UNIVERSITY OF CALIFORNIA, SAN DIEGO

Understanding the global architecture of gene regulation in human cells through analysis of chromatin signatures

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy

in

Bioinformatics

by

Gary Chung Hon

Committee in charge:
Professor Bing Ren, Chair
Professor Wei Wang, Co-Chair
Professor Vineet Bafna
Professor Trey Ideker
Professor James Kadonaga

2009

The Dissertation of Gary Chung Hon is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California, San Diego

2009

DEDICATION

To my mom and dad, for instilling in me as a child the wonder of science and the thrill of learning.

To Sourav and Ali, whose friendship on and off campus maintained my sanity early in my graduate career.

And of course to Amy, to whom I owe the rest of my sanity.

EPIGRAPH

When I'm working on a problem, I never think about beauty. I think only
how to solve the problem. But when I have finished, if the
solution is not beautiful, I know it is wrong.

R. Buckminster Fuller

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SCHEMES

LIST OF TABLES

ACKNOWLEDGEMENTS

This thesis would not have been possible without the tireless work of members of the Ren lab in performing all the biological experiments to create the rich genome-scale maps that I spent the past several years analyzing. I would like to especially thank Nathaniel Heintzman and David Hawkins for leading the ChIP-chip efforts, for the wonderful scientific interactions we have had both in and out of lab, and for their friendship. I would also like to thank my committee members Vineet Bafna, Trey Ideker, Jim Kadonaga, and Wei Wang for their helpful suggestions and guidance to make this thesis stronger. Finally, I am grateful to my advisor Bing Ren, whose mentorship has made graduate school an intellectually stimulating and enjoyable experience.

Chapter 2, in full, is a reprint of the material as it appears in Nature Genetics 2007. Heintzman, Nathaniel D ; Stuart, Rhona K ; Hon, Gary ; Fu, Yutao ; Ching, Christina W ; Hawkins, R David ; Barrera, Leah O; Van Calcar, Sara ; Qu, Chunxu ; Ching, Keith A ; Wang, Wei ; Weng, Zhiping ; Green, Roland D ; Crawford, Gregory E ; Ren, Bing. “Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome”, Nature Genetics, vol. 39, 2007. The dissertation author was a primary investigator and author of this paper. Specifically, the dissertation author performed the computational analysis to identify chromatin signatures and used these signatures to develop a method to identify enhancers and promoters.

Chapter 3, in full, is a reprint of the material as it appears in Nature 2009. Heintzman, Nathaniel D ; Hon, Gary C ; Hawkins, R David ; Kheradpour, Pouya; Stark, Alexander ; Harp, Lindsey F ; Ye, Zhen ; Lee, Leonard K ; Stuart, Rhona K ; Ching, Christina W ; Ching, Keith A ; Antosiewicz-Bourget, Jessica E ; Liu, Hui ; Zhang, Xinmin ; Green, Roland D ; Lobanenkov, Victor V ; Stewart, Ron ; Thomson, James A ; Crawford, Gregory E ; Kellis, Manolis ; Ren, Bing. “Histone modifications at human enhancers reflect global cell-type-specific gene expression”, Nature, vol. 458, 2009. The dissertation author was a primary investigator and author of this paper. Specifically, the dissertation

author performed the computational analysis of histone modifications including work on the cell-type specificity of different function elements, predicting enhancers genome-wide, and analyzing the influence of enhancers on gene expression.

Chapter 4, in full, has been submitted for review in Cell. Hawkins, R David ; Hon, Gary C ; Yang, Chuhu ; Antosiewicz-Bourget, Jessica E ; Lee, Leonard K ; Ngo, Que-Minh ; Ching, Keith A ; Edsall, Lee E ; Ye, Zhen ; Kuan, Samantha ; Yu, Pengzhi ; Liu, Hui ; Zhang, Xinming ; Green, Roland D ; Lobanenkov, Victor V ; Stewart, Ron ; Thomson, James A ; and Ren, Bing. “Chromatin States in Human ES Cells Reveal Key Regulatory Sequences and Genes Involved in Pluripotency and Self-renewal”. The dissertation author was a primary investigator and author of this paper. Specifically, the dissertation author performed the computational analysis of histone modifications including work on the cell-type specificity of different function elements, predicting enhancers genome-wide, and analyzing the influence of enhancers on gene expression.

Chapter 5, in full, is a reprint of the material as it appears in PLoS Computationa Biology 2008. Hon, Gary ; Ren, Bing ; Wang, Wei. “ChromaSig: A Probabilistic Approach to Finding Common Chromatin Signatures in the Human Genome”, PLoS Computational Biology, vol. 4, 2008. The dissertation author was a primary investigator and author of this paper. The dissertation author performed all the analysis presented.

Chapter 6, in full, has been submitted for review in PLoS Computational Biology. Hon, Gary ; Ren, Bing ; Wang, Wei. “Discovery and annotation of functional chromatin signatures in the human genome”. The dissertation author was a primary investigator and author of this paper. The dissertation author performed all the analysis presented.

VITA

2003	Bachelor of Arts in Computer Science, University of California, Berkeley
2003	Bachelor of Arts in Molecular and Cell Biology, University of California, Berkeley
2009	Doctor of Philosophy in Bioinformatics, University of California, San Diego

PUBLICATIONS

Gary Hon, Wei Wang, Bing Ren. Discovery and annotation of functional chromatin signatures in the human genome. Submitted.

R. David Hawkins*, Gary C. Hon*, Chuhu Yang, Jessica E. Antosiewicz-Bourget, Leonard K. Lee, Keith A. Ching, Que-Minh Ngo, Pengzhi Yu, Hui Liu, Xinming Zhang, Roland D. Green, Victor V. Lobanenkov, Ron Stewart, James A. Thomson, and Bing Ren. Chromatin States in Human ES Cells Reveal Key Regulatory Sequences and Genes in Pluripotency and Self-renewal. Submitted.

Nathaniel D. Heintzman*, Gary C. Hon*, R. David Hawkins*, Pouya Kheradpour, Alexander Stark, Lindsey F. Harp, Zhen Ye, Leonard K. Lee, Rhona K. Stuart, Christina W. Ching, Keith A. Ching, Jessica E. Antosiewicz, Hui Liu, Xinmin Zhang, Roland D. Green, Victor V. Lobanenkov, Ron Stewart, James A. Thomson, Gregory E. Crawford, Manolis Kellis, Bing Ren. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature. 2009.

Gary Hon, Bing Ren, Wei Wang. ChromaSig: A Probabilistic Approach to Finding Common Chromatin Signatures in the Human Genome. PLoS Comput Biol. 2008 Oct.

Nathaniel D Heintzman, Rhona K Stuart, Gary Hon, Yutao Fu, Christina W Ching, R David Hawkins, Leah O Barrera, Sara Van Calcar, Chunxu Qu, Keith A Ching, Wei Wang, Zhiping Weng, Roland D Green, Gregory E Crawford, Bing Ren. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. 2007 Feb 4.

Teresa Reguly, Ashton Breitkreutz, Lorrie Boucher, Bobby-Joe Breitkreutz, Gary Hon, Chad Myers, Ainslie Parsons, Helena Friesen, Rose Oughtred, Amy Tong, Chris Stark, Yuen Ho, David Botstein, Brenda Andrews, Charles Boone, Olga Troyanskya, Trey Ideker, Kara Dolinski, Nizar Batada, Mike Tyers. Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. J Biol. 2006 Jun 8.

ABSTRACT OF THE DISSERTATION

Understanding the global architecture of gene regulation in human cells

through analysis of chromatin signatures

by

Gary Chung Hon

Doctor of Philosophy in Bioinformatics

University of California, San Diego, 2009

Professor Bing Ren, Chair

Professor Wei Wang, Co-Chair

There are over 200 cell types in the human body, each with a unique gene expression program precisely controlled by regulatory elements encoded in the genome such as promoters, enhancers, and insulators. Methods to identify functional genomic elements have widely focused on sequence. While these methods have been successful in finding promoters and insulators, identifying other regulatory elements, namely enhancers, is still an open problem. Our understanding of human transcription is incomplete because we do not have a complete catalog of enhancers. Recently, it has become increasingly clear that an epigenetic layer of information, especially in the form of post-translational histone modifications, marks different functional regions of the genome.

In Chapter 1, I use high-resolution maps of histone modifications in 1% of the human genome to show that active enhancers are marked by a chromatin signature distinct from promoters, and that this signature can be used to predict other active enhancers. In Chapter 2, I extend this method to predict active enhancers genome-wide in HeLa cells, showing that enhancers are epigenetically more dynamic than promoters or insulators. Marked enhancers are highly enriched near cell-type specifically expressed genes. This key positioning of active enhancers suggests they likely drive cell-type specific gene expression. In Chapter 3, to study a biological system more relevant to human development, I then apply this technique to embryonic stem cells before and after differentiation. Most enhancers display marked changes in chromatin states in a manner that correlates with differential expression of their predicted target genes. In addition, a set of poised enhancers are marked by a distinct chromatin signature near genes important for cell fate determination, underscoring the importance of these regulatory elements in regulating differentiation. Finally, in Chapters 4 and 5, I address the problem of what other chromatin signatures exist besides those at promoters and enhancers. I develop an unbiased *de novo* pattern-finding method called ChromaSig to find commonly occurring chromatin signatures. Applying ChromaSig to genome-wide maps of histone modifications, I find a novel chromatin signature marking exons and other marking distinct classes of repeat elements associated with distinct modes of gene repression.

Chapter 1 : The language of chromatin signatures

Abstract

Proper control of eukaryotic gene expression involves integration of various regulatory signals including sequences encoded in the genome as well as the chromatin that encapsulates them. Accumulating evidence suggests that epigenetic modifications of chromatin plays key roles in this process. Here, I review how the epigenomics field is rapidly progressing from descriptive observations of chromatin modifications at regulatory elements to powerful predictive models enabling use of chromatin signatures to enumerate novel functional elements that have escaped previous detection.

Introduction

Each of the over 200 cell types in the human body contains a nearly identical copy of the genome sequence. Yet the gene expression pattern for each distinct cell type is unique [1,2]. While it is obvious that this uniqueness arises from differences in how transcription is controlled, it is unclear what the mechanisms of this control are, and especially how this control is orchestrated on a global scale. This knowledge will be critical if we are to understand how a cell rewires its transcriptome upon stimulus, especially during early development, as well as how improper control of transcription causes diseases such as cancer.

At the simplest level, gene regulation requires precise control of gene activation and repression. While gene activity is easily assessed by profiling RNA, RNA abundance alone does not specify how transcription is controlled. In eukaryotes, this process involves a host of regulatory elements including non-coding RNAs, enhancers, and silencers [3], most of which have remained undiscovered. However, recently it has become increasingly clear that the epigenetic modifications of the genome, especially of histone tails, are an information-rich upstream indicator of the transcriptional status of every genomic locus in a cell. With recent technological advances making it routine to survey the epigenome on a large scale, the epigenetics field is rapidly moving from examining individual genes to all genes in the human genome. With this extension comes the recent shift from descriptive to predictive models relating chromatin signatures and the regulatory elements they mark, giving new global insights into gene regulation and development by allowing dissection of these processes in unprecedented detail.

Chromatin signatures at gene structures

First through ChIP studies focusing on individual promoters, then through ChIP-chip spanning subsets of the genome [4,5,6,7], and most recently through genome-wide techniques using ChIP-chip or ChIP-Seq [8,9,10], it is now abundantly clear that one of the hallmarks of actively transcribed protein-coding promoters is H3K4me3 (Figure 1-1). Newer technologies have offered higher resolution views, showing clearly that this modification is found on the nucleosomes flanking nucleosome-free regions that coincide with the transcription start sites (TSSs) of actively transcribed genes [4,11,12].

Unlike most epigenetic marks, it is known precisely how the H3K4me3 chromatin signature is deposited at active promoters. Examining a panel of histone methyl-transferases in yeast, Briggs et al observed that only knock-out of the Set1 methyltransferase results in complete loss of H3K4me3 [13]. Then, using ChIP-chip, Ng et al observed that Set1 occupied actively transcribed regions, and specifically is recruited to the active form of RNA polymerase II (RNAPII) bearing a serine-5 phosphorylated tail [14]. This is consistent with observations that both RNAPII and H3K4me3 simultaneously mark most promoters in the human genome and that the genes belonging to these promoters undergo at least transcription initiation [15].

Another canonical histone modification found in genic regions is H3K36me3, which has long been associated with the gene bodies of actively transcribed genes. Owing to the low resolution of traditional ChIP, ChIP-qPCR, and non-overlapping tiled microarrays used in ChIP-chip, this modification was long thought to be just a signal of elongation that is enriched non-specifically throughout the entire transcribed region [15]. Recent observations using higher resolution techniques have instead found that enrichment of H3K36me3 is much higher at exons than introns [16] (Figure 1-1). The profound observation that a chromatin signature marks exons has leant further support to the view that transcription and splicing are coupled events, implying that the complex processes regulating splicing may be controlled at the chromatin level.

Like H3K4me3, the H3K36me3 chromatin signature is intimately tied to RNAPII. During initiation, the carboxy terminal domain (CTD) of RNAPII is phosphorylated at serine-5, which recruits the Set1 protein to catalyze trimethylation of H3K4 [14]. During elongation, serine-5 phosphorylation of the CTD is replaced by serine-2 phosphorylation [17] and, as a result, Set1 is dissociated and H3K4me3 is not deposited in the gene body [14]. Instead, serine-5 phosphorylation recruits Set2 which results in trimethylation of H3K36me3 in the gene body [18].

Thus, it is clear that RNAPII-transcribed elements are generally marked by consistent chromatin signatures. An open question is whether the same chromatin signatures also exist for genes transcribed by the other polymerases. There are several polymerases known to exist in eukaryotes, each of which transcribe a distinct class of functional elements. It is possible that these distinct polymerases deposit distinct epigenetic modifications during transcription. However, the chromatin signatures observed for RNAPII-transcribed genes may arise because of the highly regulated nature of RNAPII transcription. Given that RNAPI and RNAPIII generally transcribe ubiquitously expressed elements such as rRNAs and tRNAs, less regulation of this process is required, which may require fewer epigenetic modifications.

Towards predictive chromatin signatures

A central barrier to our understanding of the human genome is an incomplete annotation of the elements encoded in it. Many human functional elements have been assigned on the basis of sequence homology with other species under the assumption that sequence conservation equates functional conservation [19,20]. These techniques by definition miss human or lineage-specific elements, which are arguably the most important in defining the human genome. Desperately needed are general, cost-efficient methods to identify functional elements in the human genome using only measurements from human cells. The observations that chromatin signatures are found at well-annotated places of the genome and that their presence correlates with activity suggests that examination of the human epigenome can reveal the functional elements contained within it.

The vast majority of epigenomic studies have focused on the descriptive view that functional loci contain chromatin signatures. For example, active promoters are marked by H3K4me3. A much stronger statement would be that H3K4me3 only marks active promoters. This predictive view suggests that the presence of the chromatin signature alone can predict the presence of a specific class of functional element. This second view is much more rigorous, offers a computational strategy to identify functional elements, and outlines specifically how to test hypotheses of function.

Several studies have shown that novel promoters can be identified on the basis of the H3K4me3 chromatin signature. Work from our lab, as well as by other groups, have shown that this mark can be used in conjunction with others to efficiently identify promoters for known and novel protein-coding genes [4,21,22]. Focusing on the 1% of the human genome studied by the ENCODE pilot project [23], we identified 198 places bearing the promoter chromatin signature [4]. While the vast majority of these are recovered by known annotations including CAGE tags, 6 were novel. Finally, using luciferase reporter assays we verified that several of these novel chromatin signature-based promoter predictions showed promoter activity *in vivo*. Similarly, taking advantage of the observations that miRNAs are transcribed by the same machinery as protein-coding genes and have promoters marked by nucleosome-depleted TSSs flanked by H3K4me3, Ozsolak et al were able to precisely map the locations of miRNA TSSs [24].

The most exciting applications of predictive chromatin signatures is in the identification of previously elusive regulatory elements. For example, isolated examples of non-protein-coding RNAs (ncRNAs) such as HOTAIR, which regulates expression of HOX cluster genes [25], have suggested a crucial role of ncRNAs in development. However, studies of ncRNAs are hindered by small catalogs of known ncRNA genes. To address this problem, Guttman et al took advantage of the observation that many non-protein-coding genes are also transcribed by the same machinery as coding genes, with RNAPII as the central component. Since RNAPII deposits H3K4me3 at promoters during initiation and H3K36me3 during elongation to mark the direction of transcription, Guttman et al searched for this chromatin signature in several mouse strains [26]. This approach successfully identified over 1000 ncRNAs including well-known members such as HOTAIR. Subsequent analysis revealed that these ncRNAs show complex expression and regulatory patterns similar to those previously observed for protein-coding genes, suggesting they are functional during mouse development.

Predictive chromatin signatures at enhancers

The epigenetic events associated with genes and gene-proximal elements has been thoroughly investigated, either through using cost-effective closed experimental systems that exclusively survey genic regions [15,27,28] or through analysis of only these regions even when using open experimental systems that survey the entire genome [8,9,29]. But the epigenome outside of genic regions has remained largely unexplored, even though epigenetic events outside of genes likely contribute to controlling gene expression.

In eukaryotes, transcription is tightly regulated by the activity of transcription factors, the vast majority of which bind to enhancers far from the genes they activate. As such, identifying active enhancers on a genome-wide scale has been an open problem. Our lab had previously shown that active transcriptional enhancers are marked by a distinct and predictive chromatin signature, central to which is strong enrichment of H3K4me1 [4] (Figure 1-1). Recently, we have used this well-defined chromatin signature to map 55,000 enhancers genome-wide in several human cell lines [22]. Unlike promoters and insulators, the epigenetic modifications marking enhancers are highly cell-type specific in a manner that correlates strongly with cell-type specific gene expression. These results tie the global architecture of chromatin signatures outside genes to regulation of gene expression.

But unlike H3K4me3 and H3K36me3 which mark genic regions, H3K4me1 is relatively understudied. It is unknown what enzyme is responsible for depositing the H3K4me1, or even if this mark arises from *de novo* addition of a methyl group to an unmodified H3K4 residue or demethylation from di or tri-methylated states. The latter would require an intermediate state containing either H3K4me2 or H3K4me3 but not the mono-methylated form that also shows no promoter activity. It is possible that these intermediate states are short-lived and are averaged out over the large population of cells used in high-throughput studies. While we have observed a handful of promoter-distal hypersensitive loci marked with stronger enrichment of H3K4me2 than H3K4me1 that may be places being demethylated to the mono-methylated form, this evidence is anecdotal at best and does not convincingly demonstrate the phenomenon on a large scale.

It will be intriguing to learn what large complex, if any, is physically associated with the molecule responsible for depositing H3K4me1. This complex will likely be the key regulator determining enhancer activity. While H3K4me3 and H3K36me3 are linked to RNAPII, it is unlikely that RNAPII plays a similar role with maintenance of H3K4me1 given that the vast majority of enhancers are not enriched for RNAPII. This factor or set of factors must satisfy several conditions.

First, given that enhancers are typically found far from gene regions, the factor cannot be limited to gene regions. Second, being a general factor required for enhancer activity, it must be ubiquitously expressed in all cell types. Third, the factor must be capable of binding inactive, unmarked enhancers. Pioneer factors such as FoxA1, which can bind repressed enhancers enveloped by heterochromatin and open them for activity, or proteins associated with pioneer factors, would satisfy these constraints [30,31].

Thus far, the only epigenetic modification predictive of active enhancers is H3K4me1. Finding other predictive modifications or modifiers of enhancer activity has been an active area of research. Using a technique called GMAT that involves ChIP followed by SAGE-like sequencing, Roh et al identified thousands of acetylation islands marked by H3K9ac or H3K14ac [32]. But since that H3K9ac is known to mark the activity of promoters more than enhancers [22], the majority of these acetylation islands were close to transcription start sites. Although promoter-specific acetylations have been discovered [33], thus far there have been no reports of acetylations specific to enhancers. Instead, acetylation of histones have generally been associated with active chromatin regions that marking both promoters and enhancers [22,33].

Using ChIP-Seq to map a panel of histone modifications [9], and Barski et al observed that enhancers were marked by H3K4me3. This apparent enrichment of H3K4me3 at enhancers could be caused by secondary physical interactions between H3K4me3-marked promoters and H3K4me1-marked enhancers as predicted by the looping mechanism of enhancer activity [34]. Indeed, analysis of this data reveals that H3K4me1 enrichment is stronger at enhancers than H3K4me3. H3K4me3 enrichment at enhancers could also be an artifact of ChIP-Seq, which is biased to hypersensitive regions that would mark enhancers [35]. In ChIP-chip studies where the ChIP sample is hybridized together with a genomic control sample, H3K4me3 is rarely observed above background levels at enhancers. Normalization procedures that take into account input control may relieve the observed H3K4me3 enrichment.

Systematic discovery of chromatin signatures

Increasingly, we are coming to appreciate the epigenome as a cell-type specific interpretation of the genetic code, specifying the activity of every part of the genome. The observations that chromatin signatures are predictive of a variety of transcribed elements including promoters, exons, miRNAs, and ncRNAs as well as untranscribed regulatory elements such as enhancers leads one to suspect that novel chromatin signatures may also mark other elements of unique function.

Nucleosome depletion is common among many active regulatory elements including genic promoters and miRNA promoters. We have also observed that enhancers marked by H3K4me1 are also depleted for core histone H3 [4,22]. Like the distribution of nucleosomes around an active TSS, nucleosomes may be well-positioned flanking a region of nucleosome depletion around other active genomic regions. Indeed, we observe a bimodal distribution of H3K4me1 enrichment at predicted enhancers, and most interestingly enrichment of transcription factors is strongest inside the nucleosome free region. Similar observations have been observed at CTCF-bound insulators [36]. Recently developed technologies such as DNase-Seq are enabling the efficient enumeration of all nucleosome free regions in the human genome [37], and systematically examination of the histone modifications around these regions will likely yield novel chromatin signatures of enhancers and other regulatory elements.

Not all functional regions of the genome are expected to be marked by DNase I hypersensitivity, in particular places of the genome that are repressed. One way to find consistent chromatin signatures marking regulatory elements outside of known regulatory regions or annotations is to apply an unbiased search on multiple dimensions of the epigenome simultaneously. We have recently developed a computational technique called ChromaSig to identify chromatin signatures that are commonly found in the epigenome. Consistent with observations in yeast [38], we find that many histone modifications are highly redundant, resulting in only a handful of distinct chromatin signatures in the human genome [39]. In particular, we find that many inactive regions of the genome not marked by DNase I hypersensitivity are simultaneously marked by multiple repressive chromatin modifications, and in particular we also observe distinct classes of repressed regions marked by H3K9me3 or H3K27me3 (unpublished).

The epigenome is constantly changing in response to the cell's many stimuli. In addition to defining how gene expression is presently controlled, the epigenome also details how the cell is ready to respond to environmental or developmental cues to alter its transcriptional output. This poised phenomenon has been well-documented at promoters where a bivalent epigenetic state ensures a poised transcriptional state critical for development [40], and likely also applies to enhancers [4,22,30] and, by extension, to other regulatory elements. Using unbiased approaches to identify which parts of the epigenome change during cellular response will reveal key regulatory elements involved in the process. Most interesting will be identifying which parts of the epigenome are marked both before and after stimulation, but where the marks have significantly changed either in terms of modification types or spatial distribution. These poised elements may be those most critical in defining the cellular response.

Conclusions

To dissect the human genome, we must first enumerate all the regulatory elements encoded by it. Although we know that many classes of functional elements exist, current approaches to map these elements are not general, efficient, accurate, genome-scale, and cell-type specific. A major obstacle in finding these elements from the genome sequence alone is that there are no natural breaks in the sequence that delimit phrases or functional elements. The epigenome is an interpretation of the genome. But while the alphabet of the epigenome is larger that of the genome, analysis of the epigenome is a much more tractable endeavor as the words of histone modification peaks are well-spaced throughout the genome. Furthermore, as the fundamental unit of this chromatin epigenome is the nucleosome, the epigenome is effectively orders of magnitude shorter than the genome, telling the story of the genome in a more compact way without skipping the important features. Well-defined, predictive chromatin signatures offer an elegant framework to comprehensive map all the functional elements in the human genome.

References

1. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, Schultz PG, Hogenesch JB (2002) Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci U S A 99: 4465-4470.

2. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 101: 6062-6067.

3. Maston GA, Evans SK, Green MR (2006) Transcriptional Regulatory Elements in the Human Genome. Annu Rev Genomics Hum Genet 7: 29-59.

4. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39: 311-318.

5. Koch CM, Andrews RM, Flicek P, Dillon SC, Karaoz U, Clelland GK, Wilcox S, Beare DM, Fowler JC, Couttet P, James KD, Lefebvre GC, Bruce AW, Dovey OM, Ellis PD, Dhami P, Langford CF, Weng Z, Birney E, Carter NP, Vetrie D, Dunham I (2007) The landscape of histone modifications across 1% of the human genome in five human cell lines. Genome Res 17: 691-707.

6. Kim TH, Barrera LO, Qu C, Van Calcar S, Trinklein ND, Cooper SJ, Luna RM, Glass CK, Rosenfeld MG, Myers RM, Ren B (2005) Direct isolation and identification of promoters in the human genome. Genome Res 15: 830-839.

7. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B (2005) A high-resolution map of active promoters in the human genome. Nature 436: 876-880.

8. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553-560.

9. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K (2007) High-resolution profiling of histone methylations in the human genome. Cell 129: 823-837.

10. Pan G, Tian S, Nie J, Yang C, Ruotti V, Wei H, Jonsdottir GA, Stewart R, Thomson JA (2007) Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. Cell Stem Cell 1: 299-312.

11. Ozsolak F, Song JS, Liu XS, Fisher DE (2007) High-throughput mapping of the chromatin structure of human promoters. Nat Biotechnol 25: 244-248.

12. Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, Wei G, Zhao K (2008) Dynamic regulation of nucleosome positioning in the human genome. Cell 132: 887-898.

13. Briggs SD, Bryk M, Strahl BD, Cheung WL, Davie JK, Dent SY, Winston F, Allis CD (2001) Histone H3 lysine 4 methylation is mediated by Set1 and required for cell growth and rDNA silencing in Saccharomyces cerevisiae. Genes Dev 15: 3286-3295.

14. Ng HH, Robert F, Young RA, Struhl K (2003) Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. Mol Cell 11: 709-719.

15. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA (2007) A chromatin landmark and transcription initiation at most promoters in human cells. Cell 130: 77-88.

16. Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J (2009) Differential chromatin marking of introns and expressed exons by H3K36me3. Nat Genet.

17. Komarnitsky P, Cho EJ, Buratowski S (2000) Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. Genes Dev 14: 2452-2460.

18. Krogan NJ, Kim M, Tong A, Golshani A, Cagney G, Canadien V, Richards DP, Beattie BK, Emili A, Boone C, Shilatifard A, Buratowski S, Greenblatt J (2003) Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II. Mol Cell Biol 23: 4207-4218.

19. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536-540.

20. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola

AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.

21. Won KJ, Chepelev I, Ren B, Wang W (2008) Prediction of regulatory elements in mammalian genomes using chromatin signatures. BMC Bioinformatics 9: 547.

22. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching KA, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenkov VV, Stewart R, Thomson JA, Crawford GE, Kellis M, Ren B (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature.

23. ENCODE_Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306: 636-640.

24. Ozsolak F, Poling LL, Wang Z, Liu H, Liu XS, Roeder RG, Zhang X, Song JS, Fisher DE (2008) Chromatin structure analyses identify miRNA promoters. Genes Dev 22: 3172-3183.

25. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, Chang HY (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129: 1311-1323.

26. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458: 223-227.

27. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG, Gifford DK, Melton DA, Jaenisch R, Young RA (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122: 947-956.

28. Miao F, Natarajan R (2005) Mapping global histone methylation patterns in the coding regions of human genes. Mol Cell Biol 25: 4650-4661.

29. Cui K, Zang C, Roh TY, Schones DE, Childs RW, Peng W, Zhao K (2009) Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. Cell Stem Cell 4: 80-93.

30. Lupien M, Eeckhoute J, Meyer CA, Wang Q, Zhang Y, Li W, Carroll JS, Liu XS, Brown M (2008) FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. Cell 132: 958-970.

31. Eeckhoute J, Lupien M, Meyer CA, Verzi MP, Shivdasani RA, Liu XS, Brown M (2009) Cell-type selective chromatin remodeling defines the active subset of FOXA1-bound enhancers. Genome Res 19: 372-380.

32. Roh TY, Cuddapah S, Cui K, Zhao K (2006) The genomic landscape of histone modifications in human T cells. Proc Natl Acad Sci U S A 103: 15782-15787.

33. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, Zhao K (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet 40: 897-903.

34. Bulger M, Groudine M (1999) Looping versus linking: toward a model for long-distance gene activation. Genes Dev 13: 2465-2477.

35. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol 27: 66-75.

36. Fu Y, Sinha M, Peterson CL, Weng Z (2008) The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. PLoS Genet 4: e1000138.

37. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE (2008) High-resolution mapping and characterization of open chromatin across the genome. Cell 132: 311-322.

38. Liu CL, Kaplan T, Kim M, Buratowski S, Schreiber SL, Friedman N, Rando OJ (2005) Single-nucleosome mapping of histone modifications in S. cerevisiae. PLoS Biol 3: e328.

39. Hon G, Ren B, Wang W (2008) ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. PLoS Comput Biol 4: e1000201.

40. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES (2006) A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell 125: 12.

Figures

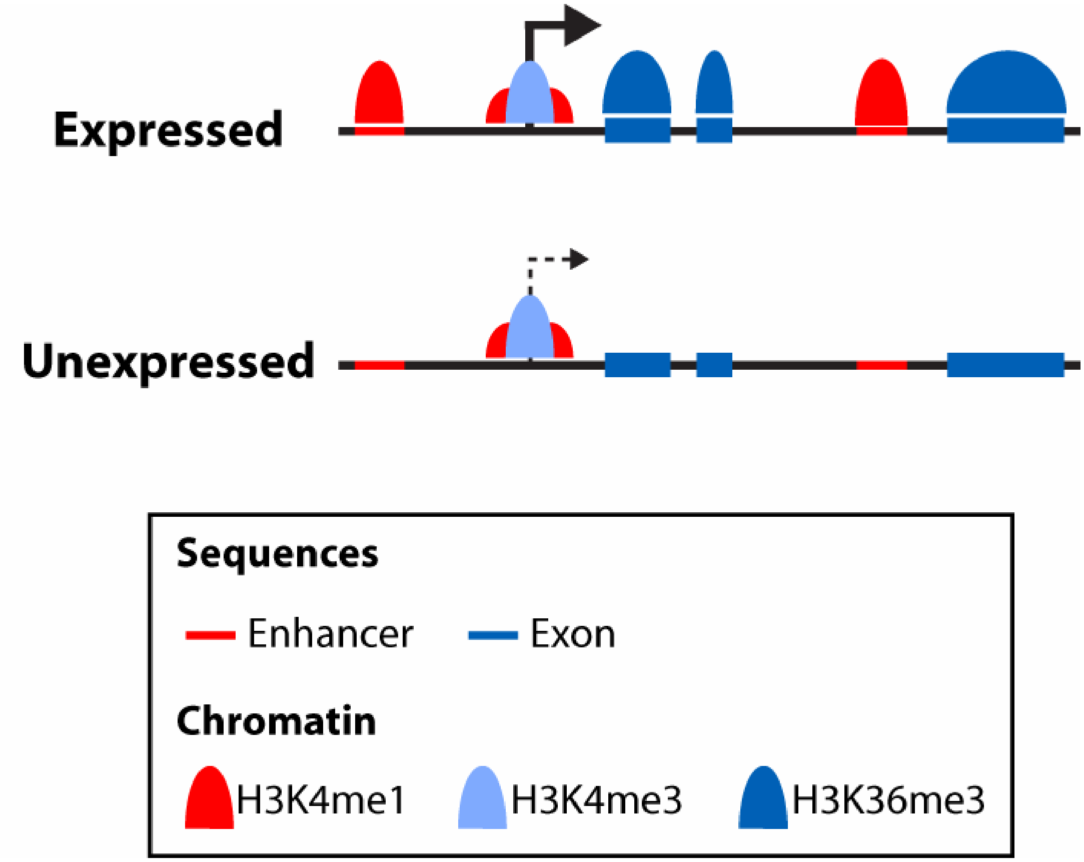


Figure 1-1: Chromatin signatures in the human genome.

Chapter 2 : Distinct and predictive chromatin signatures mark active promoters and enhancers in 1% of the human genome

Abstract

Gene regulation in eukaryotes is implemented by at least two distinct classes of activating elements: gene-proximal promoters and distal enhancers. While promoters have been extensively studied, enhancers have not, largely owing to our inability to identify them on a large scale. Both of these elements are defined by sequences encoded in the genome, but their activities vary in a cell-type dependent manner. While many recent studies have linked activation and repression of promoters with chromatin structure through histone modifications, an open question is whether other regulatory elements such as enhancers are similarly affected. Here, I report that promoters and enhancers are associated with distinct chromatin signatures that can be employed to predict these classes of regulatory elements in the human genome. Using a combination of chromatin immunoprecipitation and microarray experiments (ChIP-chip), my lab generated high-resolution maps of histone modifications in 1% of the human genome. Examining the histone modification features at known promoters and enhancers, I find that active promoters are marked by a peak of H3K4me3 at the TSS with flanking enrichment of H3K4me1. In contrast, enhancers are marked by H3K4me1 but not H3K4me3. I then developed a computational prediction algorithm employing the distinct chromatin signatures to identify new promoters and enhancers. This allowed accurate prediction of over 200 promoters and 400 enhancers in 1% of the human genome. This approach correctly predicted 84% of the regulatory elements bound by the transcription factor STAT1 as well as a novel enhancer for the carnitine transporter SLC22A5 gene. These results reveal chromatin patterns for distinct classes of transcriptional regulatory elements, offering insights into the functional relationships between chromatin modifications and regulatory activity in human cells and providing a new resource for the functional annotation of the human genome.

Introduction

Control of gene expression requires precise regulation of gene activation and repression. In eukaryotes, several distinct classes of regulatory elements encoded within the genome control transcription [1]. Promoters, which are found at the 5' ends of genes immediately surrounding the transcription start sites (TSS), serve as the point of assembly of the transcriptional machinery and initiate transcription [2,3]. Enhancers, which are often located far from promoters and can also be hundreds of kilobases away from the genes they regulate, are bound by transcription factors and coactivators to activate gene expression at promoters through what is thought to be a looping mechanism [4,5,6,7]. Insulators, which are bound by the CTCF protein, serve as insulators to block enhancer activation [8,9]. Importantly, all of these elements are defined by static sequence elements encoded directly within the genome [1]. For example, promoters often contain well-defined sequence elements such as TATA boxes, initiator elements, and downstream promoter elements that are recognized by the co-factors of RNA polymerase II [10]. In contrast, enhancers often contain motifs recognized by one of the thousands of transcription factors encoded by a eukaryotic genome , while insulators are marked by the well-defined CTCF motif [9].

Since the sequencing of the human genome, many groups have searched for these regulatory elements using sequence alone [11,12,13]. While these efforts have been largely successful for promoters and insulators, it has been much more difficult to find enhancers. There are several reasons for this difficulty: enhancers are often defined by short motifs that are often highly degenerate [14,15], these elements are usually far from the well-studied parts of the genome [5], and because the lack of large-scale maps of enhancers has made construction of adequate training sets difficult. But, even if all of these elements could be enumerated using sequence information alone, lacking cellular context it would still be unclear how these elements contribute to controlling gene expression in a cell-type dependent manner.

Increasingly, it has become clear that the epigenetic features of the chromatin around regulatory sequences are a barometer of their activity [16,17,18]. This is especially true of histone modifications [3,19,20]. For example, recent epigenetic work largely focusing on promoters has revealed that tri-methylation of lysine 4 of histone H3 (H3K4me3) marks active promoters, while other promoters marked by H3K27me3 are generally repressed. Interestingly, in stem cells, promoters marked by both these modifications are poised to become either activated or repressed upon differentiation [21]. These modifications are also highly conserved across species including human [3], mouse [21], and yeast [19]. Thus, it is clear that deciphering the regulatory information encoded in the genome will require a thorough understanding of the relationships between the transcriptional activities of these different types of cis-regulatory sequence elements and the epigenetic features of the chromatin surrounding them.

Significant progress in the fields of epigenetics and chromatin biology suggests a histone code of ever-increasing complexity with profound implications of chromatin in a variety of biological processes [22]. While some studies suggest that distal regulatory elements such as enhancers may be marked by similar histone modification patterns [23,24,25,26], the distinguishing chromatin features of promoters and enhancers have yet to be determined, hindering our understanding of a predictive histone code for different classes of regulatory elements. Here, I present high-resolution maps of multiple histone modifications and transcriptional regulators in 1% of the human genome, revealing that active promoters and enhancers are associated with distinct chromatin signatures that can be used to predict these regulatory elements in the human genome.

Results

Genome-scale maps of histone modifications

To generate large-scale maps of histone modifications, my lab performed ChIP-chip analysis [27] in 1% of the human genome (totaling 30 megabase pairs) selected by the ENCODE Consortium [28]. They mapped the patterns of core histone H3 and five histone modifications: pan-acetylation of histone H3 lysine 9/14 (H3ac), pan-acetylation of histone H4 lysine 5/8/12/16 (H4ac), and mono-, di-, and tri-methylated histone H3 lysine 4 (H3K4me1, H3K4me2, H3K4me3). They also examined binding of two components of the basal transcriptional machinery (RNAPII and TAF1) and the transcriptional coactivator p300 to identify active promoters and enhancers, respectively. Three biological replicate ChIP-chip experiments were carried out for each marker in HeLa cells before and after treatment with interferon-gamma (IFNg), as p300 is known to be involved in the cellular response to this cytokine [29]. ChIP samples were amplified, labeled, and hybridized to tiling oligonucleotide microarrays covering the non-repetitive sequences of the ENCODE regions at 38-bp resolution.

I performed both within-array and between-array normalization of the ChIP-chip data using existing methods (normalizeWithinArrays, normalizeBetweenArrays, lmFit) from the R package limma from Bioconductor [30]. After scanning and image extraction, Cy5 (ChIP DNA) and Cy3 (input) signal values were normalized within each array by applying either intensity-dependent Loess correction based on control probes or median-scaling normalization. To combine replicates, I used quantile-normalization between arrays and a linear model-fitting strategy to estimate an average log ratio for each probe. The results from this normalization are average enrichments for each marker at every probe, giving a high-resolution map of histone modifications and transcriptional regulator binding for 1% of the human genome.

To validate the ChIP-chip results, my lab performed conventional ChIP against RNAPII and tested for enrichment at 121 sites in the ENCODE regions using quantitative real-time PCR, indicating an accuracy of 97%, a specificity of near 100%, and a sensitivity of 82% for the method. These values are comparable to other ChIP-chip studies [3,9,20,25] and confirm that the ChIP-chip data is very reliable.

Descriptive chromatin signatures at promoters

Looking at individual genes, there is clear enrichment of various chromatin features at promoters. For example, the promoter of the actively transcribed RFX5 gene contains strong enrichment for several activating histone modifications including H3K4me2, H3K4me3, H3ac, and H4ac (Figure 2-1). Consistent with this, the RFX5 promoter is also enriched for transcriptional machinery including RNAPII and TAF1. These data recapitulate results from previous studies [19,20].

To explore chromatin features shared among many human promoters, I then examined ChIP-chip profiles along 10 kb regions centered at well-annotated promoters in the ENCODE regions and performed computational clustering to classify each promoter on the basis of histone modification patterns. I examined only those TSSs corresponding to well-annotated RefSeq [31] transcripts for which my lab had collected expression data. Furthermore, to prevent interference from neighboring genes, I excluded TSSs within 10 kb of each other from the analysis, resulting in a pool of 208 TSSs for clustering. From gene expression profiling experiments and analysis with the MAS5 method, 104 TSSs were defined as active promoters and 104 as inactive promoters.

I divided each of the 10 kb regions around these TSSs into 100 bins of 100 bp in size, and assigned an enrichment value to each bin by averaging the log ratios for the microarray probes within that bin, i.e. the furthest upstream bin contains the averaged log-ratio for all probes -5000 to -4900 bp from the target. As highly repetitive genomic regions are not represented on the microarrays, I used linear interpolation to give values to empty bins, and boundaries were interpolated to zero if necessary. This process was repeated for each marker within each window. I then used K-means clustering with a

Euclidean distance metric over 10000 iterations [32], clustering multiple windows simultaneously by concatenating windows for all markers and weighting all windows equally. Clusters were visualized with Treeview [32].

I observed four distinct classes of promoters in untreated HeLa cells (Figure 2-2a). On a coarse scale, there are essentially two classes of promoters: P2-4 which are highly expressed and marked by a variety of active histone modifications as well as RNAPII and TAF1; and P1 which contain lowly expressed promoters that are generally not enriched for these active epigenetic marks. On a finer scale, expression levels of transcripts within each class generally increase from class P1 to P4, and interestingly this correlates with increased enrichment of all five histone modifications, RNAPII, and TAF1. The patterns observed in HeLa cells treated with IFNg are almost identical (not shown). The transition from H3K4me3 to H3K4me2 to H3K4me1 moving downstream from active promoters into coding regions echoes the pattern seen in small scale studies in human cells [33] and globally in yeast [19,20]. These results confirm previous observations in other organisms that histone modifications are linked to promoter activity.

Interestingly, this analysis revealed a bimodal distribution of all histone modifications centered around peak binding of RNAPII and TAF1 at the TSS, implying depletion of nucleosomes at this position. ChIP-chip data for histone H3 support this conclusion (Figure 2-2a-b). These findings indicate that the nucleosome free region (NFR) observed at promoters in yeast and fly is indeed characteristic of active human promoters, supporting an evolutionarily constrained role for this phenomenon in transcriptional regulation. The degree of nucleosome depletion appears to be related to the level of gene expression, as depletion is not observed in class P1, suggesting that the formation and maintenance of NFRs at active promoters is a regulated process.

Descriptive chromatin signatures at enhancers

Next, I investigated the chromatin features marking transcriptional enhancers. As previous studies have demonstrated that p300 and related acetyltransferases are present at enhancers [23,24], my lab mapped p300 binding in HeLa cells in triplicate ChIP-chip experiments. To identify genomic loci bound by p300, I used the Mpeak tool [3] to find p300-enriched peaks for each normalized replicate array as well as the averaged array. I then defined a putative p300 target to be a peak on the averaged array that 1) has FDR < 0.10 and 2) is within 1 kb of at least one peak with FDR < 0.10 from every normalized replicate array. Using these stringent criteria, I identified 124 binding sites in untreated cells and 182 sites in treated cells.

The p300 binding sites exhibit several known and expected features of enhancers. First, over 75% of p300 binding occurs more than 2.5 kb from Gencode known gene 5'-ends [34], consistent with previous observations that enhancers can act from a distance to activate genes [5]. Second, transcriptional regulatory elements such as enhancers have long been known to exhibit increased nuclease sensitivity [35], so our collaborator Greg Crawford mapped the DNaseI hypersensitive sites (DHSs) in triplicate in HeLa cells along the ENCODE regions using a recently developed DNase-chip method [35]. A significant number of distal p300 sites (69.7%, p < 1e-16) overlap with DHSs, representing ~12% of the distal DHSs identified. Third, over 60% (p < 1e-16) of the distal p300 sites are within 1 kb of a sequence strongly conserved across seven other vertebrates, as defined by a phastCons score > 0.8 [36]. Fourth, a significant number of the distal p300 sites (44.4%, p = 4.6e-15) contain independently predicted regulatory modules (PReMods) identified based on clustering of conserved transcription factor binding motifs [37]. These lines of evidence provide strong support that the distal p300 binding sites represent a subset of enhancers.

Using the distal p300 binding sites to anchor 10 kb regions surrounding each putative enhancer, I performed computational clustering as described above to generate three classes of enhancers (Figure 2-2c-d; classes are arbitrarily named E1-E3 to simplify discussion). Several striking patterns emerge that distinguish enhancers from promoters. Interestingly, H3K4me1 is strongly enriched in a broad pattern at nearly all enhancers at the peak of p300 binding. In contrast, active promoters display a marked depletion of H3K4me1 at the TSS and enrichment more than 1 kb downstream and upstream. Furthermore, enhancers lack enrichment of H3K4me3, which is strongly enriched at promoters. H4ac, H3ac, and H3K4me2 are present in varying degrees at both promoters and enhancers, though the bimodal distribution of these modifications observed at active promoters is less pronounced at enhancers. TAF1 and RNAPII are also present at some enhancers, though more weakly than at promoters, suggesting docking of the transcriptional machinery at enhancers or physical interaction between enhancers and active promoters as proposed in various models of enhancer action [5,38]. This analysis also reveals depletion of histone H3 at enhancers, suggesting that nucleosome depletion is a general feature of both promoters and enhancers, consistent with their DNaseI hypersensitivity. But in spite of some similarities between the histone modification profiles of active promoters and enhancers, the sharp contrasts of their H3K4me1 and H3K4me3 profiles represent distinct chromatin signatures for these different classes of regulatory elements.

From descriptive to predictive models

Thus far, I have shown that active promoters and enhancers each have distinct chromatin signatures. Next, I investigated the possibility that these signatures alone are predictive of these functional elements. If this is true, then searching for these chromatin signatures would be one way of finding enhancers and promoters active in a given cell type.

Training sets were constructed with histone modification profiles surrounding known TSSs and p300 binding sites in untreated HeLa cells and were used to develop a computational prediction algorithm to locate promoters and enhancers in the ENCODE regions based on similarity to the training set chromatin profiles (Figure 2-3a).

In summary, the computational prediction model I developed consists of two stages: 1) use descriptive histone modification profiles of established transcriptional regulatory elements to identify novel elements and 2) apply discriminative filters to classify the predictions as either promoters or enhancers based on their correlation to the distinct chromatin signatures of these elements. An advantage of this two-stage descriptive-discriminative model is that the initial large set of predicted regulatory elements is filtered to remove predictions that do not sufficiently resemble the specific chromatin signatures, resulting in an approach that balances sensitivity and specificity to generate a set of high-confidence putative regulatory elements. Below I explain the development and implementation of the model.

First, I will define some of the elements used in designing the prediction algorithm. The training sets consisted of subsets of the concatenated windows in the TSS cluster (Figure 2-2a) and p300 binding site cluster (Figure 2-2c). Class P1 was excluded due to its uninformative histone modification profiles, resulting in three training sets each for TSSs and p300 binding sites. Training sets were developed using only the data from untreated HeLa cells, leaving the IFNγ-treated cell data as an independent data set for validation of the method. The test set consisted of all 10 kb windows in the ENCODE regions tiled into 100 bp bins, where each window is a concatenation of the histone modification patterns in that 10 kb region (100 average log-ratio bins of 100 bp, as described above for clustering). The data generated in treated cells served as an additional independent test set. I generated prediction sets by scanning the test sets with the 6 training set patterns, scoring the average correlation of each test set window with the training sets (shape parameter) and the sum of the absolute value of all bins in the central 2 kb of each test set window (intensity parameter). The statistical distributions of the shape and intensity parameters were approximated by the normal distribution. A test set window was retained in a prediction set if: 1) it was in the top 1% of the shape distribution, 2) it was in the top 10%

of the intensity distribution, and 3) its average correlation with the training set was higher than all neighboring windows within 1 kb.

Following this enumeration of potential regulatory elements by scanning genomic regions with chromatin signatures, predictions within each functional class (promoter and enhancer) were pooled, and if multiple predictions occurred within 1.5 kb, only the prediction with the highest average correlation was retained. To reduce false positives and ambiguous predictions, I implemented two filters to generate descriptive and discriminative sets of predictions. First, to be retained in the descriptive set, a prediction must have a correlation of at least 0.4 with the average profile of one of the training sets (this threshold was determined by examining the sensitivity and specificity of recovery for active Refseq promoters or known p300 binding sites over a range of correlation values). Additionally, to be retained in the descriptive set, the prediction must correlate more strongly to the average profile of one class of training set than any other training set as computed over H3K4me1 and H3K4me3, as these markers are the operative elements of the distinct chromatin signatures for promoters and enhancers. Based on the maximum correlation determined in this stage, the predictions are unambiguously classified as promoters or enhancers, generating high-confidence prediction lists for both classes of regulatory elements.

Since the experimental data spanned six different histone modifications, it was not immediately clear which modifications were of greatest utility for computational prediction of different functional elements. To address this issue, I performed cross-validation to assess the predictive power of each combination of histone modifications. Each training set was divided into ten groups, each group having a different 10% of the full training set withheld. I then used the average profiles of a given combination of histone modifications within each training set group to scan the test set in both strand orientations, tallying how many withheld members of the training set were recovered in the prediction set, how many withheld members were missed, and how many total predictions were made (Figure 2-4). This procedure was performed for each individual histone modification and all possible combinations to determine which combination recovers the greatest percentage of withheld training set members with the fewest relative predictions. While combining several histone modifications generally improved performance, increasing the number of modifications did not necessarily increase the algorithm's predictive power, as adding a histone modification with no information content can actually decrease performance by introducing noise. It should be noted that, owing to the degree of redundancy that exists among certain histone modifications, more than one combination may perform well, so the selection of an optimal combination is somewhat arbitrary. For example, using all 6 histone modifications, 96% of the training set promoters and 78% of the training set p300 sites are recovered, while using the optimal combinations for each training set recovers 95% active promoters and 85% of the p300 binding sites in the training sets. The combination of H3K4me1 and H3K4me3 alone offered a good balance of sensitivity and specificity, and were often present in the optimal combinations. As such, the combination of these two marks was used in subsequent analysis.

Chromatin signatures are predictive of promoters

Using the above approach, a total of 198 active promoters were predicted in the ENCODE regions in untreated HeLa cells, clustered as described previously into four classes (named PI-PIV to distinguish them from the known promoters presented in Figure 2-2) (Figure 2-3b). In HeLa cells treated with IFNg, I predicted 208 promoters, with greater than 90% overlap between the untreated and treated prediction sets (Figure 2-3c), supporting the accuracy of the method in identifying promoters in an independent data set. The untreated prediction set contains 140 (79%) of the 177 active RefSeq promoters within the ENCODE regions and 32 (21%) of 155 inactive RefSeq promoters, and 180 predictions (91%) map to known Gencode gene 5'-ends (Figure 2-3d), indicating a high degree of sensitivity and accuracy of promoter prediction. Promoter predictions in treated cells are distributed very similarly (Figure 2-3e). Comparison with the recent RIKEN human CAGE data set [39] reveal that the vast majority of the predicted promoters are supported by multiple CAGE tags. Even predicted promoters that do not map to a known Gencode 5'-end are largely supported by multiple CAGE tags (50% in untreated cells, 27% in treated cells) or DHSs (83% in untreated cells, 73% in treated cells). It is possible that the inactive promoters identified in this analysis correspond to transcripts expressed at levels below the detection threshold, or these promoters may be poised for activation. Six promoter predictions in untreated HeLa cells (nine predictions in treated cells) do not correspond to any known or putative 5'-ends, but all overlap with DHSs, suggesting that they may represent novel promoters.

Chromatin signatures are predictive of enhancers

From the above results, it is clear that chromatin signatures are predictive of active promoters in human cells. However, the true test of the utility of this predictive method is whether it can also predict enhancers. Using the same method, I predicted 389 enhancers in untreated HeLa cells (Figure 2-5a; enhancer predictions are classified EI-EIV to distinguish them from the p300 binding sites presented in Figure 2-2). As an independent test, I also predicted 324 enhancers in treated cells, with an overlap of 89% between prediction sets. Although the prediction algorithm was trained on histone modification patterns from untreated cells, predictions in treated cells accurately identified 77% of the distal p300 binding sites in treated cells, suggesting that the method is not over-fitting the training data.

Several lines of evidence support the function of these predictions as enhancers. First, over 85% of the predictions are located more than 2.5 kb from known gene 5'-ends (Figure 2-5c), consistent with their predicted function. Second, they are evolutionarily conserved, with 53.3% ($p < 1e-16$) containing a strongly conserved sequence. Third, many predicted enhancers overlap with predicted transcriptional regulatory modules (PReMods) (36.3%, $p = 1.7e-4$). Fourth, a significant proportion of the enhancer predictions (55.3%, $p < 1e-16$) overlap with DHSs, including the well-known HS2 enhancer in the b-globin locus26 (Figure 2-6). Of the 587 TSS-distal DHSs in HeLa cells, 175 (29.8%) are predicted enhancers; the other distal DHSs likely represent additional regulatory elements such as repressors or insulators, or sequences that contribute to chromatin organization. Finally, 86 enhancer predictions in the untreated set (and 116 in the treated set) map to distal p300 binding sites (Figure 2-5d-e) and many others appear to be enriched in p300 binding, but below the threshold of the stringent target selection (Figure 2-5a).

Many predicted enhancers lack p300 binding. Since p300 is only one member of a class of 200 transcriptional co-activators [40], one possibility is that some p300-independent enhancers are bound by another co-activator. To address this possibility, my lab performed additional ChIP-chip experiments to examine binding of TRAP220 (MED1), a component of the Mediator complex that has been shown to occupy enhancers as well as promoters [23,24]. Of 162 TRAP220 binding sites identified in the ENCODE regions, 78 (48.1%) are located far from known 5'-ends of transcripts and may represent potential enhancers. Almost 63% of these distal TRAP220 sites are recovered by the enhancer predictions (Figure 2-5d), and 18 of them are bound by TRAP220 but not p300, confirming the identity of these predicted enhancers. This result suggests that the chromatin-based prediction model is not limited only to enhancers marked by p300. Overall, the majority of predicted enhancers (63.5%) are supported by DNaseI hypersensitivity, binding of p300, binding of TRAP220, or a combination of these features (Figure 2-5f).

Predicted enhancers show *in vivo* enhancer activity

The computational and high-throughput validations described, while suggestive, do not provide conclusive evidence that the predicted enhancers truly function *in vivo* as enhancers. To confirm the potential of this chromatin-based approach to identify enhancers that regulate the activity of target human promoters, my lab examined a novel predicted enhancer located 6 kb upstream of the SLC22A5 (OCTN2) gene (Figure 2-7a). SLC22A5 is a widely expressed gene that codes for a carnitine transporter [41]. While substantial research has been devoted to the role of SLC22A5 in carnitine transport, fatty acid metabolism and related human diseases, very little is known about the transcriptional regulation of this gene. To test if the predicted enhancer regulates SLC22A5, my lab cloned a region of the SLC22A5 locus (L) containing the promoter and predicted enhancer (E) into a luciferase reporter construct and compared its activity to that of the locus without the predicted enhancer (LDE) in transiently transfected untreated HeLa cells. The deletion of the predicted enhancer caused a 2.5-fold reduction in reporter activity (Figure 2-7b), supporting the necessity of this site for full activity of the SLC22A5 promoter. To test whether that the predicted enhancer is sufficient to enhance the SLC22A5 promoter activity, my lab then cloned the predicted enhancer downstream of the luciferase gene in a construct containing the proximal SLC22A5 promoter (PS). The construct from the promoter-enhancer construct (PSE) showed 4.2-fold greater reporter activity than the construct containing only the promoter (Figure 2-7b), confirming that the predicted enhancer is sufficient to increase the activity of this promoter in a position-independent manner. These results suggest that the putative SLC22A5 enhancer identified by a chromatin signature is indeed critical for optimal transcriptional activation of this gene.

To further assess the accuracy of the enhancer and promoter predictions, I compared the predictions to a list of *in vivo* STAT1 binding sites independently mapped in the ENCODE regions, hypothesizing that STAT1 sites are likely to occupy both promoters and enhancers. My lab performed ChIP-chip for STAT1 in HeLa cells before and after IFNg treatment, and validated the results using quantitative real-time PCR. As expected, no STAT1 binding was detected in cells prior to treatment. However, there were 13 high-confidence STAT1 sites in IFNg-treated cells. Seven STAT1 sites map to promoter predictions, four of which map to known TSSs: IRF1 (a known STAT1 target), RPS9, c21orf59, and IFNAR2. All of these genes are expressed in HeLa cells, supporting the accuracy of the active promoter predictions. Four STAT1 sites map to enhancer predictions, while the remaining two are not recovered by any prediction. In all, the prediction model is capable of detecting the majority (>84%) of this independently generated collection of putative regulatory elements.

To validate the novel promoter and enhancer predictions at STAT1 sites, my lab performed reporter assays to examine their functional properties. In all, they examined 2 predicted novel promoters (one of which corresponded to two STAT1 binding sites) (Figure 2-8a), four STAT1 enhancer predictions (Figure 2-8b), and the two non-predicted STAT1 sites (Figure 2-8c). To test for promoter activity, regions containing the STAT1 sites were amplified from genomic DNA and cloned upstream of the luciferase gene in vectors lacking a promoter (Figure 2-8d); to test for enhancer activity, the same fragments were cloned downstream of the luciferase gene into vectors containing the SV40 minimal promoter (Figure 2-8e). Clones were transiently transfected into HeLa cells and assayed for reporter activity before and after treatment with IFNg.

Both STAT1 promoter predictions stimulated reporter activity in the absence of the SV40 promoter when cloned in the upstream position (Figure 2-8d), validating their function as promoters. Three STAT1 enhancer predictions (STAT1.08-.10) stimulated strong reporter activity when cloned in the downstream position (Figure 2-8e) but required the presence of the SV40 promoter, consistent with the positional-independence and promoter-dependence of enhancer activity. The fourth enhancer prediction (STAT1.11) exhibited only weak enhancer activity, though the STAT1 site in this region is further away from the prediction (710 bp) than any of the other STAT1 sites that examined (average ~240 bp). The effect of IFNg is variable among the different sites in both ChIP-chip binding profiles and reporter activity, though there seems to be a relationship between inducibility of p300 binding and reporter activity. The non-predicted sites (STAT1.12, -.13) displayed no functional activity and were not marked by either of the distinctive histone modification patterns (Figure 2-8c), supporting the specificity of the model. It is still possible that these sites are actually regulatory elements that cannot be tested in this system due to their function or a requirement for native chromatin context, but it is worth noting that these are the only two STAT1 sites that did not exhibit DNaseI hypersensitivity.

Since STAT1 only binds after treatment with IFNg, it is surprising that enhancer and promoter chromatin signatures exist at STAT1 binding sites prior to treatment. In fact, all four STAT1-bound enhancers were predicted in both untreated and treated HeLa cells. This implies pre-formation of enhancer chromatin structure to facilitate subsequent transcription factor binding, and suggests that chromatin signatures can identify enhancers in a “poised” state prior to their activation.

Discussion

In summary, analyzing maps of five histone modifications, four general transcriptional factors, and nucleosome density at high resolution in 30 Mbp of the human genome, I identify chromatin features that distinguish promoters from enhancers. While both kinds of regulatory elements share some features such as nucleosome depletion and enrichment of histone acetylation and H3K4me2, the distinct patterns of H3K4me1 and H3K4me3 enrichment at active promoters and enhancers define chromatin signatures that can be used to locate novel regulatory elements in the human genome. The H3K4me1 enhancer signature is present in HeLa cell chromatin at multiple loci whose enhancer activity was functionally validated, including a putative novel enhancer for the SLC22A5 gene.

In recent years, the genome sequences of a growing number of organisms have been obtained, but extracting functional information from these nucleotide sequences remains a challenge, as our knowledge of transcription factor binding motifs is incomplete and current sequence-based computational tools are limited in their ability to predict the regulatory function of genomic sequences. Here, I present a strategy to identify transcriptional regulatory elements on the basis of their epigenetic characteristics, independent of motifs or other sequence features. These chromatin signatures can be used to effectively identify enhancers on a large scale. Notably, even though the prediction model was trained only on data from untreated HeLa cells, the sensitivity of the model in data from IFNg-treated cells supports the utility of this approach in analyzing independent data sets. The results of the functional assays confirm the ability of the prediction model to identify the location and function of novel promoters and enhancers.

In the future, genome-wide maps of chromatin state in conjunction with approaches such as this will allow rapid identification of enhancers, and possibly other regulatory elements, in large eukaryotic genomes such as human. Furthermore, such approaches will be able to identify function genomic elements on a cell-type specific basis, which cannot be determined by approaches relying on genome sequence alone, but which is essential to understanding how these elements function *in vivo*. Extension of this model to additional cell types and other components of chromatin architecture will be useful in determining the mechanisms of enhancer maintenance and function in regulating tissue-specific gene expression, findings which will be particularly important to our knowledge of how epigenetic factors and distal transcriptional regulatory elements contribute to human development and disease.

