by our strict definition of human DHS gains. Also, long-range interactions may confound the simple way we assigned DHSs to the nearest gene. Future studies involving chromatin conformation capture (e.g., 3C, 4C, 5C) could be used to better map DHSs to target gene(s). Lastly, expression differences between species may result from transcription factor binding characteristics that do not alter chromatin structure. Comparative ChIP-seq studies for specific transcription factors will be necessary to determine the extent of this phenomenon. We conducted gene ontology enrichment analysis for both species-specific DHSs using GREAT (McLean et al., 2010) and differentially expressed genes using GO (http://david.abcc.ncifcrf.gov/), but did not
Figure 2.14: Comparison of chromatin DHS gains and DHS losses with genes that are upregulated and downregulated in human vs. chimpanzee lymphoblastoid cell lines (LCLs). Yellow represents chromatin and expression matches that occur more often than random permutations, while blue represent less often. P value indicated in each box.

Table 2.7: Testing for selection, P values (Two-tailed Mann-Whitney test). To test significance, for a given set of regions (Fibroblast: human DHS gain, human DHS loss, chimpanzee DHS gain, chimpanzee DHS loss, Common; LCLs: LCL human DHS gain, LCL human DHS loss, Common) we ordered the P values for selection on both hg19 and panTro2, then did a Mann-Whitney test to see if one branch has higher P values than the other.

<table>
<thead>
<tr>
<th></th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fibroblast Chimp Loss</td>
<td>0.00033</td>
</tr>
<tr>
<td>Fibroblast Chimp Gain</td>
<td>0.00182</td>
</tr>
<tr>
<td>Fibroblast Human Loss</td>
<td>0.00089</td>
</tr>
<tr>
<td>Fibroblast Human Gain</td>
<td>0.02283</td>
</tr>
<tr>
<td>Fibroblast Common</td>
<td>0.93697</td>
</tr>
<tr>
<td>LCL Human Decrease</td>
<td>0.95189</td>
</tr>
<tr>
<td>LCL Human Increase</td>
<td>0.97755</td>
</tr>
<tr>
<td>LCL Common</td>
<td>0.36911</td>
</tr>
</tbody>
</table>

find many highly enriched categories in either analysis. This indicates that chromatin gains and losses occur near many different types of unrelated genes representing a broad spectrum of gene ontologies.

2.2.5 *Species-specific DHSs show evolutionary selection and constraint*

The functional interpretations of Common and species-specific DHSs outlined above naturally lead to predictions about the operation of natural selection. We used HyPhy (Pond et al., 2005) to test for signatures of positive selection within DHS gains and DHS losses.
on either the human or chimpanzee lineage (Haygood et al., 2007) (Materials and Methods). Consistent with a functional change unique to humans, both human DHS gains and losses showed significantly more evidence for positive selection on the human branch than on the chimpanzee branch (Mann-Whitney $P = 0.03$ for gains and $P = 0.0009$ for losses, Fig. 2.12c, Tab. 2.7). Similarly, both chimpanzee DHS gains and losses showed increased positive selection on the chimp branch ($P = 0.002$ for gains and $P = 0.0004$ for losses, Fig. 2.12c, Tab. 2.7). Signatures of selection for Common DHSs were not significant on either branch. These results provide evidence that positive selection contributes to species-specific changes in chromatin, both gains and losses, and in the altered use and activity of gene regulatory elements.

Despite this connection with evolutionary pressures, only two DHS gains or losses in fibroblasts overlap previously defined human accelerated conserved non-coding sequences (HACNSs), chimpanzee accelerated conserved non-coding sequences (CACNSs), or human accelerated regions (HARs) (Prabhakar et al., 2006, 2008; Pollard et al., 2006a,b). More generally, few DHSs from any human cell-type we have analyzed, including embryonic stem cells, correspond to genomic regions of accelerated turnover. This lack of overlap may be due to the absence of DNase-seq data from specific developmental cell-types since HACNSs, CACNSs, and HARs have been associated with developmental gene regulation, or to regions of accelerated turnover representing a different type of genetic element not detected by DNase mapping.

We examined sequence conservation in DHS gains, losses, and Common sites using evolutionarily constrained regions defined by PhastCons (Pollard et al., 2006a; Siepel et al., 2005) and GERP (Cooper et al., 2005) algorithms with Genome Structure Correction (GSC) overlap test statistic (Birney et al., 2007; Parker et al., 2009; Bickel et al., 2010). By PhastCons analysis, we found that Common DHSs were the most conserved, a characteristic of regions under negative selection (Fig. 2.15). Common regions also had the greatest overlap with evolutionarily conserved elements, as defined by GERP (Fig. 2.12d; Materials
and Methods). The presence of Common DHSs in most human cell-types (Fig. 2.7c) with presumably greater functional demands may contribute to their higher conservation levels relative to gains and losses. Additionally, losses in both species were more conserved and overlapped more with conserved elements than gains (Fig. 2.15) suggestive of relaxed selection and positive selection, respectively. These trends held true even when noncoding genomic regions were partitioned based on their relationship to genes (promoter, intron, intergenic; Fig. 2.12d). In general, higher degree of conservation within specific regions of the genome can result from local differences either in mutation rate or selection (Graur and Li, 2000). Given that localized decreases in mutation rate below background are unusual, our data suggest that sequence conservation within Common DHSs is primarily driven by negative selection to maintain function.

A large fraction of DHS gains (≈70%), losses (≈60%), and Common (≈40%) sites did not overlap any highly conserved elements (Fig. 2.12d). Thus, many DHSs present
in all three species, and possibly many or all of 27 human cell-types, are not highly conserved. Understanding how these regions function in all species and cell-types without high sequence conservation poses an interesting challenge for evolutionary genomics.

### 2.2.6 DHS gains near DHS losses may indicate binding site turn-over

Previous studies have shown that individual transcription factor binding sites (TFBS) turn over rapidly during evolution (Bradley et al., 2010; Schmidt et al., 2010; Odom et al., 2007). Transposon-mediated shifts in the position of enhancers have also been documented between mouse and human (Mikkelsen et al., 2010). While these showed evidence of TFBS positional change, the turnover of entire DHSs have not been shown previously.
We identified ten possible instances of regulatory-element shuffling where a human DHS gain maps near (<50 kb) a human DHS loss (Fig. 2.16). These regions were found near genes associated with obesity (MCR4, Fig. 2.17), imprinting (GNAS, Fig. 2.17), and glial cell formation (METRNL, Fig. 2.17). We also found cases of nearby (<50 kb) human and chimpanzee DHSs that were independently gained (Fig. 2.17, Fig. 2.16). One region mapped within an intron of the SRGAP2 gene (Fig. 2.17), which is involved in neuronal guidance during brain development. Overall, the number of DHS gains and losses that mapped within close proximity to each other was not largely enriched or depleted based on randomized permutation tests, thus we cannot disprove that these findings are due to chance observations. Further detailed functional analyses are needed to determine the biological significance, if any, of these closely mapped regulatory changes.

**Figure 2.17:** Potential regulatory element shuffling associated with obesity (MC4R, imprinted gene (GNAS), glial cell formation (METRNL), neuronal guidance (SRGAP2).
2.2.7 DHSs deleted in the human or chimp lineage

Our analyses above focused exclusively on DHSs mapped to genome sequences shared between all three primate species. Recently, segments of DNA broadly conserved among mammals were found deleted specifically in the human (hCONDELs) or chimpanzee (cCONDELs) genome (McLean et al., 2011). It has been proposed that these largely gene-desert regions contain regulatory elements that contribute to species-specific phenotypes (Parker et al., 2009). We found human and chimp DHSs mapped to 6% of cCONDELs and 11% of hCONDELs supporting their role in species-specific gene expression. Many human fibroblast DHSs that overlap cCONDELs were also present in other human cell-types (Fig. 2.19) indicating that some CONDELs contain regulatory elements with pleiotropic consequences. Specific motifs may confer species-specific DHSs.

We analyzed TFBS motifs found within DHS gain, loss, and common sites across species to identify motifs associated with differences in hypersensitivity. To quantify differences, we determined log ratios of the best position weight matrix (PWM) score in a DHS between species (Materials and Methods). Most motif scores from the JASPAR database were distributed evenly between species (log ratio near zero) indicating no species-specificity trend for that motif. However, log ratios of AP1 motif scores deviated from zero and correlated with species-specific DHSs (Fig. 2.18). For example, in human DHS gains, AP1 motif match scores were higher in the human sequences and lower in the orthologous regions in chimp and macaque (Fig. 2.18a). In contrast, AP1 motif scores in human DHS losses were higher overall in both chimpanzee and macaque sequences compared to human (Fig. 2.18b). Common regions showed even distributions of AP1 motif scores across all three species (Fig. 2.18c). This trend was also found in chimpanzee where chimp DHS gains had higher AP1 motif scores in chimp sequences compared to orthologous regions from human and macaque (Fig. 2.18d), and chimpanzee DHS losses had higher AP1 motif scores in human and macaque (Fig. 2.18e). In a representative human DHS gain, we see that
Figure 2.18: (a–e) Scatterplots showing the enrichment of AP1 motif matches in species with increased hypersensitivity. Each x represents a single DHS. (a–c) Positive values on each axis indicate better motif matches on the human branch. For these regions, points in the upper-right quadrant are regions where the AP1 motif scores better in human than either chimp or macaque, whereas the lower left represent AP1 motif scores worse in human. The number of DHSs in these quadrants are indicated. (d–e) For chimp gain and loss regions, positive values for each axis indicate a better motif match in the chimp branch. (f) The AP1 motif from JASPAR and an example alignment of a representative human gain region representing a point along the diagonal in the upper-right quadrant in panel a. (g) Boxplots summarizing the results from AP1 and three other motifs. The boxplots show the distribution of (combined) log-ratios (relative to the appropriate species). P values for differences relative to common regions are significant (asterisk) in all 4 comparisons: human DHS gains, $P < 10^{-31}$; human DHS losses $P < 10^{-3}$; chimp DHS gains, $P < 10^{-13}$; chimp DHS losses, $P < 10^{-8}$ (Materials and Methods). In AP1, the significant trends illustrate the same principal observed in panels a–e. Most other transcription factors (such as SP1 and SOX10) have plots that show no pattern. ZEB1, a repressor, has an inverse relationship.
FIGURE 2.19: Chimp Condel (cCondel) regions heatmap. Shown are DNase-seq signal intensities (maximum parzen scores) of the 344 cCONDEL regions. The color scale (log) ranges from red, representing the highest signal intensities, down to black, representing little/no signal intensity. All 15 primate samples (fibroblasts and LCLs) used in our DNase-seq analysis and 27 different ENCODE cell lines are represented. Ubiquitous DHS sites that overlap cCONDELs are clustered on the right, and more common DHS sites are found in the middle. Note that these regions are do not have signal in chimpanzee data (CL or CF samples), and a small number also do not contain signal in macaque data (QF samples).
the human allele results in a better match to the canonical AP1 motif than the non-human primate alleles (Fig. 2.18f).

These results suggest species-specific sequence changes within the AP1 motif promote hypersensitivity in some species-specific DHSs in the human, chimp, and macaque genomes. AP1 was the clearest example of this from motifs represented in JASPAR (Fig. 2.18g). Other transcription factors may be acting similarly, but less frequently. For example, we also found that ZNF354C and NFE2L2 showed similar trends to AP1. In these cases, motif scores positively correlated with the presence of a species-specific DHSs. In contrast, ZEB1 displayed the opposite trend where higher motif scores correlated with the lack of a species-specific DHS (Fig. 2.18g). While the mechanism is not yet clear, our findings and ZEB1’s known role as a repressor (Papadopoulou et al., 2010) is suggestive of its ability to induce a closed chromatin state via binding to CtBP and HDAC (Zhang et al., 2001).

2.3 Discussion

Precise measurements of transcript abundance enabled by RNA-seq experiments have revealed extensive differences in gene expression among closely related species (Brawand et al., 2011) with 10–20% of transcripts within a given tissue found differentially expressed between humans and chimpanzees (Blekhman et al., 2008; Babbitt et al., 2010). Many transcripts are tissue-specific, and given the relatively small number of cell-types explored, the total number of differentially expressed genes is likely to be considerably larger. An important goal of molecular evolution research is to understand how differences in transcript abundance have evolved, both because the changes are extensive and because some may underlie the origin of functionally significant traits (Carroll, 2008; Wray, 2006; Orgogozo et al., 2007).

Most gene expression differences across species likely have a genetic basis, but it is difficult to relate expression changes to variation in genome sequences. While many
non-coding sequence differences are unlikely to impact transcription, for the subset that do, it is often not clear what genes are directly affected. In addition, a non-coding regulatory mutation may only affect gene expression in a subset of tissues or developmental stages, so many functional consequences have gone unrecognized given the limited number of studies performed thus far. Further complicating analysis, transcription is influenced by environmental factors and by epigenetic modifications. But the lack of a complete regulatory element map across species and tissues is perhaps the most important impediment to understanding gene expression differences in terms of genome sequence evolution. Changes in transcript abundance may be caused by genetic differences within individual regulatory elements in cis that affect transcription factor binding affinity (Degner et al., 2012), or within transcription factors that affect binding to many regulatory elements in trans. Even when the genetic basis is known to be in cis, there is no reliable method for identifying the causal mutations from sequence comparisons. As a result, distributions of positive and negative selection genome-wide correlate poorly with changes in transcript abundance (Holloway et al., 2007; Good et al., 2006; Babbitt et al., 2010).

In this study, we showed that analyzing chromatin accessibility using DNase-seq provides a powerful approach to link genome sequence changes to species- and tissue-specific differences in gene expression. Chromatin accessible DHSs have three properties that make them especially valuable for evolutionary analyses of gene expression (Boyle et al., 2008; Song and Crawford, 2010). First, DHSs identify all known functional classes of regulatory elements, including core promoters, enhancers, repressors, boundary elements, and locus control regions, thus revealing all cis components of transcription through a single genome-wide assay. Second, DHSs are only found when a regulatory element is active or poised, which means that DNase-seq can be used to identify evolutionary changes in tissue- and developmental stage-specific regulatory elements. And third, DHSs represent only 2% of the genome, making it possible to focus analyses on regions that are involved in transcriptional regulation and ignore regions that are not.
We performed DNase-seq on fibroblasts from three primate species and identified more than two thousand regulatory elements apparently gained or lost since the divergence of humans and chimpanzees. Turnover of regulatory elements was enriched near genes that display species-specific expression differences, indicating that gains and losses in DHSs have functional consequences on transcript abundance. To our knowledge, this is the first evidence correlating changes in DNase chromatin accessibility and gene expression across species at a genome-wide scale. We found most expression differences occurred without a detectable change in a nearby regulatory element. One possibility is that mutations within DHSs affect transcription factor binding without causing large changes in overall chromatin accessibility. Future experiments are needed to identify the specific sequence changes that regulate expression at long distances and/or via post-transcriptional mRNA stability mechanisms.

Most regulatory element changes occurred within intergenic regions and introns and were predominantly associated with cell-type-specific DHSs. These results are consistent with expected differences in the extent of pleiotropy: loss of core promoter elements will more likely affect transcription in many tissues and stages of development, while loss of distal enhancers will more likely affect transcription in a subset of tissues. Lower rates of change in core promoter elements and in regulatory elements actively utilized in multiple tissues suggest negative selection is operating to maintain regulatory elements with more critical functions.

Analysis of the DNA sequences within regulatory elements provides evidence for the operation of natural selection within these elements. Sequence within DHSs utilized across all three species show lower rates of substitution than surrounding DNA, which is a proxy for neutral evolution by drift, consistent with negative selection operating to maintain their function. In contrast, regulatory element gains on the human and chimpanzee branches have significantly elevated rates of substitution, consistent with positive selection for altered function, while regulatory element losses show slightly elevated rates, perhaps due to
relaxed selection. Explicit tests for positive selection using branch-specific likelihood ratio tests (Haygood et al., 2007) reveal that the highest association is with regulatory element gains and the lowest with common regulatory elements utilized in all three species. Thus, the genome-wide distribution of both negative and positive selection within regulatory elements correlates in predicted ways with the evolutionary conservation and change in their function. Although we are not aware of any previous evidence for such a relationship, it seems likely in principle that the operation of natural selection is often tied to gains, losses, and conservation of regulatory elements. Most instances of inferred positive selection we identified do not overlap previously described HARs (Pollard et al., 2006a) or HACNs (Prabhakar et al., 2008) highlighting that our DHS gains and losses represent a novel set of differential regulatory elements may have played a role in adaptation during human evolution.

Many studies have documented evolutionary gains and losses of individual transcription factor binding sites or H3K4me3 histone marks among related species (Bradley et al., 2010; Schmidt et al., 2010; Odom et al., 2007; Cain et al., 2011), but this is the first evidence showing gains and losses of entire DHSs. Since we only examined two cell-types and applied conservative identification criteria, the full extent of regulatory element changes between humans and chimpanzees is likely to be considerably greater than we report. Nonetheless, the instances of turnover we identified suggest regulatory element gains and losses are a common class of functional change within evolving genomes.

