Title
Computational methods for studying gene regulation and genome organization using high-throughput DNA sequencing

Permalink
https://escholarship.org/uc/item/8ns0n0qb

Author
Bonora, Giancarlo A

Publication Date
2015

Peer reviewed|Thesis/dissertation
Computational methods for studying gene regulation and genome organization using high-throughput DNA sequencing

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

by

Giancarlo A. Bonora

2015
ABSTRACT OF THE DISSERTATION

Computational methods for studying gene regulation and genome organization using high-throughput DNA sequencing

by

Giancarlo A. Bonora

Doctor of Philosophy in Bioinformatics

University of California, Los Angeles, 2015

Professor Kathrin Plath, Co-Chair

Professor Matteo Pellegrini, Co-Chair

The full sequencing of the human genome ushered in the genomics era and laid the foundation for a more comprehensive understanding of gene regulation and development. But, since the DNA sequence represents only one aspect of the genomic information housed within the nucleus, the question of exactly how it is utilized to direct developmental programs and tissue-specific gene expression is still an open one. However, rapid advances in high-throughput DNA sequencing (HTS) technologies over the past decade have allowed biologists to begin to tackle the question on a genomic scale. HTS has been coupled to bisulfite conversion of DNA for assessing cytosine methylation (bisulfite sequencing), to chromatin immunoprecipitation for ascertaining genomic locations bound by specific factors or found in a particular chromatin state.
(ChIP-seq), to the isolation of transcripts for the measurement of gene expression (RNA-seq), and to methods of chromosome conformation capture for the identification of genome-wide DNA-DNA interactions (4C-seq and Hi-C). The focus of my doctoral research has been the development of novel bioinformatics approaches to analyze the data produced by these technologies in order to shed light on how distinct cell identities are established and maintained. Here, I present highlights of this work in six chapters. Chapter 1 presents a study investigating DNA methylation changes going from the differentiated to pluripotent state, which shows that changes predominantly occur late in the process and are strongly associated with changes to chromatin state. Chapter 2 introduces methylation-sensitive restriction enzyme bisulfite sequencing (MREBS) as a method for assessing precise differential DNA methylation at cost comparable to RRBS, while providing additional information over a coverage area more comparable to WGBS. Chapter 3 presents a study showing that inhibition of ribonucleotide reductase decreased DNA methylation genome-wide by enhancing the incorporation of a cytidine analog into DNA. Chapter 4 describes a study showing that, for genes important to leaf senescence, temporal changes in expression closely matched changes to two histone modifications. Chapter 5 reviews cutting-edge research exploring the link between regulatory networks and genome organization. Chapter 6 describes a study showing that regulators responsible for cell identity contribute to cell type-specific genome organization.
The dissertation of Giancarlo A. Bonora is approved.

Jason Ernst

Siavash Kurdistani

Matteo Pellegrini, Committee Co-Chair

Kathrin Plath, Committee Co-Chair

University of California, Los Angeles

2015
This is dedicated to:

My mother and father, Jennifer and Aldo,

My darling wife, Aubrey Jean,

And Enzo Bravo!
# TABLE OF CONTENTS

ABSTRACT OF THE DISSERTATION ........................................... ii

ACKNOWLEDGMENTS ................................................................ix

VITA .................................................................................... xxvii

OVERVIEW ............................................................................. 1

References ........................................................................... 8

CHAPTER 1: DNA methylation dynamics during somatic cell reprogramming to pluripotency 11

1.1 Abstract ........................................................................... 12
1.2 Introduction ...................................................................... 13
1.3 Results ............................................................................ 18
  1.3.1 Study design and experimental approach ....................... 18
  1.3.2 DNA methylation profiles diverge from that seen in the somatic state during the course of reprogramming .................. 19
  1.3.3 Identifying and clustering differentially methylated regions of the genome during somatic cell reprogramming ................ 21
  1.3.4 DNA demethylation during reprogramming is associated with increased expression by genes related to the pluripotent state while DNA re-methylation is associated with reduced expression of developmental and somatic genes ..... 24
1.3.5 DNA methylation at reprogramming factor binding sites and its association with histone modifications ................................................................. 28

1.3.6 DNA methylation at sites of reprogramming factor co-binding and its association with histone modifications .............................................. 32

1.3.7 DNA methylation distributions in different chromatin states ....................... 37

1.3.8 Significant changes in DNA methylation with respect to chromatin state ........ 40

1.3.9 DNA methylation changes in regions undergoing transitions in chromatin state during somatic cell reprogramming ........................................ 42

1.3.10 Modeling DNA methylation using histone modifications .......................... 43

1.4 Discussion ........................................................................................................ 50

Materials and methods .......................................................................................... 56

Figures .................................................................................................................... 63

Tables ...................................................................................................................... 86

References ............................................................................................................ 95

CHAPTER 2: Determining differential DNA methylation using methylation-sensitive restriction enzyme bisulfite sequencing

2.1 Abstract ............................................................................................................ 102

2.2 Introduction ....................................................................................................... 103

2.3 Results

2.3.1 Study design and data sets ........................................................................ 106
2.3.2 MREBS DNA methylation estimates follow the same trends as WGBS and RRBS in different chromatin states .......................................................... 107

2.3.3 MREBS read counts provide additional information for determining differential DNA methylation ............................................................................................................. 109

2.3.4 MREBS DNA methylation estimates correlate and read counts anti-correlate with WGBS and RRBS DNA methylation estimates ........................................ 110

2.3.5 CpG-level DNA methylation can be modeled using MREBS data ............... 111

2.4 Discussion ........................................................................................................... 116

Materials and methods ........................................................................................... 118

Figures ..................................................................................................................... 123

Tables ...................................................................................................................... 133

References .............................................................................................................. 140

CHAPTER 3: A high-throughput screen of inactive X chromosome reactivation identifies the enhancement of DNA demethylation by 5-aza-2’-dC upon inhibition of ribonucleotide reductase

3.1 Abstract ............................................................................................................. 144

3.2 Introduction ....................................................................................................... 145

3.3 Results

3.3.1 An siRNA screen for XCR in the presence of a low 5-aza-2’-dC dose identifies the ribonucleotide reductase pathway ............................................... 148

3.3.2 Inhibitors of ribonucleotide reductase elicit XCR with 5-aza-2’-dC present .... 151
3.3.3 RNR inhibition increases incorporation of 5-aza-2’-dC into DNA .......................... 153
3.3.4 Rrm2 inhibition enhances demethylation caused by 5-aza-2’-dC ......................... 154
3.3.5 Hydroxyurea and 5-aza-2’-dC synergistically inhibit myeloid leukemia cell
line proliferation in a dose-dependent fashion ..................................................... 156

3.4 Discussion ........................................................................................................... 160

Materials and methods ............................................................................................ 164

Figures ....................................................................................................................... 172

References .................................................................................................................. 189

CHAPTER 4: A genome-wide chronological study of gene expression and two histone modifications, H3K4me3 and H3K9ac, during developmental leaf senescence

4.1 Abstract ............................................................................................................... 195

4.2 Introduction ......................................................................................................... 196

4.3 Results

4.3.1 RNA-seq gene expression analysis ................................................................. 199

4.3.2 ChIP-seq analysis for H3K4me3 and H3K9ac .................................................. 201

4.3.3 Changes in histone marks during leaf senescence .......................................... 202

4.3.4 Temporal patterns in histone mark acquisition and loss ................................. 205

4.3.5 Histone mark gene coverage correlates with gene expression during leaf
senescence ................................................................................................................. 208

4.4 Discussion .......................................................................................................... 210
CHAPTER 5: A mechanistic link between gene regulation and genome architecture in mammalian development

Abstract ........................................................................................................................................... 252

Introduction ...................................................................................................................................... 252

The segregated nucleus: compartmentalization of nuclear function ................................................. 253

Long-distance relationships: cell type-specific inter-TAD interactions point to a role for gene regulatory factors in higher order genome organization .................................................. 254

