

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Mechanistic Insights Into Gene Regulation By Transcriptional Enhancers using CRISPR
Genome Engineering

Permalink

<https://escholarship.org/uc/item/4jw4k2gb>

Author

Rivera, Chloe Monica

Publication Date

2015

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Mechanistic Insights Into Gene Regulation By Transcriptional Enhancers using
CRISPR Genome Engineering

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

in

Biomedical Sciences

by

Chloe Monica Rivera

Committee in charge:

Professor Bing Ren, Chair
Professor Webster Cavenee
Professor Joseph Ecker
Professor Christopher Glass
Professor Maïke Sander

2015

Copyright

Chloe Monica Rivera, 2015

All rights reserved.

The Dissertation of Chloe Monica Rivera is approved, and it is acceptable in
quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2015

DEDICATION

I'd like to dedicate this doctoral thesis to my grandparents-- Pablo, Mercedes, Robert and Mona. They always strived to provide for their families so that we would have better opportunities than they were given. I know they are so proud to have the first doctor in our family. I'm so grateful for this chance to challenge myself and to explore science. Now on to the next opportunity...

TABLE OF CONTENTS

SIGNATURE PAGE.....	iii
DEDICATION.....	iv
TABLE OF CONTENTS.....	v
LIST OF FIGURES.....	ix
LIST OF TABLES.....	xi
ACKNOWLEDGEMENTS.....	xii
VITA.....	xiv
ABSTRACT OF THE DISSERTATION.....	xix
CHAPTER 1- INTRODUCTION: MAPPING HUMAN EPIGENOMES.....	1
ABSTRACT.....	1
INTRODUCTION.....	1
EPIGENOME MAPPING TECHNOLOGIES.....	4
<i>Mapping DNA methylation.....</i>	<i>5</i>
<i>Mapping chromatin modification states.....</i>	<i>11</i>
<i>Mapping of chromatin structures.....</i>	<i>16</i>
Nucleosome mapping.....	16
Chromatin accessibility.....	19
Mapping of Higher Order Chromatin Architectures.....	21
UTILITY OF EPIGENOME MAPS.....	25
<i>Annotation of cis-regulatory elements from chromatin profiles.....</i>	<i>26</i>
<i>Annotation of long-range chromatin interactions.....</i>	<i>30</i>

<i>Dynamic chromatin landscapes during human development.....</i>	<i>33</i>
<i>Understanding disease variants with epigenomics.....</i>	<i>38</i>
THE ROAD AHEAD.....	40
FIGURES.....	45
TABLES.....	48
AUTHOR'S CONTRIBUTIONS.....	51
ACKNOWLEDGEMENTS.....	51
CHAPTER 2 - CRISPR REVEALS A DISTAL SUPER-ENHANCER REQUIRED FOR	
SOX2 EXPRESSION IN MOUSE EMBRYONIC STEM CELLS.....	52
ABSTRACT.....	52
INTRODUCTION.....	53
RESULTS.....	55
<i>Identification of a distal Sox2 super-enhancer in mouse ESCs.....</i>	<i>55</i>
<i>Epigenetic characterization of the Sox2-SE^{distal} in mouse ESCs.....</i>	<i>56</i>
<i>A simple and efficient strategy to disrupt enhancers with CRISPR.....</i>	<i>57</i>
<i>Sox2-SE^{distal} is responsible for over 90% of Sox2 transcription.....</i>	<i>58</i>
<i>Sox2-SE^{distal} deletion alters transcription of key pluripotency genes.....</i>	<i>59</i>
DISCUSSION.....	61
METHODS.....	63
<i>Cell culture and transfection experiments.....</i>	<i>63</i>
<i>Design of CRISPR constructs for enhancer deletion.....</i>	<i>64</i>
<i>PCR and Sanger sequencing for genotyping deletion clones.....</i>	<i>65</i>

<i>RNA isolation and allele-specific RT-qPCR.....</i>	<i>66</i>
<i>Strand-specific RNA-seq.....</i>	<i>67</i>
<i>ChIP-seq.....</i>	<i>68</i>
<i>Bioinformatic analysis.....</i>	<i>68</i>
<i>Data Availability.....</i>	<i>69</i>
FIGURES.....	71
AUTHOR'S CONTRIBUTIONS.....	78
ACKNOWLEDGEMENTS.....	78
APPENDIX.....	79

CHAPTER 3 - FUNCTIONAL IDENTIFICATION OF TRANSCRIPTIONAL

ENHANCER TARGET GENES TO REVEAL THE RULES OF GENE	
REGULATION.....	82
ABSTRACT.....	82
INTRODUCTION.....	83
<i>Definition of transcriptional enhancers and proposed models of</i>	
<i>function.....</i>	<i>83</i>
<i>Methods of identifying the promoter targets of enhancers.....</i>	<i>85</i>
<i>Summary of enhancer knockout studies.....</i>	<i>87</i>
RESULTS.....	89
<i>Sox2-SE^{distal} controls the Sox2 promoter at the level of PolII</i>	
<i>recruitment.....</i>	<i>89</i>
<i>ENCODE Enhancer KO Project study approach and progress.....</i>	<i>91</i>

<i>Deletion of Tet1 Promoter and super enhancer in mESCs.....</i>	92
<i>Functional characterization of PRDM14 SE reveals a singular promoter target.....</i>	94
DISCUSSION.....	96
METHODS.....	99
<i>Cell culture and transfection experiments.....</i>	99
<i>Design of CRISPR constructs for enhancer deletion.....</i>	100
<i>PCR and Sanger sequencing for genotyping deletion clones.....</i>	100
<i>Strand-specific RNA-seq.....</i>	101
<i>Allelic 4C-seq.....</i>	103
<i>ChIP-seq.....</i>	104
FIGURES.....	105
TABLES.....	114
AUTHOR'S CONTRIBUTIONS.....	115
ACKNOWLEDGEMENTS.....	115
APPENDIX.....	116
APPENDIX.....	120
REFERENCES.....	134

LIST OF FIGURES

Figure 1. Timeline of Sequencing Based Technologies for Mapping Human Epigenomes.....	45
Figure 2. Comparison of Human DNA Methylation Technologies.....	46
Figure 3. Hierarchical Principles of Nuclear Organization.....	47
Figure 4. Identification of a distal mouse ESC-specific super-enhancer near Sox2 gene.....	71
Figure 5. Characterization of the distal Sox2 super-enhancer.....	72
Figure 6. Deleting Sox2-SE ^{distal} using CRISPR.....	73
Figure 7. Sox2-SE ^{distal} regulates Sox2 expression in cis.....	74
Figure 8. Sox2-SE ^{distal} deletion affects key pluripotency genes.....	76
Figure 9. Genotyping of enhancer deletion clones.....	79
Figure 10. Deletion of Sox2-SE ^{distal} impairs ES cell proliferation.....	81
Figure 11. Sox2-SE ^{distal} is required for PolII recruitment and H3K27ac deposition at the Sox2 promoter.....	105
Figure 12. Knocking out Sox2-SE ^{distal} alters the local chromatin interactions with the Sox2 promoter.....	106
Figure 13. Experimental design for ENCODE 100 enhancer knockout project pipeline.....	107
Figure 14. Super enhancer and typical enhancer CRISPR deletion pipeline progress.....	108

Figure 15. QPCR genotyping results for Super enhancer and typical enhancer CRISPR deletion clones.....	109
Figure 16. Impact of Tet1 promoter and SE deletion on gene expression.....	110
Figure 17. Deletion of PRDM14-SE dramatically decreases PRDM14 gene expression.....	112
Figure 18. Histone modification features of human transcriptional enhancers.....	129
Figure 19. Chromatin states at poised, active, and decommissioned enhancers.....	129
Figure 20 ChIP-chip analysis of H3K4 methyltransferases and demethylases in K562 cells.....	130
Figure 21. Analysis of histone modifying enzyme interaction with validated PU.1 target enhancer.....	131
Figure 22. Generation of a human haploid cell gene trap library.....	132
Figure 23. LM-PCR based strategy for mapping of viral insertions.....	133

LIST OF TABLES

Table 1. Large Scale National and International Epigenomic Consortia.....	48
Table 2. Major Conceptual Advances Convey the Utility of Epigenome Maps.....	49
Table 3. Distinctive Chromatin Features of Genomic Elements.....	50
Table 4. FPKM for all mouse RefSeq genes in 4 representative mESC clones from RNA-seq.....	81
Table 5. A list of 136 genes upregulated after deleting Sox2 enhancer.....	81
Table 6. A list of 142 genes downregulated after deleting Sox2 enhancer.....	81
Table 7. A list of genes bound by Sox2 at promoters.....	81
Table 8. Summary of the results of enhancer knockout studies in mammalian systems.....	114
Table 9. List of CRISPR sgRNAs used for super enhancer knockouts.....	116
Table 10. List of CRISPR sgRNAs used for typical enhancer knockouts.....	117
Table 11. Primers used to genotype super enhancer knockout clones by PCR and qPCR.....	118
Table 12. Primers used to genotype typical enhancer knockout clones by PCR and qPCR.....	119

ACKNOWLEDGEMENTS

I'd first like to thank Dr. Bing Ren for being a supportive mentor and excellent role model, scientifically and otherwise. I'd like to thank my other faculty mentors at UCSD including Dr. Ben Yu, Dr. Bruce Hamilton, and my thesis committee.

My sincerest thanks to all current and past members of the Ren lab for their generous advice and friendship. I would like to especially thank Dr. Yan Li and Dr. Haruhiko Ishii, my scientific dream team. And I would like to thank Dr. Andrea Local and Dr. Gary Hon for being the center of all interactions – the enhancers.

I'd like to acknowledge my collaborators and scientists who shared reagents, data, or advice instrumental to my thesis work, especially Dr. Feng Zhang for the CRISPR reagents.

For their unwavering confidence in me and limitless patience as I finished this scientific journey, I'd like to acknowledge my family and friends. Mom, Dad, Evan, Marissa, and Brandon – your loving support has meant everything.

Chapter 1, in full, is an adaptation of materials that appears in Rivera, CM, Ren, B. Mapping Human Epigenomes. *Cell*. 2013 Sept.

Chapter 2, in full, is an adaptation of materials that appears in

Li Y, Rivera CM, Ishii H, Jin F, Selvaraj S, Lee AY, Dixon JR, Ren B. CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells. *PLoS One*. 2014 Dec.

VITA

EDUCATION

University of California, San Diego **2008-2015**
Doctor of Philosophy in Biomedical Sciences, Genetics and Genomics Track

University of California, Los Angeles **2004-2008**
Bachelor of Science in Microbiology, Immunology, and Molecular Genetics

BIOMEDICAL RESEARCH EXPERIENCE

Doctoral Candidate **2008-2015**
Biomedical Sciences Program, UCSD

PI: Dr. Bing Ren, Ph.D.

- Improved a low-throughput sequencing method to a high-throughput method to identify the genomic locations of retroviral insertions using the HiSeq2000 Illumina next-generation sequencing platform
- Characterized the epigenomic, transcriptional, and nuclear architecture landscapes of myeloid leukemia and embryonic stem cell lines (via ChIP-seq, RNA-seq, and HiC) to predict transcriptional enhancer elements and their target promoters genome-wide
- Validated predicted transcriptional enhancers, enhancer target genes, and regulatory element chromatin interactions with CRISPR-mediated genome engineering

Summer Research Intern **Summer 2008**
Immunology Drug Discovery, Biogen Idec

Project Leader: Dr. Elizabeth Mertsching, Ph.D.

- Screened novel small molecule inhibitors of B cell activation and proliferation with rheumatoid arthritis as lead indication
- Developed and optimized mechanistic assays for target protein and pathway inhibition

Undergraduate Research Assistant **2006-2008**
Department of Molecular and Medical Pharmacology, UCLA

PI: Dr. Harley Kornblum, M.D., Ph.D.

- Observed the self-renewal mechanisms of neural stem cells (NSC) are regulated by reactive oxygen species produced by NADPH oxidase (NOX) enzymes through complementary studies
- First, demonstrated high concentrations of reactive oxygen species (H_2O_2) increases NSC self-renewal as quantified by an *in vitro* neurosphere assay
- Secondly, pharmacologic inhibition of NOX enzymes results in decreased NSC self-renewal

Summer Undergraduate Researcher **Summer 2006**
Emory School of Medicine

PIs: Dr. Jorg Goronzy, M.D., Ph.D. and Dr. Cornelia Weyand, M.D., Ph.D.

- NSF Summer Undergraduate Research Experience (SURE) Program
- Designed a novel co-culturing method of primary T-cells and fibroblast-like synoviocytes from patients with rheumatoid arthritis
- Determined T-cell (LFA1) and synoviocyte (ICAM2) cell surface receptor interactions are necessary for the proliferation of fibroblast-like synoviocytes

LEADERSHIP AND TEAMWORK EXPERIENCE

Director of Student Committee **2014-2015**
Rady Micro-MBA Summer Course

- Pioneered the first Rady Micro-MBA Summer Course, a 10-day course covering 1st year MBA topics molded for PhD students and postdocs, which 120 students completed this year
- Prepared a written proposal highlighting the benefit of the program to all stakeholders to gain buy-in from the Graduate Student Association, Dean of Graduate Studies, and the Dean of Rady School of Management
- Surveyed 400+ prospective course participants to understand price sensitivity towards course fee and successfully negotiated reduced fees with Assistant Dean of the Rady School of Management

Vice President & Engagement Manager **2013-2015**
UCSD Advanced Professional Degree Consulting Club

- Awarded Student Organization of the Year Distinction for 2013-2014 based on demonstrated leadership, contributions to the UCSD community, and development of our members

- Directed the inaugural UCSD Consulting Firm Showcase by establishing a collaborative relationship with four San Diego consulting clubs, pitching the event idea to consulting firms, and delegating event logistics resulting in 82% of the attendees rating the event as a “5-Very Helpful” on a 1-5 scale
- As Engagement Manager, I coached teams of five PhDs and one Project Lead on project organization and management, constructing client-ready documents, and verbal and written communication skills. Completed Projects:
 - 1) Market Entry Strategy for a Risk Management Software Startup
 - 2) White Paper in collaboration with Bainbridge Consulting: Adoption of Wearable Wireless Devices for Inpatient Hospital Care

Co-President & Career Development Seminar Coordinator 2008-2015
UCSD Biomedical Sciences Student Council

- Initiated, designed, and executed a new career exploration seminar series to support student interests in alternate careers by lobbying for a \$10,000 budget, organizing training workshops, and inviting biomedical industry professionals to speak at career panels
- As Co-President I managed a \$20,000 annual budget, supervised all student committees, and served a liaison to communicate the needs of the students to the Department Chair

Bioscience Management Bootcamp Scholar 2013
Keck Graduate Institute

- Grant sponsored business training program for 40 awardees out of 300 applicants
- Completed Harvard Business School case-based courses in finance, project management, IP, and business strategy
- Collaborated in a team of five PhDs to assess the commercialization opportunity for a novel prostate cancer diagnostic
- Measured diagnostic’s potential impact on clinical decision-making, estimated savings to insurance companies (payers), and established value and reference-based pricing models
- Team received Best Presentation Award out of 8 groups

Business Consultant Intern 2013
Oxbridge Biotech Roundtable Consulting

- Conducted market research to characterize the magnetic bead industry in San Diego with a team of seven PhD student and postdoctoral consultants
- Surveyed 300+ academic and industry researchers and interviewed important thought leaders such as departmental directors and core facility directors

- Presented market entry strategy to the CEO including recommendations to limit product features and applications

Graduate Student Representative **2011**
UCSD Biomedical Sciences Admissions Committee

- Four students out of 30 student class selected to serve on the admissions committee based on organizational skills and ability to motivate student participation
- Evaluated 500+ applications, selected students to interview, and organized recruitment weekend activities

Director of Teaching **2007-2008**
UCLA City Lab

- City Lab is an outreach program to expose underprivileged high school students to molecular biology, the scientific process, and critical thinking
- Developed and presented lectures utilizing inquiry-based learning for a young audience
- Trained TAs on one-on-one mentoring and teaching skills

Vice President & Rush Director, Phi Sigma Rho Sorority **2005-2008**

- National social sorority for science and engineering majors
- As Vice President I directed an effort to petition to join the UCLA Pan-Hellenic Council which resulted in Phi Sigma Rho successfully joining the Council as an Associate Member
- Negotiated the conditions to lease a \$5M house which served as the Phi Sigma Rho UCLA chapter's first ever residence on sorority row

AWARDS AND DISTINCTIONS

UCSD Student Organization of the Year Award for the Advanced Professional Degree Consulting Club	2014
McKinsey and Company Summer Insight Program Selectee	2013
KGI Bioscience Management Bootcamp Scholarship and Best Presentation Award	2013
Cold Spring Harbor Asia High Throughput Biology Poster Award	2012
BIOCOM Life Science Immersion Program Grantee	2012
NIH Genetics Training Grant Trainee	2010

NIH Initiative for Maximizing Student Diversity Scholar	2008
Elma Gonzalez Outstanding Senior UCLA MARC Scholar Award	2008
UCLA Dean's Honor for Community Service – City Lab	2008
UCLA Dean's List	2006-2007
NIH Minority Access to Research Careers (MARC) Scholarship	2006-2008
UCLA Center for Academic and Research Excellence Scholar	2005-2006
Jean C. Cook Scholar	2004

PUBLICATIONS

* denotes co-authors

Li Y*, **Rivera CM***, Ishii H*, Jin F, Selvaraj S, Lee AY, Dixon JR, Ren B. CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells. *PLoS One*. 2014 Dec.

Rivera, CM, Ren, B. Mapping Human Epigenomes. *Cell*. 2013 Sept.

ABSTRACT OF THE DISSERTATION

Mechanistic Insights Into Gene Regulation By Transcriptional Enhancers using
CRISPR Genome Engineering

by

Chloe Monica Rivera

Doctor of Philosophy in Biomedical Sciences

University of California, San Diego, 2015

Professor Bing Ren, Chair

Transcriptional regulation is the key mechanism responsible for differential gene expression in the 300+ human cell types during both development and disease. Here I use both genomic and epigenomic tools to characterize the role of cis-regulatory elements in transcriptional regulation.

In Chapter 1, I discuss the current epigenome mapping toolkit and utilities of epigenome maps. I focus particularly on mapping of DNA methylation, chromatin modification state and chromatin structures, and emphasize the use of epigenome maps to delineate human gene regulatory sequences and developmental programs. I also provide a perspective on the progress of the epigenomics field and challenges ahead.

In Chapter 2, I explore the role of a large cell-type restricted regulatory element, called a super enhancer (SE), in controlling the expression of a master regulatory transcription factor of mouse embryonic stem cells. This SE is occupied by Oct4, Sox2, Nanog, and the mediator complex, and physically interacts with the Sox2 locus via DNA looping. Using a simple and highly efficient double-CRISPR genome editing strategy the entire 13-kb SE was deleted and the transcriptional defects in the resulting monoallelic and biallelic deletion clones were characterized with RNA-seq. These results showed that the SE is responsible for over 90% of Sox2 expression, and Sox2 is the only target gene along the chromosome. These results support the functional significance of a SE in maintaining the pluripotency transcription program in mouse ESCs.

In Chapter 3, my work continues to dissect the impact of SE deletion on transcription from the Sox2 locus. ChIP-seq and allelic 4C-seq assays are used to understand the impact on RNA polymerase II recruitment, open chromatin, and local interactions at the Sox2 promoter. Second, the functional study of enhancers is expanded to a selection of 100 enhancers and SEs in mouse

embryonic stem cells again using a CRISPR deletion strategy. The purpose of this study is to determine which enhancers predicted by chromatin signatures are functional, to identify their target gene(s), and to determine the extent to which each enhancer contributes to the expression of their target gene(s).

CHAPTER 1 – INTRODUCTION: MAPPING HUMAN EPIGENOMES

ABSTRACT

As the second dimension to the genome, the epigenome contains key information specific to every type of cells. Thousands of human epigenome maps have been produced in recent years thanks to rapid development of high throughput epigenome mapping technologies. In this review, we discuss the current epigenome mapping toolkit and utilities of epigenome maps. We focus particularly on mapping of DNA methylation, chromatin modification state and chromatin structures, and emphasize the use of epigenome maps to delineate human gene regulatory sequences and developmental programs. We also provide a perspective on the progress of the epigenomics field and challenges ahead.

INTRODUCTION

More than a decade has passed since the human genome was completely sequenced, but how genomic information directs spatial and temporal specific gene expression programs remains to be elucidated (Lander 2011). The answer to this question is not only essential for understanding the mechanisms of human development, but also key to studying the phenotypic variations among human populations and the etiology of many human diseases.

However, a major challenge remains: each of the more than 200 different cell types in the human body contains an identical copy of the genome but expresses a distinct set of genes. How does a genome guide a limited set of genes to be expressed at different levels in distinct cell types?

Overwhelming evidence now indicates that the epigenome serves to instruct the unique gene expression program in each cell type together with its genome. The word “epigenetics”, coined half a century ago by combining “epigenesis” and “genetics”, to describe the mechanisms of cell fate commitment and lineage specification during animal development (Holliday 1990; Waddington 1959). Today, the “epigenome” is generally used to describe the global, comprehensive view of sequence-independent processes that modulate gene expression patterns in a cell, and has been liberally applied in reference to the collection of DNA methylation state and covalent modification of histone proteins along the genome (Bernstein et al. 2007; Bonasio et al. 2010). The epigenome can differ from cell type to cell type, and in each cell it regulates gene expression in a number of ways - by organizing the nuclear architecture of the chromosomes, restricting or facilitating transcription factor access to DNA, and preserving a memory of past transcriptional activities. Thus, the epigenome represents a second dimension of the genomic sequence and is pivotal to for maintaining cell-type-specific gene expression patterns.

Not long ago, there were many points of trepidation about the value and utility of mapping epigenomes in human cells (Madhani et al. 2008). At the time,

it was suggested that histone modifications simply reflect activities of transcription factors (TFs), so cataloging their patterns would offer little new information. However, some investigators believed in the value of epigenome maps and advocated for concerted efforts to produce such resources (Feinberg 2007; Henikoff et al. 2008; Jones and Martienssen 2005). The last five years have shown that epigenome maps can greatly facilitate the identification of potential functional sequences thereby annotating of the human genome. Now, we appreciate the utility of epigenomic maps in the delineation of thousands of lincRNA genes and hundreds of thousands of cis-regulatory elements (Ernst et al. 2011; Guttman et al. 2009; Heintzman et al. 2007; Mouse ENCODE Consortium et al. 2012; Xie et al. 2013b; Zhu et al. 2013), all of which were obtained without prior knowledge of cell-type-specific master transcriptional regulators. Interestingly, bioinformatic analysis of tissue-specific cis-regulatory elements has actually uncovered novel TFs regulating specific cellular states.

Propelled by rapid technological advances, the field of epigenomics is enjoying unprecedented growth with no sign of deceleration. An expanding cadre of researchers is working to explore exciting frontiers in epigenomics. Many international consortia have been formed to tackle the fundamental problems in epigenomics by sharing resources and protocols (Table 1) (Beck et al. 2012b; Bernstein et al. 2010). Consequently, the number of epigenomic datasets and publications has grown exponentially in recent years. The resulting epigenomic maps have linked genomic sequences to many nuclear processes

including splicing, replication, DNA damage response, folding, chromatin packaging, and cell type specific gene expression patterns.

In this review, we focus on recent progress in several areas of epigenomics. First, we describe the remarkable advances in epigenomic technologies, especially the next-generation sequencing based applications, which have fueled the growth of the field. Second, we discuss the utility of epigenomic maps, emphasizing the power of these maps in annotating transcription units, cis-regulatory elements in the context of development and disease pathogenesis. We also explore new biological insights gained through integrative analysis of epigenomic maps in mammalian cell systems, highlighting the study of pluripotency and lineage specification of embryonic stem cells. Finally, we provide a perspective on the road ahead regarding meeting the technical challenges and addressing unanswered questions in the field. As for advancements in our understanding of the mechanistic roles of sequence-specific TFs and non-coding RNAs, chromatin and DNA modifying enzymes, and chromatin binding proteins in establishing, maintaining, and removing epigenetic marks, we refer readers to recent excellent reviews (Badeaux and Shi 2013; Calo and Wysocka 2013; Lee and Young 2013; Pastor et al. 2013; Rinn and Chang 2012; Smith and Meissner 2013).

EPIGENOME MAPPING TECHNOLOGIES

The unprecedented genome-wide scope and nucleotide precision with which we can now map human epigenomes was enabled by disruptive technology advancement, namely the application of microarrays and next generation sequencing (NGS). A slew of molecular biology assays previously used to measure a single locus are now integrated with these platforms. Powerful parallel short-read sequencing technologies have proven increasingly high throughput, fast, accurate, and cost-effective at rates faster than Moore's Law. By far, the greatest advantage of NGS is its ability to survey the entire genome in an unbiased and comprehensive manner. Due to this monumental shift in assay capacity, researchers can ask if conclusions drawn from locus-centered studies extend to other parts (even unknown parts) of the genome. In defiance of the traditional scientific method emphasizing hypothesis testing, global profiling promotes hypothesis-free exploration of new observations and correlations. Overall, the accomplishments and discoveries of the epigenomics field have hinged in large part on the iterative inventions and improvements of technologies (Figure 1). Here we discuss cutting-edge epigenome mapping techniques that integrate next generation sequencing platforms and disclose their advantages and disadvantages.

Mapping DNA Methylation

Of the four nucleotides composing DNA, cytosine is by far the most dynamic. Cytosine can be methylated at its 5th carbon (5mC) and in the human

genome 60-80% of 28 million CpG dinucleotides are methylated (Lister et al. 2009; Ziller et al. 2013). Stressing the importance of DNA methylation during development, deletion of cytosine methyltransferases responsible for *de novo* (DNMT3A, DNMT3B) or maintenance (DNMT1) of methylation through cellular divisions results in embryonic and neonatal lethality in mice (Li et al. 1992; Okano et al. 1999). Detection of DNA methylation at individual loci and with promoter-focused studies established the important repressive roles of DNA methylation in imprinting, retrotransposon silencing, and X chromosome inactivation (Bird 2002). Global DNA methylation technologies now measure DNA methylation abundance at all cytosines at base resolution in the human genome. The elucidation of complete human methylomes progressed the narrow view of 5mC as only a stable repressive mark to an epigenetic mark that is dynamically deposited and removed, can exist in non-CpG sequence contexts, and is enriched at the bodies of actively transcribed genes (Hellman and Chess 2007; Lister et al. 2009).

The DNA methylation toolkit includes three main molecular biology based techniques: digestion of genomic DNA with methyl-sensitive restriction enzymes enrichment of methylated fragments, affinity-based enrichment of methylated DNA fragments, and chemical conversion methods (Bock 2012; Laird 2010). The choice of how to assay DNA methylation depends on the resolution and genome coverage needs, and both parameters ultimately dictate the experimental cost (Figure 2). Endonuclease digestion based DNA methylation assays (MRE-seq,

etc.) determines CpG methylation at base resolution by digesting genomic DNA with several methylation-sensitive enzymes. The genome-wide CpG coverage of MRE-seq is limited by the cutting frequency of the chosen restriction enzyme(s) and biased towards the enzyme recognition sequences, while the accuracy is dependent on complete digestion. Affinity-based enrichment assays capture methylated sonicated DNA with an antibody (MeDIP-seq) or a methyl binding domain (MBD-seq) (Down et al. 2008; Serre et al. 2010). When sequencing enriched DNA fragments, at least one cytosine is certainly methylated but the exact site or combination of sites could not be directly determined. Therefore, the resolution of affinity-based assays is highly dependent on the DNA fragment size, CpG density, and immunoprecipitation quality of the reagent. The results of both restriction enzyme and affinity based sequencing methods are qualitative levels of enrichment rather than absolute. On the other hand, because affinity and restriction enzyme based methylation assays enrich or capture methylated DNA regions the sequencing costs are restricted.

Bisulfite sequencing is a chemical conversion method that directly determines the methylation state of each cytosine in a binary fashion and is widely accepted as a gold standard for mapping DNA methylation. Treatment of genomic DNA with sodium bisulfite chemically converts unmethylated cytosines to uracil. After PCR and assuming nearly complete bisulfite conversion, all unmethylated cytosines become thymidines and remaining cytosines correspond to 5mC. Initially, individual loci were assayed from BS treated genomic DNA with

locus specific PCR primers followed by Sanger sequencing (Clark et al. 1994). In a step towards increasing genomic coverage, reduced representation bisulfite sequencing (RRBS) combines restriction digestion with bisulfite sequencing for specific interrogation of high CpG density regions such as clusters of CpGs at promoters called CpG islands (Meissner 2005). At the pinnacle of the genomic coverage spectrum, bisulfite treatment coupled with whole genome sequencing (variably referred to as MethylC-seq, BS-seq, or WGBS) features nucleotide resolution and quantitative rates of methylation for all cytosines (Cokus et al. 2008; Lister et al. 2008). Careful analysis of MethylC-seq data should consider PCR biases due to unbalanced GC content of methylated and unmethylated fragments, mapping inefficiencies of bisulfite treated DNA, and estimate the bisulfite conversion rate using a spike-in control (Krueger et al. 2012; Laird 2010).

The DNA methylation field collectively experienced an epiphany upon the discovery of 5-hydroxymethylcytosine (5hmC) as an intermediate of demethylation of 5mC to cytosine (Kriaucionis and Heintz 2009; Tahiliani et al. 2009). The ten-eleven translocation (TET) family of proteins, TET1, TET2, and TET3, oxidize 5mC through 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC) intermediates before being replaced by cytosine via base excision repair pathways or a yet unidentified decarboxylase (Pastor et al. 2013). As it became clear that four variants of cytosine exist, a clarification to MethylC-seq results also came to light. 5mC and 5hmC, but not

5fC and 5caC, are both resistant to bisulfite conversion and therefore cannot be distinguished from each other in MethylC-seq data (Huang et al. 2010; Jin et al. 2010). In order to understand the role of DNA demethylation, new techniques would need to be developed to accurately differentiate cytosine methylation states.

In an exciting advancement in the field, an assortment of methods was published for the detection of all cytosine methylation states across the genome. The first versions use antibodies to either directly IP 5hmC (hMeDIP-seq) or chemically modify 5hmC making it more immunogenic (anti-CMS, hMe-Seal, GLIB, JBP-1) (Ficz et al. 2011; Pastor et al. 2011; Robertson et al. 2011; Song et al. 2011; Williams et al. 2011). While these methods are an important advancement affinity-based techniques are hindered by low resolution and qualitative signal, as mentioned above. Last year, both oxidative bisulfite sequencing (oxBS-seq) and TET assisted bisulfite sequencing (TAB-seq) were introduced as single-base resolution methods for measuring 5hmC (Booth et al. 2012; Yu et al. 2012). In oxBS-seq, 5hmC nucleotides are sensitized to bisulfite treatment after a chemical reaction to specifically oxidize all 5hmC and to 5fC. Then, DNA is successively treated with sodium bisulfite such that remaining cytosines must originally be 5mC. The sequencing results of oxBS-seq therefore accurately capture 5mC levels and subtraction of MethylC-seq data reveals true 5hmC sites. Of concern with oxBS-seq are the successive chemical treatments that can induce DNA damage and skew the results. The second approach, TAB-seq, directly assays

5hmC location and abundances. First, 5hmC is tagged with a glucose molecule using the T4 bacteriophage enzyme beta glucosyltransferase. Next, genomic DNA is treated with purified TET enzyme to oxidize 5mC to 5caC while glucosylated-5hmC is protected. Finally, after bisulfite treatment and sequencing only 5hmC is read as cytosine, while all cytosines and cytosine variants are detected as thymidine. TAB-seq's accuracy is especially dependent on efficient oxidation and conversion by TET, such that a bottleneck is the tedious process of purifying catalytically active TET enzyme (Adey and Shendure 2012; Adli and Bernstein 2011; Ng et al. 2013; Song et al. 2012). In a final round of progress for DNA methylation detection, methods to detect 5fC (fC-seq, fCAB-seq, 5fC-DP-Seq) and 5caC (5caC-seq) were recently described, and the only unturned stone is measuring 5caC at base resolution (Raiber et al. 2012; Shen et al. 2013; Song et al. 2013).

Methods without enrichment steps, like MethylC-seq, oxBS-seq, and TAB-seq, require an immense amount of sequencing and are costly (Figure 2). In order to quantitatively measure the methylation rate at each cytosine, high-quality single-nucleotide resolution 5mC methylome data sets are expected to have 30X sequencing depth. A recent study by Meissner and colleagues finds that only ~20% of CpGs are differentially methylated between 30 diverse human cell and tissue types tested. Therefore, up to 80% of MethylC-seq data are the result of superfluous sequencing of DNA fragments without CpGs or containing uninformative, constitutively methylated CpGs across 30 cell and tissue types

examined (Ziller et al. 2013). Sequencing costs can be limited by targeted DNA methylation mapping as accomplished by bisulfite padlock probes (BSPPs) and Illumina's Infinium 450K BeadChIP technology (Bibikova et al. 2011; Diep et al. 2012). Theoretically, base resolution methylomes of rare cytosine variants such as 5caC should have 1000X sequence coverage (Pastor et al. 2013). The extreme sequencing depth is necessitated by the rarity of cytosine variants; for example, in mouse embryonic stem cells (mESCs) 5hmC accounts for about 0.1% and 5caC accounts for 0.0003% of cytosines (Ito et al. 2011; Yu et al. 2012). Unfortunately, the prohibitive cost of genome-wide base resolution human methylomes has limited the number of available datasets and presumably slowed progress towards novel insights. Resolution, genomic coverage, and monetary cost serve as interdependent considerations for generating methylome data where one parameter cannot be changed without affecting the others (Figure 2). Maximizing resolution and coverage while keeping costs low, currently an unrealistic situation, may eventually be possible with innovative sequencing platforms or a yet undiscovered alternative to bisulfite treatment. We look forward to the democratization of nucleotide resolution DNA methylation technologies and the resulting novel findings from a diverse collection of methylomes.

Mapping Chromatin Modification States

Chromosomal DNA is packaged into nucleosomes with DNA wrapped around histone octamers consisting of H2A, H2B, H3, H4 subunits and their variants. The histone tails and globular domains of histone proteins are subject to over 130 post-translational modifications (PTMs) and over 700 distinct histone isoforms have been detected in human cells (Tan et al. 2011; Tian et al. 2012). Well-studied covalent modifications on histones include methylation, acetylation, phosphorylation, and ubiquitination. State-of-the-art mass spectrometry based proteomic technologies unveiled many novel histone PTMs such as crotonylation, succinylation, malonylation and others (Tian et al. 2012). Histone modifications serve both activating and silencing roles in transcription, generally by controlling the accessibility of DNA and by serving as binding substrates that recruit or exclude protein complexes (Kouzarides 2007). For example, H3K27ac is found at both active promoters and enhancers, H3K36me3 identifies actively transcribed gene bodies, and H3K27me3 marks heterochromatic or repressed regions (Li et al. 2007) (Table 3). Determining the genome-wide distribution of a histone mark can lead to clues about its role in transcriptional regulation and provoke follow-up mechanistic studies to further understand the PTMs deposition, removal, and role in development and disease.

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) maps the genome wide binding pattern of chromatin associated proteins, which includes modified histones. To perform this method, DNA-protein complexes containing a specific protein of interest are immunoprecipitated from crosslinked,

sonicated chromatin. DNA is purified from the enriched pool and adaptors are ligated for subsequent PCR and sequencing. The digital sequences of enriched DNA, called reads, are computationally aligned to the reference genome to define punctate peaks or broad blocks of modified histones or protein occupancy. Since its development in 2007, researchers have used ChIP-seq extensively to survey the genomic profiles of histones and their modifications, TFs, DNA and histone modifying enzymes, transcriptional machinery, and other chromatin-associated proteins (Barski et al. 2007; Johnson et al. 2007; Robertson et al. 2007). Furthermore, multi-dimensional datasets are now available for cell lines, primary cells, tissues, and embryos from an increasing number of species. The ENCODE and Roadmap Epigenomics Projects have contributed enormous data repositories by performing thousands of ChIP-seq experiments in hundreds of human cell types.

As the number of ChIP-seq data sets began to grow exponentially, lab-to-lab protocol variability threatened the quality of results and downstream cross-study analysis. To stave off data inconsistencies, both the ENCODE and Roadmap consortia published optimized standard operating procedures for ChIP-seq. Of major concern was the quality of antibodies for which ChIP is undeniably dependent. Both consortia assessed histone modification antibody quality using dot blot immunoassays against histone tail peptides to ensure specific binding and minimal cross reactivity (Egelhofer et al. 2011). The ENCODE Project's best practices include a rigorous two stage antibody validation process with a

combination of immunoassays, immunofluorescence patterns, and functional assays (Landt et al. 2012). These screening methods dramatically improved the data quality and reduced costs and lost time from failed ChIP-seq experiments. However, the membrane binding conditions in immunoassays screens are not identical to immunoprecipitation in solution using magnetic beads. Lamentably, the gold standard for ChIP-grade antibody classification remains actually performing the assay itself. To this point, Bernstein and colleagues designed a method called ChIP-string to screen for effective antibodies against chromatin regulator proteins (Ram et al. 2011). In this approach, multiplexed meso-scale ChIP-seq experiments survey protein enrichment at ~500 representative loci using the nCounter probe system. High quality antibodies are distinct from IgG patterns and the occupancy distribution correlates with a logical set of chromatin states. Validating antibody reagents for enrichment specificity and robustness ensures good quality ChIP-seq data sets with high signal to noise ratios.

Given that ChIP-seq is a mature technology, the technical restrictions of the technique are well defined by its users. These restrictions include the need for large amounts of starting material, limited resolution, and the dependence on antibodies. Improvements to ChIP-seq have been developed to address these limitations and expand the possibilities of its use. Collecting enough starting material for ChIP-seq can be challenging because experiments typically require 1 million (histone modifications) to 5 million (TFs and chromatin modifiers) cells. While this is feasible when studying fast dividing cell lines, the challenge arises

when studying primary cells and rare populations such as cancer stem cells or progenitor cells. ChIP-seq samples of 50,000 cells or less are possible with the ChIP-nano protocol (Adli and Bernstein 2011).

Key method modifications achieve effective chromatin fragmentation in small volumes, ensure minimal sample handling and loss by washing samples in columns, and reduce background signal. Another procedure, called ChIP-exo, improves the limited resolution from fragmentation heterogeneity after chromatin is prepared by sonication along (Rhee and Pugh 2011). As its name suggests, sonicated and immunoprecipitated DNA is treated with a 5'-to-3' exonuclease to digest DNA to the footprint of the crosslinked protein such that sequencing results are nucleotide resolution. This type of high resolution protein binding data is most beneficial for uncovering motifs of specific binding proteins and the effect of sequence variants on protein binding affinity. Profiling genome-wide DNA-protein interactions with ChIP-seq is technically challenging when studying novel proteins or protein isoforms, such as a histone variant, that lacks a robust or specific antibody. In this case, an obvious approach is to transiently or stably express a protein of interest (POI) with a tag or epitope that can be readily ChIP'ed. Controls are necessary to ensure the fusion protein's localization is not altered by non-endogenous expression levels, protein instability, steric interference, or other effects of the tag itself.

A ChIP step can be added to other genomic profiling approaches for integrated epigenomic profiling. First, two ChIP steps in a row, or Sequential-

ChIP-seq, can uncover histone PTMs on the same molecule or chromatin associated proteins in the same complex. Several groups combined bisulfite sequencing with ChIP giving rise to BisChIP-seq and ChIP-BS-seq (Brinkman et al. 2012; Statham et al. 2012). Long distance DNA interactions mediated by a specific protein can be profiled using chromatin interaction analysis by paired-end-tag sequencing, or ChIA-PET (Fullwood et al. 2009). We anticipate other inventive uses of ChIP technology to continue to uncover undiscovered roles of histone modifications and histone variants in transcriptional regulation.

