Computational epigenomics: gene regulation, comparative methodologies, and epigenetic patterns

by

Angela Yen

B.S., Massachusetts Institute of Technology (2010)
M.Eng., Massachusetts Institute of Technology (2011)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2016

Author .....
Department of Electrical Engineering and Computer Science
May 19, 2016

Certified by .....
Manolis Kellis
Professor
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Chair, Department Committee on Graduate Students

Computational epigenomics: gene regulation, comparative methodologies, and epigenetic patterns

by

Angela Yen

Submitted to the Department of Electrical Engineering and Computer Science
on May 19, 2016, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

One of the fundamental aims of biology is to determine what lies at the root of differences across individuals, species, diseases, and cell types. Furthermore, the sequencing of genomes has revolutionized the ways in which scientists can investigate biological processes and disease pathways; new genome-wide, high-throughput experiments require computer scientists with a biological understanding to analyze and interpret the data to improve our understanding about life science. This provides us with a key opportunity to use computational techniques for new biological discoveries.

While genetic variation plays an important role in influence phenotype, sequence alone cannot account for all differences: for example, different types of cells in an individual have varying function and attributes, but identical genetic makeup. This highlights the importance of studying epigenetic changes, which are dynamic chemical changes to and around the DNA. While the DNA of every cell in an individual is the same, the epigenetic context for that DNA varies from cell to cell. In this way, these epigenetic differences play a crucial role in gene regulation, with epigenetic changes both causing and recording regulatory mechanisms.

In this thesis, we combine the power of computational, statistical, and data science approaches with the new wave of epigenetic data at a genome-wide level in a number of ways. First, in chapter 2, we demonstrate the importance of computational analysis at an epigenomic level by identifying an epigenomic signature of the olfactory receptor gene family that gives insight into the mechanism behind monogenic gene regulation. Next, in chapter 3, we explain our development of ChromDiff, a novel statistical and information theoretic computational methodology to identify chromatin state differences in groups of samples. In our methodology, we use correction for external covariates to isolate the relevant signal, and as a result, we find that our method outperforms existing computational methods, with further validation through randomized simulations. In chapter 4, we apply our methodology to characteristics including sex, developmental age, and tissue type, we unveil relevant chromatin states and genes that distinguish the groups of epigenomes, with further validation of our

results through differential expression analysis and gene set enrichment. In chapter 5, we show the power of integrative analysis through the combination of DNA methylation data with chromatin state profiles, cell types, sample groups, experimental technologies, and histone mark data to reveal insightful epigenetic patterns and relationships. Finally, in chapter 6, we identify "hidden" or "unknown" covariates in epigenomic data by using agnostic principal component analysis on our samples to discover similarities between our known covariates and the identified components.

In summation, our research highlights the importance of both algorithm development and method application for epigenomic questions, reaffirming the importance of interdisciplinary research that brings together cutting-edge techniques in computer science with appropriate biological hypotheses and data. While questions and analysis must be carefully paired in an informed manner to produce meaningful, interpretable, and believable results in computational biology, our work here provides a sampling of the vast potential for scientific discovery at the intersection of the fields of computer science and biology.

Thesis Supervisor: Manolis Kellis
Title: Professor

Acknowledgments

My thesis research would not have been possible without the help of many people. Firstly, thanks to my research advisor, Manolis Kellis, for guiding me on the path to becoming an independent researcher; your support and encouragement has been invaluable at every step of the way. I would also like to thank Pardis Sabeti and Bonnie Berger, who were both thesis committee members and rotation supervisors; your mentorship and feedback was much appreciated throughout graduate school. Similarly, thanks to Nancy Lynch and David Gifford, who served as my Research Qualifying Exam committee: your perspectives were helpful to my research as well.

Additionally, this work has largely been the product of many collaborations. To Stavros Lomvardas, Angeliki Magklara, and Eleanor J. Clowney: thank you for your patience and guidance throughout the olfactory regulation project and our other collaborations; Stavros, your unfailing support and encouragement has also been much appreciated. Thank you to all our Epigenome Roadmap collaborators, especially Misha Bilenky, as well as Daniel MacArthur and Taru Tukiainen for the fun and interesting collaboration on X chromosome escape genes.

I also can not overstate how much I have learned from my labmates over the years. Some of the many past and present Kellis Lab members I must especially thank include Wouter Meuleman for your mentorship and friendship over the years; Bob Altshuler for your support and advice; Pouya Kheradpour and Gerald Quon for your ever insightful research feedback; Lucas Ward for your guidance and patience; Richard Cowper-Sallari for your optimism and encouragement; Abhishek Sarkar, Xinchen Wang, and Kunal Bhutani for being wonderful officemates (with snacks, tea, and Atlas!); Anshul Kundaje for your insightful help and generous spirit; Khoi Nguyen, Irwin Jungreis, and Max Wolf for your help with thesis preparations; Jianrong Wang for your help with enhancer linking; and Jason Ernst, for your epigenomics expertise.

Thank you to the various funding sources that have supported my graduate research, including grants through my advisor, Manolis Kellis, as well as the Siebel Scholarship and the NSF Graduate Research Fellowship Program.

Lastly, I would like to thank my family and friends, who I feel played perhaps the most critical role in supporting me throughout graduate school. To my husband, Gabriel, I am more thankful to you than you will ever know; thank you for always believing in me, lifting me up through the hard times, and celebrating with me through the good times. I owe a huge thanks to my parents: Mom and Dad, I will never be able to repay you for all the love, sacrifice, support, and encouragement you have given me throughout my life, but please know that I will always appreciate everything that you have done for me. To Phil, thank you for being not only my brother but my longest friend; it means a lot to know that I can always count on you for encouragement, love, and support. To Carlos, Clemencia, and Ana: thank you for welcoming me into your family with open arms; I feel lucky to be a part of your family. To my friends: there are not enough words for me to say what all of you mean to me, but I feel ever grateful to be friends with such good, caring, loving, funny, joy-filled, intelligent, supportive, genuine, and hilarious people; special thanks go out to Simone, Lindsay, Kent, May, Cathy, Michael, Dustin, Sara, Manasi, and Becca for your patience in dealing with my grad school struggles and problems. Lastly, I must thank my dog, Atlas: you are the best graduate school companion anyone could have asked for; I will miss spending every day with you at the office, but all things must come to an end...

Contents

1 Introduction **21**

1.1 Motivation 21

1.2 Background 23

1.2.1 Gene expression: central dogma of molecular biology 23

1.2.2 Regulation of gene expression 25

1.2.3 Functional and regulatory regions 26

1.2.4 Epigenetic state and modifications 27

1.2.5 Epigenetic variation 30

1.2.6 Human biology and model organisms 30

1.3 Relevant experimental methods 31

1.3.1 DNA sequencing and mapping 31

1.3.2 Genotyping 32

1.3.3 Chromatin immunoprecipitation 35

1.3.4 DNA methylation profiling 35

1.3.5 Transcriptome sequencing 38

1.3.6 DNA footprinting 39

1.4 Summary of research contributions 40

1.4.1 An epigenomic mechanism for gene family regulation 40

1.4.2 A genome-wide computational method for group-wise chromatin state comparisons 42

1.4.3 Epigenomic insights from comparisons based on development age, sex, and tissue type. 43

1.4.4 Integrative analysis of DNA methylation data across chromatin state, gene expression, platform, and celltypes. 44
1.4.5 Identification of unknown covariates in epigenomic samples with mutual information analysis against true covariates. 45

2 An epigenomic mechanism for regulation of the olfactory receptor gene family 47

2.1 Introduction 48
2.1.1 Problem Statement 48
2.1.2 Background and previous work 48
2.1.3 Approach 50
2.2 Computational methods 51
2.2.1 Data processing, normalization, and quality control 51
2.2.2 Detection of heterochomatic domains 53
2.2.3 Clustering and ranking of genes 56
2.3 Results 58
2.3.1 Quality controls 58
2.3.2 Whole-genome analysis of H3K9me3 and H4K20me3 in olfactory epithelial tissue 59
2.3.3 Heterochromatic signature for chemoreceptors 62
2.3.4 Heterochromatic macrodomains cover OR clusters 64
2.3.5 Further experimental validation 70
2.4 Contributions 78

3 A computational method for chromatin state comparisons across groups of epigenomes 81

3.1 Introduction 82
3.1.1 Problem Statement 82
3.1.2 Background and previous work 83
3.1.3 Approach 84
3.2 Methods 85

3.2.1 Overview of comparison of epigenomic features 85
3.2.2 Chromatin state annotations 90
3.2.3 Information theoretic representation of raw feature values . . . 90
3.2.4 Gene annotations 92
3.2.5 Covariate correction of ChromDiff feature values 93
3.2.6 Group-wise comparison statistics 94
3.2.7 Gene set enrichment calculations 95
3.2.8 Expression data analysis 97
3.2.9 Randomized simulations 98
3.3 Results 99
3.3.1 Identified genes are enriched for differential expression 99
3.3.2 ChromDiff outperforms other method for epigenomic comparison 102
3.3.3 ChromDiff identifies relevant genes and chromatin states independent of gene size and chromatin state. 107
3.3.4 Regulatory ChromDiff identifies additional and new genes when studying linked enhancers and regulatory regions. 109
3.4 Software download 111
3.5 Contributions 111

4 Comparisons of epigenomes reveal distinguishing chromatin states and genes 113

4.1 Introduction 113
4.2 Methods 114
4.2.1 Gene cluster identification 114
4.2.2 Sampling distinguishing features 114
4.2.3 Sampling significant distinguishing genes 114
4.2.4 Ordering of rows or columns 115
4.2.5 Dominant (most abundant) chromatin state heatmaps 115
4.2.6 Gene expression heatmaps 115
4.2.7 Chromatin state enrichment for X chromosome gene sets . . . 116

4.2.8 Sex-based chromatin state enrichment . . . 117
4.2.9 Violin plots for chromatin state coverage . . . 117
4.3 Results . . . 118
4.3.1 Overview . . . 118
4.3.2 Epigenetic sex differences consistent with X Chromosome inactivation . . . 118
4.3.3 Active chromatin states are enriched on genes that escape ChrX inactivation, while repressive chromatin states are enriched on inactive genes. . . . 120
4.3.4 Comparison of brain and gastrointestinal tissues reveal epigenomic changes in neuronal genes . . . 126
4.3.5 Blood samples distinguished by enhancer activity differences . 130
4.3.6 Comparison of samples based on developmental ages link to cancer genes . . . 135
4.3.7 ChromDiff identifies changes at linked enhancers based on tissue type. . . . 140
4.3.8 Studying chromatin state changes at both enhancers and DNase hypersensitive sites identifies ChrX genes. . . . 142
4.3.9 Subtypes of blood samples highlight enhancer differences. . . . 144

5 Integrative analysis of Roadmap Epigenomics data 149
5.1 Introduction . . . 149
5.2 Methods . . . 150
5.2.1 Data processing of RNA-seq, ChIP-seq, and DNase-seq . . . 150
5.2.2 Data processing of DNA methylation data . . . 151
5.2.3 Chromatin state learning . . . 151
5.2.4 Relationship between chromatin states and methylation. . . . 152
5.2.5 Generating randomized simulations of methylation profiles. . . 152
5.2.6 Calculating significant differences in simulated vs real data. . . 153
5.2.7 Clustering of celltypes based on epigenetic information. . . . 153

5.2.8 Comparison of DNA methylation platforms 154
5.3 Results 154
5.3.1 Relationship between DNA methylation and chromatin states 154
5.3.2 DNA Methylation profiles for chromatin states across epigenomes 157
5.3.3 Comparison of DNA methylation platforms 162
5.3.4 Clustering of celltypes based on pairwise epigenomic similarity. 171

6 Identification of unknown covariates in epigenomic data with comparisons to known metadata 175
6.0.1 Introduction 175
6.1 Methods 177
6.1.1 Representation of each epigenome 177
6.1.2 Principal component analysis 178
6.1.3 Mutual information between principal components and covariates 180
6.2 Results 182
6.2.1 Principal component analysis with gene body representation . 182
6.2.2 Linked regulatory region analysis 188

7 Conclusion 197
7.1 Summary of results 198
7.2 Future Work 200

List of Figures

1-1 The central dogma of molecular biology states that DNA is transcribed into RNA, which is then translated into proteins.[1] 22

1-2 Epigenetic modifications include methylation of the DNA strand (left) and chemical modifications to the tails of the histone proteins (right).[2] 23

1-3 RNA Polymerase reads the nucleotides of one strand of the DNA to produce the complementary RNA in a process known as transcription.[3] 24

1-4 The process of translation reads mRNA to assemble a chain of amino acids to form a protein.[4] 25

1-5 Triplets of mRNA nucleotides act as codons to map to amino acids.[5] 25

1-6 Illustration of functional and regulatory genomic elements. 27

1-7 A nucleosome is made up of 8 histone proteins and has DNA wrapped around it. The combination of the DNA and the nucleosome is called chromatin.[6] 29

1-8 Mapping sequenced reads of DNA. 32

1-9 *De novo* assembly of reads of DNA. 33

1-10 Explanation of allele-specific oligonucleotides. 34

1-11 Explanation of Chromatin Immunoprecipitation. 36

1-12 Bisulfite-sequencing-based experimental techniques directly measure methylated CpG sites in the genome. 37

1-13 RNA-sequencing technology measures gene expression. 38

2-1 This figure demonstrates the sliding window algorithm with an example window of size 4. 54

2-2 Hidden Markov Models are composed of hidden states (circles X1, X2, X3), observed emissions (squares y1, y2, y3, y4), transition probabilities (a arrows), and emission probabilities (b arrows).[7] 55

2-3 Explanation of k-means clustering. 57

2-4 Quality check plots reveal data relatively normal data devoid of bias. 59

2-5 Genome-wide mapping of H3K9me3 and H4K20me3 reveal a tissue-dependent heterochromatinization of the ORs in the mouse olfactory epithelium (MOE). 61

2-6 Unsupervised clustering of chromosome 2 genes. 62

2-7 Olfactory receptor genes have strongest H3K9me3 and H4K20me3 signal. 64

2-8 Comparison of blocks identified with the LOCKS protocol (top blue row) and the MA2C protocol (lower blue row) reveals similar macrodomains. 66

2-9 Part of a cluster of OR genes overlaps peaks and blocks of H3K9me3 and H4K20me3 in OE. 67

2-10 H3K9me3 and H4K20me3 is highly present at OR genes in the OR cluster in MOE (but not liver), but not beyond the OR cluster, based on ChIP-qPCR. 68

2-11 Part of a silenced OR cluster on chromosome 11 is interrupted by a small group of transcriptionally active non-OR genes. 69

2-12 The ORs acquire a highly compacted chromatin structure in the MOE. 71

2-13 OR silencing occurs independent of and developmentally prior to OR expression. 72

2-14 The active OR allele is not enriched for H3K9me3 or H4K20me3, but it is marked with H3K4me3. 75

2-15 Tissue-Specific OR Modifications Are Associated with OR-like Transgene Expression. 77

2-16 Previous model and proposed new model for olfactory gene regulation. 78

3-1 A novel method for comparative analysis of epigenomic groups. 87

3-2 Leveraging linked regulatory regions for identification of epigenomic differences at relevant regulatory regions across groups. 91
3-3 Even with the conservative Benjamini-Yekuteli multiple hypothesis correction, we still identify significantly different features and genes in 7 of 12 cases. 96
3-4 Epigenomically distinguishing genes are enriched for differential expression. 100
3-5 ChromDiff outperforms dPCA in identification of comparison-specific genes. 105
3-6 ChromDiff identifies more specific results than dPCA. 106
3-7 A variety of chromatin states are identified in simulations and applications. 107
3-8 Genes of various sizes are identified in randomized simulations, while longer genes are identified in biological comparisons. 108
3-9 Regulatory ChromDiff identifies new genes when using promoters, linked enhancers, and linked DNase Hypersensitive sites, compared to the gene body approach. 110
3-10 Proportion of distinguishing genes identified by Regulatory ChromDiff with a) promoters, b) enhancers, and c) DNase Hypersensitive regions, compared to the gene body ChromDiff approach. 110

4-1 X chromosome inactivation distinguishes male and female samples. . . 119
4-2 Distinguishing autosomal genes are associated with changes in bivalent and enhancer regions. 121
4-3 Number of escape, inactive, and variable genes on the X chromosome. 122
4-4 Comparison of chromatin state coverage at escape genes and inactive genes reveals distinctive chromatin state biases. 123
4-5 Active flanking promoter, transcribed, weakly transcribed, and enhancer chromatin states enriched in escape genes compared to inactive genes. 124

4-6 Heterochromatic, bivalent promoter, and polycomb repressed chromatin states are enriched in inactive genes compared to escape genes. 125

4-7 Comparison of chromatin state coverage at ChrX genes in female and male samples reveals chromatin state patterns consistent with X chromsome inactivation. 127

4-8 Transcriptional differences dominate brain and GI tissue comparison. 128

4-9 Brain and gastrointestinal differences reveal changing chromatin state differences in gene clusters. 129

4-10 Epigenomic differences specific to blood samples lie at blood cancer genes. 134

4-11 Polycomb targets distinguish adult and fetal samples. 138

4-12 Many genes exhibiting changes between adult and fetal samples are only associated with one chromatin state. 139

4-13 Differences at brain and digestive in linked enhancer activity identify genes with differential expression that relate to neuronal development. 143

4-14 Differences in polycomb repression and heterochromatin is identified at DNase hypersensitive sites linked to ChrX genes in a sex-based comparison. 145

4-15 Epigenomic changes at linked enhancers identify celltype-specific gene expression and genes. 147

5-1 Chromatin states and DNA methylation dynamics. 152

5-2 DNA methylation values at chromatin state regions in 15-state model. 155

5-3 DNA methylation of real chromatin state compared to background regions. 156

5-4 DNA methylation profiles of chromatin states across celltypes. 159

5-5 Across celltypes, promoter and bivalent regions are methylated most differently from background, while quiescent and polycomb repressed regions most frequently have DNA methylation similar to background. 160

5-6 Distribution of DNA methylation values for each chromatin state based on sample group. 161
5-7 Real and simulated distribution of DNA methylation values for each chromatin state in the IMR90 feta llung fibroblasts cell line. 163
5-8 Comparison of DNA methylation technologies RRBS and SBS. 165
5-9 Comparison of DNA methylation technologies SBS and mCRF. . . . 167
5-10 Comparison of DNA methylation technologies RRBS and mCRF. . . 168
5-11 DNA methylation levels in 15-state model across technologies. 170
5-12 Distribution of DNA methylation values for each chromatin state based on sample group and experimental technology. 171
5-13 DNA methylation at enhancer modules across celltypes. 172
5-14 Similarity between epigenomes based on histone mark presence in enhancer regions. 174

6-1 Illustration of dimension reduction with principal component analysis. 178
6-2 Illustration of singular value decomposition. 179
6-3 Top principal components of epigenomes based on gene body features. 183
6-4 Variance and pair-wise projections of principal components. 184
6-5 Mutual information reveals structure between principal components and covariates. 189
6-6 Principal component analysis applied to epigenomes based on enhancer features. 191
6-7 Principal component analysis in DNase hypersensitive sites. 193
6-8 Principal component analysis in promoter regions. 195

List of Tables

2.1 OR genes significantly cluster together based on the histone mark profile. 63

3.1 Analysis of biological comparisons reveal differences in chromatin state. 89

3.2 ChromDiff can capture epigenomic differences even when there are no differentially expressed genes. 102

3.3 ChromDiff identifies chromatin state differences for Adult/Fetal, Female/Male, and CellLine/PrimaryCulture comparisons, while dPCA does not. 104

4.1 Enriched gene sets for cluster A of brain and gastrointestinal comparison. 131

4.2 Enriched gene sets for cluster B of brain and gastrointestinal comparison 131

4.3 Enriched gene sets for cluster C of brain and gastrointestinal comparison. 131

4.4 Enriched gene sets for cluster D of brain and gastrointestinal comparison. 132

4.5 Enriched gene sets for cluster E of brain and gastrointestinal comparison 132

4.6 Enriched gene sets for cluster F of brain and gastrointestinal comparison. 133

4.7 Enriched gene sets for brain and gastrointestinal comparison. 133

4.8 Enriched gene sets for cluster A from comparison of blood and non-blood samples. 135

4.9 Enriched gene sets for cluster B from comparison of blood and non-blood samples 136

4.10 Enriched gene sets for cluster C from comparison of blood and non-blood samples. 136

4.11 Enriched gene sets for cluster D from comparison of blood and non-blood comparison. 137
4.12 Enriched gene sets for adult and fetal comparison. 140
4.13 Enriched gene sets for cluster A from comparison of adult and fetal samples. 141
4.14 Enriched gene sets for cluster B from comparison of adult and fetal samples. 141

Chapter 1

Introduction

1.1 Motivation

One of the fundamental aims of biology is to determine what lies at the root of differences across individuals, species, diseases, and cell types. Furthermore, the sequencing of genomes has revolutionized the ways in which scientists can investigate biological processes and disease pathways; new genome-wide, high-throughput experiments require computer scientists with a biological understanding to analyze and interpret the data to improve our understanding about life science. This provides us with a key opportunity to use computational techniques for new biological discoveries.

One key insight into the question of biological variation comes from the central dogma of molecular biology: DNA in genes is transcribed into RNA, which is then translated into proteins, as shown in Fig. 1-1. This process of turning gene DNA into proteins is also known as gene expression. This central dogma naturally suggests that sequence changes in the DNA could play an important role in phenotypic variation, and research has shown this to be true. Specifically, research like the 1000 Genomes Project[8] has compared the genomes of individuals to identify genetic variations, such as Single Nucleotide Polymorphisms (SNPs) and insertions and deletions of genetic sequence (InDels), across individuals and ethnic groups. Furthermore, comparisons of genomes of different species have produced evolutionary models of how species have evolved from one another.[9] Meanwhile, Genome-Wide Association Studies (GWAS)

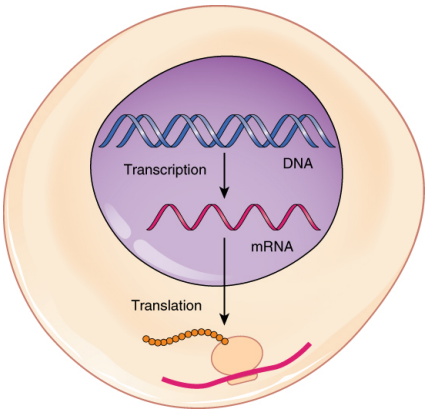

Figure 1-1: The central dogma of molecular biology states that DNA is transcribed into RNA, which is then translated into proteins.[1]

have identified genetic variations that correlate with disease populations.[10]

While genetic variation is critical for individual, species, and disease variation, it cannot account for all differences: for example, different types of cells in an individual have varying function and attributes, but identical genetic makeup. This can be explained by the important role of gene regulation, which determines when, where, and how much each gene is expressed. Even though the genetic sequence provides the building blocks for an organism, gene regulation provides the instructions for how to put those blocks together.

One way to study gene regulation is through epigenetic changes, which are dynamic chemical changes to and around the DNA. While the DNA of every cell in an individual is the same, the epigenetic context for that DNA varies from cell to cell. In this way, these epigenetic differences play a crucial role in gene regulation, with epigenetic changes both causing and recording regulatory mechanisms. Specifically, the two main types of epigenetic modifications are 1) DNA methylation, a chemical change applied directly to the DNA, and 2) histone modifications, chemical changes applied to the histone proteins that the DNA is wrapped around, as illustrated in Fig. 1-2.[11,12]

Therefore, we can look to epigenomics, the study of genome-wide epigenetic changes, to explain cell type differences. The study of epigenomic differences between cell types has led to insight into dynamic regulatory processes and cell type

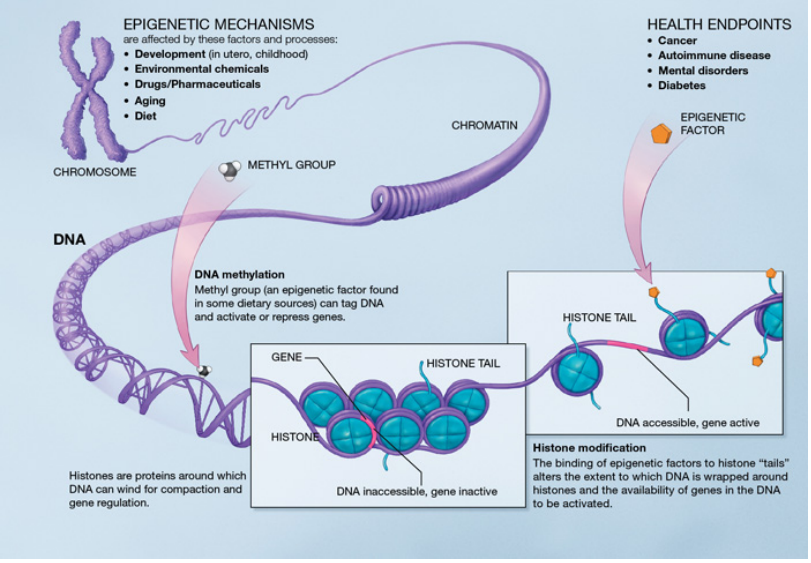

Figure 1-2: Epigenetic modifications include methylation of the DNA strand (left) and chemical modifications to the tails of the histone proteins (right).[2]

differentiation.[10,13–15] Furthermore, epigenetic changes have also been shown to vary due to genetic sequence[16] and individuals.[17]

In this thesis, we combine genetic, epigenetic, and expression information to fill in missing links between genetic variation, epigenetic changes, phenotypic variation, and gene regulation. To do this, we develop and apply computational methods to further our understanding of epigenomic regulatory mechanisms and variation. Specifically, we approach this problem with three main aims: 1) to identify gene family-specific epigenetic modifications and corresponding regulatory mechanisms; 2) to develop computational algorithms that identify meaningful epigenetic variation genome-wide; and 3) to better understand the relationship between epigenetic changes, gene regulation, and the resulting phenotype in different biological scenarios.

1.2 Background

1.2.1 Gene expression: central dogma of molecular biology

The process of gene expression, or the conversion of gene information into a gene product, makes up the central dogma of molecular biology. As illustrated in Fig. 1-1,

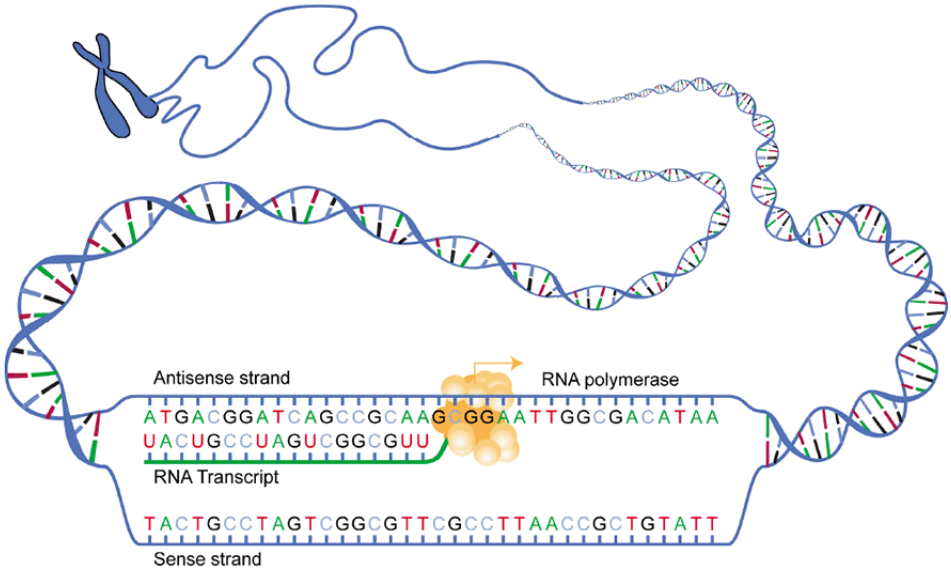

Figure 1-3: RNA Polymerase reads the nucleotides of one strand of the DNA to produce the complementary RNA in a process known as transcription.[3]

this process is classically broken up into two steps: transcription and translation. The starting point is the DNA (Deoxyribonucleic acid) in our cells, which is made up of four types of nucleotides: Adenine, Guanine, Thymine, and Cytosine. This encoding of DNA information is ideal for computational analysis, as these nucleotides act as the "bits" the make up the genetic "code". Furthermore, due to Hydrogen bonding, there is a natural pairing of these nucleotides. Specifically, Adenine and Thymine complement each other, while Guanine and Cytosine complement each other. In this way, DNA provides a redundant and robust encoding of our genetic information.

Through the process of transcription, the protein RNA Polymerase unzips the double-sided DNA to "read" the individual nucleotides. Using this information, it produces the complementary single-sided mRNA (messenger ribonucleic acid), as shown in Fig. 1-3. The mRNA is produced using this same pairing template, except with the exception of Adenine pairing with Uranine rather than Thymine.

For the translational step, the cellular machinery reads the mRNA three bases at a time (each triplet of nucleotides is called a codon). Each of these codons maps to one of 22 amino acids, as shown in Fig. 1-5. As the machinery reads each codon, the corresponding amino acid is added to the amino acid chain using complementary base

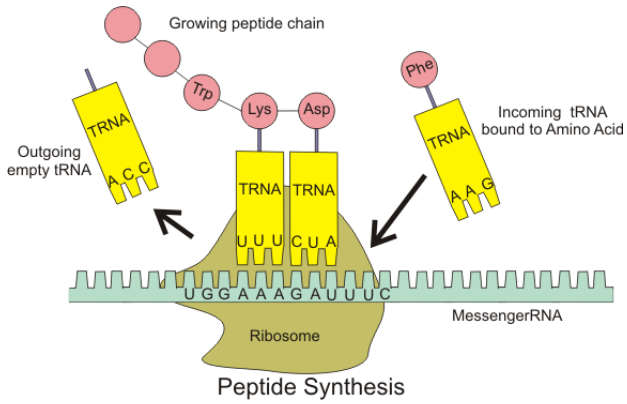

Figure 1-4: The process of translation reads mRNA to assemble a chain of amino acids to form a protein.[4]

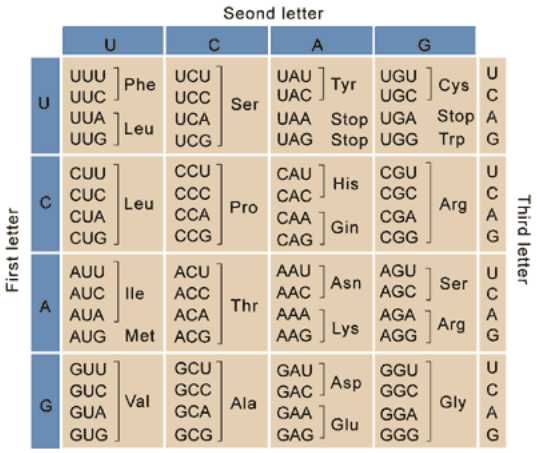

Figure 1-5: Triplets of mRNA nucleotides act as codons to map to amino acids.[5]

pairing with tRNA (transfer ribonucleic acid), in this way constructing the complete protein, as shown in Fig. 1-4.

1.2.2 Regulation of gene expression

While the overall process of producing proteins from the genetic code is considered gene expression, controlling the time period or quantity in which genes are expressed is often referred to as gene regulation.

The underlying mechanisms of gene regulation are complex and vary widely in different contexts. While the genes provide the genetic "code" necessary for biological processes, gene regulation acts as the "control" level. Just as computer programs

must decide which sub-functions to run in which contexts, gene regulation ensures that specific genes are expressed in specific cell types during specific time points: this enables the same initial stem cells to differentiate into the hundreds of distinct cell types in an adult human.

Gene regulation can imply that expression of a gene is increased or decreased, and different types of gene regulation can occur at different points along the path of gene expression. Since gene expression is the act of transcription (DNA to RNA), followed by translation (RNA to proteins), some mechanisms of gene regulation occur at the transcriptional level, while some occur at the post-transcriptional level.

1.2.3 Functional and regulatory regions

While the model of gene expression focuses on one particular gene, each gene makes up only a tiny fraction of the entire set of genetic information in an organism, which is known as a genome. Through the advent of next-generation sequencing technologies, we are now able to gather information about the entire genome and its environment.

However, with this increased amount of data, we also have the new challenge of discerning which regions of the genome are functional, important, and meaningful. While protein-coding genes are clearly important due to their role in gene expression, many other genetic sequences and regions also play a role in gene expression and regulation, as shown in Fig. 1-6.[18]

For example, promoter regions are regions at the the beginning of genes at which proteins bind to for initiation of transcription, which effects gene regulation at the transcriptional level. Enhancer regions, on the other hand, can influence the transcription of genes that are distant in terms of the numbers of nucleotides, by being close in 3-dimensional space. Regions where transcription factors bind can also influence transciptional regulation through the effects of the transcription factors.[18]

The genome-wide annotation of these regions is possible by combining information about evolutionary conservation, experimental assays, and computational methods for pattern recognition and inference.[9,11,18–21] Identification of these regions can elucidate biological mechanisms and pathways, narrow the focus and increase power for

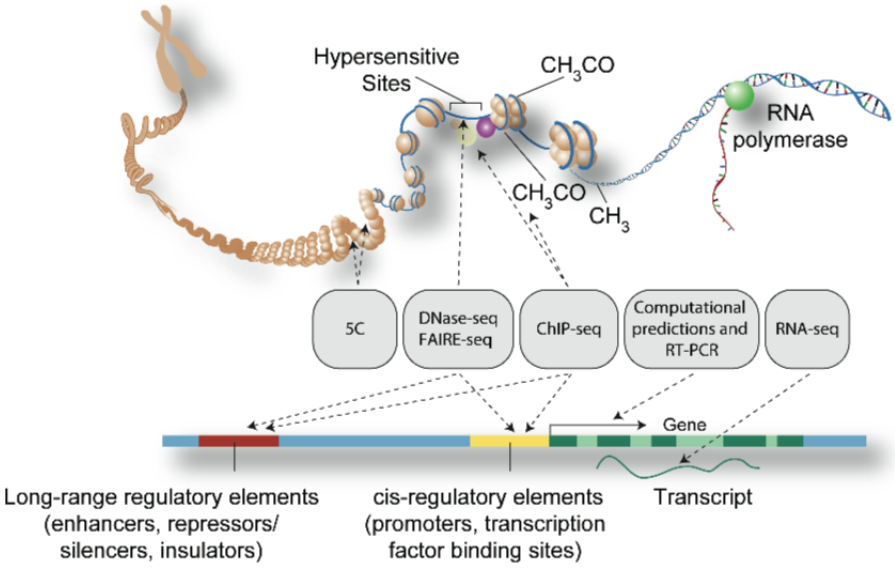

Figure 1-6: Functional and regulatory genomic elements can either lie close to the relevant gene, as in cis-regulatory elements, or be distant in nucleotide space, as with long-range regulatory elements. The identification of these regions is accomplished with a variety of experimental and computational techniques.[18]

computational models, and improve interpretation of biological findings and signals.

1.2.4 Epigenetic state and modifications

Epigenetic modifications provide one perspective that we can use for an improved understanding of functional regions and gene regulation. Specifically, epigenetic modifications are chemical changes to the environment of DNA and generally fall into two categories: 1) DNA methylation, which directly occur on the DNA strand, and 2) histone modifications, which are chemical changes to histones, which are the proteins that DNA is wrapped around, as shown in Fig. 1-2.

The overall epigenetic state may control and/or record gene regulation in different circumstances, and they can be heritable across generations and dynamic across cell types. On a cellular level, epigenetic state can play a causal role in the regulation of genes - for example, a modification might serve as a "sign" that the surrounding genes should be expressed. On the other hand, the epigenome might show the history of how the genome has been used through different developmental stages; just like hunters

can find clues about nearby animals through tracks in the dirt, scientists can see the history of a cell by observing the locations and types of epigenetic modifications.

Therefore, epigenomics, which specifically studies epigenetic state on a genome-wide scale, can facilitate discoveries of large-scale patterns of gene regulation, such as regulation of entire gene families or differentiation patterns for a cell type, as we will explore in this thesis.

1.2.4.1 DNA methylation

DNA methylation is the addition of a methyl (-CH3) group to DNA, as shown on the left side of Fig. 1-2. Specifically, two of DNA's four nucleotides, Cytosine and Adenine, can be methylated. In this thesis, we will focus on mammalian organisms, in which only Cytosine can be methylated. Furthermore, in mammals, methylation primarily occurs at CpG dinucleotides - that is, when a Cytosine and Guanine occur next do each other in DNA.[12] Based on the patterns of DNA methylation in mammals, there are a number of possible experimental techniques to quantify DNA methylation across the genome, ranging in the type of methylation they capture, as well as the scope of sites they quantify.[22]

1.2.4.2 Nucleosome positioning

DNA is tightly wrapped around protein sets called nucleosomes, analogous to how yarn is wrapped around a spool. These nucleosomes are octamers of histone proteins, and the combination of nucleosomes and the DNA wrapped around it is called chromatin, as illustrated in Fig. 1-7.

Nucleosome positioning can also play an epigenetic role in pre-transcriptional gene regulation. Specifically, regions of the DNA that wrap around nucleosomes are less accessible and more closed to transcription factors. On the other hand, the regions of DNA that link the nucleosomes are more accessible and open to transcription factors. The state of the DNA being more or less accessible due to nucleosome positioning is often referred to as an "open chromatin state" or "closed chromatin state," respectively. In general, it has been shown that chromatin states are often

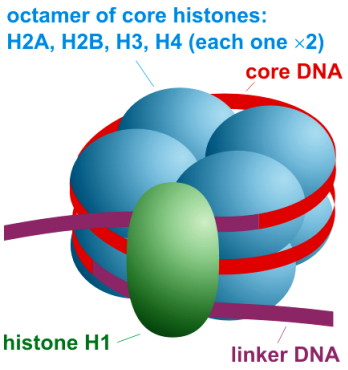

Figure 1-7: A nucleosome is made up of 8 histone proteins and has DNA wrapped around it. The combination of the DNA and the nucleosome is called chromatin.[6]

correlated with the transcription state of the corresponding genes; they can act as instructions for the genes present in the surrounding DNA, or they can record the "history" of the transcriptional state.

1.2.4.3 Histone modifications and chromatin state

Histone modifications are chemical changes made to either the core or the long tail of certain histone proteins in the histone octamer. The naming mechanism of histone modifications provides detail about the location and type of modification. There are five major classes of histones, and the name of the histone modification starts with the class of histone (e.g. H3). This is followed by the single-letter amino acid abbreviation, such as K for Lysine, and the number for the position of the amino acid in the protein. The final part of the naming procedure is the type of modification that was applied to the amino acid, such as Me3 for trimethylation. Examples of these modifications are illustrated in Fig. 1-2.

Similarly to nucleosome positioning, histone modifications can also record or control gene regulation. Furthermore, combinations of histone modifications have been able to provide a more detailed and nuanced multi-state description of epigenetic state. Methods for combinatorial chromatin state annotations include ChromHMM,[23] Segway,[24] and HMMSeg,[25] which use underlying combinatorial techniques such as Hidden Markov Models and Bayesian networks. The resulting analyses enabled by chromatin state analysis has provided fruitful findings about epigenomic variation

and lineage-specification.[17,26–30]

1.2.5 Epigenetic variation

One way to grasp the power of epigenetics is to note that the DNA in each cell of an organism is nearly identical, but the types of cell in an organism vary greatly in function, appearance, response, and activity. As different cell types have different epigenetic state, the study of epigenetics on a genome-wide scale is a natural tool to study cell type variation. Cell type variation is achieved through the process of cell differentation, where pluripotent stem cells can turn into highly-specific cell types. By comparing cells at increasingly progressed stages of differentiation, we can observe what epigenetic changes co-occur with differentiation stages, thereby providing mechanistic insights into differentiation.

In addition, epigenetic differences can also be studied at many other levels, such as variations across genotypes, individuals, anatomical groups, and tissue types, which we will also explore in this thesis. As epigenetic state is also influenced by the underlying genetics, care must be taken in unraveling genetic influences from other factors.

1.2.6 Human biology and model organisms

Much of the motivation underlying the study of these biological phenomenom lies in improving our understanding of human biology for the eventual betterment of human health. With that goal in mind, studies of human samples[19,31,32] are hugely important when possible. A number of analyses presented in this thesis will be done directly on human samples, whether they be cell lines, primary cells, or primary tissues.[11]

However, studying model organisms, such as yeast and mice, in addition to studying humans, has proven to be an incredibly powerful technique. As there are obvious ethical limitations on human experimental techniques, studying these model species with a larger toolbox of techniques can reveal biological findings that can then be confirmed in humans. Similarly, if the interactions of factors in human are too complex to

immediately unravel, some model organisms, such as yeast, provide similar but simpler systems that are a crucial stepping stone for understanding humans. Therefore, in this thesis, we will also present work based on mouse samples, as well as validation through in-vivo follow-up experiments.

1.3 Relevant experimental methods

As outlined above, genome-wide studies of genetic and epigenetic effects are possible due to the sequencing of the human genome,[33,34] as well as next-generation experimental techniques.[35,36] Here, we will outline some of the techniques most critical to our research.

1.3.1 DNA sequencing and mapping

The Human Genome was initially completed in 2004 through Sanger Sequencing,[34] which also spurred the development of faster and cheaper next-generation sequencing technologies.[35–37] These technologies take advantage of the complementary base-pairing of nucleotides. Specifically, the input to the sequencing process is fragments of DNA, and by using adaptor and primer sequences, in combination with dNTPs, flourophores, colour-coding, complementary base-pairing, and imaging, these technologies can produce "reads", the nuleotide sequence of these input DNA fragments.[37]

It is worth noting that after the reads have been generated, they still need to be assembled together, as each read is only a sequence fragment on its own. However, since most non-repetitive reads have a distinctive sequence, these reads can be combined back together or mapped back to the corresponding part of the genome. (Note that reads that map to repetitive sequence will have a much harder time being assembled, due to their non-distinctive sequence, as shown by the pink read in Fig. 1-8.) Assembling the reads can be done either with a reference genome, referred to as "read mapping", or without a reference genome, which is called *de novo* assembly.

When using a reference genome, the read sequence can be compared to the entire reference genome sequence, as shown in Fig. 1-8. While allowing for sequence mis-

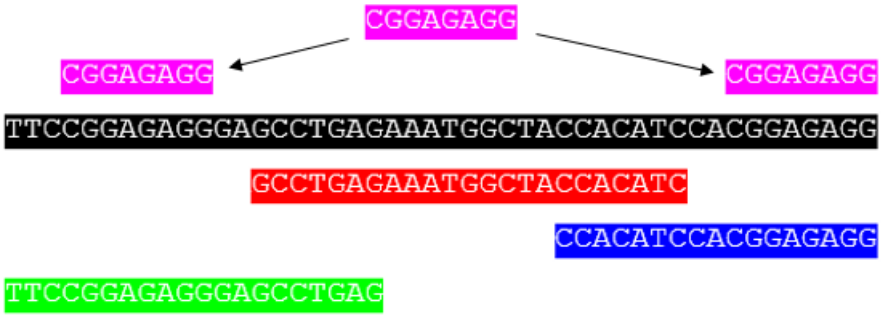

Figure 1-8: The fragments of DNA sequence (reads) can be "mapped" back to a reference genome (black) by comparing the sequence of nucleotides, as shown by the green, red, and blue fragments. However, if the read corresponds to a repetitive or common sequence, as is shown by the pink read, it will be much more difficult to map.[38]

matches due to variation or sequencing error, computational methods can find the best sequence match. Then, the read can be mapped to that location of the genome, as shown by the red, blue, and green reads in Fig. 1-8.

On the other hand, if no reference genome is given, the genome can be newly, or *de novo* assembled, as shown in Fig. 1-9. This is possible because there are multiple copies of the same genome to begin with; since the fragments are cut up randomly, many of the fragments should overlap each other. By using computational methods to identify overlapping sequence from different reads, the reads can be pieced back together like a puzzle, as shown in step 2 of Fig. 1-9, producing a final genome sequence.

1.3.2 Genotyping

In addition, to identify genetic variation, it can be more cost-effective to use array technologies, such as single-nucleotide polymorphisms (SNP) with microarrays,[40] rather than sequencing technologies. These "genotyping" approaches check for predetermined regions and known types of variation, rather than identifying large consecutive sequencees of the genome, as the proportion of the genome that varies among humans is less than 1

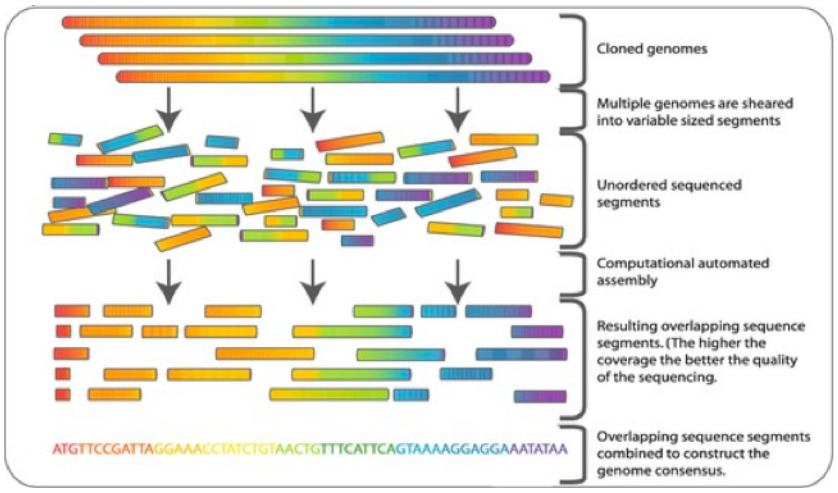

Figure 1-9: For *de novo* assembly, the overlapping parts of different fragments can be pieced back together, like puzzle pieces, to assemble a new, complete genome sequence.[39]

- The substitution of one nucleotide for another nucleotide, also called a single-nucleotide polymorphism (SNP)
- The insertion of a genetic sequence
- The deletion of a genetic sequence
- A larger genomic region that is present an abnormal number of times in the same genome, also called copy number variation (CNV)

Much of the genotyping efforts have focused primarily on SNPs, as they are the easiest to identify and the biological phenomenom of Linkage Disequilibrium (LD) allows biologists to use a SNP to infer much of the surround genetic information with a high probability. To identify the nucleotide at a particular SNP, scientists can use a "minisequencing" approach, utilizing many of the same techniques as sequencing,[40] but only identifying a single nucleotide "minisequence". It is worth noting that this approach works even if the researcher is interested in a particular genomic position, but does not know what the probable underlying nucleotides are.

However, in practice, the vast majority of people will have only one of two possible nucleotides at a particular SNP, and these two nucleotides are called the "alleles" for that SNP. (For example, in Fig. 1-10, the top sequence has the A allele, while the bottom sequence has the T allele.) Therefore, other experimental techniques can take

Figure 1-10: The use of allele-specific oligonucleotides allow only the perfect complementary sequence to completely bind to the present allele (top), while the complementary sequence for the mis-matched allele will bind more weakly (bottom). After washing away the probes with weaker binding, only the perfect complementary oligonucleotide (top) will remain. Then, through the use of a detectable tag on the probe, the underlying allele(s) can be identified based on the remaining probe(s).[41]

advantage of knowledge of the underlying alleles. Specifically, since scientists know the two probable nucleotides that can occur at a particular SNP, they can generate sequences of nucleotides with the appropriate complementary nucleotides.

These allele-specific sequences can be used as starting points, or "primers", for sequence extension. After the primer binds to the corresponding sequence, extension of the primer will only occur if the allele is a perfect match. In this case, dNTPs (both tagged and untagged) can be combined with flourescence imaging to identify which allele-specific primers are extended, since locations with flourescence can be mapped to the underlying allele-specific primer.[40]

Similarly, these allele-specific sequences can also be used as "probes" by tagging them with a radioactive, enzymatic, or fluorescent tag. Since perfectly complementary allele-specific oligonucleotides bind more strongly to the present allele, as shown in Fig. 1-10, the probes with mismatches can be washed away, while the perfect matches can remain due to stronger binding. As a result, the presence of the underlying allele(s) can be identified through measurement of the tags on the remaining allele-specific sequences.

1.3.3 Chromatin immunoprecipitation

As mentioned above, chromatin is the combination of the nucleosome and the DNA wrapped around it. Chromatin immunoprecipitation (ChIP) allows researchers to identify what regions of the genome have a certain histone modification, among other uses. Specifically, the ChIP protocol isolates out DNA fragments that are bound to specific types of proteins through approximately four steps, as shown in Fig. 1-11. The first (optional) step is to bind the proteins with DNA (if this is not already done). Next, the DNA is cut up into fragments, also known as shearing. Then, by using an antibody that specifically recognizes a protein of interest, only proteins of interest and their attached DNA are isolated. Finally, the DNA fragments are separated from the proteins, producing the final DNA fragments.

To identify the underlying sequence and genomic location of the produced DNA fragments, sequencing or microarray technologies can be used. When combining the ChIP protocol with sequencing technologies, as described in Section 1.3.1, the number of reads that map to each nucleotide of the genome provide a continuous signal of enrichment across the genome. On the other hand, microarray technologies can be used when the scientist is only interested in pre-determined regions of the genome, similarly to the techniques described in Section 1.3.2. In this case, the isolated DNA fragments are tagged with flourescence and washed over a microarray chip that contains a matrix of probes. Based on the coloring of the cells on the chip, one can identify the intensity of the signal for each probe. Since each probe sequence maps back to the genome, this results in an enrichment signal for each probe region.

1.3.4 DNA methylation profiling

Multiple experimental techniques to identify regions of DNA methylation have been developed, though they range in the type of methylation they capture, as well as the scope of sites they quantify.[22]

For example, bisulfite-based methylation experiments use the chemical bisulfite to convert unmethylated cytosines but not methylated cytosines to uracil,[43] as shown

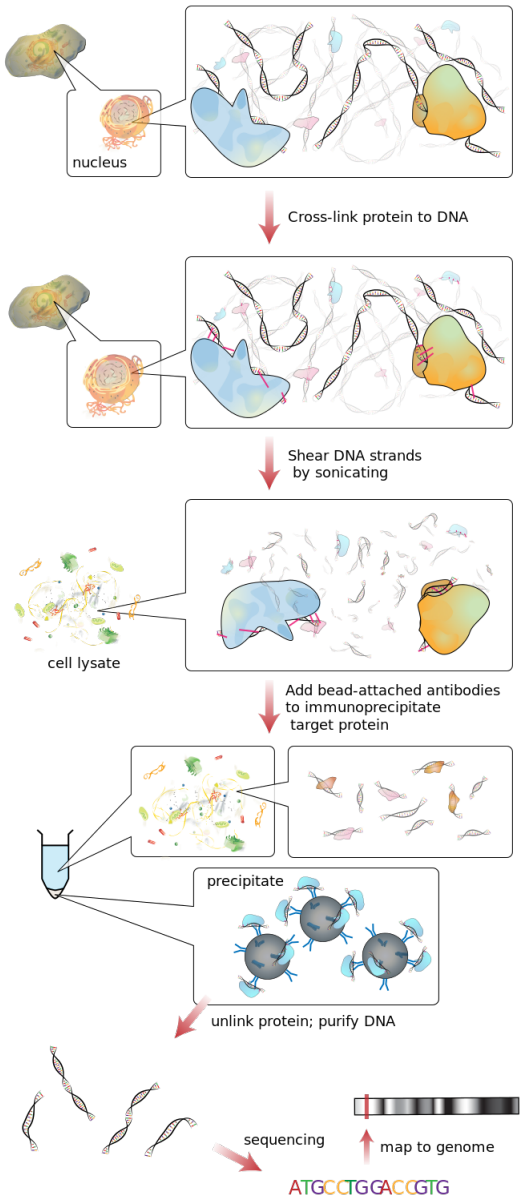

Figure 1-11: Chromatin Immunoprecipitation (ChIP) first cuts up the DNA into fragments, and then utilizes an antibody that detects and "pulls down" a certain type of protein. In this way, it only retains DNA fragments that were attached to the protein of interest, resulting in identifying genomic regions with a certain kind of chromatin or DNA-binding. As a result, it can be used to identify regions of histone modfiications, as well as transcription factor binding sites.[42]

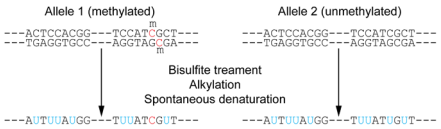

Figure 1-12: Bisulfite-sequencing-based experimental techniques directly measure methylated CpG sites in the genome. Specifically, bisulfite conversion changes any unmethylated Cytosine upstream adjacent to a Guanine into a Uracil, while methylated Cytosines upstream adjacent to Guanines are preserved as Cytosines. Then, the resulting converted nucleotides are sequenced, and by measuring the proportion of Cytosines to total Cytosines and Guanines at each CpG site, we can calculate the proportion of DNA methylation that occurred at that site on a 0 (never methylated) to 1 (always methylated) scale.[44]

in Figure 1-12. As a result, sequencing of the resulting reads allow us to identify the location of uracils (which represent unmethylated cytosines) and cytosines (which represent methylated cytosines).

Whole genome bisulfite sequencing is considered the "gold standard" of DNA methylation technologies, as it directly converts any unmethylated Cytosine in a CpG site into a Uracil, while preserving methylated Cytosines, on a genome-wide scale, resulting in both precise base-pair resolution and wide coverage. However, the cost of a single WGBS experiment can be prohibitively expensive. Therefore, other bisulfite assays, such as RRBS (Reduced Representation Bisulfite Sequencing) focus on the methylation state of pre-determined CpG sites for a lower costLi:2015aa. These options highlight a common trade-off in genomic experiments, where experimental value is determined by a combination of cost and thoroughness.

Additionally, enrichment-based assays such as MeDIP-seq (methylated DNA immunoprecipitation sequencing) and MBD-seq (methylated DNA binding domain sequencing) take an alternative approach of enriching methylated regions of DNA;[43] these can also be combined with techniques such as MRE-seq (Methylation-sensitive Restriction Enzyme digestion-sequencing), which enrich for unmethylated regions to generate a more complete picture.[45] Then, computational approaches such as methyl CRF can combine these experiments (specifically, meDIP-seq and MRE-seq) to computationally predict DNA methylation values at a base-pair resolution.

Comparisons of DNA methylation assays have revealed considerations such as cost,

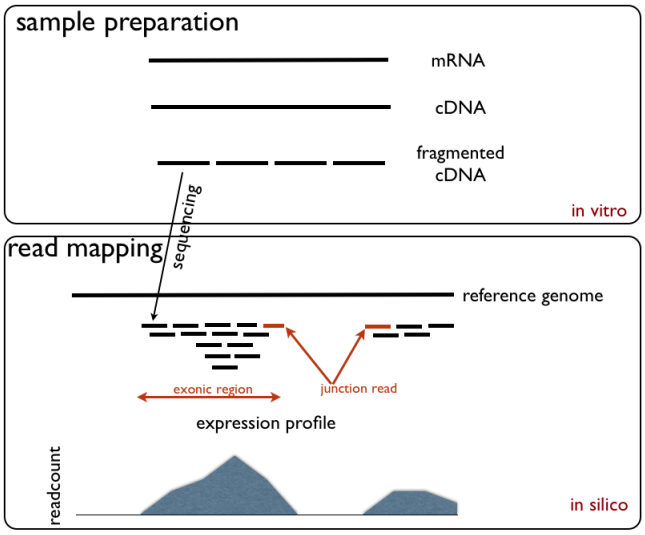

Figure 1-13: RNA-sequencing technology measures gene expression. RNA-Sequencing starts with RNA from a sample of interest, for which reverse transcriptase generates the complementary DNA (cDNA). After shearing the cDNA into fragments, these fragments can be sequenced into reads (top). Lastly, the reads are computationally mapped back to the reference genome to produce a continuous signal across the genome (bottom).[46]

coverage, resolution, concordance, and quantification.[45] This suggests that the ideal experimental assay depends on the cost considerations, planned downstream analyses, and biological context. In this thesis, we will also explore concordance between DNA methylation platforms.

1.3.5 Transcriptome sequencing

One important experimental technique to study gene regulation is RNA-seq, which allows us to measure the expression of a gene by quantifying levels of mRNA. This method does not perfectly capture all aspects of gene expression, as levels of mRNA may not perfectly correlate with amount of resulting protein due to post-transcriptional regulation. However, it does provide an informative genome-wide quantification of gene activity.

Specifically, RNA-Seq is another technique that takes advantage of next-generation sequencing technologies to profile the transcriptome.[47] In the past, genomic tiling mi-

croarrays were common to approximate the transcriptome, as they were high throughput and relatively inexpensive, and could reach a high resolution with specialized chips.[48–50] However, drawbacks of genomic tiling microarrays include assumptions about the genomic sequence, problems of cross-hybridization and complicated normalization.[47] As a result, RNA-Seq has quickly become the dominant method of transcriptome profiling.

The RNA-seq protocol first converts a population of RNA into a library of complementary DNA (cDNA) fragments. Since mRNA in eukaryotes typically end with a long sequence of Adenosine nucleotides (also called a "poly-A tail"), a complementary sequence of many T's can be used as a poly-T primer to bind to the poly-A tail. (Alternatively, random hexamers can also be used as primers.) By adding reverse transcriptase, the transcriptase can then use the primer to generate the rest of the sequence complementary to the original mRNA, which is the cDNA.

