THE INFLUENCE OF CHROMATIN STATE ON PLURIPOTENCY, TRANSLOCATIONS, AND GENE REGULATION IN HUMAN CELLS

Zhuzhu Z. Zhang

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum of Bioinformatics and Computational Biology.

Chapel Hill
2014

Approved by:
Terry Magnuson
Terrence S. Furey
Gregory E. Crawford
Yufeng Liu
Joseph G. Ibrahim
©2014
Zhuzhu Z. Zhang
ALL RIGHTS RESERVED
ABSTRACT

ZHUZHU Z. ZHANG: The influence of chromatin state on pluripotency, translocations, and gene regulation in human cells
(Under the direction of Terrence S. Furey)

A human body consists of more than a thousand cell types, each having a unique identity and function. Despite their distinct functions, all the different cells contain the same genetic information encoded in the human genome. One of the fundamental questions in biology is- how does a single genome provide the instruction for different cell types? We have learned that only a small part of the genomic information is used in each cell, and that the usage of the genome varies in different cells. The precise regulation of the genome usage is the key to cell identity.

The usage of the genomic information highly depends on whether the DNA sequence containing specific information is directly accessible to DNA-binding proteins such as transcription factors that read and translate the encoded information. Most of the genomic DNA wrapped around histone proteins, forming the nucleosomes. Most DNA-binding proteins cannot bind to their target DNA sequences if a nucleosome is present. Therefore, the nucleosome-depleted regions, “open chromatin”, represent regions of the genome that are accessible and the genomic information that is used in the cell. In this dissertation, I study the landscape and function of open chromatin, and its role in defining cell identity and function.

I examine the open chromatin architecture in a specific case of cell identity – the
reestablishment of pluripotency in terminally differentiated cells by reprogramming the cell fate
(Chapter II). Induced pluripotent stem cells (iPSCs) are reprogrammed from differentiated
somatic cells. Compared to their naturally existing counterpart embryonic stem cells (ESCs),
iPSCs have very similar but in many cases slightly altered developmental potential when
differentiating into other cell types. The cause of the different development potential is poorly
understood. In this study, I show that the regulatory landscape defined by open chromatin is
highly similar between hESCs and hiPSCs but differs at a set of key development genes. More
importantly, the chromatin differences do not appear to affect the transcription profiles at the
pluripotent state, but instead impact the regulation of transcription upon differentiation. These
results suggest that the accessibility of genomic information controlled by chromatin structure
does not only regulate the cell identity at its current state, but also influence the precise
regulation of its developmental potential.

In addition, I describe a high-throughput method I developed for functional annotation of
the regulatory elements marked by open chromatin (Chapter III). Using this approach, I
identified 3,428 open chromatin regions associated with enhancer activities in a reporter assay,
demonstrating the feasibility of functional characterization of several thousand regulatory
elements in a single experiment of this design. At last, I investigate the role of chromatin
structure in the development of cancer (Chapter III). The results indicate that characteristic
chromatin features marked by specific histone modifications may highlight genomic loci that are
susceptible to chromosomal translocation in hematologic malignancies.
To my dearest parents Siji Zhang and Yisheng Song, 
and my grandparents Wexiong Zhang and Guiru Chu
ACKNOWLEDGMENTS

I am deeply grateful to the many people who have made this dissertation possible. First, I would like to sincerely thank my mentors, Dr. Terry Furey and Dr. Greg Crawford. I could not have finished this dissertation without your immeasurable support, guidance, and encouragement. I also thank Dr. Jason Lieb for providing an environment rich with resources, exciting ideas, and engaging science. I am thankful to the members of my committee, Dr. Terry Magnuson, Dr. Yufeng Liu, and Dr. Joe Ibrahim, for the helps and suggestions on my scientific training, and especially Dr. Terry Magnuson and Dr. Yufeng Liu for their advice on my career development. I have had the privilege to work with many other amazing scientists, including but not limited to Josh Chenoweth and Paul Tesar on the stem cell project, Elliott Margulies and Hatice Ozel Abaan on the high-throughput functional assay, Tom Misteli and Bharat Burman on the chromosomal translocation project, and Wei Sun and Naim Rashid on various statistical topics. Your knowledge and scientific visions taught me to think and do science.

I am extremely thankful to all the members of Lieb lab. You have supported me in countless ways over the years. I cannot imagine how I could have survived graduate school without all of you. You have given me advice, encouragement, support, and friendship. You are truly the best colleagues and friends one can ever hope for. I am so fortunate and proud to be a member of the group. I want to specifically thank Sebastian Pott and Erin Osborne. Seb has been an incredible mentor and friend, and every advice Erin gave me has profoundly advanced me
scientifically. I am grateful to all my friends, especially Janet Doolittle, Alicia Midland, Meng Jin, Saijun Xue, and “XiaoZei”, for your company and moral support. I thank Laurie Mesibov for showing me how kind and strong a soul could be. I would also like to thank the experience of graduate school itself. Unquestionably it has been the most significant and meaningful years in my professional and personal life. It has taught me how to be a scientist, and more importantly, how to be the best of myself.

Last but not least, I cannot possibly say enough to thank my parents Siji Zhang and Yisheng Song. You are the BEST parents and friends, you made who I am today. You have been so loving and supportive; you are always there for me. Even in the roughest times, I am comforted knowing that you are by my side. I am so sorry that I have been away all these years. I hope you know that I really love what I do, and I am forever indebted to you for your unconditional love and support. And my dearest grandparents, especially my grandpa Wenxiong Zhang, I know how much you wanted to see me finishing this Ph.D.. I want to let you know that I have always had you in my heart. Now I am finally getting here, I can only imagine how happy and proud you would be. Tianji Cai, thank you for being a great partner, and most importantly, for being a best friend over the past 10 years.
TABLE OF CONTENTS

LIST OF TABLES ........................................................................................................................................... xii

LIST OF FIGURES ........................................................................................................................................... xiii

LIST OF ABBREVIATIONS AND SYMBOLS ................................................................................................. xv

CHAPTER I: INTRODUCTION ......................................................................................................................... 1

1.1 Chromatin dynamics and transcription regulation .................................................................................. 2

1.1.1 Open chromatin is a hallmark of active regulatory elements ............................................................ 2

1.1.2 Genome-wide mapping of open chromatin ........................................................................................... 4

1.2 Chromatin in human pluripotent stem cells ............................................................................................ 6

1.2.1 iPSCs differ from ESCs in subtle but potentially important ways ....................................................... 6

1.2.2 Complete reconfiguration of chromatin structure during reprogramming is fundamental to the pluripotent identity of iPSCs ........................................................................................................ 9

1.3 Functional annotation of regulatory elements in the human genome .................................................. 10

1.3.1 Functional annotation of the human genome requires developing high-throughput methods to efficiently predict and test regulatory element function on a genome-wide scale ........................................................................................................ 10

1.3.2 Enhancers activate gene expression in an orientation-independent manner ..................................... 12

1.4 Chromosomal translocations in oncogenesis ......................................................................................... 12

1.4.1 Chromosomal translocations are genetic hallmarks of most cancer cells ......................................... 12

CHAPTER II: INCOMPLETE REPROGRAMMING OF DNASE HYPERSENSITIVE SITES IN HUMAN IPS CELLS IS ASSOCIATED WITH ALTERED REGULATION OF DEVELOPMENTAL GENES UPON DIFFERENTIATION ................................................................................................. 15

2.1 OVERVIEW .............................................................................................................................................. 15
CHAPTER III: INITIAL DEMONSTRATION OF A STRATEGY FOR EN MASSE FUNCTIONAL CHARACTERIZATION OF HUMAN ENHANCER ELEMENTS

3.1 OVERVIEW ........................................................................................................... 62

3.2 INTRODUCTION ..................................................................................................... 63

3.3 RESULTS .................................................................................................................. 65

3.3.1 A high-throughput method to test enhancer function of FAIRE-enriched DNA fragments. ......................................................................................................................... 65

3.3.2 3,428 DNA elements associated with orientation-independent enhancer activity were identified............................................................................................................................... 69

3.3.3 Fragments positive for enhancer function tend to lie distal to transcription start sites.................................................................................................................................................. 73

3.3.4 Genes close to identified enhancer elements were relatively highly expressed ....... 74

3.3.5 DNA motifs in identified enhancer elements.......................................................................................................................... 74

3.4 DISCUSSION .............................................................................................................. 77

3.5 METHODS ................................................................................................................ 80

CHAPTER IV: SPECIFIC HISTONE MODIFICATION PATTERNS MARK TRANSLOCATION BREAKPOINTS

4.1 OVERVIEW .............................................................................................................. 84

4.2 INTRODUCTION ..................................................................................................... 85

4.3 RESULTS .................................................................................................................. 88

4.3.1 Systematic comparative analysis to identify chromatin features that are enriched at recurrent chromosomal translocation sites in hematopoietic malignancies ...... 88

4.3.2 Translocation genes are enriched for H3K4me1 in CD34+ cells............................... 90

4.3.3 Histone marks and DNase-I hypersensitivity defined subpopulations of translocation genes in CD34+ cells......................................................................................................................... 92

4.3.4 Histone modifications around specific breakpoints .............................................. 94
LIST OF TABLES

Table 3.1. Sequencing depth, number of mapped reads, and enriched FAIRE sites in six libraries................................................................. 71

Table 4.1. 74 translocation genes ................................................................. 102

Table 4.2. Breakpoints in 19 translocation genes........................................ 103
LIST OF FIGURES

Figure 1.1. Open chromatin regions ................................................................. 3
Figure 1.2. DNase I hypersensitivity and FAIRE detect open chromatin regions .......... 5
Figure 2.1. Experiment description and data overview ........................................ 20
Figure 2.2. Approximately 4% DHS sites are differentially open ......................... 22
Figure 2.3. dDHS sites occurred in clusters ...................................................... 24
Figure 2.4. dDHS regions show differences in local chromatin environment in differentiated somatic cells prior to reprogramming ...................................................... 29
Figure 2.5. dDHS sites clustered around developmental genes ................................ 33
Figure 2.6. Incomplete chromatin reconfiguration may foreshadow altered gene regulation upon differentiation .......................................................... 37
Figure 2.7. Incompletely reprogrammed DHS sites around GABR locus foreshadows altered expression after in vitro differentiation .................................................. 40
Figure 2.8. hiPSC and hESC lines have similar expression profiles .......................... 56
Figure 2.9. Identification of hESC-, hiPSC-, and fibroblast-specific DHS sites ............ 56
Figure 2.10. Differential DHS (dDHS) sites and dCOREs cluster nonrandomly across the genome and are connected to developmental genes .................................. 57
Figure 2.11. The local chromatin environment in differentiated somatic cells prior to reprogramming - AG04450 ................................................................. 58
Figure 2.12. Motif enrichment, genomic location, and gene connectivity of dDHS sites. .... 59
Figure 2.13. Location of dDHS sites surrounding key developmental genes ............. 60
Figure 2.14. Differential DHS sites at gene loci foreshadow expression misregulation upon in vitro differentiation .......................................................... 61
Figure 3.1. Experiment design for high-throughput characterization of enhancer activities of open chromatin regions isolated by FAIRE ........................................ 67
Figure 3.2. Collecting GFP-positive cells in fluorescence activated cell sorting (FACS). .... 69
Figure 3.3. Most enhancer positive regions appeared to be distal regulatory elements that drove the expression of nearby genes .............................................. 72
Figure 3.4. 3,428 enhancer positive fragments were enriched in transcription factor binding sites ................................................................. 76
Figure 4.1. Comparison of histone modifications and DNase hypersensitivity signals in 74 translocation genes and their 5076 control genes. ................................................................. 91

Figure 4.2. Comparison of histone modifications in individual translocation genes.............. 93

Figure 4.3. Histone modification levels in 10 windows around the breakpoint in genes HOXA11 and RARA.............................................................................................................. 97

Figure 4.4. Comparison of histone modifications and DNase hypersensitivity signals in breakpoint regions of 19 translocation genes and control regions of their control genes........ 104

Figure 4.5. Histone modification levels around the breakpoint in gene ABL1 ..................... 105

Figure 5.1. Schematic representation of reporter constructs to be employed for identification of regulatory function. ................................................................. 110
LIST OF ABBREVIATIONS AND SYMBOLS

ChIP: chromatin immunoprecipitation

DHS: DNase I hypersensitivity

ENCODE: The Encyclopedia of DNA Elements

FAIRE: formaldehyde-assisted isolation of regulatory elements

FISH: fluorescence in situ hybridization

kb: kilobase

Mb: megabase

PCA: principle components analysis

RNA-seq: RNA-sequencing

RPKM: Reads per kilo base per million

RRBS: reduced representation bisulfite sequencing

TF: transcription factor

TSS: transcription start site
CHAPTER I: INTRODUCTION

An adult human body contains more than $10^{13}$ cells [1] that consist of thousands of different cell types each with a specific identity performing a specialized function. All cells inherit and share the same genome that is contained in the zygote. Therefore, a fundamental question in biology is – how does a single genome encode the information required to produce the diversity of cell types? More specifically, how is the information accessed and used in different cell types, and how is this access regulated?

The human genome is composed of 3 billion nucleotides of DNA. The genomic DNA, 2 nanometers in diameter and in total approximately 2-meter long, is tightly packaged in the nucleus that is on average 6 micrometers in diameter. To achieve the ~10,000-fold compaction, genomic DNA wraps around histone proteins forming nucleosomes, with the nucleosome core particle (Figure 1.1) consisting of about 146 base pair (bp) of DNA and an octamer of histone proteins. Such “beads on a string” structure is the basic building block of chromatin that is organized by many levels of folding. The formation of nucleosomes is not only the first level of chromatin organization, it also controls the access of information encoded in DNA in each cell. Most proteins that regulate transcription require direct binding to nucleosome-free DNA to function [2-5]. Therefore, nucleosome depletion, or the formation of regions of “open chromatin”, is a mark of transcription factor binding events and the underlining regulatory elements that are specifically accessed and used in each cell type [2-4,6-8].
To further understand the regulatory role of chromatin dynamics at the nucleosome level in determination and specification of cell identity and function, I examined genome-wide chromatin states in human embryonic stem cells (hESCs) and induced pluripotent stem cells (hiPSCs). More specifically, I investigated how subtle differences in the open chromatin landscape between the two human pluripotent cell types may lead to different regulation of transcription during cellular differentiation (Chapter II). In addition, I developed a high-throughput method to test the function of regulatory elements dictated by open chromatin on a genome-wide scale in a single experiment (Chapter III). Finally, I explored the role of chromatin structure and predisposition of specific histone modifications in chromosomal translocation prior to oncogenesis (Chapter IV). As an introduction to this dissertation, I review the function and identification of open chromatin, detail our current understanding of epigenetic reprogramming during hiPSCs production, discuss technologies and obstacles of functional annotation of human regulatory elements, and close with a brief description into chromosomal translocation in oncogenesis.

1.1 Chromatin dynamics and transcription regulation

1.1.1 Open chromatin is a hallmark of active regulatory elements

The key to the establishment of the cell identity and function is the regulation of the usage of information coded in the genome. In eukaryotes, such regulation is controlled largely through the regulation of the chromatin state and DNA accessibility. Transcription factors and nucleosome compete for DNA-encoded information (Figure 1.1). Nucleosome disruption at active regulatory regions is a conserved feature of eukaryotic chromatin [2-4,9-13]. The binding of sequence-specific regulatory factors and the recruitment of chromatin remodeling activities at promoters, enhancers, silencers, and insulators is typically associated with nucleosome eviction
from chromatin in eukaryotic cells [14] [5,15]. Other factors that influence the nucleosome stability and positioning include the inherent properties of DNA sequence [16], the activity of chromatin remodeling complexes such as SWI/SNF (SWItch/Sucrose NonFermentable) [17], transcriptional initiation and elongation by RNA polymerases [18], and the post-translational modification of histone proteins [19-21] and incorporation of histone variants [19,22,23]. In any given cell type, those mechanisms work together to establish an overall chromatin architecture in which regions of open chromatin dictate the accessible part of genome and active regulatory elements. Genome-wide mapping and characterization of open chromatin in different cell types can provide insights in how the variation in accessibility and utilization of genomic information contribute to the specification of different cell identities and functions.

**Figure 1.1. Open chromatin regions.** In most cases, a DNA-binding protein (light blue sphere) such as a transcription factor cannot access its binding site (purple box) if it is occupied by a nucleosome (dark blue sphere). Therefore, nucleosome depletion is a mark of transcription factor binding and active transcription regulation.
1.1.2 Genome-wide mapping of open chromatin

Regions of open chromatin can be identified by deoxyribonuclease (DNase I) hypersensitive site mapping by preferential digestion of nucleosome-free regions [2-4,6,24,25], or FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) which probes open chromatin through differences in formaldehyde cross-linking efficiency between nucleosomal and non-nucleosomal DNA [7,14,26-28].

DNase I is an endonuclease that cleaves DNA at the phosphodiester linkages of its backbone. When the DNA is nucleosomal as it is in chromatin, regions with nucleosome depletion is cut preferentially by DNase I which is commonly referred as hypersensitive to DNase I cleavage [2-4,6]. DNase I hypersensitivity (DHS) was originally reported at the transcription start sites (TSSs) of heat shock genes upon activation in *Drosophila*, and has since been used to detect regions of open chromatin and the underlying regulatory elements. DHS sites mapping can be carried at individual loci or genome-wide via coupling with microarrays [24,25,29] or, as in most recent studies, with high-throughput sequencing [7,8,25,30,31].

FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements) [14,26,28] uses differences in formaldehyde cross-linking efficiency between nucleosomal and non-nucleosomal DNA to not only detect but also isolate the nucleosome-depleted DNA. Formaldehyde crosslinks proteins and DNA in direct contact [32,33]. Within each nucleosome core on average 146 bp DNA wraps around an octamer of histone proteins, providing approximately 10 to 15 histone-DNA interaction sites [34] for potential crosslinking by formaldehyde. In contrast, most DNA-binding proteins bind to DNA sequences of 5 to 15 bp [35,36], resulting few potential crosslinking sites. Therefore, the difference in crosslinking efficiency is used the simple procedure of FAIRE to isolate nucleosome-depleted genomic DNA [14,26,28] - chromatin is
first cross-linked with formaldehyde, sheared by sonication, and then a phenol-chloroform extraction is performed in which nucleosome-free fragments preferentially segregate and are enriched in the aqueous phase. The genomic DNA fragments in the aqueous phase can then be mapped by microarrays [14,29] or high-throughput sequencing [7,27,31].

Figure 1.2 DNase I hypersensitivity and FAIRE detect open chromatin regions that identify putative regulatory elements in diverse cell types (A) DNase-seq and FAIRE-seq. (B) DNase I hypersensitivity and FAIRE data from seven cell lines surrounding the HNF4A locus (145 kb) shows both ubiquitous and cell-type specific open sites that are especially prevalent in HepG2 cells. Pol II, CTCF, and Myc ChIP-seq peaks that overlap open chromatin are highlighted. (Song et al. (2011) Genome Research [7])
More recently, a new assay for probing open chromatin regions, ATAC-seq, was reported based on direct \textit{in vitro} transposition of sequencing adaptors into native chromatin [37]. Transposons have been shown to preferentially integrate into nucleosome-depleted regions of the genome [38]. In ATAC-seq, adaptors for high-throughput sequencing are integrated into regions of open chromatin by hyperactive Tn5 transposase. The adaptor-tagged genomic regions can then be amplified and sequenced. ATAC-seq is an exciting new method for mapping open chromatin, as its simple two-step protocol requires less time than DNase I or FAIRE and is reported to perform on as few as 500 cells [37], while DNase I [30] and FAIRE [28] typically require at least 1 million cells.

1.2 Chromatin in human pluripotent stem cells

1.2.1 iPSCs differ from ESCs in subtle but potentially important ways.

Pluripotent stem cells possess the unique capacity to indefinitely self-renew and to develop into all cell types in an adult body. They can potentially be used for tissue replacement after injury and regenerative medicine for curing disease such as Parkinson’s. The naturally occurring pluripotent stem cells are embryonic stem cells that are derived from an early-age embryo [39]. In 2006, an artificial type of pluripotent stem cells, induced pluripotent stem cells, were generated from mouse somatic cells by forced expression of a defined set of factors [40]. In 2007, human iPSCs were generated using similar strategies [41,42]. For use in regenerative medicine and disease research, iPSCs are proposed to have two advantages over their natural counterpart: they avoid the ethical concerns regarding the use of human embryos, and they can be derived in a patient-specific manner to avoid immune rejection after transplantation.

iPSCs and ESCs are highly similar in many respects, including cell morphology [7,8,25,27,29,31,40-47], expression of pluripotency markers [7,14,26,28,40-56], histone
modification patterns [56-58], teratoma formation [40-42,59,60], and ability to differentiate into all germ layers [40-45,61-64]. Viable and fertile adult mice have been produced from mouse iPSCs, demonstrating their pluripotent potential [65-67]. On the other hand, studies have also reported genetic and epigenetic differences between iPSCs and ESCs.

