

**Insights into the Regulatory Activities of Chromatin Through
Computational Analyses of Whole-Genome Data**

Stephen Aurelien Hoang
Mechanicsville, Virginia

B.S., University of Virginia, 2007
M.S., University of Virginia, 2011

A Dissertation presented to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Biochemistry and Molecular Genetics

University of Virginia
August, 2013

To my grandparents.

*The investigation of nature is an infinite pasture-ground, where all may graze, and
where the more bite, the longer the grass grows, the sweeter is its flavor,
and the more it nourishes.*

—*T. H. Huxley (1871)*

Acknowledgements

My deepest thanks go to Stefan Bekiranov, my mentor and friend. He was an intellectual guide for the work presented in this dissertation. His sense of duty as a mentor has been inspiring, and is responsible for my growth as a scientist. My gratitude extends beyond what I can reasonably express here.

I also owe much to Marty Mayo for making great contributions to my understanding of cancer biology. I am deeply grateful to him and his lab for generating the enormous amounts of data that form the basis of Chapter four.

My sincere thanks also go to Patrick Grant and Erdem Sendic for the opportunity to work on an exciting project that is not presented in this dissertation, but was my first excursion into yeast epigenomics. That collaboration set me on the path to an independent project, which is detailed in Chapter five.

Thanks also to Kunal Poorey, Marcin Cieřlik, and Xiaojiang Xu—the members of the Bekiranov Lab with whom my term overlapped. I have benefited greatly from their friendship and technical assistance. In particular, Xiaojiang made large contributions to the work presented in chapters two and three. Marcin deserves equal credit for the work presented in Chapter four. He was also a great sounding board for ideas that I had pertaining to Chapter five (and many other ideas that I’ve had over the last few years).

Finally, I would like to thank my family, friends, and soon-to-be wife, Clare, for being supportive, and altogether wonderful.

Abstract

In the last decade, advances in DNA sequencing technology have enabled the routine generation of whole-genome datasets that profile chromatin composition and conformation. Quantitative analyses of these data have required the development of innovative computational data analysis methodologies. These analysis techniques have been used to generate novel biological hypotheses about the regulatory activities of chromatin, which can be experimentally validated. This dissertation presents several case studies of such analyses. In the first study, machine learning models, which predict transcription as a function of multivariate histone modification levels, are used to predict the association of symmetrically dimethylated arginine 3 on histone H4 (H4R3me2s) with transcriptional repression. Methodological approaches to constructing similar models are also explored in depth. In the next study, analyses of changes in a panel of histone modifications during the epithelial-mesenchymal transition (EMT) reveal a high degree of regulatory coordination among genes within distinct functional classes and pathways. Changes at enhancers associated with these genes also show coordination with respect to transcription factor binding. These observations lead to the hypothesis that histone modifications enable and sustain transcriptional feedback loops distinctly associated with each phenotypic endpoint in EMT. In the final study, a novel approach for analyzing unbiased chromatin interaction data (Hi-C data) is presented. This approach utilizes network analysis techniques to infer conformational features of the genome. Using these techniques, assessments are made of the degree of hierarchical

organization in the budding yeast genome. Furthermore, a novel correlation between replication timing and degree of inter-chromosomal interactions is observed. The studies presented in this dissertation demonstrate the utility of computational data analysis in generating novel systems-level hypotheses about the regulatory behavior of chromatin. Since many of the analysis techniques presented in this work have not been otherwise applied to chromatin data, many domain-specific technical considerations are also discussed. This dissertation provides a variety of novel insights into the regulatory activities of chromatin; and perhaps more importantly, it provides several analytical frameworks for distilling systems-level insights from whole-genome chromatin data.

Contents

| | |
|--|-----------|
| Acknowledgements | i |
| Abstract | ii |
| Contents | iv |
| List of Figures | vii |
| List of Tables | ix |
| 1 Introduction | 1 |
| 1.1 Genomic regulation at the level of chromatin | 2 |
| 1.1.1 Overview | 2 |
| 1.1.2 Histone modifications | 5 |
| 1.1.3 Higher-order chromatin organization | 12 |
| 1.2 High-throughput methods in chromatin biology | 18 |
| 1.2.1 Overview | 18 |
| 1.2.2 ChIP-seq and Hi-C | 20 |
| 1.3 Computational analysis of whole-genome chromatin data | 22 |
| 1.3.1 Overview | 22 |
| 1.3.2 Machine learning applied to histone modification data | 23 |
| 1.3.3 Methods for whole-genome chromatin conformation data | 29 |
| 1.3.4 Network analysis of Hi-C data | 31 |
| 1.4 Dissertation rationale | 38 |
| 2 Regression analysis reveals transcriptional association H4R3me2 | 39 |
| 2.1 Introduction | 39 |
| 2.2 Model construction rationale | 42 |
| 2.2.1 Basic workflow | 42 |
| 2.2.2 Estimating input amplitudes for regression models | 44 |
| 2.3 Results and discussion | 45 |
| 2.3.1 Multilinear model | 45 |
| 2.3.2 Multilinear model terms | 47 |
| 2.3.3 MARS model | 51 |
| 2.3.4 MARS model terms | 52 |
| 2.3.5 Model comparison | 56 |
| 2.3.6 <i>In silico</i> knockout analysis | 58 |

| | | |
|----------|---|------------|
| 2.3.7 | H4R3me2 is globally repressive in ML and MARS models . . . | 60 |
| 2.3.8 | Experimental studies demonstrate H4R3me2 represses gene expression | 66 |
| 2.4 | Conclusions | 67 |
| 2.5 | Methods | 68 |
| 2.5.1 | Calculation of amplitudes | 68 |
| 2.5.2 | Selection of transcription start and stop sites | 71 |
| 2.5.3 | Building the multilinear model using stepwise linear regression | 72 |
| 2.5.4 | Building the MARS model | 74 |
| 2.5.5 | Amplitude robustness and relative error of mark amplitude estimates | 75 |
| 2.6 | Chapter acknowledgements | 78 |
| 3 | Quantification of histone modification ChIP-seq enrichment | 79 |
| 3.1 | Introduction | 79 |
| 3.2 | Methods | 83 |
| 3.2.1 | Gene selection | 83 |
| 3.2.2 | Tag repeat filter | 84 |
| 3.2.3 | Tag counting | 84 |
| 3.2.4 | Iterative model-based enrichment estimation | 86 |
| 3.2.5 | Non-iterative model-based enrichment estimate | 88 |
| 3.2.6 | Evaluation of template models | 88 |
| 3.2.7 | MARS model construction and evaluation | 88 |
| 3.3 | Results and discussion | 89 |
| 3.3.1 | Overview of model construction | 89 |
| 3.3.2 | Template model error analysis | 89 |
| 3.3.3 | Enrichment estimation and model performance | 91 |
| 3.3.4 | Enrichment profiles and gene length | 97 |
| 3.3.5 | Regulatory information embedded in spatial deposition patterns | 98 |
| 3.4 | Conclusions | 102 |
| 3.5 | Chapter acknowledgements | 103 |
| 4 | Epigenetic reprogramming in the epithelial-mesenchymal transition | 104 |
| 4.1 | Introduction | 104 |
| 4.2 | General strategy | 106 |
| 4.3 | Results and discussion | 109 |
| 4.3.1 | Chromatin profiling reveals EMT-related gene clusters | 109 |
| 4.3.2 | EMT clusters are enriched for many EMT-associated functions and phenotypes | 114 |
| 4.3.3 | Regulation of EMT signaling pathways is chromatin-mediated | 117 |
| 4.3.4 | Epigenetic switches at enhancers correlate with gene expression | 121 |
| 4.3.5 | Transcriptional control of EMT-GCs through epigenetic reprogramming of enhancers | 124 |
| 4.3.6 | Links between enhancer clusters and gene clusters suggest a chromatin-mediated transcriptional feedback | 128 |

| | | |
|----------|---|------------|
| 4.4 | Conclusions | 129 |
| 4.5 | Methods | 132 |
| 4.5.1 | Cell culture | 132 |
| 4.5.2 | ChIP-seq | 133 |
| 4.5.3 | Microarray and Gene Expression Analysis | 134 |
| 4.5.4 | ChIP-seq data processing | 134 |
| 4.5.5 | Scaled Differential Enrichments | 135 |
| 4.5.6 | Definition of Putative Enhancer Loci | 135 |
| 4.5.7 | Enhancer-associated histone modifications | 136 |
| 4.5.8 | Gene assignment and filtering of enhancer loci | 136 |
| 4.5.9 | Gene segmentation | 137 |
| 4.5.10 | Differential epigenetic profiles | 138 |
| 4.5.11 | Signal quantification and scaling | 138 |
| 4.5.12 | Annotation with GO-terms | 139 |
| 4.5.13 | Mining for genes associated with EMT | 140 |
| 4.5.14 | Functional similarity scores | 140 |
| 4.5.15 | Selection of optimal clustering | 141 |
| 4.5.16 | Clustering of gene and enhancer loci | 142 |
| 4.5.17 | TF-binding sites within promoters and enhancers. | 142 |
| 4.6 | Chapter acknowledgements | 142 |
| 5 | The network architecture of the <i>Saccharomyces cerevisiae</i> genome | 144 |
| 5.1 | Introduction | 144 |
| 5.2 | Results and discussion | 146 |
| 5.2.1 | Inter-chromosomal cliques replicate early, and are enriched for cohesin | 146 |
| 5.2.2 | Community detection | 154 |
| 5.2.3 | Community detection is robust to interaction noise | 155 |
| 5.2.4 | Inter-chromosomal network has three major compartments | 156 |
| 5.2.5 | Subcommunities of the inter-chromosomal network are modular | 159 |
| 5.2.6 | The complete network highlights high-level organization | 165 |
| 5.2.7 | Conclusion | 169 |
| 5.3 | Methods | 169 |
| 5.3.1 | Data sources and processing | 169 |
| 5.3.2 | Network construction and clique/community detection | 171 |
| 5.3.3 | Enrichment analyses | 171 |
| 5.4 | Chapter acknowledgements | 172 |
| 6 | Conclusion | 173 |
| 6.1 | Learning from large histone modification datasets | 174 |
| 6.2 | Refining perspectives on chromatin regulation during EMT | 177 |
| 6.3 | Future approaches for network analysis of genomic interaction data | 180 |
| | References | 184 |

List of Figures

| | | |
|------|---|-----|
| 1.1 | Large and small-scale properties of a network | 36 |
| 2.1 | Flowchart of multilinear and MARS model construction | 43 |
| 2.2 | Comparison of predicted and observed gene expression | 46 |
| 2.3 | Gene expression false color plots | 50 |
| 2.4 | MARS response plots | 54 |
| 2.5 | Box plots of MLM and MARS knockouts | 61 |
| 2.6 | Box plots of amplitudes across expression | 63 |
| 2.7 | Box plots of predicted gene expression before and after knockout . . . | 64 |
| 2.8 | Enriched sites across MLM knockout quintiles | 65 |
| 2.9 | Relative error of mark enrichment models | 78 |
| 3.1 | Illustration of enrichment estimation methods | 90 |
| 3.2 | Comparison of enrichment estimation methods by MARS model statistics | 93 |
| 3.3 | Example of a highly enriched 5' region on a large gene | 96 |
| 3.4 | Example of 5' enrichment overlapping the 3' end of a neighboring gene | 96 |
| 3.5 | Average histone modification enrichments stratified by gene length . . | 99 |
| 3.6 | Gene expression stratified by gene length | 101 |
| 4.1 | Experimental design and data | 108 |
| 4.2 | Correlation of histone modifications at enhancers | 109 |
| 4.3 | Gene segmentation and differential signal quantification | 110 |
| 4.4 | EMT-related gene clusters (EMT-GCs) are differentially expressed and show antipodal patterns of chromatin remodeling | 111 |
| 4.5 | Epigenetic clustering groups functionally similar genes and identifies EMT-related clusters | 114 |
| 4.6 | Heat map of differential enhancer clusters | 122 |
| 4.7 | Activation and repression of enhancers correlates with changes in gene expression | 123 |
| 4.8 | Activated and repressed enhancers associated with EMT-GCs and different sets of transcription factors | 124 |
| 4.9 | AP-1 and c-Myc binding site enrichment in gene clusters via enhancers | 126 |
| 4.10 | Evidence for broad feedback regulation by AP-1 and NF- κ B family members, and c-Myc | 127 |
| 5.1 | Cohesin enrichment vs. inter-chromosomal maximal clique size | 148 |

| | | |
|------|---|-----|
| 5.2 | Cohesin loader enrichment vs. inter-chromosomal maximal clique size | 149 |
| 5.3 | Cohesin enrichment vs. intra-chromosomal maximal clique size | 150 |
| 5.4 | Cohesin loader enrichment vs. intra-chromosomal maximal clique size | 151 |
| 5.5 | Replication timing vs. inter-chromosomal maximal clique size | 152 |
| 5.6 | Replication timing vs. intra-chromosomal maximal clique size | 152 |
| 5.7 | Expression vs. inter-chromosomal maximal clique size | 153 |
| 5.8 | Community partition of the inter-chromosomal network | 158 |
| 5.9 | Intermediate solutions to community detection in the inter-chromosomal network | 160 |
| 5.10 | Modularity of the inter-chromosomal communities | 162 |
| 5.11 | Partition of the inter-chromosomal centromeric community | 163 |
| 5.12 | Partitions of inter-chromosomal community 6 and 0 | 164 |
| 5.13 | Level 0 partition of the complete network | 167 |
| 5.14 | Partition of the complete network | 168 |

List of Tables

| | | |
|-----|---|-----|
| 2.1 | Multilinear model terms and statistics | 48 |
| 2.2 | MARS model terms | 52 |
| 2.3 | Impact of marks in MARS model | 55 |
| 2.4 | Impact of two marks in MARS model | 56 |
| 2.5 | Impact of three marks in MARS model | 57 |
| 2.6 | ML model knockout analysis | 59 |
| 2.7 | MARS model knockout analysis | 60 |
| 2.8 | Distribution of mark amplitudes | 70 |
| 3.1 | CV(RMSD) for whole gene templates plus a 2000 bp intergenic overhang | 91 |
| 4.1 | GO-terms most significantly enriched in GC16 | 112 |
| 4.2 | GO-terms significantly enriched in > 4-fold upregulated genes | 113 |
| 4.3 | Referenced GO-terms enriched in the EMT-GCs | 116 |
| 4.4 | Referenced pathways enriched in the EMT-GCs | 118 |
| 5.1 | Transcriptional regulators significantly enriched in subcommunity 25 . | 165 |

Chapter 1

Introduction

Large-scale datasets are a relatively new standard in the field of molecular biology. The ability to generate gigabytes of genomic data in a single experiment has vastly increased the necessity and potential of new data analysis techniques to make new discoveries. The subfield of chromatin biology in particular has seen tremendous development as a result of this data revolution. A wide array of DNA sequencing-based assays have enabled unbiased quantitative characterization of the composition and conformation of chromatin. These advances have driven an expansion in scope from locus-specific to genome-wide studies. The results of these studies have given broad insight into how eukaryotic genomes are regulated. Importantly, these insights have required the development and novel application of diverse and sophisticated data analysis methodologies.

This dissertation will examine several case studies in the application of data analysis techniques to whole-genome chromatin datasets; specifically, multi-dimensional histone modification, and Hi-C datasets. These studies will examine basic questions about the regulatory activities of chromatin in a general context, and in the context of human disease. The remainder of the Introduction will provide the requisite scientific, historical, and technical background for the subsequent chapters.

1.1 Genomic regulation at the level of chromatin

1.1.1 Overview

Eukaryotic organisms store DNA molecules that are on the order of meters in total length into nuclei that are on the order of micrometers in diameter [Alberts et al., 2002]. This remarkable compaction occurs while DNA is in complex with many different proteins, in a state known as *chromatin*. Principle among these proteins are the highly conserved histone family of proteins. An octamer composed of two subunits each of the histone proteins H2A, H2B, H3, and H4, in complex with approximately 146 bp of DNA form a canonical nucleosome—the fundamental constituent of chromatin [Luger et al., 1997]. The chemical and physical regulation of chromatin at the scale of nucleosomes is critical for the proper regulation of chromosomal DNA in the dense nuclear environment.

The astounding complexity and robustness of the cell requires that the information encoded in the genome be accessed in a controlled and coordinated manner. The functional regulation of chromatin is a critical mechanism for tuning the accessibility of the information in the underlying DNA sequence. There are a variety of processes that govern this regulation, including covalent modifications to histones and DNA, deposition of histone variants, nucleosome positioning, and the physical enforcement of specific spatial conformations by proteins bound to chromatin [Berger et al., 2009, Guillou et al., 2010]. The composite of these features at a given genomic locus defines the “state” of chromatin at that locus. Covalent modifications to chromatin influence its local electrochemical and steric properties, which has consequences for its compaction (and thereby its stiffness), nucleosome occupancy, and preference for binding various proteins [Bannister and Kouzarides, 2011]. These modifications are regulated dynamically by proteins in the milieu. Thus, the crosstalk between chromatin

states and the milieu in which the genome exists determines how the information stored in DNA is accessed and expressed [Johnson and Dent, 2013]. In this way, modifications to chromatin are in principle critical for the emergent organization of genomic information.

At a basic level, chromosomal DNA has two roles: a template for self-duplication during replication, and a template for the production of RNA during transcription. As a transcriptional template, DNA serves as a dictionary of possible RNA outputs. Chromatin states alter the accessibility of transcribed regions at a given locus, and thus set bounds on the transcriptional state space of the cell. Thus, the regulatory states of chromatin help enforce “grammars” for the spatio-temporal expression of RNA information that are suitable for the maintenance of life. Chromatin regulation also imposes spatio-temporal control over DNA replication; however, the global functional consequences of this regulation are somewhat less clear than in the case of transcription. However, it has been suggested that chromatin regulation enforces spatial conformations of the genome that enable efficient replication [Guillou et al., 2010].

Increasing evidence suggest that the three-dimensional conformation of the genome has important functional consequences, including a major role in cell type specification. Some of the first evidence for cell-type-specific conformations was the observation that long-range enhancer elements across human cell types show extremely high variability in histone posttranslational modifications (PTMs), relative to genic loci. Moreover, these variations correspond to cell-type-specific gene expression profiles [Heintzman et al., 2009]. There is considerable evidence that these enhancers are involved in the long-range regulation of gene expression by physically interacting with gene promoters via chromatin loops [Chepelev et al., 2012, Li et al., 2012a]. Furthermore, enhancer activity corresponds to the modification states of their histones [Creyghton et al., 2010]. Recent studies have also demonstrated that the genomic regions that are near

in space tend to have similar patterns of histone modifications [Khrameeva et al., 2012]. Together these findings suggest a model where cell-type-specific gene expression is regulated by physical interactions with enhancers; and permissible spatial interactions are influenced by the modification states of histones at enhancer loci. This model is further supported by studies that show the stability of genomic interactions within cell types relative to comparisons across cell types [Noordermeer et al., 2011, Wang et al., 2013].

Chromatin regulation is often discussed in the context of *epigenetics*. Classically defined—in the tradition set by Conrad Waddington—“An epigenetic trait is a stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence” [Berger et al., 2009]. Furthermore, these traits can be heritable through meiosis or mitosis. Thus, epigenetic mechanisms are responsible for passing certain traits across generations, and for the maintenance of cellular phenotypes across divisions (e.g., maintaining tissue identity over time). Mechanisms of chromatin regulation are often referred to as “epigenetic,” although this may not be true in the strict sense of the classical definition. For example, specific histone PTMs and variants can be transiently involved in DNA damage repair, and thus, are not likely to convey heritable traits [van Attikum and Gasser, 2009]. Conversely, there are examples of genomic regulation by histone PTMs (and accompanying enzymes), such as gene silencing across generations, that are in agreement with the classical definition *per se* [Hall et al., 2002, Francis et al., 2009].

Within the chromatin field, there are varying opinions on the usage of the term “epigenetic,” ranging from the aforementioned definition, to general mechanisms of chromatin regulation that influence phenotype and genome function [Ptashne, 2013]. Due to the ubiquity of the latter usage in the current literature, this dissertation will adhere to this less strict (and more modern) definition.

1.1.2 Histone modifications

Posttranslational modifications to histones, in the form of acetylation and methylation, were first discovered in the 1960s, and were at that time postulated to influence transcription of RNA [Allfrey et al., 1964]. Today we know that the types of histone PTMs that exist extend well beyond acetylation and methylation, and include phosphorylation, deimination, O-GlcNAcylation, ADP ribosylation, ubiquitylation, and sumoylation [Bannister and Kouzarides, 2011]. Enzymes, known generally as “readers,” “writers,” and “erasers,” are respectively responsible for binding to, depositing, and removing these PTMs. Each PTM has an influence on chromatin structure and/or the recruitment of enzymes that influence chromatin-based processes. Furthermore, these PTMs can occur at many different sites on the histone octamer in the nucleosome, principally on the unstructured histone N-terminal tails. The large number of modification types and possible modification sites leads to an enormous number of combinatorial possibilities for histone PTMs, each with potentially unique functional consequences. This observation lead to the so-called “histone code hypothesis,” which posits that histone PTMs act in a combinatorial and/or sequential manner to regulate a variety of processes on chromatin [Strahl and Allis, 2000, Jenuwein and Allis, 2001]. Much of what is presented in this work will leverage this hypothesis; however, despite the diversity of modification types, this work will focus on histone acetylation and methylation, as they are the best characterized histone PTMs.

Acetylation and methylation

Histone acetylation occurs most often on the lysine residues of the N-terminal tails of histones, although it can also occur on lysine residues in the globular portion of histone proteins [Tjeertes et al., 2009]. Acetylation of these residues is a dynamic process mediated by histone acetyltransferases (HATs), and histone deacetylases (HDACs).

HATs fall into two major categories: type-A (of which there are subcategories), and type-B. The major functional difference between the two types is that type-A HATs acetylate histones when they are integrated into chromatin, whereas type-B HATs acetylate free histones in the cytoplasm. Acetylation by type-B HATs is important for integrating newly synthesized histones into chromatin [Parthun, 2007]. The enzymatic activity of type-A HATs plays a major role in the regulation of chromatin, by dynamically altering the physical interactions between histone proteins and DNA. This dynamic regulation also requires HDAC activity, which opposes the activity of HATs.

Acetylation neutralizes the positive charge of lysine, and thereby destabilizes the interaction between histone proteins and negatively charged DNA. It is thought that this destabilization permits a more open and active chromatin structure, providing transcriptional machinery and other DNA-binding proteins greater access to DNA. Indeed, the eviction of hyperacetylated histones from DNA, *in vivo* and *in vitro*, is well established [Li et al., 2007, Chandy et al., 2006, Zhao et al., 2005]. Various genome wide studies also support the “activating” role of histone acetylation in a correlative fashion. For example, strong genome-wide correlations between gene expression and promoter enrichment of various histone acetylations and are well documented [Roh et al., 2006, Wang et al., 2008]. Acetylation of lysine 16 on histone H4 (H4K16ac) has also been shown to prevent the compaction of chromatin, which further supports the role of histone acetylation in maintaining open and active chromatin [Shogren-Knaak et al., 2006].

Methylation of histones occurs on lysine and arginine residues, and like acetylation, tends to occur on the N-terminal tails. Methylation is mediated by histone methyltransferases (HMTs) and histone demethylases (HDMs). Histone methylation was once thought to be irreversible, however, in the early 2000s this dogma began to be called into question [Bannister et al., 2002]. And in 2004, the first lysine-specific

HDM (*LSD1*) was discovered [Shi et al., 2004]. Since this initial discovery, several other lysine-specific HDMs have been identified [Tsukada et al., 2006, Whetstine et al., 2006]. By contrast, mechanisms of arginine demethylation have been more elusive. Although *JMJD6* has been reported to be an arginine HDM [Chang et al., 2007], these results have not been repeated, and no other arginine-specific HDMs have been found [Bannister and Kouzarides, 2011]. Arginine methylation can be antagonized by citrullination of arginine, however this is not strictly a reversal of methylation [Cuthbert et al., 2004, Wang et al., 2004]. Nevertheless, histone methylation is now recognized to be a dynamic process, but is comparably more stable than acetylation. Unlike acetylation, methylation does not affect the charge of the modified amino acid side chain. Methylation is also fundamentally more complicated than acetylation in that lysine residues can be mono-, di-, or trimethylated, and arginine residues can be mono- or dimethylated. Additionally, dimethylation of arginines can be symmetric or asymmetric [Bannister and Kouzarides, 2011].

Consistent with the higher degree of biochemical complexity, histone methylation also has a more complex relationship with chromatin activity than histone acetylation. For example, the well-studied modification H3K4me3 is enriched at the promoters of activated genes in a punctate fashion; whereas, H3K27me3 shows broad enrichment over transcriptionally silent heterochromatic regions [Barski et al., 2007]. A recent study has also demonstrated that in some cases of punctate enrichment at gene promoters, H3K27me3 has a positive correlation with gene expression [Young et al., 2011]. Furthermore, both H3K4me3 and H3K27me3 are mutually exclusive with their acetylated counterparts, and there is evidence for regulatory tension between the two states [Tie et al., 2009, Guillemette et al., 2011].

These examples highlight the complex nature of histone methylation with respect to chromatin regulation. The methylation of two different residues of the same histone tail can have antipodal correlations with gene expression. Moreover, the correlation of

a given modification with gene expression can also be dependent on the distribution of the modification in genomic space. These complexities are compounded by the combinatorial context in which the modifications occur.

Combinations and dependence

One of the first concrete examples of histone modifications in a functional combinatorial context was the discovery of so-called “bivalent domains” [Bernstein et al., 2006]. These domains are characterized by broad enrichments of H3K27me3 that overlap more punctate H3K4me3 enrichment at developmentally important genes that are expressed at low levels in human embryonic stem cells (hESCs). Loss of H3K27me3 is observed upon differentiation, which coincides with the upregulation of the genes with which the bivalent domains are associated. The prevailing interpretation of this phenomenon is that the bivalent domains keep genes silent, but poised, enabling rapid induction during differentiation. This induction is believed to be triggered by loss of the repressive mark.

Similar poised and activated states have been observed at promoter-distal enhancer loci. Early whole-genome studies have shown that enhancer elements can be identified by an abundance H3K4me1 [Heintzman et al., 2007]. Later studies demonstrated that active enhancers can be distinguished from poised enhancers by the presence of H3K27ac [Creyghton et al., 2010]. In hESCs, poised and active enhancers were identified by the presence or absence of the mutually exclusive H3K27me3 and H3K27ac marks; where poised, but inactive enhancers are marked by H3K4me1 and H3K27me3, and active enhancers are marked by H3K4me1 and H3K27ac [Rada-Iglesias et al., 2011]. In this study, active enhancers were linked to genes expressed in hESC, whereas poised enhancers were linked to genes involved in early embryogenesis. The presence of H3K36me3 and H3K9me3 has also been used to classify active and poised enhancers, respectively [Zentner et al., 2011].

The sequential gain and loss of histone PTMs at multivalent domains alludes to the dependence relationships in histone PTM deposition. Indeed, there are several examples of crosstalk and dependencies between histone modifications [Latham and Dent, 2007, Kouzarides, 2007, Bannister and Kouzarides, 2011]. In light of these relationships, the regulation of chromatin by histone modifications can in part be considered a directed network, where connections between histone modifications represent dependence relationships. An example of a small dependence network, which is conserved from yeast to humans, is the requirement of H2BK123 ubiquitylation by the Rad6-Bre1 complex, for the methylation of H3K4 by the COMPASS complex, and the methylation H3K79 by Dot1 [Lee et al., 2007, Kim et al., 2009]. Dependence structures like this greatly limit the possible number of chromatin states at a given genomic locus, thus imposing some constraints on what is otherwise an astronomically large state space.

Even with these constraints, conventional low-throughput biochemical techniques alone have little hope of fully deconvolving the complexity of synergistic and dependent relationships among histone modifications (i.e., validating or deciphering the histone code). High-throughput assays coupled with computational data analysis methods will be required to identify probable relationships between histone modifications in an unbiased manner. Conventional biochemical techniques can then be used to validate and refine the computational analyses. This process can be iterated to converge on true relationships in the histone code.

Human disease

In humans, many enzymes that modify histones have been linked to cancer and a variety of developmental disorders. Though this work will only focus on chromatin modifications as they relate to cancer, it is worth noting that mutations in genes that encode several well-known histone modifying enzymes have been linked to de-

developmental disorders that are accompanied by mental retardation. These genes include *EHMT1*, *CREBBP*, *EP300*, *MLL2*, *NSD1*, *NSD2*, *GLP*, and *JARID1C*, among others [van Bokhoven and Kramer, 2010, Butler et al., 2012]. Interestingly, many of the genetic mutations in these histone modifying enzymes cause developmental disorders when acquired in the germline, but lead to cancer when acquired in somatic cells [Butler et al., 2012]. There is great interest in understanding the etiological role of these enzymes in cancer, due to the great potential for designing drugs that target them. Indeed, several HDAC inhibitors are FDA approved for the treatment of several hematological and solid cancers, and trials for many more are ongoing [Khan and La Thangue, 2012]. There is a significant effort to expand epigenetic anticancer drugs to target different types of chromatin modifying enzymes [Arrowsmith et al., 2012].

Dysregulated histone modification mechanisms are believed to contribute to carcinogenesis by enabling carcinogenic gene expression profiles, and/or influencing genomic structural integrity [Bannister and Kouzarides, 2011]. For example, both the over- and underexpression of *EZH2* has been shown to contribute to carcinogenesis. A member of the developmentally important polycomb group proteins, *EZH2* is responsible for catalyzing the trimethylation of histone H3, lysine 27 (H3K27me3), which leads to transcriptional silencing. The elevated catalytic activity of *EZH2* has been shown to promote oncogenesis in a variety of solid tumors [Varambally et al., 2002, Kleer et al., 2003, Wang et al., 2012]. However, there are also examples of loss-of-function mutations in *EZH2* which promote myeloid malignancies [Ernst et al., 2010]. In each of the cases, the activity of *EZH2* is associated with the derepression or repression of a cohort of genes. The seemingly contradictory roles of *EZH2* as a oncogene and tumor suppressor, highlight the importance of stable *EZH2* activity in maintaining non-oncogenic gene expression programs.

Cancer progression is often associated with the dysregulation of mechanisms that are important in development, which is the case with *EZH2* activity [Hanahan and

Weinberg, 2011, Izrailit and Reedijk, 2012]. Therefore, the link between histone modifications and carcinogenesis is not altogether surprising, given that developmental process have been shown to be accompanied by dramatic changes in histone modification profiles. Indeed, different cell types in general have been shown to have unique histone modification profiles [Heintzman et al., 2009]. These findings give rise to the hypothesis that the ability to reprogram chromatin at the level of histone modifications may be necessary to enable stable changes in cellular phenotype. Cancer cells undergo selection in acquiring drug resistance, which constitutes a phenotypic shift, implying the acquisition of changes in histone PTM patterns [Podlaha et al., 2012]. Though several specific mechanisms have been investigated [Khan and La Thangue, 2012], this provides a broad rationale for why HDAC inhibitors are effective in the treatment of cancer. Presumably, with a reduced ability to reprogram their epigenomes, cancer cells lose some of the phenotypic plasticity that they require to maintain their fitness.

One of the hallmark phenotypic shifts that occurs during cancer progression is the epithelial-mesenchymal transition (EMT). This process occurs normally during development and wound healing, but also pathologically in cancer progression. Cancer cells that undergo EMT are more stem-like, resistant to apoptosis, and are able to escape the epithelium and invade into different tissues [Thiery et al., 2009, Polyak and Weinberg, 2009]. Not surprisingly, this de-differentiation and change in phenotype has been associated with many different types of epigenetic reprogramming, including changes in histone PTMs, DNA methylation, and microRNA abundance [Wu et al., 2012, Stadler and Allis, 2012, Wang et al., 2013].

There are several examples of histone PMTs directly involved in EMT. Canonically, the so-called "master switch" transcription factors, including Twist, Snail, Slug, and ZEB2, are responsible for regulating genes that are critical for the transition. Snail and Slug have been shown to recruit the Sin3a/HDAC1/HDAC2 complex to the promoter of *CDH1*, which leads to transcriptional silencing of the gene through deacetylation

of the promoter [Bolós et al., 2003, Peinado et al., 2007]. Loss of E-cadherin, which is the protein that is encoded by *CDH1*, is a hallmark of EMT. This demonstrates that histone acetylation regulates critical parts of the transcriptional program that distinguishes the epithelial and mesenchymal phenotypes.

In another study it was shown that in TGF β -induced EMT, there are bulk changes in several histone modifications, including global losses of H3K9me2, and gains in H3K4me3 and H3K36me3 abundance [McDonald et al., 2011]. These changes were found to be largely localized to large organized heterochromatin H3K9 methylated regions (LOCKS). Furthermore, it was found that many of the epigenetic and phenotypic changes that occur after the induction of EMT are dependent on the lysine-specific histone demethylase *LSD1*. The epigenetic reprogramming of cancer cells during EMT will be further explored in Chapter four.

The mechanistic investigation of histone modifying enzymes and effector proteins in disease processes is a highly active field of research. As previously stated, the interest in this field is driven largely by the druggability of these protein targets. In addition to this pragmatic motivation, the recent development of whole-genome methods to interrogate a wide variety of chromatin features has led to deep insights into the relationship between chromatin regulation and cellular phenotype. Although they contain rich information about chromatin states, histone modifications represent only one aspect of a panoply of features that determine genomic regulation. In order to gain a more holistic view of the relationship between chromatin regulation and phenotype, diverse whole-genome datasets must be considered simultaneously.

1.1.3 Higher-order chromatin organization

Although many fantastic insights have come from studying chromatin in the context of the linear genome, it has become clear that a great deal of regulatory information is encoded the higher-order structure of chromatin. The term “higher-order” refers

to non-random spatial arrangements of chromatin. The non-random organization of interphase chromosomes was first observed by Carl Rabl in the late 19th century [Rabl, 1885, Cremer and Cremer, 2010]. Since this initial observation, details of chromosomal organization at many different resolution scales have been described. Importantly, functional roles have been ascribed to each of these levels of organization.

At a coarse-grained level, the genome organizes into chromatin territories (CTs), where interphase chromosomes occupy semi-distinct spaces within the nucleus [Cremer et al., 1982, Cremer and Cremer, 2010]. In finer detail, metazoan chromosomes organize into fractal globule conformations, which can be described as a polymer that shows regions of local collapse. Fractal globule polymers have a power law decay in contact probability as a function of distance on the polymer chain. This decay has been empirically validated in the genomes of several organisms [Lieberman-Aiden et al., 2009, Sexton et al., 2012]. At still finer resolution, there are structures known as “topologically associating domains” (TADs) [Nora et al., 2012, Dixon et al., 2012, Dekker et al., 2013]. At the individual locus level, chromatin can organize into the so-called “30 nm fiber” structure, which is speculated to be the typical conformation of heterochromatin [Grigoryev and Woodcock, 2012]. There are also long-range functional interactions, such as enhancer-promoter interactions, which can be on the scale of megabases in distance [Li et al., 2012a].

Understanding the molecular mechanisms that mediate higher-order structures is an active area of research. Notably CTCF, and the cohesin and condensin families of proteins have been implicated as major structural components of higher-order chromatin conformations (see [Wood et al., 2010] for a review). However, both molecular-level and systems-level details of how these proteins mediate global chromatin conformation are lacking. Despite the mechanistic uncertainties in the formation of higher-order structures, chromatin organization at every level described above has been linked to functional processes, such as transcription and DNA replication [Schneider

and Grosschedl, 2007, Lieberman-Aiden et al., 2009, Ryba et al., 2010, Dixon et al., 2012, Li et al., 2012a].

Chromosome territories

The organization of chromatin at the level of CTs was one of the earliest higher-order arrangements identified [Cremer et al., 1982]. Among eukaryotes, there is notable variability in how CTs manifest. For example, the interphase budding yeast genome has a rosette structure, where the centromeres of all chromosomes pack densely in one area near the nuclear periphery, from which the chromosome arms emanate [Jin et al., 2000, Bystricky et al., 2004, Duan et al., 2010]. At the same level of organization, mammalian genomes in interphase have a more amorphous structure, but develop rosette conformations at prometaphase [Bolzer et al., 2005]. Furthermore, heterochromatin and genes transcribed at low levels tend to be associated with the nuclear lamina, whereas genes that are situated toward the interior of the nucleus are more highly transcribed [Schneider and Grosschedl, 2007]. These examples represent the most coarse-grained functional arrangements of chromatin.

Enhancers

Enhancers are a type of *cis*-regulatory element that positively regulates gene expression. Typically, they are hundreds of base pairs in size and are dense with transcription factor binding sites. They can be located in intronic or intergenic regions; within, adjacent to, or hundreds of kilobases away from their target genes. The activity of long-range (i.e., not adjacent to its target promoter) enhancers requires the looping of chromatin so that the enhancer can be in close spatial proximity to its target promoter. Presumably, enhancers recruit and localize transcription factors and essential transcriptional machinery to the promoters of their target genes, thereby facilitating active transcription. One of the first studies showing direct evidence of these long-

range interactions reported the association of an enhancer element with the β -globin (*HBB*) locus located 50 kb away [Carter et al., 2002]. High-throughput methods have since shown that similar long-range interactions are ubiquitous throughout the genome [Chepelev et al., 2012, Li et al., 2012a].

A particularly striking example of a functional long-range enhancer is the enhancer of *SHH*, which is located approximately 1 Mb away in an intron of the *LMBR1* gene. Mutations in the enhancer were shown to cause misexpression of *SHH*, which results in the development of preaxial polydactyly [Lettice et al., 2002]. It was later demonstrated that there is in fact a chromatin loop that allows the enhancer in *LMBR1* to spatially co-localize with the promoter of *SHH* [Li et al., 2012a]. This study also demonstrated the interaction of a intergenic non-coding region with the promoter of *IRS1*. In an earlier genome-wide association study (GWAS), a single nucleotide polymorphism (SNP) in the non-coding region was shown to be associated with an increased risk of diabetes and coronary artery disease [Kilpeläinen et al., 2011]. Though *IRS1* is known to be involved in type-2 diabetes, the GWAS was unable to associate the SNP with *IRS1* due to the distance between the gene and the SNP. Numerous GWAS efforts have demonstrated similar disease associations with SNPs in unannotated intergenic space, with little evidence of what role these SNPs may play in the etiology of the associated disease. It is likely that many of these isolated disease-associated SNPs are in long-range enhancer elements [Visel et al., 2009].

As previously described, histone modifications are involved in modulating enhancer activity, although they have not been shown to mediate the spatial interactions *per se*. There are, however, studies that have shown that on average, genomic loci that interact through space show similar patterns in histone modifications [Khrameeva et al., 2012]. Furthermore, the tissue specificity of long-range interactions, and their importance in maintaining tissue specific gene expression programs, is becoming increasingly appreciated [Ong and Corces, 2011]. These interactions are also being

implicated in the regulation cancer-specific gene regulation [Wang et al., 2013]. These findings, coupled with the dramatic phenotypic effects of abnormal enhancer function, highlight the practical importance of understanding the relationships between histone modifications and long-range enhancer function.

The fractal globule and intermediate structures

The identification of finer-scale long-range interaction structures has benefited greatly from the chromosome conformation capture family of assays, which includes 3C, 4C, 5C, and Hi-C, as well as the related ChIA-PET assay (Hi-C and ChIA-PET will be discussed further in the next section) [Dekker et al., 2002, Simonis et al., 2006, Dostie and Dekker, 2007, Lieberman-Aiden et al., 2009, Fullwood et al., 2009a, Fullwood et al., 2009b]. These methods have facilitated the analysis of chromatin interactions at varying degrees of throughput—from a single pair of loci (3C), to all interaction pairs throughout the genome (Hi-C). The fractal globule conformation of the genome was identified using genome-wide interaction data generated by the Hi-C method [Lieberman-Aiden et al., 2009]. In that study, the genome was shown to broadly compartmentalize into high-interaction and low-interaction regions, corresponding to open and closed chromatin, respectively. Quantification of the interaction states of chromatin were later shown to correlate very strongly with genome-wide replication timing maps [Ryba et al., 2010]. The fractal globule conformation of the genome was inferred from the specific patterning of open and closed chromatin states. It is important to note that the fractal globule only provides a high-level model for the arrangement of chromatin. Some details of this model have been brought into question [Sexton et al., 2012].

Beneath the fractal globule in the hierarchy of higher-order chromatin structure, chromosomes are divided into discrete regions of topologically associating domains (TADs) [Nora et al., 2012, Dixon et al., 2012]. These regions show a high frequency of within-TAD interactions, but are relatively insulated from neighboring regions. TADs

can be hundreds of kilobases in length, and are a fundamental structural building block of chromosomes [Dixon et al., 2012, Sexton et al., 2012]. Constraints on long-range interactions seem to be imposed by TADs. These interaction constraints suggest that they may be involved in regulating enhancer-promoter interactions. Indeed, there is some evidence that this may be the case [Shen et al., 2012]. This implicates TADs as a critical factor in the formation of substructures involved in transcriptional regulation. In support of this model, CTCF is enriched at the boundaries between TADs [Dixon et al., 2012]. CTCF is a protein whose function has long been associated with the formation of insulating boundaries between long range interactions. Many genome-wide studies have suggested that the function of CTCF goes well beyond the insulation of enhancer-promoter interactions, which may only be an indirect consequence of CTCF activity. Nevertheless, its enrichment at boundaries between TADs supports the view that TADs are functional units of higher-order chromatin structure.

There are examples in many different organisms of spatial clustering of functional genomic elements in the nucleus. As previously mentioned, centromeres in budding yeast cluster in space. There are also two clusters of tRNAs in budding yeast: one near the cluster of centromeres, and another near the nucleolus [Duan et al., 2010]. In higher eukaryotes, co-expressed genes sometimes co-localize into structures known as transcription factories [Rieder et al., 2012]. The first evidence of transcription factories came from microscopy studies that observed that the majority of transcription in the cell occurs at several hundred punctate foci in the nucleus [Jackson et al., 1993]. At the time it was thought that transcription would be distributed somewhat uniformly throughout the nucleus. In a study of human umbilical vein endothelial cells, the promoters of spatially separated genes that are induced by TNF were found to co-localize upon TNF treatment [Papantonis et al., 2010]. This raised the intriguing possibility that perhaps transcription factories are localized to a fixed point in space to which genes are recruited. It also supported the idea that co-regulated genes tend

to occupy the same transcription factories. Indeed, it was later shown that in mouse erythroid cells, genes that are regulated by KLF1 group into active KLF1-enriched transcription factories [Schoenfelder et al., 2010]. Though there is considerable evidence supporting the existence of transcription factories in higher eukaryotes, many details of how they form and function remain unknown [Rieder et al., 2012].

A broad and outstanding goal in the chromatin field is to understand how higher-order structures form, and more specifically, how modifications to chromatin influence this organization. This goal is feasible, principally due to the development of high-throughput methods to analyze chromatin composition and conformation. Some of these methods will be discussed in the following section.

1.2 High-throughput methods in chromatin biology

1.2.1 Overview

In the last 20 years, high-throughput genomic technologies have rapidly transformed the field of molecular biology into a data-intensive discipline. Chief among these transformative technological capabilities is the ability to sequence large amounts of DNA rapidly and cost-effectively. Several chromatin-based assays, such as chromatin immunoprecipitation (ChIP), have been coupled with DNA sequencing (i.e., ChIP-seq) to provide quantitative genome-wide maps of chromatin states. In addition to ChIP-seq, which provides genome-wide maps of proteins in complex with DNA, there are a variety of other sequencing-based methods that quantify other properties of chromatin. RNA transcription can be quantified using RNA-seq, CAGE, and RNA-PET. Transcription factor binding sites can be identified using ChIP-seq and DNase-seq. DNA methylation can be quantified using MeDIP-seq. Long-range interactions can be identified using

Hi-C and ChIA-PET. Chromatin structure, including nucleosome occupancy and DNA accessibility, can be assayed with DNase-seq, FAIRE-seq, ChIP-seq and MNase-seq. This list is not a comprehensive catalog of sequencing-based chromatin assays, but it represents some of the principle techniques used in the Encyclopedia of DNA Elements (ENCODE) project [Dunham et al., 2012], which is the zeitgeist of chromatin biology with respect to high-throughput methods. The ENCODE project was designed to provide an enormous collection of datasets that profile chromatin in a variety of ways in many different cell types. The hope of ENCODE is that these data will be used as a resource to broadly facilitate the understanding of genomic regulation at the level of chromatin. With contentious claims, and costs approaching \$200M, this effort is not without its controversies [Graur et al., 2013, Eddy, 2013]; however, ENCODE symbolizes the revolution in chromatin biology sparked by high-throughput genomic technologies.

This revolution began with the development of oligonucleotide microarrays in the mid-1990s [Schena et al., 1995, Lashkari et al., 1997]. Array (chip) technologies facilitate the high-throughput quantification of DNA molecules by hybridizing them to a glass slide coated with a lawn of known, genome-matched oligonucleotide sequences. Variations on array technology can be used for quantifying transcriptional output, or can be coupled with ChIP to quantify levels of protein-DNA complexes (ChIP-chip). Shortly after the development of array-based technologies, so-called massively parallel DNA sequencing was developed, which allowed researchers to sequence DNA molecules and quantify their abundance in a high-throughput manner [Brenner et al., 2000]. Currently, the falling cost and superior data quality of sequencing-based technologies are causing the displacement of array-based technologies.

1.2.2 ChIP-seq and Hi-C

All sequencing based chromatin assays follow a general procedure: (1) Isolate genomic fragments of interest, (2) sequence the fragments, and (3) map the sequences back to a reference genome. The chromatin assays mentioned in section 1.2.1 only vary significantly with respect to the first step. This section will discuss two of these variations in detail: ChIP-seq and Hi-C. The sequencing step is fairly straightforward, and is preformed by one of the many high-throughput sequencing platforms available (reviewed in [Pareek et al., 2011]). The output of a single sequencing “run” is a series sequence fragments typically ranging from 100 to 1,000 bp, with a total sequence output ranging from megabases to gigabases [Loman et al., 2012]. These fragments correspond to genomic loci that are enriched according to the protocol of the particular assay (e.g., ChIP-seq, Hi-C). Computationally mapping these sequences back to a reference genome provides quantitative, locus specific enrichments of the fragments that were sequenced.

The ChIP assay was developed to identify DNA-protein complexes within the nucleus [O’Neill and Turner, 1995, O’Neill and Turner, 1996]. Briefly, DNA is crosslinked and fragmented, after which an antibody is used to enrich for DNA fragments that are bound to a specific protein. Coupled with high-throughput sequencing, ChIP (i.e., ChIP-seq) generates genome-wide maps of proteins bound to DNA. Critically, ChIP-seq can and has been used to generate maps of a variety histone modifications [Barski et al., 2007, Wang et al., 2008]. In principle, any histone modification (or any DNA-bound protein), for which there is a specific antibody, can be mapped using ChIP-seq. Mapped ChIP-seq data amounts to integer values for each base across the genome, corresponding to the number of reads that mapped to the bases. Thus, ChIP-seq is an invaluable tool for the quantitative analysis of histone modifications in genome-wide chromatin regulation. Further details on quantitative methods will be

provided in section 1.3.

As mentioned in section 1.1.3, the chromosome conformation capture (3C) assay was developed to identify long-range interactions in the genome. The Hi-C assay is an extension of 3C that is able to interrogate all interactions throughout the genome in an unbiased manner [Lieberman-Aiden et al., 2009, Duan et al., 2010]. The basic method involves crosslinking DNA, digesting with a restriction enzyme, and enriching for fragments of DNA that were close enough in space to be crosslinked. This produces a library of fragment pairs that were spatially very near in the nucleus. After several rounds of ligation and digestion, linear fragments of DNA are produced from each pair. Each end of these fragments originates from a different member of the pair. Paired-end sequencing is then performed on these linear fragments, and the resulting sequences are mapped back to the reference genome. Pairs that map to distant positions in the genome (relative to the length of the sequenced fragment) represent genomic loci that are near in three-dimensional space, but not near in the linear sequence of the genome. The resulting pairs of interacting genomic regions can be used to construct three-dimensional models of the conformation of the genome [Lieberman-Aiden et al., 2009, Duan et al., 2010].

Clearly, having a reference genome sequence is critical for these methods. The initial sequencing of the human genome, which provided the first human reference genome, was published in 2001 [Lander et al., 2001]. Though the genomes of viruses and bacteria had been sequenced earlier, the sequencing of the human genome was a watershed event that pushed a large portion of the molecular biology field into the realm of data-intensive science. Indeed, analyzing datasets generated by ChIP-seq and Hi-C (and any other sequencing-based assay) is computationally challenging. However, analysis of these data have given deep insights into chromatin regulation. Some of the computational and mathematical techniques that are used for analyzing these data—especially those that are relevant to this dissertation—are described in the

following section.

1.3 Computational analysis of whole-genome chromatin data

1.3.1 Overview

The analysis of sequencing-based chromatin assays has several different stages and degrees of sophistication. The first and most formulaic steps are the initial quality control and mapping of the raw sequence data to the reference genome. The steps that follow usually involve gathering summary statistics and visualization of the data to gain some intuition about the data. Once basic characterizations of the data have been made, subsequent analysis approaches are highly dependent on the research goal. This section will discuss some of the important analysis methods that have been successfully applied to whole-genome chromatin data. While this field is vast, this section will have two main foci: (1) methods that have been applied to ChIP-seq data with the goal understanding the regulation of chromatin through histone modifications, and (2) methods that have been applied to Hi-C data with the goal of understanding higher-order chromatin organization.

Often, simple summary statistics of a ChIP-seq dataset are sufficient to address a specific hypotheses about the relationship between, for example, a particular histone modification and a specific genomic feature. These simple cases often arise when a single ChIP-seq experiment is considered, or where multiple experiments are considered, but are treated independently. However, this section, and indeed the majority of this dissertation, is concerned with analyzing histone modification ChIP-seq data in a high-dimensional context—where many different histone modifications are considered simultaneously. This approach allows direct investigation of the histone code hypothesis

from histone modification ChIP-seq data. Furthermore, it allows the study of various biological processes in the context of the histone code hypothesis. For example, Chapter four describes epigenetic reprogramming in the epithelial-mesenchymal transition with respect to combinatorial changes across a large panel of histone modifications. This section will briefly introduce some of the machine learning approaches that are used for this type of analysis, and present examples of their application to histone modification ChIP-seq data.

Hi-C and other large chromatin conformation datasets present different challenges. The problem of inferring genomic structure from contact frequencies between genomic regions is not straightforward. Accordingly, many diverse computational approaches have been applied to this problem, ranging from physical models and molecular dynamic simulations, to network analysis approaches. Basic approaches for identifying meaningful long-range interactions from Hi-C data will be discussed; however, physical models are beyond the scope of this work, and will not be included in this overview. Network-based methods will be the primary topic of Chapter five, and so a deeper discussion of network concepts will be presented in this section.

1.3.2 Machine learning applied to histone modification data

Machine learning is the branch of computational statistics that includes techniques that are designed to learn general relationships in a dataset that can be applied to new datasets. These techniques require a large number of observations in order to learn these relationships, and are thus well suited to analyze whole-genome datasets. Some of the machine learning techniques that have been successfully applied to histone modification ChIP-seq data include regression, classification, clustering, hidden Markov models, and bayesian networks. This section will briefly describe each of these techniques and their application to high-dimensional histone modification ChIP-seq datasets.

Regression and classification

Regression and classification are closely related supervised learning methods. Supervised methods generate models that map input data onto some “labeled” (i.e., defined) output data. In general, given a set of predictor variables \mathbf{x} , and response variables \mathbf{y} , supervised methods attempt to find a function $f: \mathbf{X} \rightarrow \mathbf{Y}$. The simplest example is univariate linear regression, where the goal is fitting a line to data in an x - y plane. The independent variable, x (called the predictor variable), is *labeled* with a value, y (called the response variable). Classification is an analogous procedure, with the major difference being that labels are categorical rather than quantitative. Models are constructed on “training data,” but can be applied to new, unobserved data. For example, in the case of univariate linear regression, the resulting linear equation can then be used to predict y for any x .

Regression methods are well suited to model the relationship between histone modifications and other genomic variables such as gene expression. Indeed, several studies have used regression to study this exact relationship, one of which will be the topic of Chapter two [Karlič et al., 2010, Xu et al., 2010]. These studies used quantified levels of histone modifications from ChIP-seq as predictor variables, and gene expression levels as response variables. They have convincingly demonstrated that histone modifications are highly predictive of gene expression. Furthermore, it has been shown that models built with data from one cell type are applicable to other cell types, demonstrating that the relationships between histone modifications and gene expression are general with respect to cell types [Karlič et al., 2010]. An interesting finding that will be discussed further in Chapter two is that some modifications that are not predictive of gene expression in a univariate context, are highly predictive in the context of many other modifications [Xu et al., 2010]. This finding suggests important combinatorial relationships between modifications—an idea that evokes concepts put forth in the histone code hypothesis [Strahl and Allis, 2000, Jenuwein

and Allis, 2001].

Classification methods have also been used extensively in the analysis of whole-genome chromatin data. Support vector machines (SVMs) are some of the most commonly used methods for this application. Briefly, SVMs attempt to discover an n -dimensional hyperplane that best classifies categorically labeled data points that have $n + 1$ dimensions; thus, generating a linear classifier. A general method using SVMs has been successfully applied to a panel of histone modification ChIP-seq profiles to identify enhancer elements [Fernández and Miranda-Saavedra, 2012]. This procedure effectively discovers histone modification profiles that predict the presence of enhancer elements. DNase hypersensitive sites have also been predicted using histone modifications in a similar way [Arvey et al., 2012]. SVMs and other classification methods are a powerful approach in determining the combinatorial patterns of histone modifications that are associated with underlying genomic features.

Cluster analysis

In contrast to regression and classification, cluster analysis (clustering) is an unsupervised learning technique; i.e., the data points are unlabeled. Many types of cluster analysis exist, including K-means, and hierarchical clustering. While the details of these methods will not be discussed here, the goal for all clustering techniques is to generate sensible groupings (clusters) of data points, where each cluster contains data points that are relatively homogenous with respect to their features. In biology, cluster analysis is commonly used to summarize and visualize genomic data sets. For example, cluster analysis is commonly used to group genes that show similar expression profiles across a series of experimental conditions.

Though clustering is a powerful approach for summarizing and visualizing epigenetic data, clustering can also give deep insights into the functional associations of the data when combined with other datasets (e.g., expression, annotations, Gene Ontology,

pathways). Clustering genomic regions in an unbiased way with respect to their histone modification profiles, groups them using only knowledge of their chromatin state. Discovering statistical associations between the resulting clusters and independent datasets is a convincing way implicate combinatorial histone modification profiles in functional processes. This strategy forms the basis for the study presented in Chapter four. Several studies have used similar strategies. For example, clustering genomic loci by histone modifications at various developmental stages in *Xenopus* revealed a stepwise progression in epigenetic reprogramming during development [Schneider et al., 2011]. Another study demonstrated dramatic reprogramming of enhancer-associated histones during the differentiation of human embryonic stem cells [Hawkins et al., 2011]. Clustering breast cancer samples by global levels of histone modifications revealed that levels of histones correlate with tumor phenotype, several prognostic markers, and patient outcome [Elsheikh et al., 2009]. Strikingly, this study only observed global histone modifications, highlighting the potential power of this strategy for identifying functional associations from high-resolution whole-genome data.

Hidden Markov models and dynamic Bayesian networks

Hidden Markov models (HMMs) and dynamic Bayesian networks are in a class of modeling techniques known as probabilistic graphical models. Generally speaking, these models attempt to discover the conditional dependence structure between random variables. Hidden Markov models attempt to discover an underlying sequence of “hidden” (unobserved) states of a system by sequentially analyzing output variables, i.e., the observed data. Since the states are unknown when training the model, HMMs are an example of unsupervised learning. The Markovian aspect of an HMM is that the state of the system at a given position in the sequence is only dependent on the observed data at the given position, and the state of the system at the previous position in the sequence. Dynamic Bayesian networks (DBNs) are a generalization

of HMMs that allow the hidden state of the system to be modeled as a series of interrelated random variables, rather than a single hidden state variable. Ultimately, the output of HMMs or DBNs is a sequence of states that are learned from the sequential observation of data.

In the context of unbiased whole-genome histone modification data, these methods segment the genome into states that are defined by the site-specific enrichment for each modification in the dataset used to train the model. Thus, an HMM or DBN discovers the chromatin states that exist across the genome. In a 2010 study by Ernst and Kellis, an HMM was used to segment the human genome based on 38 different histone modifications including acetylations, methylations, and histone variant H2A.Z [Ernst and Kellis, 2010, Barski et al., 2007, Wang et al., 2008]. The results were striking, in that a wide variety of functional genomic features were associated with distinct chromatin states discovered by the HMM. Some of these features include promoters, transcriptionally active regions, repressed regions, putative enhancers, and repeat sequences. This demonstrated that many annotated features of the genome could be recapitulated using only histone modification information—a remarkable finding. Furthermore, states associated with some features were also associated with specific ontologies. For example, multiple states were found that corresponded to promoters. Each of these states corresponded to different sets of genes enriched for distinct Gene Ontology (GO) terms. This suggests that histone modifications encode functional information along with feature type information. Similar findings were obtained in a study utilizing a more general DBN model [Hoffman et al., 2012]. Each of these studies resulted in software tools for HMM (ChromHMM [Ernst and Kellis, 2012]) and DBN (Segway [Hoffman et al., 2012]) segmentation of chromatin data. These tools were used in several published analyses of ChIP-seq datasets generated by the ENCODE consortium [Ernst et al., 2011, Dunham et al., 2012, Hoffman et al., 2013, Ernst and Kellis, 2013].