Then, by cutting up the cDNA sequences into fragments and sequencing the resulting cDNA sequencing, one can obtain reads, as described in Section 1.3.1. Again, the alignment of these sequenced cDNA fragments to the genome results in a genome-wide quantitative measure at the single-nucleotide level for the amount of transcript present. Advantages of RNA-Seq over other transcriptomic methods include single-base precision, no need for previous knowledge about the genomic sequence, low background signal, less RNA sample required at the outset, a larger possible range of expression, high reproducibility, and lower cost.[47]

1.3.6 DNA footprinting

DNA footprinting experiments provide information about the accessibility of DNA. Specifically, one experimental technique that can be used to gain information about nucleosome positioning is DNAse-Seq, which identifies DNase 1 hypersensitive sites. The protocol uses the enzyme DNase I to selectively digest DNA that is not bound to nucleosomes, whereas DNA regions tightly wrapped in nucleosome and higher-order structures are more resistant to digestion. Due to the fact that many regulatory regions operate through binding by transcription factors, these transcription factors

would likely displace nucleosomes, necessitating an "open chromatin state." Therefore, DNase-Seq can generally be utilized as an assay to identify potential regulatory and functional regions.[45]

1.4 Summary of research contributions

My thesis work is to develop and apply computational methods to further our understanding of epigenomic regulatory mechanisms and variation. Specifically, I will approach this problem with three main aims: 1) to identify gene family-specific epigenetic modifications and corresponding regulatory mechanisms; 2) to develop computational algorithms that identify meaningful epigenetic variation genome-wide; and 3) to unveil new insights regarding the relationship between epigenetic changes, gene regulation, and the resulting phenotype in varied biological scenarios.

1.4.1 An epigenomic mechanism for gene family regulation

As multicellular organisms develop from an initial single zygote into a complex system, cellular differentiation turns less specialized cells into more specialized cells. For example, pluripotent cells are unspecialized, and therefore, have the potential to differentiate into any cell type in the organism. Differentiation changes a cellâĂŹs size, shape, activity, and other physical characteristics, largely through the strict regulation of gene activity. This can be especially effective through the coordinated regulation of genes within the same gene family, such as olfactory receptor (OR) genes.

Olfactory receptor neurons, the neurons responsible for our sense of cell, are one type of specialized cell that has a strict âĂŹone neuron - one receptorâĂŹ rule: specifically, each olfactory neuron expresses exactly one olfactory receptor (OR) gene, while all the other OR genes are silenced. This means that each olfactory neuron has the genetic capacity to detect any odor molecules, but the receptors are regulated so every neuron actually detects exactly one smell. The chosen olfactory receptor gene that is expressed in the neuron largely defines the functional essence of that neuron.

The combined power of all the olfactory neurons is what enables the brain to detect a wide variety of smells. In this project, we identified the regulatory role of epigenetic modifications for the monogenic expression of olfactory receptor genes in mice.

Olfactory receptor gene regulation is especially crucial in mice, as their sense of smell is even more discriminating than humans; mice have over 1300 olfactory receptor genes (approximately 5% of their genes), while humans have only about 900 OR genes. Furthermore, mice are biologically very similar to humans, so findings in mice can often be generalized to insights in humans as well. Additionally, mice clearly provide experimental advantages over humans due to limits on data collection for humans. The lifespan of mice, as well as the increased experimental power provided by such a model organism, made mice a clear choice of model organism for this study.

We found that in the mouse olfactory epithelium, OR genes are specifically and sensitively correlated with the histone modifications H3K9me3 and H4K20me3; these marks were much less present in our control tissue, liver. We also found that other familes of chemoreceptors, such as vomeronasal receptors and formyl peptide receptors were also marked with the same histone modifications, although to a lesser degree, suggesting a similar mechanism for those gene families as well.

This epigenetic pattern revealed an epigenomic mechanistic explanation for the monogenic and monoallelic regulation behind OR genes. Specifically, the cell-type and developmentally dependent deposition of these marks along the OR clusters suggests a repressive effect on the genes. Then, these marks are removed at a single OR allele during OR choice, to allow for expression at a single OR gene and allele in each OR neuron.

In contrast to the previous view of OR choice, our data suggest that OR silencing occurs developmentally prior to OR expression, indicating that it is not the product of an OR-elicited feedback signal. In essence, the repressive state is used as a conservative starting state for this strict regulatory mechanism. Overall, this suggests a new role for chromatin-mediated silencing as the molecular foundation upon which singular and stochastic selection can be applied.

1.4.2 A genome-wide computational method for group-wise chromatin state comparisons

These large-scale epigenomic and regulatory comparisons have been shown to provide fruitful insights about gene regulation and other biological processes.[12,14,51] For example, our work described in Chapter 2 was based on the differences of histone mark patterns between liver tissue and olfactory epithelium, providing a specific example of the power of identifying epigenomic differences. Of course, epigenomic variation can be studied not only at the tissue-specific level, but also at the allelic, individual, species, or case-control level, each of which can provide different biological insights.[13,15,16,19,19,52–55,55,56,56–60]

As scientists further discovered the biological importance of epigenomic data, we also began to recognize the computational challenges that it provides. Specifically, epigenomic data includes many types of information, such as the presence of various histone marks, DNA methylation, and chromatin accessibility. These many types of data produce an exponential number of combinations that are necessary to consider, as these marks are likely to complement each other in a complex regulatory logic system.[61–64] To tackle this issue, a number of computational segmentation methods have been developed, utilizing various machine learning and statistical methods, such as hidden markov models and Bayesian networks.[23–25] By generating a "summary" chromatin state from many histone marks, we can perform a type of dimensionality reduction that retains the most important information, while also providing a biological interpretation of the histone mark data.

While progress with epigenomic comparisons and and chromatin state segmentations has been shown to be fruitful, the power of these two approaches have not yet been combined. Specifically, the key question of how to systematically identify chromatin state differences between groups of epigenomes has remained unanswered.

To address this question, we developed ChromDiff, a probabilistic and information theoretic computational method to systematically identify chromatin state differences on a genome-wide scale. To make rigorous comparisons between groups of samples, we

also needed to address the fact that the increasing amount of available data is also, by necessity, being generated in less controlled conditions. Therefore, our methodology integrates correction for external covariates, such as sample type, sex of donor, and production lab, to better isolate the most relevant and meaningful differences. It also leverages both genic and regulatory regions to identify the most relevant features for each comparison. We validated our method by showing that it outperformed existing methods for group-wise epigenomic comparisons, while also proving its specificity with a lack of findings in randomized simulations.

Our method is broadly applicable to study the role of epigenomic variation in various phenotypes, including celltype, anatomy, development, donor sex, or disease. We made it publicly available at `http://compbio.mit.edu/ChromDiff` and `http://github.com/angieyen/ChromDiff` so that future scientific studies may use it to uncover further epigenomic insights.

1.4.3 Epigenomic insights from comparisons based on development age, sex, and tissue type.

With the power of ChromDiff, we were able to study chromatin state changes with the wealth of epigenomic data available from the Epigenomics Roadmap Project. Specifically, we compared diverse groups of epigenomes, including groups based on tissue type, sample state, sex of the sample, and developmental age of the donor. By comparing annotations from the core chromatin state model from the integrative analysis,[11] we identified relevant genes and chromatin states for epigenomic groups, such as ChrX genes for the comparison of female and male samples.

Furthermore, we were able to validate our results by using matched gene expression data, as well as pathways and gene sets. We found that many genes that epigenetically distinguished between the groups did have corresponding gene expression changes, but that many more did not, suggesting that differential gene expression and chromatin state comparisons are powerful complementary approaches. Additionally, identification of enriched gene sets showed the biological relevance of the genes iden-

tified by ChromDiff, with blood comparisons isolating gene sets related to leukemia, a comparison of brain comparisons highlighted genes related to Alzheimer's disease and oligodendrocyte differentiation, while fetal samples were distinguished from adult samples with gene sets relating to fetal differentiation and Alzheimer's disease. Similarly, comparisons based on linked regulatory regions show different enhancer activity in blood cell subtypes, as well as a broad heterochromatic signature in female cells due to X Chr inactivation.

Overall, our results highlight the important genes and epigenomic states that can be identified using chromatin state comparisons of groups of epigenomes. In this way, the study of statistically meaningful chromatin state patterns in groups of epigenomes provides biological insights relating to celltype, gene regulation, and development.

1.4.4 Integrative analysis of DNA methylation data across chromatin state, gene expression, platform, and celltypes.

Epigenomics allows us to study the dynamic markings and states surrounding the DNA which varies not only across individuals, but also across celltypes. For the Epigenomics Roadmap project, we employed computational and data science approaches to detect meaningful patterns in a wide variety of epigenomic datasets across 127 cell types. As epigenomics includes the study of many chemical states surrounding DNA, including DNA methylation, histone makes, and DNase footprinting, we also integrated data across many experimental assays to identify relationships between epigenomic marks.

With this approach, we uncovered a number of findings relating to DNA methylation, especially focusing on how DNA methylation varies across the genome based on chromatin state and celltype. For example, we identified the DNA methylation state of chromatin states, showing that active promoter and bivalent states tend to be hypomethylated, while transcriptional states tend to be hypermethylated, as is consistent with existing literature.[65] However, we also found that enhancers tended

to have variable DNA methylation states, and that DNA methylation in the same chromatin state varied across cell types and proliferation status.

We quantified these patterns by comparing the DNA methylation of real chromatin state regions with that of randomized genomic regions across the genome that were matched for size and number. Based on the resulting distributions, we identified chromatin states with distinctive DNA methylation values compared to the genomic background, such as active promoter states and bivalent regions. On the other hand, we also identified that chromatin states such as quiescent, weakly polycomb repressed, and weakly transcribed regions had DNA methylation patterns most similar to random genomic regions.

Importantly, we also quantified the correlation and differences between multiple experimental technologies used to gather DNA methylation measurements. Specifically, we identified high consistency between methylation values given with bisulfite-sequencing based experiments, while a computational inference technique using enrichment-based assays resulted in weaker correlation and directional biases. As a result, we also generated platform-specific DNA methylation distributions for chromatin states, showing that platforms affected results in both variance and median value.

Lastly, we demonstrated our ability to order and cluster samples based on epigenomic similarity, both by looking at DNA methylation state in various enhancer modules, and calculating pair-wise similarity between epigenomes based on various histone marks. In sum, we presented biological hypotheses and patterns from our data-driven integrative analysis of multiple large-scale experimental datasets.

1.4.5 Identification of unknown covariates in epigenomic samples with mutual information analysis against true covariates.

With the pressing need for large-scale genomic data, many collaborative efforts have resulted in public resource datasets that researchers around the world can use to supplement or drive their own research studies. However, these datasets also present

new issues, due to their generation in less consistent and controlled experimental circumstances. For this reason, covariate correction such as the logistic regression integrated into our ChromDiff method is crucial to meaningful and fruitful analyses.

However, metadata describing the experimental and sample characteristics may be inconsistent, poorly documented, or unavailable, especially when combining multiple datasets. Here, we utilize principal component analysis to potentially identify the most important unknown covariates, inspired by the use of PCA in genetic studies to identify and correct for signal due to the population or ethnicity.[66] Then, we use our true known metadata to identify the relationship between the "unknown" principal component covariates and the known metadata covariates, using Component Selection Using Mutual Information.

We find that the top principal components map closely to covariates relating to sample group, cell type, and anatomy, confirming our previous results that clustering samples based on epigenomic signal primarily drives groups based on celltype. We additionally find that less highly-ranked covariates often share large amounts of mutual information with covariates based on processing lab, composition of sample, and developmental age of donor. Overall, we can conclude that these covariates play an important role in driving the signal of these epigenomic datasets, and to a large extent, using dimensionality reduction methods such as PCA can produce covariates that do a reasonable job as a "stand-in" covariates for known metadata. However, we also find that top principal components explain only a small proportin of the overall variance in the data. This may mean that much of the signal of interest will remain after correcting for the top principal components, but it may also indicate that, depending on the property determining the epigenomic groups, it may be better not to correct for principal components, especially when there is a risk of removing the signal of interest.

Chapter 2

An epigenomic mechanism for regulation of the olfactory receptor gene family

In olfactory neurons, there is a strict rule that each neuron must express exactly one allele of one of the 1300 olfactory receptor genes. However, the mechanism behind this monogenic expression is not yet fully understood. In this chapter, I will use olfactory receptor genes as an example gene family to identify epigenomic mechanisms for gene family-specific regulation.

Specifically, in the olfactory epithelium of mice, olfactory receptor genes are marked in a highly dynamic fashion with the molecular landmarks of constitutive heterochromatin. The cell-type-dependent deposition of H3K9Me3 and H4K20Me3 along the clusters of OR genes is differentiation-dependent, and these marks are most likely reversed during the process of OR choice for monogenic and monoallelic expression. In contrast to the previous view of olfactory receptor choice, which suggested that the silencing of the OR genes results from a feedback signal initiated by OR gene expression, our data suggests that OR silencing takes place before OR expression. This implies a new molecular role of chromatin-mediated silencing as the foundation upon which singular and stochastic selection can be applied, shown here in OR genes, but generally applicable.

2.1 Introduction

2.1.1 Problem Statement

In this chapter, we will analyze epigenomic data to reveal an epigenomic mechanism to explain regulation of a gene family. Specifically, olfactory receptor (OR) genes are the genes that code for the receptors that detect smells. OR gene regulation is a topic of general interest, as OR genes are regulated in an unusual way: specifically, in each olfactory neuron, exactly one OR gene is expressed, while all the other OR geners must be silenced. This means that while each neuron has the genetic capacity to detect any smell, the receptor genes are regulated so every neuron actually detects exactly one smell. The combined power of all the neurons enable detection of a variety of smells.

The sense of smell is especially important to mice, as they are scavengers by nature, and they must take advantage of their powerful sense of smell to find food. Furthermore, mice are a well-studied model organism for humans, with many genetic similarities that allow findings in mice to often be applied to humans. Of course, with more experimental options for mouse than for human, it made mice an obvious choice for our study.

This project was a partnership with Prof. Stavros Lomvardas's group of UCSF's Neuroscience Department, and we worked together to discover and understand what the mechanism is behind monoallelic and monogenic olfactory receptor genee regulation in mice.

2.1.2 Background and previous work

2.1.2.1 Olfaction

Olfactory perception, or the sense of smell, takes place through the detection of volatile chemicals in the olfactory epithelium; the detection of these chemicals is then transmitted to the brain, which processes the information. In contrast to other

sensory systems, olfaction requires a large family of 1000 OR genes olfactory receptor (OR) genes, and these genes undergo a strict "one neuron-one receptor" rule.

That is, olfactory sensory neuron (OSN) are responsible for the detection of odors through olfactory receptors, and in each OSN, exactly one allele of one OR gene is expressed.[67,68] This means that each olfactory neuron can detect exactly one kind of odor, dependent o n which of the ~1000 olfactory receptors is expressed. Once OSNs detect the chemicals, they transmit signals through their axons to the olfactory bulb, the region of the brain involved in olfaction. The axons of olfactory neurons that express the same receptor meet up in the same glomerulus, a spherical structure in the olfactory bulb;[69–71] this is possible because the ORs play a role both in odor detection, as well as guiding the axons to the proper glomeruli, effectively determining the OSN's identity in this way.[72–75] As ORs are important in both the wiring and physiology of olfaction, their proper expression is especially crucial.

The monoallelic and monogenic expression of OR genes is a difficult task: exactly one allele must be expressed at high levels, while the other ~1000 genes must be kept silent. The repression of the non-chosen OR genes must be extremely effective, since even a low level of transcription would result in thousands of inappropriately expressed OR molecules, due to the high number of OR genes; each individual receptor type would have low representation, the total OR activity of non-chosen alleles could be comparable to the activity of the chosen allele, possibly resulting in sensory confusion.

2.1.2.2 Previous work on olfactory regulation

In the mouse, about 1400 olfactory receptors are expressed in total in the main olfactory epithelium (MOE); they appear to be organized in a spatial and temporal fashion determined by positional clues.[70,76,77] Within each zone of expression, however, there are still several hundred alleles that could be expressed; only one of these alleles is actually transcribed in a seemingly stochastic fashion.[78] Previous experiments have implied that the production of OR protein elicits a feedback signal that prevents the expression of any other OR alleles, while stabilizing the expression of the chosen OR.[79–81] Additionally, the OR coding sequence seems to play an important role

in the OR regulation, as there has been evidence to show that the coding sequence represses heterologous promoters.[82] Furthermore, both enhancers and promoters contain regulatory information.[80,83] In the past, the Lomvardas lab had shown that a specific enhancer, the H enhancer, interacts with active OR alleles, suggesting that this enhancer might be instructive for OR expression.[84] However, genetic ablation of the H enhancer only disrupted the expression of three proximal ORs, which makes it unlikely that it is singularly responsible for orchestrating OR choice.[85,86] Therefore, the overall molecular mechanisms responsible for monoallelic and monogenic gene regulation are still unknown.

2.1.2.3 Chromatin-mediated silencing

Chromatin-mediated silencing is an effective form of transcriptional repression, and transcriptionally inactive chromatin is known as heterochromatin. Facultative heterochromatin is chromatin of silenced genes, and it is generally represented by hypoacetylation and di-methyl or tri-methyl groups on lysine 27 and/or dimethyl groups on lysine 9 of histone H3.[87] Since facultative heterochromatin often silences genes in some environments and not in others, it is dynamic and appears to be developmentally regulated.[88,89] On the other hand, constitutive heterochromatin is usually found in structural regions, such as pericentromeric and telomeric repeats, and it remains tightly packed during the cell cycle and stable during differentiation.[90,91]

2.1.3 Approach

In our project, we tested the hypothesis that chromatin-mediated silencing prevents the expression of OR genes in the sensory neurons. The Lomvardas lab generated Chromatin ImmunoPrecipitation on chip (ChIP-chip) data, which provides genomewide data for presence of epigenetic modifications, as described in Section 1.3.3. I then leveraged computational methods to analyze this ChIP-chip data for quality control, normalization, identification of regions with histone marks, and statistical quantification of significance. Finally, the Lomvardas lab performed additional

experiments to explain and validate our findings.

2.2 Computational methods

To analyze this genome-wide dataset, we utilized computational approaches. First, we processed, normalized, and conducted quality tests on the data. Then, we identified enrichments, patterns, and clusterings through a variety of methods such as sliding windows, k-means clustering, and Hidden Markov Models.

2.2.1 Data processing, normalization, and quality control

2.2.1.1 Quality control

Since ChIP-chip is an experimental method, the possibility of experimental biases or mistakes is always a potential cause for error. Therefore, I generated graphs to assess the quality of each set of data in a number of ways, through a standard set of techniques.[92] ChIP-chip data provides both an amount for ImmunoPrecipitation (IP) - which is the type of DNA the protocol specifically pulled down - and "input" - which is our control; this allows us to compare the distributions of the two sample types and identify enrichments in the immunopreciptated data compared to the control.

Ideally, the distribution of the data is easiest to model if it closely matches a normal distribution, so we visualized the distributions for the IP and the control, to study how normal the distributions appeared. To better quantify how close to normal the distributions are, I used quantile-quantile plots (also known as Q-Q plots), which compares the actual distribution with a normal distribution based on the quantiles of the data.

Additionally, we also used a Mean-Average (MA) plot to identify any potential dependencies between the ratio of IP/input (on the y-axis) and the average of IP and input signal (on the x-axis). This visualization targets the potential problem where ratios of IP/input tend to increase with increasingly strong average signals. Therefore, if this dependency does not exist, the cloud of points would be along a horizontal line,

since the average ratio does not change with the average signal. Finally, I calculated the standard deviation for various signal intensities, again checking for major skews.

2.2.1.2 Dataset normalization

Since ChIP-chip data is an experimental method, noise will inevitably be present in the data; this requires normalization within one set of data, as well as across multiple sets of data, as they are being compared to one another. We utilized a number of different number of normalization techniques to determine which best corrected for our particular experimental noise and biases. I used pre-existing normalization methods such as quantile normalization, which is a conservative normalization that fits the experimental data to a standard distribution, variance stabilization and normalization,[93] which normalizes for the varying intensities of microarrays, and global normalization,[94] which uses the median and standard deviation of log intensity ratios to correct the data for comparison across datasets.

Additionally, I developed a tailored form of normalization to suit our use of replicates and data states, which I call "weighted global normalization;" this method was similar to the standard global normalization, except that it weighted the data for each of the states (H3K9me3 in OE, for example) equally, in spite of how many replicates there are for a given state. Specifically, each sample of data is subtracted by its median and divided by its mean absolute deviation (MAD), as in usual global normalization. This gives us the l_i, the globally-normalized $log(IP/input)$ ratio, where IP_i is the original immunopreciptation signal and $input_i$ is the original input signal, while m_i and d_i are the median and mean absolute deviation, respectively, for dataset i.

$$l_i = (\log(\frac{IP_i}{input_i}) - m_i)(\frac{1}{d_i})$$

Then, we calculate the average median and average MAD within each state. That is, let m_s and d_s represent the average mean and MAD for a single state, where n_s is the number of replicates for state s. Then, in the formula, one sums over all i's for

$i \in S$ where S is the set of indices of the datasets in the state s.

$$m_s = \frac{\sum_i m_i}{n_s}$$

$$d_s = \frac{\sum_i d_i}{n_s}$$

Next, these values are averaged across all four states to obtain the weighted global median m_g and the weighted global MAD d_g. Since we expect the replicates within each state to have a similar distribution, this allows each state to contribute the same weight to the global distribution, regardless of the number of replicates in each state. Therefore, m_g and d_g are calculated as follows, given that n_g is the total number of states:

$$m_g = \frac{\sum_s m_s}{n_g}$$

$$d_g = \frac{\sum_s d_s}{n_g}$$

Then, analogous to global normalization, these "weighted global" statistics are used to scale all the dataset values back through multiplication of the data by the weighted global MAD and addition of the weighted global median. Therefore, X_i, the post-normalized log ratio for dataset i is calculated as follows.

$$X_i = (l_i)(d_g) + m_g$$

2.2.2 Detection of heterochomatic domains

We chose to detect heterochromatin domains through two general approaches: sliding window and hidden markov models (HMMs).

2.2.2.1 Sliding window approach

Since ChIP-chip data gives us an analog signal rather than a digital one, the data must be interpreted into regions that have the presence of the histone modification and regions that do not. Many techniques can be used to turn the probe data into

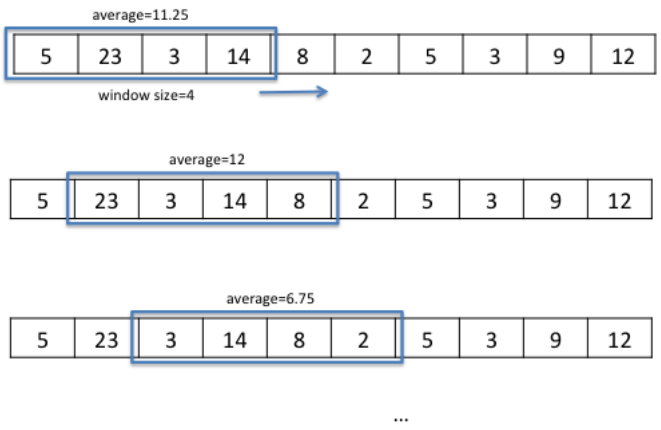

Figure 2-1: This figure demonstrates the sliding window algorithm with an example window of size 4.

finite binary peaks of enrichment. One powerful method for this is the sliding window approach,[95] which slides a window of fixed size across the genome, averaging over the probes present in that window. If the resulting average in that window meets the enrichment threshold, that window is considered a peak, as shown in Fig. 2-1.

Variations on this general approach have also been developed for specific experimental data. For example, the Model-based analysis of 2-color arrays[96] (MA2C) specifically corrected for sequence-specific biases based on GC probe content, as there are experimental biases for regions rich in Guanine and Cytosine nucleotides. Another consideration was the recent identification of large regions of chromatin K9 modifications, or LOCKs,[97] as this suggested that we might find large regions of modifications, or heterochromatin domains, as OR genes are often already clustered together in the genome.

Therefore, while much research focuses on finding narrow peaks via peak-calling, we also specifically searched for broad, large regions of enrichment, or what we call blocks. This was accomplished by using both the MA2C[96] and LOCKsWen:2009aa protocol with adjustments to the parameters for our data and goals.

In the LOCKs methods, averaging was performed across 500 basepair windows, while the minimum block size was 10,000 basepairs. In the MA2C pipeline, we used 2 sets of parameters: one to find smaller "peaks," and the other to find broader

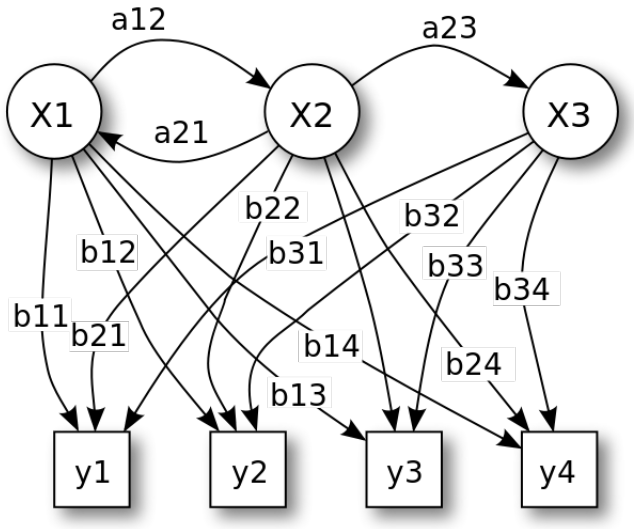

Figure 2-2: Hidden Markov Models are composed of hidden states (circles X1, X2, X3), observed emissions (squares y1, y2, y3, y4), transition probabilities (a arrows), and emission probabilities (b arrows).[7]

"blocks." For the peaks, we used a window to be 500 bp, with a FDR <5%, while the "blocks" were found by using a window of 10,000 basepairs, with a requirement of at least 20 probes in a window, with no more than 1,000 basepairs between adjacent probes.

2.2.2.2 Hidden Markov Model approach

I also used Hidden Markov Models (HMMs) to detect domains of histone modifications.[98] A HMM is a statistical model that is made up of 4 parts: 1) hidden states, 2) observed emissions, 3) probabilities of transition between hidden states and 4) probabilities of observed emissions for each hidden state, as shown in Fig. 2-2. First, in the parameter learning steps, HMM can learn the transition and emission probabilities based on some observed emissions. Then, given those probabilities, the HMM can learn the most probable hidden states underlying observed emissions. These estimated hidden states can then be used to interpret the observed emissions.

In this biological context, the observed emission is the intensity of the ChIP-chip

signal, and we learn either two or three states. These states can roughly be interpreted as enriched and repressed states (for the two-state HMM), or enriched, neutral, or repressed (for the three-state HMM). To train the HMM, we used unsupervised learning with random initializations to train the model on our data, and then we found the assignment of states that maximizes the probability of it being produced by the model.

As a result, the HMM produced "state calls" across the genome, effectively labeling every genomic region as enriched, repressed, or neutral based on the histone modification signal. These regions called as enriched could then be used analogously to the enriched peaks and blocks found by the sliding window approach.

2.2.3 Clustering and ranking of genes

2.2.3.1 Gene body representation

To represent each OR gene, we based our approach on a previous method used to identify histone modifications at human enhancers.[99] Specifically, for each gene and modification, I centered a 2,000 bp window at the translation start site. Each 2kb window consisted of 20 buckets of 100 basepairs each, where every probeâĂŹs $log(IP/input)$ ratio was added to the appropriate bucket, and all values in a bucket were averaged, including data from replicate experiments. Since there were many modifications, the values for each modification - H4K20me3 in OE tissue, H3K9me3 in OE tissue, H3K9me3 in liver tissue, and H4K20me3 in liver tissue - were concatenated.

2.2.3.2 Clustering

Once we had generated the representation for each gene, we wanted to group these genes based on signal similarity across buckets and sample types. To do this, we used a standard k-means clustering algorithm.[100] As shown in Fig. 2-3, k-means clustering begins by randomly initializing k means to k datapoints, where k is chosen by the user as the number of resulting clusters desired. Then, each datapoint is assigned to the nearest mean, and based on the assignments, the overall mean for that cluster is

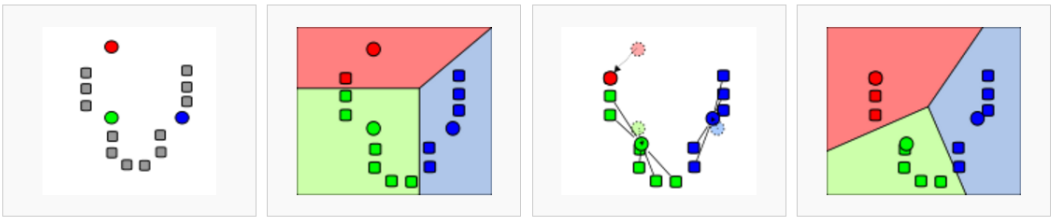

Figure 2-3: In this example of k-means clustering, the first step is to randomly initialize 3 means. (In this example, k=3.) Then, each point is assigned to the nearest of the 3 means, dividing the area into 3 sections. Next, the means are re-calculated based on all the points in each section. Then, these steps of point assignment and mean re-calculation are iteratively repeated until the algorithm converges.[101]

re-calculated. These steps of assignment and re-calculation are repeated iteratively until the algorithm converges. Specfically, the algorithm has converged when the repeating the steps of assignment and re-calculation result do not change the cluster means or assignments.

In this project, K-means clustering allowed us to identify potential patterns in signal across the four states. Cluster 3.0[100] was used to group the genes into four clusters. By tracking which genes were OR genes, we were able to calculate how many OR genes and non-OR genes were in each cluster to identify which clusters captured a signal specific to OR genes. Additionally, the resulting clusters also revealed whether different subclasses of OR genes corresponded with different patterns in histone enrichment signal.

2.2.3.3 Ranking

Lastly, we also ranked the genes by average enrichment for the histone modifications in olfactory epithelial (OE) tissue, to identify which genes were mostly strongly enriched for these heterochromatic marks. To do this, we used the 20 buckets for the gene representation, as described in Section 2.2.3.1, for each OE state (H3K9me3 in OE and H4K20me3 in OE). By averaging across all 40 buckets, and then ranking the genes from highest average value to lowest average value, we generated a sorted list of genes based on histone mark enrichment in olfactory epithelial tissue.

2.3 Results

Our data show that, in the olfactory epithelial tissue, an unusual form of heterochromatic silencing is present at olfactory receptor (OR) genes. Our ChIP-on-chip experiments show a very strong signal for H3K9me3 and H4K20me3 both specifically and sensitively at OR genes in olfactory epithelial tissue. The cell-type and differentiation-dependent presence of these trimethyl histone modifications at clusters of OR genes results in compacted and inaccessible heterochromatic macrodomains.

Surprisingly, these heterochromatic marks are found developmentally before OR transcription, implying that it is not the product of a feedback signal from OR expression. At an active OR allele, we find a significant reduction for the H3K9me3 and H4K20me3 modifications, with a strong signal instead for the H3K4me3 histone modification, often associated with active gene expression.

Lastly, I found that insertion of a reporter transgene within a heterochromatic macrodomain results in OR-like expression of this transgene instead of ubiquitous expression, as the transgene is silenced in most of the olfactory neurons. With this evidence, we believe that stochastic escape from heterochromatic silencing might be the basis of monogenic and monoallelic OR gene expression.

2.3.1 Quality controls

The quality control plots confirmed to us that most of our datasets were of good quality. In Fig. 2-4, we show example plots for dataset 39280702, which includes ImmunoPrecipitated DNA for the H3K9me3 histone modification in mouse olfactory epithelial (MOE) tissue, as well as control Input DNA from the same sample.

In the top row, we find that the IP and Input samples produce distributions fairly close to normal distributions, which is confirmed by the middle row of plots, which show how closely the distributions match a perfect normal distribution, represented by the red lines. In the bottom row, we find that there is not too much of a bias based on signal intensity; when signal intensity increases, this does not significantly change

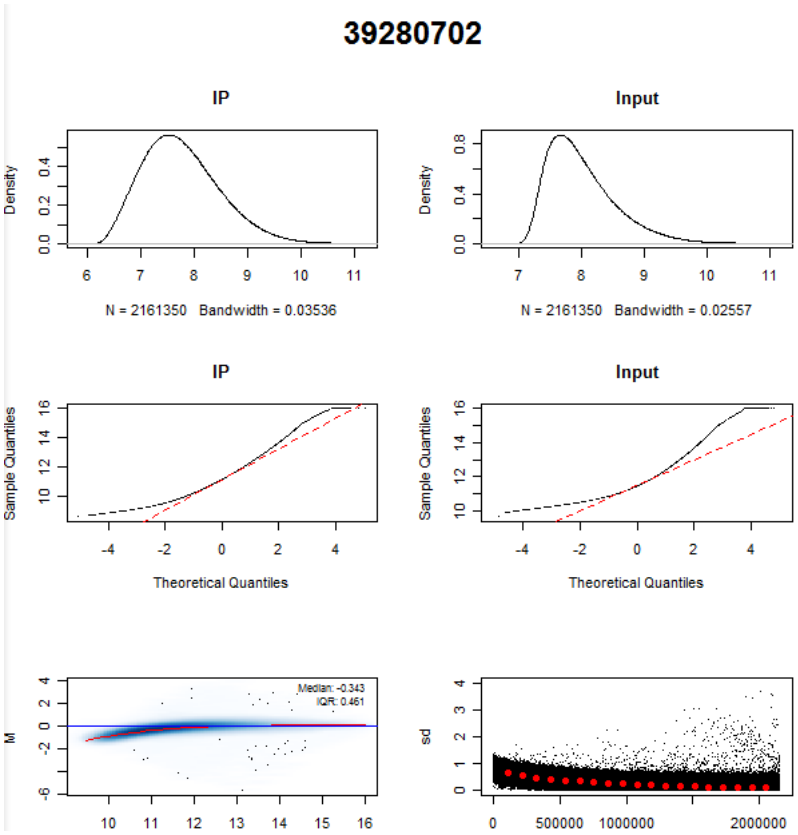

Figure 2-4: Quality check plots reveal data relatively normal data devoid of bias. To check for data quality, we plotted the distribution of the data (top row) and compared the distributions to normal distributions (second row). In the bottom left plot, we checked for biases in the log(IP/input) ratio based on the average signal of IP and input. The bottom right plot compares standard deviation of the data based on rank of signal intensities, to check for variance that is dependent on signal intensity.

the log(IP/input) ratio (bottom left) or the variance (bottom right) of the data. On the other hand, if there were no signal intensity bias, the data would produce a perfectly horizontal regression line, while a strong bias would result in very sloped lines.

2.3.2 Whole-genome analysis of H3K9me3 and H4K20me3 in olfactory epithelial tissue

Using the gene representation described in the methods section, I was able to observe the presence of histone marks at genes all across the genome. Using heatmaps, we represented histone enrichment with red and histone absence with gree, organizing

genes by chromosomal positions. Specifically, in Fig. 2-5, we present the histone mark signal for all genes in chromosomes 2, 7, and 9, in chromosomal order. The rows that correspond to OR genes are annotated in blue, while other chemoreceptor genes are annotated in orange. The left two columns show our tissue type of interest, mouse olfactory epithelial (MOE) tissue, while the two columns on the right show our control tissue, liver. These tissue types were chosen because olfactory receptor gene regulation is known to be strict in the relevant tissue, olfactory epithelium, but not in other tissues, including liver. These heatmaps are an effective way to qualitatively study the correlation between the heterochromatic marks and OR genes, as most OR genes lie in OR clusters on chromsomes 2, 7, and 9.

It is immediately obvious that the histone modification enrichment is specifically and sensitively correlated with OR genes in MOE tissue, as can be especially seen in the OR clusters in Fig. 2-5. Most genes, independently of their transcription status, appear to be devoid of both modifications in both tissues. However, in the the mouse, there is significant enrichment (red) for both H3K9me3 and H4K20me3 on OR genes (annotated in blue). Additionally, it should be noted that the presence of these marks is present in a tissue-specific manner; that is, the correlation is very strong in OE tissue (left columns) and much less strong in our control liver tissue (right columns).

Additionally, there is also some enrichment for H3K9me3 and H4K20me3 at non-OR chemoreceptor genes, although it is not strong as the enrichment at OR genes. Specifically, clusters of Vomeronasal Receptor (VR) and Formyl-Peptide Receptor (FPR) genes shown in Figure 3-4 reveal presence of heterochromatic markers similar to that of OR genes, but at a slightly lower level. Vomeronasal receptors have notably similar function to olfactory receptor genes, as VR genes encode receptors that detect pheromones. Note that both ORs and VRs are hypomethylated in the liver in agreement with published observations that report the complete absence or the low abundance of these marks on OR genes in numerous cell types.[102–104]

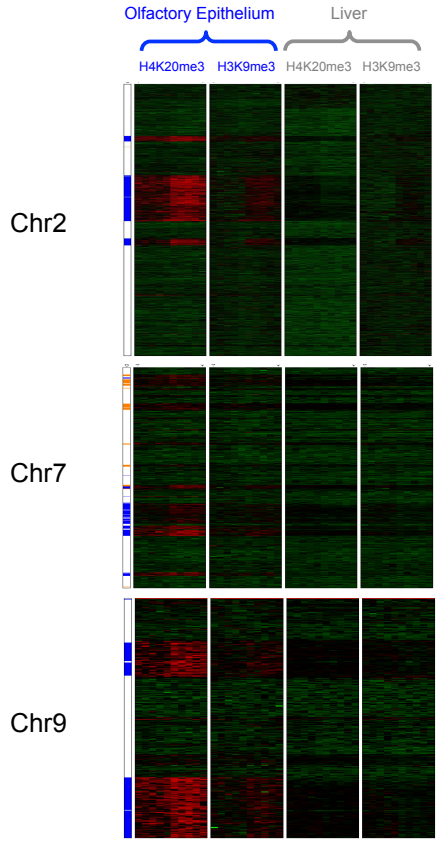

Figure 2-5: Genome-wide mapping of H3K9me3 and H4K20me3 reveal a tissue-dependent heterochromatinization of the ORs in the mouse olfactory epithelium (MOE). Histone modification signal for each mouse gene of chromsomes 2, 7, and 9, are shown for both olfactory epithelial tissue (left) and liver tissue (right). Each row visualizes -1kb to +1kb of the TSS of a gene, with average $log_2(IP/input)$ for 100 bp windows displayed; genes are vertically ordered based on their chromsomal position. OR genes are indicated on the left in blue, while other chemoreceptor genes are indicated in orange.

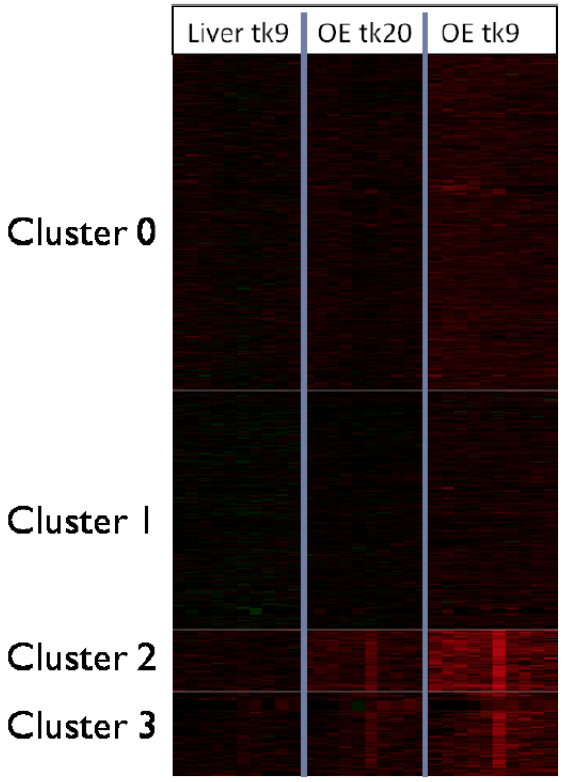

Figure 2-6: Unsupervised 4-means clustering of all chromosome 2 genes reveals a cluster of genes with low levels of H3K9me3 and H4K20me3 in general (cluster 1), a cluster with moderate levels of H3K9me3 in OE tissue (cluster 0), and clusters with enrichment for H3K9me3 and H4K29me3 in OE but hypomethylation in liver (clusters 2 and 3).

2.3.3 Heterochromatic signature for chemoreceptors

Using clustering methods as described in Section 2.2.3.1, we performed an unsupervised 4-means clustering on the genes in chromosome 2 to identify potential epigenetic signatures of OR genes. The results of the signals in the 4 clusters are shown below in Fig. 2-6. The clusters roughly correspond to tiers of strength of enrichment for the histone marks. Specifically, cluster 1 has low levels of the histone mark across all samples, while cluster 0 has some enrichment for H3K9me3 in olfactory epithelial (OE) tissue. Cluster 3 shows some enrichment in olfactory epithelium for both H3K9me3 and H4K20me3, while cluster 2 shows the strongest enrichment for both marks in OE tissue.

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
OR genes	33	15	161	163
non-OR genes	895	649	9	71
Binomial p-value (two-sided)	6.34E-44	2.92E-39	1.42E-104	7.61E-65
Relationship of OR genes and cluster	depleted	depleted	enriched	enriched

Table 2.1: OR genes significantly cluster together based on the histone mark profile. Specifically, clusters 0 and 1 are significantly depleted for olfactory receptor genes (and enriched for non-olfactory receptor genes), while clusters 2 and 3 are significantly enriched for olfactory receptor genes. P-values are calculated with the two-sided binomial test, providing statistical evidence for distinct epigenetic profiles for olfactory receptor genes.

By tracking which genes were olfactory receptor genes, I was able to identify that OR genes were strongly clustered together, as shown in Table 2.1. Almost all the OR genes are present in the 2 clusters that correlate with greater histone mark signal; furthermore, the cluster with the strongest signal is nearly solely composed of OR genes. Using a two-sided binomial test, we identified that clusters 0 and 1 (with low histone mark signal) are significantly depleted for OR genes, while clusters 2 and 3 (with stronger histone mark signal in OE) are significantly enriched for OR genes. All these findings show that the H3K9me3 and H4K20me3 modifications are strongly associated with OR genes in OE tissue. The consistent epigenetic pattern of OR genes indicates that these histone modifications are likely involved in OR gene regulation.

By studying the histone mark patterns in Fig. 2-5 and gene clustering in Fig. 2-6, we qualitatively confirmed that the "pattern" for OR genes was simply a strong presence for the heterochromatic marks. To directly quantify the relationship between the histone mark signal and the olfactory receptor genes, I ranked all genes in mouse based on the average signal intensity of H3K9me3 and H4K20me3 in OE tissue. This revealed a strong correlation between genes enriched for the histone modifications and OR genes, as shown in 2-7. (We randomly sampled 1/15 genes for the representation of all genes, due to visualization limitations.)

When ranking all genes on the left of Fig. 2-7, OR genes (shown in blue) are clustered at the top, showing that they are the genes most enriched for H3K9me3 and H4K20me3 in the MOE. In a zoomed-in view of the top 1,000 genes in Fig. 2-7

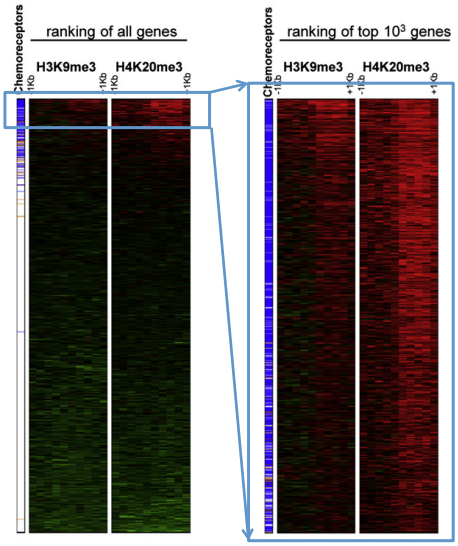

Figure 2-7: Olfactory receptor genes have strongest H3K9me3 and H4K20me3 signal. Genes most strongly enriched for H3K9me3 and H4K20me3 in OE tissue are mostly olfactory receptor genes (blue), with some chemoreceptor genes (orange), as is visualized by the ranking of all genes (left), and the zoomed-in ranking of the top 1,000 genes (right).

on the right, OR genes constitute the majority of genes with significant enrichment for both trimethyl-marks. Using the Mann-Whitney-Wilcoxon rank-sum test, we determined there was significantly more histone mark signal for OR genes compared to non-OR genes ($p < 10^{-7}$).

Many of the non-OR genes that are strongly enriched for H3K9me3 and H4K20me3 are other types of chemoreceptors (shown in orange), namely VRs and Formyl-Peptide receptors (FPRs), which matches our previous identification of chemoreceptor clusters in Fig. 2-5. These VR and FPR genes are generally clustered in extremely AT-rich isochores and likely follow the same regulatory logic as ORs, which explains their similar, but lower-level, heterochromatinization.[105–107]

2.3.4 Heterochromatic macrodomains cover OR clusters

To identify regions across the genome with a strong signal for the histone modifications H3K9me3 and H4K20me3, I utilized both the sliding window and hidden markov model approaches, but the sliding window approach better suited our data

and biological aims.

2.3.4.1 Hidden Markov Models

As described in Section 2.2.2.2, I analyzed the data with Hidden Markov Models (HMMs) to identify heterochromatic domains associated with olfactory receptor genes. However, with unsupervised HMMs, the maximum posterior probability assignments resulted in each state covering similar proportions of the genome assigned (50% for 2-state HMMS or 33% for 3-state HMMs). Since our goal was to identify strongly-marked domains, this result was problematic. The identified domains very sensitively included olfactory receptor genes and other regions of enrichment, but unfortunately, suffered in specificity. Due to the high proportion of the genome called in each state, the states did not correspond well with biological significance.

To tailor the HMM to our purposes, I then adjusted the initialization parameters to make the enriched state have a lower probability, with the hopes of increasing specificity. However, this caused the unsupervised learning step to struggle with a lack of data for the enriched state. Therefore, since both sliding window techniques, discussed below revealed more biologically meaningful domains, we used the sliding window domains for the analysis.

2.3.4.2 Sliding window method comparisons

First, we compared the results obtained from targeting narrow "peaks" and broad "blocks", using the MA2C algorithm,[96] as described in Section 2.2.2.1. We found that in the MOE, the "peaks" for the two histone modifications were strongly clustered together in broadly enriched genomic regions throughout the OR clusters in an almost continuous arrangement. Therefore, we modified the parameters to find broad "blocks" of enrichment, as described in the methods.

Using both the MA2C[96] and LOCKs[97] protocol, we were able to identify large domains of histone modifications, which we visualized with the Integrative Genome Browser.[108] Our results clearly showed that both methods found very similar domains of histone modification, as evidenced by Fig. 2-8. This comparison further supports

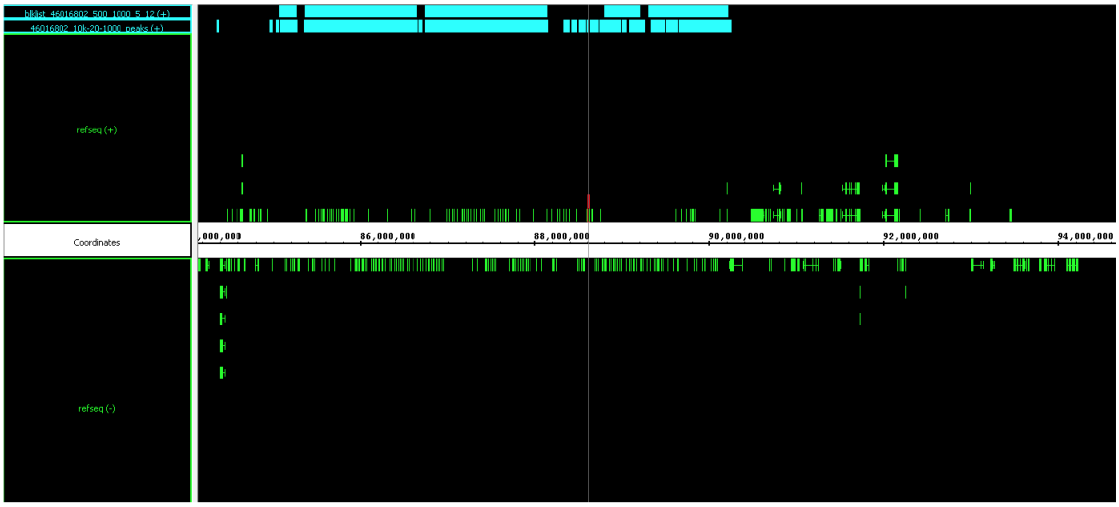

Figure 2-8: Comparison of blocks identified with the LOCKS protocol (top blue row) and the MA2C protocol (lower blue row) reveals similar macrodomains.

our findings, as it suggests that the biological signal from the data is robust and reproducible across different methodologies.

2.3.4.3 Heterochromatic macrodomains in olfactory epithelial tissue

From our identified blocks, we confirmed that H3K9me3 and H4K20me3 form heterochromatic macrodomains. These regions cover megabases of clustered OR genes in the MOE, as shown by the example cluster on chromosme 2 in Fig. 2-9. Specfically, we found that 95% of ORs fall in H4K20me3 blocks (1376 out of 1441 OR genes) and 77% of ORs fall in H3K9me3 blocks (1109 out of 1441 OR genes), corresponding to a strong statistical enrichment ($p < 10^{-7}$).

2.3.4.4 Little to no epigenetic domains in liver tissue

We also analyzed our control tissue, liver, in the same way. As expected, the low signal for H3K9me3 and H4K20me3 in liver resulted in very few peaks or blocks identified, and the identified regions rarely overlapped with OR genes, as shown in Fig. 2-9. In fact, the few peaks or blocks that were found were also not close together, and ChIP-qPCR confirmed that the actual histone mark presence at these locations was, in fact, very low, as shown in Fig. 2-10. It is unsurprising that there were a

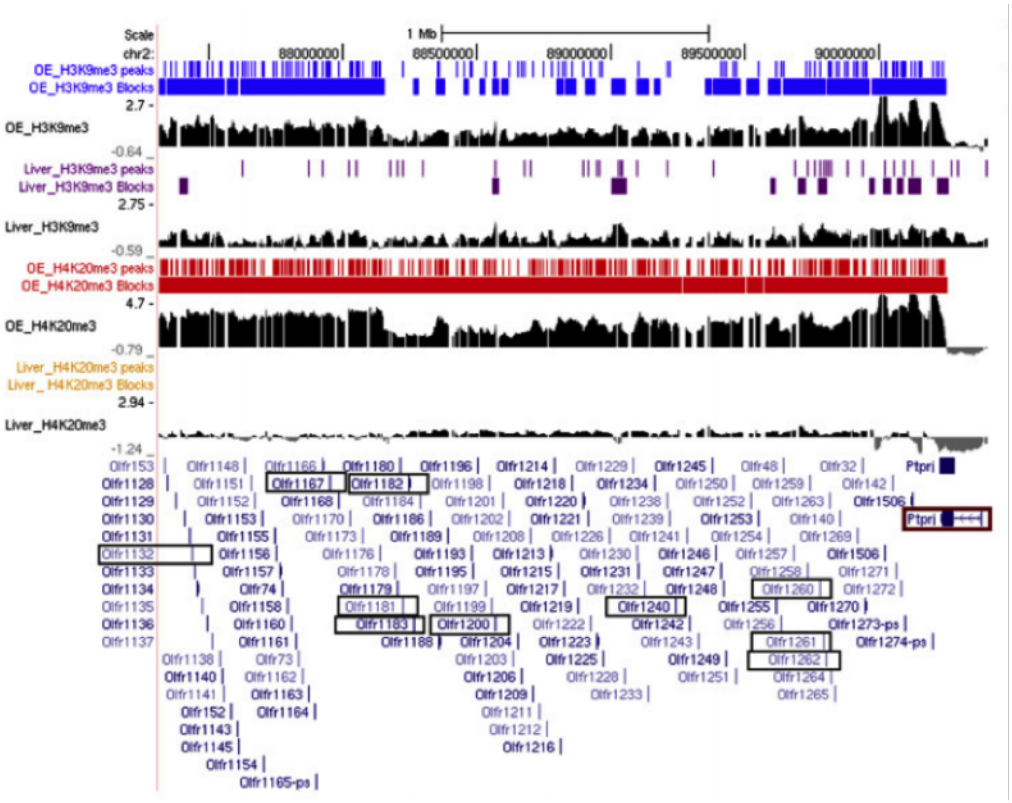

Figure 2-9: Part of a cluster of OR genes overlaps peaks and blocks of H3K9me3 and H4K20me3 in OE. Here, we show part of the biggest OR cluster located on chromosome 2, which contains 240 genes and spans a 5 MB region. The thin blue (H3K9me3) or red (H4K20me3) bars represent significant narrow peaks (FDR $\leq 5\%$) identified in the MOE, while the thick blue or red bars represent the broad blocks identified. In the liver, there are only a few sporadic H3K9me3 peaks and blocks (purple).

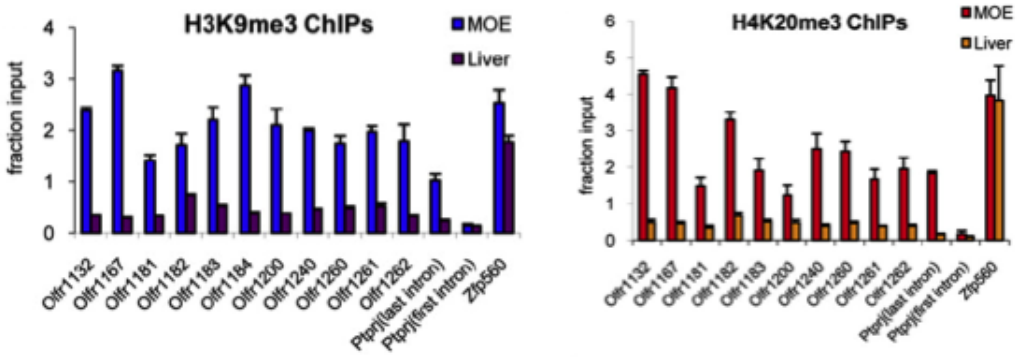

Figure 2-10: H3K9me3 and H4K20me3 is highly present at OR genes in the OR cluster in MOE (but not liver), but not beyond the OR cluster, based on ChIP-qPCR. The Ptprj gene (marked by red rectangle in Fig. 2-9) stands at the border of the OR cluster, which coincides with the border of the heterochromatic block. The most proximal intron (to the OR cluster) is enriched for both marks, while its most distal intron, located 43 kb downstream, is free of these modifications. Zfp560 serves as positive control.

few spurious peaks or blocks found, since these sliding window algorithms somewhat base their enrichment threshold relative to the signal in the entire dataset; therefore, if there was a low signal all across the genome in liver, then peaks and blocks would be called for regions that showed stronger enrichment than the rest of the genome in liver, but that still corresponded to low enrichment when compared to the strong enrichment at OR genes in OE tissue.

We further validated our ChIP-chip results by quantitative PCR (qPCR) for multiple OR gene clusters in both tissues. Whereas ChIP-chip can give a noisy signal across the entire genome, ChIP-qPCR can give a more precise signal for a very specific location. qPCR results for representative genes that were boxed in Fig. 2-9 are shown in Fig. 2-10.

2.3.4.5 Heterochromatic boundaries align with OR clusters and active non-OR genes

We also noted that the borders of the heterochromatic marks strongly coincided with the borders of OR loci, as shown in Figures 2-9 and 2-11. The reported binding

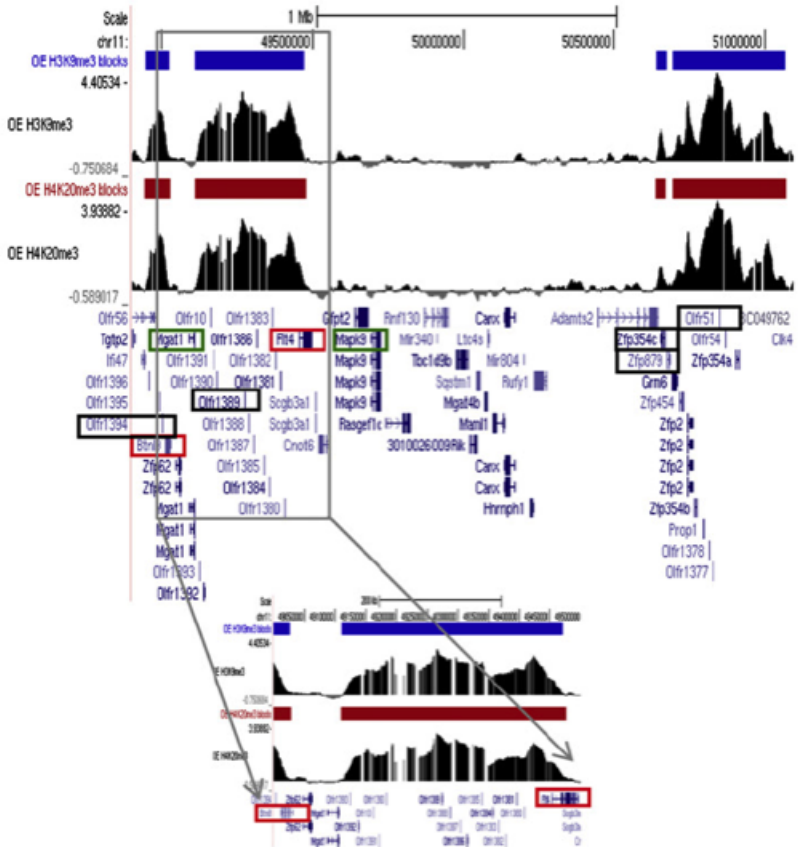

Figure 2-11: Part of a silenced OR cluster on chromosome 11 is interrupted by a small group of transcriptionally active non-OR genes. Genes that are marked by a green rectangle are transcriptionally active in the MOE, and genes marked by red rectangles do not have detectable transcripts. A zoomed-in picture of the cluster (bottom) shows that genes Btnl9 and Flt4, which are transcriptionally inactive, are partially methylated. Two sets of primers were used in ChIP-qPCR for each of these genes: one at the beginning (most proximal to the neighboring OR gene) and one at the end of the gene (most distal from the neighboring OR).

of CTCF outside of OR clusters[109] or other insulating elements,[110] may play a role in the borders of OR heterochromatin aligning with OR clusters. Additionally, the data shows that the presence of transcriptionally active non-OR genes in an OR cluster interrupts the heterochromatin blocks, until the next OR gene reconstitutes the heterochromatin (Figure 2-11). On the other hand, transcriptionally inactive non-OR genes in OR clusters are partially covered by the histone modifications, which implies that in the absence of a competing need for transcription or insulating activity, the heterochromatin can extend over non-OR genes within an OR cluster.