Mapping of Chromatin structures

Nucleosome Positioning

Moving up the organizational hierarchy of genomic organization, we now look beyond the DNA and histone modifications to the positioning of nucleosomes along the genome. Our epigenome at its most basic level is repeating units of 147 base pairs wrapped 1.7 times around each nucleosome with varying distances of linker DNA between each unit. Even this extremely simplistic model is complex because nucleosome positioning can both inhibit and promote factor binding (Bell et al. 2011). First, nucleosomes can be positioned to obstruct or reveal specific DNA sequences. Secondly, because modifications on histone tails serve as binding platforms for transcriptional regulators, nucleosome positioning regulates factor recruitment. And finally, nucleosomes are suggested to inhibit transcription by slowing progression of RNA polymerase

II as it transcribes through a gene body. From a medical perspective, it will be important to determine the possible role of aberrant nucleosome positioning as caused by disease associated SNPs, insertions, deletions, and translocations.

Our understanding of the regulation of nucleosome positioning came from studies of smaller genomes, such as those in yeast and fly (Jiang and Pugh 2009). Nucleosome positioning along DNA is influenced by favorable DNA sequence composition, the actions of ATP-dependent nucleosome remodelers, and strongly positioned nucleosomes and DNA bound proteins such as TFs and RNA Pol II are barriers to nucleosome position shifting (Mavrich et al. 2008; Narlikar et al. 2013; Yuan et al. 2005). While we understand the main determinants of nucleosome positioning, the exact contribution of each is unclear and currently under debate.

The most common method for profiling genome-wide nucleosome positioning is micrococcal nuclease digestion of chromatin followed by high-throughput sequencing (MNase-seq). When native, uncrosslinked chromatin is digested with MNase, the linker DNA is cleaved while DNA wrapped around histone octamers or bound by TFs is protected. After purifying the DNA, approximately 150bp fragments corresponding to mononucleosomes are size selected on a gel and sequenced.

Given the relative size of the human genome as compared to model organisms and that most of it is nucleosomal, mapping nucleosome positioning in the human genome is no small feat. Even with only 10-fold genome coverage in

CD4+ T Cells, it was clear that nucleosomes are depleted at active transcriptional start sites (TSSs) and enhancers with ordered positioning radiating outward (Schones et al. 2008). In an extraordinary sequencing effort, Gaffney et al. generated paired and single-end MNase-seq data in seven lymphoblastoid cell lines yielding 240X coverage of a single cell type and found that ~80% of the genome has non-random, albeit weakly positioned nucleosomes (Gaffney et al. 2012). MNase-seq based studies have provided great insights into the global distribution and dynamics of nucleosomes in the human genome. However, there are limitations of these datasets to consider. First, MNase digestion at the ends of nucleosomes is inconsistent such that exact position of nucleosomes can only be estimated as the center of fragments of different lengths. The results are unfortunately not single-base resolution. Moreover, MNase has AT sequence preferences and MNase protected 150bp fragments are inferred to be mononucleosomes but could in fact be created by other proteins as well (Brogaard et al. 2012).

To circumvent the weaknesses of digestion based detection of nucleosome positioning, Widom and colleagues used chemical modification of engineered histones to cleave DNA wrapped around nucleosomes (Brogaard et al. 2012). In short, DNA is precisely cleaved in a reaction with hydroxyl radicals if it interacts with a mutated residue on H4 while wound around a nucleosome. Using this method, nucleosome maps in yeast show remarkable accuracy and consistency. Interestingly, single-base pair resolution of yeast nucleosomes reveals a 10bp

periodic sequence preference of flexible dinucleotides throughout the 147 bases in contact with the histone octamer. This data suggests that the role of sequence composition in the rotation positioning of the nucleosomes is stronger than previously appreciated. This method could be extended to the human genome for precise, high-resolution nucleosome positioning studies.

Another MNase-independent method for mapping nucleosome positioning uses DNA methyltransferase accessibility to footprint nucleosome positions, called nucleosome occupancy and methylome sequencing (NOMe-seq) (Kelly et al. 2012). The unique activity of the DNA methyltransferase *M.CviPI*, which methylates cytosine only in the GpC context, is exploited to record the DNA's nucleosomal status because nucleosomal GpCs are protected from methylation. Next, the DNA is bisulfite treated to convert unmethylated cytosines to thymine. Following sequencing, cytosines in the CpG context were originally methylated (5mC or 5hmC) and cytosines in the GpC context were nucleosome depleted. An important advantage of NOMe-seq is the dual epigenomic information, both nucleosome position and DNA methylation abundance, comes from a single molecule rather than possibly co-occurring in a population of cells.

Chromatin Accessibility

Nucleosomes are the basic repeated structural unit of the genome and, as previously mentioned, proteins that compete for DNA binding can affect their positioning. Biochemically active regulatory elements, including promoters,

enhancers, silencers, and insulators, are bound by sequence specific regulatory TFs. Open chromatin is therefore an overarching characteristic of biochemically active genomic regions. Open chromatin can be assayed genome-wide by DNase hypersensitivity followed by sequencing (DNase-seq) or formaldehyde-assisted identification of regulatory elements followed by sequencing (FAIRE-seq) (Boyle et al. 2008; Giresi et al. 2007). DNase-seq takes advantage of the protection conferred by tightly wound nucleosomes from DNaseI endonuclease digestion. Accordingly, limited digestion of native chromatin releases nucleosome-depleted fragments. Sequencing and mapping of these fragments identifies DNaseI-hypersensitive sites (DHSs) corresponding to regulatory regions. Protein and complex binding within a DHS create 6-40bp sequences of DNaseI protection, called digital DNase footprints, which can be called after deep sequencing of DNase-seq libraries (Neph et al. 2012b). Beyond being DNaseI hypersensitive, open chromatin regions are also sensitive to shearing by sonication and this concept is exploited by the FAIRE-seq assay. Initially, chromatin isolated from formaldehyde crosslinked cells is subject to sonication. Subsequently, the DNA from open chromatin regions is isolated from the aqueous phase following phenol-chloroform extraction. Similar to DNase-seq, mapping the FAIRE-seq enriched fragments genome-wide demarcate regulatory elements.

Myriad open chromatin maps are now available due to the adoption of DNase-seq by the ENCODE Project Consortium and Roadmap Epigenomics Program. In total, these consortia produced hundreds of maps encompassing

349 cell types including pluripotent stem cells, stem cell progenitors, cultured primary cells, and human fetal tissue from various types and gestational stages (; (Maurano et al. 2012). Meta-analysis of these datasets determined DHSs span 2.1% of the genome per cell type on average and, impressively, all ~4,000,000 sites collectively cover ~40% of the genome (Maurano et al. 2012). Mapping nucleosome-depleted chromatin is a comprehensive way to identify the global catalog of regulatory elements and factor binding sites without specific antibodies or prior knowledge of cell type specific transcriptional regulators.

Mapping of Higher Order Chromatin Architecture

Lower-order chromatin structures such as the 11nm fiber, also called “beads on a string”, are followed by higher-order structures like the 30nm fiber and 700nm mitotic chromosomes (Felsenfeld and Groudine 2003). Incidentally, genome compaction brings regions that are linearly distant via short and long-range DNA-DNA interactions. Mapping the 3D structure of the nucleus is important because, like histone marks and chromatin accessibility, chromosome conformations influence mammalian gene regulation. For example, chromatin in close proximity to the lamina of the inner nuclear membrane, or lamin-associated domains (LADs), tends to be heterochromatic and transcriptionally repressed (Akhtar et al. 2013; Guelen et al. 2008). In contrast, high local concentrations of RNA polymerase II bound promoters, called transcription factories, correlate with robust gene expression (Brown et al. 2008). Historically, fluorescent

microscopy imaging, though limited by resolution and throughput, has served as the gold standard for observing nuclear structure. In fact, both LADs and transcription factories were discovered in this fashion.

A paradigm shift from FISH based methods towards measuring physical DNA interactions has revolutionized our ability to broadly map nuclear architecture. The introduction of chromosome conformation capture (3C) by Dekker and colleagues provided an alternative to mapping of distances between two loci (Dekker 2002). Succinctly, chromatin from formaldehyde crosslinked cells is digested with a restriction enzyme followed by DNA ligation under extremely dilute conditions to favor joining of ends in close proximity to each other. After reversal of crosslinks and DNA purification, the ligation frequency between two restriction fragments, measured by qPCR, indicates their interaction frequency. Such interaction frequency is generally related to the spatial distance, though this relationship could be complex, non-linear, and influenced by chromatin accessibility (Dekker et al. 2013; Williamson et al. 2012).

Differing in the number of loci tested and selection of loci, the suite of 3C-based assays include 3C, circular chromosome conformation capture (4C), chromosome conformation capture carbon copy (5C), tethered conformation capture (TCC), ChIA-PET and Hi-C (de Wit and de Laat 2012; Sajan and Hawkins 2012; van Steensel and Dekker 2010). Of these, 3C and 5C are locus-centric methods, meaning the assayed regions are selected *a priori*. Sequence specific primers are designed for each locus of interest, for instance promoters or cis-

regulatory elements. As stated previously, 3C measures the interaction frequency of one locus with another (one-to-one), between an anchor and a bait sequence, by qPCR. 5C attains higher throughput capacity by utilizing thousands of anchor and bait primers that can span an entire chromosome (many-to-many) (Dostie et al. 2006). The 4C method measures the genome-wide interaction frequency of a single anchor site. Inverse PCR of ligated and circularized interacting DNA fragments detects all interacting loci (one-to-all) (Zhao et al. 2006).

Hi-C measures the entire genome's interaction frequency with itself as a large matrix without enrichment for specific loci (all-to-all) (Lieberman-Aiden et al. 2009). The key innovation of Hi-C is the ability to enrich for ligation junctions of interacting fragments. After digestion of crosslinked chromatin with a 6-base cutter, the overhang ends are filled in with a biotinylated base. As with all 3C-based assays, the DNA fragments are extremely dilute in the presence of ligase to promote intramolecular ligation events. The ligated sample is sonicated and a streptavidin pulldown captures all junctions of interacting DNA fragments, which are then sequenced. Hi-C interaction signal-to-noise ratio is dependent on the rate of intramolecular ligation events. TCC, a Hi-C variation, performs the ligation step on the streptavidin bead to promote these favorable ligation events (Kalhor et al. 2012). ChIA-PET is another variation of Hi-C which features an immunoprecipitation step to map DNA interactions involving a POI (Fullwood et al. 2009). This approach has been useful for understanding proteins involved in

nuclear organization such as TFs, RNA polymerase, CTCF, and cohesin (Demare et al. 2013; Fullwood et al. 2009; Gerstein et al. 2012; Handoko et al. 2011; Li et al. 2012).

Due to its unbiased genome-wide scale, Hi-C promotes the discovery of novel interactions and structures. The size or level of nuclear architecture assayed with Hi-C is dependent on experimental resolution, which in turn is limited by both restriction enzyme cutting frequency and sequencing depth. At 1Mb resolution, Dekker and colleagues defined higher-order chromosome compartments that average about 5Mb in size across the genome (Lieberman-Aiden et al. 2009). Termed A and B compartments, these sequentially interchanging structures correlate with euchromatic and heterochromatic genomic regions, respectively, in a cell-type specific manner. With increase sequencing depth, we attained 40kb resolution Hi-C interaction data. In this way, our lab identified cell-type and species invariant 1 Mb regions of high local interaction frequency, called topological domains or topologically associated domains (TADs), separated by non-interacting boundary elements (Dixon et al. 2012). Boundary regions correlate with the limits of heterochromatin blocks and are enriched for CTCF binding sites, housekeeping genes, and certain transposon elements. It is may be possible to reach fragment-length resolution, or about 4kb with 6-base cutters, using Hi-C to recognize interacting regulatory elements with deeper sequencing and improved data analysis algorithms.

What has become clear using microscopy and now 3C based assays is a gradient or spectrum of DNA folding architecture (Figure 3). These interactions start as one-to-one looping of regulatory elements including enhancers, promoters, and insulators. A collection of these interactions flanked by boundary regions form local interacting neighborhoods or TADs (Dixon et al. 2012; Nora et al. 2012). Using 5C, sub-TADs were recently recognized as finer resolution segments of TADs which can be cell-type specific (Phillips-Cremins et al. 2013). Together many TADs contribute to a larger chromosome compartment and, finally, to an entire chromosome territory (Lieberman-Aiden et al. 2009). In summary, the advent of 3C technologies has revealed a novel and complex layer of genome annotation. Many questions remain regarding how nuclear architecture is regulated and how it mechanistically influences transcription, cellular identity, and possibly disease.

Utility of Epigenome Maps

In addition to rapid development of new epigenome mapping technologies, another important factor fueling the remarkable progress of epigenomics is the formation of many research consortia worldwide. Most notable among these consortia are the US Roadmap Epigenome Project, the ENCODE project, and the International Human Epigenome Consortium (IHEC) (Table 1). Modeling after the hugely successful Human Genome Project (Lander 2011), these consortia standardized experimental protocols, recommended data analysis procedures,

and most importantly, publicly released large amounts of data sets prior to publication. Consequently the number of epigenome maps generated has grown exponentially, from a handful in early 2007 to several thousands as of today. As detailed below, these maps have proved to be most valuable in annotation of the transcription units, cis-regulatory sequences, and other genomic features in the human genome. Integrative analysis of epigenomic maps has facilitated study of the gene regulatory programs involved in pluripotency, adipogenesis, and cardiomyocyte differentiation (Gifford et al. 2013; Hawkins et al. 2010; Mikkelsen et al. 2010; Pastor et al. 2011; Rada-Iglesias et al. 2012; Wamstad et al. 2012; Williams et al. 2011; Xie et al. 2013b; Zhu et al. 2013). Further, comparison with genome-wide association studies (GWAS) revealed enrichment of disease associated sequence variants in putative cis-regulatory elements, providing insights into the pathogenesis of many common human diseases.

Annotation of cis-Regulatory Elements from Chromatin Profiles

A major challenge confronting biomedical researchers is the absence of functional annotation of the human genome, where the vast majority (98.5%) do not code for proteins yet harbor most of the disease-associated genetic variations. Several types of functional sequences are known to exist in the non-coding parts of the human genome including cis-regulatory elements. These sequences, including promoters, enhancers, and insulators, govern gene expression by recruiting sequence specific TFs that modulate local chromatin

structure and assembly of transcriptional machinery. Since these functional DNAs lack consistent and recognizable sequence features, their identification and characterization had until recently been quite difficult.

Epigenome maps have proven to be a powerful tool for annotating the functional features in the human genome. This is possible because several classes of DNA elements are associated with characteristic histone modifications (Table 3). Early studies revealed that H3K4me3 predominantly associate with promoters and H3K36me3 with gene bodies. Thus, new gene units could be identified with the use of chromatin profiles. Indeed, this strategy led to the identification of several thousands of long non-coding intergenic RNA genes (lincRNA) (Guttman et al. 2009). Similarly, our lab showed that transcriptional enhancers are characterized by the presence of H3K4me1 but not H3K4me3, a combination that accurately predicted tens of thousands of new enhancers in the human genome (Heintzman et al. 2007; Heintzman and Ren 2009). Despite initial hesitation that this signature may not universally demarcate enhancers in different cell types or in more than one species, H3K4me1 marks enhancers in all tested cell types from human to zebrafish to fly (Aday et al. 2011; Heintzman and Ren 2009; modENCODE Consortium et al. 2010). Importantly, as enhancers have been annotated in many human cell types a persistent feature of enhancers is their cell type specificity consistent with their role in determining cellular identity. Other general properties of enhancers that are validated genome-wide, including DNaseI hypersensitivity, combinatorial TF binding, H3.3 and H2A.Z

histone variant enrichment, bound RNA Pol II, and RNA production (eRNAs) (Buecker and Wysocka 2012). Combinations of these features along with H3K4me1 allow refined and highly accurate genome-wide enhancer prediction (Rajagopal et al. 2013).

Initially, chromatin states in two human cell lines, HeLa cells and K562 cells, were mapped and used to predict 55,000 candidate enhancers in the human genome (Heintzman et al. 2009). This study demonstrated that chromatin modifications at enhancers, in particular H3K4me1 and H3K27ac, are cell type specific and correlate with cell type specific gene expression throughout the genome, suggesting a potentially critical role for enhancers in lineage-specific gene regulation. Subsequent studies confirmed this result in additional cell types. Rada-Iglesia and colleagues characterized the chromatin modification profiles in the human embryonic stem cells (hESC) and identified approximately 7,000 candidate enhancers featuring binding of TFs, presence of H3K4me1 mark and depletion of nucleosomes (Rada-Iglesias et al. 2011). Interestingly, they also classified these enhancers into “active enhancers” and “poised enhancers”, which differ mainly in the presence or absence of H3K27ac mark. Active enhancers are near genes expressed in hESCs, while poised enhancers are next to genes inactive in hESC but turned on during differentiation (Rada-Iglesias et al. 2011). Independently, Creighton and colleagues, by examining the chromatin modification patterns in the mESCs and several differentiated mouse cell types, also found that H3K27ac can distinguish the active enhancers from poised ones

(Creyghton et al. 2010). While tens of thousands of enhancers are typically marked in a cell type, a small subset of enhancers (<1%) form large domains up to 50kb marked by H3K4me1, H3K27ac, mediator, and master transcription factor binding (Whyte et al. 2013). These “super-enhancers” are posited to regulate key genes important for cell identity. For example, in mESCs super-enhancers are associated with genes necessary for pluripotency and in myotubes they are associated with skeletal muscle development. Lastly, enhancers are turned off or “decommissioned” by the removal of H3K4me1 and H3K27ac marks when they no longer pertain to the cellular state. For example, enhancers that control genes important for pluripotency must be decommissioned for proper differentiation (Whyte et al. 2012). By profiling key chromatin marks in different cell types, researchers can now annotate the location and activity state of regulatory elements across the genome to define a cell-type’s regulome.

Extending the idea that chromatin modification patterns are associated with different functional sequences, a comprehensive epigenomic study of nine human cell lines showed that the human genome could be segmented into regions carrying one of 15 different combinations of chromatin modification marks in each cell type (Ernst et al. 2011). The authors implemented a multivariate hidden markov model called ChromHMM to unbiasedly infer these chromatin states from the chromatin modification profiles. This approach found that each chromatin state corresponds to a specific category of genomic features, including active or poised promoters, enhancers, insulators and silenced domains

(Table 3). Similar results have been obtained with other machine-learning algorithms (Hoffman et al. 2013; Won et al. 2013). More recently, the ENCODE consortium profiled the chromatin modification state in 46 human cell types, and found that as much as 56% of the genome is associated with specific histone modification patterns indicative of biochemical activities (Consortium et al. 2012). Ongoing studies from large epigenome mapping consortia such as the NIH Epigenome Roadmap consortium are certain to enhance human genome annotation even further (Bernstein et al. 2010).

Annotation of Long-Range Chromatin Interactions

Rather than the regulome simply being a collection of elements, it must ultimately link together enhancers, insulators, promoters, and other features in 3-dimensional space. Epigenomic maps can annotate cell-type specific long-distance interactions between regulatory elements using two approaches. First, long-range looping can be mapped by physical interaction of distant DNA loci with 3C-based assays. Considering their capabilities, 5C and ChIA-PET currently provide the best balance of resolution and reasonable coverage in the human genome for this purpose (Dekker et al. 2013; Smallwood and Ren 2013). Several 5C and ChIA-PET datasets exist which are valuable for identifying interactions in selected human cell types, but diverse interaction maps are not yet available. The second method correlates regulatory elements and target promoters across many cells types to computationally infer pairs that regulate and likely interact with

each other. These *in silico* predictions require the annotation of cell-type specific enhancers and active promoters with chromatin signatures, DHSs, and expression profiles (Ernst et al. 2011; Sheffield et al. 2013; Shen et al. 2012; Thurman et al. 2012). A clear advantage of the correlative approach is the necessary data in hundreds of cell types had already been produced by epigenomic consortia. For example, Thurman *et al.* defined almost 600,000 regulatory pairs from 79 cell types by correlating ENCODE DNase-seq and RNA-seq data (Thurman et al. 2012). Some of the predicted interaction correlations defined by this approach were validated by both 3C and ChIA-PET but more extensive cross-validation is necessary. Furthermore, such results are limited to enhancer-promoter communication and cannot address the role of insulators or novel interacting regions.

While these two approaches use different concepts to define chromatin interactions, they independently come to similar conclusions about nuclear architecture. The main theme is consistent with the role of enhancers in the determination of cellular identity: enhancer-promoter interactions are cell-type specific (Sanyal et al. 2012; Shen et al. 2012). As previously suggested from single locus studies, enhancers target promoters at long distances and do not always target the nearest gene (Ong and Corces 2011). This is confirmed by a broader study using 5C analysis of 1% of the human genome in three cell types which determined that ~50% of distal regulatory elements interact with the closest active gene (Sanyal et al. 2012). Surprisingly, enhancers and promoters

can interact promiscuously with more than one element indicating a complicated web of transcriptional regulation (Sanyal et al. 2012; Shen et al. 2012; Thurman et al. 2012). In fact, Thurman *et al.* use a correlative approach to find half of TSSs regulate 10 or more distal sites and half of putative enhancers regulate more than one TSS (Thurman et al. 2012). 5C physical interaction data also confirms 49% of TSSs interact with more than one distal site but finds only 10% of enhancers interacting with more than one promoter (Sanyal et al. 2012). Another concept challenged by interactome annotations, is the classic definition of CTCF-bound insulators as having an enhancer-blocking function to limit the range of targeted enhancer activation to fine tune transcription (Phillips and Corces 2009). Notably, enhancer-promoter interactions often span hundreds of kilobases surpassing one or more CTCF sites (Demare et al. 2013; Sanyal et al. 2012). 5C analysis predicts that almost 60% of all interactions skip a co-bound CTCF/cohesin locus (Sanyal et al. 2012). Another unexpected result is frequent insulator interactions with promoters and enhancers suggesting a possible role for CTCF in transcriptional activation (Demare et al. 2013; Handoko et al. 2011; Mali et al. 2013; Sanyal et al. 2012). Finally, there is an emerging role of CTCF and cohesin co-bound regulatory elements in long-range, constitutive DNA interactions in contrast to cohesin and mediator co-bound regions which facilitate short range, cell-type specific enhancer-promoter interactions (Demare et al. 2013; Phillips-Cremins et al. 2013; Sanyal et al. 2012).

On the whole, great progress has been made towards understanding how chromatin and nuclear organization jointly regulate global gene expression patterns. Still, it remains to be determined how up to 20 enhancers choose to target a particular promoter, what the functional consequence is for a promoter to be activated by more than once enhancer, and if these results are merely an artifact of averaging interactions and epigenetic maps over a population of cells. Evidence that enhancer interactions frequently bypass insulators and that CTCF can loop to transcriptional start sites suggests we must refine our understanding of CTCF-bound insulators. Certainly the conclusions from these epigenomic studies will encourage mechanistic studies to uncover the details of transcriptional regulation by long-range interactions. Continuation of this work at high resolution and in many cell types will progress the dimensions of chromatin maps from a linear chart to a 3-dimensional model of genomic annotation.

Dynamic Chromatin Landscapes During Human Development

Advances in the robustness, throughput, and accuracy of epigenomic technologies over the past decade have co-occurred with a revolution in stem cell biology and regenerative medicine. The first completed epigenomes documented the features of cultured, immortalized or cancerous cell lines. While this enabled correlation of epigenomic features with genomic elements, the resulting epigenomic landscapes were a static view. Now that lineage-specific

differentiation protocols produce populations of cells with increasing purity, it is possible to study epigenomic dynamics during development. Such studies give insight into how cell state transitions are influenced by chromatin states at promoters and enhancers, DNA methylation dynamics at enhancers, and the expansion of repressive domains. Finally, using chromatin signatures to define lineage specific enhancers can lead to models of TF networks that regulate cell type specification.

Unique chromatin signatures mark promoters and enhancers in pluripotent cells. Some promoters in ESCs are bivalent or co-marked with active and repressive histone marks. Sequential ChIP confirmed the presence of both H3K4me3 and H3K27me3 on the same DNA molecule or promoter allele (Bernstein et al. 2006). Due to polycomb-mediated silencing, bivalent genes are not expressed and specifically mark developmental genes that are activated in downstream cell states and are “poised” in ESCs. Consistent with this idea, differentiation of hESCs into endoderm, mesoderm, and ectoderm lineages resolves 85% of bivalent promoters into monovalent states in a lineage-specific manner (Gifford et al. 2013). H3K4me1/H3K27me3 marked poised enhancers are a special class found mostly in ESCs and tend to regulate bivalent promoters of developmental genes (Rada-Iglesias et al. 2011). Chromatin state maps can have more informative power than expression data alone, because they can identify both active genes and poised genomic elements that foreshadow a cell type’s differentiation potential.

DNA methylation can also be informative for identifying enhancers and classifying their activity. Quantitative comparisons of the first human methylomes in hESCs and fibroblasts indicated regions with dynamic DNA methylation levels. A region that is relatively hypermethylated in one cell type and hypomethylated in another is called a differentially methylated region (DMR) (Lister et al. 2009). DMRs are enriched at regulatory elements as evidenced by their overlap with DNaseI sites, TF binding sites, and enhancer chromatin marks (Hon et al. 2013; Ziller et al. 2013). Active enhancers correlate with the hypomethylated DMR state, also called low-methylated regions (LMRs), and the formation and maintenance of LMRs is dependent on TF binding (Stadler et al. 2011). However, whether the depletion of methylation upon TF binding occurs passively during cell cycles or through active demethylation is still unclear. Recently generated methylomes of cytosine variants find 5hmC, 5fC, and 5caC enriched at enhancers in ESCs suggest the latter is possible (Pastor et al. 2011; Shen et al. 2013; Szulwach et al. 2011; Yu et al. 2012). Even inactive enhancers can be identified by hypomethylation. We recently discovered “vestigial enhancers”, defined as regions depleted of DNA methylation and enhancer chromatin marks in adult tissues but which exhibit enhancer activity earlier in development (Hon et al. 2013). These results suggest that the methylome retains a memory of its previous cellular identities, although how and why is still unclear. Overall, dynamic DNA methylation patterns at enhancers provide yet another epigenetic feature for their annotation. In fact, unexpected regions of hypomethylation

in transposable elements function helped to annotate these regions as tissues-specific enhancers (Xie et al. 2013a).

Cellular differentiation during development is characterized by gradual expansion of repressed domains. In ESCs, H3K27me3 is distributed across 8% of the genome and is most commonly observed at bivalent promoters of developmental genes (Zhu et al. 2013). First observed in IMR90 fibroblasts but since confirmed in many differentiated cell types, H3K27me3 peaks spread from bivalent promoters in ESCs to form large blocks covering up to 40% of the genome in differentiated cells (Hawkins et al. 2010; Zhu et al. 2013). These large repressed regions specifically span developmental genes, and presumably function to silence genes that do not pertain to the specified lineage. Correspondingly, chromatin accessibility at regulatory DNA is progressively lost during differentiation (Stergachis et al. 2013). The epigenome of ESCs is uniquely open and accessible, which is consistent with the role of these cells as pluripotent progenitors of all three germ layers. It is now clear that during differentiation and development, the epigenome is progressively restricted like its plasticity potential.

Another major utility of epigenomic studies is the ability to identify cell state or stage specific master regulators and construct transcriptional networks. DNase-seq based methods are useful for mapping all TF binding sites genome-wide in one assay. DNase footprint analysis in 41 human cell types from ESCs to primary adult cells identified 45 million footprints and subsequent *de novo* motif

analysis discerned over 600 unique motifs (Neph et al. 2012b) . This set of motifs recovered 90% of the TRANSFAC, JASPAR, and UniPROBE motif database entries. However, nearly half of the motifs found in this study are novel and bound by an unknown protein indicating there is a substantial amount of work to be done to have a complete understanding of sequence-dependent transcriptional regulators. Similar to this approach, master transcriptional regulators can be identified by annotating putative enhancers using histone marks signatures followed by motif analysis of these regions. In this way, novel master regulators of human neural crest (NR2F1), cardiac development (Meis2), and adipogenesis (PLZF) have been successfully identified (Mikkelsen et al. 2010; Paige et al. 2012; Rada-Iglesias et al. 2012). Enhancer-driven regulatory networks can be constructed by correlating TF expression data with motif analysis at putative enhancers as predicted by histone marks or open chromatin (Ernst et al. 2011; Neph et al. 2012a). Alternatively, TF networks can be directly assayed using ChIP-seq. The ENCODE consortium characterized the binding patterns of 119 TFs in five cell lines by ChIP-seq which yielded ordered predictions of TF hierarchies, rules of combinatorial TF binding, and TF binding preferences for distal or proximal sites (Gerstein et al. 2012). However with over 1000 TFs and many cell types still not assayed, the complexity of experimentally and computationally generated transcription factor regulatory networks will increase and requires further investigation.

Understanding Disease Variants with Epigenomics

The combination of data from the Human Genome Project and International HapMap Project enabled geneticists to study polygenic traits and diseases using GWAS (Frazer et al., 2009). Investigating the genetic basis of common diseases is possible by testing for variants that significantly associate with cases over controls in a population. GWAS are now a common tool in genetic epidemiology with over 1,600 studies published associating 11,000+ SNPs with hundreds of distinct diseases and traits (Welter, 2013). It is often difficult when interpreting GWAS results to identify the functional variant(s) underlying an association. The difficulty arises from the fact that a SNP associated with a particular phenotype may itself be the functional variant directly causing or contributing to the phenotype, or it may simply be in linkage disequilibrium (LD) with the functional variant. Traditional GWAS alone cannot differentiate between these two scenarios. Moreover, even after identifying a true functional variant, it can be very difficult to determine the molecular mechanism by which that variant alters genome function and contributes to disease. For example, 93% of SNPs associated with human phenotypes by GWAS are located outside of protein coding regions (Maurano et al, 2012). This finding highlights the need for annotation of the vast noncoding sequence of the human genome. Epigenomic maps are likely to play a major role in identifying specific functional variants that cause or contribute to human disease.

Assimilation of GWAS and epigenomic data can give clues about the involvement of regulatory elements and their target genes in the pathogenesis of common disease. Annotating enhancers by their chromatin state in nine cell types, Ernst et al. found that disease-associated SNPs are enriched within cell-type-specific enhancers from the appropriate disease cell type (Ernst et al., 2011). By correlating the locations of over 5,000 noncoding common disease-associated SNPs with almost 4 million DHSs from hundreds of cell types, Maurano et al. found 40% of GWAS SNPs fall within these regions (Maurano et al., 2012). Due to the cell-type specific property of regulatory elements, correlation of all GWAS SNPs in DHSs from a specific disease can facilitate de novo prediction of pathogenic tissues. Using this approach, GWAS SNPs associated with Crohn's disease are enriched in the cell-type specific DHSs of TH17 and TH1 cells, and both of these cell types have been independently associated with this disease (Maurano et al., 2012).

Variants can affect both local open chromatin and gene expression and these loci are called DNaseI sensitivity quantitative trait loci (dsQTLs) (Degner et al., 2012). Identification of dsQTLs in haplotyped lymphoblastoid cells from 70 individuals found dsQTLs frequently fall within TF-binding sites, affect nucleosome positioning as determined by MNase-seq, and overlap with GWAS disease-associated SNPs (Degner et al., 2012 and Gaffney et al., 2012). Interestingly 80% of dsQTLs lie in chromatin regions previously predicted to be

functional in lymphoblastoid cell lines by epigenetic marks, including 41% in predicted enhancers, suggesting they may be the functional variant.

Reference epigenomes in diverse tissues and the annotation of millions of putative enhancers can help researchers make hypotheses about which SNPs in LD with GWAS SNPs may contribute to disease. Following up GWAS studies by genotyping regulatory elements around GWAS SNPs may more definitively identify functional SNPs. Importantly, disease-associated SNPs that reside in putative regulatory elements must be functionally validated. SNPs in enhancers may affect their activity by disrupting TF binding, coactivator recruitment, looping interactions with target promoters, or transcription of the target promoter. To this end, each of these possible outcomes can be tested with traditional approaches such as reporter assays, gel shift assays, and allele-specific ChIP. In addition, emerging technologies such as TALENs, CRISPR/Cas9, 3C-based technologies, and new massively parallel reporter assays promise to vastly improve the rigor and throughput with which we can test these hypotheses (Melnikov et al., 2012 and Patwardhan et al., 2012).

The Road Ahead

Six decades ago, Watson and Crick put forward a model of DNA double helix structure to elucidate how genetic information is faithfully copied and propagated during cell division (Watson and Crick, 1953). Several years later, Crick famously proposed the “central dogma” to describe how information in the

DNA sequence is relayed to other biomolecules such as RNA and proteins to sustain a cell's biological activities (Crick, 1970). Now, with the human genome completely mapped, we face the daunting task to decipher the information contained in this genetic blueprint. Twelve years ago, when the human genome was first sequenced, only 1.5% of the genome could be annotated as protein coding, whereas the rest of the genome was thought to be mostly “junk” (Lander et al., 2001 and Venter et al., 2001). Now, with the help of many epigenome maps, nearly half of the genome is predicted to carry specific biochemical activities and potential regulatory functions (ENCODE Project Consortium, et al., 2012). It is conceivable that in the near future the human genome will be completely annotated, with the catalog of transcription units and their transcriptional regulatory sequences fully mapped.

However, to reach this goal a couple of technical issues should be resolved. First, current approaches to epigenomic analysis still demand a large number of cells, limiting the cell types and developmental stages that can be examined. Robust nanoscale techniques for chromatin modification profiling, methylcytosine detection or chromatin accessibility mapping will be necessary to cover the full spectrum of cellular states. A few nanoscale ChIP-seq, whole-genome bisulfite sequencing methods have been reported, enabling epigenomic analysis of rare cell populations such as the germ cells (Adey and Shendure, 2012, Adli and Bernstein, 2011 and Ng et al., 2013). We expect that the application of these methods to the rare cell types and to embryonic stages will significantly

broaden our knowledge of the dynamic human epigenome, and facilitate the discovery of additional cell-type-specific regulatory elements and transcription units. Second, a lot of the epigenome maps currently available are from tissues consisting of heterogeneous cell populations, and do not accurately reflect the epigenomic state of a specific cell type. Today, to isolate individual cell types from a tissue in large quantity for epigenomic analysis by FACS is challenging. With the advance in nanoscale epigenomic analysis technique, it will likely be routine to obtain cell types in small quantity without affecting their cellular state and produce epigenome maps for specific cell populations in a tissue type. Eventually, the solution will likely be single-cell epigenome analysis techniques. Currently, it is not yet possible to profile the chromatin state or DNA methylation status of a single cell, but this could become a reality with further development of single-molecule techniques such as SCAN (single chromatin molecule analysis in nanochannels) (Murphy et al., 2013) or SMRT (single-molecule, real-time) sequencing (Clarke et al., 2009 and Flusberg et al., 2010).

Ultimately, to truly understand how the genome programs human development and how certain sequence variants cause human disease, we ought to be able to predict from the sequence when and at what level a gene is expressed in different cell types. To achieve this goal, we need to overcome substantial challenges in two areas. First, we must gain a better understanding of the functional relationships between DNA methylation, chromatin modification state, or higher order chromatin structure and gene regulation. Epigenomic

studies have established correlations between DNA hypomethylation, certain chromatin modification signature, and chromatin structure at cis-regulatory elements such as enhancers and promoters, but to determine whether an epigenetic state is necessary for transcription or merely coincidental will require additional mechanistic studies. To this end, new technologies are required to allow manipulation of the epigenetic state of specific loci. The recent development of TALE factors and CRISPR/Cas9 as a way to target histone modification enzymes or DNA methyltransferases to specific genomic sites is a great example (Gaj et al., 2013, Ramalingam et al., 2013 and Wang et al., 2013). Second, we need to better define the target genes of the distal regulatory elements such as enhancers. This is a challenge because many enhancers are located hundreds and thousands of base-pairs away from their target genes, and it is not infrequent that an enhancer and its target gene are separated by other irrelevant genes (Smallwood and Ren, 2013). Techniques such as ChIA-PET, 4C, 5C, and Hi-C have been invented to identify looping interactions between enhancers and target promoters (de Wit et al., 2013, Li et al., 2012, Sanyal et al., 2012 and Wei et al., 2013). However, it is currently unclear whether mere spatial proximity is sufficient for functional regulation. Thus, mapping physical interactions alone is unlikely to be adequate to resolve the target genes for enhancers. An alternative, complementary approach defines target genes for an enhancer by searching for nearby genes sharing similar chromatin state or accessibility across diverse cell types (Ernst et al., 2011, Shen et al., 2012 and

Thurman et al., 2012). Despite its simplicity, this approach does not work particularly well when a gene is regulated by multiple tissue-specific enhancers in different tissues. A hybrid approach combining both spatial proximity and chromatin state information will likely more accurately define the target genes for enhancers.

In summary, the last few years have witnessed an explosion of the epigenomic field. Thousands of epigenome maps from hundreds of human cell or tissue types have been produced, adding crucial insights into the second dimension of the genomic sequence. Although these human epigenomes have illuminated potentially functional elements in the genome and improved our understanding of the human developmental programs, it is also clear that much more remains to be explored. Better characterization of epigenome variations in human populations and in patients will be critical for us to fully appreciate the epigenetic factors in human health and disease etiology. Indeed, projects are already underway to profile thousands more epigenomes from both healthy and diseased individuals (American Association for Cancer Research Human Epigenome Task Force and European Union Network of Excellence Scientific Advisory Board, 2008 and Bernstein et al., 2010). We anticipate an even greater revolution in our understanding of the human epigenome in the coming years.

FIGURES

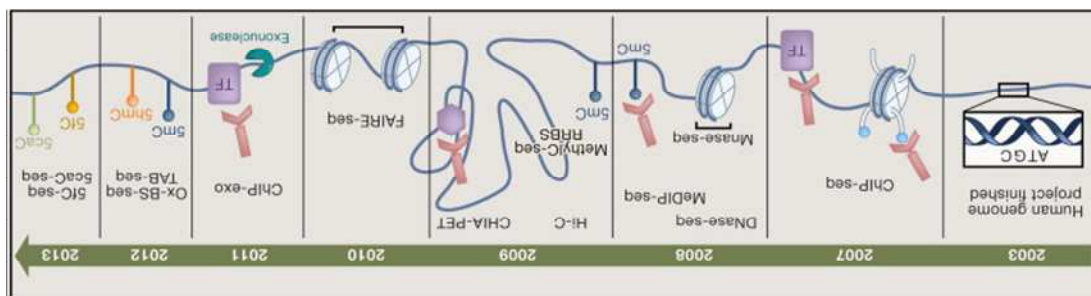


Figure 1. Timeline of Sequencing-Based Technologies for Mapping Human Epigenomes.

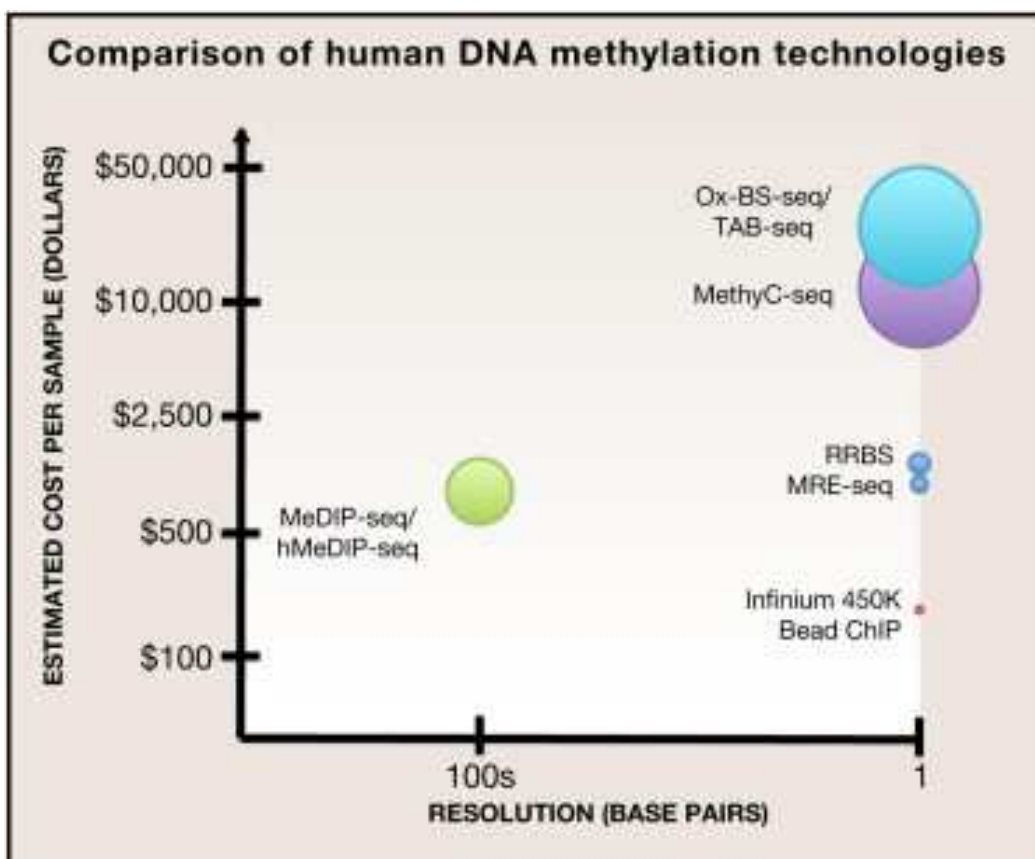


Figure 2. Comparison of Human DNA Methylation Mapping Technologies
The size of the circle is proportional to the genome-wide CpG coverage of the technology.