This approach will also be valuable to the functional annotation of the human genome, as it provides a novel and effective means to locate active transcriptional enhancers that have thus far eluded identification on a large scale. Given the degree of structural and functional conservation of chromatin and histone modifications from yeast to humans, these predictive chromatin signatures may be useful in annotating promoters and enhancers in the genomes of a variety of organisms.

Acknowledgements

Chapter 2, in full, is a reprint of the material as it appears in Nature Genetics 2007. Heintzman, Nathaniel D ; Stuart, Rhona K ; Hon, Gary ; Fu, Yutao ; Ching, Christina W ; Hawkins, R David ; Barrera, Leah O; Van Calcar, Sara ; Qu, Chunxu ; Ching, Keith A ; Wang, Wei ; Weng, Zhiping ; Green, Roland D ; Crawford, Gregory E ; Ren, Bing. “Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome”, Nature Genetics, vol. 39, 2007. The dissertation author was a primary investigator and author of this paper. Specifically, the dissertation

author performed the computational analysis to identify chromatin signatures and used these signatures to develop a method to identify enhancers and promoters.

References

1. Maston GA, Evans SK, Green MR (2006) Transcriptional Regulatory Elements in the Human Genome. Annu Rev Genomics Hum Genet 7: 29-59.

2. Kim TH, Barrera LO, Qu C, Van Calcar S, Trinklein ND, Cooper SJ, Luna RM, Glass CK, Rosenfeld MG, Myers RM, Ren B (2005) Direct isolation and identification of promoters in the human genome. Genome Res 15: 830-839.

3. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B (2005) A high-resolution map of active promoters in the human genome. Nature 436: 876-880.

4. Atchison ML (1988) Enhancers: mechanisms of action and cell specificity. Annu Rev Cell Biol 4: 127-153.

5. Blackwood EM, Kadonaga JT (1998) Going the distance: a current view of enhancer action. Science 281: 60-63.

6. Carter D, Chakalova L, Osborne CS, Dai YF, Fraser P (2002) Long-range chromatin regulatory interactions in vivo. Nat Genet 32: 623-626.

7. Gaszner M, Felsenfeld G (2006) Insulators: exploiting transcriptional and epigenetic mechanisms. Nat Rev Genet 7: 703-713.

8. Bell AC, West AG, Felsenfeld G (1999) The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell 98: 387-396.

9. Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenkov VV, Ren B (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell 128: 1231-1245.

10. Smale ST, Kadonaga JT (2003) The RNA polymerase II core promoter. Annu Rev Biochem 72: 449-479.

11. Kheradpour P, Stark A, Roy S, Kellis M (2007) Reliable prediction of regulator targets using 12 Drosophila genomes. Genome Res 17: 1919-1931.

12. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature 434: 338-345.

13. Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, Lander ES (2007) Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. Proc Natl Acad Sci U S A 104: 7145-7150.

14. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31: 374-378.

15. Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Pruss M, Reuter I, Schacherer F (2000) TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res 28: 316-319.

16. Jones PA, Baylin SB (2007) The epigenomics of cancer. Cell 128: 683-692.

17. Allfrey VG, Faulkner R, Mirsky AE (1964) Acetylation and Methylation of Histones and Their Possible Role in the Regulation of Rna Synthesis. Proc Natl Acad Sci U S A 51: 786-794.

18. Agalioti T, Lomvardas S, Parekh B, Yie J, Maniatis T, Thanos D (2000) Ordered recruitment of chromatin modifying and general transcription factors to the IFN-beta promoter. Cell 103: 667-678.

19. Liu CL, Kaplan T, Kim M, Buratowski S, Schreiber SL, Friedman N, Rando OJ (2005) Single-nucleosome mapping of histone modifications in S. cerevisiae. PLoS Biol 3: e328.

20. Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, Lee TI, Bell GW, Walker K, Rolfe PA, Herbolsheimer E, Zeitlinger J, Lewitter F, Gifford DK, Young RA (2005) Genome-wide map of nucleosome acetylation and methylation in yeast. Cell 122: 517-527.

21. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES (2006) A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell 125: 12.

22. Jenuwein T, Allis CD (2001) Translating the histone code. Science 293: 1074-1080.

23. Hatzis P, Talianidis I (2002) Dynamics of enhancer-promoter communication during differentiation-induced gene activation. Mol Cell 10: 1467-1477.

24. Wang Q, Carroll JS, Brown M (2005) Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking. Mol Cell 19: 631-642.

25. Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, Huebert DJ, McMahon S, Karlsson EK, Kulbokas EJ, 3rd, Gingeras TR, Schreiber SL, Lander ES (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. Cell 120: 169-181.

26. Roh TY, Cuddapah S, Cui K, Zhao K (2006) The genomic landscape of histone modifications in human T cells. Proc Natl Acad Sci U S A 103: 15782-15787.

27. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA (2000) Genome-wide location and function of DNA binding proteins. Science 290: 2306-2309.

28. ENCODE_Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306: 636-640.

29. Horvai AE, Xu L, Korzus E, Brard G, Kalafus D, Mullen TM, Rose DW, Rosenfeld MG, Glass CK (1997) Nuclear integration of JAK/STAT and Ras/AP-1 signaling by CBP and p300. Proc Natl Acad Sci U S A 94: 1074-1079.

30. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80.

31. Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33: D501-504.

32. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95: 14863-14868.

33. Kouskouti A, Talianidis I (2005) Histone modifications defining active genes persist after transcriptional and mitotic inactivation. EMBO J 24: 347-357.

34. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R (2006) GENCODE: producing a reference annotation for ENCODE. Genome Biol 7 Suppl 1: S4 1-9.

35. Crawford GE, Davis S, Scacheri PC, Renaud G, Halawi MJ, Erdos MR, Green R, Meltzer PS, Wolfsberg TG, Collins FS (2006) DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. Nat Methods 3: 503-509.

36. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034-1050.

37. Blanchette M, Bataille AR, Chen X, Poitras C, Laganiere J, Lefebvre C, Deblois G, Giguere V, Ferretti V, Bergeron D, Coulombe B, Robert F (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res 16: 656-668.

38. Bulger M, Groudine M (1999) Looping versus linking: toward a model for long-distance gene activation. Genes Dev 13: 2465-2477.

39. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest AR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, Chalk AM, Chiu KP, Choudhary V, Christoffels A, Clutterbuck DR, Crowe ML, Dalla E, Dalrymple BP, de Bono B, Della Gatta G, di Bernardo D, Down T, Engstrom P, Fagiolini M, Faulkner G, Fletcher CF, Fukushima T, Furuno M, Futaki S, Gariboldi M, Georgii-Hemming P, Gingeras TR, Gojobori T, Green RE, Gustincich S, Harbers M, Hayashi Y, Hensch TK, Hirokawa N, Hill D, Huminiecki L, Iacono M, Ikeo K, Iwama A, Ishikawa T, Jakt M, Kanapin A, Katoh M, Kawasawa Y, Kelso J, Kitamura H, Kitano H, Kollias G, Krishnan SP, Kruger A, Kummerfeld SK, Kurochkin IV, Lareau LF, Lazarevic D, Lipovich L, Liu J, Liuni S, McWilliam S, Madan Babu M, Madera M, Marchionni L, Matsuda H, Matsuzawa S, Miki H, Mignone F, Miyake S, Morris K, Mottagui-Tabar S, Mulder N, Nakano N, Nakauchi H, Ng P, Nilsson R, Nishiguchi S, Nishikawa S, Nori F, Ohara O, Okazaki Y, Orlando V, Pang KC, Pavan WJ, Pavesi G, Pesole G, Petrovsky N, Piazza S, Reed J, Reid JF, Ring BZ, Ringwald M, Rost B, Ruan Y, Salzberg SL,

Sandelin A, Schneider C, Schonbach C, Sekiguchi K, Semple CA, Seno S, Sessa L, Sheng Y, Shibata Y, Shimada H, Shimada K, Silva D, Sinclair B, Sperling S, Stupka E, Sugiura K, Sultana R, Takenaka Y, Taki K, Tammoja K, Tan SL, Tang S, Taylor MS, Tegner J, Teichmann SA, Ueda HR, van Nimwegen E, Verardo R, Wei CL, Yagi K, Yamanishi H, Zabarovsky E, Zhu S, Zimmer A, Hide W, Bult C, Grimmond SM, Teasdale RD, Liu ET, Brusic V, Quackenbush J, Wahlestedt C, Mattick JS, Hume DA, Kai C, Sasaki D, Tomaru Y, Fukuda S, Kanamori-Katayama M, Suzuki M, Aoki J, Arakawa T, Iida J, Imamura K, Itoh M, Kato T, Kawaji H, Kawagashira N, Kawashima T, Kojima M, Kondo S, Konno H, Nakano K, Ninomiya N, Nishio T, Okada M, Plessy C, Shibata K, Shiraki T, Suzuki S, Tagami M, Waki K, Watahiki A, Okamura-Oho Y, Suzuki H, Kawai J, Hayashizaki Y (2005) The transcriptional landscape of the mammalian genome. Science 309: 1559-1563.

40. Lonard DM, O'Malley BW (2006) The expanding cosmos of nuclear receptor coactivators. Cell 125: 411-414.

41. Schomig E, Spitzenberger F, Engelhardt M, Martel F, Ording N, Grundemann D (1998) Molecular cloning and characterization of two novel transport proteins from rat kidney. FEBS Lett 425: 79-86.

Figures

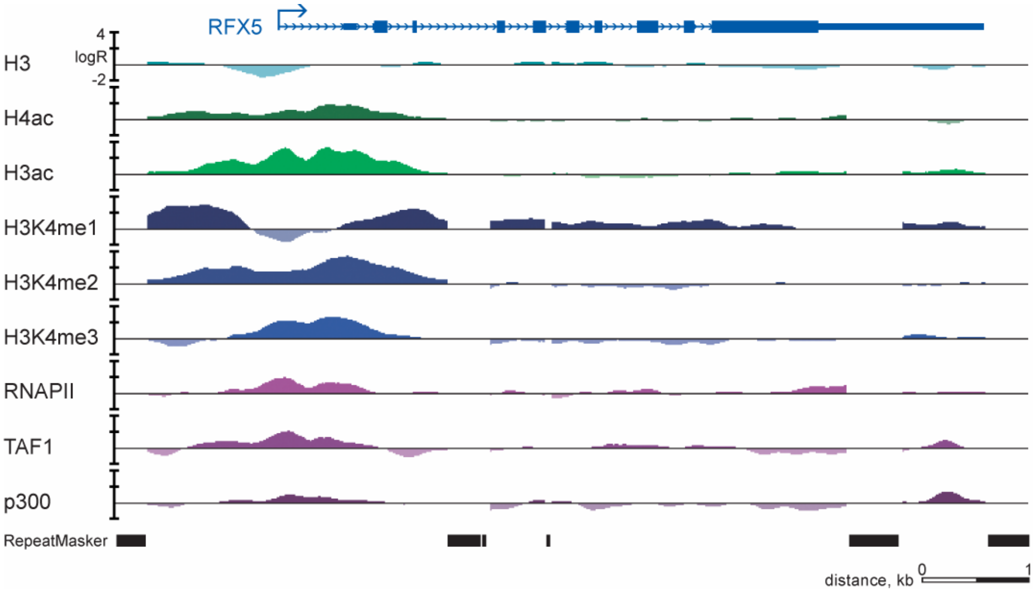


Figure 2-1: A snapshot of ChIP-chip data at the highly expressed RFX5 gene.

ChIP-chip profiles for six chromatin marks, along with RNAPII, TAF1, and p300 are shown as log-ratio of ChIP over input.

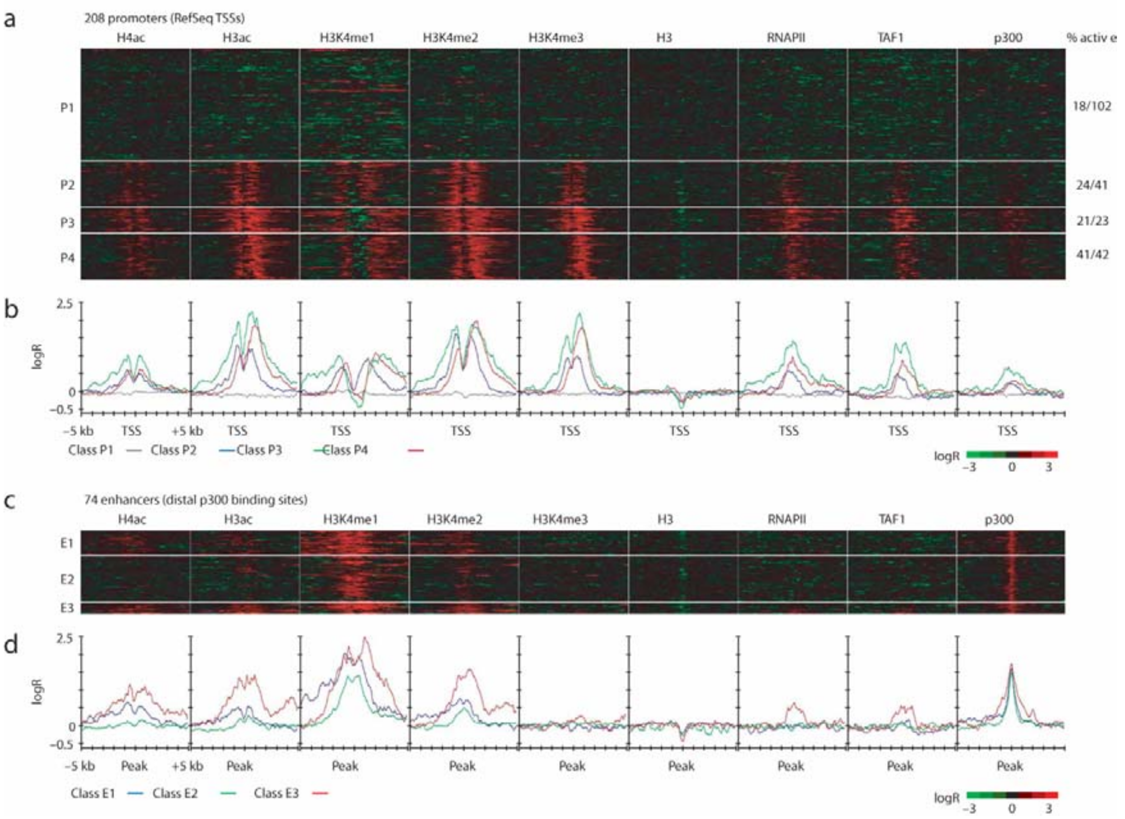


Figure 2-2: Distinct chromatin signatures at promoters and enhancers.

(a) Heat-map representing ChIP-chip enrichment of six chromatin modifications, along with RNAPII, TAF1, and p300 across 10 kb regions centered at 208 promoters in the ENCODE regions. The fraction of promoters belonging to actively expressed genes is shown on the right. (b) Average profile of ChIP-chip enrichment for each mark at promoters. (c) Heatp-map of ChIP-chip enrichment centered at 74 promoter-distal enhancers defined by p300 binding sites. (d) Average profile of ChIP-chip enrichment for each mark at enhancers.

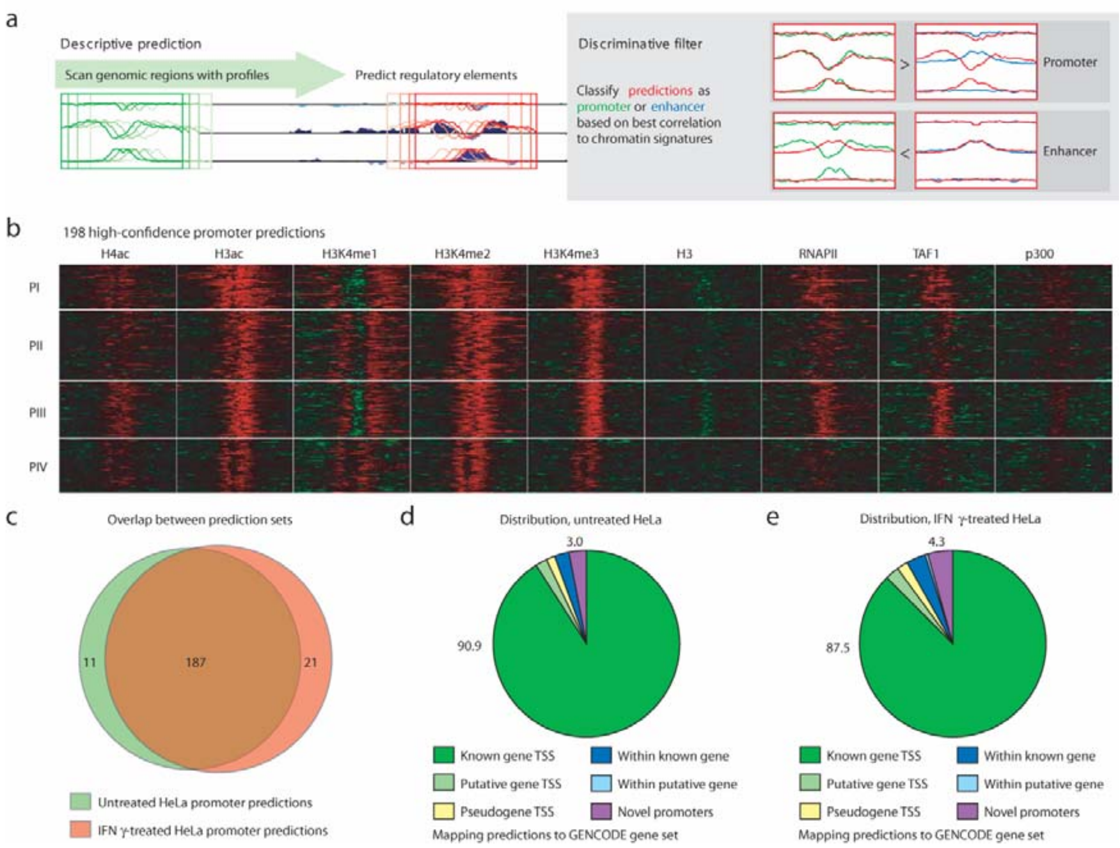


Figure 2-3: Chromatin signatures predict active human promoters.

(a) Schematic of the prediction method whereby training sets of chromatin signatures are used to scan contiguous genomic regions, in conjunction with a series of filtering steps, to predict possible promoters and enhancers. (b) Heat-map representing the ChIP-chip enrichment for 198 high-confidence active promoter predictions in HeLa cells. (c) Overlap between promoters predicted in untreated and IFNγ-stimulated HeLa cells. (d-e) The genic distribution of predicted promoters in (d) untreated and (e) treated HeLa cells as compared to known genes, putative genes, and pseudogenes.

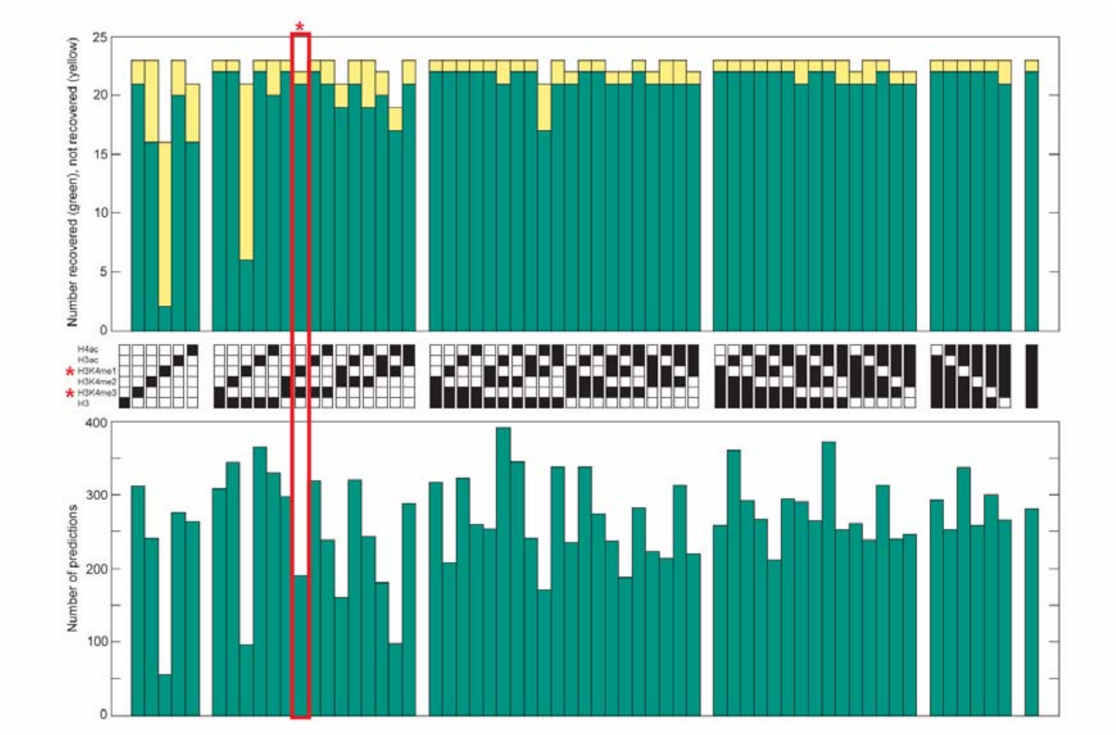


Figure 2-4: Example cross-validation results for promoter prediction.

To determine which set of histone modifications best describes each promoter and enhancer training set, all possible combinations of modifications (depicted as black squares in the middle panel) were used to scan the test set, and the recovery of training set elements was tallied for each combination. This example shows the process for training set class P3 (see Figure 2-2A). The top panel shows the number of training set predictions recovered by each combination (green) and those not recovered (yellow); the total number of elements varies slightly because outliers removed from each group change depending on each combination. The bottom panel shows the total number of predictions made for each combination. The optimal combination, in this case H3K4me1 and H3K4me3, is chosen because it uses the fewest modifications to recover the greatest relative number of training set members with the fewest relative predictions (red box and asterisks). Other combinations may also perform well but are not chosen because of the inclusion of a redundant or non-informative modification or slight loss of sensitivity; the final selection is somewhat arbitrary.

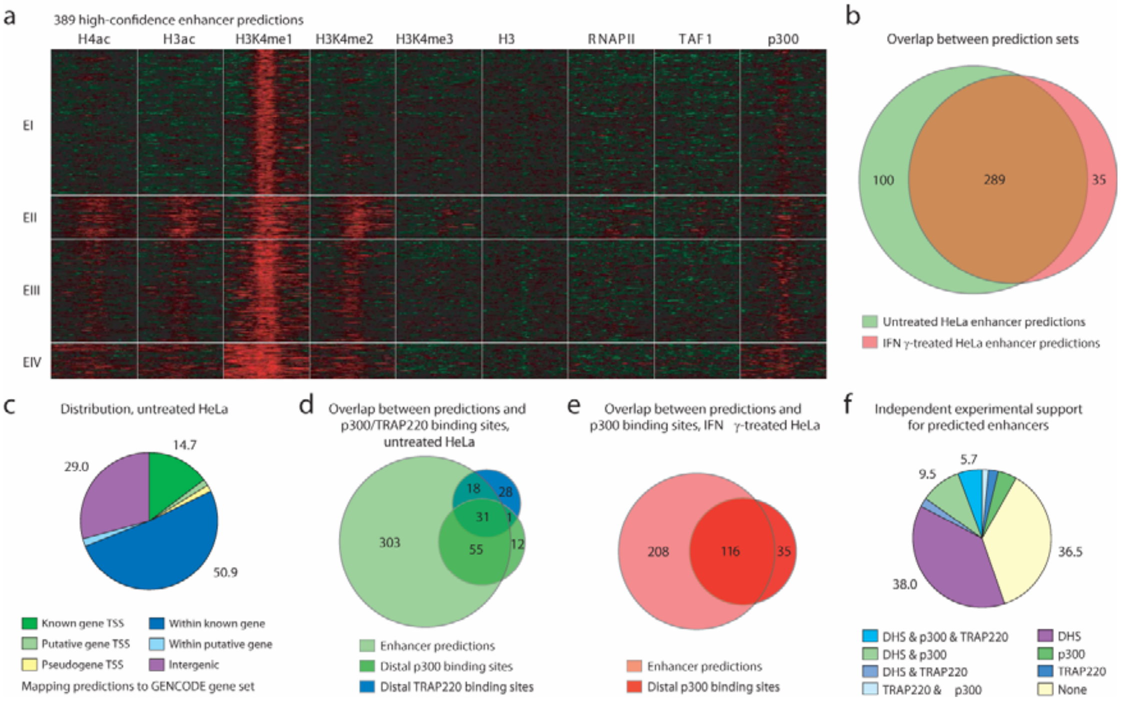


Figure 2-5: Chromatin signatures predict active human enhancers.

(a) Heat-map representing the ChIP-chip enrichment of 389 high-confidence enhancer predictions in untreated HeLa cells. (b) Overlap between enhancers predicted in untreated and IFNγ-stimulated HeLa cells. (c) The genic distribution of predicted enhancers in untreated HeLa cells as compared to known genes, putative genes, and pseudogenes. (d) Overlap of predicted enhancers in untreated HeLa cells with binding of the co-activators p300 and TRAP220 (MED1). (e) Overlap of predicted enhancers in treated HeLa cells with binding of p300. (f) Overlap of predicted enhancers with enhancer hallmarks including DNase I hypersensitivity (DHS), p300 binding, and TRAP220 binding.

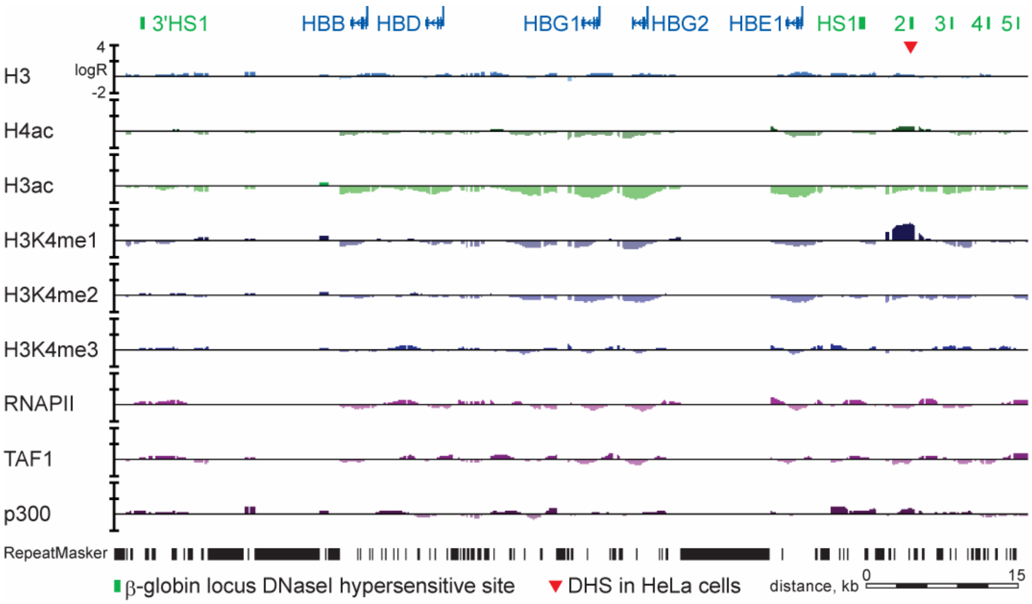


Figure 2-6: Recovery of the known β–globin HS2 enhancer.

The well-known β-globin HS2 enhancer displays DNase I hypersensitivity (DHS) in HeLa cells, and is also marked by H3K4me1. This enhancer is also predicted using the chromatin-signature based method.

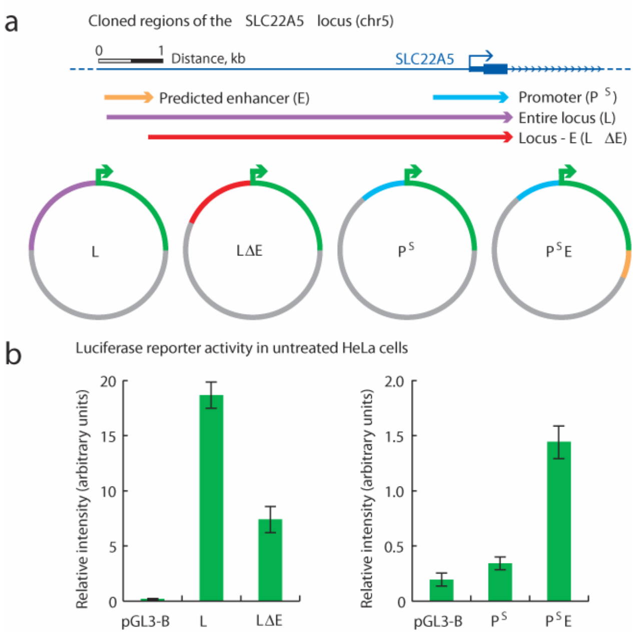


Figure 2-7: Functional validation of an enhancer to the SLC22A5 gene

(a) Schematic of the SLC22A5 promoter and upstream region. An enhancer predicted by chromatin signatures (E) is 6 kb upstream of the TSS. Shown below are various reporter constructs to assess the activity of the enhancer. (b) On the left is relative luciferase activity for the construct containing the entire locus including E compared to the same construct lacking E. On the right is relative luciferase activity of the SV40 core promoter with and without E cloned downstream.

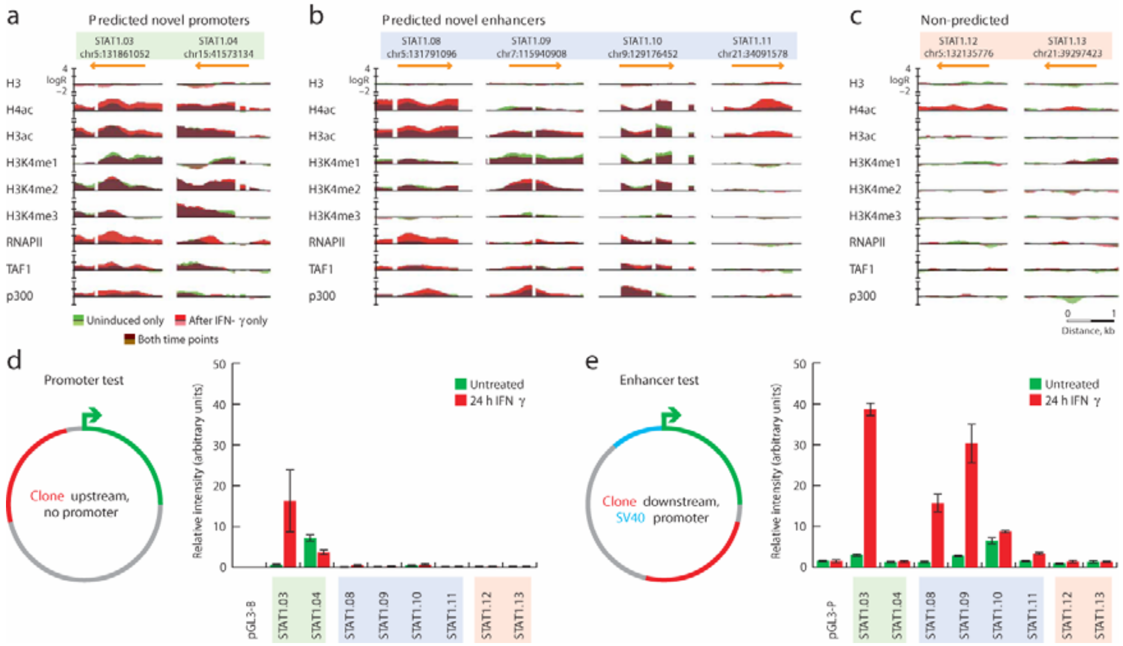


Figure 2-8: Validation of predicted novel promoters and enhancers.

ChIP-chip enrichment of (a) two predicted novel promoters bound by STAT1, (b) four predicted novel enhancers bound by STAT1, and (c) two unrecovered STAT1 binding sites. (d) On the left is the reporter construct to test promoter activity, whereby each locus is cloned directly upstream of the luciferase gene. On the right is relative luciferase activity for each locus. (e) On the left is the reporter construct to test enhancer activity, whereby each locus is cloned downstream of the luciferase gene. On the right is the relative luciferase activity for each locus.

Chapter 3 : Histone modifications at human enhancers reflect global cell-type-specific gene expression

Abstract

The human body is composed of diverse cell types with distinct functions. While it is known that lineage specification depends on cell specific gene expression, which in turn is driven by promoters, enhancers, insulators and other cis-regulatory DNA sequences for each gene [1,2,3], the relative roles of these regulatory elements in this process is not clear. My lab and I have previously developed a chromatin immunoprecipitation-based microarray method (ChIP-chip) to identify promoters, enhancers and insulator elements in the human genome [4,5,6]. Here, I use the same approach to identify promoters, enhancers and insulator elements in multiple cell types and investigate their roles in cell type-specific gene expression. I observed that chromatin state at promoters and CTCF-binding at insulators are largely invariant across diverse cell types. By contrast, enhancers are marked with highly cell type-specific histone modification patterns, strongly correlate to cell type-specific gene expression programs on a global scale, and are functionally active in a cell type-specific manner. These results defined over 55,000 potential transcriptional enhancers in the human genome, significantly expanding the current catalog of human enhancers, and highlight the role of these elements in cell type-specific gene expression.

Introduction

The human body consists of more than 200 different cell types. While the genomes of all cells are virtually identical, each cell performs distinct functions due to the unique set of genes it expresses, which ultimately is specified by how each cell precisely regulates its transcriptional output. Transcriptional regulation of eukaryotic gene expression is a complex process that requires precise spatial and temporal coordination of a host of regulatory inputs, including DNA sequence elements, transcription factor and coactivator binding, and chromatin structural features, all of which cooperate to activate transcription from promoter sequences located at the 5’-end of each gene [1,2,3,7].

Complicating our understanding of this process, however, is our incomplete knowledge of the distal *cis*-regulatory elements responsible for appropriate modification and maintenance of gene expression patterns, including enhancers that recruit a complex array of transcription factors and chromatin-modifying enzymes to activate gene transcription, and insulators that regulate enhancer-promoter interactions [2]. The relative roles and contributions of each class of regulatory element in cell type-specific gene expression remain to be resolved.

As regulatory elements are fairly static at the DNA sequence level, their activity is critically dependent on the dynamic chromatin state at the epigenetic level. Recently, chromatin immunoprecipitation-based approaches have revealed that specific histone acetylation and methylation events are localized to functional sequences in the genome, though most studies have focused on promoters [1,8]. For example, trimethylation of histone H3 lysine 4 (H3K4me3) and acetylation of histones H3 and H4 are generally associated with active promoters, while H3K9me3 and H3K27me3 are found at silenced promoters [5,9,10,11,12]. It is generally understood that chromatin modifications play an important role in dynamic transcriptional regulation at promoters, but many questions remain as to how chromatin state affects the activity of other cis-regulatory elements like enhancers.

My lab and I have previously reported that transcriptional enhancers throughout 1% of the human genome (the ENCODE regions [13]) are distinctly marked by monomethylation of histone H3 lysine 4 (H3K4me1), enabling prediction of novel enhancers on a large scale in the human genome based on this chromatin signature [4]. Here, I examine promoters, enhancers, and insulators in five diverse human cell types, discovering that the localization patterns of the insulator-binding protein CTCF and the chromatin signatures at promoters remain largely invariant across cell types, while the chromatin modifications at enhancers are cell type-specific. I extended an enhancer prediction strategy to the entire human genome in two cell types, generating the first genome-wide maps of transcriptional enhancers based on chromatin signatures. These maps reveal global properties of enhancers and support the involvement of many enhancers in cell type-specific gene expression.

Results

Expanded maps of histone modifications

Previously, I demonstrated enhancers could be determined by distinct chromatin signatures of H3K4me1 and H3K4me3 at these functional elements [4]. Focusing on HeLa cells, my lab performed ChIP-chip in the ENCODE regions for various acetylated forms of histone H3. I found that three additional histone modification marks, namely H3K9Ac, H3K18Ac and H3K27Ac are also part of the chromatin patterns at promoters and enhancers. All three acetylation marks localize to active transcription start sites (TSSs), and remain absent, as do other chromatin modifications, at inactive promoters (Figure 3-1a). These results agree with individual promoter studies observing acetylation or hyper-acetylation at active promoters [14,15], as well as with large-scale histone modification studies in yeast [16,17]. TSS-distal p300 binding sites show clear enrichment of H3K18Ac and H3K27Ac, while H3K9Ac is much reduced (Figure 3-1b). These results suggest that H3K9Ac is preferentially associated with active promoters, while H3K18Ac and H3K27Ac are associated with both promoters and enhancers.

Enhancers are marked by cell type-specific chromatin modification profiles across diverse cell types

My lab performed ChIP-chip analysis as previously described [4] to determine binding of CTCF (insulator-binding protein), the coactivator p300, and patterns of specific histone modifications in 5 diverse human cell lines: cervical carcinoma HeLa, immortalized lymphoblast GM06690 (GM), leukemia K562, embryonic stem cells (ES), and BMP4-induced ES cells (dES), focusing on 1% of the human genome selected by the ENCODE Consortium as common targets for genomic analysis [13].

Modulation of chromatin state is a key component of tissue-specific gene expression programs [16,18]. Given the diversity of these five cell lines and the critical role of promoters in regulating gene expression, I hypothesized that the chromatin modifications at promoters would uniquely define each cell type, but I actually observed the opposite. At promoters of 414 genes in the ENCODE regions, I found that the chromatin signatures at promoters are remarkably similar across all cell types (Figure 3-2a). To quantify this, I defined a cell type's enrichment profile as the sum of the log ratio enrichment values of H3K4Me1, H3K4Me3, and H3K27Ac for each promoter. I then calculated the Pearson correlation coefficient between enrichment profiles from different cell types. The enrichment profiles are highly correlated between all pairs of cell types, with an average correlation coefficient of 0.71. This observation also holds at the larger set of Gencode promoters (not shown). Additionally, it has been well-documented that CpG promoters are associated with house-keeping genes, which are ubiquitously expressed and therefore more likely to retain a constant chromatin state. Analyzing each of these different types of promoters, I observe that the correlation of histone modifications at CpG promoters is 0.62 while that at non-CpG promoters is still 0.48, both of which are significantly more correlated than expected at random. The generally invariant nature of the chromatin marks at promoters suggests that epigenetic features at this class of regulatory element are not the dominant drivers of cell type-specific gene expression patterns.

Insulator elements play key roles in restricting enhancers from activating inappropriate promoters, thereby defining the boundaries of gene regulatory domains [19]. Nearly all insulators that have been experimentally defined in the mammalian genome require the insulator binding protein CTCF to function [20]. A previous genome-wide location analysis of the insulator binding protein CTCF in human fibroblasts indicated that predicted insulators (those sites in the genome bound by CTCF) are closely correlated with the distribution of genes, and are highly conserved throughout evolution, consistent with their key role in transcription regulation [6]. Intriguingly, the overlap of predicted insulators in two cell lines in that study (IMR90 lung fibroblast and U937 hematopoietic progenitor cells) was a remarkable 67%, suggesting cell-type invariance. To further investigate this possibility, I investigated CTCF binding sites in the ENCODE regions in each of the five cell types. On average, 517 predicted insulators were recovered in each cell type, with a remarkable average of 82.8% shared between pairs of cell types. Indeed, the CTCF enrichment profiles at 729 non-redundant CTCF binding sites are nearly identical across all five cell types studied here and IMR90 cells (Figure 3-3a), and the average Pearson correlation coefficient between all pairs of profiles is 0.72, comparable to the value observed at promoters. The consistency of CTCF binding appears to extend to the entire genome (Figure 3-4). These results support insulators as being largely cell-type invariant, to a greater degree than previously appreciated. Additionally, none of the histone modifications that I examined were consistently present at predicted insulators.

I then investigated transcriptional enhancers in the ENCODE regions, using two methods. First, my lab performed ChIP-chip in HeLa, K562, and GM cells to identify 411 binding sites for the transcriptional coactivator protein p300, a co-activator known to localize at some enhancers [14,21]. I observed that chromatin modification patterns at distal p300 binding patterns are highly cell type-specific (Figure 3-3b), with an average pair wise Pearson correlation coefficient of -0.07, in sharp contrast to the similarities across cell types at promoters. Consistent with previous findings, these putative enhancers are highly enriched in H3K4me1 but not H3K4me3, and most are also marked by H3K27ac (Figure 3-3b). As p300 marks only a subset of enhancers, I then used a chromatin signature-based prediction algorithm to identify additional enhancers in all five cell types as previously described [4], predicting a total of 1423 enhancers in the ENCODE regions (Figure 3-2b). In addition to the characteristic H3K4me1 enrichment, predicted enhancers are frequently marked by acetylation of H3K27, DNaseI hypersensitivity and/or binding of transcription factors and coactivators, and many contain evolutionarily conserved sequences (Figure 3-5). Unlike promoters and predicted insulators, but similar to p300 binding sites, the chromatin modification patterns at predicted enhancers are largely cell type-specific (Figure 3-2b), with an average Pearson correlation coefficient between all pairs of cell types of just 0.14 . These results agree with previous findings that H3K4me1 is distributed in a cell type-specific manner relative to other histone modifications [18].

Genome-wide prediction of enhancers based on chromatin signatures

The results above indicate that enhancers are the most variable class of transcriptional regulatory element between these five cell types, suggesting that enhancers are of primary importance in driving cell type-specific patterns of gene expression. To identify enhancers on a global scale, my lab performed ChIP-chip throughout the entire human genome as previously described [5,6] to map enrichment patterns of H3K4me1 and H3K4me3 in HeLa cells. I predicted 38716 enhancers on the basis of chromatin signatures in the HeLa genome, of which 36589 (94.5%) were verified by replicate experiments on a condensed enhancer microarray (Figure 3-6). Based solely on their chromatin signatures, these predictions correctly recovered several previously characterized enhancers, including the b-globin HS2 enhancer [22], a distal downstream enhancer for the PAX6 gene [23], and a distal upstream enhancer for the PLAT (t-PA) gene [24] (Figure 3-7a).

The features of enhancers in the HeLa genome are consistent with what I observed previously in the ENCODE regions [4]. Most predicted enhancers (23686, 64.7%, p = 6.6e-208) exhibit strong evolutionary conservation with a PhastCons score > 0.8 [25]. The genomic distribution of the predicted enhancers are distinct from those of promoters: except for a small fraction that overlap with Known Gene 5’-ends, CAGE tags, or CpG islands, the predicted enhancers are distal to promoters, with predominantly intronic (37.9%) or intergenic (56.3%) localization (Figure 3-7b). Most predicted enhancers (61.4%) are marked by moderate or high levels of acetylation of H3K27 (Figure 3-6). The co-activator p300 and Mediator component MED1, known to bind enhancers, are found at 10741 (29.4%) and 5764 (15.8%) enhancer predictions, respectively (see Methods). Additionally, 19776 (54.1%) of the predicted enhancers exhibit significant DNaseI hypersensitivity. Collectively, I found that 23722 (64.8%) predicted enhancers are supported by some combination of DHS and/or binding of p300 and/or MED1 (Figure 3-7c). Further, the predicted enhancers seem to be distinct from other distal regulatory elements. Only 2666 (8.0%) enhancers are found near a collection of 23267 TSS-distal CTCF sites called in HeLa, IMR90, and CD4 T cells [6,12,26] (1.53-fold enrichment, p = 7.81e-120). Comparison to a genome-wide binding profile of the repressor NRSF/REST27 (which binds mainly transcriptional silencer elements) revealed that only 39 (0.11%) predicted enhancers overlap with distal NRSF/REST binding sites, significantly lower than that expected at random (3.23-fold depletion, p =3.21e-12). These findings indicate that the map of predicted enhancers is strongly enriched for true enhancer elements.

To show that predicted enhancers truly function as enhancers *in vivo*, my lab then verified the functional potential of numerous predicted enhancers in HeLa cells using luciferase reporter assays as previously described [4]. Of nine predicted enhancers that evaluated, seven (78%) were active in reporter assays while none of the random fragments tested were active (Figure 3-7e). The median activity of the enhancers was significantly different from random (p = 3.25e-4). These results offer experimental evidence for the potential function of the predicted enhancers and support the suitability of using chromatin signatures to identify genomic regions with enhancer function.

Histone modification-based prediction of promoters on a genome-scale

Using the genome-wide ChIP-chip enrichment profiles of H3K4me1 and H3K4me3, I used the histone modification-based prediction method to make 13116 promoter predictions (Figure 3-8a). I found that 9835 (75%) predicted promoters overlap with 5’-ends of UCSC Known Genes [27] (Figure 3-8b). I also compared the promoter predictions to the RIKEN human CAGE data set [28] and observed that 11001 (83.9%) overlap with multiple CAGE tags. Further, the prediction model correctly located 76% of active RefSeq transcription start sites [29] (Figure 3-8c) and even 31.5% of inactive TSS, consistent with recent studies demonstrating the presence of similar chromatin landmarks at most promoters in the human genome [11]. I also examined the overlap of predicted promoters with CpG islands (as annotated at the UCSC Genome Browser [30]), sequence elements conventionally understood to be associated with many promoters. The vast majority of promoter predictions (11186, 85.1%) overlap CpG islands, representing almost half (43.3%) of the genome's CpG islands (Figure 3-8d). These findings agree with a previous genome-wide promoter analysis [5] and are comparable to the specificity and sensitivity of the same prediction model in the ENCODE regions [4].

Predicted enhancer activity is confined within CTCF-defined domains

Most of the predicted enhancers (92%) are located greater than 10 kb from the nearest transcription start site (TSS), posing a challenge in assigned enhancers to their appropriate target genes. I partly resolved the enhancer/target gene relationship by using genome-wide location data for the insulator binding protein CTCF [6,12,26]. To determine if CTCF binding sites can be used to define the boundaries of regulatory domains within which enhancers and gene promoters may interact, I examined the effects of the loss of CTCF on global gene expression. A recent study showed that siRNA-mediated CTCF depletion in HeLa cells resulted in upregulation and downregulation of expression of numerous genes [26]. I hypothesized that upregulation of some genes was caused by increased interactions of their promoters with nearby enhancers that had been blocked by CTCF prior to its depletion (Figure 3-9, upper panel), in line with the current understanding of CTCF function. If so, the expectation is finding more predicted enhancers in the vicinity of upregulated genes and fewer enhancers near genes with unchanged or downregulated expression. To test this hypothesis, first I identified insulator-delineated domains in the genome, defining the set of insulators as the union of published CTCF binding sites from IMR90, HeLa, and CD4+ T cells [6,12,26], since I observed consistent CTCF enrichment at nearly all putative insulators across cell types in the ENCODE regions (Figure 3-3a) and genome-wide (Figure 3-4). Then I counted predicted enhancers within insulator-delineated domains adjacent to subsets of genes that were upregulated, downregulated, or unchanged by depletion of CTCF in HeLa cells. Indeed, I observed on average a 2.2-fold enrichment of enhancers within domains adjacent to upregulated genes compared to downregulated genes (Figure 3-9, bottom panel), and a 1.4-fold depletion of enhancers in domains adjacent to downregulated genes relative to genes whose expression is unchanged by CTCF depletion. These results support the putative function of the predicted enhancers and the phenomenon of CTCF-dependent blocking of enhancers by insulators on a global scale.

Identification of conserved and novel sequence motifs in predicted enhancers

Collaborators from Manolis Kellis' laboratory evaluated the predicted enhancers for conserved motif-like sequence patterns using several hundred shuffled TRANSFAC motifs across 10 mammals in a phylogenetic framework that tolerates motif movement, partial motif loss, and sequencing or alignment discrepancies (see Methods). Predicted enhancers showed conservation for 4.3% of instances (at Branch-Length-Score > 50%, see Methods), substantially greater than for the remaining intergenic regions (2.9%, p < 1e-100) and even promoter regions (3.9%, p = 1e-57). Additionally, testing a list of 123 unique TRANSFAC motifs as reported previously20 (see Supplemental Materials), they found that 67 (54%) are over-conserved and 39 (32%) are enriched in predicted enhancers. They also performed *de novo* motif discovery in enhancer regions using multiple alignments of 10 mammalian genomes [31,32], revealing 41 enhancer motifs, of which 19 match known transcription factor motifs while 22 are novel (Table 3-1). These motifs show conservation rates between 7% and 22% in enhancers (median 9.3%), compared to only 1.1% for control shuffled motifs of identical composition. Furthermore, over 90% of these motifs appear to be unique to enhancers, as only 4 motifs are enriched in promoter regions and 12 are in fact depleted in promoters (Table 3-1). These findings indicate that predicted enhancers contain unique regulatory sequences that may be specific to enhancer function.

Chromatin modifications at predicted enhancers are globally correlated with cell type-specific gene expression

I found that predicted enhancers in HeLa cells are much more highly clustered in the genome than expected at random (Wilcoxon p < 1e-300) (Figure 3-7d), consistent with observations in Drosophila [33] and similar analysis from multiple cell types in the ENCODE regions (Wilcoxon p = 1.1e-27). To investigate the association of predicted enhancers with HeLa-specific gene expression, I used Shannon entropy [34] to rank genes by the specificity of their expression levels in HeLa as compared to three other cell lines (K562, GM06990, IMR90) (Figure 3-10), then plotted the distribution of enhancers around genes within insulator-delineated domains (Figure 3-11a). I observed a striking enrichment of predicted enhancers in the domains of HeLa-specific expressed genes relative to non-specific expressed genes and HeLa-specific repressed genes (Figure 3-11a), supporting the role of these predicted enhancers in regulating HeLa-specific gene expression. Noting that most predicted enhancer enrichment occurred within 200 kb of promoters, I counted predicted enhancers within this window (within the same insulator-defined domain) around each promoter and compared counts around the different classes of expressed genes. I observed a 1.83-fold enrichment (p = 4.71e-279) of predicted enhancers around HeLa-specific expressed genes relative to random, while predicted enhancers are actually depleted around non-specific (p = 5.43e-15) and repressed (p = 4.63e-2) genes.

If chromatin modifications at predicted enhancers in HeLa are playing an important role in regulation of HeLa-specific gene expression, then the patterns in another distinct cell type should be markedly different. To test this hypothesis, my lab performed genome-wide ChIP-chip for H3K4me1 and H3K4me3 in K562 cells. Using the chromatin-signature based method, I predicted 24566 putative enhancers in this cell type. Indeed, the vast majority of enhancers predicted in K562 and HeLa cells are unique to either cell type (Figure 3-11b) even though most expressed genes are common between the cell types (Figure 3-11c). Quantitative comparison of chromatin modifications at 55454 marked enhancers in HeLa and K562 cells shows a Pearson correlation coefficient of -0.32, and the cell type-specificity of the chromatin modification profiles throughout the genome is visually striking (Figure 3-11d). Furthermore, these differences seem to have regulatory implications, as domains with HeLa-specific expressed genes are enriched in HeLa enhancers but depleted in K562 enhancers, and vice-versa (Figure 3-11e), strongly supporting the relationship between cell type-specific gene expression patterns and chromatin modifications at predicted enhancers. For example, the MET proto-oncogene has been implicated in a variety of carcinomas (including cervical) [35,36] and is 84-fold more highly expressed in HeLa cells than in K562. Ten enhancers are marked with the enhancer chromatin signature near MET in HeLa cells versus just one enhancer in this region in K562. Conversely, the adjacent CAPZA2 gene is 7-fold more highly expressed in K562 cells, and three enhancers are marked near this gene in K562 versus just one enhancer in HeLa.

To assess the cell type-specificity of functional activity of predicted enhancers, my lab cloned several regions predicted to be enhancers specifically in K562 cells (and not in HeLa cells) and subjected them to reporter assays in HeLa cells as described above and previously [4]. Of nine K562-specific enhancers, only two (22%) were active in HeLa cells as compared to 78% of the HeLa-specific enhancers (Figure 3-7e, Figure 3-12), and the median activity of the K-562 specific enhancers was not significantly different from random (p = 0.11). These findings suggest that the enhancer chromatin signature is a reliable marker of cell type-specific enhancer function.

To expand my investigation across additional cell lines, I also focused on differentially expressed genes between pairs of cell lines in the ENCODE regions. I counted the number of enhancers near the differentially expressed genes in the neighboring domains defined by CTCF sites. I found that enhancers are enriched near differentially expressed genes as compared to the same genes that are differentially repressed in another cell type, and this enrichment is largely confined within CTCF binding sites that directly flank the gene's TSS (Figure 3-13b). On average within this block, there are 0.82 enhancers per differentially down-regulated gene, while there are 1.83 enhancers per differentially up-regulated gene (Figure 3-13c). This 2.2-fold difference suggests that cell-type specific expression is influenced by enhancers and that the action of enhancers is distance-dependent and favoring proximal promoters. When I focused only on the enhancer closest to the differentially expressed gene rather than all enhancers within a CTCF block, I find a smaller difference between the distributions of enhancers in up- and down-regulated genes (Figure 3-13d). The smaller 1.76-fold difference observed here further emphasizes that multiple enhancers, and not just the single closest enhancer, are likely required to regulate differential gene expression of a single promoter.

Subsets of predicted enhancers are bound by transcription factors in other cell types

The overlap of a small but significant fraction of enhancer predictions shared by HeLa and K562 (Figure 3-11b) suggests that some enhancers may be active in multiple cell types or conditions. I compared the HeLa enhancer predictions with the results of several genome-wide studies of binding sites for sequence-specific transcription factors in different cell types, namely estrogen receptor (ER) [37], p53 [38], and p63 [39] in MCF7, HCT116, and ME180 cells, respectively. Interestingly, significant percentages of binding sites for each transcription factor (from 21.4% to 32.6%) overlap with predicted enhancers in HeLa cells (Figure 3-14a). This is in sharp contrast to a significant depletion of the repressor NRSF/REST [40] at the predicted enhancers and a minimal overlap with CTCF-binding sites.

The enhancer chromatin signature correlates with rapid gene induction in response to interferon gamma

To examine the potential role of enhancers in regulating inducible gene expression, my lab treated HeLa cells with the cytokine interferon-gamma (IFNγ) and identified binding sites for the transcription factor STAT1 throughout the genome using ChIP-chip. As a signal-dependent, latent cytoplasmic transcription factor, STAT1 is generally understood to bind its target DNA sequences only after IFNγ induction [41] although recent work has suggested that some STAT1 binding may occur prior to induction [42]. In IFNγ-treated HeLa cells, I identified 1969 STAT1 binding sites, with 85.8% of STAT1 binding sites occurring distal to UCSC Known Gene 5'-ends [27]. Comparison of these distal STAT1 binding sites with recent ChIP-seq analysis of STAT1 binding in uninduced HeLa cells [42] indicate that only 6.5% of induced STAT1 binding sites may be occupied by STAT1 prior to induction. Thus, most STAT1 binding sites identified here are very unlikely to be bound by STAT1 prior to induction.

I observed that 429 distal STAT1 binding sites overlapped enhancers that were predicted in HeLa cells prior to induction (Figure 3-14a). The H3K4me1 enhancer chromatin signature is clearly present at these STAT1 binding sites, which I designated as STAT1 group I, while no evidence of this signature is visible at the remaining 1260 distal STAT1 binding sites, designated STAT1 group II (Figure 3-14b). Intriguingly, I observed significant relative induction of expression of genes in the domains of STAT1 group I binding sites after just 30 minutes of IFNγ-induction, while induction levels remained relatively unchanged for genes in the domains of other distal STAT1 group II binding sites during this time (Figure 3-14c). These findings suggest that an enhancer chromatin signature confers increased regulatory responsiveness to a STAT1 binding site, in agreement with my previous discovery of functional enhancers in HeLa cells that were marked by the enhancer chromatin signature but were not active until they were bound by STAT1 [4].

Discussion

Recent experiments have confirmed the H3K4me1 signature at distal enhancers on a large scale and supported the role of cell type-specific chromatin states in directing the recruitment of transcription factors [18,43,44,45,46], underscoring the importance of genome-wide enhancer identification in deciphering the mechanisms of global gene regulatory networks. Toward that end, I generated the first genome-wide maps of chromatin signatures at enhancers in two human cell types, HeLa and K562, revealing that enhancers are epigenetically distinct between these cell types, and discovering a global correlation between cell type-specific chromatin modification profiles at enhancers and cell type-specific gene expression programs.

The cell type-specificity of enhancer activity in reporter assays is intriguing. While the majority (78%) of HeLa-specific enhancers that my lab evaluated for function were active in reporter assays, enhancers identified on the basis of their chromatin signatures in K562 cells showed minimal or no activity in HeLa cells. Additional experiments may reveal the epigenetic and DNA sequence-based mechanisms for this specificity, in particular the role of H3K4me1 in enhancer function and maintenance in regulating target gene expression. Also intriguing was the presence of the enhancer chromatin signature at hundreds of distal STAT1 binding sites prior to induction with IFNγ, and the observation that genes near these enhancer-marked STAT1 binding sites were rapidly and significantly upregulated upon IFNγ-treatment while genes near other distal STAT1 binding sites were not. The basis of this apparent increased regulatory response conferred by the enhancer signature remains to be fully investigated.

Many novel DNA sequence motifs appear to be enhancer-specific, though further experiments are needed to establish the function of these novel motifs. As several of the identified motifs correspond to factors that have been demonstrated to bind the predicted enhancers in various cell types, the motif data offer a very useful resource for additional experiments investigating patterns of activator-mediated gene expression in diverse cellular contexts. The predicted enhancer maps will also be of great utility in annotating the function of potential regulatory elements identified in other experiments, as demonstrated by the significant overlap of enhancer predictions with experimentally determined TFBS in diverse cell types.

These findings offer the first genome-wide evaluation of the relationship between chromatin modifications at transcriptional enhancers and global programs of cell type-specific gene expression. Subsequent experiments in diverse cell types and additional physiological contexts will provide further insight into the relationships between specific enhancers and their target genes, leading to increased understanding of transcriptional regulatory mechanisms and revealing novel therapeutic and diagnostic targets in human disease.

Methods

Microarrays

ChIP-chip spanning the ENCODE regions was performed as described previously [4]. Genome-wide, ChIP samples were hybridized to the NimbleGen genome-wide tiling microarray set (NimbleGen Systems, Inc.) as previously described [5,6] and to custom condensed enhancer microarrays (NimbleGen Systems, Inc.) using standard methods. The condensed enhancer microarrays consisted of tiled 10 kb windows around each of 38716 primary predicted enhancers and standard controls. DNase-chip was performed and the data analyzed as previously described [47]. Gene expression data for HeLa, K562, and GM cells were obtained using HU133 Plus 2.0 microarrays (Affymetrix), as described previously [5].

ChIP-chip data analysis

For ENCODE arrays, ChIP-chip data were normalized and analyzed as before [4]. On genome-wide arrays, several platforms were used. For ChIP-chip of histone modifications H3K4Me1 and H3K4Me3 in HeLa cells on Nimblegen genome-wide tiling arrays (38 array set, hg17), I normalized the raw data from each array using both the median and loess algorithms from the Bioconductor R package (treating each probe equally). For each array, I chose the normalization method that gave the most balanced distribution of random probes about a log ratio of 0. For ChIP-chip of histone modifications H3K4Me1, H3K4Me3, and H3K27Ac in K562 cells on Nimblegen HD2 Economy genome-wide tiling arrays (12 array set, hg18), I normalized the raw data from each array using MA2C [48], and mapped the normalized data to hg17 coordinates using the UCSC Genome Browser liftOver tool.

On all arrays, ChIP-chip targets for CTCF, p300, MED1, and STAT1 were selected with the Mpeak program [5].

Expression array analysis

I used the GCRMA package [49] to normalize Affymetrix mRNA expression arrays for HeLa, GM, and K562 cell types. For every pair of these cell types, I also use GCRMA to find differentially expressed and repressed genes using a p-value cutoff of 0.01 in conjunction with a fold change cutoff of 2.0.

The expression data for ES and dES cell types was done using the two-channel Nimblegen platform. Gene expression raw data were extracted using NimbleScan software v2.1. Considering that the signal distribution of the RNA sample is distinct from that of the genomic DNA (gDNA) sample, the signal intensities from RNA channels in all eight arrays were normalized with the Robust Multiple-chip Analysis (RMA) algorithm [49]. Separately, the same normalization procedure was performed on those from the gDNA samples. For a given gene, the median-adjusted ratio between its normalized intensity from the RNA channel and that from the gDNA channel was then calculated as follows: Ratio = intensity from RNA channel/(intensity from gDNA channel + median intensity of all genes from the gDNA channel). Collaborators from James Thomson's laboratory have found that this median-adjusted ratio gives the most consistent results when compared to other published human ES cell expression data, such as SAGE library information available from the Cancer Genome Anatomy Project (CGAP). Consequently, this median-adjusted ratio as the measurement for the gene expression level is used. Due to differences in platform, it is only possible to use this expression data to compare ES and dES cell types. As a conservative measure of differential expression, I use a fold-change cutoff of 2.