Bayesian networks

Bayesian networks (BNs), like HMMs and DBNs, are a type of probabilistic graphical model. The goal of a BN is to discover the conditional dependencies between random variables in a set. BNs are represented as directed acyclic graphs (DAGs), where each node represents a probability function for a random variable, and each edge represents dependence between two variables. The probability function of a given node is a conditional probability, conditioned on the values of its parent nodes. Though in principle training a BN allows one to learn the dependence structure between of random variables, the relationships may not be causal, but merely correlative [Chickering, 1995].

Several groups have attempted to use BNs to learn the dependence structure between histone modifications (see section 1.1.2 for a discussion of dependence among histone modifications) [Yu et al., 2008, Lv et al., 2010]. Although the results have produced sensible dependence relationships between some histone modifications that are well characterized, the resulting BNs are difficult to interpret. The inherent conflation of causal and correlative relationships in BNs, complicates the biological interpretation of the results. This issue is exacerbated by the high degree of correlation between many histone modifications. Furthermore, the methods used to estimate histone mark levels have produced artifacts in some of these studies (see Chapter three for further details). Though the first application of BNs to histone modifications was in 2008 [Yu et al., 2008], few insights into histone modification dependence have been gained using this method. The difficulties of interpreting BNs warrant rigorous experimental validation, which efforts to date have lacked. Thus, future application of BNs to questions in histone modification dependence relationships will greatly benefit from collaboration between wet lab and computational scientists.

1.3.3 Methods for whole-genome chromatin conformation data

Every chromosome conformation capture technique—including 3C, 4C, 5C, Hi-C, and ChIA-PET—gives information about the contact frequency between pairs of regions in the genome. The problem of reconstructing features of chromosomal organization from this data is computationally challenging. There are two strategies that are often used to analyze these data: (1) Identification of interaction domains, and other types of interaction structures through statistical analysis, and (2) construction of three-dimensional models based on interaction restraints or polymer physics [Dekker et al., 2013]. The latter approach is beyond the scope of this work, and will not be discussed further. Moreover, this discussion will be in the context of Hi-C, which is the only completely unbiased chromosome conformation capture method. However, much of what will be discussed also applies the high-throughput methods that are not biased by a single locus, which includes 5C and ChIA-PET.

The results of chromatin conformation capture experiments represent an average over a population of cells. Thus, the intensity of a signal is proportional to the interaction frequency across the population of cells, for a given pair of genomic loci. Using this reasoning, researchers have demonstrated that there is a high degree of variability in genomic conformations across a population of cells [Kalhor et al., 2012, Tjong et al., 2012]. Indeed, in many whole-genome experiments, non-zero signals are observed at virtually all pairs of loci, which reflects the conformational heterogeneity in a population of cells. Therefore, a common objective is to identify locus pairs that display significant interaction over the background. A simple method that has been used to identify regions that display strong though-space interactions is to formulate a probability model for background interactions and calculate the p-value for every interaction based on this null model [Duan et al., 2010, Sanyal et al., 2012].

For example, it is reasonable to assume that the probability of inter-chromosomal interactions is uniform across the genome, where the probability of any particular interaction is $p = 1/N$, where N is the total number of possible pairs. Under this assumption, the probability of observing k inter-chromosomal interactions between any two genomic regions is given by the binomial distribution:

$$\Pr(K = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad (1.1)$$

where n is the total number of observed inter-chromosomal interactions. The p-value for a given pair of genomic regions is thus calculated by

$$\text{p-value} = \sum_{i=k}^n \Pr(K = i) \quad (1.2)$$

Calculating the significance of intra-chromosomal interactions is more complicated than inter-chromosomal interactions. This is because the contact probability of intra-chromosomal pairs is inversely correlated with genomic distance, owing to the polymer nature of chromatin. A reasonable approach is to stratify interactions into ranges of intra-chromosomal distances. Significance can then be calculated for each range independently using the method described above.

Various approaches have also been implemented to identify interaction compartments, or clusters of interacting loci. One of the first methods to identify these compartments in Hi-C data was through the principle components of interaction matrices, where element (i, j) of an interaction matrix M_{ij} represents strength of the interaction between genomic regions i and j . The sign of the first principle component eigenvector of these matrices was found to broadly distinguish euchromatic and heterochromatic regions [Lieberman-Aiden et al., 2009]. It was later discovered that this eigenvector correlated very strongly with replication timing [Ryba et al., 2010]. Another group used an HMM approach to discover TADs from Hi-C data [Dixon et al.,

2012]. Briefly, they calculated a “directionality index” for each genomic region, which reflects the degree of upstream or downstream interaction bias of a locus. Using an HMM they were able to segment the genome into states of “upstream bias,” “downstream bias,” or “no bias.” The sequence of these states were then used to identify interaction boundaries between adjacent regions of the genome. For example, a boundary would be predicted after the sequence *downstream bias* \rightarrow *no bias* \rightarrow *upstream bias* if it was directly followed by the same sequence. Furthermore, a *downstream bias* \rightarrow *no bias* \rightarrow *upstream bias* sequence would also represent a TAD, since it shows a preference for interaction within itself, but not for adjacent up- and downstream regions.

Methods utilizing network analysis are also beginning to be applied to interaction data to identify interaction domains. However, only biased interaction data has been analyzed in this way. One study utilizing ChIA-PET analyzed the structure of the interaction network generated by Pol II centric interactions in human cancer cell lines [Sandhu et al., 2012]. Another study constructed a gene-gene interaction network from Hi-C data, also in human cancer cell lines [Wang et al., 2013]. The following section will provide an overview of relevant network analysis topics, and their meanings in the context of Hi-C analysis. Chapter five will describe the first application of network analysis to unbiased Hi-C data.

1.3.4 Network analysis of Hi-C data

In biology, network analysis is most closely associated with gene regulation networks or protein-protein interaction networks. However, networks are also a natural way to represent interactions between genomic regions, since networks abstractly represent pairwise relationships between objects. In the case of Hi-C data, genomic loci are represented as nodes in a network, and interactions are represented as edges. Representing chromatin interaction data as networks is also particularly attractive because

of the rich collection of network analysis methods that can be used to infer properties of chromatin conformation. Large-scale structural analysis of networks can be used to identify interaction domains, or chromosome territories. Small-scale analysis—on the level of individual nodes and edges—may be used to identify regions of the genome that are important in seeding higher-order structure. Some large and small-scale analysis techniques and their potential application in analyzing Hi-C data will be discussed.

For this discussion it is useful to describe some of the mathematical concepts and vocabulary used in network analysis. First, in mathematics, “networks” are more often referred to as “graphs,” and “nodes” as “vertices.” A graph G is an object defined by two sets, (V, E) , where V is a set of vertices, and E is a set of edges where each element of the set is a set of vertices, $\{v_i, v_j\}$. Furthermore, an edge can have an associated weight that gives the strength of the connection between two vertices. An *adjacency matrix* is a way of representing connections between the vertices of a graph. The adjacency matrix A of a graph $G(V, E)$ is an $n \times n$ matrix where $n = |V|$. Each entry a_{ij} of A is equal to the weight of the edge between v_i and v_j , or 1 in the case of an unweighted graph. That is, for an unweighted graph

$$A_{ij} = \begin{cases} 1 & \text{if } v_i \rightarrow v_j \\ 0 & \text{otherwise} \end{cases} \quad (1.3)$$

The adjacency matrix is a useful mathematical representation of a graph, which is widely used in network analysis. For example, spectral analysis of adjacency matrices is useful in partitioning graphs [Pothén et al., 1990]. These concepts will be useful for the following discussion.

Small-scale analysis

In network analysis it is often useful to calculate the “influence” of nodes in a network through measures of *centrality*. Influential nodes have outstanding centrality values, and are often referred to as “hubs.” These nodes form the structural lynchpins of a network as a whole. There are several centrality measures that can be used to quantify the influence of nodes in a network. Each provides different perspectives on the factors that contribute to the influence of a node. Three common and relatively intuitive centrality measures are *degree* centrality, *closeness* centrality, and *betweenness* centrality. These centrality measures will be discussed here in the context of unweighted networks; however, weighted network generalizations also exist [Barrat et al., 2004]. Importantly, each measure can be interpreted in the context of a genomic interaction network to formulate biological hypotheses from interaction data.

Degree centrality is the simplest and most intuitive measure of centrality. The degree centrality of a node v is simply defined as the number of edges incident to v (i.e., the degree of node v);

$$C_{degree}(v) = deg(v) \tag{1.4}$$

Degree centrality in a Hi-C-derived interaction network can be used to identify key structural regions of the genome. A genomic region with high degree centrality brings together many other genomic regions, and thus may be involved in nucleating structures such as transcription factories. Since Hi-C data is generated from a population of cells, a genomic region with high degree centrality may also represent regions that interact promiscuously, such as certain types of enhancer elements [Li et al., 2012a].

The closeness centrality of a node is a measure of how near the node is to all other nodes in the network. For a node v it is defined as the inverse sum of the shortest

paths from v to all other nodes in a graph G [Sabidussi, 1966];

$$C_{closeness}(v) = \frac{1}{\sum_{t \in V} d_G(v, t)} \quad (1.5)$$

where $d_G(v, t)$ is the length of the shortest path from node v to node t in graph G . Intuitively, closeness centrality and degree centrality are correlated; i.e., the more nodes a given node is connected to, the more likely it is to have a short path to any other node. Similarly, to have a low closeness centrality a node must be relatively distant from highly connected nodes. In the context of a genomic interaction network, closeness centrality may be a useful measure in distinguishing highly isolated regions of the genome from highly exposed regions of the genome. An intuitive case would be distinction between nuclear lamina-associated regions of the genome, which are relatively quiescent, and regions of the genome that are closer to the interior of the nucleus, which are relatively active.

Betweenness centrality is a measure of how important a node is in connecting other nodes in the network. The betweenness centrality for a node v is defined as the sum of the ratios of the number of shortest paths that pass through v and the number shortest paths between all pairs of nodes [Anthonisse, 1971, Freeman, 1977];

$$C_{betweenness}(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (1.6)$$

where σ_{st} is the number of shortest paths from node s to node t , and $\sigma_{st}(v)$ is the number of shortest paths from node s to node t that pass through node v . This measure could be useful in identifying genomic regions that link multiple higher-order structures, since there will be a large number of shortest paths between the structures that pass through these nodes. In principle these regions could represent the interface between two interaction domains or chromosome territories. There is

also an analogous measure of edge betweenness, that could be used to find similarly important interactions.

Section 1.3.3 described how significant interactions can be identified from interaction frequencies in Hi-C data; however this only takes into account the loci that constitute the interacting pair. A higher-order analysis, that takes into account shared neighbors, can be done using network approaches. *Equivalence measures* quantify the similarity of node pairs with respect to the common neighbors that they share. These measures could be useful for identifying pairs of nodes that are members of an interaction domain. One measure of node equivalence is the Jaccard coefficient, defined by

$$J_{uv} = \frac{|N_u \cap N_v|}{|N_u \cup N_v|} \quad (1.7)$$

where N_u is the set of nodes that neighbor node u , and N_v is the set of nodes that neighbor node v . Another equivalence measure is the cosine similarity, defined by

$$\phi_{uv} = \frac{|N_u \cap N_v|}{\sqrt{|N_u| |N_v|}} \quad (1.8)$$

which has the useful property $0 \leq \phi \leq 1$ for all pairs of nodes with degree > 0 . Equivalence measures for all pairs can be analyzed together—via hierarchical clustering, for example—to determine large-scale characteristics of a network.

Large-scale analysis

The goal of large-scale analysis is to characterize the global structure of networks. One of the most important large-scale aspects of a network is its community structure. A community is a group of nodes that has dense connections within the community relative to their connections with nodes outside of the community. Figure 1.1 shows a network divided into communities, which are represented by node color, and where node size reflects betweenness centrality magnitude. In an interaction network built

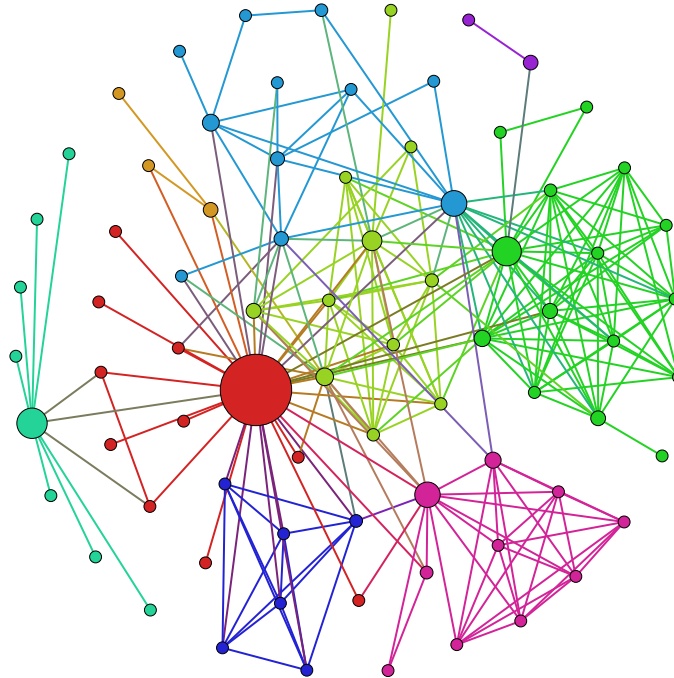


Figure 1.1: **Large and small-scale properties of a network.** Node size represents betweenness centrality magnitude. Node color represents community identity. Edge color is a mixture of the nodes connected by the edge.

from Hi-C data, communities can represent the higher-order structures of the genome, such as chromosome territories. Like centrality, many definitions for communities exist, making the problem of community detection not well posed [Newman, 2011]. However, there are a number of useful approaches for partitioning networks into communities that have successfully recapitulated known community structure in benchmark networks, and have given useful information about the structure of *de novo* networks [Lancichinetti et al., 2008, Lancichinetti and Fortunato, 2009].

A simple method for detecting communities is hierarchical clustering, using equivalence measures—such as the Jaccard coefficient, or the cosine similarity—as a distance metric. In this procedure, the metric represents the strength of the connection between pairs of nodes. Nodes are then grouped, first into pairs, and then hierarchically into larger groups, based on some linkage criterion. Many linkage criteria exist, however the basic principles of the clustering procedure are the same, regardless of the link-

age criterion. The variety of distance metrics and linkage criteria require that the community detection procedure be tuned in order to find the “best” solution. Since clustering is an ill-defined problem, evaluating the community partitions identified through hierarchical clustering (i.e., finding the “best” solution) is an *ad hoc* procedure. Methods that better define partition quality are thus more desirable.

A variety of methods exist for community detection based on modeling equilibrium dynamic processes that occur on the network, such as random walks [Zhou and Lipowsky, 2004]. These methods detect communities by attempting to optimize a quantity that indicates the quality of a partition with respect to community structure. One of the most widely used measures for this purpose is known as *modularity* [Newman, 2004], which is defined as

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \quad (1.9)$$

where A_{ij} is the adjacency matrix of the network, k_i is the degree of node i , m is the total number of edges in the network, c_i is the community to which node i belongs, and δ is the Kronecker delta. Thus, modularity is a measure of the difference between the fraction of edges within communities, and the expected fraction if the edge distribution in the graph was random. Optimization of modularity has been proven to be NP-complete [Brandes et al., 2008]; however, several algorithms for finding approximate solutions exist [Clauset et al., 2004, Medus et al., 2005, Blondel et al., 2008, Li et al., 2010]. Chapter five will further discuss community detection by modularity optimization, and the application of the technique to Hi-C data to uncover the network architecture of the budding yeast genome.

1.4 Dissertation rationale

This dissertation provides a broad perspective on the application of computational analyses to many questions in chromatin biology that were highlighted in this introduction. Together, the studies discussed in Chapters two through five are composed of a diverse panel of computational techniques and biological questions. Chapter two discusses the application of machine learning methods to histone modification ChIP-seq data, and demonstrates that the activity of certain modifications are only apparent in a combinatorial context. Chapter three discusses issues in quantifying histone modification ChIP-seq data, and presents a method for improving the results of analyses based on these data. Chapter four presents a high-dimensional analysis of changes in histone modification levels during EMT. This chapter includes novel analysis methodologies, as well as novel insights into epigenetic reprogramming in EMT. Finally, Chapter five presents a novel application of network analysis techniques to Hi-C data. It also discusses novel insights into the functional organization of the *Saccharomyces cerevisiae* genome gained from these analyses.

The unifying thread of this dissertation is the innovative use of sophisticated computational techniques to gain new insights into chromatin biology. Approaches similar to those outlined in this dissertation will only become more critical as more whole-genome chromatin data is generated. Indeed, as databases grow evermore comprehensive, many difficult questions in chromatin biology will be made approachable, and many new hypotheses will be engendered through the tools of computational data analysis.

Chapter 2

Regression analysis reveals transcriptional association H4R3me2

2.1 Introduction

Histones are subjected to numerous modifications, including methylation, acetylation and phosphorylation. Over one hundred modification/residue combinations have been discovered [Rando, 2012]. They regulate a number of important processes on DNA, including transcription [Li et al., 2007, Berger, 2007]. Extensive studies comparing histone modification and transcription levels have established that histone methylation is associated with either transcriptional repression or activation. A number of marks have been classified as “activating” with respect to transcription, such as trimethylated H3 lysine 4 (H3K4me3) and trimethylated H3 lysine 36 (H3K36me3), as well as “repressing,” such as trimethylated H3 lysine 27 (H3K27me3). These modifications can be recognized by effector proteins (readers), which can render

chromatin in either “open,” transcriptionally permissive conformations or “closed,” DNA-inaccessible conformations, respectively [Li et al., 2007, Berger, 2007].

A simple question that emerges is: Why does the cell require more than one hundred modifications to maintain two (i.e., open and closed) or a handful of chromatin states? The histone code hypothesis was developed to address this question. The histone code hypothesis suggested that distinct functional consequences result from histone modifications, and that a given outcome is encoded in the precise nature and pattern of marks [Strahl and Allis, 2000, Jenuwein and Allis, 2001]. A challenge to the hypothesis has been the identification of multiple readers for a single modification, thereby confounding “a simple one-mark-to-one-module type of decoding” [Ruthenburg et al., 2007]. A framework that keeps the histone code hypothesis intact and addresses this criticism is the phenomenon of multivalency—the cooperative engagement of several linked substrates by a species with more than one discrete interacting surface [Berger, 2007, Ruthenburg et al., 2007]. In other words, chromatin regulatory proteins and their associated complexes write, read and erase multiple histone modifications simultaneously. It has been suggested that multivalency may be widespread in chromatin regulation. Indeed, a number of recent studies are uncovering patterns of coexisting histone marks, extensive crosstalk among different modifications as well as multiple effector proteins on the same complex [Ruthenburg et al., 2007, Latham and Dent, 2007, Suganuma and Workman, 2008].

Using ChIP-chip and ChIP-seq, bivalent domains of H3K4me3 and H3K27me3 were observed at genes encoding developmentally important transcription factors in embryonic stem cells [Bernstein et al., 2006, Mikkelsen et al., 2007, Ku et al., 2008]. It is suggested that these genes are transcriptionally silent but poised for activation during development. Indeed, in differentiated cells the vast majority of bivalent domains (93/97) resolved into either K4me3 (active genes) or K27me3 (repressed genes). Consistent with the idea of widespread multivalency, it is notable that two

“opposing” marks were assayed on a genomic scale and were found to occur in bivalent domains. It raises the following question: If many more marks were mapped, would we find widespread multivalencies?

To help address these questions we applied two machine learning methods, *stepwise multilinear regression* and *Multivariate Adaptive Regression Splines* (MARS) [Friedman, 1991a, Friedman, 1991b], to genome-wide ChIP-seq maps of 20 histone lysine and arginine methylations and histone variant H2A.Z in CD4⁺ T cells [Barski et al., 2007]. We hypothesized that inclusion of two (bivalent) and three (trivalent) interacting cross-terms in the model can reveal (1) putative cross-regulation or multivalent relationships between histone modifications and (2) a global view of the epigenetic regulatory network. Specifically, we first estimate the enrichment level of each modification using a novel model-based approach, which accounts for the characteristic spatial distribution of each modification across genes. With the enrichment levels as inputs, and normalized log₂ gene expression levels as output, we built the multilinear (ML) model from a set of 21 single or monovalent inputs, 210 bivalent inputs and 1330 trivalent inputs. For the MARS model, the 21 monovalent amplitudes were supplied as input and the bi- and trivalent interacting terms were added as part of the model optimization procedure. Using 10-fold cross validation and requiring terms to appear in 5 of 10 training models, our best ML model contained 7 monovalent, 8 bivalent, and 8 trivalent terms. Using the Generalized Cross Validation (GCV) score to protect against overfitting, we trained a MARS model that had 7 monovalent, 10 bivalent, and trivalent terms. We were able to identify a number of highly significant multivalent terms, suggesting that multivalency and cross talk among histone modifications may be widespread. We were surprised that both models predicted H4R3me2 to be among the most repressive histone methylations, given that its ChIP-seq enrichment levels showed no response to increasing gene expression in a univariate context [Barski et al., 2007]. H4R3me2 has, however, been shown to be repressive in a number of

other site-specific biochemical studies [Wang et al., 2008, Hou et al., 2008, Litt et al., 2009, Zhao et al., 2009]. Along with our findings, this suggest that the global activity of H4R3me2 is multivalent.

2.2 Model construction rationale

2.2.1 Basic workflow

A diagram of our analysis workflow is shown in Figure 2.1. The inputs to our analysis are the read enrichment profiles across the human genome generated by Barski et al. [Barski et al., 2007]. In Figure 2 of their paper, they also calculated composite plots where they stratified gene expression by quartiles, aligned the transcription start sites (TSS) of all the genes, and calculated the normalized read counts as a function of position relative to the TSS. These plots show that (1) each mark has a relatively unique profile and (2) the shape of each mark’s profile displays a relatively weak dependence on gene expression level. Based on these observations, we modeled each modification’s profile at every gene as a product of a gene-dependent enrichment multiplied by a position-dependent average profile or “template.” From gene-centric read enrichment profiles we calculated the average spatial distribution of each mark across the promoter region, the scaled gene body, and downstream 3’ region as detailed in section 2.5. We then calculated all 210 possible bivalent products and all 1330 possible trivalent products from the 21 single mark amplitudes. Thus, we have 21 single modification states that are inputs to the MARS, model and a total of 1561 possible modification states that are inputs to our multilinear model.

These amplitudes were calculated for 11,796 Ensembl genes [Hubbard et al., 2009]. Because the gene expression data was 3’ biased, we could not distinguish the expression levels of different isoforms, which included multiple TSS. We selected the TSS that had the largest number of significant modifications as detailed in section 2.5. Human

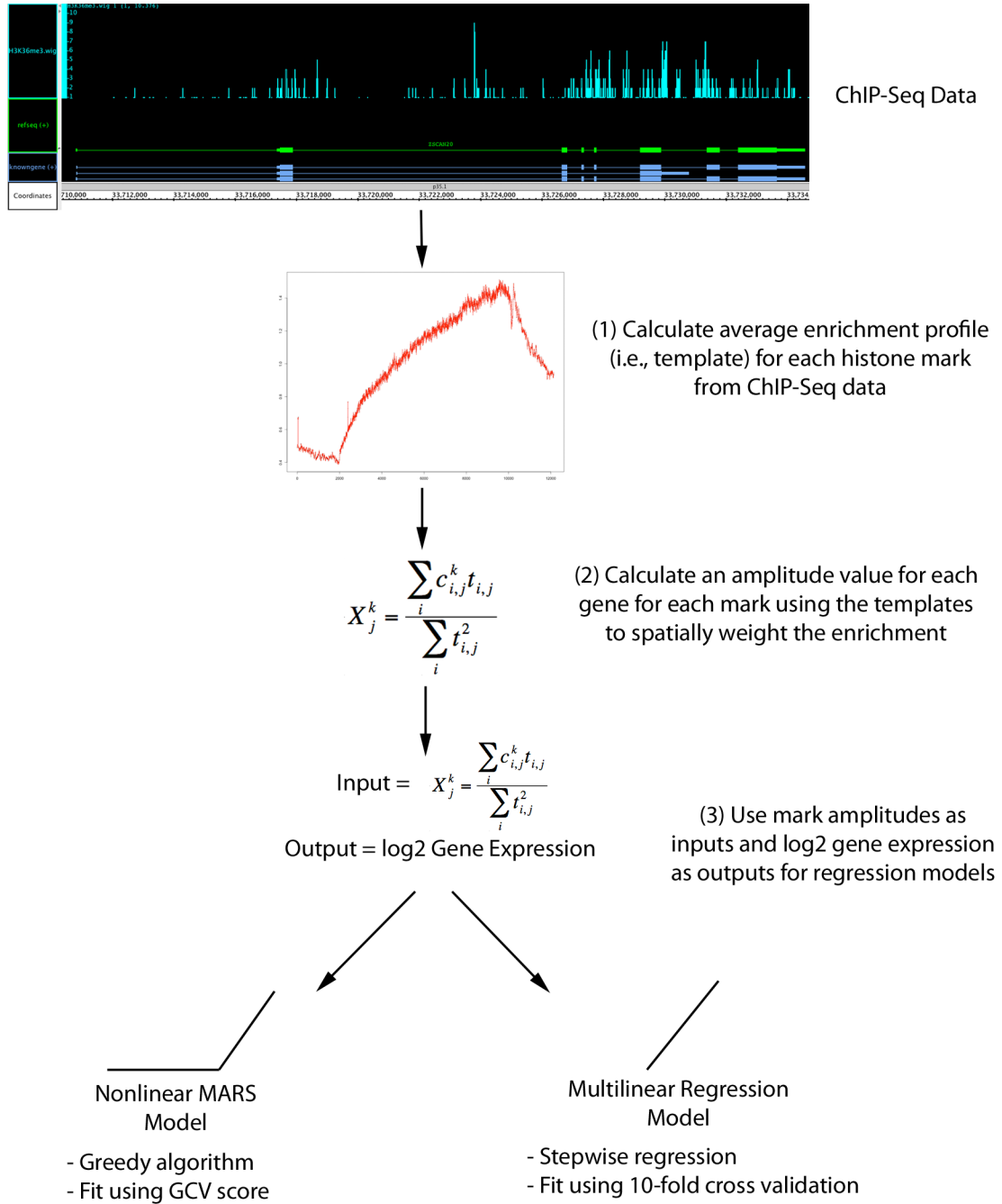


Figure 2.1: **Flowchart of multilinear and MARS model construction.** Chart describes the analysis steps in model construction. Starting with histone mark/variant ChIP-Seq data, template profile and amplitude calculation, and finally construction of regression models using mark amplitudes as inputs and \log_2 gene expression as outputs.

CD4⁺ T cell gene expression data which was used to generate the output of our models was collected from the Genomics Institute of the Novartis Research Foundation’s SymAtlas [Su et al., 2004]—a compendium of gene expression data in human and mouse tissues. Only genes that had Ensembl, UCSC, and RefSeq IDs were included in this study. Of the 18,647 genes that met these criteria, 11,796 had expression data associated with them [Su et al., 2004]. Because multiple Affymetrix probes can interrogate a single gene, the total number of expression data points for the 11,796 genes was 17,635, which constituted the output of the multilinear and MARS models.

2.2.2 Estimating input amplitudes for regression models

Two groups have estimated ChIP-seq histone modification/variant enrichment levels across genes in order to applying machine learning techniques, specifically, linear regression [Karlić et al., 2010] and Bayesian networks [Yu et al., 2008]. They count tags only in a region surrounding the transcription start site (i.e., ± 1 kbp [Yu et al., 2008] or ± 2 kbp [Karlić et al., 2010] of the TSS). A major problem with this method is that many marks do not have promoter/5’-biased enrichment patterns. A striking example is H3K36me3, which has increasing enrichment along gene bodies, which peaks near the 3’ ends [Barski et al., 2007]. Yu et al. calculated correlation coefficients between their 5’-biased mark enrichment estimates and gene expression levels, and found little to no correlation between H3K36me3 and gene expression [Yu et al., 2008]. This is an unexpected result as H3K36me3 has been characterized as an activating mark in a number of biochemical studies [Latham and Dent, 2007, Suganuma and Workman, 2008], and its levels have been shown to have a strong positive correlation with gene expression [Barski et al., 2007]. This discrepancy is likely due to the 5’ bias of their amplitude estimation method. To address this problem, we estimated the enrichment levels of each mark by calculating a weighted average across the whole gene and its flanking region as described in section 2.5. We use the average enrichment

pattern across the flanking regions and the body of scaled genes as a weighting function. However, given the large variation in gene lengths, exon/intron number, and mark deposition patterns, we also assessed the robustness and relative error of our amplitude estimation procedure. See section 2.5 for details on amplitude estimation and the robustness evaluation.

2.3 Results and discussion

2.3.1 Multilinear model

We fit the gene expression data to the multilinear (ML) model shown in equation 2.6. As described in section 2.5, we used stepwise linear regression to build the ML model. There were 21, 210 and 1330 possible terms in the first, second and third sum of equation 2.6, respectively. The final model contained 24 terms. The average training and testing MSE was 3.1213 and 3.1525 respectively for this model. The average adjusted R^2 for the training and testing data is 0.4689 and 0.4574, respectively. The fact that the train and test values are close suggests that the model was not over trained. Using all of the data, we calculated an adjusted R^2 of 0.4687 and MSE value of 3.1228. Crudely, this suggests that our model explains almost 50% of the gene expression variation after adjusting for the number of degrees of freedom. In Figure 2.2A, we show a scatter plot of the actual versus the model \log_2 gene expression levels whose associated Pearson correlation coefficient was 0.687 (p-value $< 2.2e-16$). This value is consistent with that of Karlic et al. who modeled gene expression as a function of 38 histone methylation and acetylation modifications and H2A.Z using a linear model with no interaction (multivalent) terms [Karlić et al., 2010]. Given the absence of many other mRNA regulatory factors including other histone modifications, transcription factors, and miRNAs, a relatively significant percentage of the variation is explained by these models.

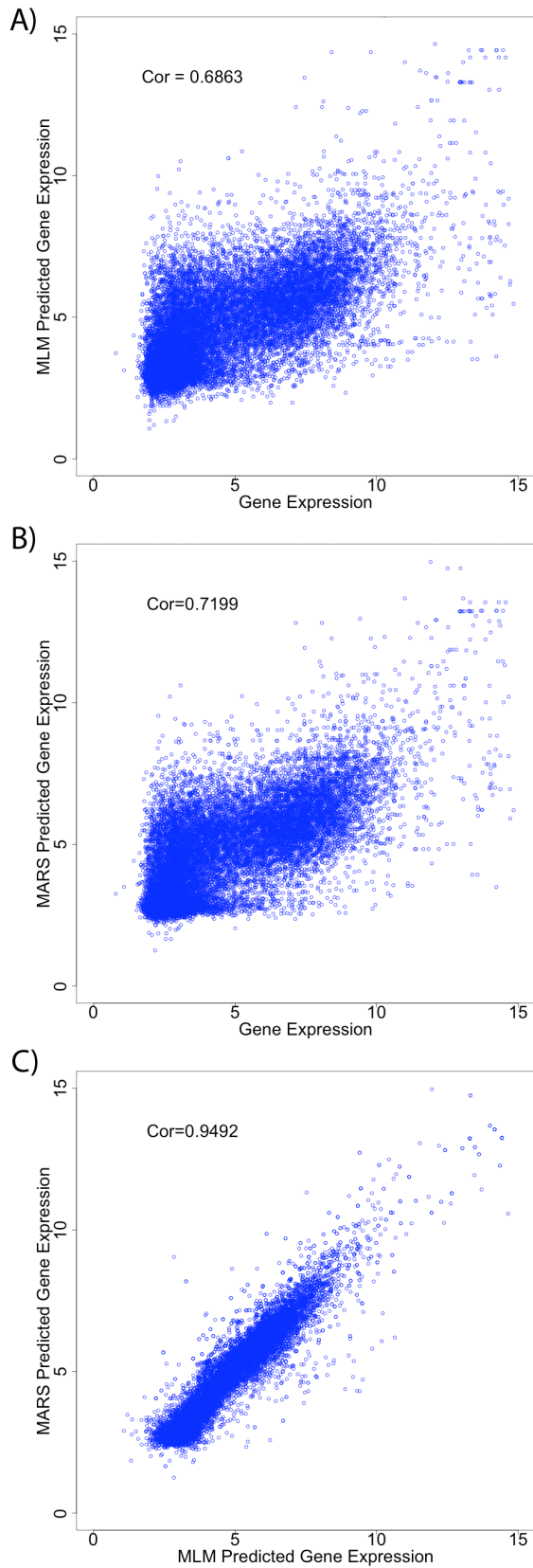


Figure 2.2: **Comparison of predicted and observed gene expression.** Scatter plots of (A) the multi-linear model (MLM) predicted gene expression versus observed gene expression; (B) the MARS predicted gene expression versus observed gene expression; and (C) the MARS predicted gene expression versus multilinear model predicted gene expression. The corresponding Pearson correlation coefficient is shown within each plot.

In the final model, 7 of the 21 (33%) single terms were found to be significant. In addition, there were 8 (4%) significant paired terms and 8 (0.6%) significant triplet terms. The terms appearing in the full model are displayed in Table 2.1 where we show each surviving term's β coefficient, the β coefficient's Z-score (i.e., the number of standard deviations away from $\beta = 0$), the term's p-value, and a robust impact factor. The robust impact factor is defined as the product of the fitting coefficient (β) and the inter-quartile range (75th percentile – 25th percentile) of the mark amplitudes. It is a robust measure of a term's impact on gene expression while the Z-score and p-value are measures of its significance. A positive (negative) β coefficient, Z-score, and impact factor indicate an activating (repressing) term in the model. The table is sorted by impact factor with activating or repressive marks labeled with “a” or “r” superscripts, respectively. We labeled the marks according to (1) the sign of their monovalent term in the ML model, or (2) the response of the mark's levels with increasing gene expression when it did not contribute a monovalent term (these marks are starred).

2.3.2 Multilinear model terms

Of the 7 monovalent terms (Table 2.1), H3K4me3, H3K36me3, H3K79me1, H3K79me3 and H4K20me1 were activating. Of these, only H3K4me3, H3K36me3 and H4K20me1 display a clear overall activating trend from composite plots [Barski et al., 2007]. Based on their composite plot analysis, Barski et al. conclude that H3K79me1 levels alone shows no overall trend with gene expression, while it has the strongest activating contribution in our ML model. This is consistent with a recent finding that H3K4me3 and H3K79me1 are the most predictive of gene expression levels in low CpG content promoters [Karlić et al., 2010]. Two arginine methylations, H4R3me2 and H3R2me1, were the only repressive monovalent marks in the model. In contrast, Barski et al. found no overall activating or repressing trend for these two methylations from their

Table 2.1: Multilinear model terms and statistics

| Term | β (trim mean) | Z (trim mean) | p (median) | Impact (trim mean) |
|--|---------------------|---------------|------------|--------------------|
| H3K79me1 ^a | 6.741 | 18.234 | 0 | 1.331 |
| H3K36me3 ^a | 4.087 | 17.802 | 0 | 0.922 |
| H3K79me3 ^a | 3.078 | 23.916 | 0 | 0.598 |
| H4K20me1 ^a | 0.977 | 21.446 | 0 | 0.450 |
| H3K4me2 ^{a*} -H3R2me1 ^r | 18.270 | 7.850 | 1.66e-15 | 0.437 |
| H3K27me2 ^{r*} -H3R2me1 ^r | 70.468 | 15.280 | 0 | 0.381 |
| H3K9me2 ^{r*} -H3K27me1 ^{r*} -H4K20me1 ^a | 37.041 | 5.643 | 9.47e-09 | 0.156 |
| H3K4me3 ^a | 0.133 | 5.729 | 5.20e-09 | 0.151 |
| H2BK5me1 ^{a*} -H3K36me3 ^a | 1.286 | 3.800 | 7.95e-05 | 0.115 |
| H2BK5me1 ^{a*} -H4K20me1 ^a -H3R2me1 ^r | 1.531 | 6.034 | 1.14e-09 | 0.030 |
| Intercept | 4.026 | 64.131 | 0 | 0 |
| H3K9me3 ^{r*} -H3K36me3 ^a | -2.747 | -12.439 | 0 | -0.010 |
| H3K4me2 ^{a*} -H3K36me3 ^a -H3K79me3 ^a | -1.274 | -3.743 | 1.02e-04 | -0.018 |
| H3K36me3 ^a -H3K79me2-H3R2me2 | -5.855 | -3.497 | 1.73e-04 | -0.022 |
| H3K27me3 ^{r*} -H3K79me2-H3K79me3 ^a | -26.341 | -6.380 | 5.78e-11 | -0.026 |
| H3K4me1 ^{a*} -H3K9me2 ^{r*} -H4K20me1 ^a | -4.563 | -3.578 | 2.65e-04 | -0.041 |
| H3K9me1 ^{a*} -H3K27me1 ^{r*} -H4K20me1 ^a | -2.350 | -5.078 | 1.92e-07 | -0.077 |
| H2BK5me1 ^{a*} -H4K20me1 ^a | -0.600 | -9.627 | 0 | -0.095 |
| H3K36me1-H3K79me1 ^a -H3K79me3 ^a | -27.840 | -8.478 | 0 | -0.115 |
| H3K4me2 ^{a*} -H3K9me1 ^{a*} | -1.578 | -3.340 | 3.57e-04 | -0.123 |
| H4R3me2 ^r | -11.121 | -13.233 | 0 | -0.301 |
| H3K27me2 ^{r*} -H3K36me3 ^a | -31.772 | -8.911 | 0 | -0.311 |
| H3K27me2 ^{r*} -H3K79me1 ^a | -56.535 | -9.535 | 0 | -0.449 |
| H3R2me1 ^r | -11.937 | -16.504 | 0 | -0.596 |

Terms appearing in the final multilinear model, and associated statistics. Trim mean of the β coefficients, Z-scores and impact factors (β multiplied by amplitude interquartile range). Trim mean is defined as the mean of the population excluding the lowest and highest 5% of the data. The superscript labels each mark as activating (a) or repressive (r). Unstarred marks correspond to monovalent terms in the model, starred marks do not have a monovalent contribution in the model, but correlated/anti-correlated with gene expression based on univariate analysis, and uncolored marks do not have clear correlation with gene expression. Rows are sorted by the impact term value.

composite plot analysis. Marks that showed an activating trend from composite plots but did not appear as monovalent terms in our ML model included H3K4me1/2, H3K9me1 and H2BK5me1. Marks that showed a repressive trend from composite plots but were absent at the monovalent level in our ML model were H3K27me2/3, and H3K9me2/3. H3K27me1 and the variant H2A.Z did not appear as monovalent terms in the ML model, and displayed complex non-monotonic enrichments as a function of increasing expression from their composite plots. Finally, marks that neither appeared at the monovalent level in the ML model nor showed any trend with respect to gene expression level were H3K79me2, H3R2me2 and H4K20me3.

While the majority of monovalent terms are activating, the majority of multivalent terms (11 of 16) are repressive. Half of the 16 multivalent terms involve a mix of activating and repressing modifications according to either the sign of their monovalent term in the ML model or composite plot trends. This is interesting given the discovery of bivalent domains of H3K4me3 and H3K27me3. Indeed, two of the three highest impact and most significant repressive terms are bivalent. They both include H3K27me2 together with H3K79me1 and H3K36me3 respectively. The highest impact activating multivalent term is also bivalent and composed of an activating mark, H3K4me2, and a repressive mark, H3R2me1. Thus, at the bivalent level, the linear model terms suggest that there is significant overlap between opposing marks (i.e., activating and repressive) and that one of them tends to “override” the other, similar to the observation that H3K27me3 overrides H3K4me3 in ES cells [Bernstein et al., 2006, Mikkelsen et al., 2007].

To further investigate the extent to which bivalent terms in the linear model point to the ability of one mark to override or oppose another overlapping mark, we generated false-color plots, shown in Figure 2.3, of gene expression levels as a function of bivalent amplitudes (on the y-axis) and one of the monovalent amplitudes (on the x-axis). We discretize the amplitudes into a 10,000 square grid (i.e., 100 x-axis and y-axis bins) and calculate the average gene expression level within every box. The colors red, yellow, green, cyan, blue and magenta represent equidistant increasing gene expression values from the minimum to the maximum levels. As illustrated in Figure 2.3A, points along lines emanating from the origin moving outward represent increasing H3K27me2 (x-axis amplitude) and constant H3K36me3, with higher slopes corresponding to a higher level of fixed H3K36me3 amplitude. Conversely, points moving upward along vertical lines correspond to fixed H3K27me2 and increasing H3K36me3, with increasing position of the vertical line on the x-axis corresponding to higher level of fixed H3K27me2 amplitudes. This allows us to visualize a given

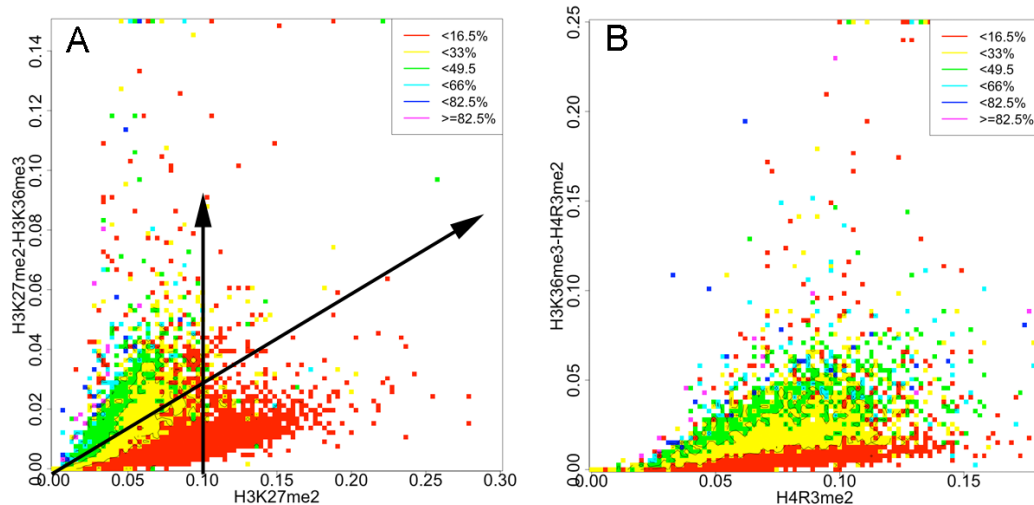


Figure 2.3: Gene expression false color plots. Gene expression (color scale) as a function of bivalent (y-axis) and monovalent (x-axis) enrichment amplitudes for (A) H3K27me2-H3K36me3 versus H3K27me2 and (B) H3K36me3-H4R3me2 versus H4R3me2. The y-axis represents the product of the amplitudes of both marks and the x-axis represents one component of the pair. Gene expression values were binned into a 10,000 square grid with level represented by color. Vertical lines represent a constant value of the x-axis mark amplitude (i.e., H3K27me2 in (A) and H4R3me2 in (B)), while a line emanating from the origin represents a constant value of H3K36me3 in (A) and (B) with the slope corresponding to H3K36me3 level. Plot (A) shows mark avoidance, as there are few genes with high levels of both marks while (B) shows a trend toward mark concurrence. These plots also demonstrate how H3K36me3 strongly overrides H4R3me2 (increasing radial slope corresponds to increasing gene expression in (B)) but has more difficulty overriding the repressive activity of H3K27me2.

modification's impact on transcription and how its regulation of transcription is continuously altered by the increasing co-occurrence of a second modification.

We found that increasing H3K27me2 corresponds to decreasing expression, as expected. We also found that increasing expression with fixed H3K27me2 and increasing H3K36me3. However, for relative H3K27me2 amplitudes exceeding 0.9, gene expression remains at low levels independent of H3K36me3 levels. This suggests that high levels of H3K27me2 are capable of overriding the gene activating potential of H3K36me3.

ML models have been applied in physical and statistical studies where a common outcome is that the single terms dominate the model in both their relative impact

and statistical significance—this is theoretically expected in many cases. In these systems, the double, triple, etc. product terms tend to make small and diminishing (in the order of the number of products) corrections to the single terms. As shown in Table 2.1, we found the expected trend, where the highest impact and most significant activating terms all monovalent. The two most significant repressive terms, H3R2me1 (highest impact) and H4R3me2, are also monovalent. However, we note that the impact of the highest impact bivalent terms is relatively large.

2.3.3 MARS model

Normalized \log_2 gene expression was modeled as a function of histone modification enrichment using the nonlinear Multivariate Adaptive Regression Splines (MARS) method [Friedman, 1991a, Friedman, 1991b]. The model was built with the *earth* package in R as described in section 2.5.

The MARS model contained 24 terms: 23 basis functions, and a constant. The MSE and R^2 for this model are 2.8387 and 0.5183, respectively. Figure 2.2B shows a scatter plot of the actual versus the MARS-predicted \log_2 gene expression levels, whose associated Pearson correlation coefficient is 0.7199. There were a total of 7 monovalent terms, with 5 unique single marks; 10 bivalent terms, with 7 unique pairs; and 6 trivalent terms, with 4 unique triplets. Table 2.2 displays each term’s hinge functions, the term’s fitting coefficient, and the number of probe sets impacted by each term. The basis functions can often have a value of zero for a wide range of amplitudes; for probe-sets in this range, the associated basis function has no impact. Thus, the number of impacted probe-sets is a measure of the global impact of each term. We also directly assessed the impact of each term on gene expression as discussed below.

Table 2.2: MARS model terms

| Coefficient | Hinge function | Non-zero genes |
|-------------|--|----------------|
| 5.531222 | 1 | 17635 |
| -10.31971 | $h(\text{H3K27me2}-0.0611382)$ | 6892 |
| 1.325129 | $h(\text{H3K79me3}-0.0948497)$ | 7218 |
| -10.70662 | $h(0.0948497-\text{H3K79me3})$ | 10417 |
| -58.18392 | $h(0.0611382-\text{H3K27me2})$ | 10743 |
| -3.112329 | $h(0.559645-\text{H4K20me1})$ | 13551 |
| -3.327909 | $h(0.545052-\text{H3K36me3})$ | 16872 |
| 15.2676 | $h(0.125391-\text{H4R3me2})$ | 17316 |
| 118.4095 | $h(\text{H3K27me2}-0.0611382) \times h(\text{H3K9me1}-0.674218)$ | 67 |
| 39.42698 | $h(\text{H2BK5me1}-1.49934) \times h(0.0611382-\text{H3K27me2})$ | 106 |
| -1.089521 | $h(0.545052-\text{H3K36me3}) \times h(\text{H4K20me1}-0.673666)$ | 3136 |
| 0.7436266 | $h(1.24429-\text{H3K79me2}) \times h(\text{H4K20me1}-0.559645)$ | 4055 |
| 47.71129 | $h(\text{H3K79me1}-0.055087) \times h(0.0948497-\text{H3K79me3})$ | 4268 |
| -258.7383 | $h(0.055087-\text{H3K79me1}) \times h(0.0948497-\text{H3K79me3})$ | 6149 |
| 16.1397 | $h(\text{H3K27me2}-0.0611382) \times h(0.674218-\text{H3K9me1})$ | 6825 |
| 1390.367 | $h(0.0611382-\text{H3K27me2}) \times h(0.0913244-\text{H3K27me3})$ | 9906 |
| 3.235207 | $h(0.545052-\text{H3K36me3}) \times h(0.673666-\text{H4K20me1})$ | 13736 |
| -37.13981 | $h(0.438075-\text{H3K36me3}) \times h(0.125391-\text{H4R3me2})$ | 15588 |
| 5.809395 | $h(1.49934-\text{H2BK5me1}) \times h(0.0611382-\text{H3K27me2}) \times h(\text{H4K20me3}-0.477337)$ | 100 |
| -100.9374 | $h(0.0611382-\text{H3K27me2}) \times h(0.0913244-\text{H3K27me3}) \times h(\text{H4K20me1}-3.93081)$ | 203 |
| -0.2791515 | $h(1.59266-\text{H3K4me3}) \times h(1.24429-\text{H3K79me2}) \times h(\text{H4K20me1}-0.559645)$ | 2971 |
| 74.37772 | $h(0.545052-\text{H3K36me3}) \times h(0.0625376-\text{H3K79me1}) \times h(0.673666-\text{H4K20me1})$ | 6329 |
| -237.028 | $h(0.0611382-\text{H3K27me2}) \times h(0.0913244-\text{H3K27me3}) \times h(3.93081-\text{H4K20me1})$ | 9703 |
| 90.39805 | $h(1.49934-\text{H2BK5me1}) \times h(0.0611382-\text{H3K27me2}) \times h(0.477337-\text{H4K20me3})$ | 10537 |

Coefficients and hinge functions within a row are multiplied and added to the products of other rows to form the MARS model. The rightmost column indicates the number of genes for which the hinge function takes a non-zero value. Terms appear in the order in which they were built into the model by the greedy algorithm.

2.3.4 MARS model terms

Of the 5 unique monovalent marks in the MARS model, 3 are activating, including H3K36me3, H3K79me3 and H4K20me1. These results are in agreement with the ML model and the data from [Barski et al., 2007]; although H3K79me3 is a complicated mark, which is enriched in the promoters of active genes and the bodies of repressed genes. H3K27me2 has a repressive trend in the model, which agrees with the univariate analysis [Barski et al., 2007]. Interestingly H4R3me2 appears to have no discernible behavior in the univariate analysis by Barski et al.; however, both the MARS and ML models select it as a repressive monovalent mark.

The MARS model also shows nonlinear trends in \log_2 gene expression as a function of mark amplitude. Figure 2.4 shows plots of predicted \log_2 gene expression as a

function of one or two mark amplitudes with all others fixed to their median value. These plots reveal whether a mark is activating or repressive. Not surprisingly, the dominant non-linear trend is saturation of predicted gene expression with increasing mark amplitude. This trend is clearly evident in both monovalent and bivalent plots shown in Figure 2.4.

To determine the global (full model) impact of individual marks appearing in the model, predictions were made with high enrichments (95th percentile) and low enrichments (5th percentile) of a given mark while fixing all other mark amplitudes at their median value. The difference between these predictions (Table 2.3—only non-zero values shown) provides an estimate of each individual mark’s impact, where positive (negative) values correspond to activating (repressive) activity. This analysis shows general agreement with the univariate analyses from [Barski et al., 2007]. However, H2BK5me1, which does not appear as a monovalent term in the model, is activating in the analysis from [Barski et al., 2007], and repressive in the MARS model. Several marks that showed no activating or repressive trend in [Barski et al., 2007] made significant contributions to the MARS model including, H3K79me1 (activating in the model), H4R3me2 (repressive), and H4K20me3 (repressive). H4K20me3 is generally associated with heterochromatin [Latham and Dent, 2007], possibly explaining why it has a slightly repressive trend in the MARS model. Interestingly, H4R3me2 has the second highest repressive impact (-0.45) and affects all of the probe sets—more than any other term in the MARS model. Moreover, it has been shown that in some cases DNA methylation, which is associated with gene silencing, is dependent on H4R3me2 [Zhao et al., 2009].

Based on the response plots (Figure 2.4), most marks appearing in a bivalent term seem to modulate each other modestly, with the exception of H4K20me1-H3K36me3, which shows nonlinear synergistic behavior. Synergies were assessed by making model predictions while varying each of the unique interaction terms in the model.

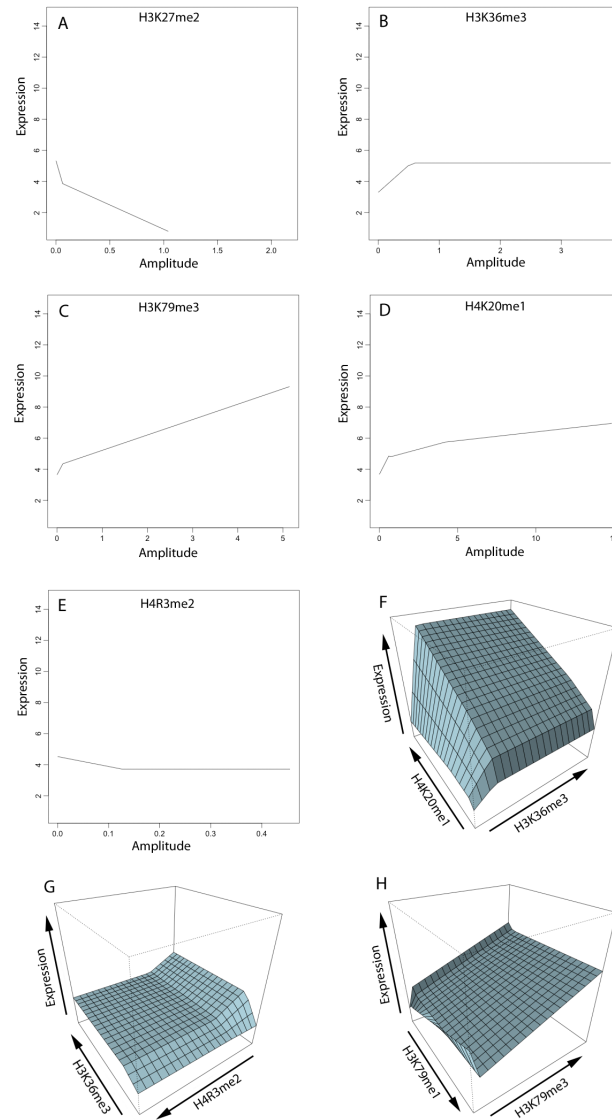


Figure 2.4: **MARS response plots.** Predicted gene expression versus amplitude for either one (2 D plots) or two marks (3 D plots) for (A) H3K27me2 (B) H3K36me3 (C) H3K79me3 (D) H4K20me1 (E) H4R3me2 (F) H4K20me1-H3K36me3 (G) H3K36me3-H4R3me2 and (H) H3K79me1-H3K79me3. Each axis represents the full range of expression and amplitude values. The trend of plots represents activating (positive slope) or repressive (negative slope) behavior. Many individual marks (A)-(E) and pairs (F)-(H) show some saturation effects and nonlinear behavior that could not be captured with a linear model; H3K36me3 (B), H4K20me1 (D) and H4R3me2 (E) show particularly distinct saturation effects. The combination H4K20me1-H3K36me3 (F) shows a dramatic nonlinear, synergistic activating effect. In contrast, the two marks in the combination H3K36me3-H4R3me2 (G) show opposing effects in that H3K36me3 activates and H4R3me2 represses gene expression.

Table 2.3: Impact of marks in MARS model

| Mark | Predicted impact (95 th -5 th percentile) |
|----------|---|
| H3K27me2 | -1.118 |
| H4R3me2 | -0.446 |
| H3K27me3 | -0.348 |
| H2BK5me1 | -0.281 |
| H4K20me3 | -0.055 |
| H3K79me1 | 0.324 |
| H4K20me1 | 1.473 |
| H3K79me3 | 1.520 |
| H3K36me3 | 1.650 |

The difference in mean predicted gene expression between the high (95th percentile) and low (5th percentile) amplitude values for a given mark while fixing all other mark amplitudes to their median values. Rows are sorted by predicted impact.

A model prediction was made with every combination of high enrichment (95th percentile) and low enrichment (5th percentile) for each element of a multivalent pair or triplet, while all other marks were held at their median values (Tables 2.4 and 2.5). The H4K20me1-H3K36me3 combination is an example of a strong synergistic, activating bivalent pair. High levels of both together correspond to highly active genes. The trivalent combination, H3K36me3-H3K79me1-H4K20me1, also shows strong synergistic activation, further suggesting that co-occupancy of H4K20me1 and H3K36me3 positively contributes to gene expression. Furthermore, this pair affects a large number of probe-sets, approximately 80% of those included in the model.

We also found one bivalent and trivalent combination composed of activating and repressive marks, H3K36me3-H4R3me2 and H3K27me2-H3K27me3-H4K20me1. As shown in Tables 2.4 and 2.5, we found that increasing each mark's level independently results in the expected activating or repressive response. High levels of the activating and repressive marks result in a moderating effect on predicted gene expression with values falling between those of high activating-low repressive and low activating-high repressive mark amplitude combinations. This reinforces the results of the ML model where we found the tendency of one mark to oppose another overlapping mark.

Table 2.4: Impact of two marks in MARS model

| Bivalent MARS term | low-low | low-high | high-low | high-high |
|--------------------|---------|----------|----------|-----------|
| H3K27me2-H3K9me1 | 4.796 | 4.796 | 3.866 | 3.377 |
| H3K27me2-H3K27me3 | 5.212 | 3.723 | 3.677 | 3.677 |
| H2BK5me1-H3K27me2 | 5.009 | 3.677 | 3.807 | 3.677 |
| H3K36me3-H4R3me2 | 3.530 | 3.468 | 5.618 | 4.556 |
| H3K79me2-H4K20me1 | 3.709 | 5.211 | 3.709 | 4.879 |
| H3K79me1-H3K79me3 | 3.009 | 5.938 | 4.721 | 5.228 |
| H3K36me3-H4K20me1 | 3.217 | 4.163 | 4.599 | 7.218 |

The mean predicted gene expression using high (95th percentile) and low (5th percentile) amplitude values of each mark described in the leftmost column. The permutations of high and low values in each column correspond to the mark order in the leftmost column. The rows are sorted by the values in the last column.