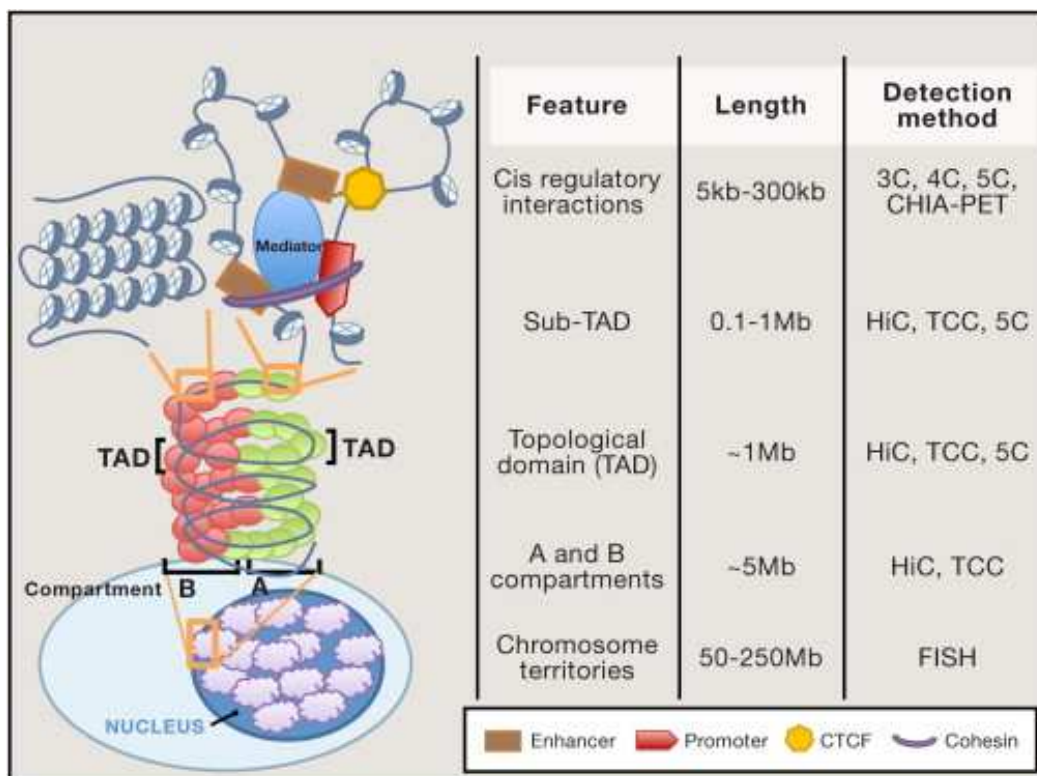


Figure 3: Hierarchical Principles of Nuclear Organization.

TABLES

Table 1. Large-Scale National and International Epigenomic Consortia.

Project Name	Start Date	Affiliations	Completed and Expected Data Contributions	Selected Publication	Access Data
Encyclopedia of DNA Elements	2003	NIH	Dnase-seq, RNA-seq, ChIP-seq, and 5C in 100s of primary human tissues and cell lines	ENCODE Project Consortium et al., 2012	http://encodeproject.org/ENCODE/
The Cancer Genome Atlas (TCGA)	2006	NIH	DNA methylomes in 1,000s of patients samples from more than 20 cancer types	Garraway and Lander, 2013	http://cancergenome.nih.gov/
Roadmap Epigenomics Project	2008	NIH	Dnase-seq, RNA-seq, ChIP-seq, and Methy/C-seq in 100 s of normal primary cells, hESC, and hESC derived cells	Bernstein et al., 2010	http://www.epigenomebrowser.org/
International Cancer Genome Consortium (ICGC)	2008	15 countries, includes TCGA	DNA methylation profiles in thousands of patient samples from 50 different cancers	The International Cancer Genome Consortium, et al., 2010	http://dcc.igcc.org/web
International Human Epigenome Consortium (IHEC)	2010	7 countries, includes BLUEPRINT, Roadmap	Goal: 1,000 Epigenomes in 250 cell types	American Association for Cancer Research Human Epigenome Task Force; European Union, Network of Excellence, Scientific Advisory Board, 2008	http://ihec-epigenomes.org

Table 2. Major Conceptual Advances Convey the Utility of Epigenome Maps.

Before Next-Gen Sequencing	The Next-Gen Sequencing Era	Future ^a
DNA Methylation		
Repressive mark at imprinted loci, transposons, and in x chromosome inactivation	Metastable mark	Comprehensive definition of epigenomic variation across all human cell types
Found in active gene bodies	Active DNA demethylation by TET family proteins through 5hmC, 5fC, and 5caC	Define epigenomic variation in populations
Only dynamic in primordial germ cells and during early embryogenesis; in CpG context	Non-CpG methylation exists especially in ESCs oocytes, and adult brain	Discovering epigenetic signatures of disease
	Tissue-specific methylation at distal regulatory elements	Technology advances to enable single-cell methylomes
Histone Modifications		A deeper understanding of epigenomic inheritability and reprogramming
Marks that correlate with promoters and gene bodies	Over 130 different histone modifications have been identified	High-throughput functional validation of predicted enhancers
H3K9me3 is a mark of heterochromatin	Identification of novel ncRNAs by promoter and gene body chromatin signatures	Epigenome engineering
H3K27me3 is a mark of facultative heterochromatin	Active enhancers are marked by H3K27ac or H4K16ac	Improved computational analysis and visualization tools
Proposal of the Histone Code Hypothesis	Poised enhancers are marked by H3K4me1 alone or in combination with H3K27me3	
Bivalent Promoters marked by H3K4me3/ H3K27me3	Expansion of repressive chromatin blocks during differentiation	
Unique chromatin signature of enhancers defined as H3K4me1	Combinations of chromatin marks define a limited number of chromatin states	
Chromatin Structure		
Nucleosome maps only in yeast, fly	Nucleosomes mapped in the human genome	
DHSs correlate with TF-binding sites and regulatory elements	Nucleosomes are well positioned around regulatory regions	
	DHSs predict enhancer-promoter pairs and cell types affected in disease	
Nuclear Architecture		
Identification of chromosome territories, LADS, and transcription factories with FISH	Identification of sub-TADs, TADs and chromosome compartments	
Few validated enhancer-promoter interacting pairs	Chromosome-wide maps of enhancer, promoter, and insulator interacting pairs	
^a Text in this column refers to DNA Methylation, Histone Modifications, Chromatin Structure, and Nuclear Architecture.		

^aText in this column refers to DNA Methylation, Histone Modifications, Chromatin Structure, and Nuclear Architecture.

Table 3. Distinctive Chromatin Features of Genomic Elements

Functional Annotation	Histone Marks	References
Promoters	H3K4me3	Bernstein et al., 2005; Kim et al., 2005; Pokholok et al., 2005
Bivalent/Poised Promoter	H3K4me3/H3K27me3	Bernstein et al., 2006
Transcribed Gene Body	H3K36me3	Barski et al., 2007
Enhancer (both active and poised)	H3K4me1	Heintzman et al., 2007
Poised Developmental Enhancer	H3K4me1/H3K27me3	Creyghton et al., 2010; Rada-Iglesias et al., 2011
Active Enhancer	H3K4me1/H3K27ac	Creyghton et al., 2010; Heintzman et al., 2009; Rada-Iglesias et al., 2011
Polycomb Repressed Regions	H3K27me3	Bernstein et al., 2006; Lee et al., 2006
Heterochromatin	H3K9me3	Mikkelsen et al., 2007

AUTHORS' CONTRIBUTIONS

CMR and BR equally conceived and wrote this review.

ACKNOWLEDGEMENTS

We apologize to those authors whose works are not covered here owing to space limitations. We thank Peter Jones, Bradley Bernstein, Fred Tyson, John Satterlee and members of the Ren Lab for their comments on earlier versions of this manuscript. This work is supported by funds from the Ludwig Institute for Cancer Research, NIH (grants U01ES017166 and U54 HG006997), and the California Institute for Regenerative Medicine (RN2-00905) to B.R. C.M.R. was supported in part by the UCSD Genetics Training Program through an institutional training grant from the National Institute of General Medical Sciences, T32 GM008666.

Chapter 1, in full, is an adaptation of materials that appears in
Rivera, CM, Ren, B. Mapping Human Epigenomes. *Cell*. 2013 Sept.

CHAPTER 2 - CRISPR REVEALS A DISTAL SUPER-ENHANCER REQUIRED FOR SOX2 EXPRESSION IN MOUSE EMBRYONIC STEM CELLS

ABSTRACT

The pluripotency of embryonic stem cells (ESCs) is maintained by a small group of master transcription factors including Oct4, Sox2 and Nanog. These core factors form a regulatory circuit controlling the transcription of a number of pluripotency factors including themselves. Although previous studies have identified transcriptional regulators of this core network, the *cis*-regulatory DNA sequences required for the transcription of these key pluripotency factors remain to be defined. We analyzed epigenomic data within the 1.5 Mb gene-desert regions around the Sox2 gene and identified a 13kb-long super-enhancer (SE) located 100kb downstream of Sox2 in mouse ESCs. This SE is occupied by Oct4, Sox2, Nanog, and the mediator complex, and physically interacts with the Sox2 locus via DNA looping. Using a simple and highly efficient double-CRISPR genome editing strategy we deleted the entire 13-kb SE and characterized transcriptional defects in the resulting monoallelic and biallelic deletion clones with RNA-seq. We showed that the SE is responsible for over 90% of Sox2 expression, and Sox2 is the only target gene along the chromosome. Our results support the functional significance of a SE in maintaining the pluripotency transcription program in mouse ESCs.

INTRODUCTION

Sox2 is one of the three core transcription factors (Oct4, Sox2, Nanog) responsible for maintaining ESC pluripotency. These core pluripotency factors form auto-regulatory loops and transcriptionally induce a cohort of other key pluripotency genes (Boyer et al. 2005; Chambers and Tomlinson 2009; Chen et al. 2008). Besides ESCs, Sox2 is also expressed in various types of adult stem cells (Arnold et al. 2011; Sarkar and Hochedlinger 2013). Recently, it has been reported that Sox2-expressing cancer stem cells may drive tumor initiation and relapse in medulloblastoma and skin tumor (Boumahdi et al. 2015; Vanner et al. 2014). Therefore, elucidating the transcriptional regulation mechanisms of Sox2 gene is important for the understanding of both pluripotency and tumorigenesis.

Enhancers play a critical role in regulating metazoan gene transcription (Bulger and Groudine 2011; Jin et al. 2014b; Levine et al. 2014; Ong and Corces 2011). Recent genome-wide analyses have revealed that enhancers are very abundant in the genome. Additionally, highly specific enhancer landscapes are responsible for cellular identity as they regulate distinct transcriptional programs in different cell types (Consortium et al. 2012; Jin et al. 2011; Mouse ENCODE Consortium et al. 2012). Clusters of enhancer constituents that together generate a domain of MED1-, and transcription factor-binding, and H3K27ac enrichment have been identified as super-enhancers (SE) or stretch enhancers (Hnisz et al. 2013; Parker et al. 2013; Whyte et al. 2013). Compelling genomic evidence has predicted that SEs play a particularly important role in the control

of cell identity and diseases (Hnisz et al. 2013; Parker et al. 2013; Whyte et al. 2013), but direct functional evidence is lacking.

Until recently, validating the *in vivo* function of a putative enhancer has proven technically difficult. Reporter assays confirm that the sequence can function as an enhancer exogenously, but it does not demonstrate whether a distal target gene can be activated by the enhancer *in vivo*. Knocking out the enhancer sequence is the gold standard approach, but conventional methods using homologous recombination are inefficient and labor-intensive. The recently developed CRISPR/Cas9 system (Cong et al. 2013; Mali et al. 2013) has proven to be a highly efficient genome editing technique and offers a promising method to validate enhancer functions *in vivo*. A recent study used CRISPR to validate the importance of a SE near GATA2 in chronic myeloid leukemia by demonstrating it is responsible for 80% of GATA2 expression (Gröschel et al. 2014). With this recent advance, we have attempted to examine the role of SEs in pluripotency and normal development.

In the mouse genome, Sox2 is located within a ~1.5Mb gene desert region. It has been postulated that regulatory sequences may lie in such gene deserts and modulate gene expression over long distances (Nobrega et al. 2003). In this study, we examined the gene desert near Sox2 and identified a putative distal super enhancer that is marked by active histone marks only in ESCs. This 13kb-long enhancer is occupied by multiple pluripotency factors and forms a long-distance DNA loop with the Sox2 promoter from 100kb away. Using a double-

CRISPR excision strategy, we studied the *in vivo* function of this SE and demonstrated that it is responsible for over 90% of Sox2 gene expression in mouse ESCs. Our results therefore provide direct evidences on the key roles of a SE in governing ESC pluripotency.

RESULTS

Identification of a distal Sox2 super-enhancer in mouse ESCs

We examined ENCODE ChIP-seq data from 23 different mouse tissues or cell types focusing on the Sox2 locus (Figure 4). Using H3K27ac, a histone mark for active enhancers (Creyghton et al. 2010; Rada-Iglesias et al. 2011), we observed a distal SE that is only present in mouse ESC lines, which is approximately 100kb downstream from the Sox2 locus. This sequence spans a relatively large 13kb region, and corresponds to a recently defined super-enhancer (Whyte et al. 2013) or stretch enhancer (Parker et al. 2013). Indeed, this region (referred to as Sox2-SE^{distal} hereafter) is one of the SEs previously identified in mouse ESCs (Whyte et al. 2013). Interestingly, although Sox2 is also expressed in several brain tissues, Sox2-SE^{distal} is not active in those tissues. Instead, a different region upstream of the Sox2 locus is marked by H3K27ac in brain tissues (Figure 4), suggesting that Sox2 is regulated by different mechanisms in ESCs and neural tissues involving different *cis*-regulatory elements. Indeed a part of the region marked by H3K27ac in brain tissues has

been shown to drive Sox2 expression in telencephalic neural stem cells (Zappone et al. 2000).

Epigenetic characterization of the Sox2-SE^{distal} in mouse ESCs

We next examined the epigenetic features at Sox2-SE^{distal}. First, we observed strong H3K27ac signal at this SE in two additional mouse ES cell lines (F123 and J1), indicating that usage of this SE is conserved across different mice strains. Previously published ChIP-seq data shows that Sox2-SE^{distal} is occupied by Oct4, Sox2 and Nanog (Figure 5). In this study, we also observed strong binding of the transcriptional co-activator p300 and the mediator complex (Med1) (Figure 5).

To test if Sox2-SE^{distal} regulates Sox2 gene through long-range DNA looping, we analyzed DNA looping interactions using Hi-C data generated from mouse ESCs (Jin et al. 2014a) and observed specific DNA looping signal between Sox2-SE^{distal} and the Sox2 promoter (Figure 5). This long-range chromatin interaction was also observed in two other independent studies using 5C (Fullwood et al. 2009; Gribnau et al. 2003; Hnisz et al. 2013; Phillips-Cremins et al. 2013; Ram et al. 2011; Splinter et al. 2012; van de Werken et al. 2012a; Whyte et al. 2013) and ChIA-PET (Kieffer-Kwon et al. 2013) in mouse ESCs. Taken together, these results suggest that Sox2-SE^{distal} could be involved in the core pluripotency circuitry through regulating Sox2 gene expression.

A simple and efficient strategy to disrupt enhancers with CRISPR

To test the *in vivo* function of Sox2-SE^{distal}, we sought to delete the entire enhancer sequence from the endogenous locus. Given that the SE spans a ~13kb region, deletion of this region using conventional methods would be very inefficient. Therefore, we explored whether the recently developed CRISPR technology (Cong et al. 2013; Mali et al. 2013) could be utilized to delete this large non-coding sequence in mouse ESCs.

In the CRISPR/Cas9 system, a small guide RNA (sgRNA) targets the Cas9 protein to a specific genomic location and generates a double-strand break. Random mutations can be generated when the DNA break is repaired via non-homologous end joining (NHEJ). The efficiency of DNA cutting by the CRISPR/Cas9 system may reach over 90%, therefore it is possible to introduce multiple mutations in the same cell with high efficiency (Wang et al. 2013). We postulated that if we introduce two sgRNAs flanking the Sox2-SE^{distal} into a cell, two DNA breaks may be generated simultaneously on the same chromosome, leading to the deletion of this long SE during DNA repair by NHEJ (Figure 6A).

We therefore used this simple approach to delete Sox2-SE^{distal} in the hybrid F123 mouse ESC line (see method and materials) derived from a cross between two inbred strains, CAST and 129 (2009; Gribnau et al. 2003) (Figure 6A). This cell line was chosen for its high SNP density and *a priori* haplotypes. Given these characteristics, this cell line is especially amenable to robust allelic analysis of CRISPR-based deletions and downstream functional assays.

After co-transfecting two sgRNA constructs, 8 out of 16 ESC clones had the desired enhancer deletion on at least one allele (Figure 6B). Copy number analysis revealed that Sox2-SE^{distal} was deleted from both alleles (biallelic clones) in 2 clones, and from one allele in the remaining 6 clones (monoallelic clones) (Figure 6C). Further genotyping analysis using Sanger sequencing confirmed that among the 6 monoallelic clones, five had a deletion on the 129 allele and one on the CAST allele (Figure 6D, Figure 9 (Appendix)). Taken together, these results suggest that our double-CRISPR strategy can delete large genomic sequences with very high efficiency (31.25%, see methods). In fact, a recent study also took a similar approach in mouse erythroleukemia cells and achieved a similar deletion frequency (Canver et al. 2014).

Sox2-SE^{distal} is responsible for over 90% of Sox2 transcription

We next performed allele specific RT-qPCR analysis (see methods) to quantify the effect of Sox2-SE^{distal} on Sox2 transcription. In wild type (WT) clones, both CAST and 129 alleles contribute to roughly half of total Sox2 expression (Figure 7A). Strikingly, in the biallelic SE knockout clone, Sox2 expression is almost completely lost, while in monoallelic SE knockout clones, total Sox2 expression level is reduced by roughly 50% (Figure 7A). Importantly, in the monoallelic clones, the level of Sox2 expression from the knockout allele is only ~7% of the intact allele on average (Figure 7A). These results reveal that a single

SE, Sox2-SE^{distal}, regulates transcription exclusively in *cis* and contributes to over 90% of total Sox2 expression.

In order to further test the *cis* and *trans* effects of the enhancer deletion on the transcriptome, we performed RNA-seq in four representative clones with different scenarios of enhancer deletion (WT, biallelic mutation, 129 monoallelic mutation, and CAST monoallelic mutation). Based on the sequences of RNA-seq reads and SNPs called for both parental alleles, we can determine the transcriptional outputs from the 129 or CAST allele for every gene (Figure 7B). Consistent with the RT-qPCR data, the RNA-seq results also supported our conclusion that Sox2-SE^{distal} is required for Sox2 gene transcription. Importantly, none of the other genes on the same chromosomes are directly affected by Sox2-SE^{distal} deletion (Figure 7B-C), suggesting Sox2 is the only target gene of Sox2-SE^{distal}. It is noteworthy that Sox2 is the only active protein-coding gene sharing the same TAD (Dixon et al. 2012) with Sox2-SE^{distal} (Figure 7B).

Sox2-SE^{distal} deletion alters transcription of key pluripotency genes

As expected, Sox2-SE^{distal} deletion affected the proliferation and morphology of mESC culture, due to reduced expression of Sox2 and several other pluripotency factors (Figure 8A-B, Figure 10 (Appendix)). We further analyzed the effects of Sox2-SE^{distal} deletion on the transcriptome using the RNA-seq data in different types of mutant clones (Table 4 (Appendix)). Comparing the biallelic knockout clone to the WT clone, we identified a signature of 136 up-

regulated genes and 142 down-regulated genes (including Sox2) using a 3-fold cutoff (Figure 8C, Table 5-S3 (Appendix)). Importantly, transcription levels of most of these genes also show significant change in the monoallelic SE knockout clones (Figure 8D), although to a lesser extent (Figure 8E). These results suggest that partial loss of Sox2 expression also results in significant *trans* effects on gene transcription in mouse ESCs. 21 out of the 142 down-regulated genes have Sox2 binding sites at promoters, while only 6 up-regulated genes have Sox2 binding at promoters (Figure 8F, Table 7 (Appendix)), suggesting some of these genes are direct targets of Sox2 transactivation. However, we cannot exclude the possibility that some genes are affected by Sox2-SE^{distal} deletion through trans-interactions or other indirect mechanisms.

Gene ontology (GO) analysis revealed distinct functions between up-regulated and down-regulated signature genes upon Sox2-SE^{distal} deletion. Consistent with the role of Sox2 as a core pluripotency factor, down-regulated signature genes are key genes involved in stem cell maintenance and early development, including Nanog, Klf4, Esrrb, Tet1, and Kit. (Figure 8G). On the other hand, up-regulated signature genes are involved in late development and specialized cellular functions such as cell migration and cell communication (Figure 8G).

We also performed an unbiased gene set enrichment analysis (GSEA) comparing our signature genes to hundreds of gene signatures from molecular signature database (MSigDB). Remarkably, we found that the Sox2-SE^{distal}

knockout signature anti-correlates with the signature in mouse ESCs lacking the Suz12 gene (Pasini et al. 2007). Suz12 is a key component of polycomb repressive complex 2 (PRC2), which has been shown to be required for ESC differentiation (Boyer et al. 2006). Consistently, we observed a significant fraction of up-regulated signature genes that are down-regulated in Suz12^{-/-} mouse ESCs, while the expression levels of a significant fraction of down-regulated signature genes are elevated in Suz12^{-/-} mouse ESCs (Figure 8H). It has been previously shown that Suz12 binds to a significant fraction of key developmental regulators in mESCs, and those genes are often co-occupied by Oct4, Sox2 and Nanog (Lee, 2006). Our results therefore agree with a model that PRC2 antagonizes the core pluripotency factors in regulating target gene expression during differentiation.

DISCUSSION

We identified a mouse ESC-specific regulatory region that controls Sox2 expression through long-range DNA looping. Sox2-SE^{distal} is an exceptionally large (~13kb) cluster of H3K27ac enrichment thereby fitting the criteria of recently defined super-enhancers or stretch enhancers (Parker et al. 2013; Whyte et al. 2013). It has been proposed that SEs may play a key role in maintaining cell identity, and our results provide unambiguous functional data supporting this hypothesis. Interestingly, we also observed a different SE proximal to Sox2 that is also marked by histone H3K27ac exclusively in brain tissues (Figure 4). This could be an SE that drives Sox2 expression in brain. It will be interesting to

further explore how the Sox2 gene transitions to different enhancer usage during the development of central nervous system (CNS).

Importantly, we have demonstrated the feasibility of using a simple double-CRISPR excision strategy to delete a 13kb non-coding region (Figure 6A). Because of the high efficiency of the CRISPR/Cas9 system, two double-strand breaks can be generated simultaneously on the same chromosome, leading to the loss of an entire 13kb region between the two targeting sites. The usage of a hybrid cell line (F123) also allows allele-specific measurements of enhancer activity. Upon deletion of Sox2-SE^{distal} using CRISPR, we found Sox2 expression is reduced by over 90%. As expected, Sox2-SE^{distal} deletion reduces the transcription of a number of key pluripotency genes and also induces a differentiation gene signature. We believe CRISPR will be a powerful tool for future study of in vivo enhancer function.

It is also striking that Sox2-SE^{distal} regulates expression of Sox2, which is located 100kb away, without affecting other nearby genes. This implies that enhancers engage in highly selective long-range interactions, and their regulation should not be simplified by assuming they regulate the nearest gene. One possible explanation for this specificity lies in the specific looping interaction between Sox2 and Sox2-SE^{distal}. Our analysis of Hi-C data shows significant interaction between Sox2 and Sox2-SE^{distal} (Figure 5). It is also noteworthy that, while there are other genes near Sox2-SE^{distal}, only the Sox2 promoter is marked by a strong peak of H3K4me3 within the same topological domain (Dixon et al.

2012;), which could explain why only the Sox2 gene is targeted by Sox2-SE^{distal}. A similar phenomenon was also observed in a recent study on GATA2 super-enhancer (Gröschel et al. 2014).

It came to our attention that a part of the region we deleted overlaps with 5' end of the Peril ncRNA transcript (Sauvageau et al. 2013). As a result, the Peril transcript was lost from the alleles with Sox2-SE^{distal} deletion (data not shown). However, Sauvageau et al. did not observe significant loss of Sox2 expression in mouse embryos when they knocked out Peril. Therefore, we conclude that the effect of the Sox2-SE^{distal} deletion on Sox2 expression is not due to loss of Peril transcript.

Understanding how genomic information directs tissue-specific gene expression remains a major challenge. Identification of cis-regulatory elements from epigenomic information and validation by efficient genome editing techniques will be a paradigm of how to study gene regulation.

METHODS

Cell culture and transfection experiments

The F1 *Mus musculus castaneus* × S129/SvJae mouse ESC line (F123 line) was a gift from the laboratory of Dr. Edith Heard and was previously described (Gribnau et al. 2003). The cells were cultured as described previously (Selvaraj et al. 2013). Importantly, cells were passaged twice on 0.1% gelatin-coated feeder-free plates before harvesting. F123 cells were plated at a density of 0.5

million/mL on 0.1% gelatin-coated feeder-free plates 24 hours before transfection. Cells were triple transfected using the Mouse ES Cell Nucleofector Kit (Lonza) and Amaxa Nucleofector with 7.5ug of each CRISPR plasmid and 5ug of pBABE-Puro. Post-transfection, cells were immediately plated on puromycin-resistant MEF feeders (GlobalStem) for recovery. 48 hours after transfection 2 µg/mL of puromycin (Sigma) was supplemented to the cell culture media for 3 days to select for transfected cells. Alkaline phosphatase staining was performed using mESC culture in the presence of MEF feeder cells with the Alkaline Phosphatase Staining kit (Stemgent). For growth curve analysis, control and mutant ES cells were plated at 1×10^5 cells in triplicates on 12-well plates and counted for six consecutive days. The cells were passaged on day 4 by making 1:3 dilutions. The cell numbers on day 5 and 6 were multiplied by three to adjust for the dilution.

Design of CRISPR constructs for enhancer deletion

Target-specific CRISPR guide RNAs were designed to optimize uniqueness and have limited off-targets using an online tool (<http://crispr.mit.edu/>). Corresponding oligonucleotides were ordered (IDT) and subcloned into the pX330 plasmid (Addgene), expressing a human codon-optimized SpCas9 and chimeric guide RNA expression plasmid, following previously published guidelines (Cong et al. 2013). The sequences flanking mEnh-Sox2^{distal} targeted by the CRISPR constructs are:

5'-CRISPR: GACGCTTCCGTTCTTGGAGT**AGG**

3'-CRISPR: TTGGATTCCCGACAACAAGCT**TGG**

PCR and Sanger sequencing for genotyping deletion clones

Genomic DNA was isolated from clonal CRISPR targeted F123 cell lines using a DNeasy Blood & Tissue Kit (Qiagen). To screen deletion clones, SE-spanning PCR primers were designed which flank the outside of the CRISPR sgRNAs and amplify a 13kb+ region. Given efficient CRISPR cutting and repair of DNA through non-homologous end joining, a ~420bp product is expected. For PCR, 20ng of genomic DNA was used and amplified with Q5 High-Fidelity DNA Polymerase (NEB).

Subsequently, a qPCR-based genotyping method was designed to quantify the DNA copy number of the targeted regions relative to control regions. qPCR copy number primers were designed inside the targeted region to quantify the efficiency of CRISPR deletion as creating a biallelic or monoallelic deletion. qPCR was performed with 10ng of genomic DNA in triplicate using SYBR green master mix (Roche) on the LightCycler 480 (Roche). Relative levels of DNA copy number were calculated using the $\Delta\Delta C_t$ method with copy number normalized to GAPDH intron 1 (chromosome 6) as a control.

Finally, Sanger sequencing was used to confirm the qPCR genotyping result and further characterize the deletion junction. The ~420bp products from PCR were TOPO-cloned using a Zero Blunt TOPO PCR cloning kit

(Qiagen). Purified plasmid was Sanger sequenced (Retrogen) using the M13 F primer. Sanger sequencing data of one of the biallelic (confirmed by copy number analysis) SE deletion clone were not available because we failed to expand it upon freeze/thaw. The following equation was used to calculate deletion efficiency: $\text{Deletion frequency} = ((\# \text{ monoallelic clones}) + 2 * (\# \text{ biallelic clones})) / (2 * (\# \text{ of total clones}))$.

The following are the primers used:

mEnh-Sox2^{distal} PCR Deletion-spanning primers

F: CCTCAAAGTGCATCTCGTCA

R: GTCCCCTCGGCTTCTGTACT

mEnh-Sox2^{distal} copy number qPCR primers

F: CAAATGCGCCTCTCACTTTA

R: TTCCCAGCAGTGCTTGTATG

qPCR copy number control primers: GAPDH intron 1 (chromosome 6)

F: CTGGCACTGCACAAGAAGAT

R: GGGTTCCTATAAATACGGACTGC

M13F Sanger Sequencing Primer

F: GTTTTCCCAGTCACGACGTTGTA

RNA isolation and allele-specific RT-qPCR

Total RNA was isolated by TRIzol (Life Tech) extraction and DNaseI (Life Tech) treated. The SuperScript III First-Strand Synthesis SuperMix for RT-qPCR

kit (Life Tech) was used for synthesis of cDNA from 1ug of RNA. RT-qPCR was performed in triplicate using SYBR green master mix (Roche) on the LightCycler 480 (Roche). Relative levels of gene expression were normalized to beta actin (*Actb*). RT-qPCR primers for *Actb* are (Fwd: 5'-CTGGCTCCTAGCACCATGAAGATC; Rev: 5'-TGCTGATCCACATCTGCTGG). The primer sequences used in this study are Pou5f1 (F: 5'-GTTGGAGAAGGTGGAACCAA, R: 5'-CCAAGGTGATCCTCTTCTGC), Nanog (F: 5'-ATGCCTGCAGTTTTTCATCC, R: 5'-GAGGAAGGGCGAGGAGAG), Esrrb (F: 5'-ACATTGCCTCTGGCTACCAC, R: 5'-CGGGCAGTTGTACTCGATG), Klf4 (F: 5'-CAGGCTGTGGCAAAACCTAT, R: 5'-CGTCCCAGTCACAGTGGTAA), Kit (F: 5'-CGTGAAGTCCATGTGGCTAA, R: 5'-CGTCTCCTGGCGTTCATAAT).

To assay the allele-specific expression from Sox2 locus, we used a common forward primer to measure the Sox2 gene expression level (Sox2-F: 5'-ACATGTGAGGGCTGGACT). A universal reverse primer (R-universal: 5'-CCTTTTGAGCATTATCAGATTTTTC) is used to measure the total Sox2 expression level. We used two different reverse primers to measure Sox2 expression from the CAST (R-CAST: CTTTTGAGCATTATCAGATTTTCT) or 129 allele (R-129: CTTTTGAGCATTATCAGATTTTCC).

Strand-specific RNA-seq

PolyA⁺ RNA was captured from ten micrograms of total RNA using Oligo(dT)₂₅ Dynabeads (Invitrogen). The resulting RNA fraction was subjected to first-strand cDNA synthesis using random primers and dT oligos in the presence

of actinomycin D. Samples were cleaned up by applying the reaction mix to mini G50 columns. The eluted sample was then subjected to second-strand synthesis with DNA polymerase I, DNA ligase, and dUTPs in the presence of RNase H. This ensures that the cDNA strand complementary to the mRNA has dNTPs and that the other strand is specifically labeled with dUTPs. The cDNA was purified by QIA-quick columns and sheared using a Bioruptor (Diagenode). The fragmented cDNA was subsequently end repaired, A-tailed, and then ligated to Illumina TruSeq adapters. The samples were run on a 2% high-resolution agarose gel and fragments were size-selected from 200bp-450bp. The fragments were then treated with UDGase to remove the labeled strand, and mRNA complementary cDNA strand was amplified by PCR and sequenced (2x100bp) using the Illumina HiSeq2000.

ChIP-seq

ChIP-Seq was carried out as previously described (Hawkins et al. 2010) with 500 µg chromatin and 5 µg antibody with the following antibodies, H3K27ac (Abcam, ab4729) and H3K4me3 (Millipore, 04-745). ChIP and input library preparation and sequencing procedures were carried out as described previously according to Illumina protocols with minor modifications (Illumina, San Diego, CA).

Bioinformatic analysis

All the sequencing data were mapped to mm9. ChIP-seq and Hi-C data analysis were performed as described previously (Jin et al. 2014a). For RNA-seq analysis, reads were mapped using Novoalign and USeq. FPKM for known mouse genes were generated using in-house scripts. To identify signature genes, a pseudo number of 4 was added to the FPKM value of each gene and a 3-fold cutoff was used to identify significantly changed genes comparing wild type and enhancer knockout clones. CAST and 129 genotypes were generated, and strain-specific reads at CAST and 129 SNPs were separated as described previously (Selvaraj et al. 2013). To improve the accuracy of calling genetic variants between the two mouse strains, we also removed a subset of variants that showed inherent biases from all downstream analyses. Briefly, we generated dummy reads across each SNP/Indel and mapped them to the genome: SNPs that showed >5% bias and Indels showing >10% bias are ignored. We also checked for copy number variation across the two alleles and ignored variants that showed significant bias (5% FDR). Finally, we removed variants identified as poor heterozygous calls (5% FDR) (Selvaraj et al. 2013). Functional analysis of the signature genes was performed using DAVID and GSEA.

Data availability

The raw data for ChIP-seq and RNA-seq experiments performed in this study have been uploaded to GEO with accession number (GSE60763). The following epigenomic data used in this study were published previously:

H3K27ac ChIP-seq in different mouse tissues, and p300 ChIP-seq in mESCs (Shen et al. 2012) (GSE29218); ChIP-seq of Med1, Oct4, Sox2 and Nanog in mouse ES cells (Whyte et al. 2013) (GSE44288); Hi-C data in mouse J1 ES cells (Dixon et al. 2012) (GSE34156) and F123 ES cells (Selvaraj et al. 2013) (GSE48592) were combined for the analysis in this study.

FIGURES

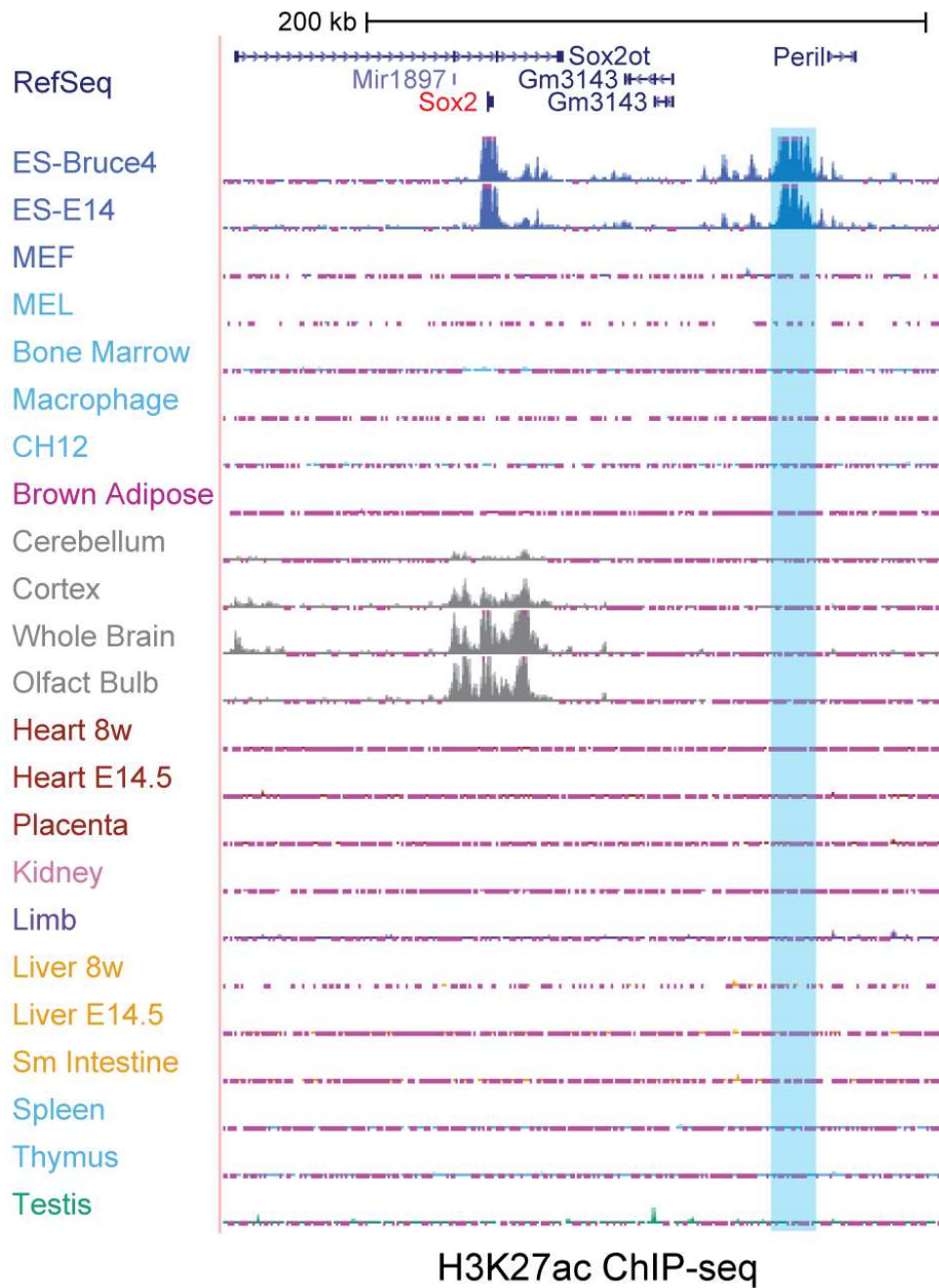


Figure 4. Identification of a distal mouse ESC-specific super-enhancer near Sox2 gene. In a genome-browser snapshot, 23 tracks of H3K27ac ChIP-seq data are shown. The top 2 tracks are from two different mouse ESC lines. Shadowed region shows the location of the mEnh-Sox2^{distal} enhancer.