The logic behind enhancer–promoter–exon looping ........................................................................... 256

The linchpins of looping: architectural proteins and chromatin contacts ........................................... 257

Completing the loop and looping ahead: future directions ............................................................... 259

Acknowledgements ........................................................................................................................... 259

References and recommended reading ............................................................................................ 259
CHAPTER 6: Long-range chromatin contacts in embryonic stem cells reveal a role for the pluripotency factors and Polycomb proteins in genome organization

Summary ........................................................................................................................................... 263

Introduction ......................................................................................................................................... 263

Results

Experimental approach to studying chromatin contacts ................................................................. 264

A pluripotency-specific organization of the mouse genome ......................................................... 264

Open/closed chromatin as the foundation of genome organization in ESCs ......................... 266

Genomic regions enriched forOct4/Sox2/Nanog and Polycomb proteins frequently colocalize in ESCs ......................................................................................................................................................... 268

Spatial segregation of Nanog and H3K27me3 in the ESC nucleus ........................................... 270

Changes in open/closed chromatin character mirror changes in genome organization during differentiation ................................................................................................................................................................................. 272

The preferential colocalization of Polycomb-enriched genomic regions is Eed dependent . 274

Discussion .......................................................................................................................................... 275

Experimental procedures ................................................................................................................. 276

Acknowledgements ......................................................................................................................... 276

References ......................................................................................................................................... 276

Supplemental information and inventory ..................................................................................... 278

Supplemental figures and legends ................................................................................................. 281

Supplemental experimental procedures ....................................................................................... 293

Supplemental references ................................................................................................................ 307
# LIST OF FIGURES

1.1 Genome-wide DNA methylation during somatic cell reprogramming ................................................................. 63
1.2 Differentially methylated regions during somatic cell reprogramming ................................................................. 65
1.3 DNA methylation levels and histone modification enrichment around EARLY-specific (‘100’) reprogramming factor binding sites ........................................................................................................ 68
1.4 DNA methylation levels and histone modification enrichment around ESC-specific (‘001’) reprogramming factor binding sites ........................................................................................................ 71
1.5 DNA methylation levels and histone modification enrichment around constitutively bound (‘111’) reprogramming factors ........................................................................................................ 72
1.6 Clustering of reprogramming factors reveals that co-binding is associated with distinct DNA methylation and histone modification patterns ....................................................................................... 73
1.7 The DNA methylation and histone modification enrichment patterns associated with transcription factor cluster groups C, D, and E ....................................................................................................... 75
1.8 DNA methylation levels with respect to chromatin state .......................................................................................... 76
1.9 Chromatin state transitions occur throughout somatic cell reprogramming but DNA methylation changes are predominantly occur late in the process ................................................................. 79
1.10 DNA methylation levels can be modeled using histone modifications ................................................................. 81
1.11 Modeled DNA methylation levels in different chromatin states ............................................................................... 84
2.1 WGBS, RRBS, and MREBS for samples representing two stages of somatic cell reprogramming ......................................................................................................................................................... 123
2.2 DNA methylation estimates based on WGBS, RRBS and MREBS data in different chromatin states ......................................................................................................................................................... 125
2.3 Chromatin state coverage by DNA methylation estimates by WGBS, RRBS, and MREBS .......................................................... 127
2.4 Differential DNA methylation levels modeled using MREBS data ......................... 129
2.5 Examples of modeled differential DNA methylation around gene loci .................. 131
3.1 High-throughput siRNA and chemical screens identify RRM2 depletion and Resveratrol as mediators of XCR ............................................................ 172
3.2 Optimization of 5-aza-2′-dC concentration for genome-wide siRNA screen .......... 175
3.3 Batch effects in siRNA screen and robust z-score normalization ......................... 176
3.4 Validation of gene hits identified by genome-wide siRNA screening .................... 177
3.5 Chemical screen results and validation ................................................................ 178
3.6 Inhibition of RNR enhances DNA incorporation of 5-aza-2′-dC to elicit XCR ....... 179
3.7 RNR inhibition increases 5-aza-2′-dC-mediated DNA methylation in MEFs .......... 180
3.8 Analysis of DNA methylation in MEFs treated with combinations of RNR inhibition and 5-aza-2′-dC .................................................................. 182
3.9 DNA methylation status of the luciferase transgene in MEFs treated with combinations of RNR inhibition and 5-aza-2′-dC ..................................................... 184
3.10 Hydroxyurea and 5-aza-2′-dC treatment of myeloid leukemia cell lines .............. 186
3.11 Extended data on the methylation analysis of K562 cells ..................................... 188
4.1 Gene expression differences during leaf senescence ............................................ 221
4.2 Coincidence of H3K4me3 and H3K9ac marks ...................................................... 223
4.3 Histone modifications in leaves from differently aged plants ............................. 225
4.4 H3K4me3 and H3K9ac histone marks in SURGs and SDRGs not identified as K4- SURGs or K4-SDRGs .................................................................................. 227
4.5 K-means clustering for H3K4me3 gain peaks and correlation to gene expression ........ 229
4.6 K-means clustering for H3K4me3 loss peaks and correlation to gene expression ........ 231
4.7 K-means clustering for H3K9ac gain peaks and correlation to gene expression ........ 232
4.8 Average H3K4me3 and H3K9ac read count profiles for genes associated with H3K9ac_GAIN Clusters #1 and #2 .......................................................... 233
4.9. K-means clustering for H3K9ac loss peaks and correlation to gene expression ........ 234
4.10 Average H3K4me3 and H3K9ac read count profiles for the 387 K4-SURGs .......... 235
4.11 Breadth of histone modifications during leaf senescence ................................. 236
4.12 Correlation matrices and dendrograms for RNA-seq RPKMs and H3K4me3 gene coverage ........................................................................................................... 238
4.13 H3K4me3 and H3K9ac marks for three genes .................................................... 239
5.1 Gene regulatory factors shape inter-TAD chromatin interactions within the pluripotent nucleus ........................................................................................................ 255
5.2 Architectural proteins act combinatorially to organize chromatin at different length-scales ................................................................................................................ 258
6.1 Long-Range Chromatin Contacts of the Pou5f1 Bait Region in ESCs ................. 265
6.2 Long-Range Chromatin Contacts Change during Differen- tiation and Are Reset upon Reprogramming of Somatic Cells to iPSCs .............................................. 266
6.3 Interactions between Regions with Similar Open/Closed Chromatin Character Are an Intrinsic Aspect of Chromosome Conforma- tion in Mouse ESCs ...................... 267
6.4 Regions of Shared Transcriptional Network Occupancy Preferentially Interact ....... 269
6.5 Nanog and H3K27me3 Segregate in the ESC Nucleus ........................................ 271
6.6 Changes in Open/Closed Chromatin Character between ESCs and MEFs Correspond to Changes in Interaction Preferences ................................................................. 272

6.7 Eed Is Required for the Colocalization of Polycomb-Occupied Genomic Regions ...... 273

6.S1 Reproducibility, validation and quality control of 4C-seq data .................................. 281

6.S2 4C-seq replicate data sets cluster by cell type, revealing pluripotency- specific chromatin contacts ......................................................................................................................... 283

6.S3 Genome-wide analysis of interactomes in ESCs ......................................................... 285

6.S4 Relationship between PC1 score and individual feature enrichment ................. 287

6.S5 Comparison of the spatial interactomes between ESCs and MEFs ....................... 289

6.S6 Additional validation and characterization of interaction preferences in Eed+/+ and Eed-/- ESCs ......................................................................................................................... 291
LIST OF TABLES