We show that sequence differences among species within particular motifs may result in species-specific DHSs, which suggests one way non-coding regulatory variants can alter chromatin structure. In particular, mutations that produce better matches to the activator protein 1 (AP1) motif on either the human or chimpanzee genome correlate with the presence of species-specific DHSs, a result detected in human DHS gains and losses as well as chimpanzee gains and losses. Sequence changes increasing the affinity for AP1 motif more likely drive species-specific changes in chromatin structure rather than species-specific coding mutations within the AP1 components, FOS and JUN proteins, altering the
sequence-binding preference of AP1. Since only a minority of species-specific DHS gains and losses has differential AP1 motif scores (Fig. 2.18), this indicates the majority of factors that govern species-specific DHSs remain to be discovered. AP1 has been implicated in many aspects of cellular function ranging from proliferation, transformation, differentiation, oncogenesis, apoptosis, hormone activation, to tumor suppression (Shaulian and Karin, 2002; Karamouzis et al., 2007; Eferl and Wagner, 2003; Biddie et al., 2011). We provide evidence that other factors act similarly to AP1 or in the opposite direction as repressors, such as ZEB1. Mapping DHSs across a more diverse set of primate samples, as well as using additional de novo motif discovery and performing ChIP-seq to reveal binding sites, will be an important part of identifying additional factors that confer changes in chromatin structure across species.

2.4 Methods

2.4.1 Cell-types and tissue culture

We obtained two cell-types from Coriell for this study: skin fibroblast cells and lymphoblastoid cell lines (LCLs). Primary skin fibroblast cells from three human, three chimpanzee, and three macaque individuals. LCLs, which are B cells immortalized with Epstein-Barr Virus, were obtained from the same three human and three chimpanzee individuals that fibroblasts were isolated from (Tab. 2.1). EBV does not reliably transfec macaque lymphocyte cells, so matched macaque LCLs cells were not available for this study. Importantly, other recent genome-wide studies that used macaque LCLs were of B-Lymphocyte cells transformed with rhesus herpes papio virus, a close relative of human EBV (Cain et al., 2011). Cells from all species were grown in standard growth media. Fibroblast growth media consisted of Gibco’s MEM (10370-021), L-Glutamine (25030-081), Pen/Strep (15140-122), and 10% FBS (Hyclone SH30070). LCLs growth media consisted of Gibco’s RPMI (21870) media with L-Glutamine, Pen/Strep, and 15% FBS. We harvested fifty
million cells for each individual biological replicate and allocated 35 million cells for
DNase assays (DNase-seq and DNase-chip), 10 million for genomic DNA (used as control
for DNase-chip array hybridization), and 5 million for RNA DGE-seq expression analysis.

2.4.2 DNaseI HS library preparation

DNase-seq libraries we generated as previously described (Boyle et al., 2008; Song and
Crawford, 2010) and sequenced via Illumina’s GAII sequencer. DNase-chip library prep-
arations, used for validation of our DNase-seq results were performed as previously de-
scribed (Crawford et al., 2006; Shibata and Crawford, 2009) and were hybridized to 1%
ENCODENimblegen arrays (Birney et al., 2007). Custom arrays were designed to cover
the orthologous regions from chimpanzee and macaque. DNase-chip array intensities
were compiled and significant DHSs were called using ChIPOTle (Buck et al., 2005)
($P < 0.000001$ peak cutoff).

2.4.3 DNase-seq aligned to native genomes and lifted to hg19

DNase-seq data generated from each species was aligned to the native genome (human
hg19, chimpanzee panTro2, and macaque rheMac2) using BWA [63]. To directly compare
three different primate species requires that they be aligned to a single reference sequence.
Because both the chimpanzee (panTro2) and macaque (rheMac2) reference sequences were
built from the existing human reference, we converted all sequences to human coordinates.
To do this, we converted each 20-mer DNase-seq sequence from panTro2 or rheMac2 to
hg19 with liftOver [23], using a match setting of 80 percent. After conversion to hg19
coordinates, we used F-seq (Boyle et al., 2008) to identify DNaseI hypersensitive (DHS)
sites. The F-seq scores from the top 100,000 peaks from each sample were used to determine
how well chromatin openness correlates among all 15 samples (Fig. 2.1). This analysis
uses a pairwise Pearson correlation to compare the similarity among samples.
2.4.4 Identifying species-specific hypersensitive sites

We used the bioconductor edgeR package to define species-specific hypersensitive regions (Robinson and Oshlack, 2010). EdgeR is designed to detect differences in count data among groups of samples. Briefly, it compares within-group variances to between-group variances using a negative binomial model, and selects entries with significant between-group differences. It was designed for differential expression data such as DGE-seq or RNA-seq, but it is similarly applicable to read counts generated by DNase-seq. One key advantage of edgeR is a normalization procedure specifically designed for high-throughput sequencing studies (Robinson and Oshlack, 2010). To locate significant differences in DNase-seq signal between species, we first identified the union set of the top 100,000 DHSs (as scored by F-seq) from all 15 samples (9 fibroblasts and 6 LCLs). We used bx-python (https://bitbucket.org/james_taylor/bx-python) to analyze the data. We divided these regions into windows, attempting to maximize the resolution of the windows while minimizing the number of windows required:

**Defining windows.** We divided the union set into overlapping windows of 300 bp. DHSs smaller than 300 bp were expanded to 300 bp. Regions larger than 300 bp were tiled with overlapping windows; the overlap varies depending on the size of the hypersensitive region to tile. We start by finding the number of windows that would fit completely inside the defined HS site using the default overlap (100 bp). If these windows discard fewer than 10% of the bases on each edge of the HS site, we tile the site using these windows, because the initial peak calls tend to run a bit wide. If using the default overlap would cause us to lose more than this edge threshold (10%), we add another window and adjust the overlap so the windows exactly cover the entire HS region. These windows cover almost all of the HS bases in the original, while minimizing the number of non-HS bases considered for the downstream analysis. After defining the windows, we liftOver each window back to chimp
and macaque and discard any windows that don’t map. We also eliminated any windows mapping to the Y chromosome, because we have female samples. This resulted in 1.2 million windows across all DHSs.

Zero counts filter. We next counted the number of reads (DNase cuts) that mapped within each window for each sample. To be certain to compare only regions that are actually found in all 3 primates, we omitted any windows that did not have at least one read count in at least one individual from each species. We believe this is a reasonable filter since there is a basal level of DNase digestion throughout the genome, and regions without any mapping reads in one species are likely caused by alignment ambiguities (such reads are lost in the liftOver process) or reference problems.

Significant differences. We then used edgeR to call windows with significantly different counts in each pairwise comparison at a false discovery rate of 1%. This defined species-specific HS windows. For example, to define human-fibroblast-specific (human DHS gain) regions, we selected any windows with significantly more counts in humans compared (pairwise) to both chimp and macaque fibroblasts. We then merged any significant neighboring windows.

Reciprocal liftOver filter. To confirm that DNase sites detected in only a single species were not due to a liftOver artifact or copy-number variation (CNV), we performed a reciprocal liftOver. Reciprocal liftOver discards regions that do not map 1:1 between species assemblies. After defining a species-specific or Common DHSs (Common DHS described below) in human coordinates, we the lifted these windows back to panTro2 or rheMac2. We then compared number of reads from the original native genome alignment to the read count on the lifted (hg19) reference. We found that, in general, these scores correlate well. However, some windows differ in counts on each reference, indicates
liftOver artifacts or CNVs. Windows with native counts exceeding the reference counts could result from counts from the native assembly being lifted to multiple places on hg19. Likewise, windows with more reads on hg19 than on the equivalent native assembly could be caused by several regions from the native genome lifting to a single location on hg19. Since these may possibly artificially appear as chromatin gains or losses, we filter out these scenarios by requiring that the number of reads on hg19 for chimp/macaque data be within 10% of the number of reads on the native assembly.

*Sharp peaks and assembly filters.* Some regions mapped to areas in the genome that are clear artifacts, most likely CNVs. These artifacts are expanses where one of the species has much higher levels of hypersensitivity than the others, and are often located near centromeres or telomeres. A DHS resulting from a massive expansion in one of our samples compared to reference genome appears as a large (kb to Mb-scale) DHS encompassing the entire CNV and/or DHSs that cluster closely together. We manually curated a list of such areas and filtered data mapping within these regions (Assembly filter). We also eliminate abnormally sharp peaks, which are likely caused by PCR artifacts, reference assembly inaccuracies, or small differences in copy number between our samples and the reference genomes. If any 30 bp window within a region contains 75% or more of the reads in the entire region, these regions are filtered out as sharp peaks. Ultimately, we confirmed the regions by visually inspecting many using both the Integrative Genomics Viewer (IGV) (Robinson et al., 2011) and the UCSC Genome Browser (Karolchik et al., 2012). These filters apply to DHS gains, losses, and common regions.

2.4.5 *Skin biopsy location and Fibroblast heterogeneity*

Chang et al. (2002) showed that human skin- and non-skin- fibroblast samples collected from different locations along the body plane showed notable differences in transcriptional profiles. Aware of this issue, we made an effort to use fibroblast cultures made from only
skin samples and from the same region of the upper arm. All 3 replicates of macaque Fibroblasts and one human Fibroblast were confirmed from Coriell to be from skin biopsies from the upper arm (the other two locations were unknown). Since our analysis poses a strict requirement of DHSs to be present in all three human individuals to be called a human DHS gain, having at least one human sample with the biopsy site confirmed to be from the upper arm ensures that the human gains are not a result of human DHS gains being a result of, for example, all human fibroblasts isolated from lower leg. Likewise, to be called a human chromatin loss, DHSs have to be absent in all three human samples (but present in all chimp and macaque samples); this biopsy location bias is again mitigated by at least one human sample being from the upper arm.

The Yerkes National Primate Center, from where the chimpanzee skin fibroblasts were obtained, unfortunately does not document the exact location of biopsy. While the standardized skin punch protocol calls for the location of the skin biopsies to be from the ear pinna (personal communication with Fawn Conner-Stroud from Yerkes), we cannot be 100% sure that the samples were isolated from this location. As described above, human DHS losses are a result of signal being present in all three chimpanzee and macaque samples, supporting that these regions are not due to chimp biopsy location. We also want to reiterate that our skin fibroblast results are highly similar in LCL lines that are all uniformly derived from blood samples.

We find that none of our chromatin gains and losses overlap the Hox genes described in the Chang et al. (2002) paper. In addition, a more recent analysis by Rinn et al. (2006) using more comprehensive microarrays and more biopsy sites, identified 337 expression array probes (299 unique genes) that were shown to be highly associated with five different general biopsy site locations. We have compared this list of genes to both species-specific chromatin gains/losses, as well as species-specific gene expression, and find <3% of the species-specific and common DHS overlap with the 299 position specific genes. Similarly, species-specific and commonly expressed genes also show <3% overlap (Tab. 2.8).
Table 2.8: Genes located closest to human-specific DHS gain/loss and common regions as well as genes identified as differentially expressed among the 3 primate species are not highly associated with genes used to identify Fibroblast biopsy locations in Rinn et al. 2006

<table>
<thead>
<tr>
<th>DHS</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>human DHS gain (836 closest genes)</td>
<td>20</td>
<td>2.4%</td>
</tr>
<tr>
<td>human DHS loss (286 closest genes)</td>
<td>5</td>
<td>1.7%</td>
</tr>
<tr>
<td>chimpanzee DHS gain (676 closest genes)</td>
<td>20</td>
<td>3%</td>
</tr>
<tr>
<td>chimpanzee DHS loss (211 closest genes)</td>
<td>6</td>
<td>2.8%</td>
</tr>
<tr>
<td>common DHS (1259 closest genes)</td>
<td>18</td>
<td>1.4%</td>
</tr>
<tr>
<td>EXPRESSION</td>
<td></td>
<td></td>
</tr>
<tr>
<td>human upregulated genes (1047)</td>
<td>27</td>
<td>2.6%</td>
</tr>
<tr>
<td>human downregulated genes (881)</td>
<td>19</td>
<td>2.2%</td>
</tr>
<tr>
<td>chimpanzee upregulated genes (785)</td>
<td>20</td>
<td>2.5%</td>
</tr>
<tr>
<td>chimpanzee downregulated genes (788)</td>
<td>19</td>
<td>2.4%</td>
</tr>
<tr>
<td>commonly expressed genes (1365)</td>
<td>10</td>
<td>0.7%</td>
</tr>
</tbody>
</table>

2.4.6 Identifying DNaseI hypersensitive sites in all species (Common sites)

In order to ensure that our tests for selection were meaningful, we wanted to compare chromatin gains and losses with a set of regions that were open in all species. Rather than simply choose DHSs that have the highest scores, we wanted to mirror the level of hypersensitivity to that of the species-specific regions. This is important because species-specific DHSs are not necessarily the strongest DHSs. We also wanted to select a set of regions similar in size to our sets of gains and losses to retain computational tractability.

To select a set of matched Common DHSs, we required that each window be similarly open in all 9 samples from all 3 species. To be considered similarly open in a given sample, the number of counts must lie between the 20% and 80% quantiles for that sample in the corresponding species-specific regions. For example, we used the human DHS gains identified by edgeR to define the distribution for each of the three human samples, and similarly for chimp and macaque. As such, our set of Common regions is the set of all windows with DNase counts within this range for each of the 9 samples (Fig. 2.2d). To reduce the number of Common regions we found to the most representative set (those that most closely match the average signal intensity of the differential DHSs), we narrowed the quantiles until we found a set of around one thousand Common regions, which we reasoned
would be a sufficient number to examine summary statistics. To ensure that our results are not biased for a specific set of Common regions, we repeated our experiments on a significantly larger set (11,000) of Common DHSs using less stringent criteria (10%-90% quantile). This larger set is even more enriched for promoter regions but does not change our conclusions (data not shown).

After identifying an initial set of potential Common DHSs, we filtered out any that appeared to be appendages to other hypersensitive sites. Without this step, many Common DHSs would map to the edge of strong hypersensitive sites. To ensure that a Common DHS is a standalone DHS, we examined the neighboring windows surrounding the initial set of Common DHSs. If a Common DHS window contained fewer than 80% of the number of reads in the adjacent window on either side, we filtered it out as most likely an appendage to a stronger DHS. This resulted in a final list of 1259 Common DHSs matched in intensity to the species-specific DHSs. We also ran this filter on DHS gains and losses, and found that very few of the gains (3–5%) and losses (3–8%) get flagged as appendages. Of these, many of them are flagged as a result of FDR threshold issues that simply didn’t quite highlight a neighboring window, and we would actually still consider this a legitimate gain region. Because gain/loss appendages are relatively rare and are largely due to threshold issues, we elected to retain them in our final list.

2.4.7 Why are there more species-specific increases than decreases?

In every comparison, we reported more species-specific gains than losses. The most important factor in determining the size of these lists is the level of significance we set by choosing a FDR. To obtain lists that match in length, we could simply adjust the FDR value for the lists to yield about the same number of regions. Instead, we decided to keep the FDR constant and select varying numbers of DHSs for each category.

However, it is still constructive to consider the disparity. In other words, at a constant FDR (1%), why are there more species-specific increases than decreases? This is possibly
a result of purifying selection. Because DHSs are regulatory (and therefore tend to be conserved), a loss of a DHS probably confers a fitness disadvantage. In this case, we would expect to see more gains than losses. It is also possible that the prevalence of increases is simply a result of the way we constructed the significance test. A DHS is a sparse signal (there are more closed regions than open regions). Combined this with the asymmetry of the evolutionary tree: the chimp and human are more similar to each other than either is to the macaque. A human-specific increase requires both macaque and chimp to be closed (the default), while a human-specific decrease requires both macaque and chimp to be open. This latter scenario will happen less often because the relationship between chimp and human is closer than either to the macaque.

In short, the greater number of gains than losses in our analysis may reflect purifying selection on DHSs; however, it may also simply be a result of the way we constructed the test, particularly due to using an outgroup species to polarize the chromatin structure changes.

2.4.8 Testing for selection

We tested for evidence of positive selection using the DHSs indicated as DHS gains, losses, and commons defined by edgeR and common analyses (see above). A branch model test (Zhang and Dietrich, 2005) in HYPHY (Pond et al., 2005) was used to assess evidence for positive selection on each the human and chimp branches. HYPHY uses a likelihood ratio test to compare two opposing models. For the null hypothesis, we specified a composite model that allowed for negative selection, neutral evolution, or relaxed constraint specifically on the branch of interest (i.e. the human branch), with negative or neutral evolution across the rest of the tree. The alternative hypothesis modeled positive selection only on the branch of interest, with negative or neutral evolution on the rest of the tree. For each region, HYPHY performed a likelihood ratio test comparing these hypotheses and output a P value that can be interpreted as a level of evidence for positive selection. In
order to test the likelihood of either the null or alternative hypothesis, we specify both the alignment of the region of interest, as well as a background sequence alignment assumed to be evolving neutrally (Wong and Nielsen, 2004). For the alignment of the region of interest, we used alignments of human, chimp, macaque, and orangutan precomputed at UCSC. For the background sequence, we collected a separate set of local introns for each region to test, following Haygood et al. (2007). To define these background alignments, we started with the UCSC knownGene definition of intron annotations, and then filtered out all first introns, splice junctions, and hypersensitive sites (in any of the 15 samples in this study). In this way, we aimed to select sequences that are evolving neutrally. After defining this set of neutral introns, we used an expanding window centered on the region of interest to collect introns in a region up to 100 kb surrounding the center. We added introns to this collection sequentially as the window expanded until we reached an alignment of 2000 nucleotides. If we were unable to find 2 kb of background introns within 100 kb of sequence, we discarded these regions (this happens rarely). Introns are commonly assumed to be evolving neutrally (Graur and Li, 2000; Gilbert, 1978), particularly when our filtering steps are taken into account; however, there are still likely to be regulatory sequences present in our background model, either due to sequences containing DNaseI HS sites in other cell-types not tested or due to unannotated or mis-annotated transcripts. In order to further correct for this possibility, we performed each likelihood-ratio test 50 times, using 50 different bootstrapped versions of the background model. We then averaged these P values to assign a final P value for each region. This method has the effect of possibly discarding any elements under selection in some of the bootstrap replicates, increasing our ability to detect positive selection even if we inadvertently chose some background regions under selection. To test significance, for a given set of regions (e.g. human DHS gains) we ordered the P values for selection on both hg19 and panTro2, then did a Mann-Whitney test to see if one branch has higher P values than the other (Tab. 2.7).