2.3.5 Further experimental validation

To further study and validate the findings of the ChIP-chip and ChIP-qPCR data, the Lomvardas lab performed more experiments to investigate the relationship between the heterochromatic histone modifications and the olfactory receptor genes in olfactory epithelial tissue.

2.3.5.1 Accessibility of heterochromatic OR genes

To determine if the histone modifications present at OR genes in mouse olfactory epithelial (MOE) tissue resulted in functional differences of the chromatin, we analyzed the accessibility of the DNA at different loci. This was accomplished through the DNase protocol introduced in Section 1.3.6.

As demonstrated by Fig. 2-12, we found that silent OR genes in MOE tissue were much less digested, and therefore, less accessible, than transcriptionally active genes; on the other hand, silent non-OR genes had intermediate accessibility. For comparison, OR loci in liver were similar to non-OR genes in terms of DNase I accessibility. These findings were also supported by other experiments, such as southern blot analysis with a degenerate OR probe (not shown here).

2.3.5.2 Isolated olfactory sensory neurons (OSNs) are enriched for H3K9me3 and H4K20me3

Since the MOE tissue is composed of multiple cell types,[111] we performed experiments to confirm that our results in OE tissue specifcially reflected the state of the olfactory sensory neurons (OSNs). To accomplish this, we performed fluorescence-activated cell sorting (FACS) experiments followed by ChIP-qPCR. This allowed us to we isolate mature OSNs from OMP-IRES-GFP mice and, as seen in Figure 2-13a, the OR genes tested have high levels of enrichment for both H3K9me3 and H4K20me3 in OSNs. Each OR gene was expressed in 0.1% of the OSNs, which supports the idea that the majority of OR genes would need to be silenced.

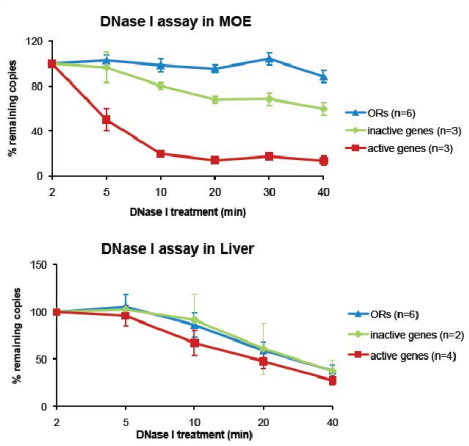

Figure 2-12: The ORs acquire a highly compacted chromatin structure in the MOE. DNase I accessibility assay with nuclei from both MOE and liver is presented here. Nuclei were treated with DNase I, DNA was isolated at various time points (2 to 40 min) and equal amounts were used for qPCR. The amount of DNA measured at each interval was expressed as a fraction of the DNA present at 2 min of enzyme treatment and was plotted over time. We assayed several ORs as well as other genes that are active or inactive in the MOE or liver, and their mean is shown here, with representative data from one experiment. In MOE, the ORs appear to be more resistant, suggesting they are less accessible.

2.3.5.3 OR silencing occurs independent of OR expression

To determine whether the heterochromatic silencing was independent of or a result of OR expression, we sorted sustentacular cells from the MOE;[112] sustentacular cells are present in OE tissue and have common developmental ancestors with the OSNs, but they do not express ORs. As shown in Fig. 2-13b, we found similar levels of H3K9me3 and H4K20me3 in the sustentacular cells as in the OSNs, suggesting that marking of OR genes with H3K9me3 and H4K20me3 occurs even in the absence of OR expression. This raises the possibility that trimethylation of lysines 9 and 20 takes place before OR activation.

2.3.5.4 OR silencing occurs developmentally prior to OR expression

To further investigate the possibility of heterochromatic silencing before OR expression, we performed ChIP-qPCR analysis in progenitor cells, starting with the most multipotent cells of the MOE, the HBCs.[113] Our results, as shown in Figure 2-13c,

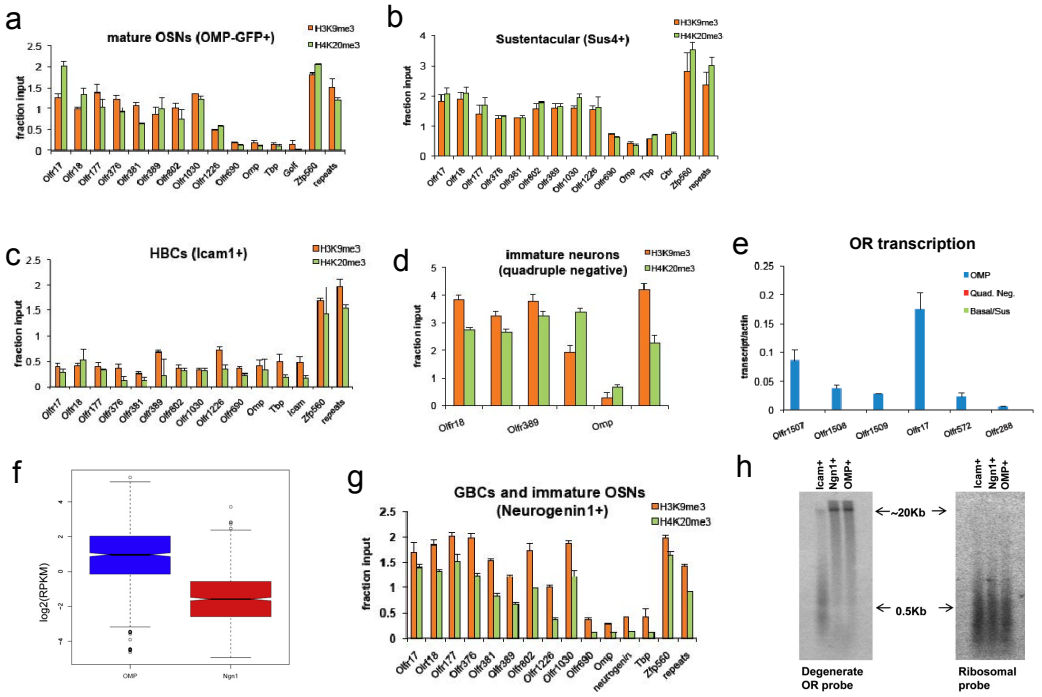

Figure 2-13: OR silencing occurs independent of and developmentally prior to OR expression. Values are the mean of triplicate qPCR, and error bars represent the SEM. a) Mature OSNs are enriched for H3K9me3 and H4K20me3 at OR genes, based on ChIP-qPCR. Golf, Tbp and Omp are negative controls, while Zfp560 and major satellite repeats are positive controls. b) Isolated sustentacular cells, which do not express ORs, show enrichment for H3K9me3 and H4K20me3, suggesting that OR silencing occurrs independentally from OR expression, based on ChIP-qPCR. Cbr is used as an additional negative control. c) HBCs, which are multipotent progenitor cells of the MOE, are not enriched for H3K9me3 and H4K20me3 on OR genes, based on ChIP-qPCR experiments. d) Immature neurons and progenitors from the MOE are enriched for H3K9me3 and H4K20me3, as shown with ChIP-qPCR of cells isolated by collecting OMP+, ICAM+, iLR+, and Sus4+ cells (quadruple negative). e) Progenitor cells (quadruple negative and basal/sustentacular cells) do not transcribe OR genes, as shown with isolated RNA, while OSNs (OMP+ cells) do. f) Progenitor Ngn1+ cells show 8-fold less OR gene expression than OMP+ cells, as shown with a boxplot of $log_2(RPKM)$. g) Progenitor Ngn1+ cells show similar levels of H3K9me3 and H4K20me3 as mature OSNs, confirming the deposition of these marks developmentally prior to OR gene expression, based on ChIP-qPCR analysis. h) Immature and progenitor neurons ICAM+, Ngn1+, and OMP+ cells show reduced accessibility at OR genes based on DNase I digestion with a degenerate OR probe compared to a ribosomal probe.

indicate that there is no enrichment for H3K9me3 and H4K20me3 on OR genes, although there is a strong signal for H3K9me2 (not shown), suggesting that in this multipotent cell, ORs are repressed via mechanisms that differ from repression in OSNs.

Additionally, we checked the chromatin state of OR genes in other progenitor cells from the MOE that are negative for OMP, ICAM-1, iLR, and SUS4, as shown in Figure 2-13d. The result was that the enrichment for H3K9me3 and H4K20me3 appeared to be as high as in the OMP+ cells in Figure 2-13a, even though, according to RT-PCR, this population does not express ORs (Figure 2-13e). Again, this suggests that the trimethylation of OR genes occur developmentally before OR expression.

To study a cell population that is more well-defined, we studied a Neurogenin1-GFP (Ngn1-GFP) BAC transgenic reporter mouse from GENSAT.[114] RT-PCR analysis showed that these cells represent a mixed population of progenitors and immature neurons. We found that Ngn1+ cells had 8-fold lower mRNA levels than the mature OSNs for 1185 OR genes (Figure 2-13f), and, importantly, in the Ngn1+ cells, 95% of OR genes have transcript levels similar to the transcript levels of silent genes. Therefore, the low levels of OR mRNA in these cells likely reflects a small percentage of contaminating mature OSNs.

When we performed FACs and ChIP-qPCR on the Ngn1+ cell population, we found high levels of enrichment for H3K9me3 and H4K20me3 on OR genes, demonstrating similar heterochromatic signature with the mature OSNs (Figure 2-13g). Therefore, the ChIP-qPCR data from the quadruple negative cells and Ngn1+ cells are consistent with H3K9me3 and H4K20me3 having been deposited on OR genes before OR expression.

We wanted to test the significance of the epigenetic transition from di-methylation to trimethylation at the OR genes during MOE differentiation, so we performed southern blot analysis on ICAM1+, Ngn1+ and OMP+ cells. Figure 2-13h demonstrates that the differentiation of HBCs to Ngn1+ cells coincides with increased protection of OR genes from DNase I digestion, suggesting that this epigenetic transition results to a less accessible OR chromatin structure retained in mature OSNs.

2.3.5.5 Epigenetic switch accompanies choice of active OR allele

To investigate the state of the single active OR allele in OSNs, we used FACS to select neurons expressing the olfactory receptor P2 from P2-IRES-GFP knocked-in mice. We isolated 40,000 GFP+ neurons and GFP- neurons, which, respectively, do and do not express the P2 allele, from P2-IRES-GFP heterozygote mice. We found that the enrichment for H3K9me3 and H4K20me3 is significantly reduced on the active OR allele, in comparison to the strong presence of the marks on P2 where it is not the active allele (Figure 2-14b-d) based on ChIP-qPCR. Specifically, the active allele state is captured by the GFP primer in the GFP+ cells (Figure 2-14b), which shows reduced H3K9me3 presence compared to the inactive alleles, as captured by the p2WT primer in GFP+ cells (Figure 2-14b) and the GFP and p2WT primers in GFP- cells (Figure 2-14c). Though the presence of these marks was reduced on the active allele, they were not completely removed; control experiments indicate that this is due to 30% contamination of the population, which is unsurprising since we were selecting for an extremely rare population (0.05% of total cells in the MOE).

To obtain an even purer population, we double-sorted the GFP+ cells, resulting in a $> 95\%$ GFP+ population, using MOR28-IRES-GFP heterozygote knock-in mice; this was only possible because MOR28-IRES-GFP mice provide more GFP+ cells. As seen in Figure 2-14e-f, ChIP-qPCRs from this extremely pure population provides even stronger evidence that H3K9me3 is absent from the transcriptionally active allele, MOR28, as shown with the GFP primer in GFP+ cells (Figure 2-14e). In contrast, the inactive MOR28 (MOR28WT) shows high levels of H3K9me3 in both GFP+ and GFP- cells (Figure 2-14e-f).

To further probe the epigenetic state of the single active allele, we measured H3K4me3 presence on the active P2 allele; H3K4me3 is a histone mark commonly associated with active promoters[115] that has a mutually exclusive distribution with H3K9me3 and H4K20me3.[116] As expected, H3K4me3 cannot be detected on OR promoters using chromatin preparations from the whole MOE (data not shown), but in Figure 2-14g there is enrichment for H3K4me3 on the active P2 promoter and

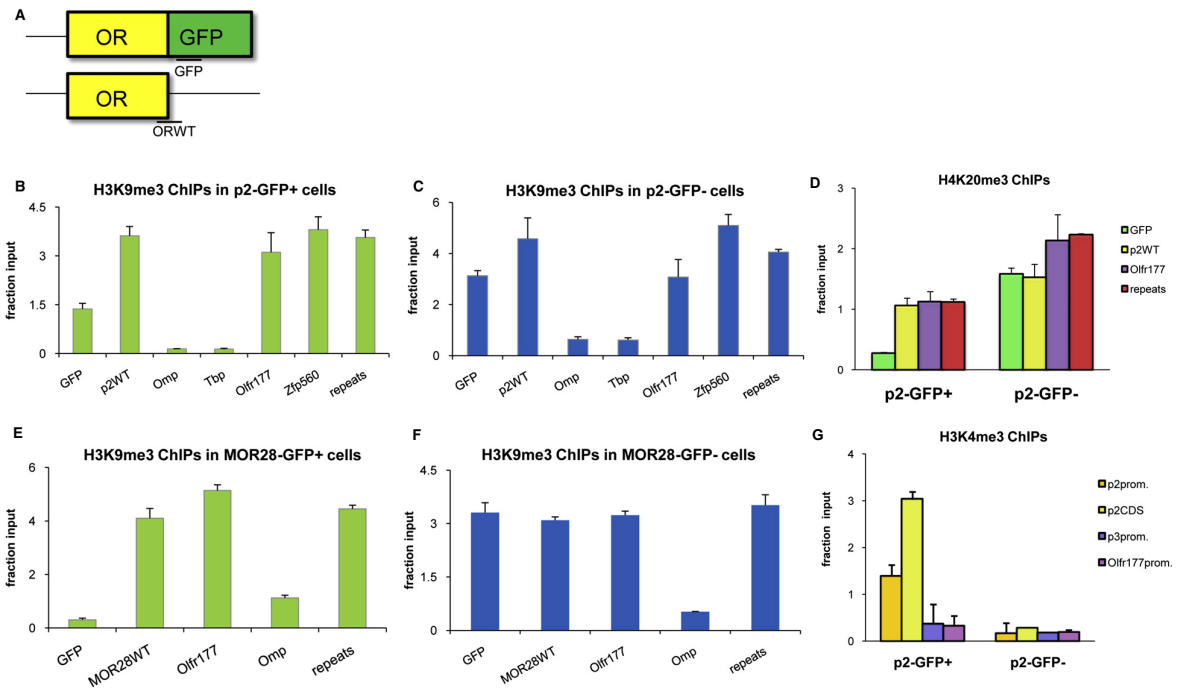

Figure 2-14: The active OR allele is not enriched for H3K9me3 or H4K20me3, but it is marked with H3K4me3. Heterozygote P2-IRES-GFP and MOR28-IRES-GFP mice were used to isolate GFP+ and GFP- cells by FACS, followed by ChIP-qPCR. Values are the mean of triplicate qPCR, and error bars represent the SEM. (A) GFP Primers specifically monitor the active P2 allele, whereas the p2WT primers specifically amplify the inactive P2 allele. (B and C) (B) H3K9me3 is reduced on the active P2 allelle in the GFP+ cells, as shown with the GFP primer, c) but not in the GFP- cells in which this P2 gene is inactive. The inactive allele, amplified specifically by the p2WT primers, shows high enrichment for H3K9me3 in both GFP+ and GFP- populations. (Omp and Tbp are negative controls, and Zfp560 and repeats are positive controls.) (D) H4K20me3 is also reduced on the active p2 allele (found with the GFP primer) in GFP+ cells but present on other inactive alleles. (E and F) A more purified population of cells shows an even greater reduction of H3K9me3 on the active allele in GFP+ cells, with high H3K9me3 presence on the other inactive alleles. These GFP+ cells from MOR28-IRES-GFP heterozygous mice were subject to a second round of FACS to yield a $> 95\%$ pure population. (G) There is significant enrichment for H3K4me3 throughout the active P2 gene in the GFP+ cells, but not on the neighboring P3 gene or a distant OR (Olfr177). As expected, there was no H3K4me3 on the P2 gene or any other OR gene in the GFP- cells.

CDS in the GFP+ population, but not in the GFP- population, where P2 is not expressed. This supports the idea that selection of the P2 allele is associated with the removal of H3K9me3 and H4K20me3 and deposition of H3K4me3. Furthermore, although H3K4me3 is strongly present the active P2 allele, it is missing from the neighboring P3 and P4 genes (Figure 2-14g) in both GFP+ and GFP- cells, despite their proximity and sequence similarity.

2.3.5.6 Heterochromatic marks induce silencing and OR-like expression in LacZ transgene

Our data suggested that heterochromatinization of OR loci universally represses OR genes. To test this hypothesis, we examined an unusual transgenic mouse, where a OMP-LacZ transgene had been inserted proximal to a singular OR gene. Normally, we would expect OMP-LacZ or OMP-GFP independent transgenes to be expressed in the majority of olfactory neurons, and numerous such transgenes are.[82,117] On the other hand, this transgene is silent in 99.9% of olfactory neurons and has a sporadic and mostly zonal expression reminiscent of that of the neighboring OR.[118]

By mapping the exact insertion site of this transgene, we found that it resides approximately 55kbs from Olfr459, as shown in Figure 2-15a. ChIP-qPCR experiments showed that the insertion site is heterochromatinized in both the wild type and transgenic mice, as shown in Figure2-15b; ChIP-qPCR also indicates that the reporter is itself marked by H3K9me3/H4K20me3 in an tissue-specific fashion, in contrast to the endogenous OMP promoter, which is unmethylated (Figure 2-15c).

To examine whether the insertion of the OMP transgene resulted in monoallelic expression, we compared the number of cells with LacZ expression between homozygous (+/+) and heterozygous (+/-) OMP-LacZ mice. This was done through the use of X-gal, which turns cells with LacZ product (*beta*-galactosidase) blue. As seen in Figure 2-15d, OMP-LacZ homozygotes have approximately 1.8 fold more cells with LacZ expression than heterozygotes, consistent with a monoallelic expression pattern.

Finally, to test whether the transgene is under the transcriptional control of the proximal OR locus, we crossed this transgenic mouse to the Emx2 knockout mice,

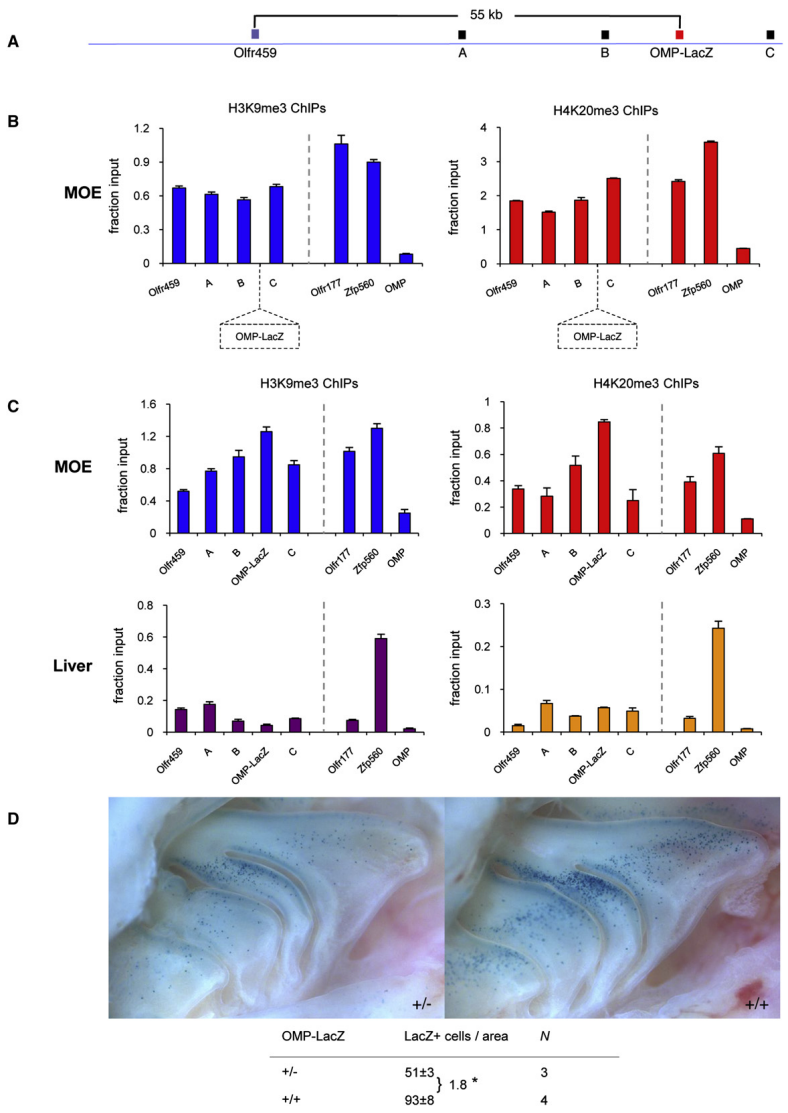

Figure 2-15: Tissue-Specific OR Modifications Are Associated with OR-like Transgene Expression. (A) Graphic representation of the Olfr459 locus and the OMP-LacZ insertion site located 55 kb away. Positions marked A, B, and C depict assayed regions in the qPCR analysis below. (B) ChIP-qPCRs with chromatin from the MOE of wild-type mouse show that the Olfr459 is enriched for H3K9me3 and H4K20me3, and both modifications appear to extend to the insertion site. (C) ChIP-qPCR analysis of the MOE and liver from OMP-LacZ-positive animals. Both H3K9me3 and H4K20me3 show MOE-specific deposition on Olfr459, the OMP-LacZ transgene, and the regions proximal to these loci. Experiment was performed in two biological replicates with similar results. Values shown here are the mean of triplicate qPCRs. Error bars represent the SEM. (D) Homozygote (+/+) OMP-LacZ mice have 1.8 times as many stained (blue) cells as heterozygote (+/-) OMP-LacZ mice, as shown by the X-gal stains of lateral whole moounts of nasal cavities. N, number of biological replicates. $^*p < 10^-4$, Student's t test.

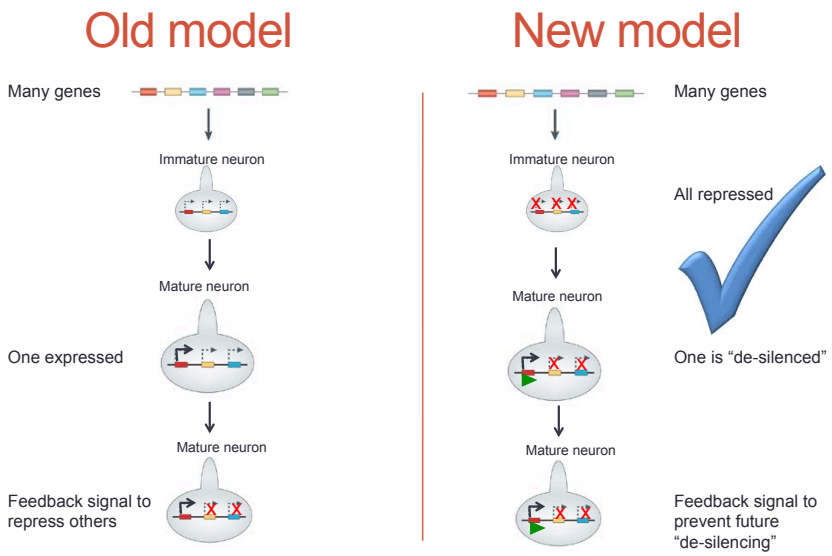

Figure 2-16: Previous model and proposed new model for olfactory gene regulation. Instead of repression of olfactory genes after expression of a single olfactory gene, our results support a new model where olfactory genes are epigenetically repressed at an early developmental stage. Later, a single allele of one olfactory gene is de-repressed, allowing expression of one gene to occur in each neuron, which triggers a feedback signal to prevent de-repression of any other genes.

as Emx2 is required for the expression of Olfr459,[119] so the offspring will not have expression of Olfr469. We found that expression of the OMP-LacZ transgene is also abolished in the offspring, suggesting that this transgene conforms to the regulatory logic of the neighboring OR (not shown).

2.4 Contributions

Taken together, our results strongly suggest displacing the old model of olfactory gene regulation with the the new model, as visualized in Figure 2-16. Specifically, we found that the presence of histone modifications H3K9me3 and H4K20me3 result in chromatin-mediated silencing of Olfactory Receptor (OR) genes independent of and developmentally prior to OR expression. Further, the transcriptional activity of a single OR allele in an olfactory neuron is likely then made possible through the de-repression of that allele, with the repressive marks replaced with the active histone modification H3K9me2. Then, this transcriptional activity most likely triggers the

previously-researched feedback signal that prevents the de-repression of any other OR alleles in that neuron. As such, our research provides insight into an epigenomic mechanism of regulation for an entire gene family of olfactory receptor genes.

Chapter 3

A computational method for chromatin state comparisons across groups of epigenomes

As we saw in Chapter 2, epigenomic data can provide critical information about gene regulation, cell differentiation, and cell type differences. However, with the advent of Next Generation Sequencing technologies, it is no longer feasible to manually analyze these biological datasets at the genome-wide level, as there are 3 billion basepairs in a human genome. Computational methods are an efficient and widely useful way to address this scientific need, and computational methods have been developed and widely adopted for a number of biological applications, such as sequence alignment and assembly,[120] gene identification,[121] motif discovering,[122] and protein structure prediction.[123]

However, there is still a lack of computational methods to analyze chromatin state information at the genome-wide level. To address this, we developed ChromDiff, a novel computational algorithm for group-wise chromatin state comparisons across the genome. With the development of ChromDiff and other methods, computational techniques have untold potential to unlock key insights into biological studies.

3.1 Introduction

Epigenomic datasets provide critical information about the dynamic role of chromatin states in gene regulation, but a key question of how chromatin state segmentations vary under different conditions across the genome has remained unaddressed. Here, we present ChromDiff, a group-wise chromatin state comparison method that generates an information theoretic representation of epigenomes and corrects for external covariate factors to better isolate relevant chromatin state changes.[124] Our methodology should be broadly applicable for epigenomic comparisons and provides a powerful new tool for studying chromatin state differences at the genome scale.

3.1.1 Problem Statement

Epigenomic datasets provide a rich resource for understanding genome activity across both genes and regulatory regions in response to developmental, environmental, or genetic signals. Epigenomic marks, including histone modifications and DNA methylation, have been shown to be highly dynamic across cell types.[12,14,51] Furthermore, epigenetic differences have been strongly associated with changes in mammalian development,[13,15] as well as gene activation and repression patterns across cell types.[52–54] Epigenomic signatures have also resulted in the identification of new regulatory elements and functional annotations.[19,55,56]

In addition to cell type differences, comparative epigenomics analyses have been applied across individuals, disease status, and species. Studies of natural epigenomic variation across individuals have shown wide-spread differences across individuals of different genotypes, and between the two alleles of the same individual.[16,57] Epigenomic comparisons across disease and control samples have been linked to differences in disease manifestation in monozygotic twins,[58] while ongoing efforts such as ICGC[59] aim to better understand the role of epigenomic alterations in cancer. Comparative epigenomics analysis across species has also proved informative, identifying conserved epigenetic marks, even in regions that fall in unconserved genetic sequences,[19,55,56]

and tools such as the Comparative Epigenome Browser (CEpBrowser) allow for direct exploration of multi-species epigenome comparisons.[60]

As our understanding of epigenomics has progressed, previous methods have leveraged histone combinations to partition the epigenome into various chromatin states, such as ChromHMM,[23] Segway,[24] and HMMSeg.[25] The resulting analyses enabled by chromatin state analysis has provided fruitful findings about epigenomic variation and lineage-specification.[17,26–30] However, no methods have yet been developed to enable group-wise chromatin state comparisons based on these combinatorial segmentations.

3.1.2 Background and previous work

Comparative epigenomic analyses initially focused on peak-calling, enrichments, domains, or comparisons for a single histone modification with various normalization and modeling approaches.[125–129] As the availability of data increased rapidly in recent years, methods tackling combinatorial approaches to histone modification data to identify patterns across many histone marks for one biological condition or sample have been developed,[61–64] including the aforementioned segmentation methods.[23–25]

However, scalable combinatorial methods to directly discover patterns between chromatin state changes and biological conditions are still limited. MultiGPS addresses the analogous question of comparing transcription factor binding Chromatin Immunoprecipitation Sequencing (ChIP-Seq) experiments across groups,[130] and therefore tailors the approach to punctate signals that are not relevant for histone mark data. To our knowledge, only one method, dPCA,[131] compares epigenomic signal across multiple histone marks under multiple conditions; it does so by performing Principal Component Analysis (PCA) analysis on the differences of the replicate averages. While dPCA has been shown to be useful, it is constrained by the limitations of PCA analysis, such as sensitivity to scaling the data. Furthermore, dPCA does not provide any options to correct for external covariate factors. Covariate correction is a crucial part of comparative analysis when using datasets with variation due to batch effects, donor variability, sample differences, and experimental differences. Additionally, the importance of covariate correction will only increase in coming years, with

the release of more public and resource datasets that will increase statistical power but will also be generated in less controlled circumstances. Lastly, dPCA compares the histone mark signal based on differences in means, but does not take advantage of existing advanced techniques that interpret combinatorial histone mark signals into segmentations based on Hidden Markov Models (HMMs) or bayesian networks.

3.1.3 Approach

In this paper, we propose a highly-scalable method, ChromDiff, for directly discovering potential relationships between chromatin states, genes, and biological conditions; in doing so, ChromDiff generates a novel information-theoretic representation for epigenetic information and employs covariate correction to enable large-scale analysis of samples while controlling for a wide variety of circumstances, including batch effects and donor variability. As a result, ChromDiff is a general statistical pipeline for comparing combinatorial chromatin states of groups of epigenomes, which we then apply to leverage the breadth of data from the Roadmap Epigenomics and ENCODE projects[11,19] and the diversity of chromatin state annotations provided by ChromHMM.[11,23]

Specifically, by utilizing the chromatin state annotation for every epigenome, we use these discrete states to quickly compare any subset of epigenomes to one another through a probabilistic representation of the chromatin states that builds upon information theory. By utilizing chromatin states that were jointly learned over all epigenomes, we are able to use a general model, but apply it to many specific biological questions. To account for various differences in sample and data generation, we also utilize the metadata of the epigenomes to correct for covariate factors, thereby better isolating differences due to a single biological attribute. This covariate correction allows ChromDiff to leverage the same set of epigenomic data for various, specific biological conditions, while controlling for other variables. Furthermore, ChromDiff is compatible with any general chromatin state segmentation, regardless of the method behind it, which makes it amenable to various existing methods, including ChromHMM,[23] Segway,[24] and HMMSeg,[25] as well as future methods that have yet to

be developed.

Similarly to other methods, ChromDiff produces sets of regions with epigenomic differences across conditions. However, our method additionally utilizes group-wise comparisons to gain statistical power, while building upon a general chromatin state model segmentation. As a pipeline, ChromDiff also provides additional features of gene set enrichment calculations and gene expression comparisons. Furthermore, ChromDiff clusters the distinguishing genomic regions into groups that exhibit similar epigenomic signatures, thereby highlighting clusters with distinct gene set enrichment and gene expression patterns. These results suggest that the identified clusters may share regulatory mechanisms and functional pathways. In this way, ChromDiff provides novel, thorough insights on the complex relationship between these general chromatin states, biological attributes, and specific clusters of genes. This method, therefore, not only enables the identification of genomic regions relevant to an epigenomic comparison based on group-wise differences, but also provides a global understanding of how chromatin states are involved in a wide variety of biological situations.

Generally, we believe our method will be broadly applicable to new epigenomic datasets, and that epigenomic comparisons across multiple marks and multiple samples will be widely used to uncover the molecular processes underlying cellular differentiation, gene regulation, and human disease.

3.2 Methods

3.2.1 Overview of comparison of epigenomic features

To capture epigenomic differences between groups of epigenomes, we focus on the set of chromatin states associated with each protein-coding gene (Figure 3-1), while generating an information theoretic-encoding of these chromatin states and correcting for external factors to isolate differences due to the comparison. We leverage the multiple samples available in each pairwise group comparison to evaluate the statistical sig-

nificance of such recurrent changes, and the multiple genes to evaluate the statistical significance of biological pathways. However, our methods are generally and broadly applicable to various regulatory genomic regions, beyond the gene-centric approach taken here.

Specifically, we define epigenomic features by calculating the probability of chromatin state assignment for each gene across each epigenome, integrated over the body of that gene (Figure 3-1) in our original ChromDiff method. For example, we apply our method to gene NRXN1 in neurosphere cultured cells (Figure 3-1a), based on the 15-state ChromHMM annotation of the Roadmap Epigenomics project.[11] First, we identify the probability that NRXN1 should be assigned to each of 15 chromatin states, integrated over the entire length on the gene body, resulting in 15 different features for NRXN1. This encoding is based on information theory, as we retain the probability distribution of each chromatin state within each gene and sample type. Therefore, this representation drastically reduces the dimension of the chromatin state data while preserving the information necessary to calculate important information theory metrics, including the entropy of each gene and sample type, as well as the divergence between one gene and another gene or background. As information theory has been shown to have applications to fields as diverse as signal processing, neurobiology, machine learning, and cryptography, it provides a theoretical foundation for our method.

ChromDiff recalculates this encoding for every gene and every epigenome, resulting in a matrix of 299,025 features (columns) and 127 epigenomes (rows) (Figure 3-1b). We then utilize logistic regression to correct for feature covariates including production center, sex of donor, sample state (solid or liquid) and sample type (cell line, primary cell, tissue, etc.) by setting the value of each covariate factor that we are not testing to be the response residuals from the logistic regression model (Figure 3-1c). This step is crucial, due to the wide variety of differences among the epigenomes; by controlling for variables that we are not currently investigating, ChromDiff is better able to identify genes with chromatin state changes that specifically correspond to the current comparison. As a result, each feature value indicates whether that gene is

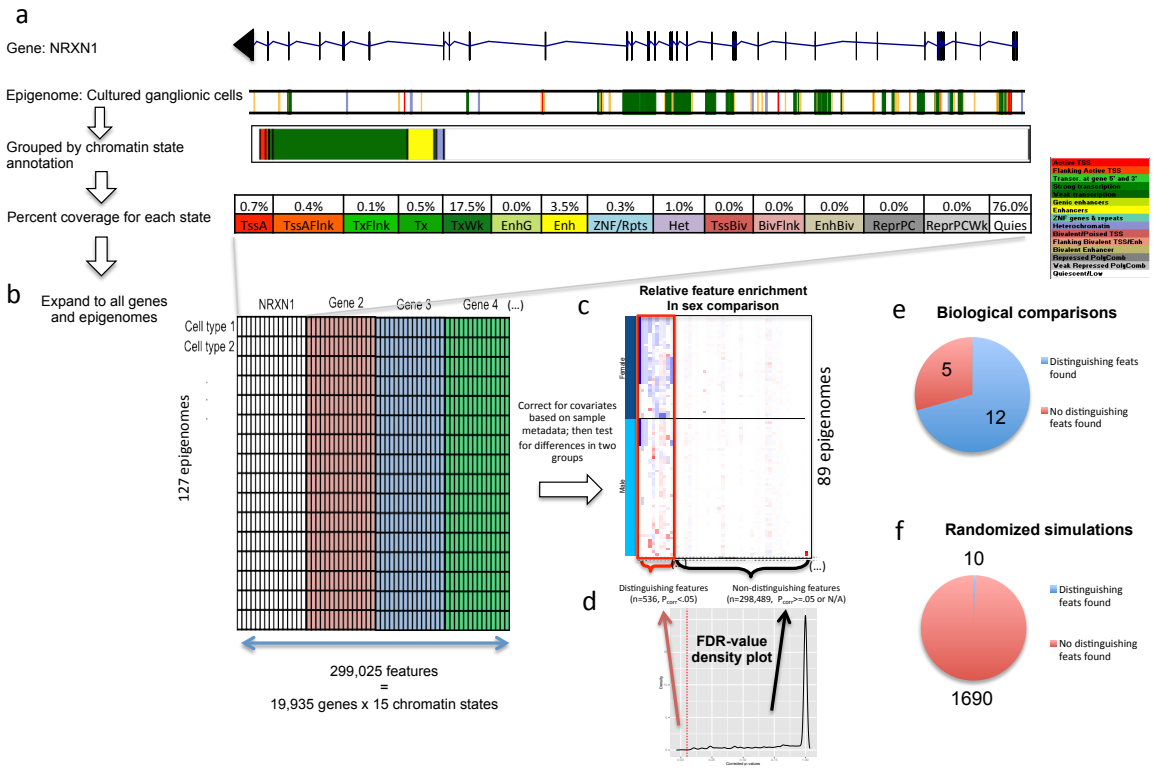

Figure 3-1: A novel method for comparative analysis of epigenomic groups. a. Starting with a single gene (NRXN1) and epigenome (cultured ganglionic cells), we represent the epigenome as the percent coverage of each chromatin state at that gene. b. Then, we repeat the process for all 127 epigenomes and 19,935 protein-coding genes, resulting in a matrix of 127 epigenomes by 299,025 features. c. After normalizing and correcting each column in the matrix for covariate factors, we compare the female and male epigenomes of the original 127 epigenomes to identify features that exhibit different behavior in female and male epigenomes. d. The density plot of corrected p-values from all features shows 536 out 299,025 features that distinguish between the two groups. e. Of the real biological comparisons that we tried, we found distinguishing epigenomic differences over 70% of the time. f. Distinguishing features were found for these randomized groupings only 10 out of 1700 times, or less than 1% of the time.

annotated as that chromatin state more or less often than expected, after correcting for covariates.

Finally, our pipeline uses these corrected feature values to recognize significant differences between two groups of samples (in this case, male and female samples), using a non-parametric Mann-Whitney-Wilcoxon test, Student's t-test, or F-test, and correcting for multiple hypothesis testing using Bonferroni, Benjamini-Hochberg, or Benjamini-Yekuteli multiple hypothesis correction. Based on these statistical results, ChromDiff reports all features (chromatin state and gene combinations) that are significantly different between the two groups at a corrected p-value cutoff of $p<0.05$ (Figure 3-1d).

Though ChromDiff can be applied to any genomic region, we initially focused on gene bodies, as they make our methodology and results easier to validate and interpret. This approach allowed us to incorporate into the ChromDiff pipeline multiple tools for downstream analysis of the resulting genes that are found to show epigenomic differences in our comparisons. Firstly, to recognize the biological processes associated with epigenomic differences, we studied the ontology enrichments of genes associated with different significant features (See Gene Set Enrichment Calculations). Secondly, for each comparison, we compared the expression level of genes with significant epigenomic features, to evaluate whether epigenomic differences are also reflected in gene expression differences. Lastly, we used hierarchical clustering to recognize clusters of features and genes that show consistent differences between samples (Section 3.2); using these clusters, we are able to find cluster-specific gene sets with specific gene set enrichment and expression behavior.

In addition to developing our ChromDiff pipeline, we also applied it to the epigenomic data from the Roadmap Epigenomics project; here, we present the results from applying ChromDiff for 17 group-wise comparisons (Table 3.1). Specifically, we used the segmentation of the Roadmap Epigenomics 15-state ChromHMM model, the statistical test of the non-parametric Mann-Whitney Wilcoxon test, and the multiple hypothesis correction of Benjamini-Hochberg FDR correction (See Methods). In total, we found significant features in over 70% of the biological groupings we tested

Group 1	Group 2	Property	Distinguishing features found	Height cutoff
CellLine	PrimaryCulture	type	TRUE	39
PrimaryCulture	PrimaryCell	type	TRUE	69
PrimaryCulture	PrimaryTissue	type	TRUE	83
PrimaryCell	PrimaryTissue	type	TRUE	83
Adult	Fetal	age	TRUE	80
Female	Male	sex	TRUE	80
BRAIN	MUSCLE	anatomy	FALSE	N/A
BRAIN	ESC	anatomy	TRUE	49
BRAIN	SKIN	anatomy	TRUE	44
MUSCLE	ESC	anatomy	FALSE	N/A
MUSCLE	SKIN	anatomy	FALSE	N/A
ESC	SKIN	anatomy	FALSE	N/A
BRAIN	GI	specialgi	TRUE	50
SKIN	GI	specialgi	TRUE	48
ESC	GI	specialgi	TRUE	48
MUSCLE	GI	specialgi	FALSE	N/A
SOLID	LIQUID	solid_liquid	TRUE	88

Table 3.1: Of the 17 comparisons analyzed that spanned many groups and metadata properties, 12 comparisons identified differences in chromatin state between the two groups. For those cases, gene clusters were selected based on a manually chosen height cutoff for hierarchical clustering.

(12/17) (Figure 3-1e, Table 3.1).

To validate our results and methodology, we performed randomized simulations that quantified how likely we would have obtained results with randomized data. By repeatedly shuffling the epigenomes for each of the 17 biological comparisons we tested (Section 3.2.9), we found the shuffled groups resulted in "significant features" less than 1% of the time (10/1700 simulations) (Figure 3-1f). This suggests that ChromDiff is able to pick up a real, biologically-meaningful signal from our biological comparisons (Figure 3-1e).

After validating our method based on gene bodies, we extended it to Regulatory ChromDiff, an updated method that leverages linked regulatory regions (Figure 3-2) to identify epigenomic changes at regulatory regions. Specifically, rather than using the gene body to calculate the percent coverage, we use the chromatin state coverage at linked regulatory regions as the regulatory chromatin state representation of that gene (Figure 3-2a) . Then, our method proceeds as before, calculating the regulatory

feature representation for each gene and epigenome (Figure 3-2b), correcting the feature values based on covariates, and testing for differences in feature values between the two groups (Figure 3-2c).

3.2.2 Chromatin state annotations

Our method is applicable to any chromatin state annotations that it is given. In our case, we have used the chromatin state annotations associated with the 15-state model from the Roadmap Epigenomics project based on the five core histone marks H3K4me3, H3K4me1, H3K36me3, H3K9me3, and H3K27me3,[11] including the annotations for epigenomic data from ENCODE.[19]

3.2.3 Information theoretic representation of raw feature values

3.2.3.1 ChromDiff: Gene body approach

In what we call ChromDiff, each feature is a combination of a gene and chromatin state, and for feature gene X and chromatin state Y, we calculate the probability of each gene X's assignment to chromatin state Y in each cell type, based on genomic coverage of the maximum posterior probability chromatin state annotation.[11] Specifically, the raw feature value for $feature_{X,Y}$ in epigenome Z is calculated as follows:

$$feature_{X,Y} = \frac{\sum_{i=start_X}^{end_X} 1_{a_i==Y}}{end_X - start_X}$$

where $start_X$ and end_X, are the basepair locations of gene start and gene end of gene X, and a_i indicates the chromatin state annotation at basepair i. (This equation implies usage for the applicable chromosome for gene X.)

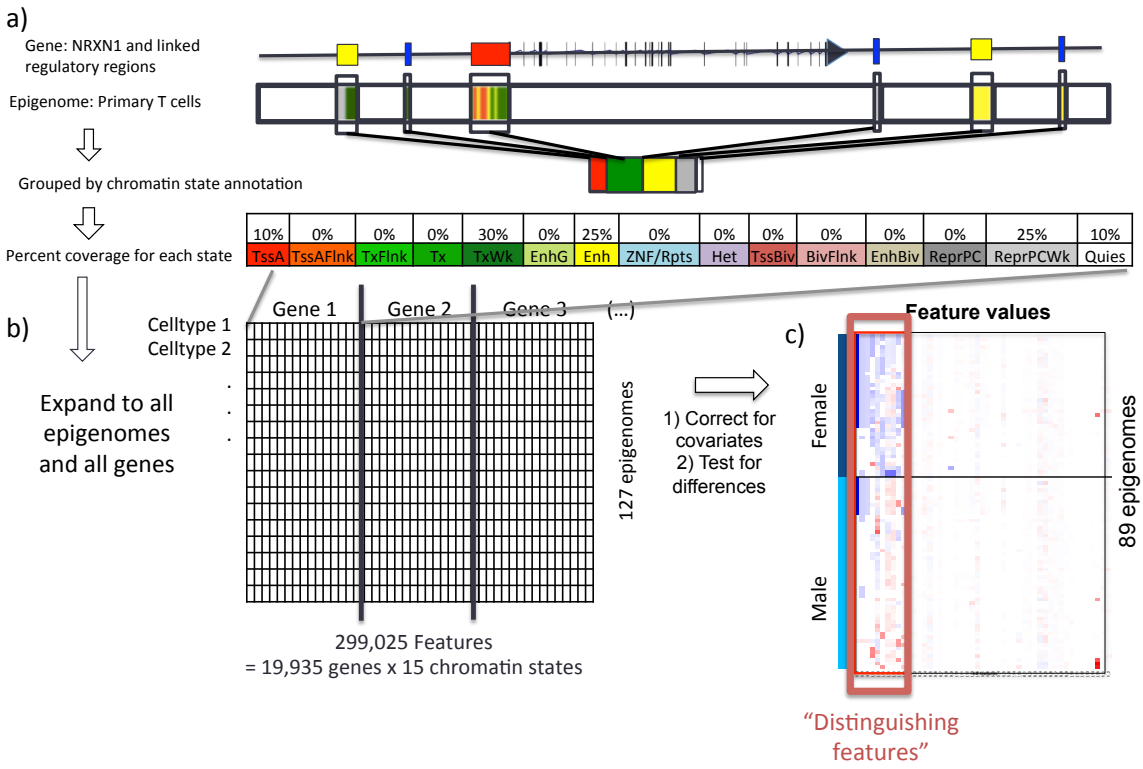

Figure 3-2: Leveraging linked regulatory regions for identification of epigenomic differences at relevant regulatory regions across groups. a. We use regulatory regions such as promoters (red), enhancers (yellow), or DNase Hypersensitive sites (blue) that are linked to a particular gene (NRXN1) to represent the regulatory chromatin state of that gene. For example, for a single epigenome (such as cultured ganglionic cells), we represent the regulatory chromatin state representation of that gene as the percent coverage of each chromatin state at the linked regulatory regions in that epigenomic sample. b. Then, we repeat the process for all 127 epigenomes and 19,935 protein-coding genes, resulting in a matrix of 127 epigenomes by 299,025 features. c. After normalizing and correcting each column in the matrix for covariate factors, we compare the female and male epigenomes of the original 127 epigenomes to identify regulatory features that exhibit statistically significantly different values between female and male epigenomes.

3.2.3.2 Regulatory ChromDiff: Linked regulatory region approach

Short-range and long-range regulatory interactions between regulatory regions with both cis and trans effects, such as enhancers and promoters, have been shown to play an important role in gene activity.[132] For this reason, we also expanded our method to allow the comparisons based on regulatory regions, including linked enhancers based on inference across gene expression data through gene modules and enhancer modules,[133] linked DNase Hypersensitivity sites based on proximity,[134] and proximal promoter regions, as defined by the 2kb centered around the TSS.

By studying the chromatin state of regulatory regions linked to genes, we can identify regulatory differences that point to biological mechanisms underlying epigenomic differences. By looking not just at the chromatin state of the gene body, but also at linked regulatory regions, we find many epignomic differences that were not identified based on the gene body alone.

In what we call Regulatory ChromDiff, we integrate regulatory regions by calculating the raw feature value $feature_{X,Y}$ for each gene X and chromatin state Y in epigenome Z as follows:

$$feature_{X,Y} = \sum_{i \in I} \frac{\sum_{i=start_j}^{end_j} 1_{a_i==Y}}{end_j - start_j}$$

where $start_j$ and end_j, are the basepair locations of the start and end regulatory regions j linked to gene X, and a_i indicates the chromatin state annotation at basepair i. (This equation implies usage for the applicable chromosome for gene X.)

3.2.4 Gene annotations

We used all protein-coding genes with corresponding gene ids and positions as given in Gencode GENCODE v10[21] for compatibility with the Roadmap Epigenomics Consortium,[11] with the exception of genes encoded on chromosome Y. Gene symbols for gene set enrichment calculations were also taken from the Gencode GENCODE

annotations.

3.2.5 Covariate correction of ChromDiff feature values

For each feature value and each comparison, we fitted a logistic regression model to our raw feature values across all the epigenomes, excluding any covariate factors that were explicitly being tested by the comparison. As our raw feature values were bounded as fractional values between 0 and 1, logistic regression allowed for appropriate correction. Specifically, we used glm, the generalized linear model functions available in the stats package in R.[135] As defined below, we used the deviance residuals from our fitted logistic model as our corrected feature values.

Formally, if we are comparing traits A and B of property C, then we have N_c explanatory variables for our model, where N_c is the number of explanatory variables after excluding any that correspond to property C. As in standard logistic regression, we are modeling the β's in the following formula:

$$\hat{feature}_{X,Y} = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{N_c} \beta_i x_i)}}$$

such that x_i corresponds to the value of the ith explanatory variable for $feature_{X,Y}$ for which we are adjusting. Therefore, the corrected feature values take on the value of the deviance residual, which is:

$$sign(feature_{X,Y} - \hat{feature}_{X,Y}) \cdot \sqrt{\hat{d_{X,Y}}^2}$$

where:

$$\hat{d_{X,Y}}^2 = 2 \cdot (feature_{X,Y} \cdot \log \frac{feature_{X,Y}}{\hat{feature}_{X,Y}} + (1 - feature_{X,Y}) * \log \frac{1 - feature_{X,Y}}{1 - \hat{feature}_{X,Y}})$$

A final detail to note is that logistic regression requires the conversion of fractional

values to "successes" and "failures", so we convert the $feature_{X,Y}$ fractions into $feature_{X,Y} \cdot N$ successes and $(1 - feature_{X,Y}) \cdot N$ failures, where N is the length of the gene, or $end_X - start_X$.

The four covariate property factors we corrected for are: 1) sex of the sample donor, 2) laboratory that processed the sample, 3) sample type, and 4) whether the sample was a solid or liquid sample. We converted the categorical covariate factors into "continuous" explanatory variables; if there were c categories for a certain factor, this resulted in c explanatory variables for that factor by converting boolean variables into binary values. For any samples that were mixtures of n multiple categories, each corresponding explanatory variable was given a value of 1/n.

More specifically, the four laboratories that contributed to the data (BI, UCSD, UCSF-UBC, and UW) resulted in four explanatory variables, with one for each lab. For any epigenome that was completely generated at a single lab, that labâĂŹs explanatory variable was given a value at 1, while the other labs were given 0. When multiple labs contributed to an epigenome, the corresponding lab covariate factor was calculated as $\frac{1}{L_e}$, where L_e is the number of labs that contributed to epigenome e. Similarly, for sex, if there were male and female donors for a given epigenome, a value of $\frac{1}{2}$ was given for the female and male covariate values; otherwise, the variable for the donor's sex was given a 1 and the other sex variable was a 0. For sample types, the value for a covariate factor was 1 for the correct sample type and 0 for the other types, since each epigenome was annotated as one of the possible five possible sample types: Cell Line, Derived Cell Line, Cancer Cell Line, Primary Cell, and Primary Tissue. For the sample state covariate factors, each sample was either annotated as Solid, Liquid, or Neither, which was translated into three corresponding explanatory variables with values of 1 for the correct annotation and 0 for the others.

3.2.6 Group-wise comparison statistics

For comparison of two groups, our method currently supports three statistical tests: the two-sided t-test, the f-tests, and the Mann-Whitney-Wilcoxon test. Any of these tests can be used to identify features that show statistically different feature values

across the two groups. In the presented results, we used only the Mann-Whitney-Wilcoxon test.

3.2.6.1 Multiple hypothesis correction

Our pipeline also supports Bonferroni, Benjamini-Hochberg, or Benjamini-Yekuteli correction[135–137] on all of our p-values based on the number of features tested for each comparison. In this analysis, we used only the Mann-Whitney test with Benjamini-Hochberg correction. Features that had a corrected p-value of less than .05 after correction, based on the number of total features tested, were considered significant.

Though our features are slightly dependent, due to the connection between each gene and its fifteen features, Benjamini-Hochberg FDR correction is always valid for dependent tests that uphold Positive Regression Dependency in a Subset (PRDS) and Benjamini-Hochberg also performs well in many practical cases and simulations.[138–142] For completeness, we also compared the number of distinguishing features found with p-values correction based on the more conservative Benjamini-Yekuteli procedure,[137] which always controls the FDR under any dependency or distribution environment, and we found that most of our biological comparisons still result in significant chromatin state differences (Figure 3-3).

3.2.7 Gene set enrichment calculations

Once we have identified genes that are associated with at least one significant feature, we calculated hypergeometric[143] p-values, effectively using the Fisher's exact test, with Storey's FDR q-value correction[144] using gene sets from MSigDB.[145] The databases used by MSigDB were C2 (curated gene sets) and C5 (GO gene sets) gene sets, downloaded by gene symbols. We used the gene symbols associated with each gene, as provided by GENCODE gene IDs. This analysis was performed on all the significant genes identified in a comparison, as well as the clusters of sampled genes identified (Sections 4.2.3 and 4.2.1).

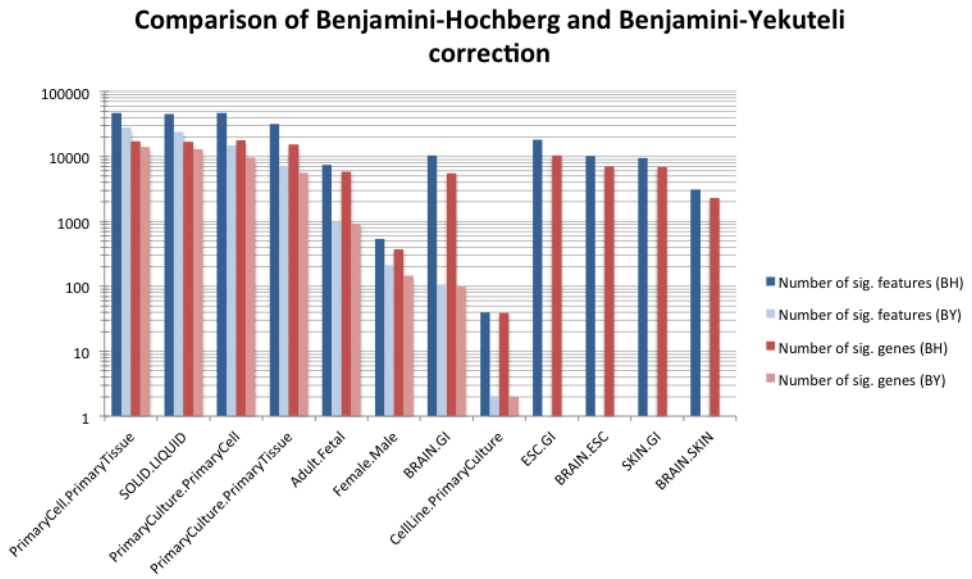

Figure 3-3: Even with the conservative Benjamini-Yekuteli multiple hypothesis correction, we still identify significantly different features and genes in 7 of 12 cases. As Benjamini-Yekuteli multiple hypothesis correction is applicable for any distribution or dependency structure, we quantified the effect that it would have on our significant results. We found that in 7 of the 12 comparisons (and in all four comparisons presented in this paper), ChromDiff would still identify many significant features and genes using Benjamini-Yekuteli correction.

3.2.8 Expression data analysis

3.2.8.1 Expression data

RNA-Seq data from the Roadmap Epigenomics and ENCODE projects[11,19] was used, when available. Specifically, the per-gene RPKM values provided by Roadmap Epigenomics[11] for ENSEMBL-defined protein-coding genes were used directly.

3.2.8.2 Significant gene expression differences

P-values were calculated using the two-sided Mann-Whitney test on all expression values of relevant genes in one group against the other group. This analysis was performed on all the significant genes identified in a comparison, as well as the clusters of sampled genes identified (Sections 4.2.3 and 4.2.1).

3.2.8.3 Proportion of genes with differential gene expression

For each ChromDiff-identified gene (a gene associated with a significant feature for a given comparison), a two-sided Mann-Whitney test was used on gene expression values between each grouping of epigenomes. Benjamini-Hochberg FDR correction[136] was used on these p-values, and genes with adjusted p-values of less than .05 were considered to have significant differential gene expression. This allowed us to calculate the percent of significant genes that had differential gene expression. This analysis was repeated for all the genes that were not identified by ChromDiff, for comparative purposes and to calculate the odds ratios (Section 3.2.8.4).

3.2.8.4 Odds ratio for differential gene expression and distinguishing genes

We calculated the odds ratio, 95% confidence interval, and corresponding p-values for the relationship between epigenomically distinguishing genes and differentially expressed genes.[146,147] Specifically, we can define a as the number of distinguishing genes with differential expression, b as the number of distinguishing genes without differential expression, c as the number of non-distinguishing genes with differen-

tial expression, and d as the number of non-distinguishing genes without differential expression. The log odds ratio, or ln OR, is therefore $\ln \frac{a \cdot d}{b \cdot c}$.

The standard error, or SE, of the log odds ratio is calculated as $\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$. The 95% confidence interval of the log odds ratio is defined as $(\ln \mathrm{OR}) \pm 1.96 \cdot \mathrm{SE}$. Using a p-value threshold of .05, there is a significant relationship between the distinguishing genes and differentially expressed genes when the 95% confidence interval for the log odds ratio does not include 0.

Specifically, the p-value can be calculated from the z-score, as $z = -\frac{\ln OR}{SE}$, and the corresponding p-value is $e^{-.717 \cdot z - .416 \cdot z^2}$.

3.2.8.5 Covariate correction for gene expression

Covariate correction was performed in the same way as described above (Section 3.2.5), except that linear regression was used instead of logistic regression, due to the unbounded nature of RPKM values.

3.2.9 Randomized simulations

To confirm the biological relevance of our tests, we performed 100 randomization tests for each biological comparison. During each randomization trial, we randomly shuffled the labels on the epigenomes we were testing, thereby retaining groups of matched size to the original comparison. Then we performed the same covariate correction and statistical testing as described above, and counted the number of significant distinguishing features found, if any.