1.1  Bisulfite sequencing library total mapped reads, mean CpG coverage depth, CpG
dimer coverage, and mean methylation per cell type ................................. 86
1.2  Significance values for tests of difference of expression values for genes associated
with differential DNA methylation cluster groups between the four cell types ......... 87
1.3  Gene ontology enrichment analysis results for genes associated with the differential
DNA methylation cluster group 3 .................................................................. 88
1.4  Gene ontology enrichment analysis results for genes associated with the differential
DNA methylation cluster group 4 .................................................................. 89
1.5  Gene ontology enrichment analysis results for genes associated with the differential
DNA methylation cluster group 5 .................................................................. 90
1.6  CpG dimer-level correlations between WGBS DNA methylation estimates and nine
histone modifications and two histone variants .............................................. 91
1.7  DNA methylation model coefficients based on WGBS data .......................... 92
1.8  DNA methylation model coefficients based on RRBS .................................. 93
1.9  DNA methylation model statistics for all four cell types .............................. 94
2.1  MRE endonuclease recognition sequence frequency within the mm9 genome .... 132
2.2  Bisulfite sequencing library mapped reads and mean CpG coverage depth .......... 133
2.3  CpG dimer coverage per bisulfite sequencing library .................................. 134
2.4  CpG dimer coverage for differential analysis per bisulfite sequencing library .... 135
2.5  CpG dimer-level correlations between bisulfite sequencing libraries .............. 136
2.6 CpG dimer-level correlations between differential values for all bisulfite sequencing library pairs ................................................................. 137

2.7 Differential DNA methylation model coefficients .............................................. 138

2.8 Differential DNA methylation model metrics ...................................................... 139

4.1 Sequencing and alignment summary for RNA-seq libraries .................................. 240

4.2 Gene Ontology enrichment for Senescence Up-Regulated Genes (SURGs) and Senescence Down-Regulated Genes (SDRGs) .............................................. 241

4.3 Sequencing and alignment summary for ChIP-seq libraries .................................. 242

4.4 Significance of pairwise comparisons in gene expression (p-values) .................... 243

5.1 Summary of chromosome conformation capture (3C)-based methods .................. 253
Funding
My tuition and stipend were supported by the following UCLA pre-doctoral fellowships: Graduate Division Dissertation Year Fellowship (2014/15), Philip Whitcome Pre-doctoral Fellowship (2011–15), Genetics Training Program Fellowship in Genetic Mechanisms (2010/11), Chancellor’s Recruitment Fellowship (2009/10), and the Boyer Recruitment Fellowship through the ACCESS program (2009). I was also partially supported by grants to K.P., who is supported by the NIH (DP2OD001686 and P01 GM099134), CIRM (RN1-00564, RB3-05080, and RB4-06133), the Jonsson Comprehensive Cancer Center and the Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research at UCLA. And I was provided funds to purchase an Apple MacBook, peripherals, software licenses, and to attend conferences by M.P.

Chapter contributions
Chapter 1 is a version of a manuscript in preparation for publication with the provisional title:

DNA methylation dynamics during somatic cell reprogramming to pluripotency
Giancarlo Bonora\textsuperscript{1,3,4,5,*}, Constantinos Chronis\textsuperscript{1,4,*}, Petko Fiziev\textsuperscript{1,3,4,5}, Jason Ernst\textsuperscript{1,2,3,4,6}, Matteo Pellegrini\textsuperscript{1,2,3,5,#}, and Kathrin Plath\textsuperscript{1,2,3,4,6,#}

*equal contribution, #correspondance: matteop@mcdb.ucla.edu; kplath@mednet.ucla.edu

\textsuperscript{1}Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, \textsuperscript{2}Molecular Biology Institute, \textsuperscript{3}Bioinformatics Interdepartmental Program, \textsuperscript{4}Department of Biological
Chemistry, 5Department of Molecular, Cell, and Developmental Biology, 6Jonsson Comprehensive Cancer Center at the David Geffen School of Medicine of the University of California, Los Angeles, CA, USA.

G.B. participated in project planning and data interpretation, performed bioinformatics analysis, and wrote the manuscript. C.C. performed the experimental data generation, participated in bioinformatics analysis, data interpretation, project planning, and assisted with writing the manuscript. PF performed bioinformatics analysis. JE provided guidance for bioinformatics analysis. M.P. provided guidance for bioinformatics analysis, participated in data interpretation, and provided supervision. K.P. conceived of the study, supervised the project, participated in data interpretation, and assisted with the writing of the manuscript.

Chapter 2 is a version of a manuscript in preparation for publication, provisionally entitled:

Differential DNA methylation assessment using methylation-sensitive restriction enzyme bisulfite sequencing

Giancarlo Bonora1,3,4,5.*, Marco Morselli1,2,5.*, Liudmilla Rubbi1,5*, Constantinos Chronis1,4, Kathrin Plath1,2,3,4,6#, and Matteo Pellegrini1,2,3,5,#

*equal contribution, #correspondance: kplath@mednet.ucla.edu; matteop@mcdb.ucla.edu

1Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, 2Molecular Biology Institute, 3Bioinformatics Interdepartmental Program, 4Department of Biological Chemistry, 5Department of Molecular, Cell, and Developmental Biology, 6Jonsson Comprehensive Cancer Center at the David Geffen School of Medicine of the University of California, Los Angeles, CA, USA.
G.B. participated in project planning and data interpretation, performed bioinformatics analysis, and wrote the manuscript. M.M. participated in project planning and data interpretation, and generated experimental data. L.R. participated in project planning and data interpretation, generated experimental data, and provided experimental guidance. C.C. produced in experimental data. K.P. participated in data interpretation and provided supervision. M.P. conceived of the study, supervised the project, interpreted the data, and provided guidance for bioinformatics analysis.

Chapter 3 is a version of an article that has been submitted for publication:

A high-throughput screen of inactive X chromosome reactivation identifies the enhancement of DNA demethylation by 5-aza-2'-dC upon inhibition of ribonucleotide reductase

Alissa Minkovksy1*, Anna Sahakyan1*, Giancarlo Bonora1*, Robert Damoiseaux2, Elizabeth Dimitrova3, Liudmilla Rubbi4, Matteo Pellegrini4, Caius G Radu3, and Kathrin Plath1#

*equal contribution, #correspondance: kplath@mednet.ucla.edu

1David Geffen School of Medicine, Department of Biological Chemistry, Jonsson Comprehensive Cancer Center, Molecular Biology Institute, and Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, 2Department of Molecular and Medicinal Chemistry and California NanoSystems Institute, 3Ahmanson Translational Imaging Division, Department of Molecular and Medical Pharmacology, 4Department of Molecular, Cell, and Developmental Biology, University of California Los Angeles, USA.

A.M. participated in project planning, experimental data generation, data interpretation, and manuscript writing. A.S. participated in project planning, experimental data generation, and data
interpretation. G.B. performed all bioinformatics analysis, and participated in data interpretation and manuscript writing. R.D. provided experimental guidance. E.D. participated in experimental data generation. L.R. participated in experimental data generation. M.P. provided guidance for bioinformatics analysis. C.G.R. provided experimental guidance. K.P. conceived of the study and supervised the project, and participated in data interpretation and manuscript writing.

Chapter 4 is a version of an article that has been accepted for publication and is in press:

A Genome-wide chronological study of gene expression and two histone modifications, H3K4me3 and H3K9ac, during developmental leaf senescence
Brusslan JA\(^1\)*#, Bonora G\(^2\)*, Rus-Canterbury AM\(^1\), Tariq F\(^1\), Jaroszewicz A\(^2\), Pellegrini M\(^2\).
*equal contribution; #correspondance: Judy.Brusslan@csulb.edu
\(^1\)Biological Science Department, California State University, Long Beach. \(^2\)Department of Molecular, Cell, and Developmental Biology, University of California Los Angeles.
J.B. conceived of the study, supervised the project, analyzed and interpreted the data, and led the manuscript writing. G.B. performed bioinformatics analysis, interpreted data, and wrote portions of the manuscript. A.R. and F.T. generated experimental data. A.R. processed RNA-seq data. M.P. provided guidance for bioinformatics analysis, interpreted data, and provided supervision.