2.3.5 Model comparison

Like the ML model, the MARS model explains about half of the variation in gene expression. Moreover, the ML and MARS model predicted gene expression profiles are highly correlated (Figure 2.2C). However, the Pearson correlation coefficient between predicted and actual \log_2 gene expression is slighter better for the MARS model. This is impressive given that both models contain the same number of terms, 24, and the MARS model was built using one round of a greedy algorithm while the ML model was built by selecting the best model from multiple rounds of a stepwise algorithm. We note that the stepwise algorithm is a more powerful and computationally expensive optimization procedure. These observations suggest that methods like MARS that are capable of modeling the nonlinear relationship between histone modification and gene expression levels should outperform models that assume this relationship is linear. Moreover, many of the bi- and trivalent terms in the ML model may not have a biological origin, but may be compensating for the nonlinearities in the data. Specifically, as shown in Table 2.1, the ML model contains two bivalent terms (H3K4me2-H3K9me1 and H2BK5me1-H4K20me1) containing activating marks with a

Table 2.5: Impact of three marks in MARS model

| Trivalent MARS term | low-low- low | low-low- high | low-high- low | low-high- high | high-low- low | high-low- high | high- high-low | high-high- high |
|---------------------------|-----------------|------------------|------------------|-------------------|------------------|-------------------|-------------------|--------------------|
| H2BK5me1 | | | | | | | | |
| H3K27me2 | 4.827 | 4.827 | 3.677 | 3.677 | 3.845 | 3.719 | 3.677 | 3.677 |
| H4K20me3 | | | | | | | | |
| H3K27me2 | | | | | | | | |
| H3K27me3 | 4.739 | 7.352 | 3.373 | 4.616 | 3.327 | 4.570 | 3.327 | 4.570 |
| H4K20me1 | | | | | | | | |
| H3K4me3 | | | | | | | | |
| H3K79me2 | 3.709 | 4.944 | 3.709 | 4.722 | 3.709 | 5.712 | 3.709 | 5.172 |
| H4K20me1 | | | | | | | | |
| H3K36me3 | | | | | | | | |
| H3K79me1 | 4.000 | 3.561 | 3.649 | 4.595 | 4.058 | 6.616 | 5.031 | 7.651 |
| H4K20me1 | | | | | | | | |

The mean predicted gene expression using high (95th percentile) and low (5th percentile) amplitude values of each mark described in the leftmost column. The permutations of high and low values in each column correspond to the mark order in the leftmost column. The rows are sorted by the values in the last column.

negative (repressive) fitting coefficient; one bivalent term composed on two repressive marks (H3K27me2-H3R2me1) with a positive (activating) fitting coefficient; and a trivalent term composed of activating marks (H3K4me2-H3K36me3-H3K79me3) with a negative (repressive) fitting coefficient. These terms have no clear biological origin and are more likely artifacts of imposing a linear model on data that is inherently nonlinear.

Both regression methods produced a model with 24 terms. However, there are only 4 common terms between the models, all of which are monovalent terms: H4R3me2, H3K79me3, H4K20me1 and H3K36me3. Both models agree that of these marks H4R3me2 is the only repressive mark, while the others are activating. Considering the linear model contains 7 monovalent terms out of a possible 21, and that the MARS model contains 5, the degree of overlap in the monovalent terms is quite high.

No overlapping multivalent terms existed between the two models. The differences between the multivalent components of each model could be the result of the way the models were built. The space of possible terms increases rapidly with valency, and the search space over which the stepwise regression procedure converges on a final

model is much larger than that of the MARS procedure. Since the MARS model is built using a greedy algorithm, it is constrained in the number of multivalent terms it can potentially include. Thus, the potential for overlap becomes less likely as valency increases.

2.3.6 *In silico* knockout analysis

In order to generate experimentally testable predictions, a knockout analysis of the ML and MARS models was performed. This procedure assessed the effect of removing a specific modification on gene expression. Predictions of gene expression were made by setting the amplitude of a single mark to zero in each model and holding all others at their experimental values. This process was repeated for all marks to determine the global effect of each histone modification. All pairwise knockouts were also performed.

Tables 2.6 and 2.7 show the single and pairwise knockouts that have the highest impact on global gene expression in the ML and MARS models, respectively. Knockouts are represented as \log_2 fold changes of wild type over knockout predictions; positive values indicate activating marks, and negative values indicate repressive marks. Both sets of knockouts identify H3K36me3 and H4K20me1 to be the among strongest globally activating marks, and H4R3me2 and H3K27me2 to be the among strongest globally repressive marks. They also indicate that pairwise combinations of H3K79 methylations, H3K36me3 and H4K20me1 are among the strongest globally activating pairs. Combinations of H3K27 methylations and H4R3me2 are indicated to be among the strongest globally repressive pairs. Figure 2.5 shows box plots of the \log_2 fold changes for each single knockout of the ML and MARS models. The general trend of knockouts in both models is similar; however, the MARS model includes fewer marks, and thus many of the knockouts (trivially) have no effect. Most of the marks absent from the MARS model have a very modest knockout effect in the ML model (median \log_2 expression ratio magnitude < 0.1), with the exceptions of H3R2me1 and

H3K4me2. We also note that the results of the knockout analysis are also robust with respect to bin size (see section 2.5).

Table 2.6: ML model knockout analysis

| Knockouts | \log_2 fold change (predicted WT/predicted KO) |
|-------------------|--|
| H4R3me2 | -0.782 |
| H3R2me1 | -0.394 |
| H3K27me2 | -0.235 |
| H3K9me1 | -0.183 |
| H3K4me3 | 0.108 |
| H3K4me2 | 0.285 |
| H3K79me3 | 0.344 |
| H4K20me1 | 0.359 |
| H3K79me1 | 0.428 |
| H3K36me3 | 0.546 |
| H3K27me2-H3R2me1 | -1.192 |
| H3R2me1-H4R3me2 | -1.175 |
| H3K27me2-H4R3me2 | -1.017 |
| H3K9me1-H4R3me2 | -0.964 |
| H3K36me1-H4R3me2 | -0.912 |
| H3K79me2-H4R3me2 | -0.859 |
| H3K27me3-H4R3me2 | -0.837 |
| H3K4me2-H3K36me3 | 0.857 |
| H3K79me1-H3K79me3 | 0.903 |
| H3K36me3-H4K20me1 | 0.907 |
| H3K36me3-H3K79me3 | 0.916 |
| H3K36me3-H3K79me1 | 0.976 |

The \log_2 fold change (predicted WT/predicted KO) in average gene expression for single and double knockouts in the multilinear model. In silico knockouts were performed by setting mark amplitudes to zero while fixing all other marks at their experimental values and making model predictions for each gene. The top 5 most repressive and activating fold changes for single and double knockouts are shown. Rows are sorted according to \log_2 fold change for single and double knockouts separately.

Table 2.7: MARS model knockout analysis

| Knockouts | \log_2 fold change (predicted WT/predicted KO) |
|-------------------|--|
| H3K27me2 | -0.742 |
| H4R3me2 | -0.506 |
| H3K27me3 | -0.244 |
| H2BK5me1 | -0.158 |
| H3K79me2 | -0.046 |
| H3K4me3 | 0.054 |
| H3K79me3 | 0.421 |
| H4K20me1 | 0.715 |
| H3K36me3 | 0.941 |
| H3K27me2-H3K27me3 | -2.333 |
| H2BK5me1-H3K27me2 | -1.329 |
| H3K27me2-H4R3me2 | -1.248 |
| H3K27me2-H4K20me3 | -0.973 |
| H3K27me2-H3K79me2 | -0.789 |
| H3K36me3-H3K4me3 | 0.996 |
| H3K36me3-H4R3me2 | 1.011 |
| H3K79me3-H4K20me1 | 1.136 |
| H3K36me3-H4K20me1 | 1.327 |
| H3K36me3-H3K79me3 | 1.362 |
| H3K79me1-H3K79me3 | 1.553 |

The \log_2 fold changes (predicted WT/predicted KO) in average gene expression for single and double knockouts in the MARS model. In silico knockouts were performed by setting mark amplitudes to zero while fixing all other marks at their experimental values and making model predictions for each gene. The top 5 most repressive and 4 activating fold changes for single as well as the top 5 most repressive and activating double knockouts are shown. Rows are sorted according to \log_2 fold change for single and double knockouts separately.

2.3.7 H4R3me2 is globally repressive in ML and MARS models

Strikingly, H4R3me2 was the most globally repressive mark in the ML model according to the knockout analysis, with an average predicted fold change (WT/knockout) in gene expression of 0.55. It was also the second most repressive mark in the MARS model knockout analysis, with an average fold change of 0.70. This was a highly unexpected result for H4R3me2 given the unresponsiveness of its ChIP-seq enrichment

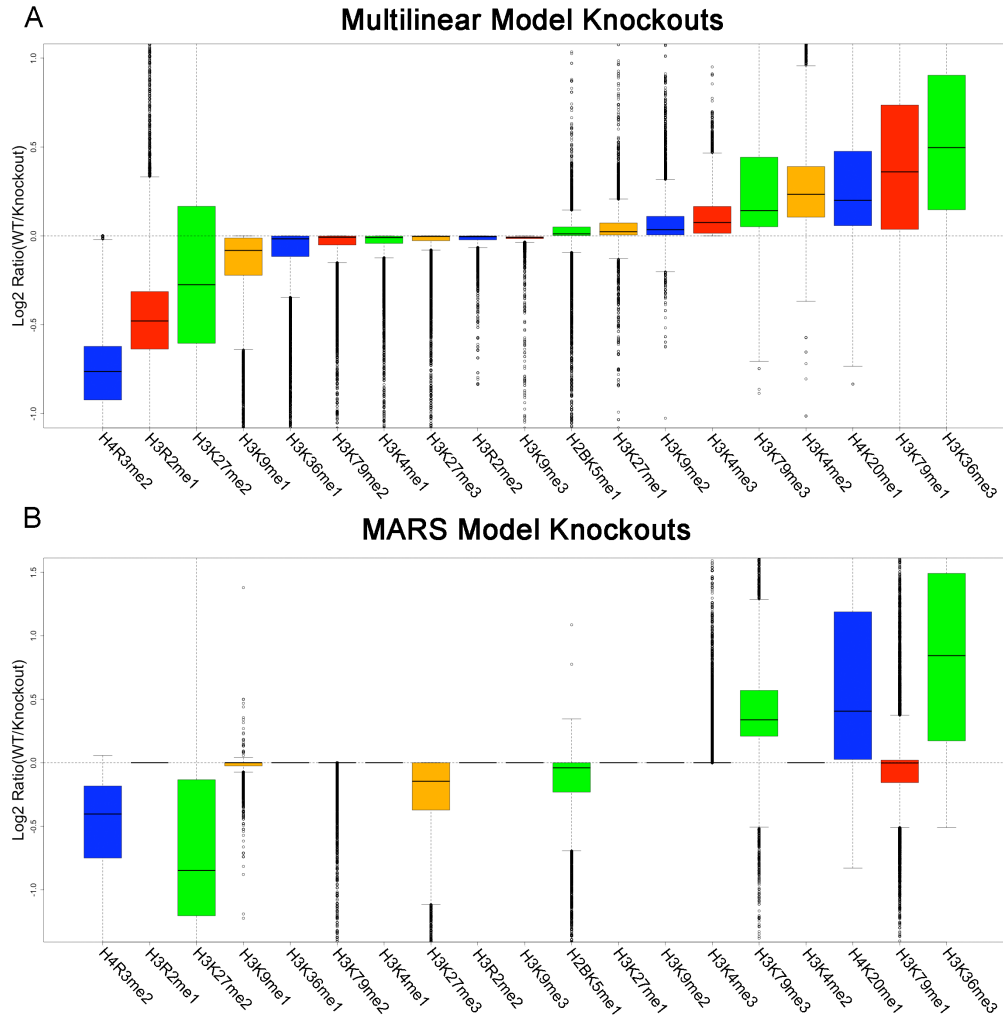


Figure 2.5: **Box plots of MLM and MARS knockouts.** Box plots representing the predicted log₂ fold change (WT/KO) in gene expression after knocking out (setting mark amplitude to zero) a single mark while holding all other amplitudes at their experimental values in both the multilinear (A) and MARS (B) models. Negative shifts indicate repressive marks and positive shifts indicate activating marks. Both models show general agreement in knockout effects. Interestingly, both models choose H4R3me2 to be among the most globally repressive marks, whereas previous studies comparing H4R3me2 levels to gene expression have shown little to no correlation, suggesting the repressive character of H4R3me2 becomes apparent in a multivariate analysis of multiple modifications.

levels to increasing gene expression [Barski et al., 2007]. Indeed, we used the data generated by Barski et al., but came to diametrically opposed conclusions regarding H4R3me2's influence on gene expression. Moreover, this conclusion is not altered by the selection of bin size, as H4R3me2 is the most highly repressive mark in the 6 k

bin-based MARS model, and second most repressive in the 8 k and 10 k bin-based MARS models (see section 2.5 for robustness analysis).

In order to make sense of the apparently contradictory behavior of H4R3me2, we first note that our analyses differed from Barski et al. in two major ways: (1) We estimated mark amplitudes using a model based weighted average, and (2) we modeled gene expression response as in the context of all of the marks simultaneously. A trivial explanation would be that our amplitude estimation procedure alone yielded a response of H4R3me2 levels to gene expression, which was then reflected in the ML and MARS fitting coefficients. We directly tested this by generating box plots of our amplitudes stratified by quartiles of gene expression shown in Figure 2.6. Consistent with the composite plots from [Barski et al., 2007], we observe little to no response of H4R3me2 amplitudes with increasing gene expression. For comparison, we generated the same box plots for H3K27me2 and observed a dramatic decrease in its amplitude with increasing gene expression level, as expected for this mark. Thus, we ruled out the mark amplitude estimation procedure as an explanation. Thus we are left with the more interesting case: H4R3me2’s repression of gene expression is revealed by accounting for the simultaneous impact of the other histone methylations.

In order to better understand H4R3me2’s effect on gene expression, we performed a comparative analysis with H3K27me2, which is highly repressive in the ML and MARS model knockout analysis, as well as in univariate analyses. Specifically, we divided the ML model wild type over knockout \log_2 fold changes by quintiles. We then calculated box plots of predicted gene expression in the WT and KO cases as shown in Figures 2.7A and 2.7B, respectively. The first 20% of the data (QU1) represents the genes most up-regulated by knocking out the mark (i.e., the largest de-repression of gene expression). The last 20% of the data (QU5) represents the genes least up-regulated by knocking out H4R3me2 and genes down-regulated by knocking out H3K27me2. The trends in WT and KO gene expression across the stratified data are opposite for

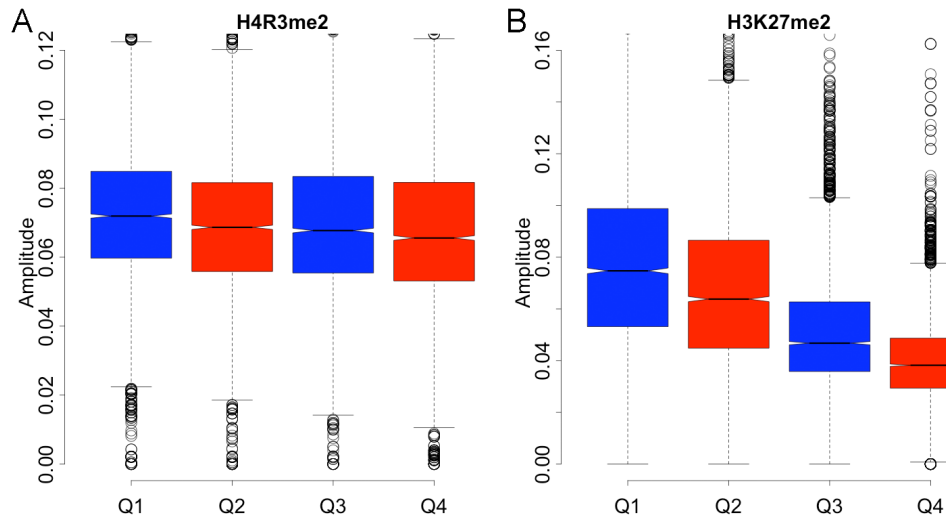


Figure 2.6: **Box plots of amplitudes across expression.** Box plots of H4R3me2 (A) and H3K27me2 (B) amplitudes across the genes, stratified by quartiles of gene expression, where Q1 and Q4 represent the lowest and highest gene expression groups, respectively.

H4R3me2 and H3K27me2. For H4R3me2, the median \log_2 gene expression in the WT is relatively low (4) in QU1 and increases slightly (to 4.7) in QU5, with the knockout showing a similar trend (Figure 2.7A). In contrast, the H3K27me2 WT median starts out considerably higher in QU1 (5.9) and plummets 8-fold (to 2.9) in QU5 with the knockout again showing a similar trend (Figure 2.7B). Thus, H4R3me2 tends to be consistently acting on relatively low expressed genes. Its removal is predicted to increase their expression 1.7-fold, on average. H3K27me2, on the other hand, has the highest impact on middle to high expressed genes, and little impact on relatively silent genes.

We also calculated the proportion of significantly enriched marks found in quintile-stratified knockouts (Figure 2.8). Interestingly, the profiles of site proportions across the stratified data clustered into activating marks (Figure 2.8A, D), arginine methylations (Figure 2.8B, E), and repressive marks (Figure 2.8C, F). As expected, we found that the proportion of H4R3me2 sites is highest in the highest QU1 of the H4R3me2 knockout, and monotonically decreases across subsequent groups (Figure 2.8B). In the H4R3me2

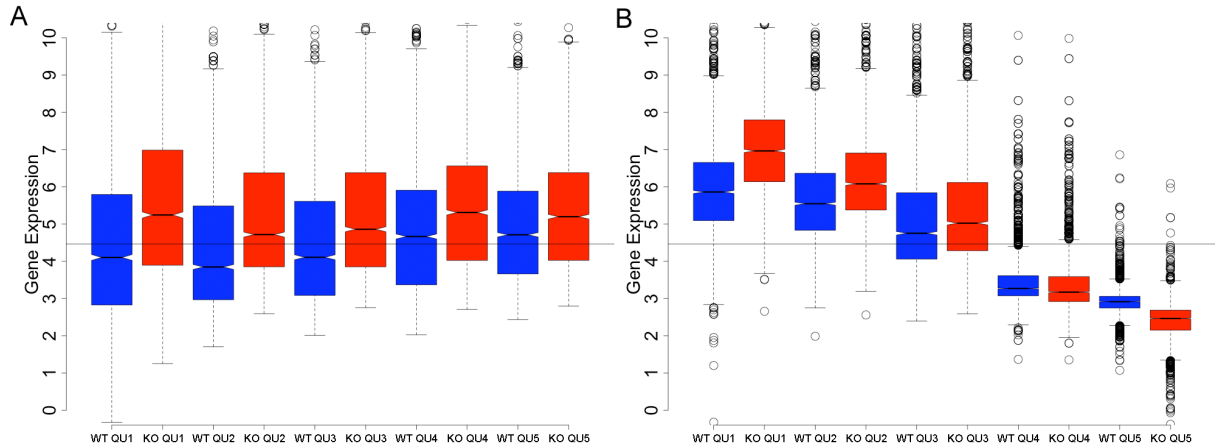


Figure 2.7: **Box plots of predicted gene expression before and after knockout.** Box plots of predicted gene expression before and after knockout of (A) H4R3me2 and (B) H3K27me2. Plots are stratified along the x-axes by quintiles of \log_2 fold change

knockout, the activating marks site profiles (Figure 2.8A) tend to be relatively low and flat or mildly increasing with decreasing \log_2 fold change. The repressive mark site profiles in the mid-range (Figure 2.8C). Thus, arginine methylation itself appears to drive the impact of H4R3me2 knockout on gene expression. In stark contrast, in the H3K27me2 knockout, the proportion of H3K27me2 and other repressive mark sites increase monotonically with decreasing knockout impact (Figure 2.8F). Activating marks showing the exact opposite trend (Figure 2.8D). The arginine methylation sites cluster together in the H3K27me2 knockout as well, and show a preference for QU5 (Figure 2.8E). Interestingly, we found that the largest impact of knocking out H3K27me2 tends to be in genes where H3K27me2 levels are relatively low and activating mark levels are relatively high. For these genes, H3K27me2 appears to be modulating or reducing gene expression from high to moderate levels (Figure 2.7B).

Taken together these analyses suggest that H4R3me2 and the other arginine methylations tend to be somewhat uncorrelated with established activating marks and repressive marks. Consequently, its absence at genes does not imply the presence of activating marks and high levels of expression (Figure 2.3B). Conversely, high levels of H4R3me2 can coincide with modest to relatively high levels of activating marks

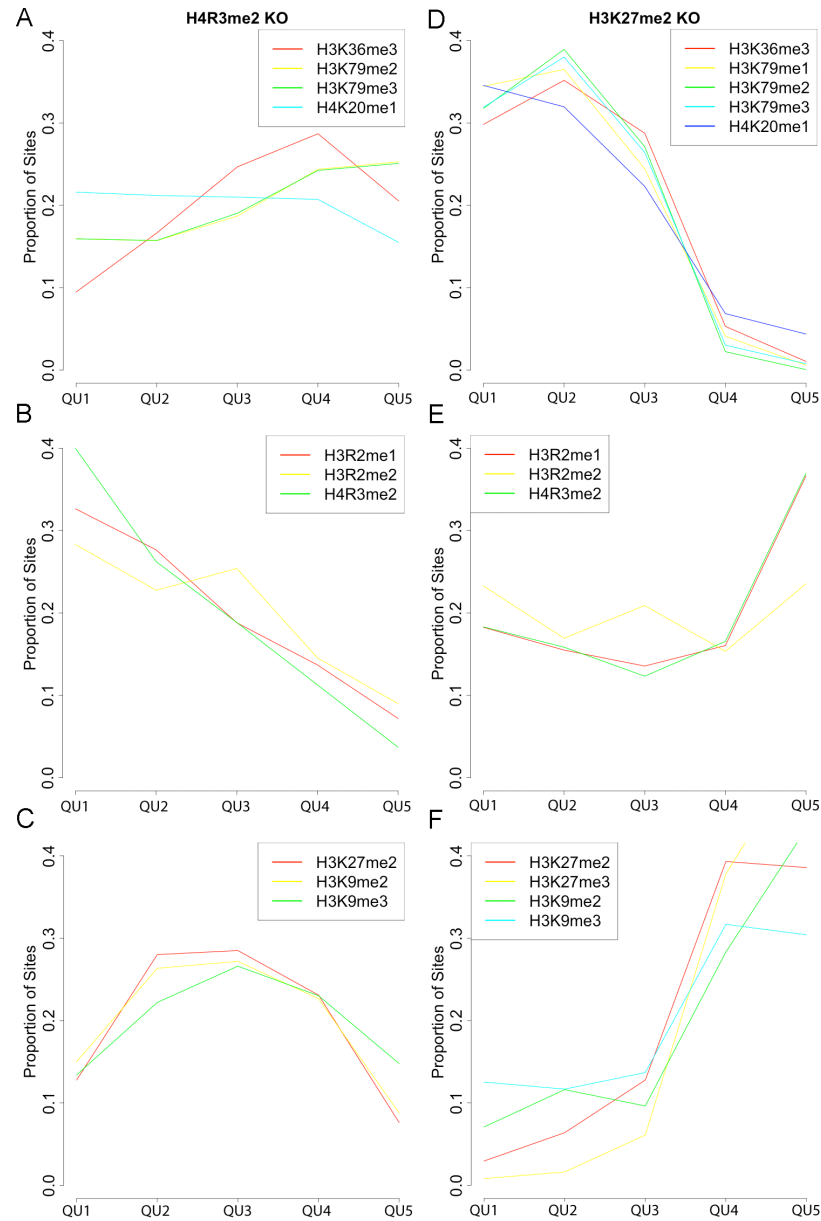


Figure 2.8: **Enriched sites across MLM knockout quintiles.** Plots show the proportion of significantly enriched sites identified by MACS (y-axis) for marks shown in the legend across the data divided by quintiles of \log_2 fold change (WT/KO) in gene expression predicted by the MLM for H4R3me2, (A)-(C), and H3K27me2, (D)-(F), knockouts. Proportions of sites were clustered using k-means clustering. For both knockouts activating marks clustered together, (A) and (D), as did arginine methylations, (B) and (E), and repressive marks, (C) and (F). H4R3me2 knockout effect only shows a strong correlation with other arginine methylations (B), while the H3K27me3 knockout effect shows strong anti-correlation with the activating marks (A) and strong positive correlation with other repressive marks (F).

like H3K36me3, which tend to override H4R3me2 (see Table 2.4, Table 2.7 and Figure 2.3B). Thus, its levels show no strong trend with overall gene expression, consistent with [Barski et al., 2007]. Instead, dimethylation of H4R3 consistently tends to further repress low to modestly expressed genes—nearly 2-fold on average. For this reason, the ML and MARS models predicted H4R3me2 to have strong repressive activity.

2.3.8 Experimental studies demonstrate H4R3me2 represses gene expression

The antibody used for the ChIP-seq experiment recognizes symmetric dimethylated H4R3 (H4R3me2s), which is deposited by the arginine methyltransferase PRMT5. A number of experimental studies have shown that PRMT5 and H4R3me2s repress gene expression [Wang et al., 2008, Hou et al., 2008, Litt et al., 2009, Zhao et al., 2009]. In an experiment that is a direct analogue of our knockout analysis, silencing of PRMT5 in mouse cell lines resulted in globally more de-repressed than repressed genes [Pal et al., 2004], supporting our result that H4R3me2s is globally repressive. PRMT5 is a member of the multi-subunit mSin3A and NuRD histone deacetylase complexes [Pal et al., 2003], suggesting H4R3me2 is associated with deacetylation and hence gene inactivation [Pal et al., 2003, Pal et al., 2004, Wysocka et al., 2006]. Interestingly, both the mSin3A-PRMT5 containing complex and recombinant PRMT5 methylate H4R3, and show an in vitro preference for methylating hypo- versus hyperacetylated histone H4R3 [Pal et al., 2003]. PRMT5 was also shown to interact with the MBD2/NuRD complex and that PRMT5 and MBD2 are recruited to CpG islands in a methylation-dependent manner, with H4R3 methylated at these loci [Le Guezennec et al., 2006]. These results are consistent with our finding that H4R3me2 tends to further repress modest to low expressed genes, which are likely hypoacetylated, or contain methylated CpG islands in their promoter regions, or both.

In a recent study, H4R3me2s was shown to be required for subsequent DNA methylation [Zhao et al., 2009]. Indeed, H4R3me2s was shown to be a direct binding target of the DNA methyltransferase DNMT3A. Loss of H4R3me2s through shRNA knockdown of PRMT5 resulted in reduced DNMT3A binding, loss of DNA methylation and six-fold induction of the fetal γ -globin gene [Zhao et al., 2009].

2.4 Conclusions

Current genomic strategies for assessing whether a particular histone modification is activating or repressive involve (1) mapping it to the genome using ChIP-chip or ChIP-seq and either (2) comparing the expression distribution of genes with and without the mark [Bernstein et al., 2005], or (3) generating composite plots of average mark levels of genes stratified by gene expression level. Using the latter approach, Barski et al. concluded that H4R3me2 is neither activating nor repressive because its levels showed no response with increasing gene expression level [Barski et al., 2007]. Using ChIP-seq data of 20 histone lysine and arginine methylations and histone variant H2A.Z in CD4⁺ T cells, we built models of gene expression as a function of histone modification/variant levels using Multilinear (ML) Regression and Multivariate Adaptive Regression Splines (MARS). The response of monovalent (non-interacting) terms in the ML and MARS model indicate whether a given modification is activating or repressive. For most of the 20 histone methylations, our assignments agree with previous analyses [Barski et al., 2007]. However, according to our in-silico knockout analysis, among the 20 methylations, H4R3me2 is predicted to be among the most globally repressive. A number of experimental studies show that PRMT5-catalyzed symmetric dimethylation of H4R3 is associated with repression of gene expression [Wang et al., 2008, Hou et al., 2008, Litt et al., 2009, Zhao et al., 2009]. This includes a recent study, which demonstrated that H4R3me2 is required for DNMT3A-mediated DNA methylation

[Zhao et al., 2009]—a known global repressor of gene expression. Consequently, this study serves as the first demonstration that H4R3me2 represses gene expression using genomic data, and shows that the regulatory role of some modifications like H4R3me2 can only be revealed by approaches that simultaneously analyze multiple activating and repressive modifications. Our findings point to a disconnect between traditional biochemical (e.g., silencing) and genomic approaches in assessing the activating or repressive potential of an individual modification. Indeed, assuming the biochemical studies are correct and H4R3me2 is repressive, one would conclude from the analysis in [Barski et al., 2007] that the antibody they used for H4R3me2 did not work. Our results suggest that it worked extremely well. Taken together, our findings have broad implications for ChIP-seq experimental design, analysis, and interpretation.

2.5 Methods

2.5.1 Calculation of amplitudes

Enrichment levels, or amplitudes, for each of the 21 histone modifications were estimated for each gene using a spatially weighted average of the mapped ChIP-seq tag counts (see Table 2.8 for the range of amplitude values). The gene list used in this study was compiled from the NCBI36 Homo sapiens database (Ensembl 54, downloaded June 24, 2009). For each mark j , an average enrichment template, t_{ij} across the 5' flanking region (i.e., -2 kbp before the transcription start site), the body of a scaled gene (a gene divided into a fixed number of bins), and the 3' flanking region (i.e., transcription stop site to +2 kbp), was first calculated as a function of relative genomic position i . For both the 5' flanking region and 3' flanking region, the coordinate i represents each nucleotide position relative to the transcription start and stop sites, respectively. Within gene bodies (i.e., transcription start site to stop site), the coordinate i represents the position in the gene body, which is divided into 8138

segments, or bins, which corresponds to the median gene length. For genes whose lengths are greater than 8138 bp, tag counts were averaged across bases within each of the 8138 bins. For genes whose lengths are less than 8138 bp, tag counts were repeated in order to generate 8138 bins. For genes not divisible by 8138 or a divisor thereof, fractions of base pairs within a bin were rounded to the nearest integer value; therefore bins containing the majority of the fraction received the full tag count value of the corresponding base pair, while the bin containing the minority received no part of the bisected base pair's value. The median value was used for the bin number to minimize biases introduced in scaling. Large bins would tend to over-smooth large genes, while small bins would tend to overrepresent copied values from small genes. The t_{ij} or template for mark j , was finally computed by (1) aligning the transcription start and stop sites of every scaled gene and then (2) calculating the average bin-averaged tag count across genes for every coordinate i . All templates were then normalized so that their average across bins was 1, such that:

$$\frac{1}{N} \sum_i t_{ij} = 1 \quad (2.1)$$

where N is the number of bins. In other words, the template is the averaged and normalized enrichment profile across all scaled genes. Because the template appears to have a characteristic shape for a given mark j across the length of scaled genes, we developed a model of relative enrichment which assumes the actual profile of any given mark is given by a product of a gene-dependent amplitude, X_j^k , for a gene, k , and the mark's template t_{ij} . In other words, gene k 's tag count profile for mark j across genomic coordinate i , c_{ij}^k , is well approximated by $X_j^k t_{ij}$. Using least squares, we minimized the difference between the model and the actual tag count profiles:

$$Q_j^k = \sum_i (c_{ij}^k - X_j^k t_{ij})^2 \quad (2.2)$$

to arrive at the following equation for mark j 's amplitude at gene k :

$$X_j^k = \frac{\sum_i c_{ij}^k t_{ij}}{\sum_i t_{ij}^2} \quad (2.3)$$

We note that in the special case where the template is constant as a function of genomic position i , X_j^k reduces to a simple average of tag counts across bins

$$X_j^k = \frac{1}{N} \sum_i c_{ij}^k \quad (2.4)$$

which is the appropriate estimate of tag “depth” for a mark whose tag distribution is uniform across a gene.

Table 2.8: Distribution of mark amplitudes

| Mark | Min | 5% | 25% | 50% | 75% | 95% | Max |
|----------|-----|--------|--------|--------|--------|--------|---------|
| H2AZ | 0 | 0.0092 | 0.0405 | 0.126 | 0.2807 | 0.4591 | 1.8056 |
| H3K4me3 | 0 | 0.0219 | 0.1108 | 0.5678 | 1.247 | 2.002 | 11.931 |
| H3K9me3 | 0 | 0.005 | 0.0106 | 0.0166 | 0.0289 | 0.0508 | 50.3513 |
| H3K36me1 | 0 | 0.0452 | 0.0713 | 0.0878 | 0.1057 | 0.1237 | 2.6032 |
| H3K4me1 | 0 | 0.0259 | 0.1031 | 0.2287 | 0.4179 | 0.6962 | 3.244 |
| H2BK5me1 | 0 | 0.0244 | 0.0634 | 0.1594 | 0.3162 | 0.5618 | 3.7164 |
| H3K4me2 | 0 | 0.0137 | 0.0595 | 0.1554 | 0.2461 | 0.3377 | 1.7125 |
| H3K9me1 | 0 | 0.0297 | 0.1055 | 0.2298 | 0.3477 | 0.4621 | 1.6088 |
| H3K9me2 | 0 | 0.0202 | 0.0364 | 0.0507 | 0.0711 | 0.0985 | 0.3064 |
| H3K27me1 | 0 | 0.0275 | 0.0691 | 0.1455 | 0.2091 | 0.256 | 2.6492 |
| H3K27me2 | 0 | 0.0222 | 0.0371 | 0.0521 | 0.0763 | 0.1024 | 2.1456 |
| H3K27me3 | 0 | 0.0152 | 0.0255 | 0.0365 | 0.0891 | 0.1558 | 4.8271 |
| H3K36me3 | 0 | 0.051 | 0.0965 | 0.1998 | 0.3217 | 0.4409 | 3.4295 |
| H3K79me1 | 0 | 0.0053 | 0.022 | 0.1237 | 0.2194 | 0.3066 | 0.998 |
| H3K79me2 | 0 | 0.0026 | 0.0096 | 0.0479 | 0.17 | 0.3635 | 1.8122 |
| H3K79me3 | 0 | 0.0077 | 0.0177 | 0.0576 | 0.2113 | 0.5024 | 3.8415 |
| H4K20me1 | 0 | 0.0155 | 0.0534 | 0.1909 | 0.5139 | 1.2093 | 13.9068 |
| H3R2me1 | 0 | 0.066 | 0.0933 | 0.1146 | 0.1431 | 0.1741 | 2.2685 |
| H3R2me2 | 0 | 0.0399 | 0.0576 | 0.0684 | 0.0819 | 0.0998 | 0.57 |
| H4K20me3 | 0 | 0.0115 | 0.0196 | 0.0264 | 0.0356 | 0.049 | 9.9591 |
| H4R3me2 | 0 | 0.0382 | 0.0559 | 0.0686 | 0.083 | 0.0975 | 0.4536 |

Mark amplitudes at various quantiles, as well as minimum and maximum values.

2.5.2 Selection of transcription start and stop sites

Many Ensembl genes contain multiple start and stop sites. Given that we only have 3' biased gene expression data, there are cases where we cannot unambiguously assign an Affymetrix probe set to one transcription start or stop site which we need for our estimate of mark enrichment. Consequently, we chose the transcription start sites that were associated with the highest number of significantly enriched histone modifications as representing the most likely expressed transcript. If a selected start site had multiple stop sites, we chose a stop site using the same scheme. In cases where multiple transcription start sites had the same number of significant marks, the most upstream transcription start site was chosen. When multiple stop sites for a given start site had the same number of significant marks a stop site was arbitrarily selected.

To determine the number of significantly enriched marks for a particular transcription start and stop site, we first calculated the distribution of mark amplitudes for all Ensembl genes. The left tail relative to the mode of the distribution of amplitudes for a particular mark was used to build a Gaussian null model as a background noise model for that mark. The mode of the amplitude distribution was used as the mean of the null model, and the standard deviation of the null model was derived using the following equation:

$$\sigma_j = \sqrt{\frac{1}{n-1} \sum_{k^*} (X_j^{k^*} - \mu_j)^2} \quad (2.5)$$

where μ_j is the mode of the amplitude distribution and the sum is over genes k^* whose amplitude is less than or equal to the mode, $X_j^{k^*}$, and n is the number of genes that satisfy this inequality. This null model was used to determine the p-value by calculating the integral of the Gaussian from the mark amplitude to infinity for each mark at every Ensembl gene [Buck et al., 2005, Gibbons et al., 2005].

A Benjamini-Hochberg false discovery rate (FDR) [Benjamini and Hochberg, 1995]

correction was applied to the p-values using the `p.adjust` function in R, and an FDR-corrected p-value cutoff of 0.05 was used to determine significantly enriched amplitudes.

2.5.3 Building the multilinear model using stepwise linear regression

We built the multilinear model using a stepwise linear regression procedure (*stepwisefit* in MATLAB), which models gene expression as a function of histone mark enrichment according to the following equation:

$$Y^k = \beta_0 + \sum_j \beta_j X_j^k + \sum_{j<l} \beta_{jl} X_j^k X_l^k + \sum_{j<l<m} \beta_{jlm} X_j^k X_l^k X_m^k + \epsilon^k \quad (2.6)$$

where Y^k is the normalized \log_2 gene expression (using GCRMA [Wu et al., 2004]); β_j , β_{jl} , β_{jlm} are mono-, bi- and trivalent histone modification fitting coefficients; are mark j amplitudes for gene k and the ϵ^k are random errors. Briefly, an initial model is fit with randomly selected terms-defined as β coefficients multiplied by one, two or three mark amplitudes. Terms from the set that are not in the initial model and make a statistically significant contribution to the model (i.e., p-value ≤ 0.05 according to an F-test) are added during a forward step. The forward step continues until no terms from the available pool of unused terms contribute significantly to the model. A backward step is then applied whereby terms are ranked in descending order according to their p-values and removed if they are not significant (i.e., p-value > 0.05). The backward step ends when no terms in the model are insignificant. The forward and backward steps are repeated until no significant terms can be added or removed, respectively.

Because stepwise linear regression is not guaranteed to converge to a globally optimal solution (i.e., minimum adjusted R^2) for any given initial seed model, we

performed multiple rounds of multiple stepwise regressions using different randomly seeded models. In the first round, we ran *stepwisefit* on the full dataset 100 times using randomly seeded models. This resulted in 100 models with a mean of 227 terms. To assess the statistical significance of a given term’s survival rate across the models, we randomly sampled 227 of the 1561 possible terms to generate a null model. While a survival rate of 0.2 was significant ($p\text{-value} < 0.05$), to increase stringency we arbitrarily selected a cutoff of 0.35 to arrive at 167 starting terms for the next round of stepwise linear regression.

To avoid the problem of overfitting and its inflation of model complexity, we applied *stepwisefit* to 10-fold cross validation data. Specifically, for each of the 10 folds we performed 10 runs of *stepwisefit* where the initial model contained the 167 terms found in the first round plus an additional 60 randomly selected terms (i.e., we generated 100 models). Using the test data, we applied only the backward step of the stepwise procedure to assess the significance of every term that survived the training step and removed those with $p\text{-values} > 0.05$. Among the 10 runs for each fold, the model with the lowest test mean square error (MSE) was selected. This resulted in 10 models for each fold. We then required a term to appear in 5 or more of the 10 models generated within each fold to be selected for the final model. This resulted in 24 terms.

We arrived at a robust estimate of the final set of 24 coefficients by fitting the training data to a model that contained only the 24 terms. This yielded 10 sets of 24 coefficients (i.e., one for each fold of the 10-fold training data). We arrived at the final value of each fitting coefficient by calculating the trimmed mean of the 10 found in each fold. The final model’s performance was assessed by calculating the mean MSE and adjusted R^2 across the 10 test and training data folds (see Results and Discussion).

2.5.4 Building the MARS model

The relationship between gene expression level and each mark's average enrichment tends to be nonlinear including saturation of gene expression response as a function of mark level. The *earth* package in R was used to build the MARS model, which naturally accounts for non-linear responses between the input and output variables. Briefly, a MARS model is the sum of basis functions multiplied by a coefficient to be determined from a regression analysis of the function

$$Y^k = c_0 + \sum_{i=1}^n c_i b_i \left(\vec{X}_i^k \right) \quad (2.7)$$

where Y^k is \log_2 gene expression of gene k (i.e., output variable), c_0 is a constant, \vec{X}_i^k is the subset of mark amplitudes that appear in term i , and $b_i(\cdot)$ is a basis function that is made up of either one or a product of two or more hinge functions. Hinge functions are splines that take the form $h(X_j^k) = \max(0, X_j^k - X_j^{k*})$ or $h(X_j^k) = \max(0, X_j^{k*} - X_j^k)$ where X_j^{k*} is a special constant known as a *knot*. We note that the two hinge functions shown above are a symmetrical pair about the vertical line $X_j^k = X_j^{k*}$.

The MARS model was built in one forward and one reverse pass. The forward pass builds the model using a greedy algorithm. It begins with an intercept term that is equivalent to the mean of the observed response variable, which is \log_2 gene expression level in our case. The algorithm then searches for the monovalent contributions fitted as a pair of symmetrical basis functions, which maximally reduce the residual sum-of-squares (RSS) at each step. It then adds bivalent terms, which are constrained to contain one of the monovalent terms and maximally reduce the RSS at a given step until a minimum RSS reduction is reached. It adds trivalent terms, which are constrained to contain one of the bivalent terms and maximally reduce the RSS at a given step until a minimum RSS reduction is reached. The reverse pass then prevents

overfitting by removing terms to optimize a Generalized Cross Validation (GCV) score. The GCV penalizes model complexity by dividing the RSS by the effective number of degrees of freedom in the model;

$$GCV = \frac{RSS}{N_g \left(1 - \frac{T+P(B)}{N_g}\right)^2} \quad (2.8)$$

where RSS is the residual sum of squares, N_g is the number of observations or genes with expression data in our case, T is the total number of terms in the model, P is a user-defined penalty (*earth*'s default is 3 for multivalent models), and B is the total number of non-constant basis functions in the model.

2.5.5 Amplitude robustness and relative error of mark amplitude estimates

Our amplitude estimation procedure is motivated by the observation that a number of histone methylations (e.g., H3K36me3, H4K20me1, H2BK5me1, etc.) are pervasive across the body of genes, and their enrichment patterns appear to scale with gene length. However, modifications occur in the context of nucleosomes which are associated with approximately 146 bp of DNA. Thus, our gene scaling procedures average different numbers of nucleosomes depending on the gene length and the selected bin size. Consequently, we assessed the robustness of our amplitude estimation procedure by recalculating our template and amplitude using 6000 (6 k) and 10,000 (10 k) bins and compared them to those calculated using our final choice of 8138 (8 k) bins (see 2.5 for details).

We first calculated the Spearman correlation coefficients of the 6 k versus 8 k and 10 k versus 8 k amplitude estimates for all 21 histone modifications/variants. We found the values to be highly correlated, with coefficients ranging from 0.994–0.9995 and 0.9975–0.9998 across marks for the comparisons of 8 k bins to 6 k and 10 k,

respectively. We also calculated the fractional difference (i.e., difference divided by mean) between 6 k and 8 k and 10 k and 8 k amplitude estimates. The absolute value of typical (50th percentile) fractional differences range from 0.0023–0.065 and 0.0016–0.042 for the comparisons of 8 k to 6 k and 10 k, respectively. Indeed, the worst values were 0.22 and 0.16 for 8 k versus 6 k and 10 k, respectively. Thus, our estimates of mark enrichment amplitudes are relatively robust with respect to bin size. Given these results, it's not surprising that our model results and main conclusions do not depend on bin size.

An advantage of a model-based approach to estimating enrichment levels is that we can directly assess model performance by calculating residuals. To assess the fit of our template model to the data we calculate the coefficient of variation of the root mean square deviation $CV(RMSD)$ for every gene, which is defined by

$$CV(RMSD) = \sqrt{\frac{\sum_i^n (c_{ij}^k - X_j^k t_{ij})^2}{nX_j^k}} \quad (2.9)$$

where n is the number of bins in the template, and all other variables follow previous definitions. In addition to the gene amplitude calculation with the 8138-bin (8 k) template, amplitudes were also calculated with 6000 (6 k) and 10,000 (10 k) bin (plus flanking regions) templates. To assess the robustness of our amplitude estimates, we calculated the Spearman correlation coefficient between the 8 k and 6 k bin amplitudes and the 8 k and 10 k bin amplitudes. We also calculated the fractional difference between the 8 k and 6 k bin amplitudes, , and the 8 k and 10 k bin amplitudes, which is similarly defined. Finally, the $CV(RMSD)$ was calculated for all marks for the 3 sets of amplitude calculations to assess effect of bin size selection on the template model fit.

In a plot of $CV(RMSD)$ versus amplitude for every mark we found a near universal curve (Figure 2.9). This results in part because our normalization of each mark's

template (i.e., their average across bins equals 1) allows us to interpret the mark amplitude as a model-based estimate of each mark’s effective read coverage. As might be expected, below amplitude values of 1 (i.e., $1\times$ coverage) the error grows rapidly. For relatively large amplitudes (i.e., values greater than 1), the CV asymptotically reach values slightly below 2. In contrast, marks whose largest amplitudes fall well below 1 have CV values that range from 2.3–5 at the largest amplitudes encountered (i.e., 95th amplitude percentile). For most marks, the 95th amplitude percentile is below 1, indicating from our crude gene-centric measure of coverage/read density that the effective sequencing coverage might be low. We also observed a steady trend upward in the CV across marks at the 95th amplitude percentile with decreasing amplitude levels. Taken together, these results indicate that RMSD between the model and the data are on the same order as the amplitude. We also note that we found essentially the same CV(RMSD) values for 6 k and 10 k bins. Nevertheless, we found that our weighted average estimate is relatively robust and should capture enrichment level trends in histone modification/variant ChIP-seq data reasonably well.

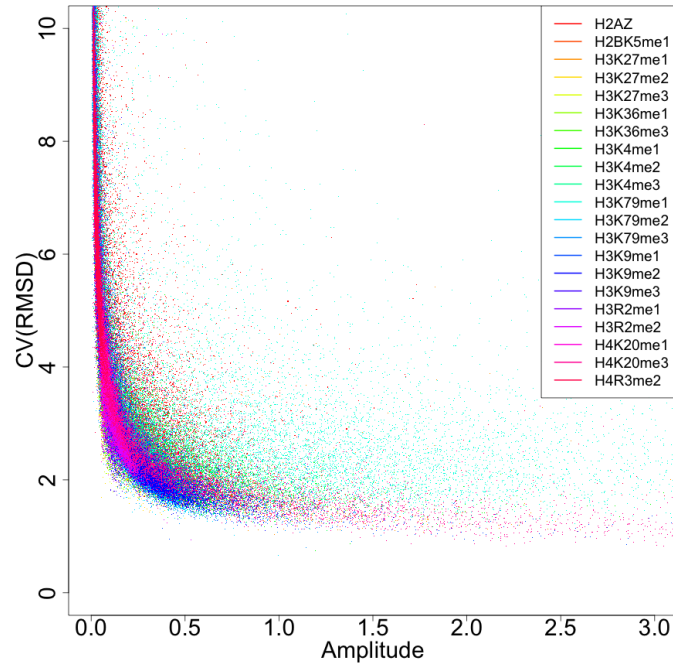


Figure 2.9: **Relative error of mark enrichment models.** $CV(RMSD)$ versus amplitude. Colors represent different marks as shown in the legend. Low amplitudes correspond to low levels/coverage, and thus high $CV(RMSD)$ values. As amplitude increases, values reach an asymptotic value.

2.6 Chapter acknowledgements

This chapter was adapted from [Xu et al., 2010]. Xiaojiang Xu deserves explicit credit for the construction and analysis of the multilinear model.

Chapter 3

Quantification of histone modification ChIP-seq enrichment

3.1 Introduction

Recent advances in high-throughput DNA sequencing technology have facilitated the generation of vast amounts of epigenomic ChIP-seq data. The availability of these datasets has provided the opportunity to utilize the power of statistical computing to model epigenetic regulatory systems. Unlike conventional biochemical approaches, the application of machine learning and data mining techniques to ChIP-seq data is capable of providing a broad, systems-level view of the epigenetic regulation. These strategies can provide insights into mechanisms of genomic control, such as the so-called “histone code” [Strahl and Allis, 2000, Jenuwein and Allis, 2001], by facilitating an integrated analysis of the many histone posttranslational modifications (PTMs) that have been described, as well as other epigenetic chromatin modifications. The histone code is a particularly attractive problem for computational applications, since it has become apparent that histone PTMs are regulated in a network fashion and are

deposited combinatorially [Latham and Dent, 2007, Wang et al., 2008, Xu et al., 2010]. However, a thorough study of the gene-biased quantification of ChIP-seq enrichment for the application of machine learning techniques has not yet been done.

Several groups have applied various machine learning techniques to epigenomic ChIP-seq data in a gene-biased fashion, including Bayesian networks [Yu et al., 2008, Cheng et al., 2011], support vector machines and regression [Cheng et al., 2011], and linear [Karlić et al., 2010, Xu et al., 2010, Cheng et al., 2011] and nonlinear regression [Xu et al., 2010]. These studies have focused on histone modifications due in part to their known role in transcriptional regulation [Kouzarides, 2002, Barski et al., 2007, Campos and Reinberg, 2009] and the availability of rich datasets exhibiting a wide variety of types of histone modifications [Barski et al., 2007]. In the case of supervised learning, the models created in these studies were built using individual genes as observations, ChIP-seq enrichments as the predictor variables, and gene expression as the response variable. In these models the predictors are histone modification/variant ChIP-seq enrichment levels for individual genes, so model quality is highly dependent on the accuracy of enrichment estimation. Since ChIP-seq enrichment levels are strongly dependent on genomic coordinate, providing a gene-wise estimate of ChIP-seq enrichment that accurately captures the relevant enrichment information across all genes—which vary in length over four orders of magnitude—is a challenging task.

The most straightforward way to estimate per-gene ChIP-seq enrichment is to simply count the number of sequence reads associated with a given gene. Indeed, counting the number of reads in the promoter region of each gene was an approach taken in some previous studies [Yu et al., 2008, Karlić et al., 2010]; however, this method has several limitations. First, tag-counting methods equally weight every position within the counting window, and thus ignore the spatial component of the enrichment data. Second, not every histone modification is 5' biased. Several modifications have greater enrichment into the body of genes, such as H3K36me3 [Barski et al., 2007, Campos

and Reinberg, 2009, Hon et al., 2009b]. This modification highlights the pitfalls associated with 5' biased enrichment estimation. It is mainly enriched in the bodies of genes, it is 3' biased, and it has a proclivity for enrichment in exons [Hon et al., 2009b, Kolasinska-Zwierz et al., 2009]. H3K36me3 also has a strong correlation with transcriptional elongation as determined by various biochemical studies [Barski et al., 2007, Kolasinska-Zwierz et al., 2009, Bannister et al., 2005, Krogan et al., 2003]. It is worth noting that a machine learning study by Yu et al. used a 5' tag counting method and found little correlation between H3K36me3 enrichment and gene expression [Yu et al., 2008]. The consequence for models that use 5' proximal enrichment estimation methods is that the effects of histone modifications with gene body or 3'-biased enrichment are underestimated or greatly obscured.

Histone methylations tend to have unique average spatial deposition patterns [Barski et al., 2007]. For example, in contrast to H3K36me3, H3K4me3 has high enrichment around the transcription start site, with depletion in the nucleosome-free region. Some modifications seem to be deposited in a specific genic region with respect to the absolute positions of nucleosomes relative to gene boundaries, while the patterns of others seem to scale with gene length. These effects can be attributed, at least in part, to the recruitment of histone methyltransferases that are dependent on the phosphorylation state of the C-terminal domain of Pol II before and during the elongation process [Hon et al., 2009b, Krogan et al., 2003, Komarnitsky et al., 2000, Ng et al., 2003]. Estimating ChIP-seq enrichment is complicated by the different modes of histone PTM deposition coupled with the wide variability in gene lengths. However, part of the information content associated with a given histone modification is encoded within the spatial distribution of the enrichment data, and so it should also be considered when estimating enrichment levels.

The selection of a genomic window used for the calculation of enrichment levels is important in capturing relevant enrichment data, since enrichment information

may be 5' or 3' proximal, intergenic, or intragenic [Cheng et al., 2011]. Choosing a window size that extends too far outside of gene boundaries may incorporate data from neighboring regions, and selecting a window size that is too small may exclude useful data. The goal in window selection is to maximize useful information content and minimize the incorporation of noise, while being generalizable across a variety of marks.

Since the quality of a statistical model is largely dependent on the quality of the observation data used to build it, refining enrichment estimation methods is important for future statistical analyses of ChIP-seq data. To resolve some of the issues involved with enrichment estimation, we compared the performance of models built using a ChIP-seq dataset of histone methylations/variants in CD4⁺ T cells generated by Barski et al. [Barski et al., 2007]. This dataset has been used in several other machine learning studies [Yu et al., 2008, Karlić et al., 2010, Xu et al., 2010]. We applied the Multivariate Adaptive Regression Splines (MARS) algorithm [Friedman, 1991a, Friedman, 1991b] to build regression models using enrichment levels of 20 histone lysine and arginine methylations plus histone variant H2A.Z. Given that gene expression levels have been shown to be highly dependent on histone modification levels, we used the Generalized Cross-Validation (GCV) [Friedman, 1991a] and R^2 metrics to assess the quality of MARS model fits and rank enrichment estimation methods.

Several different strategies were employed for estimating gene-wise enrichment levels, including tag counting and model-based approaches, which use average enrichment patterns to spatially weight enrichment of individual genes in a set of genomic windows. We also investigated the selection of window sizes for our gene-wise enrichment estimation methods. By comparing models using GCV and R^2 values, we demonstrate that the performance of regression models using histone modification enrichment levels as predictors can be greatly affected by the chosen enrichment estimation method. We conclude that methods that incorporate the spatial distribution of ChIP-seq

enrichment offer an improvement in a regression fit over tag counting methods. We also observe that whole-gene estimation windows produce superior results relative to estimations restricted to specific genic regions. Indeed, incorporating data across the entire body of the gene was the most important factor in improving the fit of our models. These improvements of gene-wise ChIP-seq enrichment estimation can improve the sensitivity and specificity of the predictions derived from machine learning models.

3.2 Methods

3.2.1 Gene selection

A list of gene transcript annotations was downloaded from the NCBI36 Homo sapiens database, Ensembl 54 (June 24, 2009), which was then filtered to only include transcripts that had Ensembl, UCSC, and RefSeq IDs. Genes that did not have corresponding expression data associated with them were removed from the list. Many of the transcripts within this list contain multiple annotated start and stop sites. Using the same procedure described in [Xu et al., 2010], we select a single Transcription Start Site (TSS) and Transcription End Site (TES) for each gene.

Some of the enrichment estimation methods described in this study calculate enrichment within a window around the TSSs and TESs, the largest of which we employed was ± 3000 bp around each site. To avoid overlap between windows where the enrichment estimate was a combination of estimates from both ends of the genes, the gene list was further filtered to only include transcripts of length 6002 bp or greater. Although not all enrichment estimation strategies have this limitation, the filtered list was used for each enrichment estimation method to allow a fair comparison of the final models. After implementing each of the aforementioned filters, the final gene list totaled 9882 genes.

3.2.2 Tag repeat filter

PCR sequence artifacts or phenomena inherent to the sequencing technology may cause repeat sequences to be produced. These artifacts manifest as large numbers of tags that map to precisely the same genomic coordinates. With the exception of H3K79 methylations, maximum repeats ranged from 231 for H3K4me2 to 4231 for H3K9me3. H3K79me1/2/3 had far fewer repeats with maximums of 23, 26, and 42, respectively. We identified these tag “pile-ups” by searching for multiple tags that mapped to the genome with precisely the same start and stop coordinates (or for differing tag lengths, within the margin of the difference in length). A cutoff of 75 repeats was chosen empirically for the modification H3K4me3 (max repeats = 1166) to filter repeat artifacts from H3K4me3 data. We assumed that the typical number of tags in these piles for a given mark crudely scaled with the total number of tags. Thus, the cutoff was scaled for other modifications by the total tag count relative to H3K4me3, and ranged from 21 (H3K79me2) to 75 (H3K4me3). H3K4me3 was chosen to determine the cutoff because it had the largest total tag count, and its tendency to form large localized peaks relative to the other modifications. This helped ensure that our cutoff was not overly stringent and was only sensitive to extreme outliers.

Using this filtering scheme H3K79me1/2/3 had data removal percentages of 0%, 0.001%, and 0.004%, respectively. H4K20me3, which had relatively large numbers of reads that mapped to repeat sequences, had 5% of all data removed. All other marks ranged from 0.01% (H3K4me1) to 0.5% (H3K9me3) of data removed. Thus, we removed extreme outliers while minimally affecting the overall dataset.

3.2.3 Tag counting

Tag counting is the summation of ChIP-seq reads within a genomic window. Any part of a read falling within a window was included in our tag counts. Following

previous studies using tag counting methods, tag count enrichments were calculated in ± 500 bp, ± 1000 bp, ± 2000 bp, and ± 3000 bp windows relative to both the annotated Transcription Start Sites (TSS) and the Transcription End Sites (TES). While evaluating which genic sub-regions to include in tag counting methods, we assessed how the inclusion of tag counts within exons as a genic sub-region category improved model performance and found their contribution was negligible. We therefore did not consider tag counts solely within exons further. To be clear, exons were not excluded from other counting methods. Another set (one for each window size) of gene-wise count-based enrichment estimates were produced by summing the TSS counts with the TES counts multiplied by a scaling factor for each of the 21 histone marks:

$$E_{jk} = C_{jk}^{TSS} + \alpha_k C_{jk}^{TES} \quad (3.1)$$

Where j is the gene, k is the modification type, E is enrichment estimate, C is the tag count in the window, and α is a scaling factor. The purpose of the scaling factor, α , is to effectively weight the contribution of the 5' and 3' ends with respect to gene expression. The scaling factor was calculated for each modification by optimizing the absolute value of the correlation between the sum of the two tag count values and gene expression:

$$|Cor(Y, C_{jk}^{TSS} + \alpha_k C_{jk}^{TES})| = \left| \frac{Cov(Y, C_{jk}^{TSS} + \alpha_k C_{jk}^{TES})}{\sqrt{Var(Y)Var(C_{jk}^{TSS} + \alpha_k C_{jk}^{TES})}} \right| \quad (3.2)$$

where Y is gene expression level. The correlation was optimized numerically with respect to α , for each modification type, k .