Enhancer prediction method

The procedure used to predict enhancers follows closely to the method outlined in Chapter 2. I first bin the tiling ChIP-chip data into 100 bp bins, averaging multiple probes that fall into the same bin. Using a sliding window on H3K4Me1 and H3K4Me3, I scan for chromatin signatures resembling a training set of enhancer patterns defined previously by the p300 binding sites in HeLa cells, keeping only those windows that correlate most with the training sets and that have significant enrichment of chromatin modifications. I then use a discriminative filter to keep only those predictions that correlate with an averaged enhancer training set more than the promoter training set. Finally, I apply a descriptive filter, keeping only those remaining predictions having a correlation of at least 0.5 with an averaged training set.

In both ENCODE and genome-wide predictions of this study, I made predictions of active promoters and enhancers as previously, with the following modifications:

- Repetitive regions of the genome are not covered by the probes on tiling arrays, contributing to gaps in coverage. Previously, I interpolated through all gaps. But this can lead to false positive predictions or biasing of the underlying background distributions when there are many gaps. To remove these concerns, here I interpolate only through gaps smaller than 1000 bp.
- In the prediction algorithm, I slide a 10 kb window across the tiled regions and compute 2 statistics for each window: the correlation with a training set and the sum of the absolute values of intensities of the middle 2 kb region of the window. The correlation part has

remained unchanged in this study. Previously, the intensity statistic appeared normally distributed, and as such I approximated it with a Gaussian distribution. In light of the larger datasets in this study, this normal assumption did not appear entirely correct. Here, I changed the intensity statistic to the sum of squares of the normalized intensities in the 2 kb region. A normalized intensity is an intensity subtracted from the mean array intensity and divided by the standard deviation of the array intensity. Since each array is properly normalized to follow a Gaussian distribution, by definition, this statistic follows a Chi-squared distribution with 42 degrees of freedom (for each window, each of the 2 modifications has 21 normalized intensities squared: 10 in each direction and one at 0).

The training set used here contained the same six groups of training sets used in Chapter 2, with the exception that the HeLa enhancer predictions used data derived from the genome-wide H3K4Me1 and H3K4Me3 arrays.

In the ENCODE regions, as in Heintzman et al, I keep predictions in the top 10% of the intensity distribution and top 1% of the correlation distribution. For the genome-wide enhancer predictions in HeLa, a ROC analysis (data not shown) indicates that a correlation cutoff of 1% and an intensity cutoff of 1% yields the best overlap with previously published predictions in the ENCODE regions. Similarly, for the genome-wide predictions for the MA2C-normalized K562 data, ROC analysis suggests using a correlation cutoff of 10% and an intensity cutoff of 1e-12.

CTCF knockdown analysis

I created three sets of genes: the 1000 most up-regulated upon siCTCF treatment, the 1000 most down-regulated, and the median 1000 unchanged genes. Then I counted predicted enhancers within five insulator-delineated domains (between CTCF binding sites) adjacent to subsets of genes that were upregulated, downregulated, or unchanged by depletion of CTCF in HeLa cells. To generate a random distribution, I also repeated this analysis for 100 sets of 1000 random genes. To obtain enhancer enrichment, I divided the observed counts with the averaged random counts. Finally, to assess significance, I assumed the random counts followed a normal distribution.

Overlap analysis

To assess the overlap of predicted enhancers with genome-wide transcription factor binding site (TFBS) data sets, I counted the number of experimentally determined TFBS within 2.5 kb of the enhancers. To determine the significance of this overlap, I compared this statistic to the distribution of statistics for 100 random sets of putative enhancers, which was approximated by a normal. Each random set had the same number of elements as the putative enhancer set. The enhancer predictions were limited to regions on the ChIP-chip array. Similarly, each random enhancer was placed uniformly at random in a sample space consisting of well-represented regions on the ChIP-chip microarray. The chromosomal distribution of each of the sets was kept constant. This careful placement of random sites ensures against artificial inflation of the overlap significance.

P-values

The p-values for correlations were obtained from using the Matlab corr function. This p-value measures the probability that there is no correlation between the two variables, against the alternative that the correlation is non-zero. The p-values for Wilcoxon rank sum tests were obtained from the Matlab ranksum function.

Gene expression and entropy analysis

Gene expression in the various cell lines was analyzed using HGU133 Plus 2.0 microarrays (Affymetrix) as described [5]. Specificity of expression was determined using a function of Shannon entropy as described [34] and the top, middle, and bottom 1000 genes from this analysis were designated as HeLa-specific expressed, non-specific expressed, and HeLa-specific repressed genes, respectively (Figure 3-10). This specificity was used for evaluation of enhancer enrichment in insulator-defined domains containing the promoters for these classes of genes (as in Figure 3-11a), where insulators were defined by CTCF binding sites. When counting enhancers around these promoters, I included all enhancers within 200 kb of a promoter as long as they were still within the same insulator-defined domain. Random distributions were generated by averaging the enrichment profiles around promoters of 100 iterations of randomly selected enhancer sets of 36589 elements. To assess enhancer and gene expression specificity between HeLa and K562 cells (as in Figure 3-11b-c), I use the MAS5 algorithm from the Bioconductor R package to generate gene expression Present/Absent calls from each cell type. Since there are two biological replicates of K562 expression, to merge calls for these replicates, probes called differently in the two replicates are labeled as Marginal. To eliminate biases from genes not expressed in either HeLa or K562, I only consider a probe if it is called Present in either HeLa or K562. I mapped Affymetrix probes to gene identifiers using the knownToU133Plus2 table from the UCSC Genome Browser, and then map the identifiers to genomic coordinates (hg17, NCBI build 35) using the knownGene table from the UCSC Genome Browser [30]. To reduce redundancy, I keep only the first gene when multiple Affymetrix probes map to the same annotated gene. The result from this filtering and mapping is a set of 11783 genes. For each such gene, I counted the number of enhancers predicted in each cell type within that gene's CTCF-domain. I sorted the genes by differential (HeLa – K562) gene expression (as defined by the RMA algorithm from the Bioconductor R package) and use a sliding window of 1000 genes to generate a profile of the average number of enhancers for each cell type as a function of average differential gene expression. This gives two profiles: one using HeLa enhancers and one using K562 enhancers. To normalize, I repeat this analysis for each cell type using 100 sets of random enhancers (placed uniformly at random on the tiling microarrays), giving 100 random enhancer-expression profiles. I then define the enhancer enrichment profile as the ratio between the number of enhancers in the observed profile and the expected number of enhancers in the averaged random profile.

Motif analysis

Enhancer regions were defined as 2 kb windows centered on each prediction, and promoter regions were defined as 1 kb windows upstream from annotated TSS. Promoters regions were excluded from enhancer regions; repeats, exons and transposons were excluded from both. Motif conservation in each region was evaluated relative to the genomes of opossum, tenrec, elephant, armadillo, cow, dog, rabbit, rat and mouse, extracted from UCSC Genome Browser and used with permission. The mammalian tree, along with branch lengths, was computed using DNAML (PHYLIP package) [50] with the F84 nucleotide model of evolution in ~500kb of randomly selected exon sequence. Known and novel motifs were discovered as previously described10, with the primary difference that instances were not required to have perfect conservation and were considered conserved if they were found across a number of species spanning at least 50% of the total branch length of the mammalian tree (Branch-Length-Score > 50%) [31,51]. Motifs were ranked based on their over-conservation, measured as the probability of observing a substantially increased number of conserved motif instances compared to that expected for motifs of identical composition, and selected all motifs with P < 1 X 10-3. A motif's enrichment was evalulated as its over-abundance, or the hypergeometric probability of observing a substantially increased number of occurrences in the intergenic and intronic regions of the human genome (regardless of evolutionary conservation) compared to motifs of identical composition, with a cutoff of P < 1E-3.

Reporter assays

Cloning and reporter assays were performed as previously described [4] and a fragment was designated as active if its relative luciferase value was greater than 2.33 standard deviations (p = 0.01) above the median random activity.

Acknowledgements

Chapter 3, in full, is a reprint of the material as it appears in Nature 2009. Heintzman, Nathaniel D ; Hon, Gary C ; Hawkins, R David ; Kheradpour, Pouya; Stark, Alexander ; Harp, Lindsey F ; Ye, Zhen ; Lee, Leonard K ; Stuart, Rhona K ; Ching, Christina W ; Ching, Keith A ; Antosiewicz-Bourget, Jessica E ; Liu, Hui ; Zhang, Xinmin ; Green, Roland D ; Lobanenkov, Victor V ; Stewart, Ron ; Thomson, James A ; Crawford, Gregory E ; Kellis, Manolis ; Ren, Bing. "Histone modifications at human enhancers reflect global cell-type-specific gene expression", Nature, vol. 458, 2009. The dissertation author was a primary investigator and author of this paper. Specifically, the dissertation author performed the computational analysis of histone modifications including work on the cell-type specificity of different function elements, predicting enhancers genome-wide, and analyzing the influence of enhancers on gene expression.

References

1. Heintzman ND, Ren B (2007) The gateway to transcription: identifying, characterizing and understanding promoters in the eukaryotic genome. Cell Mol Life Sci 64: 386-400.

2. Maston GA, Evans SK, Green MR (2006) Transcriptional Regulatory Elements in the Human Genome. Annu Rev Genomics Hum Genet 7: 29-59.

3. Nightingale KP, O'Neill LP, Turner BM (2006) Histone modifications: signalling receptors and potential elements of a heritable epigenetic code. Curr Opin Genet Dev 16: 125-136.

4. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B (2007) Distinct and predictive

chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39: 311-318.

5. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B (2005) A high-resolution map of active promoters in the human genome. Nature 436: 876-880.

6. Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenkov VV, Ren B (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell 128: 1231-1245.

7. Lemon B, Tjian R (2000) Orchestrated response: a symphony of transcription factors for gene control. Genes Dev 14: 2551-2569.

8. Mellor J (2005) The dynamics of chromatin remodeling at promoters. Mol Cell 19: 147-157.

9. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES (2006) A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell 125: 12.

10. Roh TY, Cuddapah S, Cui K, Zhao K (2006) The genomic landscape of histone modifications in human T cells. Proc Natl Acad Sci U S A 103: 15782-15787.

11. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA (2007) A chromatin landmark and transcription initiation at most promoters in human cells. Cell 130: 77-88.

12. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K (2007) High-resolution profiling of histone methylations in the human genome. Cell 129: 823-837.

13. ENCODE_Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306: 636-640.

14. Hatzis P, Talianidis I (2002) Dynamics of enhancer-promoter communication during differentiation-induced gene activation. Mol Cell 10: 1467-1477.

15. Agalioti T, Lomvardas S, Parekh B, Yie J, Maniatis T, Thanos D (2000) Ordered recruitment of chromatin modifying and general transcription factors to the IFN-beta promoter. Cell 103: 667-678.

16. Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, Lee TI, Bell GW, Walker K, Rolfe PA, Herbolsheimer E, Zeitlinger J, Lewitter F, Gifford DK, Young RA (2005) Genome-wide map of nucleosome acetylation and methylation in yeast. Cell 122: 517-527.

17. Liu CL, Kaplan T, Kim M, Buratowski S, Schreiber SL, Friedman N, Rando OJ (2005) Single-nucleosome mapping of histone modifications in S. cerevisiae. PLoS Biol 3: e328.

18. Koch CM, Andrews RM, Flicek P, Dillon SC, Karaoz U, Clelland GK, Wilcox S, Beare DM, Fowler JC, Couttet P, James KD, Lefebvre GC, Bruce AW, Dovey OM, Ellis PD, Dhami P, Langford CF, Weng Z, Birney E, Carter NP, Vetrie D, Dunham I (2007) The landscape of histone modifications across 1% of the human genome in five human cell lines. Genome Res 17: 691-707.

19. Gaszner M, Felsenfeld G (2006) Insulators: exploiting transcriptional and epigenetic mechanisms. Nat Rev Genet 7: 703-713.

20. Wei GH, Liu DP, Liang CC (2005) Chromatin domain boundaries: insulators and beyond. Cell Res 15: 292-300.

21. Wang Q, Carroll JS, Brown M (2005) Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking. Mol Cell 19: 631-642.

22. King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, Hardison RC (2005) Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. Genome Res 15: 1051-1060.

23. Kleinjan DA, Seawright A, Schedl A, Quinlan RA, Danes S, van Heyningen V (2001) Aniridia-associated translocations, DNase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of PAX6. Hum Mol Genet 10: 2049-2059.

24. Wolf AT, Medcalf RL, Jern C (2005) The t-PA -7351C>T enhancer polymorphism decreases Sp1 and Sp3 protein binding affinity and transcriptional responsiveness to retinoic acid. Blood 105: 1060-1067.

25. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034-1050.

26. Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E, Tsutsumi S, Nagae G, Ishihara K, Mishiro T, Yahata K, Imamoto F, Aburatani H, Nakao M, Imamoto N, Maeshima K, Shirahige K, Peters JM (2008) Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature 451: 796-801.

27. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D (2006) The UCSC Known Genes. Bioinformatics 22: 1036-1046.

28. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest AR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, Chalk AM, Chiu KP, Choudhary V, Christoffels A, Clutterbuck DR, Crowe ML, Dalla E, Dalrymple BP, de Bono B, Della Gatta G, di Bernardo D, Down T, Engstrom P, Fagiolini M, Faulkner G, Fletcher CF, Fukushima T, Furuno M, Futaki S, Gariboldi M, Georgii-Hemming P, Gingeras TR, Gojobori T, Green RE, Gustincich S, Harbers M, Hayashi Y, Hensch TK, Hirokawa N, Hill D, Huminiecki L, Iacono M, Ikeo K, Iwama A, Ishikawa T, Jakt M, Kanapin A, Katoh M, Kawasawa Y, Kelso J, Kitamura H, Kitano H, Kollias G, Krishnan SP, Kruger A, Kummerfeld SK, Kurochkin IV, Lareau LF, Lazarevic D, Lipovich L, Liu J, Liuni S, McWilliam S, Madan Babu M, Madera M, Marchionni L, Matsuda H, Matsuzawa S, Miki H, Mignone F, Miyake S, Morris K, Mottagui-Tabar S, Mulder N, Nakano N, Nakauchi H, Ng P, Nilsson R, Nishiguchi S, Nishikawa S, Nori F, Ohara O, Okazaki Y, Orlando V, Pang KC, Pavan WJ, Pavesi G, Pesole G, Petrovsky N, Piazza S, Reed J, Reid JF, Ring BZ, Ringwald M, Rost B, Ruan Y, Salzberg SL, Sandelin A, Schneider C, Schonbach C, Sekiguchi K, Semple CA, Seno S, Sessa L, Sheng Y, Shibata Y, Shimada H, Shimada K, Silva D, Sinclair B, Sperling S, Stupka E, Sugiura K, Sultana R, Takenaka Y, Taki K, Tammoja K, Tan SL, Tang S, Taylor MS, Tegner J, Teichmann SA, Ueda HR, van Nimwegen E, Verardo R, Wei CL, Yagi K, Yamanishi H, Zabarovsky E, Zhu S, Zimmer A, Hide W, Bult C, Grimmond SM, Teasdale RD, Liu ET, Brusic V, Quackenbush J, Wahlestedt C, Mattick JS, Hume DA, Kai C, Sasaki D, Tomaru Y, Fukuda S, Kanamori-Katayama M, Suzuki M, Aoki J, Arakawa T, Iida J, Imamura K, Itoh M, Kato T, Kawaji H, Kawagashira N, Kawashima T, Kojima M, Kondo S, Konno H, Nakano K, Ninomiya N, Nishio T, Okada M, Plessy C, Shibata K, Shiraki T, Suzuki S,

Tagami M, Waki K, Watahiki A, Okamura-Oho Y, Suzuki H, Kawai J, Hayashizaki Y (2005) The transcriptional landscape of the mammalian genome. Science 309: 1559-1563.

29. Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33: D501-504.

30. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome Res 12: 996-1006.

31. Kheradpour P, Stark A, Roy S, Kellis M (2007) Reliable prediction of regulator targets using 12 Drosophila genomes. Genome Res 17: 1919-1931.

32. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature 434: 338-345.

33. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc Natl Acad Sci U S A 99: 757-762.

34. Schug J, Schuller WP, Kappen C, Salbaum JM, Bucan M, Stoeckert CJ, Jr. (2005) Promoter features related to tissue specificity as measured by Shannon entropy. Genome Biol 6: R33.

35. Rasola A, Fassetta M, De Bacco F, D'Alessandro L, Gramaglia D, Di Renzo MF, Comoglio PM (2007) A positive feedback loop between hepatocyte growth factor receptor and beta-catenin sustains colorectal cancer cell invasive growth. Oncogene 26: 1078-1087.

36. Tsai HW, Chow NH, Lin CP, Chan SH, Chou CY, Ho CL (2006) The significance of prohibitin and c-Met/hepatocyte growth factor receptor in the progression of cervical adenocarcinoma. Hum Pathol 37: 198-204.

37. Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, Eeckhoute J, Brodsky AS, Keeton EK, Fertuck KC, Hall GF, Wang Q, Bekiranov S, Sementchenko V, Fox EA, Silver PA, Gingeras TR, Liu XS, Brown M (2006) Genome-wide analysis of estrogen receptor binding sites. Nat Genet 38: 1289-1297.

38. Wei CL, Wu Q, Vega VB, Chiu KP, Ng P, Zhang T, Shahab A, Yong HC, Fu Y, Weng Z, Liu J, Zhao XD, Chew JL, Lee YL, Kuznetsov VA, Sung WK, Miller LD, Lim B, Liu ET, Yu Q, Ng HH, Ruan Y (2006) A global map of p53 transcription-factor binding sites in the human genome. Cell 124: 207-219.

39. Yang A, Zhu Z, Kapranov P, McKeon F, Church GM, Gingeras TR, Struhl K (2006) Relationships between p63 binding, DNA sequence, transcription activity, and biological function in human cells. Mol Cell 24: 593-602.

40. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316: 1497-1502.

41. Levy DE, Darnell JE, Jr. (2002) Stats: transcriptional control and biological impact. Nat Rev Mol Cell Biol 3: 651-662.

42. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4: 651-657.

43. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SC, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermuller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung WK, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C, Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J, Nagalakshmi U, Wu J, Lian Z, Lian J, Newburger P, Zhang X, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei CL, Ruan Y, Struhl K, Gerstein M, Antonarakis SE, Fu Y, Green ED, Karaoz U, Siepel A, Taylor J, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Cooper GM, Asimenos G, Dewey CN, Hou M, Nikolaev S, Montoya-Burgos JI, Loytynoja A, Whelan S, Pardi F, Massingham T, Huang H, Zhang NR, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Seringhaus M, Church D, Rosenbloom K, Kent WJ, Stone EA, Batzoglou S, Goldman N, Hardison RC, Haussler D, Miller W, Sidow A, Trinklein ND, Zhang ZD, Barrera L, Stuart R, King DC, Ameur A, Enroth S, Bieda MC, Kim J, Bhinge AA, Jiang N, Liu J, Yao F, Vega VB, Lee CW, Ng P, Yang A, Moqtaderi Z, Zhu Z, Xu X, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Fowler JC, Couttet P, Bruce AW, Dovey OM, Ellis PD, Langford CF, Nix DA, Euskirchen G, Hartman S, Urban AE, Kraus P, Van Calcar S, Heintzman N, Kim TH, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Xu M, Haidar JN, Yu Y, Iyer VR, Green RD, Wadelius C, Farnham PJ, Ren B, Harte RA, Hinrichs AS, Trumbower H, Clawson H, Hillman-Jackson J, Zweig AS, Smith K, Thakkapallayil A, Barber G, Kuhn RM, Karolchik D, Armengol L, Bird CP, de Bakker PI, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Woodroffe A, Davydov E, Dimas A, Eyras E, Hallgrimsdottir IB, Huppert J, Zody MC, Abecasis GR, Estivill X, Bouffard GG, Guan X, Hansen NF, Idol JR, Maduro VV, Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC, Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang H, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Koriabine M, Nefedov M, Osoegawa K, Yoshinaga Y, Zhu B, de Jong PJ (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799-816.

44. Lupien M, Eeckhoute J, Meyer CA, Wang Q, Zhang Y, Li W, Carroll JS, Liu XS, Brown M (2008) FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. Cell 132: 958-970.

45. Kontaraki J, Chen HH, Riggs A, Bonifer C (2000) Chromatin fine structure profiles for a developmentally regulated gene: reorganization of the lysozyme locus before trans-activator binding and gene expression. Genes Dev 14: 2106-2122.

46. Lefevre P, Lacroix C, Tagoh H, Hoogenkamp M, Melnik S, Ingram R, Bonifer C (2005) Differentiation-dependent alterations in histone methylation and chromatin architecture at the inducible chicken lysozyme gene. J Biol Chem 280: 27552-27560.

47. Crawford GE, Davis S, Scacheri PC, Renaud G, Halawi MJ, Erdos MR, Green R, Meltzer PS, Wolfsberg TG, Collins FS (2006) DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. Nat Methods 3: 503-509.

48. Song JS, Johnson WE, Zhu X, Zhang X, Li W, Manrai AK, Liu JS, Chen R, Liu XS (2007) Model-based analysis of two-color arrays (MA2C). Genome Biol 8: R178.

49. Wu Z, Irizarry RA (2004) Preprocessing of oligonucleotide array data. Nat Biotechnol 22: 656-658; author reply 658.

50. Felsenstein J (2005) Distributed by the author Department of Genome Sciences, University of Washington, Seattle.

51. Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, Ruby JG, Brennecke J, Hodges E, Hinrichs AS, Caspi A, Paten B, Park SW, Han MV, Maeder ML, Polansky BJ, Robson BE, Aerts S, van Helden J, Hassan B, Gilbert DG, Eastman DA, Rice M, Weir M, Hahn MW, Park Y, Dewey CN, Pachter L, Kent WJ, Haussler D, Lai EC, Bartel DP, Hannon GJ, Kaufman TC, Eisen MB, Clark AG, Smith D, Celniker SE, Gelbart WM, Kellis M (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature 450: 219-232.

52. Xi H, Shulha HP, Lin JM, Vales TR, Fu Y, Bodine DM, McKay RD, Chenoweth JG, Tesar PJ, Furey TS, Ren B, Weng Z, Crawford GE (2007) Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. PLoS Genet 3: e136.

Figures and Tables

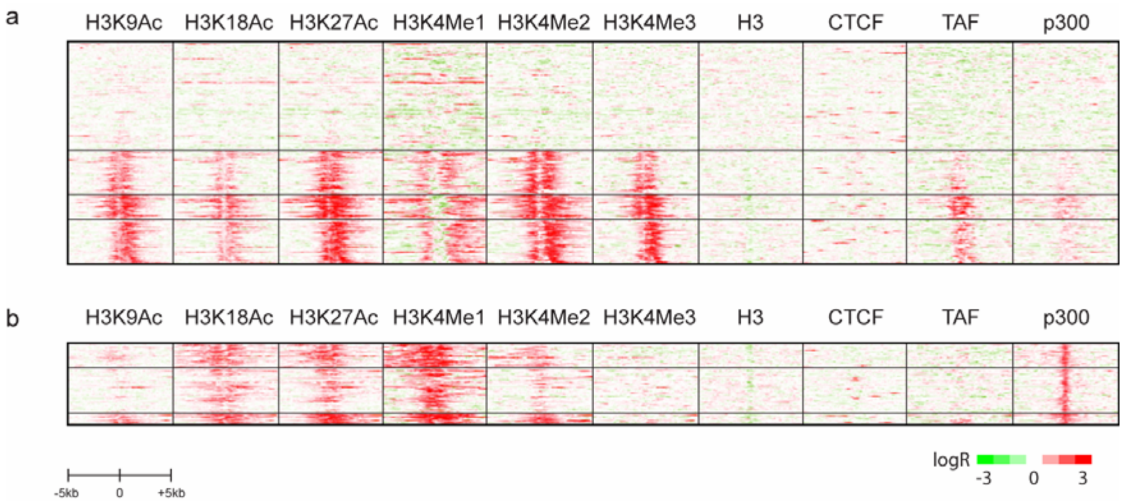


Figure 3-1: Chromatin acetylation features at promoters and enhancers in the ENCODE regions.

ChIP-chip was performed on the acetylated histones H3K9Ac, H3K18Ac, and H3K27Ac, and the enrichment was compared to the (A) promoter and (B) p300 clusters from Chapter 2 in HeLa cells [4]. Each horizontal line details the ChIP-chip enrichment of various chromatin modifications and transcription factors in 10 kb windows. For consistency in comparison, I clustered the data in the same order as Heintzman et al. [4], which used k-means clustering. All three active promoter clusters P2, P3, and P4 are highly enriched in all three acetylated histones, whereas the enhancer clusters are mostly enriched in H3K18Ac and H3K27Ac, but have only weak H3K9Ac enrichment. Average profiles of log enrichment ratios for promoters or p300 binding sites in each cluster are shown at the bottom of each panel.

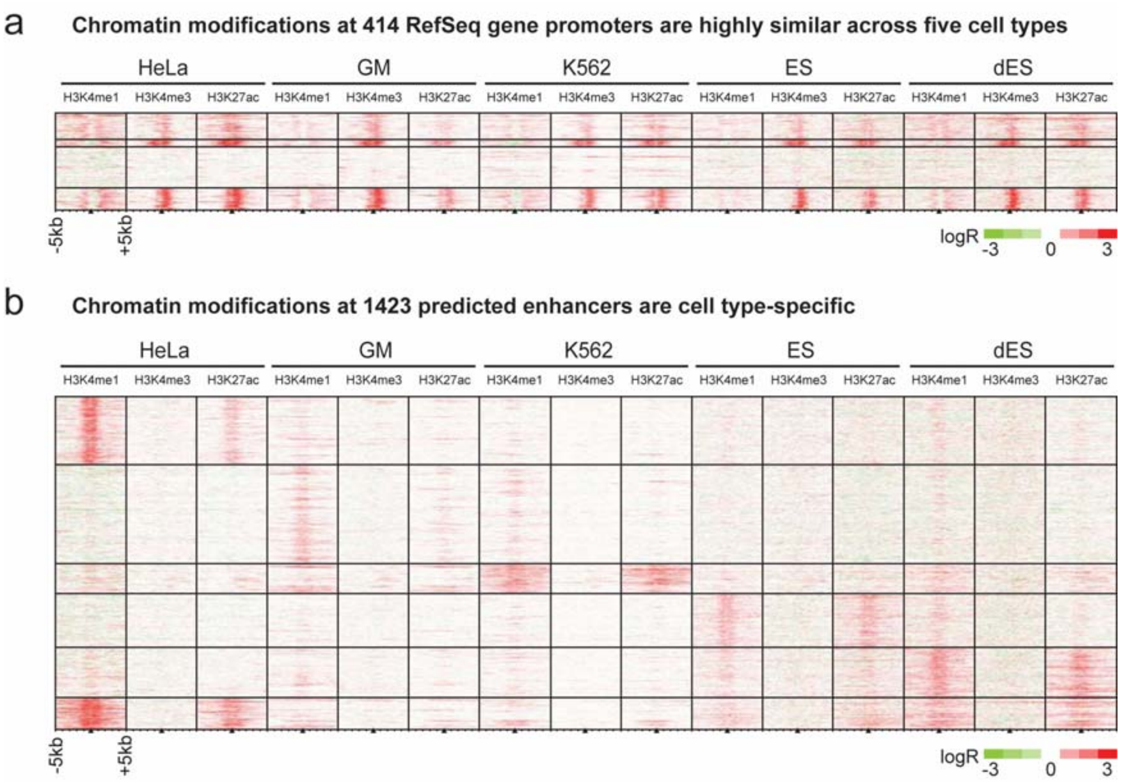


Figure 3-2: Chromatin modifications at promoters are cell type-invariant while those at enhancers are cell type-specific

My lab employed ChIP-chip to map histone modifications (H3K4me1, H3K4me3, and H3K27ac) in the ENCODE regions in five cell types (HeLa, GM, K562, ES, dES). (A) I performed k-means clustering on the chromatin modifications found +/- 5 kb from 414 promoters, and observe them to be generally invariant across cell types. (B) As in (A), but clustering on 1423 non-redundant enhancers predicted on the basis of chromatin signatures.

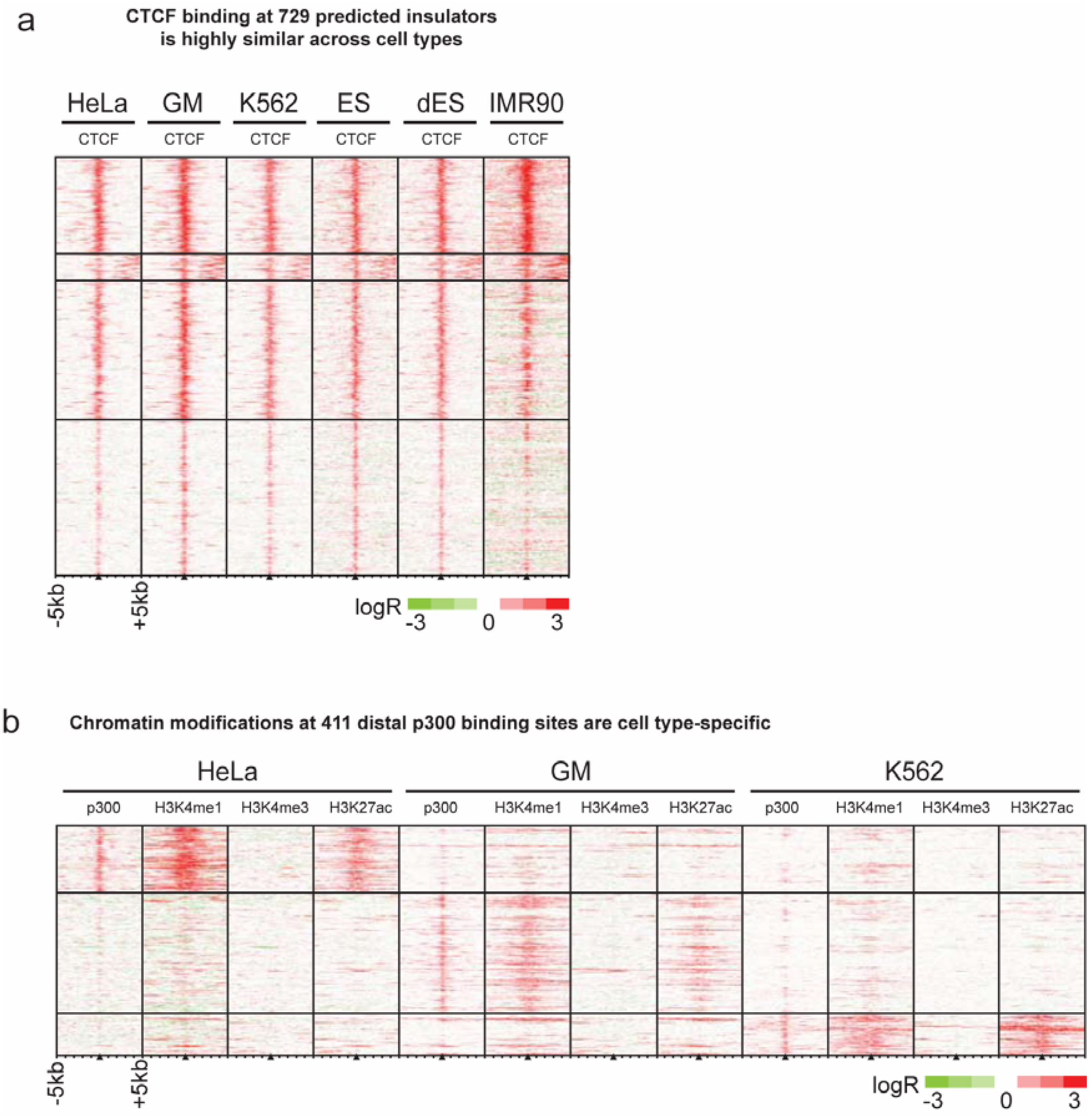


Figure 3-3: Comparison of regulatory elements in the ENCODE regions

I performed k-means clustering on CTCF enrichment at 729 non-redundant CTCF binding sites found by Mpeak [5]. For comparison I have also shown the enrichment patterns from a genome-wide study in IMR90 cells [6], which supports the cell type-invariant nature of CTCF binding. (F) I clustered the chromatin modifications at 411 non-redundant p300 binding sites in HeLa, GM, and K562 cells. Enrichment of p300 binding, which was not a criteria in the clustering, confirms the cell type-specificity of the chromatin marks at enhancers.

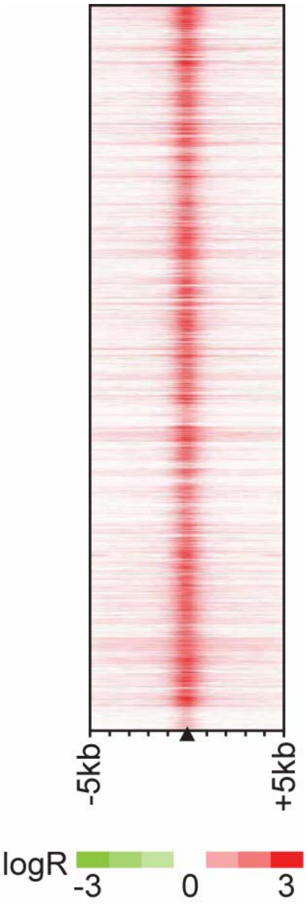


Figure 3-4: CTCF enrichment at genome-wide putative insulators in IMR90 cells

Experimentally-determined binding sites published for CTCF in IMR90, HeLa, and CD4+ T cells [6,12,26] were combined into one set of binding sites, and the ChIP-chip enrichment data at all of these sites from IMR90 cells are visualized as 10 kb windows centered at the CTCF binding sites as described above, organized by genomic position. These data support the consistency of CTCF binding across cell types.

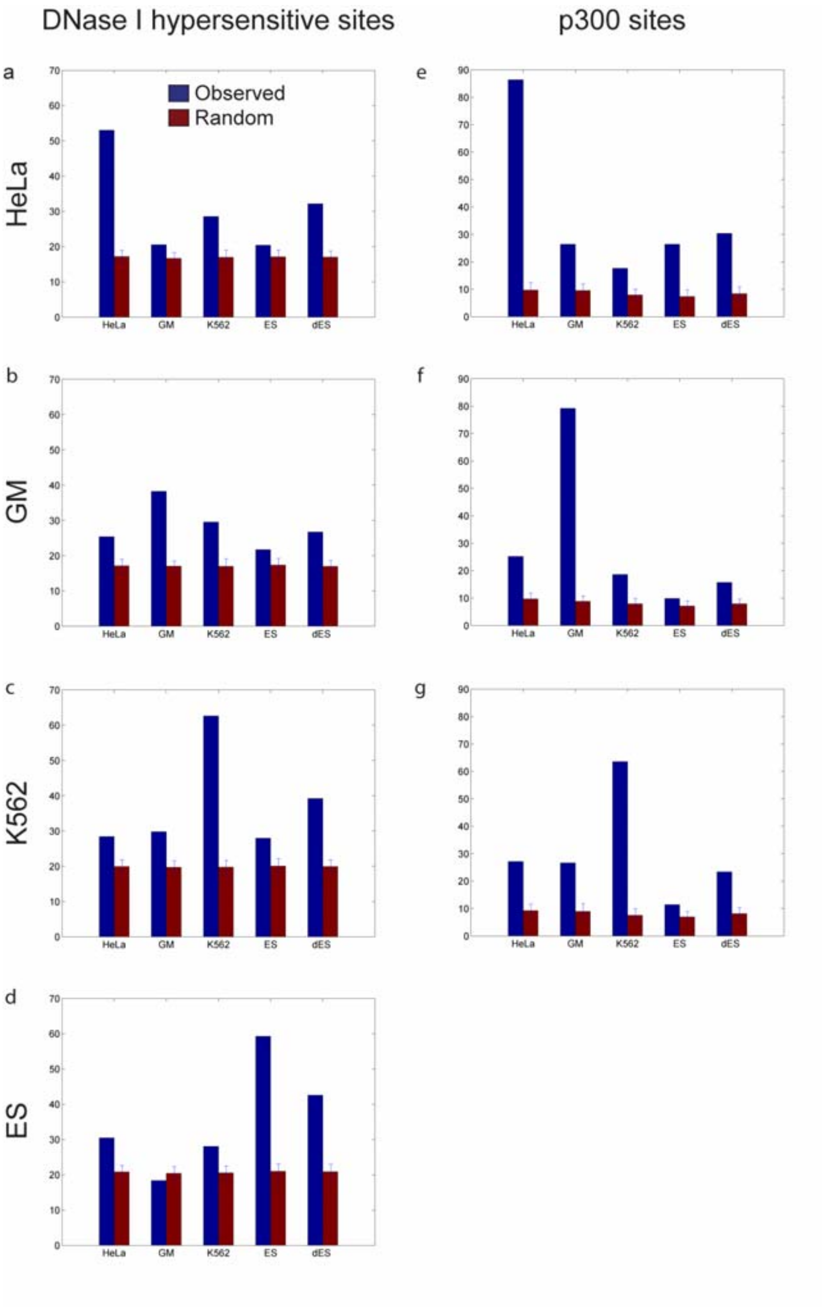


Figure 3-5: Verification of histone modification-based prediction of enhancers in the ENCODE regions

(a-d) The percentage of predicted enhancers within 2.5 kb of hypersensitive sites in HeLa, GM, K562, and ES cells as previously defined [52]. (e-g) The percentage of p300 sites mapped in HeLa, GM, and K562 cell lines within 2.5 kb of predicted enhancers.

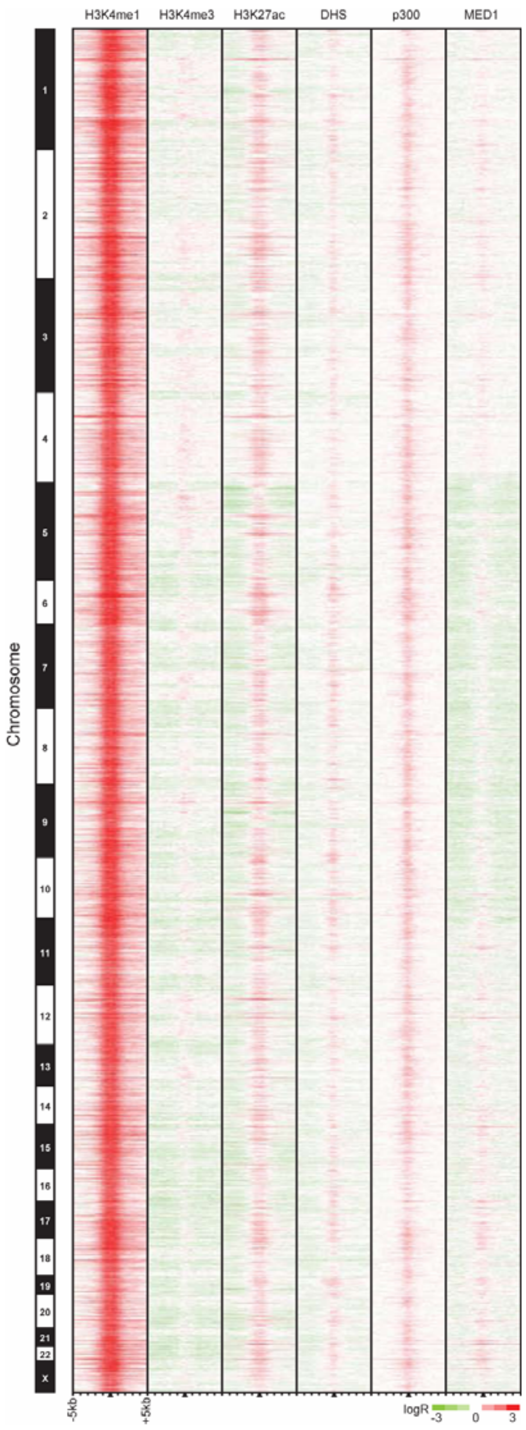


Figure 3-6: Genome-wide enhancer predictions in human cells

I predict 36589 enhancers in HeLa cells based on chromatin signatures for H3K4me1 and H3K4me3. Enhancer predictions are located at the center of 10 kb windows as indicated by black triangles, and ordered by genomic position. Enrichment data are shown for histone modifications (H3K4me1, H3K4me3, and H3K27ac), DNaseI hypersensitivity (DHS), and binding of p300 and MED1.

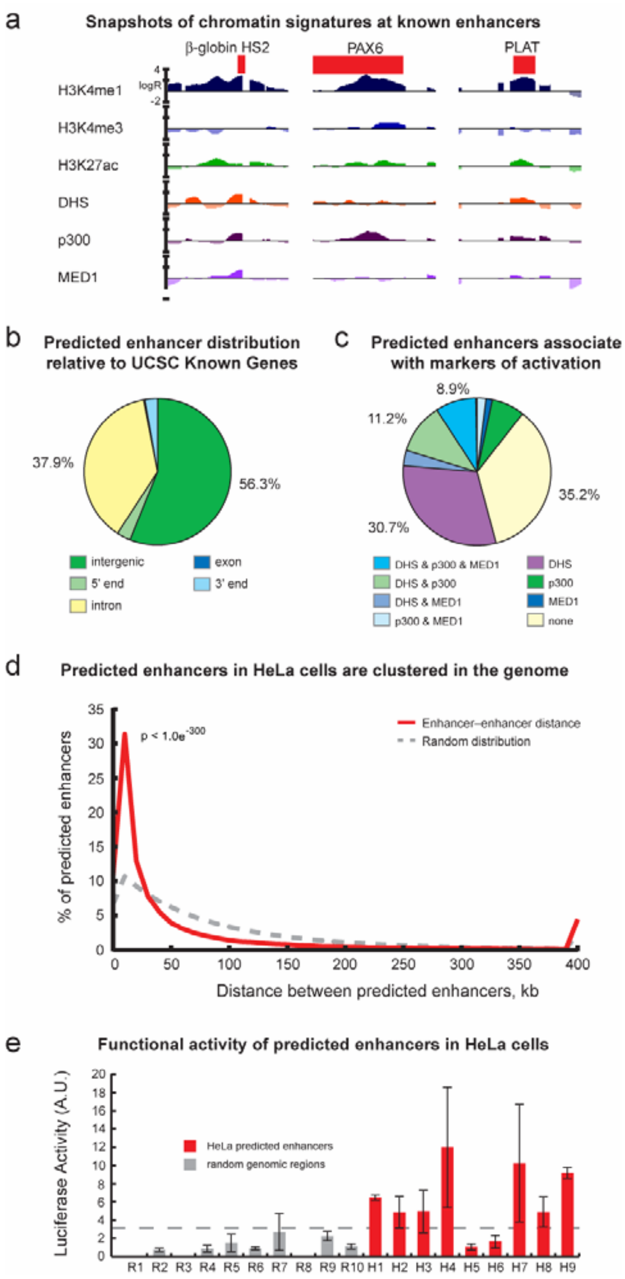


Figure 3-7: Validation of genome-wide enhancer predictions in human cells

(a) ChIP-chip enrichment profiles at several known enhancers (indicated in red) recovered by prediction: β-globin HS2, PAX6, and PLAT (5 kb windows centered on enhancer predictions. (b) Most enhancers have intergenic (56.3%) or intronic (37.9%) localization relative to UCSC Known Gene 5’-ends. (c) Most enhancers (64.8%) are significantly marked by DNaseI hypersensitivity, binding of p300, binding of MED1, or some combination thereof. (d) Distances between predicted enhancers are significantly smaller than expected by chance, suggesting that functional enhancers cluster in the genome. (e) 7 of 9 enhancers predicted in HeLa cells were active in reporter assays (red bars) as compared to none of the random fragments selected as controls (gray), where activity is defined as relative luciferase value greater than 2.33 standard deviations (p = 0.01) above the median random activity (gray dashed line).

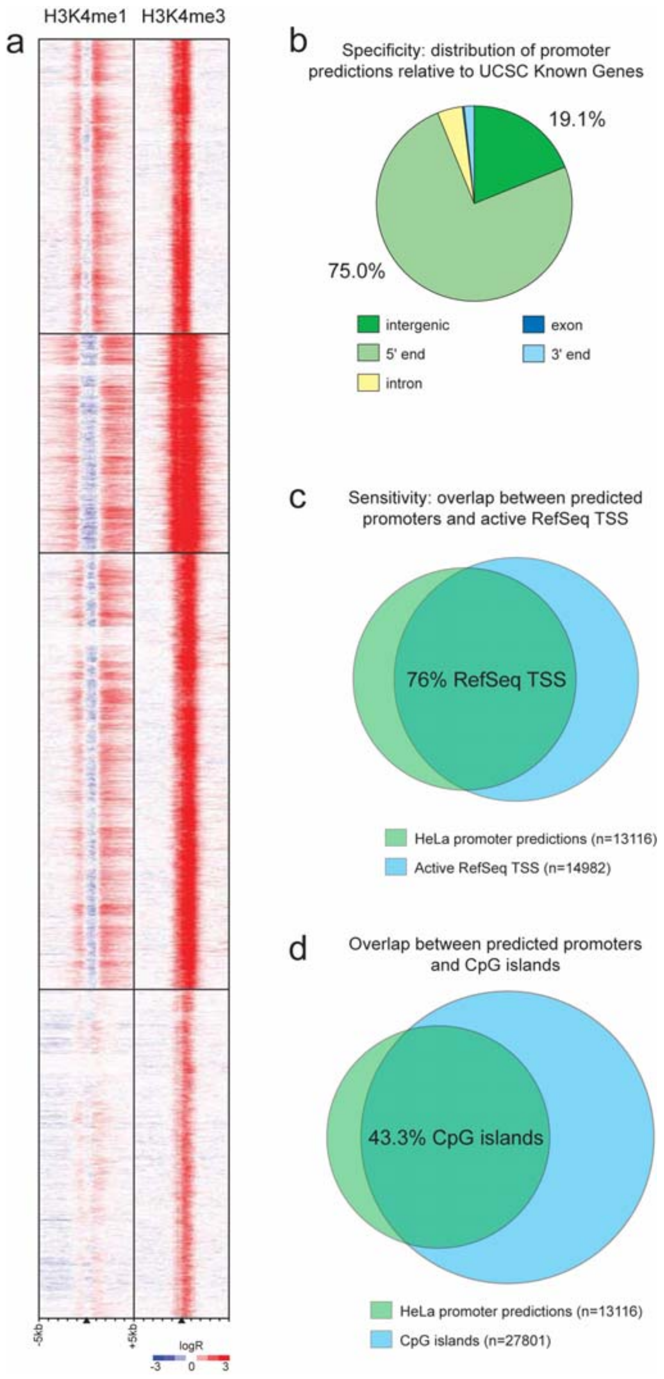


Figure 3-8: Active promoter predictions in the human genome.

(A) I predicted 13116 active promoters in HeLa cells based on chromatin signatures for H3K4me1 and H3K4me3. (B) 75% of promoter predictions map to 5' ends of UCSC Known Genes, indicating a high degree of specificity. (C) 76% of active promoters (defined as RefSeq TSS for expressed transcripts) are correctly predicted, indicating a high degree of sensitivity. (D) 85.1% of promoter predictions overlap with CpG islands (defined by UCSC Genome Browser), accounting for close to half of the CpG islands in the genome.

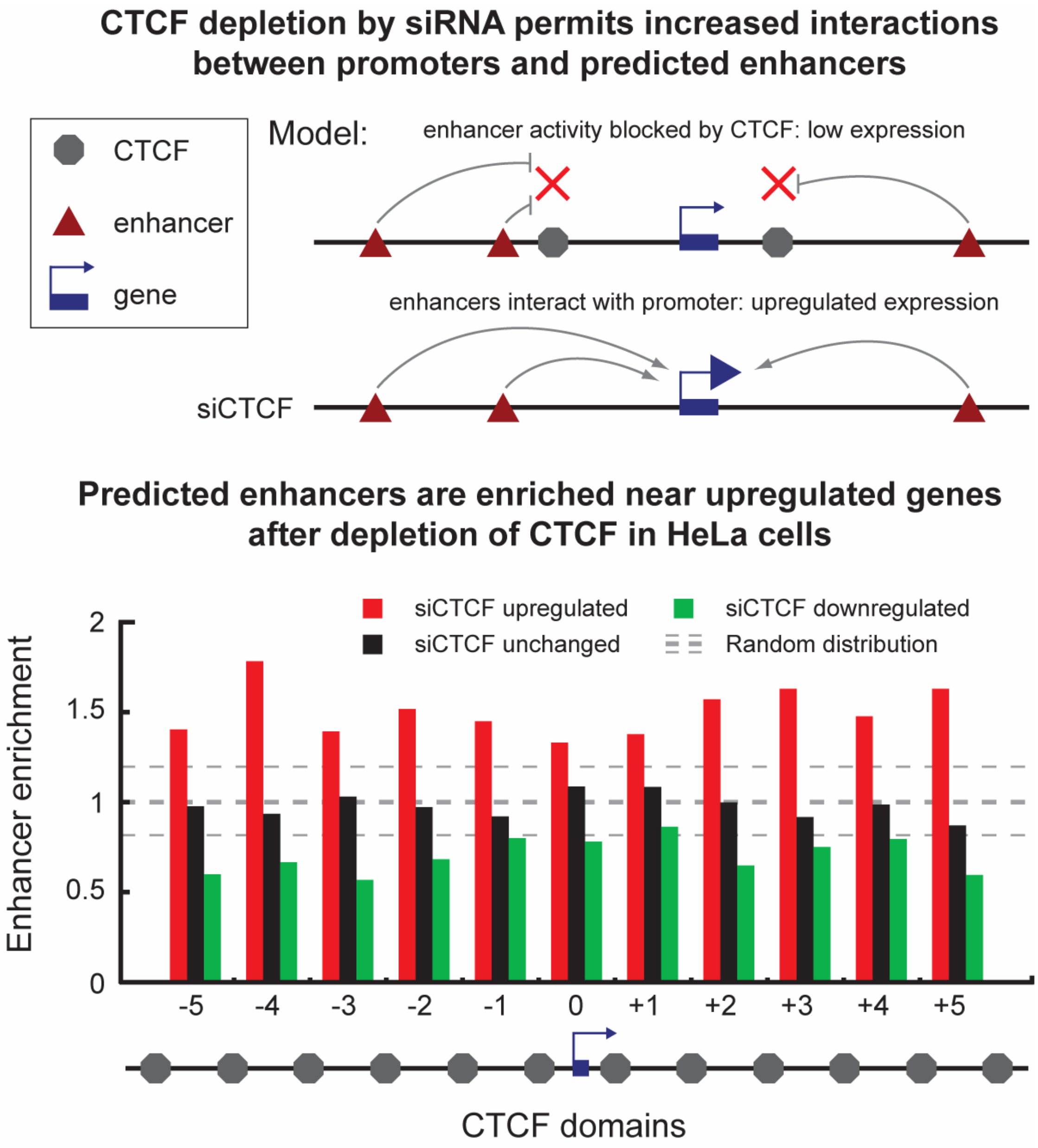


Figure 3-9: CTCF sites may serve as domain boundaries for promoter-enhancer interactions

Insulators bound by CTCF are thought to block promoter-enhancer interactions that would otherwise occur in the absence of CTCF (upper panel), a model supported by the enrichment of predicted enhancers in domains adjacent to genes that are upregulated in response to CTCF-depletion by siRNA (lower panel, red bars). Enhancers are depleted in domains adjacent to downregulated genes (green bars) relative to unchanged genes (black bars) and a random distribution (gray lines). Gene expression data are from a recently published study25. Domains are defined as the regions between CTCF sites as recently reported [6,12,26]; enhancers were counted in the five domains adjacent to each gene, upstream and downstream, and summed across respective domains to calculate enrichment relative to a random distribution.

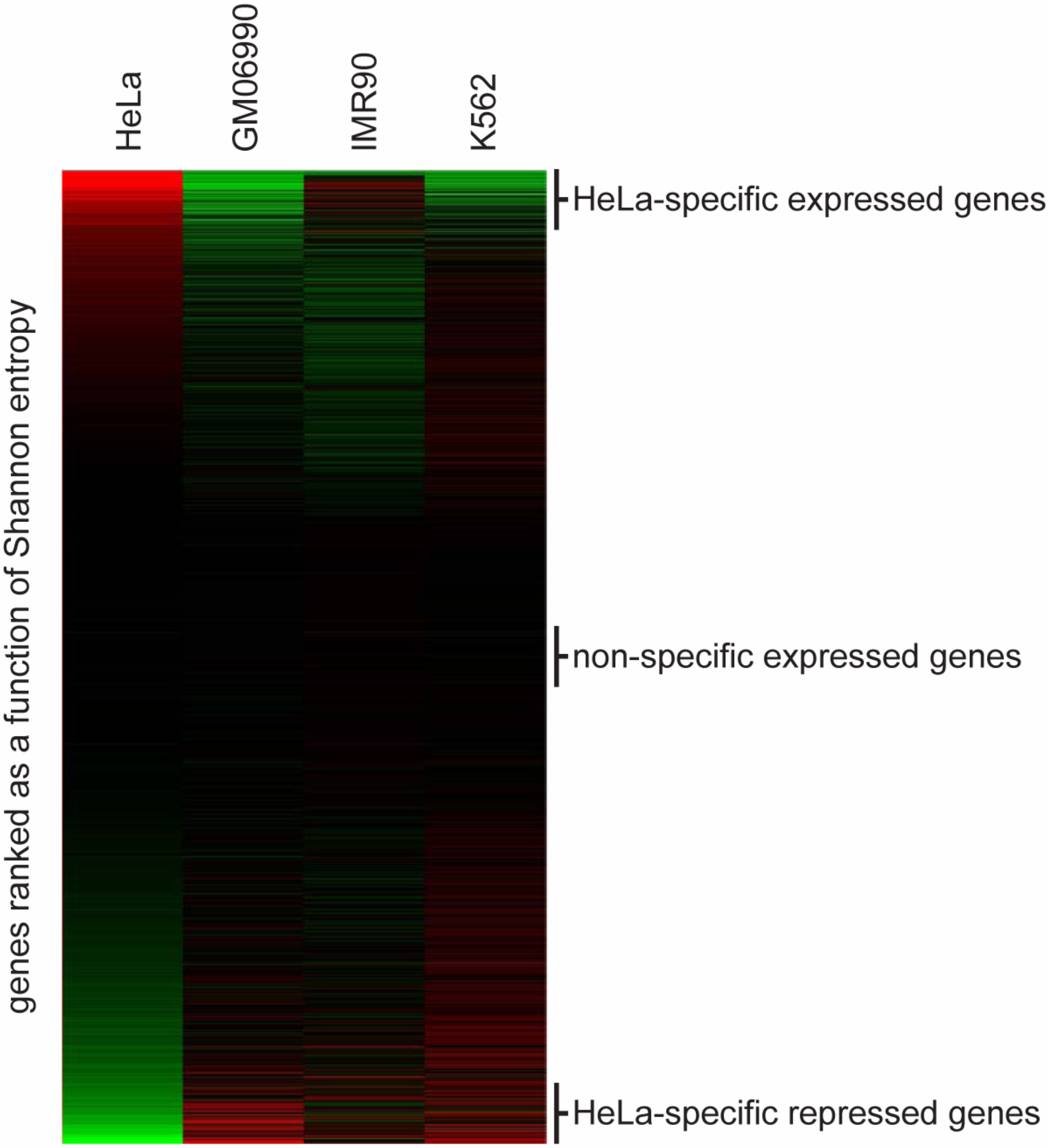


Figure 3-10: Comparison of cell type-specific gene expression in four cell types

I used Shannon entropy to rank genes by the specificity of their expression levels in HeLa as compared to three other cell lines (K562, GM06990, and IMR90 cells, representing leukemia, lymphoblast, and fibroblast lineages, respectively). The most HeLa-specific expressed genes are found at the top of the cluster, while genes that are specifically repressed in HeLa cells are found at the bottom. Genes in the middle portion of the cluster have expression levels that are similar in all four cell lines.

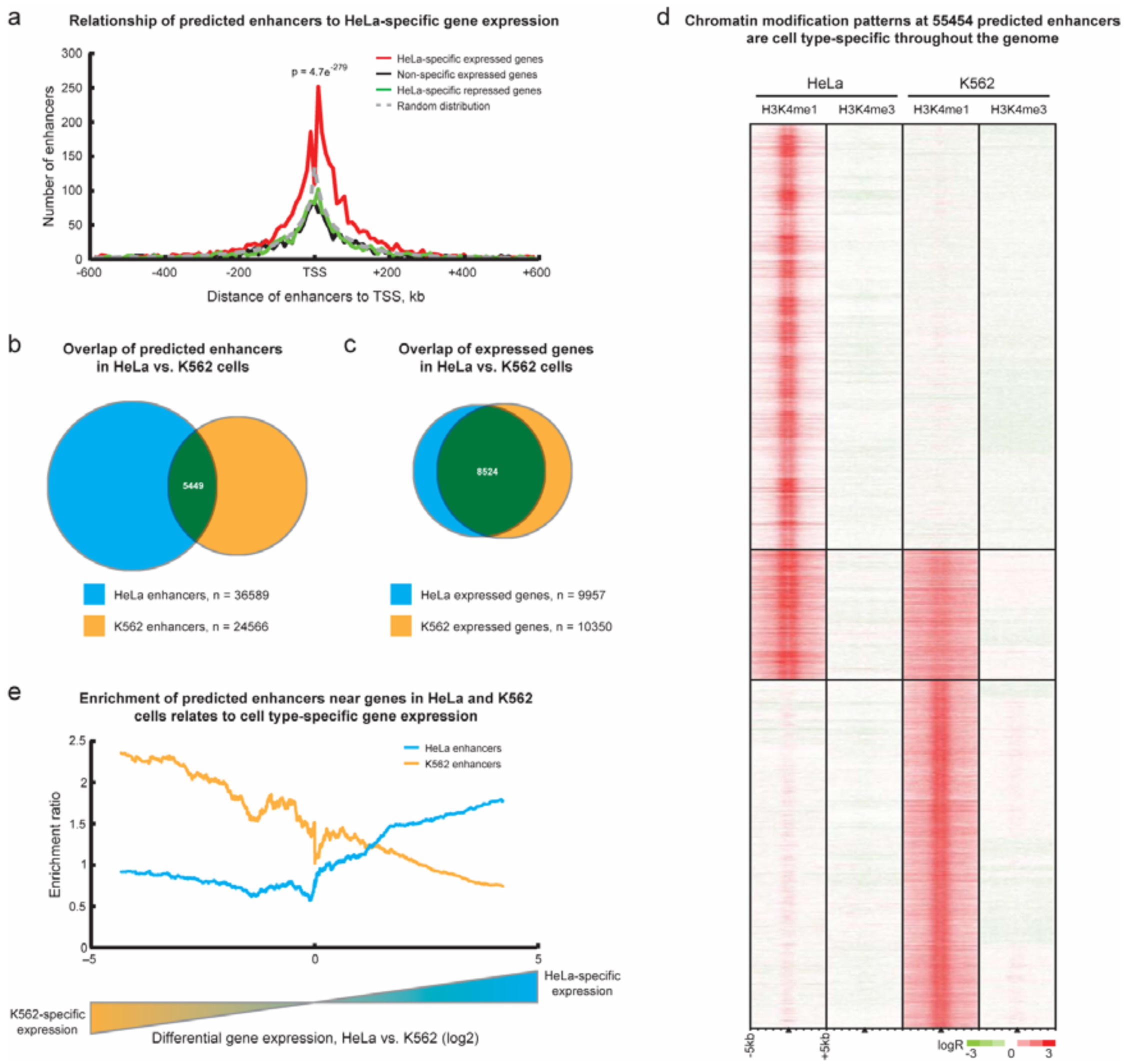


Figure 3-11: Chromatin modifications at enhancers are globally related to cell type-specific gene expression

(a) Enhancer localization relative to genes that are HeLa-specific expressed compared to K562, GM06990, and IMR90 cells (red), non-specific expressed (green), HeLa-specific repressed (black), and a random distribution (dashed grey). Predicted enhancers are enriched around HeLa-specific expressed genes within insulator-defined domains and depleted in domains of ubiquitous or non-expressed genes (p-value reflects significance of enhancer enrichment in domains of HeLa-specific expressed genes). (b) Most enhancers predicted in HeLa and K562 cells are cell-type specific while (c) most genes in HeLa and K562 cells are not specifically expressed. (d) Chromatin modification patterns are cell type-specific at the majority of 55454 enhancers predicted in HeLa and K562 cells. (e) Comparison of enhancer enrichment and differential gene expression between HeLa cells and K562 cells revealed that HeLa enhancers are enriched near HeLa-specific expressed genes (blue line) while K562 enhancers are enriched near K562-specific expressed genes (orange line).