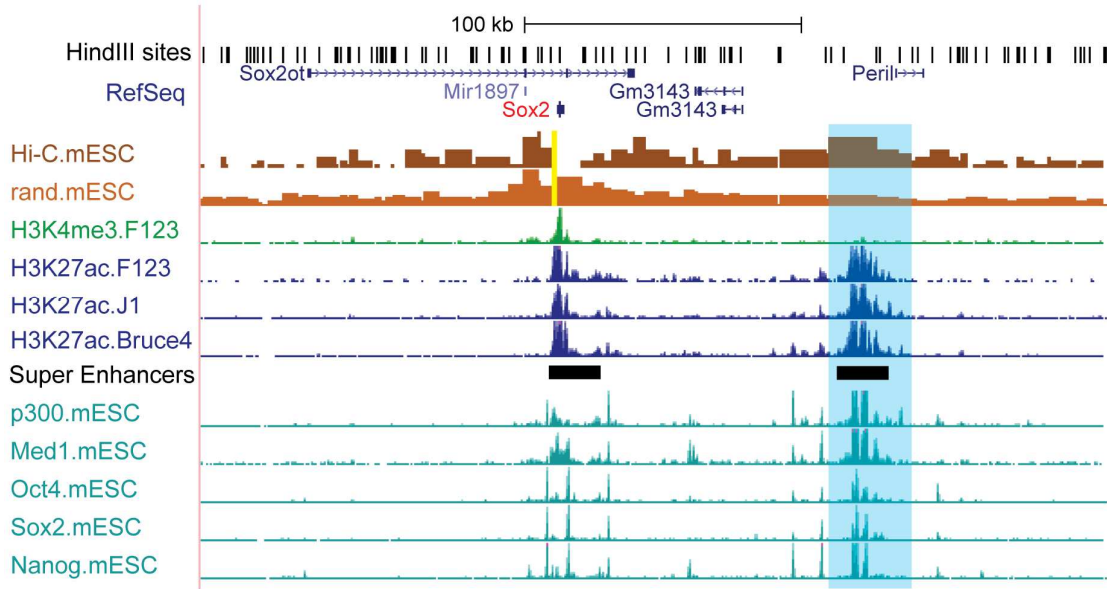


Figure 5. Characterization of the distal Sox2 super-enhancer. In a genome-browser snapshot, the top Hi-C track is a bar graph showing the Hi-C read number in mouse ESCs using Sox2 as the anchor (highlighted in yellow). Each bar represents one HindIII fragment between two HindIII cutting sites. The bar height represents the Hi-C read count from the anchor to the HindIII fragment. The 2nd Hi-C track plots the expected value for every HindIII fragment. Therefore the blue highlighted region is a significant looping interaction from the Sox2 promoter to Sox2-SE^{distal} after comparing the top two tracks. The ChIP-seq tracks show the locations of different histone marks or protein factors in mouse ESC lines. The super-enhancer track shows the locations of two super-enhancers in mESCs identified by Hnisz et al.

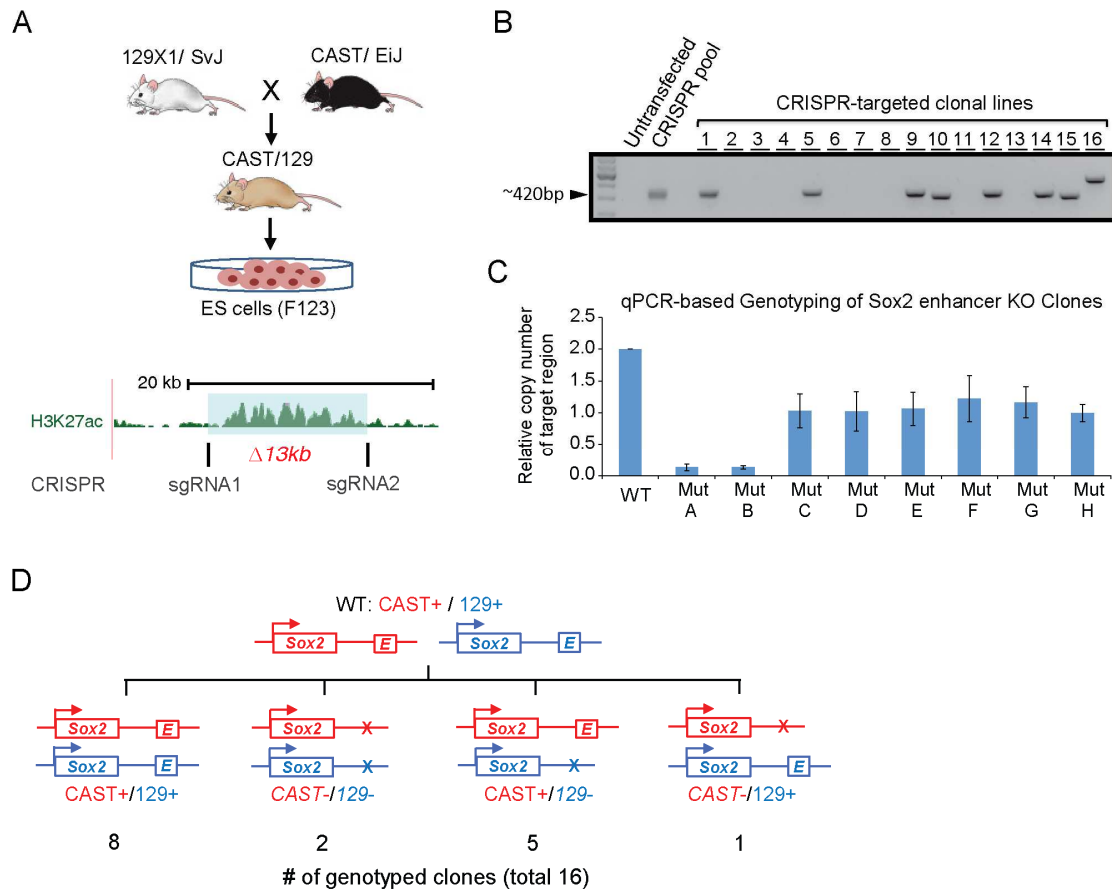


Figure 6. Deleting Sox2-SE^{distal} using CRISPR. **(A)** Schematic representation of CRISPR-based SE deletion strategy in a hybrid mouse ESC line, F123. Lower panel shows the location of two small guide RNA flanking the 13kb Sox2-SE^{distal}. **(B)** Two PCR primers flanking (see Methods) the deleted region were used to amplify the genomic DNA from mutant clones. Successful deletion (monoallelic or biallelic) generates a small PCR product of ~420bp with some length variation due to random mutations during non-homologous end-joining. 8 of the 16 clones show a product, meaning at least one copy of the SE is deleted. **(C)** qPCR was performed using two primers inside the SE (see Methods) to measure the copy number of the SE in WT ESCs and 8 mutant clones. Signal in WT clone is normalized to 2. Error bar: standard deviation. **(D)** Summary of the genotypes of the mutant ESC clones as confirmed by Sanger sequencing.

Figure 7: Sox2-SE^{distal} regulates Sox2 expression *in cis*. **(A)** Allele-specific RT-qPCR analysis of mESC clones with different genotypes. Samples are normalized to the average Sox2 signal between two WT biological replicates. Error bar denotes standard deviation. Note one biallelic Sox2-SE^{distal} deletion clone (Mut B from **Figure 3**) is missing because we fail to re-expand that clone after freeze and thawing. **(B)** Genome browser tracks showing the RNA-seq signals near the Sox2 locus. Only RNA-seq reads on SNPs between CAST and 129 genomes are plotted. Red tracks: RNA-seq reads on CAST SNPs, the signal above the base line shows transcription from positive strand and the signal below the base line shows transcription from negative strand. Blue tracks: RNA-seq reads on 129 SNPs. The bottom heat map track shows TAD defined by Hi-C. The TAD containing Sox2 gene is highlighted in yellow. **(C)** Comparison of allele-specific expression from RNA-seq in the monoallelic 129 and monoallelic CAST clones. Allele specificity is defined as the log2 ratio of RNA-seq reads from CAST and 129 allele after adding a pseudo count of 10. All genes on chromosome 3 including Sox2 are plotted. Only the Sox2 gene shows CAST specific expression in CAST+/129- clone and 129 specific expression in CAST-/129+ clone, suggesting direct regulation by Sox2-SE^{distal}.

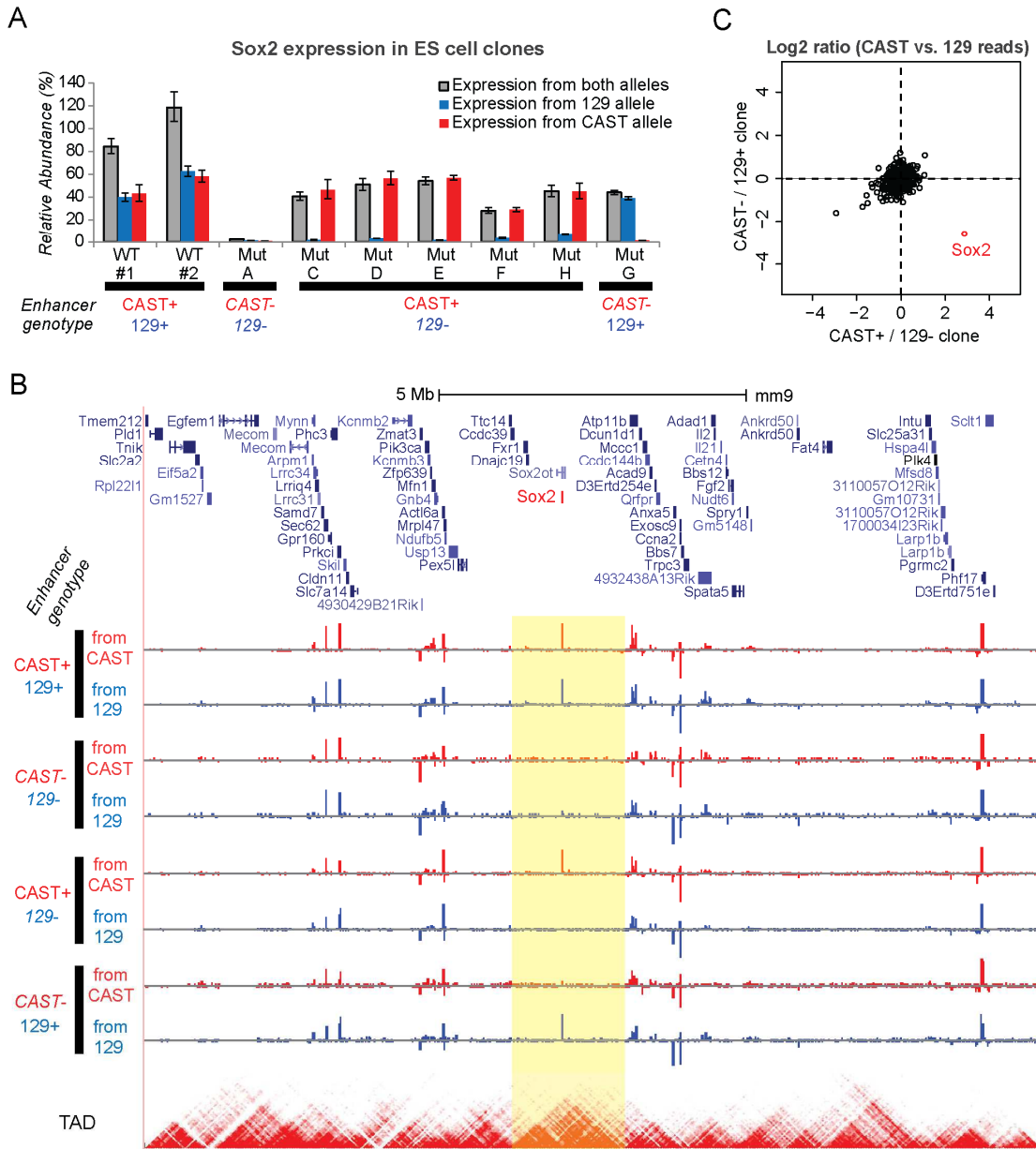
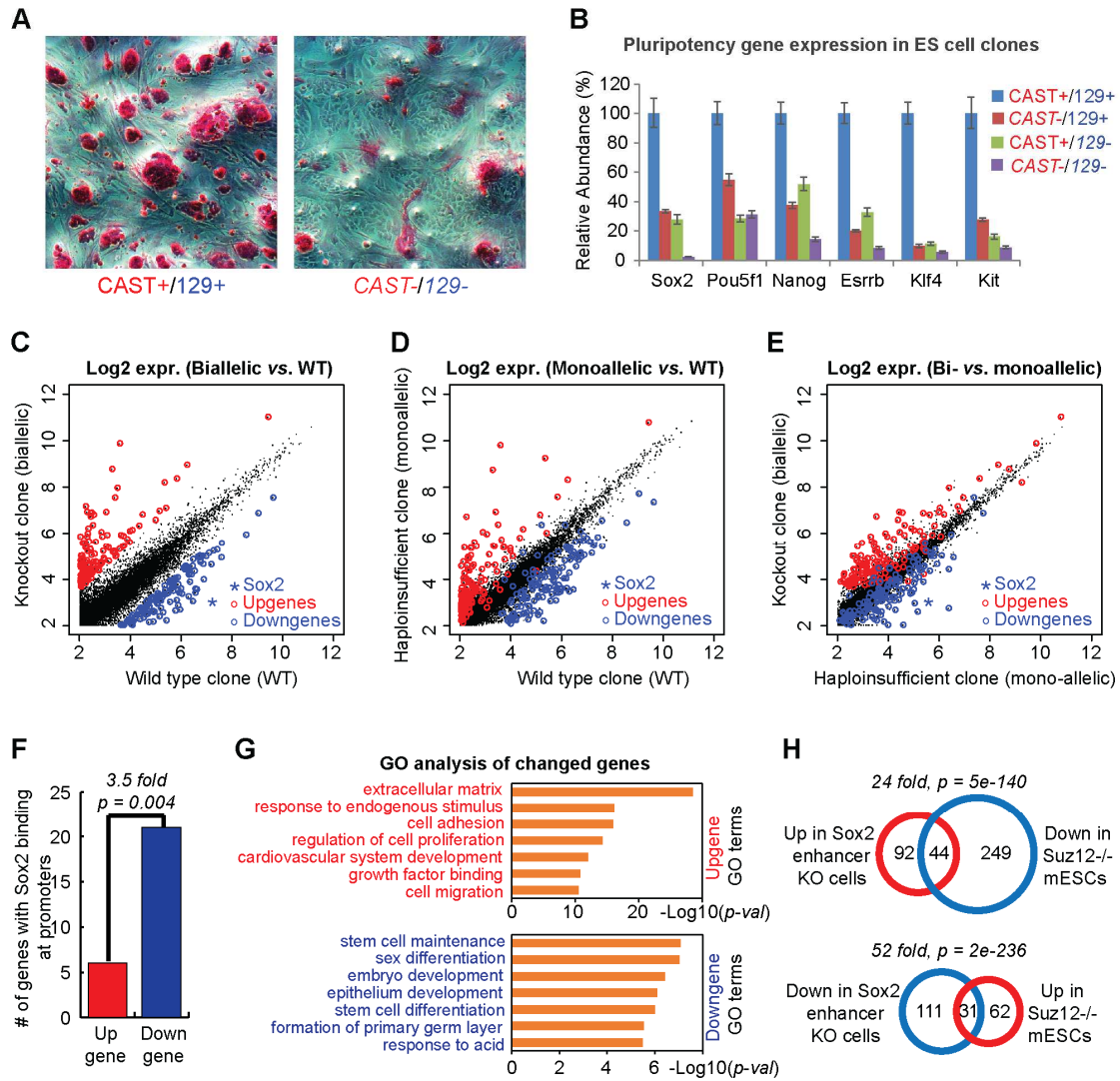


Figure 8: Sox2-SE^{distal} deletion affects key pluripotency genes. **(A)** Alkaline phosphatase staining of wild type (left) and biallelic Sox2-SE^{distal} deletion (right) clones. **(B)** RT-qPCR analysis of pluripotency genes in 4 representative wild type (WT) or Sox2-SE^{distal} deletion clones. **(C-E)** Scatter plots comparing the expression levels of all mouse genes in different Sox2-SE^{distal} deletion clones according to RNA-seq data. Knockout clone: biallelic Sox2-SE^{distal} mutant; haploinsufficient clone: monoallelic Sox2-SE^{distal} mutant. Plotted are log2 scaled FPKM values for each gene after adding a pseudo-value of 4. Signature up-regulated (red spots) or down-regulated (blue spots) genes were identified from the biallelic mutant clone **(C)**. Expression levels of these signature genes were also compared in **(D)** and **(E)**. **(F)** Bar graph comparing the up-regulated and down-regulated signature genes with Sox2 binding at promoters as called by Sox2 ChIP-seq data. P-value: Binomial test. **(G)** GO analysis identified top GO terms enriched in up-regulated or down-regulated genes upon Sox2-SE^{distal} deletion. **(H)** Venn diagrams comparing the Sox2-SE^{distal} knockout signature genes to signature genes found in Suz12 knockout mouse ESCs. P-value: binomial test.



AUTHOR'S CONTRIBUTIONS

Conceived and designed the experiments: YL CMR HI. Performed the experiments: YL CMR HI. Analyzed the data: YL CMR HI FJ SS. Contributed reagents/materials/analysis tools: FJ SS AYL JRD. Wrote the paper: YL CMR HI FJ SS BR.

ACKNOWLEDGEMENTS

This work is supported by funds from the Ludwig Institute for Cancer Research, and US National Institutes of Health (U54 HG006997) to BR. CMR was supported in part by the UCSD Genetics Training Program through an institutional training grant from the National Institute of General Medical Sciences, T32 GM008666. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

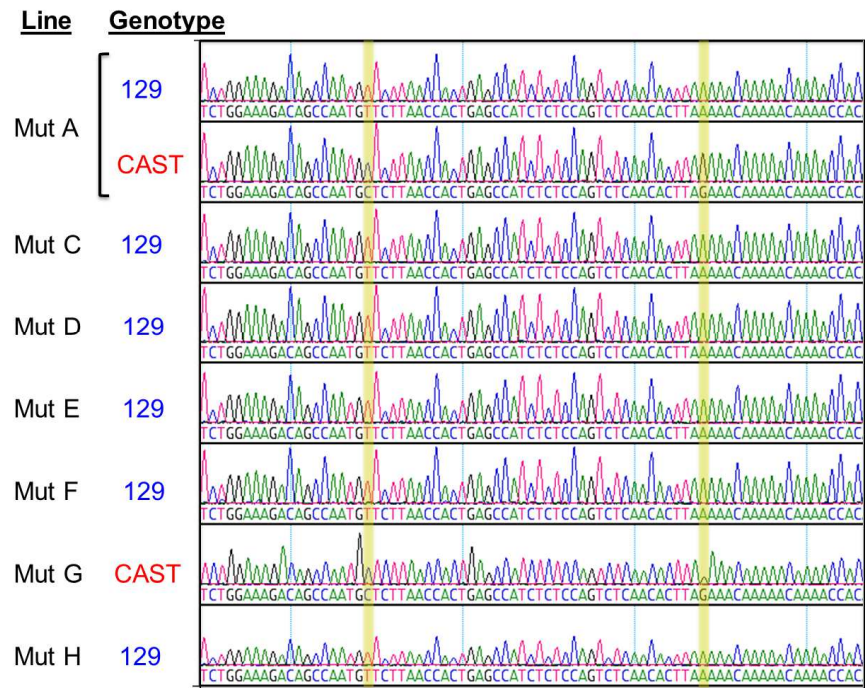
Chapter 2, in full, is an adaptation of materials that appears in Li Y, Rivera CM, Ishii H, Jin F, Selvaraj S, Lee AY, Dixon JR, Ren B. CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells. *PLoS One*. 2014 Dec.

APPENDIX

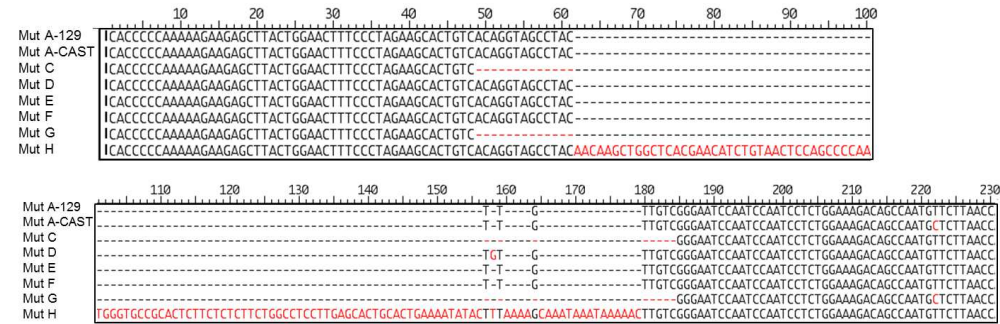
Supplemental Figures

Figure 9. Genotyping of enhancer deletion clones. **(A)** Aligned Sanger sequencing chromatograms from deletion junctions amplified from genomic DNA of targeted Sox2-SE^{distal} clones. Two SNPs (highlighted in yellow) on the 3' end of the PCR product reveal the origin of the targeted allele. Mutant A is biallelic, Mutant G is monoallelic with the CRISPRs targeted the CAST allele, and the remaining clones are monoallelic with the CRISPRs targeting the 129 allele. **(B)** Aligned Sanger sequencing results from deletion junctions amplified from genomic DNA of targeted Sox2-SE^{distal} clones. Clones had varied deletions created by the Cas9 nuclease and repaired by non-homologous end joining. Mutant H contains a short inversion between the CRISPR/Cas9 cut sites consistent with the larger PCR product observed during genotyping (see Fig. 3B). The SNP at position 222 reveals the parental original of targeted allele. Mutant A is biallelic, Mutant G is monoallelic with the CRISPRs targeted the CAST allele, and the remaining clones are monoallelic with the CRISPRs targeting the 129 allele.

A



B



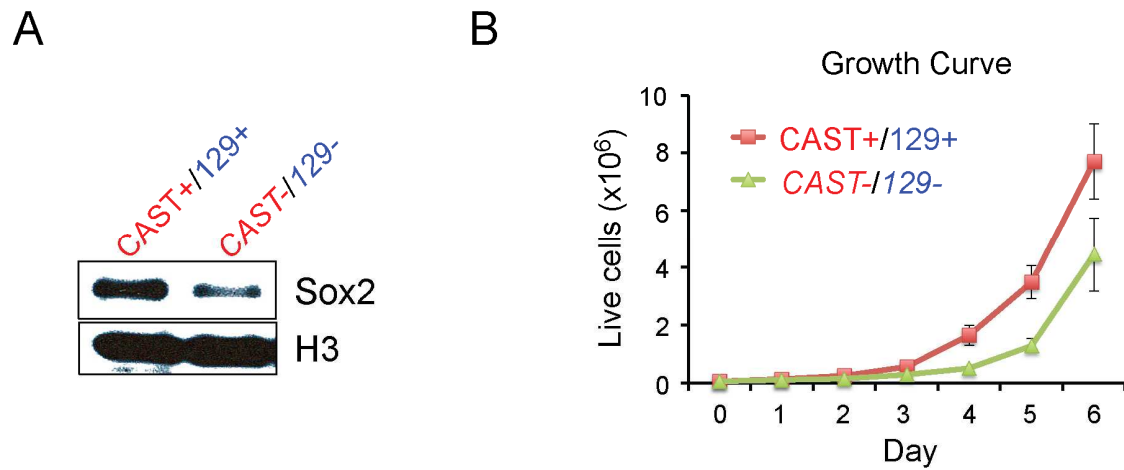


Figure 10. Deletion of Sox2-SE^{distal} impairs ES cell proliferation. **(A)** Western blot analysis of Sox2 gene in wild type and biallelic Sox2-SE^{distal} deletion mESC clones. **(B)** Growth curves of wild type and biallelic Sox2-SE^{distal} deletion mESC clones.

Supplemental Tables

Table 4. FPKM for all mouse RefSeq genes in 4 representative mESC clones from RNA-seq. (doi:10.1371/journal.pone.0114485.s003)

Table 5. A list of 136 genes upregulated after deleting Sox2 enhancer. (doi:10.1371/journal.pone.0114485.s004)

Table 6. A list of 142 genes downregulated after deleting Sox2 enhancer. (doi:10.1371/journal.pone.0114485.s005)

Table 7. A list of genes bound by Sox2 at promoters. (doi:10.1371/journal.pone.0114485.s006)

CHAPTER 3 –FUNCTIONAL IDENTIFICATION OF TRANSCRIPTIONAL ENHANCER TARGET GENES TO REVEAL THE RULES OF GENE REGULATION

ABSTRACT

Transcriptional enhancers are DNA elements that bind sequence specific transcription factors and function to boost the transcription of at least one target promoter. Large consortium based projects have identified putative cis-regulatory elements in hundreds of cell types and tissues using epigenetic signatures such as histone marks and DNase hypersensitivity. However, whether the predicted regulatory elements actually function as enhancers, the identify of their target promoters, and the measurable contribution of these enhancers to transcriptional activation are all still unclear. In this study we utilize a CRISPR-based genome engineering technique to delete typical enhancers (TEs) and super enhancers (SEs) in mouse embryonic stem cells (mESCs) from a hybrid mouse strain. This project is currently on going and the goal is to delete 100 such regulatory elements. Allelic RNA-seq analysis of reciprocal monoallelic clones is used to understand the direct impact of the deletion on genes *in cis*. Our results so far indicate several key challenges in robustly genotyping enhancer KO clones. Follow up studies, using ChIP-seq and 4C-seq are used to explore the exact stage of transcriptional activation affected by enhancer loss. In the Sox2-SE^{distal} mutants RNA PolII is no longer recruited to the promoter and hyperacetylation of histones (H3K27ac) at the promoter is lost.

Together, these results imply that Sox2-SE^{distal} plays a role in transcriptional initiation at the Sox2 promoter. As this study progresses, and more mutant lines are generated, this study will offer a set of validated enhancers ripe for further enhancer mechanism studies.

INTRODUCTION

Definition of transcriptional enhancers and proposed models of function

Enhancers can be defined as DNA elements that increase transcription from an associated promoter independent of their orientation or position relative to the proximal promoter (Blackwood and Kadonaga 1998). These *cis*-regulatory elements can be located upstream or downstream of the promoter or within an intron. As discussed in Chapter 2, clusters of enhancer constituents that generate a domain of MED1- and transcription factor-binding and H3K27ac enrichment have been identified as super-enhancers (SE) (Hnisz et al. 2013; 2011; Whyte et al. 2013). The purpose of this study in Chapter 3 is to use genomic knockouts of SEs and TEs to better understand how they mechanistically control gene expression.

As extensively reviewed, there are several models by which enhancers are thought to communicate with promoters (Blackwood and Kadonaga 1998; Bulger and Groudine 2011). The looping model suggests that distal enhancers and proximal promoters recruit specific factors which then interact, looping the DNA to form a transcription complex. How these looping interactions are formed,

maintained, and choose specificity are unclear but likely involve structural proteins such as CTCF and cohesion. Another theory is that enhancers recruit proteins that covalently modify both histones and transcription factors to regulate transcription. For example, many transcriptional coactivators that bind enhancers are histone acetyltransferases. Acetylation of histones and general transcription factors such as TFIIE and TFIIIF both promote transcription (Imhof et al. 1997). One last example of a model for enhancer function concerns nuclear localization. It has been theorized that enhancers function to target their associated promoters to punctae and active compartments in the nucleus with high local concentrations of transcriptional machinery, called transcription factories. To summarize, current studies suggest enhancers are capable of an array of functions and some, all, or none, of these may be the cause of enhancer mediated gene activation. Aberrant transcriptional activation is a common feature of cancers and other human diseases, therefore understanding the mechanisms of enhancer action may have broad impacts on the molecular etiology of disease.

Regardless of how an enhancer “finds” its target promoter, the question remains of what the enhancer contributes to the promoter microenvironment that results in increased transcriptional output. In response to this question, experimental evidence exists to support the idea that enhancers can activate transcriptional initiation and transcriptional elongation. First, enhancers are bound by RNA polymerase II, histone acetyltransferases (Heintzman et al. 2007),

general transcription factors (TAFs) (Szutorisz, 2005), and the mediator complex (Kagey et al. 2010), all of which may contribute to a fully assembled PIC at the promoter upon enhancer looping. Studies support the role of an enhancer in transcriptional pause release and elongation as well. Components of the RNA polymerase II elongation machinery, specifically ELL3, are found at enhancers in ESCs (Lin et al. 2013). Loss of ELL3 results in improper recruitment of both RNA Pol II and the super elongation complex at target promoters. A class of enhancers, called anti-pause enhancers, are bound by both Brd4 and jumonji C-domain-containing protein 6 (JMJD6) in HEK293T cells (Liu et al. 2013). The actions of this kinase and demethylase activate the P-TEFb complex, a step that is specifically required for transcriptional elongation. These studies do not necessarily conflict with each other, as it is completely possible that enhancers use diverse mechanisms to regulate promoter transcription.

Methods of identifying the promoter targets of enhancers

Regulatory elements can be predicted by scanning the genome for intragenic sequences with high conservation across species, transcription factor binding, DHS, open chromatin marks (specifically H3K27ac and H3K4me1), and transcription of short enhancer RNAs (eRNAs) (Shlyueva, 2014). However, it should be noted that each of these enhancer signatures *predict* putative enhancer regions and *correlate* with potential enhancer activity. Studies such as the ENCODE project and Epigenome Roadmap Projects have used these signatures to

predict tens of thousands of enhancers per cell type in hundreds of cell lines and primary tissues (Consortium et al. 2012). While these studies have used integrated data to generate a massive enhancer annotation resource, it remains unclear which of these predicted *cis* regulatory elements are in fact functional. Specifically we ask, in a given cell type and under a certain set of conditions, which enhancers are boosting transcriptional expression of a target gene? And, can we identify the promoter target(s) of each functional enhancer?

An extensive computational and molecular biology-based toolkit exists for researchers to predict the functional enhancer-promoter interactions. Naturally occurring variants in regulatory elements, which likely impact transcription factor binding sites, can be correlated with changes in gene expression to identify enhancer targets (Heinz et al. 2013). This approach is limited by the density of genetic variation amongst individuals, therefore, all enhancers cannot be assayed.

One approach for functionally identifying the functional targets of an enhancer is to recruit a repressor complex to the enhancer and identify changes in gene transcription. TALENs and Cas9 complexes have been used to recruit the lysine demethylase LSD1 to enhancers (Mendenhall et al. 2013) This approach is called an “epigenetic knockdown” of the target promoters. However, the resulting targets of these studies are difficult to label as interacting partners or functionally activated target promoters of the enhancer.

Using genome engineering to physically remove the enhancer sequence is considered the “gold standard” for validating enhancer function. Traditional

genome targeting methods are inefficient and inherently low throughput. The recently developed CRISPR/Cas9 system (Cong et al. 2013) has proven to be a highly efficient genome editing technique and offers a promising method to validate enhancer functions *in vivo*. We have therefore chosen to scale up our dual CRISPR cutting approach in a mid-throughput manner to delete 100 regulatory elements in mESCs.

Summary of enhancer knockout studies: Focus on the β -globin LCR

Very few studies have rigorously characterized the impact of the loss of a regulatory element on the sequential stages of transcriptional activation at a promoter. A notable example is the dissection of the role of the locus control region (LCR) that regulates the β -globin locus in erythroid cells. Overall, deletion of the LCR results in a dramatic 96% decrease in β -globin expression. Histone acetylation, DNaseI sensitivity, and transcription factor binding at the promoter are not impacted (Schübeler et al. 2001; Schübeler et al. 2000).

Preinitiation complex (PIC) formation is only partially dependent on the LCR as both general transcription factor (GTFs), including TFIIB, and total RNA polymerase II binding at the promoter are reduced 2-fold (50%) on the knockout allele. Strikingly, phosphorylated RNA polymerase II, a form of the protein when it is actively elongating, is reduced 75%. This final result indicates the β -globin LCR is necessary for transcriptional elongation but not initiation. Said differently,

the LCR “functions primarily by enhancing the transition from transcription initiation to elongation” (Sawado et al. 2003).

Finally, let’s consider the role of DNA looping or chromatin interactions between the β -globin LCR and promoter. Using FISH and 3C, it was shown that the LCR does create a long-range interaction with the β -globin promoter and this interaction selectively occurs when the β -globin gene is on (Brown et al. 2008). However, it is difficult to tease out whether this interaction is necessary for boosting promoter transcription or an indirect consequence of promoter activation. Furthermore, this doesn’t rule out the idea that an enhancer can activate a promoter without directly interacting with it.

A recent study blocked β -globin expression and enhancer-promoter looping by removing a TF called GATA1. Introduction of a synthetic zinc finger (ZF)::LDB1 fusion protein that recognizes an LCR specific sequence rescued both the looping and expression of β -globin without the presence of the GATA1 transcriptional activator (Deng et al. 2012). By bypassing the role of the transcriptional activator GATA1 this study emphasized the importance of enhancer looping and promoter contact in gene activation.

It should be appreciated that while several models of enhancer function have been proposed, there have been too few functionally characterized enhancers to make conclusions about any one model (Table 8). For example, some enhancer KO studies suggest the regulatory element activates

transcriptional initiation (e.g. *Aicd* enhancers) and some regulate transcriptional elongation (e.g. β -globin LCR).

Scientists in the field should not expect one overarching dogma for enhancer function. Rather the class of enhancer (super, typical, LCR), cell type, context (lineage-determining enhancer or signal-dependent enhancer), or exact cellular conditions could all influence the mechanism an enhancer uses to activate transcription. We are hopeful that our mid-throughput (relative to published single enhancer KO studies) enhancer KO approach will yield enough examples of enhancer-mediated activation in different genomic contexts to begin to define the rules of enhancer activation.

RESULTS

Sox2-SE^{distal} controls the Sox2 promoter at the level of PolII recruitment

We previously generated biallelic and monoallelic KO mutants of the Sox2 SE located 100kb distal to the promoter, called Sox2-SE^{distal} (Li et al. 2014). This mutant revealed that Sox2-SE^{distal} is responsible for greater than 90% of Sox2 expression. In addition, this SE does not regulate expression of any other genes *in cis* on chromosome 3.

We decided to use this system to further dissect the mechanistic impact of this SE on Sox2 promoter transcription. We first asked whether the SE is necessary for RNA PolII recruitment to the promoter, a step in transcriptional

initiation. ChIP-seq for total RNA PolII revealed RNA PolII binding is nearly completely lost at the Sox2 promoter in a Sox2-se^{distal}-/- clone (Figure 11).

We next asked whether the SE impacts the open chromatin at the promoter. To address this question, we performed ChIP-seq for H3K27ac and found this histone mark is also devoid at the Sox2 promoter in biallelic mutants (Figure 11). Taken together, this ChIP-based data suggests that the SE is necessary for transcriptional initiation, RNA PolII recruitment, and open chromatin at the Sox2 promoter.

Using allelic 4C-seq, we generated a contact profile for the Sox2 promoter in F123 mESCs (Figure 12). Overall, the contact frequencies on the 129 and CAST alleles correlate and have few differences. As expected, the strongest interaction from the Sox2 promoter is with a region 100kb upstream corresponding to Sox2-SE^{distal}. This finding is consistent with published HiC data in mESCs and our previous study in Chapter 2 (Dixon et al. 2012; Li et al. 2014). It should be noted that the 4C-seq data is much higher resolution (around 250bp) as we used 4 base cutter restriction enzymes as opposed to the HiC data which has a resolution of 20kb (Dixon et al. 2012). In WT F123 cells, there are at least three distinct contacts within 20kb upstream and downstream of the TSS. These regions correspond to previously identified Sox2 regulatory regions (SRRs) named SRR1, SRR2, and SRR 18 (Zappone et al. 2000).

The Sox2-SE^{distal} -/- cell line shows a dramatic change in local contacts from the Sox2 promoter. First, the CAST and 129 allele have slightly different

profiles. Although this is a biallelic mutant, the deletions on each allele are slightly different. The 129 allele retains interactions with the 2kb regions flanking the deleted Sox2-SE^{distal}. In contrast, the CAST Sox2 promoter loses all interactions with the SE region. Furthermore, the interactions with the promoter proximal TEs, SRR1, SRR2, and SRR18, are lost on the CAST allele but not on the 129 allele. RNA-seq and RT-qPCR results confirm that this biallelic mutant expresses Sox2 at 5% the level of WT cells. The residual interactions on the 129 allele in the biallelic mutant may explain why Sox2 expression is not completely lost in the biallelic mutant.

ENCODE Enhancer KO Project study approach and progress

Our overall goal is to delete 100 regulatory elements in the F123J hybrid mESC cell line. We partitioned this study into several phases and in this chapter we will focus only on the first phase. The goal of the first phase is to delete 20 SE and 20 typical enhancers (TE). The goal of subsequent phases will include creating deletions which tile across a singular super enhancer to better understand the modularity or cooperatively involved in their function.

We plan to attain biallelic and monoallelic (both CAST and 129) deletion clones for each candidate SE and TE. Clones will be rigorously genotyped PCR, qPCR, and Sanger sequencing. Finally, the lines will be assayed by RNA-seq to determine if the putative regulatory element controls regulation of gene(s), to identify the target gene(s) of the regulatory element, and to determine the

contribution of the regulatory element to the expression of its target(s). The experimental production pipeline for each SE or TE KO takes about 50 days (Figure 13). Our current capacity allows us to complete 5-10 KOs at a time and stagger the start of each round about every 2 weeks. Overall, Phase I of the enhancer deletion project is expected to take 4 months for completion not including subsequent time for data analysis (Figure 14).

To the date of this publication, we have designed two sgRNAs flanking 41 distinct regions which includes 20 SEs, 20 TEs, and 1 negative control region. CRISPR sgRNAs have been cloned into expression plasmids for 36 out of 41 (87%) enhancers. We have transfected F123J cells with 26 out of 41 (63%) plasmid pairs. Of the 26 transfections, 18(69%) have been expanded into multiple (average 20 each) clonal lines from a single mESC colony and are undergoing extensive genotyping (Figure 14). We have completed PCR and qPCR based genotyping for 11 candidate enhancers comprising 202 clonal cell lines. The deletion frequencies (calculation described in Methods) in these lines vary widely from 0% to 76% (Figure 15).

Deletion of Tet1 promoter and super enhancer in mESCs

The Tet1 gene encodes a DNA demethylase that catalyzes the reaction of 5-methylcytosine to 5-hydroxymethylcytosine. SgRNAs were designed to delete a region just downstream from the Tet1 promoter. This region is called as a SE when using the H3K27ac and Med1/OSN based definitions (Hnisz et al. 2013;

Whyte et al. 2013). Unfortunately, after the clones were generated and genotyped, it was discovered that the 21kb deleted region in fact also contained the Tet1 promoter. The Tet1 promoter is identifiable as a sharp H3K4me3 peak almost 7kb downstream of the annotated TSS in the most current RefSeq annotation (Figure 16). This clarification complicates the usefulness of these clones for enhancer function studies as both an enhancer and promoter are deleted.

Given that the Tet1 promoter is deleted in the mutants, we expect Tet1 expression to be absent. While we cannot test the SEs role in regulating Tet1's expression, its predicted target by proximity, we can test whether the SE regulates any other nearby genes. Unlike the Sox2 gene desert region, the Tet1 locus is found in a TAD boundary region which is enriched for active genes (Dixon et al. 2012). Therefore, this deletion provides a chance to see if SEs can regulate more than one gene. Although we cannot test whether this SE regulates Tet1, we still refer to it as TET1-SE.

We performed allelic RNA-seq on WT F123 cells and reciprocal Tet1-SE monoallelic mutants. As expected, Tet1-SE KO on the 129 allele results in the absence of Tet1 expression only from the 129 allele, and vice versa for the CAST mutant (Figure 16A, C). Interestingly, deletion of the Tet1 promoter and SE did not directly affect the expression of any gene on chromosome 10 *in cis* other than Tet1 (Figure 16B). From this study we cannot conclude whether the Tet1-SE regulates Tet1 expression. However, we can conclude that this SE does not

directly regulate the expression of any other genes on chromosome 10. These experiments are currently being repeated with new sgRNAs that delete only the Tet1-SE and avoid the promoter region.

Functional characterization of PRDM14 SE reveals a singular promoter target

PRDM14 is a sequence specific transcription factor that promotes maintenance of the ESC state by preventing induction into extraembryonic endoderm and activating expression of key pluripotency genes (Ma et al. 2011). Located near the PRMD14 gene and predicted to interact with its promoter by HiC is a SE we refer to as PRDM14-SE (Figure 17). As before, we designed two sgRNAs to completely excise this 36kb region. Out of the 16 clones screened, only one clone had a deletion by PCR screening and it was determined to be a monoallelic deletion by qPCR. RNA-seq was performed on this clone and we performed an allelic analysis of expression (Figure 17A).