A set of whole-gene tag count enrichments was calculated within a window defined by the gene boundaries plus flanking intergenic regions immediately adjacent to the annotated gene boundaries. Counts were normalized by dividing by the length of the counting window. Sets of normalized counts were calculated for the gene bodies plus

0 bp, 500 bp, 1000 bp, 2000 bp, and 3000 bp overhangs up- and downstream of the gene boundaries.

3.2.4 Iterative model-based enrichment estimation

Using a strategy similar to the one described by [Xu et al., 2010], we created a “template” t_{ik} for each mark k . The template is the normalized average enrichment profile for a given mark, within a window relative to gene coordinates, i :

$$t_{ik} = \frac{1}{N} \sum_{j=1}^N c_{ijk} \quad (3.3)$$

where c_{ijk} is the enrichment of a mark k for a gene j at genomic coordinate i and N is the total number of genes. Templates were normalized by a constant such that

$$\frac{1}{N} \sum_i t_{ik}^n = 1 \quad (3.4)$$

where t_{ik}^n is the normalized template, and N is the number of bins. We assume that the enrichment profile of a given gene can be approximated by a template t_{ik} multiplied by an enrichment level estimate X_{jk} of a mark k for a given gene j . The least squares difference Q_{jk} between the estimated enrichment profile $X_{jk}t_{ik}$ and the actual data is given by

$$Q_{jk} = \sum_i (c_{ijk} - X_{jk}t_{ik})^2 \quad (3.5)$$

By minimizing Q_{jk} with respect to the enrichment estimate X_{jk} and applying the normalization constraint given by equation 3.4 we arrive at the following enrichment estimate equation:

$$X_{jk} = \frac{\sum_i t_{ik}^n c_{ijk}}{\sum_i t_{ik}^{n^2}} \quad (3.6)$$

In addition to using a non-weighted average template as shown in equation 3.3, we minimized Q_{jk} with respect to the template t_{ik} to arrive at the following enrichment

estimate weighted tag count template equation:

$$t_{ik} = \frac{\sum_j X_{jk} c_{ijk}}{\sum_j X_{jk}^2} \quad (3.7)$$

Equations 3.6 and 3.7 can be solved iteratively, subject to the template normalization constraint given by equation 3.4. An iterative solution of these equations minimizes the least squares difference between the modeled enrichment data and the actual data c_{ijk} . In the case of the iterative solution, the template is the enrichment estimate weighted average tag count across genomic coordinate, i . The value of X_{jk} is ultimately a weighted average of enrichment across a genomic window, providing a single-value estimate of enrichment that incorporates information from the spatial distribution of the enrichment data. For our calculations the iterative process continued until the average difference between the n^{th} and the $(n + 1)^{\text{th}}$ set of enrichment estimations converged to less than 5% of the n^{th} set values.

Using this iterative model-based strategy, enrichment levels were estimated around both the TSS and TES in ± 500 bp, ± 1000 bp, ± 2000 bp, and ± 3000 bp windows, with single base pair resolution (i.e., i corresponds to a single base pair in the window). Enrichment estimates were also made with templates consisting of the TSS and TES windows combined (calculated as a single template) using the same four window sizes. In summary, a set of 5', 3', and 5'+3' enrichment estimations were made for each of the window sizes.

In another set of enrichment estimates, genes were scaled to correspond to a fixed number of bins. The scaling procedure described in [Xu et al., 2010] was used, with bin number equal to 33,346—the median gene length in the filtered gene list. The template procedure was applied to the scaled genes plus an intergenic overhang of 0 bp, 500 bp, 1000 bp, 2000 bp, and 3000 bp beyond the TSS and TES. The resolution of the genes is equal to gene length divided by bin number, while the overhang regions

have base-pair resolution.

3.2.5 Non-iterative model-based enrichment estimate

The process of iteratively solving equations 3.4, 3.6, and 3.7 is computationally expensive. A non-iterative enrichment estimation can be made with equation 3.6 using the non-weighted average template shown in equations 3.3 and 3.4. To examine the trade off between computational efficiency and template optimization, we produced one set of enrichment estimates calculated non-iteratively for every set calculated using the iterative method.

3.2.6 Evaluation of template models

Following [Xu et al., 2010], we used the coefficient of variation of the root mean square deviation, $CV(RMSD)$, to evaluate the fit of our templates:

$$CV_{ijk}(RMSD) = X_{jk}^{-1} \sqrt{\frac{\sum_i^n (c_{ijk} - X_{jk}t_{ik})^2}{n}} \quad (3.8)$$

where n is the number of indices in the template. This metric was used to compare the fit of iterative and non-iterative template models.

3.2.7 MARS model construction and evaluation

MARS models were built with each set of enrichment estimations (51 in total) using the *earth* package in R. Following [Xu et al., 2010], each model was allowed terms with up to 3 degrees of interaction. The quality of each model was evaluated using R^2 values and generalized cross validation (GCV) scores. The GCV score evaluates the fit of the model while penalizing model complexity, whereas the R^2 only considers the fit of the model to the data. A description of the MARS algorithm and GCV scores can be found in the ‘Methods’ section of [Xu et al., 2010].

3.3 Results and discussion

3.3.1 Overview of model construction

A total of 51 enrichment level estimates were made for 21 marks for 9882 Ensembl genes, corresponding to 51 different MARS models. Figure 3.1 shows a summary of each enrichment estimation method. The responses of the models are gene expression data in CD4⁺ T cells gathered from the SymAtlas database [Su et al., 2004]. In cases where multiple Affymetrix probe sets interrogated a single gene, additional observations were included in the model corresponding to each independent expression measurement with redundant enrichment data, resulting in 15,148 observations and 21 predictors per model.

3.3.2 Template model error analysis

To assess the fit of our template-based enrichment models to the enrichment data we used the CV(RMSD), as described in section 3.2. The CV(RMSD) was calculated and averaged for all genes above the 95th percentile in enrichment estimations. Table 3.1 shows the CV(RMSD) for whole gene templates plus a 2000 bp intergenic overhang, for both non-iterative and iterative methods. In 13 of the 21 marks the iterative procedure improved the CV(RMSD); however, the iterative enrichment model performs more poorly than the corresponding non-iterative model for 8 marks.

The iterative and non-iterative H4K20me3 template models had the worst CV(RMSD)s (8.11 and 5.87, respectively). Moreover, the iterative template performed much more poorly than the non-iterative. In this case, H4K20me3 is highly enriched in members of the zinc-finger (ZNF) gene family, and at low levels with a different enrichment profile across the genes in the rest of the genome [Barski et al., 2007, Ernst et al., 2010]. Thus, for H4K20me3, there are at least two classes of enrichment profiles across genes.

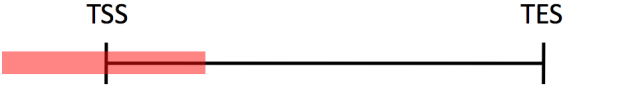
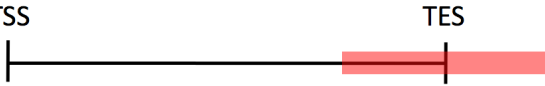
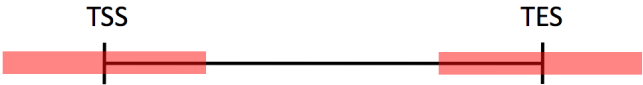

| Estimation Methods | Window Sizes | Enrichment Estimation Window Region Relative to Transcription Boundaries |
|---|---|--|
| Iterative Enrichment Estimate, Non-Iterative Enrichment Estimate, Tag Count | Relative to TSS: $\pm 500\text{bp}$, $\pm 1000\text{bp}$, $\pm 2000\text{bp}$, $\pm 3000\text{bp}$ |  |
| Iterative Enrichment Estimate, Non-Iterative Enrichment Estimate, Tag Count | Relative to TES: $\pm 500\text{bp}$, $\pm 1000\text{bp}$, 2000bp , $\pm 3000\text{bp}$ |  |
| Iterative Enrichment Estimate, Non-Iterative Enrichment Estimate, Weighted Sum of Tag Counts | Relative to TSS and TES: $\pm 500\text{bp}$, $\pm 1000\text{bp}$, $\pm 2000\text{bp}$, $\pm 3000\text{bp}$ |  |
| Iterative Enrichment Estimate, Non-Iterative Enrichment Estimate, Length-Normalized Tag Count | Gene body plus overhang region relative to gene boundaries: $\pm 0\text{bp}$, $\pm 500\text{bp}$, $\pm 1000\text{bp}$, $\pm 2000\text{bp}$, $\pm 3000\text{bp}$ |  |

Figure 3.1: Illustration of enrichment estimation methods. Summary of the methods used to make single-value estimates of gene-wise ChIP-seq enrichment. The first column lists the enrichment estimation methods. The second column lists the window sizes for which each method is applied. The last column shows a graphical representation of the estimation region for each method/window size combination relative to the transcription start sites (TSS) and transcription end sites (TES) of genes.

The iterative template is weighted by enrichment, and hence biased toward the ZNF genes. Thus it yields a poor CV(RMSD) for the majority of genes in the genome that have a different profile and have relatively low levels of H4K20me3 across their bodies. One way of resolving this problem is to apply clustering analysis to the H4K20me3 enrichment profiles across genes and identify the two or three dominant deposition profiles and apply the appropriate template to each subset of genes. Nevertheless, the iterative template method required significantly more computational resources than the non-iterative method, for only marginal improvements in the CV(RMSD) of 13 of the 21 marks, and in MARS model performance (as discussed in section 3.3.3). This suggests that the non-iterative template approach may be preferable to the iterative enrichment estimation method for many applications.

Table 3.1: CV(RMSD) for whole gene templates plus a 2000 bp intergenic overhang

| Mark | CV(RMSD) | |
|----------|---------------|-----------|
| | Non-iterative | iterative |
| H2A.Z | 3.382 | 4.442 |
| H2BK5me1 | 1.940 | 1.957 |
| H3K27me1 | 1.973 | 1.970 |
| H3K27me2 | 2.960 | 2.957 |
| H3K27me3 | 2.599 | 2.601 |
| H3K36me1 | 2.919 | 2.913 |
| H3K36me3 | 1.793 | 1.838 |
| H3K4me1 | 2.007 | 1.943 |
| H3K4me2 | 2.366 | 2.310 |
| H3K4me3 | 2.955 | 3.189 |
| H3K79me1 | 1.812 | 1.806 |
| H3K79me2 | 1.820 | 1.802 |
| H3K79me3 | 1.798 | 1.795 |
| H3K9me1 | 1.861 | 1.810 |
| H3K9me2 | 3.046 | 3.055 |
| H3K9me3 | 4.114 | 4.550 |
| H3R2me1 | 2.524 | 2.521 |
| H3R2me2 | 3.581 | 3.575 |
| H4K20me1 | 1.497 | 1.489 |
| H4K20me3 | 5.871 | 8.112 |
| H4R3me2 | 3.175 | 3.174 |

The CV(RMSD) shows the fit of template models to the enrichment data. The first column shows the mark. The second column shows the CV(RMSD) for the non-iterative template. The third column shows the CV(RMSD) for the iterative template. The CV(RMSD) is improved (lowered) by the iterative template over the non-iterative template in 13 of the 21 marks.

3.3.3 Enrichment estimation and model performance

We found a clear trend in model performance with respect to the enrichment estimation procedure used to calculate the model predictors. GCV scores range from 2.656 to 3.564 and R^2 values range from 0.517 to 0.339 across the 51 models. Figure 3.2 contains a summary of all models and their statistics. As expected, 3' estimates using small estimation windows yielded models with the poorest performance. Except for the whole gene estimates with no intergenic overhang, for equal window sizes,

models based on tag counting estimates were always outperformed by either iterative or non-iterative template-based estimates, as measured by GCV score. With the exception of 2 (whole gene tag counts with 0 and 500 bp intergenic overhangs) out of the 17 tag count-based models, both the iterative and non-iterative template-based models outperformed the tag count-based models for the same window size. Models based on whole-gene estimates outperformed all other models.

The iterative enrichment estimation method was intended to improve the fit of the template to the data; however, this does not mean that the estimated enrichment level produces a final MARS model with a better fit. Indeed, we found this to be true in our models. Of the 17 pairs of iterative and non-iterative template-based enrichment estimations, 10 produced models in which the iterative method was superior, and 7 in which the non-iterative method was superior. However, both methods produced models with similar statistics (Figure 3.1). A possible explanation for this result is that the iterative method yields a template that is an estimation-weighted average of enrichment across genomic coordinates. Thus, genes with large outlier enrichment values for a given mark could be driving the shape of the iterative template. For H4K20me3, which produced the poorest CV(RMSD)s and a largest increase in CV(RMSD) in the iterative estimate relative to the non-iterative estimate, outliers did drive poor performance. As previously discussed ZNF repeats are highly enriched for this mark while most non-ZNF genes have an extremely modest enrichment. The genes that had the largest absolute deviation between iterative and non-iterative enrichment estimates were indeed ZNF genes. This suggests that datasets with extreme outliers may be poorly represented by the iterative enrichment estimate. Incorporating robust estimation procedures (e.g., trim mean) into template and enrichment estimation calculations may improve the results of the iterative enrichment estimation method.

Selection of window size was also a factor in model performance. For the window sizes considered, larger window sizes always yielded improved model fits for any 5'

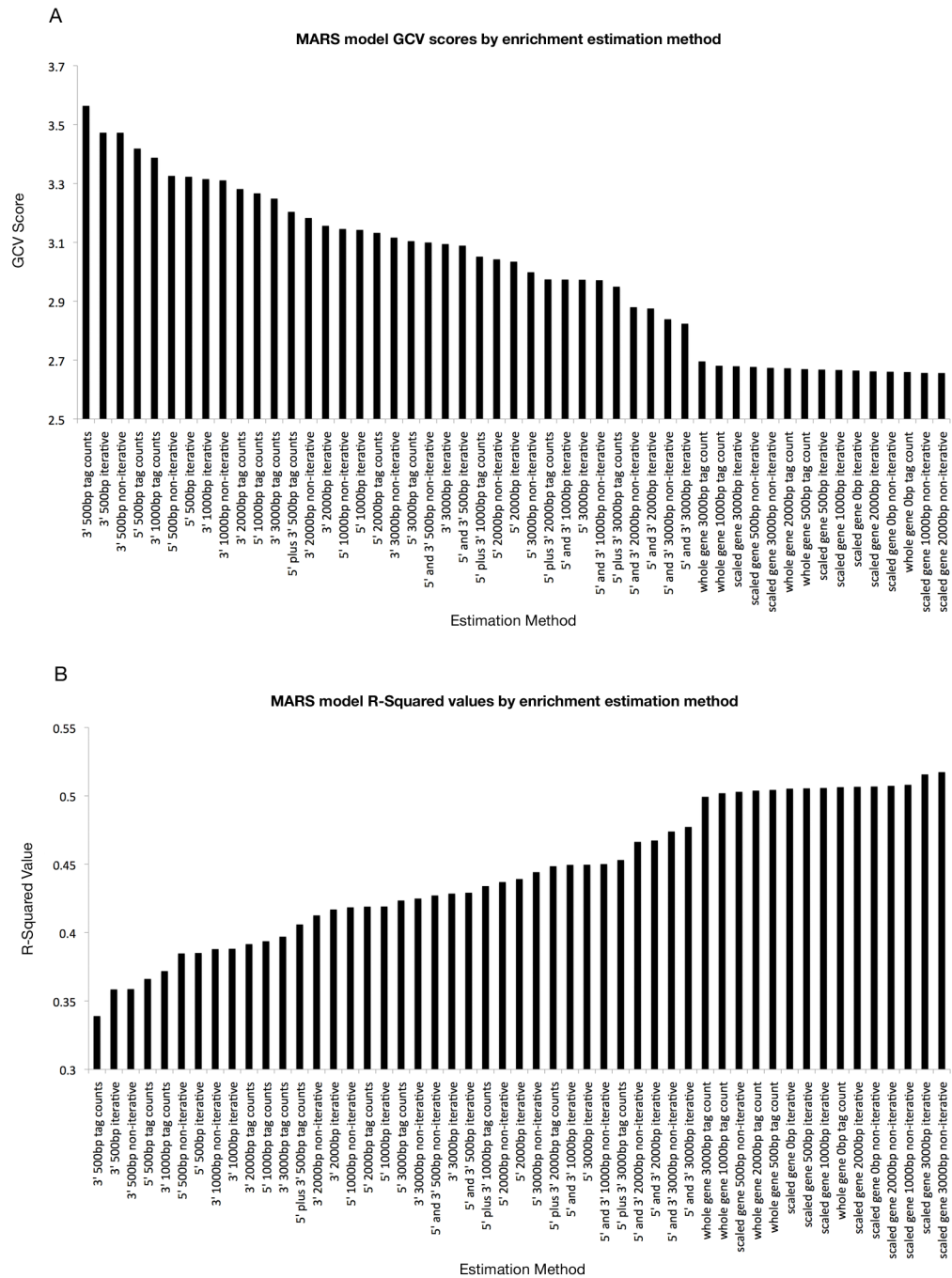


Figure 3.2: **Comparison of enrichment estimation methods by MARS model statistics.** Plots of (A) GCV and (B) R^2 values for MARS models built with each enrichment estimation method. GCV scores are sorted in descending order; small GCV scores are indicative of superior model fit. R^2 values are sorted in ascending order; large R^2 values are indicative of superior model fit. Models based on whole-gene enrichment estimates group together as the best models by both metrics.

and/or 3'-focused method. Not surprisingly, an increase in the amount of data used to calculate the predictors generally improved the performance of the models. However, this does not hold true for the whole-gene tag counting and scaled-gene methods. For these methods, the size of the overhang region is a relatively small fraction of the total genomic coverage used in the per-gene enrichment estimation.

The estimate method with the best performance based on GCV score was the whole-gene, non-iterative, template-based method with a 2000 bp intergenic overhang, which achieved a GCV score of 2.656. The whole-gene method that received the poorest GCV score was the tag count with 3000 bp intergenic overhangs, which had a GCV score of 2.696. The difference in model GCV score between the best and worst whole-gene enrichment estimation methods was only 0.04, corresponding to 1.5% difference in GCV score; the associated p-value < 0.001 . Significance was assessed by randomly permuting mark amplitudes with respect to genes. MARS was then applied to the randomly permuted data and GVC scores were calculated for the best and worst whole-gene enrichment estimation method as well as the percent difference in GCV score. A null distribution and corresponding p-value were calculated by repeating this procedure 1000 times. The worst whole-gene estimation method had a GCV score, which was 0.129 (4.6%) below that of the best method based on specific genic regions (5'+3' iterative template with 3000 bp intergenic overhang). The associated p-value < 0.001 based on the same random permutation procedure described above. This suggests that the most important factor when estimating gene-wise ChIP-seq enrichment is the inclusion of data across the entire length of gene bodies. Additionally, unlike the methods based on localized regions, the whole-gene methods do not show a strong correlation between model performance and window size; further suggesting that the enrichment data in the body of the gene contains the majority of the information content for a given gene.

The Spearman correlation between the iterative and non-iterative template-based

(2000 bp intergenic overhang) enrichment estimates was 0.994 or better across all marks. As expected, the largest deviations between the methods were in estimates of H4K20me3 in the ZNF genes. Correlations of enrichment estimations between whole-gene tag counting and template-based methods (2000 bp intergenic overhang) had a median value of 0.983, and exceeded 0.925 for all marks except for H2A.Z and H3K4me3. The correlations between the tag counting method and the iterative and non-iterative methods for H2A.Z were 0.659 and 0.675 respectively, and 0.775 and 0.771 for H3K4me3. These relatively low correlations can be attributed to the fact that on average these two marks have extremely high enrichment within a few hundred base pairs of the TSS, which rapidly falls to nearly zero beyond 2000 bp into the gene body. No other marks show such a dramatic difference between the gene body and TSS region. For extremely large genes, this means an underestimation of the enrichment using the length-normalized tag count. Indeed, many of the largest deviations between the estimation methods for these marks were for genes that were on the mega-base scale in length (Figure 3.3). Large deviations also occurred when few tags were observed within the estimation window. In these cases, differences between enrichment estimation methods can be attributed to coordinate-dependent differences in weighting. In some cases of 5' proximal marks, genes that were not enriched for the mark were flanked by genes that were (Figure 3.4). The 5' enrichment of the neighbor would sometimes bleed into the 3' region of the non-enriched gene, causing a large enrichment estimate using the tag counting method relative to the template-based methods. Since the template-based methods are a weighting scheme based on the average enrichment pattern, the intruding enrichment is down-weighted. The template-based methods are subsequently able to deconvolve enrichment signals of genes that are close neighbors, and therefore represent an advantage of these methods over tag counting.

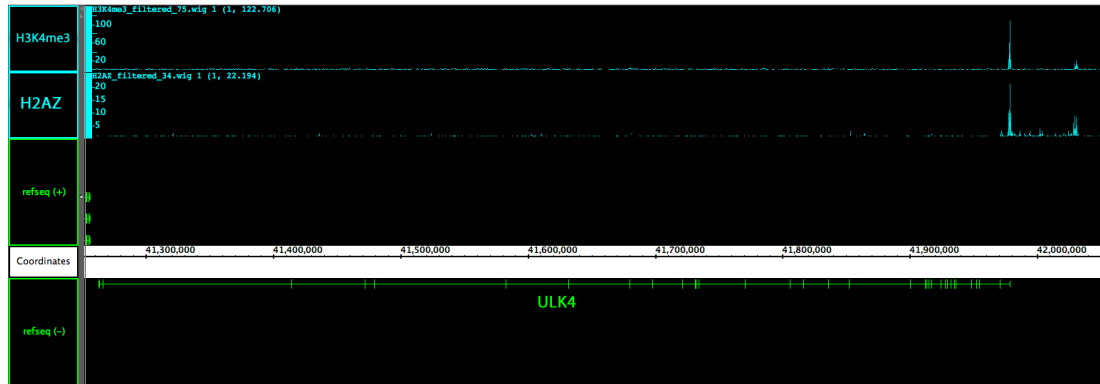


Figure 3.3: **Example of a highly enriched 5' region on a large gene.** Enrichment of H3K4me3 and H2A.Z on ULK4 is highly 5' localized. Since ULK4 is over 700 kb in length, length-normalized enrichment estimates for these marks on this gene would be underestimated relative to most genes.

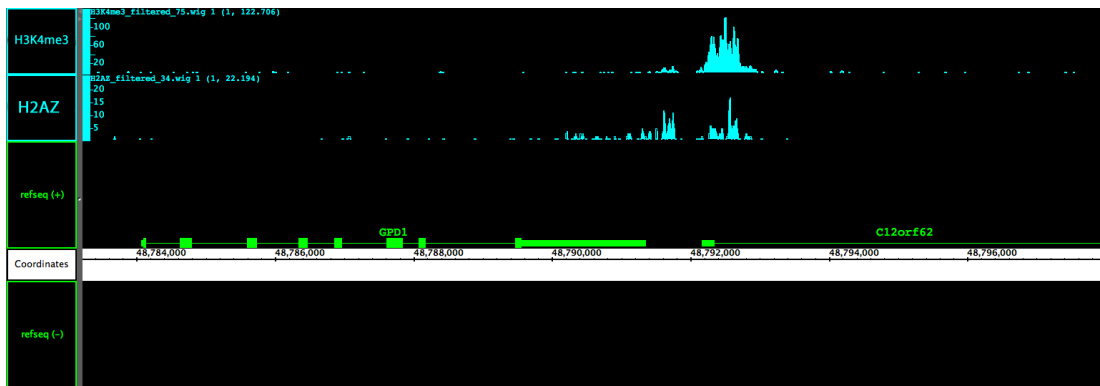


Figure 3.4: **Example of 5' enrichment overlapping the 3' end of a neighboring gene.** Five prime enrichment of H3K4me3 and H2A.Z on C12orf62 bleeds into the estimation window of GPD1, which is not enriched at its 5' end for either mark. A tag counting procedure would yield a large enrichment estimate of GPD1 relative to a template-based enrichment estimate since 3' enrichment is down-weighted for these marks using the template based procedure. Thus, for this and similar cases, the template-based enrichment estimates are better able to deconvolve neighboring ChIP-seq signals.

The accuracy and precision of amplitude estimation for all of the methods considered could be improved by subtracting background read levels and applying appropriate noise filtering. High throughput sequence analysis of input DNA samples revealed that chromatin structure affects shearing and other aspects of ChIP sample preparation, and hence introduces biases in ChIP-seq data [Teytelman et al., 2009]. This together with sequence dependent biases coming from PCR amplification of ChIP samples

argues for methods that assume an inhomogeneous background. One approach would be to use input DNA or other control samples to estimate inhomogeneous background levels; however, an accurate method, which performs this analysis remains to be developed. Indeed a recent comparison of ChIP-chip and ChIP-seq data showed that using Input-seq data as background from an unmatched sample can remove GC-content biases better than use of a matched Input-seq sample [Ho et al., 2011]. Thus, accurate background estimation and subtraction is still an area of active research. One ChIP-seq peak finding method, SICER [Zang et al., 2009], which is designed to identify significantly enriched domains in histone modification data can also be applied as a background noise filter. SICER performs the filtering based on significance. The genome is segmented into windows and those that are not members of significantly enriched islands are filtered out (i.e., set to zero). However, a significance-based filtering approach is not ideal for amplitude estimation and statistical learning applications because accurate estimates of even low, albeit insignificant, enrichments are important. High frequency noise could be removed by applying low pass filters using wavelets. Indeed, wavelet analysis has been applied to genomic tiling array ChIP-chip data for denoising, and could be generalized for ChIP-seq noise filtering [Karpikov et al., 2011].

3.3.4 Enrichment profiles and gene length

The superlative performance of the scaled-gene enrichment estimation methods was unexpected considering many of the histone modifications in this study appear to have TSS-focused enrichment [Barski et al., 2007]. It was initially unclear as to whether scaling genes to calculate the enrichment template was appropriate, considering that these modifications are physically deposited on the tails of histones which make up nucleosomes that occupy approximately 146 bp of DNA. Three of the 21 marks displayed an enrichment pattern that distinctly scaled with gene length: H3K36me3 and H3K79me2/3 (Figure 3.5). Based on the presence of marks that scale with

gene length and those that do not, we hypothesized that a template-based procedure based on absolute position of nucleosomes with the largest window size (i.e., 5'+3' template, 3000 bp window) would yield the best model. Such a model would accurately incorporate data that is based on absolute position of nucleosomes, and also capture the largest genomic region to incorporate the maximum amount of data from marks that scale with gene length. Despite this, and the fact that most of the marks do not appear to scale significantly with gene length, the estimates based on scaled genes produced models with superior performance.

To determine if H3K36me3 and H3K79me2/3—all strongly associated with gene activation—were driving the superior performance of the scaled-gene models, we rebuilt all 51 MARS models without these predictors (data not shown). Surprisingly, the scaled gene method with no intergenic overhang yielded the best model, though the 2nd and 3rd best models were based on whole-gene tag counts. This suggests that although for many marks the scaled template is less representative of the deposition pattern of very large and very small genes, the scaled template strategy offers good performance even on marks whose enrichment profiles do not appear to scale significantly with gene length.

3.3.5 Regulatory information embedded in spatial deposition patterns

Interestingly, the slopes of the enrichment profiles of the three marks that scale appear to be approximately similar across gene lengths from the TSS to approximately 1 kb into the gene body. Beyond approximately 1 kb into the gene body the slopes of the enrichment profiles begin to differ dramatically. For example, for the shortest 20%-ile of genes, average H3K79me3 enrichment rapidly decreases beyond 1 kb into the gene body. For the longest 20%-ile of genes the enrichment profile has a steady, positive slope for the same genomic window, which is about 1 kb to 6 kb into the

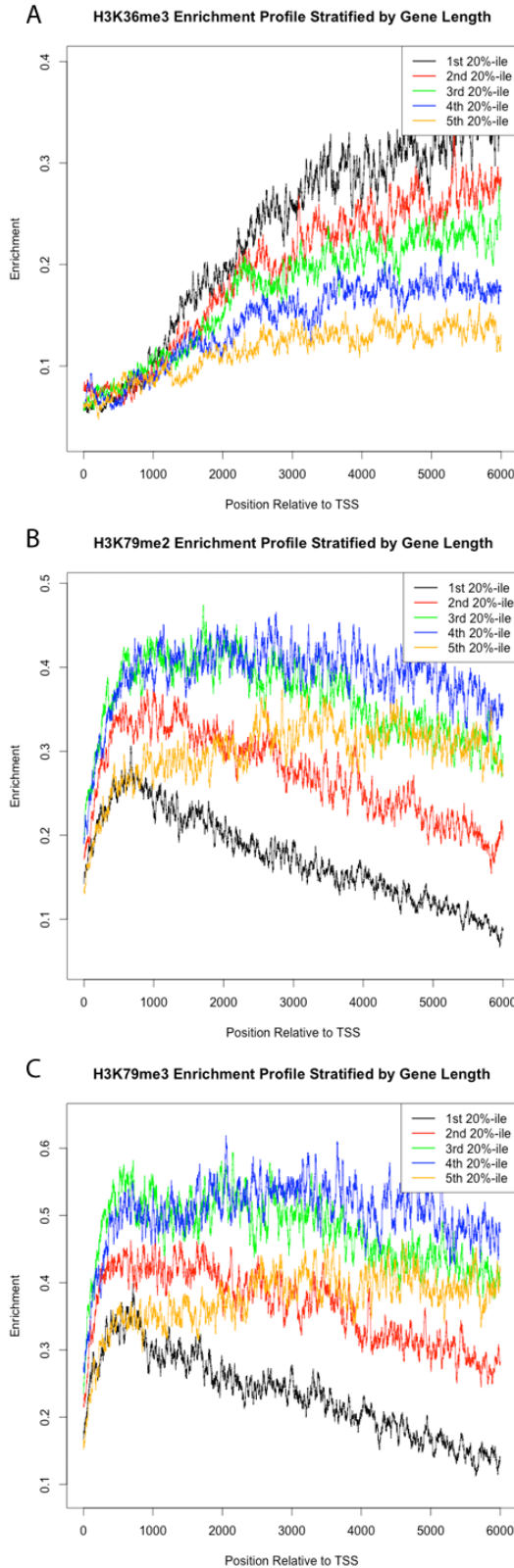


Figure 3.5: Average histone modification enrichments stratified by gene length. Plots of average enrichment profiles from the transcription start site to 6000 bp into the gene body for H3K36me3 (A), H3K79me2 (B), and H3K79me3 (C), stratified by quintiles of gene length. The variability in slope for each of these marks suggests that the enrichment pattern for each of these marks scale with gene length. For example, for the smallest 20%-ile of genes, H3K36me3 enrichment rapidly rises from the TSS to 6000 bp into the gene body; however, for each successive 20%-ile of increasing gene length, the rate of increase in enrichment is diminished for the same region.

gene body. However, from the TSS to approximately 1 kb into the gene body the enrichment profiles of these extreme length groups are nearly identical. This suggests that for these scaling marks, there is a region near the TSS, which is approximately 1 kb in size where these modifications are deposited in a length-independent manner, but beyond which the modifications are deposited in a length-dependent fashion.

Of the non-scaling marks, all but four of the modifications in this study show differences in absolute enrichment levels across gene lengths. Those that do not are H3K27me2, H3R2me2, H4R3me2, and H4K20me3. The greatest differences appear in the longest 20%-ile, which has relatively low enrichment for marks that are explicitly known to be associated with gene activation, and relatively high enrichment for genes associated with gene repression, suggesting a global reduction of transcription in large genes relative to small genes. Indeed, we find that the genes in the largest 20%-ile of gene length show significantly lower gene expression than other genes (Figure 3.6). For methods that did not use the whole gene to arrive at enrichment estimates, we rebuilt models with gene length included as a predictor to determine if the superior performance of the whole-gene estimation methods were driven by the gene length bias. The best performing model of this set was the 5'+3' non-iterative template with a 3000 bp window, which had a GCV score of 2.831. The best model based on estimates in a specific genic region and without gene length as a predictor had a GCV score of 2.824. The lack of improvement after including gene length as a predictor suggests that the performance of the whole-gene enrichment estimation methods was not driven by the gene length bias.

In addition to revealing information about transcriptional regulation, templates and enrichment estimates may also provide information on co-regulation of PTMs. For example, H4K20me1 and H3K9me1 are known to be preferentially deposited on the same nucleosomes *in vivo* [Sims et al., 2006]. The correlation (Spearman) between the templates of these two marks was 0.716, and the correlation between their enrichment

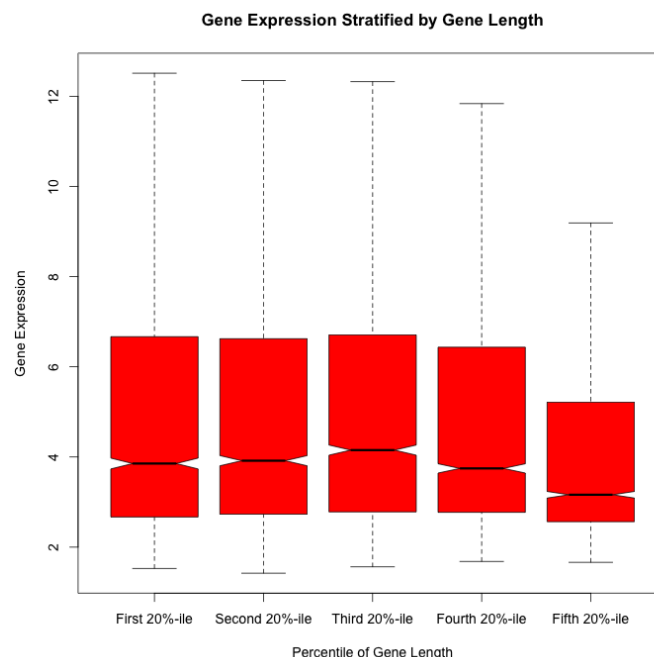


Figure 3.6: **Gene expression stratified by gene length.** Box plots of gene expression stratified by quintiles of gene length. There is a significant decrease in expression in the longest 20%-ile of genes. Along with the observation that longer genes have relatively high enrichment of repressive marks, and low enrichment of activating marks, this suggests that lower gene expression in longer genes is mediated by epigenetic mechanisms.

estimates across genes was 0.876. These strongly positive correlations of templates and enrichment estimates of marks known to co-occur suggest that co-regulatory information can be gleaned from spatial distribution and magnitude of the enrichment data. For example, our data show enrichment correlations of 0.889, 0.762, and 0.761 between H2BK5me1 and H3K79me1/2/3, respectively. Template correlations between H2BK5me1 and H3K79me1/2/3 were 0.956, 0.910, and 0.924, respectively. The high correlation between H2BK5me1 and H3K79 methylation deposition patterns and levels across the genome suggest that there may be a mechanistic link between these histone PTMs that has not yet been reported in biochemical studies (e.g., the enzymes that deposit these marks could be on the same complex). This is one of many cases where both the correlation between two marks' spatial profiles and enrichment levels

across genes is high. Using both enrichment level and spatial deposition patterns across genomes could prove to be powerful at identifying biologically relevant synergies between histone modifications, which make up the histone code.

3.4 Conclusions

Generalizing a method for estimating ChIP-seq enrichment for multiple histone modifications is complicated by the variability in the way different modifications are deposited. This variability ultimately creates different gene-wise ChIP-seq enrichment patterns, some of which scale with gene length and some which do not. Tag counting methods can yield high quality predictors for regression modeling, but ultimately some of the information content coded in the spatial distribution of the data is lost. Although many modifications are highly enriched at the 5' end of genes, much of the useful data associated with a given gene is encoded in the body of the gene. Many previous studies have attempted to estimate enrichment by only focusing on the promoter region, and in doing so, have forgone much of the relevant data.

Using the MARS regression algorithm to build regression models with enrichment levels as predictors and gene expression as responses, we compared various strategies for estimating gene-wise ChIP-seq enrichment for 20 histone methylations and histone variant H2A.Z in human CD4⁺ T cells [Barski et al., 2007]. Enrichment estimation methods were assessed and ranked by the quality of the models produced, which was measured by GCV scores. We have demonstrated that, with respect to the cis-regulatory role that the histone modifications/variant surveyed in this study play in controlling gene expression, the majority of the significant enrichment data lies within gene boundaries. Also, the incorporation of data across whole genes, as well as spatially weighting enrichment for single-value estimations of gene-wise ChIP-seq enrichment can provide significant improvement over strategies that focus on specific

genic regions. Improving methods for the quantification of ChIP-seq data for statistical modeling serves to sharpen the resolution of the models and ultimately improves the conclusions that can be drawn from them.

ChIP-seq technology facilitates the computational interrogation of genomic control networks, and the conclusions drawn by this study can serve to increase depth at which we can probe these networks using this technology. The methods outlined in this work can be applied to almost any machine learning or data mining application that uses gene-wise ChIP-seq enrichment as predictors or responses.

3.5 Chapter acknowledgements

This chapter was adapted from [Hoang et al., 2011].

Chapter 4

Epigenetic reprogramming in the epithelial-mesenchymal transition

4.1 Introduction

Differentiation and lineage commitment occurs through a highly regulated sequence of cellular changes in response to the environment [Arnold and Robertson, 2009]. A conserved de-differentiation process known as the epithelial-mesenchymal transition (EMT), occurs during physiological processes such as development and wound healing [Kalluri and Weinberg, 2009]. EMT progression involves coordinated cellular remodeling, which results in a less differentiated phenotype in order to reorganize tissue structures. Induction of EMT in epithelial cells results in loss of apical-basal polarity and the adoption of a migratory and invasive mesenchymal phenotype [Thiery, 2003]. Recent evidence suggests that inappropriate induction of EMT in tumor cells is associated with the progression of human carcinomas—reviewed in [Yang and Weinberg, 2008, Thiery et al., 2009]. During cancer progression, tumor grade, metastasis, drug resistance, tumor heterogeneity, and cancer stem cell maintenance all correlate

with deregulated EMT [Mani et al., 2008, Thomson et al., 2005, Singh and Settleman, 2010].

An increasing body of evidence indicates that the mesenchymal phenotype is established through genome-wide and locus-specific epigenetic reprogramming [McDonald et al., 2011, Dumont et al., 2008, Lombaerts et al., 2006]. This suggests that epithelial and mesenchymal phenotypes are coordinated through changes to chromatin states, and a possible role for the so-called “histone code” in EMT [Strahl and Allis, 2000, Jenuwein and Allis, 2001, Fischle et al., 2003]. According to one hypothesis, phenotypic switches depend on the chromatin-mediated stabilization of transcription factor (TF) activity [Bird, 2002, Thomson et al., 2011]. Although studies have begun to discover mechanistic roles for changes in specific histone modifications during EMT, the combinatorial nature of the reprogramming remains unclear [McDonald et al., 2011].

A number of studies have attempted to discover functional chromatin domains through a computational process referred to as “chromatin profiling” [Ernst et al., 2010, Ernst et al., 2011]. It has been established that combinatorial patterns of posttranslational histone modifications and covalent changes to genomic DNA delineate functional elements within the genome. These histone codes correlate with gene expression and function, enable the de-novo discovery of genomic features such as transcription start sites and cis-regulatory regions [Ernst et al., 2011, Ong and Corces, 2011], and also aid in specifying cell lineages [Kharchenko et al., 2011]. As a result, association between chromatin profiles and molecular function has been reported on the basis of GO-term enrichments [Ernst et al., 2010, Boyle et al., 2008, Hammoud et al., 2009, Zentner et al., 2011]. Therefore, we sought to discover patterns of histone modifications that contribute to epigenomic reprogramming during EMT, and how changes in these modifications relate to the signaling events that are known to establish the mesenchymal phenotype.

We clustered chromatin profiles, and discovered that genes and pathways involved in EMT show essentially the same changes in all sixteen histone modifications, and two variants that we profiled. We also see coordinated changes at their local enhancers. Strikingly, these genes represent a small minority of the total set of differentially expressed genes. Our results suggest that specific changes in histone modifications coordinate the regulation of genes and pathways involved in EMT. In concordance with previous research that demonstrates the epigenetic regulation of enhancer activity, we reveal distinct changes in chromatin at enhancers during EMT [Mercer et al., 2011, Hawkins et al., 2011, Creyghton et al., 2010]. Furthermore, we show that the directionality of these changes can be distinguished by enrichments for the known binding sites of two different groups of transcriptional regulators. Results from our analyses are all consistent with a model of transcriptional feedback loops mediated by shifts in chromatin states. Our data-driven and integrative computational approach reveals broad epigenetic coordination of transcription factors and signaling cascades with established roles in EMT. We put forward the hypothesis of positive feedback loops involving the NF- κ B and AP-1 TF families, and analogous repression of feedback involving *MYC*.

4.2 General strategy

Given the current research that implicates epigenetic mechanisms in the regulation of EMT, we hypothesized that epigenetic reprogramming broadly coordinates cellular processes that contribute to the phenotypic switch. Furthermore, we hypothesized that this coordination occurs in cancer cells that undergo EMT, despite their mutational landscape and genomic instability. Our goal was to discover a shared epigenetic signature between known EMT drivers and further evidence of epigenetic coordination.

To test our hypothesis, we mapped sixteen histone modifications, two histone

variants, and collected gene expression data in 3D cultures of untreated (epithelial) and cytokine-treated (mesenchymal) A549 cells (Figure 4.1A). Briefly, our model system consists of creating three-dimensional NSCLC A549 cultures by hanging droplet, and subsequently treating the spheroids with tumor necrosis factor (TNF) and transforming growth factor beta (TGF β) to induce EMT (Figure 4.1A). Similar protocols have been utilized to induce EMT in other cell types [Borthwick et al., 2012]. This model has been shown to recapitulate critical characteristics of EMT. Reprogrammed cells have a migratory phenotype, metastatic potential, stem-cell characteristics, and mesenchymal markers. In this system it has been shown that there is increase in the expression of master switch EMT transcription factors, *TWIST1*, *SNAI1*, *SNAI2* and *ZEB2*, and robust upregulation of stem-cell markers, including *KLF4*, *SOX2*, *POU5F1*, *MYCN*, and *KIT*. Furthermore, upon induction of EMT there is loss of *CDH1*, gain of *VIM*, greatly increased invasiveness, and increased ability to form lung metastases in nude mice. Importantly, in this particular system it has been demonstrated that functional characteristics of EMT are dependent on the activity of *RELA* (p65) [Kumar et al., 2013].

The set of histone marks that were mapped includes those that preferentially associate with transcription start sites, gene bodies, enhancers, or heterochromatin, as well as poorly characterized marks (Figure 4.1B) [Barski et al., 2007, Creighton et al., 2010, Heintzman et al., 2009]. We and others have shown that many of the mapped marks correlate with transcriptional activity [Xu et al., 2010]. Here we find a subset of marks correlated at enhancer loci (Figures 4.1B and 4.2). This data was used to quantify the differences in enrichment of each histone modification at gene and enhancer loci. To classify genes (and separately, enhancers) based on their differential epigenetic profiles (DEPs), we employed an unsupervised clustering technique [Newman and Cooper, 2010]. This effectively groups genes (or enhancers) that share highly similar DEPs across the eighteen chromatin marks analyzed. We then

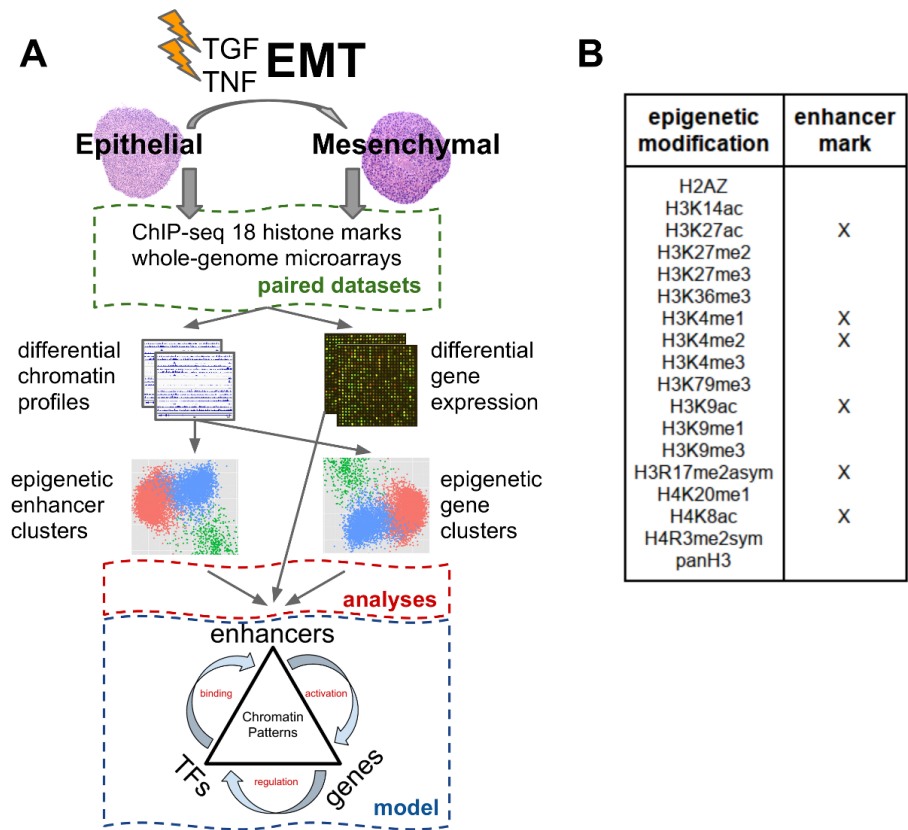


Figure 4.1: Experimental design and data. (A) Flow-chart of the experimental setup and analysis methodology. The epithelial-mesenchymal transition (EMT) was induced using TNF and TGF β in spheroid cultures. Cells before and after treatment (4 days) were collected and whole-genome gene expression and chromatin profiles of 18 histone modifications and variants were obtained. From the paired datasets we measured differential gene expression and calculated differential epigenetic profiles (DEP). The DEPs were calculated individually for gene and enhancer loci and subsequently clustered. Analyses of the resulting epigenetic gene and enhancer clusters included functional enrichment profiling, and transcription factor (TF) binding. The results were shown to be consistent with a chromatin-mediated feedback model that involves specific TFs binding activated enhancers that upregulate expression in EMT-related gene clusters. (B) Table of histone modifications assayed. Histone modifications shown to be correlated and enriched at enhancer loci are indicated.

used these gene and enhancer clusters as the foundation of our functional downstream analyses that integrate multiple sources of functional annotations and molecular data (Figure 4.1A). Specifically, unsupervised clustering enabled us to identify patterns of chromatin remodeling which we link to signaling pathways and transcription factor activity associated with EMT through comprehensive systems-level analyses.

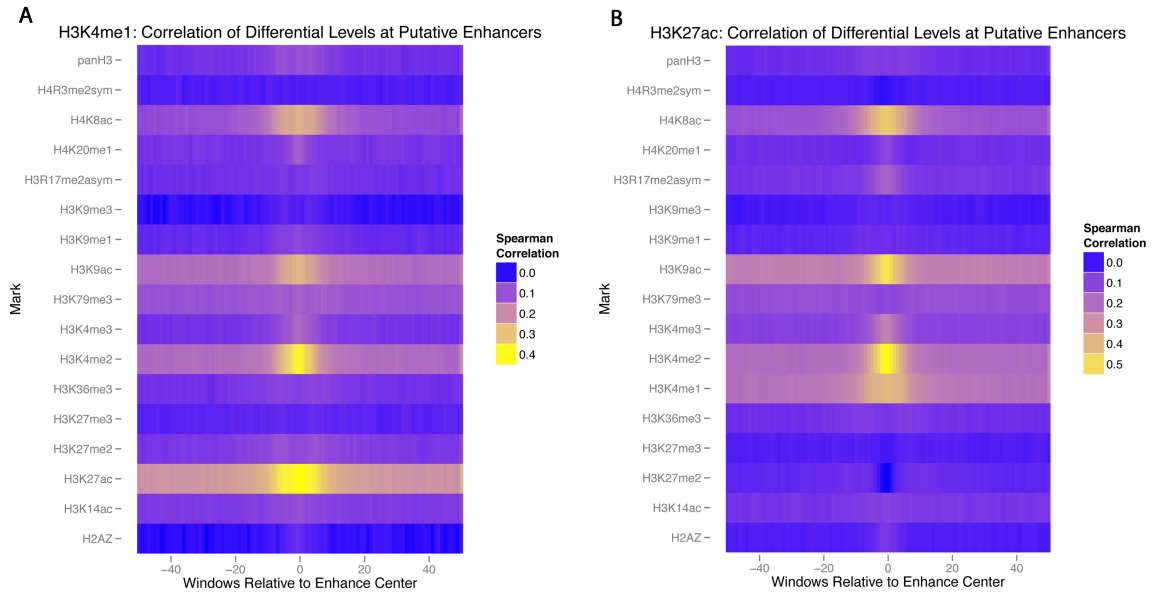


Figure 4.2: **Correlation of histone modifications at enhancers.** (A) Correlation of histone modifications with H3K4me1 at putative enhancer loci. (B) Correlation of histone modifications with H3K27ac at putative enhancer loci.

4.3 Results and discussion

4.3.1 Chromatin profiling reveals EMT-related gene clusters

Genome-wide application of our clustering methodology with the combined ChIP-seq data yielded twenty-nine non-overlapping gene clusters (GCs). Briefly, our method clusters genes based on the epigenetic profile of gains (positive difference of normalized levels of ChIP-seq enrichment between the mesenchymal and epithelial states) and losses (negative difference) of histone modifications at gene loci during EMT. Each gene locus was partitioned into four segments: promoter, transcription start site (TSS), early gene, and gene body (Figure 4.3). It should be noted that genes within a given cluster display highly similar profiles of positive and negative differences across the sixteen histone modifications and two variants (Figure 4.4A). This profile similarity likely occurs because the genes within a cluster undergo similar epigenetic regulation

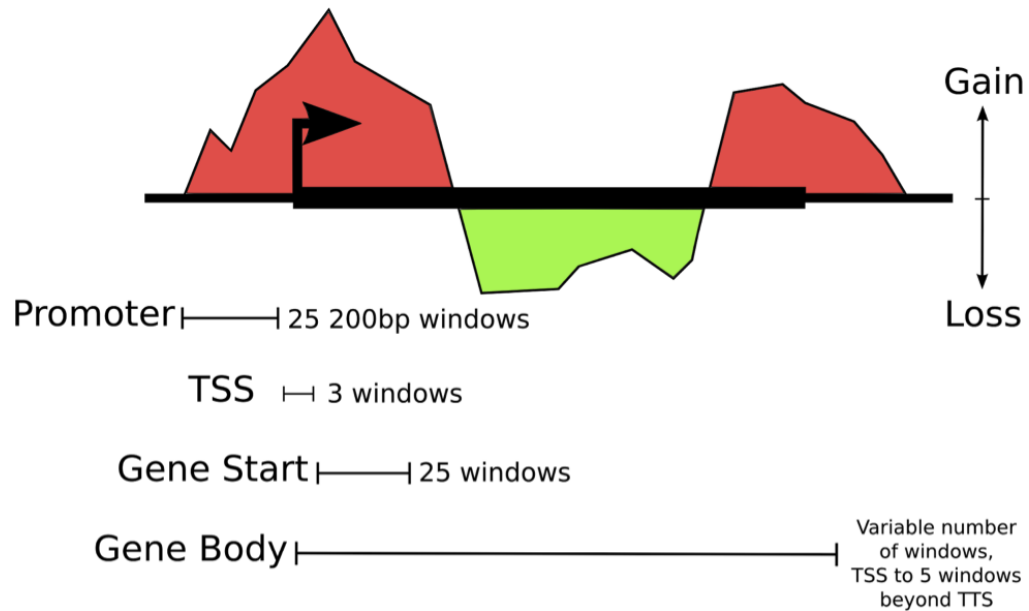


Figure 4.3: **Gene segmentation and differential signal quantification.** Gene segmentation and differential signal quantification. Gene loci were segmented into four regions: promoter, TSS, gene start, and gene body. Within each segment two values were computed for each mark: the sum of the differential gain in the mark, and the sum of the differential loss in the mark (absolute value, mesenchymal minus epithelial). These values together form the differential epigenetic profile (DEP) for each gene. Enhancers were treated similarly; however, enhancer loci were not segmented.

and recognizably distinct regulation of genes from different clusters.

To identify clusters that are associated with known EMT biology, we looked for enrichments in a subset of GO-derived molecular functions that are enriched among genes known to be involved in EMT. Two clusters, GC16 (378 genes) and GC19 (305 genes) (Figure 4.4A), are enriched for many of the same GO-terms as a literature-based reference list of EMT-associated genes, and a similar list of genes annotated with GO-terms explicitly referencing EMT. We quantify this degree of overlap and refer to it as functional similarity (Figure 4.5A). Genes within these clusters have increased expression (Figure 4.4B), and possess similar patterns of chromatin remodeling (Figure 4.4A). We have listed the most significant EMT GO terms for GC16 in Table 4.1 (e.g. “cell adhesion,” False Discovery Rate (FDR) corrected p-value < 1e-5). A third cluster, GC15 (385 genes), had a more modest functional similarity to the reference list of

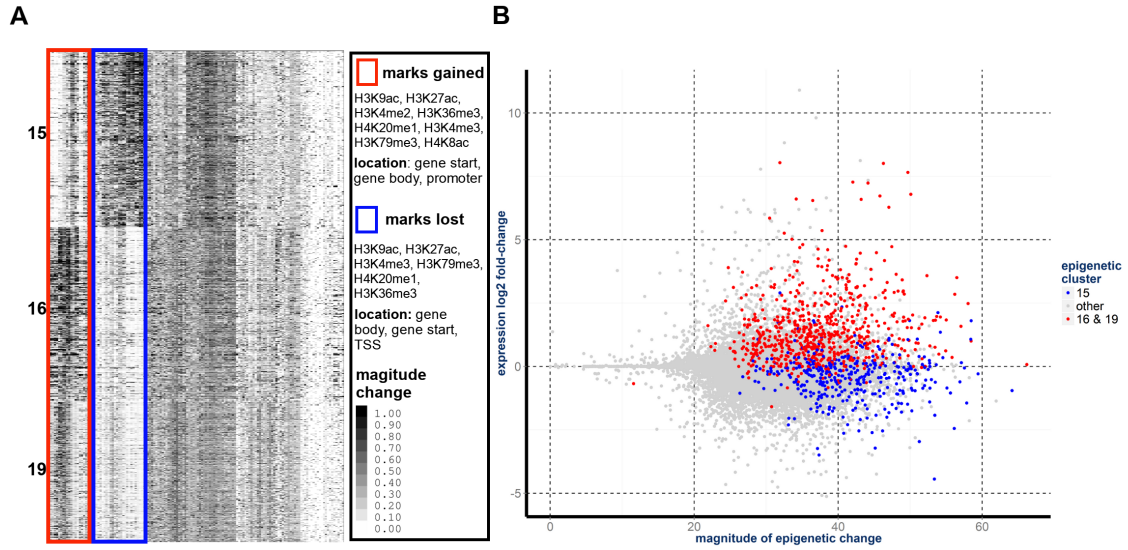


Figure 4.4: EMT-related gene clusters (EMT-GCs) are differentially expressed and show antipodal patterns of chromatin remodeling. (A) Differential epigenetic profiles (DEPs) of the EMT-GCs. Heat map shows the DEPs of genes (rows) from the EMT-GCs (other clusters are omitted). Groups of DEP columns that distinguish clusters 16 and 19 from 15 are indicated through colored boxes. Summary of the antipodal patterns of change in histone modification levels are provided in the table. The red box shows changes specific to clusters 16 and 19. The blue box shows changes specific to cluster 15. (B) EMT-GCs in the differential expression-epigenetic plane. Each dot represents a gene, colored dots are genes from the EMT-GCs: 16 and 19 (red), and 15 (blue). Differential gene expression (log₂ fold-change) is on the Y-axis. The total magnitude of epigenetic difference (sum of DEP elements) at a gene locus is on the X-axis.

EMT-associated genes, but had high functional similarity to GC16 and GC19 (Figure 4.5B). However in contrast, GC15 shows a global decrease in expression (Figure 4.4B). The similarity of GC15, GC16, and GC19 in terms of significant GO-terms suggests that genes from these three clusters are engaged in a focused and coordinated process that drives EMT. We refer to these three gene clusters as EMT-related gene clusters (EMT-GCs) and focus our attention on their characteristics and functional similarities (Figures 4.4 and 4.5). In subsequent analyses, we provide evidence that EMT is driven by genes in these clusters. Remarkably, the EMT-GCs represent only 5.2% of all 20,707 analyzed genes, compared to 18.5% that are differentially expressed at 5% FDR. Compared to differentially expressed genes, EMT-GCs show more significant

Table 4.1: GO-terms most significantly enriched in GC16

| GO-term | description | FDR |
|------------|--|-------------|
| go:0032502 | developmental process | 0 |
| go:0008219 | cell death | 0 |
| go:0006950 | response to stress | 0 |
| go:0008283 | cell proliferation | 1e-15 |
| go:0050896 | response to stimulus | 2e-15 |
| go:0009987 | cellular process | 3e-15 |
| go:0007155 | cell adhesion | 1.9e-14 |
| go:0030154 | cell differentiation | 4.7e-14 |
| go:0005515 | protein binding | 3.29e-13 |
| go:0048856 | anatomical structure dev. | 3.768e-12 |
| go:0007165 | signal transduction | 4.093e-12 |
| go:0048646 | anat. structure form. involved in morph. | 8.42e-12 |
| go:0006928 | cellular component movement | 7.1532e-11 |
| go:0007154 | cell communication | 2.02326e-10 |
| go:0048870 | cell motility | 2.09765e-10 |
| go:0005615 | extracellular space | 6.43198e-10 |
| go:0002376 | immune system process | 3.81583e-09 |
| go:0040011 | locomotion | 3.0465e-08 |
| go:0043066 | negative regulation of apopt. | 2.30123e-07 |
| go:0007568 | aging | 8.28426e-07 |
| go:0001525 | angiogenesis | 1.21924e-06 |
| go:0001816 | cytokine production | 3.45701e-06 |
| go:0008285 | negative regulation of cell prolif. | 5.14751e-06 |
| go:0005576 | extracellular region | 5.61403e-06 |
| go:0008150 | biological process | 7.64415e-06 |
| go:0042060 | wound healing | 1.05706e-05 |

A list of the most significantly enriched GO-terms in EMT-cluster 16—which has the highest functional similarity score to curated lists of EMT-associated genes. The enrichment p-values were calculated using Fishers exact test and FDR corrected.

and specific functional enrichments (compare Tables 4.1 and 4.2). Thus, analysis of chromatin profiles enabled us to narrow down the search for genes coordinated during reprogramming and enrich for EMT-regulators over differentially expressed passenger genes.