The fibroblast DHSs where we can polarize the differences using macaque all have
significant differences in the direction we expect, while the Common regions do not have significant differences. In the LCLs, where we are unable to polarize (no macaque LCLs were available), we do not detect a significant difference. This is likely due to a combination of two categories (gains and losses) that have competing selection (i.e. LCL human DHS loss = human DHS loss+chimpanzee DHS gain).

2.4.9 Testing for evolutionary constraint

We calculated the observed fraction of overlaps between DHSs and evolutionarily constrained regions using constrained regions defined by the Genome Evolutionary Rate Profiling (GERP) method (Cooper et al., 2005) executed on Enredo, Pecan, Ortheus (EPO) (Paten et al., 2008b,a) 33-way alignments. EPO alignments and GERP regions are available for download at the Ensembl browser (http://ensembl.org). We then constructed a null distribution of the fraction of expected overlaps by using the conservative Genome Structure Correction (GSC) methodology described previously (Parker et al., 2009; Bickel et al., 2010; Birney et al., 2007). We performed 1000 randomizations and calculated the mean and standard deviation from the null distribution to assess the statistical significance of the observed overlap (Fig. 2.12d).

We also used PhastCons to explore degree of sequence conservation. For each region, we extracted the mean and max PhastCons score from the primate PhastCons table at UCSC. We then compared the distribution of these scores across the regions to see how sequence conservation is related to hypersensitivity conservation (Fig. 2.15).

2.4.10 Expression analysis using digital gene expression (DGE-seq)

Total RNA purified from 5–10 million cells harvested from the same cell culture used for DNase-seq were also used to generate DGE-seq expression libraries as previously described (Babbitt et al., 2010; ’t Hoen et al., 2008). Polyadenylated RNA is captured for enrichment of mRNA and the oligo dT primer is used to make cDNA. Briefly, DGE-seq is
similar to Serial Analysis of Gene Expression (SAGE) where mRNA abundance is assessed via counting short sequences of their cDNA at specific restriction site locations. These DGE-seq libraries were sequenced using Illumina’s GAII sequencer, and averaged 10 million 20mer sequences for each sample, which were then aligned to the samples’ native reference sequence using BWA. We used EdgeR to detect differences in tag counts across species by comparing intra-species variances to inter-species variances using a negative binomial model, and selects expressed genes with significant between-species differences (Robinson and Oshlack, 2010).

2.4.11 Cross-species correlation of genome-wide expression

Unlike the analysis performed for comparing cross-species DHSs, we did not liftOver any non-human expression sequences to human. Instead, we simply compared the DGE-seq sequence counts that aligned to exonic regions within each species’ native sequence alignment. Because of the high level of homology of the exonic regions between the 3 primate species, we directly compared tag count numbers between each of the orthologous genes. For the genome-wide expression correlation comparison (Fig. 2.1), we normalized tag counts using edgeR to calculate the total library size for each sample and adjusting the tag counts accordingly so that relative differences between the depth of the sequencing reads did not influence the results. Next, we filtered out genes that did not have at least 10 combined tag counts between all of the samples to decrease the noise associated with genes that fall below the meaningful level as recommended for edgeR. Following these filters, we compared the Spearman correlation values between all of the samples and plotted the results as a heatmap with hierarchical clustering to show similarities within and between species and tissue types.
2.4.12 Identification of differentially expressed genes between species

Using edgeR, we identified genes that were significantly differentially expressed between the 3 primate species. Comparisons between species were performed on a pairwise manner comparing 3 individuals of one species against 3 individuals from a second species. The same normalization method and filters used in the expression correlation analysis was also used prior to defining the differentially expressed genes. Differential gene expression was defined as genes having a P value cutoff of 0.05. Using the Macaque expression result as an outgroup, we identified genes that displayed high or low expression specifically on the human and chimp branch. For example, genes we defined as highly expressed in human (human upregulated genes) are differentially expressed in both human/chimp and human/macaque comparisons, but not differentially expressed in the chimp/macaque comparison. Similar criteria were used to identify genes that display low expression in human compared to chimp and macaque (human downregulated genes).

2.4.13 Correlation of species-specific DHSs with expression

To firmly establish the connection between differential chromatin and differential expression, we tested for significance in overlap in both directions: First, we tested if differential DHSs tend to be located near differentially expressed genes, and second, we tested if differentially expressed genes tend to have differential DHSs nearby.

Comparing differential DNase-seq signal to differential gene expression. We compared human DHS gain, human DHS loss, chimpanzee DHS gain, and chimpanzee DHS loss regions to species-specific genes (e.g., human upregulated/downregulated genes, etc.). We first assigned each DHS its nearest gene, then intersected these with each differential expression set. The number of overlaps for every pairwise combination of DHSs vs. differentially expressed genes are shown in Tab. 2.5. We calculated P values by permuting (100,000 times) the set of all expressed Fibroblast genes, randomly selecting the number of genes for
each comparison, and intersecting those random sets with the genes nearest species-specific DHSs. This established a null distribution of overlaps in random intersects.

Comparing differential gene expression to differential DHSs. To establish the connection in the opposite direction, we did a similar analysis in the reverse direction. Using the UCSC knownGene table, we merged all isoform coordinates for each gene and found all DHSs within a surrounding 20 kb window. We calculated pairwise overlaps of DHSs between each differential DHS list and these lists of all nearby DHSs. If a gene contained a differential DHS within 20 kb, it was counted as a match. The number of overlaps for every pairwise combination of sites near differentially expressed genes vs. differential DHSs are shown in Tab. 2.6a. We then permuted genes similar to above (1,000 times), located all nearby DHSs, and tested for overlap to create a null distribution of overlap count.

2.4.14 Variation in motifs in species-specific DHSs

To connect sequence changes to species-specific DHSs, we compared JASPAR motif scores across species. We first extracted the orthologous DNA sequences for human, chimpanzee, and macaque for each of our DHS gain, loss, and common sites. We scanned and scored each of these sequences for all the position weight matrices (PWMs) in the JASPAR database. We scored a sequence for a given PWM as the highest motif score anywhere in that sequence. This resulted in a region-by-motif matrix of scores; each score is the highest score for each motif/sequence combination. To compare species, we took the log ratio of scores (human/chimp, human/macaque, and chimp/macaque). Where this score is 0, the highest score does not differ between species. Deviations from 0 indicate the direction of improvement in motif match (in a human/chimp comparison, a positive number means the best match in the human sequence scored higher than the best match in the chimp sequence). After calculating these scores and each pairwise log-ratio, we plotted the log-ratios (Fig. 2.18) to examine trends. We calculated the pairwise log-ratios for
multiple species comparisons and plotted these on different axis to check whether increases over one species correlate with increases over the other. In these two-dimensional plots, each axis quantifies a different pairwise species relationship. Points that cluster in the upper-right quadrant have higher scores than either of the other species; points that cluster in the lower-left have lower scores. The most interesting variation in these plots is whether the points congregate in the upper-right or lower left. To assess significance, we project each data point onto the $y = x$ line to reduce the dimensionality to 1. We then used the Wilcoxon rank-sum test to compare each distribution to the distribution of the common regions (Fig. 2.18g).
Clusters of Regulatory Elements in Humans

This research chapter contains unpublished research as well as a modified excerpt of a research article published by Lingyun Song et al. in the journal Genome Research in 2011 (Song et al., 2011).

3.1 Introduction

Besides differences among species, regulatory elements also govern differences across cell-types within a single species. Each cell-type contains the same genome, but activates a unique set of regulatory elements to effect a different phenotype. Therefore, identifying regulatory elements for every cell-type requires an experiment for each cell-type. Such experiments were part ENCODE project, with a goal essentially to identify the active regulatory elements in a wide range of cell-types using the DNase-seq and FAIRE assays. The ENCODE project focused only on humans, and ultimately resulted in DNase-seq data from over 100 human cell-types, but early analyses were restricted to fewer cell-types. I contributed to such an analysis of 7 diverse cell-types, presented here (Song et al., 2011).

In this initial study, we established the utility of cross-cell-type comparison of regulatory
elements. We discovered groups of regulatory elements with coordinated activity—open chromatin sites coordinately present or absent across one or more cell-types. By chance, we would not expect several nearby regulatory elements to have similar cross-cell-type patterns, so we reasoned that these clusters contained biological information. This finding was consistent with previous reports of clustered FAIRE sites called clusters of open regulatory elements (COREs) (Gaulton et al., 2010). A CORE is a sequence of regulatory elements with two properties: 1) they are located near one another along the linear genome; and 2) they are correlated across cell-types in regulatory signal intensity. For example, a CORE could be formed by a group of 15 regulatory elements within 50 kb, all present in the same 2 cell-types. The initial COREs were defined using only a single cell-type (pancreatic islets), so by considering 7 other cell-types and adding in DNase-seq data, we were able to identify more subtle cross-cell-type COREs.

We hypothesized that COREs could represent individual gene modules. Because we had matched gene expression data for the same cell types, we were able to explore this hypothesis by comparing the cross-cell-type regulatory element activity with the cross-cell-type gene expression. Therefore, I explored the relationship of COREs to gene expression and other high-throughput data signals for the same cell-types. I developed software to identify clusters of regulatory elements with similar patterns and then link these clusters to the expression of nearby genes.

3.2 Results

3.2.1 Clusters of regulatory elements drive cell-type specific expression

To detect COREs systematically in our data set, we calculated pairwise correlations of open chromatin signals using data from all seven cell-types (see Methods) (Fig. 3.1A,B). Using these correlations, we designed a Hidden-Markov Model (HMM) to define 181 high-confidence COREs (see Methods) (Tab. 3.1). COREs varied in size from 32 kb to 6.6
Table 3.1: Cell lines with significantly increased total open chromatin across each CORE were identified by T tests. For each CORE, it was determined if the cell type with the top expressed gene (including those within 10 kb of upstream or downstream of the CORE), the highest Pol II signal, and the highest CTCF signal matched the cell line(s) with significant open chromatin. In 8 instances, there were no cell types with significantly more open chromatin. For these eight COREs, we considered the single cell types with the highest median chromatin level and compiled them in the appropriate 1 cell type column. The table thus shows what correspondence with open chromatin was present for all181 COREs. Some COREs did not contain any annotated genes for which expression data was available.

<table>
<thead>
<tr>
<th>Open Chromatin (OC)</th>
<th>Contains Annotated Gene</th>
<th>Contains No Annotated Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 cell type</td>
<td>&gt;1 cell type</td>
</tr>
<tr>
<td>OC matches expression and Pol II</td>
<td>21</td>
<td>31</td>
</tr>
<tr>
<td>OC matches expression only</td>
<td>9</td>
<td>14</td>
</tr>
<tr>
<td>OC matches Pol II only</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>Neither expression nor Pol II</td>
<td>19</td>
<td>3</td>
</tr>
<tr>
<td>OC matches CTCF</td>
<td>40</td>
<td>45</td>
</tr>
<tr>
<td>OC matches none</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>60</td>
<td>54</td>
</tr>
</tbody>
</table>

Mb.

We hypothesized that COREs represent coordinated nucleosome depletion events caused by multiple regulatory elements participating in the regulation of a nearby gene or genes. We determined in which cell line(s) each CORE was active using the Wilcoxon rank sum test (see Methods) (Fig. 3.1C). Ninety-five COREs (52%) had increased open chromatin within primarily one cell-type, while 78 COREs (43%) were characterized by increased open chromatin within at least two cell-types (Tab. 3.1). In the remaining eight COREs (5%), no cell-type had significantly more open chromatin signal relative to the others, as defined by our threshold. We examined genes inside or within 10 kb of each CORE and found that, for 75 of the 114 COREs with at least one gene with expression data, the cell-type with the highest expression also contained significantly more open chromatin (Tab. 3.1). Among the 67 COREs not associated with any genes for which we had expression data, the aggregate Pol II signal was greatest in a cell-type with significantly enriched open chromatin (Tab. 3.1). Considering all COREs, the highest cumulative CTCF signal was in a cell-type with enriched open chromatin 77% (140/181) of the time. These relations with expression and Pol II and CTCF binding also hold when COREs are analyzed in
FIGURE 3.1: Open chromatin patterns form clusters of open regulatory elements (COREs). (A) Pairwise correlations between 500 open chromatin sites from chromosome 2 show three blocks of correlated sites (see Methods). Each row and column represents an open chromatin region found by both DNase-seq and FAIRE-seq in at least one of the seven cell-types. Red indicates high correlation, white indicates no correlation, and blue indicates negative correlation. Vertical and horizontal lines show CORE boundaries. (B) DNase-seq (y-axis fixed at 0.1) and FAIRE-seq (y-axis fixed at 0.04) signals for a 90-kb subsection of CORE 98 containing the GYPC gene. GYPC is the only gene in this CORE. Highlighted are open chromatin sites found in all cell-types, only GM12878 and K562 together, and GM12878 and K562 individually. (C) Boxplots show the distributions of open chromatin levels within open chromatin sites with CORE 98. GM12878 and K562 both have significantly higher levels of open chromatin (*, Wilcoxon rank sum test). (D) Relative expression levels (y-axis) of GYPC show increased expression in GM12878 and K562 cell lines. (E) Open chromatin sites within CORE 98 also show higher normalized Pol II ChIP-seq read counts in GM12878 and K562 cell-types. (F) Normalized CTCF ChIP-seq read counts do not show significant differences between GM12878 and K562 and other cell-types CORE98. (G) Pol II and CTCF signals in this 90-kb region (shown in B) provide preliminary annotations of similar and differential open chromatin sites.

aggregate(Fig. 3.2). In 18 of 175 COREs, open chromatin levels did not correspond to gene expression, CTCF binding, or Pol II binding. For these cases, we may not have correctly identified a distant target gene associated with the CORE, or the open chromatin regions may not be acting as enhancers.
To demonstrate the utility of identifying COREs, we show one typical CORE in detail (Fig. 3.1A). This CORE extends over 1.2 Mb, but GYPC is the only annotated gene found in this region, so we focused on the 90 kb surrounding the gene (Fig. 3.1B). GYPC encodes both glycophorin-C and glycophorin-D, which function in membrane stability of human erythrocytes and lymphocytes (Walker and Reid, 2010). Open chromatin is detected at the GYPC TSS in nearly all cell-types. However, nucleosome-depletion events unique to GM12878 and K562 occur 10 kb upstream of the TSS and within the transcribed region of GYPC (Fig. 3.1B). These are accompanied by increased FAIRE and DNaseI signal (Fig. 3.1C), expression of GYPC (Fig. 3.1D), and higher Pol II signal (Fig. 3.1E–G). Thus, this CORE is typical, showing coordinated nucleosome depletion, gene expression, and transcription factor binding, and identifies several key regions that may be responsible for the cell-type specific regulation of this gene. Many other COREs reveal putative regulatory elements surrounding genes with cell-type specific functions. For example, CORE 40 surrounds a single gene with variable expression that is ES-specific, *OTX2*.

**Figure 3.2**: At a summary level, Expression, PolII, and CTCF signal intensity all correspond with the particular open cell-types in each CORE.
**Figure 3.3:** An ES-specific CORE. This example shows the correlation between Chromatin structure, RNA level, PolII Read Count, and CTCF Read Count.

**Figure 3.4:** An NHEK-specific CORE. This example shows the correlation between Chromatin structure, RNA level, PolII Read Count, and CTCF Read Count.
(Fig. 3.3). In another CORE, CORE 70 defines a 600-kb region on chromosome 1 that contains several genes with specific functions in keratinocytes, including late cornified envelope (LCE) genes and small proline rich proteins (SPRR). Within CORE 70, NHEK cells exhibit significantly more nucleosome depletion, higher expression of these genes, greatly increased amounts of Pol II signal, and greater CTCF signal (Fig. 3.4). In other cases, unannotated genes within COREs can be associated with a specific cell-type. For example, CORE 60 is a 325-kb region encompassing the sparsely annotated RNF152 (ring finger protein 152) gene. Increased nucleosome depletion, expression of this gene, and Pol II signal are observed specifically in HUVEC and NHEK cells.

3.2.2 Expansion to 40 human cell-types

After additional DNase-seq data had been collected for 40 human cell-types, we also identified COREs in this data. Using an independent method, we confirmed that when expanding to 40 cell-types, COREs can still be identified (Fig. 3.5). By adding in additional cell-types, we should be able to identify more subtle COREs.
3.3 Discussion

One of the major challenges in biology is mapping regulatory elements to the genes they regulate. Commonly, we assume the nearest gene is the most likely target, but this oversimplifies the problem. In this study, we showed that clusters of open regulatory elements defined by multiple sites spanning tens of thousands of bases show good correspondence between open chromatin and levels of gene expression, Pol II signal, and/or CTCF signal. Analysis suggests that COREs encompass noncoding DNA elements that act coordinately to regulate genes important for cell type identity and function. In many cases, the identification of COREs can guide candidate target gene selection. Methods like 3C, 4C, 5C, Hi-C, and ChIA-PET (van Steensel and Dekker, 2010) will continue to be important in solving this difficult problem, but using correlation between regulatory element activity and gene expression is a promising alternative (see Chapter 4).