Alleleic RNA-seq analysis revealed that although this clone was genotyped as a mono-allelic deletion, the SE is completely deleted on the 129 allele and there is likely a microdeletion on the CAST allele. Two lines of evidence support this new genotype assignment. First, SEs are transcribed and their transcripts are detected by our PolyA+ RNA-seq protocol. In the KO clone, no reads are seen at the SE on the 129 allele. For the CAST allele, however, reads can be seen across the SE except for a 3kb region which is distinctly devoid of transcripts suggesting a micro-deletion on the 3' end of the SE. The second piece of evidence is the RNA-

seq signal from the PRDM14 locus. We observe that the 129-based PRDM14 expression is decreased to 0.45% relative to WT levels. In addition, the expression of PRDM14 on the CAST allele is also dramatically decreased to 10% of WT levels (Figure 17B). As enhancers are thought to function exclusively *in cis*, this data suggests there is also a deletion on the CAST allele. To carefully identify direct *cis* targets of the PRMD14-SE, RNA-seq analysis of reciprocal monoallelic mutants is needed.

The inconsistencies in genotyping mutants with large CRISPR deletions also occurred in several Tet1 mutants (data not shown). To review our genotyping methods, we use PCR with sgRNA flanking primers. An efficient product (generally less than 1000bp) is generated if one or both alleles are deleted. The PCR is not quantitative and we generally cannot distinguish between biallelic and monoallelic mutants. Following this step, qPCR is used to assay one 100bp region in-between the sgRNA cut sites. While we generally try to design the qPCR assay region in very center of the KO region, there are cases where repeats and SNPs prevent such a design. QPCR primers close to either sgRNA are not ideal. If a microdeletion created by one sgRNA at the same region of the qPCR primers, this clone will incorrectly be assigned a biallelic clone. Conversely, if the microdeletion occurs at a region away from the qPCR primers, such a clone will be mislabeled as monoallelic. In conclusion, a single qPCR primer set which assays only a 100bp region cannot accurately genotype large CRISPR deletion clones due to the frequent occurrence of microdeletions of 1-

5kb. Our CRISPR genotyping pipeline will need to seek new and additional approaches to ensure robust genotyping of CRISPR KO clones. However, these artifacts are less likely to occur when genotyping small deletions with qPCR based methods. Therefore, the genotyping pipeline for TEs is likely appropriate.

DISCUSSION

The purpose of this study in Chapter 3 is to use genomic deletions of SEs and TEs to better understand how they mechanistically control gene expression. The results of this study, while preliminary and incomplete, generated enough data to make some initial insights. These insights shed light on the mechanisms enhancers use to activate promoter transcription and reinforce our view of the CRISPR/Cas9 system as an efficient tool for generating enhancer KOs.

Taken together, the RNA PolII and H3K27ac ChIP-based data suggest that the SE is necessary for transcriptional initiation, RNA PolII recruitment, and open chromatin at the Sox2 promoter. It will certainly be of interest to determine more generally if SEs as a class function exclusively regulate transcriptional initiation. The Sox2 promoter 4C-seq data revealed extensive local interactions with TEs defined as SRR1, 2, and 18. A recent study deleted these SRR regions singularly, and in a triple mutant, and found no significant effect on Sox2 expression (Zhou, 2014). While this is an intriguing result, it is not clear whether there is any effect on Pol II recruitment, Pol II travel rates, or Pol II elongation.

After deleting Sox2-SE^{distal} in F123 mESCs with 31.25% deletion efficiency, it was not immediately clear if it is possible to attain similar rates of enhancer mutant clones at other loci. Now that we have genotyped 11 different lines, we first observe that CRISPR cutting is similarly efficient in the F123 and F123J sub-lines (Figure 15, compare Sox2 vs. Sox2(F123)). CRISPR deletion efficiency seems to cluster at two ends of the spectrum, from very poor efficiency to very high efficiency. For example, 55% of the sgRNA pairs have deletion efficiencies under 10% and nearly all remaining sgRNA pairs (36%) have deletion efficiencies over 60%. This may suggest that sgRNA sequences have a binary impact – meaning they either cut or they don't.

Notably, TE4's set of clonal lines had a deletion efficiency of 76%. All clones that contained a deletion in the PCR assay were shown to in fact be biallelic deletions in by qPCR (Figure 15). While the high biallelic rate (100%) is remarkable, it could be a hindrance to our study design that requires reciprocal monoallelic mutants to identify direct enhancer targets *in cis*. Further genotyping of TE clones is necessary before drawing any conclusions or changing our experimental approach.

The Sox2-SE^{distal} and PRDM14-SE KOs both suggest that these SEs only target one gene, Sox2 and PRDM14 respectively. This is slightly surprising given the broad domain SEs encompass and the many active genes surrounding the SE (in the PRDM14 case). Recent studies report the existence of so called "super enhancer domains" (Downen et al. 2014). CTCF and cohesion co-bound sites that

loop to each other form these insulated domains. This loop forms an insulated “neighborhood” containing SEs and their potential target genes, as predicated by enhancer-promoter interactions. In fact, perturbation of these CTCF boundary sites disrupts proper expression of nearby genes (Dowen et al. 2014). We found that the Sox2 and PRMD14 genes and SEs both line within SE domains (Sox2 data not shown, Figure 17). As more SE KO clones are generated it will be interesting to ask whether this is a general phenomenon.

Enhancer deletion studies are useful in that they reveal target promoter and these mutant lines can serve as tools to understand the mechanisms behind transcriptional activation. Such enhancer KO studies can also potentially lead to new therapeutic targets. One such example, is a gene targeting therapy for patients with β -thalassemia which recently entered clinical trials. Patients with β -globin thalassemia lack proper β -globin expression and hemoglobin formation leaving many patients dependent on regular blood transfusions. Another form of globin call γ -globin, similar to β -globin and expressed specifically in the fetal stages of development, has been shown to ameliorate β -globin thalassemia symptoms when re-expressed in mouse models or due to natural variation in sub-groups of patients (Bauer et al. 2012). A series of studies discovered that a transcriptional repressor, called Bcl11A, inhibits γ -globin expression in adults (B (Bauer et al. 2013). Furthermore, it was determined that Bcl11A is regulated by an erythroid specific enhancer. Disruption of this enhancer restores γ -globin expression and hemoglobin formation (Bauer et al. 2013). Sangamo, a

biotherapeutic company, has entered Phase I clinical trials to determine effectiveness of deleting the Bcl11A erythrocyte specific enhancer in the hematopoietic stem cells of patients with a pair of synthetic zinc finger nucleases. This example highlights the usefulness of identifying cell type specific enhancers, benefits of understanding their potential contribution to disease pathology, and the impact of efficient genome engineering tools.

METHODS

Cell culture and transfection experiments

The F1 *Mus musculus castaneus* × S129/SvJae mouse ESC line (F123J sub-line) was a gift from the laboratory of Dr. Rudolf Jaenisch and was previously described (Gribnau et al. 2003; Kouzarides 2007; Landt et al. 2012; Patel et al. 2011; Selvaraj et al. 2013). The cells were cultured as described previously (Selvaraj et al. 2013). Importantly, cells were passaged twice on 0.1% gelatin-coated feeder-free plates before harvesting. F123 cells were plated at a density of 0.5 million/mL on 0.1% gelatin-coated feeder-free plates 24 hours before transfection. Cells were triple transfected using the Mouse ES Cell Nucleofector Kit (Lonza) and Amaxa Nucleofector with 7.5ug of each CRISPR plasmid and 5ug of pBABE-Puro. Post-transfection, cells were immediately plated on puromycin-resistant MEF feeders (GlobalStem) for recovery. 48 hours after transfection 1 µg/mL of puromycin (Sigma) was supplemented to the cell culture media for 3 days to select for transfected cells.

Design of CRISPR constructs for enhancer deletion

Target-specific CRISPR guide RNAs were designed to optimize uniqueness and have limited off-targets using an online tool (<http://crispr.mit.edu/>). Corresponding oligonucleotides were ordered (IDT) and subcloned into the pX330 plasmid (Addgene), expressing a human codon-optimized SpCas9 and chimeric guide RNA expression plasmid, following previously published guidelines (Cong et al. 2013). CRISPR sgRNA sequences used to delete SEs and TEs can be found in the Appendix (Table 9, Table 10).

PCR and Sanger sequencing for genotyping deletion clones

Genomic DNA was isolated from clonal CRISPR targeted F123J cell lines using a DNeasy Blood & Tissue Kit (Qiagen). To screen deletion clones, PCR primers were designed which flank the outside of the CRISPR sgRNAs. PCR genotyping primers are listed in the Appendix (Table 11, Table 12). For PCR, 20ng of genomic DNA was used and amplified with Q5 High-Fidelity DNA Polymerase (NEB).

Subsequently, a qPCR-based genotyping method was designed to quantify the DNA copy number of the targeted regions relative to control regions. qPCR copy number primers were designed inside the targeted region to quantify the efficiency of CRISPR deletion as creating a biallelic or monoallelic deletion. qPCR was performed with 10ng of genomic DNA in triplicate using SYBR green master

mix (Roche) on the LightCycler 480 (Roche). Relative levels of DNA copy number were calculated using the $\Delta\Delta C_t$ method with copy number normalized to GAPDH intron 1 (chromosome 6) as a control. qPCR genotyping primers are listed in the Appendix (Table 11, Table 12).

Finally, Sanger sequencing was used to confirm the qPCR genotyping result and further characterize the deletion junction. PCR products were gel purified and TOPO-cloned using a Zero Blunt TOPO PCR cloning kit (Qiagen). Purified plasmid was Sanger sequenced (Retrogen) using the M13 F primer. The following equation was used to calculate deletion efficiency: Deletion frequency = $((\# \text{ monoallelic clones}) + (2 * \# \text{ biallelic clones})) / (2 * (\# \text{ of total clones}))$.

The following are the primers used:

qPCR copy number control primers: GAPDH intron 1 (chromosome 6)

F: CTGGCACTGCACAAGAAGAT

R: GGGTTCCTATAAATACGGACTGC

M13F Sanger Sequencing Primer

F: GTTTTCCCAGTCACGACGTTGTA

Strand-specific RNA-seq

Total RNA was isolated by TRIzol (Life Tech) extraction. PolyA⁺ RNA was captured from ten micrograms of total RNA using Oligo(dT)₂₅ Dynabeads (Invitrogen). The resulting RNA fraction was subjected to first-strand cDNA

synthesis using random primers and dT oligos in the presence of actinomycin D. Samples were cleaned up by applying the reaction mix to mini G50 columns. The eluted sample was then subjected to second-strand synthesis with DNA polymerase I, DNA ligase, and dUTPs in the presence of RNase H. This ensures that the cDNA strand complementary to the mRNA has dNTPs and that the other strand is specifically labeled with dUTPs. The cDNA was purified by QIA-quick columns and sheared using a Bioruptor (Diagenode). The fragmented cDNA was subsequently end repaired, A-tailed, and then ligated to Illumina TruSeq adapters. The samples were run on a 2% high-resolution agarose gel and fragments were size-selected from 200bp-450bp. The fragments were then treated with UDGase to remove the labeled strand, and mRNA complementary cDNA strand was amplified by PCR and sequenced (2x100bp) using the Illumina HiSeq2000.

For RNA-seq analysis, reads were mapped using Novoalign and USeq. FPKM for known mouse genes were generated using in-house scripts. To identify signature genes, a 3-fold cutoff was used to identify significantly changed genes comparing wild type and enhancer knockout clones. CAST and 129 genotypes were generated, and strain-specific reads at CAST and 129 SNPs were separated as described previously (Hawkins et al. 2010). To improve the accuracy of calling genetic variants between the two mouse strains, we also removed a subset of variants that showed inherent biases from all downstream analyses. Briefly, we generated dummy reads across each SNP/Indel and mapped them to the genome:

SNPs that showed >5% bias and Indels showing >10% bias are ignored. We also checked for copy number variation across the two alleles and ignored variants that showed significant bias (5% FDR). Finally, we removed variants identified as poor heterozygous calls (5% FDR).

Allelic 4C-seq

4C-seq was carried out as previously described (Splinter et al. 2012). Briefly, 10 million F123J cells were harvested and crosslinked in solution with 2% paraformaldehyde. Chromatin was digested with DpnII (NEB), ligated under dilute conditions, re-digested with Bfal (NEB), and again ligated under dilute conditions. Primers were designed to inversely amplify a DpnII-Bfal restriction fragment overlapping the Sox2 promoter and also contain the Illumina Truseq adaptor sequences. The R primer was designed to read into a SNP to discriminate allelic interactions. Libraries were amplified using Expand Long Template Polymerase (Roche) and quantified using the KAPA Library Quantification qPCR kit (KAPA). 4C-seq libraries were sequenced (2x100bp) using the Illumina HiSeq2000.

Sox2 Promoter Allelic 4C-seq primers

F:AATGATACGGCGACCACCGAGATCTACACTCTTCCCTACACGACGCTCTTCCGATC
TACGTCTGGACAATGGGAGATC

R (Adaptor 21):

CAAGCAGAAGACGGCATACGAGATCGAAACGTGACTGGAGTTCAGACGTGTGCTCTT
CCGATCTCACATCTCCAATCCCCATTC

The 4C data were analyzed by the published 4Cseq Pipeline with default parameters (van de Werken et al. 2012b). The 2MB region around the bait was used for scaling and plotting. To split reads by parental alleles, a SNP in the second read was used to assign read one to an allele. Read one indicated the interacting region and each allele was analyzed separately.

ChIP-seq

ChIP-Seq was carried out as previously described (Hawkins et al. 2010) with 500 µg chromatin and 5 µg antibody with the following antibodies, H3K27ac (Active Motif, AM39133) and Total RNA Polymerase II (Santa Cruz sc-899X). ChIP and input library preparation and sequencing procedures were carried out as described previously according to Illumina protocols with minor modifications (Illumina, San Diego, CA). ChIP-seq analysis was performed as described previously in addition to indel realignment using GATK before splitting the reads into two alleles.

FIGURES

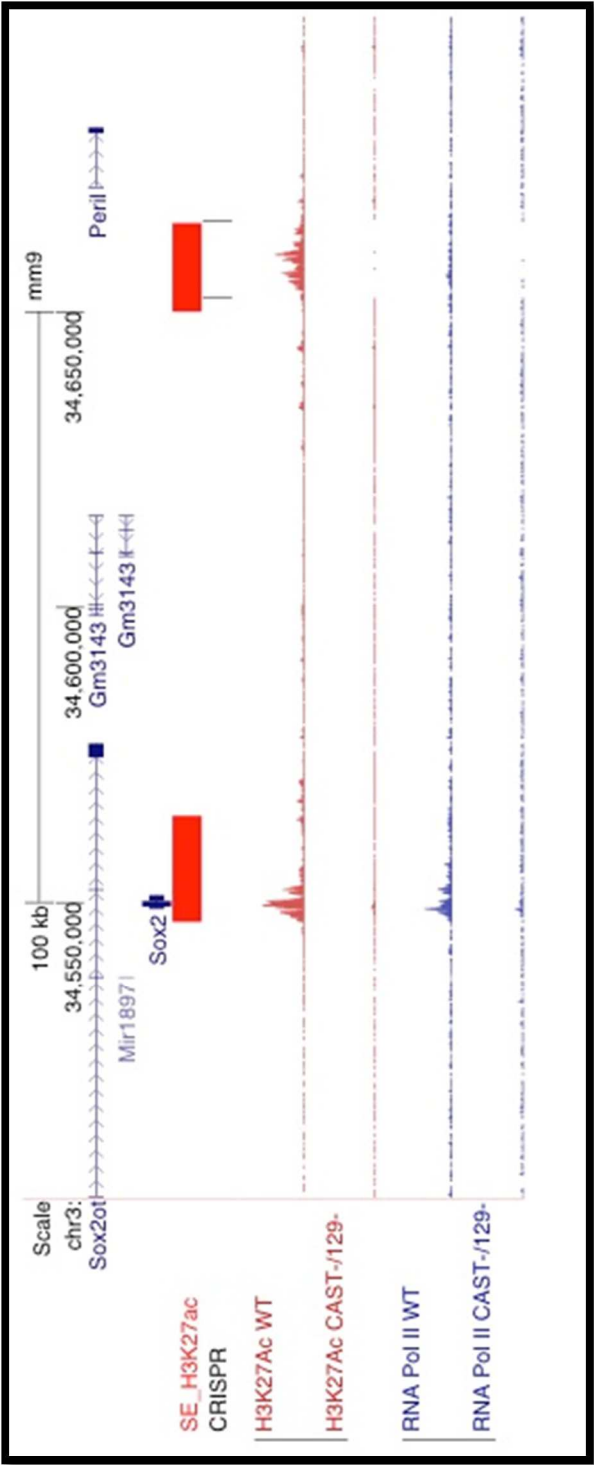


Figure 11. Sox2-SE^{distal} is required for PolII recruitment and H3K27ac deposition at the Sox2 promoter. ChIP-seq signal for H3K27ac and total RNA Polymerase in both the F123J WT and Sox2-SE^{distal} lines are displayed in the UCSC Genome Browser.

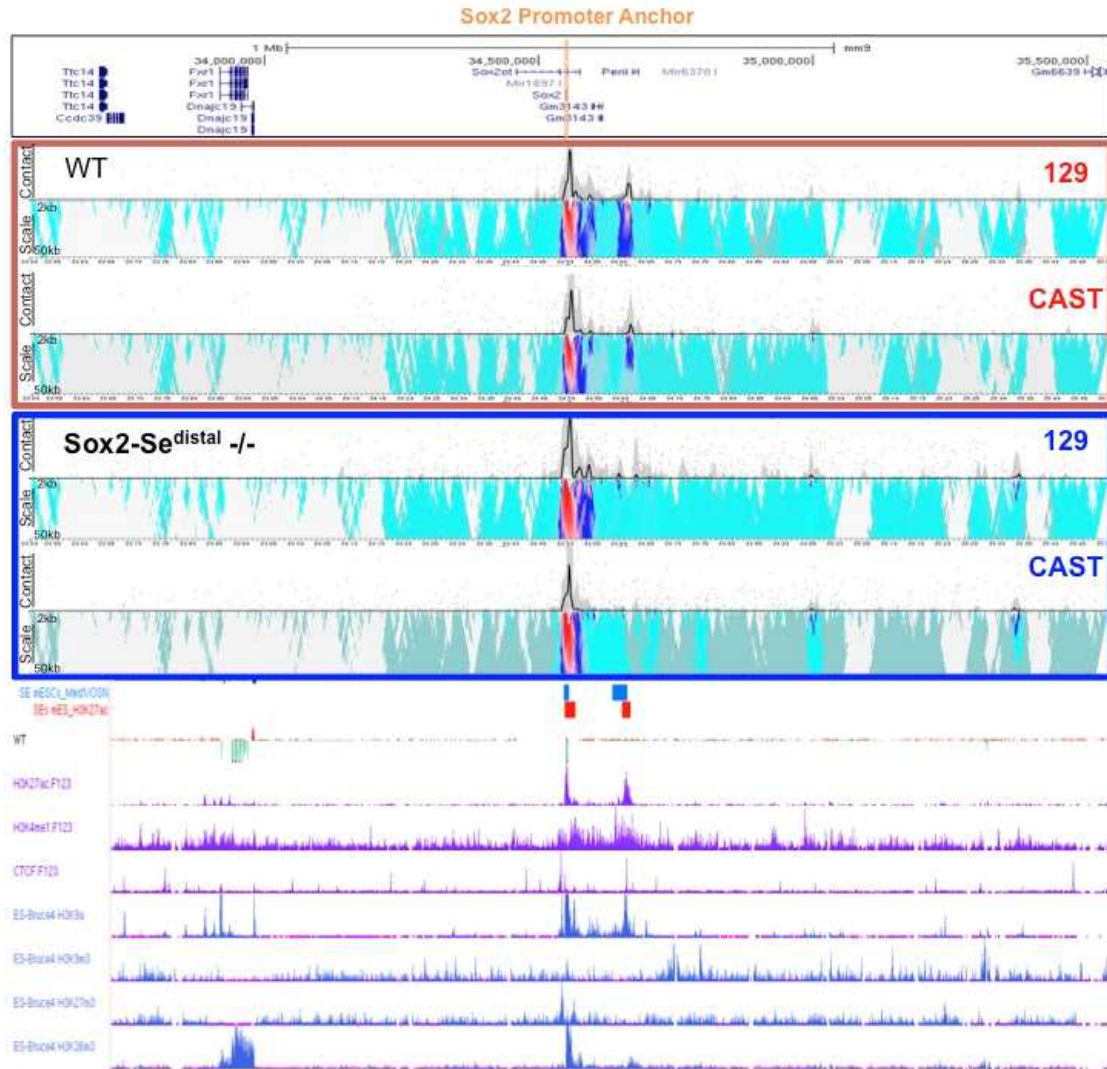


Figure 12. Knocking out Sox2-SE^{distal} alters the local chromatin interactions with the Sox2 promoter. The 4C bait fragment is highlighted in yellow and corresponds to the Sox2 promoter. 4C-seq signal is displayed in two panels, the contact and scale panels. The contact panel depicts the contact profile medians main trend (black line) and the 20th and 80th percentile medians of normalized coverage (smoothed grey lines). The scale panel uses medians of normalized coverage for sliding windows (2kb to 50kb) displayed as color-coded multiscale diagrams with values representing enrichment relative to the maximum attainable 12-kb median value. The UCSC Genome Browser tracks visualize mESC specific super enhancers called by OSN/Med1 signal or H3K27ac, WT F123] RNA-seq signal (Forward strand in green and reverse strand in red), and ChIP-seq signal for H3K27ac, H3K4me1, CTCF, H3K9ac, H3K9me3, H3K27me3, and H3K36me3.

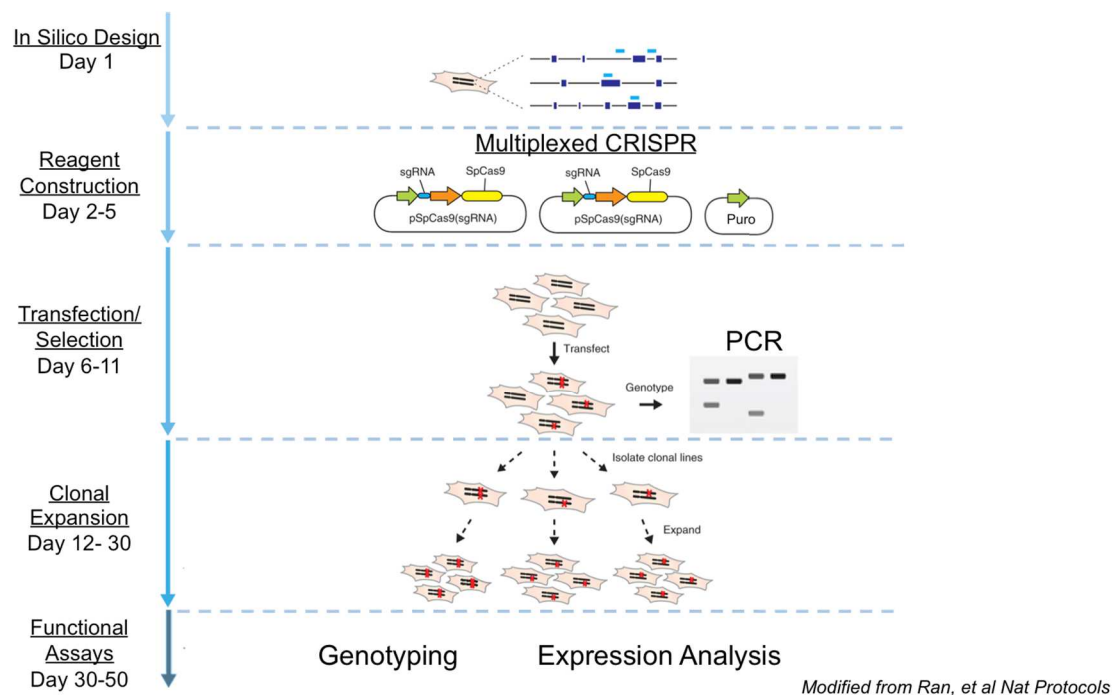


Figure 13. Experimental design for ENCODE 100 enhancer knockout project pipeline. Generation of CRISPR plasmid reagents is discussed in detail in the Methods section.

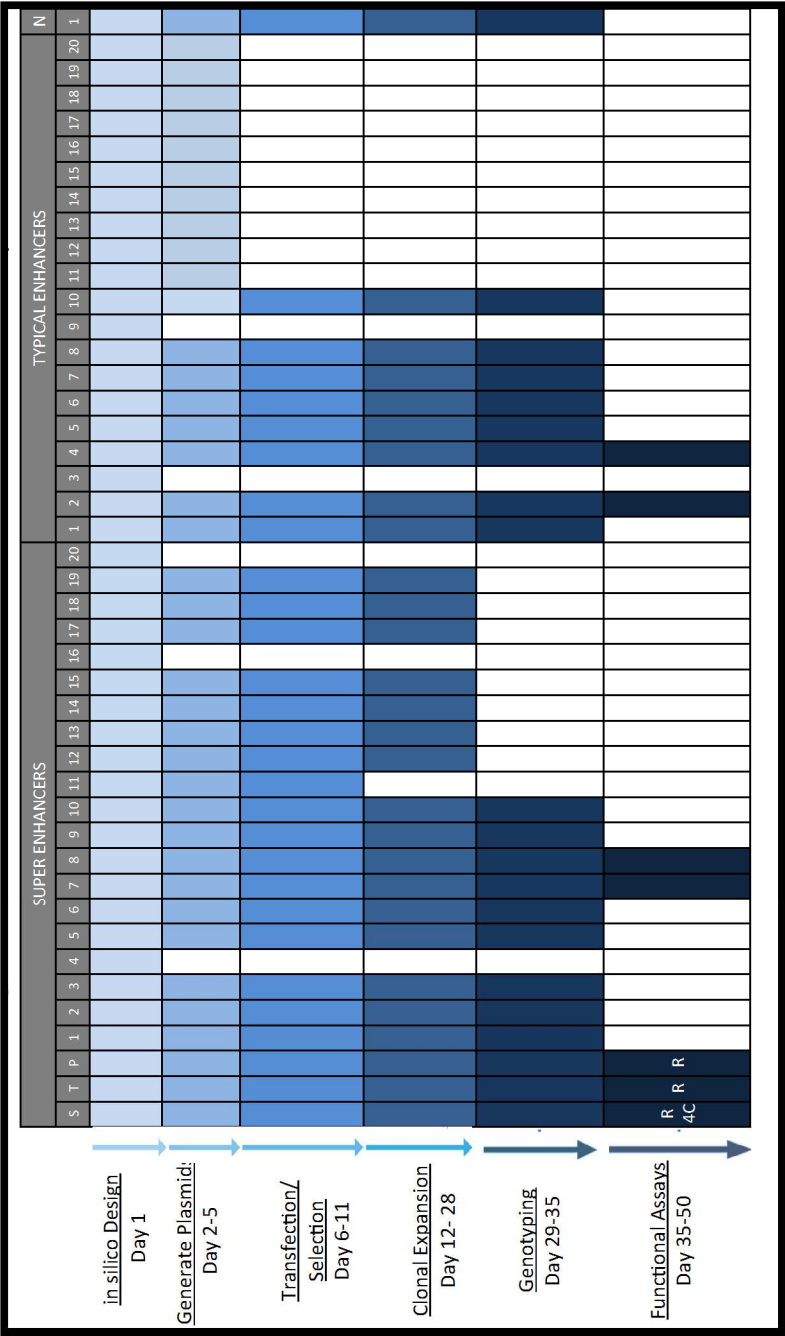


Figure 14. Super enhancer and typical enhancer CRISPR deletion pipeline progress. CRISPRs are designed for 20 super enhancers (SEs), 20 typical enhancers (TEs), and 1 negative control region. Filled blue cells indicate the step is completed and blank (white) cells indicate incomplete steps.

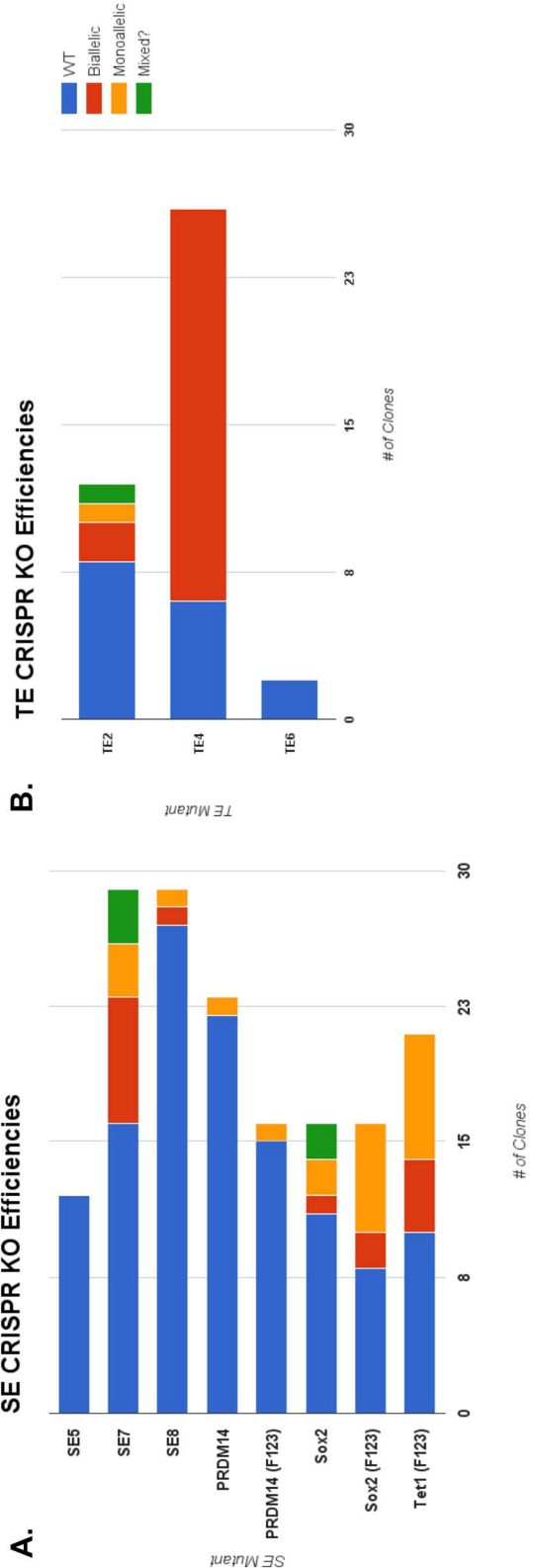


Figure 15. QPCR genotyping results for Super enhancer and typical enhancer CRISPR deletion clones.

Figure 16. Impact of Tet1 promoter and SE deletion on gene expression. (A) This UCSC genome browser shot depicts several tracks of data on chromosome 10 near the Tet1 locus. The sgRNAs designed to delete the Tet1-SE are shown with a thick black dotted line. Also noted are the regions called SEs by either H3K27ac definition (Red) or Med1/OSN occupancy (blue) (Elling et al. 2011; Fullwood et al. 2009; Hnisz et al. 2013; Sawado et al. 2003; Schones et al. 2008; Whyte et al. 2013). H3K4me1 and H3K4me3 signal for mESCs is from previously published studies (Bell et al. 2011; Carette et al. 2011; Gaffney et al. 2012; Li et al. 2014; Mathieu et al. 2000; Spicuglia et al. 2002; Whyte et al. 2012). Allelic RNA-seq data from F123 WT, Tet1-SE KO on the 129 allele, and Tet1-SE KO on the Cast allele are also shown. (B) Comparison of allele-specific expression from RNA-seq in the monoallelic 129 and monoallelic CAST clones. Allele specificity is defined as the log₂ ratio of RNA-seq reads from CAST and 129 allele. All genes on chromosome 10 including Sox2 are plotted. (C) Quantification of allelic Tet1 mRNA signal from the RNA-seq analysis.

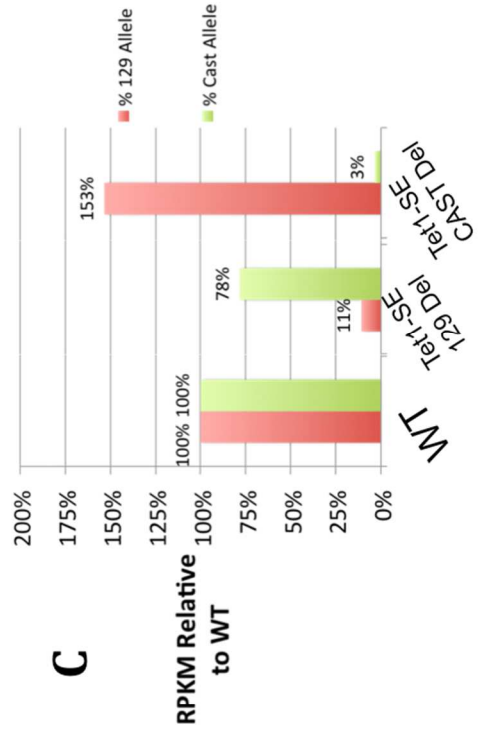
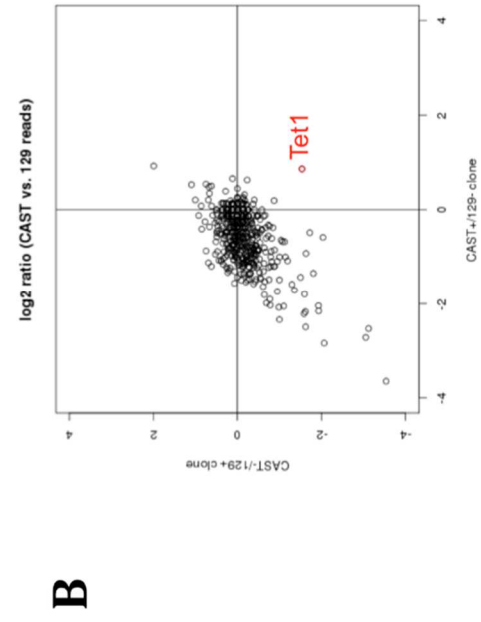
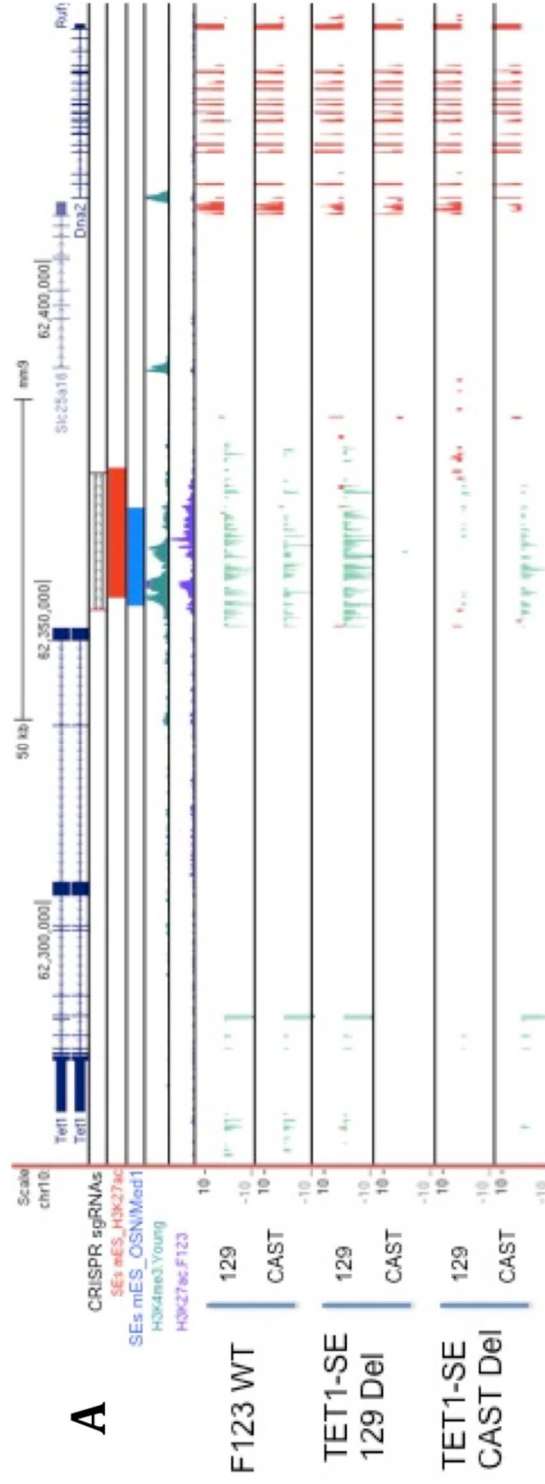
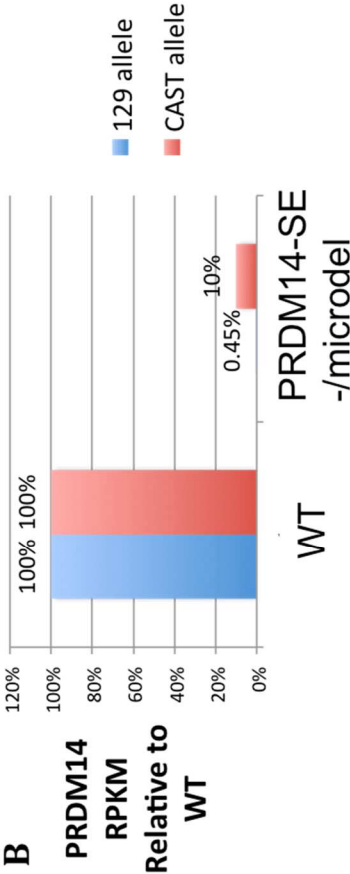
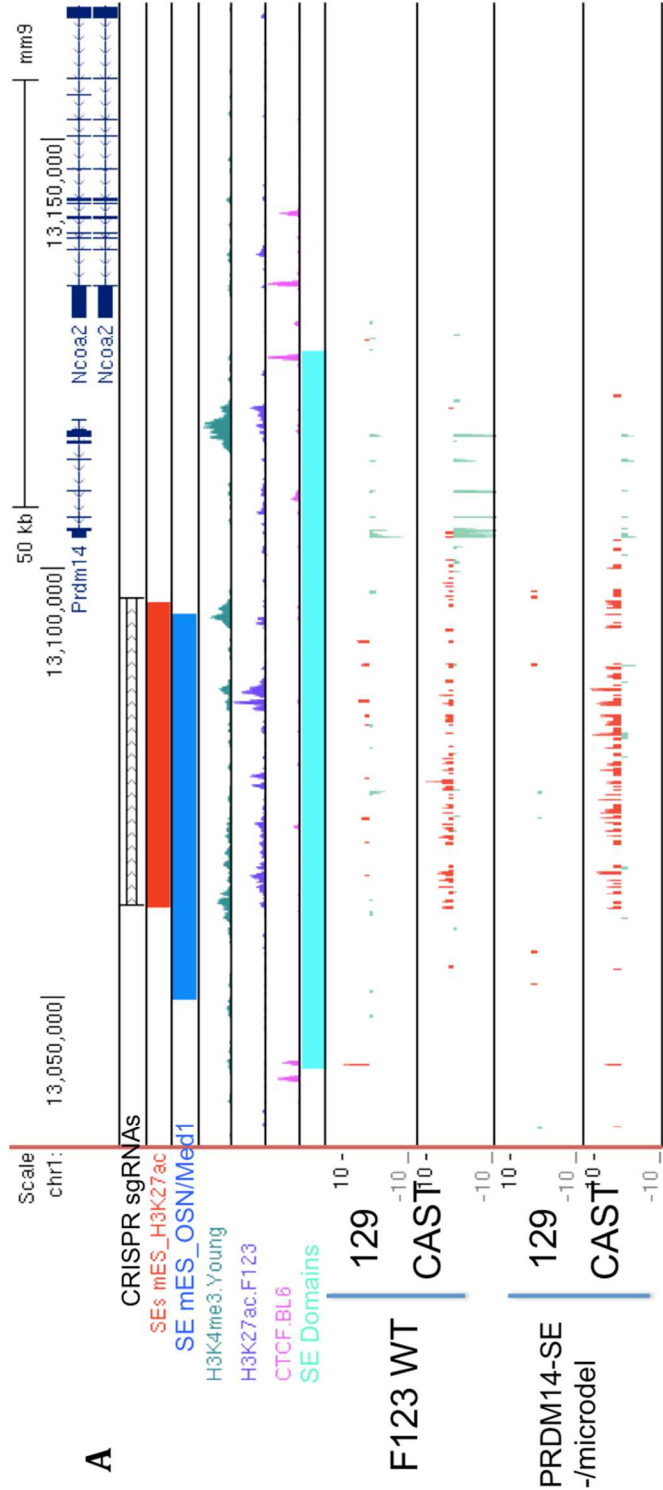


Figure 17. Deletion of PRDM14-SE dramatically decreases PRDM14 gene expression. (A) A UCSC genome browser shot of the PRDM14 locus in mESCs. Tracks denote the sgRNA location and defined SE element and SE domain. ChIP-seq data for H3K4me3, H3K27ac, and CTCF in mESCs is from previously published studies (Brogaard et al. 2012; Heinz et al. 2010; Hon et al. 2013; Jiang and Pugh 2009; Jin et al. 2010; Li et al. 2014; Shen et al. 2012; Sur et al. 2012; Whyte et al. 2012; Ziller et al. 2013). Allelic RNA-seq is displayed for both WT and PRDM14-SE KO cells. Red signal is from the F strand and green signal is from the R strand. (B) Quantification of allelic RNA-seq signal for the PRDM14 gene in WT and KO cells.