Chapter 5 is a reprint of a published article:

A mechanistic link between gene regulation and genome architecture in mammalian development
Bonora G1, Plath K2#, Denholtz M3#
#correspondence: kplath@mednet.ucla.edu; matthew.denholtz@gmail.com

1David Geffen School of Medicine, Department of Biological Chemistry, Jonsson
Comprehensive Cancer Center, and the Eli and Edythe Broad Center of Regenerative Medicine
and Stem Cell Research, University of California Los Angeles, CA 90095, USA. 2David Geffen
School of Medicine, Department of Biological Chemistry, Jonsson Comprehensive Cancer
Center, and the Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research,
University of California Los Angeles, CA 90095, USA. 3David Geffen School of Medicine,
Department of Biological Chemistry, Jonsson Comprehensive Cancer Center, and the Eli and
Edythe Broad Center of Regenerative Medicine and Stem Cell Research, University of
California Los Angeles, CA 90095, USA.

G.B. conducted the literature review and wrote the paper, with the assistance of K.P. and M.D.,
who supervised the process.

Chapter 6 is a reprint of a published article:


Long-range chromatin contacts reveal a role for the pluripotency and Polycomb networks
in genome organization

Matthew Denholtz1,2,4,7*, Giancarlo Bonora1,2,3,4,5*, Constantinos Chronis1,4, Erik Splinter8,
Wouter de Laat8, Jason Ernst1,2,3,4,6, Matteo Pellegrini1,2,3,5,6#, and Kathrin Plath1,2,3,4,5,6#
*equal contribution; #correspondence: matteop@mcdb.ucla.edu; kplath@mednet.ucla.edu

1Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, 2Molecular
Biology Institute, 3Bioinformatics Interdepartmental Program, 4Department of Biological
Chemistry, 5Department of Molecular, Cell, and Developmental Biology, 6Jonsson Comprehensive Cancer Center, 7Molecular Biology Interdepartmental PhD Program, 1-7 at the David Geffen School of Medicine of University of California, Los Angeles, CA, USA, 8Hubrecht Institute-KNAW & University Medical Center, Utrecht, The Netherlands.

M.D. and G.B. designed research, performed experiments, analyzed data, and wrote the paper. C.C. performed experiments. E.S. and W.d.L. trained M.D. in 4C-seq library preparation in 2008. J.E. provided analytic tools. M.P. provided supervision and data analysis. K.P. designed research, analyzed data, wrote the paper, and supervised the overall project.

**Very important people**

First and foremost, I would like to thank my two co-chairs and co-mentors, Kathrin and Matteo, for their guidance, support, and understanding throughout the course of my PhD work. It has been a privilege and an honor to work with and learn from such exceptional researchers, and to come to better appreciate what it takes to earn their very well-founded reputations for mastery of their fields. Furthermore, they have served as role models as to how to successfully manage having a family alongside the demands of academic research, and were both wonderfully accommodating with respect to me starting a family in the midst of my graduate studies.

I would also like to thank Siavash Kurdistani and Jason Ernst for serving on my committee, for taking the time to meet personally or to attend a lab meeting on occasion, and for providing valuable scientific and career advice. Additional thanks to Jason for his suggestions regarding the clustering of transcription factor binding sites, and to both him and Petko Fiziev for
running chromHMM to generate the chromatin states used in studies presented in Chapters 1, 2 and 6.

I would like to express my immense gratitude to Matthew Denholtz, who spearheaded the research presented in Chapter 6. He performed all the 4C-seq and FISH experiments in the study, and in large part the project was ‘his baby.’ His deep understanding of the subject matter, salient comments, and insightful questions, guided my analytical work on the project. Matt also provided me with many important pointers and valuable advice in the writing of the review presented in Chapter 5. I also need to thank him for his ongoing support and interest in my progress.

Many thanks to Kostas Chronis, who contributed enormously to three of the six chapters in this dissertation. He produced all of the ChIP-seq and RNA-seq data sets for the study presented in Chapter 6 that were not already publically available. He is also leading a massive project, of which the research presented in Chapter 1 is a part, and for which he has produced the vast amount of data presented in that chapter, including all the bisulfite sequencing, ChIP-seq, and RNA-seq, which I also utilized in Chapter 2. His thorough knowledge of the field served to inform and guide my work on these projects. I also want to thank him for his valuable comments and critical reading of the draft manuscript that served as basis for Chapter 1.

Thank you to Marco Morcelli and Liudmilla Rubbi for producing the MREBS libraries presented in Chapter 2, and for informative discussions about the project. Thanks also need to go to Liudmilla for producing the RRBS libraries presented in Chapter 3.
Thanks to Alissa Minkovksy for spearheading the project presented in Chapter 3, and to Anna Sahakyan for all her contributions.

Many thanks to Judy Brusslan for having me on the project presented in Chapter 4, and for generously offering co-first authorship on the paper.

Thanks to all other members of the Plath and Pellegrini groups, past and present, who have provided advice and support through the years. In particular, many thanks to Bernadett Papp and Vincent Pasque for taking the time to read and comment on drafts of chapters in this dissertation. Thank you to Shawn Cokus and David Casero for the numerous times that they have taken the time to advise me regarding some or other bioinformatics problem. Thanks to Luz Orozco for showing me the ropes with respect to bisulfite sequencing analysis. Thanks to Kelvin Zhang for advice on RNA-seq analysis. Thanks to Rupa Sridharan for past conversations and a recent conversation about the future.

Thank you to everyone in the Bioinformatics and Biological Chemistry offices for all their help in various ways throughout my time as a graduate student at UCLA, especially Pamela Hurley and Phuong Pham, the SAOs for the bulk of my time. I would also like to thank the ACCESS program for the warm welcome and support that I received when I started out at UCLA, in particular to Jody Spillane, and her staff, as well as Greg Payne.
Thanks to the members of the Biology Department at CSU Dominguez Hills for providing me the encouragement and support needed to pursue my goal to conduct academic research. In particular, many thanks to Hee-Kwang Choi, Helen Chun, and John Thomlinson for their continued interest in my progress. And to Laura Robles and David Nishioka for their important support in the past.

Thank you to Sydella Blatch for taking the time to provide valuable advice regarding graduate studies to a relative stranger.

A great many thanks my wonderful parents, Jennifer and Aldo, as well as my lovely siblings, Alessandro, Angela, and Daniela, for all their continued support and encouragement.

And most importantly, I will never be able to thank my darling wife, Aubrey, enough for her remarkable support and patience, and for often having to pick up a lot of slack on the domestic front during while I have been in pursuit a PhD. I could never have got to this point without her love, counseling, and care. And thank you to my little man, Enzo, for putting up with my absences and understanding that although I may be busy at times, ‘Papa always comes home!’
VITA

1992–1994  B.S. in Computer Science and Mathematics, University of Cape Town, South Africa
1995–1997  B.S. in Electrical Engineering, University of Cape Town, South Africa
1998–2006  Programmer / Information technology consultant, Cape Town/London/Los Angeles
2003–2005  B.S. in Biology, Emphasis in Biochemistry and Physiology, University of South Africa (Incomplete practical modules)
2007–2009  M.S. in Biology, Emphasis in Molecular & Cell Biology, California State University, Dominguez Hills
2007–2009  Research assistant, Laboratory of Professor Hee Kwang Choi, California State University, Dominguez Hills
2007–2009  Laboratory instructor, General Biology, California State University, Dominguez Hills
2009  Lecturer, General Biology, California State University, Dominguez Hills
2009–2015  Graduate student researcher, Laboratories of Professor Kathrin Plath and Professor Matteo Pellegrini, University of California, Los Angeles
2011  Teaching Assistant, Genetics, Life Sciences Core, University of California, Los Angeles

AWARDS AND HONORS

2014/15  Dissertation Year Fellowship, Graduate Division, University of California, Los Angeles
2011–2014  Philip Whitcome Pre-doctoral Fellowship, Molecular Biology Institute, University of California, Los Angeles
2010/11  Pre-doctoral Fellowship in Genetic Mechanisms, Genetics Training Program, University of California, Los Angeles
2009  Chancellor’s Recruitment Fellowship, ACCESS program, University of California, Los Angeles
2009  Boyer Recruitment Fellowship, ACCESS program, University of California, Los Angeles
PUBLICATIONS


Overview
When the human genome was completely sequenced over a decade ago, it confirmed that the genomics era had arrived\textsuperscript{1}. This effort laid the groundwork for a more comprehensive understanding of gene regulation and development, but only represented one dimension of the information housed within the nucleus of each cell. Since the hundreds of different cell types in the human body all contain an identical copy of the genome, many questions still remain as to how the expression of only a very specific repertoire of genes is directed and controlled in each cell type to establish and maintain their distinct identities.