3.4 Methods

Open chromatin sites found by both DNase-seq and FAIRE-seq in at least one cell-type were considered. Each site was represented by a vector of seven combined open chromatin values, one for each cell-type, and quantile-normalized across cell-types. Pairwise correlations were calculated between each site and 500 surrounding sites. We defined COREs with a two-state HMM, using an average correlation across 5 adjacent open regions as the observable. We tuned transition probabilities and emission probabilities by hand and inspected output to ensure reasonable results. This HMM yielded just 181 high-confidence blocks of varying sizes across the entire genome. With relaxed parameters, the HMM can be tuned to identify over 1,000 lower-confidence COREs.

We used the UCSC knownGene table to find genes within and outside blocks. Genes (UCSC knownGene annotation) overlapping and within 10 kb were assigned to each CORE for expression comparisons. In each CORE, cell-types with significantly more open
chromatin signal were determined using pairwise Mann-Whitney tests \( P < 0.05 \) vs. at least four cell-types). For the randomized distribution of matches, we first selected all genes in the top 20% of expression variance. We discarded any blocks not containing any of these genes. For the remaining blocks, we randomly selected (from the top 20% of genes) the same number of genes the block actually contains, and counted how many of these genes match the block chromatin structure. We summed up the number of such matches across all blocks, and repeated this analysis 1000 times for a distribution of matches. CTCF and Pol II ChIP-seq signal were calculated as normalized sequence read counts mapped inside COREs. To eliminate background signals, only reads located within the top 20K Pol II or top 10K CTCF peaks were considered.

For the expanded CORE analysis based on 40 cell-types, we took an alternative approach. The regulatory elements are still represented by an activity vector (in this case, a 40-vector). Next, rather than relying on correlations among regulatory elements, we first clustered regulatory elements by cross-cell-type pattern (using a self-organizing map, see Chapter 4). We then visualized the sequence of cluster assignments to see if there would be regulatory elements that fit the definition of a CORE.
4

Regulatory Elements in Humans Across Tissues

This research chapter is based on a research article published by Nathan C. Sheffield, et. al, in the journal Genome Research in 2013 (Sheffield et al., 2013)

4.1 Introduction

To expand on the results from Chapter 3, we increased the number of cell-types from 7 to 112, and combined DNase-seq data with expression data to explore how the patterns of regulatory signal differ across human cell-types.

Transcriptional regulation involves a complex interplay of transcription factors (TFs) binding to DNA regulatory elements to control gene expression. This interplay enables a single genome to give rise to hundreds of cell-types. Understanding transcriptional regulation requires a full accounting of regulatory elements, including: 1) their genomic locations; 2) their cell-type specificity; 3) the identity of factors that bind them; and 4) the genes they target. Ultimately, this accounting will enable us to determine how regulatory elements affect tissue-specific gene expression. In this study, we begin to address these issues by integrating chromatin accessibility and expression data from many human cell-
Here, we used an extensive collection of DNaseI-hypersensitive sites to provide new insights into tissue-specific regulatory programs. We clustered over 2 million DHSs from 112 diverse biological samples by tissue-specificity into 1,856 chromatin profiles and found each cluster to have a distinct bias relative to location, evolutionary conservation, CpG islands, and promoter proximity (distal vs. proximal).

Gene expression profiling has emerged as a powerful tool to classify tumors (Wu et al., 2010). The added resolution of regulatory information may provide a more robust way to classify cell-types. To test this, we assigned the 112 samples into tissue groups and developed classifiers to assign tissue-type based on DNaseI hypersensitivity patterns across the cell-type groups. Our models predicted tissue-type with >80% accuracy in leave-one-out experiments. We used this framework to investigate lineage of cancer cell-types with a predictor developed using only 43 individual DHSs. A similar model trained to predict the sex of each sample uncovered a set of sex-specific DHSs surrounding 3 loci on the X chromosome, one of which includes the XIST locus. These results contribute to our understanding of cancer biology and sex determination, and highlight the utility of leveraging DNase-seq data across many cell-types. DNase-seq assays typically identify >100,000 active regulatory elements in a single experiment, but unlike ChIP experiments, they do not directly reveal which TFs bind to these elements. Many TFs bind to a specific pattern of DNA bases at TF binding sites (TFBSs), often represented as a motif, which can be learned by detecting over-represented sequences in regulatory elements. As DNase-seq data from multiple cell-types can predict TF binding (Song et al., 2011), the newly available data enables a thorough analysis of many cell-types. After clustering DHSs, we used de novo motif discovery to identify relevant known and novel TF motifs and thus predict active TFs that bind to each regulatory element.

Even after identifying TF binding, a key remaining problem is to associate elements with target genes they regulate (Heintzman and Ren, 2009; Stadhouders et al., 2011). These
associations can be determined empirically by using chromatin conformation capture (3C) and derivatives to detect long-range chromatin loops (reviewed in Wei and Zhao (2011)). Unfortunately, three-dimensional chromatin information often is locus and cell-type specific, and lacks resolution at the level of individual regulatory elements. In practice, typical analyses link elements to genes using heuristics, most commonly by simply assigning them to the nearest gene. Though this is reasonable, it is not always accurate (Noonan and McCallion, 2010). Recent studies have pioneered new mapping methods using correlations between expression and other genomic features to link regulators to genes at greater distances and across gene boundaries (Akalin et al., 2009; Ernst et al., 2011). However, linking gene expression to DNaseI signal has not yet been explored. We used correlation between DNaseI and matched expression data to identify possible target genes for many regulatory elements.

The DNaseI and expression data used in this study are accessible within the UCSC Genome Browser (Rosenbloom et al., 2010). However, the linear nature of genome browsers is not ideal for viewing results of the type we present here, which include clustering, motifs, and networks. For that reason, we created a database and web interface to better visualize our analytical results (http://dnase.genome.duke.edu). Through this resource, users can view DHS chromatin accessibility profiles, locate similar sites, and view enriched motifs and predicted target genes. Resources of this type will enable biologists to synthesize meaningful conclusions from integrated experimental results. These results and resources bring us closer to the goal of explaining how chromatin structure relates to transcriptional regulation across diverse human cell-types.
Figure 4.1: Data before and after correcting for batch effects. (A) Hierarchical clustering, (B) Multidimensional scaling, and (C) First principal component. Corresponding results for the microarray data are shown in panel D, E, and F. Uncorrected data group by batch, and the first principal component captures the effect. After correction, the data group as biologically expected.

4.2 Results

4.2.1 DNaseI hypersensitive sites cluster cell-types by biological similarity

Genomic locations of 2.7 million DNaseI hypersensitive sites (DHSs) from 125 samples were described previously (Thurman et al., 2012). From this data, we selected a subset of 112 samples for which we had both DNase-seq and expression data. The 112 samples represent 72 unique cell-types and 15 unique tissue lineages (Appendix A). Data were generated in one of two labs, each using a distinct DNaseI protocol. We improved on the previously published open chromatin measurements by accounting for batch affects
Figure 4.2: (A) A 50x50 self-organizing map (SOM). Each box represents a cluster of DHSs with similar DNase-seq signal profiles across samples, color-coded by tissue (legend, left). Cluster color corresponds to the combination of cell-types in which the associated DHSs have high signal in the detailed profile. Square size indicates the number of DHSs assigned. (B) Average DHS profiles across samples for four clusters. Clusters are open in related cell-types (54 and 25) or less related cell-types (1091 and 1295). *malignant samples.
that grouped the data by lab rather than by biological signal (see Methods). We used ComBat (Johnson et al., 2007) to remove these batch effects, after which both the DNaseI and expression data clustered according to expected biological relationships (Fig. 4.1).

Using these data we first investigated DNase-seq signals for common patterns across cell-types. Previously, we briefly described an initial self-organizing map (SOM) that clustered DHSs by their profile of hypersensitivity across cell-types (Wehrens and Buydens, 2007; Thurman et al., 2012). In this study, we improved this clustering by increasing the resolution, introducing a step to merge highly similar clusters, and using the batch-corrected data to redefine SOM clusters; we defined 1,856 clusters of DHSs (see Methods). This enabled us to identify subtle patterns in the data more robustly and to group similarly acting sites more accurately.

Each DHS was assigned to the single cluster in the SOM that most closely matched its hypersensitivity profile across cell-types (Fig. 4.2A). An overall cluster profile (or average DNaseI signal in each cell-type) was defined by calculating the average hypersensitivity across the DHSs it contained (Fig. 4.2B). Throughout this manuscript, we refer to clusters using the cell-types with increased signal in this averaged DNaseI signal profile. We found that multi-cell-type clusters (those whose DHSs were open in more than one cell-type) generally involved cell-types with known relationships (e.g., Fig. 4.2B). In cases where clusters grouped cell-types without obvious biological similarity, this sharing of DHSs may indicate distant lineage relationships, reuse of regulatory elements, transformation related to cancer progression, or may simply reflect a limit in the resolution of the SOM.

4.2.2 SOM clusters capture variation in CpG-island, promoter, and conservation

We annotated each SOM cluster of regulatory elements with respect to overlap with promoters, CpG-islands, and evolutionarily conserved elements (see Methods). We found clear associations between cluster assignment and all three features, which we have illustrated together in a scatterplot (Fig. 4.3A). For example, clusters in the upper-right of the scat-
FIGURE 4.3: (A) Each cluster is plotted as a bubble. X-axis indicates the percent of the top 100 DHSs in that cluster (ranked by nearness to the cluster center) that overlap a CpG island; y-axis indicates the percent that overlap a promoter; color indicates the percent that overlap a PhastCons Conserved Element (Siepel et al., 2005). The size of the bubble indicates the number of DHSs belonging to the cluster. Red bubbles in the upper-right indicate clusters capturing primarily highly-conserved, CpG-rich promoter elements. (B) DNaseI signal profiles of five example clusters, showing the distribution of distance to the transcription start site (TSS) of the nearest gene. Cluster 99 is promoter-rich; cluster 1259 is preferentially located in an early intron; cluster 199 is highly conserved, but not associated with promoters or CpG islands; cluster 881 is primarily distal, with no regions within 500bp of a TSS. (See also Fig. 4.42)

terplot (Fig. 4.3A) are enriched for promoters, CpG-islands, and conserved elements, and have stronger DNaseI signal across many cell-types (e.g., cluster 99; Fig. 4.3B; Fig. 4.4C). Among clusters with similar promoter overlap, the distribution of the distance from DHSs to transcription start site (TSS) varies. For instance, clusters 1361 and 1259 both have 20-30% promoter overlap, but sites from cluster 1259 are more commonly found just downstream of the TSS (near 5 introns), whereas sites from cluster 1361 are further from the TSS (Fig. 4.3B). This finding suggests that DHSs with similar patterns across cell-types are likely to share relationships with sequence conservation and genomic location. A striking outlier is the non-promoter, non-CpG cluster 199, which has an uncharacteristically high conservation score; this cluster, along with other similar clusters, contains ubiquitous distal
Figure 4.4: (A) Heatmap illustrating hypersensitivity patterns across cell-types. Columns represent clusters; rows represent 112 cell-type samples. Tissue is identified by the colored legend on the right; similar tissues cluster together. Heatmap color corresponds to hypersensitivity. (B) Cell-type specificity of individual DHSs (maroon) and clusters (blue). Most clusters (and individual DHSs) represent DHSs present in a few cell-types, with a rapidly decreasing distribution as number of open cell-types increases. There is an increase in clusters open across all cell-types. (C) Scatterplot as in Fig. 4.3, except the color indicates the number open tissues. (D) Stripchart, each point corresponds to a cluster with <90% overlap with PhastCons Vertebrate conserved elements. Y-axis: clusters are divided into two groups, those with an enriched CTCF motif and those without. X-axis: the proportion of the cluster that overlaps a promoter. All highly enriched clusters with low promoter overlap are CTCF-enriched. There is a clear bimodal distribution across promoter overlap. (E) Beanplots showing the distribution of number of open tissues (x-axis) with respect to their % overlap with Vertebrate, Placental, and Primate PhastCons conserved elements, and Promoters (2Kb upstream of annotated TSS). (F) Distribution of normalized DNase-seq signal intensities.
DHSs that are highly enriched for CTCF motifs (Fig. 4.3 and Fig. 4.4D).

4.2.3 A logistic classifier predicts cell-type lineage with few DHS inputs

Since some regulatory elements are highly specific to certain cell-types, we reasoned that a subset of elements could be used as molecular markers for identifying cell-type lineage. To test this, we built a multinomial logistic classifier (Fig. 4.5) that assigns a probability among multiple classes (tissue lineages). Each cell-type was first assigned to one of 15 primary tissue-types based on known biology (Appendix A). We removed all malignant cell-types and restricted the model to the 7 tissue-types containing at least 4 samples each, resulting in a training set of 80 samples across 7 classes. Assuming SOM cluster patterns would be good candidates for differentiating lineages, we used an initial feature set consisting of 1,856 DHSs: one from each cluster that was most similar to the average SOM cluster profile. Trained classifiers assigned the highest probability to the correct tissue lineage with greater than 80% accuracy in leave-one-out cross-validation (See Methods). The final model trained using all samples chose only 43 DHSs as informative features (examples are shown in Fig. 4.5D). These 43 DHSs are thus one minimum representative set of DHSs with high tissue specificity that can be used to predict tissue identity. The classifier trained using all 80 samples only misclassified 2 (2.5%) of the 80 samples used to build it: aortic smooth muscle (AoSMC_SFM), and cardiac myocytes (HCM; Fig. 4.5A). In these two cases, the model assigned about 30% probability to the correct lineage (muscle), but a higher (albeit still weak) probability to the fibroblast class. The inability to distinguish between fibroblast and muscle lineages may reflect the biological similarity between them; it is possible to convert fibroblasts into muscle cells in vitro (Tapscott et al., 1988). In addition, regulatory element differences among the included smooth, cardiac, and skeletal muscle samples complicate the muscle lineage and may not be captured by the 43 DHSs used by the model. Samples from blood and stem cells were never misclassified.

To investigate the remaining data, we used this model to classify the 27 malignant
Figure 4.5: Predictions from a multinomial logistic regression classifier trained to predict tissue identity for a given sample with data from 43 DHSs. (A) Predictions for training data, along with known tissue of origin (left column). Colors within the heatmaps reflect the predicted probability of belonging to each of the 7 tissue classes. (B) Predictions for malignant samples not included in the training, but whose presumed tissue of origin was included in the model. * denotes malignant samples. (C) Predictions for samples whose tissue (or presumed tissue) was excluded from the training, as tissue types had <5 samples. (D) The DNaseI signal profiles of 7 (out of 43) clusters selected by the model with positive coefficients. (E) The DNaseI profile for the single sex-specific site (chrX:130926460-130926610) selected by the classifier. Enlarged barplot shows the distinction between samples divided by sex for the subset of samples included in the model.
samples as well as the 5 primary cell-types left out of the training model (Fig. 4.5B, 4.5C). 14 of the malignant samples are presumed to associate with one of the 7 lineages that were included in the model (Fig. 4.5B). For these, the model prediction agreed with this presumed lineage in 9 out of 14 cases. Among the 5 samples whose classification did not agree, 4 were derived from brain tumors, and 3 of these represented specific brain-cell sub-lineages not present in the training model, which consisted solely of astrocytes. Astrocytes are a subtype of glial cells, which are non-neuronal cells of ectodermal origin. These 3 brain tumor samples were generally not strongly assigned to any lineage (average max. probability 34%; Fig. 4.5B). The fourth misclassified brain cancer was glioblastoma, which the model confidently (86%) classified as epithelial. Glioblastoma, like astrocytes, originate from glial cells, so this misclassification may indicate differences between astrocytes and other glial cell-types, or a substantial remodeling of glial cell chromatin structure that occurs during cancer progression and results in an epithelial-like pattern. In fact, there are reported glioblastoma cases with epithelial differentiation (Rodriguez et al., 2008; Tanaka et al., 2011). This result indicates that this glioblastoma line is more similar to epithelial cell-types than to the astrocytes at the chromatin level. The only malignant sample correctly classified as brain was medulloblastoma, which is an embryonal brain cancer consisting of both neuronal and glial cells (Gilbertson and Ellison, 2008).

The remaining (non-brain) mis-classification was the K562 leukemia cell line, which we expected would associate with the hematopoietic lineage, but instead weakly associated with multiple lineages, none with probability greater than 30%. The lack of a strong assignment to the hematopoietic lineage may be due to its similarity to undifferentiated erythrocytes (red blood cells), while the hematopoietic lines used to build the model are leukocytes (white blood cells). In contrast, the leukocyte cancer cell-types (CLL, CMK, HL-60, Jurkat, and NB4) are all confidently (>75%) assigned to the hematopoietic lineage. This indicates that our blood-specific signatures are not general to all blood cell-types, but of the lymphoid lineage only. Another correctly classified sample was Ntera2, a
teratocarcinoma cell-line often used as a pluripotent stem cell model (Pleasure and Lee, 1993), which was appropriately assigned to the stem cell lineage. We similarly evaluated the lineage associations for the remaining excluded samples (Fig. 4.5C).