We find, in general terms, that the EMT-GCs are distinguished by relatively large gains (GC16, GC19) and losses (GC15) of activating histone modifications (Figure

Table 4.2: GO-terms significantly enriched in > 4-fold upregulated genes

| GO-term ID | Description | FDR |
|------------|---|-------------|
| GO:0010466 | negative regulation of peptidase activity | 0.000237071 |
| GO:0030414 | peptidase inhibitor activity | 0.000681243 |
| GO:0032355 | response to estradiol stimulus | 0.002522024 |
| GO:0045669 | positive regulation of osteoblast differentiation | 0.005647154 |
| GO:0042060 | wound healing | 0.0139314 |
| GO:0001525 | angiogenesis | 0.013598033 |
| GO:0014070 | response to organic cyclic compound | 0.013809122 |
| GO:0001666 | response to hypoxia | 0.020969663 |
| GO:0007568 | aging | 0.036470674 |

A list of significantly enriched GO-terms in the set of genes that show greater than 4-fold upregulation. Enrichments are not as specific or as strong as those associated with the EMT-GCs.

4.4A). We inspected the patterns of epigenetic remodeling to discover which of the assayed marks most uniquely identify the EMT clusters. We find that in GC15, the histone modifications H4K20me1, H3K79me3, H3K27ac, H3K4me3, and H3K9ac are lost throughout gene bodies. Overall the epigenetic changes in GC19 are very similar to GC16 with some exceptions. GC16 and GC19 show relatively strong gains of H3K4me2/3, H3K36me3, H4K20me1, H3K9ac, and H3K27ac across gene bodies. Relative to GC16, gains in GC19 are large for H3K79me3, and moderate for H3K27ac, H3K9ac, and H3K4me2/3 in gene bodies. Consistent with their chromatin changes, GC15 and GC16 display the most antipodal changes in gene expression (Figure 4.4B). By comparison, clusters other than the EMT-GCs exhibit small magnitudes of chromatin and expression changes. These observations are in agreement with many findings concerning the broad role of epigenetics in transcriptional regulation and the transcriptional effects associated with specific marks [Hoang et al., 2011, Wang et al., 2009b, Wang et al., 2009a].

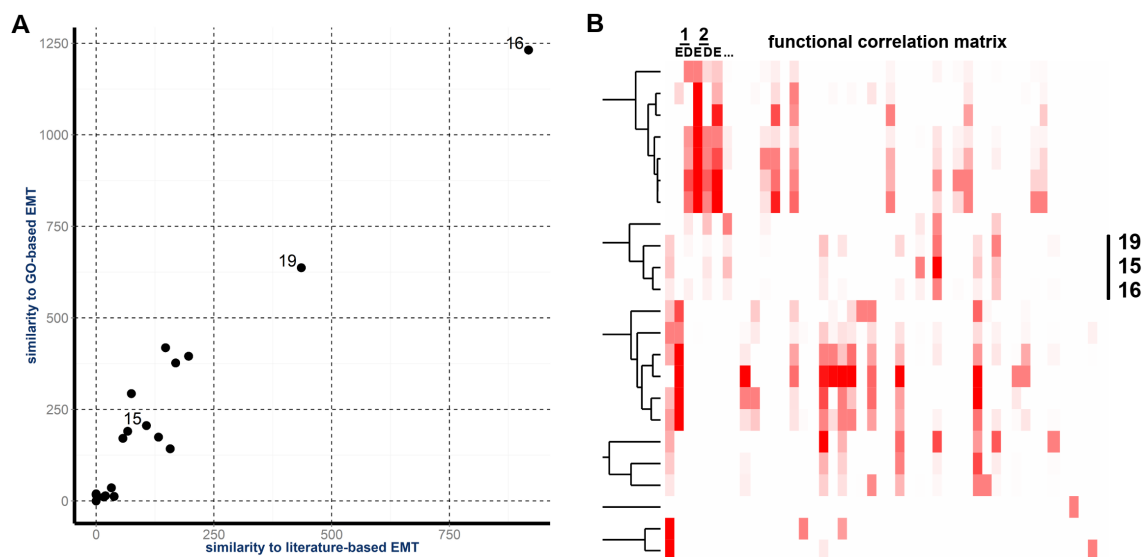


Figure 4.5: Epigenetic clustering groups functionally similar genes and identifies EMT-related clusters. (A) Assessment of EMT functions in gene clusters. Degree of functional similarity between the epigenetic gene clusters and two lists of genes associated with EMT corresponding to genes obtained by manual literature mining and those annotated with GO-terms that included EMT. Functional Similarity Scores (FSS) of each cluster to the two reference EMT gene lists are plotted. (B) Functional similarity of gene clusters. Heat map shows the hierarchical clustering of the Functional Correlation Matrix of epigenetic gene clusters. A trimmed dendrogram of the clustering is shown. Each row represents a source gene cluster while each column represents either the enrichment (E) or depletion (D) score with a target cluster. The sum of the E and D scores is the FSS for a given cluster pair. Columns are arranged numerically by cluster ID.

4.3.2 EMT clusters are enriched for many EMT-associated functions and phenotypes

In order to associate the EMT-GCs with a more comprehensive set of molecular functions and biological processes we profiled them for enrichments for all GO-terms. We removed a large fraction of spurious associations using a 1% FDR cutoff, which revealed that clusters GC16 and GC19 show strong GO enrichment profiles (50 and 23 significant terms, respectively). We found hallmark EMT-regulatory GO-terms, such as cell adhesion and migration, in GC16 and GC19 (Table 4.3). The terms “cell motility,” “basement membrane,” “stress fiber,” and “focal adhesion” are robustly

enriched in GC16 and/or GC19. GO-terms related to the physiological role of EMT such as, “wound healing” and “developmental process” also appeared in these clusters, while GC19 overlaps with the term “cell morphogenesis.” In contrast, GC15 has only five significant terms, four of which are associated with development and growth (Table 4.3). Together, these GO-based analyses reveal a broad similarity of GC15, GC16, and GC19 and association with multiple aspects of EMT, despite differences in the enrichment for specific GO-terms.

Since pathological EMT is linked to metastasis and aggressive tumors, we hypothesized that the genes in the EMT-GCs are associated with advanced cancer phenotypes. To test this hypothesis, we assessed the overlap between these clusters and the sets of genes that distinguish advanced, aggressive cancers from less advanced cancers. These genes sets were obtained from the Molecular Signatures Database 3.0 (MSigDB) [Liberzon et al., 2011]. We observe that genes overexpressed in mesenchymal versus luminal types of breast cancer [Charafe-Jauffret et al., 2006] are over-represented in GC16 and GC19 (fold enrichment over background: 9.4, FDR corrected p-value = $2.3\text{e-}30$) and (9.6-fold, $p = 1.3\text{e-}25$), respectively. Consistently, the downregulated genes from the same study are enriched in GC15 (3.7-fold, $p = 0.0002$). Further analysis revealed that GC16 shows significant enrichment for genes upregulated in the peripheral versus the central part of pancreatic tumors (5.4-fold, $p < 1\text{e-}5$) [Nakamura et al., 2007]. This cluster also contains genes that distinguish metastatic tumors from primary colorectal carcinomas (7.89-fold, $p < 1\text{e-}5$) [Provenzani et al., 2006]. In summary, significant overlaps of EMT-GCs with expression signatures of several advanced cancers suggests that tumors of epithelial origin have a common EMT-associated epigenetic mechanism that contributes to progression and metastasis.

Table 4.3: Referenced GO-terms enriched in the EMT-GCs

| GC | GO-term | Enrichment | p-value |
|----|--|------------|------------|
| 16 | wound healing | 13.568 | 0.00001057 |
| 16 | plasma membrane | 1.982 | 0.0001816 |
| 16 | receptor binding | 4.84 | 0.00024 |
| 16 | seq-spec DNA binding TF activity | 2.58 | 0.006 |
| 16 | signal transduction | 2.523 | < 1e-8 |
| 16 | cellular process | 2.651 | < 1e-8 |
| 16 | cell communication | 2.358 | < 1e-8 |
| 16 | cell motility | 4.231 | < 1e-8 |
| 16 | basement membrane | 8.739 | 0.0095945 |
| 16 | cell differentiation | 3.078 | < 1e-8 |
| 16 | aging | 6.851 | 0.00000083 |
| 16 | growth | 3.286 | 0.00008581 |
| 16 | cell death | 3.859 | < 1e-8 |
| 16 | cell proliferation | 3.901 | < 1e-8 |
| 16 | negative regulation of apoptosis | 6.253 | 0.00000023 |
| 16 | immune system process | 2.988 | < 1e-8 |
| 16 | cytokine production | 4.981 | 0.00000346 |
| 16 | developmental process | 3.105 | < 1e-8 |
| 16 | MAP kinase tyr/ser/thr phosphat activity | 34.02 | 0.026 |
| 16 | inactivation of MAPK activity | 20.46 | 0.024 |
| 16 | pos reg of NF-kappaB TF activity | 9.34 | 0.0015 |
| 19 | plasma membrane | 2.022 | 0.00142517 |
| 19 | signal transduction | 2.79 | < 1e-8 |
| 19 | cellular process | 2.108 | 0.00001248 |
| 19 | cell communication | 2.671 | < 1e-8 |
| 19 | cell motility | 3.425 | 0.0002315 |
| 19 | focal adhesion | 8.441 | 0.0034188 |
| 19 | cell differentiation | 2.532 | 0.00000486 |
| 19 | cell death | 2.519 | 0.00059504 |
| 19 | cell proliferation | 2.765 | 0.00016907 |
| 19 | immune system process | 2.302 | 0.02549018 |
| 15 | cellular process | 2.01 | 0.0000025 |
| 15 | sequence-specific DNA binding | 2.99 | 0.027 |
| 15 | developmental process | 1.93 | 0.00042 |
| 15 | cell differentiation | 1.91 | 0.052 |
| 15 | cell death | 2.22 | 0.0038 |
| 15 | anatomical structure development | 1.98 | 0.00098 |
| 15 | cell proliferation | 1.98 | 0.0069 |

GO-terms significantly enriched in GC15, GC16, and GC19. Only GO-terms directly referenced in the manuscript are shown. GO-term annotations are obtained from GOA and NCBI. Enrichment is the fold relative to the background frequency of a GO-term annotation. P-values are calculated by Fishers Exact Test and are FDR corrected.

4.3.3 Regulation of EMT signaling pathways is chromatin-mediated

Among the GO-terms enriched for GC16 and GC19 are several that correspond to a generic level of many different pathways e.g. “receptor binding,” “signal transduction,” “protein kinase activity,” and transcription factor activity (Tables 4.3 and 4.1). We hypothesized that chromatin remodeling coordinates the activity of a signaling cascade across all levels of a specific pathway. Since, GO-terms only identify functional layers shared by multiple pathways, rather than whole independent pathways, we assessed whether EMT-GCs are enriched for genes from a collection of known pathways. This analysis provides evidence for broad coordination of genes involved in EMT and cancer-related pathways through chromatin remodeling (pathways referenced in this section are listed in Table 4.4). In addition to several novel insights, we recapitulated many of the pathways and processes that represent the canonical EMT phenotype. For example, both upregulated clusters are enriched for “focal adhesion,” “ECM-receptor interaction,” “adherens junctions,” “tight junctions,” and E-Cadherin (CDH1) related pathways. GC19 shows enrichment for additional pathways involved in cell motility such as “regulation of actin cytoskeleton,” and “leukocyte transendothelial migration.”

Since we assessed the histone modification and expression levels from cells that had been exposed to TNF and TGF β over an extended time course, we expected to find delayed early and late response genes within the EMT-GCs. Some well known delayed early and late genes confirmed our hypothesis, including *EGFR* (GC16, log₂ fold-change: 2.45), *SNAI2* (GC16, log₂fc 4.06), *INHBA* (GC16, log₂fc 8.01), *INHBB* (GC15, log₂fc -3.24), *COL1A1* (GC16, log₂fc 4.25), *SKIL* (GC19, log₂fc 3.22), *TGFBR1* (GC19, log₂fc 3.53). Surprisingly, we also observed persistent epigenetic and transcriptional activation of genes associated with the immediate early response to TNF and TGF β exposure. Gene expression profiling indicates that many immediate

Table 4.4: Referenced pathways enriched in the EMT-GCs

| GC | Pathway name | Enrichment | p-value |
|----|---|------------|------------|
| 16 | Pathways in cancer | 4.618 | 0.00000627 |
| 16 | Direct p53 effectors | 8.279 | 0.00000023 |
| 16 | p53 signaling pathway | 7.1 | 0.096 |
| 16 | Focal adhesion | 5.298 | 0.00013609 |
| 16 | ECM-receptor interaction | 6.963 | 0.01242671 |
| 16 | Cytokines and Inflammatory Response | 18.189 | 0.0087528 |
| 16 | Interleukin-1 processing | 54.274 | 0.01740033 |
| 16 | T Cell Receptor Signaling Pathway | 8.32 | 0.00000663 |
| 16 | TNF-alpha/NF-kB Signaling Pathway | 4.28 | 0.03567735 |
| 16 | CD40/CD40L signaling | 13.097 | 0.04278382 |
| 16 | MAPK signaling pathway | 3.493 | 0.09603616 |
| 19 | Pathways in cancer | 5.303 | 0.00000226 |
| 19 | Focal adhesion | 6.245 | 0.00003282 |
| 19 | E-cad sig in the nasc adherens junction | 24.776 | 0.00000267 |
| 19 | Regulation of Actin Cytoskeleton | 6.012 | 0.00571942 |
| 19 | Adherens junction | 13.011 | 0.00000273 |
| 19 | junction | 14.07 | 0.00496435 |
| 19 | Canonical NF-kappaB pathway | 20.435 | 0.00422071 |
| 19 | MAPK signaling pathway | 4.918 | 0.083575 |
| 19 | Leukocyte transendothelial migration | 8.442 | 0.00006173 |
| 19 | T Cell Receptor Signaling Pathway | 8.321 | 0.00000663 |
| 19 | TGF-beta receptor signaling | 15.678 | 0.00001359 |

Pathways significantly enriched in GC16 and GC19. Only pathways directly referenced in the manuscript are shown. Pathways have been sourced from the NCBI Biosystems. Enrichment is the fold relative to the background frequency of a pathway annotation. P-values are calculated by Fishers Exact Test and are FDR corrected.

early genes (IEGs) remained upregulated rather than returning to basal levels. For example *JUN*, *MAF*, *MYCN*, and *KLF7* show strong overexpression and have an active chromatin profile (GC16 and GC19). Other IEGs including *JUNB*, *GADD45B*, *ZFP36*, *ZFP36L1*, *HES1*, *EPHA2*, *IER3*, *SOX9*, and *MAFG* show moderate overexpression, but appear in the epigenetically repressed GC15. In many cases, IEGs are induced by MAP kinase (MAPK) signaling after growth hormone stimulation [Avraham and Yarden, 2011]. These IEGs then induce the transcription of delayed early genes (DEGs). A negative feedback mechanism exists through the repressive activity of DEGs on IEG expression and MAPK signaling.

We observed that the EMT-induced cells upregulated protein phosphatases that attenuate MAPK signaling, including dual-specificity phosphatases (DUSPs). The EMT-GCs contained a significant number of these phosphatases. Specifically, GC16 and GC19 contain *DUSP1/5/6/8/10/16*, while *DUSP4* is a member of GC15. We gained additional support for the activation of MAPK attenuation through GO analysis. We found that GO-terms for “MAP kinase phosphatase activity” and “inactivation of MAPK activity” were enriched in GC16 (Table 4.3). In summary, we observed sustained IEG expression despite an enrichment of DUSP family members in the EMT clusters. The apparent continued transcription of both IEGs and DUSPs, well beyond the early response, suggests loss of negative feedback regulation of MAPK signaling in our system.

We used TNF as a proinflammatory cytokine to enhance TGF β -induced EMT in our model system, and we find that genes that propagate TNF signaling are upregulated and strongly enriched in GC16 and GC19. Specifically, the TNF/NF- κ B signaling pathway is enriched in both upregulated EMT-GCs, while GC16 is enriched for signaling from the TNF receptor, CD40. An enrichment of genes related to the “positive regulation of NF- κ B” in GC16 further supports sustained NF- κ B activity. Interestingly, cluster GC15 also contains several NF- κ B-related proteins. For example, we observed downregulation of the β -arrestin 1 and 2 genes (*ARRB1/2*, $\log_2\text{fc}$ -1.62 and -2.61, respectively). Arrestins show increased expression in differentiated cells and inhibit cellular responses to growth stimuli. Although, their role in EMT remains unclear, overexpression of either *ARRB1* or *ARRB2* in HeLa cells inhibits NF- κ B-mediated transcription. This inhibition occurs primarily through interactions and stabilization of I κ B α (*NFBI*), as well as interactions with the I κ B kinases [Witherow et al., 2004, Kovacs et al., 2009]. Clinical data shows that serum levels of arrestins are lower in patients with NSCLC, and that these decreased levels correlate with poor survival [Wu et al., 2011]. In our system it has been shown that constitutive

activity of NF- κ B is required for induction of EMT and potentiates a mesenchymal phenotype [Kumar et al., 2013]. Taken together these data indicate that constitutive NF- κ B activation during EMT occurs through the epigenetic reprogramming of genes that regulate TNF signaling.

The EMT-GCs also contain many genes that participate in the EGFR signaling pathway, including the receptors themselves. The *EGFR* gene is upregulated and contained in GC16, while *ERBB2* and *ERBB3* (GC15) are significantly downregulated ($\log_2\text{fc}$ -2.30 and -2.04, respectively). Upregulation of the active ErbB2/3 heterodimer occurs in more differentiated cancers, and therefore downregulation of *ERBB2/3* and upregulation of *EGFR* may constitute a receptor switch associated with the core basal phenotype [Foulkes et al., 2010]. Such events may affect ligand specificity and enable cellular reprogramming. Importantly, EMT is associated with resistance to EGFR inhibition [Byers et al., 2013]. This analysis indicates that epigenetic reprogramming contributes to altered EGF signaling in our model system.

Further examination of GC16 and GC19 revealed enrichment for additional pathways broadly associated with cancer and EMT, most of which overlap or crosstalk with TNF, MAPK, or EGFR signaling. For example, GC16 and GC19 are enriched for genes from large cancer-related pathways including: “KEGG: pathways in cancer,” “direct p53 effectors” and the “p53 signaling pathway.” Furthermore, the intersection of these pathways includes many highly upregulated genes from the EMT-GCs such as *SNAI2* ($\log_2\text{fc}$ 4.06), *PRDM1* ($\log_2\text{fc}$ 3.60), *JUN* ($\log_2\text{fc}$ 3.62), and *EGFR* ($\log_2\text{fc}$ 2.45). We also observed an overrepresentation of several immune response pathways in the EMT-GCs. GC16 is enriched for the “cytokines and inflammatory response” and “interleukin-1 processing” pathways, while GC19 is enriched for “T cell receptor signaling.” These findings agree with recent studies that establish a strong association of paracrine cytokine signaling and inflammatory pathways with EMT and metastatic cancer-progression [Kasai et al., 2005, Wu and Zhou, 2010, Bhola et al., 2013].

4.3.4 Epigenetic switches at enhancers correlate with gene expression

We extended our epigenetic analysis to putative enhancer loci, due to the known association between the chromatin state at enhancers and expression of proximal genes [Heintzman et al., 2009, Visel et al., 2009, May et al., 2012, McLean et al., 2010]. This analysis provided insight into the role of specific TFs in the induction of EMT. Moreover, integration of the gene and enhancer clustering showed coordinated changes in chromatin states at genes and enhancers during EMT.

We hypothesized that differential gene expression correlates with epigenetic modulation of proximal enhancers. To test this hypothesis, we identified 75,937 putative enhancers in epithelial and mesenchymal cells based on promoter-distal H3K4me1 and H3K27ac peaks, which mark enhancers in promoter-distal regions [Creyghton et al., 2010]. Next we identified additional “enhancer-associated” marks, which correlate with either H3K4me1 or H3K27ac at these putative enhancer sites (Figure 4.2). The enhancer-associated marks include H3K4me1/2, H3K27ac, H3K9ac, H4K8ac, and H3R17me2asym. Of the 75,937 putative enhancers, 30,681 were found to be differentially marked by the enhancer-associated marks between the epithelial and mesenchymal states. We then grouped these differential enhancers into thirty-eight clusters based on their differential levels of the enhancer-associated marks. Within a given cluster all enhancer marks had the same trend of either gain or loss. Correspondingly, few clusters show simultaneous gain and loss of different marks. Thus, we divided enhancer clusters into two groups: “gain” or “loss.” Within these groups, clusters show distinct magnitudes of change for specific marks (Figure 4.6).

The enhancer-associated marks are generally associated with open chromatin and active enhancers, which suggests that gain and loss clusters correspond to activation and repression, respectively. To test the association of enhancer remodeling to gene

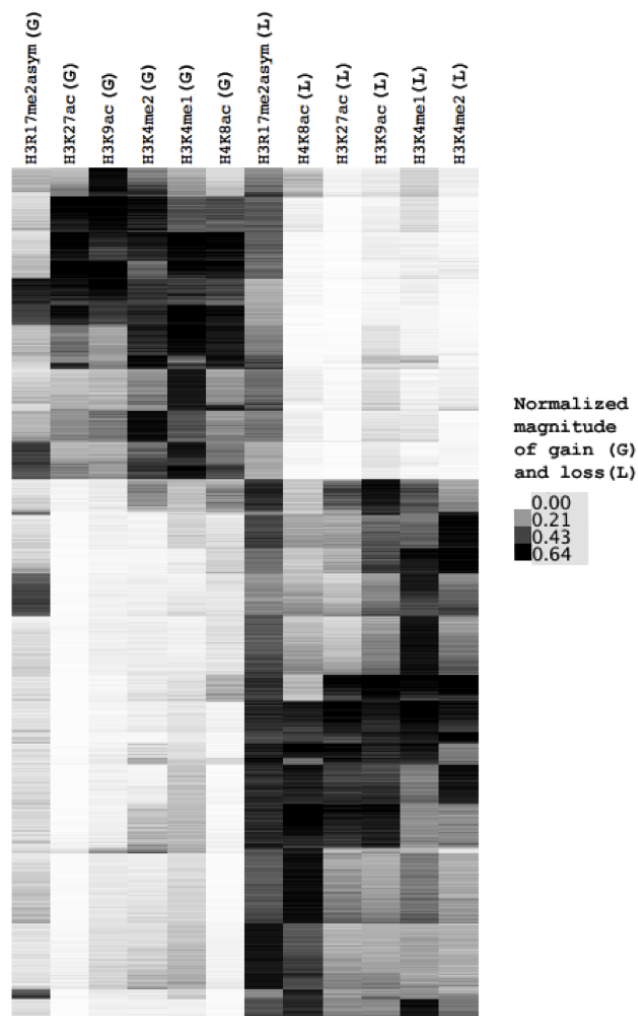


Figure 4.6: **Heat map of differential enhancer clusters.** Heat map showing differential enhancer clusters that are either activated or repressed. These clusters generally show gain (G) or loss (L) across all marks, corresponding to activation or repression, respectively. While H3R17me2asym shows correlation with differential H3K27ac levels at enhancers, it has relatively little coherence across the globally activated and repressed clusters. Additionally, of the marks that correlate with differential H3K27ac or H3K4me1 levels at enhancers, H3R17me2asym shows the weakest correlation (Figure 4.2).

expression, we assigned a “gain-loss” score to each enhancer cluster. We define this score as the mean of the difference between gains and losses across the enhancer-associated marks. These gain-loss scores of enhancer clusters are strongly correlated with the mean differential expression of genes associated with the clusters (PCC=0.89, Figures 4.8A and 4.7). Therefore, our analysis establishes a link between gain clusters

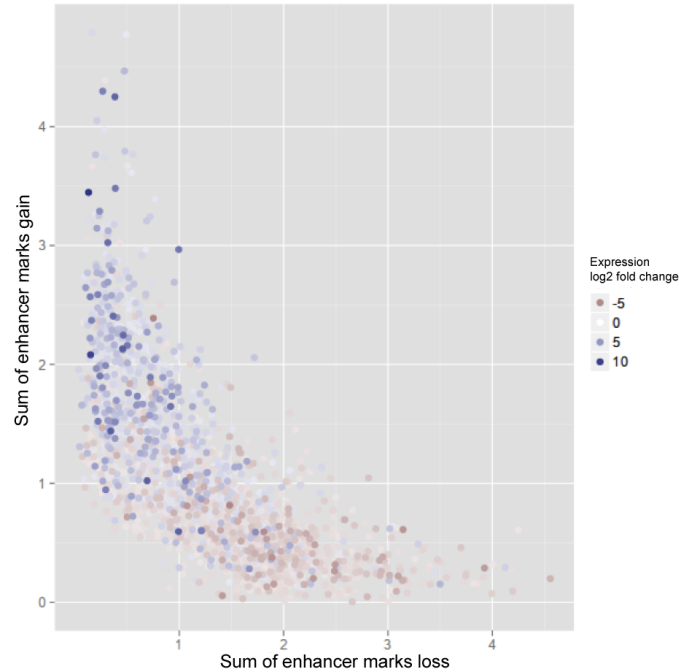


Figure 4.7: **Activation and repression of enhancers correlates with changes in gene expression.** The plot shows the correlation between differential gene expression (\log_2 fold-change, color), and the activation (Y-axis) and repression (X-axis) of proximal enhancers. Each dot represents a gene. Its position in the X-Y plane indicates whether its proximal enhancers activated (dot close to Y) or repressed (dot close to X).

and activated genes, as well as a link between loss clusters and repressed genes.

The EMT clusters also showed strong association with differential enhancers relative to other gene clusters (Figure 4.8B). Examination of these clusters revealed that GC16 and GC19 show striking enrichment for genes associated with activated enhancer clusters. Consistently, GC15 shows strong association with erased enhancer clusters. Interestingly, GC17 also shows overlap with activated enhancer clusters despite lacking noteworthy EMT functional similarity. However, this cluster contains some highly upregulated genes associated with EMT, such as *MMP1*, *MMP9*, and *MMP10*, which are upregulated 453-fold, 278-fold, and 1910-fold, respectively. Together, these observations indicate a widespread co-regulation of enhancers and genes involved in EMT through chromatin remodeling.

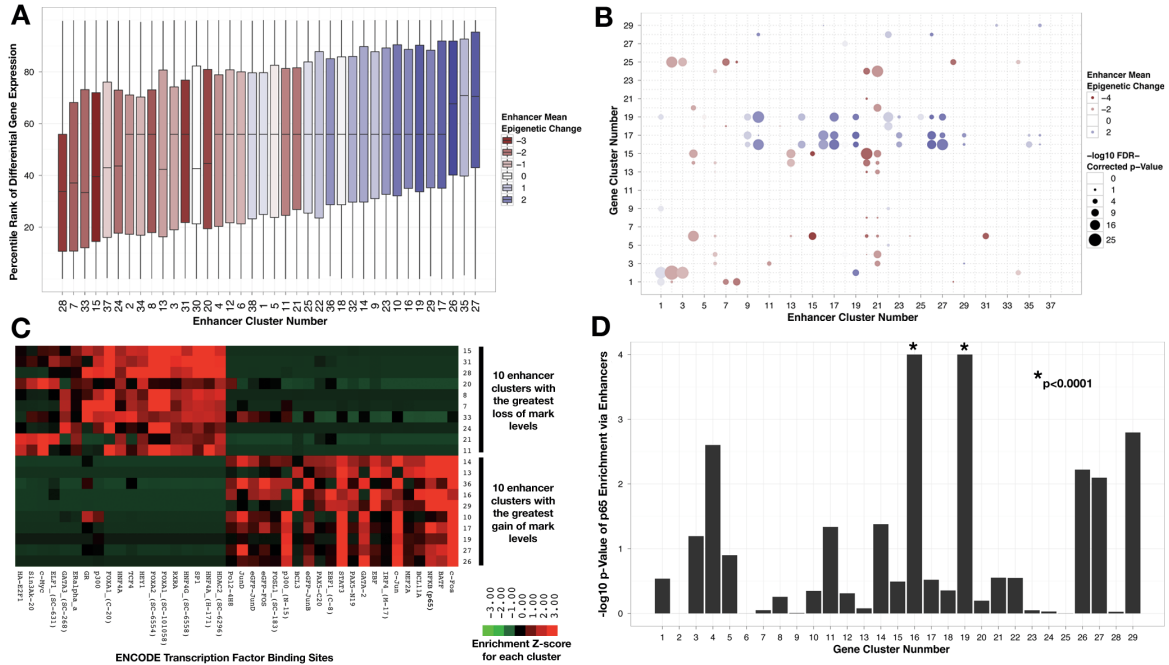


Figure 4.8: Activated and repressed enhancers associated with EMT-GCs and different sets of transcription factors. (A) Box plots of percentile ranks of differential expression for genes associated with each enhancer cluster. Boxes are colored by average magnitude of gain (blue) or loss (red) of enhancer-associated marks. (B) Overlap between gene clusters and genes linked to enhancer clusters. Bubbles are colored with respect to enhancers in the same manner as the boxes in panel A. Size of the bubbles represents the $-\log_{10}$ p-value of the overlap. (C) Association of activated and repressed enhancer clusters with transcription factor binding sites. Significance of overlap between ENCODE transcription factor binding sites (columns) and the 10 enhancer clusters the strongest activated signatures as well as the 10 equivalent repressed enhancer clusters (rows). Each spot on the heatmap is the $-\log_{10}$ p-value of the overlap, which is Z-score normalized by row. (D) Association of p65 binding sites with gene clusters via enhancers. Enrichment of p65 binding sites (ENCODE) in the enhancers assigned to each gene cluster.

4.3.5 Transcriptional control of EMT-GCs through epigenetic reprogramming of enhancers

Because modification of histone tails in enhancer regions influences DNA accessibility, we wanted to determine if the binary regulation (activation or repression) of enhancers corresponds to the binding of specific TFs during EMT. We compared the activated and repressed enhancer clusters for differences in preferential binding of specific TFs.

Transcription factors mapped were clustered by the enrichment of their binding sites in enhancer clusters with the lowest and highest gain-loss scores. As expected, the TFs sharply partition into two non-overlapping sets that correspond to enhancer activation and repression (Figure 4.8C). The presence of this sharp distinction between activated and repressed enhancers indicates that the epigenetic regulation of enhancers is tightly coupled to TF binding.

Several TFs downstream of the pathways enriched in the EMT-GCs (i.e., TGF β , TNF, and EGFR) are enriched in activated and repressed enhancer clusters. For example, p65 (*RELA*), c-Fos (*FOS*), and c-Jun (*JUN*) binding sites show significant enrichment in the activated enhancer clusters. Interestingly, in addition to c-Fos and c-Jun, many AP-1 family members are enriched in the activated enhancer clusters as well, namely fra-1 (*FOSL1*), jun-B (*JUNB*), jun-D (*JUND*), and B-ATF (*BATF*). Together with our pathway analyses, these results demonstrate a chromatin-mediated activation of enhancers that bind NF- κ B and AP-1 family members.

We used ENCODE transcription factor binding site data to determine whether NF- κ B and AP-1 binding sites associated with the EMT-GCs via binding sites at enhancers. We found a strong association of the p65 binding sites with enhancers linked to GC16 ($p < 0.0001$) and GC19 ($p < 0.0001$), but a weak association with GC15-linked enhancers ($p = 0.32$) (Figure 4.8D). Moreover, we observed a similar pattern for AP-1 family member binding sites (Figure 4.9). These results strongly suggest that genes in GC16 and GC19 are regulated through the differential epigenetic activation of enhancers that contain p65 and AP-1 family member binding sites.

In addition to the connection between EMT-GCs and activated enhancers that bind AP-1 or NF- κ B TFs, we observed other evidence that regulation of these transcription factors contribute to EMT (statistical associations shown in Figure 4.10A as black arrows). First, AP-1 and NF- κ B family members show high transcriptional upregulation, and are found in GC16 and GC19. Additionally, genes with predicted

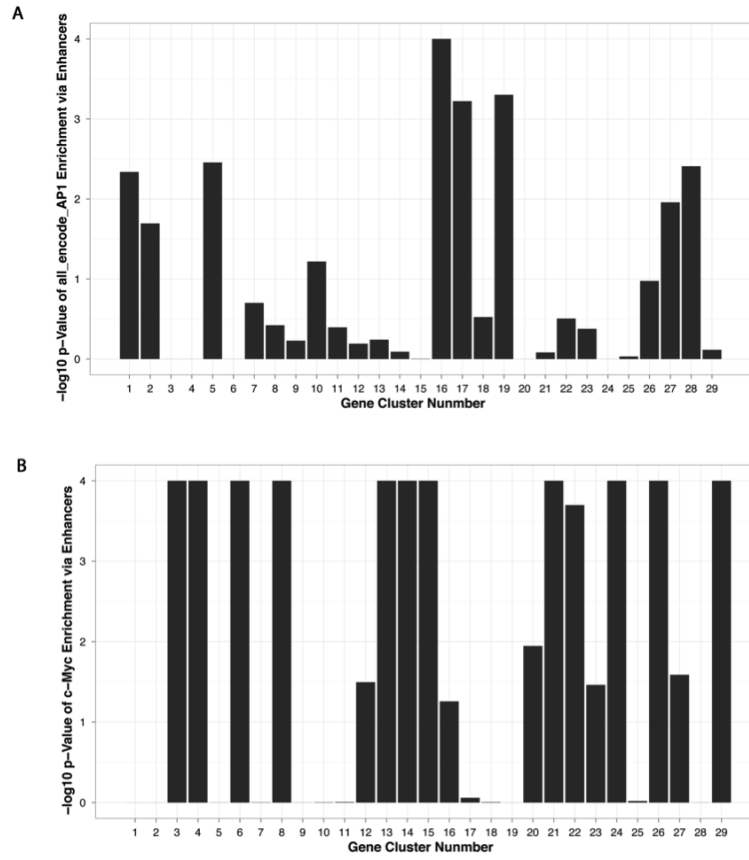


Figure 4.9: **AP-1 and c-Myc binding site enrichment in gene clusters via enhancers.** (Association of (A) AP-1 and (B) c-Myc binding sites with gene clusters via enhancers. Enrichment of each factor's binding sites in the enhancers assigned to each gene cluster.

AP-1 or NF- κ B binding sites in their promoters are enriched in GC16 (5.6-fold, $p = 0.00004$) and (8.9-fold, $p < 1e-5$), respectively. GC19 is also enriched for genes with predicted AP-1 binding sites in their promoters (2.7-fold, $p = 0.009$). Examination of GC16 revealed a strong enrichment of genes induced by NF- κ B signaling in primary human keratinocytes (19.5-fold, $p < 1e-5$) and fibroblasts (13.4-fold, $p < 1e-5$) [Hinata et al., 2003], as well as the core NF- κ B signaling proteins themselves (54.4-fold, $p = 0.003$) [Gilmore, 2006]. Taken together, these results provide evidence that AP-1 and NF- κ B are major regulators of the genes in the upregulated EMT clusters (Figure 4.10A).

Examination of the erased enhancer clusters identified c-Myc as the only enriched

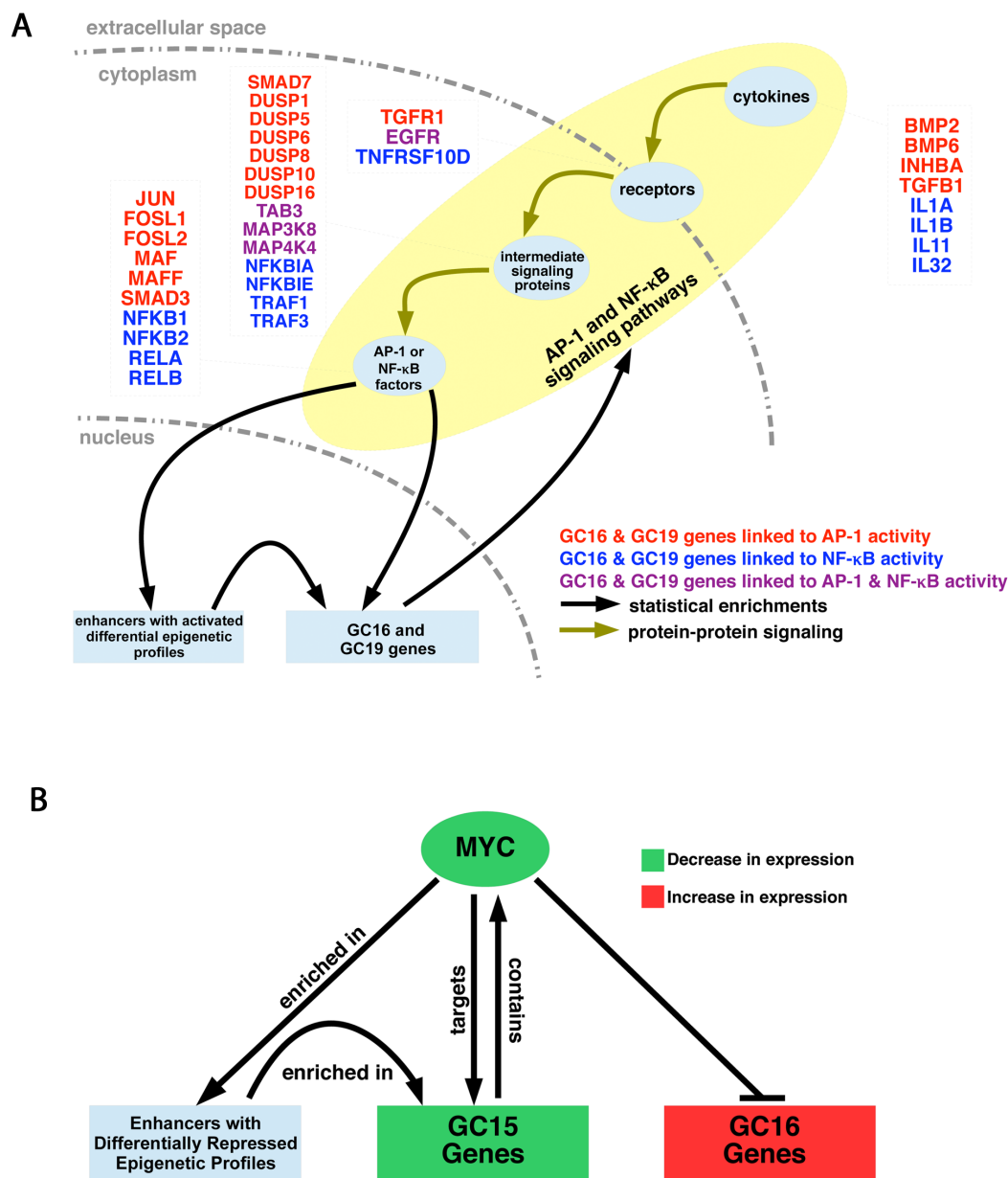


Figure 4.10: **Evidence for broad feedback regulation by AP-1 and NF- κ B family members, and c-Myc.** (A) Statistical enrichments of AP-1 and NF- κ B binding sites link these transcription factors to activated enhancers and the upregulated EMT-GCs. EMT clusters themselves are enriched for in pathways and functions associated with positive regulation of AP-1 and NF- κ B. Some genes in GC16 and GC19 that are known to regulate either AP-1 or NF- κ B are listed. (B) c-Myc binding sites are enriched in repressed enhancers and the repressed EMT gene cluster, GC15. Moreover, GC16 is enriched for genes that are repressed by c-Myc.

TF that is downstream of the pathways enriched in the EMT-GCs. Association of c-Myc binding sites to EMT-GCs via enhancers revealed a significant association with GC15, and a lack of association with GC16 and GC19. It should be noted that this analysis also demonstrates an association between enhancers with c-Myc binding sites and other gene clusters with more modest differential expression (Figure 4.9). This may be explained by the expansive role of c-Myc in gene regulation [Nie et al., 2012]. Comparison to experimental data revealed that GC15 possesses significant enrichment for validated c-Myc targets from two sources: (4.5-fold, $p = 0.002$) [Ben-Porath et al., 2008] and (2.2-fold, $p = 0.04$) [Zeller et al., 2003]. Furthermore, GC16 significantly overlaps the subset of negatively regulated c-Myc targets (5.7-fold, $p = 7.8e-7$) [Zeller et al., 2003], suggesting that c-Myc has opposing transcriptional effects on GC15 and GC16. Finally, from microarray we observed a nearly 2-fold decrease in *MYC* expression after induction of EMT in our system. We validated that *MYC* was in fact downregulated by QT-PCR and observed a significant and almost 4-fold reduction in transcript. These results suggest that decreased c-Myc activity contributes to EMT progression in our model system, through both the de-activation and de-repression of genes in the EMT-GCs (Figure 4.10B).

4.3.6 Links between enhancer clusters and gene clusters suggest a chromatin-mediated transcriptional feedback

Strikingly, AP-1 and NF- κ B transcription factors, and c-Myc themselves reside in the EMT-GCs. Thus, these TFs potentially regulate their own expression and undergo chromatin regulation that is similar to their targets. For example, a large fraction of the AP-1 family of genes reside in the EMT-GCs, including *FOSL1* ($\log_2\text{fc}$ 3.12), *FOSL2* ($\log_2\text{fc}$ 0.88), *JUN* ($\log_2\text{fc}$ 3.62), *MAF* ($\log_2\text{fc}$ 7.27), and *MAFF* ($\log_2\text{fc}$ 1.21),

which are in GC16; while *FOS* (no significant change), *MAFG* ($\log_2\text{fc}$ 1.05), *JUND* (no significant change), and *JUNB* ($\log_2\text{fc}$ 1.80) belong to GC15. Genes that encode TFs that are not AP-1 family members, but which can heterodimerize with AP-1 members also reside in the EMT-GCs, including *CEBPD* (GC15, $\log_2\text{fc}$ -3.49), *CEBPB* (GC15, $\log_2\text{fc}$ 0.89), and *CEBPG* (GC16, $\log_2\text{fc}$ 0.61). Additionally, GC16 contains three NF- κ B family members: *NFKB2* ($\log_2\text{fc}$ 1.76), *RELA* ($\log_2\text{fc}$ 1.23), *RELB* ($\log_2\text{fc}$ 2.27); while *NFKB1* ($\log_2\text{fc}$ 1.89) appears in GC19. As expected, the downregulated *MYC* gene resides in GC15. Based on these coordinated changes in chromatin state for a small set of TFs and their respective pathways, enhancer binding sites, and downstream targets, we put forward a hypothetical model that EMT is maintained by chromatin-mediated transcriptional feedback mechanisms involving the TF families that we have highlighted. This model provides a plausible explanation for sustained activity and critical role of NF- κ B in the experimental system.

4.4 Conclusions

A rapidly growing body of research demonstrates that EMT is an epigenetically regulated process (for recent reviews see [Stadler and Allis, 2012, Wu et al., 2012]). The known mechanisms of regulation involve miRNAs, chromatin structure, DNA methylation, and changes to histone modification levels. EMT in non-transformed cells has been likewise linked to remodeling of specific chromatin domains (i.e., the so-called “LOCKS”) [McDonald et al., 2011]. We therefore hypothesized that genes involved in EMT are broadly coordinated through epigenetic mechanisms. We have made four key observations in support of this: (1) Genes known to be associated with the EMT phenotype are shown to have strong, specific, and highly similar differential chromatin profiles. (2) Epigenetic regulation at gene and enhancer loci linked to EMT is consistent in terms of chromatin activation, repression and differential

gene expression. (3) Two distinct classes of enhancers associated with activated or repressed chromatin, are significantly enriched for binding sites of two different sets of TFs. (4) The upstream pathways and downstream targets of the TFs linked to activated enhancers (AP-1 and NF- κ B family members) are enriched for genes with EMT-specific epigenetic profiles. Therefore, epigenetic regulation of genes that drive EMT is coordinated and specific in our model system. These findings link chromatin remodeling to shifts in cellular signaling networks. They are also consistent with a model of positive feedback that maintains the phenotypic switch (Figure 4.10A). The constitutive activation of NF- κ B in our system, and the extensive reprogramming at NF- κ B target loci, provides further support for this data-driven hypothesis.

Although we have been able to associate combinatorial epigenetic profiles with clear functional roles, our results do not address the specific cooperative mechanism of chromatin remodeling. However, we identified a number of candidate chromatin modifying enzymes that are differentially expressed. Upregulated chromatin modifiers include the histone deacetylase *HDAC9* (\log_2 fc 3.53), methyltransferase *EZH2* (\log_2 fc 1.13), and demethylases *JHDM1D* (\log_2 fc 3.38) and *KDM1B* (\log_2 fc 1.38). Downregulated enzymes include the deacetylase *HDAC1* (\log_2 fc -1.15), methyltransferases *ELP3* (\log_2 fc -0.92), *NCOA2* (\log_2 fc -1.43), and *EHMT2* (\log_2 fc -1.10). In addition, genes and enhancers with EMT-specific chromatin remodeling patterns are enriched for targets of specific chromatin remodeling complexes. For example, ENCODE-mapped Sin3a and HDAC2 binding sites are enriched in repressed enhancers. These factors have been implicated in EMT by a study that has shown that the master switch factors SNAI1 and SNAI2 recruit the Sin3a/HDAC1/HDAC2 complex to silence *CDH1* in EMT [von Burstin et al., 2009]. The enrichment of HDAC2 binding sites at silenced enhancers suggests a broader silencing role for HDAC2 in EMT. These associations point toward select chromatin modifying complexes and enzymes as likely epigenetic drivers of EMT.

We also found that chromatin modulates, and effectively maintains the activation of pathways involved in the response to TNF/TGF β after prolonged stimulation with these cytokines. Surprisingly many canonical immediate early response genes, such as JUN, remained active transcriptionally and epigenetically. Many of the pathways downstream of TNF/TGF β show further evidence of chromatin-mediated transcriptional switching. Within the TGF β signaling pathway we observe a striking bidirectional regulation of TGF β superfamily cytokines, their receptors, and their downstream signaling components. We also see differential regulation of MAPK phosphatases, and a pronounced switch in EGF receptors. Within these examples, genes that are upregulated often have the GC16 or GC19 activated epigenetic signature, while downregulated genes have the opposite GC15 repressed differential profile. These results are consistent with previous findings that EMT involves switches among receptor tyrosine kinases that activate the MAP-ERK pathway [Thomson et al., 2008]. Thus, we conclude that modulation of critical pathways during EMT involves coordinated epigenetic activation and repression.

One of our most unexpected finding is that epigenetically active and repressed enhancer regions are enriched for the binding sites of two non-overlapping sets of specific TFs. This lends support to the model that chromatin and TF profiles jointly govern the locus specific regulation of gene expression. The magnitude of the differential epigenetic regulation that we observe at enhancers is in agreement with several studies that highlight the epigenetic plasticity of enhancers relative to promoters [Hawkins et al., 2011, Heintzman et al., 2009]. Ours results suggest that global availability of TF binding sites at enhancers distinguish epithelial and mesenchymal phenotypes. Consistently, several studies have demonstrated the cell-type specificity of enhancers and TF binding patterns [John et al., 2011, Jin et al., 2011]. There is also evidence that the observed regulation of enhancers is specific to epithelial and mesenchymal phenotypes. For example, we linked FOXA1 and FOXA2 with enhancers that are

repressed in EMT. These so-called “pioneer” factors are believed to facilitate opening of chromatin at enhancers to enable lineage specific transcriptional regulation [Lupien et al., 2008, Sekiya et al., 2009, Li et al., 2012b]. Interestingly, these TFs have been shown to promote the epithelial phenotype and block EMT in various systems [Song et al., 2010, Wan et al., 2005, Burtscher and Lickert, 2009, Mehta et al., 2012].

In summary, we have shown extensive epigenetic reprogramming at both gene and enhancer loci between the end states of the EMT. Changes to chromatin states enable the constitutive activation of transcription factors (some of which are associated with an immediate early response), their upstream signaling pathways, and target enhancers. Based on these results we put forward a hypothesis in which EMT is driven in large part by chromatin-mediated activation of transcriptional positive feedback loops. The linchpins of this feedback are two TF families: AP-1 and NF- κ B. Interestingly, of all gene clusters, GC15 and GC16 show the highest fractional composition of transcription factors, which includes a large number of AP-1 and NF- κ B family members. This suggests that epigenetic reprogramming during EMT alters the transcriptional profile of the cell by broadly altering chromatin accessibility, and by regulating genes which directly mediate transcription—a potential feedback mechanism in itself. Together, our results suggest a high-level mechanism for how complex signaling networks can be coordinated during EMT, and cellular state transitions, generally.

4.5 Methods

4.5.1 Cell culture

NSCLC lines A549 were purchased from ATCC and grown in DMEM (CellGro), 10% FBS (Invitrogen) and penicillin/streptomycin (Invitrogen). Spheroid (3D) cultures were resuspended in DMEM/10%FBS as 25000 cell aggregates using the hanging droplet technique. Newly formed spheroids were transferred onto polyhema plates

containing DMEM/2% FBS to prevent aggregates from attaching to the dish. For EMT-induction, monolayer or spheroid cultures were incubated in DMEM/2% FBS and treated with vehicle or with TNF (10 ng/mL) and TGF β (2 ng/mL) for 48 hours. The 2D and 3D cultures were then treated with vehicle or TNF and TGF β a second time for an additional 48 hours. The samples were subsequently collected and subjected to RNA isolation or ChIP-seq. TGF β (PHG 9204) and TNF (PHG 3015) were purchased from invitrogen or life technologies.

4.5.2 ChIP-seq

Chromatin immunoprecipitation (IP) followed by sequencing (ChIP-seq) assays were performed in spheroid cultures only. TGF β /TNF treated and control cells were cross-linked in 1% formaldehyde. The cross-linking reaction was quenched using 125 mM glycine, and the samples were collected for ChIP-seq analysis according to the Myers lab protocol as described in [Johnson et al., 2007]. Approximately 1.2×10^7 cells were used per IP, and the DNA was sheared to approximately 400 bp fragments by sonication with a bioruptor. After DNA recovery, we used standard Illumina protocols and reagents to prepare the ChIP-seq library (Illumina 11257047 rev A). The antibodies used for IP are listed: H2A.Z (abcam, ab4174), H3K4Me1 (Active Motif, 39635), H3K4Me2 (Active Motif, 39141), H3K4Me3 (Active Motif, 39159), H3K27Ac (Abcam, 4729), H3K27Me2 (Active Motif, 39245), H3K27Me3 (Active Motif, 39155), H3K14Ac (Active Motif, 39599), H3K36Me3 (Abcam, ab9050), H3K79Me3 (Abcam, ab2621), H3K9Ac (Active Motif, 39137), H3K9Me1 (Active Motif, 39249), H3K9Me3 (ab8898), HeR17Me2asym (Abcam, ab8284), H4K8Ac (Millipore, 17-10099), H4R3Me2asym (Abcam, ab5823), H4K20Me1 (Active Motif, 39175), pan-H3 (Active Motif, 39163).

4.5.3 Microarray and Gene Expression Analysis

Microarray analysis of was performed on technical duplicates of TGF β /TNF treated and untreated cells in both two-dimensional and spheroid cultures. Total isolated mRNA was hybridized to Affymetrix U133 plus 2.0 microarrays. The raw data was analyzed using Bioconductor [Gentleman et al., 2004]. Background subtraction was performed using GCRMA. The limma package [Smyth, 2004] was used to perform differential expression analysis, in which a 5% FDR-adjusted p-value cutoff was chosen.

Normalized expression values for all probes were propagated onto genes considered in this analysis. We used a comprehensive, but non-redundant, set of high-confidence protein-coding transcripts. We eliminated the majority of redundant transcripts coding for isoforms of a single gene, together with pseudo- and RNA-coding genes, a final list of 20,707 canonical transcripts represented by UCSC IDs and gene symbols (HGNC) [Kuhn et al., 2013]. Further, each gene was annotated with expression values from all probes that map to any of the genes' transcripts and isoforms as defined by all the transcripts known to UCSC (July 2011). In analyses of differential gene expression the probeset with the largest \log_2 fold-change ($\log_2\text{fc}$) magnitude between treated and untreated samples has been chosen to represent a set of transcripts.

4.5.4 ChIP-seq data processing

Images generated by the Illumina sequencer were initially processed using the Illumina pipeline. Sequences were mapped to the human reference genome, hg19 (GRCh37), using the BWA software with all default options [Li and Durbin, 2010]. In cases where a tag aligned to multiple sites the match with the smallest edit distance was chosen. In the event of an exact tie a single mapping site was randomly chosen. Sequences that fully or partially overlapped problematic regions were discarded. We defined problematic regions as those with known mappability issues, e.g. repetitive

sequences (from the UCSC genome browser microsatellite track, downloaded July 8, 2011) and genomic coordinates with high false positive rates of enrichments, as identified by [Pickrell et al., 2011]. All remaining mapped tags were extended to 200 bp in the 3' direction to account of the expected length of nucleosome-bound DNA.

4.5.5 Scaled Differential Enrichments

To generate chromatin enrichments the genome was segmented into 200 bp bins. The extended tags were assigned to each genomic bin they overlapped. The raw enrichment (RE) is simply the per-window overlap count. REs have been calculated for each of the mapped histone marks from both epithelial (3D untreated) and mesenchymal (3D treated) samples. To allow for comparisons of enrichment profiles between the epithelial (E) and mesenchymal (M) samples we normalized pairs of REs for each histone modification or variant. We used an in-house implementation of the normalization procedure used in the DESeq algorithm to calculate scale factors for each pair [Anders and Huber, 2010]. Scaled enrichments (SE) were obtained by multiplying REs window-wise by the appropriate scale factors. Finally, we calculated scaled differential enrichments (SDE) by subtracting (for all histone modifications separately) the epithelial SE (ESE) from the mesenchymal MSE at each genomic window; i.e., $SDE = ESE - MSE$.

4.5.6 Definition of Putative Enhancer Loci

We have adapted the methodology of [Creyghton et al., 2010] to locate putative enhancer sites using histone modifications. A set of initial putative loci was derived from the raw enrichments of two “core enhancer” marks H3K27ac and H3K4me1 that have been previously shown to be sufficient to distinguish enhancers from other genomic elements. The SICER software [Zang et al., 2009] was used to call peaks of both marks in the epithelial and mesenchymal states, using corresponding panH3 samples as a

control. Peak calls with gaps less than or equal to 600 bp were merged. The final calls were based on a FDR-corrected p-value < 0.01 . These peaks were subsequently used to delineate enhancer regions. Potential enhancer sites were anchored on the window within a given peak call that had the maximum nominal enrichment of one of the two marks, corresponding to the mark for which the peak was called. Since enhancers discovered by profiling p300 occupancy have been shown to be depleted of H3K4me3, these anchor sites were filtered to exclude those that overlapped H3K4me3 SICER peaks (called in the same manner as H3K4me1 and H3K27ac). Finally, anchor sites based on H3K4me1 peaks that were within 1 kb of sites based on H3K27ac peaks were collapsed to the H3K27ac-based site. The resulting set of 200 bp putative enhancer sites were expanded to include the flanking 1 kb, to produce a set of 75,937 putative enhancer sites, each 2200 bp in length.

4.5.7 Enhancer-associated histone modifications

Within our panel of epigenetic modifications we identified a subset of marks that are associated with enhancer activity. Marks that showed clear position-dependent correlation with either H3K4me1 or H3K27ac differential enrichment at putative enhancer loci include: H3K4me2, H3K9ac, H3R17me2asym, H4K8ac (Figure 4.2). Together with the initial two, these marks comprised our set of six enhancer-associated marks.

4.5.8 Gene assignment and filtering of enhancer loci

The initial set of 75,937 putative enhancers was further filtered to enrich for regions with significant epigenetic changes during EMT. We retained enhancers with a significant change for at least one “enhancer-associated” histone modifications. The significance calls were based on a extreme-value null-model derived from the set of all enhancers. For each enhancer a single extreme-value is retained that corresponds to the largest

magnitude of change in either the positive (“gain”) or negative (“loss”) direction. The distribution of maximal magnitudes was represented through a kernel density estimate (Gaussian kernel, bandwidth 0.025). The left tail of this distribution was used to calculate a Gaussian null model of the noise regime of the differential signals. This Gaussian null model has parameters $\mu = \hat{\mu}$ and $\sigma = \hat{\sigma}$, where $\hat{\mu}$ is equal to the mode of the kernel density estimate, and $\hat{\sigma}$ is calculated using the following equation:

$$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i}^{x_i \leq \hat{\mu}} (x_i - \hat{\mu})^2} \quad (4.1)$$

Potential enhancers that had a p-value > 0.05 were filtered, yielding a final set of 30,681 putative differential enhancers. These enhancers were assigned to genes they likely regulate using a heuristic method described by [McLean et al., 2010]. Briefly, each gene was assigned a cis-region defined as the region from the given gene’s TSS to the neighboring TSSs in either direction, or 1 Mb if the nearest TSS is further than 1 Mb. Enhancers that fall within a gene’s cis-region are assigned to that gene.