Transcription factors (TFs), which are proteins that can bind to specific DNA loci to turn genes on or off, are key players in establishing cell identity\textsuperscript{2}. For instance, somatic cell reprogramming is a method where terminally differentiated cells can be reprogrammed back into a pluripotent state by ectopic expression of four TFs known as the Yamanaka reprogramming factors\textsuperscript{3,4}.

Additionally, recent advances have helped to confirm and better appreciate the fundamental role played by an additional layer of regulatory mechanisms and information above and beyond that encoded by the DNA sequence, namely, the epigenome, which has been referred to as the second dimension of the genome\textsuperscript{5,6}. For example, cytosines within the DNA sequence can exist in different methylation states and the proteins around which the DNA strands are wound can be decorated with a dazzling array of covalent modification, both of which have been shown to be involved in imparting cell identity\textsuperscript{5-9}.

In parallel with this, newly available tools have served to uncover the relationship between genome organization and the regulation of transcription, helping to establish that genomes are organized around gene regulatory factors that govern cell identity\textsuperscript{10,11}. 
The human reference genome was first assembled using Sanger sequencing technology, as were those of model organisms such as the mouse\textsuperscript{1,12,13}. But, over the last decade the next generation of sequencing technology, high-throughput DNA sequencing (HTS), has taken center stage\textsuperscript{14}. As HTS technologies continue to develop and mature, biologists have an ever-expanding and evolving toolbox of approaches to draw on for investigating genomic phenomena\textsuperscript{14,15}.

An example of such a technology is whole-genome bisulfite sequencing (WGBS), which couples the conversion of unmethylated cytosine residues to uracil and HTS, allowing for DNA methylation estimates throughout the genome\textsuperscript{16-18}. Reduced representation bisulfite sequencing (RRBS) adds endonuclease digestion and DNA fragment size selection steps for deeper sequencing of an important subset of the genome and hence higher confidence methylation estimates at a fraction of the cost\textsuperscript{19,20}. Chromatin immunoprecipitation followed by HTS (ChIP-seq) enables researchers to determine locations of the genome bound by specific transcription factors, as well as the locations of histones marked by specific modifications\textsuperscript{21,22}. RNA-seq provides a digital measure of gene expression levels\textsuperscript{23,24}. Chromosome conformation capture (3C), a method to assess DNA-DNA interactions within a statistical population of cells, can be coupled to HTS to theoretically identify all genetic loci that come into close physical proximity with a gene or genetic locus of interest (4C-seq), or every locus in the genome (Hi-C), giving one a window into the three-dimensional (3D) genome organization of cells under study\textsuperscript{25-28}.

The analysis of the data produced by these HTS technologies has required the development of novel bioinformatics tools and analysis workflows. This endeavor has been the focus of my doctoral research under the guidance of my mentors, Professor Kathrin Plath and Professor Matteo Pellegrini. I will present highlights of this work in the subsequent six chapters,
within the context of the biological questions that these techniques were being used to address. Each chapter is a self-standing, with its own figures, tables, and references, since each is a reprint or modified version of a manuscript on which I am a first author. However, I cannot overemphasize the fact that in all cases these have been very much collaborative projects and I encourage the reader to please refer to the ACKNOWLEDGMENTS section for a full exposition of highly warranted attributions.

Chapter 1 is based on a manuscript in preparation for publication that utilized WGBS and RRBS in combination with ChIP-seq and RNA-seq to investigate DNA methylation changes at four developmental stages in the mouse, going from the differentiated to the pluripotent state by means of somatic cell reprogramming. For this study, I was responsible for processing all WGBS and RRBS data from four cell types representative of different stages of development, and conducted analysis to identify patterns of significant change in DNA methylation, and showed that changes predominantly occur late in reprogramming. Next, I integrated RNA-seq data to investigate the connection between DNA methylation and gene expression changes, and observed the demethylation of genes involved in cell division and stem cell maintenance and re-methylation of genes involved in differentiation and development. I also integrated ChIP-seq data for transcription factor binding sites to explore the connection between DNA methylation and factor binding. The binding sites of the four reprogramming factors showed different methylation patterns and motif dependence, and changes in DNA methylation were typically preceded by changes in histone modifications. Then, I went on to show that there was a strong association between DNA methylation and certain histone modifications more generally, and
built a multiple logistic regression model to predict DNA methylation based on the histone modification signals.

Chapter 2 is a version of another manuscript in preparation that introduces methylation-sensitive restriction enzyme bisulfite sequencing (MREBS) as an alternative HTS technique for assessing differential DNA methylation between two samples at cost comparable to RRBS, but providing additional differential methylation information over a coverage more comparable to that provided by WGBS. I analyzed all bisulfite HTS data sets used in the study and compared DNA methylation estimates across different chromatin states. I showed that differential DNA methylation values based on MREBS data correlate well with those based on WGBS and RRBS data at the CpG level. Conversely, differential MREBS read counts within 1kb windows around each CpG anti-correlate with WGBS and RRBS differential DNA methylation values. Seeing this, I built a multiple regression model that combined both MREBS differential DNA methylation values and MREBS differential read count data that predicted WGBS-based differential DNA methylation values more accurately than did a model based on RRBS data, but for a slightly smaller fraction of the genome (~1.5% vs. 3%). A big additional advantage of MREBS was that a model using MREBS differential read data alone to provided differential DNA estimates across close to 60% of the genome.

Chapter 3 is a version of a manuscript submitted for publication and currently under review, entitled “A high-throughput screen of inactive X chromosome reactivation identifies the enhancement of DNA demethylation by 5-aza-2’-dC upon inhibition of ribonucleotide reductase.” For the study, I processed the outputs of an siRNA and a chemical screen and
analyzed the data to identify hits. I also processed and analyzed associated bisulfite sequencing data, the results of which showed that inhibition of ribonucleotide reductase (RNR) decreased genome-wide DNA methylation levels by enhancing the incorporation of 5-aza-2’-dC into DNA. Myeloid leukemia cell lines treated with 5-aza-2’-dC and the RNR inhibitor hydroxyurea synergistically inhibited of cell growth and decreased genome-wide DNA methylation.

Chapter 4 is a version of a research article accepted for publication in *Plant Physiology* in March 2015, entitled “A Genome-wide chronological study of gene expression and two histone modifications, H3K4me3 and H3K9ac, during developmental leaf senescence”, a preview of which is already available. It describes a study that utilized ChIP-seq to determine the prevalence of said histone modifications and RNA-seq to assess gene expression during leaf aging in *Arabidopsis thaliana*. For this study, I used custom-built software to identify significant ChIP-seq signals, and, since it was a time course, I had to come up with computational methods to determine how these signals changed over time. Additionally, I integrated RNA-seq data with the ChIP-seq signals to explore the relationship between the histone modifications and gene expression. The results showed that many genes up-regulated during senescence were already weakly marked by the histone modification, but, for a subset of genes important to senescence, temporal changes in gene expression closely matched the changes in the histone marks, and, furthermore, that the extent of these histone marks over genes, as well as their intensity within the population of cells, was positively correlated with expression.

Chapter 5 is a reprint of a review that we were invited to write for *Current Opinion in Genetics & Development* and published in August 2014, entitled “A mechanistic link between
gene regulation and genome architecture in mammalian development”, which discussed cutting-edge research exploring that link between gene regulatory networks and 3D genome organization, and serves as a good introduction to the following chapter. 