Formally, let X_A and X_B be the sets of epigenomes that correspond, respectively, to the traits A and B of category C. (For example, A=Female, B=Male, C=Sex of donor.) Then define $X = X_A \cup X_B$. For each randomization trial $t_i^{A,B}$, for $1 \leq i \leq 100$, randomly draw a new X_A from X of size $|X_A|$ with uniform probability (without replacement). Then define $X_B = X - X_A$, which means, by construction, that X_B will be size $|X_B|$. Then, as described above, perform covariate correction for every category C' such that $C \neq C$, and perform the statistical Mann-Whitney-Wilcoxon

test to identify distinguishing features as usual.

We then summarized our findings as shown in Figure 3-1e, by counting the fraction of all randomized trials that resulted in any significant distinguishing features, and contrasted this with the fraction of biological comparisons that resulted in found significant distinguishing features. Specifically, Figure 3-1e depicts, in blue, the following fraction:

$$\frac{\sum_{A,B} \sum_{i=1}^{100} 1_{t_i^{A,B} \text{ found significant features}}}{\sum_{A,B} 100}$$

Specifically, Figure 3-1f depicts, in blue, the following fraction:

$$\frac{\sum_{A,B} 1 \text{ comparison of } X_A \text{ and } X_B \text{ found significant features}}{\sum_{A,B} 1}$$

3.3 Results

3.3.1 Identified genes are enriched for differential expression

The genes identified from our comparisons often exhibited different expression levels between the groups compared. To quantify this, for each of the 12 comparisons with distinguishing features (Table 3.1), we calculated how many of our identified genes had differential gene expression between the two groups of the comparison (Figure 3-4a) (Section 3.2.8.3). Three comparisons that revealed epigenomic differences did not have any differentially expressed genes: Brain/ESC, CellLine/PrimaryCulture, and ESC/GI. Furthermore, in the 9 cases with differentially expressed genes, the epigenomically distinguishing genes included proportionally more differentially expressed genes than the non-identified genes, with log odds ratios ranging from $0.13 - 2.26$ and 95% confidence intervals as shown (Fig. 6b). In all 9 cases, the increased proportion of differentially expressed genes is found to be significant, with the null hypothesis of $\ln(\text{OR}) = 0$, or equivalently $\text{OR} = 1$, falling outside the 95% confidence interval.

To more precisely identify genes directly associated with each biological feature,

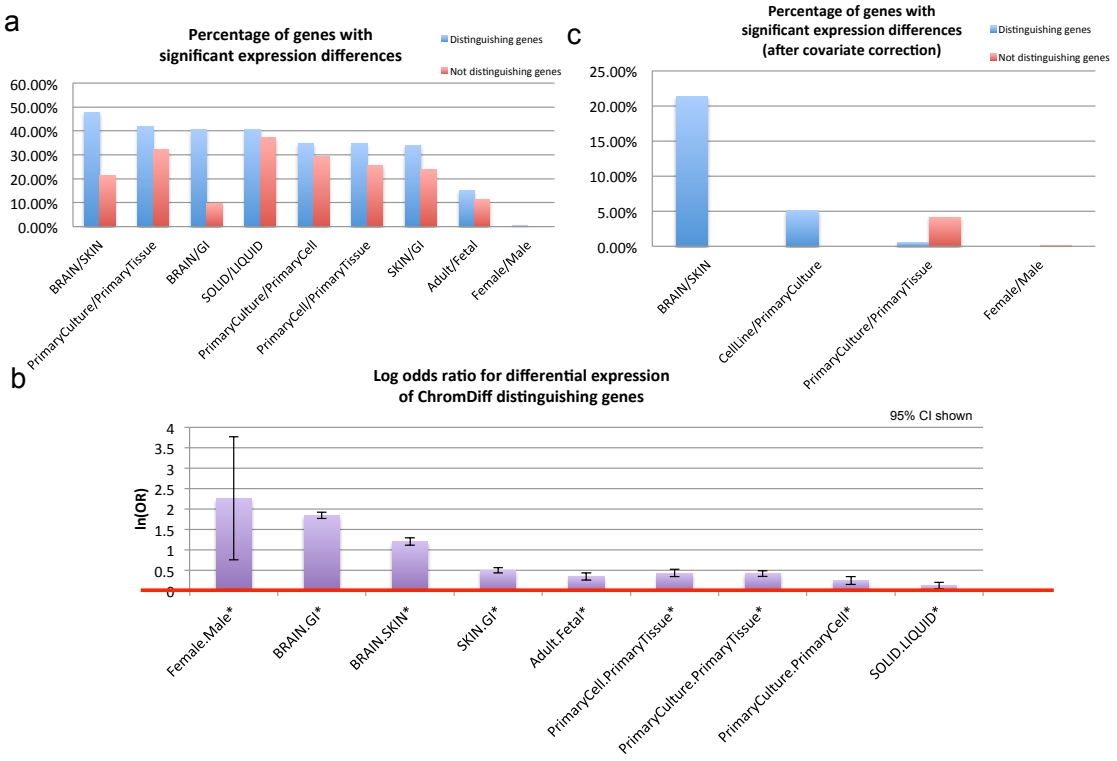

Figure 3-4: Epigenomically distinguishing genes are enriched for differential expression. By analyzing expression of genes that our method identifies as being part of distinguishing gene and chromatin state combinations, we find that that our method both recaptures differential gene expression and identifies distinguishing epigenetic context not captured by differential gene expression. (We performed this analysis on the 12 comparisons that produced distinguishing epigenomic features.) a. Overall, identified genes are enriched for differentially expressed genes, although less than 50% of the genes identified are significantly differently expressed. (The three comparisons that resulted in no differentially expressed genes are excluded.) b. In every of our nine remaining comparisons, our identified genes were enriched for differentially expressed genes overall (as designated by asterisks), based on calculation of the log odds ratio and the corresponding 95% confidence interval. (All comparisons with no differentially expressed genes were omitted.) Error bars shown designate the 95% confidence interval ($p < .003$ in all cases, 2-sided Z-test) c. After correcting for covariates in expression data, ChromDiff identifies all of the differentially expressed genes in two of the remaining four cases. (Eight comparisons yielded no significant expression differences and are therefore excluded.)

we further used linear regression to correct the RPKM values for the same covariate factors previously used on the epigenomic features (production center, sex of donor, sample state, and sample type). After covariate correction as described in Section 3.2.5, only four comparisons had any genes that were found to be differentially expressed: Brain/Skin, CellLine/PrimaryCulture, PrimaryCulture/PrimaryTissue, and Female/Male. Surprisingly, all of the differentially expressed genes found in the Brain/Skin and CellLine/PrimaryCulture comparisons were also identified by ChromDiff as distinguishing genes (Figure 3-4c). On the other hand, many of the differentially expressed genes found in the Primary Culture/Primary Tissue and Female/Male comparison were not identified based on epigenomic changes.

Lastly, due to the fact that we had more samples with chromatin state data than gene expression data, we also applied ChromDiff to the original 17 comparisons while excluding any epigenomes without gene expression data. Due to the reduced power, in this case, only four biological comparisons yielded epigenomically distinguishing features and genes (Table 3.2). However, for three of those biological comparisons, no differentially expressed genes were found after correcting for covariates, while the fourth comparison revealed that about 42% of the differentially expressed genes were identified by ChromDiff (Figure 3-4d).

Overall, we have strong evidence that comparative chromatin state and differential gene expression methodologies are complementary approaches for comparative analysis. Whether we use covariate correction or limit ourselves to epigenomes with expression data, the main result is the same: the genes with differential expression are always a minority of the entire set of ChromDiff-identified genes, implying that many genes with epigenomic differences are not differentially expressed, and would be missed by differential gene expression analysis. While differential expression analysis has proven to be and will continue to be extremely useful, the chromatin state comparison provided by ChromDiff provides another lens with which to view comparative analysis, and using these tools in combination will only improve the power of the analysis.

	PrimaryCulture vs PrimaryCell	PrimaryCell vs PrimaryTissue	Solid vs Liquid	PrimaryCulture vs PrimaryTissue
Distinguishing genes that are diff. expressed	0	0	0	129
Total distinguishing genes	198	939	3459	13727
Not distinguishing genes that are diff. expressed	0	0	0	178
Total not distinguishing genes	19546	18805	16285	6017

Table 3.2: ChromDiff can capture epigenomic differences even when there are no differentially expressed genes. In three comparisons that yield no differentially expressed genes, ChromDiff still identifies genes showing epigenomic differences. These are three out of the four comparisons that still yield epigenomically distinguishing results when ChromDiff is limited to epigenomes with expression data. (This analysis used only epigenomes with expression data that was corrected for the same covariates as the chromatin state data. Any comparisons that yielded no significant ChromDiff differences are excluded.)

3.3.2 ChromDiff outperforms other method for epigenomic comparison

To our knowledge, only one previous method exists to address comparison of epigenomic groups, and it utilizes PCA analysis on the differences between the means of the groups.[131] However, this approach presents a number of limitations. First, this approach innately ties the identification of combinatorial histone mark patterns to PCA analysis. Meanwhile, many segmentation methods have proven the usefulness of different machine learning approaches to identify combinatorial chromatin states, such as HMMs[23,25] and Bayesian networks.[24] Since ChromDiff is compatible with any segmentation, it enables the use of a variety of existing and future methodologies. Second, public data resources such as Roadmap Epigenomics and ENCODE empower researchers everywhere to make discoveries through their analyses; however, these resources also necessitate less standardized data, as the data is often generated from a variety of labs, individuals, and samples. ChromDiff corrects for covariate

factors based on the metadata that the user provides, and in doing so, it is uniquely positioned to perform comparative epigenomic analysis from these resource datasets. In contrast, PCA is designed for use with replicates, which limits the type and number of biological comparisons that can be performed from any dataset. Lastly, PCA is sensitive to relative scaling of values, while our rank-based statistical tests produce the same results regardless of scaled values.

To validate the expected improvement that ChromDiff provides for this analysis, we applied the differential PCA method[131] to the Roadmap Epigenomics data for the same five histone marks used by ChromHMM: H3K4me3, H3K4me1, H3K36me3, H3K9me3, and H3K27me3.[11] dPCA reports whether any differential principal components were found, based on a cutoff of a signal-to-noise ratio of 5, since accuracy suffers substantially for components with a lower signal-to-noise ratio.[131] In three important comparisons, dPCA fails to recover any significant principal components, thereby generating no follow-up regions: specifically, dPCA found no genes with epigenomic differences for comparisons based on sex (Female/Male), developmental age (Adult/Fetal), and type (CellLine/PrimaryCulture) (Table 3.3). These results indicate that dPCA is unable to identify differences for two biological properties (sex and developmental age); furthermore, epigenomic sex differences due to X chromosome inactivation is one of the most studied and well-understood examples of differing epigenomic state, and as such, it represents a "gold standard" that dPCA is unable to reproduce.

While dPCA is able to identify epigenomic differences for some comparisons when ChromDiff is not (Table 3.3), the genes that dPCA identifies are much less specific to the biological comparison, likely due to lack of covariate correction. For example, we noted that the gene set lastowska neuroblastoma copy number dn was in the top two enriched gene sets for every comparison (14 of 14) that produced results, and chen liver metabolism qtl cis was the other gene set in the top two for over half of the comparisons (8 of 14).

To quantify this lack of specificity from dPCA, we calculated the Jaccard similarity score of the lists of resulting genes for pairs of comparisons, and as expected, we see

Biological comparison	number of genes (dPCA)	number of genes (ChromDiff)
Adult_Fetal	0	5852
BRAIN_ESC	6488	7059
BRAIN_GI	4966	5533
BRAIN_MUSCLE	3750	0
BRAIN_SKIN	10080	2286
CellLine_PrimaryCulture	0	39
ESC_GI	7232	10457
ESC_SKIN	7847	0
Female_Male	0	369
MUSCLE_ESC	6450	0
MUSCLE_GI	3863	0
MUSCLE_SKIN	3256	0
PrimaryCell_PrimaryTissue	18231	17109
PrimaryCulture_PrimaryCell	18476	17827
PrimaryCulture_PrimaryTissue	16580	15481
SKIN_GI	12244	6830
SOLID_LIQUID	18423	17001

Table 3.3: We compare ChromDiff to dPCA, the other existing method for group-wise epigenomic comparisons, by applying both methods to the same Epigenome Roadmap dataset using the same group comparisons. Although dPCA is unable to identify any epigenomic differences, ChromDiff identifies genes with chromatin state changes in the cases of a) Adult vs Fetal, b) Cell Lines vs Primary Cultures, and c) Female vs Male, which is especially relevant as they span different biological properties of age, sample heterogeneity, and sex.

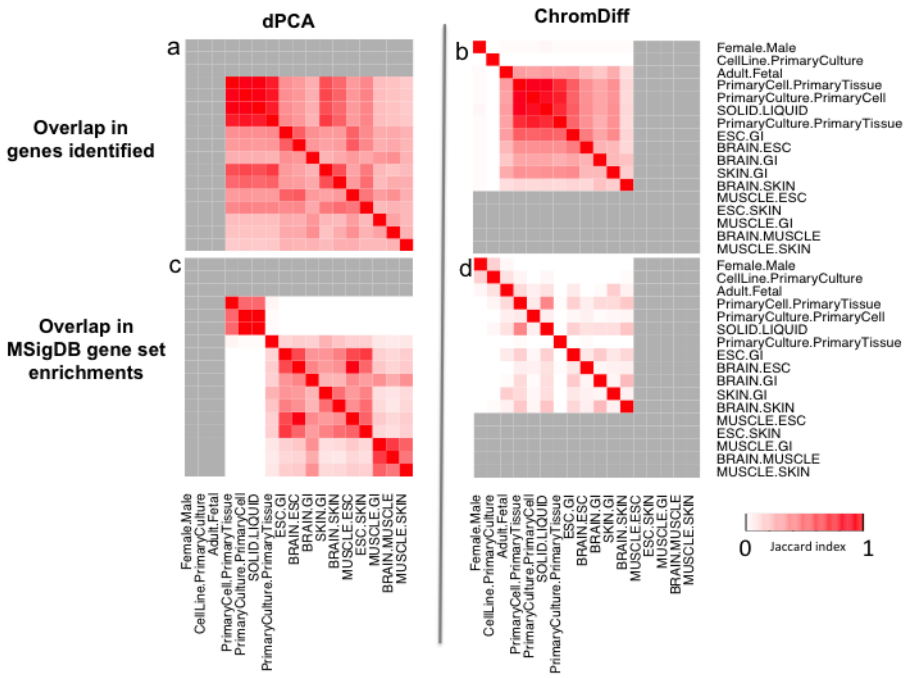

Figure 3-5: ChromDiff outperforms dPCA in identification of comparison-specific genes. a. dPCA re-discovers the same genes in many of its varying biological comparisons, while b. ChromDiff more frequently identifies different genes in different comparisons, based on the Jaccard index. Similarly, c. the enriched gene sets and pathways identified by dPCA are markedly similar for many of their comparisons, while d. ChromDiff achieves higher specificity that produces different, relevant gene set enrichments in different cases, based on the Jaccard similarity index.

higher similarity scores for the dPCA results than the ChromDiff results (Figure 3-5a,b). Even more strikingly, we see very high similarity between the enriched MSigDB gene sets for the genes identified by dPCA (Figure 3-5c), while ChromDiff returns gene set enrichments specific to that comparison (Figure 3-5d). Since some of these comparisons also share epigenomic groups (for example, the same brain epigenomes are used for Brain/ESC and Brain/GI), we filter out similarity scores for pairs of comparisons with overlapping groups (Figure 3-6a-d). After filtering, we find that dPCA has a higher average similarity score among unrelated comparisons for both gene and MSigDB results (Figure 3-6e).

These results show that for the Roadmap Epigenomics dataset, ChromDiff is more powerful than dPCA: ChromDiff can identify genes showing important epigenomic changes even when dPCA does not have enough power, and ChromDiff also more

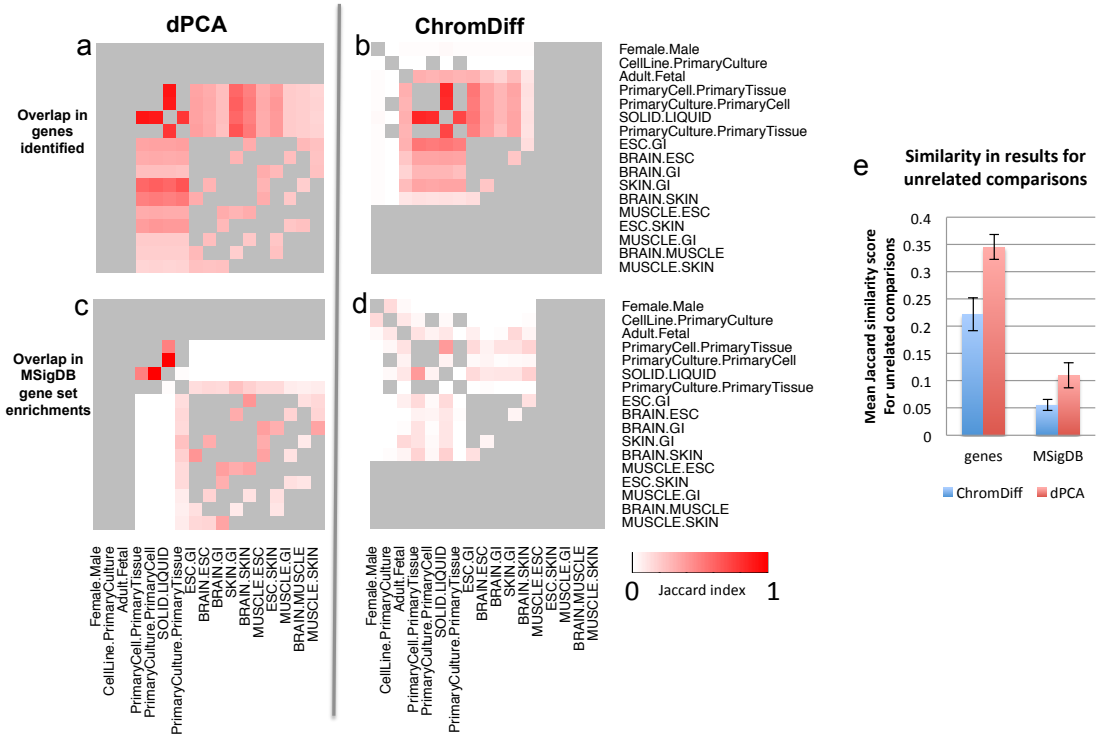

Figure 3-6: ChromDiff identifies more specific results than dPCA. After filtering out pairs of comparisons with shared epigenomic groups, we find that a. dPCA's gene results are less specific than b. ChromDiff's gene results for unrelated comparisons. Similarly, c. dPCA's gene set enrichments are less specific than d. ChromDiff's gene set enrichments for unrelated comparisons. e. We quantify this result by confirming that dPCA's results have higher mean similarity scores for unrelated comparisons than ChromDiff does, with bars displaying standard error of the sample mean. These were calculated from the 53 and 58 pairs of unrelated comparisons for ChromDiff and dPCA results, respectively, as shown in a-d.

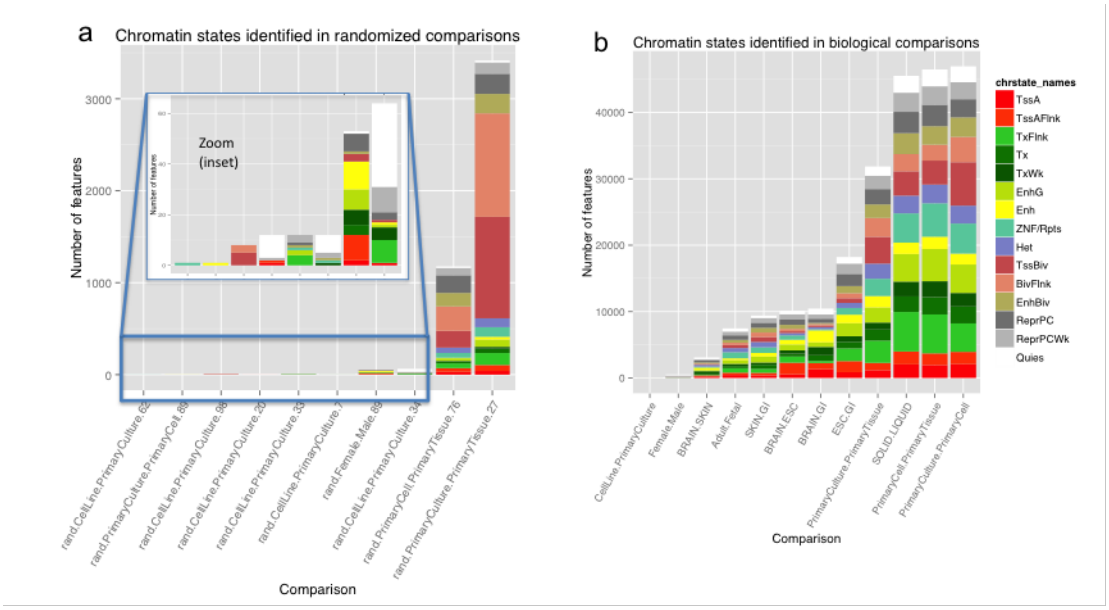

Figure 3-7: A variety of chromatin states are identified in simulations and applications. a. In randomized comparisons, we find a wide variety of chromatin state distributions in the distinguishing features found, while in b. biological comparisons, we found that all the chromatin states were well represented.

identifies specific and relevant gene sets than dPCA, likely due to its ability to correct for covariates.

3.3.3 ChromDiff identifies relevant genes and chromatin states independent of gene size and chromatin state.

If ChromDiff operated with a bias towards certain chromatin states or gene sizes, this should become evident in the results from our simulations. However, in practice, we found no consistent bias towards any chromatin state (Figure 3-7a) or gene size (Figure 3-8a) in results from our randomized simulations. This provides confidence that the distribution of chromatin states (Figure 3-7b) and gene sizes (Figure 3-8b) in our real results is based on real biological signal; for example, some brain-specific genes have been shown to be long[148] and three of our four comparisons that identified the longest genes involved brain epigenomes.

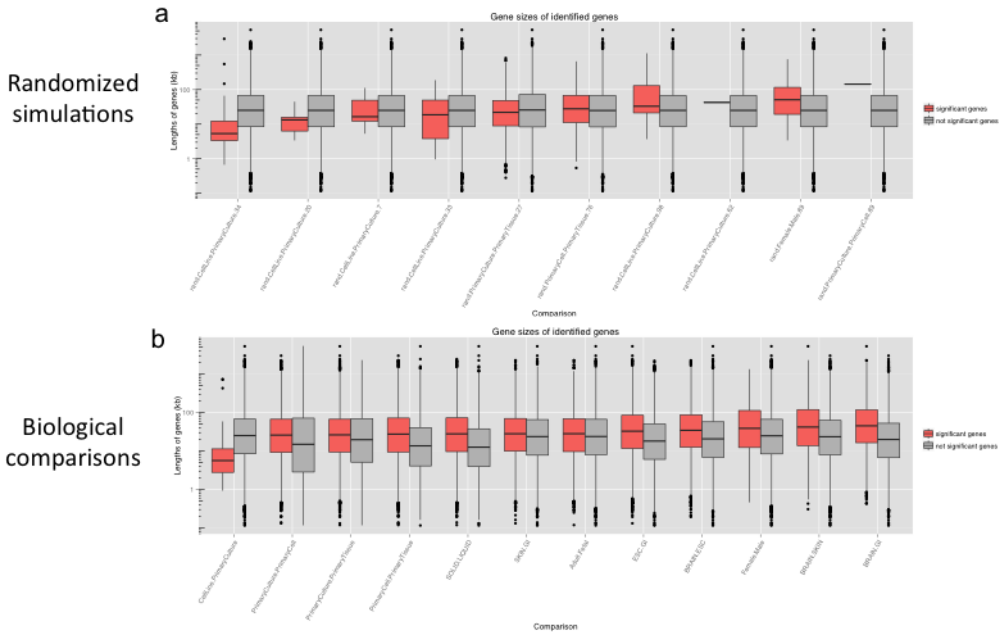

Figure 3-8: Genes of various sizes are identified in randomized simulations, while longer genes are identified in biological comparisons. a. In randomized simulations, we identify genes with a variety of gene sizes, suggesting ChromDiff does not bias for certain gene sizes. b. In biological comparisons, the genes identified were often longer, suggesting that longer genes exhibit more epigenomic changes.

3.3.4 Regulatory ChromDiff identifies additional and new genes when studying linked enhancers and regulatory regions.

As described in Section 3.2.3.2, we also developed a second version of ChromDiff, called Regulatory ChromDiff, which uses linked regulatory regions to search for epigenomic differences between sample groups. We applied this method to three different sets of regulatory regions: promoter regions, enhancer regions and DNase Hypersensitive sites. These regulatory regions were leveraged for 19 comparisons of biologically relevant sample groups, and we found that 15 of the 19 comparisons resulted in identification of significantly distinguishing features (Figure 3-9). Further, in Figure 3-9, we show the number of genes identified by Regulatory ChromDiff for each of 15 comparisons based on the three sets of regulatory regions.

We also applied the original gene body ChromDiff method to the same 15 comparisons. This allowed us to separate the identified genes into two categories: whether the gene was identified only by the Regulatory ChromDiff method, or whether it had previously been identified by the gene body approach. In Figure 3-9, we then color the gene counts correspondingly, with genes previously identified by the gene body approach colored in gray, while genes only identified by the Regulatory ChromDiff approach are shown in red (for promoters), yellow (for enhancers), and blue (for DNase Hypersensitive sites).

We can also include genes identified only by the gene body ChromDiff approach. Specifically, we look at all genes identified for a particular comparison using either ChromDiff or Regulatory ChromDiff (for a particular set of regulatory regions). Then, in Figure 3-10, we show the proportion of these genes that were identified only by ChromDiff (in white), only by Regulatory ChromDiff (in red, yellow, or blue for promoters, enhancers, or DNase Hypersensitive sites), or by both ChromDiff and Regulatory ChromDiff.

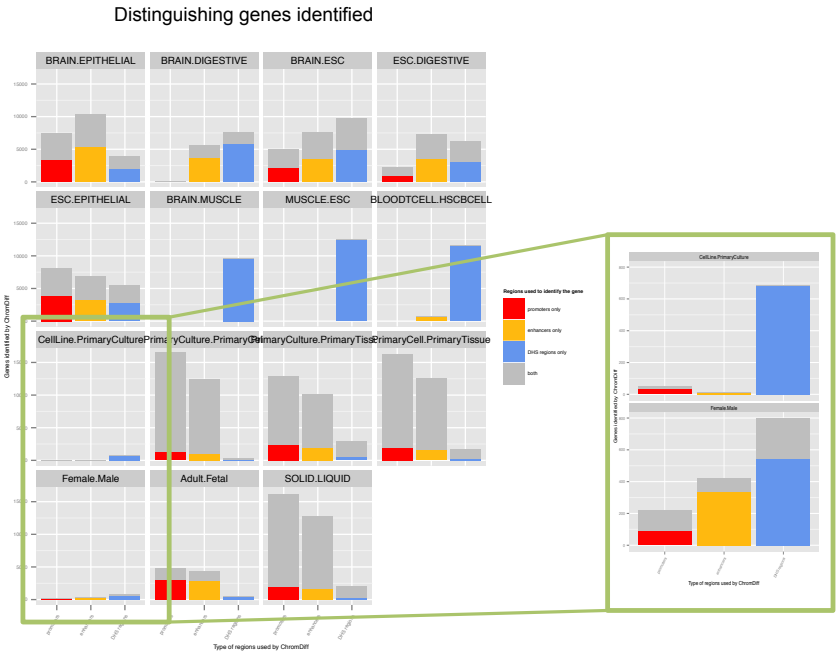

Figure 3-9: ChromDiff is applied to 15 biological comparisons and identifies new genes when using promoters (shown in red), linked enhancers (shown in yellow), and linked DNase Hypersensitive sites (shown in blue), compared to the gene body approach. The total number of distinguishing genes identified for each comparison are shown in the cumulative bars, based on the summation of the region-specific genes (shown in red, yellow, and blue) and the genes identified by both versions of ChromDiff (shown in gray).

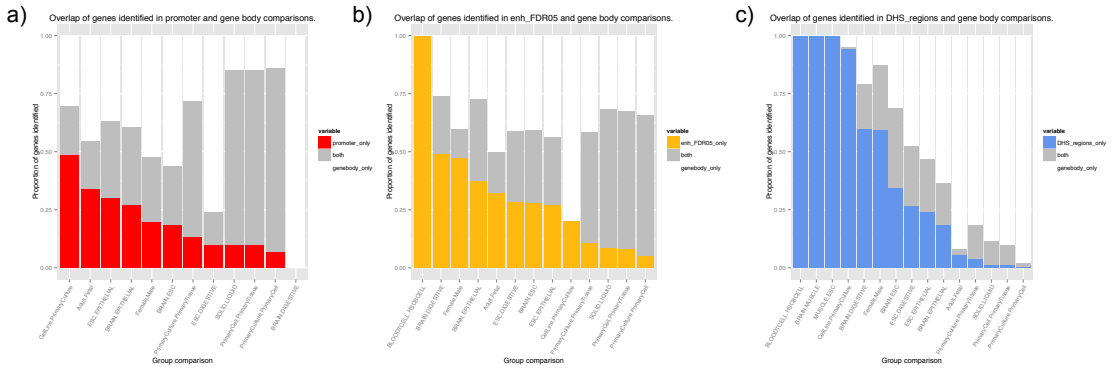

Figure 3-10: Proportion of distinguishing genes identified by Regulatory ChromDiff with a) promoters, b) enhancers, and c) DNase Hypersensitive regions, compared to the gene body ChromDiff approach. Specifically, we show the proportion of genes identified by both approaches (gray), genes identified only by the gene body approach (white), and genes identified by Regulatory ChromDiff (red, yellow, or blue).

3.4 Software download

Code for ChromDiff, results, instructions for usage, data used here, and additional information can be found at http://compbio.mit.edu/ChromDiff. It is also publicly available on Github at https://github.com/angieyen/ChromDiff. ChromDiff is freely available for download and usage under a GPL 3 license.

3.5 Contributions

Overall, our method for comparing epigenomic states between groups of epigenomes highlights chromatin state changes at genes with relevant functions by generating an information theoretic encoding of the epigenome and isolating differences corresponding to a single biological attribute with covariate correction. Applications of our method reveal that different chromatin states and genes play important distinguishing roles in different comparisons. We further find an overall enrichment for differentially expressed genes in our identified gene sets, and in some cases, ChromDiff even identifies all differentially expressed genes based solely on the epigenomic data. We validate our methodology by showing that shuffled simulations almost always yield no epigenomic differences and that our approach outperforms the only existing method for group-wise epigenomic comparative analysis, particularly in the specificity of our results. Due to these findings, we believe that our method is a powerful and innovative tool that will only increase in power to elucidate biological differences as more epigenomes become available.

The field of systems biology seeks to uncover the dynamics of gene regulatory processes in diverse biological functions including differentiation and disease. To date, analyses have focused primarily on differential gene expression analysis and epigenomic comparisons limited to histone mark signals. By comparing chromatin state annotations here, the results suggest a rich set of molecular features distinguish differential gene activity across different biological parameters. Our methods leverage both gene bodies (ChromDiff) and linked regulatory regions (Regulatory ChromDiff)

to show that genes, as well as distal and proximal regulatory regions, exhibit relevant epigenomic differences across sample type. Thus, we believe that our comparative analysis of chromatin states provides value by elucidating how the chromatin states associated with each gene vary across biological parameters and conditions.

Previous methods for comparative epigenomic analysis have already been fruitful, but ChromDiff is ideally suited for analysis of resource datasets, due to its ability to control for external covariates and perform multiple group-wise comparisons using different attributes on the same dataset. Therefore, as more epigenomes become available in less controlled settings, ChromDiff will continue to be a valuable tool and pipeline for research. Furthermore, using chromatin states rather than the raw underlying signal allows for an abstraction based on segmentation that can be used with varying underlying methodologies, and it also provides an additional lens with which to examine the interplay of gene expression, epigenomic context, and biological attributes and pathways.

Chapter 4

Comparisons of epigenomes reveal distinguishing chromatin states and genes

4.1 Introduction

By leveraging epigenetic information with computational analysis, we are able to identify novel biological insights regarding epigenomic differences across many properties. Specifically, to demonstrate the power of our method described in Chapter 3, we apply ChromDiff to identify genes and chromatin states that differentiate epigenomes across donor sex, tissue type, sample state, and donor developmental age. The results reveal that distinct types of epigenomic features vary with different biological properties and strongly validate our statistical approach. In addition, our specific comparisons result in new biological insights on the types of epigenomic features and pathways that underlie each of our comparisons. Lastly, we present evidence that chromatin state changes at linked regulatory regions, in addition to changes at gene bodies, can also allow scientists to identify the gene, pathway, and expression changes relevant for a particular characteristic.

4.2 Methods

As described in Chapter 3, we used ChromDiff to identify group-wise chromatin state differences. Here, we describe the visualization and sampling techniques used to generate our result figures.

4.2.1 Gene cluster identification

Clustering of genes was based on hierarchical clustering using the complete linkage method based on Euclidean distances, as implemented by hclust in the R stats package.[135] After obtaining the dendrogram for the hierarchical clustering, we manually identified a cutoff for each comparison while taking into account cluster homogeneity and size. The resulting clusters were annotated on the heatmaps if they included at least 5% of all elements clustered. We also performed calculations for enriched gene set enrichments and gene expression analysis on these annotated clusters, as described in Section 3.2.7 and Section 3.2.8.

4.2.2 Sampling distinguishing features

For visualizations, we sampled down to 10,000 distinguishing features when more features than this were identified. The sampled features and their associated genes were then used for all heatmaps, such as gene and chromatin state combination plots, feature enrichment plots, most abundant (dominant) chromatin state plots, and gene expression difference plots. To sample to the 10,000 distinguishing features that would be most informative, we prioritized genes that were associated with the greatest number of significant distinguishing features, breaking ties based on the p-value of the most significant associated feature. Then, all features associated with the prioritized genes were chosen, until 10,000 features were reached.

4.2.3 Sampling significant distinguishing genes

All genes corresponding to sampled distinguishing features were retained.

4.2.4 Ordering of rows or columns

When ordering any rows or columns, the dendrogram was generated based on the clustering, which was then ordered by the means of the vectors.

4.2.4.1 Ordering of heatmaps based on genes

We first ordered our dominant chromatin state heatmaps. The columns were ordered based on the column means according to the dendrogram for hierarchical clustering described in Section 4.2.1. The rows were ordered separately for the two biological groupings. First, for each group, we identified the dendrogram for the corrected feature values via hierarchical clustering of hclust[135] and ordered the dendrogram based on the row means. Then, we simply concatenated the orderings for the two groups. Finally, unless otherwise noted, these gene and epigenome orderings were carried throughout all of the heatmaps showing gene information.

4.2.5 Dominant (most abundant) chromatin state heatmaps

For each comparison, these heatmaps visualized the significant genes as columns and relevant epigenomes as rows. Each cell shows the corresponding color for which chromatin state was most present for that gene in that epigenome. Colors for chromatin state were taken from the ChromHMM state coloring as given by the integrative Roadmap Epigenomics project. Cell type and gene orderings were calculated as described above (Ordering of heatmaps based on genes).

4.2.6 Gene expression heatmaps

Row (cell type) and column (gene) orderings were copied from the significant gene majority state heatmaps, as described in Ordering of heatmaps based on genes. The color of each cell represents the corresponding $\ln(\text{RPKM} + 1)$ value for each gene and epigenome combination. The colorscale sets the minimum expression value to be red, the median to be white, and the maximum to be blue. Any epigenomes that

had no expression data were plotted as white rows, for ease of comparison with other heatmaps.

4.2.7 Chromatin state enrichment for X chromosome gene sets

To compare the chromatin state of escape and inactive X chromosome genes, we used the previously computed chromatin state feature values (as described in Section 3.2.3 and Section 3.2.5) for each X chromosome gene. Then, to compare the chromatin state profiles of the active and inactive variables in female samples, we used the one-sided Mann-Whitney-Wilcoxon test. Specifically, for each chromatin state, we averaged the chromatin state coverage across the 38 female samples for each gene. Then, we compared that average chromatin state coverage for all active genes to the average chromatin state coverage for all inactive genes. We performed both one-sided tests, to test for enrichment in active genes, as well as enrichment in inactive genes.

Next, we performed simulations to account for any chromatin state biases, such as the fact that the active and inactive genes are all from the X chromosome. Specifically, we generated 10,000 randomized simulations where we randomly shuffled the "active" or "inactive" labels on the combined set of 485 genes, while retaining the sizes of each gene set. For each of these simulated "active" and "inactive" gene sets, we calculated both Mann-Whitney-Wilcoxon p-values as described above. Then, we calculated a permutation "p-value" based on these 10,000 random simulations. Specifically, we calculated the percentile ranking of the p-value for our real data, compared to the simulated p-values. Formally:

$$p_{perm} = \frac{k+1}{N+1}$$

where k is number of simulations (out of 10,000) where $p_{sim} \leq p_{real}$, and N is the number of simulations (10,000).

Finally, we used Bonferroni multiple hypothesis correction to correct for our 30 tests, one for each of 15 chromatin states, and both possible test directions. In other words, we considered a $p_{perm} < .00166$ to be significant, as this new threshold was based on a cutoff $\frac{.05}{30}$.

We visualized these resulting p_perm values in a horizontal bar plot, with red dashed lines for the significance cutoffs on a $\log_{10} p_{perm}$ scale. We combined both possible bias directions onto a "back-to-back" bar chart to visualize which chromatin states were biased in which direction.[149]

4.2.8 Sex-based chromatin state enrichment

To compare the chromatin state profiles of the female and male epigenomes, we used the one-sided Mann-Whitney-Wilcoxon test. Specifically, for each chromatin state, we averaged the chromatin state coverage across the 38 female samples for each gene to calculate the "female average". Then, we averaged the chromatin state coverage across the 51 male epigenomes for each gene to calculate the "male average". Then, we compared the female average to the male average for each of four gene groups of interest: active, inactive, variable, and one combined group of all the active, inactive, or variable genes. We performed both one-sided tests, to test for enrichment in female samples, as well as enrichment in male samples.

Next, we performed simulations to account for any chromatin state biases in our epigenome samples. Specifically, we generated 10,000 randomized simulations where we randomly shuffled the "female" or "male" labels on the combined set of 89 epigenomes, while retaining the sizes of each epigenome set. For each of these simulated "female" and "male" epigenomic groups, we again calculated Mann-Whitney-Wilcoxon p-values as described above.

Then, we calculated a permutation "p-value" with Bonferroni multiple hypothesis correction and visualized the results in a back-to-back bar chart, as described above in Section 4.2.7.

4.2.9 Violin plots for chromatin state coverage

To visualize the chromatin state differences between gene groups or sample sex, we plotted the respective average chromatin state coverages in an overlaid boxplot and violin plot. For visualization purposes, we first converted the chromatin state coverage

values into z-scores: specifically, we calculated the overall mean and standard deviation across all relevant data points, and then normalized each value by subtracting the mean and dividing by the standard deviation.

For the individual plots, we set whiskers for the boxplots at 1.5 times beyond the outer quartiles from the median. Plot limits were set to twice the value of the most extreme whisker value in both the positive and negative values, to center the plot to 0.[149] For the combined plots that show all fifteen chromatin state plots together, we set whiskers for the boxplots at 1.5 times beyond the outer quartiles from the median. Plot limits were set to the value of the most extreme whisker of the individual plots.[149]

4.3 Results

4.3.1 Overview

Using the approaches described in Section 4.2, we applied ChromDiff to epigenomic samples from Epigenome Roadmap**??** based on sex, tissue type, sample state, and developmental age to find relevant genes, switching chromatin states, enriched gene sets, and differentially expressed genes. Based on our application of ChromDiff to female and male samples, we followed up on the epigenomic differences between escape and inactive genes on the inactive X chromosome in female cells. Finally, we leveraged both proximal and distal enhancers, promoters, and DNase Hypersensitive sites with Regulatory ChromDiff to find epigenomic differences based on sex, tissue type, and sample state. We found many new genes that were missed by only looking at the gene body, verifying the importance of using regulatory regions for systematic identification of differences across the genome.

4.3.2 Epigenetic sex differences consistent with X Chromosome inactivation

In our first comparison, we sought epigenomic differences between male and female samples. We found 536 significant epigenomic features (gene-chromatin state com-

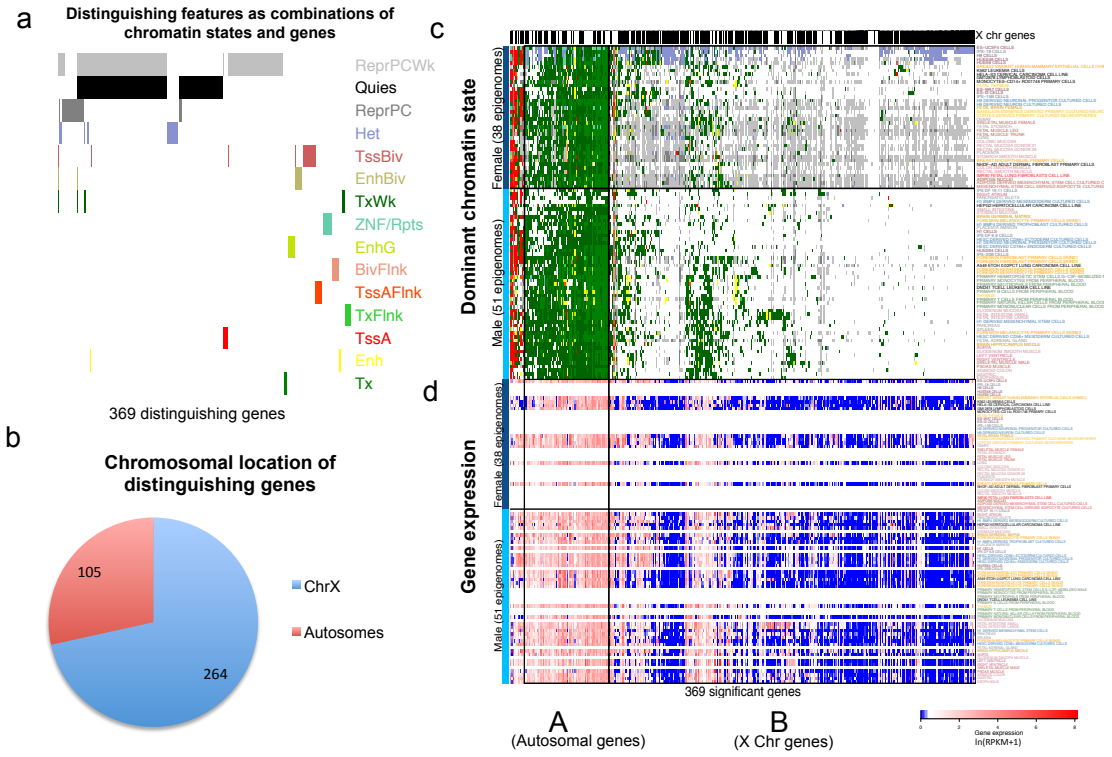

Figure 4-1: X chromsome inactivation distinguishes male and female samples. Comparison of male and female epigenomes identifies a. 536 features that are associated with 369 genes and all 15 chromatin states, where b. 264 of the 369 genes are located on the X chromosome. c. 124 of the identified X chromsome genes are mainly quiescent in male samples but weakly repressed or heterochromatic in female cell types (mostly in cluster B), while 56 genes are transcribed in female and male samples (mostly autosomal genes in cluster A), shown here by the most abundant chromatin state for these genes. d. Expression data for these genes (when available) confirms similar expression levels between male and female samples, as suggested by the chromatin state annotations.

binations) distinguishing male from female samples (that we will call 'distinguishing features'), corresponding to 369 genes (that we will refer to as 'distinguishing genes'), and encompassing all 15 chromatin states (Figure 4-1a). Most distinguishing genes are only associated with one feature (only a single chromatin state is significantly different), with the exception of 133 genes that exhibit significant differences in multiple chromatin states, mostly quiescent and weak Polycomb repression (114 of 133 genes) (Figure 4-1a).

Remarkably, over 70% of the distinguishing genes are located on the X chromosome (264 of the 369 genes) (Figure 4-1b). Many of these chromosome X genes (124 of

264 genes) are primarily quiescent in male samples and primarily heterochromatic or Polycomb-repressed in female samples, as exemplified by many of the genes in cluster B in Figure 4-1c, which visualizes the most abundant chromatin state for each distinguishing gene. For these 124 genes, the X chromosome location and epigenomic signature of Polycomb and heterochromatic repression in females is consistent with known mechanisms of X inactivation.[150] In addition, we see another epigenomic signature (exemplified by gene cluster A in Figure 4-1c) at the 56 genes that are mostly transcribed in both females and males (31 of these 56 genes are autosomal). Figure 4-2 shows that many of these transcribed and autosomal genes are associated with changes in bivalent (TssBiv, EnhBiv, BivFlnk), enhancer (EnhG, Enh), and transcribed (TxFlnk, TxWk) regions. Overall, gene expression is largely unchanged between the female and male epigenomes at the distinguishing genes, despite the epigenomic differences (Figure 4-1d), with only 2 out of the 368 distinguishing genes with expression data exhibiting significantly different expression levels. Again, this is consistent with X inactivation due to the allelic imbalance of X chromosomes for female and male donors.

4.3.3 Active chromatin states are enriched on genes that escape ChrX inactivation, while repressive chromatin states are enriched on inactive genes.

In collaboration with Taru Tukiainen in the Macarthur Lab, we further investigated the chromatin state patterns of both escape and inactivated genes on the X chromosome. Previous work has shown that, even on the copy of the X chromosome that is inactivated in a female cells, some X chromosome genes "escape" the inactivation, and are still, in fact, expressed.[151] To identify the "escape" genes on the X chromosome that manage to be expressed, even when inactivated, Taru performed single-cell experiments that measured gene expression. With these results, Taru categorized each gene as either "escape", "inactive", or "variable", depending on how consistently they escaped X chromsome inactivation (if ever). After mapping these genes

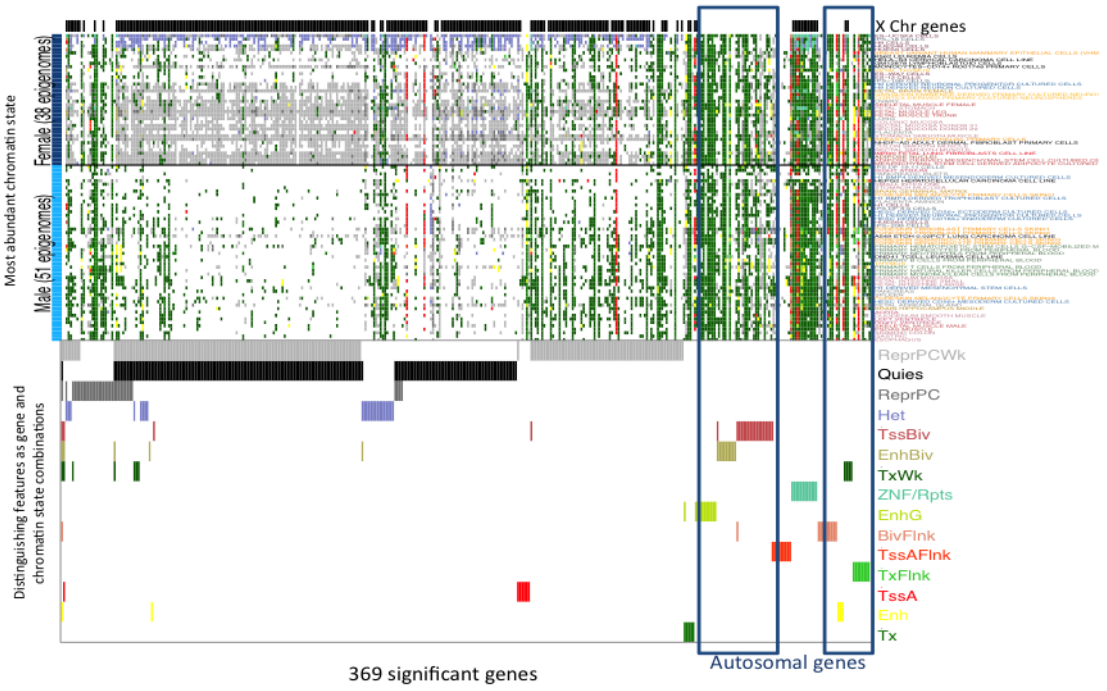

Figure 4-2: Distinguishing autosomal genes are associated with changes in bivalent and enhancer regions. While X chromosome genes are largely associated with changes in quiescent and polycomb repressed regions, transcribed autosomal genes (highlighted in blue) are largely associated with bivalent and enhancer regions, as well as flanking transcribed (TxFlnk) regions.

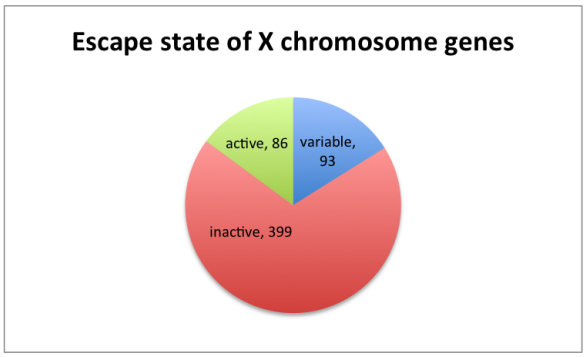

Figure 4-3: We found 86 escape genes, 399 inactive genes, and 93 variable genes on the X chromsome based on experimental expression after mapping to GENCODE v10 protein-coding genes.

to GENCODE v10 for chromatin state information, we had 99 active, 432 inactive, and 101 variable X chromosome genes, as shown in Figure 4-3.

Using these gene groups, we compared the distribution of Z-scores of coverage for each chromatin state between escape and inactive genes, as shown in Figure 4-4b. Using statistics and simulations as described above in Section 4.2.7, we specifically identified significant differences in chromatin state patterns between the two gene groups (4-4a).

While comparing the chromatin state patterns on escape genes and inactive genes, we find significant enrichment for active marks being biased for escape genes, as well as repressive marks being significantly biased toward inactive genes, as shown in Figure 4-4a. Specifically, we found a significant enrichment active flanking promoter (TssAFlnk), transcribed (Tx), weakly transcribed (TxWk), and enhancer (Enh) in escape genes compared to inactive genes. On the other hand, we found a significant enrichment for the repressive chromatin states of heterchromatic (Het), bivalent promoter (TssBiv), and polycomb repressed (ReprPC) regions in inactive genes compared to escape genes.

The differences in chromatin state coverage between escape, inactive, and variable ChrX genes can be more explicitly visualized with the distributions of Z-scores for chromatin state coverage in escape, inactive, and variable genes. Figure 4-4b shows a general trend of more active chromatin state presence at escape genes, and more

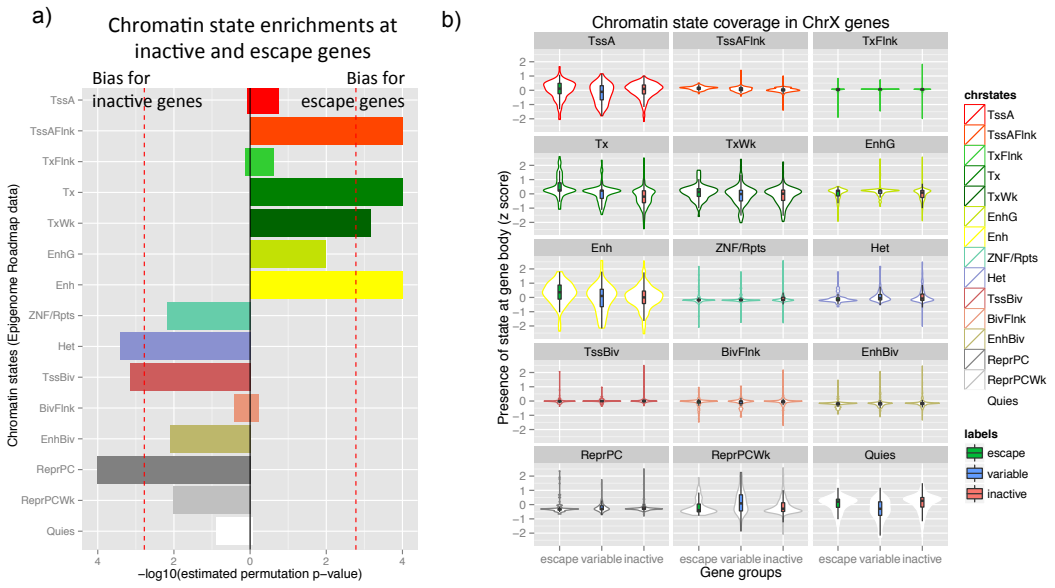

Figure 4-4: Comparison of chromatin state coverage at escape genes and inactive genes reveals distinctive chromatin state biases. Escape genes are significantly enriched for active chromatin states such as flanking active flanking promoter (TssAFlnk), transcribed (Tx), weakly transcribed (TxWk), and enhancer (Enh) regions. Conversely, inactive genes are significantly enriched for repressive chromatin states such as heterochromatic (Het), bivalent promoter (TssBiv), and polycomb repressive (ReprPC) regions. Permutation p-values are based on using the Mann-Whitney-Wilcoxon test to compare chromatin state coverage of escape genes and inactive genes, compared to a null distribution based on 10,000 shuffled simulations. Significance cutoff is based on Bonferroni correction ($p < .00166$). b) Average chromatin state coverage at escape, inactive, and variable genes illustrates more presence of active states at escape genes and more presence of repressive states at inactive genes. Average chromatin state coverage has been calculated for each gene across female samples and converted into Z-scores across all genes. Box-and-whiskers plots are overlaid by violin plots for the distribution of Z-scores in each gene group.

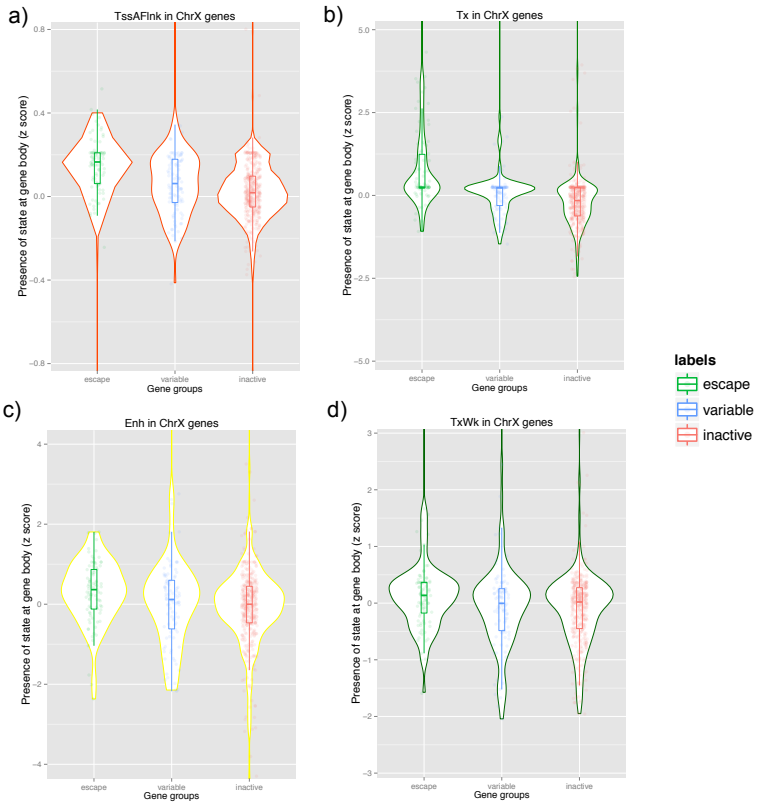

Figure 4-5: Active flanking promoter, transcribed, weakly transcribed, and enhancer chromatin states enriched in escape genes compared to inactive genes. Presence of a) active flanking promoter (TssAFlnk), b) transcribed (Tx), c) weakly transcribed (TxWk), and d) enhancer (Enh) regions is significantly higher in escape genes than in inactive genes, as described in Figure 4-4a ($p < .00166$). Meanwhile, variable genes often have intermediate levels of chromatin state coverage that lie between the coverage of escape and inactive genes.

repressive chromatin state in inactive genes. The differences that were found to be significant, as shown in Figure 4-4a, are more closely presented in Figure 4-5 and Figure 4-6, for states biased towards escape genes and inactive genes, respectively.

These results not only confirms that epigenomic mechanisms behind the heterochromatic inactivation of one X chromosome in each female cell, but also suggests an epigenomic mechanism for escape from X inactivation through the substitution of active histone marks for repressive marks.