Alterations in the differentiation potential. Human iPSCs have been reported to have lower efficiency than ESCs in differentiation to neural [68,69] and blood lineages [70]. In addition, Human iPSCs have been suggested to have altered developmental potency [68,69,71-74]; in many cases they have skewed differentiation potential towards the cell lineage of their cells of origin [71-73,75]. For example, iPSCs derived from human pancreatic islet beta cells were shown to differentiate more readily into insulin producing cells [73]. Most importantly, mice produced from mouse iPSCs seemed to have a higher death rate, with some displaying physical abnormalities [65,74]. In addition, Zhao et al. reported that in contrast to mouse ESCs, mouse iPSCs trigged an immune reaction after implanted in mice [76].

Alterations in genomic DNA sequence. Compared to human ESCs, studies show that human iPSCs have more novel copy-number variations (CNVs) [77-79], and ten-times higher predicted mutation frequency [80].

Transcriptional activities. iPSCs and ESCs have very similar global expression [81]. Nonetheless, a set of genes including some developmental genes were reported to be differentially expressed between iPSCs and ESCs [72,81-83]. However, these genes show laboratory-specific expression [56,81,84], and it remains unclear whether stable patterns of differential gene expression are shared between different iPS cell lines. In addition, the reported differential gene expression between ESCs and iPSCs cannot be explained by the histone
modifications [56], and only partially correlate with the DNA methylation patterns at the
promoters of these genes [72].

Epigenetic abnormalities. Despite the global similarities of DNA methylomes and
histone modification patterns in iPSCs and ESCs [47,48,72,85-88], aberrant methylation of CG
dinucleotides [8,31,72,88-101] and incomplete reinstating of non-CG methylation [99-101], an
epigenetic feature seen only in pluripotent cells, were observed. The methylation inconsistencies
between iPSCs and ESCs are associated with epigenetic memory of somatic progenitors [71-
73,102] and de novo methylation aberrations, which appears to impact developmental potentials
[69,71-73,102]. Residual somatic DNA methylation patterns and aberrant de novo methylations
in iPSCs were reported to transmit to differentiated cells and contributed to transcription
variation upon differentiation [101].

A few common themes regarding the genetic and epigenetic stabilities of ESCs and
iPSCs have emerged [103]. First, these genetic and epigenetic abnormalities represent both cell-
of-origin memories [71,73,83,104] and features that are unique to iPSCs [73,81]. Some of the
abnormalities seem to be common in several independently derived iPS cell lines [93,94,99,101],
while others tend to be laboratory-[84,88] or technique- specific [83,105], thus are likely
stochastic rather than functional variations. Second, the observed differences between iPSCs and
ESC can arise at different stages of reprogramming and subsequent culture of iPSCs [77-79].
Notably, some lesions, for example novel CNVs, seem to be selected against and disappear in
higher-passage iPSCs [78], indicating they are not permanent. Third, regions prone to
amplification, deletion or point mutation seem to be enriched in genes involved in cell-cycle
regulation [78,80] and cancer [77,79], suggesting that some functional genomic regions may be
more susceptible to aberrant reprogramming.
Taken together, these data indicate that iPSCs differ in subtle but potentially important ways from ESCs. Closer inspection of these differences is needed to understand how they may potentially affect the regulatory program that controls the pluripotent cell identity.

1.2.2 Complete reconfiguration of chromatin structure during reprogramming is fundamental to the pluripotent identity of iPSCs.

The pluripotent stem cell state is under the control of a transcriptional program which is likely implemented in the context of a particularly “open” chromatin state [106,107] that is unique to pluripotent stem cells. The chromatin in embryonic stem cells is at a more open state when compared with chromatin in somatic cells [55,108,109]. This globally open state of chromatin is considered important for the maintenance of pluripotency, as it allows DNA regulatory elements openly accessible to transcriptional regulators [106,110]. Reprogramming of somatic cells to pluripotent stem cells requires reconfiguration, especially re-opening, of chromatin. Incomplete re-opening of chromatin during reprogramming may leave an epigenetic memory of the original cell type, and inaccurate re-opening may result in aberrant gene expression. Thereby, genome-wide characterization of the chromatin structure in iPSCs in comparison to ESCs and parental somatic cells is essential to understand the cell identity of iPSCs.

Genome-wide maps of nucleosomes with histone H3K4me3 and H3K27me3 modifications indicate that there is little difference between ESCs and iPSCs with respect to these markers [56]. However, such histone mark maps do not completely reflect the global chromatin structure in iPSCs. Analysis of global chromatin compaction using approaches like DNase hypersensitivity will illustrate how completely the chromatin structure in iPSCs is reconfigured.
1.3 Functional annotation of regulatory elements in the human genome

1.3.1 Functional annotation of the human genome requires developing high-throughput methods to efficiently predict and test regulatory element function on a genome-wide scale.

DNA regulatory elements are short non-coding regions that can be bound by sequence-specific regulatory factors that contribute to the regulation of the expression of genes. Promoter, enhancer, insulator and silencer are major categories of regulatory elements. The annotation of the human genome has suggested the existence of a large number of regulatory elements. As part of the effort of the Encyclopedia of DNA Elements (ENCODE) consortium, over 100,000 putative regulatory elements are defined by open chromatin in each of the more than hundred human cell types and tissue samples [7,8,29,31]. The number of new sites identified has not yet saturated even as the number of cell lines analyzed has increased [7,8,31]. Given the variety of cell types present in human, the total number of regulatory elements in human genome is likely on the order of $10^6$ [8,31,111]. While DNase I, FAIRE, ChIP and other methods have made identification of putative regulatory elements routine, those methods do not provide information regarding the functional activity of each element. The rapidly increased list of putative regulatory elements has created an urgent need for genome-scale methods that can efficiently annotate the functions of these putative elements.

Regulatory elements have been discovered and characterized by gene-centric studies and systematic large-scale approaches. In gene-centric studies, reporter and transgenic assays are commonly used and rather accurate. However, these experiments are labor-intensive, and the can only test ~10 elements in a single experiment [112-124]. Systematic approaches include computationally predicting regulatory elements by transcription factor binding preference motif discovery [125-129] or sequence conservation [130], and experimental techniques such as DNase
I hypersensitive site mapping [7,8,24,25,29-31], FAIRE [7,14,27,31], and ChIP [131-135]. Computational predictions often exhibit high false positive and false negative rates [136,137], and thus the identified candidates need to be validated by experimental methods. The function of each individual element identified by DNase hypersensitivity of FAIRE is unknown and currently often inferred computationally. Data from ChIP experiments for chromatin marks that are specifically associated with promoters, enhancers, insulators or silencers can be used to directly identify these elements. For example, monomethylation of histone H3 at lysine residue 4 (H3K4me1) [29,31,137] and transcriptional coactivator p300 [24,31] binding have proven predictive for enhancer activity of genomic elements. Taken together, although a large number of putative regulatory elements can be identified by existing large-scale approaches, compatible large-scale assays that can functionally test these putative elements need to be developed.

Traditionally, regulatory elements have been functionally characterized one or a few at a time through serial transfection of individual reporters [112,117,118,121-124], but more recently DNA elements have been characterized using systematic large-scale approaches [95,97,98,138]. One such high-throughput functional method was designed to characterize promoter regions in the human genome [138]. In this study, random fragments from sheared genomic DNA were cloned into a GFP-based promoter reporter vector and tested for their ability to drive GFP reporter expression in cell culture. While they identified 858 putative promoters, this approach is highly inefficient as active regulatory elements in each cell type occupy less than 5% of the genome. Therefore most cloned fragments will not be functional. Other studies developed massively parallel reporter assays to test enhancer activities of either a few known human enhancers and a large number of their engineered variants [95,97] or a few thousand predicted enhancers and their variants [98]. Despite the success of those studies, it should be noted that the
tested enhancers were synthesized from known or predicted enhancer sequences. Such synthesis-based approaches are costly and require exact DNA sequences of tested elements, thus it is hard to adapt these to the genome-wide annotation of thousand of native enhancers in a given human cell type of which many are unknown elements.

1.3.2 Enhancers activate gene expression in an orientation-independent manner

Unlike promoters that are typically immediately upstream of genes, enhancers can be distant-acting regulatory elements that enhance transcription levels of genes by acting on their promoters. They can be located upstream, downstream or within their target genes, and can modulate expression independent of their orientation [139,140]. They do not need to be particularly close to the targeted genes, and sometimes are found extremely far from their targets [140-142], or even on a different chromosome [143,144]. Enhancers have been shown involved in many developmental and disease-relevant processes [140,145-147].

1.4 Chromosomal translocations in oncogenesis

1.4.1 Chromosomal translocations are genetic hallmarks of most cancer cells

Chromosomal translocations have been reported in several forms of human cancer [61-64,148-151] and are implicated in approximately 20% of cases of cancer morbidity [150,151]. Translocations can lead to tumorigenesis by activating oncogenes [151-154], creating fusion transcripts [151,153,155-157], or disrupting normal gene activity resulting in malignant transformation [158,159]. Chromosomal breakpoints can be detected by cytogenetics [155,160]. In addition, advances in sequencing and bioinformatics are uncovering clinically relevant translocations and rearrangements [61-64,161]. Recently, recurrent translocations were discovered in >50% of prostate cancer cases [157,162], giving us new insights into the genetic complexity of this disease.
Chromosomal translocations are genetic hallmarks of most cancer cells, and some translocations are responsible for malignant transformation [61-64,148]. In several types of cancers, translocations are the causative agent of disease by producing or activating oncogenes responsible for malignant transformation [151,153,155-157]. On a molecular level, translocations require the formation of DNA double-strand breaks (DSBs) at two or more genomic loci, followed by the illegitimate joining of broken chromosomal ends through DNA repair. Depending on the chromosome breakpoints, a translocation can result in the disruption or misregulation of normal gene function [158,159]. These molecular rearrangements, in many cases, are considered to be the primary cause of various cancers [61-64]. Indeed, over the past few decades, clinical cytogeneticists have been able to link specific chromosome breakpoints to clinically defined cancers, including subtypes of leukemias, lymphomas, and sarcomas [61-64,148-151]. Virtually all of the translocations observed in tumors have arisen through somatic events, so these are not inherited in families [149-151]. Translocations in some lymphomas and leukemias lead to the juxtaposition of promoter or enhancer elements from one gene with the intact coding region of another gene [151-154]. In other cases such as in Chronic myelogenous leukemia (CML) and many of the acute leukemias, translocations result in the recombination of the coding regions of two different genes [151,153,155-157]. This results in a fusion protein that might have a new function.

1.4.2 The mechanisms by which translocations consistently recur between certain sites of the genome are largely unknown.

Although the pathological relevance of chromosome translocations in human leukemia and lymphoma is well established [75,148,163], the mechanisms by which translocations consistently recur between certain sites of the genome are largely unknown [75,148]
There is increasing evidence that DSBs involved in translocations occur at non-random sites in the genome [75, 148, 163]. First, the spatial arrangement of chromosomes and the gene fusions that result from translocations are tissue- and cell-type specific [75, 148, 163, 165]. Second, breakpoints in common translocations have been cloned and tend to cluster in small regions of introns [75, 148, 163, 166]. Third, studies that mapped the genome-wide landscape of translocations in the absence of selection found that translocations occur primarily between transcriptionally active regions of the genome [75, 148, 161, 163, 164]. These observations suggest that certain regions of the genome are more susceptible to DNA breakage and translocations; however, what precisely defines these sites remains unknown. There has been indication that local chromatin architecture may influence the recurrent translocation frequency [167, 168]. Taken together, despite the well-documented impact of chromosomal translocations in cancer, the central unsolved questions in the field are what determines the sites of chromosomal breakage and what molecular factors influence susceptibility of DNA to breakage.
CHAPTER II:
INCOMPLETE REPROGRAMMING OF DNASE HYPERSENSITIVE SITES IN
HUMAN IPS CELLS IS ASSOCIATED WITH ALTERED REGULATION OF
DEVELOPMENTAL GENES UPON DIFFERENTIATION

2.1 OVERVIEW

Human induced pluripotent stem cells (hiPSCs) are similar to human embryonic stem cells (hESCs) in many respects, but can have altered differentiation potential and DNA methylation patterns. A central unanswered question is how completely chromatin in hiPSCs is reconfigured to resemble that of blastocyst-derived hESCs, and whether any imperfections in the chromatin remodeling process are functionally relevant. To address these questions, we mapped DNaseI hypersensitive (DHS) sites using DNase-seq, which identify regions of nucleosome depletion and transcription factor binding, in seven samples: three human ESC lines, two human iPSC lines, and their two matched parental fibroblasts. Over 95% of DHS sites were shared between hESCs and hiPSCs, but we detected thousands of reproducible differential DHS (dDHS) sites. We show that dDHS sites that are not completely reprogrammed are associated with different levels of active and repressive chromatin marks in the parental cells prior to reprogramming. dDHS sites are significantly enriched for “super-enhancers” that are critical to hESC identity, and up to 9% of dDHS sites are themselves found in clusters, dCOREs, which span tens of kilobases and significantly overlap with previously identified differentially methylated regions (DMRs). Genes associated with dDHS sites are highly enriched for key transcription factors that function in early embryonic development. While RNA levels of these genes are typically indistinguishable in undifferentiated hESCs and hiPSCs, many of these genes
show differential expression in the examined hiPSCs and hESCs upon differentiation into mesendodermal or neuroectodermal lineages. Thus, regulatory elements susceptible to incomplete reprogramming in hiPSCs can be anticipated by the chromatin state in the parental cell, and flaws in regulatory element reprogramming foreshadow gene regulation upon hiPSC differentiation. While we do not claim that the exact set of dDHSs found here exists between any given hESC and hiPSC line, we predict that any set of dDHS sites will share the properties of the set we describe here.

2.2 INTRODUCTION

iPSCs and ESCs are similar in many respects, including cell morphology, expression of pluripotency markers, histone modification patterns, teratoma formation, and ability to differentiate into all germ layers [41-44,47,169-171]. However, there are also many differences [77-80,99-101], and recent studies suggest that some of these differences may be rooted in a failure to fully recapitulate an ESC chromatin state in iPSCs. For example, despite the global similarity of human iPSC and ESC DNA methylomes, aberrant methylation of CpG islands and incomplete restoration of methylation at non-CG mega-regions were observed in hiPSCs [99-101]. The methylation inconsistencies between iPSCs and ESCs are associated with epigenetic memory of somatic progenitors [71-73,102] and de novo methylation aberrations, which appears to impact developmental potentials [69,71-73,102]. Residual somatic DNA methylation patterns and aberrant de novo methylations in iPSCs were transmitted to differentiated cells and contributed to transcription variation upon differentiation [31]. Adding to evidence for the importance of chromatin in iPSC formation, depletion of Mbd3, a core member of the NuRD (nucleosome remodeling and deacetylation) repressor complex results in reprogramming efficiencies near 100% by suppressing NuRD-mediated inhibition of downstream OCT3/4,
SOX2, KLF4 and c-MYC target genes [172]. Thus, understanding the extent to which chromatin is correctly reconfigured at gene regulatory elements is fundamental to understanding how iPSCs are formed, how any underlying differences could lead to changes in transcriptional programs upon differentiation, and how these differences impact developmental and therapeutic potential.

In this study we report genome-wide mapping and characterization of DNaseI hypersensitive (DHS) sites in human ESCs (hESCs), human iPSCs (hiPSCs), and matched parental fibroblast lines. DHS sites represent regions of nucleosome depletion and transcription factor binding to DNA, and therefore is a powerful readout of the regulatory state and potential of a given cell type [6,8,25]. While over 95% of DHS sites were shared between hESC and hiPSC lines, we identified thousands of differential DHS (dDHS) sites. Compared to completely reprogrammed genomic regions, DHS sites that were present in ESCs but were not fully reprogrammed to an open state in iPSCs displayed lower levels of active chromatin marks and higher levels of repressive marks in the parental cell line, indicating a likely mechanism behind sub-optimal reprogramming. Differential DHS (dDHS) sites such as these were often found near each other, forming dCOREs (differential Clusters of Open Regulatory Elements), and also occurred preferentially in “super-enhancers”, which were recently identified in hESCs and shown to be important for ES cell identity and function [31]. Clusters of dDHS sites were also significantly enriched around key developmental genes, including $SOX, ZIC, PAX, HOX, FOX,$ and $HMX$ gene family members. Interestingly, the transcriptional output of those genes was indistinguishable between undifferentiated hESCs and hiPSCs, but after differentiation towards mesendodermal or neur ectodermal lineages, many such genes displayed differential expression profiles. Our study shows that chromatin sites in somatic cells that are resistant to reprogramming have characteristic chromatin features, and that these sites of incomplete
reprogramming in iPSCs foreshadow differences in regulation of gene expression upon differentiation. DNase-seq screening of undifferentiated hiPSC lines may identify those that are unable to efficiently differentiate into specific somatic cell lineages and reveal new insight into the genetic and epigenetic mechanisms that regulate early human development.

2.3 RESULTS

2.3.1 Human ESCs and iPSCs have very similar accessible chromatin profiles

To assess the completeness of chromatin reconfiguration during the reprogramming of human somatic cells, we examined the landscape of DNaseI hypersensitive (DHS) sites across the genome in seven samples: three human embryonic stem cell (hESC) lines (H1, H7, and H9), two human induced pluripotent stem cell (hiPSC) lines, and their two matched parental fibroblasts (Figure 2.1.A-C). To ensure reproducibility, three biological replicates were analyzed from each line (two replicates of H1; see Methods). DHS sites mark most transcription factor binding sites and overall chromatin accessibility [6,8,25], and therefore provide a powerful indicator of global chromatin reprogramming. We also measured RNA levels in all samples using Affymetrix human exon arrays.

As expected, the accessible chromatin profiles of hiPSCs closely resembled those of hESCs, but were distinct from their parental fibroblasts (Figure 2.1.B). For example, three DHS sites surrounding NANOGEN, a key transcriptional regulator of pluripotency, were shared only by hESC and hiPSC lines, while a separate DHS site near NANOGEN was present only in fibroblasts (Figure 2.1.C). Among the 79,190 DHS sites that were common to all hiPSC lines and the 53,250 DHS sites that were common to all hESC lines, 48,870 sites were shared between them (Figure 2.1.B). Of these, 29,063 DHS sites were also present in parental fibroblasts, leaving 19,807 DHS sites that were unique to hESCs and hiPSCs. Using the same criteria, only 8,220
sites were shared between hiPSCs and their fibroblast progenitors and were not found in hESCs. This indicates that the chromatin of hiPSCs is far more similar to hESCs than to their cell type of origin.

We also examined similarity among the samples by analyzing the DHS signal intensity using principle component analysis (PCA) (Figure 2.1.D) and hierarchical clustering (Figure 2.1.E). Similar analyses were performed using RNA expression data (Supplemental Figure 2.1.A-B). In each analysis, the hESC lines grouped together with the hiPSC lines in both chromatin and gene expression space and were well separated from the parental fibroblast lines. This analysis confirmed that the genome-wide chromatin landscapes of hiPSCs are more similar to those of hESCs, and both are comparatively different from the landscapes of the parental fibroblast cells. This general similarity in DHS sites between hiPSCs and hESCs is consistent with what was previously observed using alternate methods such as DNA methylation and histone profiling [88,99,101,173].

2.3.2 Approximately 4% of DHS sites are differentially accessible between hESCs and hiPSCs

Differentially methylated regions exist between ES and iPS cells [99,101]. Some of these regions reflect methylation marks that were not erased from somatic cells during the reprogramming process (called “memory” marks), while others were de novo methylation events specific to iPSCs.