We also used SOM-based DHS features to train a predictor to discriminate between male and female samples. We found a single cluster (488) containing sex-specific hypersensitive sites (Fig. 4.5E). A single representative DHS predicted the correct sex in 40 of 43 (93%) non-malignant cell-types with known sex. This cluster (488) consists of 30 DHSs on the X-chromosome that fall primarily into 3 loci, one of which surrounds the XIST gene. The second locus includes a noncoding RNA (LOC286467) recently shown to be the only locus on the X-chromosome, besides XIST, with sex-specific Pol2 binding (Reddy et al., 2012). The third locus also includes a poorly documented noncoding RNA (LOC550643). Both the second and third loci have complex tandem repeat structures, and all three include annotated piRNAs, which are known to have vital sex-specific roles in germline cells (Girard et al., 2006). Interestingly, these two loci were identified in a recent independent study as having intense H3K4me2 signals on the metaphase X chromosome (Horakova et al., 2012). Each locus was also implicated in inactive-X-specific long-range interactions supporting a role in sex specificity. This result indicates that the SOM method can indeed capture differential regulatory element features in other biological divisions across cell-types besides tissue lineage (Fig. 4.5E).

4.2.4 DHS clusters are enriched for known and novel TF motifs

One motivation for clustering DHSs was to find groups of sites with similar activity profiles, which may indicate commonly bound transcription factors (TFs). We therefore analyzed the clusters for enrichment of TF motifs. We used de novo motif discovery to identify enriched motifs and then assigned motifs to specific factors based on the JASPAR (Portales-Casamar et al., 2010) motif database (see Methods). We found 1,279 (69%) clusters had at least one significant motif (e < 1e-6), while 918 (49%) clusters had a motif that could be assigned a
**Figure 4.6:** (A) Representative examples of de novo motif discovery results and highly significant known motif matches. (B-C) De novo motif discovery identified several enriched motifs for which there were no convincing matches to the TF databases. We sometimes found a similar motif across multiple clusters associated with similar cell-types.

<table>
<thead>
<tr>
<th>Cluster Profile</th>
<th>Motif Discovered</th>
<th>Database Match</th>
<th>Factor</th>
<th>e-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>Novel Motifs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>Similar Novel Motifs</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We found highly cell-type-specific clusters enriched for motifs known to be important for those cell-types, but clusters commonly enriched in a specific cell-type did not necessarily share similar motifs, indicating that clusters could discern subtle differences in patterns. Fig. 4.6 provides specific examples of individual clusters and their relevant motif enrichments. For example, the stem-cell-specific cluster 3 was enriched for the known pluripotency factor POU5F1 (Oct-4) motif. The hematopoietic-specific cluster 24 con-
**Figure 4.7:** (A) Extended result of IRF analysis illustrating the motifs enriched in the 2nd 100 regions in each cluster. These motifs match those from the first 100, suggesting the consistent differences in motif are biological. (B) Complete results (for all motifs) for distribution of top tissue-type for JASPAR. In this heatmap, rows are tissues and columns are transcription factors. Colors indicate the proportion of clusters with that top tissue type for that factor (yellow = high, blue = low). (C) Boxplots for each factor tested for ChIP overlap with motifs found in DHS within a cluster; for each factor, the two boxes indicate the % overlap with ChIP peaks for: LEFT, clusters where the motif was enriched, RIGHT, all included clusters where other motifs were enriched. P-value is the Mann-Whitney p-value comparing the two distributions for each factor.
tained the ETV7 (Tel2) motif, consistent with its importance in hematopoietic lineages and leukemia (Potter et al., 2000; Cardone et al., 2005); and an erythroid-specific cluster, 2215, was enriched for GATA family motifs, which are essential for erythroid development (Zhu and Emerson, 2002; Ferreira et al., 2005). Interestingly, the motif for the REST repressor was enriched in a medullo-repressed cluster (cluster 36), indicating the potential to also reveal lineage-specific repressive elements. We also found motifs in ubiquitous clusters, discussed further below.

In 39% of the cases, de novo motifs did not convincingly match known motifs in JASPAR (Fig. 4.6B), representing possible new or poorly-characterized regulators. For example, in a Urothelium-specific cluster (2090), we identified a short motif (consensus TCCAAC) without a good match in the database. Other clusters (e.g., 1694, 607, 1105, 2142) had similarly high de novo p-values without known motif matches. We found a series of clusters (of which 3 are depicted in Fig. 4.6C) that find similar motifs with a CANNTG core sequence and an appendage with ATW consensus 8 bp away. These motifs likely reflect poorly characterized or unknown TFs not yet present in JASPAR, or a complex of TFs.

4.2.5 Motifs in similar clusters reveal subtle motif differences

Interferon Regulatory Factors (IRFs) are DNA-binding proteins that regulate the entire immune response (Paun and Pitha, 2007). The DNA binding domain is highly conserved among the 9 human IRF family members (consensus 5-AANNGAAA-3), but different IRFs bind slight variations of the core sequence (Honda et al., 2006) (Fig. 4.8). IRFs may also bind in complex with SPI1, another hematopoietic factor, forming a longer TFBS (Brass et al. 1999). In our analysis, we detected IRF1/IRF2/SPI1-like motifs predominantly in clusters specific to hematopoietic cell lineages, but among these there was variation in DNaseI signal intensity among LCLs, B cell leukemia (CLL), T cells (CD4, Jurkat, and Th), megakaryocytes (CMK), and erythroleukemia (K562). We noticed slight
**Figure 4.8:** Database motifs for IRFs and SPI1 from JASPAR show both common and distinct features. MEME motifs discovered in several hematopoietic-specific clusters also have distinctions. The clusters vary in cell-type specificity among the hematopoietic cell-types, and the motif logo varies as well, while retaining some semblance of the known SPI1/IRF family motifs.

Variations on the motifs accompanying differences in DNaseI signal across hematopoietic cell-types (Fig. 4.8). This may be due to differences between IRFs and SPI1 binding, different cofactors that modulate an IRF’s binding preference, or distinct IRFs in specific hematopoietic lineages. We reason these motif variations represent biological differences in motif preference rather than statistical noise because in other cases (e.g. in the case of
CTCF), we see less variation among discovered motifs across clusters. We also see similar patterns when looking at an independent set of regions from the same clusters (Fig. 4.8).

4.2.6 Motif discovery results are consistent with experimental ChIP data

We used ChIP data from the ENCODE project to validate our discovered motifs (Fig. 4.9A). Using representative DHSs from each cluster with enriched motifs (see Methods), we compared overlap with ChIP peaks from 43 experiments (Dunham et al., 2012). We expected some incongruence in overlap between motif and ChIP results because ChIP data comes from only a subset of cell-types included in the motif analysis. For example, we compared ChIP results for a single IRF from just 3 cell-types, while our motif analysis considered 14 hematopoietic lineages. Without ChIP data for all cell-types, we expect to find many instances of a positive motif result without a corresponding ChIP signal. Additionally, ChIP reports signal at indirectly bound sites where a motif would not. Despite these limitations, there is good correspondence (Mann-Whitney p-values between 10-5 and 10-133) between motif enrichment and ChIP results. The correspondence is particularly high for CTCF (Fig. 4.9A), which is probably due to its cross-cell-type consistency; DHSs in clusters with CTCF motif enrichment and CTCF sites based on ChIP experiments have 96% median overlap, compared to 4% overlap with other clusters. A similar trend is seen for other factors tested. There is high overlap among the IRFs, SPI1, and RUNX1 ChIP and motif results, consistent with all 3 factors co-regulating hematopoietic lineages (Huang et al., 2008). The SP1-motif clusters overlap not only SP1 ChIP peaks, but also ChIP peaks for most of the other factors, consistent with the role of SP1 as a general, promoter-enriched factor with many interacting partners (Kaczynski et al., 2003).

4.2.7 Global TF trends suggest AP-1 is a chromatin-accessibility factor

We wanted to know whether individual TFs whose motifs are present in several clusters revealed biologically interesting properties about their function (Fig. 4.9B, Fig. 4.8). For
The following text is not within a table or diagram:

**Figure 4.9:** (A) Concordance (yellow=high, blue=low) between ChIP results (x-axis) and motif discovery in DNaseI clusters (y-axis). (B) The cell-type specificity for selected motifs. This heatmap shows the distribution of most-open tissues for each motif. For example, 100% of the clusters where the POU5F1 motif was found had stem cells (Stem) as the most open tissue type, whereas MYF family motifs were found predominantly in muscle clusters. (C-D) Each colored square represents a cluster with enrichment for the given motif. X-axis: overlap with CpG islands; y-axis: overlap with promoters; color: the number of tissues with at least one sample above a cutoff. Each factor shown here has a different distribution of cell-type specificity and promoter/CpG-island overlap. Size of a square indicates the number of DHSs in the cluster. (E) Number of clusters that are enriched for the most common motifs.
each TF, we summarized motif results from all clusters and identified lineage trends. We found that TFs with roles in certain cell-types were most often enriched in clusters with a small number of relevant tissue lineages. For example, the myogenic factor (MYF) family motif was enriched primarily in muscle-specific clusters, HNF4 in liver clusters, POU5F1 in stem cell clusters, and SPI1 in hematopoietic clusters (Fig. 4.9B); these are all biologically relevant enrichments (Scott et al., 1994; Nichols et al., 1998; Odom et al., 2004; Cao et al., 2006). This starkly contrasted with ubiquitously expressed transcription factors SP1, AP-1, and CTCF, which did not have a bias toward a single lineage (Fig. 4.9B). We examined the CpG-content, genomic location, and tissue specificity of clusters where each TF motif was enriched to characterize the regulatory elements that bind each factor. For example, SP1 was enriched in clusters with CpG-island promoters that are present in many cell-types (Fig. 4.9C); this may partly reflect the GC-rich SP1 motif. CTCF was enriched in clusters representing distal DHSs present in many cell-types, which is consistent with previous reports (Xi et al., 2007; Lee et al., 2012, Fig. 4.9D). In fact, we found the CTCF motif was enriched in all 12 non-promoter, highly conserved clusters (Fig. 4.4D). The absence of another motif with this property reinforces the uniqueness of function of the CTCF protein. SPI1, MYF family, and IRF family motifs were preferentially enriched in cell-type-specific distal clusters (Fig. 4.9D). Plots similar to those in Fig. 4.9C were generated for each TF in JASPAR.

The most commonly enriched motif discovered was that of Activating Protein 1 (AP-1), found in about 12% (220) of the clusters. By comparison, the second most common motif, for SP1, was found in about 8% of clusters (152 clusters; Fig. 4.9E). AP-1 is the well-studied FOS:JUN heterodimer that activates both basal and inducible expression (Angel and Karin, 1991). It has been implicated in a variety of cellular functions, including cell proliferation, immunity, apoptosis, and differentiation (Angel and Hess, 2012). We found the AP-1 motif enriched exclusively in non-promoter, non-CpG-island clusters (Fig. 4.9D). In contrast to the tissue-specific factors like MYF family members and SPI1, AP-1 was
found in both tissue-specific clusters as well as those shared among many cell-types. As detailed in the Discussion, these results suggest that AP-1 may play a general role in chromatin accessibility in many different tissues and genomic locations.

### 4.2.8 Chromatin and expression correlation predicts long-range interaction

The DNase-seq experiment naturally leads to the question of identifying target genes for DHSs. Song et al. (2011) used cross-cell-type correlation among DHSs to identify blocks of similar regulatory elements and co-expressed genes. Thurman et al. (2012) approached this by correlating distal DHSs with promoter DHSs. We reasoned that if the pattern of a DNase-seq signal across cell-types matched the pattern of expression of a gene across cell-types, this provided evidence that the gene is a regulatory target of the DHS. Therefore, we correlated DNaseI hypersensitivity with gene expression data to infer the target genes
Figure 4.11: (A) The distribution of the number of connections to genes for each DHS. (B) The distribution of the number of connections to DHSs for each gene. (C) Q-Q plot showing that protein-coding genes tend to have more associations than RNA genes. (D) Tieplot showing connections at the PHC1 locus, as in Fig. 4.10. (E) Enriched reactome pathways, and posterior probability of activation, based on two sets of genes associated with each cluster: First, genes based on assigning DHSs in the cluster to nearest gene, and second, adding genes to the first set based on DHS signal/gene expression correlation analysis.
(both protein-coding and RNA) for each of the 2.7 million DHSs (see Methods). About 530,000 of the 2.7 million sites (20%) correlated significantly with at least one gene within 100kb (permutation p-value < .05), a significant enrichment over the 5% expected by chance. Of these, 71% correlate with a single gene, but some correlate with as many as 44 genes (Fig. 4.11A). 31,000 Ensembl genes (98%) correlated with at least 1 DHS, and the median number of DHSs associated to a gene was 19 (Fig. 4.11B). Protein-coding genes tended to have more associations than RNA genes (Fig. 4.11C). Fig. 4.10A-B illustrates representative examples showing correlations of DHSs to genes that are color coded to indicate the tissue-types that are driving the correlation with gene expression (see Methods). These examples show that associated DHSs can be very far away, crossing multiple gene boundaries.

Long range regulatory interactions have been previously reported based on chromosome conformation capture (3C) experiments (Tolhuis et al., 2002). 3C data is not a perfect comparison for several reasons (Methods). Despite these limitations, we compared our results to 3C data and found the 3C and correlation results corroborate one another in 8 of 12 cases we investigated (described below).

Beta-globin locus (mouse, humanized mouse): Good accord. The beta-globin locus control region (LCR) is a collection of five well-characterized DHSs upstream of the beta-globin genes (Molete et al. 2002). The LCR is located near the epsilon-globin gene and has been shown to regulate the other globin genes (HBB/HBD) about 30-50kb away in erythrocyte but not in brain cell lineages [Tolhuis2002]. The globin genes are expressed in erythroid cells at different times in development; for example, HBE1, is embryonic, HBG1 and HBG2 are fetal, while HBD and HBB are adult forms (Molete et al. 2002). Our study is limited in detecting connections by the cell-types we characterized, and the primary cell-type driving connections at this locus is K562 (representing undifferentiated erythrocytes), which is known to express the embryonic globin gene HBE1 (Jackson, 2003). Previous
3C experiments showed erythroid-specific proximity between beta-globin genes and the LCR located about 40kb from the HBE1 gene (Tolhuis et al., 2002; Palstra et al., 2003). Our results reproduced these findings with a K562-driven link between HBE1 and several downstream hypersensitive sites (Fig. 4.10A). Most notable, a highly significant correlation linked the beta-globin gene to one of the DHSs in the LCR, hypersensitive site 4 (HS4). This provides a specific association between a particular gene and a particular DHS within the LCR. In addition, there were several other hematopoietic-driven (cyan-colored) links throughout the region (Fig. 4.10A). (Tolhuis et al., 2002; Splinter et al., 2006; Fang et al., 2009; Palstra et al., 2003).

H19/IGF2 ICR (mouse): Good accord. Another well-studied example is the H19/IGF2 locus, which has been shown to have an imprinted long-range interaction (Leighton et al., 1995). In this original study, a 6.2 kb deletion affected expression differently when inherited maternally vs. paternally, but this study did not identify individual DHSs that may be involved. An Imprinted Control Region (ICR) located between H19 and IGF2 binds CTCF on the maternal but not paternal allele. When CTCF binds, an enhancer located on the other side of H19 is unable to interact with the IGF2 gene and instead only enhances H19 expression. On the paternal allele, the ICR is methylated, which blocks CTCF binding and allows the enhancer to bind the IGF2 promoter and increase IGF2 expression. While we did not detect interactions with the ICR, we did detect strong correlations between IGF2 gene and several DHSs located in the H19 enhancer region (Fig. 4.10B). The correlations were driven primarily by liver lineages, consistent with the role for IGF2 in liver cells. This interaction was detected without any knowledge about allele specificity. Enhancers upstream of H19 have been shown to interact with IGF2 in an parent-specific manner. We find a correlation between IGF2 and several DHSs in the enhancer region, driven by expression and hypersensitivity in liver or brain cells (Leighton et al., 1995; Kurukuti et al., 2006; Murrell et al., 2004).
**Alpha-globin locus (humanized mouse): Good accord.** Result is similar to the beta-globin locus. The 3C experiments show an LCR inside C15orf35 (a.k.a. NPRL3), which connect to HBA1 and HBA2 (adult globins). Our samples, including K562, do not have expression from the arrays on these genes; however, the adjacent embryonic globin gene HBZ is expressed highly in K562 cells; from this gene we see significant correlations to the DHSs in the LCR. This is a series of K562-specific DHSs in NPRL3 introns Vernimmen et al. (2009).

**OCA2/rs12913832 in HERC2 locus (human): Perfect accord.** SNP rs12913832 (chr15: 28365618), in a HERC2 intron, is relevant for loop formation and gene expression of OCA2. OCA2 is involved in skin and eye pigmentation. Our data links OCA2 expression to a DHS located essentially at this SNP (only 143 bp away: chr15:28365325-28365475). This link is driven by expression and hypersensitivity in melanocytes, where the loop occurs, as well as a few other related skin or neuronal cell-types (Visser et al., 2012).