Further, we compare the chromatin state patterns at the X chromosome genes (including escape, active, and variable genes) between male and female samples, and

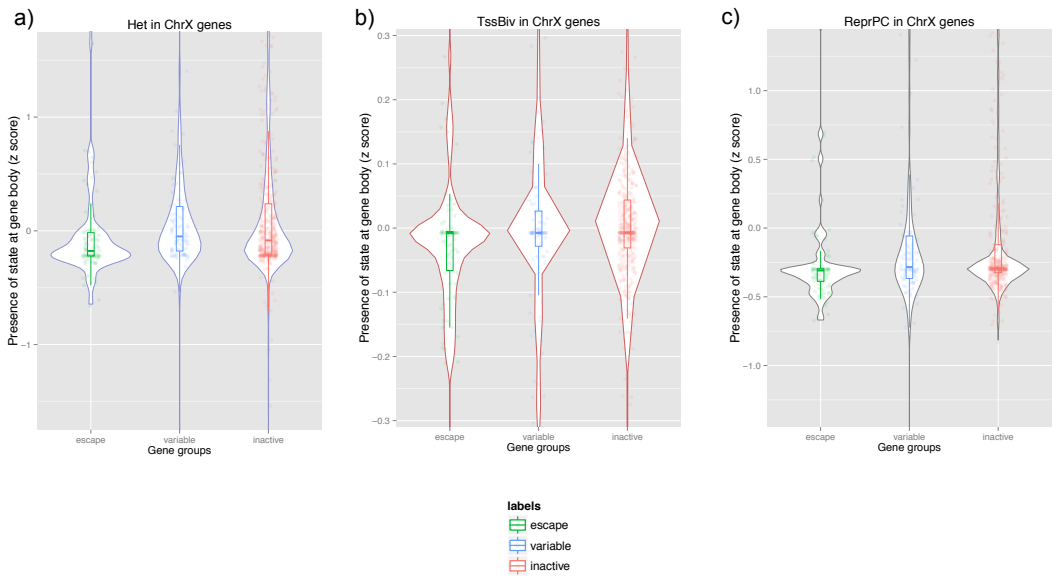

Figure 4-6: Heterochromatic, bivalent promoter, and polycomb repressed chromatin states are enriched in inactive genes compared to escape genes. Presence of the repressive chromatin states of a) heterochromatic (Het), b) bivalent promoter (TssBiv), and c) polycomb repressed (ReprPC) regions is significantly higher in inactive genes than in escape genes, as calculated in Figure 4-4a ($p < .00166$). Meanwhile, variable genes often have intermediate levels of chromatin state coverage that lie between the coverage of escape and inactive genes.

as we previously found in the genome-wide comparison of female and male samples, active marks were biased for male samples while repressive marks are biased for female samples. This, again, is in line with X chromosome inactivation, which must inactivate one X chromosome of each female cell, while the only X chromosome of male cells is completely active.[151]

4.3.4 Comparison of brain and gastrointestinal tissues reveal epigenomic changes in neuronal genes

In addition to sex-based differences, we identified tissue-specific epigenomic differences by comparing brain cells and tissues against gastrointestinal tissues, two of the anatomical groups for which we had the most epigenomic data. We found 10,455 distinguishing features, corresponding to 5,533 distinguishing genes. For visualization purposes, we have sampled down in this and future examples to 10,000 distinguishing features and their associated genes (Section 4.2.3).

Over 40% (2,274 of 5,533 genes) of the genes distinguishing brain from gastrointestinal tissues involve multiple chromatin states for each gene. Of the 5,079 genes associated with the 10,000 sampled features, six groups of genes emerge, representing genes with distinguishing features involving: (a) promoter and enhancer regions, (b) weakly transcribed and quiescent regions; (c) enhancer and weakly transcribed regions, (d) enhancer regions only, (e) polycomb repressed and active TSS regions, and (f) genic enhancer regions (Figure 4-8a, left to right; Figure 4-9). These results highlight the powerful ability of ChromDiff to identify relationships between chromatin states: these gene groups suggest combinations of chromatin states that act in coordinated ways to complement and/or reinforce one another.

In contrast to the sex-based comparison, the comparison of brain and gastrointestinal tissues identifies many genes with significant expression differences. Specifically, in 18% of the discriminative genes (1,043/5,533), the most abundant chromatin state switched between mainly transcribed in one group (Tx, TxWk, or TxFlnk) to primarily Polycomb-repressed or quiescent in the other group (ReprPC, ReprPCWk,

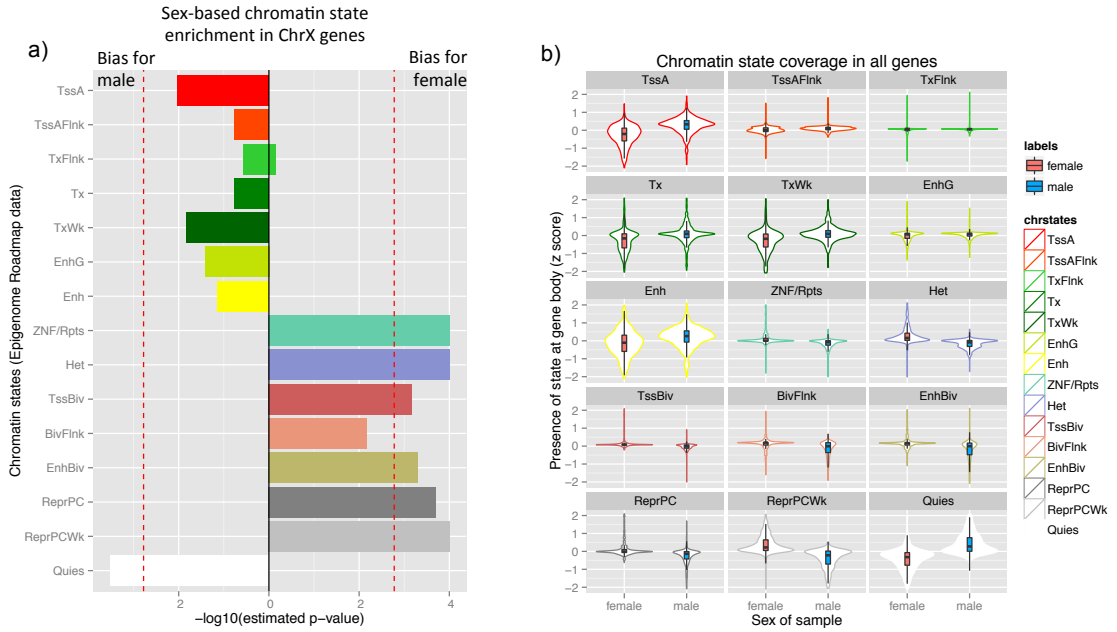

Figure 4-7: Comparison of chromatin state coverage at ChrX genes in female and male samples reveals chromatin state patterns consistent with X chromsome inactivation. a) Genes on the X chromosome are significantly enriched for repressive chromatin states such as zinc finger and repetitive (ZNF/Rpts), heterochromatic (Het), bivalent promoter (TssBiv), bivalent enhancer (EnhBiv), polycomb repressed (ReprPC), and weakly polycomb repressed (ReprPCWk) regions in female samples. On the other hand, when compared to female samples, male samples are enriched for quiescent (Quies) regions characterized by a lack of histone mark presence. This suggests that the broad chromosome-wide repression of X inactivation in female samples results in a "repressive" chromatin state at these regions, even though they were already quiescent in male samples. Permutation p-values are based on using the Mann-Whitney-Wilcoxon test to compare chromatin state coverage of X chromosome genes in female and male samples, compared to a null distribution based on 10,000 shuffled simulations. Significance cutoff is based on Bonferroni correction ($p < .00166$). b) Average chromatin state coverage of ChrX genes in female and male samples illustrates significantly higher presence of repressive states in female samples and significantly more quiescent regions in male samples ($p < .00166$). Higher levels of active states in male samples can also be observed, although they do not meet our significance cutoff. Average chromatin state coverage was separately calculated for female and male samples for each gene; these values were then converted into Z-scores across the values for both sexes. Box-and-whiskers plots are overlaid with violin plots for the distribution of Z-scores in each gene group.

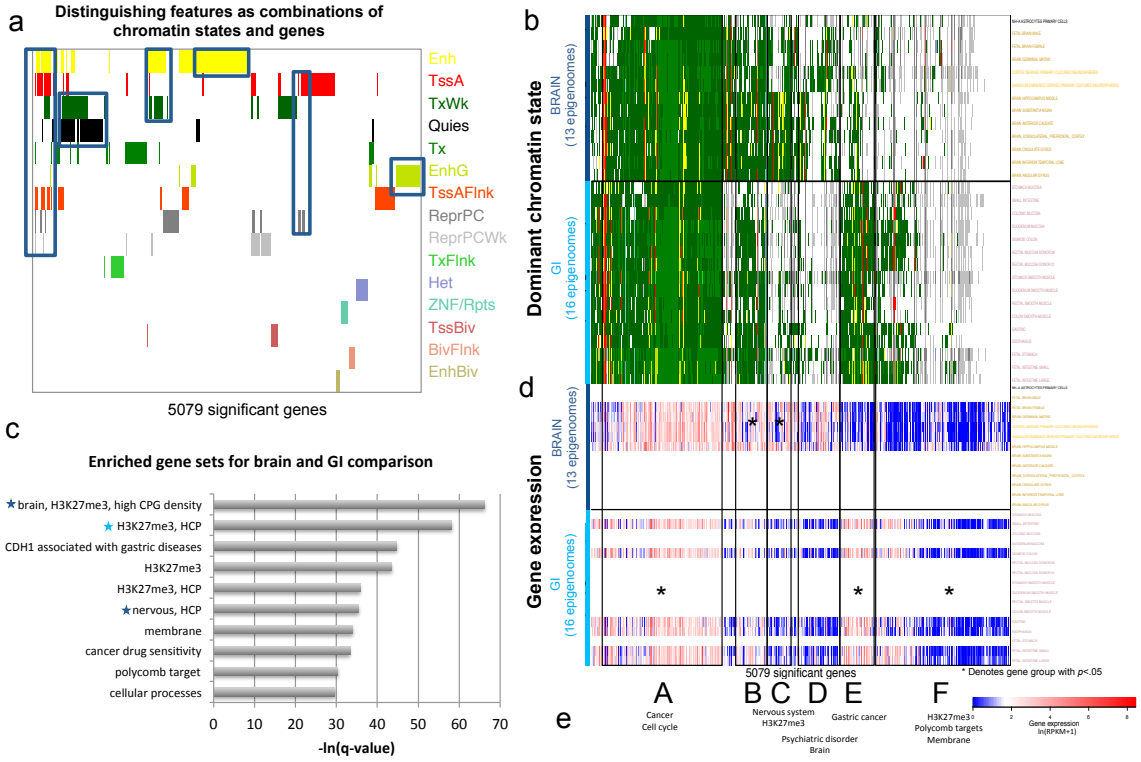

Figure 4-8: Transcriptional differences dominate brain and GI tissue comparison. Comparison of brain and gastrointestinal epigenomes reveal a. chromatin state changes that co-occur within groups of chromatin states, as well as cluster-specific transcriptional differences at associated genes based on b. most abundant chromatin state and d. gene expression data (when available). 5 of the 6 identified gene groups have significantly different expression between brain and GI samples, with asterisks indicating $p<.05$ based on the two-sided Mann-Whitney test. c. Identified genes are enriched for brain (dark blue stars) and gastric (light blue stars) specific purposes and gene sets, as well as other gene sets (black), as evidenced by the top ten gene set annotations. e. Genes in each epigenomic cluster contain different gene set annotations, such as cancer-related and cell cycle gene sets (cluster A), gastric-specific (cluster E) and brain-specific (cluster D) gene sets, genes related to the nervous system (cluster C), genes associated with histone marks (clusters C and F), and membrane genes (cluster F).

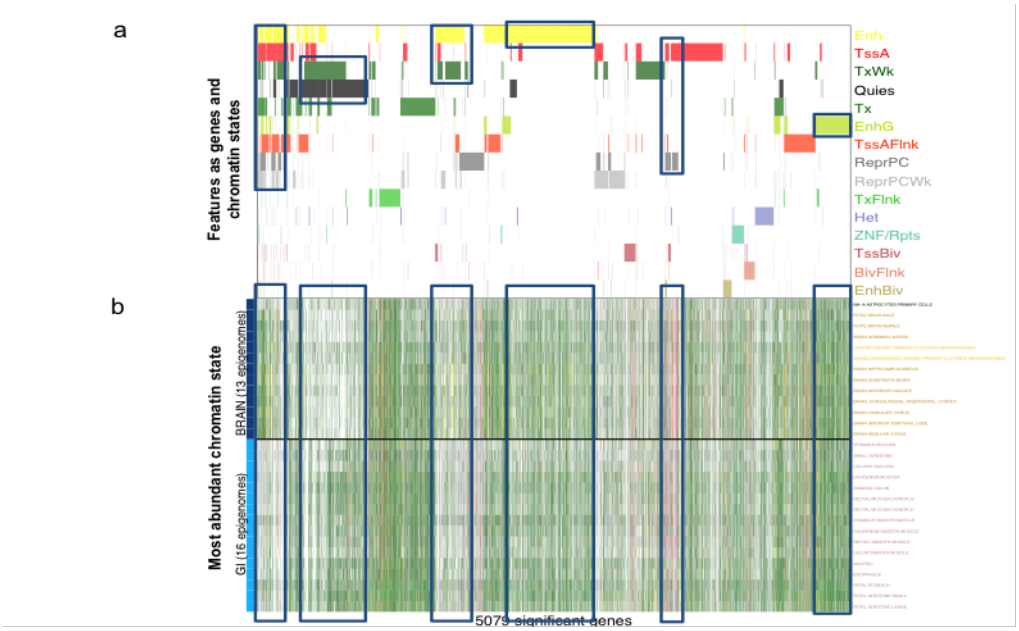

Figure 4-9: Brain and gastrointestinal differences reveal changing chromatin state differences in gene clusters. a. Various chromatin states exhibit coordinated changes at corresponding gene clusters. Specifically, from left to right, the groups of chromatin states (highlighted in blue) that change at the same genes are a) enhancer and promoter regions (Enh/TssA/TssAFlnk), b) transcribed and quiescent (TxWk/Quies), c) enhancer and transcribed (Enh/TxWk), d) enhancer (Enh), e) promoter and repressed (TssA,/ReprPC), and f) genic enhancer (EnhG) regions. b. For groups a and c, gastrointestinal (GI) tissues are mostly quiescent, while group b genes are mostly transcribed in GI tissues. Group c and d genes are more often enhancer regions in brain samples, while group e genes are annotated as promoter states in both brain and GI samples.

or Quies), as exemplified by gene clusters C, D, E, and F (Figure 4-8b). The majority of these switching genes (675/1,043) showed significant expression differences from RNA-Seq data (Figure 4-8d). Overall, 40% of all distinguishing genes showed significantly different expression (2236/5507).

For many of the remaining genes, including those in gene cluster A (Figure 4-8b), epigenomic differences did not involve the most abundant chromatin state. For example, in both brain and gastrointestinal epigenomes, 86% of genes in cluster A are annotated as primarily transcribed or enhancer states (1253/1452 genes), but the majority of cluster A genes were identified based on features that did not involve transcription or enhancer states (1177/1452).

Furthermore, our gene clustering based on epigenomic signal also revealed that different gene clusters mapped to varying gene set functions, ranging from brain-specific genes (clusters C and D), genes related to gastric cancer (cluster E), cell cycle genes (cluster A), Polycomb targets and genes marked by H3K27me3 (cluster F), and genes associated with cancer (clusters A, B, and E) (Figure 4-8e, Tables 4.1 to 4.6). The entire set of 5,533 distinguishing genes is enriched for genes known to be important for brain and gastrointestinal function, including genes with brain-specific histone modifications and targets of CDH1, which has recently been shown to be associated with gastric cancer[152,153] (Figure 4-8c, Table 4.7).

4.3.5 Blood samples distinguished by enhancer activity differences

With the resources of various blood epigenomes, we compared chromatin states at gene bodies for the liquid samples (blood) against the solid samples (tissues and other primary cells). ChromDiff found 45,513 significant differentiating features associated with 17,001 genes. The 10,000 sampled features and their associated 1,721 genes are largely dominated by transcription, enhancer, quiescent, and repression states (Figure 4-10a), and about 40% of the genes show expression differences (717/1717 of sampled genes, 6827/16827 distinguishing genes) (Figure 4-10b).

annotation	geneset	qval
cancer	NUYTTEN_EZH2_TARGETS_UP	7.73E-14
Alzheimer's	BLALOCK_ALZHEIMERS_DISEASE_UP	7.73E-14
parvin family of actin-binding protein (cytoskeleton, cell adhesion)	JOHNSTONE_PARVB_TARGETS_3_UP	1.11E-13
aggressive tumors	ONKEN_UVEAL_MELANOMA_UP	4.54E-12
apoptosis	GRAESSMANN_APOPTOSIS_BY_DOXORUBICIN_UP	4.54E-12
fibronectin response	MILI_PSEUDOPODIA_HAPTOTAXIS_DN	1.66E-11
tumor suppressor genes	LOPEZ_MBD_TARGETS	1.94E-11
TGF-beta signaling	KOINUMA_TARGETS_OF_SMAD2_OR_SMAD3	1.16E-10
cancer	REN_ALVEOLAR_RHABDOMYOSARCOMA_DN	1.50E-10

Table 4.1: Enriched gene sets for cluster A of brain and gastrointestinal comparison. Comparison of brain and gastrointestinal samples show that distinguishing genes in cluster A are related to Alzheimer's disease. (10 most strongly enriched gene sets shown.)

annotation	geneset	qval
cancer	RODRIGUES_THYROID_CARCINOMA_POORLY_DIFFERENTIATED_DN	1.54E-02

Table 4.2: Enriched gene sets for cluster B of brain and gastrointestinal comparison. Comparison of brain and gastrointestinal samples show that distinguishing genes in cluster B are related to thyroid carcinoma. (Only one significantly enriched gene set found.)

annotation	geneset	qval
HCP, H3K27me3	MIKKELSEN_MCV6_HCP_WITH_H3K27ME3	5.65E-08
nervous system	GOBERT_OLIGODENDROCYTE_DIFFERENTIATION_DN	4.36E-04
nervous system	NERVOUS_SYSTEM_DEVELOPMENT	4.80E-04
nervous system	LEIN_NEURON_MARKERS	1.03E-03
H3K27me3	BENPORATH_ES_WITH_H3K27ME3	1.15E-03
HCP, H3K27me3	MIKKELSEN_MEF_HCP_WITH_H3K27ME3	1.90E-03
nervous, HCP	MEISSNER_NPC_HCP_WITH_H3K4ME2	2.48E-03
cancer	SMID_BREAST_CANCER_BASAL_UP	2.48E-03
nervous	MARTORIATI_MDM4_TARGETS_NEUROEPITHELIUM_DN	3.50E-03
nervous	LEE_NEURAL_CREST_STEM_CELL_DN	5.35E-03

Table 4.3: Enriched gene sets for cluster C of brain and gastrointestinal comparison. Comparison of brain and gastrointestinal samples show that distinguishing genes in cluster C are related to the nervous system and H3K27me3 modifications. (10 most strongly enriched gene sets shown.)

annotation	geneset	qval
nervous	LEIN_OLIGODENDROCYTE_MARKERS	1.49E-03
psychiatric disorder	ASTON_MAJOR_DEPRESSIVE_DISORDER_DN	1.49E-03
brain	LU_AGING_BRAIN_UP	1.49E-03
HCP	MIKKELSEN_MEF_HCP_WITH_H3K27ME3	1.49E-03
cancer	BRUINS_UVC_RESPONSE_VIA_TP53_GROUP_A	1.49E-03
cell surface interactions	PID_INTEGRIN1_PATHWAY	4.93E-03
knockdown of proto-oncogene	YANG_BCL3_TARGETS_UP	4.93E-03
nervous	NERVOUS_SYSTEM_DEVELOPMENT	4.93E-03
brain, HCP	MEISSNER_BRAIN_HCP_WITH_H3K4ME3_AND_H3K27ME3	4.93E-03
nervous, HCP	MEISSNER_NPC_HCP_WITH_H3K4ME2	5.68E-03

Table 4.4: Enriched gene sets for cluster D of brain and gastrointestinal comparison. Comparison of brain and gastrointestinal samples show that distinguishing genes in cluster D are related to psychiatric disorders, brain function, and the nervous system. (10 most strongly enriched gene sets shown.)

annotation	geneset	qval
gastric cancer	ONDER_CDH1_TARGETS_2_DN	2.04E-20
cancer drug sensitivity	COLDREN_GEFITINIB_RESISTANCE_DN	8.28E-20
cancer	CHARAFE_BREAST_CANCER_LUMINAL_VS_MESENCHYMAL_UP	2.70E-14
brain	MEISSNER_BRAIN_HCP_WITH_H3K4ME3_AND_H3K27ME3	3.02E-11
mammary stem cell	LIM_MAMMARY_STEM_CELL_DN	7.31E-09
cancer	CHARAFE_BREAST_CANCER_BASAL_VS_MESENCHYMAL_UP	5.28E-07
cancer	LIU_PROSTATE_CANCER_DN	5.28E-07
breast	MCBRYAN_PUBERTAL_BREAST_3_4WK_UP	6.31E-07
cancer	DELYS_THYROID_CANCER_UP	7.87E-07
cancer	DODD_NASOPHARYNGEAL_CARCINOMA_UP	8.18E-07

Table 4.5: Enriched gene sets for cluster E of brain and gastrointestinal comparison. Comparison of brain and gastrointestinal samples show that distinguishing genes in cluster E are related to brain function and a variety of cancer types. (10 most strongly enriched gene sets shown.)

annotation	geneset	qval
H3K27me3	MIKKELSEN_MEF_HCP_WITH_H3K27ME3	8.83E-52
H3K27me3	BENPORATH_ES_WITH_H3K27ME3	1.58E-35
polycomb targets	BENPORATH_SUZ12_TARGETS	1.35E-28
H3K27me3	MIKKELSEN_MCV6_HCP_WITH_H3K27ME3	1.02E-26
polycomb targets	BENPORATH_EED_TARGETS	4.52E-26
nervous, H3K27me3	MEISSNER_NPC_HCP_WITH_H3K4ME2_AND_H3K27ME3	3.90E-22
membrane	PLASMA_MEMBRANE	7.82E-20
nervous, H3K27me3	MEISSNER_BRAIN_HCP_WITH_H3K4ME3_AND_H3K27ME3	1.34E-19
nervous	REACTOME_NEURONAL_SYSTEM	3.37E-19
membrane	PLASMA_MEMBRANE_PART	2.24E-18

Table 4.6: Enriched gene sets for cluster F of brain and gastrointestinal comparison. Comparison of brain and gastrointestinal samples show that distinguishing genes in cluster F are related to polycomb targets, psychiatric disorders, brain function, and the nervous system. (10 most strongly enriched gene sets shown.)

annotation	geneset	qval
brain, H3K27me3, high CPG density promoters (HCP)	MEISSNER_BRAIN_HCP_WITH_H3K4ME3_AND_H3K27ME3	1.97E-29
H3K27me3, HCP	MIKKELSEN_MEF_HCP_WITH_H3K27ME3	5.19E-26
CDH1 associated with gastric diseases	ONDER_CDH1_TARGETS_2_DN	3.64E-20
H3K27me3	BENPORATH_ES_WITH_H3K27ME3	1.34E-19
H3K27me3, HCP	MIKKELSEN_MCV6_HCP_WITH_H3K27ME3	2.70E-16
nervous, HCP	MEISSNER_NPC_HCP_WITH_H3K4ME2	3.79E-16
membrane	PLASMA_MEMBRANE	1.69E-15
cancer drug sensitivity	COLDREN_GEFITINIB_RESISTANCE_DN	3.38E-15
polycomb target	BENPORATH_SUZ12_TARGETS	6.72E-14
cellular processes	PEREZ_TP53_TARGETS	1.43E-13

Table 4.7: Enriched gene sets for brain and gastrointestinal comparison. Comparison of brain and gastrointestinal samples show that the total set of distinguishing genes are related to brain and nervous system function, as well as gastric diseases. (10 most strongly enriched gene sets shown.)

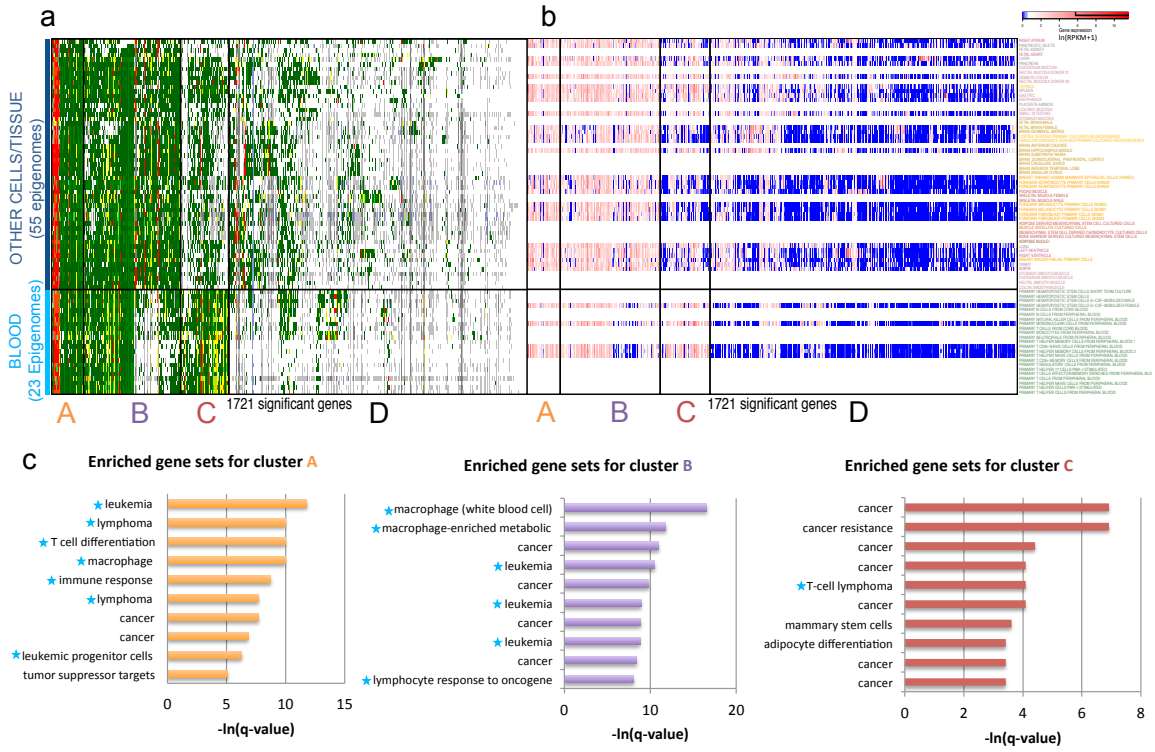

Figure 4-10: Epigenomic differences specific to blood samples lie at blood cancer genes. Comparison of blood epigenomes with other primary cells and tissues reveals distinguishing epigenomic activity, at genes marked by a. transcriptional, enhancer, repressed, and quiescent chromatin states, as shown by the most abundant chromatin state heatmap. The corresponding transcriptional activity can be seen in b. gene expression profiles for the associated genes (with the genes in the same ordering in 4a and 4b). c. Genes in clusters A and B are enriched for blood-specific gene sets (light blue stars), such as gene sets relating to immune response, lymphoma, leukemia, and macrophage activity. On the other hand, cluster C is enriched for genes relating to cancer, while cluster D is generally composed of membrane genes.

annotation	geneset	qval
leukemia (blood cancer)	MARTENS_BOUND_BY_PML_RARA_FUSION	7.62E-06
lymphoma (blood cancer)	YU_MYC_TARGETS_DN	4.70E-05
T cell differentiation	LEE_DIFFERENTIATING_T_LYMPHOCYTE	4.70E-05
macrophage (white blood cell)	CHEN_METABOLIC_SYNDROM_NETWORK	4.70E-05
immune response (hepatitis B viral clearance)	WIELAND_UP_BY_HBV_INFECTION	1.65E-04
lymphoma (blood cancer)	PASQUALUCCI_LYMPHOMA_BY_GC_STAGE_DN	4.65E-04
cancer	LINDGREN_BLADDER_CANCER_CLUSTER_2B	4.65E-04
cancer	WALLACE_PROSTATE_CANCER_RACE_UP	1.09E-03
leukemic progenitor cells (blood cancer)	TORCHIA_TARGETS_OF_EWSR1_FLI1_FUSION_DN	1.92E-03
tumor suppressor targets	SANSOM_APC_TARGETS_DN	6.01E-03

Table 4.8: Enriched gene sets for cluster A from comparison of blood and non-blood samples. Comparison of blood and non-blood samples show that the total set of distinguishing genes in cluster A are related to leukemia, lymphoma, and immune response. (10 most strongly enriched gene sets shown.)

We find four main clusters of genes (Figure 4-10a,b) that correspond to different gene set enrichments. Gene cluster A is characterized by strong enrichment for gene sets relating to immune response, T cell differentiation, and blood cancers, while cluster B is enriched for genes related to macrophage function and leukemia. Cluster C is enriched for genes relating to general cancer development, and cluster D is enriched for membrane genes, likely due to blood cell-specific membrane function[154,155] (Figure 4-10c, Tables 4.8 to 4.11).

4.3.6 Comparison of samples based on developmental ages link to cancer genes

For our final comparison, we investigated differences in adult and fetal samples based on donor metadata. All samples that were listed from a pre-birth donor were labeled as Fetal samples; all samples labeled Adult samples either came exclusively from adult donors (over 18 years old), or came partially from adult donors with no age information for other donors.

We found 7,472 significant epigenomic features distinguishing Adult and Fetal samples, spanning 5,852 unique genes. Visualization of the most abundant chromatin

annotation	geneset	qval
macrophage (white blood cell)	FOSTER_TOLERANT_MACROPHAGE_DN	6.50E-08
macrophage-enriched metabolic network	CHEN_METABOLIC_SYNDROM_NETWORK	7.42E-06
cancer	NUYTTEN_EZH2_TARGETS_UP	1.77E-05
leukemia	KRIGE_RESPONSE_TO_TOSEDOSTAT_24HR_UP	2.89E-05
cancer	CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_DN	5.63E-05
leukemia	MULLIGHAN_MLL_SIGNATURE_2_UP	1.29E-04
cancer	GOZGIT_ESR1_TARGETS_DN	1.43E-04
leukemia	ALCALAY_AML_BY_NPM1_LOCALIZATION_UP	1.51E-04
cancer	KRIEG_HYPOXIA_NOT_VIA_KDM3A	2.33E-04
lymphocyte response to oncogene	DIRMEIER_LMP1_RESPONSE_LATE_UP	3.32E-04

Table 4.9: Enriched gene sets for cluster B from comparison of blood and non-blood samples. Comparison of blood and non-blood samples show that the total set of distinguishing genes in cluster B are related to macrophage function and leukemia. (10 most strongly enriched gene sets shown.)

annotation	geneset	qval
cancer	BERTUCCI_MEDULLARY_VS_DUCTAL_BREAST_CANCER_DN	9.97E-04
cancer resistance	MASSARWEH_TAMOXIFEN_RESISTANCE_UP	9.97E-04
cancer	RODRIGUES_THYROID_CARCINOMA_ANAPLASTIC_DN	1.21E-02
cancer	KIM_WT1_TARGETS_12HR_UP	1.67E-02
T-cell lymphoma	PICCALUGA_ANGIOIMMUNOBLASTIC_LYMPHOMA_UP	1.67E-02
cancer	KRIEG_HYPOXIA_NOT_VIA_KDM3A	1.67E-02
mammary stem cells	LIM_MAMMARY_STEM_CELL_UP	2.70E-02
adipocyte differentiation	TSENG_ADIPOGENIC_POTENTIAL_DN	3.28E-02
cancer	GRAESSMANN_APOPTOSIS_BY_SERUM_DEPRIVATION_DN	3.28E-02
cancer	DODD_NASOPHARYNGEAL_CARCINOMA_UP	3.28E-02

Table 4.10: Enriched gene sets for cluster C from comparison of blood and non-blood samples. Comparison of blood and non-blood samples show that the total set of distinguishing genes in cluster C are related to lymphoma and other cancers. (10 most strongly enriched gene sets shown.)

annotation	geneset	qval
membrane	PLASMA_MEMBRANE	2.27E-14
membrane	INTEGRAL_TO_MEMBRANE	4.40E-14
membrane	INTRINSIC_TO_MEMBRANE	4.40E-14
membrane	PLASMA_MEMBRANE_PART	2.46E-13
membrane	MEMBRANE_PART	1.51E-12
membrane	INTEGRAL_TO_PLASMA_MEMBRANE	1.78E-12
membrane	INTRINSIC_TO_PLASMA_MEMBRANE	1.78E-12
membrane	MEMBRANE	2.11E-12
liver	HSIAO_LIVER_SPECIFIC_GENES	1.11E-09
MAP kinase	YOSHIMURA_MAPK8_TARGETS_UP	8.39E-09

Table 4.11: Enriched gene sets for cluster D from comparison of blood and non-blood comparison. Comparison of blood and non-blood samples reveal membrane function for gene cluster D. (10 most strongly enriched gene sets shown.)

state of each gene in each epigenome (Figure 4-11a) revealed that most significant genes had the same most abundant state in both adult and fetal epigenomes, suggesting more subtle underlying epigenomic changes at these genes. Specifically, although the most abundant chromatin state was usually transcribed or quiescent states, the underlying changing chromatin state spans all fifteen chromatin states (Figure 4-12a). Specifically, many genes that were mostly quiescent or repressed were associated with changes in transcribed, promoter, and genic enhancer states, while genes that were mostly transcribed in all epigenomes exhibited changes in ZNF and quiescent state annotations (Figure 4-12b). Transcription was also similar between adult and fetal samples at most sampled genes (Figure 4-11b), with only 15% of all distinguishing genes differentially expressed (or 887 of 5,798 distinguishing genes with expression data).

Overall gene set enrichments for the 5,852 identified genes resulted in wide-ranging biological pathways, including gene sets related to liver, Polycomb targets, and cytokines (Table 4.12). However, from the visualization of the most abundant chromatin states (Figure 4-11a), we identified two gene subgroups with distinctive epigenomic signatures and cohesive corresponding enrichments (Figure 4-11c, Tables 4.13 and 4.14). Cluster A genes are enriched for genes related to tumors and anticancer

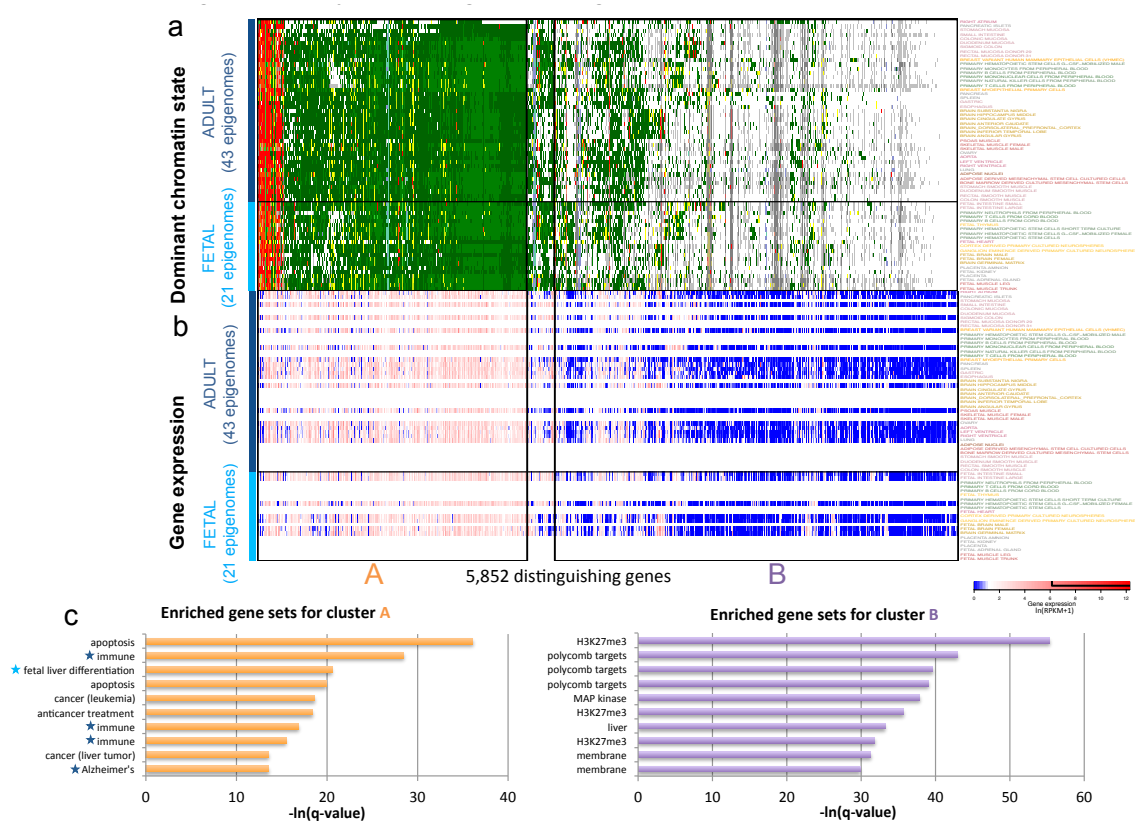

Figure 4-11: Polycomb targets distinguish adult and fetal samples. a. The most abundant state of distinguishing genes for adult and fetal epigenomes is largely unchanged between the two groups, with the most popular states being active promoters, transcriptional, repressed, or quiescent regions. b. Gene expression profiles confirm similar levels of expression between the adult and fetal epigenomes at identified genes. c. Genes in cluster A are enriched for age-related genes (notated by stars) relating to fetal cell differentiation and Alzheimer's disease, while genes in cluster B are enriched for Polycomb targets, which have been shown to exhibit different behavior in fetal and adult cells.

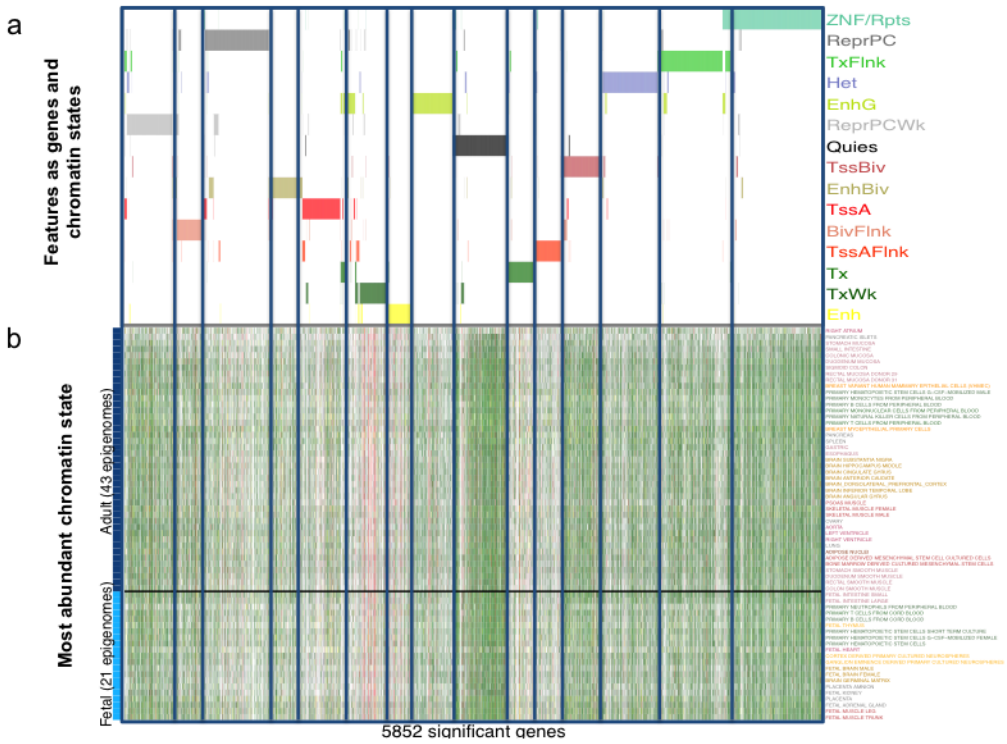

Figure 4-12: Many genes exhibiting changes between adult and fetal samples are only associated with one chromatin state. a. Visualization of features as chromatin states and genes show that most genes are identified by changes due to only one chromatin state, rather than coordinated changes of multiple chromatin states. The most common chromatin states to differ between adult and fetal epigenomes were ZNF/Rpts, ReprPC, Quies, TxFlnk, Het, and ReprPCWk regions. b. Different patterns for the most abundant chromatin state can be seen for genes associated with different chromatin states; for example, genes with changes due to the Tx chromatin state are largely quiescent in both groups, while genes with changes due to ZNF/Rpts are mostly transcribed in both groups.

annotation	geneset	qval
liver	HSIAO_LIVER_SPECIFIC_GENES	1.47E-06
cytokine	KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	1.47E-06
liver	CHIANG_LIVER_CANCER_SUBCLASS_PROLIFERATION_DN	6.07E-06
liver	YAMASHITA_LIVER_CANCER_STEM_CELL_DN	7.18E-05
MAPK8 (proliferation)	YOSHIMURA_MAPK8_TARGETS_UP	7.18E-05
H3K27me3	BENPORATH_ES_WITH_H3K27ME3	1.51E-04
liver	HOSHIDA_LIVER_CANCER_SUBCLASS_S3	2.21E-04
Polycomb targets	BENPORATH_EED_TARGETS	5.81E-04
Polycomb targets	BENPORATH_PRC2_TARGETS	2.83E-03
liver	SU_LIVER	4.88E-03

Table 4.12: Enriched gene sets for adult and fetal comparison. Comparison of adult and fetal samples show that distinguishing genes are related to liver and cytokine function. (10 most strongly enriched gene sets shown.)

treatment response, as well as genes relating to apoptosis; this result is supported by previous work that has shown that tumors have similar expression profiles to early developmental tissues.[156] Furthermore, cluster A is also enriched for genes relating to differentiation of fetal liver cells, as well as genes related to immune response and Alzheimer's disease; this is particularly relevant given the increasingly recognize role of immune processes in Alzheimer's disease[157] and that proteins known to affect fetal development also play a protective role for Alzheimer's disease.[158–160] On the other hand, genes in cluster B are enriched for membrane genes and Polycomb targets, which is relevant given the evidence that polycomb proteins distinguish fetal and adult hematopoietic stem cells.[161,162] Taken together, this validates the ability of ChromDiff to identify relevant gene sets and pathways, despite a lack of change in expression data.

4.3.7 ChromDiff identifies changes at linked enhancers based on tissue type.

We also leveraged proximal and long-range interactions between genes and regulatory regions to identify epigenomic differences between sample groups, as described in

annotation	geneset	qval
apoptosis	GRAESSMANN_APOPTOSIS_BY_DOXORUBICIN_DN	2.12E-16
immune	MARSON_BOUND_BY_FOXP3_UNSTIMULATED	4.46E-13
erythroid differentiation from fetal liver	PILON_KLF1_TARGETS_DN	1.15E-09
apoptosis	GRAESSMANN_APOPTOSIS_BY_DOXORUBICIN_UP	2.29E-09
cancer (leukemia)	DIAZ_CHRONIC_MEYLOGENOUS_LEUKEMIA_UP	8.38E-09
anticancer treatment	BUYTAERT_PHOTODYNAMIC_THERAPY_STRESS_UP	1.07E-08
immune	MARSON_BOUND_BY_FOXP3_STIMULATED	4.54E-08
immune	GALINDO_IMMUNE_RESPONSE_TO_ENTEROTOXIN	1.86E-07
cancer (liver tumor)	ACEVEDO_LIVER_TUMOR_VS_NORMAL_ADJACENT_TISSUE_UP	1.36E-06
Alzheimer's	BLALOCK_ALZHEIMERS_DISEASE_UP	1.36E-06

Table 4.13: Enriched gene sets for cluster A from comparison of adult and fetal samples. Comparison of adult and fetal samples show that distinguishing genes in cluster A are related to fetal liver differentiation, immune response, and Alzheimer's disease. (10 most strongly enriched gene sets shown.)

annotation	Geneset	qval
H3K27me3	BENPORATH_ES_WITH_H3K27ME3	8.94E-25
polycomb targets	BENPORATH_PRC2_TARGETS	2.30E-19
polycomb targets	BENPORATH_SUZ12_TARGETS	6.19E-18
polycomb targets	BENPORATH_EED_TARGETS	1.12E-17
MAP kinase	YOSHIMURA_MAPK8_TARGETS_UP	3.57E-17
H3K27me3, HCP	MEISSNER_BRAIN_HCP_WITH_H3K4ME3_AND_H3K27ME3	3.29E-16
liver	HSIAO_LIVER_SPECIFIC_GENES	3.67E-15
H3K27me3, HCP	MIKKELSEN_MEF_HCP_WITH_H3K27ME3	1.50E-14
membrane	INTRINSIC_TO_PLASMA_MEMBRANE	2.67E-14
membrane	PLASMA_MEMBRANE_PART	1.08E-13

Table 4.14: Enriched gene sets for cluster B from comparison of adult and fetal samples. Comparison of adult and fetal samples show that distinguishing genes in cluster B are related to polycomb targets and H3K27me3 modifications. (10 most strongly enriched gene sets shown.)

Section 3.2.3.2. Specifically, we compared brain and digestive tissues based on the chromatin state of linked enhancer regions. This resulted in the identification of 5590 genes with distinguishing epigenomic features, as shown in Figure 4-13a. These genes were largely identified based on differences in enhancer (yellow) and quiescent (white) states between the two groups. Based on these epigenomic signatures, we identified four main gene clusters: two clusters exhibit tissue-specific enhancer activity (gene cluster A with brain-specific enhancers, and gene cluster B with digestive-specific enhancers), while gene cluster C seems largely to be linked to transcribed regions in both brain and digestive samples and gene cluster D shows enhancer activity in some digestive samples.

Next, we compared the epigenomic state of linked regulatory regions to the transcriptional state of the relevant regions (Figure 4-13b). We found coordinated activity between enhancer and gene expression, suggesting that distinguishing enhancer and regulatory activity could recapitulate distinguishing gene expression. For example, gene cluster A is more expressed in brain samples, gene cluster B is more expressed in digestive samples, and gene cluster D is more expressed in a subset of digestive samples. Although we can not conclude directional causality, our findings suggest a regulatory mechanism of where changes in enhancer activity cause changes in linked gene expression.

Finally, we found relevant gene set enrichments for identified gene clusters as shown in Figure 4-13c. Specifically, our gene cluster that shows brain-specific enhancer and expression activity also shows strong enrichments for genes relating to nervous cell differentiation, nervous system development, and brain tumors.

4.3.8 Studying chromatin state changes at both enhancers and DNase hypersensitive sites identifies ChrX genes.

4 Using our linked regulatory regions, we can again compare our female and male samples and "recover" a signal for X Chromosome inactivation. Specifically, we now use linked DNase Hypersensitive sites to identify genes with chromatin state differences

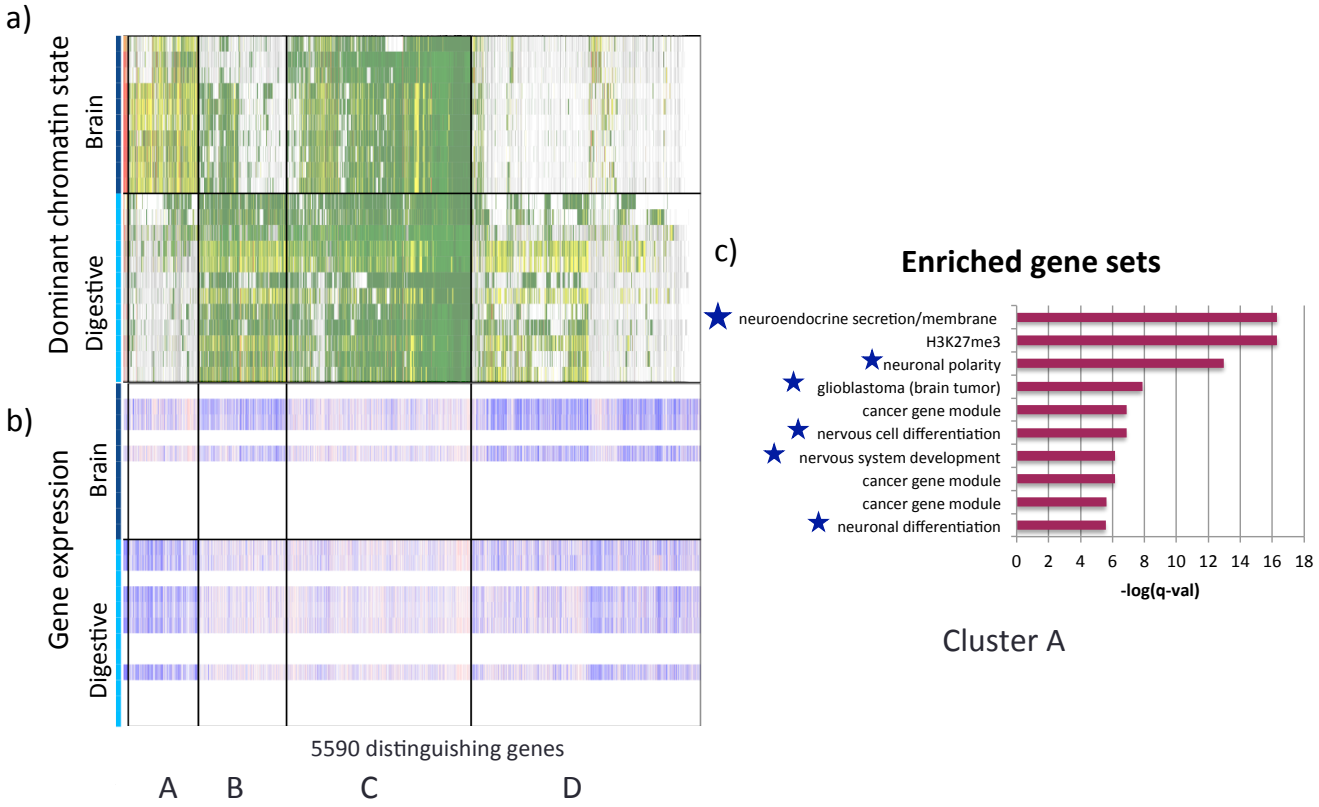

Figure 4-13: Differences at brain and digestive in linked enhancer activity identify genes with differential expression that relate to neuronal development. a) We find tissue-specific enhancer activity, as shown with the dominant chromatin state visualization at linked enhancer genes. Genes cluster into four groups based on their dominant epigenomic signatures, with cluster A showing brain-specific enhancer activity, and clusters B and D showing digestive-specific enhancer activity, while cluster C shows transcriptional and enhancer activity at both tissues in linked regulatory regions. b) Expression at these corresponding genes shows coordinated activity with the enhancer activity, suggesting a potential mechanism for increased enhancer activity leading to increased transcription. c) Gene cluster A, which showed brain-specific enhancer activity and expression, is enriched for relevant brain gene sets, such as neuronal differentiation, glioblastoma, and neuronal polarity.

between male and female samples. We again find the familiar signal of polycomb repression (gray) at X chromosome genes in female samples, while male samples appear quiescent (white) (Figure 4-14a), which is consistent with the known epigenomic repression of one X chromosome in each female cell.

In fact, an overwhelming 93.1% of the distinguishing genes were located on the X chromosome (719/772). Furthermore, by using Regulatory ChromDiff with DNase hypersensitive sites, we identify more X chromosome genes with epigenomic differences than we did when utilizing the gene body ChromDiff approach (Figure 4-14b), suggesting that linked DNase hypersensitive sites may allow increased sensitivity over gene bodies for the sex-based epigenomic comparison.

As before, we find similar levels of gene expresssion between female and male samples at these X chromosome genes, as shown in Figure 4-14c.The fact that we see epigenomic differences but similar gene expression levels is consistent with the underlying different number of X chromosomes in male and females.

As suggested by the dominant chromatin state plot, we find the majority of distinguishing features are associated with changes in polycomb repressed regions and quiescent regions. Specifically, in Figure 4-14d, we can see the number of features that were assigned to each chromatin state in a cumulative distribution plot, based on various cutoffs for the top distinguishing features. Similarly, in Figure 4-14e, we visualize the distinguishing features as gene and chromatin state combinations. We can see that most genes had coordinated changes at weakly polycomb repressed (ReprPCWk) and quiescent (Quies) regions, with some genes additionally having changes at polycomb repressed (ReprPC), weakly transcribed (TxWk), bivalent enhancer (EnhBiv) and bivalent promoter (TssBiv) regions.

4.3.9 Subtypes of blood samples highlight enhancer differences.

Finally, we used Regulatory ChromDiff with linked enhancers to compare subtypes of blood samples. Specifically, we compared hematopoietic stem cells (HSCs) and B cells to T cells and blood samples. As shown in Figure 4-15, we find differences at 709 distinguishing genes, which are approximately grouped into four gene clusters based

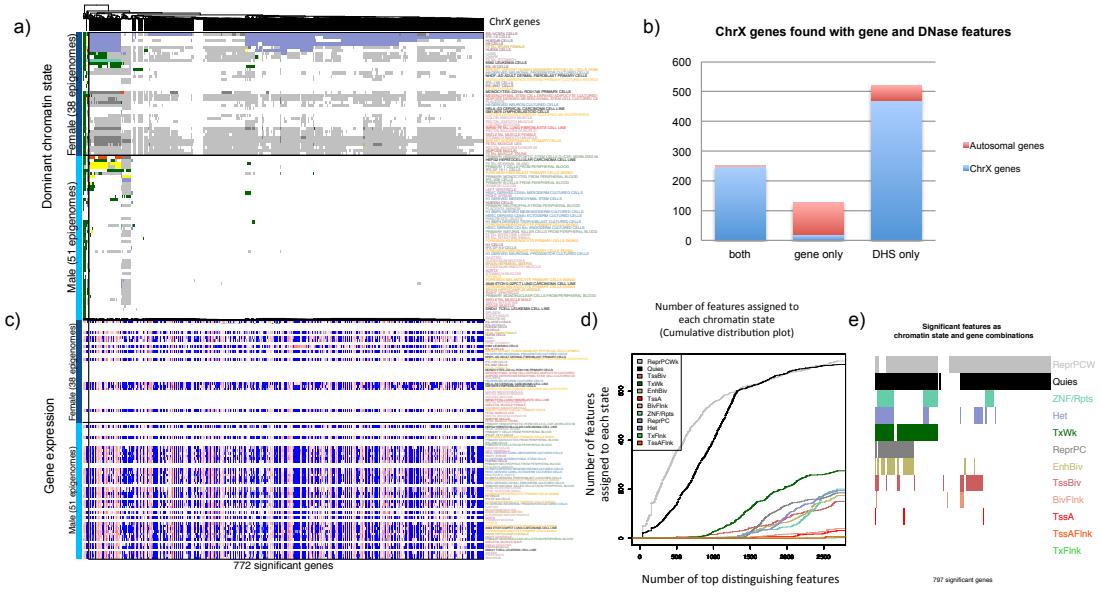

Figure 4-14: Differences in polycomb repression and heterochromatin is identified at DNase hypersensitive sites linked to ChrX genes in a sex-based comparison. a) Among the 772 distinguishing genes identified by Regulatory ChromDiff, we find a prevalent signature of polycomb repression (gray) and heterochromatin (purple) in female samples at X chromosome genes, as visualized by the dominant chromatin state plot. 719/772 of the identified genes are located on the X chromosome. b) Of the 738 X chromosome genes identified via the ene body ChromDiff or DHS Regulatory ChromDiff, 471 (64%) were identified only using the DHS approach. Furthermore, 248 (32%) were verified by both approaches, with only 18 (2%) identified by only the gene body approach. c) Gene expression levels are similar between female and male samples, likely due to differences in the number of X chromosome copies. d) Quiescent (Quies) and weakly polycomb repressed (ReprPCWk) dominate the most significant distinguishing features, as shown in a cumulative distribution plot. e) Most genes have coordinated changes at weakly polycomb repressed (ReprPCWk) and quiescent (Quies) regions, with some genes also showing changes at heterochromatic (Het), ZNF repeats (ZNF/Rpts), weakly transcribed (TxWk), polycomb repressed (ReprPC), bivalent enhancers (BivEnh), and bivalent promoters (TssBiv).

on their dominant chromatin state at linked enhancers (Figure 4-15a). We see that these gene clusters have different enhancer activity between the two groups: Clusters A and B tend to have higher enhancer activity in the T cells/blood group, compared to the B cells/HSCs group, while clusters C and D tend to have higher enhancer activity in the B cells group. Though it is hard to make broad conclusions based on the corresponding gene expression data, due to lack of samples (Figure 4-15b), we tend to see higher levels of gene expression in the samples showing higher enhancer activity, suggesting corresponding activity between enhancers and gene expression.

When we study the distinguishing features as gene and chromatin state combinations, we find the majority of features are associated with changes in the enhancer (Enh) ChromHMM state; since we used linked enhancer regions, this makes sense. However, we also some distinguishing features associated with flanking promoter (TssAFlnk), quiescent (Quies), and weakly transcribed (TxWk) regions. Finally, when we identify enriched gene sets in a cluster-specific way, we find many relevant gene sets. For example, in Figure 4-15d, we show that gene cluster B was enriched for many gene sets relating to T cell signaling, regulation, and differentiation, suggesting that epigenomic changes at linked enhancers can be used to identify celltype-specific genes.