To identify differential DHS sites between hESCs and hiPSCs, we first identified DHS sites present in all (intersection) samples from one cell type but that were missing in any (union) of
Figure 2.1. Experiment description and data overview. A) Experimental design. DNase-seq and expression analysis was performed on three human ESC lines, two skin fibroblast lines, and two iPS lines that were derived from the skin fibroblasts. Human iPS cell lines were generated with pMX-based retroviral vectors carrying OCT4, SOX2, KLF4, and MYC [41]. Three biological replicates were analyzed for each line. B) Diagram showing number of DHS sites that were shared between human ES, fibroblast, and iPS cell lines. Note that hiPSC and hESC DHS sites are highly similar, and the presence of memory DHS sites shared between ESC and
fibroblasts. C) Representative DNase-seq data (left) from all three cell types shows data is of high quality. Region surrounding **NANOG** locus (right) shows DHS sites that are similar between hiPSC and hESC lines (orange boxes), and were different to DHS sites found only in fibroblast lines (gray box). Biological replicates (R1, R2, and R3) were shown for each cell line. A DHS sites downstream of **NANOG** was detected in all cell lines. The y-axis for each DNase-seq dataset was fixed at a constant normalized read depth. D-E) Replicates of hiPSC lines clustered with replicates of hESC lines and were separated from their parental fibroblast lines in chromatin space by both PCA clustering (D) and hierarchical clustering (E). (See also Figure S1)

the other two cell types (**Supplemental Figure 2.2**). We chose this approach over our initial comparisons (described above) to be more conservative and to exclude differential sites that might arise due to thresholding artifacts. For DHS sites that are present in both hESCs and hiPSCs, we also identified those with statistically different signal intensities using the R package edgeR [172,174,175] (see **Methods**). Considering all 158,393 DHS sites that represent the union set of DHS from hESCs and hiPSCs, together these methods identified 5,799 differential DHS (dDHS) sites (~4%). These dDHS were grouped into 3 categories (**Figure 2.2A**): (1) hESC dDHS sites representing 2,600 DHS sites that were more open in hESCs than in hiPSCs and not open in the parental fibroblast lines (**Figure 2.2B**), (2) hiPSC dDHS sites representing 2,221 DHS sites that were more open in hiPSCs than in hESCs and not open in the parental fibroblast lines (**Figure 2.2C**), and (3) Memory dDHS sites representing 978 DHS sites that were shared by all hiPSC lines and all fibroblast lines but not present in any of the hESC lines (**Figure 2.2D**) (see **Table S1** for the complete list). Given the variability in DNA methylation recently detected across many hiPSCs [88], we do not expect that the dDHS sites identified with the particular lines used in our study represent a universal “signature” that is shared among all hiPSCs and hESCs. Analysis of additional lines will likely identify a different but partially overlapping set of dDHS sites, which we predict will share the same features as the set we describe below for the reasons explained below.
Figure 2.2 Approximately 4% DHS sites are differentially DHS (dDHS) between hESCs and hiPSCs. A) Genome-wide analysis identifies 5,799 hESC, hiPSC, and memory dDHS sites. Specifically, we identified 2,600 hESC dDHS sites that were identified only hESCs or were more open in hESCs, 2,221 hiPSC dDHS sites that were identified only in hiPSCs or more open in hiPSCs, and 978 memory dDHS sites that were open in both iPSCs and their parental fibroblasts but not in hESCs. B) A cluster of hESC dDHS sites (gray box) at PTPRN2 locus. C) Example of hiPSC dDHS site upstream of GLYATI1 gene. D) Example of memory dDHS site at TSS of TRIM4. (See also Supplemental Figure 2.2)
2.3.3 dDHS sites often map in clusters

We noticed a large cluster of dDHS sites at the PTPRN2 gene locus that span a 1Mb genomic region (Figure 2.2B). In this region, 69 of the 72 DHS sites were identified as hESC dDHS sites, while none of these were hiPSC dDHS sites. It has been previously suggested that clusters of dDHS in close proximity might act together to influence gene expression or other chromosomal functions [7,27]. To systematically determine whether dDHS sites tended to cluster together in the genome, we plotted the location of all dDHS sites on individual chromosomes and observed that clusters of dDHS sites grouped together in a non-random manner (Supplemental Figure 2.3A). For example, we found large domains of similarly defined hESC dDHS sites (Figure 2.3A, red arrow), hiPSC dDHS sites (Figure 2.3A, green arrow), and memory dDHS sites (Figure 2.3A, blue arrow).

To provide a more detailed characterization of dDHS domains, we next analyzed the distance between each pair of dDHS sites within each of the three categories. A significant percentage of both hESC dDHS sites (8.5%) and hiPSC dDHS sites (5.9%) were within 10kb of another hESC dDHS site or hiPSC dDHS site, respectively (Figure 2.3B). In contrast, we found very few pairs of memory dDHS sites (0.3%) within 10kb of each other. Permutated dDHS sites rarely (< 0.3%) clustered together (see Methods). We thereby defined a dCORE (differential Clusters of Open Regulatory Elements) as a stretch of dDHS sites in which each dDHS site was within 10kb of another dDHS site of the same type. In this manner, 63 hESC dCOREs, 37 hiPSC dCOREs, and 1 memory dCORE were identified (Figure 2.3C). In contrast, only 2 clusters were identified when considering sets of permutated dDHS sites (see Methods). dCOREs were detected throughout the genome (Supplemental Figure 2.3B), and the sizes ranged up to 63kb, with the largest hESC and hiPSC dCOREs consisting of 9 dDHS sites each. Given the large
number of hESC and hiPSC dCOREs relative to clusters of permuted dDHSs sites, these
dCOREs likely represent functionally linked differences in chromatin states between the two cell
types.

Figure 2.3. dDHS sites occurred in clusters. A) Distribution of hESC, hiPSC, and memory
dDHS sites on a representative locus on chromosome 8. Note that dDHS sites appear to cluster
by dDHS type. B) Distribution of dDHS sites for each chromosome (top) and relative numbers of
hESC dDHS vs. hiPSC dDHS (bottom). C) A higher than expected proportion of hESC dDHS
sites and hiPSC dDHS sites cluster were within 10kb of their closest hESC dDHS site or hiPSC
dDHS sites, respectively. Most memory dDHS sites did not cluster. D) Clusters of dDHS of the
same type define dCOREs. Each dCORE consists of at least three dDHS sites of the same type
that are within 10kb of another dDHS site. (See also Supplemental Figure 2.3)

To gain insight into the biological significance of dCOREs, we compared them to
previously identified “hotspots” of differentially methylated CG (CG-DMRs) and megabase-
scale non-CG regions (non-CG mega-DMRs) that exhibited aberrant DNA methylation patterns
in hiPSCs relative to hESCs [99,101]. Lister et al. identified 1,175 CG-DMRs that cover
approximately 1.7 MB, and 25 non-CG mega-DMRs (after liftover) that cover approximately 29MB [99]. 20 (32%) of 63 hESC dCOREs and 8 (11%) of 72 hiPSC dCOREs overlapped with CG-DMRs, and 9 (14%) of hESC dCOREs overlapped with non-CG mega-DMRs. The overlaps were statistically significant (p-values < 2.2e-16) given the overall coverage of CG-DMRs and non-CG mega-DMRs. Notably, the PTPRN2 locus described above (Figure 2.2B) contained a non-CG mega-DMR [99], and was differentially methylated at its nearby CpGs in 9 of 17 hiPSC lines relative to 7 hESC lines analyzed by Ruiz et al. [101]. In total, 7 hESC lines were used in the two DNA methylation studies, and 22 individual hiPSC lines derived from 9 distinct parental cell lines with varied reprogramming efficiencies using 3 different protocols. Therefore, we believe that overlapped regions between DMRs highlighted in the DNA methylation studies and dCOREs identified in this study are likely features common to hiPSC lines and are biological significant. It suggests that, despite the variation in the epigenetic landscape among hiPSC lines, incomplete reconfiguration of the epigenetic state is shared by hiPSCs at a set of megabase-sized genomic regions such as dCOREs. As stated above, we predict that analysis of open chromatin in additional hiPSC lines will identify a different but partially overlapping set of dDHS sites, with the overlapping sites frequent in those shared megabase-scale regions.

2.3.4 dDHS sites occur more often than expected in super-enhancers

Large genomic domains containing clusters of transcriptional enhancers termed “super-enhancers” have been recently characterized in various human cell types [176]. Super-enhancers are suggested to drive cell type-specific gene expression patterns and play key roles in the control of cell identity [176-178]. ESC super-enhancers are bound by Oct4, Sox2, Nanog, and mediator, and these regions stimulate higher transcriptional activity than typical enhancers.
To explore whether dDHS sites overlapped super-enhancers, we examined the 684 super-enhancers identified in H1 hESCs [176].

Of all 158,393 DHS sites identified in hESCs and hiPSCs, 1,955 (1.2%) overlapped with 655 (95%) H1 super-enhancers. Using this as our baseline for the expected proportion of overlap, we found that a significantly higher proportion of hESC dDHS sites (n=75, 2.8%; p = 2.33e-10) overlapped with H1 super-enhancers (n=65, 9.5% of all H1 super-enhancers). A higher proportion of hiPSC dDHS sites (n=42, 1.9%; p-value 0.0058) also overlapped with H1 ESC super-enhancers (n=32, 4.7%). While these numbers may seem small on an absolute scale, this is to be expected given the small number and genomic coverage of super-enhancers. In fact, these overlapping dDHS/super-enhancer sites have outsized significance not only statistically but biologically, occurring near key developmental genes such as SOX2, SOX2-OT (Supplemental Figure 2.6), ZIC2, ZIC5 (Supplemental Figure 2.7), SOX11, and SOX21 (Supplemental Figure 2.7). Furthermore, 6 of 63 ESC dCOREs (9.5%) and 2 of 37 iPSC dCOREs (5.4%) overlapped with H1 super-enhancers. In contrast, fewer than expected memory dDHS sites (n=2, 0.2%; p-value 7.20e-05) overlapped with H1 super-enhancers (n=2, 0.3%).

We also compared dDHS sites to super-enhancers identified in fibroblast lines NHDF-Ad, NHLF, and IMR90 [176]. Approximately 5% of memory dDHS sites overlapped super-enhancers for all three cell types, which was a significant enrichment (p-values < 3e-07) over that of all hESC and hiPSC DHS sites (~1.2%). The single memory dCORE did not overlap with any H1 super-enhancers, but did overlap with super-enhancers identified in HMEC (human mammary epithelial cells) and 15 other differentiated human cell and tissue types. Together our results show that compared to all DHS sites, ESC dDHS sites are significantly enriched for H1 super-enhancers that are crucial for hESC identity. In addition, memory dDHS sites were found
less often than expected in H1 super-enhancers, but more often than expected in fibroblast super-enhancers.

2.3.5 hESC, hiPSC, and memory dDHS sites display differences in basal chromatin state prior to reprogramming

We asked why chromatin remodeling failed or was incomplete at some genomic regions, while throughout most of the genome chromatin accessibility in hiPSCs was indistinguishable from hESCs. To test whether specific genomic features are associated with impaired reprogramming, we compared the intensity of histone modifications and transcription factor binding at dDHS sites where reconfiguration was incomplete and genomic regions that were successfully reprogrammed. Since reprogramming was initiated from the fibroblast lineage, we used ENCODE ChIP-seq data on the skin fibroblast cell line NHDF-Ad and the primary fibroblast cell line AG04450 [31] to characterize the chromatin status at different types of dDHS sites in fibroblast cells prior to reprogramming. We examined histone modifications or variants associated with active chromatin (H3K9ac, H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K36me3, H3K79me2, H4K20me1, and H2A.Z), repressed chromatin (H3K9me3, H3K27me3, and the poly-comb complex protein EZH2), and the binding of insulator protein CTCF. We compared ChIP signals in fibroblast cells at hESC dDHS sites (Figure 2.4A; Supplemental Figure 2.4A), hiPSC dDHS sites (Figure 2.4B; Supplemental Figure 2.4B), and memory dDHS sites (Figure 2.4C; Supplemental Figure 2.4C) to their relevant control regions (see Methods).

Compared to DHS sites that were successfully reprogrammed, hESC dDHS sites had decreased levels of the active chromatin marks H3K9ac (p-value < 0.001), H3K27ac (p-value < 0.001), H3K4me2 (p-value < 0.002), H3K4me3 (p-value < 0.008), and H2A.Z (p-value < 0.001), and
increased levels of all repressive marks analyzed (p-value < 0.02) (Figure 2.4A; Supplemental Figure 2.4A). This indicates that hESC dDHS sites that failed to reprogram were in a more closed chromatin state in fibroblasts, possibly resulting in less accessibility to transcription factors or chromatin remodelers that are required to reprogram the chromatin state of fibroblasts. The trend was consistent for ChIP-seq data from both the NHDF-Ad and primary AG04450 fibroblast cell lines (Supplemental Figure 2.4A). This observation is consistent with the previous report of OSKM-Differentially Bound Regions (OSKM-DBRs), which are refractory to binding of pioneer reprogramming factors Oct4, Sox2, Klf4, and c-Myc (O, S, K, and M) [179]. Soufi et al. showed that H3K9me3 is the dominant chromatin feature at these OSKM-DBRs and is an impediment to initial OSKM Binding. Indeed, 218 (8%) of 2,600 hESC dDHS sites were found in OSKM-DBRs.
Figure 2.4 dDHS regions show differences in local chromatin environment in differentiated somatic cells prior to reprogramming compared to controls. ChIP-seq data from NHDF-Ad fibroblast cell line was compared to regions identified as hESC dDHS (A), hiPSC dDHS (B) and memory dDHS (C). For each panel, CTCF is shown on the left, followed by active histone marks in the middle, and repressive marks to the right. A) hESC dDHS (blue line) were compared to control DHS sites (black line) that were present in hiPSCs and hESCs, but not
fibroblasts (see gray inset for pictorial representation). Direction of significant enrichment is shown by red (higher) and green (lower) arrows. Asterisks indicate significant p-values based on random permutation analysis (* p-value < 0.05; ** p-value < 0.005). Black arrows indicate trends that did not reach significance. Relative to control sites, hESC dDHS sites had lower levels of active histone marks and higher level of repressive histone marks. B) hiPSC dDHS (blue line) were compared to control DHS sites (black line) that were identified in a previous study but are not present in any of the three cell types. Relative to control sites, hiPSC dDHS sites had higher levels of both active histone marks and repressive histone marks. C) Memory dDHS (blue line) were compared to control DHS sites (black line) that were only present in fibroblasts. Relative to control, memory dDHS sites had higher levels of active histone marks. (See also Supplemental Figure 2.4)

To our surprise, hiPSC dDHS sites displayed higher levels of active H3K4me2 chromatin marks (p-value < 0.01), repressive EZH2 and H3K27me3 chromatin marks (p-values < 0.001), and insulator protein CTCF (p-value < 0.001) in ChIP-seq data from NHDF-Ad fibroblast cells (Figure 2.4B; Supplemental Figure 2.4B). Higher levels of both active chromatin marks (H3K27ac and H3K4me3) and repressive chromatin marks (H3K27me3) were also observed in AG04450 ChIP-seq data (Supplemental Figure 2.4B). For this comparison, hiPSC dDHS sites were measured against genomic regions previously identified as a DHS site in at least one of >100 samples from various cell types [180] but that did not represent a DHS site in hESCs, hiPSCs, and fibroblasts (see Methods). In other words, the hiPSC dDHS sites, which were closed in somatic cells and then inappropriately opened during reprogramming, had higher levels of either repressive or active histone marks in the starting somatic cell. This conclusion was made in comparison to regions of closed chromatin in somatic cells that remained properly closed during reprogramming. The presence of either chromatin state in fibroblasts may make these regions more amenable to chromatin remodeling than other regions of closed chromatin.

Compared to fibroblast DHS sites that were successfully closed in iPSCs, memory dDHS sites had higher levels of H2A.Z (p-value < 0.001) and H3K4me2 (p-value 0.03) active chromatin marks, higher levels of insulator protein CTCF (p-value < 0.001), and slightly higher
level of EZH2 (p-value 0.02) in ChIP-seq data from NHDF-Ad fibroblast cells (Figure 2.C; Supplemental Figure 2.4C). Similar trends were observed in ChIP-seq data from the primary fibroblast cell line (Supplemental Figure 2.4C). This indicates that compared to DHS sites that closed during reprogramming, the memory dDHS sites that were more resistant to being reconfigured to a closed state harbored more active chromatin marks prior to reprogramming.

2.3.6 hESC and hiPSC dDHS sites contain KLF4 motifs, and most memory dDHS sites contain an AP-1 motif

To test whether the identified dDHS sites contained unique sequence features we performed de novo motif analysis on each class of dDHS sites. The motif for KLF4 was detected in both hESC dDHS sites and hiPSC dDHS sites (Supplemental Figure 2.5A), indicating a possible role for one of the four reprogramming factors in over- or under-reprogramming certain regulatory elements. In addition, the DNA motif for AP-1 was significantly enriched in memory dDHS sites (Supplemental Figure 2.5A). AP-1 is a transcription activator complex composed of Fos and Jun family members, and has been shown to function as a pioneer factor that facilitates access to chromatin [181]. Approximately 70% of all memory sites contained an AP-1 motif. In addition, 55% of FOSL2 and 44% of JUND ChIP-seq sites overlapped with memory sites (see Methods), which were significantly higher than the < 10% overlap of FOSL2 and JUND ChIP-seq sites with hESC dDHS sites and hiPSC dDHS sites (p-value < e-100). The enrichment of the AP-1 motifs and binding sites in memory dDHS suggests that AP-1 may contribute to the persistence of open chromatin at those loci.
2.3.7 Clusters of differential DNase Hypersensitive sites are connected to loci encoding key developmental transcription factors

The observation that dDHS sites tended to occur in clusters suggests that they might cooperatively regulate the transcription of a nearby gene. Only a small fraction (7%) of dDHS sites was at or near gene transcription start sites (TSS), indicating most dDHS sites represent distal regulatory elements such as enhancers (Supplemental Figure 2.5B). To connect all DHS and dDHS sites to their target genes, we used a previously developed methodology based on correlation between the DHS site signal intensity and the expression level of a gene across 112 human cell line and tissue types, which incorporates data from hESC and hiPSC lines [180]. Using this connection matrix, we found 45,143 of the 158,393 DHS sites from either hESCs or hiPSCs were connected to 24,529 genes (out of 36,638 total coding and non-coding genes on the Affymetrix exon array), representing a total of 96,383 connections (some individual DHS sites were connected to multiple genes, and some genes were connected to multiple DHS sites). Among the connected DHS sites, 2,483 (5.5%) were dDHS sites, which were involved in 5,734 connections. The median connectivity of dDHS sites was 2 genes, which was the same as that of other hESCs or hiPSC DHS sites.
Figure 2.5. dDHS sites clustered around developmental genes. A) List of the top 100 genes that were most connected to hESC, hiPSC, or memory dDHS sites. Genes were divided by different gene ontology categories and rank ordering for each gene is indicated in parentheses. B) hESC dDHS sites clustered around developmental genes. Genes highlighted in red were also connected to hESC dCOREs. C) hiPSC dDHS sites clustered around a non-overlapping set of developmental genes. Genes highlighted in red were also connected to hiPSC dCOREs. (See also Supplemental Figure 2.5)
We sought to identify genes that were connected to more dDHS sites than expected and thus were potentially differentially regulated in hESCs and hiPSCs. For each gene we counted the number of connected DHS and dDHS sites. To determine enrichment of dDHS connections compared to the background frequency, we calculated FDR-adjusted p-values using the exact binomial test (see Methods). All genes were ranked by their FDR-adjusted p-values, and smaller p-values indicate greater enrichment of dDHS sites connected to a gene. The top 100 ranked genes based on these p-values are listed in Figure 2.5A (see Table S3 for the complete list).

We manually grouped the top 100 genes that were enriched for dDHS site connections into six general categories based on GO term annotations (see Methods, Figure 2.5A). The most prominent feature of this list was 25 genes that were transcription factors and other key regulators of early embryonic development. This includes SOX2, one of the four reprogramming factors that is also critical in early neural differentiation, and PAX6, a master transcription factor that is important for the development of the eye, nose, central nervous system and pancreas. Additional developmental genes represented in the top 100 that were enriched for dDHS sites were HOXA, ZIC, FOXA and HMX gene family members (Supplemental Figure 2.6, Supplemental Figure 2.7).