4.5.9 Gene segmentation

The calculation of the raw epigenetic profile is based on four segments delineated for each gene. The sizes of all but one segment are fixed. The remaining one accommodates the variable length of genes. The fixed size segments are: promoter (PR), transcription start site (TSS) and gene start (GS). The whole gene (WG) segment is variable in size but is at least 1.2 kb long. We define the sizes and boundaries of segments based on windows, which have a fixed size of 200 bp and their boundaries are independent of genomic landmarks such as TSSs. The location of the TSS defines the reference window, which together with its two adjacent windows defines the TSS segment. The two remaining fixed-size segments, PR and GS, have a size of 25 windows (5 kb). The PR and GS segments are located immediately upstream and downstream of

the TSS segment respectively, while the WG segment begins at the TSS reference window and extends 5 windows (1 kb) beyond the window containing the transcription termination site. Enhancers were treated as single-segment, contiguous 11-window (2200 bp) regions (see section 4.5.6).

4.5.10 Differential epigenetic profiles

We calculated differential epigenetic profiles (DEP) at both gene and enhancer loci. We base the DEPs on scaled differential enrichments (SDEs, see section 4.5.5) for all mapped histone modifications at gene loci, and enhancer associated marks at putative enhancer loci. The calculation is a multistep procedure that results in a profile (fixed-sized feature vector) that summarizes the multivariate differences in histone modification levels between the paired samples at each locus. In the first step, gene loci are split into segments (see section 4.5.9), while enhancers are kept whole. Next, within all segments, SDEs for each considered histone modification are quantified (see section 4.5.11).

4.5.11 Signal quantification and scaling

The genome-wide SDEs quantify epithelial to mesenchymal differences for each mark at 200 bp resolution across the genome (Scaled Differential Enrichments). Each gene segment is comprised of a set of bookended windows (Figure 4.3). For each histone modification, and within each segment, we reduce the SDE to two numeric values, which intuitively capture the level of gain and loss of the mark in the epithelial to mesenchymal direction. Strictly speaking, we independently calculate the absolute value of the sum of the positive (gain) and negative (loss) values of the SDE within a segment. Hence, we obtain a gain and loss value for all histone modifications within each segment of a gene (or an enhancer region). The differential epigenetic profile (DEP) of each gene (or enhancer) is a vector of gains and losses of multiple histone

modifications at all segments (single segment for enhancers). In the case of gene loci we quantify all histone marks, and in the case of enhancer loci only the enhancer-associated modifications (see section 4.5.7). DEPs are arranged into a DEP matrix individually for genes and enhancers (Figure 4.4A). Each row represents a DEP for a gene (or enhancer) and each column represents a segment-mark-direction combination (features). Columns (features) were non-linearly scaled using the following sigmoid equation:

$$z = \frac{2}{1 + e^{\frac{-2x}{u}}} \quad (4.2)$$

Where z is the scaled value, x is the raw value, and u is the value of some upper percentile of all values of a feature. We have chosen the 95th percentile. Intuitively, this corrects for differences in the dynamic range of changes to histone modification levels and for differences in segment size. Scaled values (DEP elements) are within the 0-1 range. The scaling is approximately linear for about 95% of the data points.

4.5.12 Annotation with GO-terms

Each gene was comprehensively annotated with Gene Ontology terms combined from two primary annotation sources: EBI GOA (retrieved 20110905) and NCBI gene2go (retrieved September 4, 2011). These annotations were merged at the transcript cluster level (see section 4.5.3), which means that GO-terms associated with isoforms were propagated onto the canonical transcript. Every protein-coding gene was re-annotated with terms from two GO-slms provided by the Gene Ontology consortium. The re-annotation procedure takes specific terms and translates them to generic ones. We used the map2slim tool and the two sets of generic terms: “PIR” (Protein Informatics Resource) and “generic terms.” In addition to GO, we have included two other major annotation sources: NCBI BioSystems, and the Molecular Signature Database 3.0 (MSigDB).

4.5.13 Mining for genes associated with EMT

We attempted to construct a representative list of genes relevant to EMT. This list was obtained through a manual survey of relevant and recent literature. We extracted gene mentions from recent reviews on the epithelial-mesenchymal transition. A total of 142 genes were retrieved and resolved to UCSC transcripts. A second set of genes associated with EMT was based on GO annotations.

4.5.14 Functional similarity scores

We developed a score to quantify functional similarity for any two sets of genes. Strictly speaking, the functional similarity score (FSS) measures the degree of overlap between the two lists of GO-terms enriched for the two sets. First, we obtain two lists of significantly enriched GO-terms for the two sets of genes. The enrichment p-values were calculated using Fisher's Exact Test and FDR-adjusted for multiple hypothesis testing. For each enriched term we also calculate the fold change; i.e., whether it is enriched or depleted relative to the background frequency. The similarity between any two sets is given by

$$FSS(A, B) = \sum_c^C \log(p_c^A \times p_c^B) + \sum_d^D \log(p_d^A \times p_d^B) \quad (4.3)$$

Where A and B are two lists of significantly enriched GO-terms (here FDR-corrected $p < 0.01$). C and D are sets of GO-terms that are either enriched or depleted in both lists, but not enriched in A and depleted in B and vice-versa. Intuitively, this score increases for every significant term that is shared between two sets of genes, with the restriction that the term cannot be enriched in one, but depleted in the other cluster. If one of the sets of genes is a reference list of EMT-associated genes this functional similarity score is, in general terms, a measure of relatedness to the functional aspects

of EMT.

4.5.15 Selection of optimal clustering

We have followed a heuristic benchmarking approach to select a suitable unsupervised clustering method to group genes based on differential epigenetic profiles, while maximizing the biological interpretability of DEPs. Because there is no correct solution to unsupervised machine learning tasks, we evaluated clustering solutions based on their interpretability in the domain of the epithelial-mesenchymal transition. Intuitively, a “good” clustering method groups genes with similar functions together. Therefore, we expected a small number of the clusters to be enriched for genes related to the EMT process (see section 4.5.13). However, such straightforward approach would have the drawback of being strongly biased towards what is known, whereas the goal of unsupervised machine learning is to uncover what is not. To alleviate this problem rather than calculating enrichments for genes known to be involved in EMT we calculated the FSS that measures the degree of functional similarity between a cluster and a reference set of genes associated with EMT. Our goal was to find a combination of gene segmentation, data scaling and machine learning algorithm that performs well in grouping functionally related genes together. We evaluated three markedly different unsupervised learning methods: hierarchical clustering, AutoSOME [Newman and Cooper, 2010], and WGCNA [Langfelder and Horvath, 2008], a number of ways to partition gene loci into segments, and scaling methods. Based on the distribution of EMT FSSs and a number of semi-quantitative indicators such as cluster size, differential gene expression we chose a final methodology (see sections 4.5.16, 4.5.9, and 4.5.11).

4.5.16 Clustering of gene and enhancer loci

DEP matrices (see section 4.5.11) associated with each of the 20,707 canonical transcripts (genes) and each of the 30,681 final enhancers were clustered using AutoSOME with the following settings: -P -g10 -p0.05 -e200. The output of AutoSOME is a crisp assignment of genes (or enhancers) into clusters and each cluster contains genes (enhancers) with similar DEPs. For visualization, columns (features) were clustered using hierarchical Ward clustering and manually rearranged if necessary.

4.5.17 TF-binding sites within promoters and enhancers.

Transcription factor binding sites were obtained from the ENCODE transcription factor ChIP track of the UCSC genome browser (downloaded December 15, 2011) [Kuhn et al., 2013]. This dataset contains a total of 2,750,490 binding sites for 148 different factors pooled from variety of cell types from the ENCODE project. The enrichment of each transcription factor in each enhancer and gene cluster was calculated as the cardinality of the set of enhancers or promoters (5400 bp, centered on the window containing the transcription start site) that have a nonzero overlap with a given set transcription factor binding sites. The significance of the enrichment was calculated using a one-tailed Fisher’s Exact Test (cluster membership vs. TF enrichment).

4.6 Chapter acknowledgements

This chapter was adapted from a manuscript currently in review. The authors of this manuscript include Marcin Cieřlik, Stephen Hoang, Natalya Baranova, Sanjay Chodaparambil, Manish Kumar, David Allison, Xiaojiang Xu, J. Jacob Wamsley, Lisa Gray, David Jones, Marty Mayo, and Stefan Bekiranov. Marcin Cieřlik and Stephen Hoang were the lead authors of this work, and share equal credit for all of

the bioinformatics work and the preparation of the manuscript. Marty Mayo and Stefan Bekiranov share equal credit as primary investigators for this work. The other authors are responsible for all of the wet lab work involved in this study. In particular, Natalya Baranova is responsible for generating the vast majority of the ChIP-seq data.

Chapter 5

The network architecture of the *Saccharomyces cerevisiae* genome

5.1 Introduction

The non-random spatial organization of chromosomes in the eukaryotic nucleus is strongly associated with various types of genomic regulation. Spatial compartmentalization has been shown, in many organisms, to correspond to transcriptional regulation, DNA replication, and chromatin states [Lieberman-Aiden et al., 2009, Nora et al., 2012, Ryba et al., 2010, Tolhuis et al., 2002]. Thus, techniques to understand the structure-function relationships in the genome will be critical to advance our understanding of genomic regulation.

Chromosome conformation capture (3C) technology has enabled the identification of long-range interactions between genomic loci [Dekker et al., 2002]. High-throughput methods, such as Hi-C and ChIA-PET, have built on the 3C framework, and are capable of comprehensively mapping spatial interactions throughout the genome [Lieberman-Aiden et al., 2009, Fullwood et al., 2009b]. These techniques have enabled

investigation of the spatial organization of whole genomes. Data generated by these technologies can be challenging to analyze due to their high complexity, and low signal-to-noise ratios [Simonis et al., 2007]. However, several groups have used these data to characterize genomic folding principles, interactions between regulatory elements, and functional territories composed of distant genomic regions [Lieberman-Aiden et al., 2009, Dixon et al., 2012, Fullwood et al., 2009a, Li et al., 2012a]. A variety of strategies have been employed to analyze these data, including polymer-based physical models, molecular dynamic simulations, hidden Markov models, and three-dimensional reconstructions [Dixon et al., 2012, Dorier and Stasiak, 2010, Di Stefano et al., 2013, Duan et al., 2010]. Each approach has limitations, and new approaches will be required to explore the full richness of these datasets (see [Dekker et al., 2013] for a Review).

Genomic interaction data is essentially composed of pairwise relationships between genomic regions. Since networks abstractly represent pairwise relationships between objects, this type of data has an inherent network structure. Thus, networks can be used to generate highly intuitive representations of this type of data. Networks are also a convenient and highly flexible framework for storing, analyzing, and integrating interaction data. Furthermore, information of biological interest that is contained in interaction data, such as compartmental characteristics, can be extracted by analyzing the architectural properties of an interaction network. As is necessary to analyze genome-scale datasets, efficient algorithms have been developed to identify some of these network properties in very large networks [Blondel et al., 2008, Tomita et al., 2006].

Here we demonstrate how intuitive, biologically meaningful analyses of large genomic interaction datasets can be achieved purely through network abstractions. Although some groups have begun to employ networks for analyzing gene-gene and other types of interactions from Hi-C data [Wang et al., 2013], and transcription

factor-biased ChIA-PET data [Sandhu et al., 2012], to our knowledge no network-based methods have been applied to unbiased genomic interaction data.

In this study, we generate and analyze network models constructed from an unbiased genome-wide interaction dataset generated in *Saccharomyces cerevisiae* by [Duan et al., 2010]. We investigate two structural properties of these networks, namely, communities and cliques. Briefly, communities are sets of densely connected nodes within a network, and cliques are sets fully connected (all to all) nodes in a network. We focus on these structural network properties, because they directly correspond to spatial grouping in a genomic interaction network. We investigate how these network features correspond to regulatory properties of the genome, such as replication timing, and protein binding data. We also explore the use of community detection techniques in analyzing the hierarchical interaction structure of the genome. The analyses presented here represent a general framework and proof-of-principle for using networks to infer genomic organization from unbiased genomic interaction data.

5.2 Results and discussion

5.2.1 Inter-chromosomal cliques replicate early, and are enriched for cohesin

We created network models of genomic interactions, where nodes represent genomic loci, and edges represent statistically significant interactions between loci ($< 1\%$ FDR, unless otherwise stated). Several groups have noted the highly stochastic nature of these interactions in vivo [Simonis et al., 2006, Osborne et al., 2004]. Only a relatively small fraction of a population of cells exhibit a given interaction in an experiment [Dekker et al., 2013, van Steensel and Dekker, 2010, Gibcus and Dekker, 2013]. For this reason, we employ methods of network analysis that are robust to

the addition of “noisy” edges. One such procedure for detecting regions of strong interaction is clique detection. Cliques are sets of nodes that show complete interaction (all connected to all). Because of their specific topology, large cliques are unlikely to form at random in relatively sparse networks. Therefore, genomic regions that are members of large cliques likely represent sets of regions that exhibit relatively robust and stable interactions.

The known functions of cohesin make it an excellent candidate for a mediator of stable inter-chromosomal interactions. In budding yeast, cohesin has a well established role in mediating inter-chromosomal cohesion between newly replicated sister chromatids [Sherwood et al., 2010, Uhlmann, 2009]. There is also evidence that mutations in cohesin pathway proteins can lead to disruption in chromatin condensation and organization [Gard et al., 2009]. In mammalian cells, cohesin has been shown to be necessary to establish and maintain functional through-space chromatin interactions that influence transcriptional regulation [Hadjur et al., 2009, Mishiro et al., 2009, Nativio et al., 2009]. Although sister chromatid cohesion is well known, other types of cohesin-mediated inter-chromosomal interactions are not well studied in budding yeast. Therefore, we chose to investigate cohesin enrichment at inter-chromosomal cliques to (1) look for evidence that cohesin is involved in establishing stable inter-chromosomal interactions, and (2) to evaluate the biological relevance of cliques.

In the inter-chromosomal network, we calculated the maximum clique size for each genomic fragment, which is the size of the largest clique of which a given fragment is a member. At each of these fragments we also assessed the enrichment levels of the cohesin subunits Scc1 and Smc3, as well as the cohesin loader subunits Scc2 and Scc4. Since cohesin proteins mediate inter-chromosomal interactions, we expected to see high levels of these factors in large cliques. Indeed, there is a clear trend of increasing levels of both cohesin and its loader with increasing inter-chromosomal clique size (Figures 5.1 and 5.2). By definition, every fragment in an inter-chromosomal clique represents

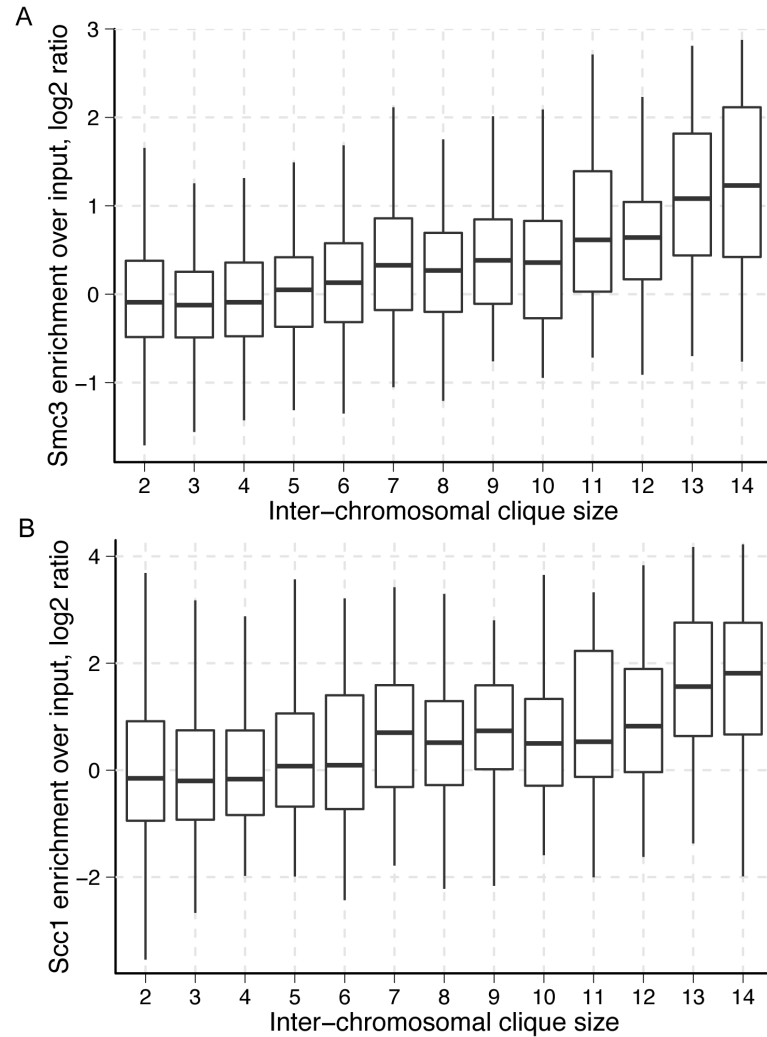


Figure 5.1: Cohesin enrichment vs. inter-chromosomal maximal clique size. Enrichment of cohesin subunits (A) Smc3 and (B) Scc1 with respect to maximal fragment clique size. The maximal clique size for a fragment is the size of the largest clique to which a genomic fragment belongs. Each member of an inter-chromosomal clique represents a fragment from a different chromosome. Thus, cohesin enrichment increases with number of interacting chromosomes.

a different chromosome. This suggests that in addition to its role in sister chromatid cohesion, cohesin may be directly involved in maintaining through-space interactions where many chromosomes come together in a single region in space. Interestingly, there is no trend between any of these factors and intra-chromosomal clique size (Figures 5.3 and 5.4). This result suggests that cohesin has a less prominent role in directly mediating intra-chromosomal interactions.

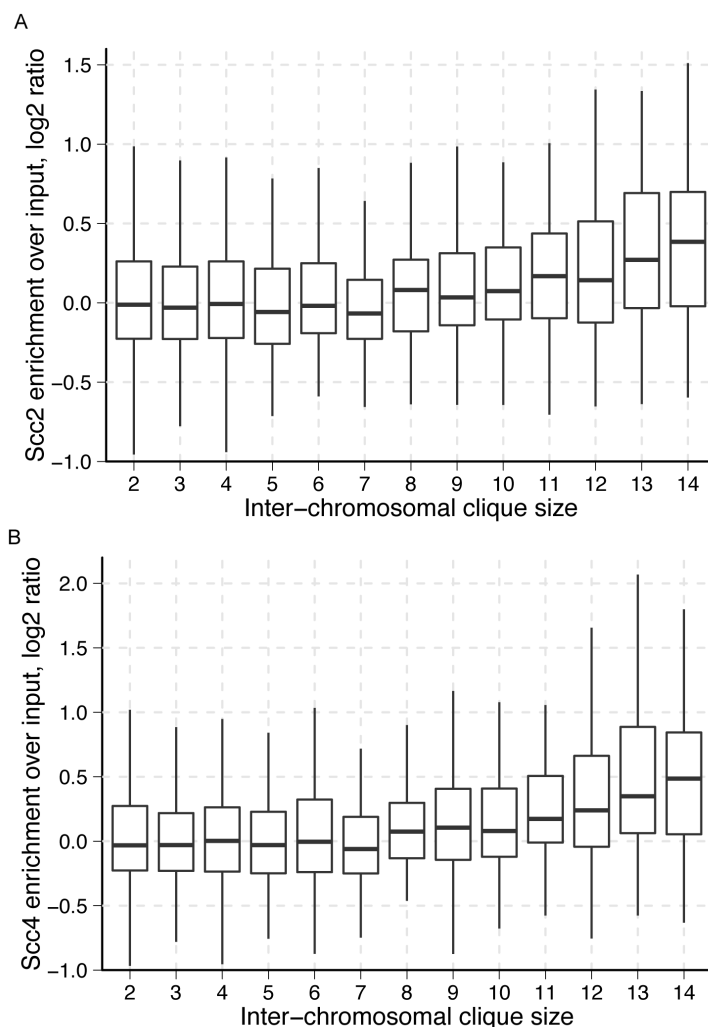


Figure 5.2: **Cohesin loader enrichment vs. inter-chromosomal maximal clique size.** Enrichment of cohesin loader subunits (A) Scc2 and (B) Scc4 with respect to maximal fragment clique size. Like cohesin itself, cohesin loader enrichment increases with number of interacting chromosomes.

Cohesin has also been shown to be recruited to sites of active replication in budding yeast [Tittel-Elmer et al., 2012]. Moreover, in mammalian systems it has been shown that the level of chromosomal interaction correlates strongly with replication timing [Ryba et al., 2010]. It has also been postulated that cohesin mediates chromosomal conformations that are favorable for efficient replication [Guillou et al., 2010]. Since we see both high cohesin enrichment and a high degree of inter-chromosomal interactions in large cliques, we expected to see a strong relationship between inter-chromosomal

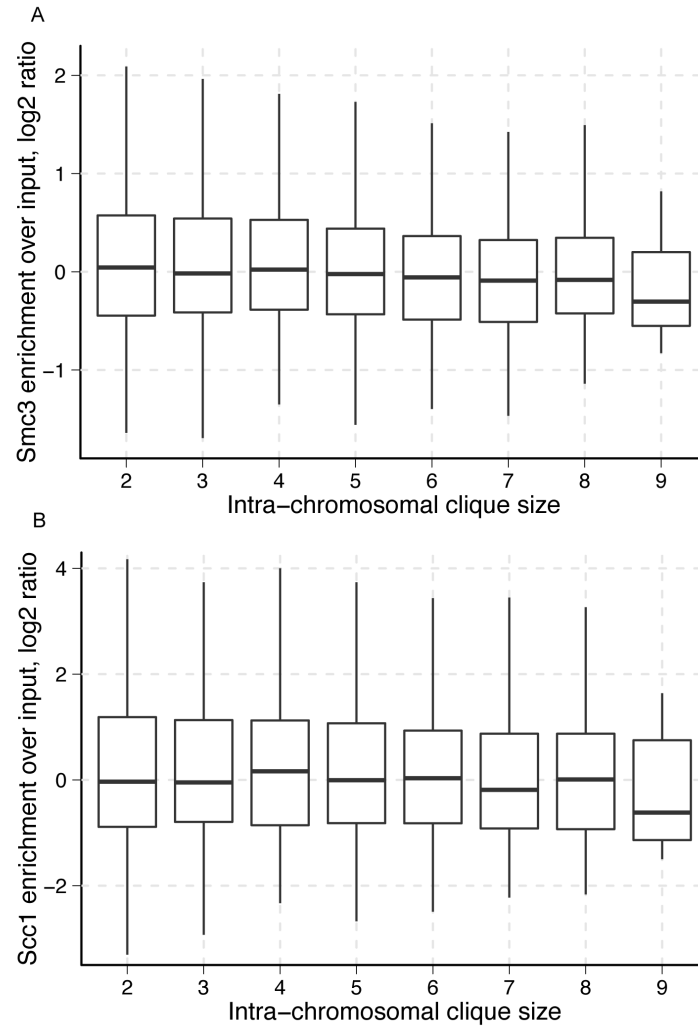


Figure 5.3: **Cohesin enrichment vs. intra-chromosomal maximal clique size.** Enrichment of cohesin subunits (A) Smc3 and (B) Scc1 with respect to maximal fragment clique size in the intra-chromosomal network. This plot includes intra-chromosomal cliques across all chromosomes. Unlike the inter-chromosomal cliques, cohesin enrichment and intra-chromosomal clique size are independent.

clique size and replication timing. Indeed, that is what we observed (Figure 5.5, n.b., higher % replication indicates earlier replication, see [McCune et al., 2008] for details). However, like cohesin enrichment, we observed independence between intra-chromosomal clique size and replication timing (Figure 5.6). We also observed independence between clique sizes and gene expression (Figure 5.7). These findings suggest that regions of the genome that form stable interactions involving many different chromosomes tend to replicate early. This type of interaction can be expected

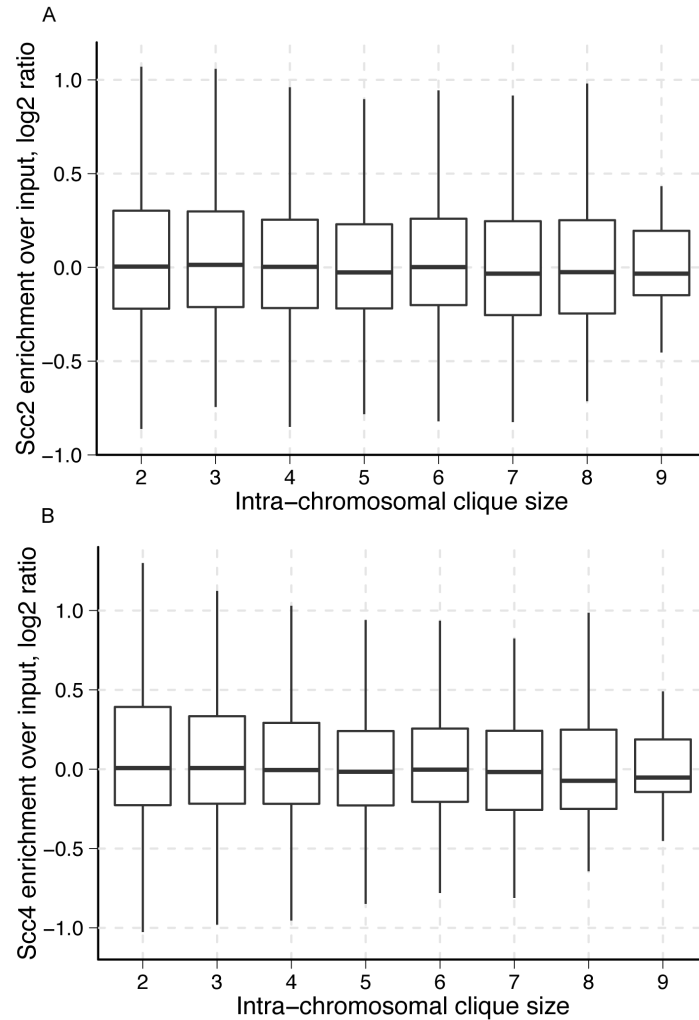


Figure 5.4: **Cohesin loader enrichment vs. intra-chromosomal maximal clique size.** Enrichment of cohesin loader subunits (A) Scc2 and (B) Scc4 with respect to maximal fragment clique size in the intra-chromosomal network. Like cohesin itself, cohesin loader enrichment and intra-chromosomal clique size are independent.

to occur in centromeric regions in budding yeast, due to the known rosette organization of the genome, where chromosome arms extend from a centromeric cluster near one spindle pole [Duan et al., 2010, Jin et al., 2000, Bystricky et al., 2004]. Moreover, centromeric regions are well established as regions of early replication in budding yeast [McCarroll and Fangman, 1988, Feng et al., 2009]. Though these relationships have been established, to our knowledge, a direct relationship between number of inter-chromosomal interactions and replication timing has not been shown.

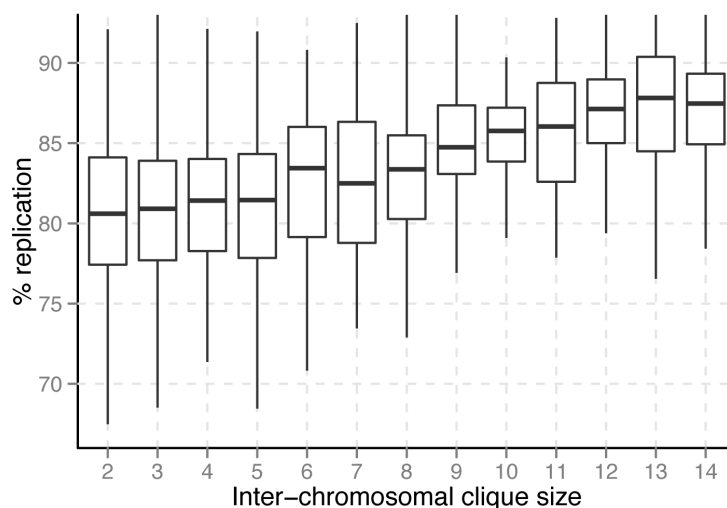


Figure 5.5: **Replication timing vs. inter-chromosomal maximal clique size.** Higher % replication indicates earlier replication. Larger inter-chromosomal clique size clearly trends with earlier replication. Sites where many chromosomes make stable contacts tend to replicate early.

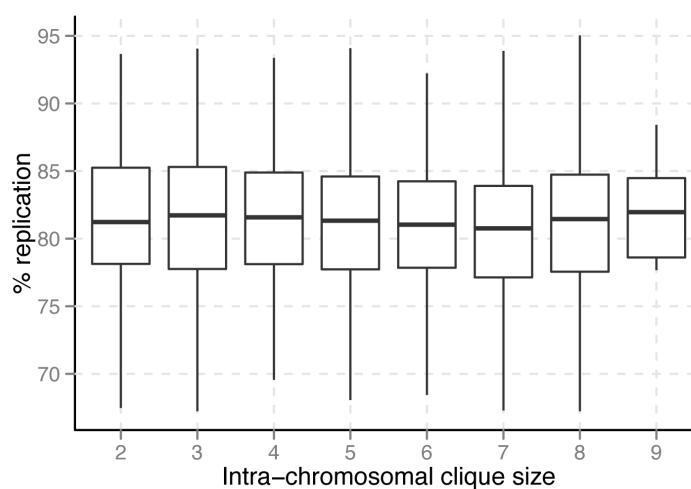


Figure 5.6: **Replication timing vs. intra-chromosomal maximal clique size.** Unlike inter-chromosomal cliques, intra-chromosomal clique size and replication timing are independent.

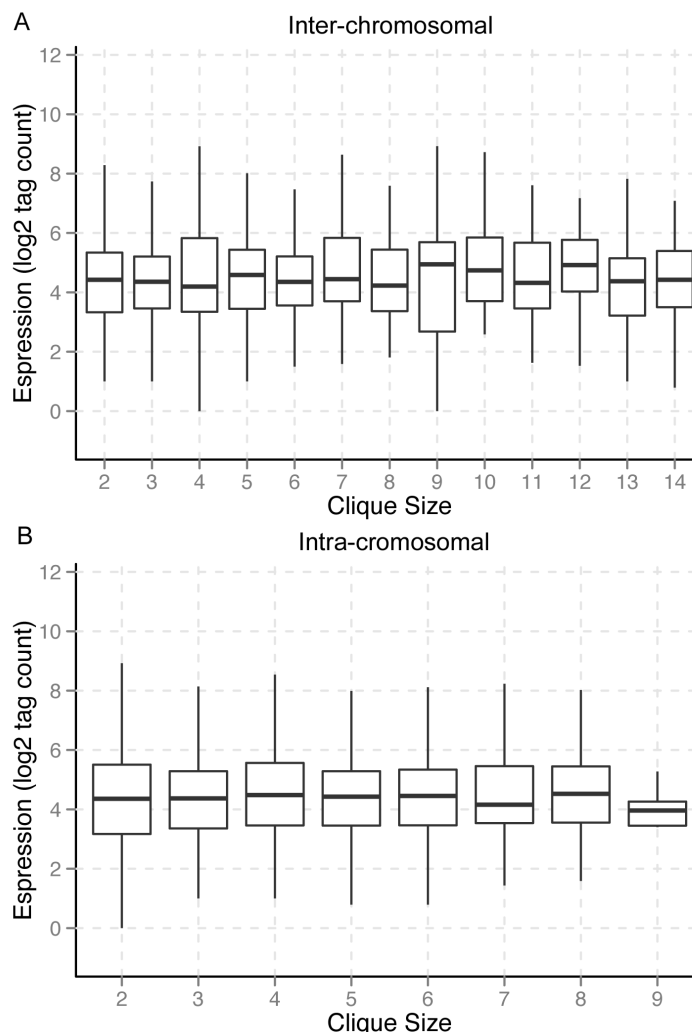


Figure 5.7: **Expression vs. inter-chromosomal maximal clique size.** Gene expression level is independent of the inter-chromosomal clique size of its genomic locus.

The clear trends between inter-chromosomal clique size, cohesin enrichment, and replication timing demonstrate that biologically relevant information can be gleaned from the structural properties of genomic interaction networks. Relatively complex information about the interaction behavior of a genomic region can be obtained through simple characterizations of an interaction network. Since clique size is a relatively simple aspect of the inter-chromosomal network architecture, these findings demonstrate the potential for more sophisticated network analyses.

5.2.2 Community detection

Communities are groups of densely connected nodes in a network. In a genomic interaction network, communities represent dense clusters of interacting genomic loci, e.g., chromosome territories. Therefore, the community structure of the genomic interaction network is of great interest, since it reflects how the genome is spatially compartmentalized. The budding yeast genome has been shown to have some degree of compartmentalization, including the clustering of the rDNA locus on chromosome XII [Léger-Silvestre et al., 1999], and the clustering of tRNAs [Thompson et al., 2003, Haeusler et al., 2008]. By comparison, metazoan genomes show a very high degree of spatial compartmentalization, including the formation of topologically associating domains (the so-called TADs), and transcription factories [Nora et al., 2012, Dixon et al., 2012, Drier and Stasiak, 2010]. The degree to which transcription factory structures form in yeast genomes is unclear [Taddei et al., 2010, Tanizawa et al., 2010]. In principle, community detection methods can be used to identify these types of structures.

Detecting communities involves partitioning the network so that nodes within communities are densely connected, and nodes between communities are sparsely connected. A commonly used metric for the quality of a partition is its modularity given by [Newman, 2004]

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \quad (5.1)$$

where A_{ij} is the adjacency matrix of the network, m is the sum of the edge weights in the network k_i is the sum of the edge weights attached to node i , c_i is the community to which node i belongs, and δ is the Kronecker delta. Many community detection procedures take the approach of attempting to maximize modularity. However, optimizing modularity has been shown to be an NP-complete problem, and is thus

computationally intractable [Brandes et al., 2008]. Therefore, all of the modularity-based algorithms to detect communities are heuristic methods that approximate modularity maximization. Furthermore, there is a resolution limit associated with modularity, where communities below a certain size cannot be detected. This minimum community size is a function of the total number of edges in the network, and the ratio of outgoing edges to internal edges in the community being detected [Fortunato and Barthélemy, 2007]. However, methods have been developed that mitigate this limitation [Blondel et al., 2008, Khadivi et al., 2011].

To detect communities in our genomic interaction network, we implemented the so-called Louvain algorithm [Blondel et al., 2008]. This method hierarchically merges communities to maximize modularity. We selected this method of community detection for several reasons. First, the method has been shown to produce partitions with better global modularity than many other competing algorithms. Second, in terms of speed, the algorithm performs well on very large networks, having been successfully applied to networks with billions of nodes and hundreds of millions of edges. Third, the resolution limit does not strictly apply to this method. Finally, due to the hierarchical nature of the solution, intermediate steps toward the global solution could potentially give insight into the hierarchical community structure of a network.

5.2.3 Community detection is robust to interaction noise

Since the modularity resolution limit is a function of the total number of edges in the network, “noise” edges in the network increase the minimum detectable community size [Fortunato and Barthélemy, 2007]. However, the coarse-grained community structure of the network should be robust to noise. To confirm this, we generated pairs of networks, one with an edge FDR threshold of 1% and the other with an FDR threshold of 0.01%. It is worth noting that the latter is a subnetwork of the former. We then calculated the community membership recapitulation as the fraction

of genomic regions within a given community in the smaller network that are found within a single community in the larger network. We followed this procedure for one pair of networks made with inter-chromosomal interactions only, and another pair made with the union of inter-chromosomal and intra-chromosomal interactions (i.e., a complete interaction network).

The inter-chromosomal-only networks had 31,832 and 13,537 edges at the 1% and 0.01% FDR thresholds, respectively. The mean community recapitulation rate for these networks was 84% across the communities in the smaller network. The complete interaction networks had 59,132 and 31,426 edges at the 1% and 0.01% FDR thresholds, respectively. This pair yielded a mean community recapitulation of 93%. These calculations suggest that, at a coarse-grained scale, community detection is highly robust to the selection of significance thresholds for network edges. The subsequent analyses are done on networks with a 1% FDR edge threshold. We selected this less stringent threshold in order to incorporate larger portions of the genome.

5.2.4 Inter-chromosomal network has three major compartments

The inter-chromosomal network contains 2955 nodes and 31,832 edges. The partition solution to this network has three hierarchical levels (see [Blondel et al., 2008] for a detailed explanation of hierarchical structure of the solution): level 0, level 1 and level 2. Level 2 is the highest level of the partition hierarchy, and corresponds to the global maximum modularity found by the algorithm. At this level, the inter-chromosomal network partitions into three communities that pass our size filter (see Methods). These communities represent 98.7% of the nodes in the network. Community 0 contains 61.9% of the nodes in the network, community 6 contains 23.1%, and community 1 contains 13.7%. These three major communities roughly correspond distance from centromeric regions (Figure 5.8A). Community 1 corresponds to centromere-proximal

regions; community 6 tends to flank community 1 regions; and community 0 tends to comprise large portions of the chromosome arms, relatively far from the centromeres.

We looked at enrichment of several chromosomal features and transcriptional regulators in each of the three high-level communities. Community 1 contains all of the centromeres, so not surprisingly it has a highly significant enrichment for centromeres ($p = 9.36e-14$). It also has a significant enrichment for tRNAs ($p = 0.008$), which is consistent with the observation of a centromere-proximal spatial cluster of tRNAs [Duan et al., 2010]. Community 1 is also the only community of the three that has a significant enrichment for any of the 200 transcriptional regulators that we tested. Moreover, out of the 200 proteins, *Irr1*, a cohesin subunit, is the only one that is significantly enriched ($FDR = 4.21e-10$). This highly significant localization of cohesin in the centromeric community, and the enrichment of cohesin at large inter-chromosomal cliques, suggest that cohesin may play a role in maintaining the rosette configuration of the genome by creating inter-chromosomal links between different chromosomes in the centromeric community.

The centromere-distal communities had less dramatic enrichments. Community 0 does not contain enrichments for the chromosomal features, or any of the transcriptional regulators we assessed. This is not surprising, considering this community accounts for over half of the genome, and is the most sparsely connected of the three. Although, community 0 tends to be more centromere-distal than community 6, community 6 contains a significant enrichment for telomeres ($p = 0.0088$). This suggests possible looping associations between telomeres and telomere-distal regions of chromosomes. The size of communities 0 and 6 contribute to their non-specificity; that is, they are low-resolution communities. Therefore, we sought to explore the hierarchical community structure of the genome.

Communities at each successive hierarchical level of the detection algorithm represent aggregations of communities in the preceding level. Therefore, the intermediate

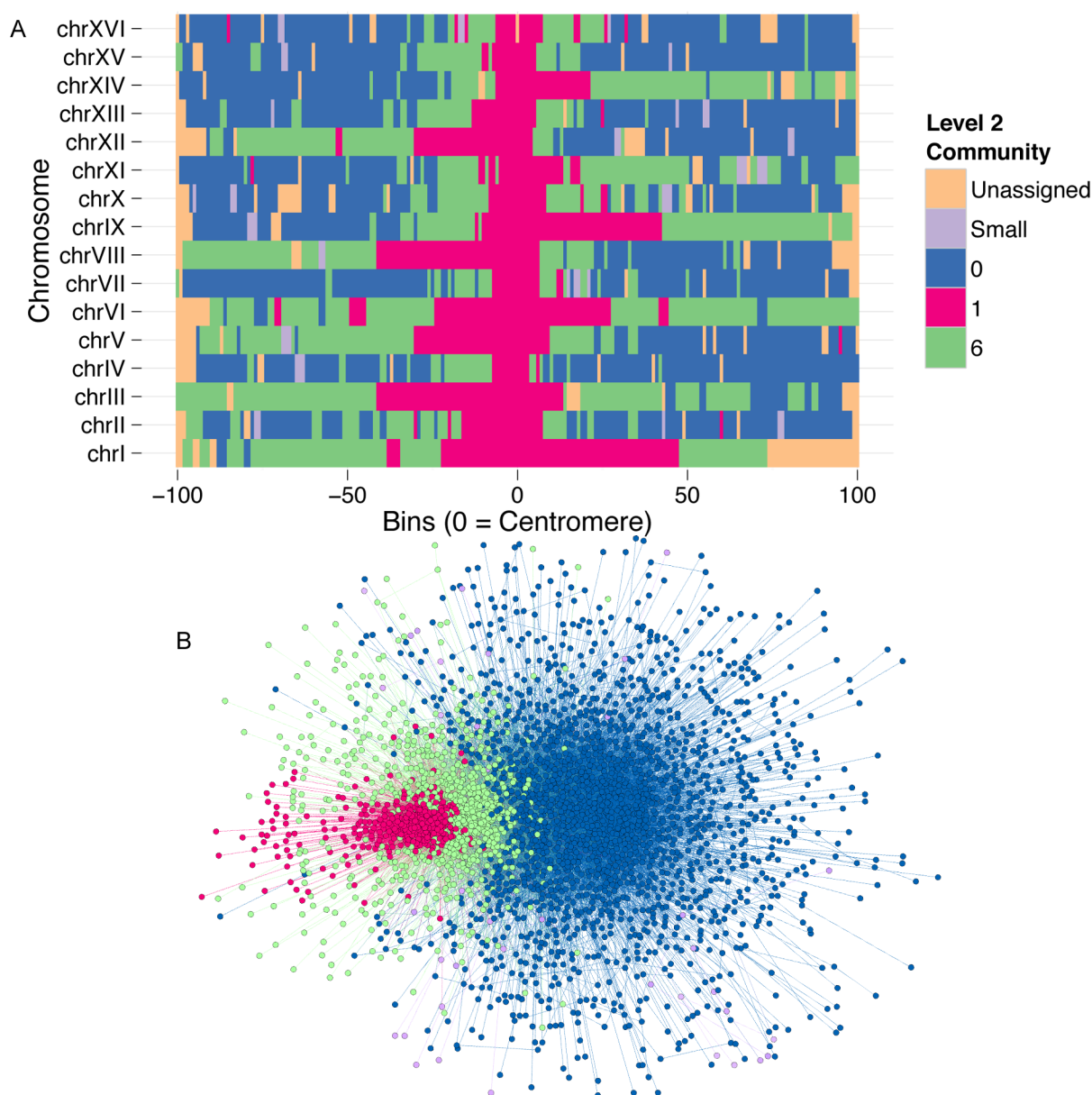


Figure 5.8: Community partition of the inter-chromosomal network. Partition showing the final solution of the community detection algorithm on the inter-chromosomal network. (A) Scaled chromosomes which are centered on centromeres show somewhat symmetrical community assignment about the centromere. (B) A force-directed network representation of the community partition shows the layered interaction structure of the genome. Together, these figures show the rosette configuration of the genome, where centromeres cluster, and chromosome arms extend in one direction away from the centromeres. Interaction domains are roughly stratified by the distance from the centromeres.

partitions of the inter-chromosomal network, should give information about the hierarchical structure of the network. However, the partition levels of this network give little indication of hierarchical structure. At level 1, there are three communities nearly identical to the three communities in level 2 (Figure 5.9A). In order of size they represent 61.9%, 22.7%, and 13.7% of the total number of nodes in the network. Therefore, most of the communities that were merged from level 1 to level 2 were below the size filter (see Methods). At level 0, we observe the unfolding of one small community, 16, which contains 1.3% of the total nodes in the network (Figure 5.9B). Interestingly, this community is strongly enriched for fragments that overlap telomeric regions ($p = 6.3e-9$). This is consistent with other studies that have shown the strong inter-chromosomal association of telomeres [Duan et al., 2010, Hediger et al., 2002, Schober et al., 2008]. Overall, the lack of separation of the major communities at lower levels in the hierarchy suggests that there is little hierarchical structure in this network. Indeed, a qualitative inspection of a force-directed layout of this network supports this conclusion (Figure 5.8B).

5.2.5 Subcommunities of the inter-chromosomal network are modular

One possibility for the lack of evidence for hierarchical structure in the inter-chromosomal network is that the intermediate solutions to the detection algorithm do not have the ability to resolve subcommunities. Moreover, force-directed layouts of the large network may not impose a geometry that allows visual discernment of community structure, especially for subtle communities. To further investigate the possibility of hierarchical structure in the inter-chromosomal network we performed community detection on the three subnetworks that represent each of the three major communities in the inter-chromosomal network. We treated each of these subnetworks as independent networks. In this section, we will only refer to the communities generated by the

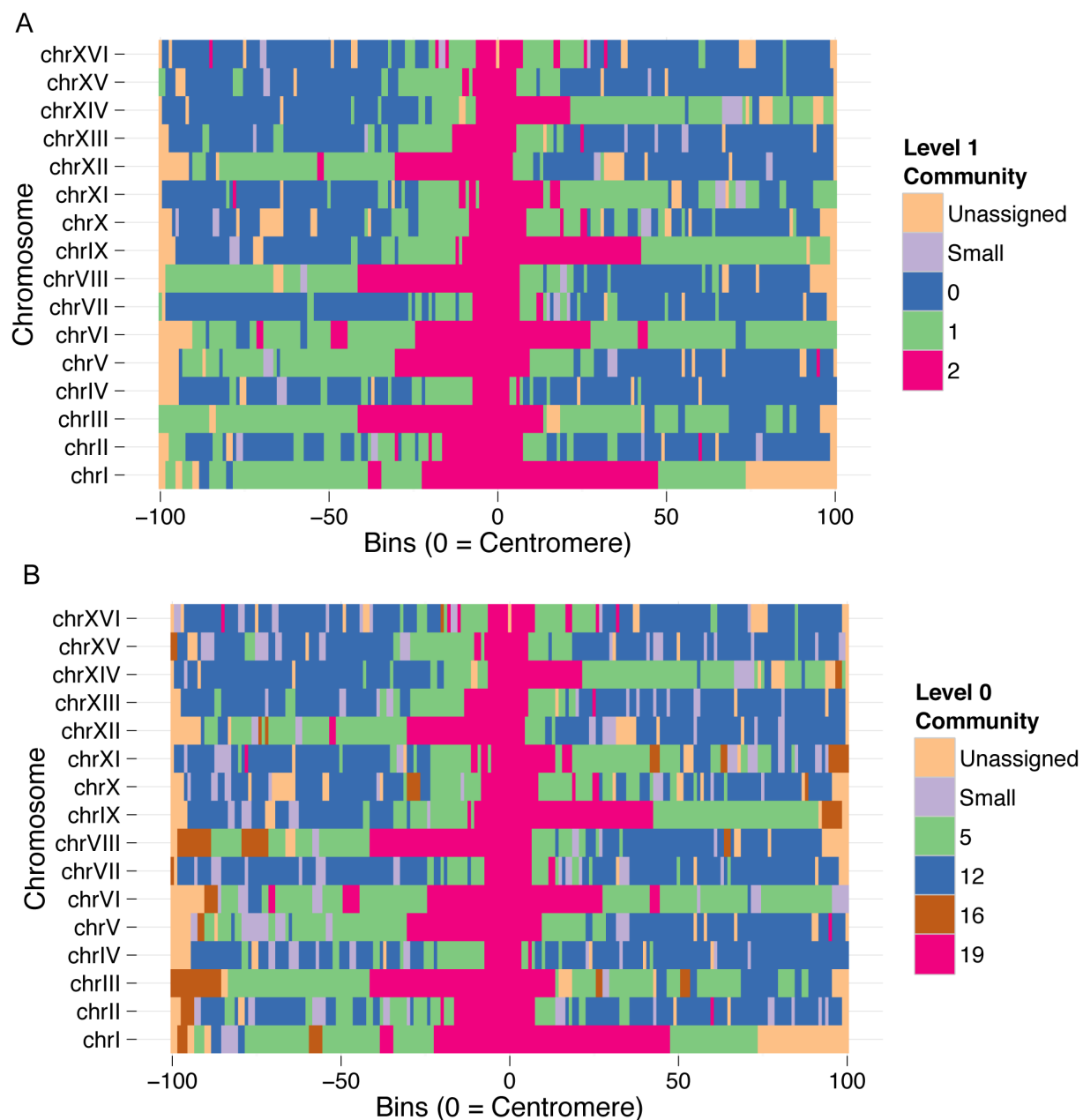


Figure 5.9: **Intermediate solutions to community detection in the inter-chromosomal network.** (A) The level 1 partition of the inter-chromosomal network is similar to the level 2 partition (Figure 5.8A), which is the final partition. (B) At the level 0 partition, community 16 emerges, which contains several telomeric fragments. Most of the community merges from level 0 to 2 involve small communities. Together, the intermediate solutions give relatively little insight into hierarchical community structure.

final partition of these subnetworks. Also, these communities-within-communities will henceforth be referred to as “subcommunities.” To assess the presence of hierarchical community structure in this data, we looked for evidence of modular structure in the subnetworks, and biological meaning in the subcommunities.

Since partitions of random networks can have highly variable modularity, a modularity value on its own does not have a meaning [Reichardt and Bornholdt, 2006]. Therefore, to assess the degree of modularity of each of the three communities we compared their modularity to random networks of equal size (edge number) and order (node number). We partitioned, and calculated the modularity of 10,000 random networks for each of the three subnetworks. We compared the modularity of the non-random community partitions to the empirical random modularity distributions using a standard score (Figure 5.10). All three subnetworks had modularity greater than their 10,000 matched random networks. Thus, these three communities have some degree of non-random subcommunity structure with $p < 0.0001$.

The subnetwork induced by community 1 (the centromeric community) of the inter-chromosomal network had the greatest degree of modularity over random (Figure 5.10, $Z = 16.46$). Thus, of the three subnetworks the centromeric network shows the strongest evidence for hierarchical organization. The community assignments for this subnetwork represent large, linearly contiguous segments of chromosomes (Figure 5.11A). This is remarkable because information about the linear orientation of fragments in the inter-chromosomal network is encoded through inter-chromosomal interactions. Thus, linearly contiguous subcommunity assignments are made purely through similarities in inter-chromosomal interactions. Linearly continuous community assignments are an indicator of a high degree of community structure within community 1. Unlike community 1 as a whole, none of the subcommunities within this subnetwork showed significant enrichment for binding sites of the 200 transcriptional regulators. However, the subcommunities distinguish themselves with respect to replication timing

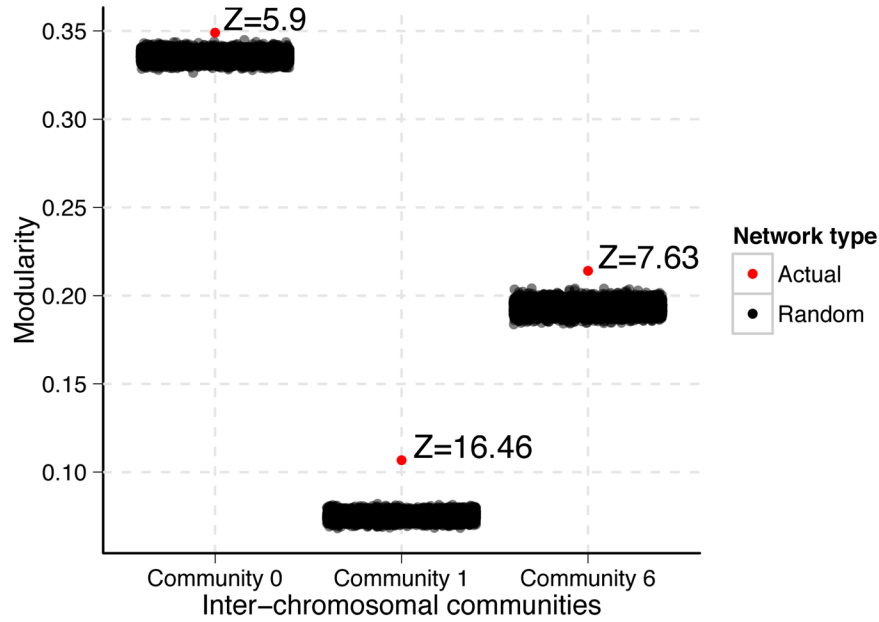


Figure 5.10: **Modularity of the inter-chromosomal communities.** The modularity over random of the subnetworks induced by each of the three major inter-chromosomal communities. The red point represents the modularity of the partition of the subnetwork. The black points represent the modularity of the partitions of 10,000 random subnetworks of equal size and order. The standard score of each red point relative to the black points are given. All three communities show non-random modularity.

and cohesin enrichment levels. Strikingly, ordering the subcommunities by median replication timing or by median cohesin enrichment produces the same result (Figures 5.11B and 5.11C).

Community 6 of the inter-chromosomal network showed the second largest modularity over random ($Z = 7.63$). Like the partition of community 1, we see large contiguous chromosomal segments assigned to a single subcommunity (Figure 5.12A). However, there are also many subcommunities that are highly fragmented and interleaved, potentially indicating a low degree of hierarchical structure. This qualitative assessment is consistent with this community's modularity over random, relative to that of community 1. Community 6 tends to flank centromere-proximal regions (community 1), but is also enriched for telomeric regions. Consistently, we find that subcommunity 2 is significantly enriched for telomeres ($p = 5.6e-5$). Along with the highly significant

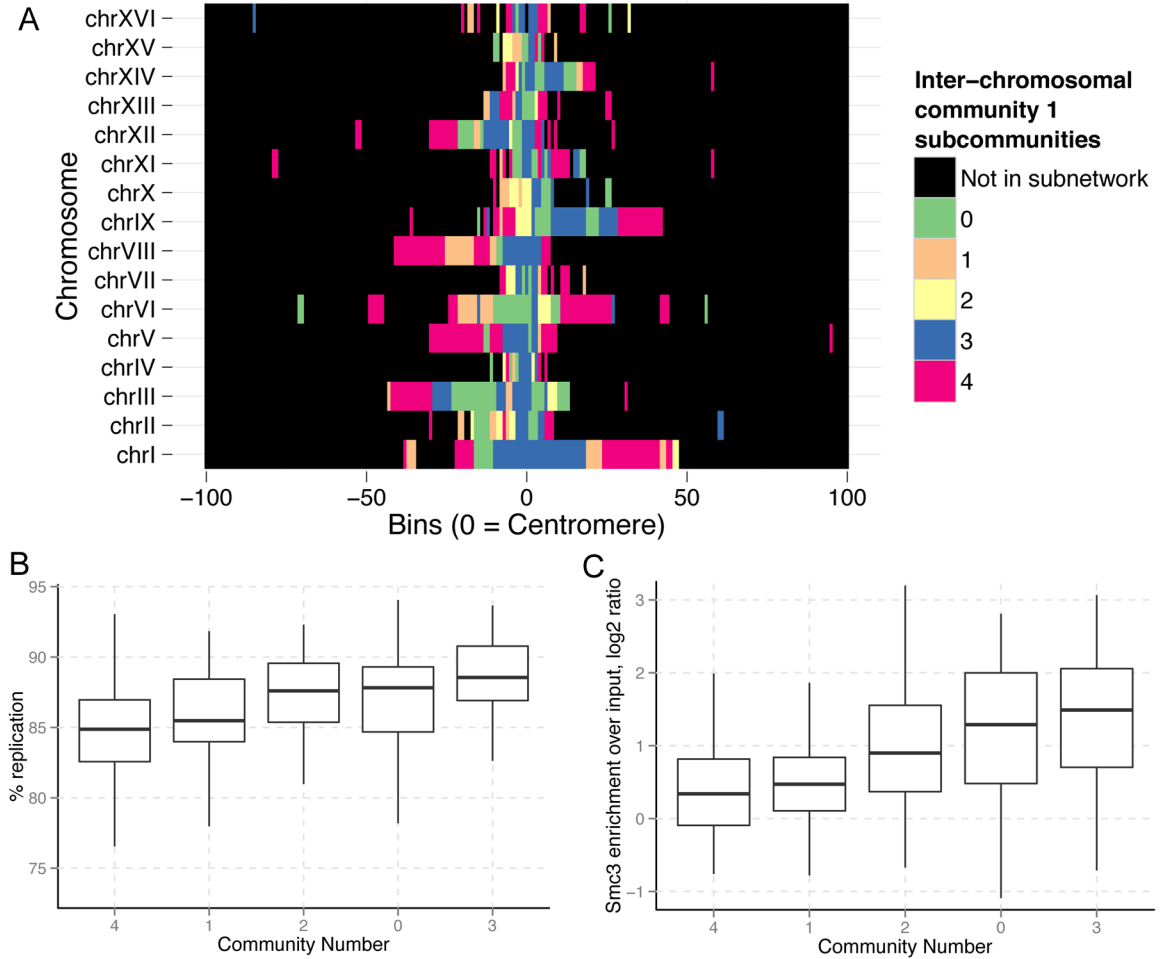


Figure 5.11: Partition of the inter-chromosomal centromeric community. The subnetwork induced by community 1 of the final inter-chromosomal partition was repartitioned. (A) The community assignments show long linear stretches that belong to a single community, which demonstrates that linear orientation information is encoded in inter-chromosomal contact information. This subnetwork partitions into communities that can be distinguished by (B) replication timing, and (C) cohesin enrichment.

grouping of telomeres in the level 0 partition of the whole inter-chromosomal network, this demonstrates that inter-chromosomal interactions between telomeres form highly distinct clusters in this dataset.