**Th2 locus (mouse): Poor accord.** (IL4 connecting to HS sites in the 3’ end of RAD50) In mouse, these studies report a 3C connection in T cells but not B cells or fibroblasts. The Th2 LCR (inside the 3’ end of RAD50) connects with the IL genes (IL13, IL4, and IL5). We see no corresponding correlation in the region because IL13, IL4, and IL5 are only lowly or uniformly expressed in all included samples; this includes all T-cells, hence we see no connection to the RAD50 intron LCR (Spilianakis et al., 2005; Spilianakis and Flavell, 2004).

**SHH locus (mouse): Good accord.** Sagai et al. (2005) identified 3 elements near the Sonic Hedgehog (SHH) gene that are conserved from mammals to fish, and called them mammal-fish conserved sequences 1-3 (MFCS1-3). They first showed that mutations in MFCS1 cause abnormal limb development in mice; Amano et al. (2009) later showed with both
3C and FISH that this enhancer loops around to regulate SHH expression specifically in limb-bud tissues in development. In a separate study, Sagai et al. (2009) identified 3 additional nearby conserved elements, and then tested these 3 as well as MFCS2-3 in a developmental assay. They found that 4 of these conserved elements drive SHH expression in epithelial linings during mouse development. They did not find a tissue-specific affect for MFCS2. (Amano et al., 2009; Sagai et al., 2005, 2009).

We do not have corresponding embryonic epithelial or limb-bud representatives in this dataset; however, it has been shown that the hedgehog pathway is upregulated in breast cancers (Kubo et al., 2004; Koga et al., 2008). Consistent with this, we find the SHH gene to be most highly expressed in MCF-7 cells. Sites correlated to the SHH gene are mostly driven by MCF-7-specific DHSs, of which there are several throughout the SHH-LMBR1 locus. We do not find any correlated MCF7-specific DHSs near MFCS1, or the other 4 elements found to drive epithelial expression; however, there is a series of 3 MCF7-specific DHSs near MFCS2 that do correlate with SHH expression. Of these 3 DHSs, 2 are essentially within MFCS2, 1 is about 5kb away. This result may indicate that this conserved element, the only of the 6 with still unknown developmental tissue specificity, may drive expression in mammary gland morphogenesis, where SHH expression is known to be important (Kubo et al. 2004).

POU5F1 locus (mouse): Good accord. This study reported an ES-specific connection between Oct4 and an enhancer located about 2kb upstream. In human, we detect significant correlation between Oct4 expression and several upstream enhancers. These may correspond to the same event seen in mouse. Despite some question as to the accuracy of the arrays we used in measuring Oct4 expression, the profile appears at least slightly elevated in ES cells, driving the correlation to several highly specific nearby ES cell DHSs, including one about 3kb from the end of Oct4 (Kagey et al., 2010).
**Nanog locus (mouse): Poor accord.** The Nanog expression data from the array is of limited reliability due to cross-hybridization. (Kagey et al., 2010).

**Phc1 locus (mouse): Good accord.** PHC1 is an ES-specifically expressed in our data, as expected. Although we did not identify a single DHS that connects at the same distance in the mouse data, we do see several significant DHSs ranging from a few kb to more than 80kp upstream. These are all potential regulators of PHC1. This may reflect rearrangements between mouse and human. Kagey et al. (2010) reported 3C results for areas surrounding 4 pluripotency regulators, including PHC1, in mouse ES cells. We detected correlations with individual hypersensitive sites several kb upstream of PHC1 that were consistent with the 3C results. In each case, these associations were driven by pluripotency-specific DHSs and expression. In addition to a correlation with the nearest upstream enhancer, we also detected several other long range (75-100kb) pluripotency-specific sites that correlate with PHC1 expression. In contrast, a gene flanking PHC1 were connected to DHSs that are driven primarily through correlations in endothelial cells. We were unable to explore the connections surrounding the other genes due to issues with microarray measurements (See Methods) (Kagey et al., 2010).

**Lefty locus (mouse): Poor accord.** Our expression data for Lefty has no variation across cell-types (Kagey et al., 2010).

**IFNG locus (human, mouse): Good accord.** In this study Hadjur et al. (2009) used 3C to show the IFNG locus interacts with regulatory elements located both upstream and downstream the IFNG gene. These are specific to Th1 cells. In our data, we included Th1 cells and found similar results, correlating Th1-specific IFNG gene expression with Th1-specific DHSs throughout the region.
**LCN2 locus, LCR in Ciz1 (mouse): Poor accord.** This study demonstrates a loop between an glucocorticoid-response element in the LCN2 promoter and the nearby Ciz1 gene, which occurs in hepatocytes but not pituitary cells. Hakim et al. (2009) reported that horomone induction activates Ciz1 expression and suggested that the loop structure exists prior to hormone induction to facilitate rapid transcriptional response. Consistent with this hypothesis, we do not see a correlation between DHSs near LCN2 and expression of Ciz1 because the expression of Ciz1 is uniform. We would not expect to see a correlation with expression unless we included hormone induced samples of the same variety, which we did not (Hakim et al., 2009).

**4.2.9 A web resource for exploring results**

The results presented here begin to provide more detailed and informative annotations for 2.7 million DHSs contributing to gene regulation in 112 samples across 72 diverse cell-types. To facilitate the further exploration of these data by the research community, we have created a web resource (http://dnase.genome.duke.edu/) to query, display, and extract data. The resource allows queries by regulatory element, by gene, by genome coordinates, by transcription factor, or by cell-type specificity. For researchers starting from a single regulatory element, the web interface provides a list of other regulatory elements with similar cell-type-profiles via the SOM clustering. For each SOM cluster, the user can view enriched motifs, genomic distribution, CpG and conserved element overlap, and associated genes and pathways. For any gene of interest, users may view expression, download sets of connected regulatory elements, and explore the clusters to which these connected elements belong. The web resource also enables data to be downloaded in text format for input into genome browsers or external computational pipelines.
4.2.10 Clusters are enriched for biologically relevant pathways

Functional pathways require the coordinated regulation of multiple genes, and their activity may vary depending on cell-type under the direction of cell-type-specific regulatory elements. We hypothesized that the similar, cell-type specific regulation in some of our SOM clusters corresponds to the activity of genes in the same pathway. We associated genes with SOM clusters and determined whether there was enrichment in clusters for genes in common pathways. Initially, for each SOM cluster, we mapped each DHS to its nearest protein-coding gene. 585 clusters were enriched for at least 1 pathway (Fig. 4.11). For example, immune-related genes are commonly enriched in lymphoblast-specific clusters. In several myoblast/myotube-specific SOM clusters, the DHSs are near genes that relate to striated muscle contraction. To determine how our DHS-gene mappings affected this analysis, we assigned to each DHS all of its connected genes, or if no connections had been found, to its nearest gene. We then repeated the pathway enrichment analysis using SOM cluster gene sets based on these DHS gene assignments. Adding genes to SOM clusters based on our DHS to gene mappings had a minor effect, but the lack of a correct answer made it impossible to test whether this offered an improvement.

4.3 Discussion

Our global clustering of DHSs revealed novel open chromatin pattern relationships among a diverse set of human cell-types. Many clusters grouped cell-types of common lineage, enabling accurate lineage classifications based on only a few DHSs. We also identified several biologically relevant pathway enrichments for genes near particular clusters. In future work, we could further delineate among clusters by adding ChIP data for TFs or histone marks, DNA methylation, or DNaseI footprinting (Hesselberth et al., 2009; Boyle et al., 2011; Pique-Regi et al., 2011). Creating clusters from a larger set of cell-types and developmental stages along with epigenetic data could be a powerful way to characterize
cell-type lineage.

The primary experimental data for cell-type-specific TF binding comes from ChIP of TFs known in advance (Dunham et al., 2012). We showed that characterizing hypersensitivity across cell-types also yields convincing de novo motif discovery results, including identifying novel regulators and new roles for known regulators. This approach provides an unbiased (no a priori knowledge/antibodies required) complement to ChIP, and motifs we discovered in over 1,000 clusters provide a rich resource for further investigation. Our results invite follow-up study into the function of AP-1, and motifs not yet found in the databases. This resource will also be useful for motif scanning to narrow results based on DHS profiles.

In our motif analysis, the AP-1 motif was the most commonly detected, both in ubiquitous and cell-type specific clusters of DHSs. Since its subunits (FOS and JUN) are ubiquitously expressed, the cell-type specificity is probably conferred by other factors. This is consistent with a role for AP-1 as a pioneer factor that opens DNA for other factors, or it may be an otherwise general and universal chromatin-accessibility factor. This hypothesis is consistent with experimental results confirming a role for AP-1 in diverse pathways (Angel and Hess 2011). Recent evidence also corroborates the general role of AP-1 in forming accessible chromatin; for example, Shibata et al. (2012) found AP-1 motifs to be associated with chromatin accessibility differences among primates. Similarly, Biddie et al. (2011) showed that inhibiting AP-1 impedes formation of accessible chromatin and reduces glucocorticoid receptor (GR) binding, suggesting that AP-1 has a role in transcriptional pathways mediated by GR. There is also evidence in neurons that AP-1 functions as a general chromatin-accessibility factor, with tissue specificity conferred by cofactors or post-translational modification (Angel and Karin, 1991; Weber and Skene, 1998). These results are consistent with our finding, which further suggests that this role for AP-1 may be even more general.

Our motif results also highlighted the uniqueness and prominence of CTCF. It is well
known that CTCF is an extremely conserved and important factor Phillips and Corces (2009). Consistent with this, we found the CTCF motif highly significantly enriched in all 12 highly conserved clusters with low promoter overlap (Fig. 4.4). These clusters typically had extreme motif discovery e-values, with greater than 90% of the sequences containing the motif.

Using correlations between DNaseI signal and gene expression levels, we predicted mappings between >500M potential regulatory elements and their target gene. We showed that correlation results were often supported by 3C results where these data were available. However, the agreement was not perfect, which is understandable: most importantly, this may be due to either looping interactions or individual DHSs creating poised states without actually affecting expression (Margaritis and Holstege, 2008). Nevertheless, open chromatin correlation offers a complement to lower resolution, time-consuming, and expensive chromatin capture-based experiments. This increased level of resolution is necessary for some follow-up studies, such as increasing resolution of chromatin interaction data, or examining particular SNPs that occur in regulatory elements. Since regulatory mutations likely contribute to complex diseases (Epstein 2009), this type of data will be of clinical interest going forward. By narrowing down vast stretches of non-coding DNA to individual DHSs, we can look for individual SNPs specifically within these sites. As such, DNaseI/expression correlation is a powerful additional source of information to inform models of transcriptional regulation.

4.4 Methods

4.4.1 DNase-seq protocols

DNase-seq data was generated using two different protocols, referred to as the Duke and UW protocols (Song and Crawford, 2010; Sabo et al., 2004; Crawford et al., 2004). Briefly, for both, DNA is digested with a small concentration of DNaseI, and the ends of the
resulting fragments are short-tag sequenced, where the 5' end corresponds to a DNaseI cut site. The Duke protocol ligates an adapter to high molecular weight DNA fragments containing DNaseI-digested ends and then uses MmeI to create 20bp DNA molecules for sequencing. In contrast, the UW protocol requires two nearby DNaseI cuts creating fragments that are size selected and sequenced. We used DNase-seq data from 112 cell lines, 45 generated at Duke and 67 generated at UW. Data for at least two replicates were generated for each sample, tested for reproducibility, and combined for each cell-type. All DNase-seq and expression data are publicly available on the UCSC Genome Browser (http://genome.ucsc.edu). Read data are available at the Short Read Archive (Duke: SRX100886–SRX100920 and SRX189386–SRX189433; UW: SRX191006–SRX191058 and SRX201249–SRX201305).

4.4.2 Data normalization and processing

DHSs from all samples were combined as described previously (Thurman et al., 2012). For each cell-type, we calculated the number of DNaseI cuts in each DHS. Counts were quantile-normalized and capped at the 99th quantile by setting the top 1% of scores to the 99th quantile score. Counts were re-scaled by dividing by the maximum count for that cell-type, yielding a normalized score between 0 and 1. Artifactual differences between protocols (Fig. 4.1) were corrected using ComBat (Johnson et al., 2007). Because the data was too large for the ComBat model, we divided it into subsets of about 300,000 DHSs and corrected the batch affect individually on each subset.

4.4.3 Microarray protocol and normalization

We used Affymetrix Human Exon 1.0 ST microarrays to measure gene expression. We estimated gene-level expression by normalizing 332 microarray replicates measuring 140 cell lines (data available at GEO; Duke: GSE15805; UW: GSM651582, GSM472913, GSM651582) that included all samples for which we had DNase-seq data.
First, probesets flagged as cross-hybridizing were removed from the analysis (Salomonis et al., 2010). We assigned remaining probesets to genes based on the GENCODE v10 annotation (July 2011) available from Ensembl (Harrow et al. 2006; Flicek et al. 2011). We reasoned that constitutive exons would provide the most robust expression estimates, so we flagged constitutive probesets (those present in all protein-coding transcripts), and used only these for genes with at least 4. For all other genes, including all non-protein-coding genes, we used all probesets that mapped to the gene. After assigning probesets to genes, we normalized the data using Affymetrix Power Tools (APT) with the chipstream command rma-bg, med-norm, pm-gcbg, med-polish. This chipstream calls for an RMA normalization with gc-background correction using antigenomic background probes. Using hierarchical clustering and multidimensional scaling, we found the arrays primarily clustered based on protocol reagents and lab. To make the arrays comparable, we used ComBat to correct for this batch affect (Johnson et al., 2007). After correction with ComBat, the arrays grouped according to expected biological similarity (Fig. 4.1).

4.4.4 Classifying regulatory elements with a self-organizing map

A self-organizing map (SOM) was constructed using the kohonen R package (Wehrens and Buydens 2007) that was modified to handle more data. Our SOM consisted of a hexagonal 50x50 grid (2500 total clusters, or clusters). Since SOMs typically identify many similar clusters, the initially learned SOM was refined by merging similar clusters, resulting in 1,856 unique final clusters. Self-organizing maps

After building the initial SOM, we then refined the clusters to merge those that were sufficiently similar. To do this, we first defined a merge priority list using a hierarchical clustering on the cluster centers. This list defines the order in which the clusters are combined in the hierarchical clustering. We then iterated through this list and considered whether to join the two closest (in euclidean space) cluster centers. For a proposed cluster-join step, we calculated the Mahalanobis distance from all points in the smaller cluster
(fewer DHSs) to their cluster center, and also to the proposed joining cluster center. If the median (50th quantile) distance to the joining cluster center was less than the first quartile (25th quantile) distance to the original cluster center, then the clusters were merged, taking the center of the larger cluster. The Mahalanobis distance is a slight variation on Euclidean distance that considers the covariance matrix of points that belong to a cluster. When calculating the Mahalanobis distance of a point to a new cluster, the result can be thought of as a probability contour of generating that point from the density of that cluster. After considering each of the 2,499 merges proposed by the hierarchical clustering, we merged 644 SOM clusters, resulting in 1,856 unique final clusters.

4.4.5 CpG-island, promoter, and conserved element overlap

For each cluster, we extracted the 100 DHSs closest to the cluster center, as assessed by Mahalanobis distance, and tested these for overlap with promoters, CpG-islands, and conserved elements. Promoters were defined as 2kb upstream of the TSS for the UCSC RefGene annotation (Kent et al., 2002). CpG-island annotations (Bock et al. 2007) and PhastCons vertebrate conserved elements (Siepel et al., 2005) were downloaded from UCSC genome browser. We used R bioconductor packages GenomicRanges (Aboyoun, P, Pages, H, and Lawrence, 2012) and rtracklayer (Lawrence et al., 2009) to do the overlap analyses.

For this and many subsequent analyses, we selected only 100 representative DHSs for each cluster. We selected the 100 sites that were closest to the cluster center in Mahalanobis distance. Our rationale for choosing the top 100 is to enhance signal-to-noise. The SOM reduces the dimensionality of the data by clustering millions of sites into just a few thousand clusters. Every site must be assigned to a cluster, so sites that do not match one of the common patterns will essentially be randomly assigned. To eliminate this noise, selecting the top 100 sites in each cluster ensures that we are using only sites that are actually representative of the cluster pattern. For clusters with fewer than 100 sites, all sites are
4.4.6 Tissue lineage identity classifier

We used multinomial logistic regression to classify samples by tissue-type on the basis of hypersensitivity. Each non-malignant sample was assigned to one of 15 tissue lineage classes. Non-malignant samples from classes with too few (<5) samples were not used, leaving 80 samples distributed across the remaining classes: Brain (5), Endothelial (12), Epithelial (14), Fibroblast (27), Hematopoietic (12), Muscle (5), and ES (5).

For features, we identified the single hypersensitive site closest to each cluster center based on the Mahalanobis distance. We fit a multinomial logistic regression model using the glmnet R package (Friedman et al., 2010) with leave-one-out (79-fold) cross validation. We used misclassification frequency as the distance model and used LASSO regularization ($\alpha = 1$) for sparsity. We chose the lambda (regularization) parameter that minimized the misclassification error during cross-validation. Classifications for the malignant cell-types were predicted using a model trained with data from all 80 cell-types. For the sex classifier, we used a similar model, after filtering malignant samples and those with unknown sex.