2.3.8 hESC and hiPSC dDHS sites are connected to distinct sets of key developmental genes

Each of the top 100 connected genes appeared to be enriched for either hESC dDHS sites or hiPSC dDHS sites but not both (Fig. 5B-C). For example, of the 9 dDHS sites mapped in the ~110 kb HOXA gene locus, all were comprised of only hESC dDHS sites. Similar enrichment for hESC dDHS was found at SOX2, ZIC2 and ZIC5 (Fig. 5B), while SOX1, SOX5 and ZIC3 genes were enriched for hiPSC dDHS sites (Fig. 5C). To test whether hESC or hiPSC dDHS sites were both ever observed in the same gene, we repeated the exact binomial test described
above on hESC, hiPSC, and memory dDHS sites separately. We found 78 genes enriched for hESC dDHS sites, 193 genes enriched for hiPSC dDHS sites, and 344 genes enriched for dDHS memory sites (FDR-adjusted p < 0.05) (see Table S4 for the complete list). Only 2 genes were held in common between the hESC and hiPSC enriched gene sets, indicating that each class of dDHS generally associates with a distinct set of genes. Similarly, hESC dCOREs and hiPSC dCOREs were connected to different sets of developmental genes (Fig. S3C-F; see Table S5 for the complete list), including several of the genes enriched for hESC dDHS and hiPSC dDHS sites (Fig. 5B-C, highlighted in red). Consistent with this, each class of dDHS site were functionally distinct (Table S6). hESC dDHS-associated genes were strongly enriched for GO categories related to development. In contrast, genes enriched for hiPSC and memory dDHS sites were not significantly associated with specific GO terms. Thus, DNase sites that are open in ESCs but not iPSCs tend to be connected to genes with developmental functions, while DNase sites that are open only iPSCs but not ESCs are not functionally coherent.

2.3.9 Genes highly connected to dDHS sites are differentially regulated upon differentiation

Previous studies have shown that RNA transcript levels in hESCs and hiPSCs are very similar [41,88]. We also found that overall transcript abundance in hESCs and hiPSCs was highly concordant based on our exon array data (R = 0.98, Pearson correlation). In addition, all 494 genes that were highly connected to any type of dDHS site had less than two-fold differences in expression between hESCs and hiPSCs (Figure 2.6A). This could suggest that the dDHS sites between hESCs and hiPSCs do not have major functional consequences. However, we noticed that developmental genes located near dDHS sites were not appreciably expressed in either pluripotent cell type. Therefore, we wondered whether functional consequences of dDHS sites are manifested only after the cells are induced to differentiate into somatic cell derivatives.
We analyzed published expression data from the same hESC and hiPSC lines used in our DHS analysis that were differentiated into mesendodermal and neurectodermal lineages [182]. For hESCs, we identified 2,084 and 6,509 genes that were significantly up- or down-regulated during ectodermal or mesendodermal differentiation (see Methods). We then divided these genes into four categories: hESC dDHS site enriched genes; hiPSC dDHS site enriched genes; memory dDHS site enriched genes; and genes not enriched for dDHS sites (see Figure 2.2A). We found 20, 22, 23, and 2,019 genes in each group, respectively, upon neurectodermal differentiation, and 42, 69, 72, and 6,326 genes, respectively, upon mesendodermal differentiation.
Figure 2.6. Incomplete chromatin reconfiguration may foreshadow altered gene regulation upon differentiation. A) Global expression is highly similar between hiPSC and hESC lines (R=0.983). The top 100 genes that were most connected to dDHS sites are indicated as involved in development (red) and other gene ontology categories (yellow). B) Differentially expressed genes were identified in hESCs during in vitro differentiation into neuroectoderm (top) and mesendoderm (bottom). For differentially expressed genes surrounding hESC dDHS sites...
(n=20), these genes displayed significantly lower fold changes in gene expression during hiPSC differentiation compared to hESC differentiation (red box plots, p-values indicated). Differentially expressed genes around hiPSC dDHS (green boxes), memory dDHS (blue boxes), or non-dDHS sites (white boxes) do not display differences in fold expression changes. C) Comparison similar as (B) but based on genes that are differentially expressed during hiPSCs differentiation into neuroectoderm (top) and mesendoderm (bottom). Genes enriched for hiPSC dDHS (green boxes) and memory dDHS (blue boxes) were improperly regulated in hiPSCs during neurectodermal differentiation, but not mesendodermal differentiation. (See also Figure S6, Figure S7)

To test whether genes connected to hESC dDHS (sites only open in ESCs) were differentially regulated during hiPSCs differentiation, we compared the fold change in RNA levels of these genes during mesendodermal and neurectodermal differentiation between hESCs and hiPSCs (Figure 2.6B). We found that genes enriched for hESC dDHS (sites open only in hESCs not hiPSCs) had significantly lower fold-changes in hiPSCs compared to hESCs during both neurectodermal (p-value 0.0133) and mesendodermal (p-value 0.0019) differentiation (red bars, Figure 2.6B and examples in Supplemental Figure 2.7A). Genes connected to hiPSC dDHS sites, memory dDHS sites, and those not significantly connected to dDHS sites had similar changes in RNA abundance in hiPSCs and hESCs.

Conversely, we tested whether genes connected to hiPSC dDHS sites (sites open only in hiPSCs) would be upregulated during differentiation of hiPSCs (Figure 2.6C). We identified 1,130 and 4,151 differentially expressed genes upon neurectodermal or mesendodermal differentiation in hiPSCs (Figure 2.6C). As above, we divided genes into four categories: hESC dDHS site enriched genes; hiPSC dDHS site enriched genes; memory dDHS site enriched genes; and genes without enrichment for dDHS sites (see Figure 2.2A). We found 8, 4, 13, and 1,105 genes in each group, respectively, upon neurectodermal differentiation, and 20, 46, 42, and 4,043 genes, respectively, upon mesendodermal differentiation. We calculated gene expression fold change in each group upon differentiation in hESCs and hiPSCs (Figure 2.6C). Despite the
small numbers per group, as predicted by our hypothesis, genes enriched for hiPSC dDHS sites and memory dDHS sites displayed a greater fold change in expression during neurectodermal differentiation in hiPSCs compared to hESCs (Figure 2.6C, green and blue bars and examples in Supplemental Figure 2.7B). In contrast, genes associated with hESC dDHS sites and genes without dDHS enrichment showed similar changes in transcription levels in hiPSCs and hESCs. Interestingly, during mesendodermal differentiation, transcript level changes were similar in hESCs and hiPSCs for all groups of genes, including those enriched for hiPSC dDHS or memory dDHS sites (Figure 2.6C, bottom). This suggests that hiPSC dDHS sites may influence regulation during differentiation into some somatic lineages, but not others.
Figure 2.7. Incompletely reprogrammed DHS sites around GABR locus foreshadows altered expression after *in vitro* differentiation.  A) Locus on chromosome 4 (chr4:46,039,200-47,216,540) that harbors DHS sites which are specifically missing in hiPSC line NIHi7 at promoters of *GABRA2*, *GABRA4*, and *GABRB1* genes. B) Expression data
showing that these genes are not upregulated in NIHi7 (red box) after mesendodermal differentiation. Other pluripotent cell lines differentiated into mesendodermal lineage express higher levels of these genes. C) Expression data showing no change in expression for these genes in any hiPSC or hESC after differentiation towards neuroectoderm. D) Schematic diagram depicting hESC dDHS sites (gray triangles) having a functional impact in individual hiPSC lines on gene expression of key developmental genes only after differentiation. While this diagram indicates a deficiency in a single cell lineage (gray box), analysis of additional cell lineages (dotted line) will be needed to understand the full impact of dDHS sites.

In addition to differentiation lineage-specific differences between hiPSCs and hESCs, we also detected lineage-specific differences between the two hiPSC lines. For example, differential chromatin accessibility was observed in only one of the hiPSC lines (NIHi7) surrounding the GABR locus. DHS sites were present at the TSSs of GABRA2, GABRA4, and GABRB1 in all hESC lines and hiPSC line NIHi11, but were missing in hiPSC line NIHi7 (Figure 2.7A). During mesendodermal differentiation, the expression of these three genes were up-regulated in the hESC cells lines and NIHi11, but was not changed in NIHi7 (Figure 2.7B). In contrast, during neurectodermal differentiation these genes exhibited no change in expression for any of the cell lines (Figure 2.7C). Thus, genes lacking appropriate DHS sites in specific hiPSC lines exhibit altered gene regulation that are only detected upon differentiation into certain lineages. This supports our hypothesis that a lack of DHS sites at a given locus in all or a subset of individual hiPSCs can foreshadow improper gene regulation during in vitro differentiation (Figure 2.7D).
2.4 DISCUSSION

Here we report a genome-wide comparison of chromatin accessibility between human ESCs, fibroblasts, and iPSCs derived from these fibroblasts. While DHS sites of reprogrammed hiPSCs are remarkably similar to hESCs, approximately 4% of DHS sites exhibited differential accessibility. These dDHS sites are associated with genes that include key developmental factors. Despite the high transcriptional concordance in undifferentiated cells, we observed that some of these dDHS-enriched developmental genes were inappropriately regulated during differentiation (see Figure 2.7D). Our data are consistent with the hypothesis that chromatin landscape required for gene regulation in somatic lineages pre-exists in undifferentiated pluripotent stem cells [183].

In human iPSC studies, especially those for their potential usage in regenerative medicine, it is critical to screen for and use fully reprogrammed iPSC clones. However, typically only a small fraction (less than 1%) [184] of transfected cells turn into fully reprogrammed iPSC clones. To date, immunohistochemistry, transcription profiling, and DNA methylation assays at promoters of key pluripotent markers are routinely used for primary validation and characterization of iPSC lines. Our study shows that the different chromatin architecture in iPSCs may have little impact on gene transcription at pluripotent states but can significantly influence the regulation of gene transcription including some key developmental factors upon differentiation. We do not expect that every hiPSC and/or hESC line will have the exact same set of “signature” dDHS sites. Instead, each pluripotent line will likely have its own set of chromatin differences that share characteristics similar to the ones described here. Therefore, evaluation of open chromatin configurations may be useful for identification of iPSCs with specific differentiation potentials.
2.4.1 Differential DNase sites occur in clusters, an arrangement of regulatory elements seen at key lineage-specific genes across cell types

We found 9% and 6% of hESC and hiPSC dDHS sites, respectively, in clusters (“dCOREs”) that span tens of kilobases. Similar observations of clustered regulatory elements have been reported in pancreatic islets [27]. More recently, blocks of active regulatory element super-enhancers that were typically >10kb in length were described in mouse and human ESCs, as well as in 86 diverse human cell types [176-178]. Approximately 10% of ESC dCOREs overlapped with super-enhancers identified in H1 hESCs. Therefore dCOREs likely represent a regulatory paradigm that exists across different cell types and is specifically associated with genes that control and define cell identity. The presence of dCOREs suggests that reprogramming of chromatin architecture at individual elements within these clusters may occur in a coordinated fashion. In addition, given the cell type-specific nature of super-enhancers, we speculate that the dCOREs in different hiPSC lines may vary depending on their cell types of origin. Future studies that characterize dDHSs and dCOREs between iPSC lines generated from different cell types can facilitate the understanding of the process and mechanism of chromatin reprogramming and potentially help improve the reprogramming efficiency and fidelity.

2.4.2 Complete chromatin reconfiguration at a given locus varies according to the local chromatin environment.

Compared to other reprogramming processes such as germ cell reprogramming, nuclear transfer, and cell fusion, iPSC generation has a much lower efficiency and a higher error rate [185]. While the mechanism by which reprogramming is achieved during iPSC induction is not yet fully elucidated, in addition to the sequence-specific transcription factors used to make the iPSCs, nucleosome organization and histone modification status have been implicated. During
the first 48 hours of induction of the Yamanaka reprogramming factors OCT3/4, SOX2, KLF4 and e-MYC (OSKM), O, S, and K act as pioneer transcription factors and bind to DNase1 resistant distal elements that lack covalent histone modifications [179]. These distal elements subsequently gain the H3K4me2 mark during reprogramming as the somatic enhancer signature is transformed into the ESC signature [186]. These events precede promoter activation and large-scale transcriptional changes that ultimately delineate somatic cells from iPSCs.

It is therefore important to identify and characterize determinants of differential DHS accessibility between iPSC and ESC to truly understand the nature of reprogramming, the characteristics of iPSCs, and to ultimately improve the efficiency and fidelity of in vitro reprogramming. We showed that the complete reconfiguration of chromatin structure might depend on the local chromatin environment of the starting cell population. This supports reports showing that chromatin modifying enzymes or drugs improve the efficiency of iPSC induction. For instance, addition of the chromatin-associated transcriptional repressor UTF1 to conventional Yamanaka factors (OCT3/4, SOX2, KLF4 and C-MYC) significantly improved the formation of iPSC clones [187]. In addition, either treatment with inhibitors of DNA methyltransferase (DNMT), or knocking down the MBD3 subunit of the NuRD repressor complex increased reprogramming efficiency [172,174]. It is possible that optimizing combinations of these and other chromatin remodelers will further improve the efficiency and fidelity of iPSC generation.
2.4.3 Chromatin that does not properly open or that remains improperly closed in iPSCs harbors features inherently resistant to reprogramming

Both hESC dDHS sites and memory dDHS sites are regions at which the chromatin accessibility in hiPSC lines mimicked their progenitor lines and differed compared to hESC lines. It is possible that those genomic domains are generally more resistant to chromatin reconfiguration during reprogramming. It is not clear whether these refractory sites contribute to the low efficiency of iPSC generation in particular, or whether these loci are also problematic in other somatic reprogramming approaches.

Regions that are refractory to pioneer reprogramming factor binding have also been identified [179]. These megabase-sized regions are enriched for H3K9me3 and termed OSKM-Differentially Bound Regions (OSKM-DBRs). In the small proportion of somatic cells that reprogram following OSKM induction these OSKM-DBRs become accessible. Our study showed that genomic regions not reprogrammed to match the DHS profiles in hESC were enriched for repressive H3K9me3 and EZH2 in somatic skin fibroblasts. Therefore, a subset of dCOREs identified by DNaseI hypersensitivity may reflect OSKM-DBRs.

The presence of specific sequence motifs at dDHS might provide a mechanistic link between site-specific reprogramming efficiency and transcription factors that act as drivers of differentiation and de-differentiation. Indeed, we have found over-representation of the KLF4 motif at both hESC and hiPSC dDHS sites possibly indicating a direct role of Yamanaka factors themselves in failed reprogramming. In contrast, memory dDHS sites were enriched for AP-1 motifs suggesting that persistent binding of AP-1 factors might hinder the chromatin remodeling at these locations.
Genetic differences between the hESC and hiPSC lines we analyzed might also contribute to the differences in chromatin accessibility between these samples. Recent analysis of DHS in 70 lymphoblastoid cell lines has revealed DNase I sensitivity quantitative trait loci (dsQTLs) [188]. The stringent criteria we used to identify dDHS makes it likely that very few of the dDHS regions identified in this study are due to dsQTLs. Future studies using iPS cells from larger numbers of individuals will be important to distinguish the precise contribution of dsQTLs and their impact on reprogramming potential.

2.4.4 Incompletely reprogrammed sites may become functionally relevant only upon differentiation of iPS cells.

Pluripotent stem cells are revealing new insights into the chromatin regulatory mechanisms that control development. DHS sites that define the landscape of differentiated cells are remarkably well represented even before the initiation of differentiation. A recent study that produced DHS maps from ESCs and 48 definitive somatic cell types concluded that definitive cell types share on average 37% of the DHS sites found in undifferentiated ESCs [183]. Therefore, during development, the majority of ESC DHS sites are lost, and smaller numbers of lineage-restricted DHS sites are activated de novo.

In this study we examined a limited number of hESCs and hiPSCs cell lines and found that ~4% of the DHS sites showed differential accessibility between the two cell types. Despite these chromatin differences, transcript levels between hESCs and hiPSCs were almost indistinguishable. It seems that that dDHS sites are not essential for self-renewal, but are instead crucial for accurate gene regulation upon cellular differentiation. Indeed, most of the dDHS-enriched genes or dCORE-connected genes such as ZIC2 and ZIC5, are not involved in pluripotency maintenance but are transcribed and involved in various cellular differentiation
processes. These findings echo results that found similar functions for aberrantly methylated genes [101]. In addition, memory dDHS sites are enriched for fibroblast super-enhancers that may harbor key regulatory elements only for lineage specific gene expression. The incomplete reprogramming at those lineage-specific regulatory elements are likely most problematic for differentiation, not the stem cell state.

2.5 METHODS

**Cell culture.** Human ESC lines H1 (WA01-Feeder Independent), H7 (WA07-Feeder Independent), and H9 (WA09-Feeder Dependent) were obtained from WiCell. Two skin fibroblast cell lines PF43 (AG20443) and PF95 (AG08395) were obtained from Coriell and cultured under recommended conditions using DMEM supplemented with 10% Fetal Bovine Serum, 2mM L-glutamine, and 0.1mM (0.7µl/100ml final media volume) 2-mercaptoethanol, 50 units/ml penicillin, and 50 g/ml streptomycin. The reprogramming and derivation of human iPS cell lines NIHi7 (from PF95) and NIHi11 (from PF43) has been described [189]. H1 and H7 cells were cultured on Matrigel in mTeSR™1 medium. Three replicates of NIHi7, three replicates of NIHi11, and two replicates of H9 hES cells were cultured on a feeder layer of irradiated CF1 mouse embryonic fibroblasts in hES media (DMEM-F12 (Invitrogen), 20% KSR (Invitrogen), 20ng/ml FGF2 (R&D Systems), 2mM L-glutamine, 0.1mM 2-mercaptoethanol, and 1x non-essential amino acids.) A third replicate of H9 hES cells was grown on Fibronectin in hES media.

**DNase-seq.** DNase-seq was performed as previously described [7]. Data was generated for two biological replicates of H1 cells, and three biological replicates of H7, H9, NIHi7, NIHi11, PF43, and PF95 cells. Human Affymetrix exon arrays were processed by Sheffield *et al.*, and used to connect DHS sites with target genes [180].
**DNase-seq data preprocessing and normalization.** Sequencing data were processed and mapped to NCBI build 37/hg19 human reference genome, and DHS sites were mapped using standard pipeline, as described previously [7]. Data are publicly available on the UCSC Genome Browser (http://genome.ucsc.edu) and at the Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo) under accession number GSE32970. DHS sites mapped in each replicates were ranked by their p-values. Top 100K DHS sites from each sample were used for analysis. For each cell type, the number of mapped sequence tags was counted at each base (base count) in each DHS site. Counts were normalized and scaled using the following formula:

Normalized base count = base count x 10^9 / (total number of mapped tags in all DHS x total length of all DHS)

**Microarray protocol and normalization.** Affymetrix Human Exon 1.0 ST microarray data were generated as described by Song et al. [7] from four replicates of H1 cells, three replicates of H7, H9, PF43, and PF95 cells, and two replicates of NIHi7, NIHi11. Exon arrays were processed and normalized by Sheffield et al. [180]. Data are publicly available on the UCSC Genome Browser (http://genome.ucsc.edu) and GEO under accession number GSE15805. For each gene, the average expression level in all hESC samples and hiPSC samples were calculated separately. Average expression levels in hESCs and hiPSCs were plotted using function smoothScatter in R Bioconductor Package geneplotter (v1.38.0) in **Fig. 6A**.

**Overall comparison of DHS sites from hESCs, hiPSCs, and fibroblasts.** Intersect sets of top 100K DHS sites from replicates of hESCs (53,250 intersect sites), hiPSCs (79,190 intersect sites), and fibroblasts (84,263 intersect sites) were generated using BEDTools suite (version 2.16.1) [53]. The intersect sets of DHS sites from three cell types were compared to
each other using BEDTools suite. The results were plotted using Venn Diagram Plotter (http://omics.pnl.gov/software/VennDiagramPlotter.php) as shown in Figure S2.

**Clustering analysis.** A union set of 212,986 DHS sites of top 100K DHS sites from all hESC, hiPSC, and fibroblasts was created using BEDTools suite. The number of mapped sequence tags in each union site was counted, and then normalized by the length of the site and the number of all mapped tags in each sample. Normalized tag counts were used to cluster samples in the chromatin space. RNA expression levels measured on Affymetrix exon arrays were used to cluster samples in the transcription space. Principle component analysis (PCA) and hierarchical clustering were performed using function prcomp and hclust available in the R statistical package (http://www.r-project.org). Hierarchical clustering results were visualized using function ColorDendrogram available in R package “sparcl” [190].