Inter-chromosomal community 0 shows the weakest modularity over random ($Z = 5.9$). Accordingly, it has very few large contiguous subcommunities (Figure 5.12B). The only chromosomal feature enrichment that we observed was a significant enrichment for tRNAs in subcommunity 25 ($p = 0.0033$). Consistent with others [Duan et al., 2010],

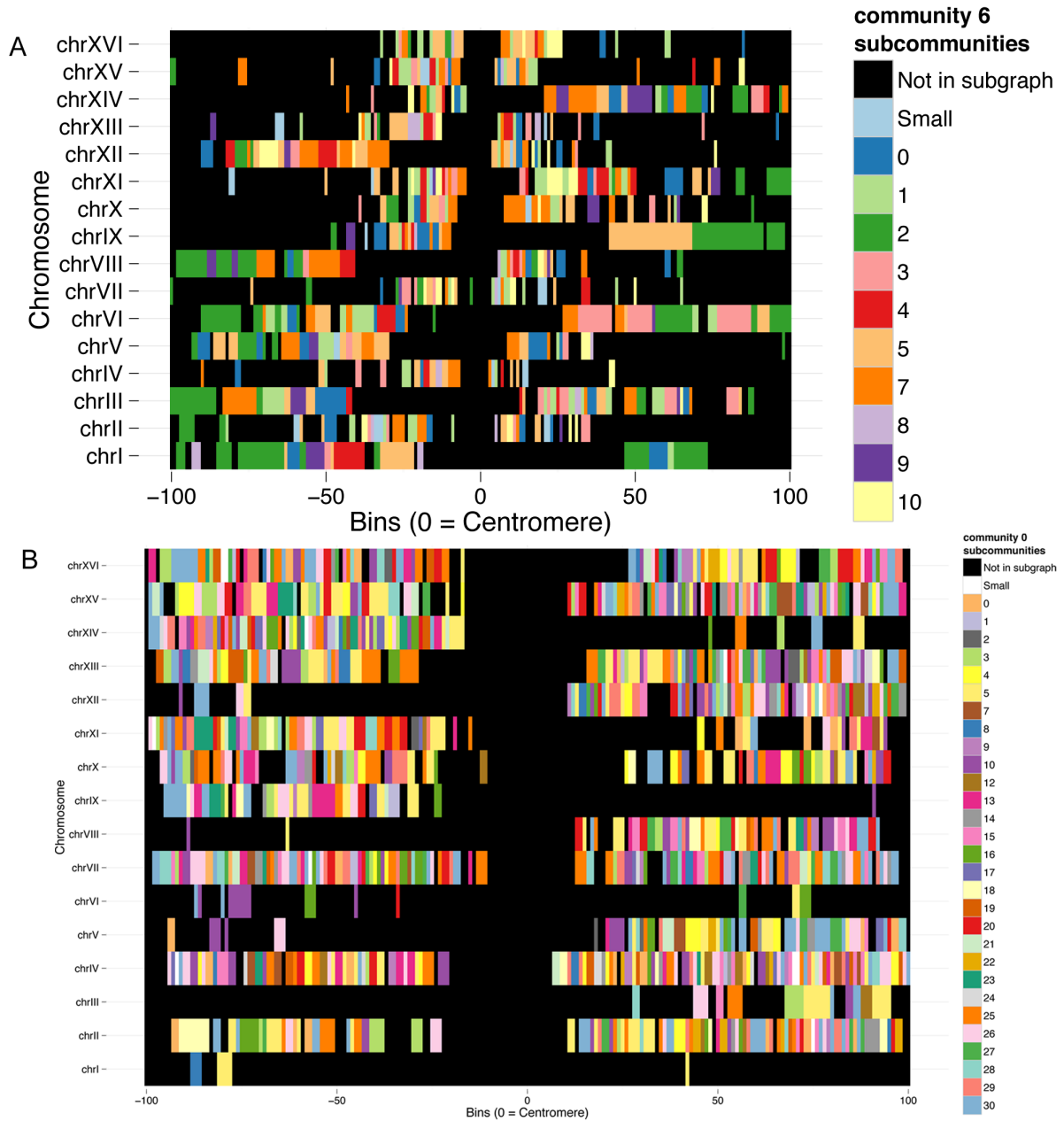


Figure 5.12: **Partitions of inter-chromosomal community 6 and 0.**(A) The partition of the subnetwork induced by community 6 shows several large continuous community assignments, indicating some modular community structure. (B) The partition of the community 0 subnetwork is highly fragmented, indicating very little modular community structure.

we find two regions of significant tRNA clustering, one at centromeric community in the full inter-chromosomal network, and here in subcommunity 25. Based on the findings in [Duan et al., 2010], this grouping of tRNAs is presumably proximal to the

nucleolus. Intriguingly, of all of the communities in this subnetwork, this is the only subcommunity that shows enrichment at a 1% FDR threshold for any transcriptional regulator of the 200 tested. Even more striking, this community is enriched for 24 of the 200 factors (Table 5.1). Together, these finding suggest biological meaning to the partition of subcommunity 25. However, overall inter-chromosomal community 0 has subtle community structure.

Table 5.1: Transcriptional regulators significantly enriched in subcommunity 25

| TF | Total frags in subcommunity 25 | Total Frags in community 0 | Community 0 Frags with TF | Subcommunity 25 and TF frag overlap | Expected overlap | FDR |
|--------|--------------------------------|----------------------------|---------------------------|-------------------------------------|------------------|-------------|
| Gcr1 | 234 | 1830 | 389 | 79 | 49.74098361 | 0.000196441 |
| Isw2 | 234 | 1830 | 815 | 139 | 104.2131148 | 0.000196441 |
| Rtt103 | 234 | 1830 | 763 | 127 | 97.56393443 | 0.001717116 |
| Tfa2 | 234 | 1830 | 829 | 135 | 106.0032787 | 0.001717116 |
| Hpa3 | 234 | 1830 | 470 | 86 | 60.09836066 | 0.001717116 |
| Ssl2 | 234 | 1830 | 1148 | 175 | 146.7934426 | 0.001717116 |
| Pcf11 | 234 | 1830 | 477 | 87 | 60.99344262 | 0.001717116 |
| Rgr1 | 234 | 1830 | 1015 | 158 | 129.7868852 | 0.001831788 |
| Otu1 | 234 | 1830 | 427 | 79 | 54.6 | 0.002047512 |
| Ino4 | 234 | 1830 | 696 | 116 | 88.99672131 | 0.002428898 |
| Rpb2 | 234 | 1830 | 938 | 147 | 119.9409836 | 0.002428898 |
| Set2 | 234 | 1830 | 760 | 124 | 97.18032787 | 0.002428898 |
| Not5 | 234 | 1830 | 147 | 35 | 18.79672131 | 0.002428898 |
| Tfc8 | 234 | 1830 | 334 | 63 | 42.70819672 | 0.005488213 |
| Htb2 | 234 | 1830 | 670 | 110 | 85.67213115 | 0.005522941 |
| Irc20 | 234 | 1830 | 661 | 109 | 84.52131148 | 0.005522941 |
| Rsc9 | 234 | 1830 | 993 | 152 | 126.9737705 | 0.005522941 |
| Ccr4 | 234 | 1830 | 110 | 27 | 14.06557377 | 0.005901224 |
| Sif2 | 234 | 1830 | 318 | 60 | 40.66229508 | 0.006163883 |
| Ioc2 | 234 | 1830 | 976 | 149 | 124.8 | 0.007186381 |
| Rpb7 | 234 | 1830 | 1142 | 169 | 146.0262295 | 0.007453414 |
| Rpd3 | 234 | 1830 | 1035 | 156 | 132.3442623 | 0.007453414 |
| Zms1 | 234 | 1830 | 315 | 59 | 40.27868852 | 0.007453414 |
| Sfp1 | 234 | 1830 | 833 | 130 | 106.5147541 | 0.009669288 |

Subcommunity 25 is significantly enriched for 24 different TFs, and is the only subcommunity in the community 0 subnetwork to be enriched for any TFs at 1% FDR threshold.

5.2.6 The complete network highlights high-level organization

Next, we partitioned the network containing both intra- and inter-chromosomal edges (which will be referred to as the “complete network”) into communities. The

interpretation of this network has a major caveat associated with it. The FDRs of intra-chromosomal and inter-chromosomal links were calculated using different assumptions and probability models (see [Duan et al., 2010] for details). Therefore, the “actual” significance of an edge is likely different for an intra-chromosomal and inter-chromosomal edge at the same FDR value. A network incorporating both types of edges will thus be distorted, having an imbalance of one type of edge over the other. Nevertheless, this network can give some insights into the organizational principles of the genome

The solution to the complete network partition has two hierarchical levels: level 0 (Figure 5.13), and level 1 (Figure 5.14). Like the inter-chromosomal network, the differences between the levels are largely restricted to relatively small communities. At level 1, the partition shows the tendency for centromeric regions across all chromosomes to colocalize into a single community. Outside of this centromeric community, many chromosomes or chromosome arms tend to form isolated communities. Notably, chromosome VIII and chromosome XII have different community associations for each chromosomal arm. The segregated interactions of the arms of chromosome XII has been previously observed, where the rDNA locus acts as an interaction boundary for the up- and downstream regions of the chromosome [Duan et al., 2010]. In a force-directed representation of this network, community 0 which represents the region downstream of the rDNA locus on chromosome XII appears to be one of the most isolated regions in the genome (Figure 5.14B).

The tendency for whole chromosomes to group into single communities would suggest that the influence of the intra-chromosomal edges overwhelms that of the inter-chromosomal edges. As a corollary, communities that span different chromosomes in this partition may represent robust inter-chromosomal interactions. Other than community 5, which contains the centromeric regions, community 9 shows the highest degree of cross-chromosomal membership. Large portions of chromosomes III, V, and

VIII belong to community 9, as well as small portions of many other chromosomes, suggesting a relatively high degree of inter-chromosomal interaction in these regions.

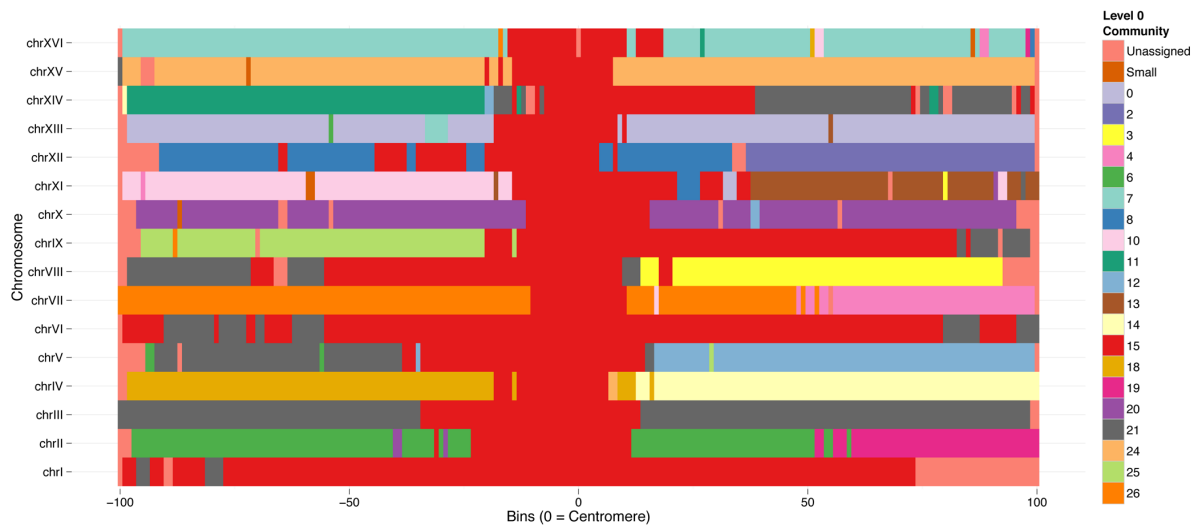


Figure 5.13: Level 0 partition of the complete network. The level 0 partition of the network containing both inter- and intra-chromosomal interactions shows very similar community structure to the level 1 (and final) partition. This indicates that there is little hierarchical community structure information in the intermediate solution to the final partition.

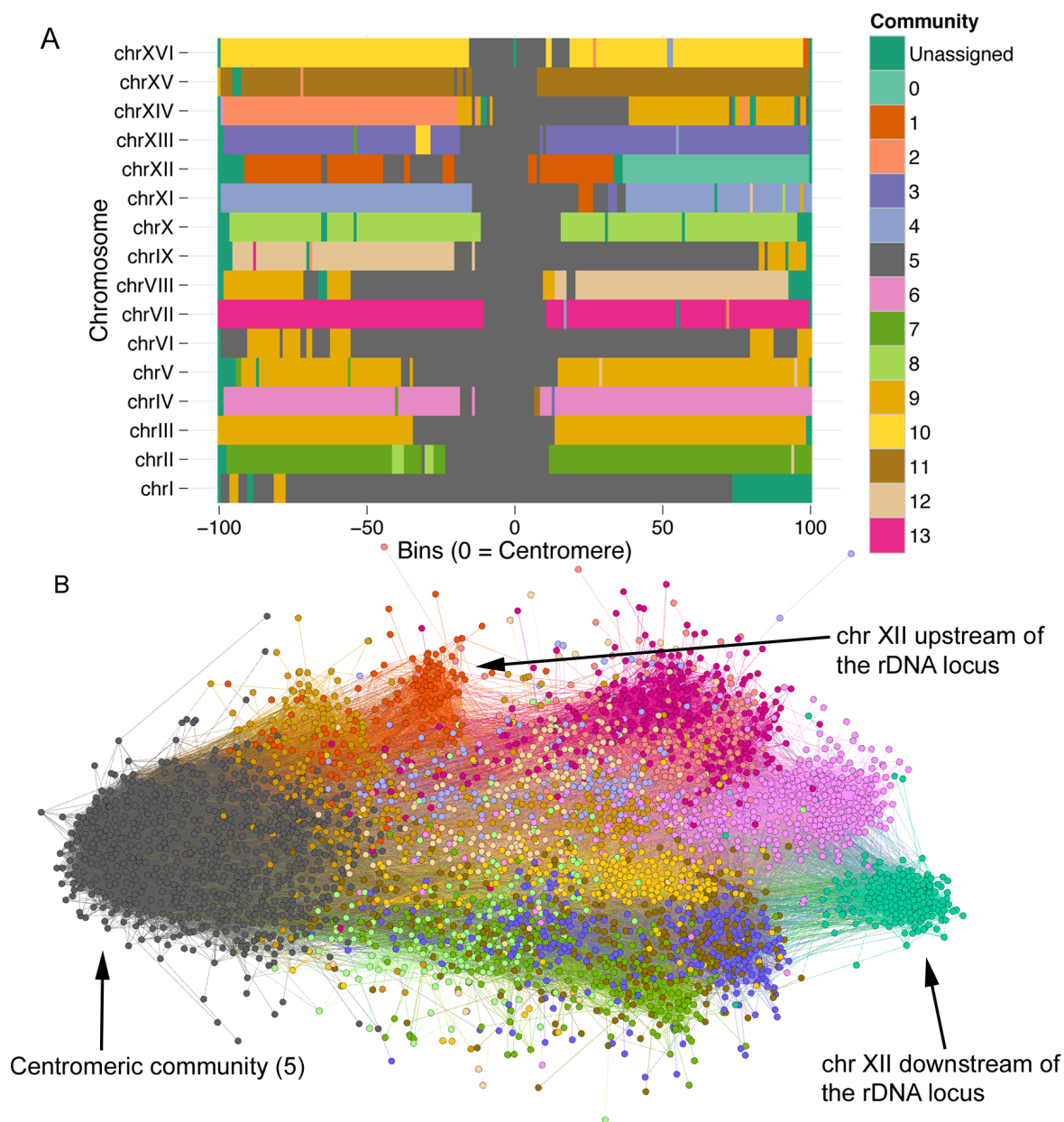


Figure 5.14: **Partition of the complete network.** Final solution to the community partition of the network containing inter- and intra-chromosomal interactions. (A) Scaled chromosomes centered on the centromeres, shows unification of all chromosomes at the centromeres into a single community. Outside of the centromeric community, most chromosomes are broadly belong to a single community. Chromosome XII shows split assignment relative to the rDNA locus (unassigned). (B) A force-directed layout of the complete network shows that each side of chromosome XII is isolated from the other, and they are pushed to the edges of the network.

5.2.7 Conclusion

Here we present a novel approach for detecting spatial groupings from unbiased genomic interaction data. Using this approach we are able to show biologically meaningful spatial associations between genomic elements. Surprisingly, very simple measures of network architecture, such as clique size, are able provide novel insights into the functional organization of the genome. Much of what is presented here is a proof-of-principle, where we have simplified the procedure for constructing the networks. There are many different approaches to network construction. For example, one could apply a method for weighting the edges in the network to improve community detection sensitivity. Furthermore, there are numerous methods to characterize network structure that could be applied to this data, some of which are described in section 1.3.3. Chapter six will describe in greater detail some of the techniques that could be used to build upon this study.

5.3 Methods

5.3.1 Data sources and processing

The chromosome interaction data was generated by Duan et al. [Duan et al., 2010]. The data used to build the networks presented here are from the HindIII fragment interactions that were also confirmed by EcoRI interactions. FDR calculations for these interactions were also taken from Duan et al. Only interactions that achieved the stated FDR thresholds were included in the networks. In addition, only HindIII fragments that met the mappability criteria set out by Duan et al. were included in the networks.

Replication timing data was obtained from McCune et al., Supplemental data 1 [McCune et al., 2008]. In these data, replication timing is represented as a percentage

of a pooled sample of S phase cells for which a locus has replicated. Thus, higher percentages represent earlier replication. The replication percentages for each HindIII fragment were calculated as the mean of the replication percentage that overlapped the given fragment.

ChIP-seq data for cohesin (Smc1, Scc1) and cohesin loader (Scc2, Scc4) subunits were obtained from Hu et al. [Hu et al., 2011]. Raw sequence reads for both ChIP and whole cell extract fractions were mapped to the UCSC sacCer3 genome assembly using Bowtie 2 with default settings [Langmead and Salzberg, 2012]. The number of mapped reads overlapping each HindIII fragment was calculated and assigned to the fragment. The enrichment levels presented in this work were calculated as the \log_2 ratio of ChIP vs. control for each HindIII fragment.

Processed gene expression for ORFs were obtained from Nagalakshmi et al. [Nagalakshmi et al., 2008]. Binding sites for 200 different transcriptional regulators in yeast came from Venters et al. [Venters et al., 2011]. From this data, probe sets that passed 5% FDR significance cutoff were considered binding sites. Only the binding site data at 25C was used in this study. HindIII fragments that intersected (any fraction) one or more binding sites of a given factor were labeled as containing the factor.

Gene annotations were obtained from the “SGD Genes” track of the UCSC Genome Browser database (downloaded February 19, 2013). Centromere, telomere, and tRNA annotations were obtained from the “SGD Other” track of the UCSC Genome Browser database (downloaded February 19, 2013) [Meyer et al., 2013]. Genes were assigned to fragments that contained the given gene’s transcription start site. Like the binding sites of transcriptional regulators, fragments were labeled as containing or not containing centromeres, telomeres or tRNAs, based on a non-zero overlap criteria. All feature intersections were calculated using BEDtools [Quinlan and Hall, 2010].

All coordinate-based datasets that did not correspond to the sacCer3 assembly of the *Saccharomyces cerevisiae* genome were lifted over to sacCer3 using the UCSC

Genome Browser liftOver tool [Meyer et al., 2013].

5.3.2 Network construction and clique/community detection

The networks were built using the NetworkX Python module [Hagberg et al., 2008], where mappable HindIII fragments were represented as nodes, and interactions meeting the FDR threshold were included as edges with weight = 1. The networks presented here represent the largest connected component of the networks induced by the interaction data. All network visualizations were created with the Gephi software [Bastian et al., 2009].

Clique detection was performed using the *find_cliques* function in NetworkX. Each node was assigned a maximum clique size, which is the size of the largest clique to which the node belongs. An in-house implementation of the Louvain algorithm was used to perform community detection [Blondel et al., 2008]. The communities detected at each level of the solution are numbered sequentially from zero, though the numbering is arbitrary. We found that the algorithm often tends to create a small number of very small communities (relative to the size of the communities that make up the vast majority of the networks) at the edges of the networks. These are often chains of nodes connected by single edges, which are not robust communities. Therefore, we chose to filter communities that contained < 10 nodes. The subcommunities were detected by applying a second round of community detection to the subnetwork that represents each community detected in the total network.

5.3.3 Enrichment analyses

Enrichments for protein binding sites and genomic features (centromeres, telomeres, tRNAs) were calculated using the two-tailed Fisher’s exact test. The categories for

the contingency table used to calculate the result of the test were, fragments that contain a given feature, and fragments that belong to a given community. Thus, the test calculates the probability that fragment feature assignment and fragment community assignment are independent. In the case of the transcriptional regulators, the FDR was calculated by applying the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995] to the set of 200 factors for each community.

5.4 Chapter acknowledgements

This chapter was adapted from a manuscript in preparation, authored by Stephen Hoang and Stefan Bekiranov. Marcin Cieřlik, Patrick Grant, and Jeff Smith provided constructive conversations that aided in the completion of this work. In particular, Patrick Grant deserves acknowledgment for initiating a collaboration that was the progenitor for this work.

Chapter 6

Conclusion

Computational analyses of large genomic datasets are highly effective in generating testable hypotheses, and constructing theoretical frameworks for biological reasoning. The insights gained from computational studies can be used to guide experimental studies through direct attempts to validate computationally-derived hypotheses, or by focusing the context of experimental designs. In ideal cases, there is an iterative workflow between computational analyses and experimental data generation. In each round of iteration, computational results and experimental results refine and focus each other to ultimately converge on scientific truth. As demonstrated by the studies presented in this dissertation, the power of high-throughput genomic technologies have placed the field of chromatin biology squarely within the bounds of this mode of scientific discovery.

As publicly available data becomes more comprehensive, increasingly specific and complex biological insights will be made accessible through purely computational means. For scale, the Sequence Read Archive (SRA) [Leinonen et al., 2011] contains over 1.6 petabases of sequence data, over 740 terabases of which are publicly available. The Gene Expression Omnibus (GEO) [Barrett et al., 2013] contains over 32,000 genomic studies. Databases such as these provide the raw material for computationally-

driven theoretical biology. Many of the studies presented by the ENCODE consortium serve as examples of this hypothesis-generation model in chromatin biology [Dunham et al., 2012]. However, computational studies must be complimented by experimental work in order to draw definitive biological conclusions. Even so, the separation of theory—derived from analysis of data—from experimentation—guided by theory—represents an increasingly prominent paradigm in molecular biology, and one that will become evermore prominent as more data is generated. Many of the results presented in this work are manifestations of the theoretical component of this paradigm.

With these notions under consideration, the remainder of this chapter will discuss possible future directions of the studies presented in this dissertation. These extensions include the utilization of publicly available data, experimental design proposals for the generation of new data, and possibilities for new data analysis methodologies.

6.1 Learning from large histone modification datasets

Chapters two and three explore the use of machine learning techniques to understand the biological activity of histone modifications and variants. Studies of this kind are particularly well suited to publicly available data. Specifically, the large number of experimental observations and high dimensionality that is required for these types of studies are often difficult to generate outside of large, well-funded efforts. Furthermore, when studying general properties of histone modifications, publicly available datasets provide a large pool of test data to assess learned models.

One of the most extensive and well-studied histone modification/variant ChIP-seq datasets was generated by Keji Zhao’s lab at the NIH [Barski et al., 2007, Wang et al., 2008]. This dataset consists of genome wide maps of 38 different histone modifications, and variant H2A.Z in human CD4⁺ T cells. Indeed, a portion of

this dataset [Barski et al., 2007] is the basis of chapters two and three. Many other machine learning studies have been performed on this dataset, including a regression study somewhat similar to the one presented in Chapter two [Karlić et al., 2010]. At least two studies have used these data to segment the genome into chromatin states based on combinatorial patterns of histone modifications [Ernst et al., 2010, Hon et al., 2009a]. These data have also been used in attempts to learn the dependence relationships between modifications [Yu et al., 2008]. Many other studies have used more simple data mining techniques to discover correlations between the modifications and various genomic features, such as enhancers [Visel et al., 2009], introns, and exons [Schwartz et al., 2009, Huff et al., 2010]. These examples list only a subset of the data analysis-based studies engendered by the Zhao lab data. The diversity and depth of the insights gained from these data is a testament to the power of mining large publicly available chromatin datasets. Although the majority of publicly available datasets do not approach this biological breadth and level of experimental consistency, comparable datasets will become more available over time, especially through efforts like ENCODE [Dunham et al., 2012] and the Roadmap Epigenomics Project [Chadwick, 2012].

The vast majority of histone ChIP-seq datasets, including the Zhao dataset, are generated in a static population of asynchronous cells. As chapters two and three in part illustrate, studies on these datasets have proven useful in elucidating the composition of chromatin domains, functional correlations between modifications and chromatin activity, and the relationship between various genomic features and histone modifications. However, since histone modifications are dynamic, and vary during phenotypic transitions, it would be valuable to study modifications over time courses, captured during dynamic processes. Studying the sequence of chromatin modification events and corresponding functional changes during a phenotype transition would be extremely powerful in gaining insight into the causal roles of histone modifications.

Understanding the mechanistic roles of histone modifications in this way is a top priority in the chromatin field [Henikoff and Shilatifard, 2011]. Furthermore, computational methods have great promise for gaining insight into these questions.

Many studies, including the one presented in Chapter four, evaluate differences in histone modifications between two biological states. However, these studies provide little explanation for how, and in what sequence, the changes that created the differences occurred. Several studies have tracked histone modifications in many sequential biological states, such as stages of hematopoiesis [Abraham et al., 2013], and *C. elegans* development [Gerstein et al., 2010], but studies that track intermediate changes between states are lacking. One recent study tracked several histone modifications (and several transcription factors) during circadian cycles of mouse liver cells at four hour resolution using ChIP-seq [Koike et al., 2012]. Datasets like this can be utilized to learn dependent relationships between histone modifications by utilizing machine learning techniques, like hidden Markov models and dynamic Bayesian networks (see section 1.3.2). Using these methods, one can learn the conditional probability of transitioning from one modification state to another, thus learning dependence between modifications. Furthermore, since this dataset contains transcription factor maps, the relationship between chromatin states and transcription factor binding could also be investigated in a similar way. It should be noted that the relationships learned using these methods may not be general, but specific to a process; e.g., circadian regulation in the case of [Koike et al., 2012]. This illustrates how computational techniques can be a powerful approach for *guiding* experimental investigations by generating specific and testable hypotheses. Moreover, computational techniques can reasonably guide experimental studies of combinatorial histone modification effects that would be otherwise implausible due to the large combinatorial space of histone modification states.

6.2 Refining perspectives on chromatin regulation during EMT

Many hypotheses about chromatin regulation during EMT were generated through the analyses presented in Chapter four. Principle among these is that the maintenance of the gene expression program that defines the mesenchymal phenotype is effectuated by chromatin-mediated transcriptional feedback involving AP-1, NF- κ B, and c-Myc. Several experimental strategies could be employed to validate and refine this model, including perturbation experiments, mapping of additional factors, and generation of time-resolved data.

In this system it has already been demonstrated that the inhibition of NF- κ B prevents EMT from occurring [Kumar et al., 2013]. A useful step toward validating the computational predictions made in Chapter four would be to see if changes in gene expression under NF- κ B-inhibited conditions are most dramatically impacted in the EMT gene clusters found in Chapter four. It would also be useful to see if perturbing AP-1 family members or c-Myc also prevents EMT, and/or has analogous impacts on gene expression. Many AP-1 family members exist, and many are upregulated during EMT, so finding critical targets may be challenging. However, some AP-1 family members are massively upregulated (e.g., *MAF* up 155-fold), which can serve to prioritize targets. Since the ChIP-seq data was gathered at the end points of EMT (full epithelial or mesenchymal states), it would be useful to see if perturbing these factors induces an mesenchymal-to-epithelial transition (MET), after EMT has been induced. Based on the computationally-derived hypothesis, inhibiting NF- κ B or AP-1 family members, or inducing c-Myc activity should result in a less mesenchymal phenotype in EMT-induced cells. Furthermore, since the putative feedback loops involve proteins upstream of the transcription factors, targeting “topologically vulnerable” proteins (from a signaling network perspective), such as cell surface receptors, would also be

an effective strategy for perturbing the feedback.

The hypothesis put forward by the computational results also states that the transcriptional feedback is chromatin-mediated, meaning that coordinated changes in chromatin permit the transcription factors to execute a transcriptional program that is specific to EMT. Thus, targeting the activity of chromatin modifying enzymes is a second mode of experimental perturbation that could be used to validate the computational predictions. In particular, H3K9, H3K27, and H4K8 acetylation are strongly correlated with transcription factor-specific enhancer activation. Thus, attractive targets for perturbation experiments include HATs, such as *PCAF*, *CREBBP*, *EP300*, *TIP60*, and *MYST2*. HDACs could be targeted in a similar way. Indeed, evidence was presented in Chapter four that suggests HDAC2 may be involved in altering the chromatin-mediated transcriptional program by silencing specific enhancers. Validation of this prediction would expand the known role of HDAC2 in EMT, from silencing of the *CDH1* gene, to silencing a wide array of epithelial-specific genes through deactivation of their enhancers.

In order to frame perturbation experiments in the context of the data-driven hypotheses, new ChIP-seq experiments should be done under perturbed conditions. Fortunately, the entire panel of histone modifications does not need to be recapitulated for each experiment. Based on the observed changes and global correlations, a subset of approximately five factors should be sufficient to assess changes in epigenetic programming. Based on the observed reprogramming and the requirement of normal NF- κ B signaling, one interesting question is: Are epigenetic changes in p65 binding sites dependent on normal p65 activity? One possibility is that p65 binds to chromatin and initiates reprogramming of its binding sites. Another possibility is that p65 binding sites are reprogrammed, which then allows p65 to bind. A third possibility is that a subset of final (mesenchymal state) binding sites are initially reprogrammed which leads to self-propagating systemic changes that eventually lead to the genome-wide

reprogramming that we observe. The latter case is most consistent with positive feedback maintenance of NF- κ B activity. These cases could be resolved in a locus-specific manner utilizing ChIP-seq of p65 and histone modifications under normal and perturbation conditions. Analogous experiments could be devised for AP-1 family members and c-Myc.

In addition to the aforementioned transcription factors and histone modifying enzymes, data on certain so-called “pioneer factors” would be potentially informative. Pioneer factors can bind condensed heterochromatin to initiate the establishment of active chromatin [Magnani et al., 2011]. Interestingly, we observed epigenetic silencing of sites associated with the FOXA family of pioneer factors. Consistently, it has been shown that loss of FOXA1/2 is required for EMT in pancreatic cancer [Song et al., 2010]. These results suggest that pioneer factors, generally, may be important in guiding EMT-specific epigenetic reprogramming.

Time course ChIP-seq experiments of any and all of the factors listed above would provide extremely useful data toward understanding the epigenetic reprogramming that occurs during EMT. As previously mentioned, the dataset used in Chapter four only includes endpoints; i.e., the epithelial or mesenchymal states, and nothing in between. Thus, hypotheses about changes that occur *during* EMT are inferences based on observing the differential of the endpoints. Using time course data, the sequence of events that occur on chromatin could be observed directly. These observations would facilitate prediction of causal relationships, and would greatly improve the potential for mechanistic predictions.

These examples of possible experiments demonstrate how the analysis of one large and general dataset can be used to guide new, and more focused experiments. Ideally, this strategy of simultaneously focusing analysis and experimentation would narrow in on deeper and deeper insights into epigenetic reprogramming in EMT.

6.3 Future approaches for network analysis of genomic interaction data

One of the most obvious extensions of the work presented in Chapter five is to improve the sensitivity and resolution of the methodology. Even with improved methods it is difficult to determine the performance of a method, due to the dearth of studies of this kind. Thus, much work needs to be done to understand how network properties of Hi-C data correspond to properties of genomic conformation. This would be a largely exploratory exercise, applying network analysis to a variety of datasets to understand the advantages and limitations of network analysis, and to build benchmarks for its application. Such studies could potentially be highly rewarding, as there are abundant existing and forthcoming network analysis methods, which may be able give deep insight into genomic organization. For example, measures of node and edge centrality could easily be applied to this data (see Section 1.3.3); however, the biological meanings of such measures are somewhat less intuitive than the identification of interaction clusters presented in Chapter five. Learning how these measures can be interpreted in a biological context would be valuable in advancing this analysis methodology. Like most studies in computational biology, the computational method will improve iteratively with experimental data generated.

One of the principle limitations of the scheme that was applied in Chapter five was the use of uniformly weighted edges. Edge weights give information about strength of interaction between nodes. In genomic interaction data, this could reflect be the significance of the measured interaction between genomic regions, or the contact frequency between regions. Using the contact frequency alone would eliminate the need for setting significance thresholds for interactions. This is desirable because, in principle, all interaction data could be used to build the network, and no assumptions would have to be made to construct a probability model of the interaction frequencies.

Furthermore, this approach could also improve the resolution of the community detection procedure.

A general observation of the results in Chapter five is that the communities detected in the interaction networks are relatively large—size on the order of a chromosome arm for the complete network (see Section 5.2.6). The conclusion of the chapter alluded to methods of network construction that could improve community detection resolution. One such scheme, proposed by Khadivi et al., involves applying edge weights that accentuate community structure on the basis of network topology alone [Khadivi et al., 2011]. The basic principle is to upweight edges that are within communities and downweight edges that are between communities. This is achieved using two network measures known as edge betweenness centrality (EBC) and the common neighbor ratio (CNR). Edge betweenness centrality is analogous to node betweenness centrality defined by equation 1.6, and is defined as

$$B(e_{ij}) = \sum_{u \neq v \in V} \frac{\sigma_{uv}(e_{ij})}{\sigma_{uv}} \quad (6.1)$$

where $\sigma_{uv}(e_{ij})$ is the number of shortest paths that connect nodes u and v that include edge e_{ij} , and σ_{uv} is the total number of shortest paths that connect nodes u and v . Intuitively this value will be large for edges that connect different communities, since shortest paths that connect nodes from the distinct communities will often pass through these edges. Similar to the way EBC highlights between community edges, the CNR highlights within community edges. Given the adjacency matrix A of a network, the CNR of two nodes i and j is defined by

$$C_{ij} = \frac{2(A_{ij} + \sum_k A_{ik}A_{jk})}{\sum_k A_{ik} + \sum_k A_{jk}} \quad (6.2)$$

This is the ratio of neighbors shared by nodes i and j to the total number of neighbors belonging to i and j . By definition, nodes within a single community will have a high

CNR compared to nodes from different communities. The EBC and CNR can be combined to produce the following edge weighting scheme that emphasizes community structure:

$$W(e_{ij}) = \frac{b_{ij}^{-\alpha} \times C_{ij}^{\beta}}{\sum_{\substack{uv \\ u \neq v}} b_{uv}^{-\alpha} \times C_{uv}^{\beta}} \quad \text{for } e_{ij} \in E \text{ and } \alpha, \beta > 0 \quad (6.3)$$

where $W(e_{ij})$ is the weight of edge e_{ij} , b_{ij} is equivalent to $B(e_{ij})/\max_{uv}\{B(e_{uv})\}$. The constants α and β can be selected heuristically [Khadivi et al., 2011]. This weighting scheme expands the bounds on the well-known community detection resolution limit for methods based on modularity optimization [Fortunato and Barthélemy, 2007, Khadivi et al., 2011]. This technique could be applied to studies similar to the one presented in Chapter five in order to resolve finer communities. In principle, one could also devise a weighting scheme that is a composite of network topology and, for example, interaction frequency. Even with these improvements to network construction, there is an inherent resolution limit for community detection, imposed by the sensitivity of the assay and biological variability.

It is possible to design experiments that evaluate the sensitivity of community detection applied to Hi-C data—at least at a coarse-grained level. In budding yeast, it has been shown that certain mutations in cohesin cause defects in the formation of certain higher-order chromatin structures, such as the nucleolus and tRNA clusters [Gard et al., 2009]. This study used light microscopy to assess these defects. In principle, if coupled with Hi-C, the results of microscopy-based analyses could be compared with community detection analyses. The consistency between observations through microscopy and community detection results would provide a relatively low-resolution benchmark for validating and evaluating various community detection techniques. Resolution could potentially be improved by coupling microscopy with site-specific chromosome conformation capture techniques like 3C and 4C. The resolution of the benchmark technique also provides a point of reference for assessing the resolution

of community detection methods. For example, if communities are detected below the resolution limit of the benchmark technique, different approaches could be used to validate the existence of a corresponding structure. This is yet another scenario where computational methods and experimental methods inform each other, resulting in mutual improvement.

In addition to benchmarking, and exploring variations on network construction, an obvious next step is to apply these methods to the genomes of multicellular organisms, which have a higher degree of organizational complexity. Unlike the yeast genome, many of these genomes have fractal globule conformations [Lieberman-Aiden et al., 2009, Sexton et al., 2012], and have specific domains of association [Nora et al., 2012, Dixon et al., 2012]. These structures naturally form interaction communities, making community detection algorithms a potentially powerful tool for studying the spatial organization of these genomes. Somewhat counterintuitively, the higher complexity of these genomes may make the interpretation of community detection results more straightforward. Furthermore, there is a relatively large number of studies that have performed three-dimensional analyses on these genomes, which provides a rich set of results to which network-based results can be compared.

Network analysis of genomic interaction data holds tremendous promise since the structural analysis of networks, and the three-dimensional analysis of genomes are both highly active fields of research (See [Newman, 2011] for a review of network structure analysis, and [Dekker et al., 2013] for a review of genomic conformation analysis). Much of the interest and development in the network field is driven by the accessibility of large datasets that can be coerced into network structures. Genomic interaction datasets are an excellent example of such data, though network analysis has not been broadly applied to them. The continuing improvements in genome-wide interaction assays, coupled with the active network analysis field, presents an opportunity for synergy between data acquisition technology and analysis methodology.

References

- [Abraham et al., 2013] Abraham, B. J., Cui, K., Tang, Q., and Zhao, K. (2013). Dynamic regulation of epigenomic landscapes during hematopoiesis. *BMC genomics*, 14(1):193.
- [Alberts et al., 2002] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *Molecular Biology of the Cell*. Garland Science, New York, NY, USA, 4th edition.
- [Allfrey et al., 1964] Allfrey, V. G., Faulkner, R., and Mirsky, A. E. (1964). Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 51:786–94.
- [Anders and Huber, 2010] Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106.
- [Anthonisse, 1971] Anthonisse, J. M. (1971). The rush in a directed graph. Technical report, Stichting Mathematisch Centrum, Amsterdam.
- [Arnold and Robertson, 2009] Arnold, S. J. and Robertson, E. J. (2009). Making a commitment: cell lineage allocation and axis patterning in the early mouse embryo. *Nature reviews. Molecular cell biology*, 10(2):91–103.
- [Arrowsmith et al., 2012] Arrowsmith, C. H., Bountra, C., Fish, P. V., Lee, K., and Schapira, M. (2012). Epigenetic protein families: a new frontier for drug discovery. *Nature reviews. Drug discovery*, 11(5):384–400.
- [Arvey et al., 2012] Arvey, A., Agius, P., Noble, W. S., and Leslie, C. (2012). Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome research*, 22(9):1723–34.
- [Avraham and Yarden, 2011] Avraham, R. and Yarden, Y. (2011). Feedback regulation of EGFR signalling: decision making by early and delayed loops. *Nature reviews. Molecular cell biology*, 12(2):104–17.
- [Bannister and Kouzarides, 2011] Bannister, A. J. and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell research*, 21(3):381–95.

- [Bannister et al., 2002] Bannister, A. J., Schneider, R., and Kouzarides, T. (2002). Histone methylation: dynamic or static? *Cell*, 109(7):801–6.
- [Bannister et al., 2005] Bannister, A. J., Schneider, R., Myers, F. A., Thorne, A. W., Crane-Robinson, C., and Kouzarides, T. (2005). Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. *The Journal of biological chemistry*, 280(18):17732–6.
- [Barrat et al., 2004] Barrat, A., Barthélemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–52.
- [Barrett et al., 2013] Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., and Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets—update. *Nucleic acids research*, 41(Database issue):D991–5.
- [Barski et al., 2007] Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–37.
- [Bastian et al., 2009] Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. In *International AAAI Conference on Weblogs and Social Media*. AAAI.
- [Ben-Porath et al., 2008] Ben-Porath, I., Thomson, M. W., Carey, V. J., Ge, R., Bell, G. W., Regev, A., and Weinberg, R. A. (2008). An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. *Nature genetics*, 40(5):499–507.
- [Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, 57(1):289–300.
- [Berger, 2007] Berger, S. L. (2007). The complex language of chromatin regulation during transcription. *Nature*, 447(7143):407–12.
- [Berger et al., 2009] Berger, S. L., Kouzarides, T., Shiekhata, R., and Shilatifard, A. (2009). An operational definition of epigenetics. *Genes & development*, 23(7):781–3.
- [Bernstein et al., 2005] Bernstein, B. E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D. K., Huebert, D. J., McMahon, S., Karlsson, E. K., Kulbokas, E. J., Gingeras, T. R., Schreiber, S. L., and Lander, E. S. (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, 120(2):169–81.

- [Bernstein et al., 2006] Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S. L., and Lander, E. S. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2):315–26.
- [Bhola et al., 2013] Bhola, N. E., Balko, J. M., Dugger, T. C., Kuba, M. G., Sánchez, V., Sanders, M., Stanford, J., Cook, R. S., and Arteaga, C. L. (2013). TGF- β inhibition enhances chemotherapy action against triple-negative breast cancer. *The Journal of clinical investigation*, 123(3):1348–58.
- [Bird, 2002] Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes & development*, 16(1):6–21.
- [Blondel et al., 2008] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.
- [Bolós et al., 2003] Bolós, V., Peinado, H., Pérez-Moreno, M. A., Fraga, M. F., Esteller, M., and Cano, A. (2003). The transcription factor Slug represses E-cadherin expression and induces epithelial to mesenchymal transitions: a comparison with Snail and E47 repressors. *Journal of cell science*, 116(Pt 3):499–511.
- [Bolzer et al., 2005] Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Müller, S., Eils, R., Cremer, C., Speicher, M. R., and Cremer, T. (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS biology*, 3(5):e157.
- [Borthwick et al., 2012] Borthwick, L. A., Gardner, A., De Soyza, A., Mann, D. A., and Fisher, A. J. (2012). Transforming Growth Factor- β 1 (TGF- β 1) Driven Epithelial to Mesenchymal Transition (EMT) is Accentuated by Tumour Necrosis Factor α (TNF α) via Crosstalk Between the SMAD and NF- κ B Pathways. *Cancer microenvironment : official journal of the International Cancer Microenvironment Society*, 5(1):45–57.
- [Boyle et al., 2008] Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., and Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–22.
- [Brandes et al., 2008] Brandes, U., Delleng, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., and Wagner, D. (2008). On Modularity Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20(2):172–188.
- [Brenner et al., 2000] Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S. R., Moon, K., Burcham, T., Pallas, M., DuBridge, R. B., Kirchner, J., Fearon, K., Mao, J., and Corcoran,

- K. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature biotechnology*, 18(6):630–4.
- [Buck et al., 2005] Buck, M. J., Nobel, A. B., and Lieb, J. D. (2005). ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. *Genome biology*, 6(11):R97.
- [Burtscher and Lickert, 2009] Burtscher, I. and Lickert, H. (2009). Foxa2 regulates polarity and epithelialization in the endoderm germ layer of the mouse embryo. *Development (Cambridge, England)*, 136(6):1029–38.
- [Butler et al., 2012] Butler, J. S., Koutelou, E., Schibler, A. C., and Dent, S. Y. R. (2012). Histone-modifying enzymes: regulators of developmental decisions and drivers of human disease. *Epigenomics*, 4(2):163–77.
- [Byers et al., 2013] Byers, L. A., Diao, L., Wang, J., Saintigny, P., Girard, L., Peyton, M., Shen, L., Fan, Y., Giri, U., Tumula, P. K., Nilsson, M. B., Gudikote, J., Tran, H., Cardnell, R. J. G., Bearss, D. J., Warner, S. L., Foulks, J. M., Kanner, S. B., Gandhi, V., Krett, N., Rosen, S. T., Kim, E. S., Herbst, R. S., Blumenstein, G. R., Lee, J. J., Lippman, S. M., Ang, K. K., Mills, G. B., Hong, W. K., Weinstein, J. N., Wistuba, I. I., Coombes, K. R., Minna, J. D., and Heymach, J. V. (2013). An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 19(1):279–90.
- [Bystricky et al., 2004] Bystricky, K., Heun, P., Gehlen, L., Langowski, J., and Gasser, S. M. (2004). Long-range compaction and flexibility of interphase chromatin in budding yeast analyzed by high-resolution imaging techniques. *Proceedings of the National Academy of Sciences of the United States of America*, 101(47):16495–500.
- [Campos and Reinberg, 2009] Campos, E. I. and Reinberg, D. (2009). Histones: annotating chromatin. *Annual review of genetics*, 43:559–99.
- [Carter et al., 2002] Carter, D., Chakalova, L., Osborne, C. S., Dai, Y.-f., and Fraser, P. (2002). Long-range chromatin regulatory interactions in vivo. *Nature genetics*, 32(4):623–6.
- [Chadwick, 2012] Chadwick, L. H. (2012). The NIH Roadmap Epigenomics Program data resource. *Epigenomics*, 4(3):317–24.
- [Chandy et al., 2006] Chandy, M., Gutiérrez, J. L., Prochasson, P., and Workman, J. L. (2006). SWI/SNF displaces SAGA-acetylated nucleosomes. *Eukaryotic cell*, 5(10):1738–47.
- [Chang et al., 2007] Chang, B., Chen, Y., Zhao, Y., and Bruick, R. K. (2007). JMJD6 is a histone arginine demethylase. *Science (New York, N.Y.)*, 318(5849):444–7.

- [Charafe-Jauffret et al., 2006] Charafe-Jauffret, E., Ginestier, C., Monville, F., Finetti, P., Adélaïde, J., Cervera, N., Fekairi, S., Xerri, L., Jacquemier, J., Birnbaum, D., and Bertucci, F. (2006). Gene expression profiling of breast cell lines identifies potential new basal markers. *Oncogene*, 25(15):2273–84.
- [Cheng et al., 2011] Cheng, C., Yan, K.-K., Yip, K. Y., Rozowsky, J., Alexander, R., Shou, C., and Gerstein, M. (2011). A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome biology*, 12(2):R15.
- [Chepelev et al., 2012] Chepelev, I., Wei, G., Wangsa, D., Tang, Q., and Zhao, K. (2012). Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell research*, 22(3):490–503.
- [Chickering, 1995] Chickering, D. M. (1995). A transformational characterization of equivalent Bayesian network structures. In Bensard, P. and Hanks, S., editors, *UAI’95 Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 87–98, San Mateo, CA. Morgan Kaufmann Publishers Inc.
- [Clauset et al., 2004] Clauset, A., Newman, M., and Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70(6):066111.
- [Cremer et al., 1982] Cremer, T., Cremer, C., Baumann, H., Luedtke, E. K., Sperling, K., Teuber, V., and Zorn, C. (1982). Rabl’s model of the interphase chromosome arrangement tested in Chinese hamster cells by premature chromosome condensation and laser-UV-microbeam experiments. *Human Genetics*, 60(1):46–56.
- [Cremer and Cremer, 2010] Cremer, T. and Cremer, M. (2010). Chromosome territories. *Cold Spring Harbor perspectives in biology*, 2(3):a003889.
- [Creyghton et al., 2010] Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A., and Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*, 107(50):21931–6.
- [Cuthbert et al., 2004] Cuthbert, G. L., Daujat, S., Snowden, A. W., Erdjument-Bromage, H., Hagiwara, T., Yamada, M., Schneider, R., Gregory, P. D., Tempst, P., Bannister, A. J., and Kouzarides, T. (2004). Histone deimination antagonizes arginine methylation. *Cell*, 118(5):545–53.
- [Dekker et al., 2013] Dekker, J., Marti-Renom, M. A., and Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature reviews. Genetics*, 14(6):390–403.

- [Dekker et al., 2002] Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *Science (New York, N.Y.)*, 295(5558):1306–11.
- [Di Stefano et al., 2013] Di Stefano, M., Rosa, A., Belcastro, V., di Bernardo, D., and Micheletti, C. (2013). Colocalization of Coregulated Genes: A Steered Molecular Dynamics Study of Human Chromosome 19. *PLoS Computational Biology*, 9(3):e1003019.
- [Dixon et al., 2012] Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–80.
- [Dorier and Stasiak, 2010] Dorier, J. and Stasiak, A. (2010). The role of transcription factories-mediated interchromosomal contacts in the organization of nuclear architecture. *Nucleic acids research*, 38(21):7410–21.
- [Dostie and Dekker, 2007] Dostie, J. and Dekker, J. (2007). Mapping networks of physical interactions between genomic elements using 5C technology. *Nature protocols*, 2(4):988–1002.
- [Duan et al., 2010] Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y. J., Lee, C., Shendure, J., Fields, S., Blau, C. A., and Noble, W. S. (2010). A three-dimensional model of the yeast genome. *Nature*, 465(7296):363–7.
- [Dumont et al., 2008] Dumont, N., Wilson, M. B., Crawford, Y. G., Reynolds, P. A., Sigaroudinia, M., and Tlsty, T. D. (2008). Sustained induction of epithelial to mesenchymal transition activates DNA methylation of genes silenced in basal-like breast cancers. *Proceedings of the National Academy of Sciences of the United States of America*, 105(39):14867–72.
- [Dunham et al., 2012] Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B.-K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., Shores, N., Simon, J. M., Song, L., Trinklein, N. D., Altschuler, R. C., Birney, E., Brown, J. B., Cheng, C., Djebali, S., Dong, X., Ernst, J., Furey, T. S., Gerstein, M., Giardine, B., Greven, M., Hardison, R. C., Harris, R. S., Herrero, J., Hoffman, M. M., Iyer, S., Kellis, M., Kheradpour, P., Lassmann, T., Li, Q., Lin, X., Marinov, G. K., Merkel, A., Mortazavi, A., Parker, S. C. J., Reddy, T. E., Rozowsky, J., Schlesinger, F., Thurman, R. E., Wang, J., Ward, L. D., Whitfield, T. W., Wilder, S. P., Wu, W., Xi, H. S., Yip, K. Y., Zhuang, J., Bernstein, B. E., Green, E. D., Gunter, C., Snyder, M., Pazin, M. J., Lowdon, R. F., Dillon, L. A. L., Adams, L. B., Kelly, C. J., Zhang, J., Wexler, J. R., Good, P. J., Feingold, E. A., Crawford, G. E., Dekker, J., Elinitzki, L., Farnham, P. J., Giddings, M. C., Gingeras, T. R., Guigó, R., Hubbard, T. J.,

Kellis, M., Kent, W. J., Lieb, J. D., Margulies, E. H., Myers, R. M., Starnatoyannopoulos, J. A., Tennebaum, S. A., Weng, Z., White, K. P., Wold, B., Yu, Y., Wrobel, J., Risk, B. A., Gunawardena, H. P., Kuiper, H. C., Maier, C. W., Xie, L., Chen, X., Mikkelsen, T. S., Gillespie, S., Goren, A., Ram, O., Zhang, X., Wang, L., Issner, R., Coyne, M. J., Durham, T., Ku, M., Truong, T., Eaton, M. L., Dobin, A., Tanzer, A., Lagarde, J., Lin, W., Xue, C., Williams, B. A., Zaleski, C., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Batut, P., Bell, I., Bell, K., Chakraborty, S., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Li, G., Luo, O. J., Park, E., Preall, J. B., Presaud, K., Ribeca, P., Robyr, D., Ruan, X., Sammeth, M., Sandu, K. S., Schaeffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Hayashizaki, Y., Hubbard, T. J., Reymond, A., Antonarakis, S. E., Hannon, G. J., Ruan, Y., Carninci, P., Sloan, C. A., Learned, K., Malladi, V. S., Wong, M. C., Barber, G. P., Cline, M. S., Dreszer, T. R., Heitner, S. G., Karolchik, D., Kirkup, V. M., Meyer, L. R., Long, J. C., Maddren, M., Raney, B. J., Grasfeder, L. L., Giresi, P. G., Lee, B.-K., Battenhouse, A., Sheffield, N. C., Showers, K. A., London, D., Bhinge, A. A., Shestak, C., Schaner, M. R., Kim, S. K., Zhang, Z. Z., Mieczkowski, P. A., Mieczkowska, J. O., Liu, Z., McDaniell, R. M., Ni, Y., Rashid, N. U., Kim, M. J., Adar, S., Zhang, Z., Wang, T., Winter, D., Keefe, D., Iyer, V. R., Sandhu, K. S., Zheng, M., Wang, P., Gertz, J., Vielmetter, J., Partridge, E. C., Varley, K. E., Gasper, C., Bansal, A., Pepke, S., Jain, P., Amrhein, H., Bowling, K. M., Anaya, M., Cross, M. K., Muratet, M. A., Newberry, K. M., McCue, K., Nesmith, A. S., Fisher-Aylor, K. I., Pusey, B., DeSalvo, G., Parker, S. L., Balasubramanian, S., Davis, N. S., Meadows, S. K., Eggleston, T., Newberry, J. S., Levy, S. E., Absher, D. M., Wong, W. H., Blow, M. J., Visel, A., Pennachio, L. A., Elnitski, L., Petrykowska, H. M., Abyzov, A., Aken, B., Barrell, D., Barson, G., Berry, A., Bignell, A., Boychenko, V., Bussotti, G., Davidson, C., Despacio-Reyes, G., Diekhans, M., Ezkurdia, I., Frankish, A., Gilbert, J., Gonzalez, J. M., Griffiths, E., Harte, R., Hendrix, D. A., Hunt, T., Jungreis, I., Kay, M., Khurana, E., Leng, J., Lin, M. F., Loveland, J., Lu, Z., Manthravadi, D., Mariotti, M., Mudge, J., Mukherjee, G., Notredame, C., Pei, B., Rodriguez, J. M., Saunders, G., Sboner, A., Searle, S., Sisui, C., Snow, C., Steward, C., Tapanari, E., Tress, M. L., van Baren, M. J., Washietl, S., Wilmington, L., Zadissa, A., Zhengdong, Z., Brent, M., Haussler, D., Valencia, A., Raymond, A., Addleman, N., Alexander, R. P., Auerbach, R. K., Balasubramanian, S., Bettinger, K., Bhardwaj, N., Boyle, A. P., Cao, A. R., Cayting, P., Charos, A., Cheng, Y., Eastman, C., Euskirchen, G., Fleming, J. D., Grubert, F., Habegger, L., Hariharan, M., Harmanci, A., Iyenger, S., Jin, V. X., Karczewski, K. J., Kasowski, M., Lacroute, P., Lam, H., Larnarre-Vincent, N., Lian, J., Lindahl-Allen, M., Min, R., Miotto, B., Monahan, H., Moqtaderi, Z., Mu, X. J., O'Geen, H., Ouyang, Z., Patacsil, D., Raha, D., Ramirez, L., Reed, B., Shi, M., Slifer, T., Witt, H., Wu, L., Xu, X., Yan, K.-K., Yang, X., Zhang, Z., Struhl, K., Weiss-

- man, S. M., Tenebaum, S. A., Penalva, L. O., Karmakar, S., Bhanvadia, R. R., Choudhury, A., Domanus, M., Ma, L., Moran, J., Victorsen, A., Auer, T., Centarin, L., Eichenlaub, M., Gruhl, F., Heerman, S., Hoeckendorf, B., Inoue, D., Kellner, T., Kirchmaier, S., Mueller, C., Reinhardt, R., Schertel, L., Schneider, S., Sinn, R., Wittbrodt, B., Wittbrodt, J., Jain, G., Balasundaram, G., Bates, D. L., Byron, R., Canfield, T. K., Diegel, M. J., Dunn, D., Ebersol, A. K., Frum, T., Garg, K., Gist, E., Hansen, R. S., Boatman, L., Haugen, E., Humbert, R., Johnson, A. K., Johnson, E. M., Kutuyavin, T. M., Lee, K., Lotakis, D., Maurano, M. T., Neph, S. J., Neri, F. V., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Rynes, E., Sanchez, M. E., Sandstrom, R. S., Shafer, A. O., Stergachis, A. B., Thomas, S., Vernot, B., Vierstra, J., Vong, S., Wang, H., Weaver, M. A., Yan, Y., Zhang, M., Akey, J. A., Bender, M., Dorschner, M. O., Groudine, M., MacCoss, M. J., Navas, P., Stamatoyannopoulos, G., Stamatoyannopoulos, J. A., Beal, K., Brazma, A., Flicek, P., Johnson, N., Lusk, M., Luscombe, N. M., Sobral, D., Vaquerizas, J. M., Batzoglou, S., Sidow, A., Hussami, N., Kyriazopoulou-Panagiotopoulou, S., Libbrecht, M. W., Schaub, M. A., Miller, W., Bickel, P. J., Banfai, B., Boley, N. P., Huang, H., Li, J. J., Noble, W. S., Bilmes, J. A., Buske, O. J., Sahu, A. O., Kharchenko, P. V., Park, P. J., Baker, D., Taylor, J., and Lachovskiy, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.
- [Eddy, 2013] Eddy, S. R. (2013). The ENCODE project: missteps overshadowing a success. *Current biology : CB*, 23(7):R259–61.
- [Elsheikh et al., 2009] Elsheikh, S. E., Green, A. R., Rakha, E. A., Powe, D. G., Ahmed, R. A., Collins, H. M., Soria, D., Garibaldi, J. M., Paish, C. E., Ammar, A. A., Grainge, M. J., Ball, G. R., Abdelghany, M. K., Martinez-Pomares, L., Heery, D. M., and Ellis, I. O. (2009). Global histone modifications in breast cancer correlate with tumor phenotypes, prognostic factors, and patient outcome. *Cancer research*, 69(9):3802–9.
- [Ernst and Kellis, 2010] Ernst, J. and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature biotechnology*, 28(8):817–25.
- [Ernst and Kellis, 2012] Ernst, J. and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nature methods*, 9(3):215–6.
- [Ernst and Kellis, 2013] Ernst, J. and Kellis, M. (2013). Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. *Genome research*, pages gr.144840.112–.
- [Ernst et al., 2011] Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shores, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., and Bernstein, B. E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–9.