TABLES

Table 8. Summary of the results of enhancer knockout studies in mammalian systems.

Mammalian System	Enhancer Target Gene	Insights into enhancer function from KO experiments	Reference
Mouse (mESCs), Human	Sonic Hedgehog (<i>Shh</i>)	Distal enhancer 1Mb away, in intron of another gene, <i>Shh</i> expression completely lost (in situ hybridization) specifically in limb bud	(Lettice et al. 2003; Sagai et al. 2005)
Mouse (MEL)	B-globin	LCR, KO decreases B-globin expression by 96%, Decreases RNA Pol II and GTF promoter binding by 50%, Decreases P-RNA PolII 75%, important for elongation	(Bauer et al. 2013; Sawado et al. 2003)
Mouse (T cells)	TCR β	E β , Required for RNA PolII and CBP recruitment to promoter, important for initiation	(Spicuglia et al. 2002)
Mouse	c-Myc	Myc-335 enhancer 500 kb upstream of Myc gene, 2-fold decrease in Myc expression in intestines	(Sur et al. 2012)
Mouse	CD4	Enhancer 13kb upstream of CD4 (E4p), deletion reduced CD4 expression and H327ac and H3k4me3 at promoter in T cells. Suggests enhancer has a role in promoter remodeling and initiation. When E4p is conditionally deleted after activation of T cells and the CD4 gene, CD4 expression is maintained without the enhancer, suggests enhancers are not necessary for maintenance of gene expression after epigenetic activation of the locus	(Chong et al. 2010)
Mouse, Human	Bcl11A	Enhancer in intron 2, decrease Bcl11A expression 95%	(Bauer et al. 2013)
Mouse	Aicd and Apobec1	Delete E1 and E2 (188kb away) reduces NIPBL and RNA Pol II binding at promoter	(Kieffer-Kwon et al. 2013)

AUTHOR'S CONTRIBUTIONS

Chloe Rivera conceived, designed, performed, and analyzed all experiments and results except those noted below. Dr. Haruhiko Ishii performed ChIP-seq on F123 WT and Sox2-SE^{distal}-/- cells. Dr. Jason Li also generated CRISPR KO lines in F123J and assayed their genotype. Dr. Jason Li and Zhen Ye prepared RNA-seq libraries for Tet1 and PRDM14 SE KO clones. Yunjiang Qiu analyzed the ChIP-seq data, allelic 4C-seq data, and allelic RNA-seq data.

ACKNOWLEDGEMENTS

This work is supported by funds from the Ludwig Institute for Cancer Research, and US National Institutes of Health/ National Human Genome Research Institute (U54 HG006997-03S) to BR. CMR was supported in part by the UCSD Genetics Training Program through an institutional training grant from the National Institute of General Medical Sciences, T32 GM008666.

APPENDIX

Supplemental Tables

Table 9. List of CRISPR sgRNAs used for super enhancer knockouts.

ID	Chrom	Start	Stop	sgRNA_L_s	sgRNA_R_s
Sox2	chr3	34652359	34665321	GACGCTTCCGTTCTTGAGT	GTTGGATTCCCACAACAAGC
Neg1	chr3	34,616,316	34,630,179	GAATTCGTGTTAACGCCAGAC	GTTGTCTCGAACCACCTTGTT
Tet1	chr10	62,345,494	62,366,873	GAGTAACCGGAAGTAACCCGG	GCTATCCCGGCTTCCAATTG
PRDM14	chr1	13,060,528	13,096,476	GTCTGGCACCCATATAACGAT	GCCACGGTGATTGGCTGGACC
Med13L	chr5	118,882,891	118,935,390	GAGCAGTCATTGACCGTTAAG	GGGGGATGGTTACTATCGCC
1	chr12	111184138	111191254	GTTTCAGTAACTATTGGCACCC	GACAGCTCAGCCCCCTCGTCTT
2	chr1	88336186	88344983	GATTCTGTTGGATCTACCAGC	GGCCATTCTAGAAGGACGAC
3	chr4	133368629	133377500	GTGTGGAAGCTTAGCCCATAC	GATCCTCCCTAGGGGAATTC
4	chr5	110894893	110903971	GACAAGGGAGGAGCCCTTATA	GCCTTTGCACAGGTTCCGGTC
5	chr4	118740463	118750403	GGGCCCCACACATGGTAGTA	GTGACAGCTGTGATCCCGAGT
6	chr15	81341768	81355291	GCGCACATTTGCGCGCGGTT	GTTGGATACATCTATGTGGC
7	chr4	55476178	55494550	GTCCCAACCCTGATTGGACCC	GCTGAACCTGAATAATAGCG
8	chr18	38526534	38545997	GTTTGGACGGAACACGCTAAC	GAGTAACGGAGTATGAATGGT
9	chr11	117782809	117803373	GTCATTTCCAGACCCCGCCAC	GTCAACAAAGGCACCCGGTAG
10	chr4	118678043	118701531	GTATCATATCATTAACTCG	GAATACCTTCTATTCCGGGCG
11	chr8	37590739	37614380	GTGGATATGCTTTACGCTTCC	GATCTGAGTTTCGGGTGCTAC
12	chr7	3191738	3215914	GACGTTGCCCACTTCTAGCAC	GCCTAAGACCTAGCAAATTGC
13	chr4	106511199	106542774	GCATCCACTTCCACCCGGTC	GAAATGGGCAGCAGTCGAAA
14	chr4	130177240	130216596	GCTTAGGCCCTGGATAAATCC	GATGGGTGATGTTTACTCCGC
15	chr12	88133697	88175015	GGCATATACTCAGTTCCCCC	GCTACGGCCTGTCCTTACAT
16	chr4	148085445	148128748	GCCATTCATGGAGACTCCGCA	GAGCTGCGCGGCTGTCTTAA
17	chr5	118884350	118935020	GTGGAGGATTCCGGTCTTAC	GTGGTCAGGCCAGACGCCATC
18	chr2	152002081	152017780	GCTTGTTTtagTACCGTCAT	GGTTAGTAAGAGTGGGCGTA
19	chr2	154243688	154261039	GGGAGCCAATTGAAAGCGCT	GGTGCCTAAAAAGGTCTCGG
20	chr4	141113966	141131666	GCCGGCAGTATCGTGACTC	GGCTCCGGGTATCAACTCGC

Table 10. List of CRISPR sgRNAs used for typical enhancer knockouts.

TE_ID	Chrom	Start	Stop	sgRNA_L_s	sgRNA_R_s
1	chr8	74857330	74858302	GACTAGGAGACTTCCCGCTCA	GCCCAGTTCCTGGCAATCCGG
2	chr6	122665619	122666581	GCTACAGAACCTACTTGGGC	GATGCTCAGCCAGTCTATGG
3	chr9	99053600	99054600	GGGAAGTGAAGGCTCCCCTA	GATTGGTTCGTGAAGCAGAT
4	chr3	137496052	137497014	GCCAAAGAATCAACCCCTG	GTAGGACACTCGTGTCTAGGC
5	chr9	118991100	118992100	GTAGCATAGTAAACCGGCTAT	GGTGCTCACCCCGCATCCTA
6	chr16	48303488	48304467	GTGTTCTTTGCACAGCGTCAT	GTATCCCTAGCCTAATCTATG
7	chr4	108306800	108307800	GTGGCCAGCCAGGGATTAGCC	GGTTTTTAAGTCTTATTCGT
8	chr6	55741713	55742726	GATGCCCTCCATCACTACAC	GCTTTAACTACGTCGGTGCC
9	chr11	49619300	49620300	GAAGGGGTCCGTAGGGAAGTT	GACGTGCCATCACAGATACAC
10	chr9	9268943	9269894	GCCCGCTTTTCAGATCACAAT	GTGTAAGGTGGTAGCTAACTG
11	chr8	97710983	97711993	GGAGTTTCTGGTGCGGTTCA	GTCATCCGGGCAAAATAGGCT
12	chr14	79509623	79510586	GCCAACCATTAGGTAGCCAT	GATCGGGAAGCTGATGATCCT
13	chr17	7444695	7445609	GTCTCAGATAGGTCCGATAA	GTACTAAGCAATCTCCATAC
14	chr19	20998236	20999229	GTAGACGGCTGGTATCATGA	GTATAGTGACACCTAGTAGAT
15	chr14	30763757	30764720	GGGAGGACCCGAGCCTATGC	GAAAACTGTTGCAGCCCCGGC
16	chr18	31728208	31729158	GGGTTTAACTGTGCCAATG	GAAGCCAACCTACGAACGAACA
17	chr15	61826135	61827093	GAGTGGTGGAAGTTCGAGTCC	GCGAATCTCTCTTAACGGGGT
18	chr10	33949762	33950731	GAAGGTGAGATTTTCGGAGCT	GCCTACTGGTATAGAGTAATG
19	chr17	71121493	71122443	GGCTTGTTGTTTACCTAAAC	GTTCTTAAGGACATTATACGT
20	chr1	120586239	120587188	GCCGAGTTAATGCAGCACGGA	GTCTAGGCGGCTGCAACACAG

Table 11. Primers used to genotype super enhancer knockout clones by PCR and qPCR.

Primers for PCR Genotyping		
ID	F	R
Sox2	CCTCAAAGTGCATCTCGTCA	GTCCCCTCGGCTTCTGTACT
Neg1	TGCACACACATACTCCACAAAG	GCCACATGTGTTAGCTGGTG
1	TTGAGGAGCTCTAACAGGTAGGA	GGAGGGTCAGCTGCATACTT
2	GACTGTCCACCACAGCTGAA	AAAGAATTTGAAAAGGGGGTTT
3	TTATGTTGTTGGGGGAGCAC	CCTGCAGTGACCAGGAATTT
4	TGCCCAGAACCCTAGAGAAA	CACACACTGCCACCACAAAT
5	ATGGATGGAAGAGAGGCAGA	TCCCCTGTTTCATTTCACGATT
6	GTATCCCAGGACAGCCTTCA	GGATCAGCACAGGGGAATAG
7	ACAAGCAGGGCATGTCAAC	GTGGGGCTTGTGTTTCAGAT
8	GGCCACTTCCTGTGAAAATG	GCCCTAGATGATGTCCCTGA
9	CAACTTCATCTTGTTACTCCTGAAGA	CTGGGATCTGAGCTTGGTCT
10	GAAACCCAACAGAGCTTGGA	CCATCTCAGCATGGGAAAAT
11	CCACACACCCACCTCTACCT	CCAGCTGGTCACCCTAAAC
12	CTTTGGTTGGTAGGAGAGGT	CAAGGCTCAAGTACGCCAAG
13	GTAGGGAGGAAGCCAGCTCT	CTGTGGTCTGGAGAACAGAGAGT
14	GCTGTCTCTGTTTCAGGCTCTC	CAGACAACCTCTGGGCTCCAT
15	CAGTGGGAACTGGCAAAAC	CAGCCCCTTTACTTTGCTTG
16	GAATTGCAAATGCACACAGC	GTCCTTTGCCTGGAATGTGT
17	CAGTTGCCACATGACAATC	GCAATTGAGCGTACGAATGA
18	GCATATATCCTCAAATTTCTTGGA	CGCTACCACATTTGGCTTTT
19	GCTTTTCACCACCCAAAGTC	GGAATTAGCTGGGGCTCTTC
20	GCAATCCAGAGGTCAAATTC	CCACTTCAGCTTTGCTGCTA
Primers for qPCR Genotyping		
ID	F	R
1	AGATTGCACGCAGCCTAAGT	TCCCAGAGTGAAGGATGGAC
2	CGCATCTCTGAGGTTCTGT	CTACAACTCCCAGGGCAGAA
3	CGGCAATACAGTGCAAAGAA	TGACTCCTCAGCACCTTCCT
4	GGTGCCCTGTCAGAAGGATA	CGCTTCCAGTCTTCTTGGAG
5	CCACAAACCTCCTAGGGTCA	CCTCTGGCTTTCTAGCCTCA
6	CAAATTGAAGCCAAGCCATT	GGTTGTTGGCAGAGCTTGTT
7	GACCCAGGAACTGAATCGAA	GCCCTGCACTAGCATCTTTC
8	TCTCTCTCGGCTCTGGACTC	GGTCTGCTGCAGCTGTGTAA
9	GATCAGGTGGCTCAGTGGAG	GCTCAACAAACGTCTCTGTCC
10	GAATTTTCTCGGCCCTATC	ACAGTCTGCTGCCTCCAGAT
11	GAGAGAGGTGGCGTTCTTTG	ACGAGTGGCCTTGTTACACC
12	TGGGTGTTTAAGCAGAAATTGG	GATCTAAGGCAATGCGGAAG
13	GCTCTCTGGGGTCACTGTA	GAAGGCAGGTGGTTTCTGAT
14	AGAGGTTGGCATGGGTACAG	TTGCAATTTTGCCCTTTGAT
15	GAAGCTATCCATCGGTCCAC	GTTCTGGCTGCCTCTTTGTC
16	GAGACGCTTGGTTGAAGAGG	CCCATTAGGCAACCACTCAT
17	TCTGTGAAATCCCTGCCTTC	GCTTGCTTCTATGGCTGGAG
18	CCGGTGCTCCTTCAGTTTAG	GTTATGGGAGGAGGGATGCT
19	CCCCACAACACCTCTAAACG	TGCCAAAGAATGCAAATCAG
20	GTTATTCCCCACCCTCCACT	GGAAAGAGCGTTGAGCTGAC

Table 12. Primers used to genotype typical enhancer knockout clones by PCR and qPCR.

Primers for PCR Genotyping		
ID	F	R
1	AAATAACCACCGTGAAGTCA	AAGAAGCAGCGAAACACATT
2	GGAGGGACCACAGTCTCTTG	TCACCTTCAGAGCTTGACACA
3		
4	ACAGCGCCAGAGACAGAAGT	TGCCAACTTGACACATCTGAG
5	GAGGCAGACAGAAGCCAGTC	AGTGCAAATGGAGACGACCT
6	CCTGGCTCAGGACATAGCAT	TGGAGCTTTTCCTCAGGATG
7	CACTGCACCTGTGCATTCTT	ATGCTTACACGCCACATTCC
Primers for qPCR Genotyping		
ID	F	R
1	CTGAGGGCAGCTAGAGGAAA	CATGAGAAGAAATGGCCAGAA
2	CAGAGGCTTTTTGCTTCAGG	AGGGGAAAAAGGAGGATTTG
3	GCTGGGCATGTAAGTGTCACT	GTGGGGATGTTGCCTATTTG
4	GCCAACTCCTTTTCATTTCA	GTGTGGACCTTGGAGGACAG
5	CTCCTGGGGACAAAGAATGA	CTCTGGGGAAACTGAGTCCA
6	ACACAGCGAGTTTCCACCTC	CCAGCTTCCCTGGTTTAGTG
7	GCCACTCTGGCCAATCTTAG	GGAAAGAGTTGGGATGCTGA
8	CCTACTGAGGCTCACCCAAG	CCCACCTAGGGCATCAATAA
9	CAGGGTTCTCTCCTCCACTG	TGCCTGTAAACCCTTTCTGG
10	CAGAACAAAGCAGAGGCACA	CTGGACTCCACCTTTCTTGG

APPENDIX

Supplemental Chapter: Mechanisms of histone H3 lysine 4 methylation at transcriptional enhancers

Abstract

Precise regulation of temporal and cell-type specific gene expression in eukaryotes depends on the interaction between cis-regulatory elements and trans-acting factors. The enhancer, one such cis-regulatory element, facilitates the decondensation of chromatin and the assembly of transcriptional machinery at the proximal promoter. Recent studies from our lab identified a novel chromatin signature at enhancer elements, presence of H3K4me1 and lack of H3K4me3, which can be used to identify enhancers on a genome-wide scale. Enhancers show more variation in different cell types than promoters and insulators indicating enhancers are the most important regulatory cis-regulatory element for cell-type specific gene expression. Here, I hypothesize that sequence specific transcription factors recruitment of a H3K4 monomethyltransferase to enhancers forms a chromatin state necessary for appropriate gene expression. In order to test this hypothesis, we are conducting both an unbiased screen and a candidate approach to identify the methyltransferase complex responsible for H3K4 methylation at enhancers. Identification of this complex will enable further investigation into the role of the H3K4 monomethyl mark at enhancers including answering whether it is necessary for transcriptional activation. The proposed study will identify the chromatin modifying enzymes that regulate methylation at enhancers and determine whether this chromatin mark is an important regulator of transcription. Since aberrant transcriptional activation is a common feature of cancers, the results of this study may yield insight into the mechanism of misregulated enhancers in tumorigenesis and human development in general.

Background and Significance

Transcription and chromatin

Precisely regulated expression of protein-coding genes is essential for complex processes such as development and differentiation. Eukaryotic transcription occurs through a sequence of steps that involve recruitment of the pre-initiation complex (PIC), initiation, promoter clearance, elongation, and termination of transcription. The packaging of the DNA template into nucleosomes to form chromatin affects all these stages of transcriptional activation (Workman and Kingston 1998). Nucleosomes are composed of 147 base pairs of DNA wrapped around the histone octamer of H2A, H2B, H3, and H4. The histone tails and globular domains are subject to over 60 post-translational

modifications, including methylation, acetylation, phosphorylation, ubiquitination, sumolation, and ADP-ribosylation (Kouzarides 2007). It has been proposed that histone modifications have three main functions (Li et al. 2007):

- (i) Histone modifications can change the net charge of nucleosomes and such a charge can antagonize DNA-histone interactions. One example of this is acetylation, which adds a negative charge to DNA-histone contact points thus repelling the negative phosphate backbone of DNA.
- (ii) modifications can act as binding substrates to recruit proteins that affect chromatin function.
- (iii) modifications can directly affect chromatin structure by promoting open or closed conformations

Our lab has recently identified a unique histone modification pattern at enhancers. However, how this chromatin signature is established and its role in enhancer function are both unknown.

Transcriptional Enhancers

Enhancers are defined as DNA elements that increase transcription from an associated promoter independent of their orientation or position relative to the proximal promoter. These *cis*-regulatory elements can be located upstream or downstream of the promoter, within an intron, or even on another chromosome. The dominant model of enhancer function is direct interaction of enhancers and promoters through looping of intervening DNA sequence (Bulger and Groudine 2011). Recent advances in chromosome conformation capture technologies (3C, 4C, 5C, and Hi-C) have confirmed enhancers and promoters interact genome-wide. How this interaction results in an increase in transcription is debatable. Data from differing experiments suggest that RNA polymerase II could be transferred from the enhancers to the promoter, looping could increase the local concentration of factors that recruit polymerase at the promoter, or looping could migrate the gene locus to a favorable transcriptional territory in the nucleus (Krivega and Dean 2012).

Genetic and biochemical analyses of individual loci and more recent genome-wide studies have helped define the structural characteristics of enhancers. Enhancers are sensitive to DNA nucleases due to high rates of histone replacement and enrichment of unstable histone variants H3.3 and H2A.Z (Jin et al. 2009; Mito et al. 2007). This open chromatin conformation at enhancers is thought to contribute to preferential binding of transcription factors whose DNA binding sequences are more accessible at these sites (John et al. 2011). This leads to another key feature of enhancers - their sequences are enriched with clusters of TF binding motifs (Zinzen et al. 2009). The ability of cell-type specific TFs to activate transcription is dependent on the recruitment of coactivators, which are often broadly expressed. Transcriptional coactivator proteins found at enhancers include histone modifying enzymes, ATP-dependent histone remodelers, and subunits of the mediator complex which interact with the basal

transcriptional machinery at promoters (Buecker and Wysocka 2012). TFs regulate subsets of enhancers in a certain cell-type but coactivators are broadly recruited to enhancers. Therefore, genome-wide profiling of coactivators is a compelling strategy for enhancer mapping.

Distinct chromatin signatures define functional classes of enhancers

P300 is a transcriptional coactivator and histone acetyltransferase recruited to both promoters and enhancers. By using proximity to transcriptional start sites (TSSs) to eliminate promoter sequences, our lab was able to use the binding pattern of p300 to analyze chromatin marks at distal enhancers (Heintzman et al. 2007). Using chromatin immunoprecipitation hybridized to a DNA microarray chip (ChIP-chip) we mapped p300 and several histone marks to the human genome. Computational clustering analysis revealed that enhancers are marked by broad H3K4me1 peaks and are depleted of H3K4me3 (Figure 18). This finding is in contrast to active promoters, which are strongly correlated with H3K4me3. Furthermore, we were able to use the distinct chromatin signature at enhancers and chromatin immunoprecipitation followed by sequencing (ChIP-seq) to identify 55,000 potential transcriptional enhancers in several human cell lines (Heintzman and Ren 2009). These observations were validated by *in vivo* lacZ enhancer reporter studies in mouse (Visel et al. 2009) and by genome-wide studies in fly (Nègre et al. 2011) and zebrafish (Aday et al. 2011; Fullwood et al. 2009).

The simplicity of genome-wide profiling of enhancers has promoted further dissection of enhancer functional states. Enhancers can be divided into active, poised, and decommissioned sub-classes. The activation level of enhancers can be estimated by proximal gene expression since enhancers tend to reside near the promoter they regulate. Using this assumption, several groups identified H3K27ac as a mark of active enhancers in mESCs (Rada-Iglesias et al. 2011) and adult mouse tissues (Creyghton et al. 2010). Active enhancers recruit Pol II which transcribes short bidirectional RNAs called enhancer RNAs (eRNAs) and interestingly eRNA expression levels correlate with the expression of the target gene (Kim et al. 2010; Melgar et al. 2011; Wang et al. 2011). The function of eRNAs is unknown but one researcher speculates “the marking of enhancer elements by eRNAs as well as H3K4me1 hints at a scenario where eRNAs could interact with the chromatin to establish or maintain H3K4me1 or where the H3K4me1 promotes eRNA transcription” (Flynn and Chang 2012). Two classes of poised or primed enhancers co-exist in hESCs mESCs, and differentiated cells – a H3K4me1/H3K27me- class and a H3K4me1/H3K27me3 class (Rada-Iglesias et al. 2011; Zentner et al. 2011).

Primed developmental enhancers are modified by Polycomb-mediated methylation. Upon differentiation into a cell type where the enhancer is active, the H3K27 demethylases UTX and JMJD3 have been shown to remove H3K27me3 resulting in rapid acetylation of this lysine residue (Lee et al. 2006; Svotelis et al.

2011). Finally, during stages of differentiation a subset of enhancers must be silenced as they do not pertain to the newly acquired cellular identity. It has been suggested that LSD1, a H3K4/H3K9 demethylase, and the HDAC activity of the NURD complex are responsible for decommissioning enhancers by removing the H3K4me1/K27ac marks during differentiation of mESCs (Whyte et al. 2012). The enzymes responsible for poised enhancers and removal of enhancer methylation marks have been described (Figure 19), however the methyltransferase(s) responsible for H3K4me1 at enhancers has not been identified.

It has been shown in yeast and mouse that H3K4me3 levels are elevated at 5' ends of genes and correlate with high gene expression levels and active RNA polymerase II occupancy (Shilatifard 2012). H3K4me2 is also present at gene bodies and promoters in highly transcribed genes. These genome-wide studies and others have led to a consensus in the field that H3K4 methylation correlates with active genes and open chromatin (Ruthenburg et al. 2007). However there are many unanswered questions about the role of H3K4me1 at enhancers in gene expression. Which enzymes are responsible for the establishment of the H3K4me1 mark at enhancers? How and when are these enzymes recruited to enhancers? Is this modification necessary for enhancer function? Does this modification serve to recruit transcription factors or chromatin modifying enzymes to enhancers?

Histone 3 Lysine 4 Methyltransferases

In mammalian cells, methylation of H3K4 is a conserved feature of the SET1/MLL family (hereafter MLL complexes) of methyltransferases, which includes six members: SET1A, SET1B, MLL1, MLL2, MLL3, and MLL4 (Shilatifard 2012). MLL complexes have a Su(var)3-9/Enhancer of zeste/Trithorax (SET) domain which is conserved back to yeast and each complex has been shown to be capable of mono-, di-, and tri-methylation of H3K4 (Ruthenburg et al. 2007). However, it is clear that their activities are not redundant as deletion of MLL1 (Yu et al. 1995) or MLL2 (Glaser et al. 2006) in mice results in embryonic lethality. Deletion of MLL1 in mouse embryonic fibroblasts downregulates expression of only 5% of promoters, including the HOX genes, does not affect global H3K4 methylation levels, but does correlate with diminished H3K4me3 levels at HOX genes specifically. Another study published conflicting data demonstrating that KD of MLL1 in MCF-7 breast cancer cells causes a global decrease in all H3K4 methylation marks (Jeong et al. 2011). There is a consensus on depletion of SET1A/SET1B resulting in a global decrease in H3K4me3 but not affecting global H3K4me1 levels (Wang et al. 2009). Together this data suggests that MLL complexes have consistent specificity for lysine 4 of histone 3 but have specialized functions and targets in the cell.

MLL family enzymes are only fully active in the context of a core multi-subunit complex, which include WDR5, Rbbp5, ASH2L, and DPY30 (WRAD). The mechanism by which WRAD promotes methylation involves allosteric control,

physical interaction and recruitment to chromatin (Wysocka et al. 2005), and possibly the utilization of a second H3K4 methyltransferase active site present within WRAD itself (Patel et al. 2011). In contrast to the MLL enzymes, deletion of any of the core subunits effectively reduces the global level of H3K4 methylation (Dou et al. 2006; Patel et al. 2011). Besides the core subunits there exist several complex specific subunits which are responsible for the specific properties of the different complexes. For example, SET1/SET1B complexes contain a Wdr82 subunit that recruits the complex to H2B monoubiquitination marks at promoters after transcriptional initiation (Lee and Skalnik 2008). However, SET1A/SET1B can also be recruited in a transcription-independent manner via the Cfp1 subunit which directs binding of the complex to unmethylated CpG islands commonly found at promoters (Thomson et al. 2010). The final aspect adding to the complexity of MLL complexes are their interaction with other complexes as identified by co-purification of native MLL complexes with the H3K27 demethylase UTX (Lee et al. 2007), histone deacetylases HDAC1 and HDAC2 as part of the NURD complex (Nakamura et al. 2002), and subunits of the BRG1 chromatin remodeling complex (Rozenblatt-Rosen et al. 1998).

The literature provides some anecdotal insights or hints into which methyltransferase complex may be responsible for monomethylation of H3K4 at enhancer elements. Originally, it was thought the histone methyltransferase PR-Set7/Set8/KMT5a protein was responsible for H3K4 monomethylation but this has since been disregarded due to conflicting evidence. PR-Set7 has a SET domain unrelated to the MLL family SET domains. It was initially biochemically described from Hela extracts as a H3K4 specific methyltransferase with robust monomethyltransferase activity (Wang et al. 2001). More recent studies have dismissed these claims and instead find PR-Set7 to robustly catalyze H4K20me1 as reviewed by Reinberg and colleagues (Beck et al. 2012a). One study characterized human ASH1's *in vitro* H3K4 methyltransferase activity but three independent labs show consistent results that human ASH1 functions as an H3K36 methyltransferase and localizes to transcribed gene bodies (An et al. 2011; Tanaka et al. 2007; Yuan et al. 2005). Finally, there are also reports describing MLL1 as a H3K4 monomethyltransferase. It has been shown that the product specificity of SET domain containing HMTs is structurally dependent on a phenylalanine/tyrosine (F/Y) switch residue within the SET domain. SET HMTs with a F-residue have larger active sites that can accommodate larger side chains and therefore can sequentially methylate substrates to tri-methylation. In contrast, SET HMTs with a Y-residue have limited active site volume and are therefore primarily monomethyltransferases (Collins et al. 2005). For example, PR-Set7 has a tyrosine residue making it a monomethyltransferase, however mutation to phenylalanine causes tri-methyltransferase gain-of-function (Yao et al. 2012). The MLL family of HMTs is an exception to this rule however, as all MLL family members have a tyrosine at the switch residue yet can trimethylate H3K4. It has been suggested that upon association of MLL family HMTs with the

Ash2l or Rbbp5 core subunits an allosteric change to the active site occurs allowing trimethylation (Southall et al. 2009). Another group demonstrates MLL1 only has monomethyltransferase activity on histone peptides *in vitro* and the that novel non-SET domain activity of Rbbp5/Ash2L allows sequential methylation to trimethylation (Patel et al. 2011). This preliminary evidence from the literature suggesting MLL complexes may monomethylate H3K4 is the reason for our candidate screen of MLL family members.

Preliminary Studies

Co-localization of candidate enzymes with H3K4me1 peaks

As previously described, our study found an important distinction in the chromatin modifications at enhancers as comparison to promoters. Promoters are enriched with H3K4me3 centered at the transcriptional start site (TSS). However, enhancers generally lack this modification and instead correlate with H3K4me1 peaks. A goal of this study is to identify the enzyme(s) responsible for the establishment of this mark at enhancers. As a pilot experiment, we decided to ask whether any of the known H3K4 methyltransferases co-localize with H3K4me1 peaks at putative enhancers. In order to carry out this study, we used ChIP-chip to analyze the binding of the MLL1, MLL4, SET1, Ash2L, RbBP5, WDR5 and H3K4 demethylase enzymes LSD1, Jarid1A, and Jarid1C to 1% of the human genome. From this pilot experiment we were able to identify putative enhancers in the ENCODE regions by their H3K4 methylation pattern. One such putative enhancer is in the ACCN4 gene on chromosome 2 (Figure 20). We were able to see enrichment of both methyltransferases and demethylases at this H3K4me1 peak. As a consequence of this initial study, we made two predictions as to how the H3K4me1 mark could be established at enhancers: (1) methyltransferases capable of only mono-methylating the enhancer are recruited, and (2) methyltransferases capable of mono-, di-, or tri-, methylation are recruited with demethylases in an equilibrium which keeps H3 monomethylated at lysine 4.

One drawback of this experiment was that the enzymes binding near the H3K4me1 peak could be involved in the establishment, maintenance, or removal of this mark. This static system is just a snap shot of the histone modification pattern and enzyme binding cannot resolve between these three phases. Therefore, it was imperative we find a system in which H3K4me1 peaks can be induced to determine which proteins are recruited during establishment specifically.

A kinetic study of an in vivo enhancer activation model system

In order to identify the histone modifying enzymes involved in the establishment of H3K4me1 at enhancers it was imperative to find an induction or stimulation system in which enhancers are *de novo* mono-methylated. A limited

number of such H3K4me1 induction systems have been described in the literature. SOX2 enhancer binding in mESCs (Liber et al. 2010)

and enhancer binding of the pioneer factor FOXA1 during neural differentiation of mouse embryonic carcinoma cells (Sérandour et al. 2011) both induce H3K4me1. A third study determined PU.1-ER tamoxifen inducible expression in mouse hematopoietic progenitors causes promoter-distal PU.1 binding sites to gain H3K4me1 (Heinz et al. 2010). PU.1 is an ETS family transcription factor necessary for the generation of myeloid and lymphoid lineage differentiation from hematopoietic stem cells (DeKoter and Singh 2000). The PU.1-ER system allows culturing of mouse hematopoietic progenitors and selective induction of PU.1 expression in the nucleus and their differentiation into macrophages when cultured in the presence of 4-hydroxytamoxifen (4-OHT). After one hour of induction PU.1 bound to 7,000 genomic regions, and after 24 hours a strong increase in H3K4me1 at a large proportion of these sites. These data suggest PU.1 directly or indirectly recruits the histone modifying enzymes responsible for deposition of H3K4 methylation at enhancers.

For our purposes, the PU.1 inducible system is also a H3K4me1 inducible system. Therefore, we decided to use this system to determine which HMTs or demethylases are recruited to newly established H3K4me1 peaks at enhancers. I first replicated the data published by Heinz, *et al.* by inducing PU.1-ER cells with 4-OHT using a more detailed time-course (0, 1, 3, 6, 12, and 24 hours). I determined the induction of several PU.1 target genes by RT-qPCR was consistent with published microarray data (Weigelt et al. 2009) (Figure 21A). For this study, I selected five enhancers with greater than 4-fold H3K4me1 induction from published ChIP-seq data (Heinz et al. 2010) and luciferase enhancer reporter activity (Figure 21B, 21C). ChIP-qPCR analysis confirms that LSD1, an H3K4 and H3K9 demethylase, is recruited to two (Il1rl2 and TLR3) PU.1 target enhancers during H3K4me1 establishment (Figure 21D). A recent study suggests a seemingly inactive LSD1 is bound to 97% of H3K4me1-marked enhancers in mESCs and upon differentiation its demethylase activity removes the mono-methyl mark from silenced enhancers (Whyte et al. 2012). ChIP-grade antibodies for MLL2, MLL3, or MLL4 were unavailable. Robust and consistent ChIP-qPCR signal could not be obtained using antibodies against MLL1 or Jarid1C. Due to a lack of reliable reagents against candidate mono-methyltransferase enzymes for ChIP we sought to design a HMT antibody independent screening strategy to perform in parallel.

Establishment of a high-throughput genetic screen in haploid human cells for regulators of enhancer activity

Failed ChIP experiments resulting from poor quality antibodies against our candidate enhancer mono-methyltransferases stresses the importance of a HMT antibody-independent screening method. Another important consideration is the enzyme responsible for methylation at enhancers may not have a SET

domain or have been previously described as a methyltransferase. Here, we propose an unbiased genetic screen in haploid cells for mutations causing a global decrease in H3K4me1 levels or enhancer reporter activity.

In the past few years, two separate research groups have described gene trap mediated genetic screens in haploid mammalian cells. The first published study used a pseudo-haploid (disomy of chromosome 8) human chronic myeloid leukemia cell line (Carette et al. 2009). They transduced these cells with a gene trap virus and used GFP as a reporter for gene disruption. Gene trap vectors are cleverly designed with several key features: they are promoter-less such that the reporter is only expressed when the virus inserts within a gene, they contain a strong splice acceptor for expression when inserted within an intron, and gene trap vectors end with a polyA signal which will truncate and likely disrupt expression of a functional gene product. Purification of haploid cells with viral insertions in a gene by flow sorting for GFP⁺ cells yielded a gene trap library of cells with myriad independent insertions. This gene trap library could then be used in a negative selection screen for mutations conferring resistance to bacterial toxins. The remaining research group developed a pathogenic haploid mouse embryonic stem cell line and used a gene trap screen for ricin resistance (Elling et al. 2011). Screening of haploid gene trap libraries provides yeast-like genetics in human and mouse cells and we were provoked by the idea of using this system to identify regulators of chromatin states.

Establish a gene trap library in haploid human cells

In order to establish this system in our lab I purified haploid human cells by sorting, transduced them with gene trap vectors, mapped the insertions using an LM-PCR method coupled with next generation sequencing, and validated our viral insert mapping method via cloning and Sanger Sequencing. KBM-7 cells are a nearly-haploid chronic myeloid leukemia cell line clone with the Philadelphia chromosome and a propensity for diploidization over time in culture. Therefore, I used DAPI stained cells of mixed ploidy to identify haploid cells by DNA content. Upon gating on DAPI stained haploid cells it became evident that the small haploid cells in the population could be sorted by size without the need for a DNA stain (Figure 22A). Large-scale retroviral packaging in 293T cells using calcium phosphate transfection of retroviral and gag-pol-env plasmids was optimized for high titer virus. As a pilot experiment to test the retroviral mapping protocol, 293T cells were infected with the GFP gene trap virus and 100,000 GFP⁺ cells were collected (Figure 22B). As detailed in Figure 23 and published (Carette et al. 2011), genomic DNA was purified from infected cells and digested with a 4-base cutter methyl-insensitive restriction enzyme known to cut just upstream of the retroviral 5'LTR. Linear amplification mediated PCR (LAM-PCR) using a biotinylated primer specific for the viral LTR was used for 100 cycles. Single stranded LAM-PCR products containing the viral-genomic junction were captured via magnetic streptavidin beads. A ssDNA linker was designed with a 5'

phosphate, to facilitate ligation, and 3'OH so prevent linker concatamers and ligated to captured products with a ssDNA ligase. Finally, PCR using primers using a nested retroviral 5' LTR primer and a primer which anneals to the linker will amplify the viral-genomic junction and add Illumina adaptor sequences. Each molecular step of this strategy was optimized for efficiency and sensitivity. 8 picomoles of the 293T gene trap viral insertion library was run on the Illumina Hi Seq using a custom sequencing primer such that reads start at the first bp of the genomic sequence following the LTR sequence. Approximately 20,000 monoclonal reads corresponded to gene trap insertions in 17,000 transcripts – however this number is inflated likely 2-fold due to redundant UCSD identifiers. Nevertheless, this pilot experiment served as a proof-of-principle that gene trap insertions in human cells can be mapped after a phenotypic screen or selection screen and the published methods are repeatable.

Supplemental Figures

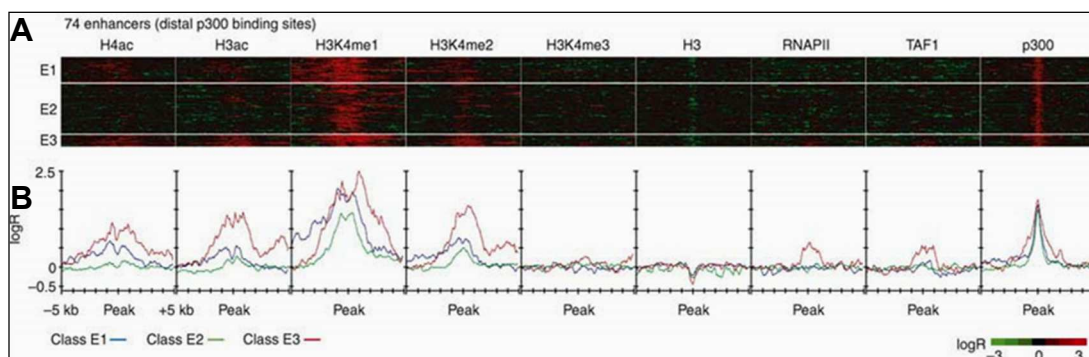


Figure 18. Histone modification features of human transcriptional enhancers. ChIP-chip was performed against six histone markers (H4ac, H3ac, H3K4me1, H3K4me2, H3K4me3, and H3) and three transcription factors (RNA pol II, TAF1, and P300) in the ENCODE regions. (a) Data shown are 10kb windows centered on promoter-distal p300 peaks at putative enhancers with computational clustering into three classes (E1, E2, and E3). (Heintzman, 2007)

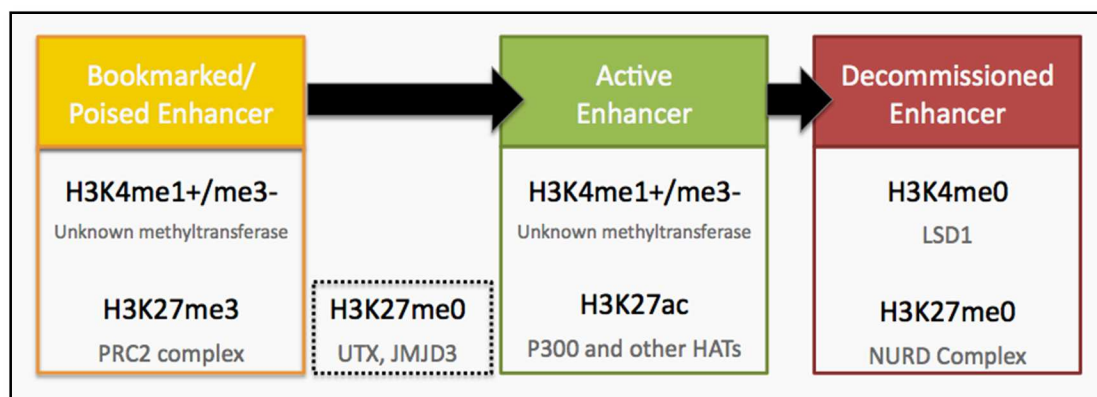


Figure 19. Chromatin states at poised, active, and decommissioned enhancers. Functional classes of enhancers can be identified by specific combinations of chromatin marks. The histone modifying enzymes experimentally determined to catalyze these marks at enhancers are listed.