Chapter 6 is reprint of a research article published in Cell Stem Cell in November 2013, entitled “Long-range chromatin contacts reveal a role for the pluripotency and Polycomb networks in genome organization”, which describes a study that employed 4C-seq, Hi-C, ChIP-seq, as well as RNA-seq data to investigate how gene regulatory networks affect, and are affected by, three-dimensional nuclear organization in mouse cells at different stages of development. I built an analysis pipeline to decode the 4C-seq data and identify DNA contacts made by specific genes. Subsequently, to better understand the relationship between genomic interactions and factors that regulate development and cell identity, I developed additional methods to examine changes in 3D interactions across different cell types, and to identify associations between genome organization and regulatory features. In order to do so, I had to develop a computational approach to integrate 30 linear genomic feature data sets, including TF binding and histone modification ChIP-seq data and RNA-seq expression values. The results allowed us to observe that interacting genomic regions often harbor similar complements of TFs and a similar chromatin state. More specifically, regions enriched for factors belonging to well-known gene regulatory networks showed a strong affinity for one another, suggesting that regulators responsible for cell identity contribute to cell type-specific genome organization.
References


Chapter 1

DNA methylation dynamics during somatic cell reprogramming to pluripotency
1.1 Abstract

Somatic cells can be reprogrammed back to a pluripotent state by the ectopic expression of four reprogramming factors. This change in cell identity is accompanied by changes to the epigenetic state of the cells, including to the methylation levels of cytosine residues along the DNA strands. Here, we set out to assess genome-wide DNA methylation levels in four cell types representative of different stages of somatic cell reprogramming, with a particular focus on reprogramming factor binding sites and chromatin state. We found that DNA methylation changes predominantly occurred late in reprogramming, and observed the demethylation of genes involved in cell division and stem cell maintenance and re-methylation of genes involved in differentiation and development. The binding sites of the four reprogramming factors showed different methylation patterns and motif dependence. Changes in DNA methylation were typically preceded by changes in histone modifications. This was seen more generally with essentially no significant changes in DNA methylation coinciding with transitions in chromatin state during the first 48 hours of reprogramming, but significant changes in DNA methylation later in the process. A strong association between DNA methylation and certain histone modifications was observed throughout the course of reprogramming, enabling CpG level methylation to be modeled using histone modification data.
1.2 Introduction

Somatic cell reprogramming is a method where terminally differentiated cells can be reprogrammed back into a pluripotent state by ectopic expression of four transcription factors known as the Yamanaka reprogramming factors, namely OCT4 (O), SOX2 (S), KLF4 (K), and MYC (M; also known as ‘c-MYC’). The resulting reprogrammed cells are termed induced-pluripotent stem cells (or iPSCs) and possess an embryonic stem cell-like character and morphology. However, the re-establishment of the full pluripotent state is slow, taking a couple of weeks, and occurs at a very low frequency, with less than 1% of the of starting differentiated cells successfully reaching the pluripotent state, implying that there are strong barriers in place within somatic cells preventing their return to a pluripotent state. There are, therefore, continuing efforts to enhance reprogramming kinetics and efficiency. Besides the goal of improving reprogramming efficiency and regenerating tissue for clinical use, somatic cell reprogramming an interesting system for studying the gene regulation and the epigenetics involved, and a powerful tool for the study of development and regulation of genes in general.

The underlying question being, ‘What governs cell identity?’

At the most fundamental level, cell identity is governed by transcription factors, which are proteins that bind the genome to turn a particular repertoire of genes on or off. The Yamanaka pluripotency factors are examples of such transcription factors, which help to turn on genes that maintain the pluripotent state in stem cells. They do so by binding to specific DNA loci along the genome, often in combination, such as the beginning of genes or at enhancer regions.
In addition to TF binding, there is information associated with the DNA beyond that encoded in the sequence itself, referred to as the epigenome, that helps to impart cell identity\textsuperscript{5,6}. DNA is packaged into the small confines of the nucleus in the form of highly compact chromatin, the core unit of which is the nucleosome\textsuperscript{7}. Each nucleosome core particle consists of a 147 bp stretch of DNA wound around an octamer of histone proteins (two sets of four proteins belonging to the H2A, H2B, H3 and H4 histone families)\textsuperscript{7}. H1 histones bind DNA at the entry and exit sites of the nucleosomes to fix it in place and compact the chromatin further\textsuperscript{7}. H3 and H4 histone family members possess long tails whose amino acid residues are subject to covalent modification. For instance, H3K4me3 refers to trimethylation of the fourth amino acid, which happens to be a lysine, and this particular modification is generally associated with regions of the genome that are transcriptional start sites\textsuperscript{6,8,9}. Different combinations of histone modifications impart different functional states to the chromatin in that region, and although the DNA within the nucleus all cell types is identical, their chromatin states will be specific to each type\textsuperscript{6,10}. In other words, histone modifications serve to annotate the regulatory state of different regions of an otherwise static genomic sequence, thereby marking genomes with a specific cell identity, in combination with TF binding\textsuperscript{6,10-12}.

DNA methylation is another layer of epigenomic information that is a modification to the nucleotide sequence itself, more specifically to the fifth position of cytosines\textsuperscript{13-15}. De novo DNA methyltransferases 3a and 3b (DNMT3A and DNMT3B) establish patterns of 5-methylcytosine (5mC) during development, while a methylation maintenance enzyme (DNMT1) is required to preserve such patterns of methylation through DNA replication cycles\textsuperscript{13,14}. Within mammals, methylation of cytosines occurs predominantly in the CpG context and symmetrical, with 60-
80% of all CpGs methylated\textsuperscript{13,14}. Methylated DNA is usually associated with repression of gene activity, with demethylated regions being associated with TF binding and gene activation, often in CpG dense regions such as CpG islands\textsuperscript{6,13,14}. A strong link between DNA methylation and histone modifications is also known to exist\textsuperscript{16}.

Returning to somatic cell reprogramming, some cells are only partially reprogrammed. Late reprogramming intermediates, termed pre-iPSCs, reach a stage just prior to full pluripotency\textsuperscript{1}. These clonal cells are self-renewing, have largely changed morphology and make colonies more typical of embryonic stem cells and iPSCs. However they still fail to express many endogenous pluripotency genes, including \textit{Pou5f1}, which encodes OCT4, and \textit{Sox2}. The emergence of intermediate cell states such as pre-iPSCs reinforces the idea of there being strong barriers in place to prevent a return to a pluripotent state. Nevertheless, pre-iPSCs can be converted to iPSCs with additional treatments that enhance the late phase of reprogramming\textsuperscript{1}. For example, inhibitors of DNA methylation, such as 5-azacytidine, or the knockdown of the DNA methylation maintenance enzyme (DNMT1), promote the conversion of pre-iPSCs to the pluripotent state\textsuperscript{1,17}. In fact, these treatments enhance reprogramming in general, which implicates DNA methylation as a potential barrier to reprogramming.

This notion has been reinforced by the fact that DNA methylation in iPSCs is reset to an ESC-like state\textsuperscript{2,18,19}. For instance, promoters of pluripotency genes are hypomethylated in the pluripotent stage relative to the differentiated state, and this demethylation was shown to be a late-stage process\textsuperscript{2,20,21}. However, for other regions of the genome, beyond promoters, changes to DNA methylation dynamics are unclear.
In this study, we were interested in assessing DNA methylation levels during the course of somatic cell reprogramming to pluripotency within a mouse model system. More specifically, we aimed to address the following main questions: First, at what stage(s) of reprogramming do DNA methylation changes take place? Second, what is the relationship between TF binding and DNA methylation levels, if any? Third, more broadly, what are the chromatin characteristics in the regions of genome that see DNA methylation change? Finally, might it be possible to predict whole-genome DNA methylation patterns based on chromatin state?