- [Ernst et al., 2010] Ernst, T., Chase, A. J., Score, J., Hidalgo-Curtis, C. E., Bryant, C., Jones, A. V., Waghorn, K., Zoi, K., Ross, F. M., Reiter, A., Hochhaus, A., Drexler, H. G., Duncombe, A., Cervantes, F., Oscier, D., Boultonwood, J., Grand, F. H., and Cross, N. C. P. (2010). Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. *Nature genetics*, 42(8):722–6.
- [Feng et al., 2009] Feng, W., Bachant, J., Collingwood, D., Raghuraman, M. K., and Brewer, B. J. (2009). Centromere replication timing determines different forms of genomic instability in *Saccharomyces cerevisiae* checkpoint mutants during replication stress. *Genetics*, 183(4):1249–60.
- [Fernández and Miranda-Saavedra, 2012] Fernández, M. and Miranda-Saavedra, D. (2012). Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines. *Nucleic acids research*, 40(10):e77.
- [Fischle et al., 2003] Fischle, W., Wang, Y., and Allis, C. D. (2003). Histone and chromatin cross-talk. *Current opinion in cell biology*, 15(2):172–83.
- [Fortunato and Barthélemy, 2007] Fortunato, S. and Barthélemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America*, 104(1):36–41.
- [Foulkes et al., 2010] Foulkes, W. D., Smith, I. E., and Reis-Filho, J. S. (2010). Triple-negative breast cancer. *The New England journal of medicine*, 363(20):1938–48.
- [Francis et al., 2009] Francis, N. J., Follmer, N. E., Simon, M. D., Aghia, G., and Butler, J. D. (2009). Polycomb proteins remain bound to chromatin and DNA during DNA replication in vitro. *Cell*, 137(1):110–22.
- [Freeman, 1977] Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1):35.
- [Friedman, 1991a] Friedman, J. H. (1991a). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1):1–67.
- [Friedman, 1991b] Friedman, J. H. (1991b). Rejoinder: Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1):123–141.
- [Fullwood et al., 2009a] Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., Chew, E. G. Y., Huang, P. Y. H., Welboren, W.-J., Han, Y., Ooi, H. S., Ariyaratne, P. N., Vega, V. B., Luo, Y., Tan, P. Y., Choy, P. Y., Wansa, K. D. S. A., Zhao, B., Lim, K. S., Leow, S. C., Yow, J. S., Joseph, R., Li, H., Desai, K. V., Thomsen, J. S., Lee, Y. K., Karuturi, R. K. M., Herve, T., Bourque, G., Stunnenberg, H. G., Ruan, X., Cacheux-Rataboul, V., Sung, W.-K., Liu, E. T., Wei, C.-L., Cheung, E., and Ruan, Y. (2009a). An oestrogen-receptor- α -bound human chromatin interactome. *Nature*, 462(7269):58–64.

- [Fullwood et al., 2009b] Fullwood, M. J., Wei, C.-L., Liu, E. T., and Ruan, Y. (2009b). Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome research*, 19(4):521–32.
- [Gard et al., 2009] Gard, S., Light, W., Xiong, B., Bose, T., McNairn, A. J., Harris, B., Fleharty, B., Seidel, C., Brickner, J. H., and Gerton, J. L. (2009). Cohesinopathy mutations disrupt the subnuclear organization of chromatin. *The Journal of cell biology*, 187(4):455–62.
- [Gentleman et al., 2004] Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80.
- [Gerstein et al., 2010] Gerstein, M. B., Lu, Z. J., Van Nostrand, E. L., Cheng, C., Arshinoff, B. I., Liu, T., Yip, K. Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., Alves, P., Chateigner, A., Perry, M., Morris, M., Auerbach, R. K., Feng, X., Leng, J., Vielle, A., Niu, W., Rhrissorakrai, K., Agarwal, A., Alexander, R. P., Barber, G., Brdlik, C. M., Brennan, J., Brouillet, J. J., Carr, A., Cheung, M.-S., Clawson, H., Contrino, S., Dannenberg, L. O., Dernburg, A. F., Desai, A., Dick, L., Dosé, A. C., Du, J., Egelhofer, T., Ercan, S., Euskirchen, G., Ewing, B., Feingold, E. A., Gassmann, R., Good, P. J., Green, P., Gullier, F., Gutwein, M., Guyer, M. S., Habegger, L., Han, T., Henikoff, J. G., Henz, S. R., Hinrichs, A., Holster, H., Hyman, T., Iniguez, A. L., Janette, J., Jensen, M., Kato, M., Kent, W. J., Kephart, E., Khivansara, V., Khurana, E., Kim, J. K., Kolasinska-Zwierz, P., Lai, E. C., Latorre, I., Leahey, A., Lewis, S., Lloyd, P., Lochovsky, L., Lowdon, R. F., Lubling, Y., Lyne, R., MacCoss, M., Mackowiak, S. D., Mangone, M., McKay, S., Mecnas, D., Merrihew, G., Miller, D. M., Muroyama, A., Murray, J. I., Ooi, S.-L., Pham, H., Phippen, T., Preston, E. A., Rajewsky, N., Räscher, G., Rosenbaum, H., Rozowsky, J., Rutherford, K., Ruzanov, P., Sarov, M., Sasidharan, R., Sboner, A., Scheid, P., Segal, E., Shin, H., Shou, C., Slack, F. J., Slightam, C., Smith, R., Spencer, W. C., Stinson, E. O., Taing, S., Takasaki, T., Vafeados, D., Voronina, K., Wang, G., Washington, N. L., Whittle, C. M., Wu, B., Yan, K.-K., Zeller, G., Zha, Z., Zhong, M., Zhou, X., Ahringer, J., Strome, S., Gunsalus, K. C., Micklem, G., Liu, X. S., Reinke, V., Kim, S. K., Hillier, L. W., Henikoff, S., Piano, F., Snyder, M., Stein, L., Lieb, J. D., and Waterston, R. H. (2010). Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science (New York, N.Y.)*, 330(6012):1775–87.
- [Gibbons et al., 2005] Gibbons, F. D., Proft, M., Struhl, K., and Roth, F. P. (2005). Chipper: discovering transcription-factor targets from chromatin immunoprecipitation microarrays using variance stabilization. *Genome biology*, 6(11):R96.
- [Gibcus and Dekker, 2013] Gibcus, J. H. and Dekker, J. (2013). The hierarchy of the 3D genome. *Molecular cell*, 49(5):773–82.

- [Gilmore, 2006] Gilmore, T. D. (2006). Introduction to NF-kappaB: players, pathways, perspectives. *Oncogene*, 25(51):6680–4.
- [Graur et al., 2013] Graur, D., Zheng, Y., Price, N., Azevedo, R. B. R., Zufall, R. A., and Elhaik, E. (2013). On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome biology and evolution*, 5(3):578–90.
- [Grigoryev and Woodcock, 2012] Grigoryev, S. A. and Woodcock, C. L. (2012). Chromatin organization - the 30 nm fiber. *Experimental cell research*, 318(12):1448–55.
- [Guillemette et al., 2011] Guillemette, B., Drogaris, P., Lin, H.-H. S., Armstrong, H., Hiragami-Hamada, K., Imhof, A., Bonneil, E., Thibault, P., Verreault, A., and Festenstein, R. J. (2011). H3 lysine 4 is acetylated at active gene promoters and is regulated by H3 lysine 4 methylation. *PLoS genetics*, 7(3):e1001354.
- [Guillou et al., 2010] Guillou, E., Ibarra, A., Coulon, V., Casado-Vela, J., Rico, D., Casal, I., Schwob, E., Losada, A., and Méndez, J. (2010). Cohesin organizes chromatin loops at DNA replication factories. *Genes & development*, 24(24):2812–22.
- [Hadjur et al., 2009] Hadjur, S., Williams, L. M., Ryan, N. K., Cobb, B. S., Sexton, T., Fraser, P., Fisher, A. G., and Merkenschlager, M. (2009). Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature*, 460(7253):410–3.
- [Haeusler et al., 2008] Haeusler, R. A., Pratt-Hyatt, M., Good, P. D., Gipson, T. A., and Engelke, D. R. (2008). Clustering of yeast tRNA genes is mediated by specific association of condensin with tRNA gene transcription complexes. *Genes & development*, 22(16):2204–14.
- [Hagberg et al., 2008] Hagberg, A., Swart, P., and Schult, D. (2008). Exploring network structure, dynamics, and function using NetworkX. In Varoquaux, G., Vaught, T., and Millman, J., editors, *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA.
- [Hall et al., 2002] Hall, I. M., Shankaranarayana, G. D., Noma, K.-I., Ayoub, N., Cohen, A., and Grewal, S. I. S. (2002). Establishment and maintenance of a heterochromatin domain. *Science (New York, N.Y.)*, 297(5590):2232–7.
- [Hammoud et al., 2009] Hammoud, S. S., Nix, D. A., Zhang, H., Purwar, J., Carrell, D. T., and Cairns, B. R. (2009). Distinctive chromatin in human sperm packages genes for embryo development. *Nature*, 460(7254):473–8.
- [Hanahan and Weinberg, 2011] Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5):646–74.

- [Hawkins et al., 2011] Hawkins, R. D., Hon, G. C., Yang, C., Antosiewicz-Bourget, J. E., Lee, L. K., Ngo, Q.-M., Klugman, S., Ching, K. A., Edsall, L. E., Ye, Z., Kuan, S., Yu, P., Liu, H., Zhang, X., Green, R. D., Lobanenko, V. V., Stewart, R., Thomson, J. A., and Ren, B. (2011). Dynamic chromatin states in human ES cells reveal potential regulatory sequences and genes involved in pluripotency. *Cell research*, 21(10):1393–409.
- [Hediger et al., 2002] Hediger, F., Neumann, F. R., Van Houwe, G., Dubrana, K., and Gasser, S. M. (2002). Live Imaging of Telomeres. *Current Biology*, 12(24):2076–2089.
- [Heintzman et al., 2009] Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W., Ching, K. A., Antosiewicz-Bourget, J. E., Liu, H., Zhang, X., Green, R. D., Lobanenko, V. V., Stewart, R., Thomson, J. A., Crawford, G. E., Kellis, M., and Ren, B. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–12.
- [Heintzman et al., 2007] Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E., and Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics*, 39(3):311–8.
- [Henikoff and Shilatifard, 2011] Henikoff, S. and Shilatifard, A. (2011). Histone modification: cause or cog? *Trends in genetics : TIG*, 27(10):389–96.
- [Hinata et al., 2003] Hinata, K., Gervin, A. M., Jennifer Zhang, Y., and Khavari, P. A. (2003). Divergent gene regulation and growth effects by NF-kappa B in epithelial and mesenchymal cells of human skin. *Oncogene*, 22(13):1955–64.
- [Ho et al., 2011] Ho, J. W. K., Bishop, E., Karchenko, P. V., Nègre, N., White, K. P., and Park, P. J. (2011). ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC genomics*, 12:134.
- [Hoang et al., 2011] Hoang, S. A., Xu, X., and Bekiranov, S. (2011). Quantification of histone modification ChIP-seq enrichment for data mining and machine learning applications. *BMC research notes*, 4(1):288.
- [Hoffman et al., 2012] Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., and Noble, W. S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature methods*, 9(5):473–6.
- [Hoffman et al., 2013] Hoffman, M. M., Ernst, J., Wilder, S. P., Kundaje, A., Harris, R. S., Libbrecht, M., Giardine, B., Ellenbogen, P. M., Bilmes, J. A., Birney, E., Hardison, R. C., Dunham, I., Kellis, M., and Noble, W. S. (2013). Integrative annotation of chromatin elements from ENCODE data. *Nucleic acids research*, 41(2):827–41.

- [Hon et al., 2009a] Hon, G., Wang, W., and Ren, B. (2009a). Discovery and annotation of functional chromatin signatures in the human genome. *PLoS computational biology*, 5(11):e1000566.
- [Hon et al., 2009b] Hon, G. C., Hawkins, R. D., and Ren, B. (2009b). Predictive chromatin signatures in the mammalian genome. *Human Molecular Genetics*, 18(R2):R195–R201.
- [Hou et al., 2008] Hou, Z., Peng, H., Ayyanathan, K., Yan, K.-P., Langer, E. M., Longmore, G. D., and Rauscher, F. J. (2008). The LIM protein AJUBA recruits protein arginine methyltransferase 5 to mediate SNAIL-dependent transcriptional repression. *Molecular and cellular biology*, 28(10):3198–207.
- [Hu et al., 2011] Hu, B., Itoh, T., Mishra, A., Katoh, Y., Chan, K.-L., Upcher, W., Godlee, C., Roig, M. B., Shirahige, K., and Nasmyth, K. (2011). ATP hydrolysis is required for relocating cohesin from sites occupied by its Scc2/4 loading complex. *Current biology : CB*, 21(1):12–24.
- [Hubbard et al., 2009] Hubbard, T. J. P., Aken, B. L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K., Jenkinson, A., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Rios, D., Schuster, M., Slater, G., Smedley, D., Spooner, W., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S., Zadissa, A., Birney, E., Cunningham, F., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Kasprzyk, A., Proctor, G., Smith, J., Searle, S., and Flicek, P. (2009). Ensembl 2009. *Nucleic acids research*, 37(Database issue):D690–7.
- [Huff et al., 2010] Huff, J. T., Plocik, A. M., Guthrie, C., and Yamamoto, K. R. (2010). Reciprocal intronic and exonic histone modification regions in humans. *Nature structural & molecular biology*, 17(12):1495–9.
- [Izrailit and Reedijk, 2012] Izrailit, J. and Reedijk, M. (2012). Developmental pathways in breast cancer and breast tumor-initiating cells: therapeutic implications. *Cancer letters*, 317(2):115–26.
- [Jackson et al., 1993] Jackson, D. A., Hassan, A. B., Errington, R. J., and Cook, P. R. (1993). Visualization of focal sites of transcription within human nuclei. *The EMBO journal*, 12(3):1059–65.
- [Jenuwein and Allis, 2001] Jenuwein, T. and Allis, C. D. (2001). Translating the histone code. *Science (New York, N.Y.)*, 293(5532):1074–80.
- [Jin et al., 2011] Jin, F., Li, Y., Ren, B., and Natarajan, R. (2011). PU.1 and C/EBP(alpha) synergistically program distinct response to NF-kappaB activation

- through establishing monocyte specific enhancers. *Proceedings of the National Academy of Sciences of the United States of America*, 108(13):5290–5.
- [Jin et al., 2000] Jin, Q., Fuchs, J., and Loidl, J. (2000). Centromere clustering is a major determinant of yeast interphase nuclear organization. *J. Cell Sci.*, 113(11):1903–1912.
- [John et al., 2011] John, S., Sabo, P. J., Thurman, R. E., Sung, M.-H., Biddie, S. C., Johnson, T. A., Hager, G. L., and Stamatoyannopoulos, J. A. (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature genetics*, 43(3):264–8.
- [Johnson and Dent, 2013] Johnson, D. G. and Dent, S. Y. R. (2013). Chromatin: receiver and quarterback for cellular signals. *Cell*, 152(4):685–9.
- [Johnson et al., 2007] Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, N.Y.)*, 316(5830):1497–502.
- [Kalhor et al., 2012] Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L. (2012). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature biotechnology*, 30(1):90–8.
- [Kalluri and Weinberg, 2009] Kalluri, R. and Weinberg, R. A. (2009). The basics of epithelial-mesenchymal transition. *The Journal of clinical investigation*, 119(6):1420–8.
- [Karlić et al., 2010] Karlić, R., Chung, H.-R., Lasserre, J., Vlahovicek, K., and Vingron, M. (2010). Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 107(7):2926–31.
- [Karpikov et al., 2011] Karpikov, A., Rozowsky, J., and Gerstein, M. (2011). Tiling array data analysis: a multiscale approach using wavelets. *BMC bioinformatics*, 12:57.
- [Kasai et al., 2005] Kasai, H., Allen, J. T., Mason, R. M., Kamimura, T., and Zhang, Z. (2005). TGF-beta1 induces human alveolar epithelial to mesenchymal cell transition (EMT). *Respiratory research*, 6:56.
- [Khadivi et al., 2011] Khadivi, A., Ajdari Rad, A., and Hasler, M. (2011). Network community-detection enhancement by proper weighting. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 83(4 Pt 2):046104.
- [Khan and La Thangue, 2012] Khan, O. and La Thangue, N. B. (2012). HDAC inhibitors in cancer biology: emerging mechanisms and clinical applications. *Immunology and cell biology*, 90(1):85–94.

- [Kharchenko et al., 2011] Kharchenko, P. V., Alekseyenko, A. A., Schwartz, Y. B., Minoda, A., Riddle, N. C., Ernst, J., Sabo, P. J., Larschan, E., Gorchakov, A. A., Gu, T., Linder-Basso, D., Plachetka, A., Shanower, G., Tolstorukov, M. Y., Luquette, L. J., Xi, R., Jung, Y. L., Park, R. W., Bishop, E. P., Canfield, T. K., Sandstrom, R., Thurman, R. E., MacAlpine, D. M., Stamatoyannopoulos, J. A., Kellis, M., Elgin, S. C. R., Kuroda, M. I., Pirrotta, V., Karpen, G. H., and Park, P. J. (2011). Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature*, 471(7339):480–5.
- [Khrameeva et al., 2012] Khrameeva, E. E., Mironov, A. A., Fedonin, G. G., Khaitovich, P., and Gelfand, M. S. (2012). Spatial proximity and similarity of the epigenetic state of genome domains. *PloS one*, 7(4):e33947.
- [Kilpeläinen et al., 2011] Kilpeläinen, T. O., Zillikens, M. C., Stančáková, A., Finucane, F. M., Ried, J. S., Langenberg, C., Zhang, W., Beckmann, J. S., Luan, J., Vandenput, L., Styrkarsdottir, U., Zhou, Y., Smith, A. V., Zhao, J.-H., Amin, N., Vedantam, S., Shin, S.-Y., Haritunians, T., Fu, M., Feitosa, M. F., Kumari, M., Halldorsson, B. V., Tikkanen, E., Mangino, M., Hayward, C., Song, C., Arnold, A. M., Aulchenko, Y. S., Oostra, B. A., Campbell, H., Cupples, L. A., Davis, K. E., Döring, A., Eiriksdottir, G., Estrada, K., Fernández-Real, J. M., Garcia, M., Gieger, C., Glazer, N. L., Guiducci, C., Hofman, A., Humphries, S. E., Iso-maa, B., Jacobs, L. C., Jula, A., Karasik, D., Karlsson, M. K., Khaw, K.-T., Kim, L. J., Kivimäki, M., Klopp, N., Kühnel, B., Kuusisto, J., Liu, Y., Ljunggren, O., Lorentzon, M., Luben, R. N., McKnight, B., Mellström, D., Mitchell, B. D., Mooser, V., Moreno, J. M., Männistö, S., O’Connell, J. R., Pascoe, L., Peltonen, L., Peral, B., Perola, M., Psaty, B. M., Salomaa, V., Savage, D. B., Semple, R. K., Skaric-Juric, T., Sigurdsson, G., Song, K. S., Spector, T. D., Syvänen, A.-C., Talmud, P. J., Thorleifsson, G., Thorsteinsdottir, U., Uitterlinden, A. G., van Duijn, C. M., Vidal-Puig, A., Wild, S. H., Wright, A. F., Clegg, D. J., Schadt, E., Wilson, J. F., Rudan, I., Ripatti, S., Borecki, I. B., Shuldiner, A. R., Ingelsson, E., Jansson, J.-O., Kaplan, R. C., Gudnason, V., Harris, T. B., Groop, L., Kiel, D. P., Rivadeneira, F., Walker, M., Barroso, I., Vollenweider, P., Waeber, G., Chambers, J. C., Kooner, J. S., Soranzo, N., Hirschhorn, J. N., Stefansson, K., Wichmann, H.-E., Ohlsson, C., O’Rahilly, S., Wareham, N. J., Speliotes, E. K., Fox, C. S., Laakso, M., and Loos, R. J. F. (2011). Genetic variation near *IRS1* associates with reduced adiposity and an impaired metabolic profile. *Nature genetics*, 43(8):753–60.
- [Kim et al., 2009] Kim, J., Guermah, M., McGinty, R. K., Lee, J.-S., Tang, Z., Milne, T. A., Shilatifard, A., Muir, T. W., and Roeder, R. G. (2009). RAD6-Mediated transcription-coupled H2B ubiquitylation directly stimulates H3K4 methylation in human cells. *Cell*, 137(3):459–71.
- [Kleer et al., 2003] Kleer, C. G., Cao, Q., Varambally, S., Shen, R., Ota, I., Tomlins, S. A., Ghosh, D., Sewalt, R. G. A. B., Otte, A. P., Hayes, D. F., Sabel, M. S., Livant, D., Weiss, S. J., Rubin, M. A., and Chinnaiyan, A. M. (2003).

- EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. *Proceedings of the National Academy of Sciences of the United States of America*, 100(20):11606–11.
- [Koike et al., 2012] Koike, N., Yoo, S.-H., Huang, H.-C., Kumar, V., Lee, C., Kim, T.-K., and Takahashi, J. S. (2012). Transcriptional architecture and chromatin landscape of the core circadian clock in mammals. *Science (New York, N.Y.)*, 338(6105):349–54.
- [Kolasinska-Zwierz et al., 2009] Kolasinska-Zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X. S., and Ahringer, J. (2009). Differential chromatin marking of introns and expressed exons by H3K36me3. *Nature genetics*, 41(3):376–81.
- [Komarnitsky et al., 2000] Komarnitsky, P., Cho, E. J., and Buratowski, S. (2000). Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. *Genes & development*, 14(19):2452–60.
- [Kouzarides, 2002] Kouzarides, T. (2002). Histone methylation in transcriptional control. *Current opinion in genetics & development*, 12(2):198–209.
- [Kouzarides, 2007] Kouzarides, T. (2007). Chromatin modifications and their function. *Cell*, 128(4):693–705.
- [Kovacs et al., 2009] Kovacs, J. J., Hara, M. R., Davenport, C. L., Kim, J., and Lefkowitz, R. J. (2009). Arrestin development: emerging roles for beta-arrestins in developmental signaling pathways. *Developmental cell*, 17(4):443–58.
- [Krogan et al., 2003] Krogan, N. J., Kim, M., Tong, A., Golshani, A., Cagney, G., Canadien, V., Richards, D. P., Beattie, B. K., Emili, A., Boone, C., Shilatifard, A., Buratowski, S., and Greenblatt, J. (2003). Methylation of histone H3 by Set2 in *Saccharomyces cerevisiae* is linked to transcriptional elongation by RNA polymerase II. *Molecular and cellular biology*, 23(12):4207–18.
- [Ku et al., 2008] Ku, M., Koche, R. P., Rheinbay, E., Mendenhall, E. M., Endoh, M., Mikkelsen, T. S., Presser, A., Nusbaum, C., Xie, X., Chi, A. S., Adli, M., Kasif, S., Ptaszek, L. M., Cowan, C. A., Lander, E. S., Koseki, H., and Bernstein, B. E. (2008). Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS genetics*, 4(10):e1000242.
- [Kuhn et al., 2013] Kuhn, R. M., Haussler, D., and Kent, W. J. (2013). The UCSC genome browser and associated tools. *Briefings in bioinformatics*, 14(2):144–61.
- [Kumar et al., 2013] Kumar, M., Allison, D. F., Baranova, N. N., Wamsley, J. J., Katz, A. J., Bekiranov, S., Jones, D. R., and Mayo, M. W. (2013). NF- κ B Regulates Mesenchymal Transition for the Induction of Non-Small Cell Lung Cancer Initiating Cells. *PLoS ONE*, Accepted.

- [Lancichinetti and Fortunato, 2009] Lancichinetti, A. and Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Physical Review E*, 80(5):056117.
- [Lancichinetti et al., 2008] Lancichinetti, A., Fortunato, S., and Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78(4):046110.
- [Lander et al., 2001] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissole, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kasprzyk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp,

- D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., and Szustakowski, J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.
- [Langfelder and Horvath, 2008] Langfelder, P. and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, 9:559.
- [Langmead and Salzberg, 2012] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–9.
- [Lashkari et al., 1997] Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O., and Davis, R. W. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences*, 94(24):13057–13062.
- [Latham and Dent, 2007] Latham, J. A. and Dent, S. Y. R. (2007). Cross-regulation of histone modifications. *Nature structural & molecular biology*, 14(11):1017–24.
- [Le Guezennec et al., 2006] Le Guezennec, X., Vermeulen, M., Brinkman, A. B., Hoeijmakers, W. A. M., Cohen, A., Lasonder, E., and Stunnenberg, H. G. (2006). MBD2/NuRD and MBD3/NuRD, two distinct complexes with different biochemical and functional properties. *Molecular and cellular biology*, 26(3):843–51.
- [Lee et al., 2007] Lee, J.-S., Shukla, A., Schneider, J., Swanson, S. K., Washburn, M. P., Florens, L., Bhaumik, S. R., and Shilatifard, A. (2007). Histone crosstalk between H2B monoubiquitination and H3 methylation mediated by COMPASS. *Cell*, 131(6):1084–96.
- [Léger-Silvestre et al., 1999] Léger-Silvestre, I., Trumtel, S., Noaillac-Depeyre, J., and Gas, N. (1999). Functional compartmentalization of the nucleus in the budding yeast *Saccharomyces cerevisiae*. *Chromosoma*, 108(2):103–113.
- [Leinonen et al., 2011] Leinonen, R., Sugawara, H., and Shumway, M. (2011). The sequence read archive. *Nucleic acids research*, 39(Database issue):D19–21.
- [Lettice et al., 2002] Lettice, L. A., Horikoshi, T., Heaney, S. J. H., van Baren, M. J., van der Linde, H. C., Breedveld, G. J., Joosse, M., Akarsu, N., Oostra, B. A., Endo, N., Shibata, M., Suzuki, M., Takahashi, E., Shinka, T., Nakahori, Y., Ayusawa, D., Nakabayashi, K., Scherer, S. W., Heutink, P., Hill, R. E., and Noji, S. (2002). Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proceedings of the National Academy of Sciences of the United States of America*, 99(11):7548–53.

- [Li et al., 2007] Li, B., Carey, M., and Workman, J. L. (2007). The role of chromatin during transcription. *Cell*, 128(4):707–19.
- [Li et al., 2012a] Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., Poh, H. M., Goh, Y., Lim, J., Zhang, J., Sim, H. S., Peh, S. Q., Mulawadi, F. H., Ong, C. T., Orlov, Y. L., Hong, S., Zhang, Z., Landt, S., Raha, D., Euskirchen, G., Wei, C.-L., Ge, W., Wang, H., Davis, C., Fisher-Aylor, K. I., Mortazavi, A., Gerstein, M., Gingeras, T., Wold, B., Sun, Y., Fullwood, M. J., Cheung, E., Liu, E., Sung, W.-K., Snyder, M., and Ruan, Y. (2012a). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1-2):84–98.
- [Li and Durbin, 2010] Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 26(5):589–95.
- [Li et al., 2010] Li, S., Chen, Y., Du, H., and Feldman, M. W. (2010). A genetic algorithm with local search strategy for improved detection of community structure. *Complexity*, 15:53 – 60.
- [Li et al., 2012b] Li, Z., Gadue, P., Chen, K., Jiao, Y., Tuteja, G., Schug, J., Li, W., and Kaestner, K. H. (2012b). Foxa2 and H2A.Z mediate nucleosome depletion during embryonic stem cell differentiation. *Cell*, 151(7):1608–16.
- [Liberzon et al., 2011] Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics (Oxford, England)*, 27(12):1739–40.
- [Lieberman-Aiden et al., 2009] Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, 326(5950):289–93.
- [Litt et al., 2009] Litt, M., Qiu, Y., and Huang, S. (2009). Histone arginine methylations: their roles in chromatin dynamics and transcriptional regulation. *Bioscience reports*, 29(2):131–41.
- [Loman et al., 2012] Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., and Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology*, 30(5):434–9.
- [Lombaerts et al., 2006] Lombaerts, M., van Wezel, T., Philippo, K., Dierssen, J. W. F., Zimmerman, R. M. E., Oosting, J., van Eijk, R., Eilers, P. H., van de Water, B., Cornelisse, C. J., and Cleton-Jansen, A.-M. (2006). E-cadherin transcriptional downregulation by promoter methylation but not mutation is related

- to epithelial-to-mesenchymal transition in breast cancer cell lines. *British journal of cancer*, 94(5):661–71.
- [Luger et al., 1997] Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F., and Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389(6648):251–60.
- [Lupien et al., 2008] Lupien, M., Eeckhoutte, J., Meyer, C. A., Wang, Q., Zhang, Y., Li, W., Carroll, J. S., Liu, X. S., and Brown, M. (2008). FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell*, 132(6):958–70.
- [Lv et al., 2010] Lv, J., Qiao, H., Liu, H., Wu, X., Zhu, J., Su, J., Wang, F., Cui, Y., and Zhang, Y. (2010). Discovering cooperative relationships of chromatin modifications in human T cells based on a proposed closeness measure. *PloS one*, 5(12):e14219.
- [Magnani et al., 2011] Magnani, L., Eeckhoutte, J., and Lupien, M. (2011). Pioneer factors: directing transcriptional regulators within the chromatin environment. *Trends in genetics : TIG*, 27(11):465–74.
- [Mani et al., 2008] Mani, S. A., Guo, W., Liao, M.-J., Eaton, E. N., Ayyanan, A., Zhou, A. Y., Brooks, M., Reinhard, F., Zhang, C. C., Shipitsin, M., Campbell, L. L., Polyak, K., Briskin, C., Yang, J., and Weinberg, R. A. (2008). The epithelial-mesenchymal transition generates cells with properties of stem cells. *Cell*, 133(4):704–15.
- [May et al., 2012] May, D., Blow, M. J., Kaplan, T., McCulley, D. J., Jensen, B. C., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Afzal, V., Simpson, P. C., Rubin, E. M., Black, B. L., Bristow, J., Pennacchio, L. A., and Visel, A. (2012). Large-scale discovery of enhancers from human heart tissue. *Nature genetics*, 44(1):89–93.
- [McCarroll and Fangman, 1988] McCarroll, R. M. and Fangman, W. L. (1988). Time of replication of yeast centromeres and telomeres. *Cell*, 54(4):505–13.
- [McCune et al., 2008] McCune, H. J., Danielson, L. S., Alvino, G. M., Collingwood, D., Delrow, J. J., Fangman, W. L., Brewer, B. J., and Raghuraman, M. K. (2008). The temporal program of chromosome replication: genomewide replication in *clb5*{ Δ } *Saccharomyces cerevisiae*. *Genetics*, 180(4):1833–47.
- [McDonald et al., 2011] McDonald, O. G., Wu, H., Timp, W., Doi, A., and Feinberg, A. P. (2011). Genome-scale epigenetic reprogramming during epithelial-to-mesenchymal transition. *Nature structural & molecular biology*, 18(8):867–74.
- [McLean et al., 2010] McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology*, 28(5):495–501.

- [Medus et al., 2005] Medus, A., Acuña, G., and Dorso, C. (2005). Detection of community structures in networks via global optimization. *Physica A: Statistical Mechanics and its Applications*, 358(2-4):593–604.
- [Mehta et al., 2012] Mehta, R. J., Jain, R. K., Leung, S., Choo, J., Nielsen, T., Huntsman, D., Nakshatri, H., and Badve, S. (2012). FOXA1 is an independent prognostic marker for ER-positive breast cancer. *Breast cancer research and treatment*, 131(3):881–90.
- [Mercer et al., 2011] Mercer, E. M., Lin, Y. C., Benner, C., Jhunjhunwala, S., Dutkowski, J., Flores, M., Sigvardsson, M., Ideker, T., Glass, C. K., and Murre, C. (2011). Multilineage priming of enhancer repertoires precedes commitment to the B and myeloid cell lineages in hematopoietic progenitors. *Immunity*, 35(3):413–25.
- [Meyer et al., 2013] Meyer, L. R., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Kuhn, R. M., Wong, M., Sloan, C. A., Rosenbloom, K. R., Roe, G., Rhead, B., Raney, B. J., Pohl, A., Malladi, V. S., Li, C. H., Lee, B. T., Learned, K., Kirkup, V., Hsu, F., Heitner, S., Harte, R. A., Haeussler, M., Guruvadoo, L., Goldman, M., Giardine, B. M., Fujita, P. A., Dreszer, T. R., Diekhans, M., Cline, M. S., Clawson, H., Barber, G. P., Haussler, D., and Kent, W. J. (2013). The UCSC Genome Browser database: extensions and updates 2013. *Nucleic acids research*, 41(Database issue):D64–9.
- [Mikkelsen et al., 2007] Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R. P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E. S., and Bernstein, B. E. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–60.
- [Mishiro et al., 2009] Mishiro, T., Ishihara, K., Hino, S., Tsutsumi, S., Aburatani, H., Shirahige, K., Kinoshita, Y., and Nakao, M. (2009). Architectural roles of multiple chromatin insulators at the human apolipoprotein gene cluster. *The EMBO journal*, 28(9):1234–45.
- [Nagalakshmi et al., 2008] Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, N.Y.)*, 320(5881):1344–9.
- [Nakamura et al., 2007] Nakamura, T., Kuwai, T., Kitadai, Y., Sasaki, T., Fan, D., Coombes, K. R., Kim, S.-J., and Fidler, I. J. (2007). Zonal heterogeneity for gene expression in human pancreatic carcinoma. *Cancer research*, 67(16):7597–604.
- [Nativio et al., 2009] Nativio, R., Wendt, K. S., Ito, Y., Huddleston, J. E., Uribe-Lewis, S., Woodfine, K., Krueger, C., Reik, W., Peters, J.-M., and Murrell, A.

- (2009). Cohesin is required for higher-order chromatin conformation at the imprinted IGF2-H19 locus. *PLoS genetics*, 5(11):e1000739.
- [Newman and Cooper, 2010] Newman, A. M. and Cooper, J. B. (2010). AutoSOME: a clustering method for identifying gene expression modules without prior knowledge of cluster number. *BMC bioinformatics*, 11:117.
- [Newman, 2004] Newman, M. (2004). Analysis of weighted networks. *Physical Review E*, 70(5):056131.
- [Newman, 2011] Newman, M. E. J. (2011). Communities, modules and large-scale structure in networks. *Nature Physics*, 8(1):25–31.
- [Ng et al., 2003] Ng, H. H., Robert, F., Young, R. A., and Struhl, K. (2003). Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Molecular cell*, 11(3):709–19.
- [Nie et al., 2012] Nie, Z., Hu, G., Wei, G., Cui, K., Yamane, A., Resch, W., Wang, R., Green, D. R., Tessarollo, L., Casellas, R., Zhao, K., and Levens, D. (2012). c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell*, 151(1):68–79.
- [Noordermeer et al., 2011] Noordermeer, D., de Wit, E., Klous, P., van de Werken, H., Simonis, M., Lopez-Jones, M., Eussen, B., de Klein, A., Singer, R. H., and de Laat, W. (2011). Variegated gene expression caused by cell-specific long-range DNA interactions. *Nature cell biology*, 13(8):944–51.
- [Nora et al., 2012] Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., and Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485(7398):381–5.
- [O’Neill and Turner, 1995] O’Neill, L. P. and Turner, B. M. (1995). Histone H4 acetylation distinguishes coding regions of the human genome from heterochromatin in a differentiation-dependent but transcription-independent manner. *The EMBO journal*, 14(16):3946–57.
- [O’Neill and Turner, 1996] O’Neill, L. P. and Turner, B. M. (1996). Immunoprecipitation of chromatin. *Methods in enzymology*, 274:189–97.
- [Ong and Corces, 2011] Ong, C.-T. and Corces, V. G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature reviews. Genetics*, 12(4):283–93.
- [Osborne et al., 2004] Osborne, C. S., Chakalova, L., Brown, K. E., Carter, D., Horton, A., Debrand, E., Goyenechea, B., Mitchell, J. A., Lopes, S., Reik, W., and Fraser, P. (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature genetics*, 36(10):1065–71.

- [Pal et al., 2004] Pal, S., Vishwanath, S. N., Erdjument-Bromage, H., Tempst, P., and Sif, S. (2004). Human SWI/SNF-associated PRMT5 methylates histone H3 arginine 8 and negatively regulates expression of ST7 and NM23 tumor suppressor genes. *Molecular and cellular biology*, 24(21):9630–45.
- [Pal et al., 2003] Pal, S., Yun, R., Datta, A., Lacomis, L., Erdjument-Bromage, H., Kumar, J., Tempst, P., and Sif, S. (2003). mSin3A/histone deacetylase 2- and PRMT5-containing Brg1 complex is involved in transcriptional repression of the Myc target gene cad. *Molecular and cellular biology*, 23(21):7475–87.
- [Papantonis et al., 2010] Papantonis, A., Larkin, J. D., Wada, Y., Ohta, Y., Ihara, S., Kodama, T., and Cook, P. R. (2010). Active RNA polymerases: mobile or immobile molecular machines? *PLoS biology*, 8(7):e1000419.
- [Pareek et al., 2011] Pareek, C. S., Smoczynski, R., and Tretyn, A. (2011). Sequencing technologies and genome sequencing. *Journal of applied genetics*, 52(4):413–35.
- [Parthun, 2007] Parthun, M. R. (2007). Hat1: the emerging cellular roles of a type B histone acetyltransferase. *Oncogene*, 26(37):5319–28.
- [Peinado et al., 2007] Peinado, H., Olmeda, D., and Cano, A. (2007). Snail, Zeb and bHLH factors in tumour progression: an alliance against the epithelial phenotype? *Nature reviews. Cancer*, 7(6):415–28.
- [Pickrell et al., 2011] Pickrell, J. K., Gaffney, D. J., Gilad, Y., and Pritchard, J. K. (2011). False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics (Oxford, England)*, 27(15):2144–6.
- [Podlaha et al., 2012] Podlaha, O., Riester, M., De, S., and Michor, F. (2012). Evolution of the cancer genome. *Trends in genetics : TIG*, 28(4):155–63.
- [Polyak and Weinberg, 2009] Polyak, K. and Weinberg, R. A. (2009). Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. *Nature reviews. Cancer*, 9(4):265–73.
- [Pothen et al., 1990] Pothen, A., Simon, H. D., and Liou, K.-P. (1990). Partitioning Sparse Matrices with Eigenvectors of Graphs. *SIAM Journal on Matrix Analysis and Applications*, 11:430–452.
- [Provenzani et al., 2006] Provenzani, A., Fronza, R., Loreni, F., Pascale, A., Amadio, M., and Quattrone, A. (2006). Global alterations in mRNA polysomal recruitment in a cell model of colorectal cancer progression to metastasis. *Carcinogenesis*, 27(7):1323–33.
- [Ptashne, 2013] Ptashne, M. (2013). Epigenetics: Core misconception. *Proceedings of the National Academy of Sciences of the United States of America*, 110(18):7101–3.

- [Quinlan and Hall, 2010] Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6):841–2.
- [Rabl, 1885] Rabl, C. (1885). Über Zelltheilung. *Morphologisches Jahrbuch*, 10:214–330.
- [Rada-Iglesias et al., 2011] Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470(7333):279–83.
- [Rando, 2012] Rando, O. J. (2012). Combinatorial complexity in chromatin structure and function: revisiting the histone code. *Current opinion in genetics & development*, 22(2):148–55.
- [Reichardt and Bornholdt, 2006] Reichardt, J. and Bornholdt, S. (2006). When are networks truly modular? *Physica D: Nonlinear Phenomena*, 224(1-2):20–26.
- [Rieder et al., 2012] Rieder, D., Trajanoski, Z., and McNally, J. G. (2012). Transcription factories. *Frontiers in genetics*, 3:221.
- [Roh et al., 2006] Roh, T.-Y., Cuddapah, S., Cui, K., and Zhao, K. (2006). The genomic landscape of histone modifications in human T cells. *Proceedings of the National Academy of Sciences of the United States of America*, 103(43):15782–7.
- [Ruthenburg et al., 2007] Ruthenburg, A. J., Li, H., Patel, D. J., and Allis, C. D. (2007). Multivalent engagement of chromatin modifications by linked binding modules. *Nature reviews. Molecular cell biology*, 8(12):983–94.
- [Ryba et al., 2010] Ryba, T., Hiratani, I., Lu, J., Itoh, M., Kulik, M., Zhang, J., Schulz, T. C., Robins, A. J., Dalton, S., and Gilbert, D. M. (2010). Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome research*, 20(6):761–70.
- [Sabidussi, 1966] Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, 31(4):581–603.
- [Sandhu et al., 2012] Sandhu, K. S., Li, G., Poh, H. M., Quek, Y. L. K., Sia, Y. Y., Peh, S. Q., Mulawadi, F. H., Lim, J., Sikic, M., Menghi, F., Thalamuthu, A., Sung, W. K., Ruan, X., Fullwood, M. J., Liu, E., Csermely, P., and Ruan, Y. (2012). Large-scale functional organization of long-range chromatin interaction networks. *Cell reports*, 2(5):1207–19.
- [Sanyal et al., 2012] Sanyal, A., Lajoie, B. R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–13.
- [Schena et al., 1995] Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270(5235):467–470.

- [Schneider and Grosschedl, 2007] Schneider, R. and Grosschedl, R. (2007). Dynamics and interplay of nuclear architecture, genome organization, and gene expression. *Genes & development*, 21(23):3027–43.
- [Schneider et al., 2011] Schneider, T. D., Arteaga-Salas, J. M., Mentele, E., David, R., Nicetto, D., Imhof, A., and Rupp, R. A. W. (2011). Stage-specific histone modification profiles reveal global transitions in the *Xenopus* embryonic epigenome. *PloS one*, 6(7):e22548.
- [Schober et al., 2008] Schober, H., Kalck, V., Vega-Palas, M. A., Van Houwe, G., Sage, D., Unser, M., Gartenberg, M. R., and Gasser, S. M. (2008). Controlled exchange of chromosomal arms reveals principles driving telomere interactions in yeast. *Genome research*, 18(2):261–71.
- [Schoenfelder et al., 2010] Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N. F., Horton, A., Andrews, S., Kurukuti, S., Mitchell, J. A., Umlauf, D., Dimitrova, D. S., Eskiw, C. H., Luo, Y., Wei, C.-L., Ruan, Y., Bieker, J. J., and Fraser, P. (2010). Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nature genetics*, 42(1):53–61.
- [Schwartz et al., 2009] Schwartz, S., Meshorer, E., and Ast, G. (2009). Chromatin organization marks exon-intron structure. *Nature structural & molecular biology*, 16(9):990–5.
- [Sekiya et al., 2009] Sekiya, T., Muthurajan, U. M., Luger, K., Tulin, A. V., and Zaret, K. S. (2009). Nucleosome-binding affinity as a primary determinant of the nuclear mobility of the pioneer transcription factor FoxA. *Genes & development*, 23(7):804–9.
- [Sexton et al., 2012] Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell*, 148(3):458–72.
- [Shen et al., 2012] Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenko, V. V., and Ren, B. (2012). A map of the cis-regulatory sequences in the mouse genome. *Nature*, 488(7409):116–20.
- [Sherwood et al., 2010] Sherwood, R., Takahashi, T. S., and Jallepalli, P. V. (2010). Sister acts: coordinating DNA replication and cohesion establishment. *Genes & development*, 24(24):2723–31.
- [Shi et al., 2004] Shi, Y., Lan, F., Matson, C., Mulligan, P., Whetstine, J. R., Cole, P. A., Casero, R. A., and Shi, Y. (2004). Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. *Cell*, 119(7):941–53.

- [Shogren-Knaak et al., 2006] Shogren-Knaak, M., Ishii, H., Sun, J.-M., Pazin, M. J., Davie, J. R., and Peterson, C. L. (2006). Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science (New York, N.Y.)*, 311(5762):844–7.
- [Simonis et al., 2006] Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature genetics*, 38(11):1348–54.
- [Simonis et al., 2007] Simonis, M., Kooren, J., and de Laat, W. (2007). An evaluation of 3C-based methods to capture DNA interactions. *Nature methods*, 4(11):895–901.
- [Sims et al., 2006] Sims, J. K., Houston, S. I., Magazinnik, T., and Rice, J. C. (2006). A trans-tail histone code defined by monomethylated H4 Lys-20 and H3 Lys-9 demarcates distinct regions of silent chromatin. *The Journal of biological chemistry*, 281(18):12760–6.
- [Singh and Settleman, 2010] Singh, A. and Settleman, J. (2010). EMT, cancer stem cells and drug resistance: an emerging axis of evil in the war on cancer. *Oncogene*, 29(34):4741–51.
- [Smyth, 2004] Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3:Article3.
- [Song et al., 2010] Song, Y., Washington, M. K., and Crawford, H. C. (2010). Loss of FOXA1/2 is essential for the epithelial-to-mesenchymal transition in pancreatic cancer. *Cancer research*, 70(5):2115–25.
- [Stadler and Allis, 2012] Stadler, S. C. and Allis, C. D. (2012). Linking epithelial-to-mesenchymal-transition and epigenetic modifications. *Seminars in cancer biology*, 22(5-6):404–10.
- [Strahl and Allis, 2000] Strahl, B. D. and Allis, C. D. (2000). The language of covalent histone modifications. *Nature*, 403(6765):41–5.
- [Su et al., 2004] Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R., and Hogenesch, J. B. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6062–7.
- [Suganuma and Workman, 2008] Suganuma, T. and Workman, J. L. (2008). Crosstalk among Histone Modifications. *Cell*, 135(4):604–7.
- [Taddei et al., 2010] Taddei, A., Schober, H., and Gasser, S. M. (2010). The budding yeast nucleus. *Cold Spring Harbor perspectives in biology*, 2(8):a000612.

- [Tanizawa et al., 2010] Tanizawa, H., Iwasaki, O., Tanaka, A., Capizzi, J. R., Wickramasinghe, P., Lee, M., Fu, Z., and Noma, K.-i. (2010). Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic acids research*, 38(22):8164–77.
- [Teytelman et al., 2009] Teytelman, L., Ozaydin, B., Zill, O., Lefrançois, P., Snyder, M., Rine, J., and Eisen, M. B. (2009). Impact of chromatin structures on DNA processing for genomic analyses. *PLoS one*, 4(8):e6700.
- [Thiery, 2003] Thiery, J. P. (2003). Epithelial-mesenchymal transitions in development and pathologies. *Current opinion in cell biology*, 15(6):740–6.
- [Thiery et al., 2009] Thiery, J. P., Acloque, H., Huang, R. Y. J., and Nieto, M. A. (2009). Epithelial-mesenchymal transitions in development and disease. *Cell*, 139(5):871–90.
- [Thompson et al., 2003] Thompson, M., Haeusler, R. A., Good, P. D., and Engelke, D. R. (2003). Nucleolar clustering of dispersed tRNA genes. *Science (New York, N.Y.)*, 302(5649):1399–401.
- [Thomson et al., 2005] Thomson, S., Buck, E., Petti, F., Griffin, G., Brown, E., Ramnarine, N., Iwata, K. K., Gibson, N., and Haley, J. D. (2005). Epithelial to mesenchymal transition is a determinant of sensitivity of non-small-cell lung carcinoma cell lines and xenografts to epidermal growth factor receptor inhibition. *Cancer research*, 65(20):9455–62.
- [Thomson et al., 2008] Thomson, S., Petti, F., Sujka-Kwok, I., Epstein, D., and Haley, J. D. (2008). Kinase switching in mesenchymal-like non-small cell lung cancer lines contributes to EGFR inhibitor resistance through pathway redundancy. *Clinical & experimental metastasis*, 25(8):843–54.
- [Thomson et al., 2011] Thomson, S., Petti, F., Sujka-Kwok, I., Mercado, P., Bean, J., Monaghan, M., Seymour, S. L., Argast, G. M., Epstein, D. M., and Haley, J. D. (2011). A systems view of epithelial-mesenchymal transition signaling states. *Clinical & experimental metastasis*, 28(2):137–55.
- [Tie et al., 2009] Tie, F., Banerjee, R., Stratton, C. A., Prasad-Sinha, J., Stepanik, V., Zlobin, A., Diaz, M. O., Scacheri, P. C., and Harte, P. J. (2009). CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing. *Development (Cambridge, England)*, 136(18):3131–41.
- [Tittel-Elmer et al., 2012] Tittel-Elmer, M., Lengronne, A., Davidson, M. B., Bacal, J., François, P., Hohl, M., Petrini, J. H. J., Pasero, P., and Cobb, J. A. (2012). Cohesin association to replication sites depends on rad50 and promotes fork restart. *Molecular cell*, 48(1):98–108.
- [Tjeertes et al., 2009] Tjeertes, J. V., Miller, K. M., and Jackson, S. P. (2009). Screen for DNA-damage-responsive histone modifications identifies H3K9Ac and H3K56Ac in human cells. *The EMBO journal*, 28(13):1878–89.

- [Tjong et al., 2012] Tjong, H., Gong, K., Chen, L., and Alber, F. (2012). Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. *Genome research*, 22(7):1295–305.
- [Tolhuis et al., 2002] Tolhuis, B., Palstra, R. J., Splinter, E., Grosveld, F., and de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular cell*, 10(6):1453–65.
- [Tomita et al., 2006] Tomita, E., Tanaka, A., and Takahashi, H. (2006). The worst-case time complexity for generating all maximal cliques and computational experiments. *Theoretical Computer Science*, 363(1):28–42.
- [Tsukada et al., 2006] Tsukada, Y.-i., Fang, J., Erdjument-Bromage, H., Warren, M. E., Borchers, C. H., Tempst, P., and Zhang, Y. (2006). Histone demethylation by a family of JmjC domain-containing proteins. *Nature*, 439(7078):811–6.
- [Uhlmann, 2009] Uhlmann, F. (2009). A matter of choice: the establishment of sister chromatid cohesion. *EMBO reports*, 10(10):1095–102.
- [van Attikum and Gasser, 2009] van Attikum, H. and Gasser, S. M. (2009). Crosstalk between histone modifications during the DNA damage response. *Trends in cell biology*, 19(5):207–17.
- [van Bokhoven and Kramer, 2010] van Bokhoven, H. and Kramer, J. M. (2010). Disruption of the epigenetic code: an emerging mechanism in mental retardation. *Neurobiology of disease*, 39(1):3–12.
- [van Steensel and Dekker, 2010] van Steensel, B. and Dekker, J. (2010). Genomics tools for unraveling chromosome architecture. *Nature biotechnology*, 28(10):1089–1095.
- [Varambally et al., 2002] Varambally, S., Dhanasekaran, S. M., Zhou, M., Barrette, T. R., Kumar-Sinha, C., Sanda, M. G., Ghosh, D., Pienta, K. J., Sewalt, R. G. A. B., Otte, A. P., Rubin, M. A., and Chinnaiyan, A. M. (2002). The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature*, 419(6907):624–9.
- [Venters et al., 2011] Venters, B. J., Wachi, S., Mavrich, T. N., Andersen, B. E., Jena, P., Sinnamon, A. J., Jain, P., Roller, N. S., Jiang, C., Hemeryck-Walsh, C., and Pugh, B. F. (2011). A comprehensive genomic binding map of gene and chromatin regulatory proteins in *Saccharomyces*. *Molecular cell*, 41(4):480–92.
- [Visel et al., 2009] Visel, A., Rubin, E. M., and Pennacchio, L. A. (2009). Genomic views of distant-acting enhancers. *Nature*, 461(7261):199–205.
- [von Burstin et al., 2009] von Burstin, J., Eser, S., Paul, M. C., Seidler, B., Brandl, M., Messer, M., von Werder, A., Schmidt, A., Mages, J., Pagel, P., Schnieke, A.,

- Schmid, R. M., Schneider, G., and Saur, D. (2009). E-cadherin regulates metastasis of pancreatic cancer in vivo and is suppressed by a SNAIL/HDAC1/HDAC2 repressor complex. *Gastroenterology*, 137(1):361–71, 371.e1–5.
- [Wan et al., 2005] Wan, H., Dingle, S., Xu, Y., Besnard, V., Kaestner, K. H., Ang, S.-L., Wert, S., Stahlman, M. T., and Whitsett, J. A. (2005). Compensatory roles of Foxa1 and Foxa2 during lung morphogenesis. *The Journal of biological chemistry*, 280(14):13809–16.
- [Wang et al., 2012] Wang, H., Albadine, R., Magheli, A., Guzzo, T. J., Ball, M. W., Hinz, S., Schoenberg, M. P., Netto, G. J., and Gonzalgo, M. L. (2012). Increased EZH2 protein expression is associated with invasive urothelial carcinoma of the bladder. *Urologic oncology*, 30(4):428–33.
- [Wang et al., 2004] Wang, Y., Wysocka, J., Sayegh, J., Lee, Y.-H., Perlin, J. R., Leonelli, L., Sonbuchner, L. S., McDonald, C. H., Cook, R. G., Dou, Y., Roeder, R. G., Clarke, S., Stallcup, M. R., Allis, C. D., and Coonrod, S. A. (2004). Human PAD4 regulates histone arginine methylation levels via demethylation. *Science (New York, N.Y.)*, 306(5694):279–83.
- [Wang et al., 2009a] Wang, Y., Zhang, X.-S., and Xia, Y. (2009a). Predicting eukaryotic transcriptional cooperativity by Bayesian network integration of genome-wide data. *Nucleic acids research*, 37(18):5943–58.
- [Wang et al., 2013] Wang, Z., Cao, R., Taylor, K., Briley, A., Caldwell, C., and Cheng, J. (2013). The properties of genome conformation and spatial gene interaction and regulation networks of normal and malignant human cell types. *PloS one*, 8(3):e58793.
- [Wang et al., 2009b] Wang, Z., Zang, C., Cui, K., Schones, D. E., Barski, A., Peng, W., and Zhao, K. (2009b). Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell*, 138(5):1019–31.
- [Wang et al., 2008] Wang, Z., Zang, C., Rosenfeld, J. A., Schones, D. E., Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Peng, W., Zhang, M. Q., and Zhao, K. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature genetics*, 40(7):897–903.
- [Whetstine et al., 2006] Whetstine, J. R., Nottke, A., Lan, F., Huarte, M., Smolnikov, S., Chen, Z., Spooner, E., Li, E., Zhang, G., Colaiacovo, M., and Shi, Y. (2006). Reversal of histone lysine trimethylation by the JMJD2 family of histone demethylases. *Cell*, 125(3):467–81.
- [Witherow et al., 2004] Witherow, D. S., Garrison, T. R., Miller, W. E., and Lefkowitz, R. J. (2004). beta-Arrestin inhibits NF-kappaB activity by means of its interaction with the NF-kappaB inhibitor IkappaBalpha. *Proceedings of the National Academy of Sciences of the United States of America*, 101(23):8603–7.

- [Wood et al., 2010] Wood, A. J., Severson, A. F., and Meyer, B. J. (2010). Condensin and cohesin complexity: the expanding repertoire of functions. *Nature reviews. Genetics*, 11(6):391–404.
- [Wu et al., 2012] Wu, C.-Y., Tsai, Y.-P., Wu, M.-Z., Teng, S.-C., and Wu, K.-J. (2012). Epigenetic reprogramming and post-transcriptional regulation during the epithelial-mesenchymal transition. *Trends in genetics : TIG*, 28(9):454–63.
- [Wu and Zhou, 2010] Wu, Y. and Zhou, B. P. (2010). TNF-alpha/NF-kappaB/Snail pathway in cancer cell migration and invasion. *British journal of cancer*, 102(4):639–44.
- [Wu et al., 2004] Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F., and Spencer, F. (2004). A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association*, 99(468):909–917.
- [Wu et al., 2011] Wu, Z., Tong, W., Tan, Z., Wang, S., and Lin, P. (2011). [The clinical significance of β -arrestin 2 expression in the serum of non-small cell lung cancer patients]. *Zhongguo fei ai za zhi = Chinese journal of lung cancer*, 14(6):497–501.
- [Wysocka et al., 2006] Wysocka, J., Allis, C. D., and Coonrod, S. (2006). Histone arginine methylation and its dynamic regulation. *Frontiers in bioscience : a journal and virtual library*, 11:344–55.
- [Xu et al., 2010] Xu, X., Hoang, S., Mayo, M. W., and Bekiranov, S. (2010). Application of machine learning methods to histone methylation ChIP-Seq data reveals H4R3me2 globally represses gene expression. *BMC Bioinformatics*, 11(1):396.
- [Yang and Weinberg, 2008] Yang, J. and Weinberg, R. A. (2008). Epithelial-mesenchymal transition: at the crossroads of development and tumor metastasis. *Developmental cell*, 14(6):818–29.
- [Young et al., 2011] Young, M. D., Willson, T. A., Wakefield, M. J., Trounson, E., Hilton, D. J., Blewitt, M. E., Oshlack, A., and Majewski, I. J. (2011). ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. *Nucleic acids research*, 39(17):7415–27.
- [Yu et al., 2008] Yu, H., Zhu, S., Zhou, B., Xue, H., and Han, J.-D. J. (2008). Inferring causal relationships among different histone modifications and gene expression. *Genome research*, 18(8):1314–24.
- [Zang et al., 2009] Zang, C., Schones, D. E., Zeng, C., Cui, K., Zhao, K., and Peng, W. (2009). A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics (Oxford, England)*, 25(15):1952–8.

- [Zeller et al., 2003] Zeller, K. I., Jegga, A. G., Aronow, B. J., O'Donnell, K. A., and Dang, C. V. (2003). An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets. *Genome biology*, 4(10):R69.
- [Zentner et al., 2011] Zentner, G. E., Tesar, P. J., and Scacheri, P. C. (2011). Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome research*, 21(8):1273–83.
- [Zhao et al., 2005] Zhao, J., Herrera-Diaz, J., and Gross, D. S. (2005). Domain-wide displacement of histones by activated heat shock factor occurs independently of Swi/Snf and is not correlated with RNA polymerase II density. *Molecular and cellular biology*, 25(20):8985–99.
- [Zhao et al., 2009] Zhao, Q., Rank, G., Tan, Y. T., Li, H., Moritz, R. L., Simpson, R. J., Cerruti, L., Curtis, D. J., Patel, D. J., Allis, C. D., Cunningham, J. M., and Jane, S. M. (2009). PRMT5-mediated methylation of histone H4R3 recruits DNMT3A, coupling histone and DNA methylation in gene silencing. *Nature structural & molecular biology*, 16(3):304–11.
- [Zhou and Lipowsky, 2004] Zhou, H. and Lipowsky, R. (2004). Network Brownian motion: A new method to measure vertex-vertex proximity and to identify communities and subcommunities. *Lecture Notes in Computer Science*, 3038:1062–1069.