**Identification of differential DHS sites.** Union sets of top 100K DHS sites of hESCs (142,412 union sites), hiPSCs (118,089 union sites), and fibroblasts (113,917 union sites) were generated using BEDTools suite. Intersect (defined above) and union DHS sites of hESCs, hiPSCs, and fibroblasts were compared to each other using BEDTools. Specifically, 2,010 dDHS sites were mapped only in hESCs by subtracting hiPSC union sites and fibroblast union sites from hESC intersect sites. Similarly, 1,393 dDHS sites were mapped only in hiPSCs by subtracting hESC union sites and fibroblast union sites from hiPSC intersect sites. 978 memory sites were identified by subtracting hESC union sites from the intersect of hiPSC intersect sites and fibroblast intersect sites. The results were plotted using Venn Diagram Plotter in Fig. S2.

R package edgeR (version 3.0.8) [175] was used to identify DHS sites that had significantly different amplitudes of DNase signal between hESCs and hiPSCs. After creating a union set of 158,393 DHS sites for hESCs and hiPSCs, each union DHS site was divided into
non-overlapping 500bp windows starting from the center of the site. From this analysis, 375,371 500bp windows were obtained. The number of mapped DNase-sequence tags in each 500bp was counted and used as input for edgeR analysis. Windows that had significantly different amplitudes between hESCs and hiPSCs were identified using a cutoff of FDR adjusted p-value 0.05. Adjacent windows that had a significant threshold were merged. Overall, in addition to the 2,010 dDHS sites that were mapped only in hESCs (as described above), 590 loci had consistently higher amplitudes in hESCs relative to hiPSCs. Together 2,600 DHS sites were defined as hESC dDHS sites. Similarly, in addition to the 1,393 dDHS sites mapped only in hiPSCs (as described above), 828 loci had consistently higher amplitudes in hiPSCs than in hESCs. Together 2,221 DHS sites were defined as hiPSC dDHS sites. Idiogram of dDHS sites ([Fig. 3A and Fig. S3A]) was generated using Idiographica (http://www.ncRNA.org/idiographica) [191].

**Identification of dCOREs.** dCOREs were defined as a stretch of dDHS sites among which each dDHS site was within 10kb of another dDHS site of the same type, using function cluster in BEDTools suite. The coordinates of each type of dDHS sites were permutated over the genomic space of the 212,986 union DHS sites from hESCs, hiPSCs, and fibroblasts (defined above). Clusters of permutated dDHS sites were identified in the same way as dCOREs. Idiogram of dCOREs was generated using Idiographica ([Fig. S3B]).

**Comparison of dDHS sites and dCOREs to DMRs, OSKM-DBRs, and super-enhancers.** Genome coordinates of CG- and non-CG mega-DMRs identified by Lister et al. [99], and coordinates of OSKM-DBRs [179] were converted from NCBI build 36/hg18 to NCBI build 37/hg19 human reference genome using UCSC genome browser tool Batch Coordinate
Conversion (liftOver) (http://genome.ucsc.edu/cgi-bin/hgLiftOver). dDHS sites and dCOREs were compared to DMRs, OSKM-DBRs, and super-enhancers using BEDTools suite.

**Histone modifications in dDHS sites in fibroblasts.** Three types of control DHS sites were used to compare to dDHS sites. 1) permuting coordinates of hESC dDHS sites in the genomic space of DHS sites that were shared by hESCs and hiPSCs but not fibroblasts (i.e. DHS sites that were successfully opened during reprogramming), 2) permuting coordinates of hiPSC dDHS sites in genomic regions that did not present a DHS site in hiPSC, hESC, or fibroblasts but were previously identified as a DHS site in at least one of the other samples of the 112 samples examined in the previous study [180] (i.e. regions that faithfully remained closed during reprogramming), and 3) permuting coordinates of memory dDHS sites in genomic space of fibroblast DHS sites that were closed in hiPSCs and hESCs (i.e. regions that was successfully closed during reprogramming). 1000 sets of control sites were generated from 1000 permutations for each type of dDHS sites.

ENCODEx ChIP-seq data of histone modifications H3K9ac, H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K36me3, H3K79me2, H4K20me1, H3K9me3, and H3K27me3, histone variant H2A.Z, and insulator protein CTCF on the skin fibroblast cell line NHDF-Ad and the primary fibroblast cell line AG04450 were downloaded from the UCSC Genome Browser. From each dataset, the number of mapped sequence tags in each dDHS site and each control site was counted and normalized by the length of the site and the total number of mapped sequence tags in that dataset. The average of normalized tag densities in each type of dDHS sites and each set of its corresponding control sites was calculated. The average normalized density of dDHS sites was plotted in red dots in Fig. S4, while the distribution of the average normalized tag densities in 1000 control sets was represented in boxplots. The difference between the average
normalized tag density in dDHS sites and the mean of the average tag densities in its 1000 set of control sites was calculated, and presented in unit of the standard deviation (SD) of the 1000 average normalized densities from 1000 control sets (Fig 4 and Fig. S4). It can be formulized as:

\[
\frac{\text{Average normalized density}_{\text{dDHS}} - \text{mean (Average normalized density}_{1000 \text{ control sets})}}{\text{SD (Average normalized density}_{1000 \text{ control sets})}}.
\]

An empirical p-value was calculated for each comparison:

\[ p\text{-value} = \frac{\# \text{ of control sets that had smaller (or bigger) average normalized density than that of the dDHS sites}}{1000} \]

**Motif analysis and overlap with ChIP-seq data.** DNA sequences of dDHS sites were extracted from the UCSC Genome Browser. *De novo* motif search in dDHS sites was performed using MEME-ChIP (http://meme.nbcr.net/meme/cgi-bin/meme-chip.cgi) [192]. Motifs identified with e-values < e-10 were reported. The frequency of AP-1 motif in memory dDHS sites was scanned using FIMO (http://meme.nbcr.net/meme/cgi-bin/fimo.cgi) [193]. ENCODE ChIP-seq peaks of FOSL2 and JUND in A549 (Human lung adenocarcinoma epithelial cell line) were downloaded from the UCSC Genome Browser. The overlap between A549 FOSL2 and JUND ChIP-seq peaks and dDHS sites was determined using BEDTools suite.

**Genome annotation.** Location of dDHS sites relative to genes was annotated using function annotatePeaks in HOMER (version v3.17) (Hypergeometric Optimization of Motif EnRichment) suite [54] with default UCSC RefGene annotation. Promoters were defined as transcriptional start sites (TSSs) ±1000 bp. Transcription termination site (TTS) regions were defined as from -100 bp to +1kb of TTS.

**Connecting dDHS sites and dCOREs to genes.** To connect all DHS sites including dDHS sites to their target genes, we used a previously developed methodology [180]. The
underlying premise of the method is that if the amplitude of a DHS site was significantly positively or negatively correlated with the expression level of a gene across many cell lines, that DHS site was then considered to be connected to that gene. For each DHS site, all genes with a TSS within 100 kb were examined. To determine correlations between DNase-seq and expression array data, 112 different cell lines and tissues across 72 diverse cell types from ENCODE and Roadmap Epigenomics project were used, including the hESC and hiPSC lines used in this study. The union set of ~2.7M DHS sites from all 112 datasets were connected to 31,429 genes, representing 781,679 connection events at a cutoff of p-value 0.05. In some cases, individual DHS sites were connected to multiple genes, and in some cases multiple DHS sites were connected to the same gene.

45,153 of the 158,393 hESCs and hiPSCs union DHS sites were connected to 24,529 genes using this connection matrix, including 2,483 of the 5,799 dDHS sites. For each of those connected genes, the number of connected union DHS sites and the number of connected dDHS sites were counted. To identify genes that were enriched for dDHS site connections, the following exact binomial test was performed on each gene using R function binom.test:

\[
\text{binom.test(x=\# of connected dDHS sites, n=\# connected union DHS sites, p=2483/45143, alternative = "greater")}
\]

All connected genes were then ranked by their FDR-adjusted p-values from the binomial test. The smaller the p-value is, the more enriched the gene is for dDHS site connections. To confirm the results of the exact binomial test, we also performed permutation-based analysis. Coordinates of dDHS sites were permutated in the genomic space of the 158,393 hESCs and hiPSCs union DHS sites. Permutated dDHS sites were connected to genes using the same connecting matrix. For each gene, the number of connected permuted dDHS sites was counted.
1000 permutations were performed. For each gene, the number of connected true dDHS sites was compared to numbers of connected permutated dDHS sites from the 1000 permutation. An empirical p-value was calculated as:

\[ p\text{-value} = \frac{\# \text{ of permutations in which the } \# \text{ of connected permutated dDHS sites was bigger than the } \# \text{ of connected real dDHS sites}}{1000} \]

The results of simulation studies of the top 20 genes from the binomial test were plotted in Fig. S5C. The permutation based p-values agreed well with the binomial test results (Table S3). We repeated the connection analysis and the bionomial test on each type of dDHS sites separately. hESC dDHS enriched genes, hiPSC dDHS enriched genes, and memory dDHS enriched genes were similarly defined at a cutoff of FDR-adjusted p-value 0.05. Similarly, we connected each type of dCOREs to genes. Genes connected to at least one dCORE were reported.

**Gene ontology analysis.** Gene ontology (GO) terms of the top 100 genes that were enriched for dDHS site connections were extracted using GATHER (http://gather.genome.duke.edu/) [194]. The 100 genes were manually grouped into six general categories based on the keywords in their GO term annotations (Fig. 5A). GO term enrichment analysis (Table S6) of hESC dDHS enriched genes, hiPSC dDHS enriched genes, and memory dDHS enriched genes were performed using DAVID (Database for Annotation, Visualization and Integrated Discovery) Bioinformatics Resources 6.7, (http://david.abcc.ncifcrf.gov/) [195,196].

**Identification of genes that were differentially expressed after mesendodermal or neurectodermal differentiation.** Expression data of undifferentiated and differentiated hESCs H1, H7, H9 and hiPSCs NIHi7 and NIHi11 measured on Agilent human One Color Gene
Expression Oligo microarrays were downloaded from http://stemcelldb.nih.gov [182]. Raw data were quantile normalized than log2 transformed. To identify genes that were significantly up- or down-regulated during mesendodermal or neurectodermal differentiation, we used microarray analysis tool Significance Analysis of Microarrays (SAM) version 4.0 [197] and a cutoff of q-value < 0.05 and fold change >2.

2.6 Data Access

All DNase-seq and expression datasets have been deposited in Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession numbers GSE32970 and GSE15805.
2.7 Supplemental data

Figure 2.8. hiPSC and hESC lines have similar expression profiles. Replicates of hiPSC lines clustered with replicates of hESC lines and were separated from their parental fibroblast lines in expression space by both PCA clustering (A) and hierarchical clustering (B).

Figure 2.9. Identification of hESC-, hiPSC-, and fibroblast-specific DHS sites. A) hESC-specific DHS sites (n=2,010) were identified by taking the intersection of all DHS sites identified from the H1, H7, and H9 cell lines, and subtracting the union set of DHS sites from both hiPSC and fibroblast lines. B) hiPSC-specific DHS sites (n=1,392) were identified similarly using method described in (A). C) fibroblast-specific DHS (i.e. memory dDHS) sites (n=978) were identified by taking the intersection of all DHS sites from hiPSCs and fibroblast lines and subtracting the union set of DHS sites from hESC lines.
Figure 2.10. Differential DHS (dDHS) sites and dCOREs cluster nonrandomly across the genome and are connected to developmental genes. A) Each hESC dDHS, hiPSC dDHS, and memory dDHS site was mapped to the genome. hESC and hiPSC dDHS sites often clustered together in close proximity. The X chromosome contained mostly hiPSC dDHS sites. B) Location of dCORES relative to each chromosome. C) Number of dDHS sites within dCORES and genes that are connected to these dDHS sites. D-F) List of genes that are connected to hESC dCOREs (D), hiPSC dCOREs (E), and memory dCORE (F).
Figure 2.11. The local chromatin environment in differentiated somatic cells prior to reprogramming – AG04450. Similar to Figure 4 in the main text, but ChIP-seq data from AG04450 fibroblast cell line was compared to regions identified as hESC dDHS (A), hiPSC dDHS (B) and memory dDHS (C). For each panel, active histone marks are on the left, while repressive marks are on the right. A) hESC dDHS (blue line) were compared to control DHS sites (black line) that were present in hiPSCs and hESCs, but not fibroblasts (see inset for pictorial representation). Direction of significant enrichment is shown by red (higher) and green (lower) arrows. Black arrows are not significantly different. Relative to control sites, hESC dDHS sites had lower levels of active histone marks and higher level of repressive histone marks. B) hiPSC dDHS (blue line) were compared to control DHS sites (black line) that were not present in any of the three cell types. Relative to control sites, hiPSC dDHS sites had higher levels of both active histone marks and repressive histone marks. C) Memory dDHS sites (blue line) were compared to control DHS sites (black line) that were only present in fibroblasts. Relative to control, memory dDHS sites had higher levels of active histone marks. The right part of figure represents distribution of random permutations.
Figure 2.12. Motif enrichment, genomic location, and gene connectivity of dDHS sites. 

A) *De novo* motif detection was performed on hESC dDHS, hiPSC dDHS, and memory dDHS sites. 

B) Location of dDHS sites relative to genes. Shown are the location of different classes of dDHS sites relative to promoter, intron, exon, transcription termination site (TTS), and intergenic regions. “Genomic” represents the total genome coverage for each classification. 

C) The top 20 genes that are most connected to dDHS sites, described in Figure 5, were analyzed by permutation testing (box plots). Each of the 20 genes were shown to have significantly more connections to dDHS sites (red circles) than the permutation datasets.
Figure 2.13. Location of dDHS sites surrounding key developmental genes. Screenshots of hESC dDHS sites (gray boxes) at SOX2, FOXA1, HMX2/HMX3, and HOXA gene loci. ESC dDHS sites around SOX2 overlapped with a hESC super-enhancer (red bar) identified in H1 ESCs.
Figure 2.14. Differential DHS sites at gene loci foreshadow expression misregulation upon *in vitro* differentiation. A) ZIC2 and ZIC5, SOX21, OTX2, and BMP4, were enriched for hESC dDHS sites (gray boxes), and were insufficiently up-regulated in hiPSCs upon mesendodermal differentiation. hESC dDHS sites at ZIC2/ZIC5 and SOX21 loci overlapped with hESC super-enhancers (red bar). B) PLB1 was enriched for hiPSC dDHS sites (gray box in the middle) and memory dDHS sites (the remaining gray boxes), and was improperly up-regulated in hiPSCs upon neuroectodermal differentiation.
CHAPTER III:
INITIAL DEMONSTRATION OF A STRATEGY FOR *EN MASSE* FUNCTIONAL CHARACTERIZATION OF HUMAN ENHANCER ELEMENTS

3.1 OVERVIEW

We developed a method for highly parallel measurement of the functional activity of putative regulatory elements isolated from open chromatin in a given cell type. Our assay consists of cloning FAIRE-enriched DNA (Formaldehyde Assisted Identification of Regulatory Elements) into GFP-based enhancer reporter vectors, transfecting cells with the reporter constructs, sorting for GFP-positive cells, recovering the inserts from the reporter, and sequencing them. Each FAIRE-enriched fragment is tested in the forward and reverse orientation relative to the reporter gene. Here we report an initial demonstration of this assay in HEK293T cells. Among 26,046 FAIRE sites identified by stringent criteria in HEK293T cells, 3,428 (13%) were enriched in GFP-positive cells in both orientations. We found that relative to all FAIRE fragments, the 3,428 regions were on average further from transcription start sites, and that motifs of eleven known transcription factors, including known enhancer-binding factors AP-1 and Foxa2, were enriched in those regions. In addition, genes that were near these 3,428 regions were expressed at higher levels than overall gene expression levels in HEK293T cells. Thus we identified 3,428 open chromatin regions associated with orientation-independent activation of a reporter in HEK293T cells, demonstrating the feasibility of functional characterization of several thousand enhancer elements in a single experiment of this design.
3.2 INTRODUCTION

Gene regulatory elements are short non-coding genomic regions that regulate the transcription of genes. These elements can act as promoters, enhancers, insulators and silencers, and are typically bound by sequence-specific transcription factors. The binding of sequence-specific regulatory factors is usually accompanied by nucleosome loss, a hallmark of regulatory activity in eukaryotic cells [5,14,15]. These nucleosome-depleted regions or “open chromatin” can be detected through methods like DNase I hypersensitivity [5,6,8,15,25] or FAIRE (Formaldehyde Assisted Identification of Regulatory Elements) [7,14,26,27]. The results of these methods can be quantitatively measured across the genome in a single experiment by quantitative PCR, microarray, or high-throughput sequencing.

FAIRE has been used to annotate regulatory elements in dozens of human cell types and tissues [7,27,31]. There are typically about 100,000 FAIRE sites per cell type. Of these, only ~10% of FAIRE sites overlap transcriptional start sites (TSSs) or are within 2 kb of a TSS [7,27]. Therefore, most putative regulatory sites identified by FAIRE are non-promoter distal elements. A large proportion of these distal elements are putative enhancers, many of which play key roles in the control of cell-type-specific gene expression programs [140,198-202]. A typical mammalian cell contains tens of thousands of active enhancers, and it has been estimated that there may be more than 1 million enhancers active across all human cell types [8,31,111]. These distal enhancers can be located upstream, downstream or within their target genes, and can modulate expression independent of their orientation [139,140,198,199,201]. Some have been found at large genomic distances from their targets [141,142], or even on a different chromosome[143,144]. Identifying and functionally annotating these distal elements is important
for understanding normal cell biology and for understanding the disease-associated sequence variation that occurs in these regulatory elements [146,147,203-206].

While DNase I, FAIRE, ChIP and other methods have made identification of putative regulatory elements routine, those methods do not provide information regarding the functional activity of each element. Determining the function of putative regulatory elements is currently a major bottleneck in the study of gene regulation. Traditionally, regulatory elements have been functionally characterized one or a few at a time through serial transfection of individual reporters [112,117,118,121-124], but more recently DNA elements have been characterized using systematic large-scale approaches [95,97,98,138]. One such high-throughput functional method was designed to characterize promoter regions in the human genome [138]. In this study, random fragments from sheared genomic DNA were cloned into a GFP-based promoter reporter vector and tested for their ability to drive GFP reporter expression in cell culture. While they identified 858 putative promoters, this approach is highly inefficient as active regulatory elements in each cell type occupy less than 5% of the genome. Therefore most cloned fragments will not be functional. Other studies developed massively parallel reporter assays to test enhancer activities of either a few known human enhancers and a large number of their engineered variants [95,97] or a few thousand predicted enhancers and their variants [98]. Despite the success of those studies, it should be noted that the tested enhancers were synthesized from known or predicted enhancer sequences. Such synthesis-based approaches are costly and require exact DNA sequences of tested elements, thus it is hard to adapt these to the genome-wide annotation of thousand of native enhancers in a given human cell type of which many are unknown elements.
In this study, we developed a highly parallel genome-wide method for measuring the enhancer activity of native putative regulatory elements identified by FAIRE in a given human cell type. To demonstrate the utility of this assay, we tested FAIRE-enriched DNA from HEK293T cells. Of 26,046 regions of open chromatin identified by stringent criteria in these cells, we identified 3,428 sites associated with orientation-independent enhancer activity. These regions were found to be more distal than typical FAIRE elements, and showed enrichment for eleven transcription factor binding motifs including for AP-1 and Foxa2. In addition, genes that were near these 3,428 regions were expressed at higher levels than overall gene expression levels in HEK293T cells. Thus, functionally identifying several thousand enhancer elements in a single experiment is feasible with a design based on ours, which represents a 100-fold increase in throughput relative to traditional serial approaches.

3.3 RESULTS

3.3.1 A high-throughput method to test enhancer function of FAIRE-enriched DNA fragments.

Each FAIRE experiment identifies approximately 100,000 active regulatory elements in any given cell type. Testing the function of each of the putative regulatory elements is a daunting challenge. We aimed to interrogate the enhancer activity of most or all of these elements in a single high-throughput reporter assay. In our assay, each putative regulatory element is tested for its ability to induce the expression of a GFP reporter gene (Figure 3.1A). FAIRE fragments were cloned en masse into a Gateway entry vector and then transferred to a GFP-based enhancer reporter vectors through the Gateway cloning system (Figure 3.1B). Two enhancer reporter vectors were constructed, which contained Gateway attR1-R2 sites (forward reporter vector) or attR2-R1 sites (reverse reporter vector) upstream of a E1B minimal promoter [207,208] and the
GFP gene. In this way, FAIRE fragments were inserted 69 bp upstream of the GFP gene in both orientations. This design allowed us to test whether the function of each insert was orientation-independent, traditionally a defining feature of enhancers [139,140,198,199,201]. Cells were transfected with either forward or reverse reporter constructs and sorted based on GFP intensity. Cells transfected with the positive control construct that contains the SV40 enhancer and cells transfected with the negative control construct that contains no insert were used to set the gating parameters in the fluorescence-activated cell sorting (FACS). Inserts in the entry vector, reporter vectors, and GFP-positive cells were recovered by PCR amplification. The raw FAIRE fragments and recovered FAIRE inserts were also sequenced and analyzed.