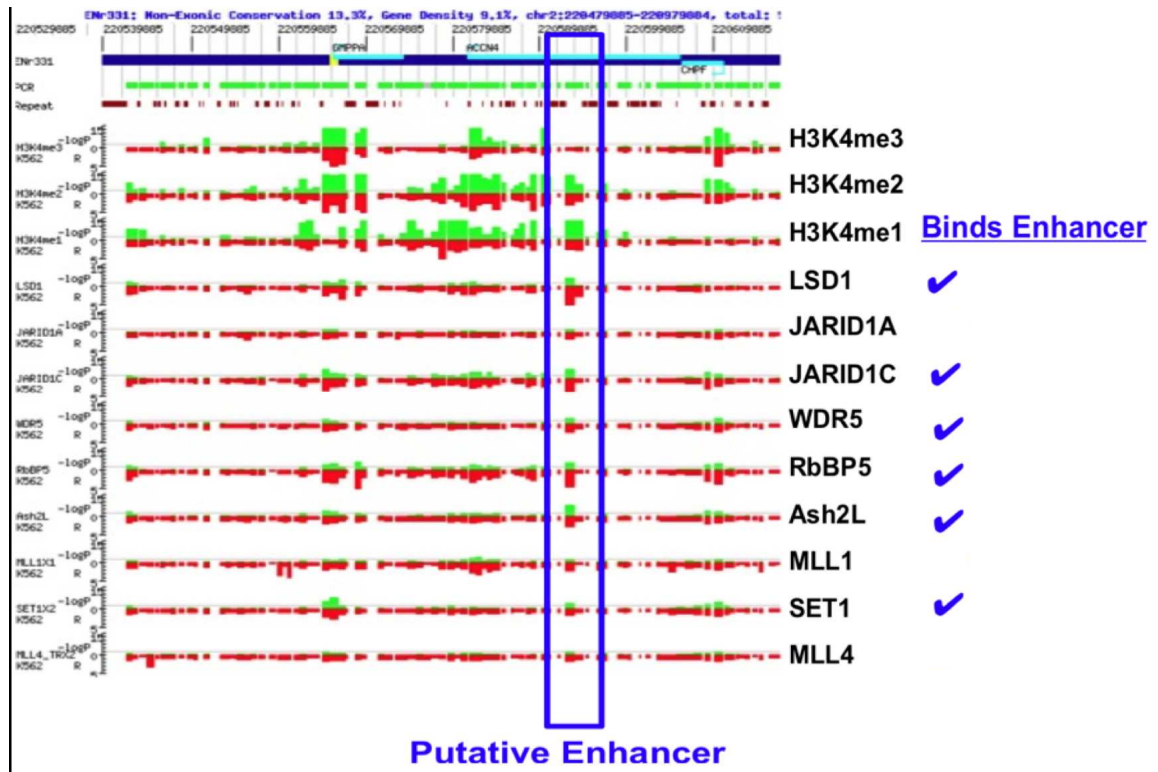


Figure 20. ChIP-chip analysis of H3K4 methyltransferase and demethylase binding patterns in K562 cells. An intergenic enhancer in the ACCN4 gene is marked by H3K4me1/2 but not me3. This enhancer also shows enrichment for LSD1, Jarid1C, WDR5, RbBP5, Ash2l, and Set1. Figure courtesy of Dr. Saurabh Agarwal.

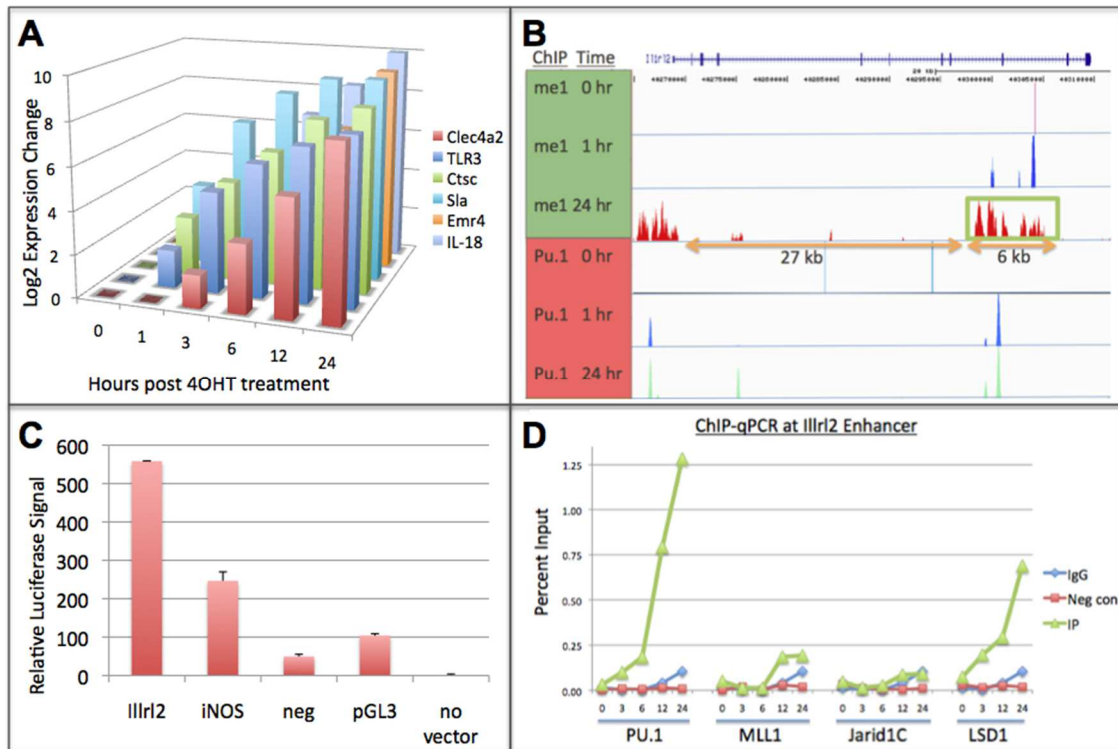


Figure 21. Analysis of histone modifying enzyme interaction with validated PU.1 target enhancer. (A) Induced RNA expression of PU.1 target genes during 4OHT treatment time course. (B) The intergenic *Illr12* enhancer has induced H3K4me1 after 24 hours of treatment. ChIP-seq data from Heinz, *et al.* (Heinz et al. 2010; Hon et al. 2013; Jin et al. 2010; Ziller et al. 2013) (C) Luciferase enhancer reporter assay was performed in the RAW 264.7 macrophage cell line. Luciferase signal was normalized to Renilla signal and the fold change over the no vector control was determined. The *iNOS* enhancer is a known macrophage enhancer (positive control), the negative control contains a random genomic sequence, and pGL3 is the luciferase vector with out a subcloned enhancer. (D) ChIP-qPCR enrichment as a percent of the input signal for PU.1, MLL1, Jarid1C, and LSD1 at indicated 4OHT time points in PU.1-ER cells. The negative control shows enrichment at a random genomic sequence devoid of K4me1 marks.

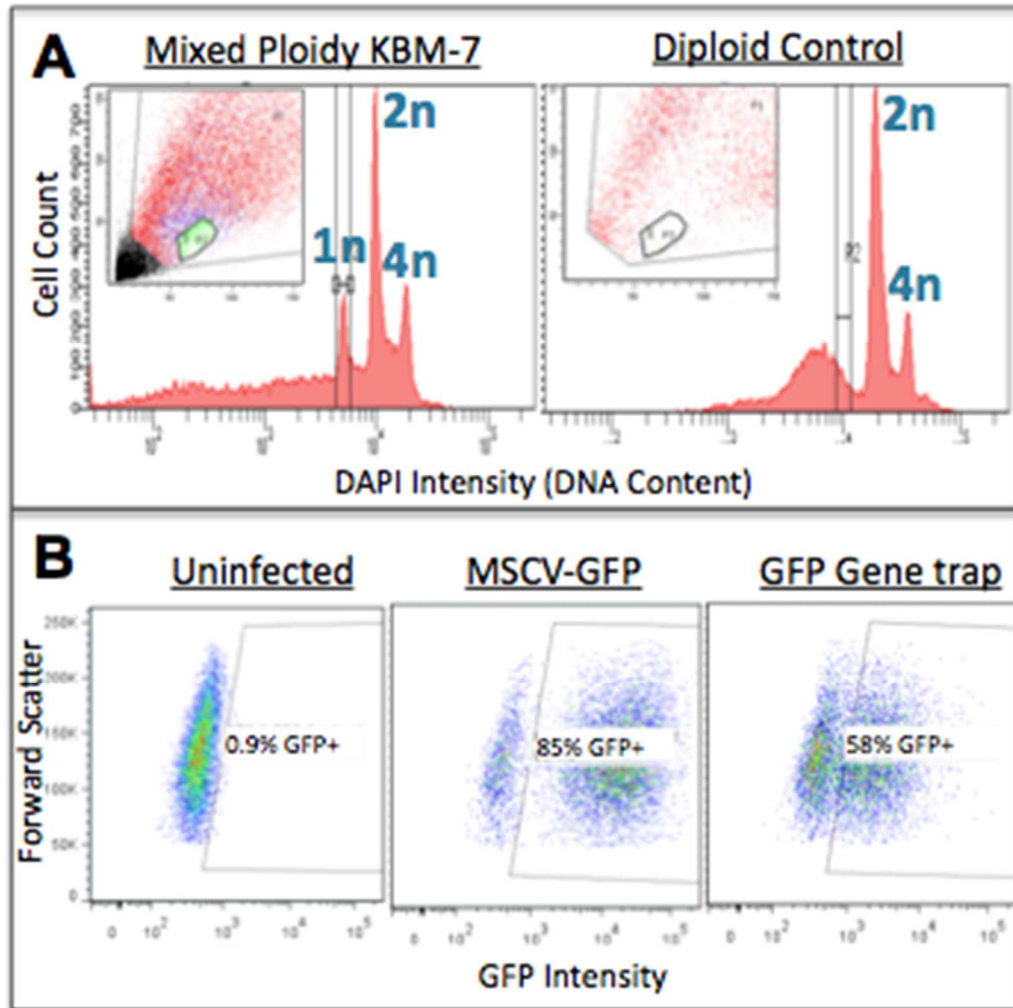


Figure 22. Generation of a human haploid cell gene trap library. (A) DAPI staining reveals the ploidy of cells. (B) Transduction rate of 293T cells with MSCV-GFP and GFP gene trap retrovirus determined by GFP.

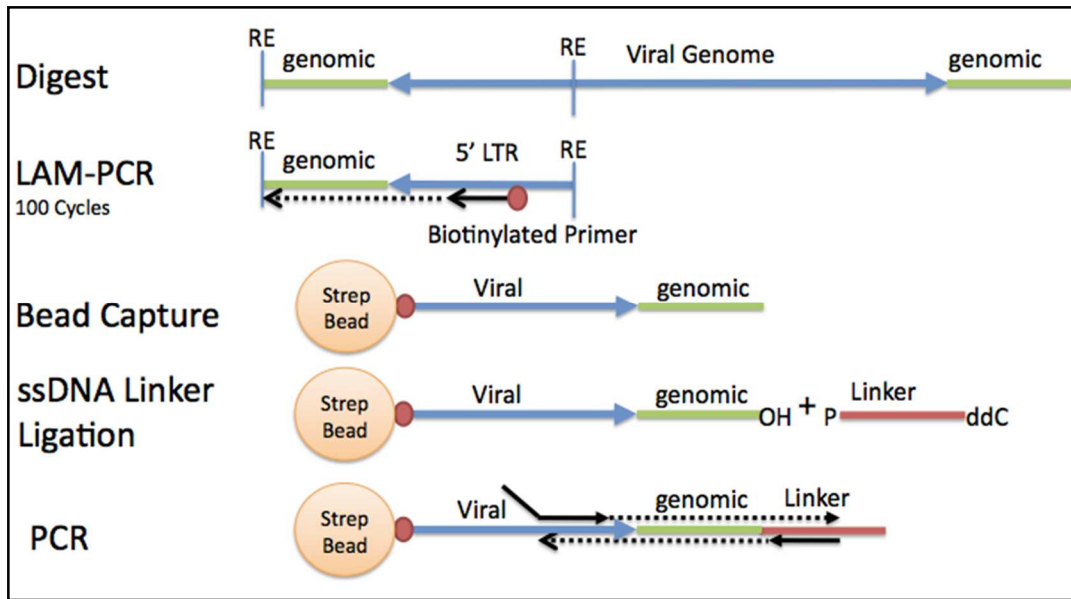


Figure 23. LM-PCR based strategy for mapping of viral insertions. Genomic DNA is purified from infected cells and digested. Linear amplification mediated PCR (LAM-PCR) uses a biotinylated primer specific for the viral LTR for 100 cycles. These single stranded products are captured by streptavidin beads and a ssDNA linker is ligated to the 5' end. A final PCR step amplifies viral-genomic junctions with Illumina sequencing adaptors.

REFERENCES

- Aday AW, Zhu LJ, Lakshmanan A, Wang J, Lawson ND (2011) Identification of cis regulatory features in the embryonic zebrafish genome through large-scale profiling of H3K4me1 and H3K4me3 binding sites. *Dev Biol* 357:450–462. doi: 10.1016/j.ydbio.2011.03.007
- Adey A, Shendure J (2012) Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. *Genome Res* 22:1139–1143. doi: 10.1101/gr.136242.111
- Adli M, Bernstein BE (2011) Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. *Nature Protocols* 6:1656–1668. doi: 10.1038/nprot.2011.402
- Akhtar W, de Jong J, Pindyurin AV, Pagie L, Meuleman W, de Ridder J, Berns A, Wessels LFA, Van Lohuizen M, van Steensel B (2013) Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell* 154:914–927. doi: 10.1016/j.cell.2013.07.018
- American Association for Cancer Research Human Epigenome Task Force, European Union, Network of Excellence, Scientific Advisory Board (2008) Moving AHEAD with an international human epigenome project. *Nature* 454:711–715. doi: 10.1038/454711a
- An S, Yeo KJ, Jeon YH, Song J-J (2011) Crystal structure of the human histone methyltransferase ASH1L catalytic domain and its implications for the regulatory mechanism. *J Biol Chem* 286:8369–8374. doi: 10.1074/jbc.M110.203380
- Arnold K, Sarkar A, Yram MA, Polo JM, Bronson R, Sengupta S, Seandel M, Geijsen N, Hochedlinger K (2011) Adult Stem and Progenitor Cells Are Important for Tissue Regeneration and Survival of Mice. *Stem Cell* 9:317–329. doi: 10.1016/j.stem.2011.09.001
- Badeaux AI, Shi Y (2013) Emerging roles for chromatin as a signal integration and storage platform. *Nat Rev Mol Cell Biol* 14:211–224. doi: 10.1038/nrm3545
- Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K (2007) High-resolution profiling of histone methylations in the human genome. *Cell* 129:823–837. doi: 10.1016/j.cell.2007.05.009
- Bauer DE, Kamran SC, Lessard S, Xu J, Fujiwara Y, Lin C, Shao Z, Canver MC, Smith EC, Pinello L, Sabo PJ, Vierstra J, Voit RA, Yuan G-C, Porteus MH, Stamatoyannopoulos JA, Lettre G, Orkin SH (2013) An erythroid enhancer of

- BCL11A subject to genetic variation determines fetal hemoglobin level. *Science* 342:253–257. doi: 10.1126/science.1242088
- Bauer DE, Kamran SC, Orkin SH (2012) Reawakening fetal hemoglobin: prospects for new therapies for the β -globin disorders. *Blood* 120:2945–2953. doi: 10.1182/blood-2012-06-292078
- Beck DB, Oda H, Shen SS, Reinberg D (2012a) PR-Set7 and H4K20me1: at the crossroads of genome integrity, cell cycle, chromosome condensation, and transcription. *Genes & Development* 26:325–337. doi: 10.1101/gad.177444.111
- Beck S, Bernstein BE, Campbell RM, Costello JF, Dhanak D, Ecker JR, Grealley JM, Issa JP, Laird PW, Polyak K, Tycko B, Jones PA, for the AACR Cancer Epigenome Task Force (2012b) A Blueprint for an International Cancer Epigenome Consortium. A Report from the AACR Cancer Epigenome Task Force. *Cancer Research* 72:6319–6324. doi: 10.1158/0008-5472.CAN-12-3658
- Bell O, Tiwari VK, Thomä NH, Schübeler D (2011) Determinants and dynamics of genome accessibility. *Nat Rev Genet* 12:554–564. doi: 10.1038/nrg3017
- Bender MA, Bulger M, Close J, Groudine M (2000) Beta-globin gene switching and DNase I sensitivity of the endogenous beta-globin locus in mice do not require the locus control region. *Mol Cell* 5:387–393.
- Bernstein BE, Meissner A, Lander ES (2007) The mammalian epigenome. *Cell* 128:669–681. doi: 10.1016/j.cell.2007.01.033
- Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* 125:315–326. doi: 10.1016/j.cell.2006.02.041
- Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, Farnham PJ, Hirst M, Lander ES, Mikkelsen TS, Thomson JA (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology* 28:1045–1048. doi: 10.1038/nbt1010-1045
- Bhandare R, Schug J, Le Lay J, Fox A, Smirnova O, Liu C, Naji A, Kaestner KH (2010) Genome-wide analysis of histone modifications in human pancreatic islets. *Genome Res* 20:428–433. doi: 10.1101/gr.102038.109
- Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth

- GP, Gunderson KL, Fan J-B, Shen R (2011) High density DNA methylation array with single CpG site resolution. *Genomics* 98:288–295. doi: 10.1016/j.ygeno.2011.07.007
- Bird A (2002) DNA methylation patterns and epigenetic memory. *Genes & Development* 16:6–21. doi: 10.1101/gad.947102
- Blackwood EM, Kadonaga JT (1998) Going the distance: a current view of enhancer action. *Science* 281:60–63.
- Bock C (2012) Analysing and interpreting DNA methylation data. *Nat Rev Genet* 13:705–719. doi: 10.1038/nrg3273
- Bonasio R, Tu S, Reinberg D (2010) Molecular signals of epigenetic states. *Science* 330:612–616. doi: 10.1126/science.1191078
- Booth MJ, Branco MR, Ficz G, Oxley D, Krueger F, Reik W, Balasubramanian S (2012) Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* 336:934–937. doi: 10.1126/science.1220671
- Boumahdi S, Driessens G, Lapouge G, Rorive S, Nassar D, Le Mercier M, Delatte B, Caauwe A, Lenglez S, Nkusi E, Brohée S, Salmon I, Dubois C, del Marmol V, Fuks F, Beck B, Blanpain C (2015) SOX2 controls tumour initiation and cancer stem-cell functions in squamous-cell carcinoma. *Nature* 511:246–250. doi: 10.1038/nature13305
- Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG, Gifford DK, Melton DA, Jaenisch R, Young RA (2005) Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells. *Cell* 122:947–956. doi: 10.1016/j.cell.2005.08.020
- Boyer LA, Plath K, Zeitlinger J, Brambrink T, Medeiros LA, Lee TI, Levine SS, Wernig M, Tajonar A, Ray MK, Bell GW, Otte AP, Vidal M, Gifford DK, Young RA, Jaenisch R (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* 441:349–353. doi: 10.1038/nature04733
- Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE (2008) High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell* 132:311–322. doi: 10.1016/j.cell.2007.12.014
- Brinkman AB, Gu H, Bartels SJJ, Zhang Y, Matarese F, Simmer F, Marks H, Bock C, Gnirke A, Meissner A, Stunnenberg HG (2012) Sequential ChIP-bisulfite

- sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. *Genome Res* 22:1128–1138. doi: 10.1101/gr.133728.111
- Brogaard K, Xi L, Wang J-P, Widom J (2012) A map of nucleosome positions in yeast at base-pair resolution. *Nature*. doi: 10.1038/nature11142
- Brown JM, Green J, Neves das RP, Wallace HAC, Smith AJH, Hughes J, Gray N, Taylor S, Wood WG, Higgs DR, Iborra FJ, Buckle VJ (2008) Association between active genes occurs at nuclear speckles and is modulated by chromatin environment. *J Cell Biol* 182:1083–1097. doi: 10.1083/jcb.200803174
- Buecker C, Wysocka J (2012) Enhancers as information integration hubs in development: lessons from genomics. *Trends Genet* 28:276–284. doi: 10.1016/j.tig.2012.02.008
- Bulger M, Groudine M (2011) Functional and Mechanistic Diversity of Distal Transcription Enhancers. *Cell* 144:327–339. doi: 10.1016/j.cell.2011.01.024
- Calo E, Wysocka J (2013) Modification of Enhancer Chromatin: What, How, and Why? *Mol Cell* 49:825–837. doi: 10.1016/j.molcel.2013.01.038
- Canver MC, Bauer DE, Dass A, Yien YY, Chung J, Masuda T, Maeda T, Paw BH, Orkin SH (2014) Characterization of genomic deletion efficiency mediated by clustered regularly interspaced palindromic repeats (CRISPR)/Cas9 nuclease system in mammalian cells. *J Biol Chem* 289:21312–21324. doi: 10.1074/jbc.M114.564625
- Carette JE, Guimaraes CP, Varadarajan M, Park AS, Wuethrich I, Godarova A, Kotecki M, Cochran BH, Spooner E, Ploegh HL, Brummelkamp TR (2009) Haploid genetic screens in human cells identify host factors used by pathogens. *Science* 326:1231–1235. doi: 10.1126/science.1178955
- Carette JE, Guimaraes CP, Wuethrich I, Blomen VA, Varadarajan M, Sun C, Bell G, Yuan B, Muellner MK, Nijman SM, Ploegh HL, Brummelkamp TR (2011) Global gene disruption in human cells to assign genes to phenotypes by deep sequencing. *Nature Biotechnology* 29:542–546. doi: 10.1038/nbt.1857
- Chambers I, Tomlinson SR (2009) The transcriptional foundation of pluripotency. *Development* 136:2311–2322. doi: 10.1242/dev.024398
- Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh Y-H, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung W-K, Clarke ND, Wei C-L, Ng H-H (2008) Integration

- of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. *Cell* 133:1106–1117. doi: 10.1016/j.cell.2008.04.043
- Chong MMW, Simpson N, Ciofani M, Chen G, Collins A, Littman DR (2010) Epigenetic propagation of CD4 expression is established by the Cd4 proximal enhancer in helper T cells. *Genes & Development* 24:659–669. doi: 10.1101/gad.1901610
- Clark SJ, Harrison J, Paul CL, Frommer M (1994) High sensitivity mapping of methylated cytosines. *Nucleic Acids Res* 22:2990–2997.
- Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotech* 4:265–270. doi: 10.1038/nnano.2009.12
- Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* 452:215–219. doi: 10.1038/nature06745
- Collins RE, Tachibana M, Tamaru H, Smith KM, Jia D, Zhang X, Selker EU, Shinkai Y, Cheng X (2005) In vitro and in vivo analyses of a Phe/Tyr switch controlling product specificity of histone lysine methyltransferases. *J Biol Chem* 280:5563–5570. doi: 10.1074/jbc.M410483200
- Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, Zhang F (2013) Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* 339:819–823. doi: 10.1126/science.1231143
- Consortium TEP, Consortium TEP, data analysis coordination OC, data production DPL, data analysis LA, group W, scientific management NPM, steering committee PI, Boise State University and University of North Carolina at Chapel Hill Proteomics groups (data production and analysis), Broad Institute Group (data production and analysis), Cold Spring Harbor, University of Geneva, Center for Genomic Regulation, Barcelona, RIKEN, Sanger Institute, University of Lausanne, Genome Institute of Singapore group (data production and analysis), Data coordination center at UC Santa Cruz (production data coordination), Duke University, EBI, University of Texas, Austin, University of North Carolina-Chapel Hill group (data production and analysis), Genome Institute of Singapore group (data production and analysis), HudsonAlpha Institute, Caltech, UC Irvine, Stanford group (data production and analysis), targeted experimental validation LBNLG, data production and analysis NG, Sanger Institute, Washington University, Yale University, Center for Genomic Regulation, Barcelona, UCSC, MIT, University

- of Lausanne, CNIO group (data production and analysis), Stanford-Yale, Harvard, University of Massachusetts Medical School, University of Southern California/UC Davis group (data production and analysis), University of Albany SUNY group (data production and analysis), University of Chicago, Stanford group (data production and analysis), University of Heidelberg group (targeted experimental validation), University of Massachusetts Medical School Bioinformatics group (data production and analysis), University of Massachusetts Medical School Genome Folding group (data production and analysis), University of Washington, University of Massachusetts Medical Center group (data production and analysis), Data Analysis Center (data analysis) (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 488:57–74. doi: 10.1038/nature11247
- Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, Jaenisch R (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci USA* 107:21931–21936. doi: 10.1073/pnas.1016071107/-/DCSupplemental/st01.xlsx
- Crick F (1970) Central dogma of molecular biology. *Nature* 227:561–563.
- de Wit E, de Laat W (2012) A decade of 3C technologies: insights into nuclear organization. *Genes & Development* 26:11–24. doi: 10.1101/gad.179804.111
- Degner JF, Pai AA, Pique-Regi R, Veyrieras J-B, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE, Stephens M, Gilad Y, Pritchard JK (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482:390–394. doi: 10.1038/nature10808
- Dekker J (2002) Capturing Chromosome Conformation. *Science* 295:1306–1311. doi: 10.1126/science.1067799
- Dekker J, Marti-Renom MA, Mirny LA (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet* 14:390–403. doi: 10.1038/nrg3454
- DeKoter RP, Singh H (2000) Regulation of B lymphocyte and macrophage development by graded expression of PU.1. *Science* 288:1439–1441.
- Demare LE, Leng J, Cotney J, Reilly SK, Yin J, Sarro R, Noonan JP (2013) The genomic landscape of cohesin-associated chromatin interactions. *Genome Res* 23:1224–1234. doi: 10.1101/gr.156570.113
- Deng W, Lee J, Wang H, Miller J, Reik A, Gregory PD, Dean A, Blobel GA (2012)

- Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* 149:1233–1244. doi: 10.1016/j.cell.2012.03.051
- Diep D, Plongthongkum N, Gore A, Fung H-L, Shoemaker R, Zhang K (2012) Library-free methylation sequencing with bisulfite padlock probes. *Nat Meth* 9:270–272. doi: 10.1038/nmeth.1871
- Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485:376–380. doi: 10.1038/nature11082
- Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C, Green RD, Dekker J (2006) Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Res* 16:1299–1309. doi: 10.1101/gr.5571506
- Dou Y, Milne TA, Ruthenburg AJ, Lee S, Lee JW, Verdine GL, Allis CD, Roeder RG (2006) Regulation of MLL1 H3K4 methyltransferase activity by its core components. *Nat Struct Mol Biol* 13:713–719. doi: 10.1038/nsmb1128
- Downen JM, Fan ZP, Hnisz D, Ren G, Abraham BJ, Zhang LN, Weintraub AS, Schuijers J, Lee TI, Zhao K, Young RA (2014) Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* 159:374–387. doi: 10.1016/j.cell.2014.09.030
- Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Gräf S, Johnson N, Herrero J, Tomazou EM, Thorne NP, Bäckdahl L, Herberth M, Howe KL, Jackson DK, Miretti MM, Marioni JC, Birney E, Hubbard TJP, Durbin R, Tavaré S, Beck S (2008) A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature Biotechnology* 26:779–785. doi: 10.1038/nbt1414
- Egelhofer TA, Minoda A, Klugman S, Lee K, Kolasinska-Zwierz P, Alekseyenko AA, Cheung M-S, Day DS, Gadel S, Gorchakov AA, Gu T, Kharchenko PV, Kuan S, Latorre I, Linder-Basso D, Luu Y, Ngo Q, Perry M, Rechtsteiner A, Riddle NC, Schwartz YB, Shanower GA, Vielle A, Ahringer J, Elgin SCR, Kuroda MI, Pirrotta V, Ren B, Strome S, Park PJ, Karpen GH, Hawkins RD, Lieb JD (2011) An assessment of histone-modification antibody quality. *Nat Struct Mol Biol* 18:91–93. doi: 10.1038/nsmb.1972
- Elling U, Taubenschmid J, Wirnsberger G, O'Malley R, Demers S-P, Vanhaelen Q, Shukalyuk AI, Schmauss G, Schramek D, Schnuetgen F, Melchner von H, Ecker

- JR, Stanford WL, Zuber J, Stark A, Penninger JM (2011) Forward and reverse genetics through derivation of haploid mouse embryonic stem cells. *Cell Stem Cell* 9:563–574. doi: 10.1016/j.stem.2011.10.012
- Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473:43–49. doi: 10.1038/nature09906
- Feinberg AP (2007) Phenotypic plasticity and the epigenetics of human disease. *Nature* 447:433–440. doi: 10.1038/nature05919
- Felsenfeld G, Groudine M (2003) Controlling the double helix. *Nature* 421:448–453. doi: 10.1038/nature01411
- Ficz G, Branco MR, Seisenberger S, Santos F, Krueger F, Hore TA, Marques CJ, Andrews S, Reik W (2011) Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* 473:398–402. doi: 10.1038/nature10008
- Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Meth* 7:461–465. doi: 10.1038/nmeth.1459
- Flynn RA, Chang HY (2012) Active chromatin and noncoding RNAs: an intimate relationship. *Curr Opin Genet Dev* 22:172–178. doi: 10.1016/j.gde.2011.11.002
- Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. *Nat Rev Genet* 10:241–251. doi: 10.1038/nrg2554
- Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, Chew EGY, Huang PYH, Welboren W-J, Han Y, Ooi HS, Ariyaratne PN, Vega VB, Luo Y, Tan PY, Choy PY, Wansa KDSA, Zhao B, Lim KS, Leow SC, Yow JS, Joseph R, Li H, Desai KV, Thomsen JS, Lee YK, Karuturi RKM, Herve T, Bourque G, Stunnenberg HG, Ruan X, Cacheux-Rataboul V, Sung W-K, Liu ET, Wei C-L, Cheung E, Ruan Y (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* 462:58–64. doi: 10.1038/nature08497
- Gaffney DJ, McVicker G, Pai AA, Fondufe-Mittendorf YN, Lewellen N, Michelini K, Widom J, Gilad Y, Pritchard JK (2012) Controls of nucleosome positioning in the human genome. *PLoS Genet* 8:e1003036. doi:

10.1371/journal.pgen.1003036

- Gaj T, Gersbach CA, Barbas CF III (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in Biotechnology* 31:397–405. doi: 10.1016/j.tibtech.2013.04.004
- Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, Min R, Alves P, Abyzov A, Addleman N, Bhardwaj N, Boyle AP, Cayting P, Charos A, Chen DZ, Cheng Y, Clarke D, Eastman C, Euskirchen G, Frietze S, Fu Y, Gertz J, Grubert F, Harmanci A, Jain P, Kasowski M, Lacroute P, Leng J, Lian J, Monahan H, O'Geen H, Ouyang Z, Partridge EC, Patacsil D, Pauli F, Raha D, Ramirez L, Reddy TE, Reed B, Shi M, Slifer T, Wang J, Wu L, Yang X, Yip KY, Zilberman-Schapira G, Batzoglou S, Sidow A, Farnham PJ, Myers RM, Weissman SM, Snyder M (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature* 489:91–100. doi: 10.1038/nature11245
- Gifford CA, Ziller MJ, Gu H, Trapnell C, Donaghey J, Tsankov A, Shalek AK, Kelley DR, Shishkin AA, Issner R, Zhang X, Coyne M, Fostel JL, Holmes L, Meldrim J, Guttman M, Epstein C, Park H, Kohlbacher O, Rinn J, Gnirke A, Lander ES, Bernstein BE, Meissner A (2013) Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. *Cell* 153:1149–1163. doi: 10.1016/j.cell.2013.04.037
- Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* 17:877–885. doi: 10.1101/gr.5533506
- Glaser S, Schaft J, Lubitz S, Vintersten K, van der Hoeven F, Tufteland KR, Aasland R, Anastassiadis K, Ang S-L, Stewart AF (2006) Multiple epigenetic maintenance factors implicated by the loss of Mll2 in mouse development. *Development* 133:1423–1432. doi: 10.1242/dev.02302
- Gosselin D, Link VM, Romanoski CE, Fonseca GJ, Eichenfield DZ, Spann NJ, Stender JD, Chun HB, Garner H, Geissmann F, Glass CK (2014) Environment drives selection and function of enhancers controlling tissue-specific macrophage identities. *Cell* 159:1327–1340. doi: 10.1016/j.cell.2014.11.023
- Gribnau J, Hochedlinger K, Hata K, Li E, Jaenisch R (2003) Asynchronous replication timing of imprinted loci is independent of DNA methylation, but consistent with differential subnuclear localization. *Genes & Development* 17:759–773. doi: 10.1101/gad.1059603

- Gröschel S, Sanders MA, Hoogenboezem R, de Wit E, Bouwman BAM, Erpelinck C, van der Velden VHJ, Havermans M, Avellino R, van Lom K, Rombouts EJ, van Duin M, Döhner K, Beverloo HB, Bradner JE, Döhner H, Löwenberg B, Valk PJM, Bindels EMJ, de Laat W, Delwel R (2014) A Single Oncogenic Enhancer Rearrangement Causes Concomitant EVI1 and GATA2 Deregulation in Leukemia. *Cell* 157:369–381. doi: 10.1016/j.cell.2014.02.019
- Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, de Klein A, Wessels L, de Laat W, van Steensel B (2008) Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* 453:948–951. doi: 10.1038/nature06947
- Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458:223–227. doi: 10.1038/nature07672
- Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CWH, Ye C, Ping JLH, Mulawadi F, Wong E, Sheng J, Zhang Y, Poh T, Chan CS, Kunarso G, Shahab A, Bourque G, Cacheux-Rataboul V, Sung W-K, Ruan Y, Wei C-L (2011) CTCF-mediated functional chromatin interactome in pluripotent cells. *Nature Publishing Group* 43:630–638. doi: 10.1038/ng.857
- Hawkins RD, Hon GC, Lee LK, Ngo Q, Lister R, Pelizzola M, Edsall LE, Kuan S, Luu Y, Klugman S, Antosiewicz-Bourget J, Ye Z, Espinoza C, Agarwahl S, Shen L, Ruotti V, Wang W, Stewart R, Thomson JA, Ecker JR, Ren B (2010) Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* 6:479–491. doi: 10.1016/j.stem.2010.03.018
- Heintzman ND, Ren B (2009) Finding distal regulatory elements in the human genome. *Curr Opin Genet Dev* 19:541–549. doi: 10.1016/j.gde.2009.09.006
- Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* 39:311–318. doi: 10.1038/ng1966
- Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 38:576–589. doi: 10.1016/j.molcel.2010.05.004

- Heinz S, Romanoski CE, Benner C, Allison KA, Kaikkonen MU, Orozco LD, Glass CK (2013) Effect of natural genetic variation on enhancer selection and function. *Nature* 503:487–492. doi: 10.1038/nature12615
- Hellman A, Chess A (2007) Gene body-specific methylation on the active X chromosome. *Science* 315:1141–1143. doi: 10.1126/science.1136352
- Henikoff S, Strahl BD, Warburton PE (2008) Epigenomics: a roadmap to chromatin. *Science* 322:853. doi: 10.1126/science.322.5903.853a
- Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, Hoke HA, Young RA (2013) Super-Enhancers in the Control of Cell Identity and Disease. *Cell* 155:934–947. doi: 10.1016/j.cell.2013.09.053
- Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E, Hardison RC, Dunham I, Kellis M, Noble WS (2013) Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* 41:827–841. doi: 10.1093/nar/gks1284
- Holliday R (1990) Mechanisms for the control of gene activity during development. *Biol Rev Camb Philos Soc* 65:431–471.
- Hon GC, Rajagopal N, Shen Y, McCleary DF, Yue F, Dang MD, Ren B (2013) Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nature Publishing Group*. doi: 10.1038/ng.2746
- Huang Y, Pastor WA, Shen Y, Tahiliani M, Liu DR, Rao A (2010) The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS ONE* 5:e8888. doi: 10.1371/journal.pone.0008888
- Imhof A, Yang XJ, Ogryzko VV, Nakatani Y, Wolffe AP, Ge H (1997) Acetylation of general transcription factors by histone acetyltransferases. *Curr Biol* 7:689–692.
- Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA, He C, Zhang Y (2011) Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* 333:1300–1303. doi: 10.1126/science.1210597
- Jeong KW, Kim K, Situ AJ, Ulmer TS, An W, Stallcup MR (2011) Recognition of enhancer element-specific histone methylation by TIP60 in transcriptional activation. *Nat Struct Mol Biol* 18:1358–1365. doi: 10.1038/nsmb.2153
- Jiang C, Pugh BF (2009) Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet* 10:161–172. doi: 10.1038/nrg2522

- Jin C, Zang C, Wei G, Cui K, Peng W, Zhao K, Felsenfeld G (2009) H3.3/H2A.Z double variant[ndash]containing nucleosomes mark “nucleosome-free regions” of active promoters and other regulatory regions. *Nature Publishing Group* 41:941–945. doi: doi:10.1038/ng.409
- Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen C-A, Schmitt AD, Espinoza CA, Ren B (2014a) A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* 503:290–294. doi: 10.1038/nature12644
- Jin F, Li Y, Ren B, Natarajan R (2014b) Enhancers. *Transcription* 2:226–230. doi: 10.4161/trns.2.5.17712
- Jin F, Li Y, Ren B, Natarajan R (2011) PU.1 and C/EBP(alpha) synergistically program distinct response to NF-kappaB activation through establishing monocyte specific enhancers. *Proc Natl Acad Sci USA* 108:5290–5295. doi: 10.1073/pnas.1017214108
- Jin S-G, Kadam S, Pfeifer GP (2010) Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. *Nucleic Acids Res* 38:e125. doi: 10.1093/nar/gkq223
- John S, Sabo PJ, Thurman RE, Sung M-H, Biddie SC, Johnson TA, Hager GL, Stamatoyannopoulos JA (2011) Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Publishing Group* 43:264–268. doi: 10.1038/ng.759
- Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316:1497–1502. doi: 10.1126/science.1141319
- Jones PA, Martienssen R (2005) A blueprint for a Human Epigenome Project: the AACR Human Epigenome Workshop. In: *Cancer Res.* pp 11241–11246
- Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS, Taatjes DJ, Dekker J, Young RA (2010) Mediator and cohesin connect gene expression and chromatin architecture. *Nature* 467:430–435. doi: 10.1038/nature09380
- Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature Biotechnology* 30:90–98. doi: 10.1038/nbt.2057
- Kearns NA, Pham H, Tabak B, Genga RM, Silverstein NJ, Garber M, Maehr R (2015) Functional annotation of native enhancers with a Cas9-histone demethylase