4.4.7 Motif analysis

We selected the 100 DHSs from each cluster that were nearest the cluster center, as assessed by Mahalanobis distance. We extracted sequences for these regions and searched them for motifs using MEME (Bailey and Elkan, 1994) with the following settings: zero or one occurrences per sequence (ZOOPS), a motif size range of 8 to 22 nucleotides, and an e-value cutoff of 3. After identifying motifs, we used the bioconductor package motIV (Mercier et al., 2011) to compare the discovered motifs to the JASPAR (Portales-Casamar et al., 2010) motif database, recording the top 5 matches in each case.

There are several areas for future improvement in the SOM analysis. First, we are restricted to TFs that alter chromatin structure. Along this line, 521 of the 2500 SOM clusters
did not show any significant motif enrichment. This highlights that a common pattern does not necessarily imply a common motif or factor. The SOM also does not distinguish direct from indirect effects; A given DHS may adopt a particular DHS configuration via looping, in which case we would not necessarily find the same motif in all DHSs with that profile. Alternatively, several TFs could drive a particular pattern, complicating motif discovery, and TF binding is combinatorial so there is more complex logic (AND and OR). Some of these problems could be alleviated by combining these results with ChIP results for TFs or histone marks, DNA methylation, or DNaseI footprinting (Boyle et al., 2011). Additional information like this could enable us to further delineate among clusters, increasing our ability to draw meaningful biological conclusions.

### 4.4.8 General motif summary results

For the general motif summary, we filtered motifs based on both MEME (e-value < 1e-5) and motIV (e-value < 1e-8) scores. Each SOM cluster was assigned the tissue with the highest DHS signal in that cluster. For each TF, we found all clusters in which that TF was found, and collected the distribution of tissue types for those clusters (Fig. 4.8). To determine how many cell-types were present in a given cluster, we counted the number of cell-types with at least 1 sample above a normalized DHS score of 0.4. This cutoff provides a reasonable approximation, as the distribution of signals has low frequency between 0.2 and .95, so numbers within this range are likely to yield similar results (Fig. 4.4G). We then found the cell-type specificity of a given TF, according to the distribution of number of open cell-types in the clusters where that TF motif is enriched.

### 4.4.9 Discovered motifs in shared clusters

We found motifs enriched in clusters defined by DHSs with high signals in multiple cell-types (Fig. 4.6B). For example: GATA and FEV/ETS motifs were enriched in a cluster open in erythrocytes (red blood cells), megakaryocytes (platelet-producing bone marrow cells),
and endothelial cell-types (cluster 743). GATA family members are expressed in all three cell-types (Zhu and Emerson, 2002; Ferreira et al., 2005; Orkin et al., 1998; Shivdasani et al., 1997; Dorfman et al., 1992). ESE1 is an ETS family transcription factor that mediates inflammation, which may explain the megakaryocyte and endothelial specificity (Pignatelli and Carnevale, 2011; Rudders et al., 2001). The hepatocyte nuclear factor (HNF)-like motifs were enriched in a cluster shared between hepatocytes and endothelial cell-types (cluster 541). It is well known that HNFs regulate development and function in hepatocytes (Odom et al., 2004), and they also appear to function in endothelial cells (Osanai et al., 2002). The motif for RUNX1, a hematopoietic differentiation factor (Okuda et al., 2001), was enriched in a cluster preferentially open in hematopoietic lines as well as a number of other cell lineages (cluster 14). The Myf motif was enriched in a muscle and neural cluster (cluster 1518). In general, the expression patterns of these factors are consistent with the tissue-specificity of the motif signals.

These are cases where clusters grouping multiple cell-lineages were enriched for motifs. There are a variety of explanations for these results. Such a motif could actually be functioning in two lineages as a result of shared recent lineage history or shared regulatory mechanism despite divergent lineage history. Different factors could bind the same motif in different lineages, or a single factor could have a different function in different lineages. The result could also be an artifact of the method rather than biologically meaningful. For example, it may reflect the inability of the SOM to separate two unique profiles; in this case, the motif may be enriched in a subset of the regions with a strong enough signal to overcome the noise.

4.4.10 Validating motif discovery with ChIP results

To assess whether the motifs we discovered could be validated by published ChIP results, we tested for overlap between our results and ChIP peaks from the ENCODE project. We selected 43 genome-wide ChIP experiments in relevant cell-types for TFs with motifs we
discovered, and compared these results pairwise with the clusters that were enriched for those same motifs. For each cluster, we counted the percent of DHSs that overlapped ChIP peaks in each experiment. This comparison yielded a matrix of overlap percentages. We used Mann-Whitney significance test to measure the expected positive comparisons (where the motif enrichment matched the ChIPped factor) against the expected negative comparisons (where the enriched motif was for a different TF) to find p-values of overlap. The overlap for each factor was: AP-1 (FOS:JUN) (5% vs 0%), GATA (7% vs 1%), HNFs (30% vs 2%), Pou5f1 (18% vs 0%), IRFs (5% vs 1%), SPI1 (26% vs 1%), RUNX (89% vs 9%), NRSF (50% vs 1%), and Sp1 (5% vs 1%).

4.4.11 Mapping regulatory elements to target genes they regulate

We calculated the Pearson correlation across samples between gene expression and normalized DNaseI scores for each DHS within 100kb of each gene. To reduce noise, we set a minimum value for DNaseI signal (0.1) and for gene expression (4). We calculated a permutation p-value by calculating a null distribution of DHS correlations for each gene to a random sample of 10,000 DHSs from different chromosomes, and considered p < 0.05 significant.

4.4.12 Comparing DHS-gene correlation to experimental 3C evidence.

DHS data combined with expression gives us a reasonable hypothesis for mapping distal regulatory elements to genes. We proposed several connections between genes and regulators and tested a few well-studied 3C loci to investigate the overlap between our predictions and experimental results. We found the results to be remarkably consistent; however, in some sense, we expect overlap to be imperfect. For some loci, we found no convincing agreement with the reported 3C looping interactions. Disagreement can be for a variety of reasons. In most cases, this may be because we did not analyze the cell-types where the 3C interaction was described. In other cases, the examples highlight the incomplete agreement
between 3C and the correlation analysis, possibly due to poised DHSs or cross-hybridizing of probes used for expression analysis causing the incorrect reporting of expression levels. Here are some of the limitations to comparing our method with 3C experiments:

1. *Combinatorial regulation complicates the correlation analysis.* First, regulatory elements are known to function in concert. For example, there are complex relationships (such as boolean AND, OR, XOR, etc.) between elements. In the event of such higher-order relationships, we would not necessarily expect simple correlation between individual elements and expression. It is possible to build models that identify the simplest combinatorial relationships, but the possibilities quickly explode into the computationally intractable. This is an area of future research.

2. *Open chromatin sites can be bound by different factors.* DNase-seq signal is one step abstracted away from the actual binding of a transcription factor. A distal regulatory element could easily be open in many cell-types, but bound by different factors in each. In this case, the identity of the bound factors may be the meaningful signal that would affect expression, while the DNase signal may not show many differences across cell-types (and hence, would not correlate with expression). This points to the benefit of using additional information, like computational motif analysis or ChIP results to supplement the DNase signal. The correlations we see are therefore driven by regulatory elements with binding sites specific to a single cell-type.

3. *Both experiments are limited by tissue, environmental, and developmental specificity.* The regulatory connections we seek are often cell-type, environment, and developmental stage specific. For either 3C or correlation analysis, we will only see connections that occur under conditions included in the study. As such, the links proposed by either method should be considered a subset of all possible links.
4. **Poised enhancers may be DNaseI-hypersensitive or create long-range loops without affecting expression.** There is some evidence that long-range regulators may be poised to prepare a gene set for rapid response. Poised enhancers could form a chromatin loop and be DNaseI-hypersensitive, but not affect expression. One recent example is the GR-induced activation of the Ciz1 gene in mouse (Hakim et al., 2009). In this study, Hakim et al. proposed that a chromatin loop forms before expression is induced in order to facilitate rapid response. Another example is in the beta-globin locus, where progenitor erythrocytes begin to form the LCR loop without expressing the beta-globins, and expression begins later down the differentiation pathway. If a poised enhancer loops to interact with a target promoter without affecting expression, the correlation and 3C results will disagree (Zentner et al., 2011; Creyghton et al., 2010; Rada-Iglesias et al., 2011).

5. **Physical proximity may reflect packing, and have nothing to do with regulation.** DNA must be packed tightly to fit within the nucleus. Evidence suggests that physical proximity is often regulatory, but it is likely that some physical proximity is simply due to space restriction in the nucleus.

6. **Regulation may not require physical proximity.** Despite recent progress, how enhancers work is still unclear. Much emerging evidence supports chromatin loops bringing distal regulators near promoters of genes they regulate. These results imply tissue-specific physical proximity (Tolhuis et al. 2002). In a recent review, all tested transcription factors contributing to loops affected transcription (Sexton et al., 2009); this result held for both forming and maintaining loops. However, this is still an open question and scientists have been hypothesizing about it for decades (Ptashne, 1986). Even if physical proximity implied regulation, it does not follow that regulation implies physical proximity. There is some evidence for another model of distal regulation: the tracking model (Hatzis and Talianidis, 2002; Herendeen et al., 1992). These models are not necessarily mutually exclusive: an
enhancer may track along the DNA to find a promoter, thereby forming a loop (Dean 2011). Regulation that happens at a distance could possibly be identified by the correlation analysis but not by 3C.

7. **Populations of cells complicate measurements.** Neither the correlation analysis nor 3C will identify short-lived interactions because of noise inherent in a population of cells. Equivalently, long-lived interactions that occur in a small subset of cells will also be difficult to detect. These limitations may manifest themselves differently in a 3C experiment. For example, there could be a regulatory region that maintains an open chromatin state in the majority of cells throughout a time-course, while it may only be involved in an enhancing loop interaction for a brief moment. In such a case, we may identify a connection using the correlation analysis, but see weak or no evidence of loop formation.

8. **3C data comes primarily from mouse.** Most of the long-range interaction studies have been done in mouse cells. While there will clearly be differences between mouse and human, the high level of conservation in many of these regions make mouse a reasonable model to begin with. Vernimmen et al. (2009) have engineered mouse cells with human sequences to test interactions at the human alpha-globin locus. Despite key differences between mouse and human (Anguita et al., 2002), they conclude the overall architecture is maintained. Nevertheless, there are 65 million years of evolution since the human mouse divergence and there is no guarantee that 3C mouse data is always representative of human.

9. **Correlation detects indirect regulation.** Correlation may detect connections that are correlated but not interacting and directly influencing regulation. For example, a DHS will correlate not only with the gene it regulates but also with every gene within 100kb that has a matching expression pattern. This is both strength and a limitation: it flags potential trans-effects at the cost of specificity.
10. **3C resolution is limited by location and frequency of restriction sites.** These can be multi-kb-sized fragments. On the other hand, the correlation results highlight specific, 150 bp regulatory elements. This difference introduces complications into comparing the results. For example, there may be multiple nearby hypersensitive sites connecting in different directions. This may indicate a transitory looping structure, which a typical 3C experiment could miss due to diluted signal.

4.4.13 **POUF51/NANOG and ES-specific connections**

In our expression data, Oct4 (POU5F1) is expressed only slightly higher in ES cells than other cell-types. Several reviews have raised questions about the cell-type expression pattern of Oct4 and pointed out some question of the reliability of Affy arrays (de Jong and Looijenga, 2006; Liedtke et al., 2008). Affy HuEx arrays are unable to differentiate between two isoforms of Oct4: Oct4A, which is expressed in the nucleus and necessary for pluripotency, and Oct4B, which is expressed in the cytoplasm and present in somatic cells (Liedtke et al., 2008). Along similar lines, Nanog also has a very questionable expression result, due to probe cross-hybridization. Most of the probes mapping to Nanog on the Affymetrix HuEx arrays we used are flagged for cross-hybridization problems. The single probeset passing the cross-hybridization filter is not in a constitutive exon, so our Nanog expression measurements are questionable. In line with this, Nanog expression measurements are generally not higher in normal stem cells, though it is upregulated in Ntera2 cells. This limitation of the Affy arrays makes it more difficult for us to identify ES-specific connections, motifs, and pathways near these genes. In the future, more accurate expression data may enable a more thorough study of ES-specific information.

4.4.14 **Beta-globin locus and olfactory receptor genes**

In our microarray data, several olfactory receptor genes (OR51AB1P, OR51B4, OR51B3P, OR51B2) are overexpressed in K562 cells. As such, these genes correlate with K562-
specific DHSs throughout the locus. It is unlikely that these genes are expressed in erythroid tissue, as they are assumed to be primarily used in epithelial tissue. This expression can be explained by a few things: First, these are small genes which may not have accurate probes on the microarrays; and second, the K562 cell line highly expresses the nearby HBE1 gene (embryonic hemoglobin), which could be driving rampant expression of these nearby olfactory genes. Because this likely represents an artifact, we removed these connections from the figure.

4.4.15 Pathway analysis

For each SOM cluster, we created a set of genes consisting of all genes assigned to any DHS in that cluster. To filter noise, we selected only the DHSs in the nearest 90th quantile to the cluster profile, with a maximum of 1,000 sites in each cluster. We then calculated pathway enrichment scores using the bioconductor package MGSA (Model-based Gene Set Analysis) (Gentleman et al., 2004; Bauer et al., 2010, 2011). MGSA considers a query set of genes along with a database of pathways or gene sets, and builds a model to predict which gene sets are likely to have generated the genes in the query set. It uses a Bayesian method to calculate a posterior probability over gene sets, allowing us to infer which gene sets are likely to be activated. We considered a query set enriched for a given pathway if the posterior probability of activation was >0.5. Pathways in the KEGG (Kanehisa et al., 2012) and Reactome 40 (Matthews et al., 2009) pathway databases were considered.
Gene regulation is fundamental for the existence of life. It underlies basic biological processes from cell division to development to environmental response. The key to these processes lies in understanding their systems of gene regulation. My work has focused on regulatory DNA itself and how it controls transcription. But regulatory DNA is just a single layer in the regulatory system that also includes alternative splicing, mRNA decay, mRNA and protein localization, and protein stability. Each of these is important for cellular systems, and each is an area of active research, and each can independently be the focus of an entire career of research. In my focus on regulatory DNA, my work has sought to understand the first layer in the regulatory system: how do regulatory DNA sequences confer differential gene expression?

We cannot yet truly decode the regulatory DNA that the cell reads, but we have made monumental progress as a result of high-throughput DNA sequencing technology. In just a few years, we have gone from spotty annotation of regulatory elements to complete genome-wide maps across species and across cell-types. As we explore this data in greater depth, we will begin to learn more about the regulatory code. The work I present here has helped to uncover some properties of gene regulation. However, there is clearly more to be
5.1 The future of cross-species DHS comparisons

In the evolutionary context (Chapter 2), this work established the utility of evolutionary comparison of DNaseI-seq data. By adding in regulatory element activity and linking this to both expression and to genotype, we took a step closer to understanding how the regulatory system encodes differential expression, and ultimately, different phenotypes. We also annotated the first set of possible human-specific regulatory elements. There are many directions for future research in this area.

*Expanding cell-types.* We have so far focused on fibroblasts, and touched on lymphoblasts. These are among the most readily available samples; they can be harvested from living animals and humans. Expanding to other cell-types will yield different results and new regions. It will be interesting to investigate differences in nervous tissue and other more biologically relevant tissues. However, the difficulty of obtaining different cell-types humanely from primates will limit the scope of these studies.

*Validating computational predictions.* Our motif analysis predicted that differences in AP-1 binding lead to the differences in DNaseI hypersensitivity we detected. We have validated several of the predictions of AP-1 differences between human and chimpanzee and shown that AP-1 is actually differentially binding according to the predictions based on motif analysis and DNase-seq signal. This exciting result shows that our computational model reflects reality. However, the model is still quite restricted; admittedly, the test uncovered predictive trends at the aggregate level, and we do not have the power necessary to identify causative variants at individual sites.
Expanding species. Additional individuals or species will enable us to focus the results. We are working to leverage the power of evolutionary differences in primates to ask questions about less widespread differences. We are in the process of adding DNase-seq data from other primates, including orangutan, gorilla, and mouse lemur, as well as expanding the tree to other placentals such as mouse, rat, and dog. Additional species will increase our ability to identify not just trends across sites, but individual specific variants that confer regulatory difference. It is also possible to increase statistical power using additional individuals from the same species, but the variation inherent in a cross-species analysis provides extended scope by exploring more variants.

5.2 The future of cross-cell-type comparison in humans

In the tissue comparison within humans (Chapters 3 and 4), we established the utility of using DNase-seq to identify TF binding sites and predict regulatory targeting. The ENCODE project continues with a third phase to continue the effort to build an Encyclopedia of DNA elements. Besides the basic directions of increasing cell-types and data-types, there are several more targeted areas of interest for future research:

Changes over time. Open chromatin experiments have clearly made progress toward their primary goal to identify regulatory elements, but identifying them is only the beginning. Regulatory elements are dynamic; they drive expression specific to tissue, developmental stage, genetic background, and environment. Distal elements in particular often act in very specific contexts. To decode the human genome, regulatory elements must not only be identified, but also characterized by context specificity. Most of these data were derived from static snapshots of a particular cell-type. While all these cell-types are ultimately related in a developmental tree (each cell-type is derived from the embryo), the relationships among cell-types we sampled is too sparse to establish much regarding the developmental hierarchy. One way to approach this question is to look at how regulatory DNA acts over
time during development. Such an experiment requires significantly more work and expense than looking at just a single time point for a given cell-type, but this area is likely to see an increase in attention as sequencing capacity increases. In addition to the developmental time scale, changes over time can be explored in other contexts, such as cell-cycle stages, response to environment or therapeutics, and allelic background (Birney et al., 2010).