It should be noted that not all cloned DNA fragments in our library originate from open chromatin regions. About 100,000 open chromatin regions are identified in a typical FAIRE experiment, covering ~1% of the genome. FAIRE DNA libraries are estimated to be 20 to 50-fold enriched for active regulatory regions based on qPCR measurements [14,26]. If one assumes that 1% of the genome is within open chromatin, a library with 20-fold enrichment of open chromatin over background by FAIRE will consist of 17% FAIRE DNA fragments. Similarly, if the proportion of the genome in open chromatin is 5% and the enrichment is 20-fold, the proportion of FAIRE fragments in the library is 50%. In any case, our starting FAIRE library is a mixture of DNA from putative active regulatory regions (FAIRE sites) and background genomic regions.
Figure 3.1. Experiment design for high-throughput characterization of enhancer activities of open chromatin regions isolated by FAIRE.  

A) Schematic representation of the enhancer reporter assay. FAIRE fragments were cloned upstream of E1B minimal promoter and GFP gene in the reporter vector. Reporter vector with no insert and vector with an SV40 enhancer were used as negative and positive controls, respectively. The enhancer activity of tested FAIRE fragments was shown by their ability to drive GFP expression.  

B) Experiment design. FAIRE fragments were isolated from HEK293T cells, cloned into Gateway entry vector, transferred into two reporter vectors in which the FAIRE insert was upstream of the GFP gene in both orientations through Gateway recombination system. These FAIRE constructs were transfected into HEK293T cells. GFP-positive cells were collected in FACS. FAIRE fragments were recovered and sequenced from six libraries: the original FAIRE DNA, Gateway entry vector, forward reporter vectors, reverse reporter vectors, GFP-positive cells transfected with forward reporter vectors, and GFP-positive cells transfected with reverse reporter vectors (See Methods).
Another consideration is that the original FAIRE reaction used to make the library was performed on approximately ~250 million cells, meaning that theoretically ~500 million copies of each genomic locus were available for inclusion in the starting library. Approximately 15% of the FAIRE DNA (2 µg) was used in the Gateway cloning. Because DNA from both open chromatin and genomic background are equally likely to be cloned into Gateway vectors, the proportion of active regulatory elements in the Gateway entry vector library should reflect the proportion in the starting FAIRE reaction. In our case, one bottleneck for preserving the complexity of the library was the number of bacterial colonies that were collected following creation of the library in the Gateway entry vector. Each colony contains one vector that carries one DNA segment. To ensure that most of the approximately ~26,000 stringent open chromatin sites were cloned into a Gateway entry vector, 300,000 colonies were collected and assayed for enhancer activity.

Three independent transfection and FACS experiments were performed in each orientation, for a total of six experiments. In each FACS experiment, cells transfected with the positive or negative control vector were used to set the FACS gating parameter. The gate was set such that no negative control cells were in the “GFP bright” region (Figure 2B). Using this setting, approximately 5% of the positive control cells were in the GFP-bright region, which we considered GFP-positive. Cells transfected with libraries of FAIRE-enriched reporter constructs were sorted using the same gating parameters. Only GFP-positive cells were collected. In each of the six experiments, approximately 5,000-8,000 GFP-positive cells were collected from a total of ~10M transfected cells (Figure 2). PCR amplification from the cell lysate was used to recover ~650-bp DNA fragments. Due to the design of the primers, these PCR amplicons contained both vector and insert DNA. These fragments were then sheared to ~100-200 bp for sequencing.
library preparation. Sequencing reads from all three experiments of each orientation were aligned to reference genome separately then pooled for downstream analyses.

Figure 3.2. Collecting GFP-positive cells in fluorescence activated cell sorting (FACS). Gates were set in PE vs. FITC scatter plots (A) and FITC vs. FSC plots (B). A threshold was chosen such that no negative control cells were in the P3 GFP bright region (B). Only cells in P3 GFP-bright region were considered GFP-positive and were collected. Using this gating, ~5% of the positive control cells were GFP-positive.

3.3.2 3,428 DNA elements associated with orientation-independent enhancer activity were identified

We identified FAIRE regions with enhancer activity by comparing the sequences recovered from the original FAIRE enrichment, the entry vector library, the reporter vector libraries, and GFP-positive cells. FAIRE regions enriched in the original FAIRE experiment were identified using ZINBA [52]. To reduce false positive rates and generate a high-confidence set of FAIRE regions, a stringent ZINBA threshold was used to identify 26,046 FAIRE sites.
Sequences from the entry vector library, reporter vector libraries, and GFP-positive cells were aligned to reference human genome NCBI GRCh37 (hg19), and genomic regions whose DNA was enriched in each library were again identified by ZINBA (Table S1). It should be noted that we used a standard transient-transfection protocol that can result in multiple reporter vectors in each transfected cell [209,210]. Therefore, multiple inserts may be present in GFP-positive cells, only one of which may have enhancer activity. Nonetheless, we reasoned that true enhancer elements would consistently induce expression of the reporter gene across all independent replicates, and that by combining data from the independent transfection experiments, we would increase the ratio of true positives to false positives. In addition, we used the following computational criteria in downstream analyses to reduce the false positive rate. Regions identified as associated with enhancer activity must 1) be enriched in the starting FAIRE library; 2) be enriched in both forward and reverse GFP-positive libraries; and 3) be present in the entry vector, forward, and reverse reporter libraries. Selecting elements with activity in both orientations not only confirms a traditional defining property of enhancers, but also represents independent observations of activity of the same DNA segment.
<table>
<thead>
<tr>
<th>Sample</th>
<th># of total reads</th>
<th># of aligned reads (Aligned %)</th>
<th># of FAIRE/enriched regions</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAIRE</td>
<td>57,046,446</td>
<td>43,787,792 (76.76%)</td>
<td>26,046</td>
</tr>
<tr>
<td>Entry Vector</td>
<td>145,273,195</td>
<td>4,289,053 (2.95%)</td>
<td>21,132</td>
</tr>
<tr>
<td>Forward Reporter Vector</td>
<td>195,447,256</td>
<td>6,885,125 (3.52%)</td>
<td>23,782</td>
</tr>
<tr>
<td>Reverse Reporter Vector</td>
<td>190,425,149</td>
<td>6,917,432 (3.63%)</td>
<td>23,941</td>
</tr>
<tr>
<td>Forward GFP-positive cells</td>
<td>60,960,333</td>
<td>10,103,201 (16.57%)</td>
<td>23,619</td>
</tr>
<tr>
<td>Reverse GFP-positive cells</td>
<td>45,021,375</td>
<td>8,907,007 (19.78%)</td>
<td>24,003</td>
</tr>
</tbody>
</table>

Table 3.1. Sequencing depth, number of mapped reads, and enriched FAIRE sites in six libraries. FAIRE fragments from six libraries were sequenced and analyzed: the original FAIRE DNA, Gateway entry vector, forward reporter vector, reverse reporter vectors, GFP-positive cells transfected with forward reporter vectors, and GFP-positive cells transfected with reverse reporter vectors.
Figure 3.3. Most enhancer positive regions appeared to be distal regulatory elements that drove the expression of nearby genes. A) Identifying enhancer positive fragments. 3,428 of 26,046 FAIRE regions were enriched in reporter vectors that were able to drive GFP expression in both orientations. B) 62% of these 3,428 enhancer positive fragments were in non-promoter intergenic regions. C) On average, enhancer positive fragments were further away from TSSs than FAIRE fragments or randomized FAIRE fragments. D) Genes near enhancer positive fragments were expressed at higher levels than other genes.
A total of 23,619 inserts (Table S2) were enriched in GFP-positive cells in the forward orientation, and 24,003 (Table S3) in the reverse orientation. Among these, 8,730 (Table S4) were held in common, meaning that they were associated with GFP-positive cells in both orientations. Of these 8,730, 3,428 overlapped with the 26,046 original stringently-called FAIRE sites and were also present in the entry and reporter libraries (Figure 3.3A). We refer to these 3,428 fragments (Table S5) as “enhancer positive” from this point.

### 3.3.3 Fragments positive for enhancer function tend to lie distal to transcription start sites

We examined the genomic locations of these 3,428 enhancer positive regions. Nine percent were located within 2-kb upstream of RefSeq genes, which was about the same proportion as for FAIRE peaks in the starting FAIRE library (9.8%). It is known that promoter elements can act as enhancers, and that some enhancer elements are near transcription start sites (TSSs) [211]. Approximately 62% of the enhancer positives were located in non-promoter intergenic regions, relative to the expected null frequency of 58% (all genomic DNA) (Figure 3.3B). We calculated the distance between each enhancer positive region to the nearest gene (Figure 3.3C, green). Similarly, we calculated such distance of FAIRE peaks (Figure 3.3C, red) and randomized FAIRE peaks (Figure 3.3C, black). FAIRE peaks were on average 70kb from the nearest TSS, while FAIRE peaks with randomized positions were 30kb away. In contrast, the 3,428 enhancer positive regions were located on average 210kb from the nearest TSSs, and the distribution of distances was strongly skewed toward longer fragments. This observation is consistent with enhancer positive regions functioning as regulatory elements that act at a distance.
3.3.4 Genes close to identified enhancer elements were relatively highly expressed

To further assess evidence for enhancer function of the 3,428 positive regions, we assigned each to its nearest RefSeq gene as its putative target and determined expression levels of these target genes. Based on this mapping, 674 genes were associated with at least one enhancer positive regions, of which 114 (16.9%) were associated with more than three, with one gene TLE4 associated with 140 regions (Table S6). TLE4 (transducin-Like Enhancer Of Split 4) was shown to be highly expressed in HEK293 cells compared to other cell types [212]. This suggests enhancer positive regions may cluster and regulate common target genes, similar to previous observations in pancreatic islets [27]. More recently, blocks of active regulatory element super-enhancers were described in mouse and human ESCs, as well as in 86 diverse human cell types [176-178]. We hypothesized that associations with enhancer positive regions would result in higher expression of target genes. To test this, we measured RNA levels in HEK293T cells using Affymetrix human exon arrays. We found that these 674 genes were expressed at significantly higher levels compared to expression levels of all genes (Figure 3.3D; t test, p-value < 2.2e-16).

3.3.5 DNA motifs in identified enhancer elements

We further interrogated the sequence composition of the identified putative enhancer elements by de novo motif analysis. Thirty-nine overrepresented motif sequences were found in those regions using the cERMIT [213] and CisFinder [214] software packages. Using STAMP [215], the closest matches in the TRANSFAC [57] or JASPAR [59] databases were returned for each de novo motif. Eleven transcription factor binding site motifs were found in the overrepresented de novo motifs (E-value < 1e-9) (Figure 4A). The motifs corresponding to the two most prominent known enhancer-binding factors were AP-1 and FoxA2. Activator protein 1
(AP-1) is a transcription factor complex comprised of members of the Fos/Jun family [216,217]. AP-1 has been shown to bind to enhancer elements and regulate gene expression in response to numerous stimuli [218-220] likely by maintaining chromatin accessibility [181]. Foxa2 is a transcriptional activator [221,222][223] and a member of the forkhead class of DNA-binding proteins that has been shown to regulate nucleosome depletion [221,224]. All of the factors that bind to the eleven enriched motifs are expressed in HEK293 cells [225,226]. An example of an enhancer-positive region containing an AP-1 binding site is shown in Figure 4B. This element appears to be open selectively in HEK293T cells, human embryonic stem cell line H1, cervical cancer cell line HeLa-S3, liver carcinoma cell line HepG2, and human umbilical vein endothelial cell line HUVEC, but not active in chronic myelogenous leukemia cell line K562 and lymphoblastoid cell line GM12878 (Figure 4B).
Figure 3.4. 3,428 enhancer positive fragments were enriched in transcription factor binding sites. A) Motifs of eleven transcription factors, including known enhancer binding factors AP-1 and Foxa2, were found enriched in enhancer positive fragments. B) An example of AP-1 binding site in enhancer positive fragment upstream of SHC3 gene.
3.4 DISCUSSION

Open chromatin regions identified by DNase I hypersensitivity or FAIRE often harbor active regulatory elements [6-8,14,25,27] that are used to regulate cell-type-specific gene expression. Using such methods, a large number of putative regulatory elements have been identified in projects such as ENCODE [29,31] and the Epigenome Roadmap [48]. Effective and practical high-throughput experimental methods that can functionally test these putative elements are urgently needed. In this study, we developed a highly parallel method for testing a large number of regulatory elements in a given cell type in a single experiment. As proof of concept, we performed this assay in the HEK-293 cell line using a library derived from FAIRE-enriched DNA fragments. We identified 3,428 FAIRE-enriched DNA elements with orientation-independent enhancer activity. Our results demonstrate the feasibility of such high-throughput functional assays, which represent orders-of-magnitude improvement in functional characterization of regulatory elements over existing methods. Moreover, this assay can be easily modified to test for promoter, silencer and insulator function, and performed in other cells of interest.

While there are many aspects of the method that could be improved (and are discussed below), it is important to point out the features of the experiment that are crucial to the success and general applicability of the assay. First, we used FAIRE DNA as starting material. To ensure that nearly all of the estimated 100,000 regulatory elements in a given cell type are cloned into the reporter vectors, and to preserve the library complexity during cloning steps, it is advantageous to start with a DNA library in which the ratio of regulatory DNA fragments to non-regulatory DNA is as high as possible. It is also critical to start with a library that contains regulatory elements that are intact and have not been broken into non-functional fragments.
FAIRE isolates regulatory regions intact, enriches active regulatory regions 20 to 50-fold [14,26], and can be performed on both cultured cells and tissues. In contrast, within a sheared or enzymatically digested genomic DNA library, or within a DNase HS library, regulatory elements are often cut or destroyed. Similarly, a library of ChIP fragments would contain only a fraction of all active regulatory regions, those that are associated with a particular regulatory factor of interest. In addition, the yield of ChIP is much lower than FAIRE and obtaining enough starting material (1-2 µg of DNA) for Gateway cloning can be challenging. Therefore, using FAIRE as starting material allowed us to build a putative regulatory element library of relatively high specificity and complexity. Second, we used the Gateway cloning system, which allows DNA fragments to be efficiently transferred between different reporter vectors. After DNA-fragments are cloned into the Gateway entry vector, the Gateway system makes downstream cloning for such high-throughput experiments fast and efficient. More importantly, it is easy to move the inserts to different reporter vectors so that in addition to the enhancer activity tested in this study, promoter, insulator or silencer function can be tested in parallel in future experiments. We also note that regulatory elements derived from tissues can be tested in comparable cell lines.

While the initial results of this pilot study provide a proof of principle, some aspects of the design should be improved in future experiments. First, the sequencing libraries derived from the entry vector and both reporter vectors were made by shearing the entire construct (insert plus vector DNA), and the inserts in GFP-positive cells were recovered by a PCR product that included a significant part of the vector backbone and GFP gene along with the insert. As a result, our sequencing libraries contained inserts, vector sequences, and hybrids of insert and vector sequences. This made the rate of alignment to human genome low (Table 1). Attempts at computational processing of the reads to remove vector sequence were not effective. To
overcome this problem for these experiments we used a brute-force approach by sequencing each library to great depth and using only sequences that mapped perfectly to the human genome. To the extent this affected our sensitivity, it is possible that some regulatory elements were not recovered, or were not adequately enriched in aligned tags to be distinguishable from background. This problem can be avoided by optimizing primer design in future experiments. Second, GFP-negative cells contain important information and should be collected. Third, we used transient transfection to introduce reporter vectors into the testing cell line. Thus, multiple reporter vectors may enter a single cell during transfection. Although we tried to overcome this problem by performing three independent transfection experiments, and adopted stringent criteria in analysis on the combined dataset, it would be beneficial to employ a single-site stable integration scheme such as a lentiviral system [49,50] in future experiments.
3.5 METHODS

**Cell culture.** Human embryonic kidney cell line HEK293T (Cat.#: HCL4517) were obtained from Open Biosystems (Part of Thermo Fisher Scientific) and cultured under recommended conditions using DMEM (Invitrogen, Cat.# 11995-065) supplemented with 10% Fetal Bovine Serum (Cellgro, Cat.# 35-016-CV), 2mM L-glutamine (Invitrogen, Cat.# 25030-081), 50 units/ml penicillin and 50 g/ml streptomycin (Invitrogen, Cat.# 15140-122). 250 million cells were collected for FAIRE experiment.

**FAIRE-seq.** FAIRE-seq was performed as previously described [14,28]. Five independent FAIRE experiments were performed, each on 50M HEK293T cells. FAIRE DNA from five experiments were pulled and used for sequencing and Gateway cloning.

**Cloning of FAIRE fragments into the Gateway entry vector.** 2 µg FAIRE DNA was used for Gateway cloning. FAIRE fragments were end repaired using Illumina library preparation reagents, ligated with adaptors containing BstX I sticky ends, and then cloned between BstX1 sites into Gateway entry vector pOTspnAttL.

**Cloning of FAIRE fragments from Gateway entry vector into Gateway reporter vectors.** 2 µg FAIRE DNA was used for Gateway cloning. FAIRE fragments were end repaired using Illumina library preparation reagents, ligated with adaptors containing BstX I sticky ends, and then cloned between BstX1 sites into the Gateway entry vector pOTspnAttL.

**Cloning of FAIRE fragments from Gateway entry vector into Gateway reporter vectors.** Gateway destination vector pGL3-GFP was constructed using pGL3-Basic luciferase reporter vector (Promega, cat.# E1751). The luciferase gene was replaced with the GFP gene between NcoI and XbaI sites, the E1B basal promoter was cloned between BglIII sites and HindIII sites 33-bp upstream of GFP gene, and the Gateway cassette of attR1-ccdB gene-attR2
was cloned into SmaI sites 69bp upstream of GFP gene. Vectors with the Gateway cassette inserted in both directions were confirmed by sequencing and named forward (attR1-R2-GFP) and reverse (attR2-attR1-GFP) vectors. FAIRE fragments were transferred into the pGL3-GFP forward and reverse vectors by Gateway recombination (Invitrogen).

For positive control forward and reverse vectors, SV40 enhancer was cloned between Kpn and Sac sites 75bp upstream of GFP gene.

Transfection of HEK293T cells using Lipofectamine. Forward and Reverse reporter vectors containing FAIRE fragments, SV40 enhancer (positive control) or no insert (negative control) were used to transfet HEK293T cells. HEK293T were plated on 6-well plates at a density of 400,000 cells in 2 ml medium per well. For each 6-well plate, 24 µg Plasmid DNA was diluted into 1.5ml Opti-MEM reduced serum medium (Invitrogen, cat.#31985-070), and mixed with 60 µl Lipofectamine 2000 (Invitrogen, cat.# 11668-019) that was diluted in 1.5ml Opti-MEM reduced serum medium, and incubated at room temperature for 20 minutes. 500 µl of DNA/Lipofectamine mix was then added to each well. Cells were incubated at 37°C for 24 hours before the FACS assay.

Detection of GFP-positive cells by fluorescence-activated cell sorting (FACS). Cells were trypsinized and resuspended in DMEM without phenol red for the FACS assay. Gates were set to exclude cellular debris. Cells transfected with negative or positive reporter vectors were sorted first. The green fluorescent cell population of interest was gated based on light scatter and fluorescence. A threshold was chosen such that no negative control cells were in the P3 GFP bright region while ~5% of the positive control cells were. Cells were sorted at a rate of 2500–3000 events/s. Only cells in P3 GFP-bright region were considered GFP-positive and were collected.
**PCR amplification and high-throughput sequencing of FAIRE inserts.** GFP-positive cells were lysed by freezing and thawing five times. Cell lysate was used for PCR amplification. Primers flanking a ~650bp region that contained the FAIRE insert, part of GFP gene and part of vector backbone were used. To avoid over-amplification, 12, 14, 16, 18, 20, 22, 24 cycles of PCR amplification were tested first, and 16 or 18 cycles were chosen for each experiment, based on the number of cycles that gave the dimmest visible band when examined on gel. PCR products were then shared into approximately 100-200bp fragments and sequenced on Illumina GAII machines.