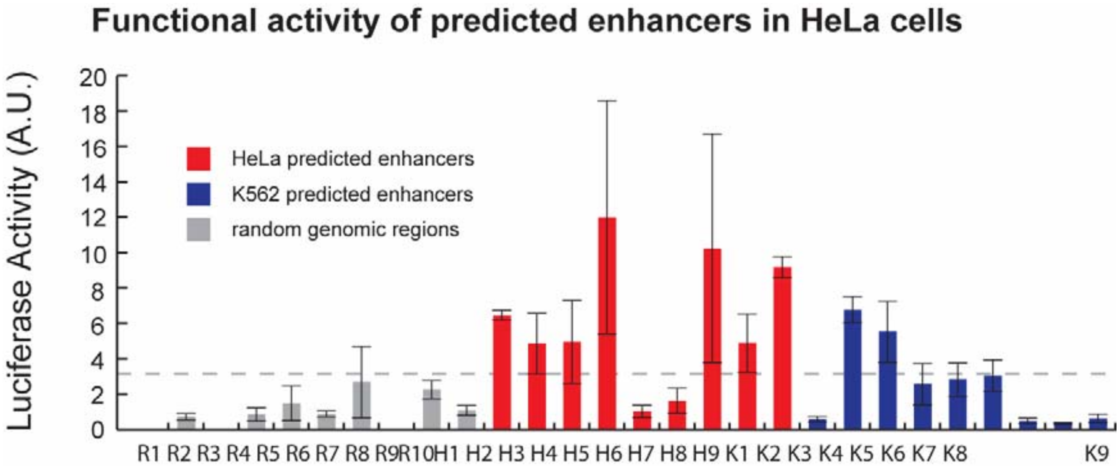


Figure 3-12: Cell type-specificity of predicted enhancer activity in reporter assays

In addition to the HeLa-specific enhancers and random regions assayed in Figure 3-7e, additional K562-specific enhancers were cloned and assayed for reporter activity in HeLa cells. Enhancers predicted specifically in K562 cells (blue bars) were much less likely to be active in HeLa cells than the HeLa-specific enhancers (red), and the median activity is not significantly different from random regions (gray). The dashed line represents a significance threshold of $p = 0.01$ as in Figure 3-7e.

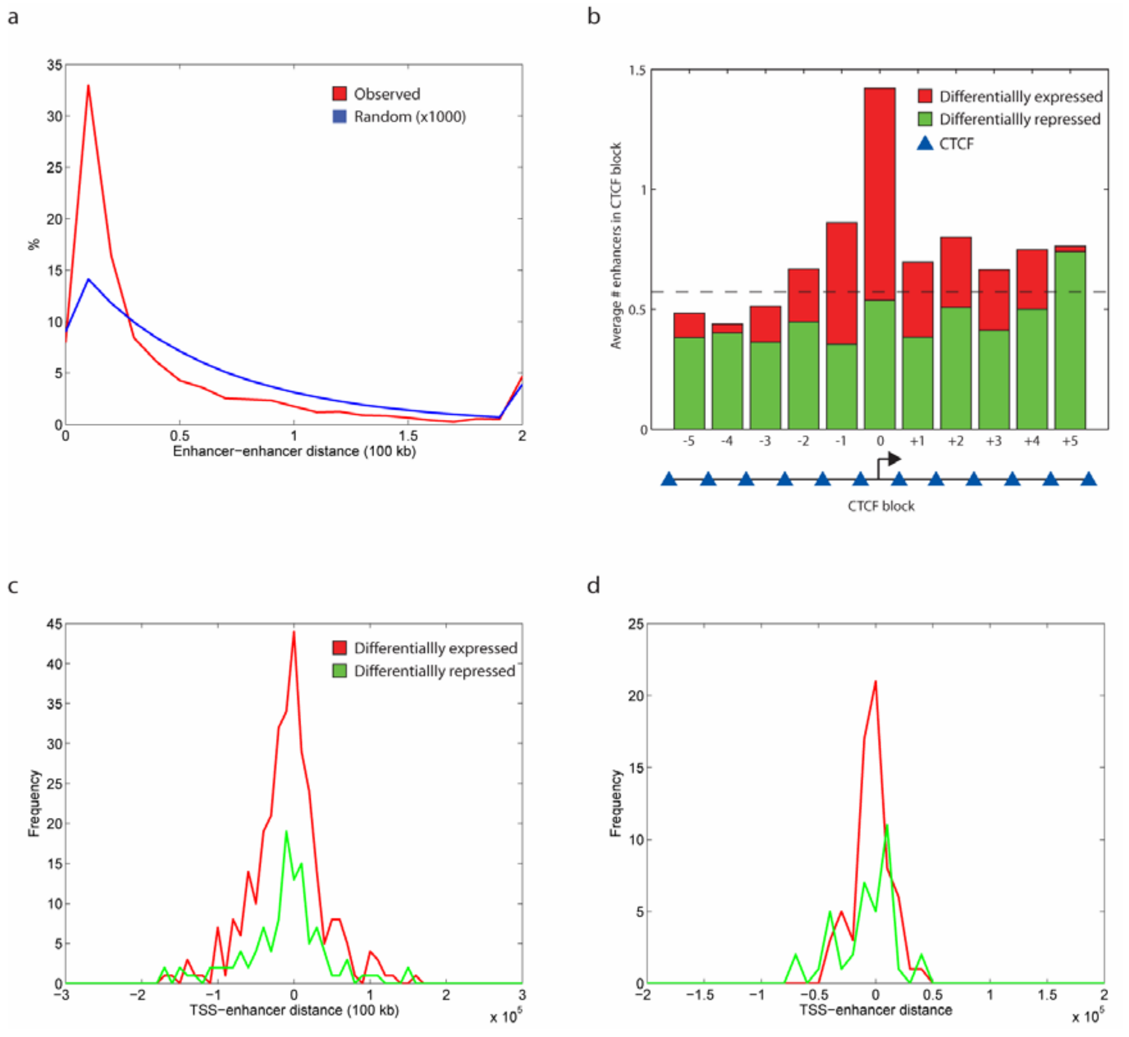
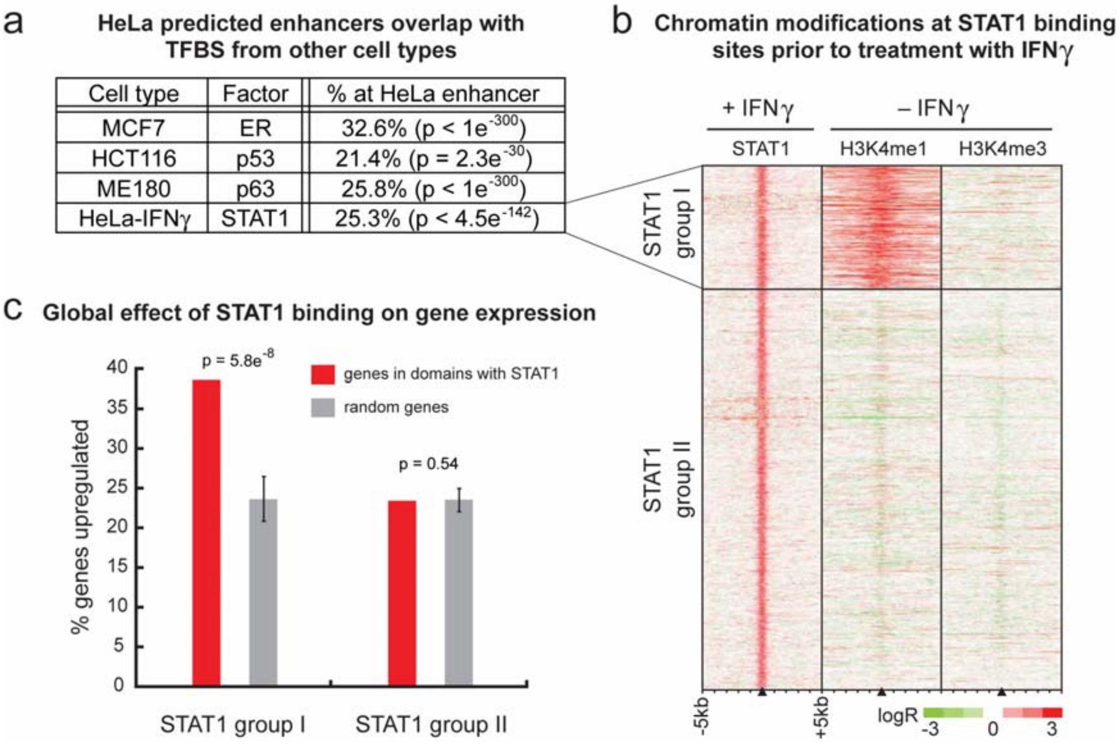


Figure 3-13: ENCODE Enhancers are clustered at differentially expressed genes.

(a) The distribution of adjacent enhancer-enhancer distances (red), as compared to 1000 sets of randomly placed sites (blue), indicates that enhancers are highly clustered. (b) A CTCF block is defined by flanking CTCF binding sites. Using the 729 consensus CTCF binding sites to define CTCF blocks, I count the average number of enhancers found in blocks relative to the TSSs of differentially expressed and repressed genes. Differentially expressed genes are enriched in enhancers when compared to differentially repressed genes, with the strongest enrichment found in CTCF block 0.The dotted line indicates the expected average number of enhancers in a CTCF block. For HeLa, GM, and K562, differential expression is defined by an RMA p-value cutoff of 0.01 and a fold change cutoff of 2.0. (c) A detailed view of the distribution of enhancers in CTCF block 0. Here, I show the distribution of enhancer-TSS distances all enhancers within this CTCF block. Negative distances indicate upstream enhancers, while positive distances indicate downstream enhancers. Enhancers are more concentrated to differentially expressed genes relative to differentially repressed genes. (d) Rather than examining the distribution of all enhancer-TSS distances in a differentially expressed/repressed gene’s CTCF block, I examine only the closest one here. While I do observe enrichment in differentially expressed genes, the effect is smaller than that observed when I consider all enhancer-TSS distances.



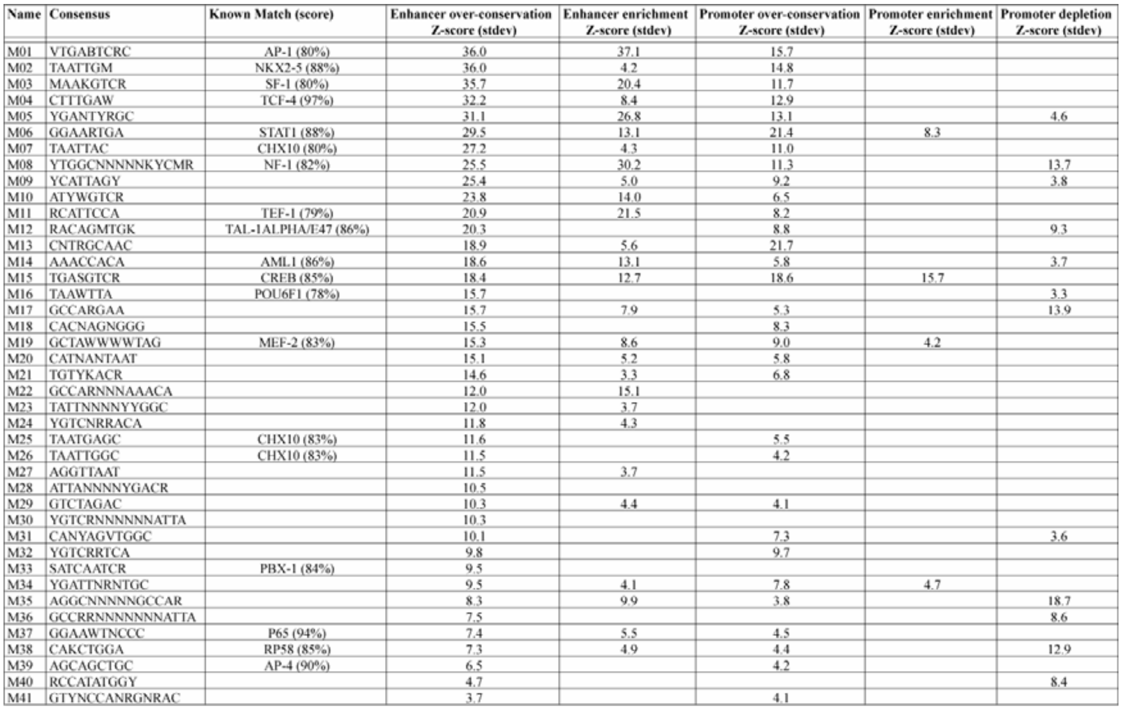
Cell type	Factor	% at HeLa enhancer
MCF7	ER	32.6% ($p < 1e^{-300}$)
HCT116	p53	21.4% ($p = 2.3e^{-30}$)
ME180	p63	25.8% ($p < 1e^{-300}$)
HeLa-IFNγ	STAT1	25.3% ($p < 4.5e^{-142}$)

Figure 3-14: Chromatin modifications are associated with increased regulatory response of transcription factor binding sites at enhancers

(a) Predicted enhancers in steady-state HeLa cells overlap with significant fractions of transcription factor binding sites (ER, p53, p63) in diverse cell types (MCF7, HCT116, ME180), as well as with STAT1 binding sites in HeLa cells treated with the cytokine interferon-gamma (HeLa-IFNγ). (b) Hundreds of STAT1 binding sites after treatment (+IFNγ) are marked by the enhancer chromatin signature in HeLa cells even prior to treatment (-IFNγ). (c) In HeLa cells treated with IFNγ (upper panel), gene expression is significantly (p = 5.8 X 10-8) more likely to be induced by STAT1 binding at sites with the enhancer chromatin signature (red, STAT1 group I) than by STAT1 binding at other distal sites (red, STAT1 group II) relative to a random distribution (gray).

Table 3-1: De novo motifs enriched in predicted enhancer regions

Known Match score represents the shared information content between novel and known motif [32]. Over-conservation is calculated as the excess conservation of a motif in enhancers or promoters relative to that expected for a random motif of identical composition. Enrichment is calculated as the over-abundance of a motif in enhancers or promoters relative to that expected for a random motif of identical composition. Enhancer-specific motifs are those lacking significant promoter enrichment. All significance values are expressed as Z-scores, corresponding to the number of standard deviations away from the mean of a normal distribution.



Name	Consensus	Known Match (score)	Enhancer over-conservation Z-score (stdev)	Enhancer enrichment Z-score (stdev)	Promoter over-conservation Z-score (stdev)	Promoter enrichment Z-score (stdev)	Promoter depletion Z-score (stdev)
M01	VTGABTCRC	AP-1 (80%)	36.0	37.1	15.7		
M02	TAATTGM	NKX2-5 (88%)	36.0	4.2	14.8		
M03	MAAKGTCR	SF-1 (80%)	35.7	20.4	11.7		
M04	CTTTGAW	TCF-4 (97%)	32.2	8.4	12.9		
M05	YGANTYRGC		31.1	26.8	13.1		4.6
M06	GGAARTGA	STAT1 (88%)	29.5	13.1	21.4	8.3	
M07	TAATTAC	CHX10 (80%)	27.2	4.3	11.0		
M08	YTGGCNNNNNKYCMR	NF-1 (82%)	25.5	30.2	11.3		13.7
M09	YCATTAGY		25.4	5.0	9.2		3.8
M10	ATYWGTCR		23.8	14.0	6.5		
M11	RCATTCCA	TEF-1 (79%)	20.9	21.5	8.2		
M12	RACAGMTGK	TAL-1ALPHA/E47 (86%)	20.3		8.8		9.3
M13	CNTRGCAAC		18.9	5.6	21.7		
M14	AAACCACA	AML1 (86%)	18.6	13.1	5.8		3.7
M15	TGASGTCR	CREB (85%)	18.4	12.7	18.6	15.7	
M16	TAAWTTA	POU6F1 (78%)	15.7				3.3
M17	GCCARGAA		15.7	7.9	5.3		13.9
M18	CACNAGNGGG		15.5		8.3		
M19	GCTAWWWWTAG	MEF-2 (83%)	15.3	8.6	9.0	4.2	
M20	CATNANTAAT		15.1	5.2	5.8		
M21	TGTYKACR		14.6	3.3	6.8		
M22	GCCARNNNAAACA		12.0	15.1			
M23	TATTNNNNYYGGC		12.0	3.7			
M24	YGTCNRRACA		11.8	4.3			
M25	TAATGAGC	CHX10 (83%)	11.6		5.5		
M26	TAATTGGC	CHX10 (83%)	11.5		4.2		
M27	AGGTTAAT		11.5	3.7			
M28	ATTANNNNYGACR		10.5				
M29	GTCTAGAC		10.3	4.4	4.1		
M30	YGTCRNNNNNNATTA		10.3				
M31	CANYAGVTGGC		10.1		7.3		3.6
M32	YGTCRRTCA		9.8		9.7		
M33	SATCAATCR	PBX-1 (84%)	9.5				
M34	YGATTNRNTGC		9.5	4.1	7.8	4.7	
M35	AGGCNNNNNGCCAR		8.3	9.9	3.8		18.7
M36	GCCRRNNNNNNNATTA		7.5				8.6
M37	GGAAWTNCCC	P65 (94%)	7.4	5.5	4.5		
M38	CAKCTGGA	RP58 (85%)	7.3	4.9	4.4		12.9
M39	AGCAGCTGC	AP-4 (90%)	6.5		4.2		
M40	RCCATATGGY		4.7				8.4
M41	GTYNCCANRGNRAC		3.7		4.1		

Chapter 4 : Chromatin states in human ES cells reveal key regulatory sequences and genes involved in pluripotency and self-renewal

Abstract

Human embryonic stem cells (hESCs) are offering a new therapeutic approach because of their unique ability to proliferate indefinitely *in vitro* and differentiate into multiple cell types. However, the molecular mechanisms of pluripotency and self-renewal remain incompletely understood. To elucidate the key regulatory sequences and genes responsible for these cellular properties, I have determined potential enhancers and insulators in the genome of human ES cells and examined the dynamics of four key chromatin modifications (H3K4me1, H3K4me3, H3K27ac and H3K27me3) at both promoters and enhancers during the differentiation of these cells. I observe that most enhancers gain or lose H3K4me1 and H3K27ac during differentiation in a manner that correlates with expression of their potential target genes. By contrast, chromatin modifications at promoters remain stable and largely invariant during hESC differentiation, with the exception of a small number of promoters where a dynamic switch between acetylation and methylation at H3K27 marks the transition between activation and silencing of gene expression. These results reveal more than 50,000 potential enhancers for early human development, and identify key genes that are involved in differentiation and maintenance of pluripotency in human ES cells.

Introduction

Human embryonic stem cells (hESCs) are derived from the inner cell mass of the blastocyst [1]. Due to their ability to self-renew while retaining the potential to differentiate into most other cell types in the body, there has been growing interest to explore hESCs in regenerative medicine, and as a model system to study early human development.

Transcriptional regulation is a fundamental aspect of the molecular mechanisms controlling self-renewal, pluripotency and lineage specification. A core transcriptional regulatory network consisting of transcription factors OCT4, SOX2, NANOG, TCF3 and their regulatory target genes is believed to control the gene expression program to maintain self-renewal and pluripotency in hESC [2,3]. In addition, chromatin state throughout the human genome also appears to play important roles in this process [4]. For example, trimethylation of both histone H3 lysine 4 and lysine 27 at gene promoters (termed bivalent domains) has been proposed as a mechanism for regulating development and proliferation [5,6,7,8,9,10].

To understand how the core transcriptional network regulates gene expression, chromatin immunoprecipitation based analysis has been used to identify the promoters bound by OCT4, SOX2, and NANOG in hESCs [2], and the results identified an extensive auto-regulatory and feed forward loop. Recently, more extensive transcription factor networks were established in mESCs using similar approaches and promoter microarrays [11]. The surprising finding is that a large number of promoters appear to be regulated by multiple transcription factors. While these studies focusing on transcriptional promoters have suggested critical roles that some genes and their promoters play in regulating self-renewal and pluripotency, other genes and genomic sequences that play important roles in this process remain to be identified. Consistent with this notion, recent studies using ChIP-Seq [12] to investigate the binding sites of 13 site-specific transcription factors in mESC provided evidence that pluripotent factors frequently act from promoter-distal genomic sequences to regulate pluripotency genes. Furthermore, we and others recently demonstrated that transcriptional enhancers, a class of promoter-distal regulatory sequences, generally play important roles in driving tissue- and cell-type specific gene expression [13,14]. Analysis of the enhancers in the ES cell genome should therefore provide insight into key genes and regulatory sequences for self-renewal and pluripotency.

To identify key genes and regulatory sequences involved in self-renewal and pluripotency, I have determined potential enhancers in the genome of human ES cells and examined the dynamics of chromatin state at both promoters and enhancers during the differentiation of these cells. I identify over 50,000 potential enhancers in the undifferentiated ES cell (hESC) and differentiated ES cell (dESC) genomes. There are remarkable differences of chromatin dynamics at human promoters and enhancers. The chromatin state at promoters is generally stable during differentiation, with a small fraction undergoing changes that primarily involve a switch between active acetylation and repressive methylation at H3K27 which define a set of genes that appear to be important for maintenance of ES cell pluripotency, and another set that are involved in differentiation. By contrast, a majority of the enhancers display striking changes in chromatin states in a manner that correlates with differential expression of their predicted target genes. In addition, I also identify a set of poised enhancers marked by a distinct chromatin signature near genes important for cell fate determination, underscoring the importance of these regulatory elements in regulating differentiation.

Results and Discussion

Genome-wide maps of chromatin state in hESC before and after differentiation

Low passage (20-50) hESCs (H1) were grown in feeder cell free medium TeSR1 by my laboratory's collaborator James Thomson, as described [15]. To differentiate the hESCs, the cells were treated with BMP4 for 4 to 6 days, giving a heterogeneous cell population which is a mixture of endoderm (lineage markers: GATA4, GATA6, SOX17), mesoderm (FOXF1, GATA5, CXCR4), and trophectoderm (CDX2, GATA2, GATA3).

My lab utilized chromatin immunoprecipitation coupled with genome-wide tiling microarrays (ChIP-chip) [16] (Figure 4-1) to map chromatin modifications in the genomes of both hESCs and dESCs at high resolution. The focus was on four modifications – H3K4me1, H3K4me3, H3K27ac and

H3K27me3. Previous chapters have demonstrated that the patterns of H3K4me1 and H3K4me3 profiles along the genome allows for identification of potential enhancers in particular cells. Additionally, the methylation at H3K27 has been demonstrated to play a critical role in silencing of gene expression in ES cells [17]. I have also recently suggested that H3K27 may also be acetylated at active gene promoters [13]. By comparing the genome-wide maps of these four chromatin modifications from hESCs to those in dESCs, I hypothesized that I would be able to identify key promoters and enhancers that contribute to maintenance of pluripotency and self-renewal.

Dynamic switch between acetylation and methylation at H3K27 during hESC differentiation

Promoters are key transcriptional regulatory sequences that integrate extracellular and intracellular inputs to control transcriptional initiation of genes. Previous studies have identified methylation of H3K4 and H3K27 at promoters to be important for the poised state of some key developmental regulator genes. These promoters are not transcribed in ES cells, but could either become activated during differentiation when the methylation mark on H3K27 is lost, or permanently silenced when the H3K4me3 modification is erased [6,17]. To determine whether additional promoters display dynamic changes in chromatin modification during ES cell differentiation, I examined modifications on H3K4 and H3K27 in both hESCs and dESC. I found that the presence of H3K4me3 reveals little information in terms of gene activation, as enrichment of this mark appears invariant during differentiation (Figure 4-2, Figure 4-4). This observation is in agreement with several recent studies finding this modification to be present at 70-80% of known TSS [7,8,13,18,19]. Interestingly, when I examined modifications to H3K27, I found a number of promoters displaying a switch between acetylation and methylation (Figure 4-4). Trimethylation of this residue (H3K27me3) is a known marker of repressed promoters [17,20], in contrast to acetylation (H3K27ac), which is generally a hallmark of active chromatin [21,22]. My results indicate that these two modifications, residing on the same residue, are mutually exclusive: H3K27me3-marked promoters show no enrichment for H3K27ac, while those marked by H3K27ac are not enriched for H3K27me3.

To quantify how these modifications switch upon differentiation, I ranked TSSs by the change in levels of active H3K27ac and repressive H3K27me3: $C_g = (\text{H3K27ac}_{\text{dESC}} - \text{H3K27ac}_{\text{hESC}}) - (\text{H3K27me3}_{\text{dESC}} - \text{H3K27me3}_{\text{hESC}})$ (Figure 4-2). Genes with low C_g exhibit a combination of H3K27ac loss and H3K27me3 gain after differentiation. Examination of gene expression reveals that in general these genes are actively transcribed in hESC and repressed in dESC. This class of genes is of particular interest as it contains the key stem cell transcription factors OCT4 (POU5F1), SOX2, and NANOG. For example, SOX2 shows hyper-acetylation at H3K27 in hESCs that is lost following differentiation and becomes repressed by H3K27me3 (Figure 4-3). Additional genes showing the same active to repressive switch include notable transcription factors and signaling molecules likely important in the regulation of ESC pluripotency and self-renewal (Table 4-1). For example, of just the few gene promoters included, a number of WNT signaling factors are revealed, including TCF7L1, FZD7, FZD8 and SFRP2. Based on the C_g metric of change in chromatin structure, OCT4, SOX2, and NANOG ranked 30, 1, and 155, respectively, among the top 1% of 22047 genes. However, based on changes in gene expression, these genes ranked 2591, 13, and 637, respectively, only among the top 12% of all genes. Thus, the specificity of a chromatin-based list in predicting key stem cell genes is likely much higher than that of an expression-based list.

By contrast, genes with high C_g show gain of H3K27ac and loss of H3K27me3 upon differentiation. These genes show the opposite expression pattern to that of low C_g genes, illustrating the close correlation between epigenetic modifications and gene expression. For example, the transcription factor gene HAND1 shows no H3K27ac in the hESC epigenome but is enveloped by H3K27me3 marked chromatin. Following differentiation, HAND1 undergoes a complete switch: losing H3K27me3, gaining H3K27ac, and becoming actively expressed (Figure 4-3). These results agree with recent findings examining H3K27me3 loss at developmentally important gene promoters [6,8,17,19].

Overall, only 5.7% of all promoters exhibit at least a 2-fold change in H3K27 chromatin state during hESC differentiation, defining a set of genes likely integral in each cell type.

Genome-wide identification of enhancers in hESCs and early development

Recent studies have suggested that enhancers play important roles in cell type- and tissue-specific gene expression [13]. To identify enhancers that regulate stem cell gene expression during differentiation, I employ a computational algorithm that identifies potentially active enhancers based on chromatin modification patterns of H3K4me1 and H3K4me3 [13,23]. This method predicts 28,809 enhancers in hESCs and 33,369 in dESCs. The distribution of the chromatin-predicted enhancers is primarily distal to the TSSs, with approximately 50% lying in intergenic regions for each cell type and just over 40% falling in intragenic regions, above what is expected at random (Figure 4-5a). Additionally, these enhancers tend to be clustered, indicating that multiple enhancers may act together to drive gene expression (Figure 4-5b). To validate the function of the predicted enhancers, my lab cloned 17 enhancers downstream of the Luciferase gene in a reporter construct and tested luciferase expression in human ES cells after transient transfection. Of the 17 putative enhancer constructs tested in this assay, 14 (82%) showed higher level of enhancer activity ($p = 0.01$) compared to random genomic regions that showed no significant reporter activities (Figure 4-5c). These results support the accuracy of the enhancer predictions.

Motif analysis of hESC enhancers identifies key transcription factors

To identify the common themes of enhancer sequences and further elucidate the transcriptional regulatory mechanisms guiding embryonic stem cells and differentiation, our collaborator Ron Stewart investigated if known transcription factor binding sites (TFBS) from the JASPAR and TRANSFAC databases were enriched at enhancers in a cell type-specific manner. They identified both hESC-specific motifs and dESC-specific motifs (Table 4-2). The high confidence hESC-specific motifs include those that are recognized by KLF4 and c-MYC, two transcription factors that are capable of reprogramming human fibroblasts to become iPS cells when transduced with OCT4 and SOX2 [24,25]. Also included in this list is a motif for FOXD3, which is known to be involved in maintaining mouse ESCs and in the hESC pluripotency gene regulatory network [8,26]. Only the TRANSFAC database contains motifs for OCT4 and SOX2, and 940 hESC enhancers contain a joint OCT4:SOX2 motif, consistent with the role of these two factors in regulating ES cell gene expression [2]. Additionally, a number of motifs are consistently found from both databases. In contrast to the hESC-specific motifs, the high confidence dESC-specific enhancer motifs represent several transcription factors known to be involved in early development or differentiation, including Brachyury (mesoderm gene expression), FOXC1 (heart field specification), the Myf family (myogenesis), and ZEB1 (epithelial-mesenchymal transitions) [27,28,29,30] (Table 4-2) . Of the transcription factor motifs classified at dESC-specific enhancers, none of the corresponding factors are known to play a role in human ESC maintenance or in reprogramming to an induced pluripotent state.

If the predicted enhancers function *in vivo*, one expectation is significant binding of transcription factors. In order to test this hypothesis, my lab employed high-throughput sequencing coupled with chromatin immunoprecipitation (ChIP-Seq) to determine the binding sites for SOX2 and NANOG. I identified 4,818 SOX2 and 20,973 NANOG binding sites (FDR = 1%) using the MACS peak finding software [31] against a background of input hESC DNA. Comparing to putative hESC enhancers, 39.1% and 35.5% of the SOX2 and NANOG binding sites were recovered, respectively, compared with 4.5% and 4.3% at putative dESC enhancers (Figure 4-5d). Additionally, a number of binding sites not recovered by hESC enhancer predictions show a weak enrichment of H3K4me1 in hESCs but not dESCs, which may reflect enhancers missed by the prediction algorithm (Figure 4-6). The presence of these key stem cell regulators at enhancers strongly suggests a central role of enhancers in defining the ES cell gene expression program. These results indicate that other transcription factors with motifs enriched in hESC enhancers such as KLF4, MYC, and FOXD3 likely bind to the predicted hESC enhancers, and play important roles in self-renewal or maintenance of pluripotency.

Dynamics of chromatin state at enhancers reveal cell type-specific usage

Since promoters that undergo dynamic changes in chromatin structure generally belong to key stem cell and developmental genes, I wondered if chromatin dynamics at enhancers would identify key sequences regulating the same processes. To assess the dynamics of chromatin modifications at human enhancers, I clustered H3K4me1, H3K4me3, H3K27me3, and H3K27ac at the predicted enhancers. Most predicted enhancers exhibit dramatic gains or losses of H3K4me1 and H3K27ac during differentiation (Figure 4-7). Of particular note is the general absence of H3K27me3 at these sequences, suggesting that this repressive modification is mainly acting on promoters. In contrast, a significant number of enhancers are associated with H3K27ac. I ranked the predicted enhancers by the change in levels of acetylation between hESCs and dESCs: $C_e = (\text{H3K27ac}_{\text{dESC}} - \text{H3K27ac}_{\text{hESC}})$ (Figure 4-8). Just as individual enhancers studies have shown the presence of hyper-acetylation [21,22,32,33,34,35,36], I find that hyper-acetylated enhancers tend to be cell-type specific. In addition, hyper-acetylated enhancers are nearer to up-regulated genes than enhancers lacking acetylation (Figure 4-8), suggesting a role of H3K27ac in modulating enhancer activity.

CTCF-organized regulatory domains predict enhancer targets

Genes regulated by enhancers marked in a cell type-specific manner likely contribute to defining the unique abilities of stem cells. However, to find these target genes, it is first necessary to link enhancers to the genes they regulate. To do this, I focused on the vertebrate insulator binding protein CTCF [37,38], which is known for its enhancer-blocking activity when bound between enhancers and promoters [39,40]. Therefore, to complete the cis-regulatory map, my lab performed ChIP-chip to map 33,302 CTCF binding sites (FDR = 1%) in the hESC genome. CTCF sites show minimal variation across multiple cell types, allowing use of the hESC genome-wide CTCF binding map in dESCs as well [13,41,42] (Figure 4-10). I then partitioned the genome into CTCF-Organized Regulatory Domains (CORDs), cis-regulatory blocks flanked by CTCF binding sites (Figure 4-9a). If the model of CTCF function is true, then I expect hESC-specific enhancers to be highly enriched in CORDs containing hESC-specific genes compared to dESC-specific genes, and vice-versa. Using the C_e ranking from Figure 4-8, I divided the predicted enhancers into three equal-sized groups that are hESC-specific, non-specific, and dESC-specific. I observed that hESC-specific enhancers are highly enriched within the CORDs containing the 1000 most hESC-specific genes. Similarly, dESC-specific enhancers are enriched within CORDs containing the 1000 most dESC up-regulated genes (Figure 4-9b). In contrast, neighboring CORDs do not show enrichment of cell type-specific enhancers (Figure 4-9c). In agreement with results from Chapter 3, these results also suggest enhancers may play a key role in regulating gene expression from promoters in the same CORD.

Through the examination of enhancer enrichment relative to all genes within their respective CORD, I observe that CORDs containing differentially expressed genes are enriched with cell type-specific enhancers, while non-differentially expressed genes remain static for enhancer enrichment (Figure 4-9d). The dynamics of chromatin reorganization upon differentiation also reveals that enhancers are generally weak, act synergistically and as the number of enhancers increases within CORDs, and differential expression increases linearly on a log scale (Pearson correlation = 0.82) (Figure 4-11).

Finally, to identify additional genes important in self-renewal, pluripotency, and differentiation, I extended this analysis to predict promoter targets within the CORDs of the top 1% of hESC-specific enhancers and dESC-specific enhancers based on C_e. These lists provide additional candidates for genes important in defining each cellular state. As confirmation, I discovered several putative enhancers in CORDs containing genes important for hESC regulation. A view of the SOX2 locus reveals a number of predicted enhancers downstream of the gene. To date, only a single enhancer has been identified in mouse ESCs, approximately 4-kb downsteam of the TSS [43]. I predicted a human enhancer in this region that is epigenetically marked only in hESCs, one of several predicted enhancer elements downstream of the gene (Figure 4-9e). Additionally, I predicted three hESC-specific enhancers upstream of FOXD3, a gene that is important for pluripotency and known to activate Nanog and Oct4 expression in mouse ESCs [26] (Figure 4-12). I also predicted several hESC-specifically marked enhancers in the CORDs containing OCT4 and NANOG, as well as a number of other genes required for ES cell pluripotency. The functional validation of these enhancers is illustrated in Figure 4-5c.

Genes regulated by cell type-specific enhancers likely contribute to defining each cellular state. Further examination of enhancer gene targets include JMJD2C, JARID2, LEFTY1, as well as other transcription factors, and MAP kinase signaling molecules in hESCs, while dESC enhancer targets reveal genes such as several HOX and GATA factors. By linking enhancers to target promoters, these results allow for the expansion of regulatory networks and provide a more precise depiction of regulatory pathways in ES cells.

Chromatin dynamics at poised enhancers correlate with cell fate commitment

One of the most intriguing aspects of embryonic stem cells is their ability to differentiate into a variety of other cell types in the body in response different environmental cues. My analysis shows that there are three classes of epigenetically-marked enhancers: those marked specifically in hESCs, those marked specifically in dESCs, and those marked in both. While the first and second groups are enriched near genes specifically expressed in hESCs and dESCs, respectively, enhancers marked in both cell types are enriched near both hESC- and dESC-specific genes (Figure 4-13a, Figure 4-14). To investigate the mechanisms that lead to the reprogramming of the hESC transcriptome, I examined this class of 8863 shared enhancers that are marked before and after differentiation, reasoning that extracellular signaling may act through some of these sequences to activate a group of key regulators for cell fate determination.

Particularly interesting within the class of 8863 enhancers marked in both cells types are those that are enriched in CORDs containing dESC-specific genes (Figure 4-13a). Many of these shared enhancers are only marked by H3K4me1 in ES cells, but upon differentiation they gain H3K27ac (Figure 4-14a). Since H3K27ac is a mark of activity, I hypothesized that these enhancers may be inactive in ES cells but poised and awaiting a regulatory signal to activate them, therefore giving rise to acetylation and differentiation. If true, then I expect these enhancers to be enriched near genes induced early during differentiation. When I examined the enrichment of shared enhancers near genes differentially up-regulated at various time points during BMP4 treatment (3, 6, 12, 24, 48, and 120 hrs), I indeed observed that this set of poised enhancers is significantly enriched in CORDs containing early response genes (Figure 4-13c). This is in contrast to the most dESC-specific acetylated enhancers from Figure 4-8 (Figure 4-14b) or the shared enhancers that lose acetylation which show no enrichment near the same genes (Figure 4-13d).

Interestingly, the enhancers in this category can be found near genes coding for the developmental transcription factors MSX1 and MEIS1, which are up-regulated at 3hrs and 48hrs respectively. Each of these genes is highly expressed in dESCs and their CORDs contain numerous shared enhancers, but H3K27ac only marks the enhancers in dESCs (Figure 4-13b). In addition, BMP4 itself as well as downstream factors SMAD3, SMAD6, SMAD7 and ID2 are also found in this category at 3hrs. This set of genes contains a number of additional transcription factors, including HAND1, GATA3, CDX2, FOXO4, LEF, JUN, and SOX9. These 7 factors along with SMAD3 all have TFBS motifs enriched in dESC-specific enhancers, suggesting these factors go on to establish the cell fate through transcriptional regulation at enhancers. Thus, these results suggest that poised enhancers contribute to ES cell pluripotency by pre-marking enhancers for genes likely responsible for early steps in cell fate commitment.

Conclusions

I have analyzed chromatin modification dynamics to identify key genes and regulatory sequences contributing to human embryonic stem cell functions. I provide a global view of chromatin dynamics upon differentiation of hESCs, a crucial step in understanding how differential gene expression is controlled. By assessing how the chromatin state changes during differentiation of hESC, I reveal a chromatin switch at a subset of H3K27 gene promoter histones, assessing how repression by H3K27me3 during differentiation is important for hESCs. This subset of specifically regulated genes includes several stem cell-specific factors.

Additionally, I describe the first genome-wide maps of enhancers in hESCs and dESCs, showing that many enhancers are functionally active, are occupied by transcription factors, and are enriched for motifs. Furthermore, the vast number of enhancers implies that most genes are highly regulated through the use of enhancers. This is supported by the majority of mapped transcription factor binding sites observed outside of promoter regions. Additional evidence of this was recently demonstrated in mouse ESCs, showing that of the transcription factors studied, the majority of binding sites are also distal to promoters, especially pertaining to ES-specific factors [12]. I find that cell type-specific enhancer chromatin modifications correlate with cell type-specific gene expression within CTCF-organized regulatory domains (CORDs). The cell type-specific enhancer regulation of genes within CORDs expands the potential of an ESC regulatory network.

I also identify a set of poised enhancers marked by H3K4me1 in hESCs and dESCs that become acetylated upon differentiation. The poised enhancer state likely allows for activation of early response genes important for the initial steps in cell fate commitment, thereby contributing to stem cell pluripotency (Figure 4-15).

Methods

CTCF binding site location

My lab used the Mpeak program to determine binding sites of CTCF peaks as previously described [41] with the following modifications: peaks consisted of at least 3 consecutive probes having a signal threshold above 1.5 standard deviations at a false discovery rate of 1%.

Quantitative assessment of chromatin change

Below I describe the procedure for calculating C_g. The procedure for computing C_e is similar. The NimbleGen gene expression data span 22047 genes. For each gene and cell type, I calculate the sum of the log2 enrichment of H3K27ac and H3K27me3 in a 10-kb window centered at the TSS, and take the difference H3K27ac – H3K27me3 representing the enrichment of H3K27ac over H3K27me3 in a single cell type. I then compute the difference of this value over the 2 different cell types (dESC – hESC), and rank all genes using this difference. Negative differences indicate ES-specific H3K27ac and dES-specific H3K27me3. Positive differences indicate ES-specific H3K27me3 and dES-specific H3K27ac.

Enhancer predictions

The procedure used to predict enhancers follows closely to that in Chapters 2 and 3. Specifically, I first bin the tiling ChIP-chip data into 100 bp bins, averaging multiple probes that fall into the same bin. Empty bins are interpolated if the distance between flanking non-empty bins is less than 1-kb, and set to 0 otherwise. I scan this binned data, keeping only those windows 1) in the top 10% of the intensity distribution and 2) having H3K4me1 and H3K4me3 profiles in the top 1% of all windows using the same training set of sites as in Chapter 2. I use a discriminative filter on H3K4me1 and H3K4me3 to keep only those sites that correlate with the averaged enhancer training set more than the promoter training set. Finally, I apply a descriptive filter on H3K4me1 and H3K4me3, keeping only those remaining predictions having a correlation of at least 0.5 with an averaged training set.

Motif Discovery

Data: 637 genes down-regulated during the first 48 hours of differentiation induced by BMP4 treatment were defined as human embryonic stem cell (hESC) specific genes while 1214 genes up-regulated 48 or more hours after BMP4 treatment were defined as differentiation specific genes. 1028 enhancers identified in hESCs were mapped to the hESC specific genes bounded by insulators, and 3221 enhancers identified in BMP4 with FGF (have to specify slightly different conditions here) differentiated cells were mapped to the differentiation specific genes bounded by insulators. Genomic sequences of these hESC and differentiation specific enhancers of 5000 kbs were extracted from the UCSC GoldenPath database of the hg17 assembly [44]. Two data sets with 1028 and 3221 random genomic sequences of 5000 kbs were also extracted from the same database as controls.

Procedure: 566 TRANSFAC [45] and 96 vertebrate transcription factors (TFs) motif matrices were downloaded from the JASPAR database [46]. MotifLocator, software based on a classical position-weight matrix scoring scheme, was downloaded from the INCLUSive database [47] and was used to search the hESC and differentiation specific enhancers for potential binding sites of the 96 TFs.

The motifs' ability to classify foreground sequences from background sequences was measured by the balanced misclassification error rate (1). The error was defined as:

$$ErrorRate = 1 - \left[(Sensitivity + Specificity)/2 \right]$$

Sensitivity was defined as the proportion of sequences in the foreground set containing a motif, and specificity was defined as the proportion of sequences in the background set without the motif [48]. The threshold for motif matching was optimized for each matrix to minimize the error rate. To identify hESC specific TFs, hESC specific enhancers were used as foreground sequences while differentiation specific enhancers were used as background sequences. Correspondingly, to identify differentiation specific TFs, the foreground and the background data sets were flipped.

The significance of the balanced misclassification error rate for a motif (p-value) for a given comparison was determined by the distribution of the error rate. This distribution was estimated by a permutation method (1). To further verify a motif's ability to classify foreground sequences from background sequences, a 95% confidence interval (95% CI) (2) of the difference between the proportion of the sequences with the motif in the foreground set and the proportion of the sequences with the motif in the background set was calculated for each of the 96 TFs. If zero is not in the 95% CI, the difference between the two sets is significant at the 5% level. Otherwise, it is not significant. The results were filtered to include motifs with p-value <0.05, specificity >2/3 and zero being outside the 95% CI. To prove the abilities of this algorithm to identify the difference between hESC specific enhancers and differentiation specific enhancers, two random genomic data sets with 1028 and 3221 sequences were compared with each other. The difference between these two data sets was much less significant than the one between hESC and differentiation enhancers, indicating the great power of the algorithm to distinguish two data sets.

Acknowledgements

Chapter 4, in full, has been submitted for review in Cell. Hawkins, R David ; Hon, Gary C ; Yang, Chuhu ; Antosiewicz-Bourget, Jessica E ; Lee, Leonard K ; Ngo, Que-Minh ; Ching, Keith A ; Edsall, Lee E ; Ye, Zhen ; Kuan, Samantha ; Yu, Pengzhi ; Liu, Hui ; Zhang, Xinming ; Green, Roland D ; Lobanenkov, Victor V ; Stewart, Ron ; Thomson, James A ; and Ren, Bing. "Chromatin States in Human ES Cells Reveal Key Regulatory Sequences and Genes Involved in Pluripotency and Self-renewal". The dissertation author was a primary investigator and author of this paper. Specifically, the dissertation author performed the computational analysis of histone modifications including work on the cell-type specificity of different function elements, predicting enhancers genome-wide, and analyzing the influence of enhancers on gene expression.

References

1. Thomson JA, Itskovitz-Eldor J, Shapiro SS, Waknitz MA, Swiergiel JJ, Marshall VS, Jones JM (1998) Embryonic stem cell lines derived from human blastocysts. Science 282: 1145-1147.

2. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG, Gifford DK, Melton DA, Jaenisch R, Young RA (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122: 947-956.

3. Marson A, Levine SS, Cole MF, Frampton GM, Brambrink T, Johnstone S, Guenther MG, Johnston WK, Wernig M, Newman J, Calabrese JM, Dennis LM, Volkert TL, Gupta S, Love J, Hannett N, Sharp PA, Bartel DP, Jaenisch R, Young RA (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell 134: 521-533.

4. Chi AS, Bernstein BE (2009) Developmental biology. Pluripotent chromatin state. Science 323: 220-221.

5. Azuara V, Perry P, Sauer S, Spivakov M, Jorgensen HF, John RM, Gouti M, Casanova M, Warnes G, Merkenschlager M, Fisher AG (2006) Chromatin signatures of pluripotent cell lines. Nat Cell Biol 8: 532-538.

6. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES (2006) A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell 125: 12.

7. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553-560.

8. Pan G, Tian S, Nie J, Yang C, Ruotti V, Wei H, Jonsdottir GA, Stewart R, Thomson JA (2007) Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. Cell Stem Cell 1: 299-312.

9. Roh TY, Cuddapah S, Cui K, Zhao K (2006) The genomic landscape of histone modifications in human T cells. Proc Natl Acad Sci U S A 103: 15782-15787.

10. Roh TY, Wei G, Farrell CM, Zhao K (2007) Genome-wide prediction of conserved and nonconserved enhancers by histone acetylation patterns. Genome Res 17: 74-81.

11. Kim J, Chu J, Shen X, Wang J, Orkin SH (2008) An extended transcriptional network for pluripotency of embryonic stem cells. Cell 132: 1049-1061.

12. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh YH, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung WK, Clarke ND, Wei CL, Ng HH (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133: 1106-1117.

13. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching KA, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenkov VV, Stewart R, Thomson JA, Crawford GE, Kellis M, Ren B (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature.

14. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio LA (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457: 854-858.

15. Ludwig TE, Bergendahl V, Levenstein ME, Yu J, Probasco MD, Thomson JA (2006) Feeder-independent culture of human embryonic stem cells. Nat Methods 3: 637-646.

16. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA (2000) Genome-wide location and function of DNA binding proteins. Science 290: 2306-2309.

17. Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, Kumar RM, Chevalier B, Johnstone SE, Cole MF, Isono K, Koseki H, Fuchikami T, Abe K, Murray HL, Zucker JP, Yuan B, Bell GW, Herbolsheimer E, Hannett NM, Sun K, Odom DT, Otte AP, Volkert TL, Bartel DP, Melton DA, Gifford DK, Jaenisch R, Young RA (2006) Control of developmental regulators by Polycomb in human embryonic stem cells. Cell 125: 301-313.

18. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA (2007) A chromatin landmark and transcription initiation at most promoters in human cells. Cell 130: 77-88.

19. Zhao XD, Han X, Chew JL, Liu J, Chiu KP, Choo A, Orlov YL, Sung WK, Shahab A, Kuznetsov VA, Bourque G, Oh S, Ruan Y, Ng HH, Wei CL (2007) Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. Cell Stem Cell 1: 286-298.

20. Squazzo SL, O'Geen H, Komashko VM, Krig SR, Jin VX, Jang SW, Margueron R, Reinberg D, Green R, Farnham PJ (2006) Suz12 binds to silenced regions of the genome in a cell-type-specific manner. Genome Res 16: 890-900.

21. Agalioti T, Lomvardas S, Parekh B, Yie J, Maniatis T, Thanos D (2000) Ordered recruitment of chromatin modifying and general transcription factors to the IFN-beta promoter. Cell 103: 667-678.

22. Lomvardas S, Thanos D (2002) Modifying gene expression programs by altering core promoter chromatin architecture. Cell 110: 261-271.

23. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39: 311-318.

24. Park IH, Zhao R, West JA, Yabuuchi A, Huo H, Ince TA, Lerou PH, Lensch MW, Daley GQ (2008) Reprogramming of human somatic cells to pluripotency with defined factors. Nature 451: 141-146.

25. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S (2007) Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 131: 861-872.

26. Pan G, Thomson JA (2007) Nanog and transcriptional networks in embryonic stem cell pluripotency. Cell Res 17: 42-49.

27. Beddington RS, Rashbass P, Wilson V (1992) Brachyury--a gene affecting mouse gastrulation and early organogenesis. Dev Suppl: 157-165.

28. Braun T, Bober E, Winter B, Rosenthal N, Arnold HH (1990) Myf-6, a new member of the human gene family of myogenic determination factors: evidence for a gene cluster on chromosome 12. EMBO J 9: 821-831.

29. Buckingham M, Meilhac S, Zaffran S (2005) Building the mammalian heart from two sources of myocardial cells. Nat Rev Genet 6: 826-835.

30. Chua HL, Bhat-Nakshatri P, Clare SE, Morimiya A, Badve S, Nakshatri H (2007) NF-kappaB represses E-cadherin expression and enhances epithelial to mesenchymal transition of mammary epithelial cells: potential involvement of ZEB-1 and ZEB-2. Oncogene 26: 711-724.

31. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, Liu XS (2008) Model-based Analysis of ChIP-Seq (MACS). Genome Biol 9: R137.

32. Hatzis P, Talianidis I (2002) Dynamics of enhancer-promoter communication during differentiation-induced gene activation. Mol Cell 10: 1467-1477.

33. Palmer MB, Majumder P, Green MR, Wade PA, Boss JM (2007) A 3' enhancer controls snail expression in melanoma cells. Cancer Res 67: 6113-6120.

34. Schubeler D, Francastel C, Cimbora DM, Reik A, Martin DI, Groudine M (2000) Nuclear localization and histone acetylation: a pathway for chromatin opening and transcriptional activation of the human beta-globin locus. Genes Dev 14: 940-950.

35. Shang Y, Myers M, Brown M (2002) Formation of the androgen receptor transcription complex. Mol Cell 9: 601-610.

36. Zhao B, Ricciardi RP (2006) E1A is the component of the MHC class I enhancer complex that mediates HDAC chromatin repression in adenovirus-12 tumorigenic cells. Virology 352: 338-344.

37. Bell AC, West AG, Felsenfeld G (1999) The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell 98: 387-396.

38. Yusufzai TM, Tagami H, Nakatani Y, Felsenfeld G (2004) CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. Mol Cell 13: 291-298.

39. Gaszner M, Felsenfeld G (2006) Insulators: exploiting transcriptional and epigenetic mechanisms. Nat Rev Genet 7: 703-713.

40. Valenzuela L, Kamakaka RT (2006) Chromatin insulators. Annu Rev Genet 40: 107-138.

41. Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenkov VV, Ren B (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell 128: 1231-1245.

42. Cuddapah S, Jothi R, Schones DE, Roh TY, Cui K, Zhao K (2009) Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Res 19: 24-32.

43. Tomioka M, Nishimoto M, Miyagi S, Katayanagi T, Fukui N, Niwa H, Muramatsu M, Okuda A (2002) Identification of Sox-2 regulatory region which is under the control of Oct-3/4-Sox-2 complex. Nucleic Acids Res 30: 3202-3213.

44. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome Res 12: 996-1006.

45. Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Pruss M, Reuter I, Schacherer F (2000) TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res 28: 316-319.

46. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res 32: D91-94.

47. Thijs G, Moreau Y, De Smet F, Mathys J, Lescot M, Rombauts S, Rouze P, De Moor B, Marchal K (2002) INCLUSive: integrated clustering, upstream sequence retrieval and motif sampling. Bioinformatics 18: 331-332.

48. Barrera LO, Li Z, Smith AD, Arden KC, Cavenee WK, Zhang MQ, Green RD, Ren B (2008) Genome-wide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs. Genome Res 18: 46-59.

49. Xi H, Shulha HP, Lin JM, Vales TR, Fu Y, Bodine DM, McKay RD, Chenoweth JG, Tesar PJ, Furey TS, Ren B, Weng Z, Crawford GE (2007) Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. PLoS Genet 3: e136.

50. Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E, Tsutsumi S, Nagae G, Ishihara K, Mishiro T, Yahata K, Imamoto F, Aburatani H, Nakao M, Imamoto N, Maeshima K, Shirahige K, Peters JM (2008) Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature 451: 796-801.

51. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K (2007) High-resolution profiling of histone methylations in the human genome. Cell 129: 823-837.

Figures and Tables

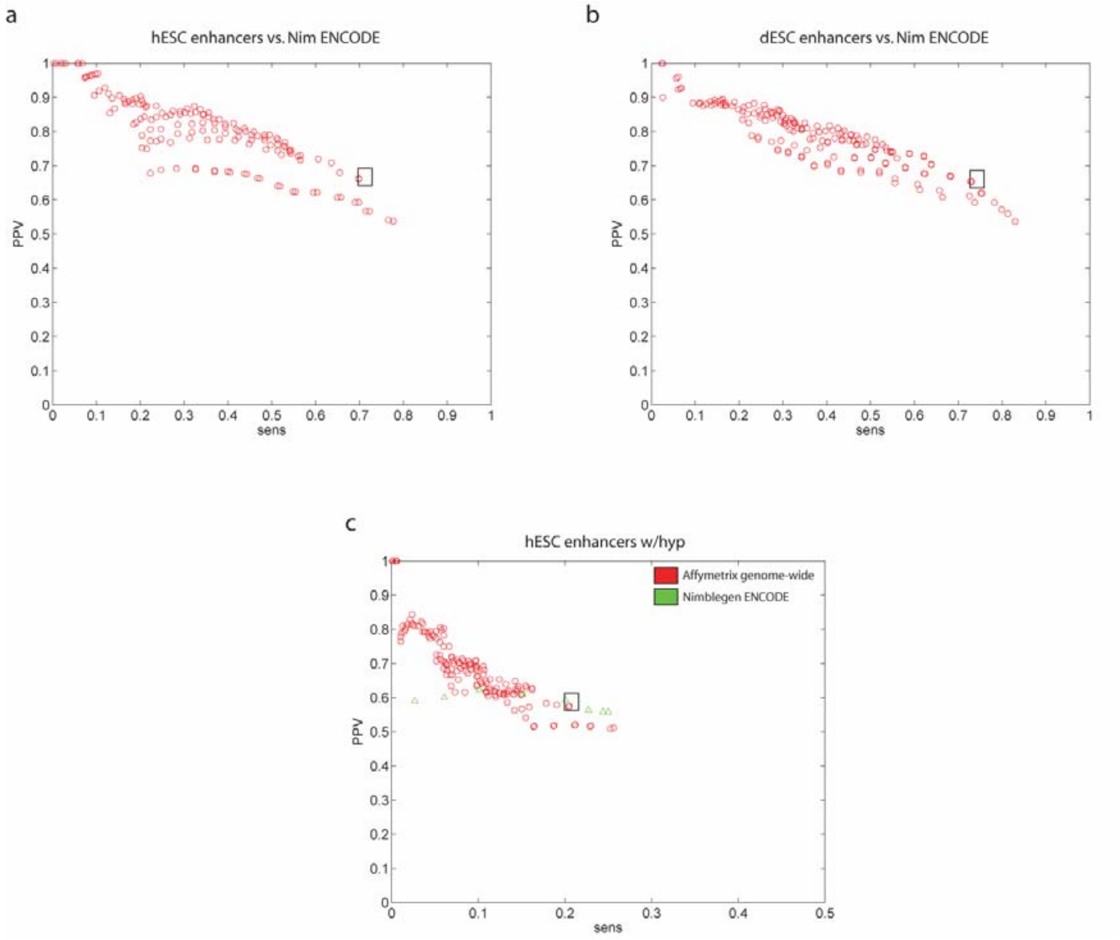


Figure 4-1: Validation of enhancers and platform comparisons.

(a) Comparison of enhancer predictions from Affymetrix genome-wide arrays to Nimblegen ENCODE arrays using the enhancer predictions from Nimblegen ENCODE arrays as a gold standard. (b) As in (a), but for differentiated ES cells. (c) As in (a) but using ENCODE hESC DNase I hypersensitivity data [49] as a gold standard. Each symbol represents a different set of parameters. The black box indicates the set of enhancers used here.

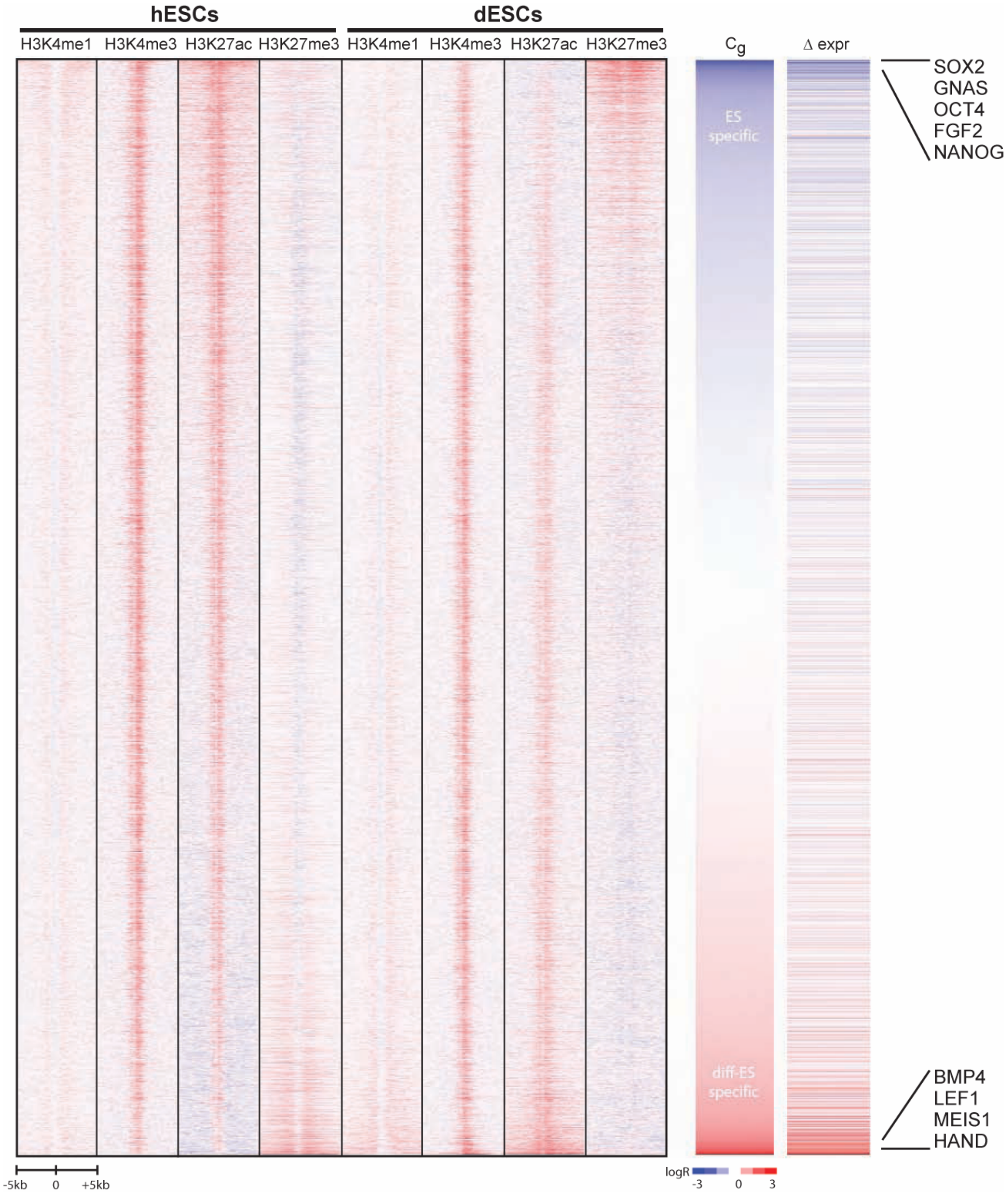


Figure 4-2: Dynamic switch of H3K27 modifications at promoters

(left) Heat-map of histone modifications within 5-kb of 22,047 TSSs, before and after differentiation. (middle) For each gene and cell type, I calculate the difference (H3K27ac – H3K27me3), and rank genes by comparing the difference of this value between the cell types (dESC – hESC). A negative value represents hESC enrichment of H3K27ac and dESC enrichment of H3K27me3 (blue C_g). A positive value represents dESC enrichment of H3K27ac and hESC enrichment of H3K27me3 (red C_g). (right) Difference in gene expression (dESC/hESC); blue is hESC-specific expression while red is dESC-specific expression. Representative genes are noted on the far right.