We saw that DNA methylation profiles diverge during the course of reprogramming. Genes whose promoters lost DNA methylation showed increased expression and were enriched in gene sets related to cell division and stem cell maintenance. Conversely, genes whose promoters gained methylation during reprogramming showed a significant reduction in expression and were enriched for differentiation and development.

We found that KLF4 binding occurs at enhancer-like states, while MYC prefers to bind open active promoter sites that show little to no DNA methylation at any stage of reprogramming and regardless of the absence or presence of a motif. OCT4 and SOX2 binding sites, on the other hand, show greater stage specificity and motif dependence, often binding in tandem. Early-specific and constitutive O/S binding sites with O/S motifs were found to occur exclusively at more methylated, enhancer-like sites, whereas ESC-specific O/S binding sites occurred exclusively in enhancer-like regions regardless of the absence or presence of motifs. Interestingly, changes in DNA methylation around factor binding sites appeared to follow changes to chromatin state.

Looking beyond factor binding sites at DNA methylation within different categories of chromatin state, significantly differentially methylated regions going to the pluripotent state were
enriched within regions of the genome that possessed an enhancer-like chromatin state character in. There was no average change in DNA methylation in regions undergoing chromatin state transitions during the first 48 hours of reprogramming. However, regions transitioning from EARLY intermediates to active chromatin states in ESCs showed a marked decrease in DNA methylation, while those transitioning away from active states showed an increase in DNA methylation.

Bisulfite sequencing-based CpG methylation estimates could be reasonably well modeled by histone modification using multiple logistic regression and only four key histone modifications were sufficient to produce a near-optimal fit.
1.3 Results

1.3.1 Study design and experimental approach

To assess global DNA methylation during the course of somatic cell reprogramming, we used whole-genome bisulfite sequencing (WGBS, also referred to as BS-seq, methyl-seq, or methylC-seq), which couples bisulfite treatment of DNA to high-throughput DNA sequencing\textsuperscript{22-24}. We used this approach to estimate the DNA methylation levels in four cell types that represented different stages during the somatic cell reprogramming process: 1) Un-induced mouse embryonic fibroblasts (or MEFs) that harbor an inducible OSKM ‘stem cell cassette’\textsuperscript{25} at the Col1A1 locus that allows for the simultaneous induction of the four reprogramming factors upon the addition of doxycycline, as starting cell type. 2) Early reprogramming intermediates were represented by MEFs that had been induced with doxycycline for 48 hours (EARLY). 3) A late reprogramming intermediate represented by the partially induced pluripotent cells that are typically obtained by retroviral expression of OSKM (pre-iPSCs, LATE). 4) Mouse embryonic stem cells (ESCs) were used to represent the fully reprogrammed state, since they are more readily available and abundant than iPSCs, but essentially equivalent (Figure 1.1A).

WGBS libraries for each of these four cell types were generated by Kostas Chronis, a post-doctoral follow in Professor Plath’s laboratory, who produced all libraries mentioned in this chapter. Libraries were sequenced deeply enough to ensure an average read coverage of 3.5–7.8X per CpG dimer after mapping to the mm9 genome using BS-Seeker\textsuperscript{26} (Figure 1.1B, Table 1.1). The LATE intermediates were less deeply sequenced than the other lines and therefore exhibited somewhat lower CpG coverage levels (Figure 1.2Biii).
In an effort to replicate the WGBS libraries for a subset of the genome, reduced representation bisulfite sequencing (RRBS) libraries were also produced. By introducing a step where the genomic DNA is first digested using a methylation insensitive restriction endonuclease (typically MspI, with the recognition sequence C\textasciitildeCGG), RRBS interrogates of a smaller, CpG rich portion of the genome (5-10%) that arguably includes most regions that are of regulatory interest, such as promoters and enhancers, while significantly reducing the amount of sequencing required\textsuperscript{27,28}.

RRBS reads were mapped to an in silico MspI-digested reduced mm9 reference genome. As expected the proportion of CpG dimers with 5X coverage (4.1–6%) was lower than that seen for WGBS libraries (44.3–80.6%), and the mean methylation levels across these dimers (0.46–0.49) was lower than that seen for the WGBS samples (0.67–0.72) due to the enrichment for CpG-rich regions, which are typically less methylated (Table 1.1).

1.3.2 DNA methylation profiles diverge from that seen in the somatic state during the course of reprogramming

To begin to address the spatiotemporal question we had regarding DNA methylation changes during somatic cell reprogramming, we initially compared the distributions and profiles of the genome-wide CpG-level DNA methylation estimates based on WGBS data. The distributions of bulk DNA methylation were comparable, albeit the distribution of the LATE intermediates was somewhat more bimodal than that of the other cell lines and their mean DNA methylation level was slightly lower (0.67) than that of the other three lines (0.70–0.72) (Figure 1.1C). This could be a consequence of the lower sequencing depth in the LATE intermediates (Figure 1.1Biii).
RRBS DNA methylation distributions were comparable but reflected the bias for typically lower methylation in CpG rich regions (data not shown).

The correlation values between the MEF DNA methylation profile and those of the other three cell types representing the sequential stages of somatic cell reprogramming (EARLY, LATE and ESCs) showed a corresponding sequential drop in value (Figure 1.1C). The DNA methylation profile 48 hours into the reprogramming process was still closely correlated to that of the fibroblasts ($r = 0.78$), with the biggest difference in correlation being between the somatic starting cells and the resulting pluripotent state (ESCs) ($r = 0.51$). However, the LATE intermediates were still some way from attaining the DNA methylation pattern seen in the pluripotent state ($r = 0.57$; Figure 1.1C). This relationship can also be seen in the heat map and hierarchical clustering of the pairwise correlations between WGBS DNA methylation profiles based on CpG dimers, with tight clustering between the MEFs and EARLY intermediates away from the LATE intermediates and ESCs (Figure 1.1Di), and is recapitulated by the RRBS samples (Figure 1.1Dii).

When looked at altogether, the pairwise correlations of the DNA methylation profiles of the WGBS and RRBS samples (based on CpG dimers with sufficient coverage between each pair of samples) tend to cluster by cell type (Figure 1.1Dii). However, the MEF and EARLY DNA samples first cluster by method, although, as a group, they cluster together and away from the LATE and ESC samples (Figure 1.1Diii). Presumably the MEF and EARLY DNA methylation profiles are so similar that differences due to protocol are more influential. For the LATE and ESC samples, on the other hand, cell type-specific differences predominate, even though the
correlations between RRBS-based DNA methylation profiles are generally stronger than that seen based on the WGBS data (Figure 1.Diii).

These results confirmed that DNA methylation changes do indeed occur during the course of somatic cell reprogramming, with DNA methylation profiles diverging progressively and the biggest changes taking place later in the process.

1.3.3 Identifying and clustering differentially methylated regions of the genome during somatic cell reprogramming

Initial scanning of the WGBS DNA methylation data in a genome browser revealed that differentially methylated regions (DMRs) did indeed exist between samples (Figure 1.2A). For example, Olig3 is demethylated in the MEF sample, and in fact still demethylated 48 hours into somatic cell reprogramming. However, by the LATE stage, as well as in ESCs, this locus is methylated again. Interestingly, Olig3 is not expressed at any stage of reprogramming, suggesting that this might perhaps be an example of global indiscriminate methylation. Conversely, Essrb, Dppa2 and Dppa4 are loci that all exhibit demethylation exclusively in ESCs. These genes are also specifically expressed in the pluripotent state. Runx1, on the other hand shows no change and remains demethylated throughout reprogramming even though it is expressed in fibroblasts and still somewhat in the LATE intermediates, but not in pluripotency. These DNA methylation patterns were visible at both CpG-level (Figure 1.2Ai), as well as when DNA methylation was averaged within 500bp windows tiled along the mouse genome (Figure 1.2Aii).
To systematically determine DMRs between sample pairs based on the WGBS DNA methylation estimates, I used the mean DNA methylation levels within 500bp windows tiled across the genome. A 500bp window was considered to be significantly and consistently differentially methylated, if it showed an absolute difference in DNA methylation of at least 50% in 3/6 of the pairwise comparisons. Based on this criteria, 120,744 of the approximately 5 million 500bp windows genome-wide were identified as DMRs, representing ~2.3%, or ~60Mb, of the mouse genome. A lower differential DNA methylation threshold of 25% resulted in 826,784 500 bp windows (~15.6%) being identified as DMRs, and using a very stringent 75% threshold led to only 7,594 DMRs (~0.14%).