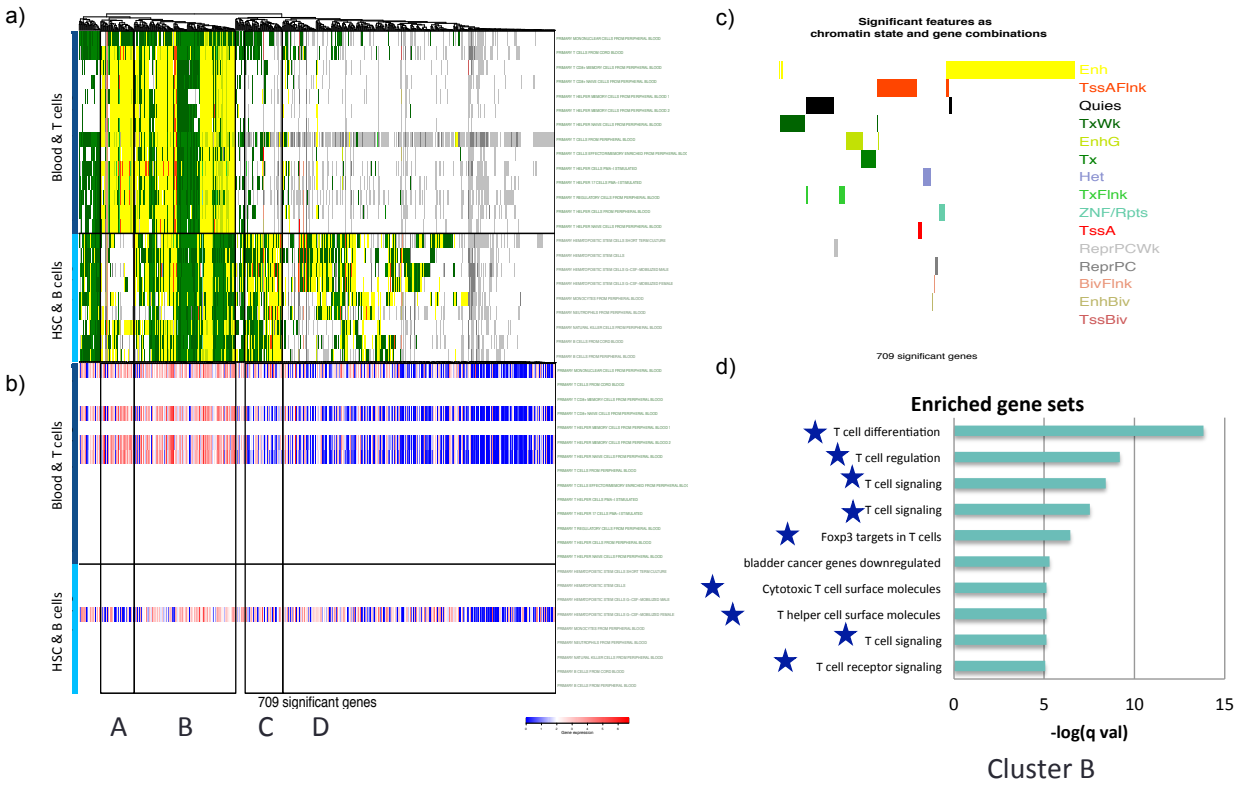

Figure 4-15: Epigenomic changes at linked enhancers identify celltype-specific gene expression and genes. a) Linked enhancer activity differs between HSCs/B cells and T cells/blood samples at 709 distinguishing genes, which can broadly be grouped into four clusters based on the dominant chromatin state at the linked enhancer regions. Clusters A and B have more enhancer activity in the T cell group, while clusters C and D have increased enhancer activity in the B cell group. b) Gene expression data, though limited, suggests coordinated gene expression and enhancer activity in a celltype-specific manner. c) Distinguishing features are largely associated with changes at enhancer regions, likely due to the fact that we targeted Regulatory ChromDiff towards linked enhancer regions. d) Gene clusters are enriched for relevant gene sets, such as enrichment of gene cluster B for gene sets relating to T cell differentiation, regulation, and signaling.

Chapter 5

Integrative analysis of Roadmap Epigenomics data

5.1 Introduction

While the primary sequence of the human genome is largely preserved in all human cell types, the epigenomic landscape of each cell can vary considerably, contributing to distinct gene expression programs and biological functions.[65,163–165] Epigenomic information, such as covalent histone modifications, DNA accessibility and DNA methylation can be interrogated in each cell and tissue type using high-throughput molecular assays.[47,164,166–168] The resulting maps have been instrumental for annotating cis-regulatory elements and other non-exonic genomic features with characteristic epigenomic signatures[28,169] and for dissecting gene regulatory programs in development and disease.[13,28,51,167,170,171] Despite these technological advances, we still lack a systematic understanding of how the epigenomic landscape contributes to cellular circuitry, lineage specification, and the onset and progression of human disease.

To facilitate and spearhead these efforts, the NIH Roadmap Epigenomics Program was established with the goal of elucidating how epigenetic processes contribute to human biology and disease. One of the major components of this programme consists of the Reference Epigenome Mapping Centers (REMCs),[172] which systematically characterized the epigenomic landscapes of representative primary human

tissues and cells. We used a diversity of assays, including chromatin immunoprecipitation (ChIP)[28,53,169,173] DNA digestion by DNase I (DNase),[167,174] bisulfite treatment,[163,164,175,176] methylated DNA immunoprecipitation (MeDIP),[177] methylation-sensitive restriction enzyme digestion (MRE),[178] and RNA profiling,[168] each followed by massively parallel short-read sequencing (-seq). The resulting data sets were assembled into publicly accessible websites and databases, which serve as a broadly useful resource for the scientific and biomedical community. Here we report the integrative analysis of 111 reference epigenomes, which we analyse jointly with an additional 16 epigenomes previously reported by the Encyclopedia of DNA Elements (ENCODE) project[28].[179]

Specifically, we identify epigenomic patterns in varied genomic regions and biological contexts by integrating a rich dataset of epigenomic information, including cell type, chromatin state, DNA methylation, and gene expression. We believe that our results here demonstrate the potential for biological insights by leveraging computational methods to study epigenomic data at the genome-wide level. We integrate information about histone marks, DNA methylation, DNA accessibility and RNA expression to infer high-resolution maps of regulatory elements annotated jointly across a total of 127 reference epigenomes spanning diverse cell and tissue types. We use these annotations to recognize epigenome differences that arise during lineage specification and cellular differentiation, to identify regulatory regions with coordinated activity across cell types, and to study the interplay of different epigenetic modifications. These analyses demonstrate the importance and wide applicability of our data resource, and lead to important insights into epigenomics, differentiation and disease.

5.2 Methods

5.2.1 Data processing of RNA-seq, ChIP-seq, and DNase-seq

For information on processing of the raw RNA-seq, ChIP-seq, and DNase-seq signal, see the Methods section of the integrative Roadmap Epigenomics paper.[11]

5.2.2 Data processing of DNA methylation data

For more information on processing of the WGBS, RRBS, and MeDIP/MRE/methylCRF DNA methylation data, see the Methods section of the integrative Roadmap Epigenomics paper.[11]

5.2.3 Chromatin state learning

To capture the significant combinatorial interactions between different chromatin marks in their spatial context (chromatin states) across 127 epigenomes, we used ChromHMMv.1.10106, which is based on a multivariate Hidden Markov Model.

We generated a "core" 15-state chromatin state model. Specifically, a ChromHMM model applicable to all 127 epigenomes was learned by virtually concatenating consolidated data corresponding to the core set of five chromatin marks assayed in all epigenomes (H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3). The model was trained on 60 epigenomes with highest-quality data (Figure 4-1k), which provided sufficient coverage of the different lineages and tissue types. The ChromHMM parameters used were as follows: reads were shifted in the 59 to 39 direction by 100 bp. For each consolidated ChIP-seq data set, read counts were computed in non-overlapping 200-bp bins across the entire genome. Each bin was discretized into two levels, 1 indicating enrichment and 0 indicating no enrichment.

The binarization was performed by comparing ChIP-seq read counts to corresponding whole-cell extract control read countswithin each bin and using a Poisson P value threshold of 131024 (the default discretization threshold in ChromHMM). We trained severalmodels in parallelmode with the number of states ranging from 10 states to 25 states.We decided to use a 15-state model (Figure 4-10k)aâĂŞf) for all further analyses since it captured all the key interactions between the chromatin marks, and because larger numbers of states did not capture sufficiently distinct interactions. The trained model was then used to compute the posterior probability of each state for each genomic bin in each reference epigenome.

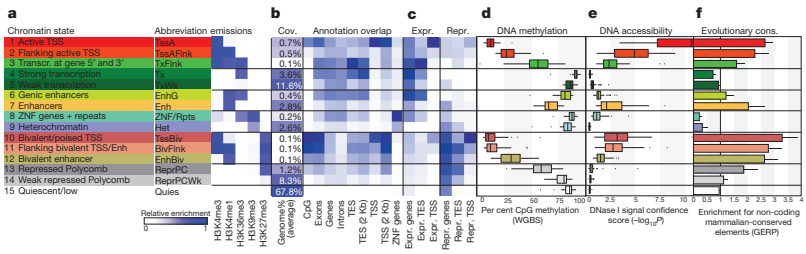

Figure 5-1: Chromatin states and DNA methylation dynamics. a, Chromatin state definitions, abbreviations and histone mark probabilities. b, Average genome coverage. Genomic annotation enrichments in H1-ES cells. c, Active and inactive gene enrichments in H1-ES cells. d, DNA methylation. e, DNA accessibility. d, e, Whiskers show 1.53 interquartile range. Circles are individual outliers. f, Average overlap fold enrichment for GERP evolutionarily conserved non-exonic nucleotides. Bars denote standard deviation.

5.2.4 Relationship between chromatin states and methylation.

The distribution of DNA methylation (percent CpG methylation from WGBS data) was computed using regions belonging to each of the 15 chromatin states based on the core set of 5 marks across all reference epigenomes for which these data sets were available (Figure 5-1d, e). CpGs with a minimum read coverage of 5 were used to calculate the average methylation percentages within genomic regions labelled with each chromatin state from the 15-state primary model. Only regions containing more than 3 CpGs with at most 200 bp between consecutive CpGs were used. Plots were generated using ggplot2 package for R (v.3.02).[135]

5.2.5 Generating randomized simulations of methylation profiles.

For each chromatin state that we were interested in, we wanted to compare how different the distribution of DNA methylation was compared to the rest of the genome. To do this, for each celltype, we took the original, real chromatin state regions called, and we generated "simulated regions" of the same length in the same celltype, to control for the effects of region length on DNA methylation averages. Furthermore, our original lengths were filtered for requiring at least 3 CpGs, so our randomized

regions also required at least 3 CpGs. Using these simulated regions, we generated a "background methylation profile" by averaging the fractional methylation values over the region.

5.2.6 Calculating significant differences in simulated vs real data.

Based on the DNA methylation values for real and randomized chromatin state regions, we wanted to quantify how similar or dissimilar the real and randomized data was. To do this, we sampled 5000 regions from the real chromatin state regions, and 5000 regions from the corresponding randomized regions, which were matched for genomic size. Then, we performed a two-sided paired Mann-Whitney-Wilcoxon test to calculate how likely these samples were drawn from different distributions. For multiple hypothesis correction, we performed Bonferroni correction for the 15 chromatin states tested..

5.2.7 Clustering of celltypes based on epigenetic information.

We performed hierarchical clustering of the samples based on epigenetic information. We initially studied various histone modifications (H3K4me1, H3K27me3, HeK36me3, and H3K4me3) at various regions (genome-wide, gene bodies, or enhancers). Based on the resulting clusters of samples, we chose to use the H3K4me1 signal at enhancer regions to calculate Pearson correlations between samples. Specifically, the H3K4me1 signal was averaged for each enhancer region in each sample type, resulting in each sample type being represented as a vector of length 502,064, for 502,064 enhancers. Based on these vectors, a Pearson correlation score was calculated for each pair of sample types, producing a 90x90 matrix for the 90 sample types.

Then, using this matrix of pairwise celltype similarity, we performed hierarchical clustering of the celltyeps, producing a dendrogram for the celltypes. By producing the optimal leaf ordering for this dendrogram, we obtained the ordering of celltypes that could then be used consistently in cell type correlation heatmaps based on other

epigenetic features. For example, even though the ordering was based on H3K4me1 signal at enhancers, the analogous process could be used to calculate pairwise Pearson correlation values between samples based on H3K4me3, H3K27me3, and H3K36me3 as well.

5.2.8 Comparison of DNA methylation platforms

For a subset of epigenomic samples, we had DNA methylation data acquired from multiple platforms. This was ideal for identifying discrepancies between platforms or platform-specific biases. For any sample that had data based on at least 2 of the platforms, we plotted respective methylation values for common sites in a scatterplot. Due to the high density of values, we used hexplots with an exponential color scale. Specifically, the hexplot counts the number of points in each hexagonal space, and the color scale shows that count based on a log color-scale.

5.3 Results

5.3.1 Relationship between DNA methylation and chromatin states

With the vast amount of chromatin state and DNA methylation information available through Roadmap Epigenomics, we were able to study the DNA methylation patterns for each chromatin state. As shown in Figure 5-2, we then calculated a probability distribution function for each chromatin state. This revealed that while most chromatin states tend to be hypermethylated (which mirrors the fact that the majority of the genome is hypermethylated), chromatin states related to promoters (TssA, TssBiv, BivLnk, TssAFlnk) tended to be unmethylated. Further, this provides strong evidence that distinct chromatin states have distinct DNA methylation patterns, even though chromatin states are learned only from histone mark information.

However, numerous attributes of chromatin state regions could influence their DNA methylation patterns. For example, a chromatin state can occurr in a broad

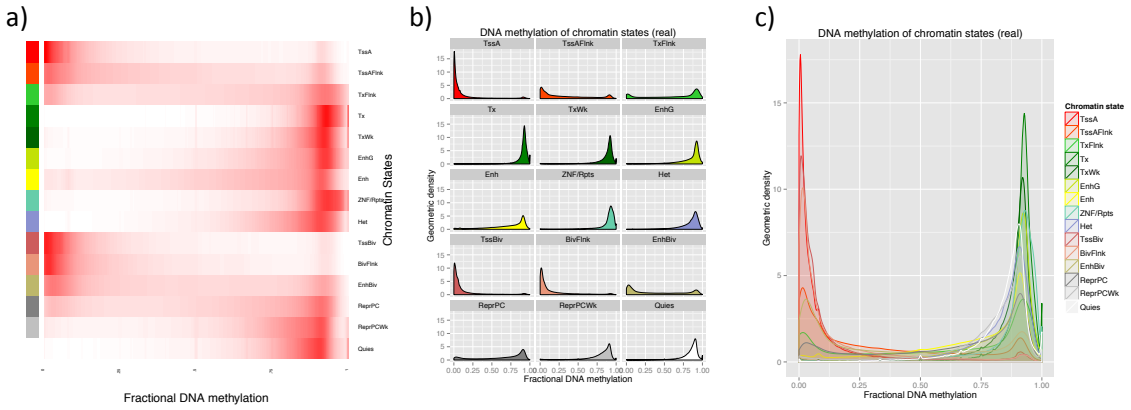

Figure 5-2: DNA methylation values at chromatin state regions in 15-state model in an a) matrix representation, b) density plot representation for each chromatin state, and c) overlapping density plot representation for each chromatin state. DNA methylation occurrs mostly in a bimodal manner, with many promoter and bivalent states (shades of red) showing hypomethylation and many transcription (green) and repressed (gray) states exhibiting hypermethylation.

or narrow pattern, and the prevalence of each chromatin state across the genome also varies. To quantify how much of a potential effect these factors could have on the DNA methylation profile, we generated "simulated" DNA methylation profiles. Specifically, we chose random regions of matched size and number to our real chromatin state regions, based on a uniform distribution across the genome. In this way, we generated a "null distribution" of DNA methylation values for each set of chromatin state regions.

In Figure 5-3a, we see that many simulated random regions tend to have high DNA methylation values, reflecting the fact that the majority of the genome is hypermethylated. We also see that the DNA methylation values of many chromatin states closely reflect the DNA methylation of random regions matched by length and number. However, for some chromatin states, we see a DNA methylation pattern very distinctive from the randomized null distribution. For example, Figure 5-3a shows that promoter sites (TssA and TssAFlnk) and bivalent regions (TssBiv, BivFlnk, EnhBiv) are much less methylated than their randomized counterparts, suggesting that hypomethylation of promoter and bivalent regions could be biologically meaningful in transrrcriptional machinery or regulatory mechanisms.

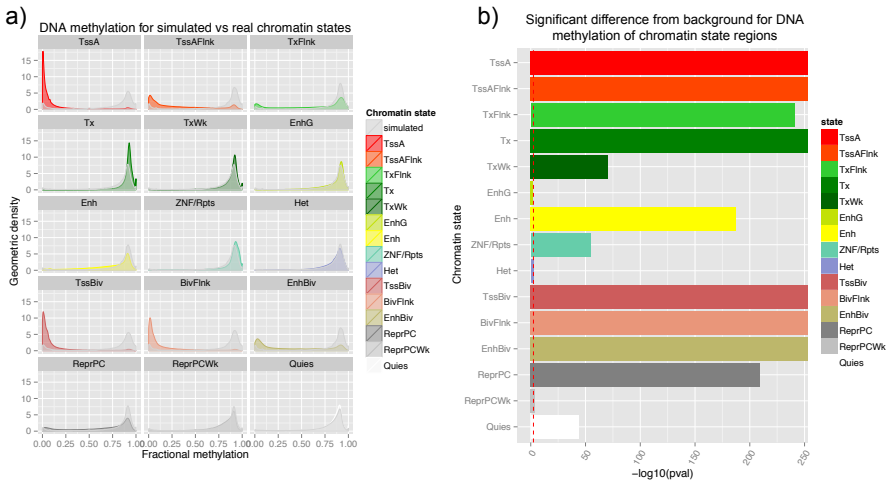

Figure 5-3: a) DNA methylation of real chromatin state and simulated regions suggests that promoter (TssA and TssAFlnk) and bivalent (TssBiv, BivFlnk, and EnhBiv) are distinctly hypomethylated compared to the rest of the genome, suggesting a biological link between the chromatin state and DNA methylation. b) Promoter, transcribed, bivalent, and polycomb repressed regions have distinctive DNA methylation patterns. Specifically, the Mann-Whitney-Wilcoxon test reveals that promoter (TssA, TssAFlnk), transcribed (TxFlnk, Tx), bivalent (TssBiv, BivFlnk, EnhBiv), and polycomb repressed (ReprPC) regions have the most distinctive DNA methylation patterns compared to background. On the other hand, genic enhancers (EnhG), heterochromatic (Het), weakly polycomb repressed (ReprPCWk), and quiescent regions (Quies) have DNA methylation patterns most similar to background.

To quantify these differences, we sampled 5000 real and randomized regions for each chromatin state, and calculated a Mann-Whitney-Wilcoxon p-value to determine how different the chromatin state and background DNA methylation distributions were. In Figure 5-3b, we show that promoter, transcribed, bivalent, and polycomb repressed regions have DNA methlyation patterns most different from the genomic background. This finding is consistent with the Figure 5-3, which showed that promoter (TssA, TssAFlnk) and bivalent (TssBiv, BivFlnk, EnhBiv) regions were the most hypomethylated, setting them apart from the genomic background which is mostly hypermethylated. Notably, we also see a distinctive DNA methylation pattern for transcribed regions (TxFlnk, Tx), enhancer (Enh), and polycomb repressed (ReprPC). Even though these regions tend to be hypermethylated overall (Figure 5-3a), the TxFlnk, Enh, and ReprPC regions are slightly less methylated overall compared to background, while transcribed (Tx) regions are slightly more methylated.

As suggested by the visualization of real vs randomized regions in Figure 5-3a, Figure 5-3b quantifiably verifies that genic enhancers, heterochromatic, weakly polycomb repressed, and quiescent regions have DNA methylation values most similar to the genomic background.

5.3.2 DNA Methylation profiles for chromatin states across epigenomes

We next studied the relationship between DNA methylation dynamics and histone modifications across 95 epigenomes with methylation data, extending previous studies that focused on individual lineages.[14,15,175,180] We found that the distribution of methylation levels for CpGs in some chromatin states varied significantly across tissue and cell type, tracking closely with the sample's developmental and differentiated stage. (Figure 5-4a).

For example, flanking promoter (TssAFlnk) regions were largely unmethylated in terminally differentiated cells and tissues such as Heart and Digestive samples, but were frequently methylated for several embryonic stem cell and embryonic-stem-

cell-derived samples (Bonferroni-corrected F-test $p < 0.01$). Enhancer (Enh) regions were highly methylated in pluripotent cells (IMR, ES, iPS, and ES derived cells), but showed a broader distribution of intermediate methylation in differentiated cells and tissues ($p < 0.01$). Bivalent enhancer (EnhBiv) reginos were unmethylated in most primary cells and tissues, but showed a broader distribution of methylation levels in pluripotent cells, possibly reflecting cell-to-cell heterogeneity ($p < 0.01$). The consistency of DNA methylation level across samples also was highly dependent on the chromatin state: for example, polycomb repressed (ReprPC) regions showed highly varying levels of methylation across samples, while heterochromatic (Het) regions showed consistently high levels of methylation in almost all epigenomes.

For completion, we also present DNA methylation profiles for all 15 chromatin states in the core model across all available epigenomes and technologies (Figure 5-4b). Again, we see that the experimental technology seems to have an effect on the resulting methylation values (WGBS in red, RRBS in blue, mCRF in green), but that general trends still emerge. For example, quiescent states (Quies) tend to be hypermethylated across all technologies, providing a useful "default" background methylation state. Compared to the flanking promoter regions (TssAFlnk), other promoter regions (TssA, TssBiv, BivFlnk) tend to be more consistently hypomethylated, and transcription regions (Tx, TxWk) again show strong evidence for consistent hypermethylation.

To quantify how different the DNA methylation patterns were at chromatin state regions for each celltype, we again used our randomized genomic regions. Specifically, for each chromatin state and celltype, we sampled 5000 real and 5000 matched randomized regions, and then compared average DNA methylation values using the Mann-Whitney-Wilcoxon test.

Then, we sorted the celltype and chromatin state combinations based on how much their real regions differed from background, and grouped them into three groups. As shown in Figure 5-5a, combinations with promoter and bivalent regions (TssA, TssAFlnk, TssBiv, BivFlnk, EnhBiv) most frequently showed up in the group with regions most different from background. This confirmed our previous findings across

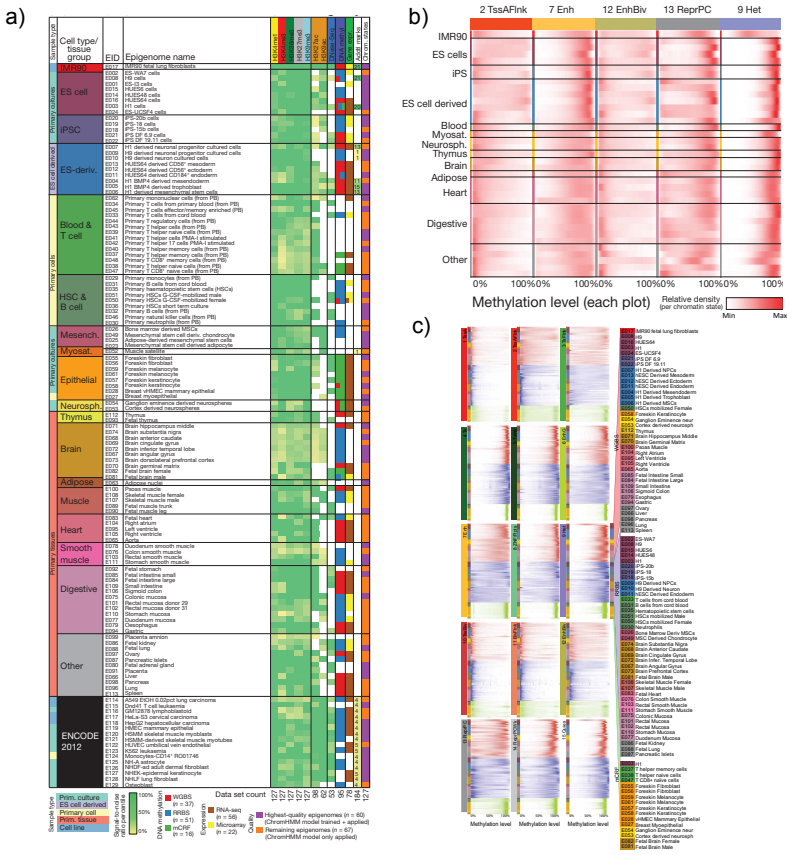

Figure 5-4: a) 127 celltypes were used to generate b) celltype-specific DNA methylation profiles of flanking promoter, enhancer, bivalent, polycomb repressed, and heterochromatic regions, showing that DNA methylation levels vary with both chromatin state and sample type. DNA methylation shown is from whole-genome bisulfite sequencing experiments (WGBS), with the geometric density of chromatin state regions at that methylation value shown on a natural log color scale across cell types. (White=min ln(density+1), Red=max ln(density+1).) Left column indicates celltype and tissue groupings; a full list of sample IDs is shown in panel c with complete annotations in panel a. c) DNA methylation variation across cell types. Density plots denote distribution of DNA methylation levels from 0% to 100% for each chromatin state across the 95 reference epigenomes profiled for whole-genome bisulfite (WGBS, red), reduced representation bisulfite (RRBS, blue), or MeDIP/MRE (mCRF, green). The respective colour (red, blue, or green) was set to the maximum ln(density+1) value for each chromatin state and respective platform, with intermediate values coloured on a natural log scale. For each panel, the subset of reference epigenomes profiled using each technology are listed, using the colours, order, and abbreviations shown in panel a.

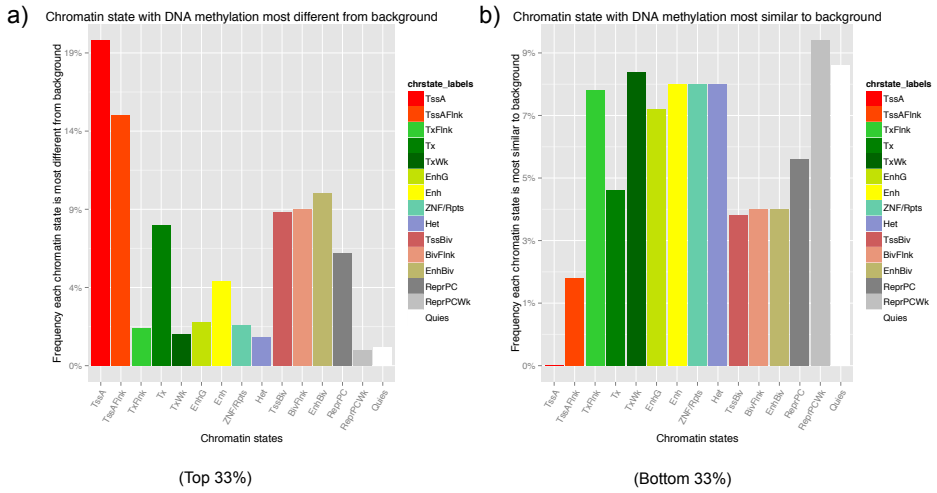

Figure 5-5: Across celltypes, promoter and bivalent regions are methylated most differently from background, while quiescent and polycomb repressed regions most frequently have DNA methylation similar to background. Combinations of celltype and chromatin state were divided into three groups based on p-values quantifying the difference between DNA methylation values of real and background regions.

celltypes that promoter and bivalent regions were DNA methylated in a way most distinctive from background, as shown in Figure 5-3c.

On the other hand, when looking at the celltype and chromatin state combinations with DNA methylation values most similar to background (Figure 5-5b), we find weakly polycomb repressed (ReprPCWk) and quiescent (Quies) regions show up most frequently. Again, this is largely consistent with our previous celltype-agnostic comparison of DNA methylation of real and simulated chromatin state regions (Figure 5-3c)

In Figure 5-6a, we merge all the chromatin state regions identified across all platforms and identify a per-epigenome DNA methylation profile for each chromatin state. This is similar to Figure 5-4b, but with one distribution for each epigenome, even when multiple platforms are involved. Though we can certainly see celltype-variation in the DNA methylation density values, it is difficult to identify the most relevant differences due to the overlapping information.

Therefore, in Figure 5-6b, we calculate a pooled DNA methylation distribution for each group and chromatin state, so that we can see some of the differences across

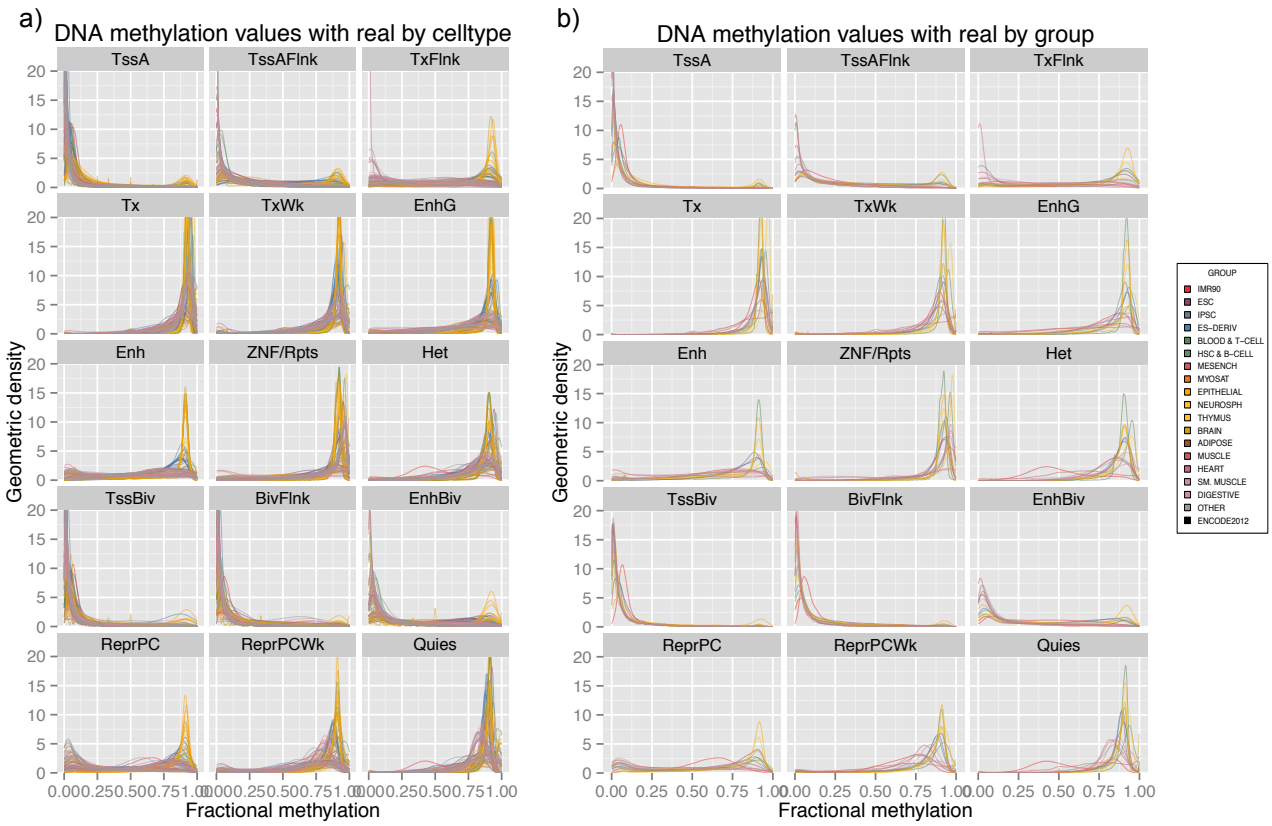

Figure 5-6: Distribution of DNA methylation values for each chromatin state based on sample group.

epigenomic group. For example, enhancer state regions seem to be highly methylated in blood cells (green), but less methylated in cultured cells such as embryonic stem cells (gray), induced pluripotent stem cells (purple), and ES-derived cell cultures (blue).

The sample IMR90 (in red) consistently sticks out as having different DNA methylation values in many chromatin states, compared to the other samples, both at the celltype-level in panel a, and the group-level in panel b (Figure 5-6). For example, IMR90 shows intermediate average methylation levels in heterochromatic (Het), polycomb repressed (ReprPC), and quiescent (Quies) regions.

To interpret whether this difference in IMR90 should be attributed to changes specifically at chromatin states, we utilized our simulated chromatin states in the IMR90 sample, as shown in Figure 5-7. Though there are still differences in DNA

methylation patterns between the real chromatin states and random genomic regions, it is also clear that the random regions of IMR90 seem to be methylated differently than the average over all epigenomes, as shown in Figure 5-3. Specifically, while most DNA methylation values tend to be highly methylated in random regions (consistent with the fact that the majority of the genome is hypermethylated), many random IMR90 regions tend to have a broad distribution of DNA methylation values, with a much higher proportion of intermediate levels of methylation (around 50%) than other epigenomes.

This suggests that, while DNA methylation of chromatin states in IMR90 may still be meaningful, the difference between IMR90 and other epigenomes is confounded by the overall differences in DNA methylation genome-wide between IMR90 and other epigenomes. In fact, existing literature supports our finding, with a previous methylome study of IMR90 and another cell line (H1) showing that only 67.7% of CG sites were methylated in IMR90, compared to 82.7% in H1 cells.[175] This analysis emphasizes the importance of many factors for comparative analysis of DNA methylation of chromatin states, including celltype, sample group, genome-wide methylation patterns, the number and size of chromatin state regions, and other comparison samples.

5.3.3 Comparison of DNA methylation platforms

Due to the integrative nature of the Roadmap Epigenomics Consortium, there was a variety of experimental technologies used to generate data. For the DNA methylation data specifically, the three technologies used were whole genome bisulfite sequencing (SBS), reduced-representationl bisulfite sequencing (RRBS), and methyl Conditional Random Field (mCRF), as described in 1.3.4.

Though the experimental platform used in each sample varied across celltypes (Figure 5-4a), there were some samples for which DNA methylation data was obtained via multiple technologies. This enabled a direct comparison of how consistent different experimental techniques are. Specifically, for each celltype with DNA methylation produced from more than one technology, we compared the fractional methylation identified by each platform for the same CpG site. Ideally, we would expect a perfect

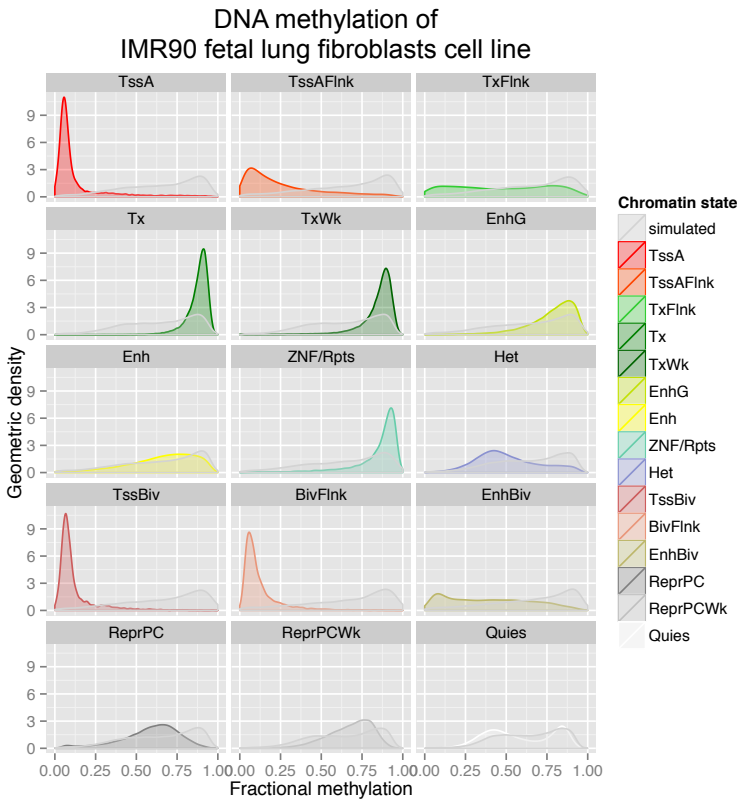

Figure 5-7: Real and simulated distribution of DNA methylation values for each chromatin state in the IMR90 feta llung fibroblasts cell line.

match between the value generated from each two technologies, resulting in a perfect $y = x$ regression line with $r^2 = 1$. In reality, of course, this was not the case, as shown in Figure 5-8, Figure 5-10, and Figure 5-9. For these comparisons, we filtered only to CpG sites that were covered by both technologies.

We see fairly strong concordance between the DNA methylation values for the two bisulfite-sequencing based approaches, as shown in Figure 5-8. The overall hex bin plots show a high density of points in the bottom left and the top right, showing that many CpG sites are consistently identified as hypomethylated or hypermethylated from both techniques. The best-fit linear regression lines identified based on the original CpG sites are also close to $y = x$, with y-intercepts close to 0 (between .02 and .1) and slopes close to 1 (between .82 and .97). Furthermore, the coefficient of determination, or r^2 value, is fairly high, ranging from $.755 - .898$. The consistency between these two technologies gives confidence to DNA methylation values identified by either, and suggests that RRBS could be a good lower-cost alternative to SBS. However, even though RRBS closely replicates the values of SBS for a lower cost, there are still benefits to SBS not depicted in this visualization, as SBS provides DNA methylation values for every CpG site in the genome, while RRBS profiles far fewer locations.

Though the bisulfite-based technologies show strong concordance with each other, when we compare the bisulfite sequencing results with the mCRF values, we see much less agreement. For example, when comparing SBS to mCRF, as shown in Figure 5-9, many CpG sites that seem to be intermediately methylated from the SBS dataset, are predicted to be hypermethylated in the mCRF dataset, as shown by the horizontal red streak in the top middle-right corner. Furthermore, each cell type seems to have its own particular biases. For example, many of the CpG sites in the H1 cells are predicted as highly methylated in SBS (between $.75 - 1$), but hypomethlyated or intermediately methylated in mCRF (as shown by the horizontal red lines). On the other hand, in the primary skin cells, mCRF predicts a range of DNA methylation values for sites that are hypomethylated in SBS, as shown by the vertical line close to the y-axis.

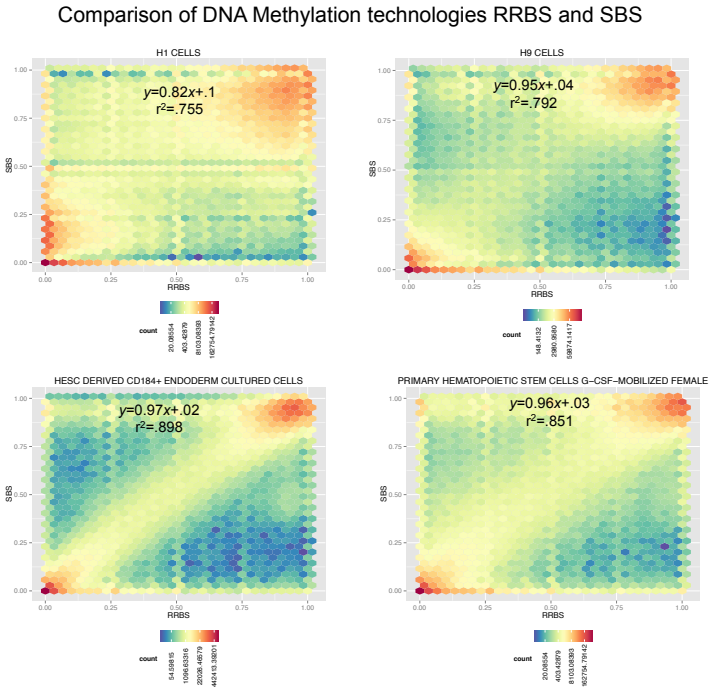

Figure 5-8: Comparison of DNA methylation technologies RRBS and SBS across four samples, including embryonic stem cells, ESC-derived cultured cells, and primary hematopoetic stem cells, shows fairly strong concordance between reduced representation bisulfite sequencing (RRBS) and whole genome bisulfite sequencing (SBS).

To quantify these discrepancies, we performed linear regression on the underlying points (one per CpG site), and the results confirmed our visual assessment. The y-intercept values, which should ideally be close to 0, were markedly higher, ranging from $.39 - .544$, indicating the best-fit line would predict an intermediate methylation value in mCRF for a completely hypomethylated CpG site in SBS (methylation value of 0). Furthermore, the slopes are much further from the ideal slope of 1, with values ranging from $.53 - .8$. Lastly, not only was the regression line far from ideal, the coefficient of determination values also suffered, with r^2 values ranging from $.39 - .544$, suggesting that the linear model was not a good fit at all for this data.

Therefore, although mCRF can identify genome-wide basepair-resolution DNA methylation values at a lower cost, the accuracy of these values suffers through the use of computational inference techniques, resulting in clear biases in the mCRF values. As a result, for those who are strongly constrained by experimental cost, the choice of RRBS or mCRF would need to be made based on the relative importance of genomic coverage and precision, as each technology has its strengths and weaknesses.

When comparing RRBS and mCRF values, we see more agreement than we did between SBS and mCRF values. As shown in Figure 5-10, we see regression lines closer to $y = x$, with higher slopes (ranging from $.85 - .95$) and lower y-intercepts (ranging from $.1 - .15$). However, the coefficient of determination still suggests a fair amount of inconsistency in values between the two technologies, with r^2 values of .618 and .762, respectively, for fetal brain tissue and embryonic stem cells.

The fact that mCRF seems to be more consistent with RRBS than with SBS suggests that the CpG sites selected in RRBS may be more consistent and easier to predict. This makes biological sence, as RRBS enriches for genomic regions with a high density of CpG sites, and mCRF is based on underlying enrichment-based data from me-DIP and methyl-Seq. Therefore, both RRBS and mCRF could be reasonable choices for experiments that are focused on genomic regions with a high-density of CpG regions, though RRBS is still likely a bit more accurate. However, despite its high cost, SBS remains the true "gold standard" of DNA methylation values with unequaled accuracy and coverage.

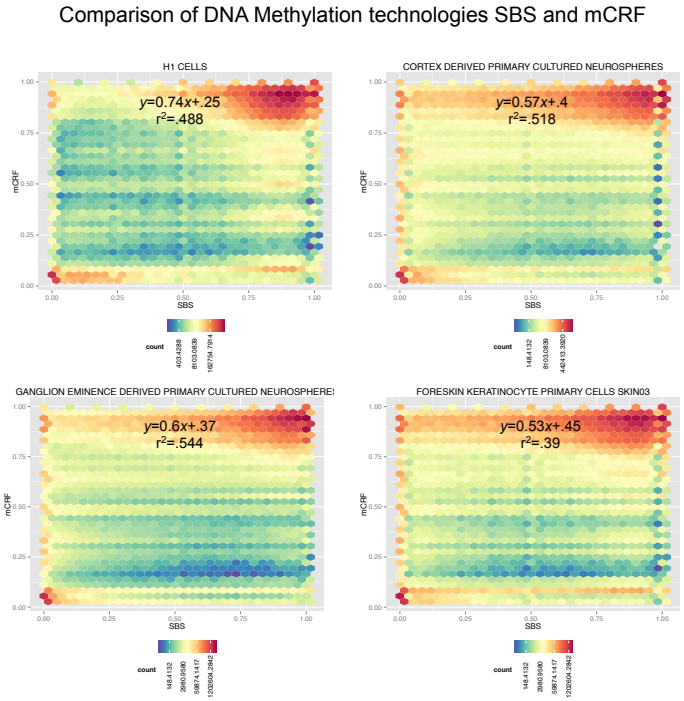

Figure 5-9: Comparison of DNA methylation technologies SBS and mCRF across four samples, including embryonic stem cells, ESC-derived cell cultures, and primary skin cells, shows a tendency for mCRF to predict intermediately-methylated sites (according to RRBS) as close to 1, as shown by the horizontal red streak on the top right of the plots.

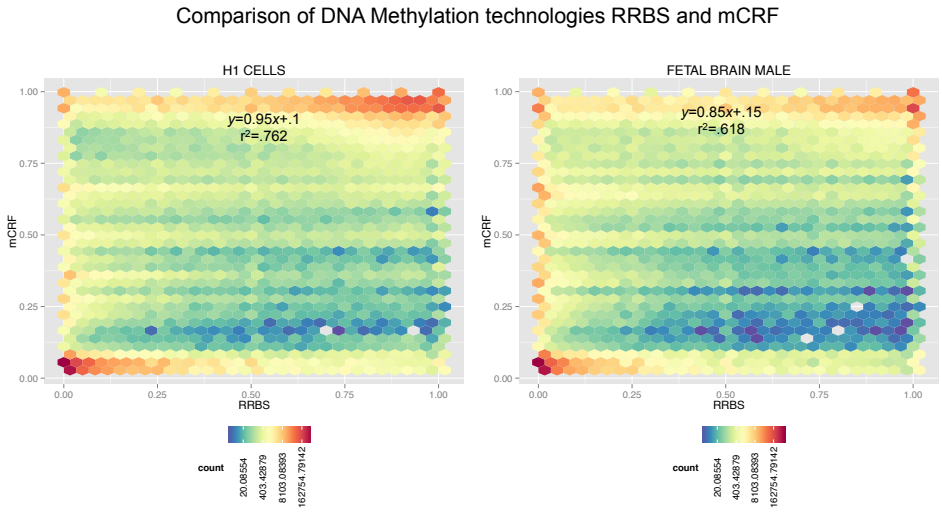

Figure 5-10: Comparison of DNA methylation technologies RRBS and mCRF across two samples (embryonic stem cells and brain tissue) shows a tendency for mCRF to predict intermediately-methylated sites (according to RRBS) as close to 0 or 1, as shown by the horizontal red streaks on the bottom left and top right of the plot. Furthermore, there is an overall tendency to mCRF values to be predicted as higher than RRBS values, as shown by the general yellow-red enrichment above the y=x line.

After comparing DNA methylation values at individual CpG sites, we wanted to see if platform-specific differences affected the average DNA methylation level for chromatin state regions. Therefore, we generated DNA methylation profiles for each chromatin state in a platform-specific way. In Figure 5-11, we represent this information in box-plot formats across the three platforms, clearly visualizing differences in DNA methylation for different platforms, as expected based on our direct comparison of DNA methylation values for CpG sites above.

Despite platform differences, we still see consistent relative patterns for chromatin states, as shown in Figure 5-11 for the three platforms: a) WGBS, b) RRBS, and c) mCRF. Specifically, promoter regions (TssA, TssBiv, BivFlnk) were consistently some of the most hypomethylated region across technologies, while Quies, ReprPCWk, and Tx were consitently hypermethylated.

Interestingly, the differences in visualizations between boxplots (Figure 5-11a-c) and density distributions (Figure 5-11d) highlights the different interpretation of in-

termediate methylation levels. For example, in panel a-c, chromatin states that have an average intermediate methylation level are TxFlnk and EnhBiv. However, the box plot visualization is based on using the average value for that chromatin state whereas the density distribution visualizes somewhat bimodal distributions. For examples, though TxFlnk and EnhBiv regions have intermediate average methylation values across platforms (ranging from about 50%-80%), the density distribution plots in panel d indicate bimodal distributions for both states, especially in the mCRF data (green). Though there are more TxFlnk and EnhBiv regions with intermediate methylated values than other states such as Quies and ZNF/Rpts, the methylation values with the highest density are still focused at the hypomethylated (0-25%) and hypermethylated (75-100%) values. This suggests a general consistency in DNA methylation within a given chromatin state region, as the average methylation of a particular genomic region is either close to 0 or 1. Nevertheless, within a certain chromatin state category, there is clearly a diversity of DNA methylation patterns, especially for certain chromatin states with higher variance, such as TxFlnk and EnhBiv regions.

Finally, we could generate platform-based DNA methylation values for each epigenomic group, as shown in Figure 5-12. Interestingly, the groups profiled using mCRF (panel b) seem to have more consistent DNA methylation values with each other than for the groups with data from SBS (panel a) and RRBS (panel c). This could be because the groups that used mCRF to obtain DNA methylation values (epithelial, neurosphere, brain, T cells, and ESCs) are more similar to one another than for RRBS and WGBS. However, since we have already shown that mCRF often gives inconsistent results compared to the Bisulfite-Sequencing based approaches (Figure 5-9), it also raises the possibility that the mCRF biases make it difficult to pick up the more nuanced differences that may occurr between groups of epigenomes.

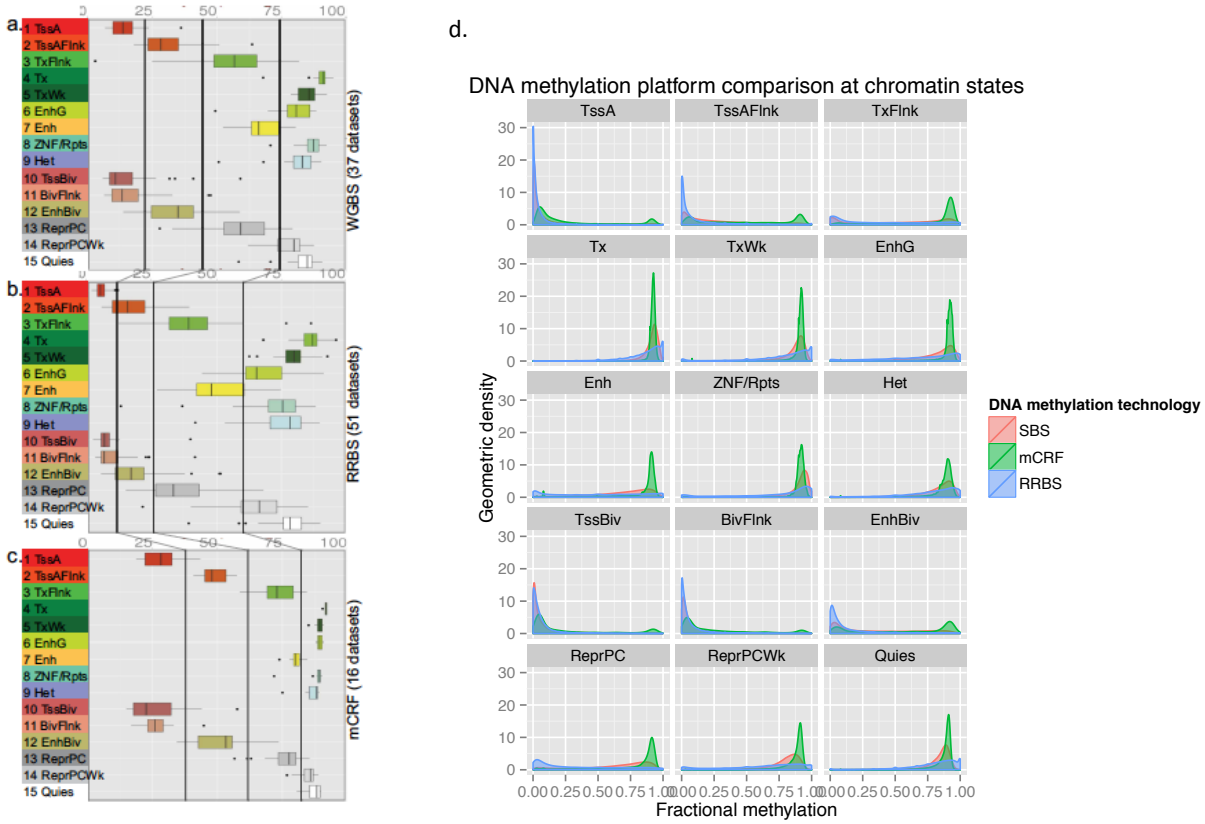

Figure 5-11: DNA methylation levels in 15-state model across technologies. We observed significant differences in the average methylation levels observed that were correlated with the different DNA methylation platforms used, but the relative relationships in average DNA methylation level were conserved across chromatin states. WGBS (panel a) values were overall higher than RRBS results (panel b) (as expected given the enrichment for CpG island in RRBS experiments). In contrast, mCRF experiments (panel c) resulted in higher levels of DNA methylation than WGBS experiments. In panel d, the overlaid distributions of DNA methylation levels in each chromatin state again shows that mCRF seems to give consistently higher levels of DNA methylation across chromatin states, as is consistent with our previous findings in Figure 5-9 and Figure 5-10. This highlights the importance of recognizing and potentially correcting for DNA methylation-platform-specific biases before performing integrative analyses.

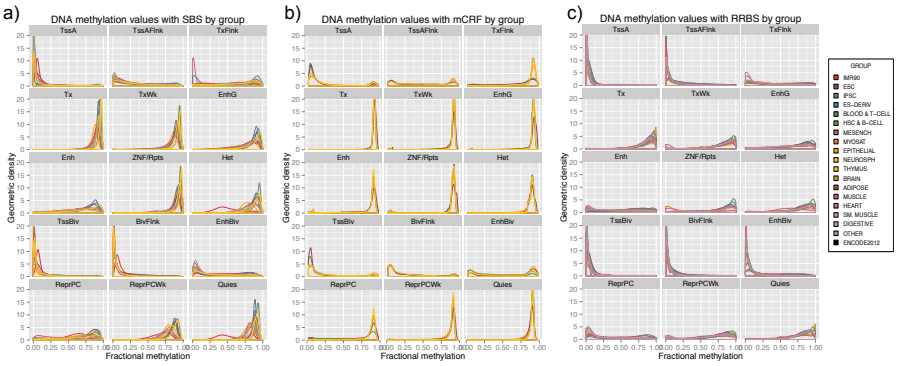

Figure 5-12: Distribution of DNA methylation values for each chromatin state based on sample group for a) whole genome bisulfite sequencing, b) methylCRF, and c) reduced representation bisulfite sequencing.

5.3.4 Clustering of celltypes based on pairwise epigenomic similarity.

With our wide variety of epigenomic data across celltypes, we were also interested in how similar celltypes are, based on epigenomic metrics. In Figure 5-13, we calculated the average DNA methylation value in each epigenome for each cluster of enhancers, and then clustered both enhancer modules and epigenomes based on these values. Again, we see the importance of the methylation technology used, with many datasets clustering based on the experimental platform used. However, we certainly see some consistency in methylation state for enhancer clusters across celltyeps, with many consistently hypomethylated enhancer clusters (on the left in blue). Some enhancer clusters also tend to show dynamic methylation values across epigenomes (on the right), with hypermethylation in the majority of samples, except for hypomethylation in samples in the digestive, smooth muscle, and brain groups. Additionally, we see the methylation values of IMR90 stand out, consistent with our findings in Figure 5-7, as it has hypomethylated values even in hypermethyatled enhancer clusters.

In addition to clustering samples based on DNA methylation values, we also clustered samples based on histone mark profiles. Specifically, for each histone mark, we calculated the average value in each enhancer in each epigenome, and then calculated pair-wise correlations between epigenomes based on these gene-mark values.

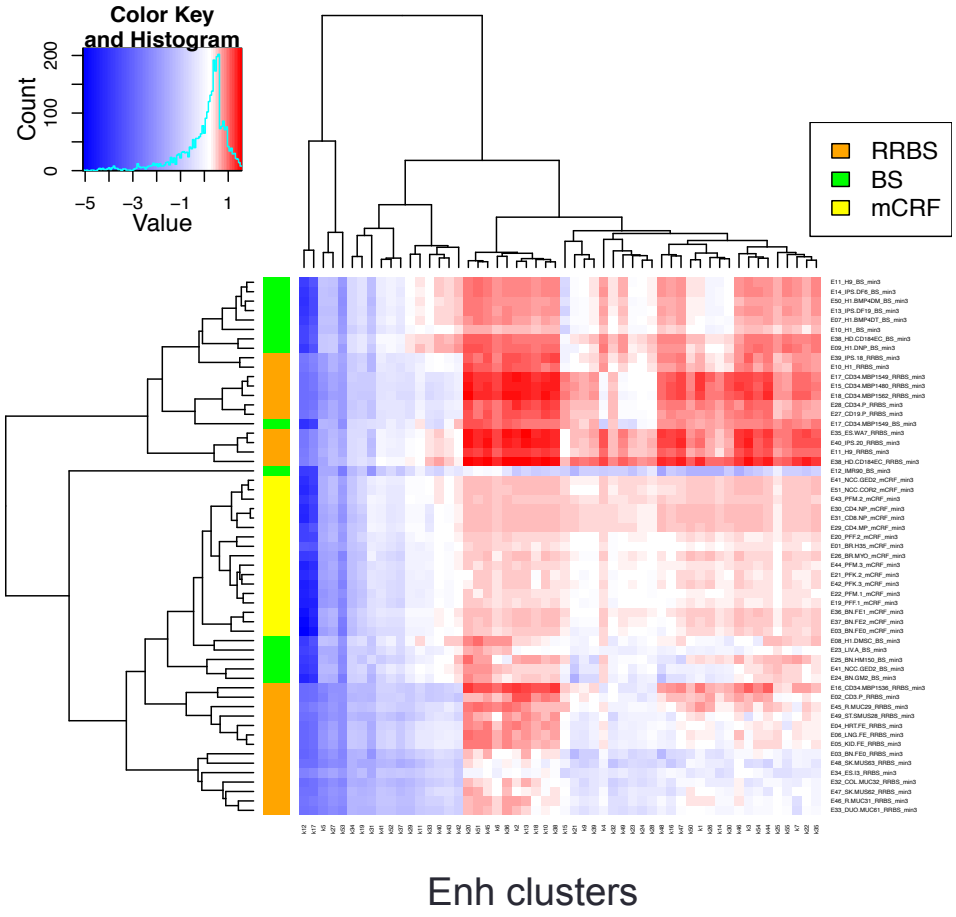

Figure 5-13: DNA methylation at enhancer modules across celltypes. Epigenomes somewhat cluster by DNA technology used, as well as sample group.

Then, we averaged the pair-wise correlation values for four histone marks: H3K4me1, H3K4me3, H3K27me3, and H3K36me3. We can order the celltypes based on these average correlation values (on both axes), and then visualize the strength of correlations between epigenomes, as shown in Figure 5-14.

These results show that histone marks do, in fact, group epigenomes into consistent clusters. However, we also see more consistent correlations within groups for certain marks, such as H3K4me1 and H3K27me3. On the other hand, H3K36me3 and H3K4me3 tend to have relatively similar correlations even across different epigenomic groups. These results highlight the ability of epigenomic marks to distinguish groups of samples from one another, highlighting a similar principle to the one we used in Chapter 3 and Chapter 4. These results further confirm the importance of studying epigenomic patterns to better understand on celltype-specific gene regulation, development, and differentiation.