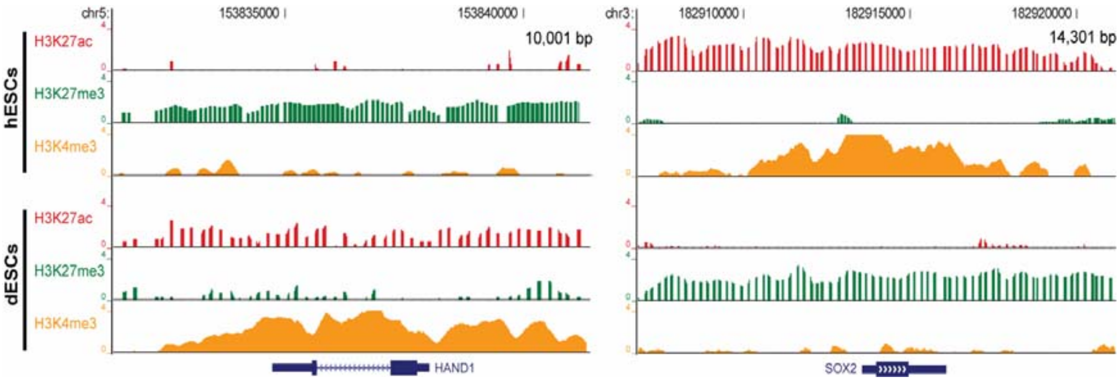


Figure 4-3: Snapshots of histone modifications around HAND1 and SOX2

UCSC Genome Browser snapshots showing the log2 ratio enrichment for H3K27ac (red), H3K27me3 (green) and H3K4me3 (orange) compared to input. Gene names are listed at the 5' end of the gene structure. (left) A 10 kb window around the HAND1 gene illustrating the presence of H3K27me3 in hESCs that switches to H3K27ac following differentiation. (right) A 14 kb window around the SOX2 gene illustrating the presence of H3K27ac in hESCs that switches to H3K27me3 following differentiation.

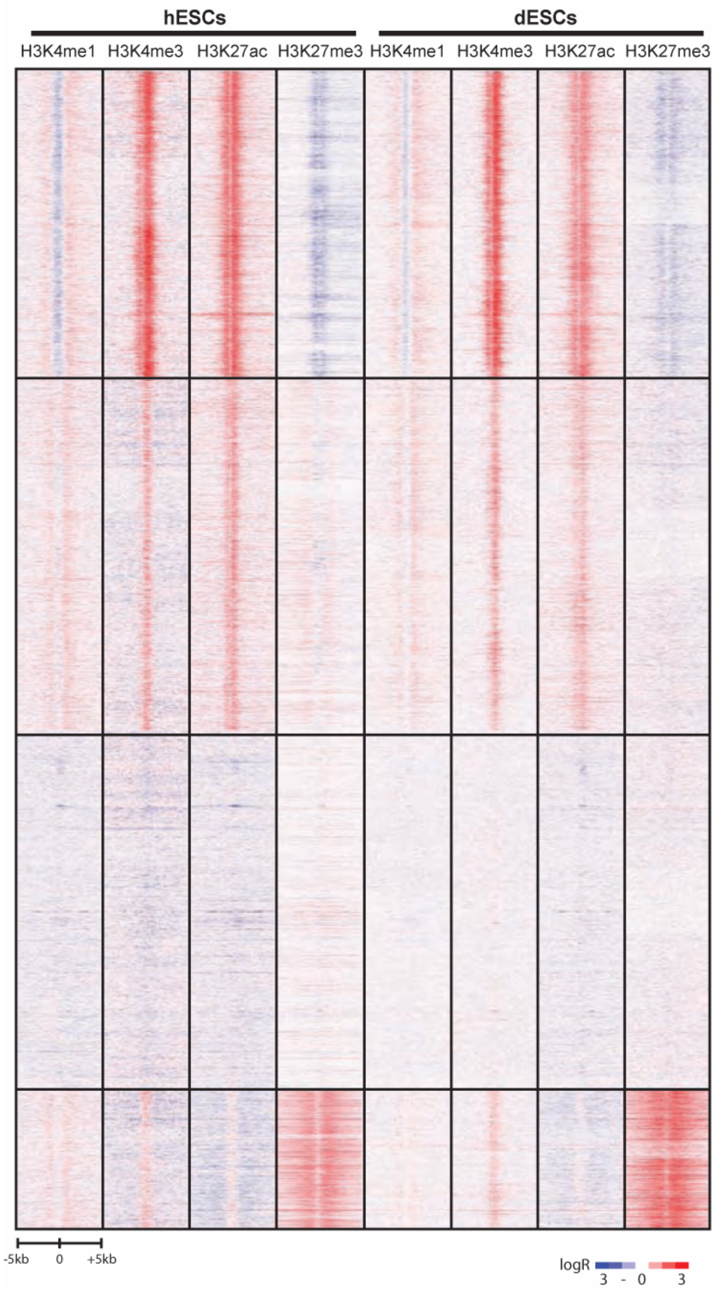


Figure 4-4: k-means cluster of modifications at 22,047 gene TSS with expression data.

The figure is organized as in Figure 4-2. k-means = 4.

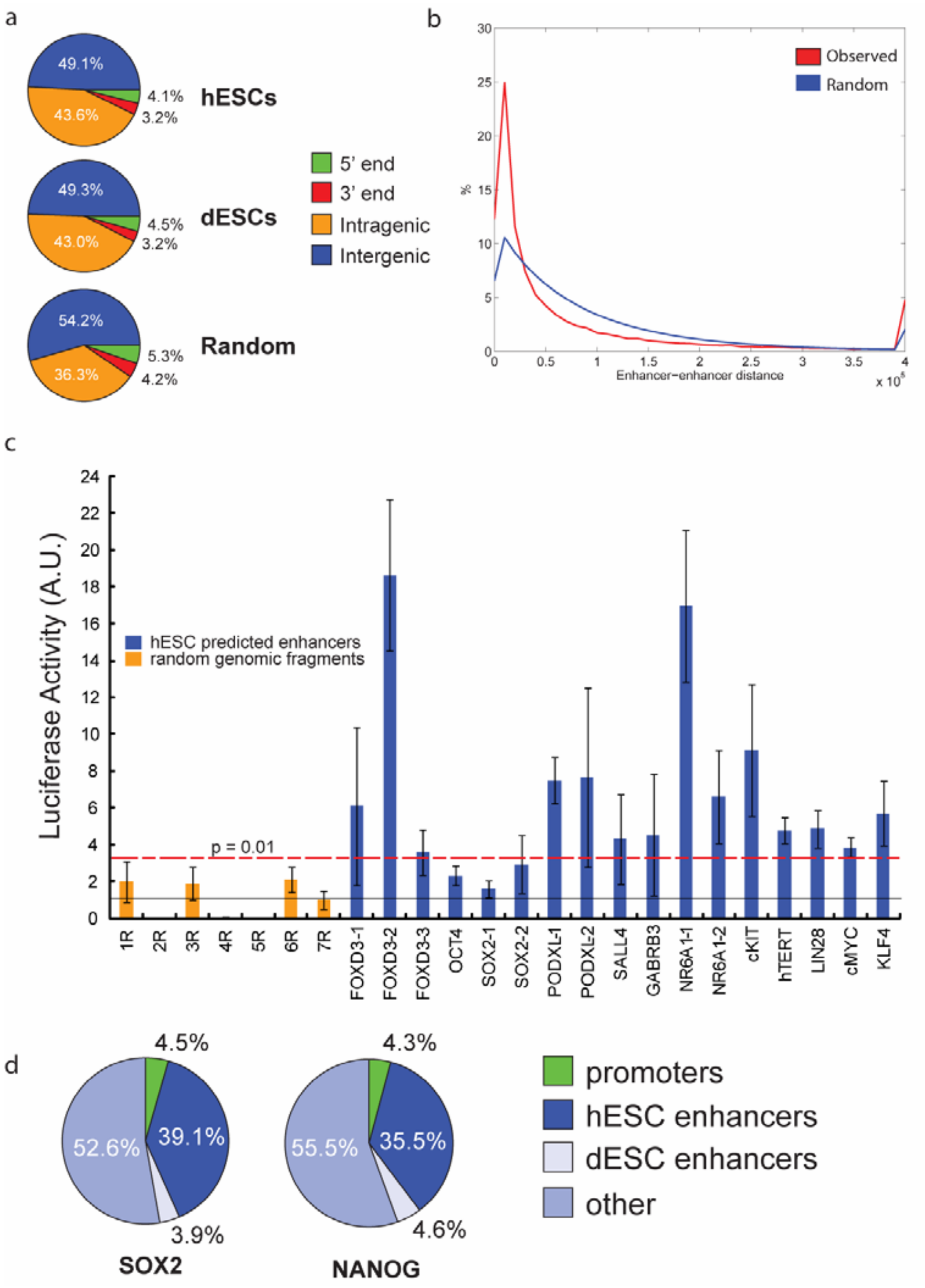


Figure 4-5: Enhancer features and functional validation

(a) Distribution of enhancers in each cell type relative to 5' and 3' ends of genes as well as intragenic and intergenic regions.(b) Distribution of distances between adjacent enhancers. (c) Reporter assays of enhancer function at 17 predicted hESC enhancers and 7 randomly chosen genomic regions, cloned downstream of a luciferase gene. The dashed red line indicates a p-value cutoff of 1%.(d) Overlap of ChIP-Seq binding sites for transcription factors Sox2 and NANOG, compared to promoters, predicted hESC enhancers, and predicted dESC enhancers.

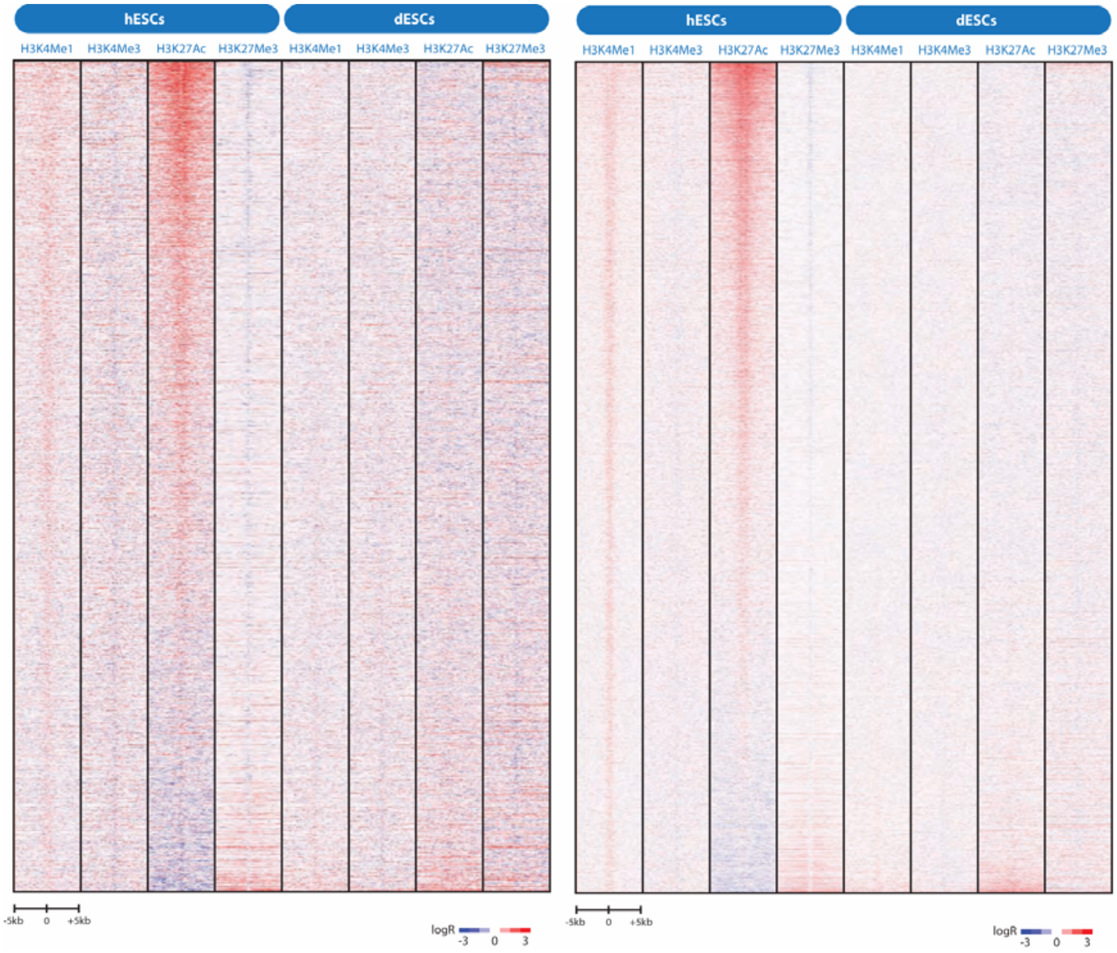


Figure 4-6: Clustering of histone modifications at distal SOX2 and NANOG binding sites not predicted as enhancers.

Binding sites are ranked by acetylation levels. The clusters illustrate that most sites present some H3K4me1 and H3K27ac, suggesting they are enhancer sites not called by the prediction algorithm.

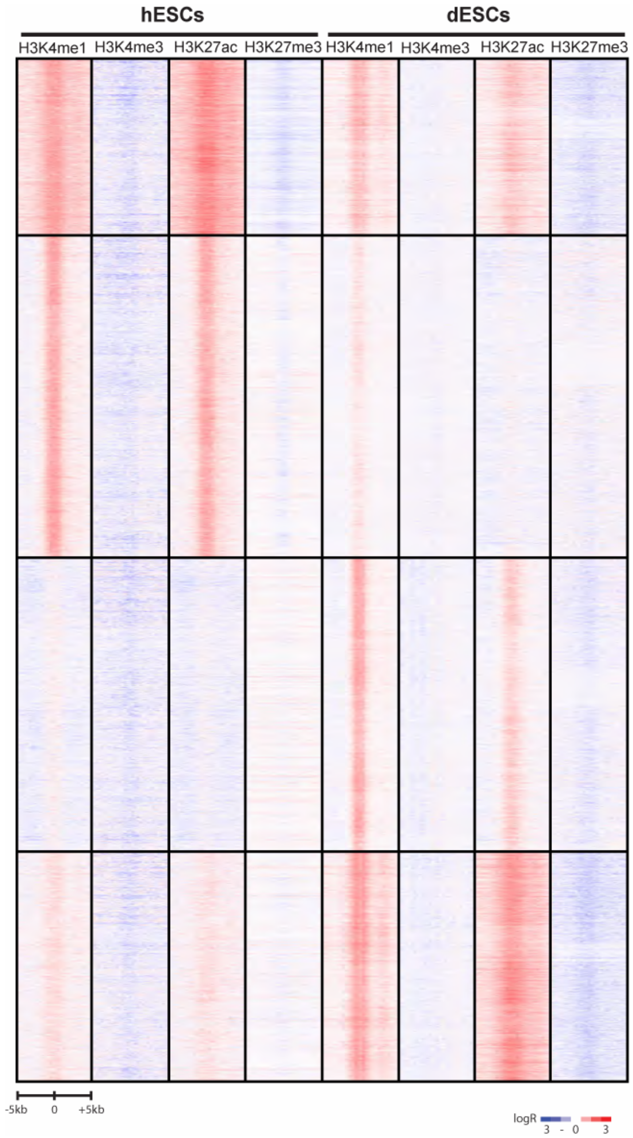


Figure 4-7: k-means cluster of predicted enhancers in hESCs and dESCs.

The figure is organized as in Figure 1.2. k-means = 4.

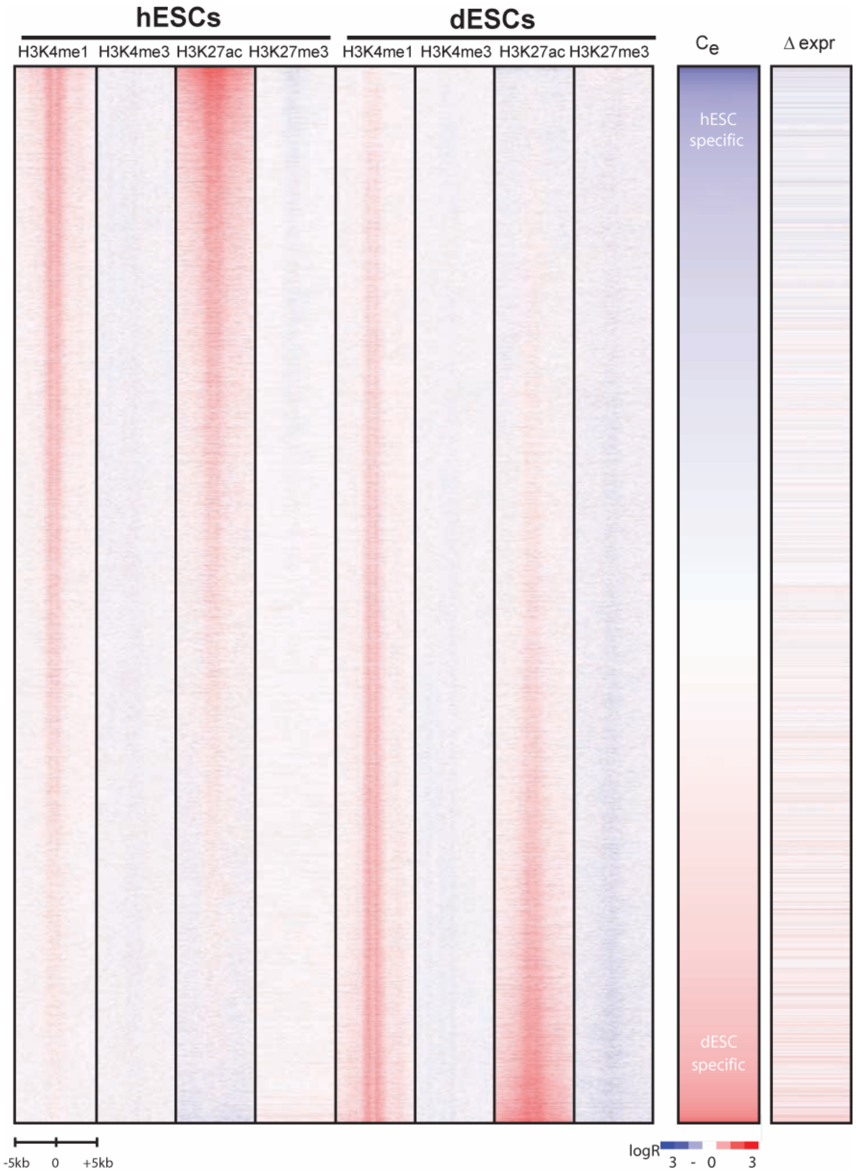


Figure 4-8: Dynamic switch of H3K27ac at enhancers

(left) A heat-map of histone modifications within 5-kb of predicted enhancers, ranked based on differences in H3K27ac (dESC – hESC). (middle) The cell-type specificity of chromatin modifications at enhancers, C_e = (H3K27ac$_{dESC}$ – H3K27ac$_{hESC}$). (right) Changes in gene expression of neighboring genes.

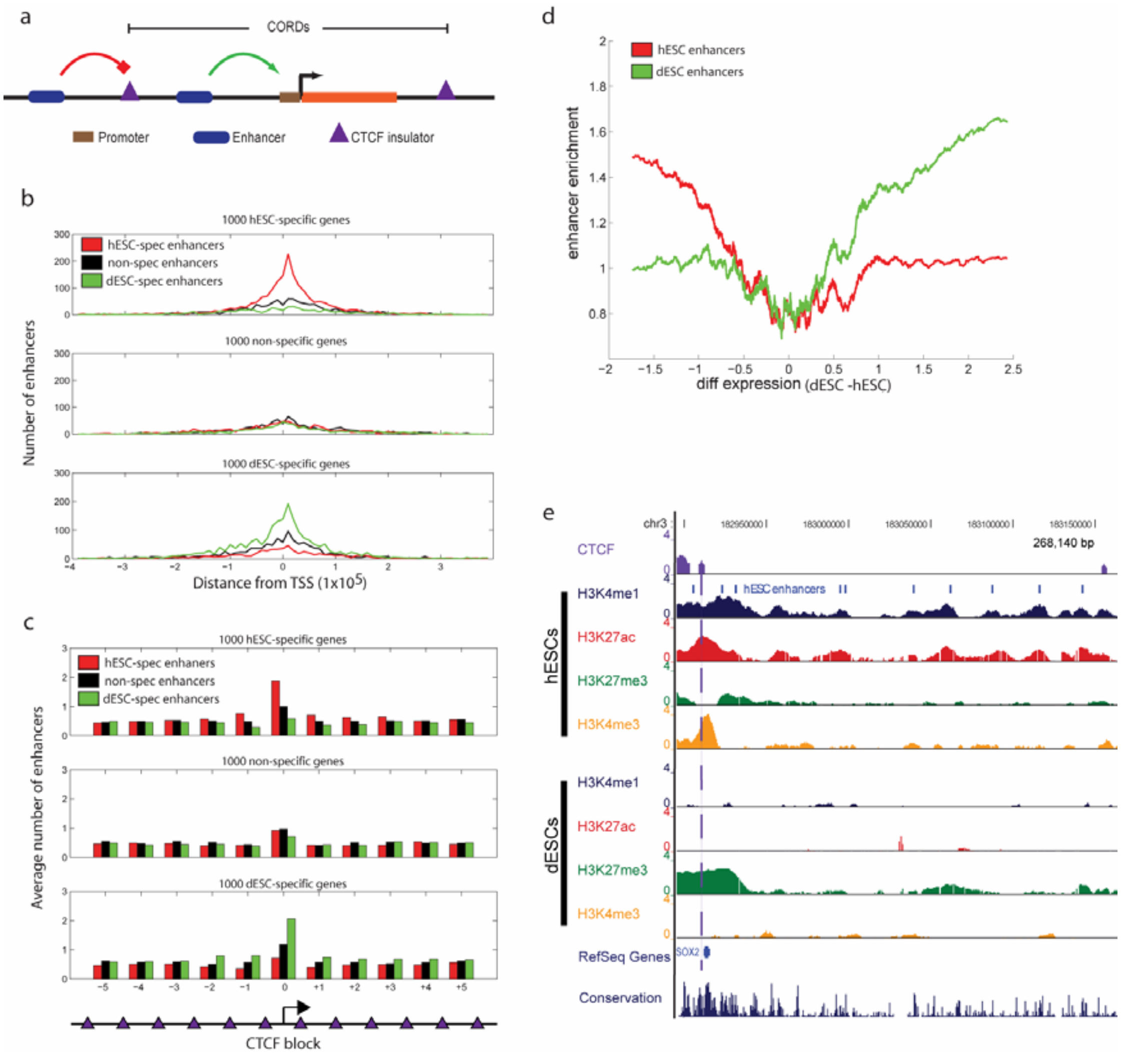


Figure 4-9: Enhancer cell-type specificity and predicted gene targets

(a) CORDs – Diagram of CTCF-organized regulatory domains. Regions bounded by CTCF containing promoters and enhancers. (b) Distribution of hESC-specific, dESC-specific, and non-specific enhancers within CTCF-defined domains containing promoters of hESC-specific, dESC-specific, and non-specific genes. (c) As in (b), but expanded to neighboring CTCF-defined domains. (d) The average number of enhancers for each cell type were counted as a function of average differential gene expression by using a sliding window of 1000 genes for all 22,047 genes from Figure 4-2, and then normalized over 100 random distributions of enhancers to obtain enhancer enrichment. (e) UCSC Genome Browser snapshots around a key ESC gene SOX2 showing the localization of hESC-specific enhancers marked by H3K4me1 (blue) and H3K27ac (red) within CTCF-defined domains (purple). The purple, dashed vertical line indicates the position of CTCF sites close to the genic region. Gene names are located at the 5’ end of the gene structure.

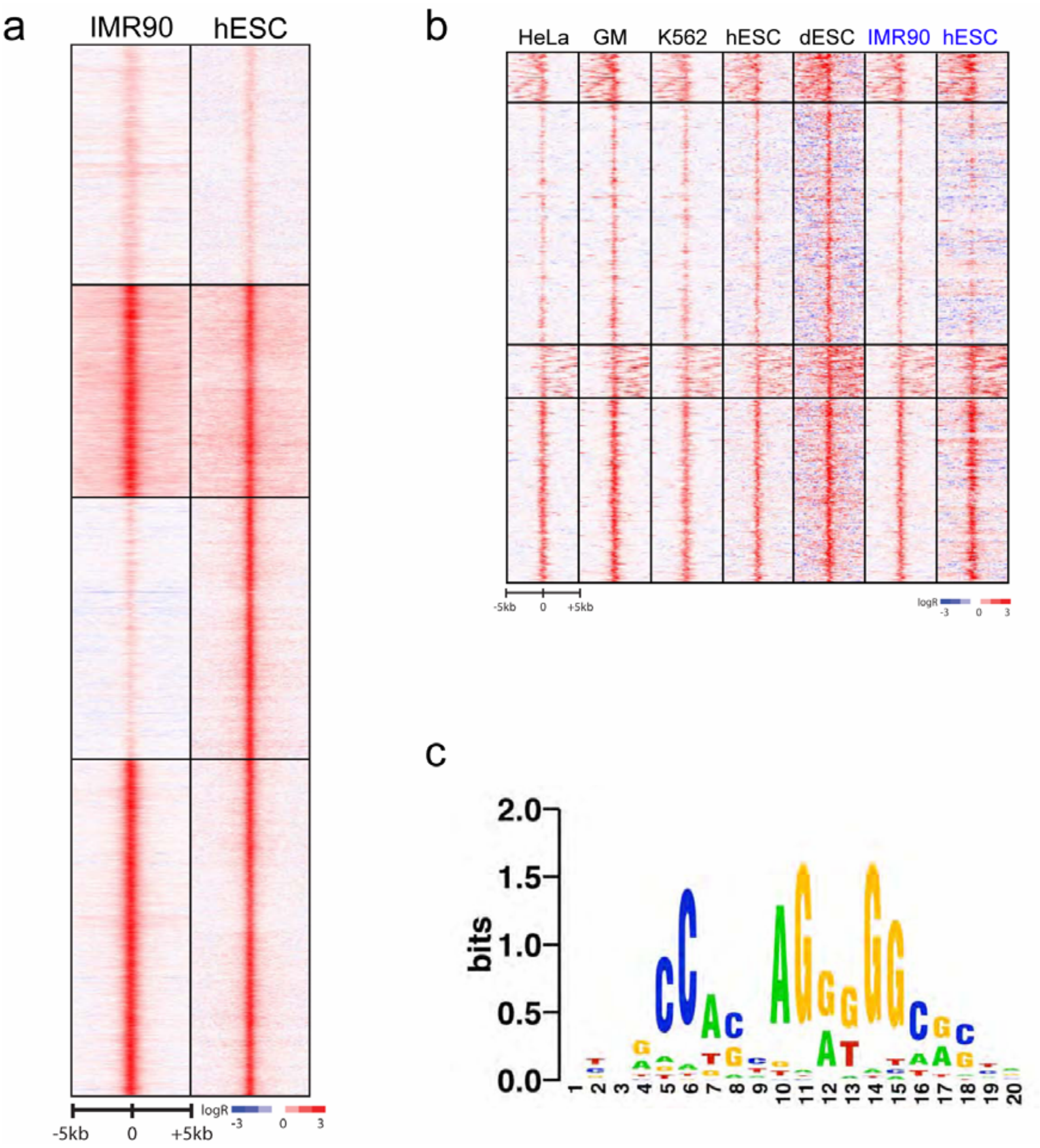


Figure 4-10: CTCF binding site analysis

(a) Clustergram at 29,880 combined CTCF binding sites recovered from IMR90 [41], HeLa [50], and CD4+ T cells [51], as represented in genome-wide CTCF binding sites found in IMR90 and hES cells. (b) k-means clustergram of CTCF binding sites in the ENCODE region from HeLa, GM06990 (GM), K562 leukemic cells, hESCs, and dESCs from Chapter 3. These are compared to ENCODE regions extracted from genome-wide data in IMR90 cells [41] and hESCs presented here (blue). (c) The increased number of CTCF binding sites provides a slightly modified motif.

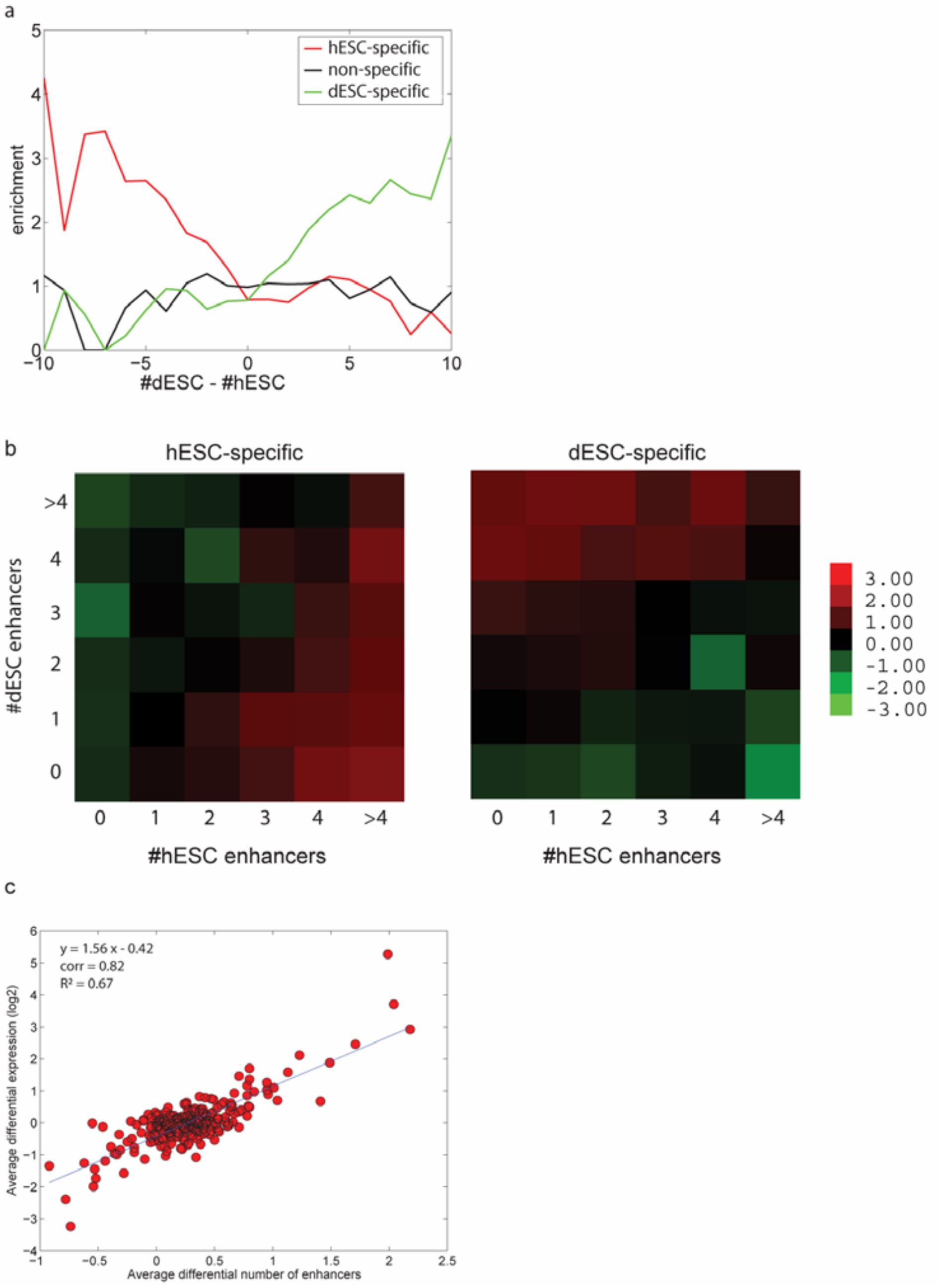


Figure 4-11: Changes in chromatin modification at enhancers correspond to changes in gene expression

(a) For each of the 1000 hESC-specific (red), dESC-specific (green), and non-specific (black) genes, I counted the number of enhancers found before and after differentiation within the CTCF-defined domain. I then plotted the distribution of this difference normalized over the distribution of all genes. (b) As in (a), except directly comparing enhancer numbers within CTCF-defined domains for hESC-specific genes (right) and dESC-specific genes (left). The enrichment ratio, as above, is shown as a heat-map (red = enrichment; green lack of enrichment). (c) Plot of differential enhancer number as a function of differential expression for all 22,047 genes, averaged into 100 gene bins.

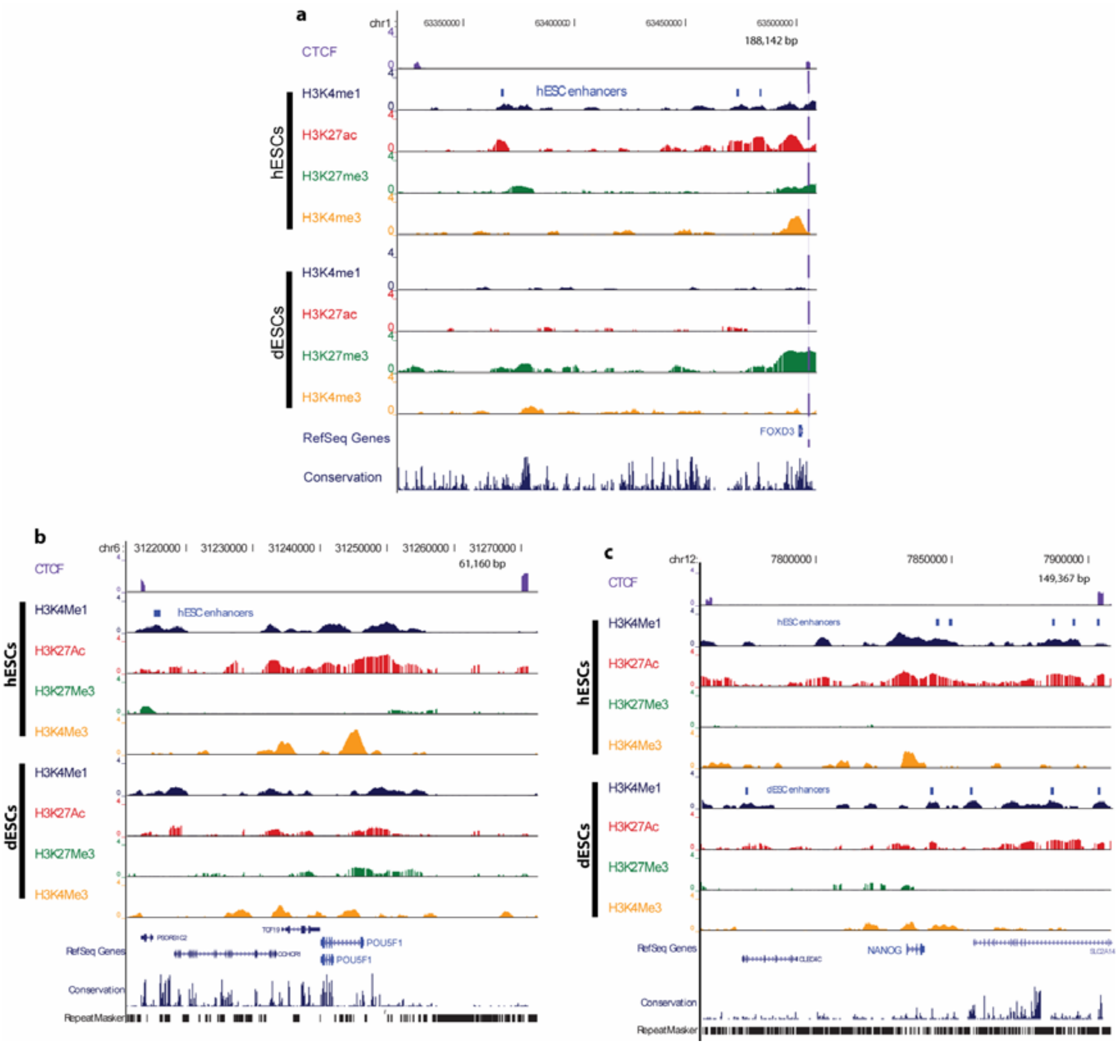


Figure 4-12: UCSC Genome browser shots of histone modifications and enhancer predictions in CTCF-defined blocks containing FOXD3, OCT4(POU5F1) and NANOG.

Figure is displayed as in Figure 4-3.

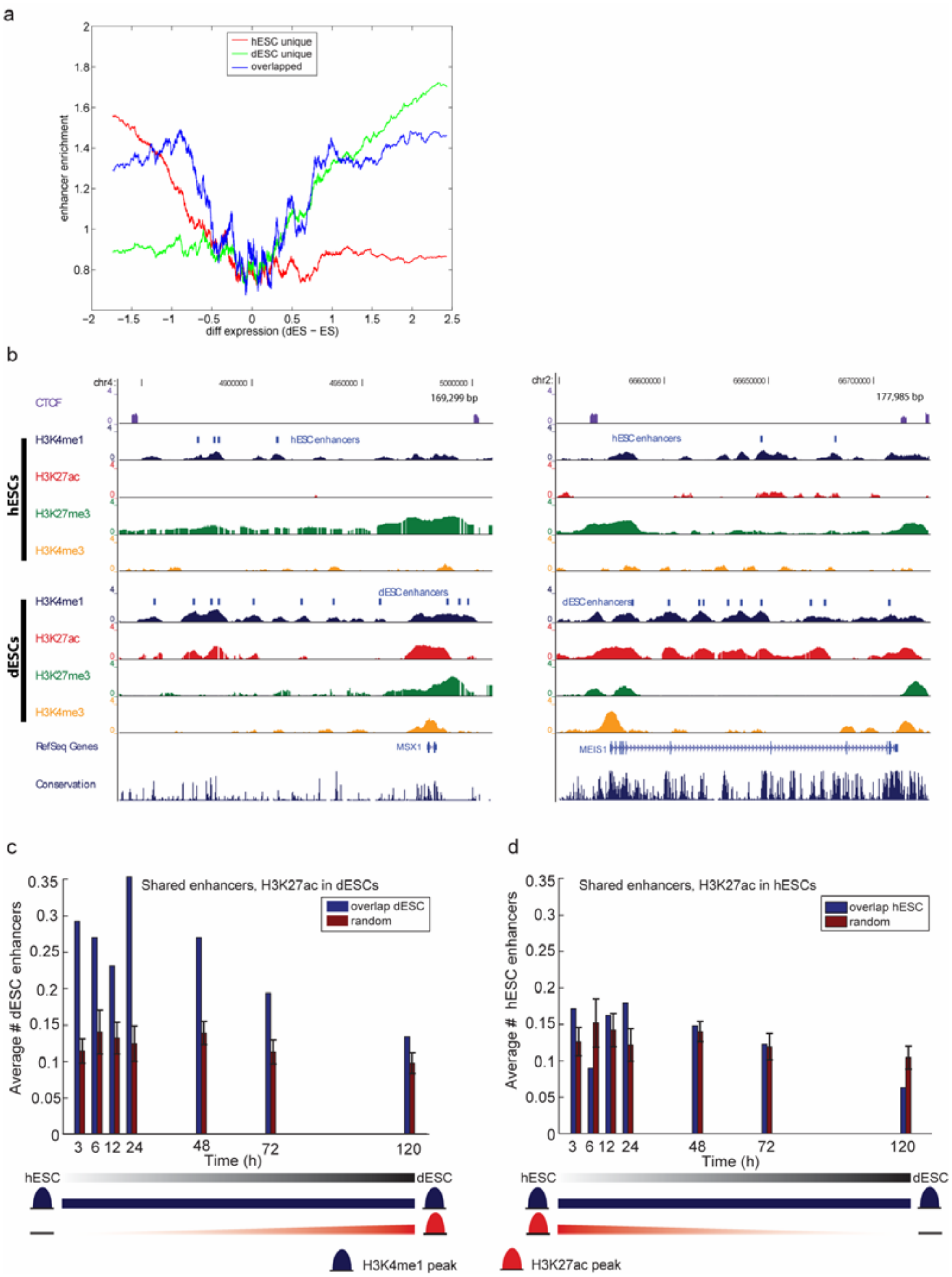


Figure 4-13: Subset of shared enhancers are poised for early response

(a) As in Figure 4-9d, but for three subsets of enhancers: those uniquely marked in hESCs (red), those uniquely marked in dESCs (green), and the remaining 8863 that are marked in both (blue). (b) UCSC Genome Browser snapshots of MSX1 and MEIS1 gene loci. (c, d) Gene expression was measured at 3, 6, 12, 24, 48, 72, and 120 hours after BMP4/bFGF treatment of hESCs. For differentially expressed genes at each time point, I counted the average number of acetylated enhancers with cell type specificity, defined as the 2000 shared enhancers with the most H3K27ac in (c) dESCs and (d) hESCs.

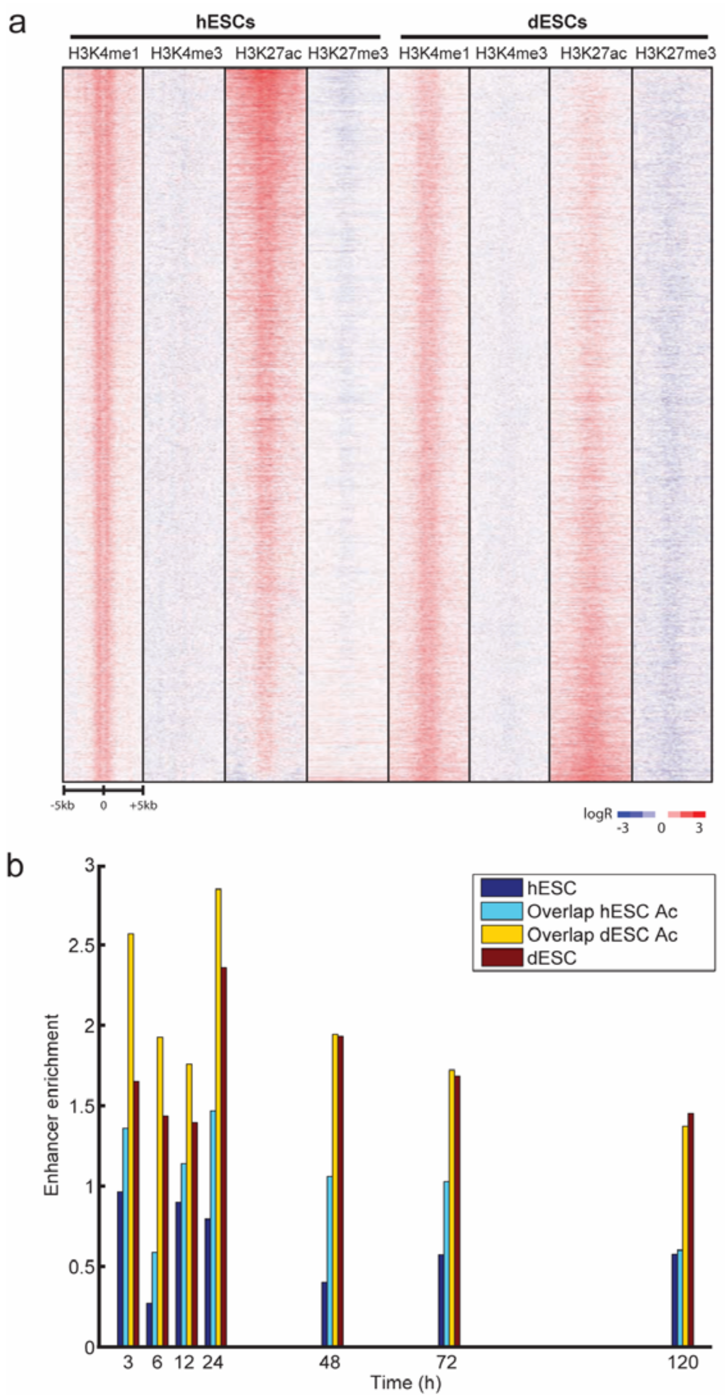


Figure 4-14: Poised Enhancers

(a)Heatmap of 8863 predicted enhancers that are shared between hES and dES cells, ranked on H3K27ac intensity. Each end of the spectrum shows that some enhancers exhibit cell type-specific acetylation, although mono-methylated in both cell types. (b) Assessment of enhancer enrichment during a time course of gene expression during BMP4 treatment. Early response genes are more enriched in shared enhancers acetylated in dESCs (yellow) compared to dESC-specific (red), hESC-specific (dark blue), or shared enhancers with hESC acetylation (light blue).

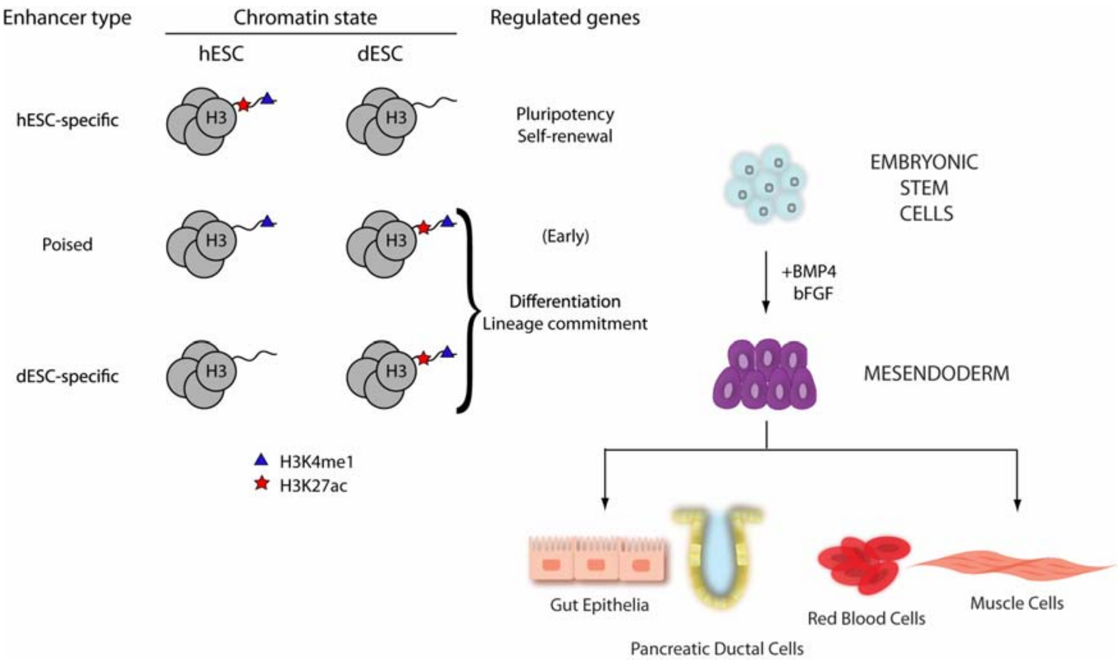
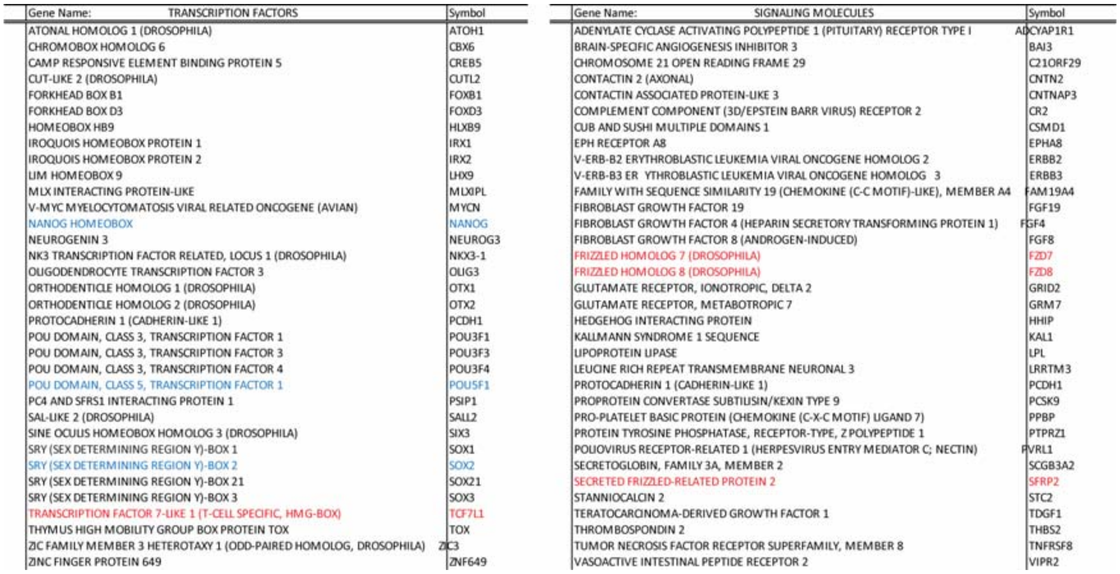


Figure 4-15: Model of cell type-specific enhancers and poised enhancers in cell fate.

This model illustrates the role of poised enhancers in hESC pluripotency and cell fate commitment. ES cells grown in the presence of BMP4 and bFGF give rise to 3 of 4 possible lineages (ectoderm excluded). Poised enhancers contribute to initiation of lineage determination by activating early response genes which go on to establish the cell fate.

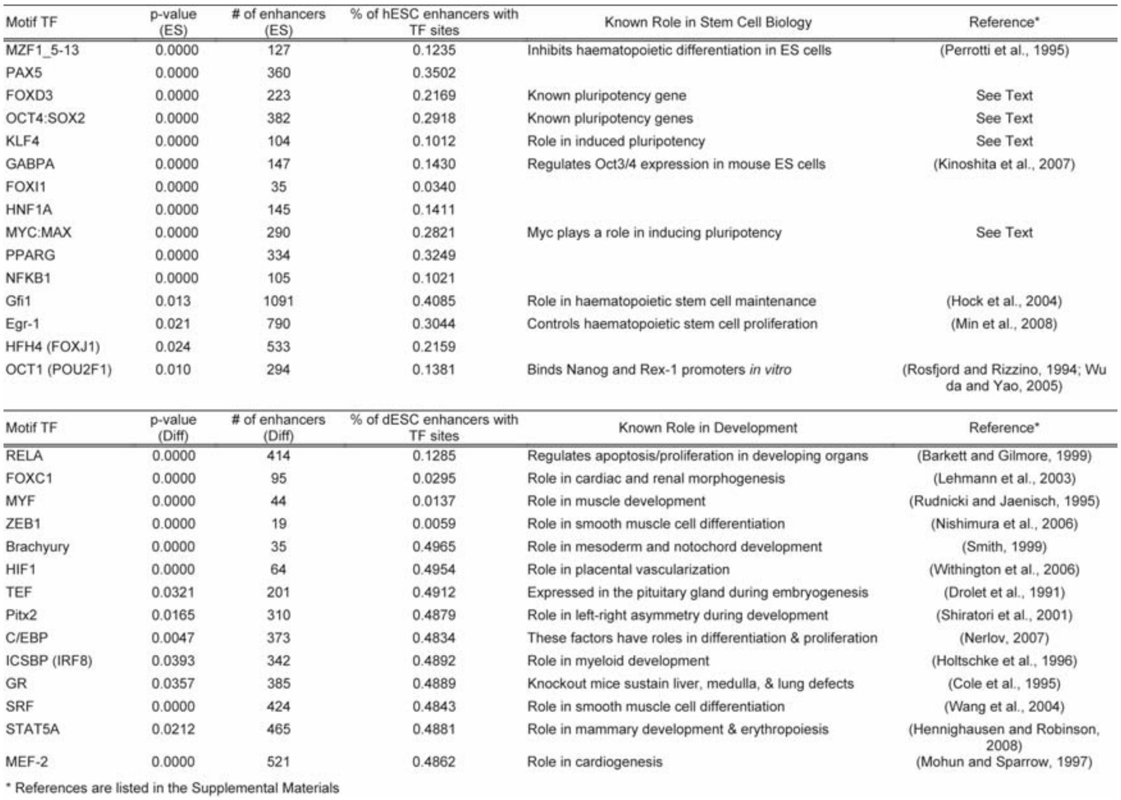
Table 4-1: Representative transcription factors and signaling molecules repressed by H3K27me3 following differentiation.



Gene Name: TRANSCRIPTION FACTORS	Symbol
ATONAL HOMOLOG 1 (DROSOPHILA)	ATOH1
CHROMOBOX HOMOLOG 6	CBX6
CAMP RESPONSIVE ELEMENT BINDING PROTEIN 5	CREB5
CUT-LIKE 2 (DROSOPHILA)	CUTL2
FORKHEAD BOX B1	FOXB1
FORKHEAD BOX D3	FOXD3
HOMEOBOX HB9	HLXB9
IROQUOIS HOMEOBOX PROTEIN 1	IRX1
IROQUOIS HOMEOBOX PROTEIN 2	IRX2
LIM HOMEOBOX 9	LHX9
MLX INTERACTING PROTEIN-LIKE	MLXIPL
V-MYC MYELOCYTOMATOSIS VIRAL RELATED ONCOGENE (AVIAN)	MYCN
NANOG HOMEOBOX	NANOG
NEUROGENIN 3	NEUROG3
NK3 TRANSCRIPTION FACTOR RELATED, LOCUS 1 (DROSOPHILA)	NKX3-1
OLIGODENDROCYTE TRANSCRIPTION FACTOR 3	OLIG3
ORTHODENTICLE HOMOLOG 1 (DROSOPHILA)	OTX1
ORTHODENTICLE HOMOLOG 2 (DROSOPHILA)	OTX2
PROTOCADHERIN 1 (CADHERIN-LIKE 1)	PCDH1
POU DOMAIN, CLASS 3, TRANSCRIPTION FACTOR 1	POU3F1
POU DOMAIN, CLASS 3, TRANSCRIPTION FACTOR 3	POU3F3
POU DOMAIN, CLASS 3, TRANSCRIPTION FACTOR 4	POU3F4
POU DOMAIN, CLASS 5, TRANSCRIPTION FACTOR 1	POU5F1
PC4 AND SFRS1 INTERACTING PROTEIN 1	PSIP1
SAL-LIKE 2 (DROSOPHILA)	SALL2
SINE OCULIS HOMEOBOX HOMOLOG 3 (DROSOPHILA)	SIX3
SRY (SEX DETERMINING REGION Y)-BOX 1	SOX1
SRY (SEX DETERMINING REGION Y)-BOX 2	SOX2
SRY (SEX DETERMINING REGION Y)-BOX 21	SOX21
SRY (SEX DETERMINING REGION Y)-BOX 3	SOX3
TRANSCRIPTION FACTOR 7-LIKE 1 (T-CELL SPECIFIC, HMG-BOX)	TCF7L1
THYMUS HIGH MOBILITY GROUP BOX PROTEIN TOX	TOX
ZIC FAMILY MEMBER 3 HETEROTAXY 1 (ODD-PAIRED HOMOLOG, DROSOPHILA)	ZIC3
ZINC FINGER PROTEIN 649	ZNF649

Gene Name: SIGNALING MOLECULES	Symbol
ADENYLATE CYCLASE ACTIVATING POLYPEPTIDE 1 (PITUITARY) RECEPTOR TYPE I	ADCYAP1R1
BRAIN-SPECIFIC ANGIOGENESIS INHIBITOR 3	BAI3
CHROMOSOME 21 OPEN READING FRAME 29	C21ORF29
CONTACTIN 2 (AXONAL)	CNTN2
CONTACTIN ASSOCIATED PROTEIN-LIKE 3	CNTNAP3
COMPLEMENT COMPONENT (3D/EPSTEIN BARR VIRUS) RECEPTOR 2	CR2
CUB AND SUSHI MULTIPLE DOMAINS 1	CSMD1
EPH RECEPTOR A8	EPHA8
V-ERB-B2 ERYTHROBLASTIC LEUKEMIA VIRAL ONCOGENE HOMOLOG 2	ERBB2
V-ERB-B3 ER YTHROBLASTIC LEUKEMIA VIRAL ONCOGENE HOMOLOG 3	ERBB3
FAMILY WITH SEQUENCE SIMILARITY 19 (CHEMOKINE (C-C MOTIF)-LIKE), MEMBER A4	FAM19A4
FIBROBLAST GROWTH FACTOR 19	FGF19
FIBROBLAST GROWTH FACTOR 4 (HEPARIN SECRETORY TRANSFORMING PROTEIN 1)	FGF4
FIBROBLAST GROWTH FACTOR 8 (ANDROGEN-INDUCED)	FGF8
FRIZZLED HOMOLOG 7 (DROSOPHILA)	FZD7
FRIZZLED HOMOLOG 8 (DROSOPHILA)	FZD8
GLUTAMATE RECEPTOR, IONOTROPIC, DELTA 2	GRID2
GLUTAMATE RECEPTOR, METABOTROPIC 7	GRM7
HEDGEHOG INTERACTING PROTEIN	HHIP
KALLMANN SYNDROME 1 SEQUENCE	KAL1
LIPOPROTEIN LIPASE	LPL
LEUCINE RICH REPEAT TRANSMEMBRANE NEURONAL 3	LRRTM3
PROTOCADHERIN 1 (CADHERIN-LIKE 1)	PCDH1
PROPROTEIN CONVERTASE SUBTILISIN/KEXIN TYPE 9	PCSK9
PRO-PLATELET BASIC PROTEIN (CHEMOKINE (C-X-C MOTIF) LIGAND 7)	PPBP
PROTEIN TYROSINE PHOSPHATASE, RECEPTOR-TYPE, Z POLYPEPTIDE 1	PTPRZ1
POLIOVIRUS RECEPTOR-RELATED 1 (HERPESVIRUS ENTRY MEDIATOR C; NECTIN)	PVRL1
SECRETOGLOBIN, FAMILY 3A, MEMBER 2	SCGB3A2
SECRETED FRIZZLED-RELATED PROTEIN 2	SFRP2
STANNIOCALCIN 2	STC2
TERATOCARCINOMA-DERIVED GROWTH FACTOR 1	TDGF1
THROMBOSPONDIN 2	THBS2
TUMOR NECROSIS FACTOR RECEPTOR SUPERFAMILY, MEMBER 8	TNFRSF8
VASOACTIVE INTESTINAL PEPTIDE RECEPTOR 2	VIPR2

Table 4-2: Transcription factor binding site motifs enriched in hESC or dESC enhancers.