Similarly, FAIRE inserts in the Gateway entry vector, forward and reverse reporter vectors were amplified using the same pair of primers, sheared, and sequenced.

**Sequencing data preprocessing.** Sequencing data were processed and mapped to NCBI build 37/hg19 human reference genome using BOWTIE [51]. Reads with zero mismatch that were mapped to less than 4 places in the genome were used in downstream analyses. It should be noted that the percent of mapped reads of Gateway vector libraries and GFP-positive cell libraries were low (between 3%-20%) because most of the sequenced sheared PCR fragments were either from the vector backbone or a hybrid of FAIRE fragment and vector backbone and therefore could not be 100% mapped to the human genome. Genomic regions that were enriched in starting FAIRE library, forward GFP-positive cells, and reverse GFP-positive cells were identified by ZINBA [52] using FDR threshold 0.05.

**Overall comparison of enriched regions in FAIRE, forward GFP-positive, and reverse GFP-positive libraries.** Enriched regions in each library were compared using BEDTools suite (version 2.16.1) [53]. The results were plotted using Venn Diagram Plotter (http://omics.pnl.gov/software/VennDiagramPlotter.php).
**Genome annotation.** Location of dDHS sites relative to genes was annotated using function annotatePeaks in HOMER (version v3.17) (Hypergeometric Optimization of Motif EnRichment) suite [54] with default UCSC RefGene annotation. The output of HOMER were processed using BEDTools suite (version 2.16.1).

**Microarray protocol and normalization.** Affymetrix Human Exon 1.0 ST microarray data were generated as described by Song *et al.* [7] from HEK293T cells. Exon arrays were processed and normalized by Sheffield *et al.* [180]. Data are publicly available on the UCSC Genome Browser (http://genome.ucsc.edu) and GEO under accession number GSE15805.

Thank you SS for the transforming scientific and life experience. You have taught me much about science, the human nature, and myself.

**Motif analysis.** DNA sequences of GFP positive regions were extracted from the UCSC Genome Browser. *De novo* motif search in dDHS sites was performed using cERMIT [213] and CisFinder [214]. The online version of CisFinder (http://lgsun.grc.nia.nih.gov/CisFinder/) was used with default parameter settings, except “clustered” motifs rather than the “elementary” ones. Top motifs from cERMIT and CisFinder were annotated using the STAMP web server (http:// www.benoslab.pitt.edu/stamp/) [215], with the “selected eukaryotic” option and searched in TRANSFAC [57] and JASPAR 2010 [59] databases.
CHAPTER IV: SPECIFIC HISTONE MODIFICATION PATTERNS MARK TRANSLOCATION BREAKPOINTS

4.1 OVERVIEW

Chromosomal translocations can disrupt normal gene function and result in fusion genes with new functions. Specific chromosome breakpoints and the formation of fusion genes are linked to various clinically defined cancers, and are shown to be the primary cause in many cases including subtypes of human leukemia and lymphoma. Genome-wide studies of chromosomal translocations suggest there exist DNA sequence-independent hotspots in the genome where translocations consistently recur. However, by which mechanisms translocations recur between certain genomic regions are still largely unknown. We hypothesized that susceptibility to chromosomal breakage is affected by chromatin structure marked by characteristic histone modifications. To test our hypothesis, we systematically analyzed public data from Roadmap Epigenomics of various histone modifications representing both active and repressive chromatin marks at 74 documented recurrent translocation breakpoint loci in hematopoietic cancers. We screened for histone modification patterns that may distinguish genes frequently involved in translocations from matched control genes in hematopoietic stem cells. We found higher levels of monomethylated histone H3 lysine 4 (H3K4me1) at translocation genes and specific patterns compared to control genes. In addition, 15 translocation genes featured enrichment of both active chromatin mark H3K4me3 and repressive chromatin mark H3K27me3, which is commonly seen in genes poised for transcription. Taken together, our results suggest
that characteristic chromatin features marked by specific histone modifications are enriched at translocation genes involved in hematologic malignancies. This finding sheds lights on understanding the genomic hotspots of chromosomal translocations in oncogenesis, and can provide guidance for experimental validation of the role of predisposition of specific histone modifications prior to chromosomal translocations.

4.2 INTRODUCTION

Balanced chromosome translocations are the most common genetic aberrations found in hematologic malignancies [61-64,75,148,149,163] and are responsible for initiating tumorigenesis in some cases [61-64]. A marked feature of translocations in leukemia and lymphoma is that they are repeatedly observed in different patients with the same tumor type [61-64,75,149,227-229] [230]. Several studies have attributed the recurrence and non-random distribution of translocations to cell lineage- and tissue-specific genome organization and the spatial interaction of chromosomes [75,87,148,163,227,230-232]. While physical proximity is essential for the formation of translocations, and can determine which chromosomes fuse on a genomic level, the local factors that predispose genomic regions to chromosomal breakage and translocations remain to be elucidated [75,148,163].

It is evident that the initiating event for translocation formation is DNA double strand breaks (DSBs) at two or more genomic loci. DSBs can be induced by cellular stress, genotoxic stress, or endogenously during transcription and DNA replication processes. DSBs result in broken ends that can recombine legitimately to remain intact or illegitimately to form translocations [75]. Therefore, it was proposed that the recurrent pattern of translocations could be due to homologous DNA sequences at breakpoints that allow them to recombine illegitimately. However, sequencing studies have shown that while breakpoints do tend to cluster
in specific regions, ranging from a few hundred bases to more than 20 kb long, these regions are not defined by any consistent homologous sequences [61-64,161]. It has also been proposed that certain DNA sequence or structure variations could render stretches of DNA more vulnerable to breakage [75,148,163]. Recent studies have linked some breakpoints to common fragile sites that are enriched in AT-dinucleotide repeats, giving these regions high DNA helix flexibility and the ability to form secondary non-B DNA structures capable of inhibiting DNA replication [75,148,163]. Similarly, it has been shown that bcl1 and bcl2 breakpoints are enriched in CpG dinucleotide sequences and that the major breakpoint region of bcl2 adopts a stable non-B DNA structure that is targeted by the RAG endonuclease during B-cell lymphogenesis [75,148,161,163]. While these studies re-affirm that breakpoints occur in defined regions within the genome that may be more susceptible to DSBs and translocations, it has become clear that correlations with DNA sequence only apply to a limited number of genomic locations.

Given the emerging role of chromatin in DNA accessibility and repair, and the fact that DSBs result in the context of chromatin, it seems plausible that the local chromatin environment could predispose certain genomic regions to breakage and translocations. In support of this view, genome-wide mapping of translocating regions after DSBs were introduced at c-myc or IgH suggest that translocations occur at higher frequency between transcriptionally active regions of the genome [75]. In addition, we recently found that the regions near translocation breakpoints in anaplastic large cell lymphoma (ALCL) are transcriptionally active prior to translocation formation. The implication of these studies is that the chromatin structure near breakpoints may be altered prior to translocation formation.

One potential mechanism for a contribution of chromatin to DNA DSBs and translocation formation is via histone modifications. Histones make up the most basic unit of chromatin, the
nucleosome, around which DNA is wrapped. The highly basic amino terminal tails of histones project away from the nucleosome and are subject to post-translational modifications that include acetylation, methylation, and phosphorylation. Histone modifications have been shown to play a key role in determining higher-order chromatin structure and modulating DNA transactions such as transcription, replication, DNA repair, and access to enzymes and regulatory proteins [20,111]. In addition, histone modifications are thought to indicate chromatin status [20,31]. Modifications such as acetylation have been linked to decondensed chromatin states and active transcription, while condensed and transcriptionally silent regions are generally enriched in H3K9me3 marks [20,31]. However, it is apparent that the histone “code” is becoming more complex with evidence that histone modifications can act combinatorily to create “facultative” states in which it is not obvious whether a region is silent or active, condensed or decondensed. Even less is clear about whether histone modifications can define regions of increased DNA breakage and translocations.

To date, histone modifications have been mapped only in a few breakpoints involved in hematopoietic cancers [48]. There has been no systematic demonstration of histone modification patterns in a comprehensive set of translocation genes involved in leukemias and lymphomas. Given the recently emerged role of chromatin structure in DNA damage and repair, we hypothesize that altered chromatin structure predisposes genomic sites to DNA breaks and translocations. To test our hypothesis, we leveraged existing databases containing information on chromosome translocations and genome-wide chromatin sequencing to map histone modification and DNaseI sites at the most common translocation sites. Our screening included active chromatin marks H3K4me1, H3K4me3, H3K27ac, H3K36me3, and DNaseI hypersensitivity sites, and repressive chromatin marks H3K9me3, and H3K27me3. We found that the majority of
translocation genes were located in chromatin regions defined by histone marks such as H3K4me1 that are traditionally associated with transcription and regulatory elements, and that some marks including H3K36me3 are more precisely found at breakpoints within these genes. Our results suggest that alterations in chromatin at the level of histone modifications are an important upstream event of translocations by making breakpoint regions vulnerable to DNA breakage and translocations.

4.3 RESULTS

4.3.1 Systematic comparative analysis to identify chromatin features that are enriched at recurrent chromosomal translocation sites in hematopoietic malignancies

To test our hypothesis that specific chromatin structure predisposes genomic regions to chromosomal breaks and translocation, we applied a systematic bioinformatics approach to examine patterns of histone modifications and open chromatin profiles at genomic loci that were frequently involved in translocation in hematopoietic malignancies. We used CD34+ cells as our model system in this study. CD34+ cells represent immature hematopoietic stem cells that can differentiate to a variety of specialized blood cells or cells of the immune system. Many studies have suggested that chromatin architecture and histone modification patterns can be inherited from progenitor cells to differentiated cells [87]. Thereby, for translocations that occur in cells at more differentiated stages, such as BCR-ABL1 in chronic myeloid, and lymphoid progenitors, myeloids, and lymphoids, it is reasonable to examine the epigenetic states at those loci in their CD34+ hematopoietic stem cells. By interrogating the chromatin structure and histone modification profiles at genomic regions that were frequently involved in hematopoietic malignancies in CD34+ cells, we could gain insights on which chromatin features may be
enriched at translocation breakpoints and may potentially make those loci more vulnerable to double strand breaks than other genomic regions.

We took advantage of the rich dataset on hematopoietic stem cells available in NIH Epigenomics Roadmap Project [48]. We chose the primary CD34+ cell line generated from a 33-year-old female (see Methods). Among other CD34+ cell lines studied in Epigenomics Roadmap Project, this primary CD34+ cell line has the most extensive datasets including DNaseI-seq measuring chromatin accessibility, ChIP-seq data for active histone modifications H3K4me1, H3K4me3, H3K36me3, H3K27ac and repressive histone modifications H3K9me3, H3K27me3, reduced representation bisulfite sequencing (RRBS) data of DNA methylation, and mRNA-seq for transcription levels. To collect information on genomic loci that were frequently involved in chromosomal translocations in hematopoietic malignancies, we queried the Mitelman Database of Chromosome Aberrations that catalogs all reported cases of chromosomal translocations. We identified 74 RefSeq genes that were reported in at least 10 cases of hematopoietic malignancies (Table 4.1). Chromatin features such as histone modification and chromatin accessibly are closely linked to expression levels and genomic features of genes. Therefore, to identify chromatin features that may be preferentially enriched at those frequent translocation genes, for each translocation gene, we firstly defined a set of normal control genes that have similar gene expression level based on mRNA-seq data, and similar genomic features including 1) length, 2) percentage of transcript that is exonic, 3) number of exons, and 4) GC%. Up to 100 control genes were selected for each translocation genes (See Methods). Altogether, 5,076 control genes were selected for all 74 translocation genes (Supplementary Table 4.1).
4.3.2 Translocation genes are enriched for H3K4me1 in CD34+ cells

We examined the chromatin accessibility and presence of histone modifications within gene bodies and their surrounding upstream and downstream 2kb regions. We will refer to those regions as gene +/- 2kb regions hereafter. The levels of histone modifications and DNaseI hypersensitivity were measured by sequencing tag density (i.e. average number of sequencing tags normalized by the sequence depth). We compared the distribution of tag densities of each histone modification and DNaseI hypersensitivity in the gene +/- 2kb regions of 74 translocation genes and their 5076 control genes (Figure 4.1). Among the histone modifications examined, the population of translocation genes had a statistically significantly higher level of active mark H3K4me1 than the population of control genes (t-test p-value 0.00451; Wilcoxon test p-value 0.00008). There was no statistically significant difference of the distribution or average levels of active marks H3K4me3, H3K36me3, H3K27ac and repressive marks H3K9me3 and H3K27me3 between the translocation genes and control genes. Similarly, the distribution and average of DNase-I hypersensitivity levels were comparable between the translocation gene set and control gene set.

Histone modification H3K4me1 has been shown to be a mark of active transcription and active enhancers. The enrichment of H3K4me1 in gene +/- 2kb regions of translocation genes in comparison to their control genes suggest a higher frequency of active enhancers present in translocation genes. A key characteristic of translocation events is the special proximity of genomic regions that contain the paired breakpoints in the nucleus, which can be measured by fluorescence in situ hybridization (FISH). Enhancers are thought to function by bringing distal regulatory elements into the proximity of gene promoters with bound transcription machinery, forming a loop-like structure. It is possible that translocation sites are enriched for active
enhancers that help bring translocation partners together and thereby facilitate the formation of a translocation.

**Figure 4.1. Comparison of histone modifications and DNase hypersensitivity signals in 74 translocation genes and their 5076 control genes.** Signal densities of H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3, H3K36me3, and DNase HS in annotated gene +/-2kb were calculated (see Methods). The distributions of each mark in translocation genes and control genes were compared in the Kernel density plots. The x-axis of each plot is average read count in each gene (# of reads per kb). The distribution of H3K4me1 in translocation genes showed a statistically significant shift (t-test p-value 0.00451) from that in control genes.

It should be noted that in this analysis translocation genes and control genes were compared as two aggregated populations. The 74 translocation genes had a broad range of expression levels in CD34+ measured by mRNA-seq. Since the level of histone modification and
chromatin accessibility is correlated with transcriptional activity, it is likely that both translocation gene sets and their control gene sets contain subpopulations of genes that have different patterns of histone modifications, as discussed in more detail below.

4.3.3 Histone marks and DNase-I hypersensitivity defined subpopulations of translocation genes in CD34+ cells

Translocation genes had varying expression levels in CD34+, which suggests that they may have distinct profiles of histone modifications and chromatin accessibility. To examine whether such subpopulations existed among 74 translocation genes, we compared each translocation gene to its corresponding set of control genes separately. For each histone mark or DNaseI hypersensitivity site, the tag density in the +/- 2kb region of each translocation gene was compared to the tag densities of its control genes (Supplementary Figure 4.1). The percentage of control genes for which tag densities were lower than that of the translocation gene was calculated (see Methods). 74 translocation genes were clustered by the patterns of their relative levels of histone modifications and DNaseI hypersensitivity compared to each of their own control genes (Figure 4.2). We find that there are subgroups of translocation genes that have higher H3K4me1, H3K4me3, and H3K27ac levels, but relatively lower H3K9me3 and H3K27me3 signals.

As shown in the Figure 4.2, different translocation genes showed different patterns of chromatin signals. A sub-population consisting of 30 of the 74 translocation genes featured enrichment of active marks, namely levels of H3K4me1, H3K4me3, H3K27Ac, and DNaseI hypersensitivity. For example, BCL2, BCL3, BCL6 all had relatively higher levels of H3K4me1, H3K3me3, and H3K27ac. Half of these 30 genes also had relatively higher signal of the active transcription mark H3K36me3 (top left cluster). About 20 genes showed relatively higher levels
of both active mark H3K4me3 and repressive mark H3K27me3. It is possible that those genes had bivalent promoter domains that are marked by the coexistence of H3K4me3 and H3K27me3 and may be “primed” for transcriptional activation.

**Figure 4.2. Comparison of histone modifications in individual translocation genes.** Average read count for each histone modification and DNase HS in each translocation gene was compared to those in corresponding control genes. The percentile rank of a translocation gene among its control genes is plotted in the heatmap. Translocation genes were clustered and ordered using the overall ranking matrix (see Methods).

Notably, majority of the 74 translocation genes had relatively higher level of H3K4me1 than its control genes. The enrichment of H3K4me1 was frequently accompanied by enrichment of H3K4me3. This is consistent with the observation above that translocation genes had overall
higher level of H3K4me1 than control genes. This suggests that among genes with similar transcriptional activity, the level of H3K4me1 may distinguish those that are more prone to translocation from those that are not. Interestingly, some translocation genes, for example ALK, that were enriched for H3K4me1 and sometimes also H3K3me3 or DNase hypersensitivity were not appreciably expressed in CD34+, indicating that those genes were in a more open chromatin state even though they were not actively transcribed. It is possible that the more open chromatin state at those loci in combination with the presence of enhancer elements make those regions more vulnerable to translocation.

4.3.4 Histone modifications around specific breakpoints

Different histone modifications locate at different regions of genes and have different shapes of signals. H3K4me3 and H3K9me3 often have narrow peaky signals near transcription start sites (TSSs) of actively and inactively transcribed genes, respectively. H3K4me1 and H3K27ac co-localize with enhancer elements, which also have relatively narrow peak signals. H3K36me3 and H3K27me3 exhibit much broader signals along the entire gene body, with H3K36me3 marking actively transcribed genes and H3K27me3 marking transcriptionally repressed genes. All reported breakpoints in hematopoietic malignancies were mapped to introns of translocation genes. Studies have shown that histone modifications may have unique patterns at intron-exon boundaries and in introns. Such location information of histone modifications was not considered in the analyses above in which averaged signal density across the +/- 2kb promoter region was used. To truly understand the local chromatin environment and histone modification patterns at precise breakpoints, we decided to explore the strength and shape of each histone modification around actual annotated translocation sites. Coordinates of breakpoints in 19 of 74 translocation genes were collected from literature (see Methods; Supplementary
Table 4.2). We checked each gene and breakpoint region in the UCSC Genome Browser. Many of those gene mRNA annotations resulted from the sequencing of fusion genes or apparently truncated transcripts that corresponded to many of the breakpoint regions, showing that annotated start sites of these transcripts were adjacent to breakpoint regions we had. This supports the accuracy of the breakpoint information.

Breakpoints were mapped to one or multiple introns in the 19 translocation genes, resulting breakpoint regions of lengths from ~1kb to ~229kb. Each breakpoint region was extended by 1kb on the 5' end and 1kb on the 3' end to also examine the surrounding chromatin environment. These will be referred to as breakpoint +/- 1kb regions hereafter. For each breakpoint +/- 1kb region, to ensure the robustness of the results, two types of control regions were selected from control genes of each translocation gene: I) regions of the same relative length compared to the whole gene (control regions I); and (ii) regions of the same length (control regions II; see Methods). Each breakpoint region or control region was divided into 10 equal-sized, non-overlapping windows. The histone modification signal density was calculated in each window and in the entire region.

The overall density plots from the 19 breakpoint regions versus control regions were similar to the density plots from all 74 translocation genes versus control genes for most of the marks (Supplementary Figure 4.1). The results using control regions I and II were also comparable. There was not a single pattern that defined well all of the regions. Breakpoint regions seemed to be divided into those with very low histone mark signals relative to their control regions and those with enriched chromatin signals. For example, breakpoint regions of ABL1 (Supplementary Figure 4.2), PBX1, and BCL11B appeared to have lower levels of all histone modifications examined compared to their control regions. In contrast, some breakpoint
regions, such as HOXA11, PML, MLL, and BCR, had relatively higher levels of certain histone modifications than corresponding control regions. One of the primary differences when comparing the gene level plots and the region level plots was the H3K36me3 signal. Several breakpoints had relatively higher H3K36me3 signals, often accompanied by higher levels of H3K4me1 (e.g. HOXA11 in Figure 4.3.A), H3K27ac, and/or H3K4me3 (e.g. RARA in Figure 4.3.B). Higher H3K4me1 or H3K4ac signals might reflect some higher order chromatin structures such as loop structures at active enhancers. These distinct histone modification patterns may also reflect the differences in chromatin marks between the 5'-gene and 3'-gene in a fusion gene pair. It is common that the 5'-gene in a fusion gene contributes an active promoter and is originally actively transcribed, while the 3'-gene is originally not transcribed. An example of such a translocation gene pair is NPM1-ALK. Indeed, NPM1 and ALK showed very different histone modification patterns at their breakpoint regions (unpublished data).
Figure 4.3. Histone modification levels in 10 windows (see Methods) around the breakpoint (red) in genes *HOXA11* (A) and *RARA* (B) compared to histone modification levels in 10 windows from similar regions within 100 control genes of each gene (blue). Compared to control regions, breakpoint regions have relatively higher levels of H3K36me3 as well as H3K4me1 (*HOXA11*), and H3K27ace and H3K4me3 (*RARA*).