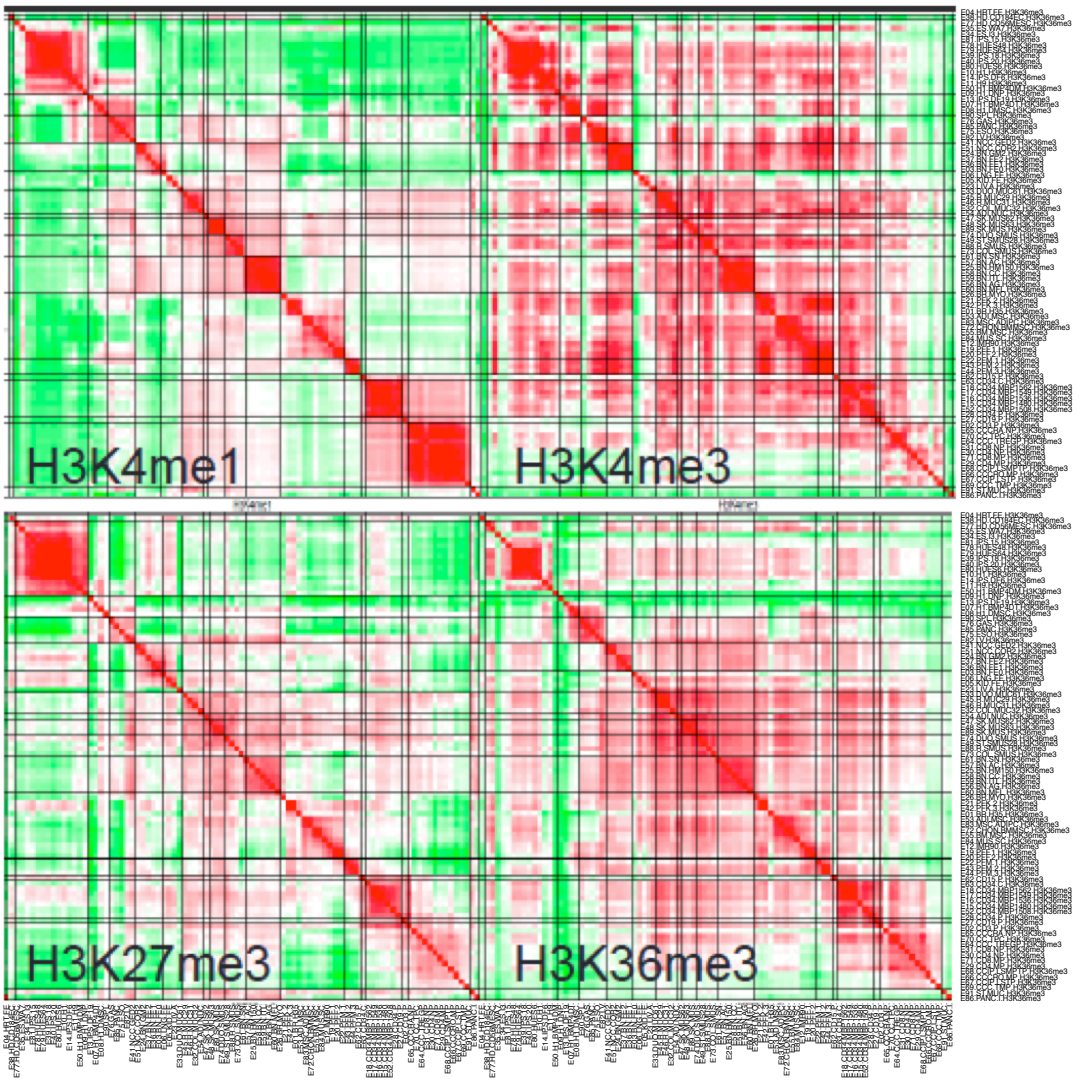

Figure 5-14: Similarity between epigenomes based on histone mark presence in enhancer regions reveals consistent sample clusters based on epigenomic data. Specifically, H3K4me1 seems to give the strongest correlations within clusters, followed by H3K27me3, H3K36me3, and lastly, H3K4me3, which has relatively high levels of similarity for many pairs of epigenomes.

Chapter 6

Identification of unknown covariates in epigenomic data with comparisons to known metadata

6.0.1 Introduction

As increasing amounts of epigenomic data are generated in less controlled environments, the importance of identifying and possibly correcting for covariates also increases. In fact, in Chapter 3, we included correction for known covariates in our methodology for this precise reason. To isolate the signal most relevant to a particular biological question or context, it is absolutely necessary to remove as many confounding factors as possible before proceeding with the analysis and making biological conclusions.

In Chapter 3, we took advantage of a rich set of metadata, including lab, donor, sample type, and sample group for each epigenome. With this information, we were able to explicitly correct for covariates in our samples using regression residuals before performing our epigenomic comparisons.

However, in many cases, a well-organized, structured, and consistent set of data may not be available. In fact, for the Epigenomics Roadmap data, we painstakingly constructed the final set of consistent and informative metadata information, coor-

dinating across many data centers and experimentalists. While grants, journals, scientists, and institutions are increasingly pushing for public release of datasets used, aggregation and standardization across these public resources will become increasingly difficult. Obtaining meaningful and consistent metadata will be one of the big challenges in these situations, possibly due to unresponsive authors, uncooperative labs, or perhaps simply poor documentation.

Of course, the discussion of "meaningful" metadata, also raises another important question: What are the most important covariates to correct for? The importance of some confounding factors are well document, such as batch effects (in any experiment)[181] and population structure (for genetic sequence data).[66,182] However, that still leaves the possibility of unknown or unidentified covariates playing an influential role in the data.

Luckily, a number of computational and mathematical approaches can be employed to identify potential covariates. One commonly used approach is to perform unsupervised dimensionality reduction to explore the data and identify possible covariates.[183–186] To take it a step further, the most important dimensions can then be treated as unkonwn covariates, with appropriate corrections applied. In particular, principal component analysis has been widely adopted for this purpose in a number of biological contexts, especially genome-wide association studies.[66] PCA has been shown to be very powerful, especially in correction for ethnic background of individuals, with the top principal components often even being more informative than self-reported ethnicity.[187]

PCA has also been applied to epigenomic data to capture the differences in data or variance resulting from profiling different histone marks.[188,189] However, to our knowledge, this approach has never been applied directly to chromatin state annotations, which captures the combinatorial behavior of histone marks. Furthermore, this approach of dimensionality reduction has never been applied to an epigenomic dataset of this size (with 127 samples) or with this breadth, as it spans sample type (primary cultures, primary cells, and primary tissues), anatomical groups, tissue type, and both adult and fetal samples. This unique dataset provides an unprecedented look

at how principal components capture covariates provided in the metadata, producing an automated way to identify covariates that contribute the most variance across the samples.

Here, we apply principal component analysis to our information theoretic representation based on chromatin state at gene bodies, as described in Chapter 3. We identify the amount of variance explained by the top principal components, identify the known covariates that top covariates correspond to, and identify the number of principal components that seem to most meaningfully contribute to data variance. In this way, we provide some insight into the feasibility of unsupervised covariate correction on epigenomic datasets and indicate which covariates can (and can not) be captured by top principal components.

6.1 Methods

6.1.1 Representation of each epigenome

To represent each epigenomic sample, we used the feature representation described in Section 3.2.3, where each epigenome is represented by chromatin state features. Specifically, our chromatin states are based on the core 15-state model from the integrative Epigenomics Roadmap analysis.[11] Meanwhile, our features utilize both the gene-body and regulatory region approach; in the gene body approach, chromatin state coverage at each gene body is directly used to calculate feature values, as shown in Figure 3-1. In our regulatory region approach, chromatins tate coverage at all linked regulatory regions for a gene is used to calculate feature values. As previously described, the result is a $N \times M$ matrix, where N is the number of epigenomes (in this case, 127), and M is the number of features (in this case, up to 299,025 features if every combination of 19,935 genes and 15 chromatin states is used).

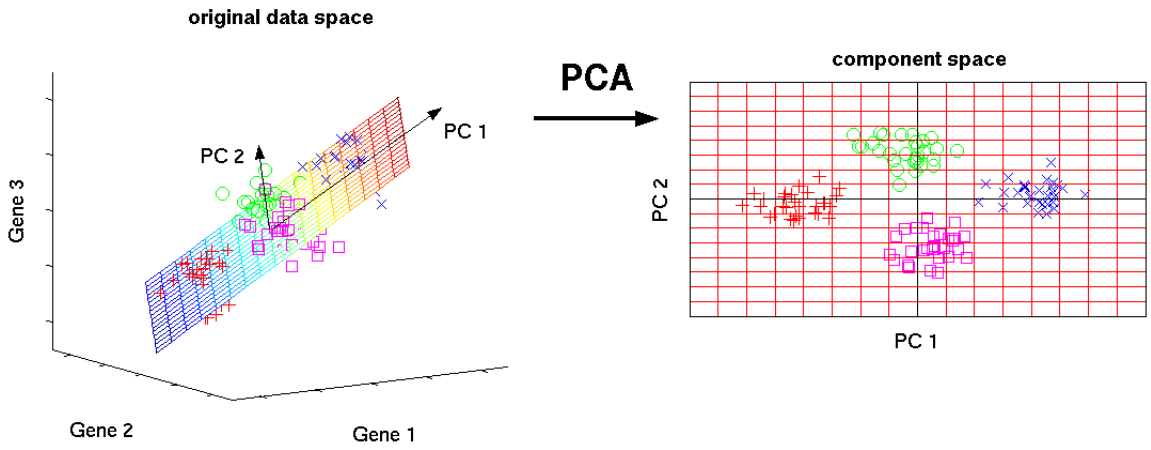

Figure 6-1: Principal component analysis can perform dimension reduction, for example, by projecting points in three-dimensional space onto a two-dimensional plane that captures nearly all the variance of the original dataset.[190]

6.1.2 Principal component analysis

As described above, principal component analysis is a dimensionality reduction approach that can reduce the M-dimensional space into a lower-dimensional space with minimal loss of information from a variance perspective. At a basic level, each principal component is a linear transformation of the original dimensions, such that the top principal components explain the most variance, and all principal components are orthogonal to one another. An illustrative example is shown in Figure 6-1, where data in 3-dimensional space is transformed into the 2-dimensional space of the top two principal components, while nearly preserving all pairwise distances between points. While this is not always possible, this visualization gives some intuition about the utility of Principal Component Analysis (PCA).

From a theoretical standpoint, there are two methods that can perform a principal component analysis, generating principal components that meet the required criteria of maximal variance and orthogonality. The two mathematical approaches are singular value decomposition (SVD) and eigenvalue decomposition of the original $N \times M$ matrix. Here, we employ the SVD approach, as implemented by the prcomp function in the stats package in R.[135]

Singular value decomposition of a matrix factors the original matrix A into a

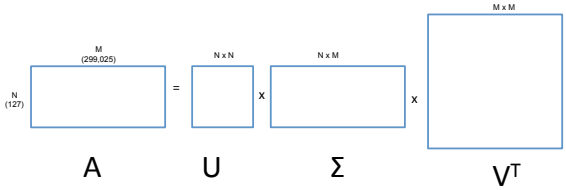

Figure 6-2: Illustration of singular value decomposition, which produces principal components as columns of V, and corresponding variances as the diagonal elements of Σ.

product of three matrices, U, Σ, and V^T, and it has been proven that a SVD exists for any matrix. Furthermore, U is a $N \times N$ unitary matrix, Σ is a diagonal $N \times M$ matrix with non-negative real numbers, and V^T is a $M \times M$ unitary orthogonal matrix. Intuitively, U and V^T can be viewed as rotation matrices, while Σ is a scaling matrix. The columns of V are called the "right singular vectors", and each column represents each principal component as the linear combination of features. Meanwhile, the elements on the diagonal of Σ are called the "singular values", where the i^{th} singular value quantifies the corresponding variance explained by the i^{th} singular vector. The "top" principal components are the principal components with the greatest variances, which are the same as the principal components with the highest singular values.

In addition, pre-processing of the data is important before performing PCA, as PCA is sensitive to transformations of the data such as centering and scaling. Here, we choose to both center (based on the mean) and scale (to unit variance) each feature in the data. Intuitively, centering our data moves our origin to the mean of the data, allowing the top principal component to explain the most variance relative to the center of the data. Without centering the data, the top principal component would be in the direction of the mean of the data due to the "variance" from the origin to the data. Meanwhile, scaling each feature to unit variance "weights" the importance of each feature equally. Intuitively, if your feature values are based on measurements, then changing the units of your measurement would drastically affect the the top principal components - for example, the variance between a measurement of 1000 (meters) and 2000 (meters) is much larger than the variance between a measurement of 1 (km) and 2 (km). In this way, features that have larger variance due to the

numerical measurements would be given a much stronger weight in determining the principal components (and explained variance).

For our situation, we chose to scale the data so that our analysis weights each gene and chromatin state combination equally, even if some of them have a smaller variance across samples and other have larger variance across samples. (This conservative choice parallels our choice to use the Mann-Whitney-Wilcoxon test in Chapter 3, which does not measure the magnitude of differences between datapoints, but only the relative ordering. However, scaling the data does not have as extreme of an effect, as the relative magnitudes of variance within one feature is preserved; only the overall variance of that feature is scaled to 1.) However, if one wanted to award more importance to genes whose chromatin state highly varies across samples, one could also choose to perform PCA on unscaled data in this application.

6.1.3 Mutual information between principal components and covariates

One of the goals of our principal component analysis is to use principal components to identify unknown covariates, especially in the absence of known covariates provided by metadata. In our case, we do have the actual known covariates, which allows us to further investigate and interpret our principal components. Specifically, in INSERT CITATION HERE OF CSUMI, the authors propose a method called Component Selection Using Mutual Information, where principal components are chosen based on their information theoretic mutual information with covariates. Mutual information is an information theory metric that has been applied to a wide variety of fields.

In our case, we can use this method to identify the covariates that are primarily responsible for our correlated with each principal component. For each pair of covariates and principal components, the mutual information between the two variables are calculated. Then that metric is normalized by the entropy of the covariates, to standardize for the fact that different covariates have different inherent variance properties.

Formally, the mutual information between covariate C and principal component P is calculated as:

MI(C, P)=Entropy(C)+ Entropy(P) - Entropy(C, P) Further, the entropy of a discrete variable X with possible values V is defined as:

$$\text{Entropy(X)} = \sum_{x \epsilon V} \text{p(x)} \log \text{p(x)}$$

and the entropy of two discrete variables X and Y with possible values V and U is defined as:

$$\text{Entropy(X, Y)} = \sum_{x \epsilon V} \sum_{y \epsilon U} \text{p(x, y)} \log \text{p(x, y)}$$

Further, the probability of X taking on each possible value x can be estimated from the data as the proportion of times in the dataset that x occurrs, and analogous calculations can be used for joint probabilities. Then, as proposed in CITATION, the normalized mutual information statistic for covariates is calculated as:

$$\text{MI}_C\text{(C, P)} = \frac{\text{MI(C, P)}}{\text{Entropy(C)}}$$

Since the covariates are all categorical discrete variables, we can apply these definitions directly. However, the principal components are continuous variables, so we discretize them for the entropy and mutual information calculations. Discretizing the principal component P is possible using a binning approach. That is, using b number of bins, we can divide the range of continuous values for P into b equal sized bins. Specifically, the size of each bin is:

$$\frac{\text{max(P)} - \text{min(P)}}{b}$$

Then for each data point (epigenome), the projected value for that epigenome onto a particular principal component is converted into a categorical value of bin_a, where a is the assigned bin from the discretization. Finally, the categorical values for each principal component can be used to calculate entropy and mutual information, based

on the discrete variable formulas above.

6.2 Results

6.2.1 Principal component analysis with gene body representation

6.2.1.1 Identification of clusters formed by top principal components

First, we identified principal components based on the gene body features, as described in Section 6.1.2, and tried to identify corresponding covariates in the metadata that these principal components corresponded to. As shown in Figure 6-3a, the first two principal components fairly strongly separate the primary blood cell samples from the rest of the tissues and celltypes. They also seem to separate the embryonic stem cells (ESC), induced pluripotent stem cells (iPSC), and cultures derived from embryonic stem cells (ESDERIV) from the rest of the samples. The separation of these three main clusters, shown in green, blue, and red, is even more clear when adding in the third principal component in a 3-dimensional visualization in Figure 6-3b.

6.2.1.2 Variance of top principal components

A common approach to deciding how many principal components to look at is to study how much variance each principal component captures. While we expect the variance explained by each subsequent principal component to decrease, the amount by which it decreases can be informative about which principal components are most important. A scree plot, as shown in Figure 6-4a, visualizes the variance captured by each principal component, and the "elbow" of the scree plot, where the slope of the line decreases, can indicate the most important principal components to look at. In this case, we observe an "elbow" around the third principal component, suggesting that the top three principal components, which we have already visualized and explained, are sufficient. However, this qualitative judgement also suggests another 2nd "elbow"

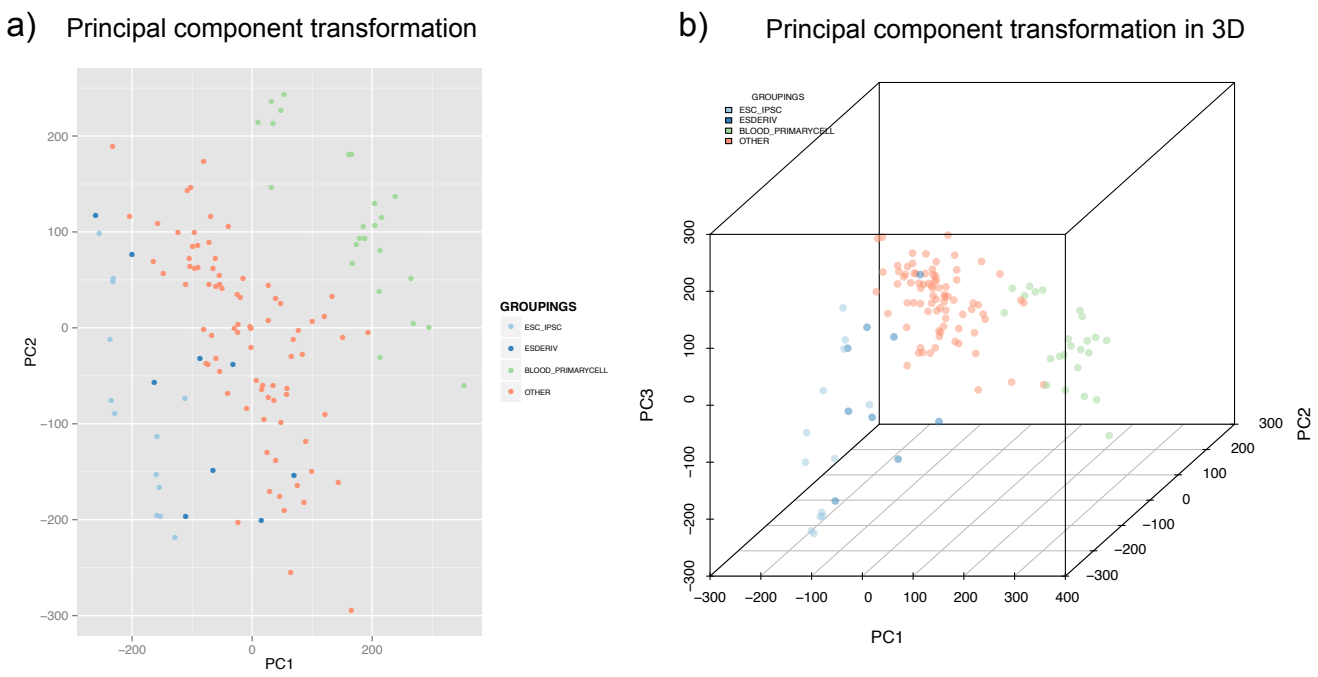

Figure 6-3: Top principal components of epigenomes based on gene body features separate out stem cells (ESC or iPSC) and ESC-derived cultures, shown in blue, and primary blood cells, shown in green. These clusters are already apparent based on the top two principal components (a), but the separation becomes even clearer when considering the three top principal components together (b).

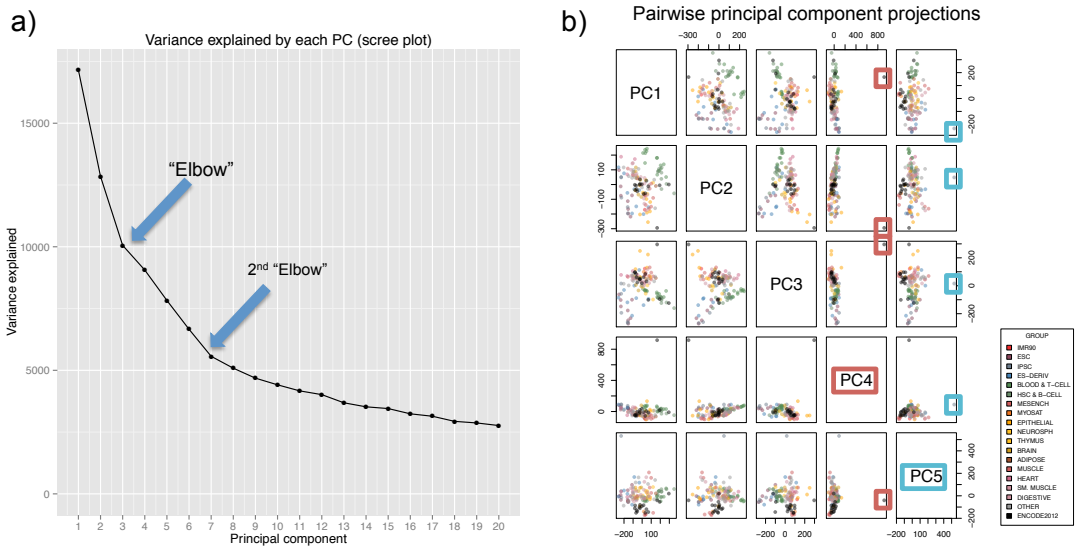

Figure 6-4: a) Scree plot of principal components, showing the variance captured by the top 20 principal components. The "elbow" of the scree plot, where the slope of the line sharply decreases, suggests a cutoff for the most imortant principal components to use. In this case, there appears to be a first elbow around the 3rd principal components, and a second elbow around 7 principal components. b) The 4th and 5th principal components separate one epigenome at a time.

around the 7th principal component.

6.2.1.3 Pairwise projections of top principal components

When we visualize the top five principal components in Figure 6-4b, we see support for focusing on the top three principal components. In particular, the 4th and 5th principal component each seem to isolate out one sample at a time, which is not very interesting from a biological perspective, nor is it very applicable for the discovery of unknown covariates, as it corresponds simply to a "covariate" that identifies that particular sample (i.e. a covariate that is set to "TRUE" for that sample and "FALSE" for all other samples).

However, it is worth noting that the first three principal components only cover about 20% of the total variance of the dataset. This makes sense, as there is clearly a lot more information and structure in our epigenomic data beyond only identifying

stem cell and blood samples, as we showed by comparing groups of epigenomes in Chapter 3.

6.2.1.4 Relationships between principal components and covariates using mutual information

Next, we use mutual information as described in Section 6.1.3 to identify relationships between principal components and covariates. In Figure 6-5a, this normalized mutual information metric is visualized as a heatmap, with the covariate categories reordered based on correlation values. The mutual information validates the previous observation that the first three principal components strongly separated the epigenomes by "Groupings", the covariate identifying 1) ESC and iPSC cells, 2) ESC derived cells, 3) primary blood cells, and 4) other epigenomes. Unsurprisingly, we also see that the "Groupings" category closely cluster with other covariates based on celltype and anatomical group, including "Anatomy", "SpecialGI" (anatomy with a Gastrointestinal grouping), "Group" (sample groups), and "Blood" (blood cells vs other cells). These five covariates have a strong mutual information relationship with the first three principal components, suggesting that the first three covariates would be a strong estimation for these covariate categories. However, some covariate categories seem to also be strongly tied to later principal components.

To visualize this, in Figure 6-5b-f, we identify the two principal components of the top 20 that have the most mutual information with each covariate, and project the epigenomes onto those two principal components, with the top PC on the x-axis, and the next best PC on the y-axis.

This also shows that some of the covariates are related, as identified by the same principal components being shown. Specificaly, both the group (Figure 6-5b) and sample state (Figure 6-5c) choose PC 3 as the strongest connection, with PC 1 as the next best component. In both cases, the approximate clusters for each category can be seen. This makes sense, as certain "group" and "state" categories correspond to each other. For example, the "liquid" state samples are overwhelmingly blood samples, the "solid" state samples are mostly tissue samples, and the samples with

state "NA" are mostly cell lines or cultured cells. This visualizes how the same principal components can be relevant for multiple covariates, due to mutual information between the covariates themselves.

In Figure 6-5d, we show that the top projection for the age covariate, principal components 6 and 3, separates adult samples from fetal samples, with some separation of other (cultured) samples. This suggests that, even though the top three principal components clearly have strong relationships with many covariates, later components can also be important for certain covariates. Again, the same components are chosen for another covariate, sample type, (although in flipped order), as shown in Figure 6-5e; again this can be explained by the overlapping categories between sample type and developmental age, as Adult samples are largely blood samples, the tissue samples are mostly fetal, and cell lines are classifed as "Other" for age.

Lastly, in Figure 6-5f, the principal components 2 and 1 shared the most information with the covariate for lab, with the projections showing clusters for UW, UCSF-UBC, and UCSD. On the other hand, there does not seem to be a tight cluster for Broad Institute (BI) samples, which are spread throughout the projection. Interestingly, this projectionj also seems to be consistent for some mied samples - for example, one green mixed sample is a mixture of samples from BI, UCSD, and UCSF-UBC, and it lies between the UCSD and UCSF-UBC cluster, with many BI samples nearby. However, this is certainly not true for all mixed samples - one of the samples mixed between UCSF-UBC and BI (in purple) strongly falls in the UCSD cluster. Two other (on the far right) fall closest to the UCSD cluster, although based on PC1 only, they group with many UCSF-UBC samples.

The mutual information between principal components and known covariates certainly suggests that using the top principal components could be a good estimate for unknown covariates. Additionally, the scree plot in Figure 6-4a suggested using the top 3 or 7 principal components, and many of the known covariates were most strongly associated the top 3 principal components; overall, most known covariates shared the most mutual information with principal components ranked in the top 7.

6.2.1.5 Identification of unknown covariates based on principal components

Some principal components did not share much information with known covariates, such as principal component 4. This suggests that perhaps PC4 captures an unknown covariate that we should be correcting for, with the possibility that correcting the data for PC4 before running ChromDiff would further isolate the signal we are testing for, with improved results.

However, this brings about the biggest problem with identifying "unknown" covariates, which is not knowing what property is being corrected for. With group-wise comparisons like ChromDiff, the groups are usually based on some property or covariate. For example, we compared liquid samples (blood primary cell) to solid samples (tissue samples), in Section 4.3.5. Because we had known covariates, we could choose to correct only for the properties that were not immediately relevant. For example, we usually corrected the data based on sample state (liquid or solid) as described in Section 3.2.5, but when comparing liquid and solid samples, we obviously did not include that in the covariate correction step. This is still not a perfect approach, as some covariates correlate strongly with one another, as we have shown. For example, we also correct for sample type (primary cell, primary tissue, and so on), which strongly correlates with sample state (most liquid samples are primary cells, most solid samples are primary tissues). Nevertheless, with known covariates, we are able to explicitly choose trade-offs between correcting for covariates to isolate the relevant signal and omitting covariates so that we don't "correct away" the important features.

Unfortunately, with unknown covariates, we are unable to interpret the meaning of each principal component. Therefore, the trade-off becomes more difficult. Should we choose not to correct for any principal components, thereby leaving the possibility of strong confounding factors in our data, leading to misleading results? Or shoudl we conservatively correct for the top principal components, knowing that we may be removing the signal that we most care about, rendering our analysis unfruitful?

We hope that our analysis of mutual information between principal components

and known covariates is helpful for this tough decision, as it will allow future researchers to estimate the covariates that the top principal components represent in similar datasets. While our exact conclusions are, of course, only perfectly accurate for the Epigenomic Roadmap dataset, we expect that similar results might be found in future epigenomic dataets represented with features based on chromatin state coverage in gene bodies. Next, we will expand our analysis and results to features based on linked regulatory regions, rather than gene bodies.

6.2.2 Linked regulatory region analysis

Just as we presented a gene-body approach and regulatory region approach for ChromDiff comparisons, we also applied our principal component analysis to our features based on linked regulatory regions. We found fairly similar, but not identical, results, and we hope that that results will be generalizable to future analyses based on linked regulatory regions.

6.2.2.1 Principal component analysis based on linked enhancers

When we used features based on chromatin state coverage at linked enhancers, the top principal components separated out the same main three clusters as before, as shown in Figure 6-6a. Specifically, the top three principal components, shown in panel a, strongly separate the cluster of primary blood cells, and the cluster of ESC's, iPSC's and ESC-derived cultures, from the other samples.

The pair-wise plots of the top 5 principal components further clarifies the main distinction of each component. As shown in Figure 6-6b, PC2 mostly separates out the primary blood cells, while PC3 mostly separates the stem cells (ESCs and iPSCs), from the ESC-derived cultures and other samples. However, the purpose of PC1 is less clear - even though the first two principal components do show the three cluster groupings, it largely seems to be due to PC2, where the projection of the data onto PC2 would still retain the main cluster ordering of blood cells (green), then other cells (red), then ESC/iPSC/ESC-derived cultures (blue). The most distinguishing

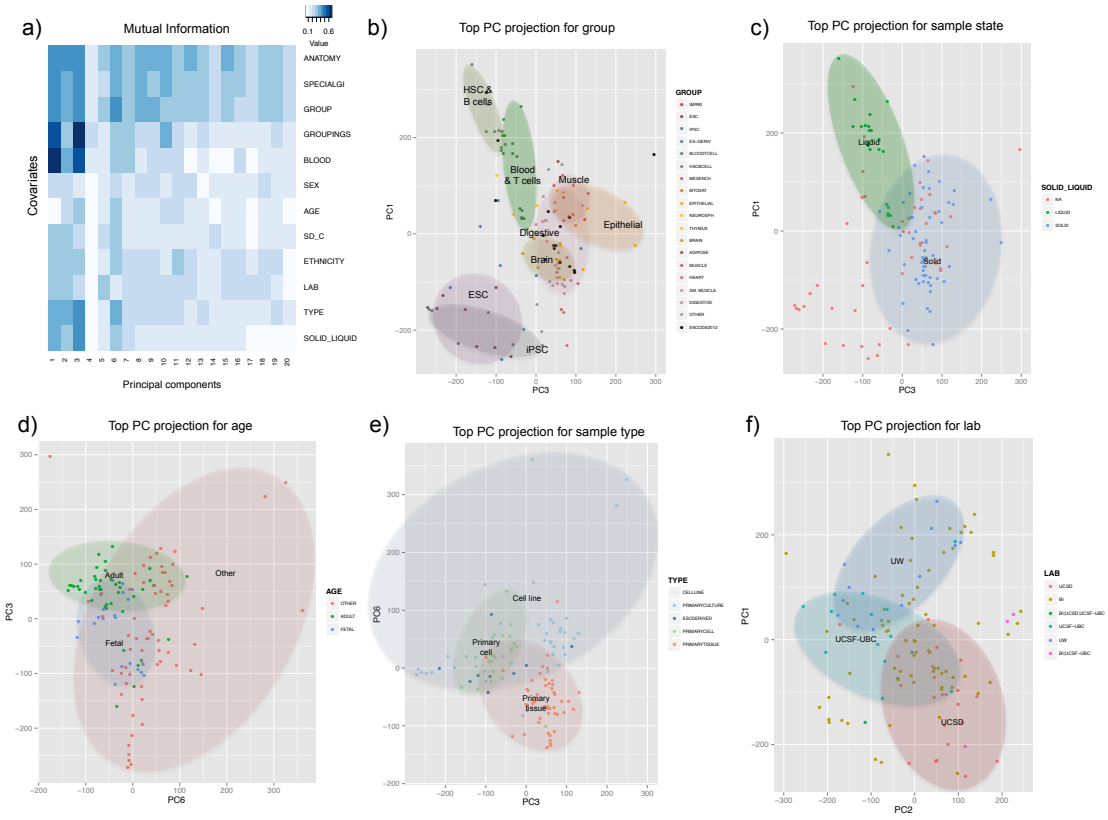

Figure 6-5: Mutual information reveals structure between principal components and covariates. a) Normalized mutual information between principal components and covariates reveals groups of covariates with similar information and which principal components strongly associate with each covariate. b-f) Epigenomes projected onto top two principal compoents for each covariate, with the top PC on the x-axis and the next PC on the y-axis. Specifically, principal components were chosen based on b) group, c) sample state, d) developmental age, e), sample type, and f) experimental lab.

separation that comes solely from the top principal component seems to be foreskin fibroblast primary cells, shown in golden yellow (Figure 6-6b).

While the pairwise plots clarified some of the distinctions made by the top three principal components, the purpose of fourth and fifth principal components was less clear. To some degree, PC4 seems to again separate out the fibroblast cells, while PC5 seems to separate the fibroblast cells with a stem cell sample (Figure 6-6b). While these principal components were not as obviously singling out a single sample as the previous example (Figure 6-4b), their purpose was not entirely clear, especially from a covariate perspective.

When we looked at the scree plot, there was not an obvious "elbow" in the plot, although it seemed like it could have been after the fifth or twelfth principal component, as shown in Figure 6-6c. Again, the proportion of the total variance explained by these top components was still limited, as the top three principal components explain less than 10% of the total variance.

Next, we visualized the mutual information between the principal covariates and the covariates (Figure 6-6d), finding that the second PC strongly correlated with blood samples, while the first three PCs were also associated with anatomy groups. In panel e, we see the top two principal components based on the lab covariate are PC4 and PC1, with clusters emerging for the main four labs. In panel f, we project the epigenomes onto PC2 and PC6, the top covariates for group, and we see the blood clusters close to one another on the left, with many cultured cell samples on the right (ESC, iPSC, ES-derived), with most of the tissue samples somewhere in the middle.

6.2.2.2 Principal component analysis based on DNase hypersensitive sites

When applying the same analysis to DNase I hypersensitive sites, we find that the top three principal components perform much of the same separation of the three main clusters based on blood primary cells, stem cell and ESC-derived cultures, and other samples (Figure 6-7a). Specifically, the pairwise plots in Figure 6-7b shows that the second principal component cleanly separates blood samples from ESC-derived and stem cell samples, while the third principal component separates the

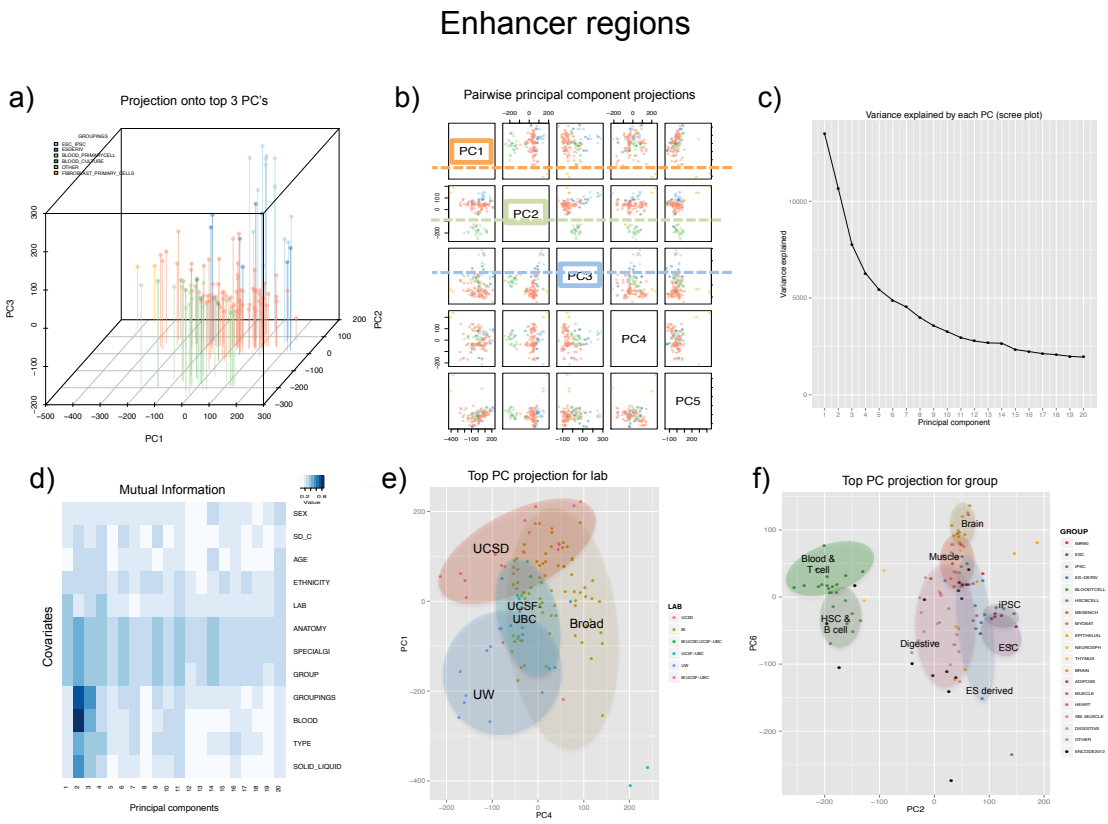

Figure 6-6: Principal component analysis applied to epigenomes based on enhancer features shows that a) the top three principal components separate out blood samples (green), stem cell and ESC-derived samples (blue), and fibroblast samples (yellow). b) Pairwise principal component plots emphasize the main distinction of the top three principal components. c) Based on variance explained by each PC, the top five principal components could provide a reasonable interpretation of the data. d) Mutal information between covariates and principal components allows us to pick the top two principal components for projectsion based on the covariates of e) lab and f) sample group.

rest of the samples from the two aforementioned groups. As shown in Figure 6-7c, the top principal components explain much more variance in the data than the previous examples, with the top three principal components explaining about 40% of the variance in the data-set. Considering that there are in total 120 principal components, the high proportion of variance explained by only three components suggest that the principal components are able to more efficiently capture the variance in this dataset, compared to the chromatin state data based on gene bodies and enhancer.

Mutual information analysis (Figure 6-7d) confirms the relevance of the top principal components for many of the known covariates. As shown in Figure 6-7e, the top principal components based on experimental lab is with PC4 and PC1, which roughly clusters the data based on the four main labs. Additionally, for the covariate that identifies whether each sample was a single donor, composite donor, or cultured sample (NA), principal components 2 and 18 share the most mutual information and approximately separate the epigenomes (Figure 6-7f).

6.2.2.3 Principal component analysis based on promoter state

When applying PCA to data based on promoter regions, the first two principal component seems to separate the same three clusters of blood samples, stem cell and ESC-derived cultures, and other samples (Figure 6-8a). Notably, most of the separation seems to be driven by the top principal component, while the second and third principal component mainly separate out a HepG2 sample, as shown in Figure 6-8a-b. Based on the amount of variance explained shown in Figure 6-8c, we find that the amount of variance explained by each principal component drops off around the third and eighth component. As shown in Figure 6-8c, the top three principal components only explain about 15% of the variance of the data.

Using mutual information, we find that the first and 4th component strongly correlate with tissue type-based covariates, with some fairly related principal components for lab as well (Figure 6-8e). Specifically, in Figure 6-8f, we see that UCSF-UBC, UCSD, and UW cluster fairly well based on a projection to principal component 2

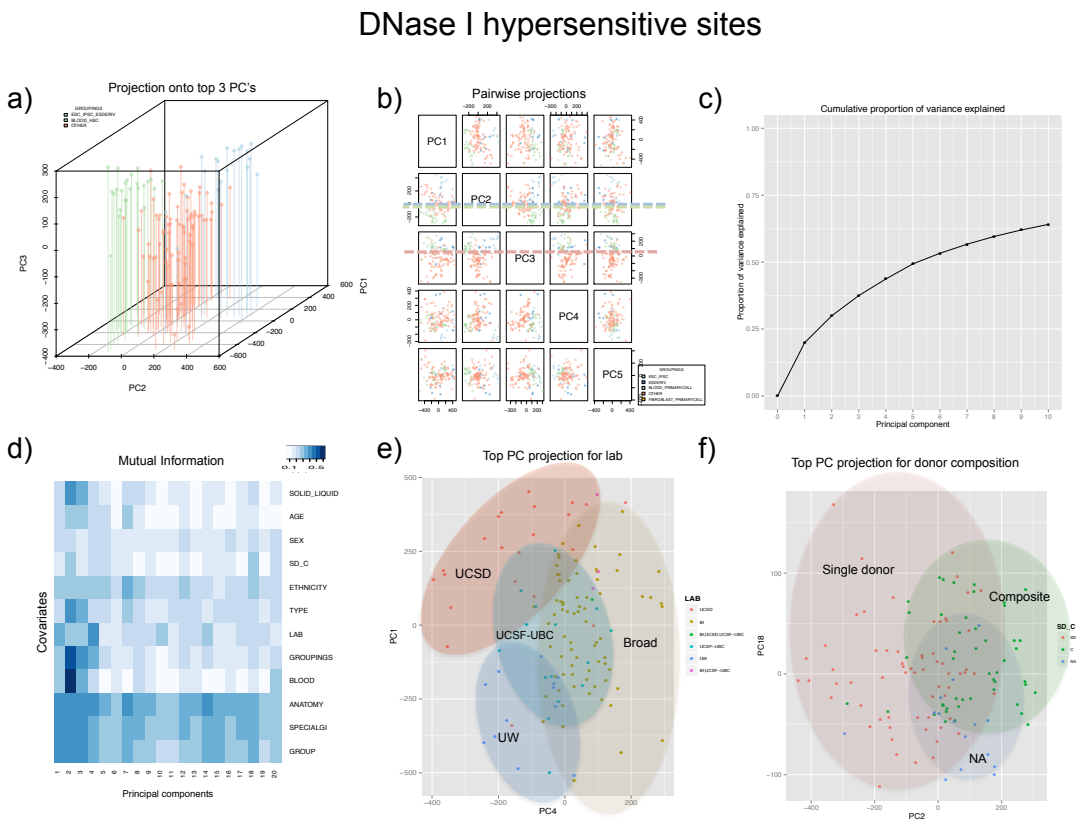

Figure 6-7: Principal component analysis in DNase hypersensitive sites shows that the top three principal components primarily separate the samples based on the same three clusters as before, based on a) a 3D projection of the data and b) pairwise principal component plots. c) However, the top 3 PCs also explain a higher proportion of the total variance in the data than in the gene body and enhancer case. d) Mutual information between covariates and principal components confirms the importance of PC2 and PC3 for clusters bsed on groupings (shown in panel a). Mutual information also quantifies the most associated principal components for e) samples based on lab (PC4 and PC1) and f) samples based on anatomy group (PC2 and PC18).

and 12. Interestingly, this projection suggests that the Broad Institute samples seems to be fairly well mixed with samples from UCSF and UCSD-UBC.

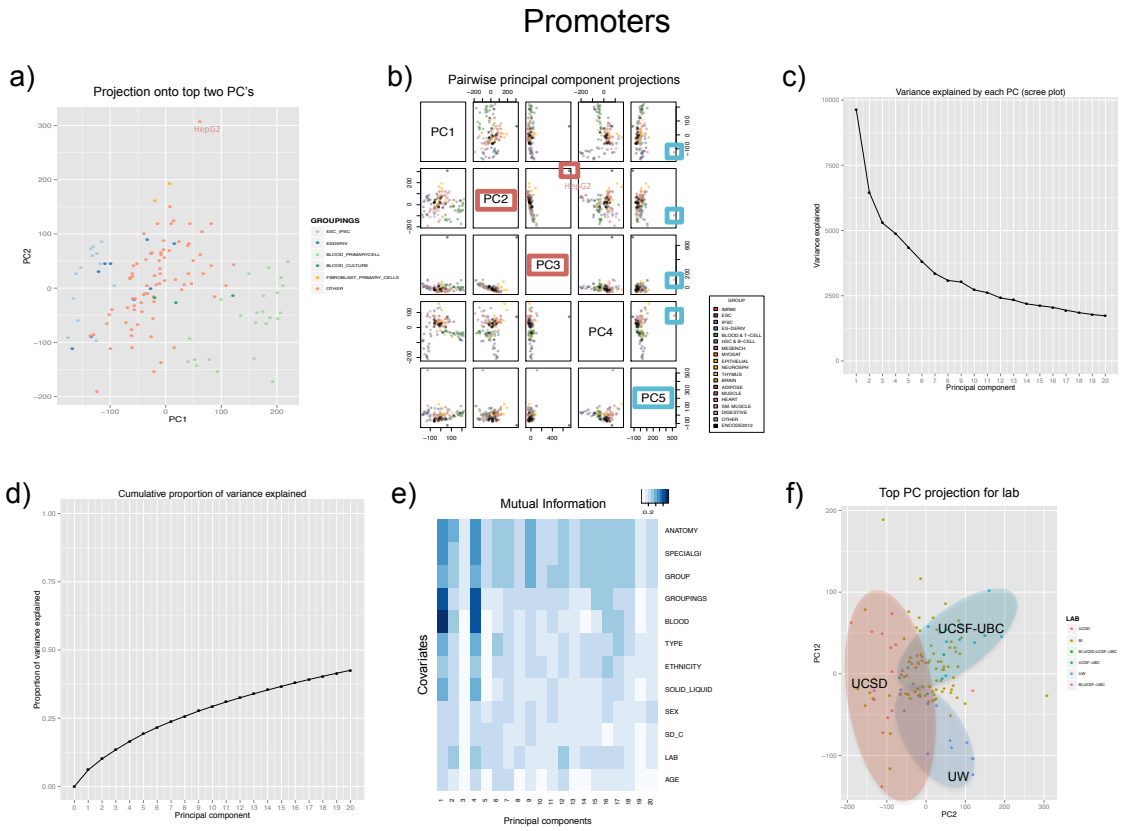

Figure 6-8: Principal component analysis based on promoter regions shows that a) the top principal component primarily separate the samples based on the same three clusters as before, b) while the second, third, and fifth principal components primarily isolate one sample at a time. c) The amount of variance explained by each principal component tapers off around principal component 3 and 8. d) Though the top three principal components separate the samples in meaningful ways, they explain a minority of the total variance in the data, about 15%. e) Principal components 1 and 4 share relatively high mutual information with tissue type and cell type covariates, while PC 2 and 12 are the strongest PCs for the lab covariate. f) We find that projection of the samples to PC 2 and 12 separates samples based on whether they originated from UCSD, UCSF-UBC, or UW, with Broad Institute (BI) samples relatively dispersed.

Chapter 7

Conclusion

In this thesis, we have utilized computational methods, algorithms, and approaches to better understand, interpret, and analyze biological data. Specifically, we have focused on the field of epigenomics, but our work has implications on gene regulation, cell type differences, and regulatory regions.

With our interdisciplinary work, our contributions span three categories: mechanisms of gene regulation, development of comparative methods, and identification of broad epigenetic patterns. Biologically, our results have provided mechanistic insights into the monoallelic and monogenic regulation of olfactory receptor genes, as well as how genes escape from X chromosome inactivation. Methodologically, we developed ChromDiff and Regulatory ChromDiff, which are novel methods for systematic identification of epigenomic differences between cell types or sample groups; our software is open-source, free, and publicly available from `http://compbio.mit.edu/ChromDiff` and `http://www.github.com/angieyen/ChromDiff`, for the benefit of the general public and future research. Additionally, we have identified general epigenetic patterns, including the celltype-specific DNA methylation of chromatin state regions, the genes with distinct epigenetic signatures based on sex, tissue type, cell type, and developmental age, and the relationship between known covariates and top principal components across 127 epigenomic samples.

7.1 Summary of results

First, we studied the monoallelic and monogenic expression of olfactory receptor genes in mice by studying their epigenetic state. We found a strong signal of H3K9me3 and H4K20me3 that was specific to olfactory receptor genes in olfactory tissue genome-wide. Using peak calling, clustering, hidden markov models, chromosome-wide visualizations, and statistics, we computationally verified that the sensitivity and specificity of this epigenetic signature genome-wide. Our results, along with additional experimental follow-up from our collaborators, supported a new model for olfactory gene regulation, where the entire family of olfactory receptor genes are epigenetically and heterochromatically repressed at an early developmental stage, prior to olfactory expression. Our model proposes that at a later stage, one allele of one olfactory gene is "de-repressed" and the repressive mark H3K9me3 is replaced with the activating mark H3K9me2. Our work provides a novel example of epigenomic regulation of genes that may prove to be generally applicable to other contexts, as well as revealing specific discoveries pertaining to the olfactory system.

Next, we developed ChromDiff, a novel computational method to identify epigenetic differences between groups of samples at the genome-wide scale. Our method combines biologically relevant information, such as genes and chromatin states, with computational techniques drawn from the fields of information theory, machine learning, and statistics. Our method uses a novel information theoretic representation of our epigenomic information, which allows for a type of dimensionality reduction on our large-scale data while retaining important information for relevant research questions. It also supports multiple statistic tests and hypothesis correction approaches to support a variety of needs, and it notably corrects for covariate effects, which will be increasingly relevant as integration of unmatched datasets becomes increasingly important. In addition to identifying genes and chromatin states that distinguish between the two groups, our pipeline also provides follow-up analyses, such as gene clustering, gene set enrichment, and differential gene expression. We also extended ChromDiff to Regulatory ChromDiff to support integration of proximal and distal

regulatory regions. ChromDiff software is free, open-source, and available for download to the general public under the GNU General Public License v3.0 at `http://compbio.mit.edu/ChromDiff` and `http://www.github.com/angieyen/ChromDiff`.

We then applied ChromDiff and Regulatory ChromDiff to 127 samples from the Epigenome Roadmap project, grouping them based on sex, developmental age, tissue type, sample type, and cell type. We also looked for epigenomic differences at gene bodies, as well as promoters, enhancers, and DNase Hypersensitive sites linked to genes. We found that our methods identified genes relevant to each comparison, such as the identification of X chromosome genes for the sex-based comparison, as well genes relating to neuronal development for the comparison of brain and digestive samples. The identified distinguishing genes were enriched for differential expression between the groups, but the majority of them did not show expression differences, suggesting the importance of epigenomic comparisons as a complementary approach to gene expression.

Another one of my main research projects was the integrative analysis of the Epigenome Roadmap data, which included histone modification, DNA methylation, chromatin accessibility, and gene expression data across 127 human primary cells and tissues. I found distinct DNA methylation pattern across chromatin states, and also studied those patterns in a celltype-specific way, finding differences between cultured samples and tissue samples, for example. We also compared data from different DNA methylation technologies to quantify biases, and we clustered samples based on epigenetic data, finding strong concordance between our clustering order and celltype. Our integrative analysis presents a more complete and unified picture of how epigenetic state varies across celltypes, marks, and platforms.

Lastly, we identified the top principal components of our epigenomic samples based on our ChromDiff feature representation. Generally, we found that the top principal components grouped the samples into a cluster of cultured samples (ESCs, iPSCs, and cells derived from ESCs), a cluster of primary blood samples (T cells, B cells, and HSCs), and a third cluster made up mostly of solid tissue samples. We also used component selection using mutual information (CSUMI) to find relationships

between known covariates and top principal components. Generally, we found that the strongest connections to the top principal components were, again, based on the anatomical grouping of the sample, but that other known covariates such as lab and developmental age were also approximated with high-ranked principal components.

7.2 Future Work

There are many future directions for the research presented in this thesis. In chapter 2, we presented a new mechanism for epigenetic gene regulation and a model for olfactory gene regulation. This work has been extended to understand how the spatial location and organization of olfactory genes contributes to olfactory regulation**??**, and we expect future work in this area to continually build a more detailed mechanistic model for gene regulation. In chapter 3 and 4, we presented a method for epigenomic comparisons, with applications to groups based on a variety of biological attributes. We expect that, as more data is generated in coming years, our general method will be broadly applicable in a variety of biological contexts, and we also hope that future computational biologists build upon our work with various representations, statistical approaches, and improved chromatin state segmentations. In chapter 5, we produced the largest epigenomic reference to date, and we expect future research to expand this reference with additional biological samples. With this reference, future researchers can use epigenomic and regulatory information to better understand the functional impact of various genomic regions, such as identifying a superenhancer that links the FTO gene and obesity**??**. Finally, in chapter 6, we begin to explore the identification of unknown covariates in epigenomic data; while our discoveries begin to quantify the explanatory power of top principal components, more work will need to be done to find the explanation for principal components that do not strongly correlate with known covariates.

With the biological discoveries, novel algorithms, and rich datasets that have been introduced in the past few years, computational epigenomics has proven itself to be an area ripe for research. Moving forward, we expect the research questions presented

in this thesis to be expanded upon and improved. Additionally, we expect that our results lay a foundation for follow-up research, especially in the area of experimental validation. With new experimental technologies ranging from reporter assays?? to epigenome editing??, the potential to validate our hypotheses in vivo is greater than ever, elucidating the path for computational epigenomics to have an impact on human health and disease treatment in coming years.

Bibliography

[1] OpenStax College. Illustration from anatomy and physiology, connexions web site., 2013.

[2] National Institutes of Health. Epigenetic mechanisms are affected by several factors and processes including development in utero and in childhood, environmental chemicals, drugs and pharmaceuticals, aging, and diet. dna methylation is what occurs when methyl groups, an epigenetic factor found in some dietary sources, can tag dna and activate or repress genes. histones are proteins around which dna can wind for compaction and gene regulation. histone modification occurs when the binding of epigenetic factors to histone âĂIJtailsâĂİ alters the extent to which dna is wrapped around histones and the availability of genes in the dna to be activated. all of these factors and processes can have an effect on peopleâĂŹs health and influence their health possibly resulting in cancer, autoimmune disease, mental disorders, or diabetes among other illnesses. national institutes of health, 2005.

[3] NHGRI Darryl Leja. Dna transcription, 2010.

[4] Boumphreyfr. Illustration of trna building peptide chain, 2009.

[5] National Institutes of Health. Codon table, 2009.

[6] Darekk2. Nucleosome organization schema, 2012.

[7] Tdunning. Hidden markov model with output, 2012.

[8] Goncalo R. Abecasis, Adam Auton, Lisa D. Brooks, Mark A. DePristo, Richard M. Durbin, Robert E. Handsaker, Hyun Min Kang, Gabor T. Marth, and Gil A. McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[9] Kerstin Lindblad-Toh, Manuel Garber, Or Zuk, Michael F. Lin, Brian J. Parker, Stefan Washietl, Pouya Kheradpour, Jason Ernst, Gregory Jordan, Evan Mauceli, Lucas D. Ward, Craig B. Lowe, Alisha K. Holloway, Michele Clamp, Sante Gnerre, Jessica Alföldi, Kathryn Beal, Jean Chang, Hiram Clawson, James Cuff, Federica Di Palma, Stephen Fitzgerald, Paul Flicek, Mitchell Guttman, Melissa J. Hubisz, David B. Jaffe, Irwin Jungreis, W. James Kent, Dennis Kostka, Marcia Lara, Andre L. Martins, Tim Massingham, Ida Moltke,

Brian J. Raney, Matthew D. Rasmussen, Jim Robinson, Alexander Stark, Albert J. Vilella, Jiayu Wen, Xiaohui Xie, Michael C. Zody, Jen Baldwin, Toby Bloom, Chee Whye Chin, Dave Heiman, Robert Nicol, Chad Nusbaum, Sarah Young, Jane Wilkinson, Kim C. Worley, Christie L. Kovar, Donna M. Muzny, Richard A. Gibbs, Andrew Cree, Huyen H. Dihn, Gerald Fowler, Shalili Jhangiani, Vandita Joshi, Sandra Lee, Lora R. Lewis, Lynne V. Nazareth, Geoffrey Okwuonu, Jireh Santibanez, Wesley C. Warren, Elaine R. Mardis, George M. Weinstock, Richard K. Wilson, Kim Delehaunty, David Dooling, Catrina Fronik, Lucinda Fulton, Bob Fulton, Tina Graves, Patrick Minx, Erica Sodergren, Ewan Birney, Elliott H. Margulies, Javier Herrero, Eric D. Green, David Haussler, Adam Siepel, Nick Goldman, Katherine S. Pollard, Jakob S. Pedersen, Eric S. Lander, and Manolis Kellis. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476–82, 2011.

[10] Joel N. Hirschhorn and Mark J. Daly. Genome-wide association studies for common diseases and complex traits. *Nature reviews. Genetics*, 6(2):95–108, 2005.

[11] Consortium Roadmap Epigenomics, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y. C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K. H. Farh, S. Feizi, R. Karlic, A. R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L. H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, and M. Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–30, 2015. Kundaje, Anshul Meuleman, Wouter Ernst, Jason Bilenky, Misha Yen, Angela Heravi-Moussavi, Alireza Kheradpour, Pouya Zhang, Zhizhuo Wang, Jianrong Ziller, Michael J Amin, Viren Whitaker, John W Schultz, Matthew D Ward, Lucas D Sarkar, Abhishek Quon, Gerald Sandstrom, Richard S Eaton, Matthew L Wu, Yi-Chieh Pfenning, Andreas R Wang, Xinchen Claussnitzer, Melina Liu, Yaping Coarfa, Cristian Harris, R Alan Shoresh, Noam Epstein, Charles B Gjoneska, Elizabeta Leung, Danny Xie, Wei Hawkins, R David Lister, Ryan Hong, Chibo Gascard, Philippe Mungall, Andrew J Moore, Richard Chuah, Eric Tam, Angela Canfield, Theresa K Hansen, R Scott Kaul, Rajinder Sabo, Peter J Bansal, Mukul S

Carles, Annaick Dixon, Jesse R Farh, Kai-How Feizi, Soheil Karlic, Rosa Kim, Ah-Ram Kulkarni, Ashwinikumar Li, Daofeng Lowdon, Rebecca Elliott, GiNell Mercer, Tim R Neph, Shane J Onuchic, Vitor Polak, Paz Rajagopal, Nisha Ray, Pradipta Sallari, Richard C Siebenthall, Kyle T Sinnott-Armstrong, Nicholas A Stevens, Michael Thurman, Robert E Wu, Jie Zhang, Bo Zhou, Xin Beaudet, Arthur E Boyer, Laurie A De Jager, Philip L Farnham, Peggy J Fisher, Susan J Haussler, David Jones, Steven J M Li, Wei Marra, Marco A McManus, Michael T Sunyaev, Shamil Thomson, James A Tlsty, Thea D Tsai, Li-Huei Wang, Wei Waterland, Robert A Zhang, Michael Q Chadwick, Lisa H Bernstein, Bradley E Costello, Joseph F Ecker, Joseph R Hirst, Martin Meissner, Alexander Milosavljevic, Aleksandar Ren, Bing Stamatoyannopoulos, John A Wang, Ting Kellis, Manolis eng 5R24HD000836/HD/NICHD NIH HHS/ ES017166/ES/NIEHS NIH HHS/ F32 HL110473/HL/NHLBI NIH HHS/ F32HL110473/HL/NHLBI NIH HHS/ K99 HL119617/HL/NHLBI NIH HHS/ K99HL119617/HL/NHLBI NIH HHS/ P30AG10161/AG/NIA NIH HHS/ R01 AG015819/AG/NIA NIH HHS/ R01 ES024984/ES/NIEHS NIH HHS/ R01 ES024992/ES/NIEHS NIH HHS/ R01 HG004037/HG/NHGRI NIH HHS/ R01 HG007175/HG/NHGRI NIH HHS/ R01 HG007354/HG/NHGRI NIH HHS/ R01AG15819/AG/NIA NIH HHS/ R01AG17917/AG/NIA NIH HHS/ R01HG004037/HG/NHGRI NIH HHS/ R01HG004037-S1/HG/NHGRI NIH HHS/ R01NS078839/NS/NINDS NIH HHS/ RC1HG005334/HG/NHGRI NIH HHS/ RF1 AG015819/AG/NIA NIH HHS/ T32 GM007266/GM/NIGMS NIH HHS/ U01 ES017154/ES/NIEHS NIH HHS/ U01AG46152/AG/NIA NIH HHS/ U01DA025956/DA/NIDA NIH HHS/ U01ES017154/ES/NIEHS NIH HHS/ U01ES017155/ES/NIEHS NIH HHS/ U01ES017156/ES/NIEHS NIH HHS/ U01ES017166/ES/NIEHS NIH HHS/ Howard Hughes Medical Institute/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. England 2015/02/20 06:00 Nature. 2015 Feb 19;518(7539):317-30. doi: 10.1038/nature14248.