- fusion. *Nat Meth* 12:401–403. doi: 10.1038/nmeth.3325
- Kelly TK, Liu Y, Lay FD, Liang G, Berman BP, Jones PA (2012) Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res* 22:2497–2506. doi: 10.1101/gr.143008.112
- Kieffer-Kwon K-R, Tang Z, Mathe E, Qian J, Sung M-H, Li G, Resch W, Baek S, Pruett N, Grøntved L, Vian L, Nelson S, Zare H, Hakim O, Reyon D, Yamane A, Nakahashi H, Kovalchuk AL, Zou J, Joung JK, Sartorelli V, Wei C-L, Ruan X, Hager GL, Ruan Y, Casellas R (2013) Interactome Maps of Mouse Gene Regulatory Domains Reveal Basic Principles of Transcriptional Regulation. *Cell* 155:1507–1520. doi: 10.1016/j.cell.2013.11.039
- Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, Markenscoff-Papadimitriou E, Kuhl D, Bito H, Worley PF, Kreiman G, Greenberg ME (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature* 465:182–187. doi: 10.1038/nature09033
- Kouzarides T (2007) Chromatin modifications and their function. *Cell* 128:693–705. doi: 10.1016/j.cell.2007.02.005
- Kriaucionis S, Heintz N (2009) The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* 324:929–930. doi: 10.1126/science.1169786
- Krivega I, Dean A (2012) Enhancer and promoter interactions-long distance calls. *Curr Opin Genet Dev* 22:79–85. doi: 10.1016/j.gde.2011.11.001
- Krueger F, Kreck B, Franke A, Andrews SR (2012) DNA methylome analysis using short bisulfite sequencing data. *Nat Meth* 9:145–151. doi: 10.1038/nmeth.1828
- Laird PW (2010) Principles and challenges of genomewide DNA methylation analysis. *Nat Rev Genet* 11:191–203. doi: 10.1038/nrg2732
- Lander ES (2011) Initial impact of the sequencing of the human genome. *Nature* 470:187–197. doi: 10.1038/nature09792
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N,

Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissole SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, la Bastide de M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglu S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kasprzyk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, Szustakowski J, International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921. doi: 10.1038/35057062

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglu S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, DeSalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shores N, Sidow A, Slattey M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, Snyder M (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE

- consortia. *Genome Res* 22:1813–1831. doi: 10.1101/gr.136184.111
- Lee ER, Murdoch FE, Fritsch MK (2007) High histone acetylation and decreased polycomb repressive complex 2 member levels regulate gene specific transcriptional changes during early embryonic stem cell differentiation induced by retinoic acid. *Stem Cells* 25:2191–2199. doi: 10.1634/stemcells.2007-0203
- Lee J-H, Skalnik DG (2008) Wdr82 is a C-terminal domain-binding protein that recruits the Setd1A Histone H3-Lys4 methyltransferase complex to transcription start sites of transcribed human genes. *Molecular and Cellular Biology* 28:609–618. doi: 10.1128/MCB.01356-07
- Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, Kumar RM, Chevalier B, Johnstone SE, Cole MF, Isono K-I, Koseki H, Fuchikami T, Abe K, Murray HL, Zucker JP, Yuan B, Bell GW, Herbolsheimer E, Hannett NM, Sun K, Odom DT, Otte AP, Volkert TL, Bartel DP, Melton DA, Gifford DK, Jaenisch R, Young RA (2006) Control of Developmental Regulators by Polycomb in Human Embryonic Stem Cells. *Cell* 125:301–313. doi: 10.1016/j.cell.2006.02.043
- Lee TI, Young RA (2013) Transcriptional regulation and its misregulation in disease. *Cell* 152:1237–1251. doi: 10.1016/j.cell.2013.02.014
- Lettice LA, Heaney SJH, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* 12:1725–1735.
- Levine M, Cattoglio C, Tjian R (2014) Looping Back to Leap Forward: Transcription Enters a New Era. *Cell* 157:13–25. doi: 10.1016/j.cell.2014.02.009
- Li B, Carey M, Workman JL (2007) The role of chromatin during transcription. *Cell* 128:707–719. doi: 10.1016/j.cell.2007.01.015
- Li E, Bestor TH, Jaenisch R (1992) Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* 69:915–926.
- Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J, Sim HS, Peh SQ, Mulawadi FH, Ong CT, Orlov YL, Hong S, Zhang Z, Landt S, Raha D, Euskirchen G, Wei C-L, Ge W, Wang H, Davis C, Fisher-Aylor KI, Mortazavi A, Gerstein M, Gingeras T, Wold B, Sun Y, Fullwood MJ, Cheung E, Liu E, Sung W-K, Snyder M, Ruan Y (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* 148:84–98. doi: 10.1016/j.cell.2011.12.014

- Li Y, Rivera CM, Ishii H, Jin F, Selvaraj S, Lee AY, Dixon JR, Ren B (2014) CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells. *PLoS ONE* 9:e114485. doi: 10.1371/journal.pone.0114485
- Liber D, Domaschewicz R, Holmqvist P-H, Mazzarella L, Georgiou A, Leleu M, Fisher AG, Labosky PA, Dillon N (2010) Epigenetic Priming of a Pre-B Cell-Specific Enhancer through Binding of Sox2 and Foxd3 at the ESC Stage. *Stem Cell* 7:114–126. doi: 10.1016/j.stem.2010.05.020
- Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J (2009) Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* 326:289–293. doi: 10.1126/science.1181369
- Lin C, Garruss AS, Luo Z, Guo F, Shilatifard A (2013) The RNA Pol II elongation factor Ell3 marks enhancers in ES cells and primes future gene activation. *Cell* 152:144–156. doi: 10.1016/j.cell.2012.12.015
- Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell* 133:523–536. doi: 10.1016/j.cell.2008.03.029
- Lister R, Pelizzola M, Downen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q-M, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462:315–322. doi: 10.1038/nature08514
- Liu W, Ma Q, Wong K, Li W, Ohgi K, Zhang J, Aggarwal AK, Rosenfeld MG (2013) Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release. *Cell* 155:1581–1595. doi: 10.1016/j.cell.2013.10.056
- Liu Z, Scannell DR, Eisen MB, Tjian R (2011) Control of embryonic stem cell lineage commitment by core promoter factor, TAF3. *Cell* 146:720–731. doi: 10.1016/j.cell.2011.08.005
- Lomvardas S, Barnea G, Pisapia DJ, Mendelsohn M, Kirkland J, Axel R (2006) Interchromosomal interactions and olfactory receptor choice. *Cell* 126:403–413. doi: 10.1016/j.cell.2006.06.035
- Ma Z, Swigut T, Valouev A, Rada-Iglesias A, Wysocka J (2011) Sequence-specific

- regulator Prdm14 safeguards mouse ESCs from entering extraembryonic endoderm fates. *Nat Struct Mol Biol* 18:120–127. doi: 10.1038/nsmb.2000
- Madhani HD, Francis NJ, Kingston RE, Kornberg RD, Moazed D, Narlikar GJ, Panning B, Struhl K (2008) Epigenomics: a roadmap, but to where? *Science* 322:43–44. doi: 10.1126/science.322.5898.43b
- Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM (2013) RNA-guided human genome engineering via Cas9. *Science* 339:823–826. doi: 10.1126/science.1232033
- Mathieu N, Hempel WM, Spicuglia S, Verthuy C, Ferrier P (2000) Chromatin remodeling by the T cell receptor (TCR)-beta gene enhancer during early T cell development: Implications for the control of TCR-beta locus recombination. *J Exp Med* 192:625–636.
- Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, Lee K, Kuttyavin T, Stehling-Sun S, Johnson AK, Canfield TK, Giste E, Diegel M, Bates D, Hansen RS, Neph S, Sabo PJ, Heimfeld S, Raubitschek A, Ziegler S, Cotsapas C, Sotoodehnia N, Glass I, Sunyaev SR, Kaul R, Stamatoyannopoulos JA (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337:1190–1195. doi: 10.1126/science.1222794
- Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, Qi J, Schuster SC, Albert I, Pugh BF (2008) A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res* 18:1073–1083. doi: 10.1101/gr.078261.108
- Meissner A (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* 33:5868–5877. doi: 10.1093/nar/gki901
- Melgar MF, Collins FS, Sethupathy P (2011) Discovery of active enhancers through bidirectional expression of short transcripts. *Genome Biol* 12:R113. doi: 10.1186/gb-2011-12-11-r113
- Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG, Kinney JB, Kellis M, Lander ES, Mikkelsen TS (2012) systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology* 1–9. doi: 10.1038/nbt.2137
- Mendenhall EM, Williamson KE, Reyon D, Zou JY, Ram O, Joung JK, Bernstein BE (2013) Locus-specific editing of histone modifications at endogenous

enhancers. *Nature Biotechnology* 31:1133–1136. doi: 10.1038/nbt.2701

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim T-K, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448:553–560. doi: 10.1038/nature06008

Mikkelsen TS, Xu Z, Zhang X, Wang L, Gimble JM, Lander ES, Rosen ED (2010) Comparative Epigenomic Analysis of Murine and Human Adipogenesis. *Cell* 143:156–169. doi: 10.1016/j.cell.2010.09.006

Mito Y, Henikoff JG, Henikoff S (2007) Histone replacement marks the boundaries of cis-regulatory domains. *Science* 315:1408–1411. doi: 10.1126/science.1134004

modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Nègre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, Washietl S, Arshinoff BI, Ay F, Meyer PE, Robine N, Washington NL, Di Stefano L, Berezhikov E, Brown CD, Candeias R, Carlson JW, Carr A, Jungreis I, Marbach D, Sealfon R, Tolstorukov MY, Will S, Alekseyenko AA, Artieri C, Booth BW, Brooks AN, Dai Q, Davis CA, Duff MO, Feng X, Gorchakov AA, Gu T, Henikoff JG, Kapranov P, Li R, MacAlpine HK, Malone J, Minoda A, Nordman J, Okamura K, Perry M, Powell SK, Riddle NC, Sakai A, Samsonova A, Sandler JE, Schwartz YB, Sher N, Spokony R, Sturgill D, van Baren M, Wan KH, Yang L, Yu C, Feingold E, Good P, Guyer M, Lowdon R, Ahmad K, Andrews J, Berger B, Brenner SE, Brent MR, Cherbas L, Elgin SCR, Gingeras TR, Grossman R, Hoskins RA, Kaufman TC, Kent W, Kuroda MI, Orr-Weaver T, Perrimon N, Pirrotta V, Posakony JW, Ren B, Russell S, Cherbas P, Graveley BR, Lewis S, Mickle G, Oliver B, Park PJ, Celniker SE, Henikoff S, Karpen GH, Lai EC, MacAlpine DM, Stein LD, White KP, Kellis M (2010) Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* 330:1787–1797. doi: 10.1126/science.1198374

Mouse ENCODE Consortium, Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, Groudine M, Bender M, Kaul R, Canfield T, Giste E, Johnson A, Zhang M, Balasundaram G, Byron R, Roach V, Sabo PJ, Sandstrom R, Stehling AS, Thurman RE, Weissman SM, Cayting P, Hariharan M, Lian J, Cheng Y, Landt SG, Ma Z, Wold BJ, Dekker J, Crawford GE, Keller CA, Wu W, Morrissey C, Kumar SA, Mishra T, Jain D, Byrsk-Bishop M, Blankenberg D, Lajoie BR, Jain G, Sanyal A, Chen K-B, Denas O, Taylor J, Blobel GA, Weiss MJ, Pimkin M, Deng W, Marinov GK, Williams BA, Fisher-Aylor KI, DeSalvo G, Kiralusha A, Trout D, Amrhein H, Mortazavi A, Edsall L, McCleary D, Kuan S,

- Shen Y, Yue F, Ye Z, Davis CA, Zaleski C, Jha S, Xue C, Dobin A, Lin W, Fastuca M, Wang H, Guigo R, Djebali S, Lagarde J, Ryba T, Sasaki T, Malladi VS, Cline MS, Kirkup VM, Learned K, Rosenbloom KR, Kent WJ, Feingold EA, Good PJ, Pazin M, Lowdon RF, Adams LB (2012) An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol* 13:418. doi: 10.1186/gb-2012-13-8-418
- Mullen AC, Orlando DA, Newman JJ, Lovén J, Kumar RM, Bilodeau S, Reddy J, Guenther MG, DeKoter RP, Young RA (2011) Master Transcription Factors Determine Cell-Type-Specific Responses to TGF- β Signaling. *Cell* 147:565–576. doi: 10.1016/j.cell.2011.08.050
- Murphy PJ, Cipriany BR, Wallin CB, Ju CY, Szeto K, Hagarman JA, Benitez JJ, Craighead HG, Soloway PD (2013) Single-molecule analysis of combinatorial epigenomic states in normal and tumor cells. *Proc Natl Acad Sci USA* 110:7772–7777. doi: 10.1073/pnas.1218495110
- Nakamura T, Mori T, Tada S, Krajewski W, Rozovskaia T, Wassell R, Dubois G, Mazo A, Croce CM, Canaani E (2002) ALL-1 is a histone methyltransferase that assembles a supercomplex of proteins involved in transcriptional regulation. *Mol Cell* 10:1119–1128.
- Narlikar GJ, Sundaramoorthy R, Owen-Hughes T (2013) Mechanisms and Functions of ATP-Dependent Chromatin-Remodeling Enzymes. *Cell* 154:490–503. doi: 10.1016/j.cell.2013.07.011
- Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyannopoulos JA (2012a) Circuitry and dynamics of human transcription factor regulatory networks. *Cell* 150:1274–1286. doi: 10.1016/j.cell.2012.04.040
- Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, Maurano MT, Humbert R, Rynes E, Wang H, Vong S, Lee K, Bates D, Diegel M, Roach V, Dunn D, Neri J, Schafer A, Hansen RS, Kutayavin T, Giste E, Weaver M, Canfield T, Sabo P, Zhang M, Balasundaram G, Byron R, MacCoss MJ, Akey JM, Bender MA, Groudine M, Kaul R, Stamatoyannopoulos JA (2012b) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 489:83–90. doi: 10.1038/nature11212
- Nègre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, Kheradpour P, Eaton ML, Loriaux P, Sealfon R, Li Z, Ishii H, Spokony RF, Chen J, Hwang L, Cheng C, Auburn RP, Davis MB, Domanus M, Shah PK, Morrison CA, Zieba J, Suchy S, Senderowicz L, Victorsen A, Bild NA, Grundstad AJ, Hanley D, MacAlpine DM,

- Mannervik M, Venken K, Bellen H, White R, Gerstein M, Russell S, Grossman RL, Ren B, Posakony JW, Kellis M, White KP (2011) A cis-regulatory map of the *Drosophila* genome. *Nature* 471:527–531. doi: 10.1038/nature09990
- Ng J-H, Kumar V, Muratani M, Kraus P, Yeo J-C, Yaw L-P, Xue K, Lufkin T, Prabhakar S, Ng H-H (2013) In vivo epigenomic profiling of germ cells reveals germ cell molecular signatures. *Developmental Cell* 24:324–333. doi: 10.1016/j.devcel.2012.12.011
- Nobrega MA, Ovcharenko I, Afzal V, Rubin EM (2003) Scanning human gene deserts for long-range enhancers. *Science* 302:413. doi: 10.1126/science.1088328
- Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J, Gribnau J, Barillot E, Blüthgen N, Dekker J, Heard E (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 485:381–385. doi: 10.1038/nature11049
- Okano M, Bell DW, Haber DA, Li E (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* 99:247–257.
- Ong C-T, Corces VG (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet* 12:283–293. doi: 10.1038/nrg2957
- Paige SL, Thomas S, Stoick-Cooper CL, Wang H, Maves L, Sandstrom R, Pabon L, Reinecke H, Pratt G, Keller G, Moon RT, Stamatoyannopoulos J, Murry CE (2012) A temporal chromatin signature in human embryonic stem cells identifies regulators of cardiac development. *Cell* 151:221–232. doi: 10.1016/j.cell.2012.08.027
- Parker SCJ, Stitzel ML, Taylor DL, Orozco JM, Erdos MR, Akiyama JA, van Bueren KL, Chines PS, Narisu N, NISC Comparative Sequencing Program, Black BL, Visel A, Pennacchio LA, Collins FS, National Institutes of Health Intramural Sequencing Center Comparative Sequencing Program Authors, NISC Comparative Sequencing Program Authors (2013) Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc Natl Acad Sci USA* 110:17921–17926. doi: 10.1073/pnas.1317023110
- Pasini D, Bracken AP, Hansen JB, Capillo M, Helin K (2007) The Polycomb Group Protein Suz12 Is Required for Embryonic Stem Cell Differentiation. *Molecular and Cellular Biology* 27:3769–3779. doi: 10.1128/MCB.01432-06

- Pastor WA, Aravind L, Rao A (2013) TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nat Rev Mol Cell Biol* 14:341–356. doi: 10.1038/nrm3589
- Pastor WA, Pape UJ, Huang Y, Henderson HR, Lister R, Ko M, McLoughlin EM, Brudno Y, Mahapatra S, Kapranov P, Tahiliani M, Daley GQ, Liu XS, Ecker JR, Milos PM, Agarwal S, Rao A (2011) Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* 473:394–397. doi: 10.1038/nature10102
- Patel A, Vought VE, Dharmarajan V, Cosgrove MS (2011) A novel non-SET domain multi-subunit methyltransferase required for sequential nucleosomal histone H3 methylation by the mixed lineage leukemia protein-1 (MLL1) core complex. *J Biol Chem* 286:3359–3369. doi: 10.1074/jbc.M110.174524
- Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee S-I, Cooper GM, Ahituv N, Pennacchio LA, Shendure J (2012) Massively parallel functional dissection of mammalian enhancers in vivo. *Nature Biotechnology* 30:265–270. doi: 10.1038/nbt.2136
- Phillips JE, Corces VG (2009) CTCF: master weaver of the genome. *Cell* 137:1194–1211. doi: 10.1016/j.cell.2009.06.001
- Phillips-Cremins JE, Sauria MEG, Sanyal A, Gerasimova TI, Lajoie BR, Bell JSK, Ong C-T, Hookway TA, Guo C, Sun Y, Bland MJ, Wagstaff W, Dalton S, McDevitt TC, Sen R, Dekker J, Taylor J, Corces VG (2013) Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* 153:1281–1295. doi: 10.1016/j.cell.2013.04.053
- Rada-Iglesias A, Bajpai R, Prescott S, Brugmann SA, Swigut T, Wysocka J (2012) Epigenomic Annotation of Enhancers Predicts Transcriptional Regulators of Human Neural Crest. *Stem Cell* 11:633–648. doi: 10.1016/j.stem.2012.07.006
- Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J (2011) A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* 470:279–283. doi: 10.1038/nature09692
- Raiber E-A, Beraldi D, Ficuz G, Burgess HE, Branco MR, Murat P, Oxley D, Booth MJ, Reik W, Balasubramanian S (2012) Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase. *Genome Biol* 13:R69. doi: 10.1186/gb-2012-13-8-r69
- Rajagopal N, Xie W, Li Y, Wagner U, Wang W, Stamatoyannopoulos J, Ernst J, Kellis M, Ren B (2013) RFECS: a random-forest based algorithm for enhancer

- identification from chromatin state. *PLoS Comput Biol* 9:e1002968. doi: 10.1371/journal.pcbi.1002968
- Ram O, Goren A, Amit I, Shores N, Yosef N, Ernst J, Kellis M, Gymrek M, Issner R, Coyne M, Durham T, Zhang X, Donaghey J, Epstein CB, Regev A, Bernstein BE (2011) Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. *Cell* 147:1628–1639. doi: 10.1016/j.cell.2011.09.057
- Ramalingam S, Annaluru N, Chandrasegaran S (2013) A CRISPR way to engineer the human genome. *Genome Biol* 14:107. doi: 10.1186/gb-2013-14-2-107
- Rhee HS, Pugh BF (2011) Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell* 147:1408–1419. doi: 10.1016/j.cell.2011.11.013
- Rinn JL, Chang HY (2012) Genome regulation by long noncoding RNAs. *Annu Rev Biochem* 81:145–166. doi: 10.1146/annurev-biochem-051410-092902
- Robertson AB, Dahl JA, Vågbo CB, Tripathi P, Krokan HE, Klungland A (2011) A novel method for the efficient and selective identification of 5-hydroxymethylcytosine in genomic DNA. *Nucleic Acids Res* 39:e55. doi: 10.1093/nar/gkr051
- Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Meth* 4:651–657. doi: 10.1038/nmeth1068
- Rozenblatt-Rosen O, Rozovskaia T, Burakov D, Sedkov Y, Tillib S, Blechman J, Nakamura T, Croce CM, Mazo A, Canaani E (1998) The C-terminal SET domains of ALL-1 and TRITHORAX interact with the INI1 and SNR1 proteins, components of the SWI/SNF complex. *Proc Natl Acad Sci USA* 95:4152–4157.
- Ruthenburg AJ, Allis CD, Wysocka J (2007) Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark. *Mol Cell* 25:15–30. doi: 10.1016/j.molcel.2006.12.014
- Sagai T, Hosoya M, Mizushina Y, Tamura M, Shiroishi T (2005) Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development* 132:797–803. doi: 10.1242/dev.01613
- Sajan SA, Hawkins RD (2012) Methods for Identifying Higher-Order Chromatin

- Structure. *Annu Rev Genom Human Genet* 13:59–82. doi: 10.1146/annurev-genom-090711-163818
- Sanyal A, Lajoie BR, Jain G, Dekker J (2012) The long-range interaction landscape of gene promoters. *Nature* 489:109–113. doi: 10.1038/nature11279
- Sarkar A, Hochedlinger K (2013) The Sox Family of Transcription Factors: Versatile Regulators of Stem and Progenitor Cell Fate. *Stem Cell* 12:15–30. doi: 10.1016/j.stem.2012.12.007
- Sauvageau M, Goff LA, Lodato S, Bonev B, Groff AF, Gerhardinger C, Sanchez-Gomez DB, Hacisuleyman E, Li E, Spence M, Liapis SC, Mallard W, Morse M, Swerdel MR, D'Ecclesiss MF, Moore JC, Lai V, Gong G, Yancopoulos GD, Friendewey D, Kellis M, Hart RP, Valenzuela DM, Arlotta P, Rinn JL (2013) Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife* 2:e01749. doi: 10.7554/eLife.01749
- Sawado T, Halow J, Bender MA, Groudine M (2003) The beta -globin locus control region (LCR) functions primarily by enhancing the transition from transcription initiation to elongation. *Genes & Development* 17:1009–1018. doi: 10.1101/gad.1072303
- Schones DE, Cui K, Cuddapah S, Roh T-Y, Barski A, Wang Z, Wei G, Zhao K (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell* 132:887–898. doi: 10.1016/j.cell.2008.02.022
- Schübeler D, Francastel C, Cimborra DM, Reik A, Martin DI, Groudine M (2000) Nuclear localization and histone acetylation: a pathway for chromatin opening and transcriptional activation of the human beta-globin locus. *Genes & Development* 14:940–950.
- Schübeler D, Groudine M, Bender MA (2001) The murine beta-globin locus control region regulates the rate of transcription but not the hyperacetylation of histones at the active genes. *Proc Natl Acad Sci USA* 98:11432–11437. doi: 10.1073/pnas.201394698
- Selvaraj S, Dixon JR, Bansal V, Ren B (2013) Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nature Biotechnology* 31:1111–1118. doi: 10.1038/nbt.2728
- Serre D, Lee BH, Ting AH (2010) MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic Acids Res* 38:391–399. doi: 10.1093/nar/gkp992
- Sérandour AA, Avner S, Percevault F, Demay F, Bizot M, Lucchetti-Miganeh C,

- Barloy-Hubler F, Brown M, Lupien M, Métivier R, Salbert G, Eeckhoute J (2011) Epigenetic switch involved in activation of pioneer factor FOXA1-dependent enhancers. *Genome Res* 21:555–565. doi: 10.1101/gr.111534.110
- Sheffield NC, Thurman RE, Song L, Safi A, Stamatoyannopoulos JA, Lenhard B, Crawford GE, Furey TS (2013) Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Res* 23:777–788. doi: 10.1101/gr.152140.112
- Shen L, Wu H, Diep D, Yamaguchi S, D'Alessio AC, Fung H-L, Zhang K, Zhang Y (2013) Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* 153:692–706. doi: 10.1016/j.cell.2013.04.002
- Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenko VV, Ren B (2012) A map of the cis-regulatory sequences in the mouse genome. *Nature* 488:116–120. doi: 10.1038/nature11243
- Shilatifard A (2012) The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis. *Annu Rev Biochem* 81:65–95. doi: 10.1146/annurev-biochem-051710-134100
- Shlyueva D, Stampfel G, Stark A (2014) Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* 15:272–286. doi: 10.1038/nrg3682
- Smallwood A, Ren B (2013) Genome organization and long-range regulation of gene expression by enhancers. *Curr Opin Cell Biol* 25:387–394. doi: 10.1016/j.ceb.2013.02.005
- Smith ZD, Meissner A (2013) DNA methylation: roles in mammalian development. *Nat Rev Genet* 14:204–220. doi: 10.1038/nrg3354
- Song C-X, Szulwach KE, Dai Q, Fu Y, Mao S-Q, Lin L, Street C, Li Y, Poidevin M, Wu H, Gao J, Liu P, Li L, Xu G-L, Jin P, He C (2013) Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* 153:678–691. doi: 10.1016/j.cell.2013.04.001
- Song C-X, Yi C, He C (2012) Mapping recently identified nucleotide variants in the genome and transcriptome. *Nature Biotechnology* 30:1107–1116. doi: 10.1038/nbt.2398
- Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Graf S, Huss M, Keefe D, Liu Z, London D, McDaniell RM, Shibata Y, Showers KA,

- Simon JM, Vales T, Wang T, Winter D, Zhang Z, Clarke ND, Birney E, Iyer VR, Crawford GE, Lieb JD, Furey TS (2011) Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res* 21:1757–1767. doi: 10.1101/gr.121541.111
- Southall SM, Wong P-S, Odho Z, Roe SM, Wilson JR (2009) Structural basis for the requirement of additional factors for MLL1 SET domain activity and recognition of epigenetic marks. *Mol Cell* 33:181–191. doi: 10.1016/j.molcel.2008.12.029
- Spicuglia S, Kumar S, Yeh J-H, Vachez E, Chasson L, Gorbach S, Cautres J, Ferrier P (2002) Promoter activation by enhancer-dependent and -independent loading of activator and coactivator complexes. *Mol Cell* 10:1479–1487.
- Splinter E, de Wit E, van de Werken HJG, Klous P, de Laat W (2012) Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: From fixation to computation. *Methods* 58:221–230. doi: 10.1016/j.ymeth.2012.04.009
- Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Schöler A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, Tiwari VK, Schübeler D (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* 480:490–495. doi: 10.1038/nature10716
- Statham AL, Robinson MD, Song JZ, Coolen MW, Stirzaker C, Clark SJ (2012) Bisulfite sequencing of chromatin immunoprecipitated DNA (BisChIP-seq) directly informs methylation status of histone-modified DNA. *Genome Res* 22:1120–1127. doi: 10.1101/gr.132076.111
- Stergachis AB, Neph S, Reynolds A, Humbert R, Miller B, Paige SL, Vernot B, Cheng JB, Thurman RE, Sandstrom R, Haugen E, Heimfeld S, Murry CE, Akey JM, Stamatoyannopoulos JA (2013) Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell* 154:888–903. doi: 10.1016/j.cell.2013.07.020
- Sur IK, Hallikas O, Vähärautio A, Yan J, Turunen M, Enge M, Taipale M, Karhu A, Aaltonen LA, Taipale J (2012) Mice lacking a Myc enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. *Science* 338:1360–1363. doi: 10.1126/science.1228606
- Svetelis A, Bianco S, Madore J, Huppé G, Nordell-Markovits A, Mes-Masson A-M, Gévry N (2011) H3K27 demethylation by JMJD3 at a poised enhancer of anti-apoptotic gene BCL2 determines ER α ligand dependency. *EMBO J* 30:3947–3961. doi: 10.1038/emboj.2011.284

- Szulwach KE, Li X, Li Y, Song C-X, Han JW, Kim S, Namburi S, Hermetz K, Kim JJ, Rudd MK, Yoon Y-S, Ren B, He C, Jin P (2011) Integrating 5-hydroxymethylcytosine into the epigenomic landscape of human embryonic stem cells. *PLoS Genet* 7:e1002154. doi: 10.1371/journal.pgen.1002154
- Szutorisz H, Dillon N, Tora L (2005) The role of enhancers as centres for general transcription factor recruitment. *Trends Biochem Sci* 30:593–599. doi: 10.1016/j.tibs.2005.08.006
- Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L, Rao A (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* 324:930–935. doi: 10.1126/science.1170116
- Tan M, Luo H, Lee S, Jin F, Yang JS, Montellier E, Buchou T, Cheng Z, Rousseaux S, Rajagopal N, Lu Z, Ye Z, Zhu Q, Wysocka J, Ye Y, Khochbin S, Ren B, Zhao Y (2011) Identification of 67 Histone Marks and Histone Lysine Crotonylation as a New Type of Histone Modification. *Cell* 146:1016–1028. doi: 10.1016/j.cell.2011.08.008
- Tanaka Y, Katagiri Z-I, Kawahashi K, Kioussis D, Kitajima S (2007) Trithorax-group protein ASH1 methylates histone H3 lysine 36. *Gene* 397:161–168. doi: 10.1016/j.gene.2007.04.027
- Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, Kerr ARW, Deaton A, Andrews R, James KD, Turner DJ, Illingworth R, Bird A (2010) CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* 464:1082–1086. doi: 10.1038/nature08924
- Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernet B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kuttyavin T, Lajoie B, Lee B-K, Lee K, London D, Lotakis D, Neph S, Neri F, Nguyen ED, Qu H, Reynolds AP, Roach V, Safi A, Sanchez ME, Sanyal A, Shafer A, Simon JM, Song L, Vong S, Weaver M, Yan Y, Zhang Z, Zhang Z, Lenhard B, Tewari M, Dorschner MO, Hansen RS, Navas PA, Stamatoyannopoulos G, Iyer VR, Lieb JD, Sunyaev SR, Akey JM, Sabo PJ, Kaul R, Furey TS, Dekker J, Crawford GE, Stamatoyannopoulos JA (2012) The accessible chromatin landscape of the human genome. *Nature* 489:75–82. doi: 10.1038/nature11232
- Tian Z, Tolić N, Zhao R, Moore RJ, Hengel SM, Robinson EW, Stenoien DL, Wu S, Smith RD, Paša-Tolić L (2012) Enhanced top-down characterization of histone post-translational modifications. *Genome Biol* 13:R86. doi:

10.1186/gb-2012-13-10-R86

- Tolhuis B, Palstra R-J, Splinter E, Grosveld F, de Laat W (2002) Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol Cell* 10:1453–1465.
- van de Werken HJG, de Vree PJP, Splinter E, Holwerda SJB, Klous P, de Wit E, de Laat W (2012a) 4C Technology: Protocols and Data Analysis, 1st ed. *Nucleosomes, Histones & Chromatin Part B* 513:89–112. doi: 10.1016/B978-0-12-391938-0.00004-5
- van de Werken HJG, Landan G, Holwerda SJB, Hoichman M, Klous P, Chachik R, Splinter E, Valdes-Quezada C, Oz Y, Bouwman BAM, Verstegen MJAM, de Wit E, Tanay A, de Laat W (2012b) Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat Meth* 9:969–972. doi: 10.1038/nmeth.2173
- van Steensel B, Dekker J (2010) Genomics tools for unraveling chromosome architecture. *Nature Biotechnology* 28:1089–1095. doi: 10.1038/nbt.1680
- Vanner RJ, Remke M, Gallo M, Selvadurai HJ, Coutinho F, Lee L, Kushida M, Head R, Morrissy S, Zhu X, Aviv T, Voisin V, Clarke ID, Li Y, Mungall AJ, Moore RA, Ma Y, Jones SJM, Marra MA, Malkin D, Northcott PA, Kool M, Pfister SM, Bader G, Hochedlinger K, Korshunov A, Taylor MD, Dirks PB (2014) Quiescent Sox2. *Cancer Cell* 26:33–47. doi: 10.1016/j.ccr.2014.05.005
- Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J,

- Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X (2001) The sequence of the human genome. *Science* 291:1304–1351. doi: 10.1126/science.1058040
- Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio LA (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 457:854–858. doi: 10.1038/nature07730
- Waddington CH (1959) Canalization of development and genetic assimilation of acquired characters. *Nature* 183:1654–1655.
- Wamstad JA, Alexander JM, Truty RM, Shrikumar A, Li F, Eilertson KE, Ding H, Wylie JN, Pico AR, Capra JA, Erwin G, Kattman SJ, Keller GM, Srivastava D, Levine SS, Pollard KS, Holloway AK, Boyer LA, Bruneau BG (2012) Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell* 151:206–220. doi: 10.1016/j.cell.2012.07.035
- Wang D, Garcia-Bassets I, Benner C, Li W, Su X, Zhou Y, Qiu J, Liu W, Kaikkonen MU, Ohgi KA, Glass CK, Rosenfeld MG, Fu X-D (2011) Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* 474:390–394. doi: 10.1038/nature10006
- Wang H, Cao R, Xia L, Erdjument-Bromage H, Borchers C, Tempst P, Zhang Y (2001) Purification and functional characterization of a histone H3-lysine 4-specific methyltransferase. *Mol Cell* 8:1207–1217.

- Wang H, Yang H, Shivalila CS, Dawlaty MM, Cheng AW, Zhang F, Jaenisch R (2013) One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* 153:910–918. doi: 10.1016/j.cell.2013.04.025
- Wang P, Lin C, Smith ER, Guo H, Sanderson BW, Wu M, Gogol M, Alexander T, Seidel C, Wiedemann LM, Ge K, Krumlauf R, Shilatifard A (2009) Global analysis of H3K4 methylation defines MLL family member targets and points to a role for MLL1-mediated H3K4 methylation in the regulation of transcriptional initiation by RNA polymerase II. *Molecular and Cellular Biology* 29:6074–6085. doi: 10.1128/MCB.00924-09
- Watson JD, Crick FH (1953) Genetical implications of the structure of deoxyribonucleic acid. *Nature* 171:964–967.
- Weigelt K, Lichtinger M, Rehli M, Langmann T (2009) Transcriptomic profiling identifies a PU.1 regulatory network in macrophages. *Biochem Biophys Res Commun* 380:308–312. doi: 10.1016/j.bbrc.2009.01.067
- Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H (2013) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 42:D1001–D1006. doi: 10.1093/nar/gkt1229
- Whyte WA, Bilodeau S, Orlando DA, Hoke HA, Frampton GM, Foster CT, Cowley SM, Young RA (2012) Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. *Nature* 482:221–225. doi: 10.1038/nature10805
- Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* 153:307–319. doi: 10.1016/j.cell.2013.03.035
- Williams K, Christensen J, Pedersen MT, Johansen JV, Cloos PAC, Rappsilber J, Helin K (2011) TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* 473:343–348. doi: 10.1038/nature10066
- Williamson I, Eskeland R, Lettice LA, Hill AE, Boyle S, Grimes GR, Hill RE, Bickmore WA (2012) Anterior-posterior differences in HoxD chromatin topology in limb development. *Development* 139:3157–3167. doi: 10.1242/dev.081174
- Won K-J, Zhang X, Wang T, Ding B, Raha D, Snyder M, Ren B, Wang W (2013) Comparative annotation of functional regions in the human genome using epigenomic data. *Nucleic Acids Res* 41:4423–4432. doi: 10.1093/nar/gkt143

- Workman JL, Kingston RE (1998) Alteration of nucleosome structure as a mechanism of transcriptional regulation. *Annu Rev Biochem* 67:545–579. doi: 10.1146/annurev.biochem.67.1.545
- Wysocka J, Swigut T, Milne TA, Dou Y, Zhang X, Burlingame AL, Roeder RG, Brivanlou AH, Allis CD (2005) WDR5 associates with histone H3 methylated at K4 and is essential for H3 K4 methylation and vertebrate development. *Cell* 121:859–872. doi: 10.1016/j.cell.2005.03.036
- Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, Zhou X, Lee HJ, Maire CL, Ligon KL, Gascard P, Sigaroudinia M, Tlsty TD, Kadlecsek T, Weiss A, O'Geen H, Farnham PJ, Madden PAF, Mungall AJ, Tam A, Kamoh B, Cho S, Moore R, Hirst M, Marra MA, Costello JF, Wang T (2013a) DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nature Publishing Group* 45:836–841. doi: 10.1038/ng.2649
- Xie W, Schultz MD, Lister R, Hou Z, Rajagopal N, Ray P, Whitaker JW, Tian S, Hawkins RD, Leung D, Yang H, Wang T, Lee AY, Swanson SA, Zhang J, Zhu Y, Kim A, Nery JR, Urich MA, Kuan S, Yen C-A, Klugman S, Yu P, Suknuntha K, Propson NE, Chen H, Edsall LE, Wagner U, Li Y, Ye Z, Kulkarni A, Xuan Z, Chung W-Y, Chi NC, Antosiewicz-Bourget JE, Slukvin I, Stewart R, Zhang MQ, Wang W, Thomson JA, Ecker JR, Ren B (2013b) Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* 153:1134–1148. doi: 10.1016/j.cell.2013.04.022
- Xu J, Bauer DE, Kerenyi MA, Vo TD, Hou S, Hsu Y-J, Yao H, Trowbridge JJ, Mandel G, Orkin SH (2013) Corepressor-dependent silencing of fetal hemoglobin expression by BCL11A. *Proc Natl Acad Sci USA* 110:6518–6523. doi: 10.1073/pnas.1303976110
- Yao J, Chu Y, An R, Guo H (2012) Understanding product specificity of protein lysine methyltransferases from QM/MM molecular dynamics and free energy simulations: the effects of mutation on SET7/9 beyond the Tyr/Phe switch. *J Chem Inf Model* 52:449–456. doi: 10.1021/ci200364m
- Yu BD, Hess JL, Horning SE, Brown GA, Korsmeyer SJ (1995) Altered Hox expression and segmental identity in Mll-mutant mice. *Nature* 378:505–508. doi: 10.1038/378505a0
- Yu M, Hon GC, Szulwach KE, Song C-X, Zhang L, Kim A, Li X, Dai Q, Shen Y, Park B, Min J-H, Jin P, Ren B, He C (2012) Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* 149:1368–1380. doi: 10.1016/j.cell.2012.04.027

- Yuan G-C, Liu Y-J, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ (2005) Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* 309:626–630. doi: 10.1126/science.1112178
- Zappone MV, Galli R, Catena R, Meani N, De Biasi S, Mattei E, Tiveron C, Vescovi AL, Lovell-Badge R, Ottolenghi S, Nicolis SK (2000) Sox2 regulatory sequences direct expression of a (beta)-geo transgene to telencephalic neural stem cells and precursors of the mouse embryo, revealing regionalization of gene expression in CNS stem cells. *Development* 127:2367–2382.
- Zentner GE, Tesar PJ, Scacheri PC (2011) Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res* 21:1273–1283. doi: 10.1101/gr.122382.111
- Zhao Z, Tavoosidana G, Sjölander M, Göndör A, Mariano P, Wang S, Kanduri C, Lezcano M, Singh Sandhu K, Singh U, Pant V, Tiwari V, Kurukuti S, Ohlsson R (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet* 38:1341–1347. doi: 10.1038/ng1891
- Zhou H, Kaplan T, Li Y, Grubisic I, Zhang Z, Wang PJ, Eisen MB, Tjian R (2013) Dual functions of TAF7L in adipocyte differentiation. *Elife* 2:e00170. doi: 10.7554/eLife.00170
- Zhou HY, Katsman Y, Dhaliwal NK, Davidson S, Macpherson NN, Sakthidevi M, Collura F, Mitchell JA (2014) A Sox2 distal enhancer cluster regulates embryonic stem cell differentiation potential. *Genes & Development* 28:2699–2711. doi: 10.1101/gad.248526.114
- Zhu J, Adli M, Zou JY, Verstappen G, Coyne M, Zhang X, Durham T, Miri M, Deshpande V, De Jager PL, Bennett DA, Houmard JA, Muoio DM, Onder TT, Camahort R, Cowan CA, Meissner A, Epstein CB, Shores N, Bernstein BE (2013) Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* 152:642–654. doi: 10.1016/j.cell.2012.12.033
- Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LTY, Kohlbacher O, De Jager PL, Rosen ED, Bennett DA, Bernstein BE, Gnirke A, Meissner A (2013) Charting a dynamic DNA methylation landscape of the human genome. *Nature* 1–5. doi: 10.1038/nature12433