We reviewed each gene and breakpoint region in the UCSC Genome Browser with annotations from the Epigenomics Roadmap project, ENCODE project, and other datasets available in the UCSC Genome Browser. Consistent with the results above, there often appeared
to be a significant increase in this signal around the actual breakpoints. H3K36me3 has been recently reported to regulate DNA mismatch repair by recruiting MutS\textalpha{}, and that a lack of H3K36me3 results in a mutator phenotype. For double-stranded breaks and translocations, DNA repair acting downstream of the paired breaks is required for translocation formation. We also noticed that based on CTCF ChIP-seq experiments in other cell types (no data set was available for CD34+ cells), there seemed to be a lot of CTCF binding in the vicinity of these breakpoints, but usually not directly in the breakpoint region. CTCF has recently been shown to be involved in alternative splicing through bringing exons in close proximity to TSSs [85]. It may suggest that chromatin structure in terms of intra- and inter-chromosomal connections facilitated by CTCF may also play a role in translocation. More quantitative analyses are needed to further investigate the statistical significance of this.

**4.4 METHODS**

**Datasets.** We used RNA-, ChIP-, and DNase-seq data on a primary CD34+ cell line generated from a 33-year-old female (donor ID: RO 01549) available in NIH Epigenomics Roadmap Project (http://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/). BED files of sequencing reads mapped to NCBI build 37/hg19 human reference genome from mRNA-seq (GSM909310), DNaseI-seq (GSM530657), and ChIP-seq data for active histone modifications H3K4me1 (GSM621451), H3K4me3 (GSM621439), H3K36me3 (GSM706843), H3K27ac (GSM772894) and repressive histone modifications H3K9me3 (GSM621436), H3K27me3 (GSM706844) were used in analyses.

**Selecting control genes for translocation genes.** For each translocation gene, we selected a set of control genes from human RefSeq genes (Release 57) [89,91] using the following criteria: each of 1) gene expression, 2) gene length, 3) percentage of transcript that was
exonic, 4) number of exons, and 5) percentage of G/C bases in each control gene was within +/- 10 percentiles of the translocation gene among all RefSeq genes. RefSeq annotations were used. Gene expression was measured by RPKM (Reads per kilo base per million) calculated from mRNA-seq data. GC% was calculated using R package Repitools v1.4.0 [92]. If more than 100 genes fulfilled the criteria, 100 genes with most similar gene expression levels were selected. Altogether, 5076 control genes were selected for all 74 translocation genes. 10 of the 74 genes had less than 100 control genes: \textit{SEPT9 (15 control genes), CBFA2T3 (20), PRDM16 (23), BCR (35), RARA (70), BCL2 (71), TCF3 (73), RUNX1 (76), ETV6 (87), HSP90AA1 (91)}.

**Measuring histone modification and chromatin accessibility levels in the gene body and promoter.** Promoters were defined as transcription start sites (TSSs) +/- 2kb. Gene bodies were defined as transcribed regions +/- 2kb. For each translocation gene and each control gene, the number of mapped sequencing tags in gene body and the number in promoter from each ChIP-seq and DNase-seq dataset were calculated and normalized by the length of the region and the number of all mapped tags in each dataset. This normalized number of mapped sequencing tags was referred as tag density per kb and used to represent the level of histone modification or chromatin accessibility.

**Comparing translocation genes and control genes.** Translocation genes and control genes were compared at both the population level and individual translocation level. The levels of histone modification and chromatin accessibility in gene bodies (defined as above) of 74 translocation genes and 5076 control genes were compared by comparing the distribution of tag densities of each histone modification and DNaseI hypersensitivity. The density plots were generated using function density available in R statistical package (http://www.r-project.org).
The levels of histone modification and chromatin accessibility in the gene body and promoter of each translocation gene were compared to those of its selected control genes. For each histone modification or DNasel, the tag density of the translocation gene was compared to the tag densities of its control genes and represented as boxplots using function boxplot in R. The percentage of control genes for which tag densities were lower than that of the translocation gene was calculated. 74 translocation genes were clustered using patterns of their relative levels of histone modifications and DNasel hypersensitivity compared to their control genes using function hclust available in R. Heatmaps were generated using R package ggplot2 v0.8.9 (http://ggplot2.org/).

Measuring histone modification and chromatin accessibility levels around breakpoints. Coordinates of breakpoint regions in 15 translocation genes were collected from literature. Each breakpoint region was extended by 1kb on the 5' end and 1kb on the 3' end (referred as breakpoint +/- 1kb region).

For each breakpoint +/- 1kb region, the breakpoint start position compared to the TSS, the fraction of the gene covered by the breakpoint region, and the breakpoint mid-point position compared to the TSS were calculated. For example, one breakpoint region could start at a position that was at 10% (of the length of the gene) downstream of TSS with a length equal to 20% of the total gene length and whose midpoint is 20% downstream of TSS. Control "breakpoint" regions from each control gene were selected in two different ways to ensure the robustness of the results: 1) Control breakpoint region of same relative size compared to length of control gene: For each control gene, a breakpoint region was defined at the same relative position as in the true breakpoint gene. For the example above, the region would start at the base pair corresponding to 10% of the control gene length from the TSS, and the size of the region
was of 20% of the control gene size. It should be noted that the length of each selected control breakpoint region was different and not equal to the size of the true breakpoint region due to differences in control gene sizes; and 2) Control breakpoint region was the same size as the true breakpoint region: For each control gene, a breakpoint region of the same absolute length as the true breakpoint region was selected. The selected region was centered at the same relative position in the control gene based the position of the center of the true breakpoint region within the translocation gene. For example, if the true breakpoint region was 5kb long and was centered at a position 20% of the length of translocation gene from the TSS, the control breakpoint region was defined as a 5kb region centered at the position 20% of the length of the control gene from the TSS. In this case, the lengths of breakpoint region and its control regions were the same.

Each breakpoint region or control region was divided into 10 equal-sized, non-overlapping windows. The tag density per kb (defined as above) of each histone modification and DNaseI was calculated in each window. Tag densities in windows of each breakpoint region were compared to those of control regions.
### 4.5 SUPPLEMENTAL DATA

<table>
<thead>
<tr>
<th>Gene</th>
<th>Gene</th>
<th>Gene</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOXA13</td>
<td>ARHGAP26</td>
<td>CEBPA</td>
<td>FUS</td>
</tr>
<tr>
<td>TLX1</td>
<td>PVT1</td>
<td>HOXA10</td>
<td>MLLT3</td>
</tr>
<tr>
<td>HOXA11</td>
<td>BCL6</td>
<td>MYO1F</td>
<td>CBFB</td>
</tr>
<tr>
<td>RUNX1T1</td>
<td>BCL2</td>
<td>JAK2</td>
<td>RUNX1</td>
</tr>
<tr>
<td>ALK</td>
<td>KIF3B</td>
<td>RBM15</td>
<td>MYC</td>
</tr>
<tr>
<td>BCL11B</td>
<td>DCPS</td>
<td>MKL1</td>
<td>HOXA9</td>
</tr>
<tr>
<td>CCND1</td>
<td>BCL9L</td>
<td>ZCCHC7</td>
<td>RPN1</td>
</tr>
<tr>
<td>FGFR3</td>
<td>MALT1</td>
<td>PML</td>
<td>SEPT11</td>
</tr>
<tr>
<td>CEBPE</td>
<td>ZBTB16</td>
<td>PBX1</td>
<td>ETV6</td>
</tr>
<tr>
<td>ID4</td>
<td>RARA</td>
<td>SEPT9</td>
<td>GSTP1</td>
</tr>
<tr>
<td>CRLF2</td>
<td>MECOM</td>
<td>TCF3</td>
<td>LMO2</td>
</tr>
<tr>
<td>RALGSP1</td>
<td>WHSC1</td>
<td>CEBPG</td>
<td>DEK</td>
</tr>
<tr>
<td>PAX5</td>
<td>SH3GL1</td>
<td>HLF</td>
<td>HSP90AA1</td>
</tr>
<tr>
<td>HOXA2</td>
<td>CBL</td>
<td>CBFA2T3</td>
<td>NPM1</td>
</tr>
<tr>
<td>PRDM16</td>
<td>BCL3</td>
<td>MLL</td>
<td></td>
</tr>
<tr>
<td>CEBPD</td>
<td>ABL1</td>
<td>NUP98</td>
<td></td>
</tr>
<tr>
<td>MYH11</td>
<td>CREBBP</td>
<td>EPS15</td>
<td></td>
</tr>
<tr>
<td>MLLT4</td>
<td>TTL</td>
<td>BCR</td>
<td></td>
</tr>
<tr>
<td>TIRAP</td>
<td>MLLT1</td>
<td>ERG</td>
<td></td>
</tr>
<tr>
<td>ELL</td>
<td>NUP214</td>
<td>AFF1</td>
<td></td>
</tr>
</tbody>
</table>

**Table 4.1:** 74 translocation genes.
<table>
<thead>
<tr>
<th>Translocation gene</th>
<th>Chromosome</th>
<th>Breakpoint Start</th>
<th>Breakpoint Stop</th>
<th>Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCR</td>
<td>chr22</td>
<td>23630284</td>
<td>23637342</td>
<td>Intron 13-17</td>
</tr>
<tr>
<td>ETV6</td>
<td>chr12</td>
<td>12006496</td>
<td>12037378</td>
<td>Intron 4 and 5</td>
</tr>
<tr>
<td>NPM1</td>
<td>chr5</td>
<td>170818804</td>
<td>170819713</td>
<td>Intron 4</td>
</tr>
<tr>
<td>DEK</td>
<td>chr6</td>
<td>18226474</td>
<td>18236682</td>
<td>Intron 9</td>
</tr>
<tr>
<td>HOXA9</td>
<td>chr7</td>
<td>27203461</td>
<td>27204496</td>
<td>Intron 1</td>
</tr>
<tr>
<td>ALK</td>
<td>chr2</td>
<td>29446395</td>
<td>29448326</td>
<td>Intron 19</td>
</tr>
<tr>
<td>BCL11B</td>
<td>chr14</td>
<td>99642533</td>
<td>99697681</td>
<td>Intron 3</td>
</tr>
<tr>
<td>ARHGAP26</td>
<td>chr5</td>
<td>142500713</td>
<td>142513531</td>
<td>Intron 18</td>
</tr>
<tr>
<td>HOXA11</td>
<td>chr7</td>
<td>27222648</td>
<td>27224054</td>
<td>Intron 1</td>
</tr>
<tr>
<td>SEPT11</td>
<td>chr4</td>
<td>77871084</td>
<td>77917577</td>
<td>Intron 1</td>
</tr>
<tr>
<td>PML</td>
<td>chr15</td>
<td>74325756</td>
<td>74326818</td>
<td>Intron 6</td>
</tr>
<tr>
<td>RARA</td>
<td>chr17</td>
<td>38487649</td>
<td>38504567</td>
<td>Intron 2</td>
</tr>
<tr>
<td>TCF3</td>
<td>chr19</td>
<td>1615821</td>
<td>1619109</td>
<td>Intron 16</td>
</tr>
<tr>
<td>PBX1</td>
<td>chr1</td>
<td>164532549</td>
<td>164761730</td>
<td>Intron 2</td>
</tr>
<tr>
<td>ABL1</td>
<td>chr9</td>
<td>133589843</td>
<td>133729451</td>
<td>Intron 1</td>
</tr>
<tr>
<td>CBFB</td>
<td>chr16</td>
<td>67116212</td>
<td>67132612</td>
<td>Intron 5</td>
</tr>
<tr>
<td>MYH11</td>
<td>chr16</td>
<td>15815278</td>
<td>15818850</td>
<td>Intron 28-32</td>
</tr>
<tr>
<td>RUNX1</td>
<td>chr21</td>
<td>36206899</td>
<td>36231770</td>
<td>Intron 21</td>
</tr>
<tr>
<td>MLL</td>
<td>chr11</td>
<td>118353137</td>
<td>118359475</td>
<td>Intron 11</td>
</tr>
</tbody>
</table>

**Table 4.2:** Breakpoints in 19 translocation genes.
Figure 4.4. Comparison of histone modifications and DNase hypersensitivity signals in breakpoint regions of 19 translocation genes and control regions of their control genes. Signal densities of H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3, H3K36me3, and DNase HS in documented breakpoints +/-1kb were calculated (see Methods). The distributions of each mark in translocation genes and control genes were compared in the Kernel density plots. The x-axis of each plot was average read count in each gene (# of reads per kb). The distribution of H3K4me1 in translocations genes showed a statistically significant shift from that in control genes.
Figure 4.5. Histone modification levels in 10 windows (see Methods) around the breakpoint in gene ABL1 (red) compared to histone modification levels in 10 windows in control regions of 100 control genes (blue). ABL1 breakpoint has no characteristic histone modification profile compared to its control regions.
CHAPTER V: DISCUSSION AND PERSPECTIVES

Our genome-wide mapping and comparison of open chromatin in human ESCs and iPSCs provide rich datasets on chromatin structures in pluripotent cells, and illustrate the impact on pluripotency and transcription regulation during cellular differentiation.

Open chromatin regions demarcate active regulatory elements. Different cell types have distinct open chromatin profiles, indicating that different subsets of regulatory elements are utilized for precise regulation of cell-type-specific gene expression. Pluripotent stem cells have a unique gene expression profile dictated by its unique chromatin architecture that directly reflects their pluripotent identity. Both previous and our studies suggest that the pluripotency of embryonic stem cells is largely obtained and maintained by acquiring a particularly “open” state of chromatin. Although hiPSC lines are capable of differentiating into different cell types and are thus pluripotent, they often behave differently than hESCs during and/or after differentiation into one or a few specific cell lineages. Such observations suggest that while cells acquired pluripotency after reprogramming, some fine-tuning of the cell identity and function is not fully achieved. Specifically, reprogrammed cells are often shown to carry epigenetic “memories” of their cell of origin. It has been suggested that some of the parental memories are inherited through reprogrammed cells carrying residuals of DNA methylation patterns of parental cells. In other words, incomplete or incorrect reset of DNA methylation, even subtle, may lead to incomplete reacquisition of pluripotency through reprogramming.
Chromatin in somatic cells also needs to be accurately re-configured to a “pluripotent” open state to achieve complete reprogramming. Our results support the idea that establishing a permissive chromatin state precedes differentiation, and that incomplete establishment of that chromatin state can impede the efficiency of iPSC differentiation. Specifically, (1) sites that are not completely reprogrammed in iPSCs tend to have characteristic chromatin profiles in the parental somatic cells, and (2) the identification of incompletely reprogrammed sites in induced pluripotent cells can be instructive in predicting variations with gene regulation upon iPSC differentiation. We expect that analysis of additional hiPSCs and hESCs will identify overlapping sets of dDHS sites with properties similar to those described here, and we predict that dDHS sites shared by hiPSC lines will be enriched in dCOREs. The exact identity of differences in individual iPS lines will likely affect the efficiency of differentiation into specific somatic lineages. Our identification of regulatory regions that are refractory to reprogramming and their properties will be critical to the understanding and use of iPSCs and embryonic stem cells.

In general, an important question when comparing iPSCs and ESCs is, which differences are functionally relevant? Genomic abnormalities occur in any cell culture including ESCs. Nonpathogenic variations may also occur in vivo such as in blood or bone marrow cells that have been used for transfusion or transplantation to treat patients for decades. Some genomic variations in iPSCs may be unpreventable and harmless for their use in regenerative medicine. Open chromatin maps have been used to guide the identification of functional SNPs that are strongly associated with type II diabetes. The rationale is, active regulatory elements reside in open chromatin regions and thus SNPs that are located in open chromatin regions may be associated with regulatory elements and thus functional. The same rationale can be used to
annotate genetic and epigenetic variation in iPSCs to differentiate functional and non-functional variation. We have shown that several differential DHS sites we identified overlap with previously identified differentially methylated regions between hESCs and hiPSCs. DNase or FAIRE in combination with quantitative PCR, microarray, or high-throughput sequencing can quantitatively assay chromatin accessibility at those loci in iPSC lines or any pluripotent cell lines in general. This can be implemented in a workflow for high-throughput screening and characterization of human pluripotent cell lines [88].

**Genome-wide high-throughput functional annotation of regulatory elements is the key to annotation of human genome**

Genome-wide experiments such as DNase I-, FAIRE-, and ChIP-seq experiments provide evidence for the genomic locations of >100,000 active DNA regulatory elements present in each human cell type [8,31]. A weakness of these assays is that they do not indicate the precise regulatory function of identified DNA elements, and this has been a major bottleneck in characterizing human genome function. Here we presented a method for the highly parallel measurement of the functional activity of putative regulatory elements isolated from open chromatin. It enables us to annotate the function of thousands of putative human regulatory elements in one single experiment, representing orders-of-magnitude improvement in functional characterization of regulatory elements over traditional reporter assays. This study is especially timely given the recent publication of massively parallel reporter assays that interrogate the enhancer activity of tens of thousands of synthesized DNA sequences[95,97,98]. These studies demonstrate the feasibility and great value of high-throughput annotation of enhancer elements. However, such synthesis-based approaches require the knowledge of exact DNA sequences of tested elements prior to the assay, and are designed to test the activities of variants of known
enhancers or putative enhancers. It is thus hard and costly to adapt such strategies to dissect
native regulatory elements that are active in a given cell type. In contrast, our design combines
the use of FAIRE to isolate intact native regulatory elements genome-wide that are active in a
given cell type, and a highly parallel enhancer reporter assay of these elements. More
importantly, our approach can be used to annotate native regulatory elements derived from
tissues in comparable cell lines. Although there were several flaws in the design of our initial
experiment as discussed in Chapter III, we are confident that with the modifications we proposed
and different reporter constructs, this approach can be broadly applied to test the function of all
putative regulatory elements that are active in the tested cell line, with the number likely on the
order of 1,000-10,000.

In addition, this method is versatile and can be easily adapted to test promoter, insulator,
or silencer function of regulatory elements (Figure 5.1). In combinations with different reporter
Gateway vectors, FAIRE fragments in one Gateway entry vector can be easily transferred to
different reporter vectors to test for promoter, enhancer, insulator, and silencer function in
parallel. In our study, we used an enhancer reporter construct (Figure 5.1B). It is feasible to
build on existing constructs and elements of known behavior to build reporters for promoters,
enhancers, silencers and insulators, as illustrated in Figure 5.1. For promoter activity, sequences
can simply be cloned upstream of the reporter (Figure 5.1A). Testing for silencer activity
(Figure 5.1C) can be challenging. One possibility is to use the well-annotated beta-globin locus.
The silencer activity of fragments can thus be measured by their ability to quench transcription
from beta-globin promoter region (spanning from -255 to +50 containing the TATA box (-30nt),
CAAT box (-70nt), and CAC box motif (-90nt), which are necessary for efficient transcription)
127-129. The native “Silencer II” element that naturally located between -302 and -294nt (5’-
TTCAATATG-3') upstream of human adult beta-globin gene can be used as a positive control. Such silencer system has been shown to work in other cell types 132. The E1B promoter used in our existing enhancer reporter construct can be used in the insulator assay (Figure 5.1D). Known insulator elements such as the chicken HS4 beta-globin 5' boundary element 17,133 can be used as positive controls. Data collected from such assays of different regulatory actives will provide a far more complete inventory of the location and function of regulatory elements in the tested cell line.

**Figure 5.1.** Schematic representation of reporter constructs to be employed for identification of regulatory function. The expected result, GFP positive (green) or negative (white) is indicated for negative controls, positive controls, and the tested fragments. * indicates use in initial experiments.
BIBLIOGRAPHY


87. Francis R, Richardson C (n.d.) Multipotent hematopoietic cells susceptible to alternative double-strand break repair pathways that promote genome rearrangements. genesdevshlporg.


91. Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a


180. Sheffield NC, Thurman RE, Song L, Safi A, Stamatoyannopoulos JA, et al. (2013) Patterns of regulatory activity across diverse human cell types predict tissue identity,


Zagranichnaya TK, Wu X, Danos AM, Villereal ML (n.d.) Gene expression profiles in HEK-293 cells with low or high store-operated calcium entry: can regulatory as well as regulated genes be identified? physiolgenomicsphysiologyorg.