[12] M. J. Ziller, H. Gu, F. Muller, J. Donaghey, L. T. Tsai, O. Kohlbacher, P. L. De Jager, E. D. Rosen, D. A. Bennett, B. E. Bernstein, A. Gnirke, and A. Meissner. Charting a dynamic dna methylation landscape of the human genome. *Nature*, 500(7463):477–81, 2013. Ziller, Michael J Gu, Hongcang Muller, Fabian Donaghey, Julie Tsai, Linus T-Y Kohlbacher, Oliver De Jager, Philip L Rosen, Evan D Bennett, David A Bernstein, Bradley E Gnirke, Andreas Meissner, Alexander eng ES017690/ES/NIEHS NIH HHS/ P01 GM099117/GM/NIGMS NIH HHS/ P01GM099117/GM/NIGMS NIH HHS/ P30 AG010161/AG/NIA NIH HHS/ P30AG10161/AG/NIA NIH HHS/ R01 AG017917/AG/NIA NIH HHS/ R01AG15819/AG/NIA NIH HHS/ R01AG17917/AG/NIA NIH HHS/ R01AG36042/AG/NIA NIH HHS/ U01 ES017155/ES/NIEHS NIH HHS/ U01ES017155/ES/NIEHS NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't England 2013/08/09 06:00 Nature. 2013 Aug 22;500(7463):477-81. doi: 10.1038/nature12433. Epub 2013 Aug 7.

[13] Wei Xie, Matthew D. Schultz, Ryan Lister, Zhonggang Hou, Nisha Rajagopal, Pradipta Ray, John W. Whitaker, Shulan Tian, R. David Hawkins, Danny Leung, Hongbo Yang, Tao Wang, Ah Young Lee, Scott A. Swanson, Jiuchun Zhang, Yun Zhu, Audrey Kim, Joseph R. Nery, Mark A. Urich, Samantha Kuan, Chia-an Yen, Sarit Klugman, Pengzhi Yu, Kran Suknuntha, Nicholas E. Propson, Huaming Chen, Lee E. Edsall, Ulrich Wagner, Yan Li, Zhen Ye, Ashwinikumar Kulkarni, Zhenyu Xuan, Wen-Yu Chung, Neil C. Chi, Jessica E. Antosiewicz-Bourget, Igor Slukvin, Ron Stewart, Michael Q. Zhang, Wei Wang, James A. Thomson, Joseph R. Ecker, and Bing Ren. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*, 153(5):1134–48, 2013.

[14] Alexander Meissner, Tarjei S. Mikkelsen, Hongcang Gu, Marius Wernig, Jacob Hanna, Andrey Sivachenko, Xiaolan Zhang, Bradley E. Bernstein, Chad Nusbaum, David B. Jaffe, Andreas Gnirke, Rudolf Jaenisch, and Eric S. Lander. Genome-scale dna methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–70, 2008.

[15] Casey A. Gifford, Michael J. Ziller, Hongcang Gu, Cole Trapnell, Julie Donaghey, Alexander Tsankov, Alex K. Shalek, David R. Kelley, Alexander A. Shishkin, Robbyn Issner, Xiaolan Zhang, Michael Coyne, Jennifer L. Fostel, Laurie Holmes, Jim Meldrim, Mitchell Guttman, Charles Epstein, Hongkun Park, Oliver Kohlbacher, John Rinn, Andreas Gnirke, Eric S. Lander, Bradley E. Bernstein, and Alexander Meissner. Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. *Cell*, 153(5):1149–63, 2013.

[16] Ryan McDaniell, Bum-Kyu Lee, Lingyun Song, Zheng Liu, Alan P. Boyle, Michael R. Erdos, Laura J. Scott, Mario A. Morken, Katerina S. Kucera, Anna Battenhouse, Damian Keefe, Francis S. Collins, Huntington F. Willard, Jason D. Lieb, Terrence S. Furey, Gregory E. Crawford, Vishwanath R. Iyer, and Ewan Birney. Heritable individual-specific and allele-specific chromatin signatures in humans. *Science (New York, N.Y.)*, 328(5975):235–9, 2010.

[17] Maya Kasowski, Sofia Kyriazopoulou-Panagiotopoulou, Fabian Grubert, Judith B. Zaugg, Anshul Kundaje, Yuling Liu, Alan P. Boyle, Qiangfeng Cliff Zhang, Fouad Zakharia, Damek V. Spacek, Jingjing Li, Dan Xie, Anthony Olarerin-George, Lars M. Steinmetz, John B. Hogenesch, Manolis Kellis, Serafim Batzoglou, and Michael Snyder. Extensive variation in chromatin states across humans. *Science (New York, N.Y.)*, 342(6159):750–2, 2013.

[18] ENCODE Project Consortium et al. A userâĂŹs guide to the encyclopedia of dna elements (encode). *PLoS Biol*, 9(4):e1001046, 2011.

[19] Bradley E. Bernstein, Ewan Birney, Ian Dunham, Eric D. Green, Chris Gunter, and Michael Snyder. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[20] Encode Consortium Mouse, J. A. Stamatoyannopoulos, M. Snyder, R. Hardison, B. Ren, T. Gingeras, D. M. Gilbert, M. Groudine, M. Bender, R. Kaul, T. Canfield, E. Giste, A. Johnson, M. Zhang, G. Balasundaram, R. Byron, V. Roach, P. J. Sabo, R. Sandstrom, A. S. Stehling, R. E. Thurman, S. M. Weissman, P. Cayting, M. Hariharan, J. Lian, Y. Cheng, S. G. Landt, Z. Ma, B. J. Wold, J. Dekker, G. E. Crawford, C. A. Keller, W. Wu, C. Morrissey, S. A. Kumar, T. Mishra, D. Jain, M. Byrska-Bishop, D. Blankenberg, B. R. Lajoie, G. Jain, A. Sanyal, K. B. Chen, O. Denas, J. Taylor, G. A. Blobel, M. J. Weiss, M. Pimkin, W. Deng, G. K. Marinov, B. A. Williams, K. I. Fisher-Aylor, G. Desalvo, A. Kiralusha, D. Trout, H. Amrhein, A. Mortazavi, L. Edsall, D. McCleary, S. Kuan, Y. Shen, F. Yue, Z. Ye, C. A. Davis, C. Zaleski, S. Jha, C. Xue, A. Dobin, W. Lin, M. Fastuca, H. Wang, R. Guigo, S. Djebali, J. Lagarde, T. Ryba, T. Sasaki, V. S. Malladi, M. S. Cline, V. M. Kirkup, K. Learned, K. R. Rosenbloom, W. J. Kent, E. A. Feingold, P. J. Good, M. Pazin, R. F. Lowdon, and L. B. Adams. An encyclopedia of mouse dna elements (mouse encode). *Genome Biol*, 13(8):418, 2012. Stamatoyannopoulos, John A Snyder, Michael Hardison, Ross Ren, Bing Gingeras, Thomas Gilbert, David M Groudine, Mark Bender, Michael Kaul, Rajinder Canfield, Theresa Giste, Erica Johnson, Audra Zhang, Mia Balasundaram, Gayathri Byron, Rachel Roach, Vaughan Sabo, Peter J Sandstrom, Richard Stehling, A Sandra Thurman, Robert E Weissman, Sherman M Cayting, Philip Hariharan, Manoj Lian, Jin Cheng, Yong Landt, Stephen G Ma, Zhihai Wold, Barbara J Dekker, Job Crawford, Gregory E Keller, Cheryl A Wu, Weisheng Morrissey, Christopher Kumar, Swathi A Mishra, Tejaswini Jain, Deepti Byrska-Bishop, Marta Blankenberg, Daniel Lajoie, Bryan R Jain, Gaurav Sanyal, Amartya Chen, Kaun-Bei Denas, Olgert Taylor, James Blobel, Gerd A Weiss, Mitchell J Pimkin, Max Deng, Wulan Marinov, Georgi K Williams, Brian A Fisher-Aylor, Katherine I Desalvo, Gilberto Kiralusha, Anthony Trout, Diane Amrhein, Henry Mortazavi, Ali Edsall, Lee McCleary, David Kuan, Samantha Shen, Yin Yue, Feng Ye, Zhen Davis, Carrie A Zaleski, Chris Jha, Sonali Xue, Chenghai Dobin, Alex Lin, Wei Fastuca, Meagan Wang, Huaien Guigo, Roderic Djebali, Sarah Lagarde, Julien Ryba, Tyrone Sasaki, Takayo Malladi, Venkat S Cline, Melissa S Kirkup, Vanessa M Learned, Katrina Rosenbloom, Kate R Kent, W James Feingold, Elise A Good, Peter J Pazin, Michael Lowdon, Rebecca F Adams, Leslie B eng R01 DK065806/DK/NIDDK NIH HHS/ R01 HG003143/HG/NHGRI NIH HHS/ R37 DK044746/DK/NIDDK NIH HHS/ RC2 HG005573/HG/NHGRI NIH HHS/ England 2012/08/15 06:00 Genome Biol. 2012 Aug 13;13(8):418. doi: 10.1186/gb-2012-13-8-418.

[21] Jennifer Harrow, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt, Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward, Rachel Harte, Michael Lin, Cédric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Bala-

subramanian, Baikang Pei, Michael Tress, Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guigó, and Tim J. Hubbard. Gencode: the reference human genome annotation for the encode project. *Genome research*, 22(9):1760–74, 2012.

[22] P. W. Laird. Principles and challenges of genomewide dna methylation analysis. *Nat Rev Genet*, 11(3):191–203, 2010. Laird, Peter W eng R01-CA118699/CA/NCI NIH HHS/ U24-CA143882/CA/NCI NIH HHS/ Comparative Study Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Review England 2010/02/04 06:00 Nat Rev Genet. 2010 Mar;11(3):191-203. doi: 10.1038/nrg2732.

[23] Jason Ernst and Manolis Kellis. Chromhmm: automating chromatin-state discovery and characterization. *Nature methods*, 9(3):215–6, 2012.

[24] Michael M. Hoffman, Orion J. Buske, Jie Wang, Zhiping Weng, Jeff A. Bilmes, and William Stafford Noble. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature methods*, 9(5):473–6, 2012.

[25] N. Day, A. Hemmaplardh, R. E. Thurman, J. A. Stamatoyannopoulos, and W. S. Noble. Unsupervised segmentation of continuous genomic data. *Bioinformatics*, 23(11):1424–6, 2007. Day, Nathan Hemmaplardh, Andrew Thurman, Robert E Stamatoyannopoulos, John A Noble, William S eng R01 GM071923/GM/NIGMS NIH HHS/ R01 GM71852/GM/NIGMS NIH HHS/ U01 HG003161/HG/NHGRI NIH HHS/ Research Support, N.I.H., Extramural England Oxford, England 2007/03/27 09:00 Bioinformatics. 2007 Jun 1;23(11):1424-6. Epub 2007 Mar 23.

[26] Jason Ernst and Manolis Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature biotechnology*, 28(8):817–25, 2010.

[27] G. J. Filion, J. G. van Bemmel, U. Braunschweig, W. Talhout, J. Kind, L. D. Ward, W. Brugman, I. J. de Castro, R. M. Kerkhoven, H. J. Bussemaker, and B. van Steensel. Systematic protein location mapping reveals five principal chromatin types in drosophila cells. *Cell*, 143(2):212–24, 2010.

[28] J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. E. Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–9, 2011.

[29] M. M. Hoffman, J. Ernst, S. P. Wilder, A. Kundaje, R. S. Harris, M. Libbrecht, B. Giardine, P. M. Ellenbogen, J. A. Bilmes, E. Birney, R. C. Hardison, I. Dunham, M. Kellis, and W. S. Noble. Integrative annotation of chromatin

elements from encode data. *Nucleic Acids Res*, 41(2):827–41, 2013. Hoffman, Michael M Ernst, Jason Wilder, Steven P Kundaje, Anshul Harris, Robert S Libbrecht, Max Giardine, Belinda Ellenbogen, Paul M Bilmes, Jeffrey A Birney, Ewan Hardison, Ross C Dunham, Ian Kellis, Manolis Noble, William Stafford eng 095908/Wellcome Trust/United Kingdom DK065806/DK/NIDDK NIH HHS/ HG004570/HG/NHGRI NIH HHS/ HG004695/HG/NHGRI NIH HHS/ HG005334/HG/NHGRI NIH HHS/ HG005573/HG/NHGRI NIH HHS/ HG006259/HG/NHGRI NIH HHS/ K99 HG006259/HG/NHGRI NIH HHS/ R01 DK065806/DK/NIDDK NIH HHS/ R01 HG004037/HG/NHGRI NIH HHS/ RC2 HG005573/HG/NHGRI NIH HHS/ Research Support, N.I.H., Extramural Research Support, U.S. Gov't, Non-P.H.S. England 2012/12/12 06:00 Nucleic Acids Res. 2013 Jan;41(2):827-41. doi: 10.1093/nar/gks1284. Epub 2012 Dec 5.

[30] S. C. Parker, M. L. Stitzel, D. L. Taylor, J. M. Orozco, M. R. Erdos, J. A. Akiyama, K. L. van Bueren, P. S. Chines, N. Narisu, B. L. Black, A. Visel, L. A. Pennacchio, and F. S. Collins. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc Natl Acad Sci U S A*, 110(44):17921–6, 2013.

[31] 1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[32] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[33] Francis S Collins, Michael Morgan, and Aristides Patrinos. The human genome project: lessons from large-scale biology. *Science*, 300(5617):286–290, 2003.

[34] International Human Genome Sequencing Consortium et al. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[35] Michael L Metzker. Sequencing technologiesâĂŤthe next generation. *Nature reviews genetics*, 11(1):31–46, 2010.

[36] Ayman Grada and Kate Weinbrecht. Next-generation sequencing: methodology and application. *Journal of Investigative Dermatology*, 133(8):e11, 2013.

[37] Daniel MacLean, Jonathan DG Jones, and David J Studholme. Application of'next-generation'sequencing technologies to microbial genetics. *Nature Reviews Microbiology*, 7(4):287–296, 2009.

[38] Luongdl. Sequence assembly, 2014.

[39] J. Commins, C. Toft, and M. A. Fares. Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. biol. procedures online (2009). accessed via springerimages., 2011.

[40] Ann-Christine Syvänen. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Reviews Genetics*, 2(12):930–942, 2001.

[41] PaleWhaleGail. Binding of aso probes, 2008.

[42] Jkwchui. Cell diagram adapted from ladyofhats' animal cell diagram. information based on illumina data sheet, as well as chip and immunoprecipitation articles and references., 2012.

[43] R. A. Harris, T. Wang, C. Coarfa, R. P. Nagarajan, C. Hong, S. L. Downey, B. E. Johnson, S. D. Fouse, A. Delaney, Y. Zhao, A. Olshen, T. Ballinger, X. Zhou, K. J. Forsberg, J. Gu, L. Echipare, H. O'Geen, R. Lister, M. Pelizzola, Y. Xi, C. B. Epstein, B. E. Bernstein, R. D. Hawkins, B. Ren, W. Y. Chung, H. Gu, C. Bock, A. Gnirke, M. Q. Zhang, D. Haussler, J. R. Ecker, W. Li, P. J. Farnham, R. A. Waterland, A. Meissner, M. A. Marra, M. Hirst, A. Milosavljevic, and J. F. Costello. Comparison of sequencing-based methods to profile dna methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol*, 28(10):1097–105, 2010. Harris, R Alan Wang, Ting Coarfa, Cristian Nagarajan, Raman P Hong, Chibo Downey, Sara L Johnson, Brett E Fouse, Shaun D Delaney, Allen Zhao, Yongjun Olshen, Adam Ballinger, Tracy Zhou, Xin Forsberg, Kevin J Gu, Junchen Echipare, Lorigail O'Geen, Henriette Lister, Ryan Pelizzola, Mattia Xi, Yuanxin Epstein, Charles B Bernstein, Bradley E Hawkins, R David Ren, Bing Chung, Wen-Yu Gu, Hongcang Bock, Christoph Gnirke, Andreas Zhang, Michael Q Haussler, David Ecker, Joseph R Li, Wei Farnham, Peggy J Waterland, Robert A Meissner, Alexander Marra, Marco A Hirst, Martin Milosavljevic, Aleksandar Costello, Joseph F eng 5U01DA025956-02/DA/NIDA NIH HHS/ 5U01ES017154-02/ES/NIEHS NIH HHS/ 5U01ES017166-02/ES/NIEHS NIH HHS/ 6U01ES017155-02/ES/NIEHS NIH HHS/ F32 CA141799/CA/NCI NIH HHS/ F32 CA141799-01/CA/NCI NIH HHS/ F32CA141799/CA/NCI NIH HHS/ R01 CA057621/CA/NCI NIH HHS/ R01 DK081557/DK/NIDDK NIH HHS/ R25 DA027995/DA/NIDA NIH HHS/ T32 CA108462/CA/NCI NIH HHS/ T32 CA108462-04/CA/NCI NIH HHS/ T32 CA108462-06/CA/NCI NIH HHS/ T32 GM007067/GM/NIGMS NIH HHS/ T32 GM008568/GM/NIGMS NIH HHS/ T32 GM008568-04/GM/NIGMS NIH HHS/ U01 DA025956/DA/NIDA NIH HHS/ U01 DA025956-01/DA/NIDA NIH HHS/ U01 ES017154/ES/NIEHS NIH HHS/ U01 ES017154-01/ES/NIEHS NIH HHS/ U01 ES017155/ES/NIEHS NIH HHS/ U01 ES017155-01/ES/NIEHS NIH HHS/ U01 ES017166/ES/NIEHS NIH HHS/ U01 ES017166-01/ES/NIEHS NIH HHS/ Comparative Study Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't 2010/09/21 06:00 Nat Biotechnol. 2010 Oct;28(10):1097-105. doi: 10.1038/nbt.1682. Epub 2010 Sep 19.

[44] Toeng. Bisulfite sequencing figure 1 small, 2007.

[45] Daofeng Li, Bo Zhang, Xiaoyun Xing, and Ting Wang. Combining medip-seq and mre-seq to investigate genome-wide cpg methylation. *Methods*, 72:29–40, 2015.

[46] Boraas. Rna-seq experiment, 2012.

[47] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[48] Lior David, Wolfgang Huber, Marina Granovskaia, Joern Toedling, Curtis J Palm, Lee Bofkin, Ted Jones, Ronald W Davis, and Lars M Steinmetz. A high-resolution map of transcription in the yeast genome. *Proceedings of the National Academy of Sciences*, 103(14):5320–5325, 2006.

[49] Paul Bertone, Mark Gerstein, and Michael Snyder. Applications of dna tiling arrays to experimental genome annotation and regulatory pathway discovery. *Chromosome Research*, 13(3):259–274, 2005.

[50] Jill Cheng, Philipp Kapranov, Jorg Drenkow, Sujit Dike, Shane Brubaker, Sandeep Patel, Jeffrey Long, David Stern, Hari Tammana, Gregg Helt, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, 308(5725):1149–1154, 2005.

[51] Jiang Zhu, Mazhar Adli, James Y. Zou, Griet Verstappen, Michael Coyne, Xiaolan Zhang, Timothy Durham, Mohammad Miri, Vikram Deshpande, Philip L. De Jager, David A. Bennett, Joseph A. Houmard, Deborah M. Muoio, Tamer T. Onder, Ray Camahort, Chad A. Cowan, Alexander Meissner, Charles B. Epstein, Noam Shoresh, and Bradley E. Bernstein. Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell*, 152(3):642–54, 2013.

[52] L. A. Boyer, K. Plath, J. Zeitlinger, T. Brambrink, L. A. Medeiros, T. I. Lee, S. S. Levine, M. Wernig, A. Tajonar, M. K. Ray, G. W. Bell, A. P. Otte, M. Vidal, D. K. Gifford, R. A. Young, and R. Jaenisch. Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*, 441(7091):349–53, 2006.

[53] A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–37, 2007.

[54] Y. B. Schwartz, T. G. Kahn, P. Stenberg, K. Ohno, R. Bourgon, and V. Pirrotta. Alternative epigenetic chromatin states of polycomb target genes. *PLoS Genet*, 6(1):e1000805, 2010.

[55] B. E. Bernstein, M. Kamal, K. Lindblad-Toh, S. Bekiranov, D. K. Bailey, D. J. Huebert, S. McMahon, E. K. Karlsson, 3rd Kulbokas, E. J., T. R. Gingeras, S. L. Schreiber, and E. S. Lander. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, 120(2):169–81, 2005.

[56] T. Y. Roh, S. Cuddapah, and K. Zhao. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev*, 19(5):542–52, 2005.

[57] LC Schalkwyk, EL Meaburn, R Smith, EL Dempster, AR Jeffries, MN Davies, R Plomin, and J Mill. Allelic skewing of dna methylation is widespread across the genome. *Am J Hum Genet*, 86:196 – 212, 2010.

[58] K. Miyake, C. Yang, Y. Minakuchi, K. Ohori, M. Soutome, T. Hirasawa, Y. Kazuki, N. Adachi, S. Suzuki, M. Itoh, Y. I. Goto, T. Andoh, H. Kurosawa, M. Oshimura, M. Sasaki, A. Toyoda, and T. Kubota. Comparison of genomic and epigenomic expression in monozygotic twins discordant for rett syndrome. *PLoS One*, 8(6):e66729, 2013.

[59] T. J. Hudson, W. Anderson, A. Artez, A. D. Barker, C. Bell, R. R. Bernabe, M. K. Bhan, F. Calvo, I. Eerola, D. S. Gerhard, A. Guttmacher, M. Guyer, F. M. Hemsley, J. L. Jennings, D. Kerr, P. Klatt, P. Kolar, J. Kusada, D. P. Lane, F. Laplace, L. Youyong, G. Nettekoven, B. Ozenberger, J. Peterson, T. S. Rao, J. Remacle, A. J. Schafer, T. Shibata, M. R. Stratton, J. G. Vockley, K. Watanabe, H. Yang, M. M. Yuen, B. M. Knoppers, M. Bobrow, A. Cambon-Thomsen, L. G. Dressler, S. O. Dyke, Y. Joly, K. Kato, K. L. Kennedy, P. Nicolas, M. J. Parker, E. Rial-Sebbag, C. M. Romeo-Casabona, K. M. Shaw, S. Wallace, G. L. Wiesner, N. Zeps, P. Lichter, A. V. Biankin, C. Chabannon, L. Chin, B. Clement, E. de Alava, F. Degos, M. L. Ferguson, P. Geary, D. N. Hayes, A. L. Johns, A. Kasprzyk, H. Nakagawa, R. Penny, M. A. Piris, R. Sarin, A. Scarpa, M. van de Vijver, P. A. Futreal, H. Aburatani, M. Bayes, D. D. Botwell, P. J. Campbell, X. Estivill, S. M. Grimmond, I. Gut, M. Hirst, C. Lopez-Otin, P. Majumder, M. Marra, J. D. McPherson, Z. Ning, X. S. Puente, Y. Ruan, H. G. Stunnenberg, H. Swerdlow, V. E. Velculescu, R. K. Wilson, H. H. Xue, L. Yang, P. T. Spellman, G. D. Bader, P. C. Boutros, P. Flicek, G. Getz, R. Guigo, G. Guo, D. Haussler, S. Heath, T. J. Hubbard, T. Jiang, et al. International network of cancer genome projects. *Nature*, 464(7291):993–8, 2010.

[60] X. Cao and S. Zhong. Enabling interspecies epigenomic comparison with cepbrowser. *Bioinformatics*, 29(9):1223–5, 2013.

[61] G. Hon, B. Ren, and W. Wang. Chromasig: a probabilistic approach to finding common chromatin signatures in the human genome. *PLoS Comput Biol*, 4(10):e1000201, 2008. Hon, Gary Ren, Bing Wang, Wei eng Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't 2008/10/18 09:00 PLoS Comput Biol. 2008 Oct;4(10):e1000201. doi: 10.1371/journal.pcbi.1000201. Epub 2008 Oct 17.

[62] Q. Song and A. D. Smith. Identifying dispersed epigenomic domains from chip-seq data. *Bioinformatics*, 27(6):870–1, 2011. Song, Qiang Smith, Andrew D eng England Oxford, England 2011/02/18 06:00 Bioinformatics. 2011 Mar 15;27(6):870-1. doi: 10.1093/bioinformatics/btr030. Epub 2011 Feb 16.

[63] T. Ye, A. R. Krebs, M. A. Choukrallah, C. Keime, F. Plewniak, I. Davidson, and L. Tora. seqminer: an integrated chip-seq data interpretation platform. *Nucleic Acids Res*, 39(6):e35, 2011.

[64] X. Zeng, R. Sanalkumar, E. H. Bresnick, H. Li, Q. Chang, and S. Keles. jmosaics: joint analysis of multiple chip-seq datasets. *Genome Biol*, 14(4):R38, 2013. Zeng, Xin Sanalkumar, Rajendran Bresnick, Emery H Li, Hongda Chang, Qiang Keles, Sunduz eng HG006716/HG/NHGRI NIH HHS/ R01 HG003747/HG/NHGRI NIH HHS/ R21 HD066560/HD/NICHD NIH HHS/ Research Support, N.I.H., Extramural England 2013/07/13 06:00 Genome Biol. 2013 Apr 29;14(4):R38. doi: 10.1186/gb-2013-14-4-r38.

[65] Peter A Jones. Functions of dna methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484–492, 2012.

[66] David Reich, Alkes L Price, and Nick Patterson. Principal component analysis of genetic data. *Nature genetics*, 40(5):491–492, 2008.

[67] L. Buck and R. Axel. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell*, 65(1):175–87, 1991.

[68] A. Chess, I. Simon, H. Cedar, and R. Axel. Allelic inactivation regulates olfactory receptor gene expression. *Cell*, 78(5):823–34, 1994.

[69] Peter Mombaerts, Fan Wang, Catherine Dulac, Steve K Chao, Adriana Nemes, Monica Mendelsohn, James Edmondson, and Richard Axel. Visualizing an olfactory sensory map. *Cell*, 87(4):675–686, 1996.

[70] Kerry J Ressler, Susan L Sullivan, and Linda B Buck. Information coding in the olfactory system: evidence for a stereotyped and highly organized epitope map in the olfactory bulb. *Cell*, 79(7):1245–1255, 1994.

[71] Robert Vassar, John Ngai, and Richard Axel. Spatial segregation of odorant receptor expression in the mammalian olfactory epithelium. *Cell*, 74(2):309–318, 1993.

[72] G Barnea, S O'Donnell, F Mancia, X Sun, A Nemes, M Mendelsohn, and R Axel. Odorant receptors on axon termini in the brain. *Science*, 304(5676):1468–1468, 2004.

[73] Paul Feinstein, Thomas Bozza, Ivan Rodriguez, Anne Vassalli, and Peter Mombaerts. Axon guidance of mouse olfactory sensory neurons by odorant receptors and the $\beta 2$ adrenergic receptor. *Cell*, 117(6):833–846, 2004.

[74] Shou Serizawa, Kazunari Miyamichi, Haruki Takeuchi, Yuya Yamagishi, Misao Suzuki, and Hitoshi Sakano. A neuronal identity code for the odorant receptor-specific and activity-dependent axon sorting. *Cell*, 127(5):1057–1069, 2006.

[75] Fan Wang, Adriana Nemes, Monica Mendelsohn, and Richard Axel. Odorant receptors govern the formation of a precise topographic map. *Cell*, 93(1):47–60, 1998.

[76] Diego J Rodriguez-Gil, Helen B Treloar, Xiaohong Zhang, Alexandra M Miller, Aimee Two, Carrie Iwema, Stuart J Firestein, and Charles A Greer. Chromosomal location-dependent nonstochastic onset of odor receptor expression. *The Journal of Neuroscience*, 30(30):10067–10075, 2010.

[77] Robert Vassar, Steve K Chao, Raquel Sitcheran, Jennifer M Nun, Leslie B Vosshall, Richard Axel, et al. Topographic organization of sensory projections to the olfactory bulb. *Cell*, 79(6):981–991, 1994.

[78] Benjamin M. Shykind. Regulation of odorant receptors: one allele at a time. *Human molecular genetics*, 14 Spec No:R33–9, 2005.

[79] Joseph W. Lewcock and Randall R. Reed. A feedback mechanism regulates monoallelic odorant receptor expression. *Proceedings of the National Academy of Sciences of the United States of America*, 101(4):1069–74, 2004.

[80] Shou Serizawa, Kazunari Miyamichi, Hiroko Nakatani, Misao Suzuki, Michiko Saito, Yoshihiro Yoshihara, and Hitoshi Sakano. Negative feedback regulation ensures the one receptor-one olfactory neuron rule in mouse. *Science (New York, N.Y.)*, 302(5653):2088–94, 2003.

[81] Benjamin M. Shykind, S. Christy Rohani, Sean O'Donnell, Adriana Nemes, Monica Mendelsohn, Yonghua Sun, Richard Axel, and Gilad Barnea. Gene switching and the stability of odorant receptor gene choice. *Cell*, 117(6):801–15, 2004.

[82] Minh Q. Nguyen, Zhishang Zhou, Carolyn A. Marks, Nicholas J. P. Ryba, and Leonardo Belluscio. Prominent roles for odorant receptor coding sequences in allelic exclusion. *Cell*, 131(5):1009–17, 2007.

[83] Andrea Rothman, Paul Feinstein, Junji Hirota, and Peter Mombaerts. The promoter of the mouse odorant receptor gene m71. *Molecular and Cellular Neuroscience*, 28(3):535–546, 2005.

[84] Stavros Lomvardas, Gilad Barnea, David J Pisapia, Monica Mendelsohn, Jennifer Kirkland, and Richard Axel. Interchromosomal interactions and olfactory receptor choice. *Cell*, 126(2):403–413, 2006.

[85] Stefan H Fuss, Masayo Omura, and Peter Mombaerts. Local and cis effects of the h element on expression of odorant receptor genes in mouse. *Cell*, 130(2):373–384, 2007.

[86] Hirofumi Nishizumi, Kouhei Kumasaka, Nobuko Inoue, Ai Nakashima, and Hitoshi Sakano. Deletion of the core-h region in mice abolishes the expression

of three proximal odorant receptor genes in cis. *Proceedings of the National Academy of Sciences*, 104(50):20067–20072, 2007.

[87] Patrick Trojer and Danny Reinberg. Facultative heterochromatin: is there a distinctive molecular signature? *Molecular cell*, 28(1):1–13, 2007.

[88] Bradley E Bernstein, Tarjei S Mikkelsen, Xiaohui Xie, Michael Kamal, Dana J Huebert, James Cuff, Ben Fry, Alex Meissner, Marius Wernig, Kathrin Plath, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2):315–326, 2006.

[89] Elena Ezhkova, H Amalia Pasolli, Joel S Parker, Nicole Stokes, I-hsin Su, Gregory Hannon, Alexander Tarakhovsky, and Elaine Fuchs. Ezh2 orchestrates gene expression for the stepwise differentiation of tissue-specific stem cells. *Cell*, 136(6):1122–1135, 2009.

[90] Barna D. Fodor, Nicholas Shukeir, Gunter Reuter, and Thomas Jenuwein. Mammalian su(var) genes in chromatin control. *Annual review of cell and developmental biology*, 26:471–501, 2010.

[91] Gunnar Schotta, Monika Lachner, Kavitha Sarma, Anja Ebert, Roopsha Sengupta, Gunter Reuter, Danny Reinberg, and Thomas Jenuwein. A silencing pathway to induce h3-k9 and h4-k20 trimethylation at constitutive heterochromatin. *Genes & development*, 18(11):1251–1262, 2004.

[92] Tobias Straub. Basic analysis of nimblegen chip-on-chip data using bioconductor/r (prot43).

[93] Ernst Wit and John McClure. *Statistics for microarrays: design, analysis and inference*. John Wiley & Sons, 2004.

[94] Wolfgang Huber, Anja Von Heydebreck, Holger Sültmann, Annemarie Poustka, and Martin Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18(suppl 1):S96–S104, 2002.

[95] W Evan Johnson, Wei Li, Clifford A Meyer, Raphael Gottardo, Jason S Carroll, Myles Brown, and X Shirley Liu. Model-based analysis of tiling-arrays for chip-chip. *Proceedings of the National Academy of Sciences*, 103(33):12457–12462, 2006.

[96] Jun S Song, W Evan Johnson, Xiaopeng Zhu, Xinmin Zhang, Wei Li, Arjun K Manrai, Jun S Liu, Runsheng Chen, and X Shirley Liu. Model-based analysis of two-color arrays (ma2c). *Genome Biol*, 8(8):R178, 2007.

[97] Bo Wen, Hao Wu, Yoichi Shinkai, Rafael A Irizarry, and Andrew P Feinberg. Large histone h3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nature genetics*, 41(2):246–250, 2009.

[98] Peter Humburg, David Bulger, and Glenn Stone. Parameter estimation for robust hmm analysis of chip-chip data. *BMC Bioinformatics*, page 343, 2008.

[99] 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, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112, 2009.

[100] Michiel JL de Hoon, Seiya Imoto, John Nolan, and Satoru Miyano. Open source clustering software. *Bioinformatics*, 20(9):1453–1454, 2004.

[101] Weston.pace. Hidden markov model with output, 2007.

[102] Susan E Celniker, Laura AL Dillon, Mark B Gerstein, Kristin C Gunsalus, Steven Henikoff, Gary H Karpen, Manolis Kellis, Eric C Lai, Jason D Lieb, David M MacAlpine, et al. Unlocking the secrets of the genome. *Nature*, 459(7249):927–930, 2009.

[103] R David Hawkins, Gary C Hon, Leonard K Lee, QueMinh Ngo, Ryan Lister, Mattia Pelizzola, Lee E Edsall, Samantha Kuan, Ying Luu, Sarit Klugman, et al. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell stem cell*, 6(5):479–491, 2010.

[104] Jessica L Larson and Guo-Cheng Yuan. Epigenetic domains found in mouse embryonic stem cells via a hidden markov model. *BMC bioinformatics*, 11(1):557, 2010.

[105] Catherine Dulac and Richard Axel. A novel family of genes encoding putative pheromone receptors in mammals. *Cell*, 83(2):195–206, 1995.

[106] Stephen D Liberles, Lisa F Horowitz, Donghui Kuang, James J Contos, Kathleen L Wilson, Jessica Siltberg-Liberles, David A Liberles, and Linda B Buck. Formyl peptide receptors are candidate chemosensory receptors in the vomeronasal organ. *Proceedings of the National Academy of Sciences*, 106(24):9842–9847, 2009.

[107] Stéphane Rivière, Ludivine Challet, Daniela Fluegge, Marc Spehr, and Ivan Rodriguez. Formyl peptide receptor-like proteins are a novel family of vomeronasal chemosensors. *Nature*, 459(7246):574–577, 2009.

[108] John W Nicol, Gregg A Helt, Steven G Blanchard, Archana Raja, and Ann E Loraine. The integrated genome browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics*, 25(20):2730–2731, 2009.

[109] Tae Hoon Kim, Ziedulla K Abdullaev, Andrew D Smith, Keith A Ching, Dmitri I Loukinov, Roland D Green, Michael Q Zhang, Victor V Lobanenkov, and Bing Ren. Analysis of the vertebrate insulator protein ctcf-binding sites in the human genome. *Cell*, 128(6):1231–1245, 2007.

[110] Jacqueline Dickson, Humaira Gowher, Ruslan Strogantsev, Miklos Gaszner, Alan Hair, Gary Felsenfeld, and Adam G West. Vezf1 elements mediate protection from dna methylation. *PLoS Genet*, 6(1):e1000804, 2010.

[111] Cynthia D Duggan and John Ngai. Scent of a stem cell. *Nature neuroscience*, 10(6):673–674, 2007.

[112] Xueyan Chen, Hengsheng Fang, and James E Schwob. Multipotency of purified, transplanted globose basal cells in olfactory epithelium. *Journal of Comparative Neurology*, 469(4):457–474, 2004.

[113] Cheuk T Leung, Pierre A Coulombe, and Randall R Reed. Contribution of olfactory neural stem cells to tissue maintenance and regeneration. *Nature neuroscience*, 10(6):720–726, 2007.

[114] Nathaniel Heintz. Gene expression nervous system atlas (gensat). *Nature neuroscience*, 7(5):483–483, 2004.

[115] Matthew G Guenther, Stuart S Levine, Laurie A Boyer, Rudolf Jaenisch, and Richard A Young. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 130(1):77–88, 2007.

[116] Kakkad Regha, Mathew A Sloane, Ru Huang, Florian M Pauler, Katarzyna E Warczok, Balázs Melikant, Martin Radolf, Joost HA Martens, Gunnar Schotta, Thomas Jenuwein, et al. Active and repressive chromatin are interspersed without spreading in an imprinted gene cluster in the mammalian genome. *Molecular cell*, 27(3):353–366, 2007.

[117] Eric Walters, Mary Grillo, A Beate Oestreicher, and Frank L Margolis. Lacz andomp are co-expressed during ontogeny and regeneration in olfactory receptor neurons of omp promoter-lacz transgenic mice. *International journal of developmental neuroscience*, 14(7):813–822, 1996.

[118] Martina Pyrski, Zheng Xu, Eric Walters, Debra J Gilbert, Nancy A Jenkins, Neal G Copeland, and Frank L Margolis. The omp–lacz transgene mimics the unusual expression pattern of or-z6, a new odorant receptor gene on mouse chromosome 6: implication for locus-dependent gene expression. *The Journal of Neuroscience*, 21(13):4637–4648, 2001.

[119] Jeremy C McIntyre, Soma C Bose, Arnold J Stromberg, and Timothy S McClintock. Emx2 stimulates odorant receptor gene expression. *Chemical senses*, 33(9):825–837, 2008.

[120] Paul Flicek and Ewan Birney. Sense from sequence reads: methods for alignment and assembly. *Nature methods*, 6:S6–S12, 2009.

[121] Michael F Lin, Irwin Jungreis, and Manolis Kellis. Phylocsf: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, 27(13):i275–i282, 2011.

[122] Timothy L Bailey, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. Meme suite: tools for motif discovery and searching. *Nucleic acids research*, page gkp335, 2009.

[123] David Baker and Andrej Sali. Protein structure prediction and structural genomics. *Science*, 294(5540):93–96, 2001.

[124] Angela Yen and Manolis Kellis. Chromatin state comparisons across sex, tissue, sample state, and developmental age, 2015.

[125] H. Xu, C. L. Wei, F. Lin, and W. K. Sung. An hmm approach to genome-wide identification of differential histone modification sites from chip-seq data. *Bioinformatics*, 24(20):2344–9, 2008. Xu, Han Wei, Chia-Lin Lin, Feng Sung, Wing-Kin eng England Oxford, England 2008/08/01 09:00 Bioinformatics. 2008 Oct 15;24(20):2344-9. doi: 10.1093/bioinformatics/btn402. Epub 2008 Jul 29.

[126] C. Taslim, J. Wu, P. Yan, G. Singer, J. Parvin, T. Huang, S. Lin, and K. Huang. Comparative study on chip-seq data: normalization and binding pattern characterization. *Bioinformatics*, 25(18):2334–40, 2009. Taslim, Cenny Wu, Jiejun Yan, Pearlly Singer, Greg Parvin, Jeffrey Huang, Tim Lin, Shili Huang, Kun eng U54 CA113001/CA/NCI NIH HHS/ U54CA113001/CA/NCI NIH HHS/ Comparative Study Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't England Oxford, England 2009/06/30 09:00 Bioinformatics. 2009 Sep 15;25(18):2334-40. doi: 10.1093/bioinformatics/btp384. Epub 2009 Jun 26.

[127] F. Johannes, R. Wardenaar, M. Colome-Tatche, F. Mousson, P. de Graaf, M. Mokry, V. Guryev, H. T. Timmers, E. Cuppen, and R. C. Jansen. Comparing genome-wide chromatin profiles using chip-chip or chip-seq. *Bioinformatics*, 26(8):1000–6, 2010. Johannes, Frank Wardenaar, Rene Colome-Tatche, Maria Mousson, Florence de Graaf, Petra Mokry, Michal Guryev, Victor Timmers, H Th Marc Cuppen, Edwin Jansen, Ritsert C eng Comparative Study Research Support, Non-U.S. Gov't England Oxford, England 2010/03/09 06:00 Bioinformatics. 2010 Apr 15;26(8):1000-6. doi: 10.1093/bioinformatics/btq087. Epub 2010 Mar 5.

[128] C. Taslim, T. Huang, and S. Lin. Dime: R-package for identifying differential chip-seq based on an ensemble of mixture models. *Bioinformatics*, 27(11):1569–70, 2011. Taslim, Cenny Huang, Tim Lin, Shili eng Research Support, N.I.H., Extramural Research Support, U.S. Gov't, Non-P.H.S. England Oxford, England 2011/04/08 06:00 Bioinformatics. 2011 Jun 1;27(11):1569-70. doi: 10.1093/bioinformatics/btr165. Epub 2011 Apr 5.

[129] Z. Shao, Y. Zhang, G. C. Yuan, S. H. Orkin, and D. J. Waxman. Manorm: a robust model for quantitative comparison of chip-seq data sets. *Genome Biol*, 13(3):R16, 2012. Shao, Zhen Zhang, Yijing Yuan, Guo-Cheng Orkin,

Stuart H Waxman, David J eng R01 DK033765/DK/NIDDK NIH HHS/ R01 HG005085/HG/NHGRI NIH HHS/ R01 HG005085-02/HG/NHGRI NIH HHS/ England 2012/03/20 06:00 Genome Biol. 2012 Mar 16;13(3):R16. doi: 10.1186/gb-2012-13-3-r16.

[130] S. Mahony, M. D. Edwards, E. O. Mazzoni, R. I. Sherwood, A. Kakumanu, C. A. Morrison, H. Wichterle, and D. K. Gifford. An integrated model of multiple-condition chip-seq data reveals predeterminants of cdx2 binding. *PLoS Comput Biol*, 10(3):e1003501, 2014. Mahony, Shaun Edwards, Matthew D Mazzoni, Esteban O Sherwood, Richard I Kakumanu, Akshay Morrison, Carolyn A Wichterle, Hynek Gifford, David K eng K01 DK101684/DK/NIDDK NIH HHS/ P01 NS055923/NS/NINDS NIH HHS/ U01 HG007037/HG/NHGRI NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. 2014/03/29 06:00 PLoS Comput Biol. 2014 Mar 27;10(3):e1003501. doi: 10.1371/journal.pcbi.1003501. eCollection 2014 Mar.

[131] Hongkai Ji, Xia Li, Qian-fei Wang, and Yang Ning. Differential principal component analysis of chip-seq. *Proceedings of the National Academy of Sciences of the United States of America*, 110(17):6789–94, 2013.

[132] Judith Marsman and Julia A Horsfield. Long distance relationships: enhancer–promoter communication and dynamic gene transcription. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1819(11):1217–1227, 2012.

[133] Jianrong Wang. personal communication.

[134] Pouya Kheradour. personal communication.

[135] R Core Team. R: A language and environment for statistical computing. *Foundation for Statistical Computing*, http://www.R-project.org/, 2013.

[136] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. 57:289–300, 1995.

[137] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[138] A. Reiner, D. Yekutieli, and Y. Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–75, 2003.

[139] K. I. Kim and M. A. van de Wiel. Effects of dependence in high-dimensional multiple testing problems. *BMC Bioinformatics*, 9:114, 2008.

[140] Sandy Clarke and Peter Hall. Robustness of multiple testing procedures against dependence. *The Annals of Statistics*, 37(1):332–358, 2009.

[141] D. M. Groppe, T. P. Urbach, and M. Kutas. Mass univariate analysis of event-related brain potentials/fields i: a critical tutorial review. *Psychophysiology*, 48(12):1711–25, 2011.

[142] D. M. Groppe, T. P. Urbach, and M. Kutas. Mass univariate analysis of event-related brain potentials/fields ii: Simulation studies. *Psychophysiology*, 48(12):1726–37, 2011.

[143] Ronald A Fisher. On the interpretation of ÏĞ2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, pages 87–94, 1922.

[144] Alan Dabney, John D Storey, and H Wickham. Q-value estimation for false discovery rate control. *R package version 1.35.0*, 2009.

[145] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–50, 2005.

[146] M. Szumilas. Explaining odds ratios. *J Can Acad Child Adolesc Psychiatry*, 19(3):227–9, 2010.

[147] D. G. Altman and J. M. Bland. How to obtain the p value from a confidence interval. *BMJ*, 343:d2304, 2011.

[148] I. F. King, C. N. Yandava, A. M. Mabb, J. S. Hsiao, H. S. Huang, B. L. Pearson, J. M. Calabrese, J. Starmer, J. S. Parker, T. Magnuson, S. J. Chamberlain, B. D. Philpot, and M. J. Zylka. Topoisomerases facilitate transcription of long genes linked to autism. *Nature*, 501(7465):58–62, 2013. King, Ian F Yandava, Chandri N Mabb, Angela M Hsiao, Jack S Huang, Hsien-Sung Pearson, Brandon L Calabrese, J Mauro Starmer, Joshua Parker, Joel S Magnuson, Terry Chamberlain, Stormy J Philpot, Benjamin D Zylka, Mark J eng P30 NS045892/NS/NINDS NIH HHS/ P30HD03110/HD/NICHD NIH HHS/ P30NS045892/NS/NINDS NIH HHS/ R01 GM101974/GM/NIGMS NIH HHS/ R01 HD068730/HD/NICHD NIH HHS/ R01 MH093372/MH/NIMH NIH HHS/ R01GM101974/GM/NIGMS NIH HHS/ R01HD068730/HD/NICHD NIH HHS/ R01MH093372/MH/NIMH NIH HHS/ T32 HD040127/HD/NICHD NIH HHS/ T32HD040127/HD/NICHD NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't England 2013/09/03 06:00 Nature. 2013 Sep 5;501(7465):58-62. doi: 10.1038/nature12504. Epub 2013 Aug 28.

[149] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.

[150] A. Wutz. Gene silencing in x-chromosome inactivation: advances in understanding facultative heterochromatin formation. *Nat Rev Genet*, 12(8):542–53, 2011.

[151] Joel B Berletch, Fan Yang, Jun Xu, Laura Carrel, and Christine M Disteche. Genes that escape from x inactivation. *Human genetics*, 130(2):237–245, 2011.

[152] H. Jing, F. Dai, C. Zhao, J. Yang, L. Li, P. Kota, L. Mao, K. Xiang, and C. Zheng. Association of genetic variants in and promoter hypermethylation of cdh1 with gastric cancer: a meta-analysis. *Medicine (Baltimore)*, 93(19):e107, 2014.

[153] X. Liu and K. M. Chu. E-cadherin and gastric cancer: cause, consequence, and applications. *Biomed Res Int*, 2014:637308, 2014.

[154] M. J. Tanner. Erythrocyte membrane structure and function. *Ciba Found Symp*, 94:3–23, 1983.

[155] L. Schrod, H. Schaefer, and R. Burger. Characterization of a t-lymphocyte membrane protein involved in t-cell function: its contribution to t-cell recognition or cellular interaction. *Immunology*, 57(4):533–8, 1986.

[156] K. Naxerova, C. J. Bult, A. Peaston, K. Fancher, B. B. Knowles, S. Kasif, and I. S. Kohane. Analysis of gene expression in a developmental context emphasizes distinct biological leitmotifs in human cancers. *Genome Biol*, 9(7):R108, 2008.

[157] S. Sekar, J. McDonald, L. Cuyugan, J. Aldrich, A. Kurdoglu, J. Adkins, G. Serrano, T. G. Beach, D. W. Craig, J. Valla, E. M. Reiman, and W. S. Liang. Alzheimer's disease is associated with altered expression of genes involved in immune response and mitochondrial processes in astrocytes. *Neurobiol Aging*, 36(2):583–91, 2015. Sekar, Shobana McDonald, Jacquelyn Cuyugan, Lori Aldrich, Jessica Kurdoglu, Ahmet Adkins, Jonathan Serrano, Geidy Beach, Thomas G Craig, David W Valla, Jonathan Reiman, Eric M Liang, Winnie S eng P30 AG019610/AG/NIA NIH HHS/ P30AG019610/AG/NIA NIH HHS/ Research Support, N.I.H., Extramural 2014/12/03 06:00 Neurobiol Aging. 2015 Feb;36(2):583-91. doi: 10.1016/j.neurobiolaging.2014.09.027. Epub 2014 Oct 2.

[158] A. J. Bingham, L. Ooi, L. Kozera, E. White, and I. C. Wood. The repressor element 1-silencing transcription factor regulates heart-specific gene expression using multiple chromatin-modifying complexes. *Mol Cell Biol*, 27(11):4082–92, 2007.

[159] T. Lu, L. Aron, J. Zullo, Y. Pan, H. Kim, Y. Chen, T. H. Yang, H. M. Kim, D. Drake, X. S. Liu, D. A. Bennett, M. P. Colaiacovo, and B. A. Yankner. Rest and stress resistance in ageing and alzheimer's disease. *Nature*, 507(7493):448–54, 2014.

[160] L. H. Tsai and R. Madabhushi. Alzheimer's disease: A protective factor for the ageing brain. *Nature*, 507(7493):439–40, 2014.

[161] M. Mochizuki-Kashio, Y. Mishima, S. Miyagi, M. Negishi, A. Saraya, T. Konuma, J. Shinga, H. Koseki, and A. Iwama. Dependency on the polycomb gene ezh2 distinguishes fetal from adult hematopoietic stem cells. *Blood*, 118(25):6553–61, 2011. Mochizuki-Kashio, Makiko Mishima, Yuta Miyagi, Satoru Negishi, Masamitsu Saraya, Atsunori Konuma, Takaaki Shinga, Jun Koseki, Haruhiko Iwama, Atsushi eng Research Support, Non-U.S. Gov't 2011/11/02 06:00 Blood. 2011 Dec 15;118(25):6553-61. doi: 10.1182/blood-2011-03-340554. Epub 2011 Oct 31.

[162] H. Xie, J. Xu, J. H. Hsu, M. Nguyen, Y. Fujiwara, C. Peng, and S. H. Orkin. Polycomb repressive complex 2 regulates normal hematopoietic stem cell function in a developmental-stage-specific manner. *Cell Stem Cell*, 14(1):68–80, 2014. Xie, Huafeng Xu, Jian Hsu, Jessie H Nguyen, Minh Fujiwara, Yuko Peng, Cong Orkin, Stuart H eng K01 DK093543/DK/NIDDK NIH HHS/ K01DK093543/DK/NIDDK NIH HHS/ P30 DK049216/DK/NIDDK NIH HHS/ U01 CA105423/CA/NCI NIH HHS/ Howard Hughes Medical Institute/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't 2013/11/19 06:00 Cell Stem Cell. 2014 Jan 2;14(1):68-80. doi: 10.1016/j.stem.2013.10.001. Epub 2013 Nov 14.

[163] Chloe M Rivera and Bing Ren. Mapping human epigenomes. *Cell*, 155(1):39–55, 2013.

[164] Vicky W Zhou, Alon Goren, and Bradley E Bernstein. Charting histone modifications and the functional organization of mammalian genomes. *Nature Reviews Genetics*, 12(1):7–18, 2011.

[165] Zachary D Smith and Alexander Meissner. Dna methylation: roles in mammalian development. *Nature Reviews Genetics*, 14(3):204–220, 2013.

[166] Peter J Park. Chip–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, 2009.

[167] Robert E Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T Maurano, Eric Haugen, Nathan C Sheffield, Andrew B Stergachis, Hao Wang, Benjamin Vernot, et al. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, 2012.

[168] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628, 2008.

[169] 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, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics*, 39(3):311–318, 2007.

[170] Shane Neph, Jeff Vierstra, Andrew B Stergachis, Alex P Reynolds, Eric Haugen, Benjamin Vernot, Robert E Thurman, Sam John, Richard Sandstrom, Audra K Johnson, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90, 2012.

[171] Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, et al. Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–1195, 2012.

[172] B. E. Bernstein, J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. A. Marra, A. L. Beaudet, J. R. Ecker, P. J. Farnham, M. Hirst, E. S. Lander, T. S. Mikkelsen, and J. A. Thomson. The nih roadmap epigenomics mapping consortium. *Nat Biotechnol*, 28(10):1045–8, 2010. Bernstein, Bradley E Stamatoyannopoulos, John A Costello, Joseph F Ren, Bing Milosavljevic, Aleksandar Meissner, Alexander Kellis, Manolis Marra, Marco A Beaudet, Arthur L Ecker, Joseph R Farnham, Peggy J Hirst, Martin Lander, Eric S Mikkelsen, Tarjei S Thomson, James A eng R01 HG003523/HG/NHGRI NIH HHS/ R01 HG004037/HG/NHGRI NIH HHS/ R01 HG005085/HG/NHGRI NIH HHS/ 2010/10/15 06:00 Nat Biotechnol. 2010 Oct;28(10):1045-8. doi: 10.1038/nbt1010-1045.

[173] Tarjei S Mikkelsen, Manching Ku, David B Jaffe, Biju Issac, Erez Lieberman, Georgia Giannoukos, Pablo Alvarez, William Brockman, Tae-Kyung Kim, Richard P Koche, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560, 2007.

[174] Sam John, Peter J Sabo, Theresa K Canfield, Kristen Lee, Shinny Vong, Molly Weaver, Hao Wang, Jeff Vierstra, Alex P Reynolds, Robert E Thurman, et al. Genome-scale mapping of dnase i hypersensitivity. *Current protocols in molecular biology*, pages 21–27, 2013.

[175] Ryan Lister, Mattia Pelizzola, Robert H Dowen, R David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, et al. Human dna methylomes at base resolution show widespread epigenomic differences. *nature*, 462(7271):315–322, 2009.

[176] Alexander Meissner, Andreas Gnirke, George W Bell, Bernard Ramsahoye, Eric S Lander, and Rudolf Jaenisch. Reduced representation bisulfite sequencing for comparative high-resolution dna methylation analysis. *Nucleic acids research*, 33(18):5868–5877, 2005.

[177] Michael Weber, Jonathan J Davies, David Wittig, Edward J Oakeley, Michael Haase, Wan L Lam, and Dirk Schuebeler. Chromosome-wide and promoter-specific analyses identify sites of differential dna methylation in normal and transformed human cells. *Nature genetics*, 37(8):853–862, 2005.

[178] Alika K Maunakea, Raman P Nagarajan, Mikhail Bilenky, Tracy J Ballinger, Cletus DâĂŹSouza, Shaun D Fouse, Brett E Johnson, Chibo Hong, Cydney Nielsen, Yongjun Zhao, et al. Conserved role of intragenic dna methylation in regulating alternative promoters. *Nature*, 466(7303):253–257, 2010.

[179] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[180] Hong Ji, Lauren IR Ehrlich, Jun Seita, Peter Murakami, Akiko Doi, Paul Lindau, Hwajin Lee, Martin J Aryee, Rafael A Irizarry, Kitai Kim, et al. Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature*, 467(7313):338–342, 2010.

[181] Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010.

[182] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463, 2010.

[183] Karthik Devarajan. Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS Comput Biol*, 4(7):e1000029, 2008.

[184] Diana Chang and Alon Keinan. Principal component analysis characterizes shared pathogenetics from genome-wide association studies. *PLOS Comput Biol*, 10(9):e1003820, 2014.

[185] Patrick R Schmid, Nathan P Palmer, Isaac S Kohane, and Bonnie Berger. Making sense out of massive data by going beyond differential expression. *Proceedings of the National Academy of Sciences*, 109(15):5594–5599, 2012.

[186] Sean Simmons, Jian Peng, Jadwiga Bienkowska, and Bonnie Berger. Discovering what dimensionality reduction really tells us about rna-seq data. *Journal of Computational Biology*, 22(8):715–728, 2015.

[187] Hansong Wang, Christopher A Haiman, Laurence N Kolonel, Brian E Henderson, Lynne R Wilkens, Loïc Le Marchand, and Daniel O Stram. Self-reported ethnicity, genetic structure and the impact of population stratification in a multiethnic study. *Human genetics*, 128(2):165–177, 2010.

[188] Fabien Filleton, Florent Chuffart, Muniyandi Nagarajan, Hélène Bottin-Duplus, and Gaël Yvert. The complex pattern of epigenomic variation between natural yeast strains at single-nucleosome resolution. *Epigenetics & chromatin*, 8(1):1, 2015.

[189] Pedro Madrigal and Paweł Krajewski. Uncovering correlated variability in epigenomic datasets using the karhunen-loeve transform. *BioData mining*, 8(1):1, 2015.

[190] Matthias Scholz. Pca transformation, 2006.