Motif TF	p-value (ES)	# of enhancers (ES)	% of hESC enhancers with TF sites	Known Role in Stem Cell Biology	Reference*
MZF1_5-13	0.0000	127	0.1235	Inhibits haematopoietic differentiation in ES cells	(Perrotti et al., 1995)
PAX5	0.0000	360	0.3502		
FOXD3	0.0000	223	0.2169	Known pluripotency gene	See Text
OCT4:SOX2	0.0000	382	0.2918	Known pluripotency genes	See Text
KLF4	0.0000	104	0.1012	Role in induced pluripotency	See Text
GABPA	0.0000	147	0.1430	Regulates Oct3/4 expression in mouse ES cells	(Kinoshita et al., 2007)
FOXI1	0.0000	35	0.0340		
HNF1A	0.0000	145	0.1411		
MYC:MAX	0.0000	290	0.2821	Myc plays a role in inducing pluripotency	See Text
PPARG	0.0000	334	0.3249		
NFKB1	0.0000	105	0.1021		
Gfi1	0.013	1091	0.4085	Role in haematopoietic stem cell maintenance	(Hock et al., 2004)
Egr-1	0.021	790	0.3044	Controls haematopoietic stem cell proliferation	(Min et al., 2008)
HFH4 (FOXJ1)	0.024	533	0.2159		
OCT1 (POU2F1)	0.010	294	0.1381	Binds Nanog and Rex-1 promoters *in vitro*	(Rosfjord and Rizzino, 1994; Wu da and Yao, 2005)

Motif TF	p-value (Diff)	# of enhancers (Diff)	% of dESC enhancers with TF sites	Known Role in Development	Reference*
RELA	0.0000	414	0.1285	Regulates apoptosis/proliferation in developing organs	(Barkett and Gilmore, 1999)
FOXC1	0.0000	95	0.0295	Role in cardiac and renal morphogenesis	(Lehmann et al., 2003)
MYF	0.0000	44	0.0137	Role in muscle development	(Rudnicki and Jaenisch, 1995)
ZEB1	0.0000	19	0.0059	Role in smooth muscle cell differentiation	(Nishimura et al., 2006)
Brachyury	0.0000	35	0.4965	Role in mesoderm and notochord development	(Smith, 1999)
HIF1	0.0000	64	0.4954	Role in placental vascularization	(Withington et al., 2006)
TEF	0.0321	201	0.4912	Expressed in the pituitary gland during embryogenesis	(Drolet et al., 1991)
Pitx2	0.0165	310	0.4879	Role in left-right asymmetry during development	(Shiratori et al., 2001)
C/EBP	0.0047	373	0.4834	These factors have roles in differentiation & proliferation	(Nerlov, 2007)
ICSBP (IRF8)	0.0393	342	0.4892	Role in myeloid development	(Holtschke et al., 1996)
GR	0.0357	385	0.4889	Knockout mice sustain liver, medulla, & lung defects	(Cole et al., 1995)
SRF	0.0000	424	0.4843	Role in smooth muscle cell differentiation	(Wang et al., 2004)
STAT5A	0.0212	465	0.4881	Role in mammary development & erythropoiesis	(Hennighausen and Robinson, 2008)
MEF-2	0.0000	521	0.4862	Role in cardiogenesis	(Mohun and Sparrow, 1997)

* References are listed in the Supplemental Materials

Chapter 5 : ChromaSig – A probabilistic approach to finding common chromatin signatures in the human genome

Abstract

Computational methods to identify functional genomic elements using genetic information have been very successful in determining gene structure and in identifying a handful of *cis*-regulatory elements. But the vast majority of regulatory elements have yet to be discovered, and it has become increasingly apparent that their discovery will not come from using genetic information alone. Recently, high-throughput technologies have enabled the creation of information-rich epigenetic maps, most notably for histone modifications. However, tools that search for functional elements using this epigenetic information have been lacking. Here, I describe an unsupervised learning method called ChromaSig to find, in an unbiased fashion, commonly occurring chromatin signatures in both tiling microarray and sequencing data. Applying this algorithm to nine chromatin marks across a 1% sampling of the human genome in HeLa cells, I recover eight clusters of distinct chromatin signatures, five of which correspond to known patterns associated with transcriptional promoters and enhancers. Interestingly, I observe that the distinct chromatin signatures found at enhancers mark distinct functional classes of enhancers in terms of transcription factor and co-activator binding. In addition, I identify three clusters of novel chromatin signatures, which contain evolutionarily conserved sequences and potential *cis*-regulatory elements. Applying ChromaSig to a panel of 21 chromatin marks mapped genome-wide by ChIP-Seq reveals 16 classes of genomic elements marked by distinct chromatin signatures. Interestingly, four classes containing enrichment for repressive histone modifications appear to be locally heterochromatic sites and are enriched in quickly-evolving regions of the genome. The utility of this approach in uncovering novel, functionally significant genomic elements will aid future efforts of genome annotation via chromatin modifications.

Introduction

In eukaryotes, DNA is packaged into nucleosomes, each consisting of an octamer of histone proteins [1,2,3]. Histones are subject to an assortment of post-translational modifications including phosphorylation, acetylation, and methylation [4,5,6]. Many of these modifications have been linked to transcriptional activation, silencing, heterochromatin formation [1,3,7,8,9], DNA damage sensing and repair [10], and chromosomal segregation [11]. Evidence is accumulating to support the hypothesis that different combinations of histone modifications confer different functional specificities [12]. For example, in *Saccharomyces cerevisiae*, the nucleosomes near active promoters are marked by H3K9ac and H3K4me3, while inactive promoters generally lack these marks [1,13,14]. In human, active promoters are associated with H3K4me3, and enhancers are associated with H3K4me1 but lack H3K4me3 [15]. With dozens of covalent modifications already detected on histones, it is conceivable that additional patterns of chromatin modifications exist, and may reveal novel functional elements of the genome.

High-throughput experimental techniques, such as chromatin immunoprecipitation on a microarray (ChIP-chip) [16,17] and its sequencing-based variant ChIP-Seq [18], have been used to map the enrichment of modified histones on a large scale [15]. This data has revealed that the profiles of chromatin modifications over large genomic regions define functional domains. In principle, analysis of the chromatin modification patterns should allow identification of different classes of functional elements associated with the different histone modifications. However, tools for finding chromatin modification patterns have been lacking [1,13,14].

Previously, supervised classification methods have been used to identify chromatin modification patterns at known functional sites [13,15,19,20,21]. For example, many studies focus entirely on well-defined transcriptional promoters [3,8,9,13,15]. But this supervised approach of focusing only on annotated loci trivializes the problem of finding commonly occurring histone modification patterns on a global scale. One of the main motivations for developing an unsupervised learning method is that it is not known *a priori* what functional elements are associated with specific histone modification patterns.

Here, I develop a novel, unbiased method for identification of histone modification patterns occurring repeatedly in the genome. I assume that a consistent repertoire of chromatin modification patterns exists, and that a pattern search algorithm should identify such patterns in an unbiased fashion without using any annotations. I treat this problem as a variant of the standard motif finding problem: given a sequence over an alphabet, find subsequences that are repeated more often than would be expected by chance. Here, rather than working with a sequence over a discrete alphabet such as nucleotides or amino acids, I analyze a sequence of real-valued enrichment of chromatin modifications over a genomic region. To perform motif finding over chromatin modifications, I develop a probabilistic method called ChromaSig. Applying ChromaSig to a panel of chromatin maps from ChIP-chip experiments performed in HeLa cells on ENCODE arrays, I recover eight distinct clusters of chromatin signatures. I recover known patterns observed at putative active promoters and enhancers [15], as well as several previously uncharacterized patterns. Furthermore, the distinct chromatin signatures found at enhancers mark distinct functional classes of enhancers in terms of transcription factor and co-activator binding. Finally, I also apply ChromaSig genome-wide to 21 chromatin marks mapped using ChIP-Seq in CD4+ T cells, recovering 16 distinct and frequently occurring chromatin signatures. ChromaSig reveals frequent and redundant cross-talk between different histone modifications at a previously unappreciated level, and reveals a unique class of quickly-evolving genome elements consistently marked by repressive histone modifications. These results support the utility of ChromaSig in discovering of novel chromatin signatures.

Methods

Overview of ChromaSig

I represent large-scale chromatin modifications maps as enrichment over consecutively tiled 100-bp bins. To find frequently-occurring chromatin signatures, ChromaSig is divided into two parts. In the first part, I find all loci of width 2-kb that are highly enriched in chromatin modifications, and therefore likely to contain chromatin signatures. But as known chromatin signatures at promoters and enhancers are typically larger than 2-kb [15], these enriched loci are likely part of a larger chromatin signature, which may be found in the vicinity of the enriched locus and oriented on either strand of DNA. Thus, I define a search region of 7-kb around each enriched locus where I search for a chromatin signature motif of size 4-kb. This choice of search region and motif sizes ensures that at least 75% of the enriched locus is covered by the motif. In the second part, ChromaSig clusters, aligns, and orients these enriched loci on the basis of chromatin modifications, using a Euclidean distance measure. A given locus *i* can either align to the motif *M*, the background *B*, or some other motif *M'*. For a given histone mark *h*, the likelihood of accepting locus *i* at location offset *l* and orientation *p* into *M* is given by:

$$L_{i,h,l,p} = \frac{\Pr\left(M \middle| \text{ locus i at l, p}\right)}{\Pr\left(B \middle| \text{ locus i at l, p}\right) + \Pr\left(M' \middle| \text{ locus i at l, p}\right)}$$

I then employ a greedy algorithm to align and orient each locus *i* to *M* by choosing the *l* and *p* that maximize the following objective function over all members of the motif: $\sum_{i \in M} \sum_{\text{all } h} L_{i,h,l,p}$.

Algorithmically, I first define the seed motif by finding a small group of loci sharing a common chromatin signature. I then expand this seed to include other loci, simultaneously refining the motif being searched. Let *D* represent the set of loci already assigned to a motif, initially empty. I sequentially visit each locus not in *D* a total of 5 times. All aligned loci having the motif are output and added to *D*, to be excluded for future rounds of pattern searching. This procedure is repeated with a new seed until no more seeds are found. An overview of the algorithm is given in Scheme 5-1 and Figure 5-1.

Chromatin modification data for ChIP-chip

I use published histone profiles for H4ac, H3ac, H3K4me1, H3K4me2, H3K4me3, and core histone H3 [15] (GEO accession GSE6273), as well as H3K9ac, H3K18ac, and H3K27ac [22] (GEO accession GSE7118). These data were obtained from ChIP-chip experiments performed in HeLa cells using oligonucleotide tiling arrays spanning the ENCODE regions, a set of 44 genomic regions with a total length of 30 Mbp. I bin the data into 100-bp bins, averaging the probes falling into each bin.

Finding loci near chromatin signatures

To reduce the search space for finding chromatin signatures, I first focus on enriched loci of width of w = 2-kb containing ChIP-chip signals significantly deviating from background. For each histone modification $h \in 1 \ldots H$, let $x_{h,i}$ be the average log-ratio of bin i. After array normalization, $x_{h,i}$ approximately follows a Gaussian distribution $N(\mu_h, \sigma_h)$. To find both histone modification rich and poor loci, I assign a χ^2 statistic to each locus of size w starting at the jth bin:

$$y_{h,j} = \sum_{k=1}^{w} z_{h,j+k}^2 \sim \chi_w^2$$

where $z_{h,j+k} = (x_{h,j+k} - \mu_h)/\sigma_h$ is a standard normal variate. I perform the above separately for each histone modification and use a p-value cutoff of 1.0E-5 to assess significant loci. To create a non-redundant list of significant loci over all histone modifications, I represent the score of a locus j as the sum of all significant $y_{h,j}$. Also, as it is likely that loci adjacent to significant loci will also be significant, I keep a statistically significant high-scoring locus only if all other loci ≤ 2.5-kb away have a lower score. Finally, I remove all loci poorly represented on the tiling microarray, here defined as containing fewer than 75% of the total number of possible probes in the locus.

Finding distinct chromatin signatures

The enriched loci above are not grouped by chromatin signature, may not be aligned, and, in the case of asymmetric patterns, may not be in the same orientation. The goal is to reverse these statements. But first, I begin with some notation. We are given a set of enriched loci from above and a seed motif of width w_M = 4-kb from initialization (described below). For a given locus, I want to determine if it contains the seed motif. But since the loci is not aligned *a priori*, I expand the search to all width w_M windows containing at least 75% of the locus, in both forward and reverse orientations. Thus, I am searching for a 4-kb motif in a 7-kb search region. For simplicity, I allow each locus to contain at most one motif.

ChromaSig refines one motif at a time. The chromatin signature of each motif is defined by the elements belonging to the motif. More specifically, it is defined as: a set of loci $\{i_1,...i_j,...i_n\}$ that contain the motif, a set of relative locations $\{l_1...l_j...l_n\}$ where l_j indicates the location offset of the motif in locus i_j, and a set of polarities $\{p_1...p_j...p_n\}$ where p_j indicates the orientation of the motif in locus i_j. Here, n is the total number of loci containing the motif, which can range from 1 to N (N is the number of loci, which is 1558 here), and p_j can be either "+" indicating the forward orientation or "-" indicating the reverse orientation. Let s_{h,i_j,l_j,p_j} (denoted by $s_{h,j}$) be the real-valued sequence of the length w_M window corresponding to locus i_j at location l_j and orientation p_j for histone modification h.

Let $s_{h,i_j,l_j,p_j}(k)$ (denoted by *$s_{h,j}(k)$*) be the value of the *k*th bin in this sequence. Given a seed pattern and a locus i_j, I search over all possible $s_{h,j}$ around i_j for an optimal match to the motif.

Define a seed motif as *m* $=\{m_1,\ldots,m_H\}$, where *H* is the number of histone modifications*, h* ranges from 1 to *H, m_h* = $\{\mu_{h,1},\ldots,\mu_{h,w_M}\}$, $\mu_{h,k}=\frac{1}{n}\sum_{j=1}^{n}s_{h,j}(k)$, and *n* is the number of aligned windows. In words, each histone modification *h* has its own length w_M pattern, which is the average of all aligned windows. Define the motif standard deviation similarly:

$$\sigma_{h,k}=\sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(s_{h,j}(k)-\mu_{h,k}\right)^2}$$

During the sampling step, I choose a locus *i* and attempt to align every length w_M window, at all possible locations *l* and orientations *p*, to the current seed motif. I compute the probability of observing a window's sequence under the motif model as

$$M_{h,l,p}=\prod_{k=1}^{w_M}P\left(s_{h,i,l,p}(k);\mu_{h,k};\sigma_{h,k}\right)$$

where $P(x;\mu;\sigma)$ is a probability defined by dividing the Gaussian probability density function by its maximum value: $P(x;\mu;\sigma)=\exp\left(-(x-\mu)^2/\left(2\sigma^2\right)\right)$.

Given a locus to be aligned to the seed, I consider two possibilities: 1) the locus aligns well to the seed and is accepted into the seed, or 2) the locus does not align well and is rejected. In the latter case, the locus may not align well because 2A) the locus matches better to a null background or 2B) the locus matches better to another motif.

To decide between these possibilities, I consider two background models. To consider 2A, I define the null background model by the mean of all bins in the entire ENCODE regions for each histone modification *h* (μ_h) and the mean of the motif standard deviations $\sigma_h = \frac{1}{w_M}\sum_{k=1}^{w_M}\sigma_{h,k}$. The probability of observing a window under the null background model is then:

$$B_{h,l,p} = \prod_{k=1}^{w_M} P\left(s_{h,i,l,p}(k); \mu_h; \sigma_h\right)$$

Ideally, I would consider 2B by aligning a locus to all other possible motifs. But since it is not known *a priori* what motifs exist, I model the probability that a window belongs to another motif by:

$$A_{h,l,p} = \prod_{k=1}^{w_M} P\left(\sigma_{another}; 0; 1\right)$$

where $\sigma_{another}$ is a user-specified parameter (here set to an empirical value of 1.75) that represents the expected quality of the match with another motif, represented as the number of standard deviations from the mean. Larger values of $\sigma_{another}$ indicate a looser background model and smaller values indicate a more stringent background model.

The $M_{h,l,p}$ represent the probabilities to add the locus to the seed at a specific location and orientation for a given histone modification, while the $B_{h,l,p}$ and $A_{h,l,p}$ represent the probabilities to exclude the locus. To determine which window aligns best to the motif model, I form the likelihood:

$$L_{h,l,p} = \frac{\Pr(\text{accept} \mid \text{Data})}{\Pr(\text{reject by 2A} \mid \text{Data}) + \Pr(\text{reject by 2B} \mid \text{Data})}$$

Applying Bayes rule,

$$\begin{aligned} L_{h,l,p} &= \frac{\Pr(\text{Data} \mid \text{accept})p_a}{\Pr(\text{Data} \mid \text{reject by 2A})p_{2A} + \Pr(\text{Data} \mid \text{reject by 2B})p_{2B}} \\ &= \frac{M_{h,l,p}p_a}{B_{h,l,p}p_{2A} + A_{h,l,p}p_{2B}} \end{aligned}$$

where p_a, p_{2A}, and p_{2B} are priors that sum to 1. Here, let $p_a = p_{2B}$ and $p_{2A} = 0.01$. When $L_{h,l,p} < 1$, the chance of rejecting a window is greater than accepting it into the motif. If this is true for all *l* and *p* for a given *h*, then there can be no favorable alignment of any window from the given locus to the motif that involves the histone modification *h*. In such a case, I unilaterally reject the locus, regardless of how well other histone modifications align. Otherwise, I find the *l* and *p* that maximizes $\sum_{h=1}^{H} \log L_{h,l,p}$, and add this aligned locus into the seed motif.

A cycle is defined to be the process of aligning each locus to the seed motif. At the end of a cycle, I construct a new seed motif containing all accepted windows in their aligned locations and orientations. At the end of 5 cycles, I output the motif and aligned loci belonging to it. To ensure generality of the chromatin signatures, I reject clusters with fewer than 20 elements or clusters having a maximum absolute log-ratio signal less than 0.5 .

Initialization

While most of the loci input to ChromaSig will not be aligned, I do expect that a small number of them will be nearly aligned. To determine the seed motif, I attempt to create seeds starting from 100 randomly chosen enriched loci. For each such locus *i*, I compute the Euclidean distance to all other loci and then use a fast approximate sorting method to find the closest ~20 loci to *i*, which forms a potential seed. Specifically, I define the leaves of a tree as the loci distances in random order and then construct a tournament tree until there are ≤ 20 parent nodes. A good seed contains both regions of high signal and low signal, with the members of the seed sharing a very similar chromatin signature. Notably, a seed saturated with signal is uninformative, as it will be difficult to align. I distinguish good seeds by using the following score:

$$seedscore = \sum_{h=1}^{H} \frac{\sum_{k=1}^{w_M/2} \left|\mu'_{h,k}\right| - \sum_{k=w_M/2+1}^{w_M} \left|\mu'_{h,k}\right|}{\sigma_h}$$

where $\mu'_{h,k}$ is $\left|\mu_h\right|$ in descending order. A high seed score indicates a motif with balanced amounts of high and low signal, together with a small standard deviation. I use the seed with the highest score to initialize ChromaSig.

Application of ChromaSig to genome-wide ChIP-Seq data

To ensure that ChromaSig is sufficiently general, I also apply it to genome-wide distributions of 21 histone marks mapped by ChIP-Seq in CD4+ T cells [18].

- **Data normalization**: I consider only those reads that map uniquely to the genome (hg18) with a maximum of 2 mismatches, and count polyclonal reads once to reduce sequencing bias. I partition the genome into 100-bp bins and count the number of reads in each bin. The number of unique monoclonal reads may be highly variable between different histone marks. For example, there are 15.4 million reads spanning H3K4me3 but only 1.9 million spanning H3K79me2. This vast difference in coverage makes it difficult to compare ChIP enrichment for different histone marks by comparing tag counts. Even for a single mark, sites of true ChIP enrichment can have a large difference in ChIP-Seq tag density [23]. To address these concerns, I normalize the number of reads in each bin $x_{h,i}$ with a sigmoid function:

$$x'_{h,i} = \frac{1}{1+e^{-(x_{h,i}-median(x_h))/std(x_h)}}$$

Where median(x_h) and std(x_h) are the median and standard deviation of the number of tags in the 100-bp bins for histone mark h, excluding spurious bins containing exactly 0 and 1 reads. By definition, $x'_{h,i}$ will be 0.5 for bins containing the median number of tags, falls to 0 as tag counts decrease, and saturates to 1 as tag counts increase.

- **Finding ChIP-Seq signal-rich loci**: As I cannot assume a Gaussian distribution of normalized enrichment, I model the background empirically using all 2-kb windows in the ENCODE regions. Furthermore, there are twice as many chromatin marks in the ChIP-Seq dataset compared to the ChIP-chip dataset, and being genome-wide the coverage is 100 times higher. To focus on the highest quality loci, I keep a statistically significant high-scoring locus only if all other loci less than 5.0-kb away have a lower score, rather than the 2.5 kb used for ChIP-chip. Furthermore, several chromatin marks including H3K9me3 and H3K36me3 are known to be enriched over large domains. To focus on chromatin signatures smaller than 10-kb, when creating a non-redundant list of significant loci, I only consider those loci $y_{h,j}$ with p-value smaller than 1E-5 and that are more than 2.5-kb away from any other significant locus in h.

- **Motif with pseudocounts**: As ChIP-Seq provides a digital readout of ChIP enrichment, many bins are empty, and it is possible that the motif mean $\mu_{h,k} = 0$ for some h and k, which results in $\sigma_{h,k} = 0$. To relieve this prohibitive constraint, I add a pseudocount of 0.5 to each position of the motif:

$$\mu_{h,k}^{seq} = \frac{1}{n+1}\left(0.5 + \sum_{j=1}^{n} s_{h,j}(k)\right)$$

$$\sigma_{h,k}^{seq} = \sqrt{\frac{1}{n}\left(\left(0.5 - \mu_{h,k}^{seq}\right)^2 + \sum_{j=1}^{n}\left(s_{h,j}(k) - \mu_{h,k}^{seq}\right)^2\right)}$$

As the number of elements in the motif increases, the contribution of the pseudocount decreases.

- **Parameters**: I run ChromaSig on ChIP-Seq data with the same parameters as for ChIP-chip data. But to focus only on the most frequently-occurring chromatin signatures, I consider only those clusters with an average normalized enrichment greater than 0.25 and with at least 500 loci.

Results

ChromaSig identifies distinct chromatin signatures

Starting with ChIP-chip data for H4ac, H3ac, H3K9ac, H3K18ac, H3K27ac, H3K4me1, H3K4me2, H3K4me3, and core histone H3 spanning the ENCODE regions, I first use a sliding window approach to identify signal-rich loci likely to contain histone modification patterns (see Methods, Figure 5-1). The goal is to find commonly-occurring patterns in this set of loci. But because this sliding-window approach is quite crude, it is unlikely that the loci will be aligned. Furthermore, a chromatin profile can be observed in two possible orientations corresponding to the two DNA strands running in opposite directions, and the sliding window approach does not account for these orientations. As such, it is unlikely that the collection of signal-rich loci is oriented optimally to preserve asymmetric chromatin signatures, such as those found at promoters [15]. I employ ChromaSig to align and orient these loci into clusters with similar chromatin signatures. Different chromatin signatures can be distinguished by different enrichment of one or more histone modifications, or they may share similar enrichment for all modifications but contain a different enrichment profile for one or more modifications. I find eight clusters spanning 1118 loci (Figure 5-2).

Loci in the same cluster share the same chromatin signature, and each cluster has a distinct chromatin signature (Figure 5-2), indicating that the method is functioning as designed. To highlight the similarities and differences of each cluster, I perform hierarchical clustering on the average profiles of each cluster (Figure 5-2). This reveals that, while some clusters are strikingly distinct from one another, others are only subtly different. On the more distinct side, CS1 is the only cluster to have strong enrichment of H3K4me3, while cluster CS8 is the only cluster to be enriched solely in H3K4me1. More subtly, the chromatin marks present at CS2 and CS3 are the same, but are consistently weaker in CS3 than CS2. Along the same lines, CS6 has narrower and weaker enrichment of H3K4me1 that distinguishes it from the other clusters bearing the H3K4me1 mark. The smallest cluster CS6 contains 44 aligned loci, suggesting that the patterns occur frequently, and may likely be found outside of the ENCODE regions. At the same time, loci in the same cluster also share similar profiles for functional marks (RNAPII, TAF250, p300), which were not the criteria used by ChromaSig. This enrichment of functional marks implies that the clusters group together functionally related genomic loci.

Comparing ChromaSig clusters to clusters from a supervised learning method

To assess the performance of ChromaSig in finding distinct chromatin signatures, I compare ChromaSig signatures to those recovered by a supervised learning approach. Using a training set of chromatin signatures at promoters and enhancers, I previously predicted 198 promoters and 389 enhancers [15]. Because this method relied on a sliding window approach that considers aligning chromatin signatures from both strands, each set of predictions should be aligned and oriented. To find distinct clusters of histone modifications on the basis of the nine chromatin marks studied here, I perform k-means clustering on the chromatin modifications near each of these two sets of predictions, giving promoter clusters SP1-4 and enhancer clusters SE1-4 (Figure 5-3A-B).

To assess the quality of ChromaSig clustering and alignment, I compare the clusters of predicted enhancers and promoters that recover at least 25% of the loci from each ChromaSig cluster (Figure 5-4, Figure 5-5). The two ChromaSig clusters CS2 and CS7 show striking similarity with clusters SE3 and SE4, respectively (Figure 5-4B, Figure 5-5B). Remarkably, even without a training set, ChromaSig employing an unsupervised learning method recovers chromatin signatures found by a supervised learning technique.

This picture changes with ChromaSig cluster CS1, which is recovered by SP3 and SP4. All three of these clusters are enriched with the same chromatin modifications, indicating that the two methods perform similarly, at least at a coarse scale. But interestingly, while the asymmetric patterns SP3 and SP4 are distinct, they appear to be mirror images of each other, and are likely the same pattern observed in opposite directions. Since ChromaSig considers strand orientation in its alignment, cluster CS1 is essentially a merge of these two mirrored clusters, forming a single distinct, consistent, and asymmetric pattern (Figure 5-4A). Thus, patterns recovered by ChromaSig are less redundant. Also, cluster CS8 contains only H3K4me1 enrichment, and the only cluster that recovers it also contains numerous loci enriched in H3K18ac and H3K27ac (Figure 5-4C). This, together with the fact that clusters CS4-6 are not recovered by any of clusters SP1-4 and SE1-4, indicate that ChromaSig can find distinct patterns not found by this supervised learning method.

ChromaSig clusters preserve pattern asymmetry, are better aligned, are less redundant, contain loci with more consistent patterns, and contain unique patterns that are not found by the supervised learning method. Most importantly, ChromaSig does not require the construction of training sets, nor does it require the specification of arbitrary parameters such as the number of clusters to find. Instead, ChromaSig finds the natural groupings of the data, creating new clusters as necessary.

ChromaSig identifies known patterns at promoters and enhancers

To date only a handful of distinct histone modification patterns have been broadly associated with specific functions. These include active promoters that are generally marked by the presence of H3K4me3 but absence of H3K4me1, and enhancers marked by the presence of H3K4me1 but absence of H3K4me3 [3,13,15]. To assess whether ChromaSig clusters of chromatin signatures correspond to specific biological functions, I first compare them to existing genome annotation.

Transcription start sites (TSS): Catalogs of transcription start sites (TSSs) are one of the most abundant and nearly complete annotations for human genomic elements. Of the 559 unique Refseq TSSs [24] in the ENCODE regions, 208 (37.2%) are proximal (hereafter defined as within 2.5-kb) to cluster CS1, far more than any other cluster (Figure 5-6A). To assess the significance of this overlap, I compare with 100 random sets of clusters of the same size, sampled from regions on the ChIP-chip array to avoid biases from probe-poor regions, giving a p-value of 3.2E-141 assuming a Gaussian distribution. The majority of Refseq TSSs are not recovered, as roughly half of them do not contain enrichment of these histone modifications [15].

Promoter and enhancer predictions: In Chapter 2, I use the same dataset but with a supervised learning approach to predict active promoters and enhancers [15]. A majority (62.6%, $p <$ 1.0E-300) of the predicted active promoters are proximal to cluster CS1 (Figure 5-6B). In addition, the enhancer predictions generally fall into clusters CS2-3 and CS6-8 (Figure 5-6C). These results indicate that cluster CS1 is highly enriched in promoters containing the active chromatin marks, while clusters CS2-3 and CS6-8 are enriched in HeLa-marked enhancers.

DNase I hypersensitivity (HS) sites: DNase I hypersensitivity is a hallmark for many types of *cis*-regulatory elements. Using a list of putative HS sites found from high-throughput, high resolution DNase-chip experiments [25], I find significant enrichment of HS sites at clusters CS1 (p = 6.7E-165), CS2-3 (p_{CS2} = 8.4E-36, p_{CS3} = 7.3E-16), and CS6-7 (p_{CS6} = 7.1E-6, p_{CS7} = 2.5E-7) (Figure 5-6D), consistent with their proposed function as promoters and enhancers. On the other hand, clusters CS4-5 shows marked depletion of HS sites (p_{CS4} = 9.7E-9, p_{CS5} = 3.7E-4).

Distinct chromatin signatures associated with distinct functions

I recover several distinct chromatin signatures associated with predicted HeLa enhancers. CS8 is only enriched in H3K4me1, while CS7 also contains H3K18ac and H3K27ac enrichment. In addition to these marks, clusters CS2-3 also have H3K4me2 enrichment, with CS2 being more acetylated than CS3. As the remaining cluster CS6 is the only one to have less than 25% of its loci recovered by predicted enhancers and also has the weakest enrichment of the enhancer hallmark H3K4me1, it may contain loci other than enhancers and I exclude CS6 from this analysis.

If the histone code hypothesis is true, I would expect functional differences between enhancers marked by different signatures. To assess if the different enhancer-like clusters also have distinct functional roles, I examine enrichment in binding sites for a variety of transcription factors and co-activators mapped in HeLa. I notice that binding sites for the transcription factor c-Myc is significantly enriched at clusters CS2 and CS3 (p_{CS2} = 4.6E-50, p_{CS3} = 3.6E-7) (Figure 5-7A). Visually comparing the chromatin modifications at these clusters which have c-Myc enrichment to clusters CS7-8 that lack c-Myc enrichment, I observe that CS2-3 have enrichment of H3ac, H4ac, and H3K4me2, while these chromatin marks are absent in E3-4. Thus, one of these marks may be important to c-Myc function. In contrast, the co-activator p300 is highly enriched at clusters CS2, CS3, and CS7 (p_{CS2} = 1.5E-75, p_{CS3} = 4.1E-8, p_{CS7} = 3.3E-8) (Figure 5-7B). Strikingly, the only cluster lacking p300 enrichment, CS8, is also the only cluster to lack enrichment of H3K18ac and H3K27ac, suggesting a connection between these chromatin marks and p300 activity. Finally, binding sites for a different co-activator MED1 are only enriched at clusters CS2 and CS7 (p_{CS2} = 5.4E-50, p_{CS7} = 4.9E-4) (Figure 5-7C), distinct from binding of p300 and c-Myc. These results suggest that enhancers marked by different chromatin signatures have unique functional roles dictated by distinct protein complexes.

ChromaSig identifies other potential regulatory sequences

Outside of promoters and enhancers, current knowledge on common histone modification patterns is sparse. ChromaSig identifies two novel signatures CS4-5 marking sites of unknown function, as well as CS6 which is only slightly recovered by enhancer predictions. To assess the possible functional significance of these genomic sites, I first analyze sequence conservation. Here, I use PhastCons scores from multiple alignments of 7 vertebrate genomes (chimp, mouse, rat, dog, chicken, fugu, and zebrafish) and human [26] to determine the amount of between-species conservation of each cluster (Figure 5-8). Conservation scores for clusters CS4-6 are generally significantly greater than that expected at random (p_{CS4} = 9.6E-5, p_{CS5} = 7.8E-2, p_{CS6} = 1.6E-3, as assessed by the Wilcoxon signed rank test compared to 10000 random sites). Turning to RegPot, which scores the regulatory potential of regions in the human genome, I find that these clusters also have greater regulatory potential than that expected at random (p_{CS4} = 3.5E-11, p_{CS5} = 2.1E-2, p_{CS6} = 1.6E-7). Together, these results suggest clusters CS4-6 contain biologically functional loci.

Clusters CS4-5 are generally depleted of all histone modifications, as well as the functional marks RNAP II, TAF1, and p300 (Figure 5-2). The overlap of cluster CS4 at Refseq TSSs (Figure 5-6A) and the lack of overlap at active promoters (Figure 5-6B) suggest that some CS4 sites may contain inactive TSSs. To assess this, I examine enrichment of clusters at promoters of expressed and unexpressed genes (Figure 5-9A-B). I observe depletion of clusters CS4-5 at the 5’ ends of expressed genes ($\mathrm{p}_{\mathrm{CS4}}$ = 7.5E-4, $\mathrm{p}_{\mathrm{CS4}}$ = 1.6E-2), and CS4 is actually enriched at promoters of unexpressed genes ($\mathrm{p}_{\mathrm{CS4}}$ = 2.4E-2). Thus, some members of CS4 may be inactive promoters. I also observe significant enrichment of cluster CS6 at promoters of unexpressed genes (p = 1.7E-3) (Figure 5-9B). This suggests that, in addition to containing enhancers, this cluster of evolutionarily conserved sequences that are marked by HS in HeLa cells may also contain inactive promoters.

As the majority of clusters CS4-5 are not explained by promoters, I next ask if these clusters recover other distal regulatory elements. The depletion of HeLa HS sites in CS4-5 (Figure 5-6D) suggests that these clusters should also be depleted of transcription factor binding sites (TFBSs). But when I examine the overlap with STAT1 binding sites in HeLa cells treated with IFN-γ (Chapter 4) [22], I observe striking enrichment with cluster CS4 (p = 5.4E-5) (Figure 5-9C). Interestingly, while ChromaSig clusters are derived from HeLa chromatin profiles, the STAT1 overlap occurs in a different cellular context, suggesting that cluster CS4 may harbor TFBSs not bound in HeLa cells.

The PreMod database [27] contains 1655 putative conserved TF modules in the ENCODE regions. As PreMod is determined by static sequence data, its sites represent TFBSs under various cellular conditions. To test the hypothesis that clusters CS4-5 mark TFBSs not bound in HeLa cells, I test the enrichment of these clusters at PreMod sites distal to HeLa HS sites. Interestingly, I find that CS4 members are enriched in these sites (p_{CS4} = 7.6E-5), suggesting that this cluster contains sites that potentially bind TFs, but not in HeLa cells (Figure 5-9D). As an independent method to support this result, I combine HS sites previously mapped in six non-HeLa cell lines [25,28]. Removing those sites near HeLa HS sites, I find significant enrichment with clusters CS4 and CS5 (p_{CS4} = 1.4E-4, p_{CS4} = 3.0E-2) (Figure 5-9E). Finally, I compare clusters CS4-5 with enhancers predicted in four cell types [22], using a previously published chromatin signature-based method [15]. Of those enhancers not marked by HS in HeLa cells, I observe significant enrichment at clusters CS4-5 (p_{CS4}= 3.7E-2, p_{CS5} = 7.7E-3) (Figure 5-9F). Together, these results suggest that ChromaSig clusters having novel chromatin signatures also contain regulatory sequences.

ChromaSig identifies distinct chromatin signatures in genome-wide ChIP-Seq data

So far, I have shown that ChromaSig can find distinct chromatin signatures using ChIP-chip data spanning the ENCODE regions. But the question remains as to whether ChromaSig is applicable on a genome-wide level or on ChIP-Seq data from next-generation sequencing. To address this, I focus on a recently published study by Barski et al. which used ChIP-Seq to map the genome-wide distributions of 21 chromatin marks in CD4+ T cells [18]. I identify 16 clusters containing distinct chromatin signatures spanning 49340 genomic loci (Figure 5-10). Using hierarchical clustering with a Euclidean distance measure to categorize the average profiles of each cluster reveals that there are essentially two main categories of genomic elements. One class, GW1-10, contains combination of the activating marks H3K4me1/2/3 and H2BK5me1. Another class, GW11-16, are more prevalently marked by the repressive marks H3K9me3, H3K27me3, and H3K36me3, and H3K79me3.

There are 5 clusters significantly enriched for promoters (Figure 5-11A), each with a distinct combination of chromatin marks. To assess significance, I compare with 100 random sets of clusters of the same size, sampled from non-repeat masked regions of the genome. In addition to being the only promoter cluster enriched in H4K20me1, GW1 contains the strongest enrichment of H3K4me3 with a corresponding wide valley of H3K4me1 enrichment, in contrast to GW7 which has weaker H3K4me3 enrichment followed by a narrower H3K4me1 enrichment profile and GW5 which contains even weaker enrichment of these marks. Of the remaining promoter-associated clusters, GW8 contains "bivalent" promoters enriched in active H3K4me3 and repressive H3K27me3 marks [29], while GW16 is mainly enriched in the repressive marks H3K9me3, H3K27me3, and H3K79me3.

Enrichment of H3K36me3 has been associated with the 3' ends of highly expressed genes [18]. Consistent with this, I observe that GW11-12, which contain the strongest enrichment of H3K36me3, are also enriched at Refseq 3'-ends (Figure 5-11B). While the vast majority of histone modifications at these two clusters are similar, it is also clear that GW11 is more enriched in H3K9me1 and H4K20me1 than GW12.

Recently, Boyle et al. mapped DNase I hypersensitive sites genome-wide in CD4+ T cells [30]. Here, I observe that clusters GW1-10, which generally contain active marks, are all enriched in DHS sites. In contrast, the remaining clusters GW11-16 marked by repressive marks all lack DHS enrichment (Figure 5-11C). Thus, GW1-10 likely contain regulatory elements functioning in CD4+ T cells. Mirroring this observation, clusters GW1-10 are also generally enriched in known regulatory elements as annotated by ORegAnno [31] (Figure 5-11D).

This analysis reveals possible functional roles for GW1-12 and GW16. Like these clusters, each remaining cluster contains loci that share a consistent chromatin signature, suggesting that each cluster contains loci that may function similarly. Interestingly, GW13-16 are all consistently marked by repressive chromatin marks, and in particular the heterochromatin mark H3K9me3. But unlike large domains of heterochromatin, GW13-16 appear to be localized to small heterochromatic loci spanning less than 5 kb. To assess possible functionality for GW13-15, I next turn to sequence conservation. Surprisingly, these clusters and GW16 are actually less conserved than expected at random ($p < 1e-15$) (Figure 5-11E). Thus, GW13-16 contain quickly evolving but consistently marked, locally heterochromatic regions of the genome, though their specific functions remain unknown.

Discussion

Large-scale maps of histone modifications provide a global view of epigenetic status and allow investigation of the influence of epigenetics in development and disease. Thanks to the development of large scale experimental approaches including ChIP-chip and ChIP-seq [16,32], datasets of histone modification profiles are rapidly accumulating. However, while numerous methods have been developed to identify the binding locations of transcription factors (TFs) from these data [19,20,21,33], methods for analysis of histone modification profiles are still lacking due to unique challenges that have not been encountered with TF data. Binding sites for TFs are generally discrete peaks and are sparsely scattered throughout the genome [19], whereas histone modifications are often repeated over many consecutive nucleosomes [1,3]. As such, finding regions of interest in a histone modification landscape is quite different from finding TF hits. While using standard peak-finding on histone modifications is possible, this approach suffers from several drawbacks. First, peak-finding ignores loci depleted of binding signal, which can be important in mapping nucleosome-depleted regions [15]. Second, analysis of histone modification data is focused on identifying a specific pattern in regions often spanning thousands of base pairs (bps) while peak finding for TFs is generally focused on much smaller regions. Third, peak finding ignores the binding profile's orientation, but the orientation of asymmetric histone patterns can be quite functionally revealing [13,15]. Finally, peak-finding treats different proteins independently, ignoring the correlation of different histone modifications, and thereby reducing the likelihood of discovering novel biological insights from the combinatorial presence of multiple histone modifications [13,15].

In this study, I introduce a strategy called ChromaSig to find commonly occurring chromatin signatures given a landscape of histone modification profiles. Using an unsupervised learning approach, ChromaSig simultaneously clusters, aligns, and orients chromatin signatures without using any training sets or external annotations. Using histone modification data alone, ChromaSig is able to distinguish subtle differences in chromatin signatures, allowing it to find natural groupings of the data without relying explicitly on heavily constraining parameters such as the number of expected clusters, which can severely hamper pattern discovery. Interestingly, even with this limited input, ChromaSig recovers chromatin signatures similar to a previously published supervised learning method that used high-quality curated training sets. In addition to discovering new chromatin signatures, the ChromaSig clusters preserve pattern asymmetry, are better aligned, and are less redundant.

ChromaSig is sensitive enough to recover known histone modification patterns for active promoters and enhancers. This recovery of known patterns further suggests that the novel patterns are real. This method is also able to clearly distinguish between different classes of enhancers based on chromatin modifications. Interestingly, I find that different functional activities of associated with enhancers, such as binding of specific co-activators and transcription factors, are linked to specific histone modifications present at the enhancers. While the mechanism for this phenomenon is unclear and will require further study, it is tempting to speculate that additional maps of chromatin marks and transcription factors in HeLa cells may uncover more specific classes of enhancer chromatin signatures associated with more specific functions, lending further support to the histone code hypothesis. This phenomenon may also occur at other genomic elements such as promoters and insulators.

ChromaSig also recovers several novel clusters CS4-5, which are simultaneously depleted of 9 chromatin modification marks and 3 general transcription factors. Such depletion may correspond to special chromatin structures that are generally resistant to immunoprecipitation. Indeed, depletion of ChIP/Input signals at these loci is also observed in independent ChIP-chip experiments against STAT1, c-Myc and other transcription factors using HeLa S3 cells [15,34]. However, I find that these sites contain evolutionarily conserved sequences and are enriched in inactive promoters and TFBSs. These observations suggest that clusters CS4-5 contain potential regulatory elements.

Application of ChromaSig genome-wide recovers only 16 distinct chromatin signatures. With the 21 different histone modifications studied here, the number of different possible combinations is 2^21. Strikingly, ChromaSig reveals that the number of frequently-occurring histone modifications is actually quite small in humans, a result mirrored in yeast [13], and some chromatin signatures occur much more frequently than others. Notably, GW1-10 are all enriched in DNase I hypersensitive sites, indicating they are likely to mark function genomic elements in CD4+ T cells. Of these, GW1/5/7/8 are highly enriched in H3K4me3, and consistent with this, are also enriched in promoters. The remaining hypersensitive clusters are enriched in known regulatory elements, some of which may be enhancers.

Consistent with this, many of these clusters contain stronger enrichment of H3K4me1 than H3K4me3. Extending from results focused on the ENCODE regions, I hypothesize that these different clusters are bound by a different combination of transcription factors and co-activators.

In recent years, numerous studies have used the genome sequence, along with high-throughput expression and transcription factor ChIP data, to characterize regulatory elements [21,35]. As the epigenetic code offers an abstraction over the genetic code, using it alone may be viable in the study of some functional genomic elements – including genes, enhancers, repressors, insulators, and other regulatory elements. As the availability of large-scale data for chromatin marks increases, the ability of methods such as the one presented here to concisely describe the underlying chromatin signatures, thereby abstracting away irrelevant or redundant data, will become increasingly more critical. Future efforts to unify both epigenetic and genetic content will be quite powerful in further identifying and characterizing regulatory elements that have eluded current methods.

Acknowledgements

Chapter 5, in full, is a reprint of the material as it appears in PLoS Computationa Biology 2008. Hon, Gary ; Ren, Bing ; Wang, Wei. “ChromaSig: A Probabilistic Approach to Finding Common Chromatin Signatures in the Human Genome”, PLoS Computational Biology, vol. 4, 2008. The dissertation author was a primary investigator and author of this paper. The dissertation author performed all the analysis presented.

References

1. Millar CB, Grunstein M (2006) Genome-wide patterns of histone modifications in yeast. Nat Rev Mol Cell Biol 7: 657-666.

2. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, Wang JP, Widom J (2006) A genomic code for nucleosome positioning. Nature 442: 772-778.

3. Roh TY, Cuddapah S, Cui K, Zhao K (2006) The genomic landscape of histone modifications in human T cells. Proc Natl Acad Sci U S A 103: 15782-15787.

4. Grant PA (2001) A tale of histone modifications. Genome Biol 2: REVIEWS0003.

5. Wang H, Wang L, Erdjument-Bromage H, Vidal M, Tempst P, Jones RS, Zhang Y (2004) Role of histone H2A ubiquitination in Polycomb silencing. Nature 431: 873-878.

6. Nathan D, Ingvarsdottir K, Sterner DE, Bylebyl GR, Dokmanovic M, Dorsey JA, Whelan KA, Krsmanovic M, Lane WS, Meluh PB, Johnson ES, Berger SL (2006) Histone sumoylation is a negative regulator in Saccharomyces cerevisiae and shows dynamic interplay with positive-acting histone modifications. Genes Dev 20: 966-976.

7. Sims RJ, 3rd, Reinberg D (2006) Histone H3 Lys 4 methylation: caught in a bind? Genes Dev 20: 2779-2786.

8. Kim TH, Barrera LO, Qu C, Van Calcar S, Trinklein ND, Cooper SJ, Luna RM, Glass CK, Rosenfeld MG, Myers RM, Ren B (2005) Direct isolation and identification of promoters in the human genome. Genome Res 15: 830-839.

9. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B (2005) A high-resolution map of active promoters in the human genome. Nature 436: 876-880.

10. Bergink S, Salomons FA, Hoogstraten D, Groothuis TA, de Waard H, Wu J, Yuan L, Citterio E, Houtsmuller AB, Neefjes J, Hoeijmakers JH, Vermeulen W, Dantuma NP (2006) DNA damage triggers nucleotide excision repair-dependent monoubiquitylation of histone H2A. Genes Dev 20: 1343-1352.

11. Cimini D, Mattiuzzo M, Torosantucci L, Degrassi F (2003) Histone hyperacetylation in mitosis prevents sister chromatid separation and produces chromosome segregation defects. Mol Biol Cell 14: 3821-3833.

12. Jenuwein T, Allis CD (2001) Translating the histone code. Science 293: 1074-1080.

13. Liu CL, Kaplan T, Kim M, Buratowski S, Schreiber SL, Friedman N, Rando OJ (2005) Single-nucleosome mapping of histone modifications in S. cerevisiae. PLoS Biol 3: e328.

14. Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, Lee TI, Bell GW, Walker K, Rolfe PA, Herbolsheimer E, Zeitlinger J, Lewitter F, Gifford DK, Young RA (2005) Genome-wide map of nucleosome acetylation and methylation in yeast. Cell 122: 517-527.

15. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39: 311-318.

16. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA (2000) Genome-wide location and function of DNA binding proteins. Science 290: 2306-2309.

17. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature 409: 533-538.

18. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K (2007) High-resolution profiling of histone methylations in the human genome. Cell 129: 823-837.

19. Zheng M, Barrera LO, Ren B, Wu YN (2005) ChIP-chip: Data, Model, and Analysis. UCLA Department of Statistics Papers.

20. Johnson WE, Li W, Meyer CA, Gottardo R, Carroll JS, Brown M, Liu XS (2006) Model-based analysis of tiling-arrays for ChIP-chip. Proc Natl Acad Sci U S A 103: 12457-12462.

21. Chris Benner FGA, Shankar Subramaniam, Christopher Glass. HOMER: An algorithm for the de novo discovery of cis-regulatory elements from high throughput data; 2006; San Diego.

22. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching KA, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenkov VV, Stewart R, Thomson JA, Crawford GE, Kellis M, Ren B (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature.

23. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316: 1497-1502.

24. Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33: D501-504.

25. Crawford GE, Davis S, Scacheri PC, Renaud G, Halawi MJ, Erdos MR, Green R, Meltzer PS, Wolfsberg TG, Collins FS (2006) DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. Nat Methods 3: 503-509.

26. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034-1050.

27. Blanchette M, Bataille AR, Chen X, Poitras C, Laganiere J, Lefebvre C, Deblois G, Giguere V, Ferretti V, Bergeron D, Coulombe B, Robert F (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res 16: 656-668.

28. Xi H, Shulha HP, Lin JM, Vales TR, Fu Y, Bodine DM, McKay RD, Chenoweth JG, Tesar PJ, Furey TS, Ren B, Weng Z, Crawford GE (2007) Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. PLoS Genet 3: e136.

29. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES (2006) A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell 125: 12.

30. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE (2008) High-resolution mapping and characterization of open chromatin across the genome. Cell 132: 311-322.

31. Montgomery SB, Griffith OL, Sleumer MC, Bergman CM, Bilenky M, Pleasance ED, Prychyna Y, Zhang X, Jones SJ (2006) ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. Bioinformatics 22: 637-640.

32. Wei CL, Wu Q, Vega VB, Chiu KP, Ng P, Zhang T, Shahab A, Yong HC, Fu Y, Weng Z, Liu J, Zhao XD, Chew JL, Lee YL, Kuznetsov VA, Sung WK, Miller LD, Lim B, Liu ET, Yu Q, Ng HH, Ruan Y (2006) A global map of p53 transcription-factor binding sites in the human genome. Cell 124: 207-219.

33. Qi Y, Rolfe A, MacIsaac KD, Gerber GK, Pokholok D, Zeitlinger J, Danford T, Dowell RD, Fraenkel E, Jaakkola TS, Young RA, Gifford DK (2006) High-resolution computational models of genome binding events. Nat Biotechnol 24: 963-970.

34. ENCODE_Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306: 636-640.

35. Wang W, Cherry JM, Nochomovitz Y, Jolly E, Botstein D, Li H (2005) Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. Proc Natl Acad Sci U S A 102: 1998-2003.

36. Kim J, Bhinge AA, Morgan XC, Iyer VR (2005) Mapping DNA-protein interactions in large genomes by sequence tag analysis of genomic enrichment. Nat Methods 2: 47-53.

Figures and Tables

```
$N$ = number of loci
$D$ = the set of all assigned loci, initially empty
Repeat while ($N \neq |D|$)
       Find a seed motif $M$ of loci $\notin D$ sharing a chromatin signature
       Repeat 5 times
              For each locus $i \notin D$
                     Compute the likelihood of adding $i$ into $M$
                     Choose to exclude $i$ from $M$, or add $i$ to $M$ in a
                            Specific location and orientation
              Update $M$
       $D = D \cup M$
       Print M
```

Scheme 5-1: Overview of ChromaSig

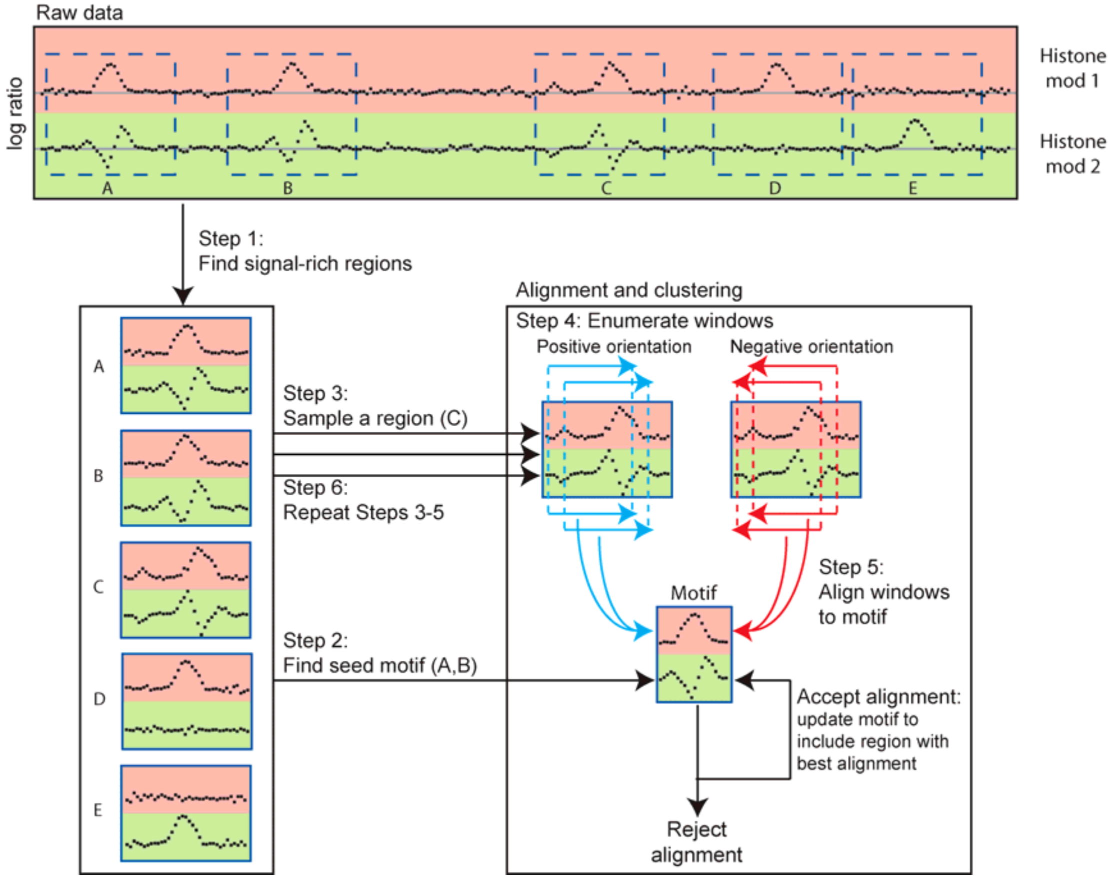


Figure 5-1: Schematic overview of ChromaSig.

In Step 1, I scan genome-scale histone modification maps to find signal-rich loci that potentially contain chromatin signatures. In Step 2, I generate a seed pattern to initialize ChromaSig. In Steps 3 through 5, I visit each enriched locus in turn, enumerate all possible 4-kb windows spanning at least 75% of the locus, and align each window to the seed. This is repeated until each locus has been visited 5 times. Loci that align well to the seed are added to the seed.

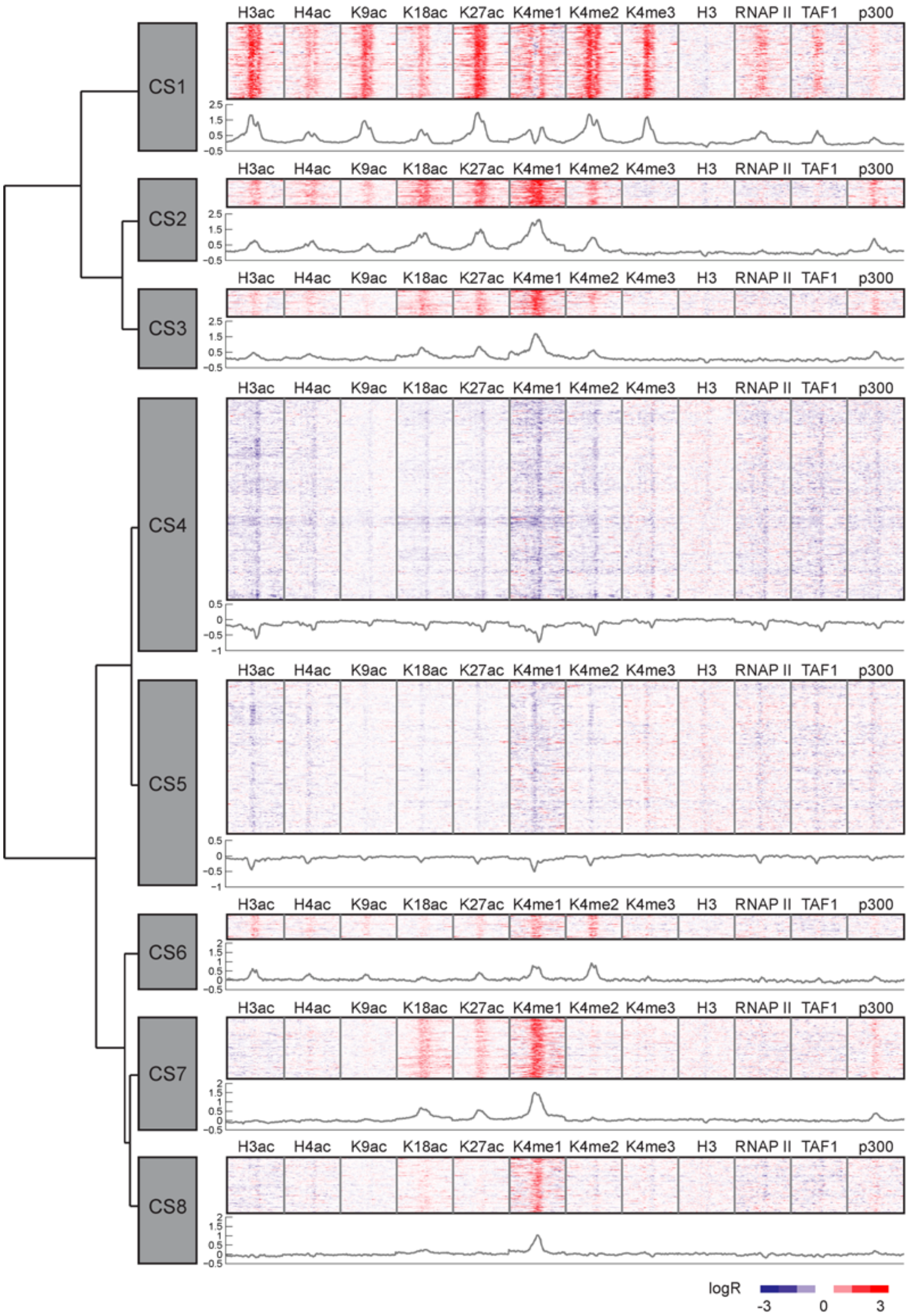


Figure 5-2: ChromaSig clusters recovered from 9 chromatin marks mapped by ChIP-chip in HeLa cells on ENCODE arrays.

Heatmaps (top) and average histone modification profiles (bottom) for each cluster output by ChromaSig. Each horizontal line in the heatmap represents chromatin marks for a single locus. The window size for each mark is 10-kb. Nine histone marks used by ChromaSig and three independent functional marks (RNAPII, TAF250, p300) are presented. To organize these clusters visually, I use hierarchical clustering with a Euclidean distance metric (left).

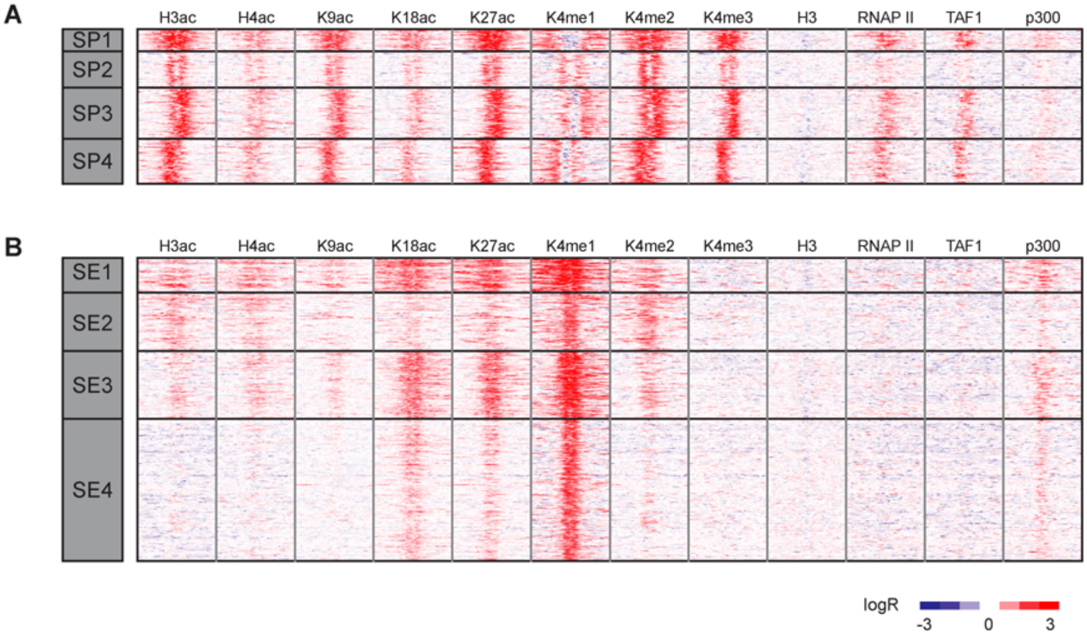


Figure 5-3: Heatmaps of promoter and enhancer predictions from Chapter 2.

Heatmaps of chromatin modifications and functional marks found at (A) promoter and (B) enhancer predictions, after performing k-means clustering on the nine chromatin marks (k=4).