Connections to other signals. This dissertation focused primarily on integrating open chromatin and gene expression data. I also brought in histone modification and ChIP data in a few targeted analyses, but there is clearly much work to be done to integrate these data types. Supplementing with other coordinated outputs, including not only sequencing-based cellular signals, but also phenotype or behavioral data will contribute to our ability to decode the regulatory signal. Each time scale, environment, and data type explores a different piece of the regulatory system.

DNase footprinting. Another area of growing interest is using DNase footprinting to identify individual transcription factor binding sites. As our sequencing capabilities increase in depth, it is becoming possible to compare DNase footprints at individual sites for different transcription factors (Boyle et al., 2011; Pique-Regi et al., 2011; Neph et al., 2012). This system may eventually enable us to annotate regulatory DNA much more thoroughly, including not only the exact bound bases but also the identity of the factors binding them.

Assigning function to regulatory elements. By combining DNase data with ChIP for histone modifications and transcription factors, we can hypothesize the general function of each element; however, we have only begun to assign precise functions. To reveal developmental specificity of an enhancer requires a developmental experiment, such as an embryonic assay in non-human model organisms, accomplished by tethering enhancer sequences to basal promoters (Pennacchio et al., 2006). These assays are currently restricted to only certain
classes of regulatory elements, and this is an area of active research.

*Regulatory element targeting.* Another key is the question of what genes a particular regulatory element affects. Neither open chromatin, nor ChIP, nor even embryonic assays directly answer this question. The current best methods to reveal gene targets are chromatin conformation capture (Wei and Zhao, 2011; Dekker et al., 2002) or ChIA-PET (Fullwood et al., 2009) experiments. These experiments examine the physical proximity of promoters to regulatory elements. This idea is based on the mounting evidence that physical proximity is associated with regulation (Sexton et al., 2009); however, its precise importance remains unclear (Fraser and Bickmore, 2007). In addition, experiments that assay the nuclear organization of the genome remain expensive, time-consuming, and difficult to interpret (Dekker, 2006). I presented a potential alternative to leverage information across cell-types to correlate cell-type- and developmental-specificity of distal elements with promoter elements (Thurman et al., 2012) or gene expression (Ernst et al., 2011). However, this method provides only indirect evidence for regulation. Additional research and new methods will be necessary before we will be able to accurately identify the targets of regulatory elements.

### 5.3 Toward synthetic biology

This study of regulatory DNA collectively amounts to filling the gap between genotype and phenotype. How does a series of As, Ts, Gs, and Cs contain the information required to produce a particular expression pattern, developmental progression, and ultimately complete multicellular organism? We are only beginning to be able to relate regulatory DNA inputs to expression outputs, but as our models of regulatory DNA become more realistic, they will eventually lead us to realize the goals of personalized medicine and synthetic biology. In medicine, understanding the regulatory system is a prerequisite to predicting a patient outcome on the basis of DNA sequence. Combined with models for
epigenetics and environment, this will enable medics to prescribe specific therapy to fit an individual regulatory program. This will be a powerful step, but the ultimate test of our understanding will be not in diagnosis or even prediction, but in synthetic biology. When we understand the rules well enough to engineer new systems from nothing more than a set of essential building blocks, that will be a significant milestone toward having a complete understanding of DNA.
<table>
<thead>
<tr>
<th>CANONICAL</th>
<th>CELL-TYPE</th>
<th>TISSUE</th>
<th>MALIG phant</th>
<th>SEX</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>A549</td>
<td>Epithelial</td>
<td>Malignant</td>
<td>M</td>
<td>epithelial cell line from a lung carcinoma</td>
<td></td>
</tr>
<tr>
<td>A549</td>
<td>Epithelial</td>
<td>Malignant</td>
<td>M</td>
<td>fetal fallopian tube epithelial blast</td>
<td></td>
</tr>
<tr>
<td>AG0450</td>
<td>Fibroblast</td>
<td>Malignant</td>
<td>M</td>
<td>fetal lung fibroblast</td>
<td></td>
</tr>
<tr>
<td>AG09309</td>
<td>Fibroblast</td>
<td>Malignant</td>
<td>F</td>
<td>adult toe fibroblast</td>
<td></td>
</tr>
<tr>
<td>AG09319</td>
<td>Fibroblast</td>
<td>Malignant</td>
<td>F</td>
<td>umbilical skin fibroblast</td>
<td></td>
</tr>
<tr>
<td>AnAf</td>
<td>Fibroblast</td>
<td>Malignant</td>
<td>F</td>
<td>aortic adventitial fibroblast cells</td>
<td></td>
</tr>
<tr>
<td>AnSM12FM</td>
<td>Muscle</td>
<td>Malignant</td>
<td>U</td>
<td>aortie smooth muscle cells</td>
<td></td>
</tr>
<tr>
<td>BE1-C</td>
<td>BE1-C</td>
<td>Malignant</td>
<td>M</td>
<td>clone of the SK-N-BE neuroblastoma cell line</td>
<td></td>
</tr>
<tr>
<td>BF</td>
<td>Fibroblast</td>
<td>Malignant</td>
<td>M</td>
<td>skin fibroblast foreskin</td>
<td></td>
</tr>
<tr>
<td>CD14</td>
<td>Hematopoietic</td>
<td>Malignant</td>
<td>F</td>
<td>CD14-positive cells from leukapheresis</td>
<td></td>
</tr>
<tr>
<td>Chorion</td>
<td>Primate</td>
<td>Malignant</td>
<td>U</td>
<td>fetal membranes from planned cesarean delivery</td>
<td></td>
</tr>
<tr>
<td>CLL</td>
<td>Hematopoietic</td>
<td>Malignant</td>
<td>F</td>
<td>chronic T-cell lymphocytic leukemia</td>
<td></td>
</tr>
<tr>
<td>CMK</td>
<td>Hematopoietic</td>
<td>Malignant</td>
<td>M</td>
<td>acute megakaryocytic leukemia</td>
<td></td>
</tr>
<tr>
<td>CoTu29</td>
<td>Skin</td>
<td>Malignant</td>
<td>M</td>
<td>malignant melanoma</td>
<td></td>
</tr>
<tr>
<td>Fpsc1L</td>
<td>Skin</td>
<td>Malignant</td>
<td>F</td>
<td>child fibroblast</td>
<td></td>
</tr>
<tr>
<td>Glo38a</td>
<td>Glioblast</td>
<td>Malignant</td>
<td>U</td>
<td>glioblastoma, these cells (aka H54 and D54)</td>
<td></td>
</tr>
<tr>
<td>GM06090</td>
<td>HEMAPOIETIC</td>
<td>Transformed</td>
<td>F</td>
<td>B-lymphoecyte, lymphoblastoid</td>
<td></td>
</tr>
<tr>
<td>GM12664</td>
<td>HEMAPOIETIC</td>
<td>Transformed</td>
<td>M</td>
<td>B-lymphoecyte, lymphoblastoid</td>
<td></td>
</tr>
<tr>
<td>GM12865</td>
<td>HEMAPOIETIC</td>
<td>Transformed</td>
<td>F</td>
<td>B-lymphoecyte, lymphoblastoid</td>
<td></td>
</tr>
<tr>
<td>GM112878</td>
<td>HEMAPOIETIC</td>
<td>Transformed</td>
<td>F</td>
<td>B-lymphoecyte, lymphoblastoid</td>
<td></td>
</tr>
<tr>
<td>GM112891</td>
<td>HEMAPOIETIC</td>
<td>Transformed</td>
<td>M</td>
<td>B-lymphoecyte, lymphoblastoid</td>
<td></td>
</tr>
<tr>
<td>GM12292</td>
<td>HEMAPOIETIC</td>
<td>Transformed</td>
<td>F</td>
<td>B-lymphoecyte, lymphoblastoid</td>
<td></td>
</tr>
<tr>
<td>GM13509</td>
<td>HEMAPOIETIC</td>
<td>Transformed</td>
<td>F</td>
<td>B-lymphoecyte, lymphoblastoid</td>
<td></td>
</tr>
<tr>
<td>GM19238</td>
<td>HEMAPOIETIC</td>
<td>Transformed</td>
<td>F</td>
<td>B-lymphoecyte, lymphoblastoid</td>
<td></td>
</tr>
<tr>
<td>GM19239</td>
<td>HEMAPOIETIC</td>
<td>Transformed</td>
<td>M</td>
<td>B-lymphoecyte, lymphoblastoid</td>
<td></td>
</tr>
<tr>
<td>GM19240</td>
<td>HEMAPOIETIC</td>
<td>Transformed</td>
<td>F</td>
<td>B-lymphoecyte, lymphoblastoid</td>
<td></td>
</tr>
<tr>
<td>HI-hESAC</td>
<td>ES</td>
<td>Stem</td>
<td>M</td>
<td>embryonic stem cells</td>
<td></td>
</tr>
<tr>
<td>HI-hESAC</td>
<td>ES</td>
<td>Stem</td>
<td>U</td>
<td>undifferentiated embryonic stem cells</td>
<td></td>
</tr>
<tr>
<td>HI-hESAC</td>
<td>ES</td>
<td>Stem</td>
<td>F</td>
<td>embryonic stem cell (hESC) HP</td>
<td></td>
</tr>
<tr>
<td>HA-c</td>
<td>Astrocety</td>
<td>Stem</td>
<td>U</td>
<td>astrocytes-cerebellar</td>
<td></td>
</tr>
<tr>
<td>HA-c</td>
<td>Astrocety</td>
<td>Stem</td>
<td>U</td>
<td>astrocytes-cerebellar</td>
<td></td>
</tr>
<tr>
<td>HA-c</td>
<td>Astrocety</td>
<td>Stem</td>
<td>U</td>
<td>astrocytes-cerebellar</td>
<td></td>
</tr>
<tr>
<td>HBMEC</td>
<td>ENDOTHELIAL</td>
<td>Transformed</td>
<td>U</td>
<td>brain microvascular endothelial cells</td>
<td></td>
</tr>
<tr>
<td>HCF</td>
<td>Endothelial</td>
<td>Transformed</td>
<td>U</td>
<td>cardiac fibroblasts</td>
<td></td>
</tr>
<tr>
<td>HCF</td>
<td>Endothelial</td>
<td>Transformed</td>
<td>U</td>
<td>cardiac fibroblasts</td>
<td></td>
</tr>
<tr>
<td>HCM</td>
<td>Muscle</td>
<td>Malignant</td>
<td>U</td>
<td>cardiac myocytes</td>
<td></td>
</tr>
<tr>
<td>HCM</td>
<td>Muscle</td>
<td>Malignant</td>
<td>U</td>
<td>conjunctival fibroblast</td>
<td></td>
</tr>
</tbody>
</table>
References


Seringhaus, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. A., Batzoglou, S.,
Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Sidow, A., Trinklein, N. D.,
Zhang, Z. D., Barrera, L., Stuart, R., King, D. C., Ameur, A., Enroth, S., Bieda, M. C.,
Kim, J., Bhinge, A. A., Jiang, N., Liu, J., Yao, F., Vega, V. B., Lee, C. W. H., Ng, P.,
Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer,
M. A., Richmond, T. A., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J.,
Fowler, J. C., Cottet, P., Bruce, A. W., Dovey, O. M., Ellis, P. D., Langford, C. F., Nix,
D. A., Euskirchen, G., Hartman, S., Urban, A. E., Kraus, P., Van Calcar, S., Heintzman,
N., Kim, T. H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C. K., Rosenfeld, M. G.,
Aldred, S. F., Cooper, S. J., Halees, A., Lin, J. M., Shulha, H. P., Zhang, X., Xu, M.,
Haidar, J. N. S., Yu, Y., Iyer, V. R., Green, R. D., Wadelius, C., Farnham, P. J., Ren, B.,
Harte, R. A., Hinrichs, A. S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig,
A. S., Smith, K., Thakkapally, A., Barber, G., Kuhn, R. M., Karolchik, D., Armengol,
L., Bird, C. P., de Bakker, P. I. W., Kern, A. D., Lopez-Bigas, N., Martin, J. D., Stranger,
B. E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., Hallgrímsdóttir, I. B., Huppert,
Idol, J. R., Maduro, V. V. B., Maskeri, B., McDowell, J. C., Park, M., Thomas, P. J.,
Young, A. C., Blakesley, R. W., Muzny, D. M., Sodergren, E., Wheeler, D. A., Worley,
Wilson, R. K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D. B., Chang, J. L., Lindblad-Toh,
K., Lander, E. S., Koribine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B.,
and de Jong, P. J. (2007), “Identification and analysis of functional elements in 1% of
the human genome by the ENCODE pilot project.” Nature, 447, 799–816.

and heritable chromatin signatures in humans.” Human Molecular Genetics, 19, R204–9.

Blekhman, R., Man, O., Herrmann, L., Boyko, A. R., Indap, A., Kosiol, C., Bustamante,
underlie human disease susceptibility,” Current Biology, 18, 883–889.


Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S.,
chromatin across the genome.” Cell, 132, 311–322.


Bradley, R. K., Li, X.-Y., Trapnell, C., Davidson, S., Pachter, L., Chu, H. C., Tonkin,


Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A.,


specific and combinatorial usage of diverse transcription factors revealed by genome-wide

enhancers from DNA sequence.” *Genome research*, 21, 2167–80.

“An enhancer deletion affects both H19 and Igf2 expression.” *Genes and Development*, 9,
2079–2089.

pitfalls for data misinterpretation in stem cell research.” *Biological chemistry*, 389,
845–50.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour,
M., Gnerre, S., Alfoldi, J., Beal, K., Chang, J., Clawson, H., Cuff, J., Di Palma, F.,
Fitzgerald, S., Flicek, P., Guttmman, M., Hubisz, M. J., Jaffe, D. B., Jungreis, I., Kent,
W. J., Kostka, D., Lara, M., Martins, A. L., Massingham, T., Moltke, I., Raney, B. J.,
Rasmussen, M. D., Robinson, J., Stark, A., Vilella, A. J., Wen, J., Xie, X., Zody, M. C.,
Baldwin, J., Bloom, T., Chin, C. W., Heiman, D., Nicol, R., Nusbaum, C., Young,
Dihn, H. H., Fowler, G., Jhangiani, S., Joshi, V., Lee, S., Lewis, L. R., Nazareth, L. V.,
R. K., Delehaunty, K., Dooling, D., Fronik, C., Fulton, L., Fulton, B., Graves, T., Minx,
P., Sodergren, E., Birney, E., Margulies, E. H., Herrero, J., Green, E. D., Haussler, D.,
*Nature*, 478, 476–82.


Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati,
P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May,
“Reactome knowledgebase of human biological pathways and processes.” *Nucleic acids
research*, 37, D619–22.


Odom, D. T., Zizlsperger, N., Gordon, D. B., Bell, G. W., Rinaldi, N. J., Murray, H. L.,
Volkert, T. L., Schreiber, J., Rolfe, P. A., Gifford, D. K., Fraenkel, E., Bell, G. I., and
factors.” Science, 303, 1378–81.

Odom, D. T., Dowell, R. D., Jacobsen, E. S., Gordon, W., Danford, T. W., Maclsaac, K. D.,
transcriptional regulation has diverged significantly between human and mouse,” Nature
Genetics, 39, 730–732.

Ohno, S. (1972), “So much ”junk” DNA in our genome.” Brookhaven symposia in biology,
23, 366–70.


Olson, M. V. and Varki, A. (2003), “Sequencing the chimpanzee genome: insights into

male-specific muscle between Drosophila species: a genetic and developmental analysis.”
Evolution & development, 9, 368–77.

factor GATA-1 in megakaryocyte development.” Stem cells, 16 Suppl 2, 79–83.

Osanai, M., Chiba, H., Kojima, T., Fujibe, M., Kuwahara, K., Kimura, H., Satoh, M.,
of endothelial Fas ligand (FasL) to prevent cancer cell transmigration: a novel defense
mechanism of endothelium against cancer metastasis.” Japanese Journal of Cancer
Research, 93, 532–41.

“The beta-globin nuclear compartment in development and erythroid differentiation.”
Nature Genetics, 35, 190–4.

(2010), “ZEB1 and CtBP form a repressive complex at a distal promoter element of the

Reviews Genetics, 10, 669–680.


Biography

Nathan C. Sheffield was born in California, USA. He earned his bachelor’s degree in Bioinformatics from Brigham Young University, Provo, Utah. At BYU, he did research in population genetics (Stacey et al., 2008) and phylogenetic systematics (Sheffield et al., 2008, 2009, 2010; Song et al., 2010; Sheffield, 2013). He completed his PhD in 2013 in Computational Biology and Bioinformatics at Duke University, Durham, North Carolina. At Duke, he received several awards, including the National Science Foundation Graduate Research Fellowship, the James B. Duke Fellowship, the Duke Primate Genomics Initiative Summer Fellowship, the NSF Nordic Research Opportunity Award, and the Duke Graduate School Teaching Mini-Grant. While working on his PhD, he authored several publications included in this dissertation (Sheffield and Furey, 2012; Song et al., 2011; Shibata et al., 2012; Sheffield et al., 2013) and also contributed to other research (Tewari et al., 2012; Natarajan et al., 2012; Myers et al., 2011; Dunham et al., 2012; Thurman et al., 2012).