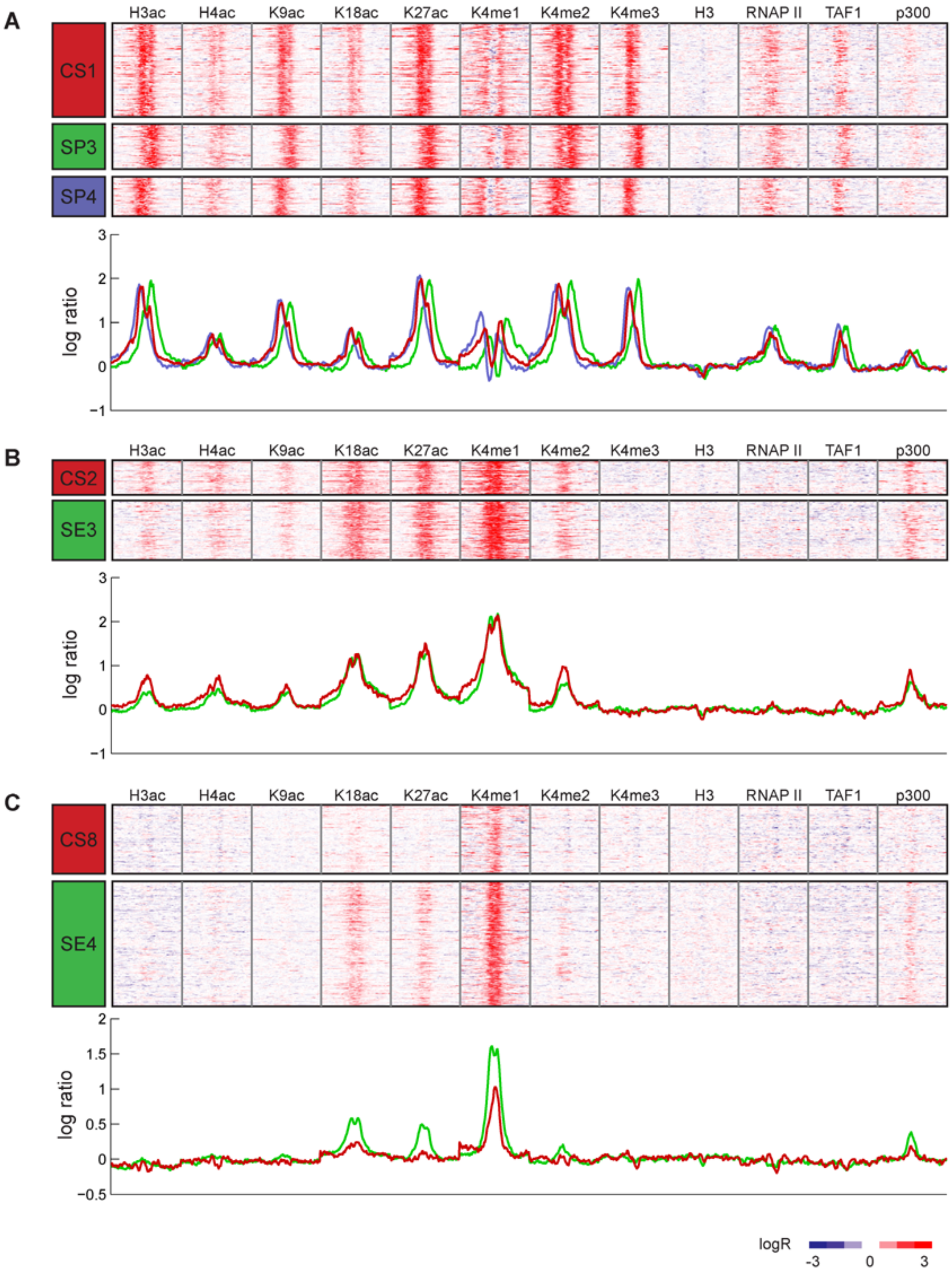


Figure 5-4: Comparison of ChromaSig to the supervised clustering method from Chapter 2.

(A) Heatmaps (top) and average histone modification profiles (bottom) for cluster CS1, together with those for SP3 and SP4, which recover CS1 (33.3% recovery by SP3 and 31.1% recovery by SP4). (B) Heatmaps (top) and average histone modification profiles (bottom) for cluster CS2, together with those for SE3, which recovers CS2 (61.2% recovery by SE3). (C) Heatmaps (top) and average histone modification profiles (bottom) for cluster CS8, together with those for SE4, which recovers CS8 (26.5% recovery by SE4). The color of each curve is indicated by the color of the cluster label.

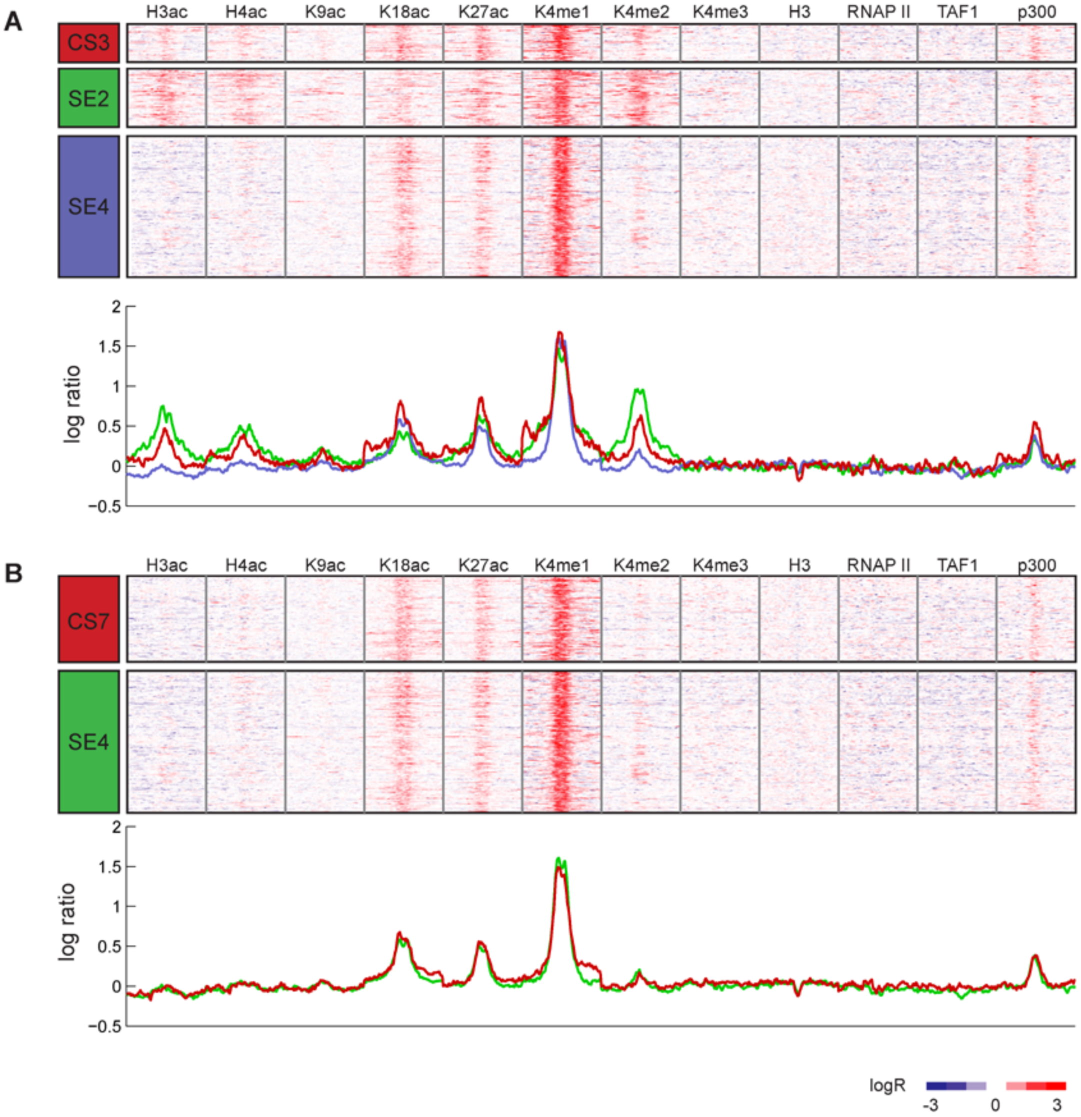


Figure 5-5: Comparison of ChromaSig clusters to clusters from Chapter 2, continued.

Heatmaps (top) and average histone modification profiles (bottom) for ChromaSig clusters (A) CS3 and (B) CS7, together with those clusters in Heintzman et al which recover the ChromaSig clusters. Comparisons for CS1-2 and CS8 can be found in Figure 3. Clusters CS4-6 are not recovered by clusters in Chapter 2. The color of each curve is indicated by the color of the cluster label.

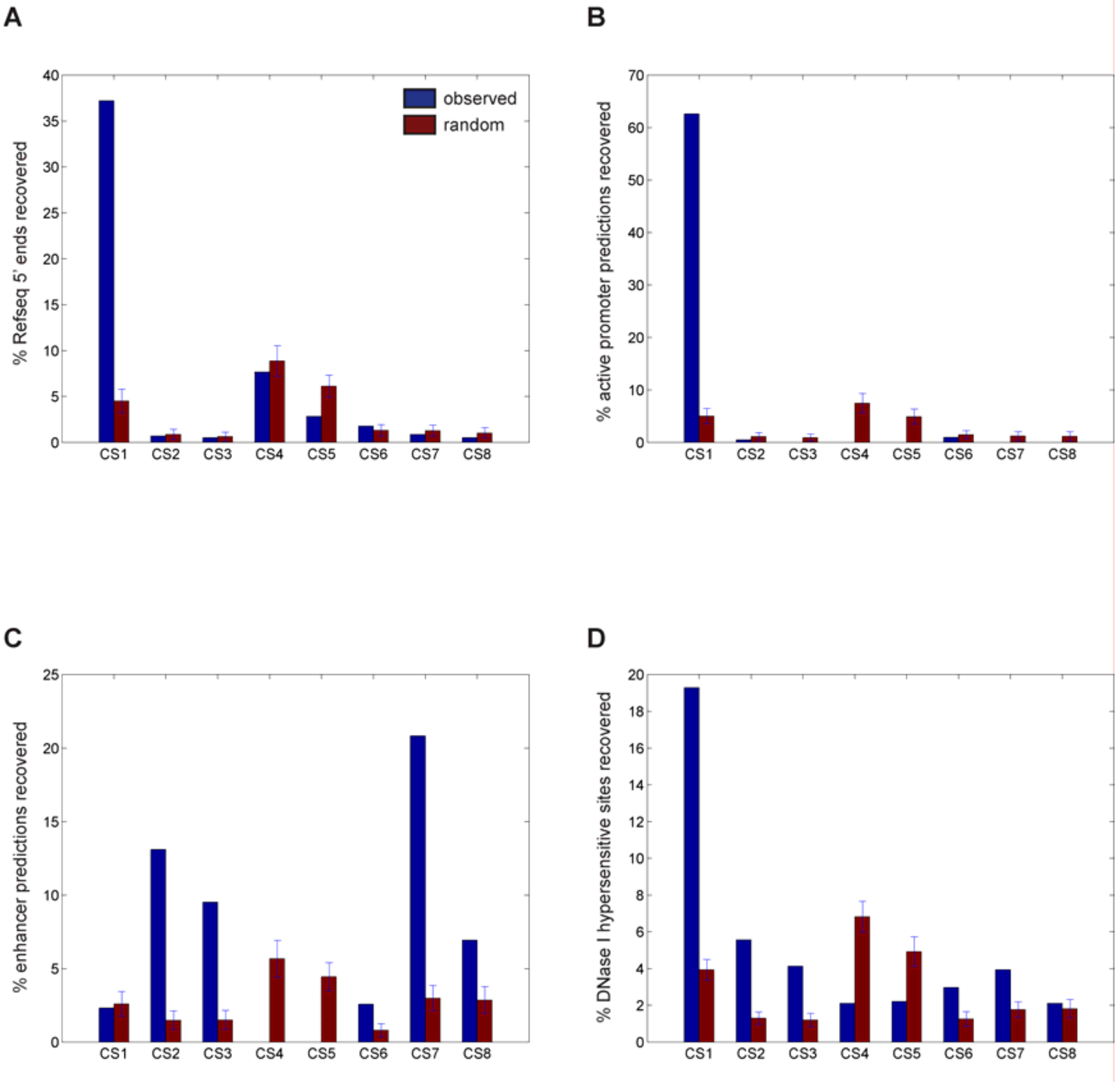


Figure 5-6: Overlap of ChromaSig clusters with known functional sites in the human ENCODE regions.

Percentage of (A) 559 unique Refseq TSSs [24], (B) 198 putative active promoters [15], (C) 389 putative enhancers [15], and (D) 1042 hypersensitive sites [25] that are found within 2.5-kb to the aligned loci, as compared to 100 sets of random sites.

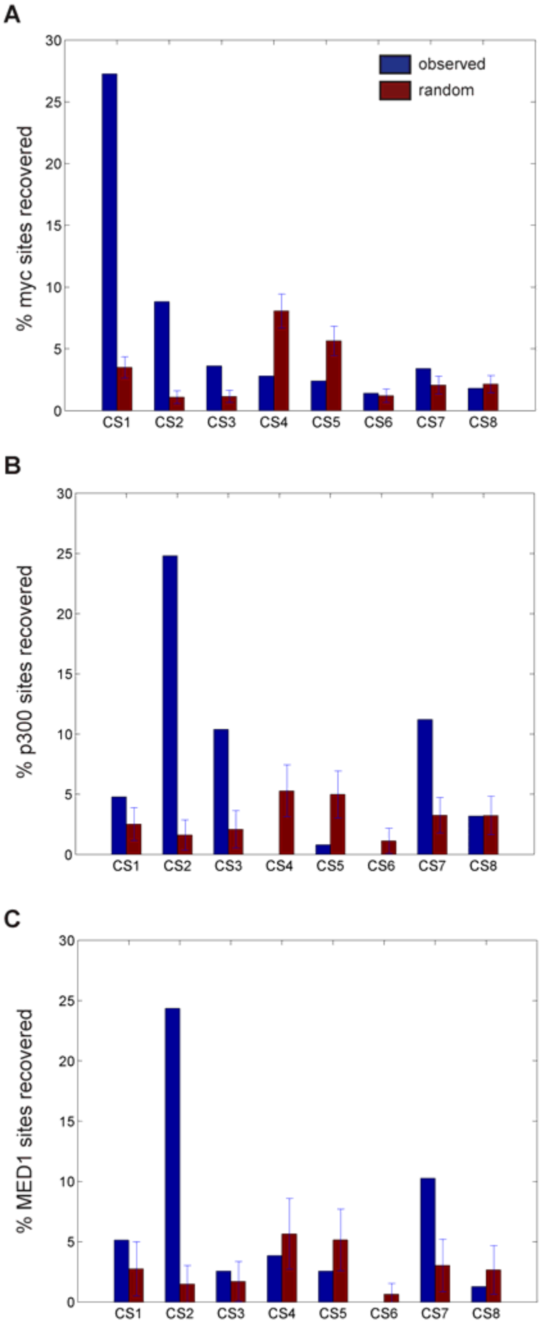


Figure 5-7: Overlap of ChromaSig clusters with transcription factors and coactivators mapped in HeLa cells in the ENCODE regions.

Percentage of (A) 499 c-Myc [36], (B) 125 p300 [15], and (C) 78 MED1 [15] binding sites found within 2.5-kb to aligned clusters, as compared to 100 sets of random sites.

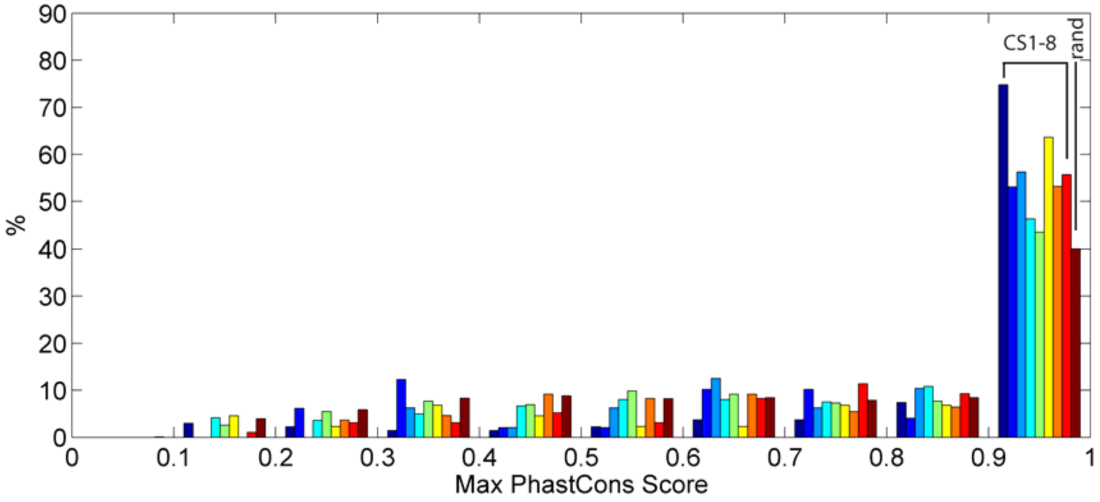


Figure 5-8: ChromaSig clusters are evolutionarily conserved.

Distribution of maximum PhastCons conservation scores [26] over a 1-kb window centered at the aligned loci, as compared to 10000 random sites.

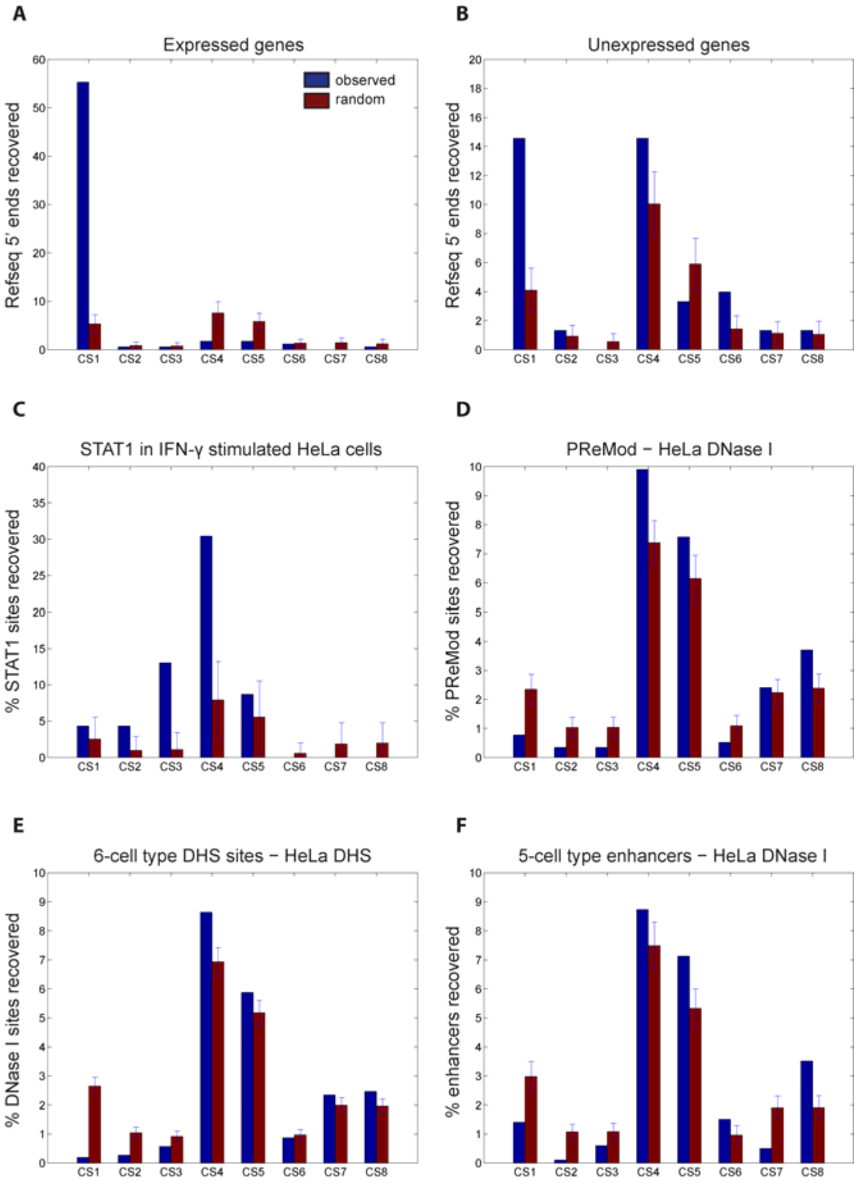


Figure 5-9: Clusters CS4-5 contain regulatory elements.

Percentage of the (A) promoters from expressed genes, (B) promoters from unexpressed genes, and (C) STAT1 binding sites in IFN-γ treated HeLa cells that are within 2.5-kb of the aligned loci. Percentage of (D) PReMod sites [27], (E) combined 6-cell type HS sites [25,28], and (F) combined 5-cell type enhancer predictions distal to HeLa HS sites that are within 2.5-kb of aligned loci. All overlaps are compared to 100 sets of random sites.

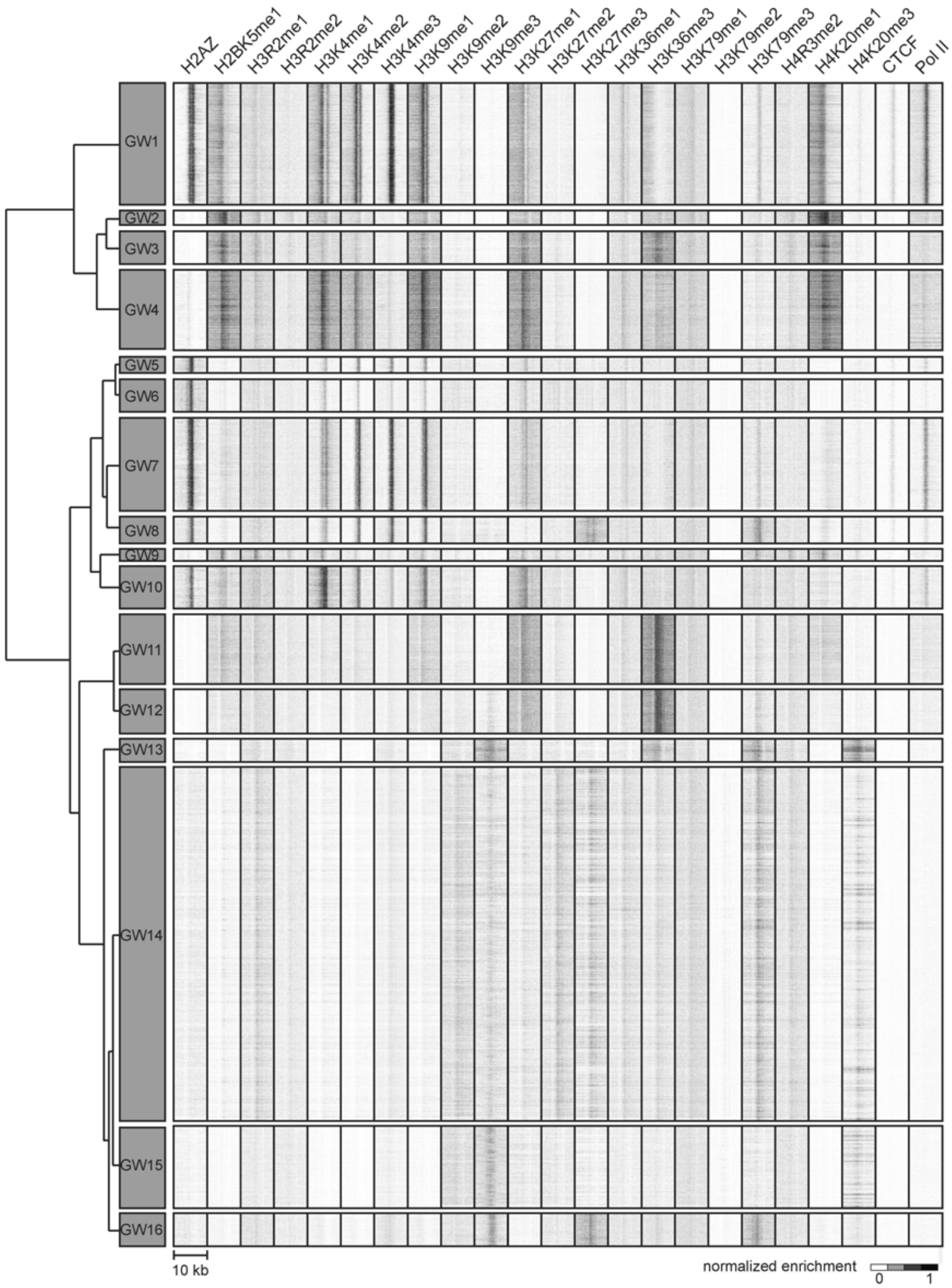


Figure 5-10: ChromaSig clusters recovered from 21 histone marks mapped by ChIP-Seq in CD4+ T cells genome-wide.

ChromaSig recovers 16 clusters spanning 49340 genomic loci. Each cluster is represented by a heatmap summarizing ChIP-Seq enrichment for all loci in the cluster. The window size for each mark is 10-kb. To organize these clusters visually, I use hierarchical clustering with a Euclidean distance metric (left).

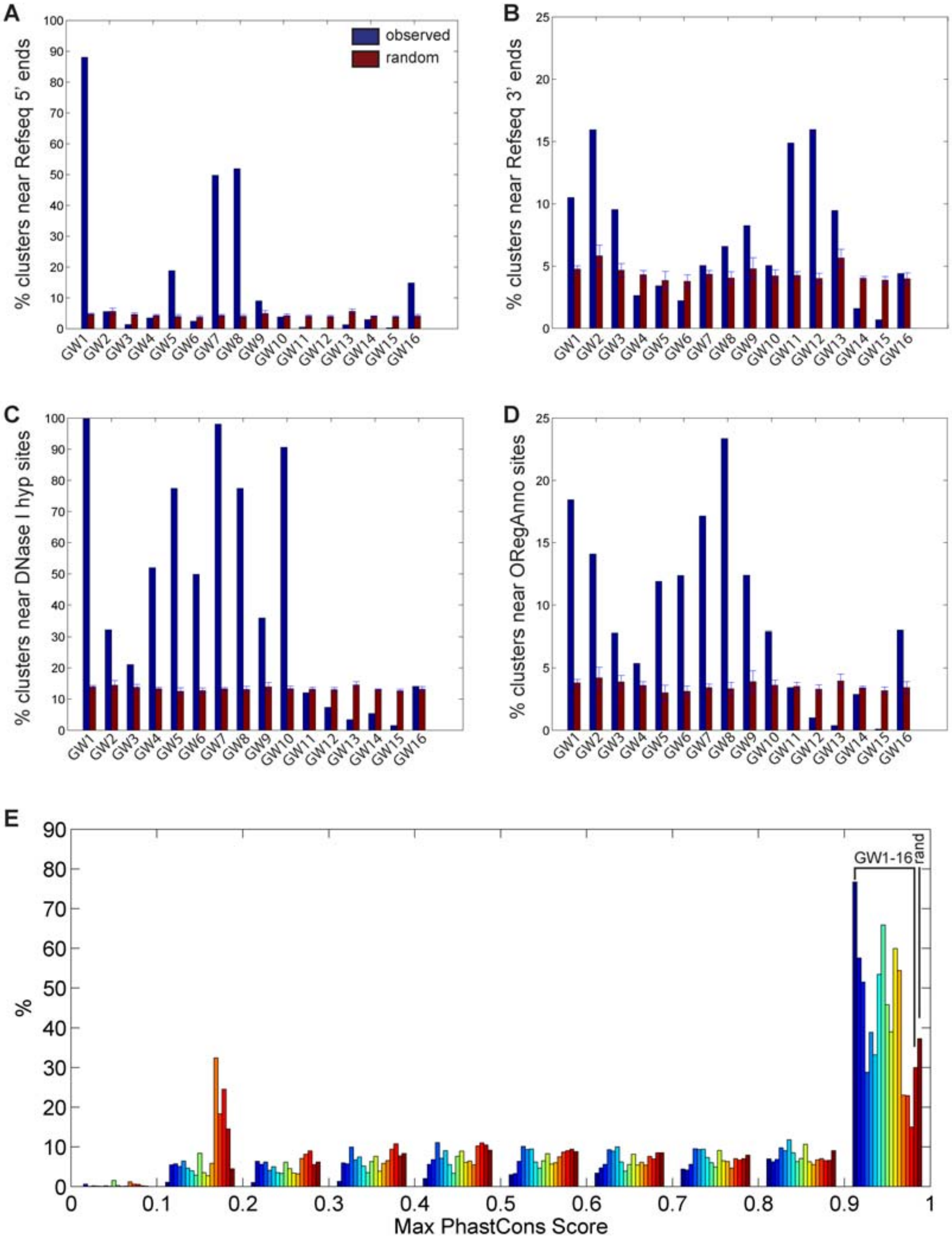


Figure 5-11: Overlap of genome-wide clusters with known annotations.

Percentage of each cluster within 2.5-kb of (A) 21211 Refseq 5′ ends [24], (B) 20754 Refseq 3′ ends [24], 95709 DNase I hypersensitive sites mapped in CD4+ T cells [30], and (D) 21959 regulatory sites from the ORegAnno database [31], as compared to 100 sets of random sites. (E) Distribution of maximum PhastCons scores [26] over a 1-kb window centered at ChromaSig aligned sites, as compared to 10000 random sites.

Chapter 6 : Discovery and annotation of functional chromatin signatures in the human genome

Abstract

Transcriptional regulation in human cells is a complex process involving a multitude of regulatory elements encoded by the genome [1]. Recent studies have shown that distinct chromatin signatures mark a variety of functional genomic elements, and that subtle variations of these signatures mark elements with different functional specificities [2,3,4]. To identify novel chromatin signatures spanning lesser-studied loci, I apply a *de novo* pattern finding algorithm [4] to genome-wide maps of histone modifications [5]. I recover previously known chromatin signatures associated with promoters and enhancers [6]. I also observe several distinct chromatin signatures with strong enrichment of H3K36me3 marking exons. Closer examination reveals that H3K36me3 is found on well-positioned nucleosomes specifically at exon 5' ends, and that this modification is a global mark of exon expression that also correlates with alternative splicing. Additionally, I observe strong enrichment of H2BK5me1 and H4K20me1 at highly expressed early exons but weaker enrichment at late exons, in contrast to the opposite distribution of H3K36me3-marked exons. Finally, I also recover frequently occurring chromatin signatures displaying strong local enrichment of repressive (H3K27me3) and heterochromatic (H3K9me2 and H3K9me3) histone modifications that mark repeat-rich regions of the genome for distinct modes repression. Together, these results highlight the rich amount of information encoded in the human epigenome and underscore its value in studying gene regulation.

Introduction

The genome sequence is a static entity defining the possible output of every cell type in the human body. In contrast, chromatin structure dynamically dictates which genomic regions are functional in a particular cell, how they function, and when. Over 100 different histone modifications are known to exist, and a single nucleosome can contain many modifications. While the number of possible combinations of histone modifications exceeds the number of nucleosomes in the human body, to date only a small number of histone modification patterns have been discovered.

Several classes of regulatory elements are marked by different chromatin signatures. Notably, I recently observed distinct chromatin signatures at active promoters and enhancers [6]. Importantly, not only are these signatures descriptive of both elements, but they are also predictive. Numerous studies have also observed that slight variations in chromatin signatures can distinguish different specificities of the same regulatory element [2,4]. For example, active promoters are generally marked by H3K4me3, repressed promoters by H3K27me3, and poised promoters by both marks [2]. Similarly, different chromatin signatures mark enhancers bound by different classes of transcription factors and co-activators [4]. In a more recent study, several chromatin signatures were also found at promoters and enhancers using genome-wide chromatin maps [3].

These observations prompted me to systematically examine the chromatin signatures that exist in known and putative regulatory elements in the human genome. The goal is to explore whether other frequently occurring chromatin signatures exist, and whether there are functional consequences of these signatures. Focusing on 21 histone modifications mapped in CD4+ T cells [5], I find only a handful of distinct chromatin signatures at promoters, and that they correlate with gene activity. I then examine signatures spanning almost 50,000 regions in the human genome that are distal to known regulatory sites. I recover 7 distinct chromatin signatures, several containing enrichment of H3K36me3 that has been recently linked to marking exons [7]. Upon further inspection, I observe that H3K36me3 is most strongly enriched at a well-positioned nucleosomes located at the 5’ ends of exons. Examination of exonic expression data reveals that stronger enrichment of H3K36me3 correlates with increasing exon activity, in a manner consistent with alternative splicing of exons. I also recover two distinct chromatin signatures rich in heterochromatic and repressive histone modifications marking distinct regions of the genome that are likely associated with different modes of repressing the genome.

Results

Chromatin signatures distinguish different classes of expressed promoters

Loci sharing common regulatory functions may share similar chromatin signatures. To systematically identify chromatin signatures genome-wide, I examine different classes of regulatory loci in turn. These loci may contain chromatin signatures, but they may not be aligned or even oriented in the same direction. I apply an unbiased clustering and alignment method called ChromaSig [4] to find consistent chromatin signatures spanning these loci while simultaneously aligning and orienting their enrichment profiles, focusing on histone modification maps profiled recently in CD4+ T cells [5]. As a proof of principle that this approach yields biologically significant results, I first studied promoters.

While chromatin signatures at promoters have been studied extensively, we still do not have a complete picture of all the distinct, commonly occurring chromatin signatures spanning all promoters. As such, our understanding of how different signatures relate to gene expression is incomplete. To address this, I apply ChromaSig to the chromatin modifications near Refseq promoters [8]. I recover 14 clusters spanning 18,533 promoters (Figure 6-1). Promoters in the same cluster share a consistent chromatin signature, and the chromatin signatures of different clusters are distinct. Some chromatin signatures at promoters are strikingly different. For example, P4 contains strong enrichment for various H3K4 methylations while P2 lacks these modifications. Other chromatin signatures are only subtly different. For example, P9 and P12 contain enrichment for the same chromatin modifications, but the pattern of enrichment is different, with P12 containing enrichment over a noticeably wider region. It is also evident that there is a high level of redundancy of histone modifications at promoters. Notably, H2AZ, H3K4me1, H3K4me2, H3K4me3, and H3K9me1 are either all found together or all absent together at promoters.

Previous studies have shown that there are at least three different classes of chromatin signatures at promoters: actively transcribed promoters marked by H3K4me3 but not H3K27me3, repressed promoters with H3K27me3 but not H3K4me3, and bivalent promoters having both these marks [2]. ChromaSig recovers all three of these previously known chromatin signatures: P8-14 have the active chromatin signature, P2 contains the repressed chromatin signature, and P4 has the bivalent signature.

Next, I wondered if different signatures correspond to different gene expression activities. On the basis of gene expression [9], I observe essentially three super-classes of promoters: P1-7 are generally repressed in CD4+ T cells, P9,11,13,14 show intermediate expression, and P8,10,12 are most highly expressed (Figure 6-1). Promoters with repressed and bivalent chromatin signatures are generally lowly expressed, while promoters with active chromatin signatures have intermediate to high levels of gene expression. Consistent with the high expression levels, P8,10,12 also contain the most enrichment of the elongation chromatin mark H3K36me3 and H4K20me1 (Figure 6-1) [5,10]. Together, these results show that ChromaSig can reliably detect distinct chromatin signatures at promoters with likely distinct functional specificities.

Distinct chromatin signatures at known regulatory elements

While transcriptional regulation occurs at the level of promoters, it is also clear that the action of promoter-distal regulatory elements is essential to controlling gene expression [1]. Like promoters, the activity of these regulatory elements is likely dependent on chromatin structure. To determine what chromatin signatures exist at distal regulatory elements, I apply ChromaSig to several classes of regulatory elements in turn: enhancers, insulators, Refseq 3’ ends, and DNase I hypersensitive sites.

Enhancers: When active, enhancers are bound by transcription factors and co-activators to increase gene expression at promoters [11,12]. Previously, I observed that enhancers are marked by strong enrichment of H3K4me1 and weak if any enrichment of H3K4me3, allowing development of a computational strategy to identify enhancers using this chromatin signature [6]. Applying this method to the genome-wide profiles of H3K4me1 and H3K4me3 in CD4+ T cells [5], I predict 32,237 promoter-distal enhancers. To validate these enhancer predictions, I compare to two hallmarks of enhancers: DNase I hypersensitivity and sequence conservation. Almost half (44.5%) of the enhancer predictions are within 1-kb of a DNase I hypersensitive site [13], and about three-fourths of the predictions are recovered by some combination of hypersensitivity and conserved DNA sequence elements from the PhastCons database [14].

I have previously observed in the ENCODE regions that different variations of chromatin modifications exist at enhancers [15]. To assess if this is true on a global scale, I apply ChromaSig to align and cluster these predicted enhancers over the entire panel of chromatin modifications. This reveals 11 distinct chromatin signatures, all of which contain stronger enrichment for H3K4me1 than H3K4me3 (Figure 6-10). Like promoters, there also appears to be much redundancy of chromatin modifications at enhancers. For example, all chromatin signatures generally share enrichment for H2BK5me1, H3K4me2, H3K9me1, H3K27me1, and H3K36me1. Interestingly, the chromatin marks H2A.Z and H4K20me1 appear to be inversely correlated: E1-5 are enriched in H2A.Z but not H4K20me1, E6 has enrichment of both marks, and E7-11 are enriched in H4K20me1 but not H2A.Z.

Insulators: CTCF is an insulator binding protein in mammals, and when bound prevents enhancers from interacting with promoters, thereby preventing activation [16]. Barski et al mapped CTCF binding in CD4+ T cells [5], and application of the Model-based Analysis of ChIP-Seq (MACS) peak finder reveals 27,110 CTCF binding sites genome-wide [17]. To focus on novel chromatin signatures, I apply ChromaSig to the 17,328 CTCF sites distal to Refseq TSSs and predicted enhancers, revealing seven distinct signatures (Figure 6-11). The only consistent feature of CTCF binding sites is enrichment of H2A.Z, consistent with previous observations [18]. However, unlike the patterns observed at promoters and enhancers, enrichment for other chromatin marks at CTCF binding sites is generally weak, suggesting that the remaining panel of chromatin marks do not functionally compliment CTCF. The exceptions are C4 and C5, which contain enrichment of H3K4me3 and RNA Pol II, and may be promoters not in the Refseq database.

Refseq 3' ends: Transcription of pre-mRNA stops at the 3' end of the gene. To find chromatin signatures at this genomic feature, I apply ChromaSig to 16,703 Refseq gene 3' ends distal to Refseq 5' ends [8]. I recover 12 distinct chromatin signatures. Like CTCF binding sites, enrichment of chromatin marks at Refseq 3' ends is generally weak. In agreement with Barski et al [5], the most consistent feature found at the majority of 3' ends is enrichment of H3K36me3, found in T1-7 (Figure 6-12). However, chromatin signatures at 3' ends are not as well aligned as those at promoters, suggesting that these chromatin signatures may occur at some other genomic feature near 3' ends, or alternatively that the 3' ends are not as well annotated as promoters.

DNase I hypersensitive sites: Recently, Boyle et al mapped nearly 100,000 DNase I hypersensitive sites genome-wide in CD4+ T cells using DNase-Seq [13]. Since DNase I hypersensitivity is a hallmark for active regulatory loci, I expected to find chromatin signatures at these sites. Applying ChromaSig to the 31,824 DNase I hypersensitive sites distal to Refseq TSSs, predicted enhancers, and CTCF binding sites, I recover 13 clusters (Figure 6-13). D1-D2 are only enriched in H3K27me1 and H3K36me3, resembling gene 3' ends. Several signatures D3-10 display characteristic enrichment of H3K4me1/2/3 which I have observed at promoters and enhancers. These may be novel promoters or enhancers missed by the enhancer prediction method. For example, D3,6,9,10 are clusters with the strongest enrichment of H3K4me3, and 31.2% of these loci are recovered by multiply-occurring CAGE tags [19], an almost 4-fold enrichment as compared to an expected 7.9% over random loci. The majority of DNase I sites D11-13 contain no noticeably strong enrichment of any chromatin mark, indicating that either these genomic elements are enriched in other chromatin modifications not profiled by Barski et al [5], or that they are not enriched in any chromatin modifications.

Distinct chromatin signatures distal to known regulatory elements

Having observed chromatin signatures at important regulatory elements including promoters and enhancers, I next asked if other chromatin signatures exist that mark loci distal to known regulatory elements. By definition, places in the genome with chromatin signatures contain strong enrichment of various histone modifications. I identify 85,318 loci with strong ChIP enrichment of histone modifications, 50,183 of which are distal to promoters [8], gene 3' ends [8], DNase I hypersensitive sites [13], CTCF binding sites [5], and sites containing an enhancer chromatin signature [6]. Applying ChromaSig to these sites, I recover 7 frequently-occurring chromatin signatures N1-7 spanning 47,874 loci (Figure 6-2). Loci in the same cluster share the same consistent chromatin signature, and each cluster is defined by a distinct chromatin signature. The recovered signatures are also distinct from the previously defined H3K4me3-rich promoter and H3K4me1-rich enhancer signatures [2,6]. Compared to chromatin signatures from randomly aligned and oriented loci, the chromatin signatures observed are significantly better aligned than expected by chance, with p-values ranging from 10^{-18} to $<10^{-300}$ (Table 6-1).

The strongest chromatin feature of these clusters is H3K36me3, known to mark the 3' ends of genes [5] and more recently exons [7], and is enriched at N1, N2, and N4. The largest clusters recovered, N5 and N6, both contain enrichment of known repressive chromatin marks [5] including H3K27me2, H3K27me3, and H3K79me3. However, N5 is also enriched in H3K9me2 and H3K9me3, which are known to mark heterochromatic regions of the genome [5].

Chromatin signatures mark exon 5' ends

I have found loci sharing frequently-occurring chromatin signatures, but it is unclear what function, if any, ties loci sharing the same chromatin signature together. To get clues to the possible function of these sites, I compare to known annotations.

H3K36me3, which has been associated with elongating RNA polymerase II, is known to be enriched within the body of transcriptionally active genes [20,21], notably at the 3' ends [5]. But since all the clustered loci are distal to gene 3' ends, the H3K36me3-rich clusters must be marking another genomic feature. Noticing that the vast majority of loci in N1-4 are intragenic (Figure 6-14), I ask if these sites are biased towards exons or introns. I observe that 57.9% of N1 sites and 63.8% of N2 sites are either inside exons or within 1-kb of exon ends, while at random only 26% of the genic regions of the genome match these criteria. To see if H3K36me3 marks exons, I examine the enrichment of this chromatin mark at exons (Figure 6-8**a-d**). To examine only those exons unambiguously marked by a chromatin signature, I only consider an exon if it is the only exon within 1-kb of a cluster locus. I observe a striking enrichment of H3K36me3 at the 5' ends of exons unambiguously marked by N1, N2, and N4. This enrichment decreases sharply upstream of the 5' end, but more gradually into the exon body. This observation also holds for exons larger than 1-kb (Figure 6-9), indicating that the result is not biased by the relatively small exon sizes in the human genome [22]. These results suggest that the clusters with strong H3K36me3 enrichment mark exon 5' ends, consistent with observations made by others [7].

H3K36me3 is a global marker of exon activity

Having observed H3K36me3 at a handful of exons, I next ask if this chromatin mark is a global indicator of exon activity. Profiling H3K36me3 at a catalog of more than 250,000 distinct exons

[23], I observe peaks of enrichment at the majority of human exons in CD4+ T cells (Figure 6-3a). In the direction of transcription, H3K36me3 enrichment increases sharply at the 5’ end of the exon, and decreases more gradually in the body of the exon, in agreement with my previous observations. In contrast, neighboring introns show no such chromatin signature (Figure 6-3**,** Figure 6-15). The presence of this chromatin mark also correlates strongly with exonic expression (Figure 6-3), with highly expressed exons having more H3K36me3 enrichment than lowly or moderately expressed exons. Altogether, these results suggest that H3K36me3 is a global marker of exon activity.

Stable nucleosome structure at exon 5’ ends

In ChIP-Seq experiments, short directional reads are sequenced directly upstream and downstream of the genomic DNA bound by the protein of interest, allowing clear distinction between sense and antisense reads. This information can be used to offer unprecedented resolution of *in vivo* binding locations of the protein [17,24]. I will use this information to more finely resolve nucleosome structure at exons. Looking at the distribution of H3K36me3 tags at all human exons, I see that reads on the sense strand are highly enriched at the 5’ ends of exons, decreasing gradually towards the 3’ exon end (Figure 6-3b). In contrast, antisense reads are distributed in the opposite way, being sharply enriched near the 3’ ends of exons.

From this information alone, it is difficult to conclude if the nucleosomes harboring H3K36me3 at exons are more fixed towards the exon 5’ or 3’ ends. Further confounding this issue is the fact that a typical nucleosome wraps between 145 and 147 bp of DNA [25], which is roughly the same size as the average human exon at 145 bp [22]. To resolve this issue, I next examine the same distribution of reads, but focusing on exons larger than 1-kb (Figure 6-3b). Again, I observe a clear enrichment of sense strand reads at 5’ exon ends. However, I also find that the highest enrichment of antisense reads is clearly inside the exon, slightly downstream of the 5’ end, while there is no enrichment at exon 3' ends. Thus, I conclude that the nucleosomes harboring H3K36me3 are more fixed towards the exon 5' end.

H3K36me3 correlates with alternative splicing

As H3K36me3 at the 5' ends of exons is a global mark of exon activity, I next wondered if the presence of this mark correlates with alternative splicing. To examine alternative splicing on a global scale, I focused on a list of 13,434 exons known to be alternatively spliced as cassette exons (UCSC Genome Browser "knownAlt" track) [26]. I examined two sets of transcripts. The "spliced in" set consists of cassette exons expressed at levels similar to neighboring upstream and downstream exons ($|\Delta \text{expr}| = 0.5$), and thus are likely to be included in a mature transcript. In contrast, the "spliced out" set consists of cassette exons expressed at lower levels than both upstream and downstream exons, and are likely excluded from the mature transcript ($\text{expr}_{\text{up,down}} - \text{expr}_{\text{alt}} > 1$). For spliced in exons, I observe that the enrichment of H3K36me3 increases gradually from upstream to alternatively spliced to downstream exons (Figure 6-4a), consistent with previous results showing a 3' bias in this chromatin mark [5]. However, H3K36me3 is noticeably depleted at spliced out exons as compared to both upstream and downstream exons (Figure 6-4b). These results suggest that, on a global scale, the presence of H3K36me3 at alternatively spliced exons correlates with inclusion of the exon in transcripts.

H2BK5me1 and H4K20me1 mark highly expressed, early exons

The initial scan revealed several classes of chromatin signatures marking exons, the largest of which are N1 and N2. Both of these contain enrichment for H3K36me3, but N1 contains stronger enrichment for H2BK5me1 and H4K20me1. This latter modification is known to be localized both at promoters and intragenic regions downstream of the promoters, with enrichment fading in the gene body [5]. These observations raise the possibility that exons marked by N1 are early exons closer to promoters while N2 are late exons closer to the 3' ends of genes. To test this hypothesis, I sorted the highly expressed exons above by distance to the transcription start site, and visualized the enrichment of histone modifications (Figure 6-5). As expected, the exons closest to the transcription start site (TSS) are all highly enriched in promoter modifications including H3K4me1, H3K4me2, and H3K4me3. In addition to H3K36me3, early exons not having these promoter marks are also enriched with H2BK5me1 and H4K20me1. This enrichment fades with increasing distance from the TSS. In contrast, H3K36me3 enrichment increases with increasing distance from the TSS, consistent with the above results (Figure 6-4a) and previous observations [5]. These results provide additional evidence for various chromatin modifications marking distinct exons in the human genome.

Distinct classes of repressive chromatin signatures

In addition to chromatin signatures N1-4, ChromaSig also identifies two chromatin signatures N5-6 having strong enrichment of repressive histone modifications (Figure 6-2). These two chromatin signatures are distinct, with N5 having stronger enrichment of heterochromatic marks H3K9me2 and H3K9me3. This subtle difference prompted me to ask if these signatures mark distinct regions of the genome. Indeed, I find that only 23.3% of N5 loci are intragenic, a notable depletion over the expected value of about 40% (Figure 6-14). In contrast, N6 loci are closer to the expected value at 36.3% intragenic.

Sequence analysis suggests that the sequences underlying N5 and N6 are distinct. First, I compare to the PhastCons database containing over 2 million conserved elements in the human genome conserved over 28 mammalian genome s[14]. I find that N5 loci are significantly depleted of conserved elements (p = 7.12E-182) while N6 is significantly enriched (p = 2.09E-26) (Figure 6-6a). Given that heterochromatic histone modifications have been known to mark repetitive regions of the genome [27] which are highly lineage-specific [22], the low conservation of N5 loci may be explained by enrichment for repetitive sequences. To test this hypothesis, I use RepeatMasker [28] to define repetitive bases within ±1-kb from each locus in N5-6. Indeed, 49.1% of N5 bases are repetitive, as compared to 32.1% of N6 bases (Figure 6-6b), suggesting that these two clusters may harbor different classes of sequences. To pursue this further, I next ask if the classes of repeats found in N5 are different from those found in N6. Counting the repetitive elements found within ±1-kb of each locus (Figure 6-6c-d), I find that N5 is significantly enriched for long terminal repeats (LTR) ($p < 1E-300$, Z-score = 39.7), while N6 is neither enriched nor depleted. For the SINE family of repeats, while both clusters are significantly depleted in Alu repeats ($p_{N5} < 1E-300$, $Z_{N5} = 81.5$; $p_{N6} = 4.76E-245$, $Z_{N6} = 33.4$), only N6 is notably enriched in MIR repeats ($p = 2.31E-177$). Similarly, L2 LINE repeats and simple repeats are notably more enriched in N6 loci than N5 loci. These results suggest that N5 and N6 have different genic distributions and mark distinct sequences of the genome.

N5 and N6 mark different domains of gene repression

I next examine whether the different genic distributions and sequence preferences of N5 and N6 relate to gene expression. It is thought that the genome is organized into different domains of transcriptional activity, with the insulator binding protein CTCF defining the boundaries of these domains [16,29]. Therefore, I partition the genome into CTCF-defined domains and determine the enrichment of N5 or N6 loci in these domains as a function promoter activity. The distributions of N5 and N6 enrichment are significantly different ($p = 5.95E-26$) (Figure 6-7a): N5 is more enriched than N6 in domains containing the most repressed genes (log expression <4), while domains containing repressed by slightly more expressed genes (log expression between 5 and 6) have higher enrichment of N6 loci than N5 loci. For moderately and highly expressed genes (log expression >6), the enrichment of both N5 and N6 loci are depleted relative to random. While it is not surprising that N5 and N6 are enriched near genes with low expression since they are both enriched in repressive histone modifications, it is remarkable that these two chromatin signatures mark distinctly different populations of lowly expressed genes. One possibility is that N5 and N6 are present in different compartments of the nucleus. To test this, I examine the localization of these loci in lamina-associated domains (LADs), previously mapped in fibroblast cells and known to contain repressed genes and gene deserts. Indeed, more than 60% of N5 loci are in LADs, compared to only 37.4% for N6 loci (Figure 6-7b). This 62% difference in enrichment is all the more surprising given that the LADs were mapped in a different cell type. Taken together, these results suggest that N5 and N6 mark distinct domains of gene expression that may be explained by their enrichment in different nuclear compartments.

Discussion

In this study, I survey the global landscape of commonly occurring chromatin signatures in the human genome. I recover known signatures at well-studied elements such as promoters and lesser-studied elements including enhancers. In addition, I find 7 distinct signatures spanning 47,874 genomic loci distal to known regulatory elements. In agreement with a previous study [7], I observe chromatin signatures marking exons, but show at a higher resolution that the 5' ends of exons are specifically modified by H3K36me3. Furthermore, I show that the enrichment level of this mark directly correlates with exonic expression, a result that had only been implied before. In addition, I recover two distinct chromatin modifications N1 and N2 marking exons in the genome-wide scan. While both are enriched in H3K36me3, N1 is uniquely enriched in H2BK5me1 and H4K20me1, which directly coincides with N1 marking early exons and N2 marking late exons.

Chromatin modifications have long been implicated in marking different transcriptional domains of the genome. For example, H3K4me3 is canonically known to mark actively transcribed promoters[2,20,30,31], while H3K27me3 is present at repressed promoters[2]. In contrast, H3K36me3 is enriched throughout the coding regions of actively transcribed genes[30,32]. These results additionally implicate chromatin modifications in regulating splicing, a process until recently thought to be decoupled from transcription both physically and temporally. In yeast, H3K36me3 is deposited by Set2, which is associated with the elongation form of RNA polymerase [33,34]. The observation that H3K36me3 marks exons, a part of gene structure in the realm of splicing rather than transcription, implies that chromatin structure play important roles in regulating splicing as well as transcription.

A large body of work on splicing regulation has been focused on how sequence-specific proteins binding directly to pre-mRNAs affect splicing [35,36]. But the static and highly degenerate natures of sequence elements associated with splicing leave unanswered the question of how cell-type specific splicing is achieved. However, recent discoveries physically linking RNA polymerase to the splicing machinery has shifted attention to the roles of the transcription machinery in regulating splicing [35,37]. This has led to two models describing co-transcriptional splicing: a kinetic model and a recruitment model[35]. While both models emphasize spliceosome activity during transcription, neither one fully explains how cell-type specific splicing is achieved. The observations that distinct chromatin signatures are present at exons, and that different signatures are associated with either inclusion or exclusion from mature mRNAs, suggest a role of chromatin state in splicing regulation. One possibility is that the writing and reading of dynamic chromatin signatures may direct splicing events. While this model is attractive, further studies will be necessary to verify this hypothesis.

I also recover several chromatin signatures enriched in repressive and heterochromatic histone modifications marking distinct populations of repetitive elements. Surprisingly, these signatures are associated with different modes of gene repression. One possible explanation for this phenomenon is that N5 loci, which contain heterochromatic chromatin modifications, are more highly enriched in nuclear lamina-associated domains than N6 loci. Thus the N5 chromatin signature may specifically repress LADs, while N6 signature represses other domains. It is possible that these two different levels of domain repression offer a way to control gene induction, with heterochromatin-rich N5 domains being more permanently repressed than heterochromatin-free N6 domains.

These results show that studying the human genome on the basis of chromatin signatures is a viable method to cataloging regulatory elements in the genome in a global, unbiased, and systematic way. Future efforts to map chromatin modifications in the human genome may allow us to define more chromatin signatures marking novel regulatory elements or different functional specificities of known regulatory elements.

Methods

Data normalization

Genome-wide distributions of histone modifications were obtained from Barski et al [5]. As in Chapter 5, I filtered reads for uniqueness and redundancy, partitioned the genome into 100-bp bins, and binned these reads. As the number of reads for each mark was highly variable, normalization was necessary to facilitate comparison. For each bin *i* and mark *h*, I normalized the number of reads in this bin $x_{h,i}$ by:

$$x_{h,i}^{norm} = \frac{1}{1 + e^{-(x_{h,i} - median(x_h))/std(x_h)}}$$

Finding ChIP-enriched loci distal to known regulatory elements.

I identified regions of width 2-kb containing enrichment for histone modifications strongly deviating (p = 0.0001) from the background distribution of ENCODE regions. I removed any enriched locus closer than 2.5 kb to another enriched locus to remove redundancy. I then removed loci within 2.5

kb to regulatory loci at promoters [8], gene 3' ends [8], CTCF binding sites [5], DNase I hypersensitive sites [13], and sites having an enhancer chromatin signature [6].

Finding chromatin signatures

I searched for chromatin signatures of size 4-kb within a region ±1-kb around the ChIP-enriched loci, using ChromaSig [4] with a background prior $p_{2A} = 0.01$ and a standard deviation factor $\sigma_{another} = 1.75$. To focus only on the most frequently-occurring chromatin signatures, I analyzed only those output clusters with at least 500 loci and an average normalized enrichment greater than 0.25 .

Chromatin signature significance

For a given cluster of size N, I defined the motif $m_{h,i}$ to be the mean normalized enrichment of the aligned loci at a specified position *i* for mark *h*. Well-aligned motifs have higher values of enrichment. For each motif, I computed the score:

$$S = \sum_{h} \max_{j} \left(m_{h,j} \right)$$

Higher values of S indicate more significant motifs. To assess significance of observing a motif spanning N loci with score S or greater, I randomly sampled 100 sets of clusters with random alignment offsets (within ±1 kb of the aligned sites) and orientations, computed the S for each random set, and modeled the random distribution of S scores as a normal. I performed this randomization either within loci in the same cluster as the original motif or over loci from all clusters.

Genome annotations

Genome annotations were downloaded from the UCSC Genome Browser [26], human genome Build 36.1 (hg18 assembly). Gene definitions were given by the Refseq Genes [8] track. Alternatively spliced exons were defined by entries in the “Alt Events” track labeled as “Cassette Exons”. The list of human loci conserved in a 28-way alignment with placental mammals was defined by the phastConsElements28wayPlacMammal table[14]. Repeat definitions were given by the RepeatMasker track [28], and lamina-associated domains mapped in Tig3 human lung fibroblasts [38] were defined by the “NKI LADs” track.

Catalogs of regulatory elements

I obtained a list of 27,110 CTCF sites by running the Model-based Analysis of ChIP-Seq [17] software with the default p-value cutoff of 1E-5. I used normalized H3K4me1 and H3K4me3 profiles to predict enhancers as in Chapter 3. ROC analysis indicated that using a p-value cutoff of 0.1 gives optimal recovery of DNase I hypersensitive sites [13], corresponding to 32,237 predicted enhancers at least 2.5-kb from Refseq TSSs.

Expression data

Transcript and exon expression data were measured in CD4+ T cells by Crawford et al [9] (GEO accession GSE4406) and Oberdoerffer et al [39] (GEO accession GSE11834), respectively.

Randomization.

To determine enrichment for a given cluster, I compared to 100 random clusters. Each random cluster contains the same number of loci as the original cluster and follows the same chromosomal distribution. Random sampling is limited to bins containing ChIP-Seq reads.

Statistical tests

To assess significance of overlap with known genome annotations, I assume that the overlap statistics for 100 random clusters follows a normal distribution. To assess significance of exon inclusion for marked versus unmarked exons, I use a two-sided Wilcoxon rank sum test to compare the median exon expression of the two sets. To assess that N5 and N6 are enriched near different classes of expressed genes, I use the paired two-sided Wilcoxon signed rank test to compare the enrichment profiles.

Acknowledgements

Chapter 6, in full, has been submitted for review in PLoS Computational Biology. Hon, Gary ; Ren, Bing ; Wang, Wei. “Discovery and annotation of functional chromatin signatures in the human genome”. The dissertation author was a primary investigator and author of this paper. The dissertation author performed all the analysis presented.

References

1. Maston GA, Evans SK, Green MR (2006) Transcriptional Regulatory Elements in the Human Genome. Annu Rev Genomics Hum Genet 7: 29-59.

2. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES (2006) A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell 125: 12.

3. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, Zhao K (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet 40: 897-903.

4. Hon G, Ren B, Wang W (2008) ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. PLoS Comput Biol 4: e1000201.

5. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K (2007) High-resolution profiling of histone methylations in the human genome. Cell 129: 823-837.

6. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39: 311-318.

7. Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J (2009) Differential chromatin marking of introns and expressed exons by H3K36me3. Nat Genet.

8. Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33: D501-504.

9. Crawford GE, Davis S, Scacheri PC, Renaud G, Halawi MJ, Erdos MR, Green R, Meltzer PS, Wolfsberg TG, Collins FS (2006) DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. Nat Methods 3: 503-509.

10. Vakoc CR, Sachdeva MM, Wang H, Blobel GA (2006) Profile of histone lysine methylation across transcribed mammalian chromatin. Mol Cell Biol 26: 9185-9195.

11. Blackwood EM, Kadonaga JT (1998) Going the distance: a current view of enhancer action. Science 281: 60-63.

12. Lonard DM, O'Malley BW (2006) The expanding cosmos of nuclear receptor coactivators. Cell 125: 411-414.

13. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE (2008) High-resolution mapping and characterization of open chromatin across the genome. Cell 132: 311-322.

14. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034-1050.

15. Hon GC, Ren B, Wang W ChromaSig: A Probabilistic Approach to Finding Common Chromatin Signatures in the Human Genome. In submission.

16. Gaszner M, Felsenfeld G (2006) Insulators: exploiting transcriptional and epigenetic mechanisms. Nat Rev Genet 7: 703-713.

17. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, Liu XS (2008) Model-based Analysis of ChIP-Seq (MACS). Genome Biol 9: R137.

18. Fu Y, Sinha M, Peterson CL, Weng Z (2008) The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. PLoS Genet 4: e1000138.

19. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest AR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, Chalk AM, Chiu KP, Choudhary V, Christoffels A, Clutterbuck DR, Crowe ML, Dalla E, Dalrymple BP, de Bono B, Della Gatta G, di Bernardo D, Down T, Engstrom P, Fagiolini M, Faulkner G, Fletcher CF, Fukushima T, Furuno M, Futaki S, Gariboldi M, Georgii-Hemming P, Gingeras TR, Gojobori T, Green RE, Gustincich S, Harbers M, Hayashi Y, Hensch TK, Hirokawa N, Hill D, Huminiecki L, Iacono M, Ikeo K, Iwama A, Ishikawa T, Jakt M, Kanapin A, Katoh M, Kawasawa Y, Kelso J, Kitamura H, Kitano H, Kollias G, Krishnan SP, Kruger A, Kummerfeld SK, Kurochkin IV, Lareau LF, Lazarevic D, Lipovich L, Liu J, Liuni S, McWilliam S, Madan Babu M, Madera M, Marchionni L, Matsuda H, Matsuzawa S, Miki H, Mignone F, Miyake S, Morris K, Mottagui-Tabar S, Mulder N, Nakano N, Nakauchi H, Ng P, Nilsson R, Nishiguchi S, Nishikawa S, Nori F, Ohara O, Okazaki Y, Orlando V, Pang KC, Pavan WJ, Pavesi G, Pesole G, Petrovsky N, Piazza S, Reed J, Reid JF, Ring BZ, Ringwald M, Rost B, Ruan Y, Salzberg SL, Sandelin A, Schneider C, Schonbach C, Sekiguchi K, Semple CA, Seno S, Sessa L, Sheng Y, Shibata Y, Shimada H, Shimada K, Silva D, Sinclair B, Sperling S, Stupka E, Sugiura K, Sultana R, Takenaka Y, Taki K, Tammoja K, Tan SL, Tang S, Taylor MS, Tegner J, Teichmann SA, Ueda HR, van Nimwegen E, Verardo R, Wei CL, Yagi K, Yamanishi H, Zabarovsky E, Zhu S, Zimmer A, Hide W, Bult C, Grimmond SM, Teasdale RD, Liu ET, Brusic V, Quackenbush J, Wahlestedt C, Mattick JS, Hume DA, Kai C, Sasaki D, Tomaru Y, Fukuda S, Kanamori-Katayama M, Suzuki M, Aoki J, Arakawa T, Iida J, Imamura K, Itoh M, Kato T, Kawaji H, Kawagashira N, Kawashima T, Kojima M, Kondo S, Konno H, Nakano K, Ninomiya N, Nishio T, Okada M, Plessy C, Shibata K, Shiraki T, Suzuki S, Tagami M, Waki K, Watahiki A, Okamura-Oho Y, Suzuki H, Kawai J, Hayashizaki Y (2005) The transcriptional landscape of the mammalian genome. Science 309: 1559-1563.

20. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA (2007) A chromatin landmark and transcription initiation at most promoters in human cells. Cell 130: 77-88.

21. Li B, Carey M, Workman JL (2007) The role of chromatin during transcription. Cell 128: 707-719.

22. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H,

Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.

23. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D (2006) The UCSC Known Genes. Bioinformatics 22: 1036-1046.

24. Kharchenko PV, Tolstorukov MY, Park PJ (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol 26: 1351-1359.

25. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 389: 251-260.

26. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome Res 12: 996-1006.

27. Grewal SI, Moazed D (2003) Heterochromatin and epigenetic control of gene expression. Science 301: 798-802.

28. Jurka J (2000) Repbase update: a database and an electronic journal of repetitive elements. Trends Genet 16: 418-420.

29. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching KA, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenkov VV, Stewart R, Thomson JA, Crawford GE, Kellis M, Ren B (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature.

30. Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, Lee TI, Bell GW, Walker K, Rolfe PA, Herbolsheimer E, Zeitlinger J, Lewitter F, Gifford DK, Young RA (2005) Genome-wide map of nucleosome acetylation and methylation in yeast. Cell 122: 517-527.

31. Liu CL, Kaplan T, Kim M, Buratowski S, Schreiber SL, Friedman N, Rando OJ (2005) Single-nucleosome mapping of histone modifications in S. cerevisiae. PLoS Biol 3: e328.

32. Bannister AJ, Schneider R, Myers FA, Thorne AW, Crane-Robinson C, Kouzarides T (2005) Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. J Biol Chem 280: 17732-17736.

33. Xiao T, Hall H, Kizer KO, Shibata Y, Hall MC, Borchers CH, Strahl BD (2003) Phosphorylation of RNA polymerase II CTD regulates H3 methylation in yeast. Genes Dev 17: 654-663.

34. Rando OJ (2007) Global patterns of histone modifications. Curr Opin Genet Dev 17: 94-99.

35. Lynch KW (2006) Cotranscriptional splicing regulation: it's not just about speed. Nat Struct Mol Biol 13: 952-953.

36. Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB (2004) Systematic identification and analysis of exonic splicing silencers. Cell 119: 831-845.

37. Maniatis T, Reed R (2002) An extensive network of coupling among gene expression machines. Nature 416: 499-506.

38. Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, de Klein A, Wessels L, de Laat W, van Steensel B (2008) Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature 453: 948-951.

39. Oberdoerffer S, Moita LF, Neems D, Freitas RP, Hacohen N, Rao A (2008) Regulation of CD45 Alternative Splicing by Heterogeneous Ribonucleoprotein, hnRNPLL. Science 321: 6.

Figures and Tables

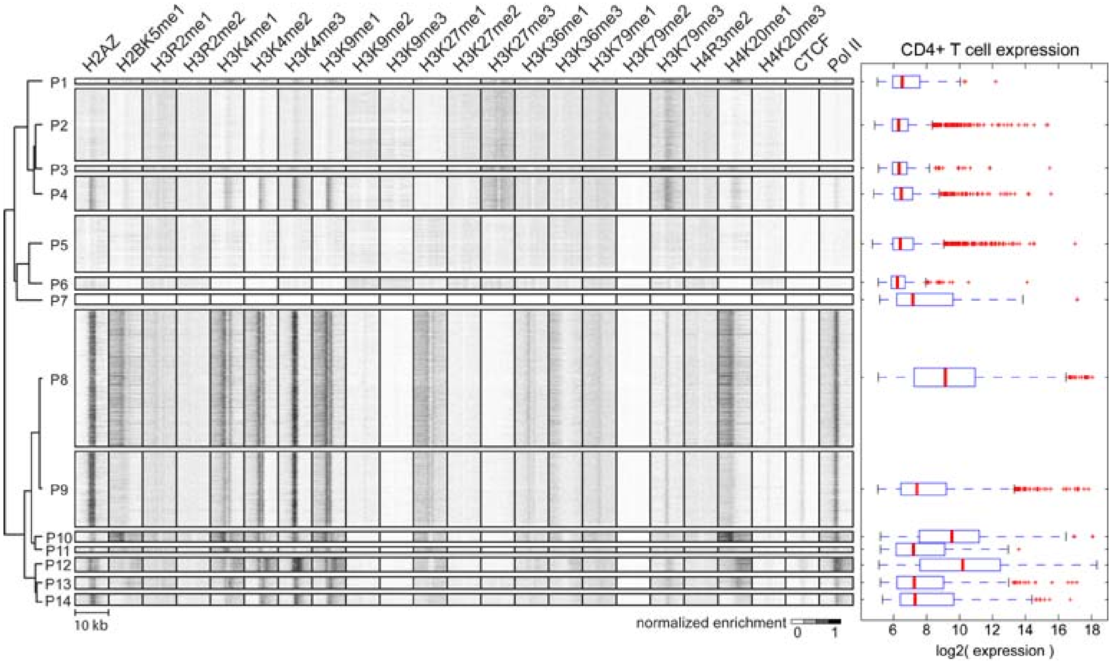


Figure 6-1: Distinct chromatin signatures spanning Refseq promoters.

(left) Applying ChromaSig to the histone modifications near 20,389 Refseq promoters recovers 14 frequently-occurring chromatin signatures spanning 18,533 promoters. The heat map represents the enrichment of H2AZ, 20 histone modifications, CTCF, and RNA polymerase II in the 10-kb region surrounding each promoter. To organize these clusters visually, I performed hierarchical clustering on the average profiles using a Pearson correlation distance metric. (right) Gene expression data for CD4+ T cells measured from a previous study [9], and re-visualized here for the different classes of promoters. Shown are the distributions of gene expression level over promoters with different chromatin signatures. Red horizontal lines indicate the median, the box extends to the lower and upper quartiles, the whiskers extend to 1.5 times the inter-quartile range, and red "+" symbols are outliers.

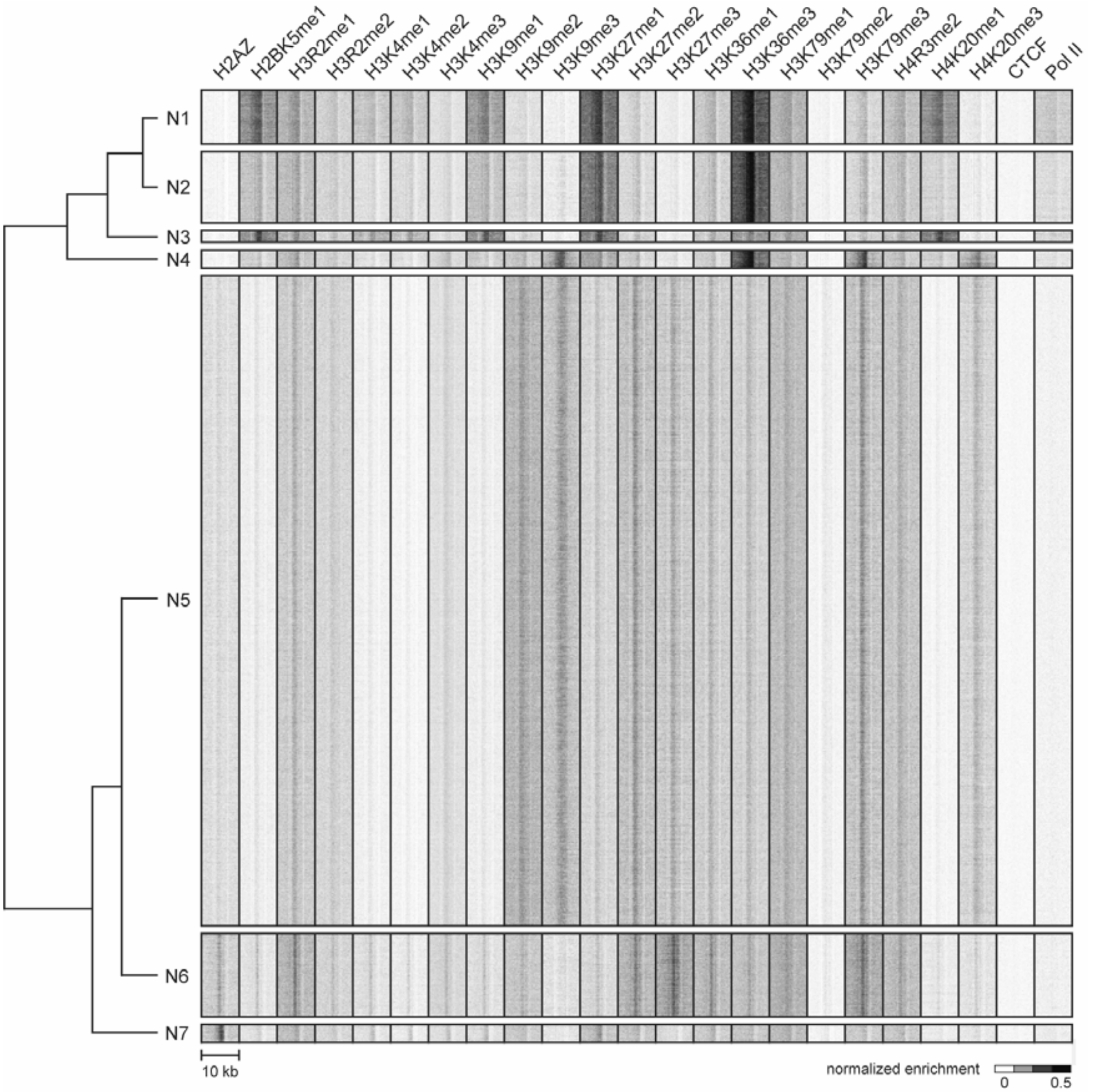


Figure 6-2: Distinct chromatin signatures spanning genomic loci distal to known regulatory elements.

I identified 50,183 genomic loci with strong ChIP enrichment of histone modifications but distal to promoters, gene 3' ends, DNase I hypersensitive sites, CTCF binding sites, and predicted enhancers. Applying ChromaSig to these loci reveals seven clusters N1-7 spanning 47,874 loci. The heat map represents the enrichment of H2AZ, 20 histone modifications, CTCF, and RNA polymerase II in the 10-kb region surrounding each locus. To organize these clusters visually, I performed hierarchical clustering on the average profiles of each ChromaSig cluster, using a Pearson correlation distance metric (left).

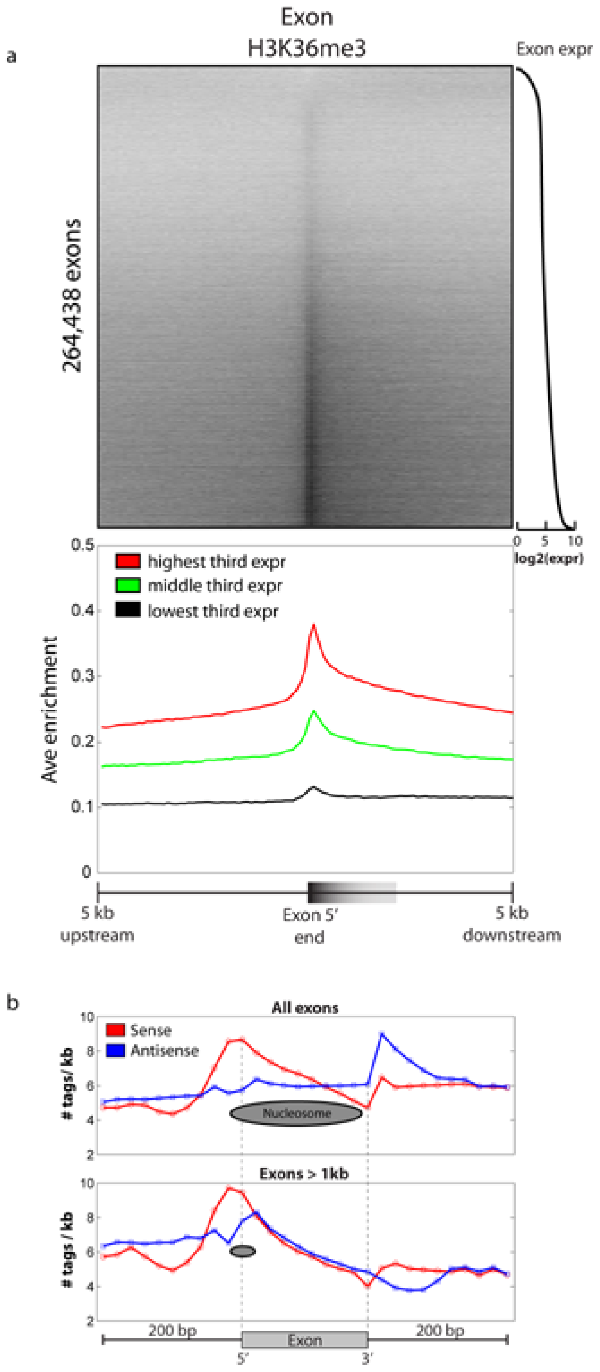


Figure 6-3: H3K36me3 marks exon 5’ ends and is a global mark of activity.

(a) The top panel is a heat map of H3K36me3 enrichment at all human exons, sorted by exonic expression (right). The bottom panel is the average H3K36me3 enrichment profile of the lowest, middle, and highest third of expressed exons from the top panel. **(b)** The top panel shows the distribution of H3K36me3 reads within and around all exons in the human genome. In red are reads on the sense strand in the direction of transcription, and in blue are antisense reads. A schematic of a positioned nucleosome is shown. The bottom panel is restricted to exons larger than 1-kb.

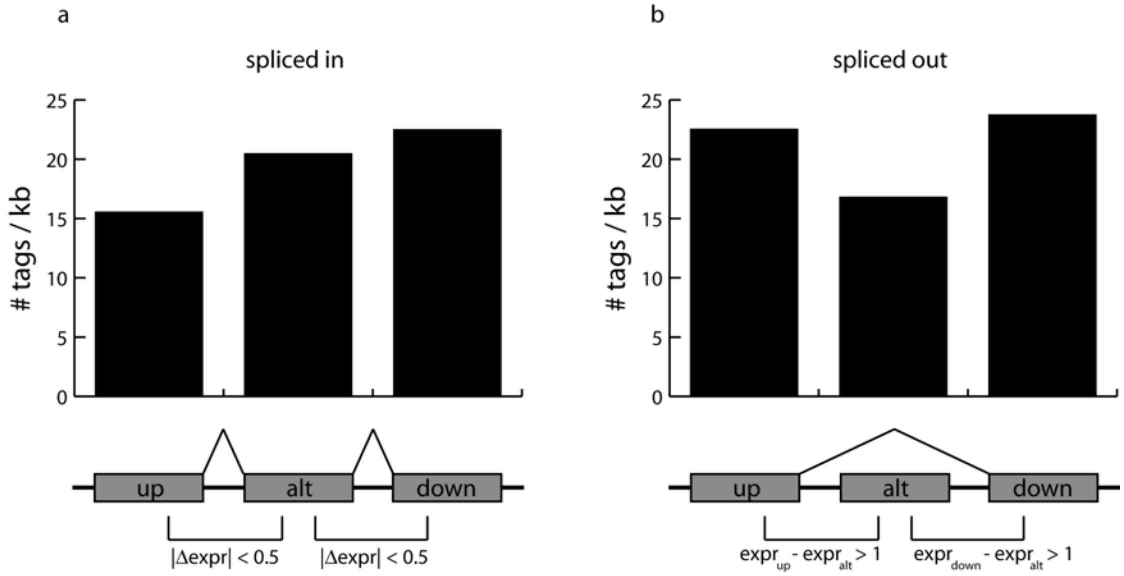


Figure 6-4: H3K36me3 enrichment correlates with alternative splicing.

The number of H3K36me3 reads per kilobase for exons near alternatively spliced cassette exons that are **(a)** spliced in or **(b)** spliced out. A cassette exon is defined to be spliced in if the difference in expression between it and its immediate upstream and downstream exons is less than 0.5 on a log2 scale. A cassette exon is defined to be spliced out if both upstream and downstream exons are at least 2-fold more expressed (1.0 on a log2 scale).

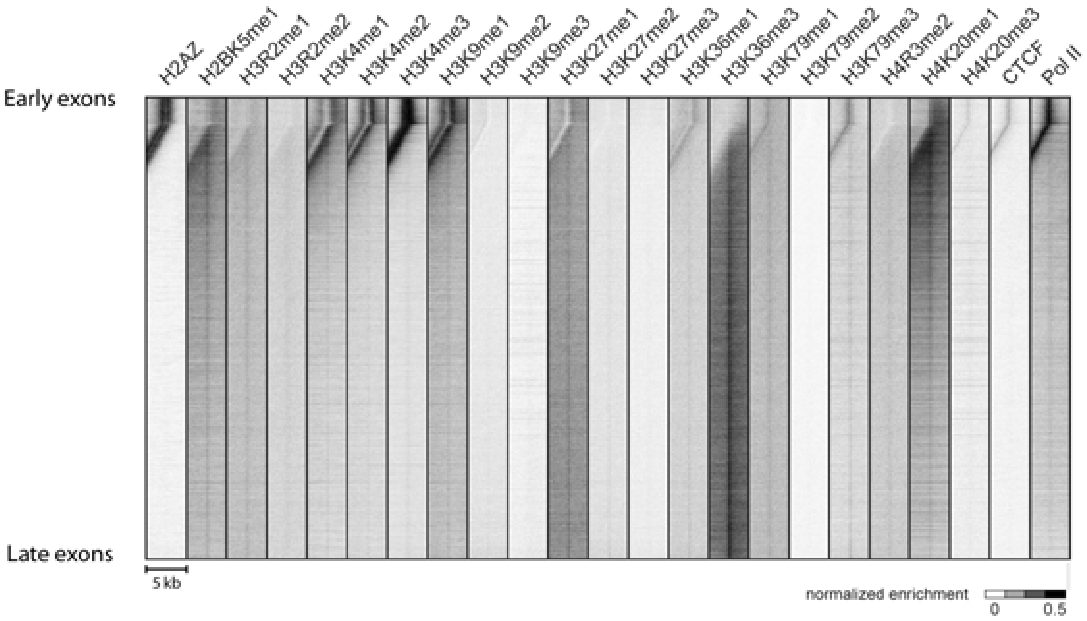


Figure 6-5: H2BK5me1 and H4K20me1 mark early exons.

Shown is a heat-map representing the enrichment of various modifications and factors in a 5-kb region surrounding the top third expressed exons. The exons are sorted by distance from the transcription start site.

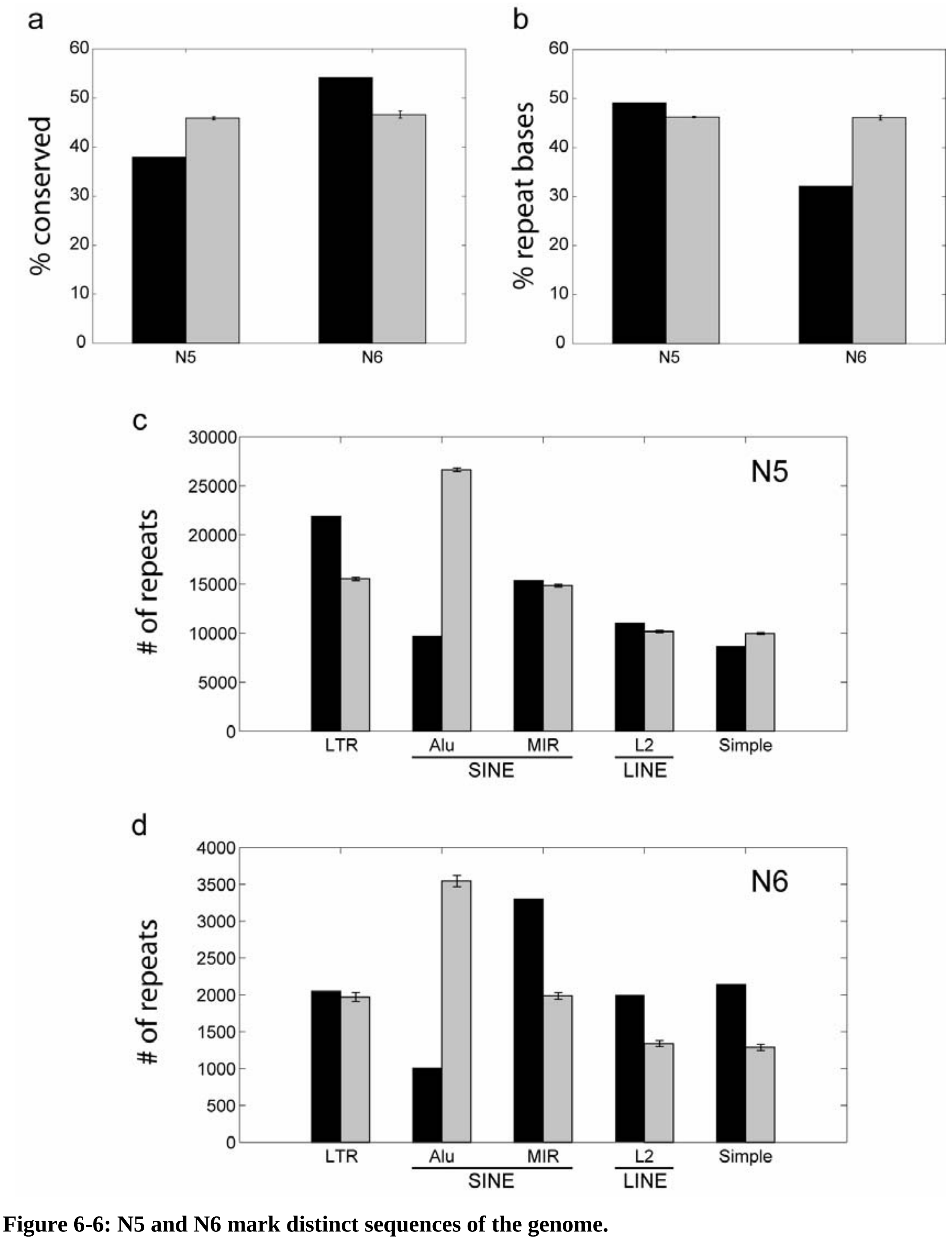


Figure 6-6: N5 and N6 mark distinct sequences of the genome.

(a) The percentage of loci in N5 and N6 within 1-kb to an evolutionarily conserved PhastCons element. **(b)** The average percentage of bases ±1 kb around each locus that are masked by RepeatMasker. **(c-d)** The number of repeat elements within ±1 kb of each locus in **(c)** N5 and **(d)** N6. Black indicates the observed value while grey indicates the expected value over random sites. The error bars indicate ±1 standard deviation. LTR, long terminal repeat; simple, simple repeat.

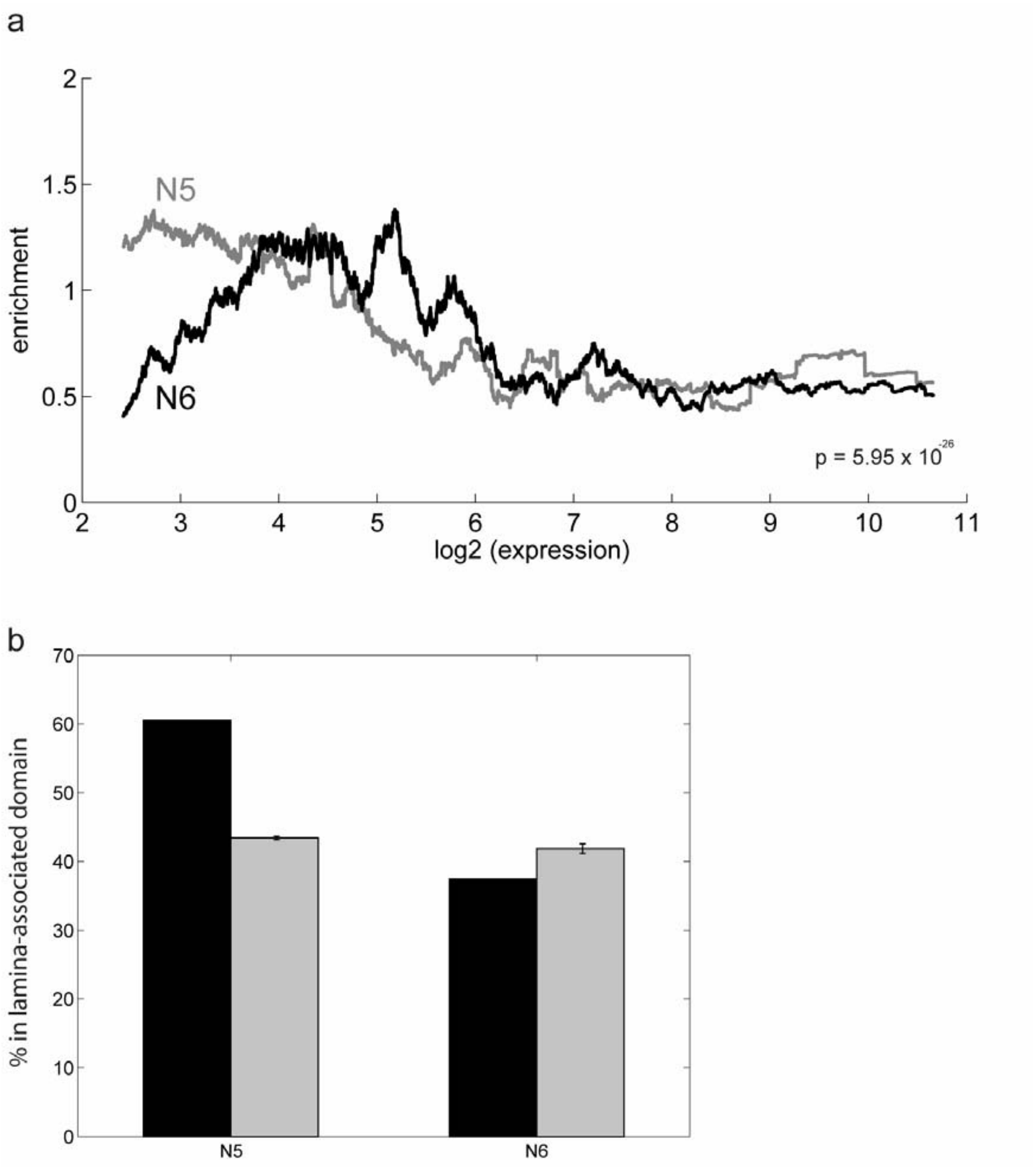


Figure 6-7: N5 and N6 mark distinct expression domains of the genome.

(a) Enrichment of N5 and N6 loci as a function of expression for genes in the same domain. I counted the number of N5 and N6 loci within the CTCF-defined domains containing human promoters, assessed enrichment as compared to that expected over random sites, and averaged over a 1000-promoter sliding window to create each profile. The signed rank p-value is indicated. **(b)** The percentage of each cluster within lamina-associated domains, previously mapped in Tig3 human lung fibroblasts (black), as compared to random sites (grey). The error bars indicate ±1 standard deviation.

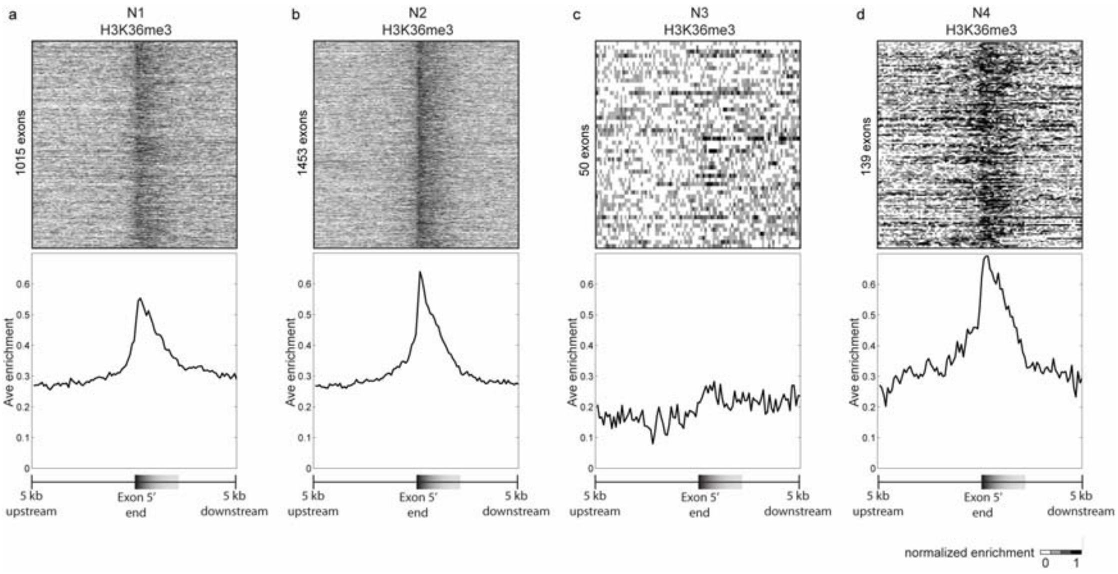


Figure 6-8: N1, N2, and N4 mark exon 5' ends.

An exon is unambiguously marked if it is the only exon within 1-kb of a genomic locus. I profiled chromatin enrichment relative to the 5' ends of unambiguously marked exons for clusters **(a)** N1, **(b)** N2, **(c)**, N3, and **(d)** N4. The top panels are heat maps representing the H3K36me3 enrichment in a 10-kb region surrounding the 5' ends of unambiguously marked exons. The bottom panels represent the average profiles of the heat maps. N3 is the negative control.

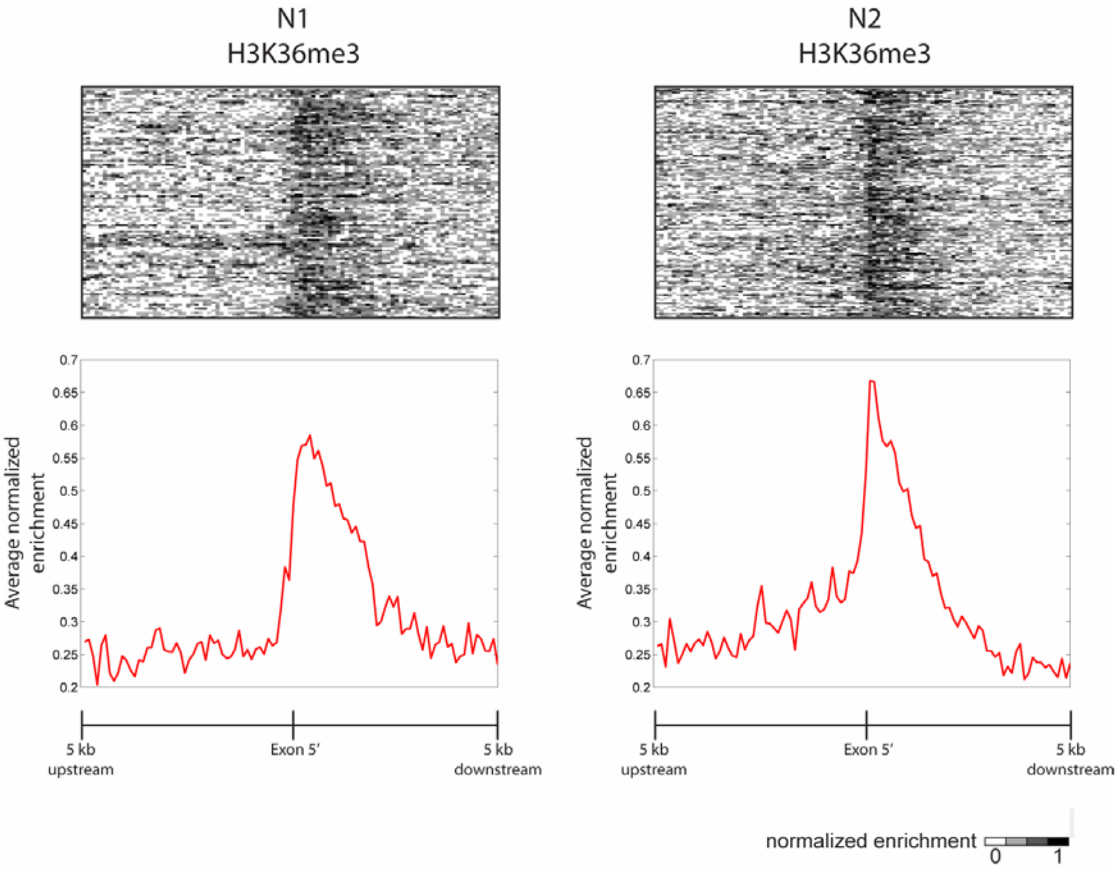


Figure 6-9: N1 and N2 mark the 5' ends of exons greater than 1-kb in length.

An exon is unambiguously marked if it is the only exon within 1-kb of a genomic locus. I profiled chromatin enrichment relative to the 5' ends of unambiguously marked exons of length >1-kb for clusters N1 and N2. The top panels are heat maps representing the H3K36me3 enrichment in a 10-kb region surrounding the 5' ends of unambiguously marked exons. The bottom panels represent the average profiles of the heat maps. Only a small number of N3- and N4-marked unambiguous exons are larger than 1-kb, and so are not shown here.

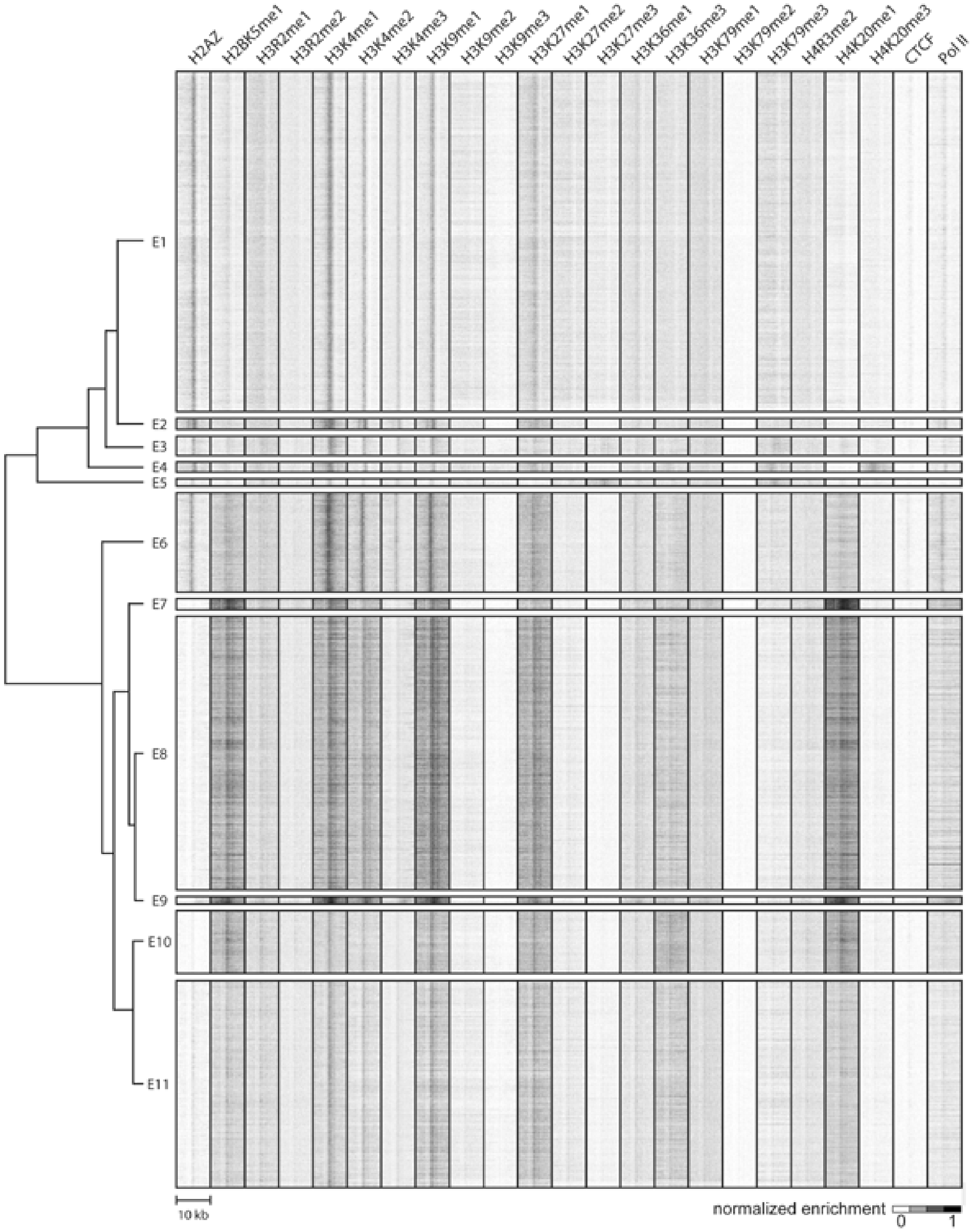


Figure 6-10: Distinct chromatin signatures spanning predicted enhancers.

On the basis of a previously published enhancer chromatin signature having strong H3K4me1 enrichment but weak H3K4me3 enrichment [6], I predicted 32,237 promoter-distal enhancers. Applying ChromaSig to these loci using the full panel of chromatin modifications mapped by Barski et al. [5], I recovered 11 clusters. The heat map represents the enrichment of H2AZ, 20 histone modifications, CTCF, and RNA polymerase II in the 10-kb region surrounding each enhancer prediction. To organize these clusters visually, I performed hierarchical clustering on the average profiles using a Pearson correlation distance metric (left).

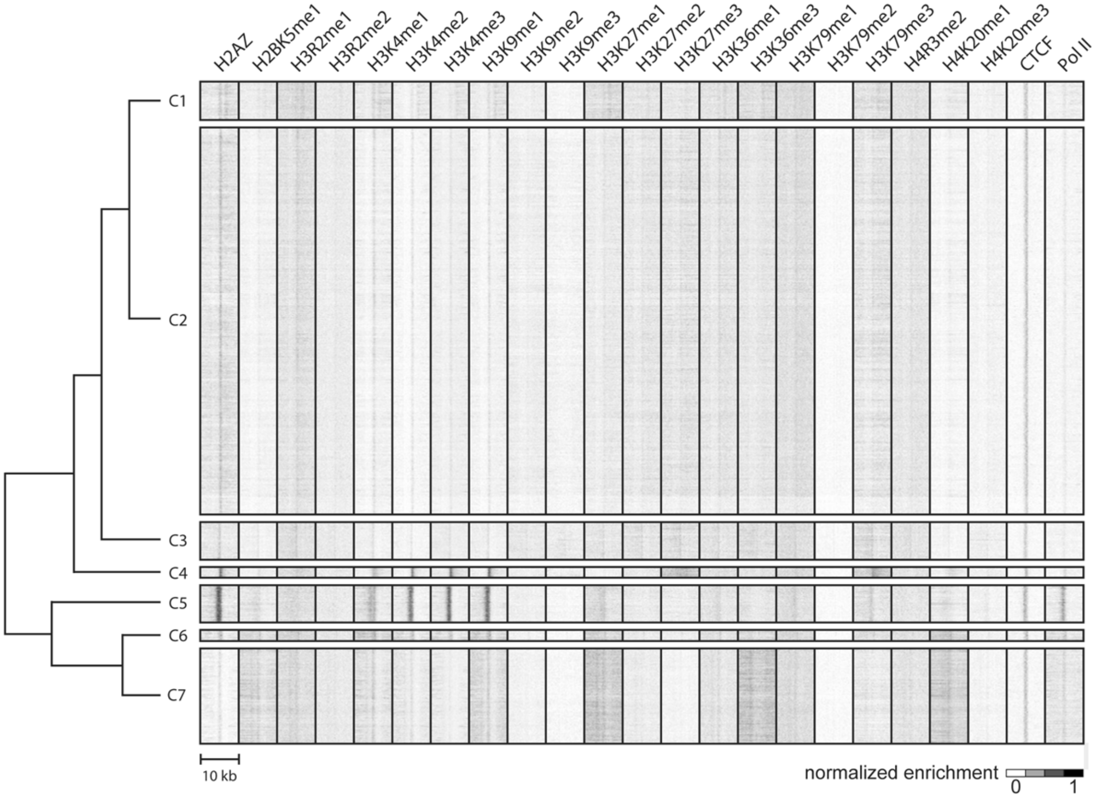


Figure 6-11: Distinct chromatin signatures spanning promoter-distal and enhancer-distal CTCF binding sites.

I used MACS [17] to identify 27,110 CTCF binding sites from the Barski et al maps [5], 17,328 of which are distal to promoters and predicted enhancers. Applying ChromaSig to the chromatin modifications around these loci, I recovered 7 clusters. The heat map represents the enrichment of H2AZ, 20 histone modifications, CTCF, and RNA polymerase II in the 10-kb region surrounding each distal CTCF binding site. To organize these clusters visually, I performed hierarchical clustering on the average profiles using a Pearson correlation distance metric (left).

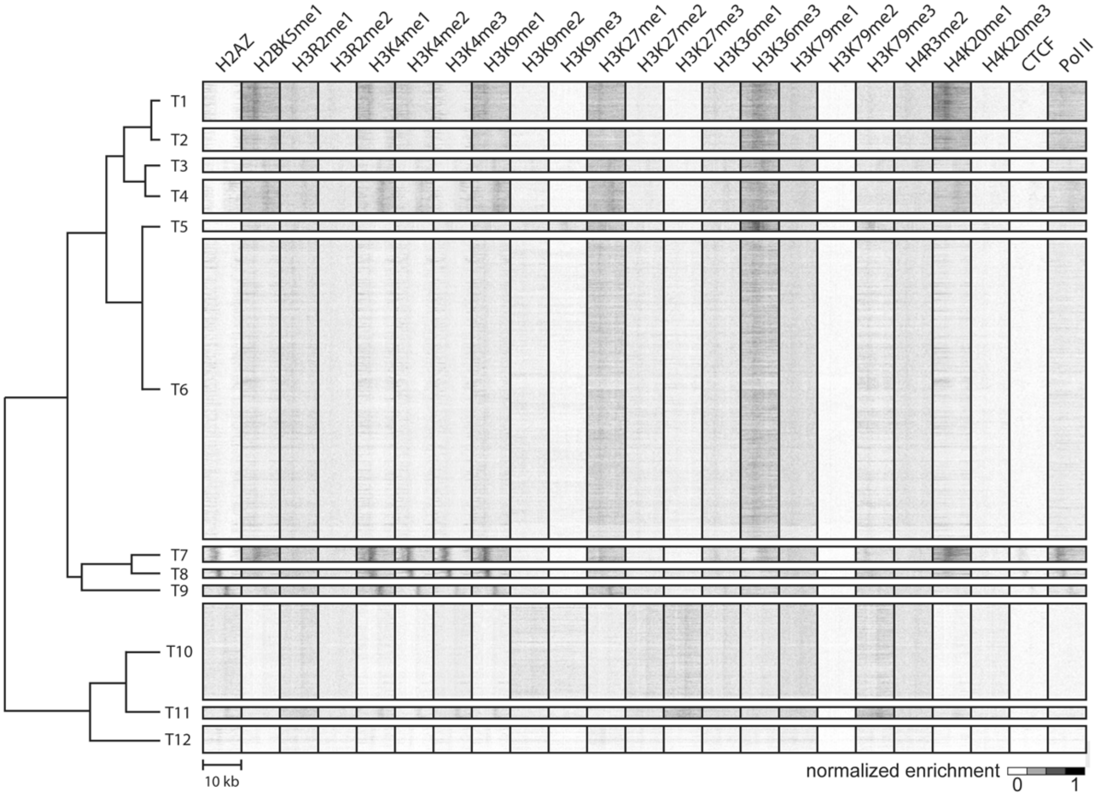


Figure 6-12: Distinct chromatin signatures spanning Refseq 3’ ends distal to Refseq promoters.

Applying ChromaSig to the histone modifications near 16,703 Refseq gene 3’ ends that are distal to Refseq TSSs, I recover 12 clusters. The heat map represents the enrichment of H2AZ, 20 histone modifications, CTCF, and RNA polymerase II in the 10-kb region surrounding each Refseq gene 3’ end. To organize these clusters visually, I performed hierarchical clustering on the average profiles using a Pearson correlation distance metric (left).

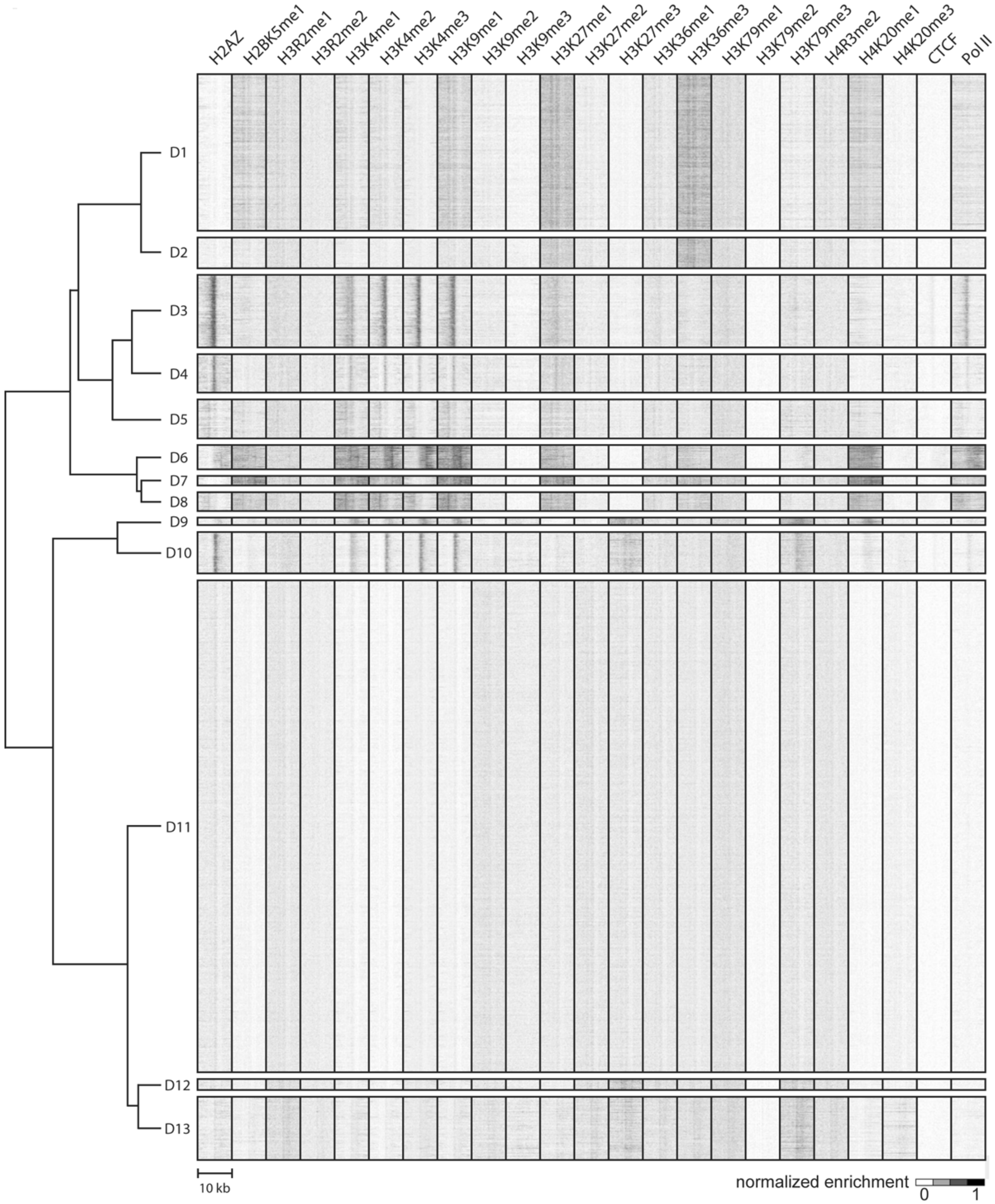


Figure 6-13: Distinct chromatin signatures spanning DNase I hypersensitive sites.

Previously, Boyle et al mapped 95,709 DNase I hypersensitive sites in CD4+ T cells, 31,824 of which are distal to Refseq TSSs, CTCF binding sites, and enhancer predictions. I applied ChromaSig to the chromatin modifications around these loci, recovering 13 clusters. The heat map represents the enrichment of H2AZ, 20 histone modifications, CTCF, and RNA polymerase II in the 10-kb region surrounding each distal DNase I hypersensitive site. To organize these clusters visually, I performed hierarchical clustering on the average profiles using a Pearson correlation distance metric (left).

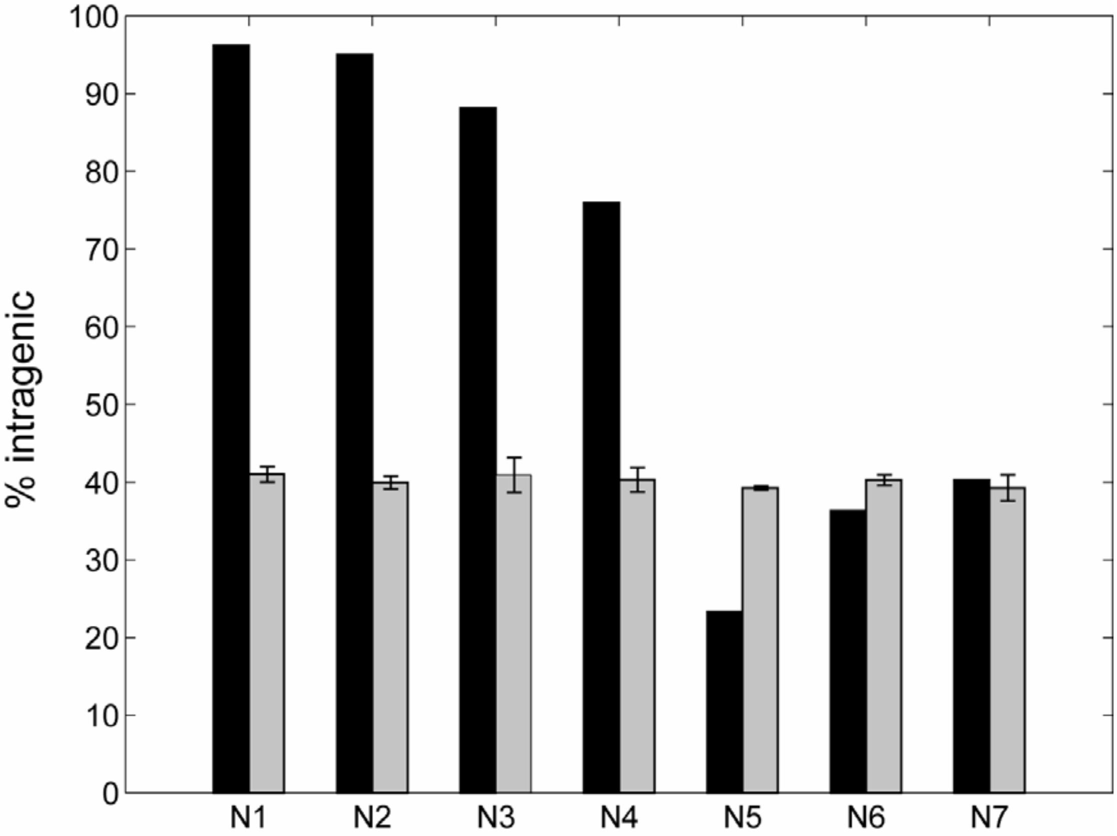


Figure 6-14: Distinct genomic distributions of chromatin signatures.

The percentage each cluster within the 5’ and 3’ ends of genes (black), as compared to random sites (grey). The error bars indicate ±1 standard deviation.

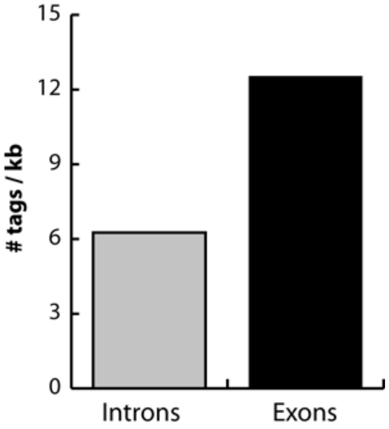
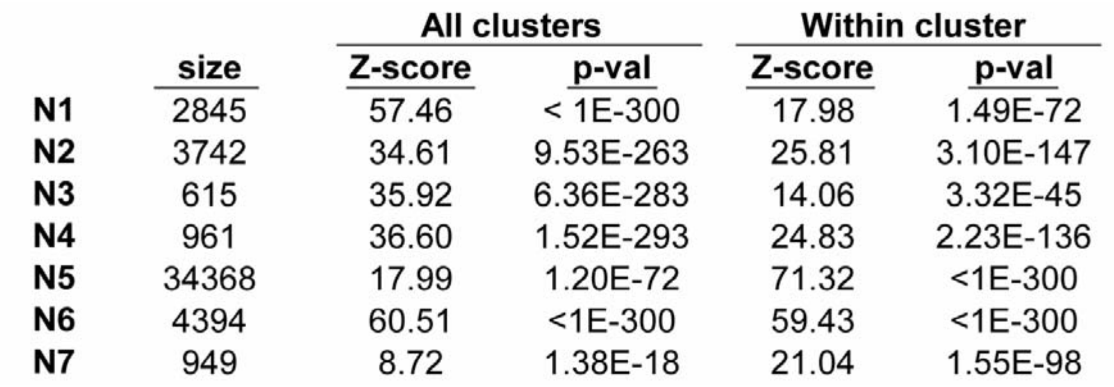


Figure 6-15: The distribution of H3K36me3 reads within exon and introns.

The number of reads found within introns and exons, normalized by the total size of each.

Table 6-1: Statistical significance of observed chromatin signatures.

Significance for each cluster is calculated by comparing to random sets of clusters sampled from within the cluster or over all clusters.



		All clusters		Within cluster	
	size	Z-score	p-val	Z-score	p-val
N1	2845	57.46	< 1E-300	17.98	1.49E-72
N2	3742	34.61	9.53E-263	25.81	3.10E-147
N3	615	35.92	6.36E-283	14.06	3.32E-45
N4	961	36.60	1.52E-293	24.83	2.23E-136
N5	34368	17.99	1.20E-72	71.32	<1E-300
N6	4394	60.51	<1E-300	59.43	<1E-300
N7	949	8.72	1.38E-18	21.04	1.55E-98

Chapter 7 : Conclusions

Large-scale, systematic studies of the epigenome are only in their infancy. Just as the genomics era was launched nearly a decade ago with the sequencing of the human genome, recent developments in mapping and characterizing the entirety of the human epigenome has launched the epigenomics era. These extra dimensions of the human genome will undoubtedly expand our current understanding of how a cell functions and develops, but may also give insights into how these processes have evolved over time. Below are several future directions likely to develop from epigenomics.

To date, chromatin signatures for a variety of functional genomic elements have been discovered. For example, promoters are marked by H3K4me3 and H3K27me3 [1,2] while exons are marked by H3K36me3 [3]. In this thesis, I have shown that H3K4me1 is a mark for enhancers. However, other functional elements in the genome including insulators, repressors, and locus control regions [4] have no well-defined chromatin signatures. While it is possible that no such signatures exist for these elements, it is tempting to speculate that one does exist. On a large scale, this could be assessed by mapping these elements and then assessing whether there is or is not a chromatin signature. For example, Johnson et al recently mapped the repressor NRSF to 1946 sites in Jurkat cells [5]. Examining the chromatin state in the same cell line could answer the question of whether a chromatin signature exists at repressors. This could be accomplished by ChIP approaches. However, as ChIP heavily relies on antibodies, and as antibodies are only available for a small subset of histone modifications, *de novo* approaches such as mass spectrometry may need to be used.

While the transcriptome specifies what genes are currently being expressed in a cell, the epigenome details the cell's more complex past history, present state, and future trajectory. Not only does the epigenome describe how gene expression is presently controlled, but it also contains information on how the cell reached its current state, as well as how the cell is ready to respond to environmental and developmental cues to alter its transcriptional output. This poised phenomenon has been well-documented at promoters where a bivalent epigenetic state ensures a poised transcriptional state [2], and likely also applies to enhancers [6,7,8]. As current research focuses towards comparing multiple lineage-related cell types, deciphering the complex epigenetic past history and future paths of the cell will yield deeper insights into gene regulation and development.

By detailing how the epigenome changes during differentiation, we can understand how control of transcription is rewired as an organism develops. In addition to answering questions about transcriptional regulation within an organism, the epigenome can also offer insight on how transcriptional regulation has evolved across species. Just as the explosion of genome sequences spurred comparative genomics, a similar explosion in epigenetic maps will motivate studies in comparative epigenomics. In recent years, comparative genomics has offered critical insights on how genome sequences, especially in terms of their transcriptional regulatory elements, have evolved. But conservation at the sequence level does not necessarily equate to functional conservation in the same cellular context. By adding a cell-type specific view of the activities of these regulatory elements, comparative epigenomics will bridge this gap. Comparison of epigenomes of the same cell types from different organisms will add a dynamic cell-type specific view to comparative genomics. Furthermore, the ability to profile the epigenomes of multiple cell types during development in several species will add yet another dimension to our understanding genome function and conservation.

The precise control of transcription is essential for proper cellular function, and dysregulation can result in disease, notably cancer. For example, distinct types of prostate cancers each have a unique gene expression profile shared only by prostate cancers of the same type, all of which are distinct from normal cells [9]. Given these differences in transcriptional output, it will be interesting to see to what extent epigenomic differences underlie transcriptional differences in cancer cells.

Extending these techniques, recently-funded consortium projects such as the NIH Epigenome Roadmap Project and the ENCODE Project will undoubtedly come close to mapping all functional elements of the human genome. In the near future, mapping of the epigenome will help us to understand how transcription is controlled in a cell, will detail the regulatory events that happen during development, and establish how these developmental events have been conserved across species.

References

1. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA (2007) A chromatin landmark and transcription initiation at most promoters in human cells. Cell 130: 77-88.

2. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES (2006) A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell 125: 12.

3. Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J (2009) Differential chromatin marking of introns and expressed exons by H3K36me3. Nat Genet.

4. Maston GA, Evans SK, Green MR (2006) Transcriptional Regulatory Elements in the Human Genome. Annu Rev Genomics Hum Genet 7: 29-59.

5. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316: 1497-1502.

6. Lupien M, Eeckhoute J, Meyer CA, Wang Q, Zhang Y, Li W, Carroll JS, Liu XS, Brown M (2008) FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. Cell 132: 958-970.

7. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching KA, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenkov VV, Stewart R, Thomson JA, Crawford GE, Kellis M, Ren B (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature.

8. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39: 311-318.

9. Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, Ferrari M, Egevad L, Rayford W, Bergerheim U, Ekman P, DeMarzo AM, Tibshirani R, Botstein D, Brown PO, Brooks JD, Pollack

JR (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proc Natl Acad Sci U S A 101: 811-816.