The 120,744 high-confidence differentially methylated windows called using the 50% threshold were then partitioned into six groups by k-means clustering in order to discern the predominant patterns of DNA methylation change during the course of somatic cell reprogramming (Figure 1.2B). For easier visualization, the average DNA methylation level across all the windows in each group was calculated for each cell type and used in the heat map. Hierarchical clustering was used to highlight groups that showed similar patterns of change (Figure 1.2B).

The first two columns of heat map (MEFs and EARLY) show that there is little in the way of large-scale changes in DNA methylation during the first 48 hours of reprogramming. However, interesting patterns of change emerge later in the process. For example, cluster groups 1 and 2 show very similar patterns and capture those regions of the genome where the LATE intermediates are demethylated relative to the other samples, as plots of the mean DNA methylation within a 20kb region centered around the DMRs in these cluster groups clearly show
(Figure 1.2Ci/ii). A plot of mean DNA methylation levels around 20,000 randomly-selected 500bp windows shows that the clusters do indeed capture sets of DMRs with a specific pattern of change, while confirming that the LATE intermediate sample shows a slightly lower mean DNA methylation than the other cell lines (Figure 1.2Cvii).

Cluster group 3 also sees the LATE intermediates lose DNA methylation relative to the starting cells and 48-hour intermediates, but in this case this demethylated state carries over to the fully pluripotent state (Figure 1.2B/Ciii).

Group 4 captures those regions that are specifically demethylated in ESCs, which might represent those barriers to somatic cell reprogramming where the LATE intermediates still need to lose DNA methylation to attain the fully reprogrammed state (Figure 1.2B/Civ). Notably, sites of demethylation appear to occur within broad regions of demethylation (Figure 1.2Civ), which could perhaps represent DNA methylation valleys.

Unlike groups 1–4, where the MEF and EARLY samples are hypermethylated relative to either or both the LATE and ESC samples, the windows in groups 5 and 6 are hypomethylated early in somatic cell reprogramming relative to the later stages. Group 5 also captures some regions showing specific hypermethylation in the LATE intermediates (Figure 1.2B/Cv). Group 6 may represent another set of genomic loci that represent barriers to full reprogramming, where the LATE intermediates have not attained the higher DNA methylation levels observed in the ESCs (Figure 1.2B/Cv).

Similar clustering trends were observed using differentially methylated windows using the lower (25%) and higher (75%) thresholds (data not shown). However, there were many more windows per group using the less stringent criteria, and many fewer windows using the more stringent criteria.
In summary, during the course of reprogramming from the somatic to pluripotent state, we identified ~2.3% of the genome to exhibit differential DNA methylation that was significant (>50%) and consistent (3/6 pairwise comparisons). Although there was little evidence for significant differential DNA methylation between the MEFs and the 48 hours intermediates, many DMRs were observed later in the process, which we grouped into distinct patterns of change including specific hypo- and hypermethylation in both the LATE intermediates and the fully pluripotent state, which are regions of the genome that could represent barriers to reprogramming.

1.3.4 DNA demethylation during reprogramming is associated with increased expression by genes related to the pluripotent state while DNA re-methylation is associated with reduced expression of developmental and somatic genes

To investigate whether there was an association between the observed differential DNA methylation patterns and gene expression, I compared the RNA-seq expression profiles of genes whose promoters (transcription start site (TSS) +/- 2,500bp) overlapped one or more of the differentially methylated windows in each of the six clusters in Figure 1.2B. If a gene was associated with more than one cluster group, it was assigned to the group with the most associated windows. If a gene was covered equally by DMRs from more than one cluster group, it was assigned to all the groups in question. Although this implies that there was the potential to associate the same gene with multiple cluster groups, this was uncommon, with the mean overlap percentage between pairs of cluster groups being 4.5% of genes.
Although cluster groups 1 and 2 accounted for the majority of the differentially methylated windows (67,246 of 120,744, or 55.7%) at the 50% threshold level (Figure 1.2B), they were associated with relatively few genes (835 and 832 genes, respectively). These genes were relatively lowly expressed, especially the group 2-associated genes, and did not show any significant difference in expression between the four cell types (Figure 1.2Di/ii, Table 1.2).

The 1,289 genes associated with the cluster group 3 LATE and ESC-specific hypomethylated windows were far more highly expressed across the cell types than seen for the group 1 and 2-associated genes, and their expression in the LATE intermediates and ESCs was significantly higher (Figure 1.2Diii, Table 1.2). In other words, for this cluster of differentially methylated windows, expression differences anti-correlated with the DNA methylation differences. Of these genes, 752 (58.3%) showed differential expression (DE) between the cell types (minimum 2X fold change between the minimum and maximum expression value across the four cell types, and the maximum expression had to be greater than the median expression level across cell types (1.02 RPKM)). 499 of these DE genes were in turn associated mouse gene identifiers in the biological pathways category of the DAVID gene ontology (GO) database. These genes were only enriched in a small set of GO terms to do with cell cycling and division, and RNA processing (Table 1.3).

The 2,098 genes associated with ESC-specific hypomethylation belonging to cluster group 4 were very significantly more expressed in ESCs (Figure 1.2Div, Table 1.2). Although the LATE intermediate DNA methylation profiles of the windows in this group are almost exactly like that of the MEFs and EARLY intermediates (Figure 1.2B/Civ), the associated genes also showed increased expression in the LATE intermediates, but not nearly to the same level of significance as seen in for the ESCs (Figure 1.2Div, Table 1.2). Of these 2,098 genes, 1,296
(61.8%) were differentially expressed between the cell types, using the same criteria as described above. These DE genes were significantly enriched for biological pathway GO terms associated with stem cell maintenance, chromatin organization, negative regulation of biosynthetic processes, and, once again, cell division (Table 1.4).

Conversely, those 1,070 genes associated with MEF and EARLY intermediate-specific hypomethylation seen in cluster group 5, show significantly higher expression levels specifically in the MEFs and EARLY intermediates (Figure 1.2Dv, Table 1.2). 639 of these genes (59.7%) were differentially expressed between the cell types (using the same criteria as described above). Associated with 442 mouse gene identifiers in DAVID in the biological pathway gene set, these genes were enriched in GO terms associated with differentiation and developmental morphogenesis (Table 1.5).

As with groups 1 and 2, the 12,506 differentially methylated windows belonging to cluster group 6 are associated with relatively few genes (413) and these are very lowly expressed across the four cell types with little difference in their distributions (Figure 1.2Dvi, Table 1.2),

Struck by the fact that the differentially methylated windows in groups 1, 2 and 6 were associated with surprisingly few gene promoters, we were interested in seeing where these DMRs fell in the genome, and whether they might instead be associated with other regulatory features, such as enhancers. Also, even though the differentially methylated windows belonging to groups 3, 4 and 5 were associated with a higher number of gene promoters, this did not preclude the possibility that they might also be associated with other cis-regulatory elements. To see where they fell in relation to gene start sites, the proportion of differentially methylated windows belonging to each cluster group found 0–5 kb, 5–50 kb, 50–500 kb, or beyond the
transcription start sites (TSSs) of UCSC known genes (either upstream or downstream) was determined using GREAT\textsuperscript{32} (Figure 1.2E).

Of note, a higher proportion of differentially methylated windows belonging to cluster groups 1, 2, and 6 fell in beyond 50kb away from gene start sites, while for cluster groups 3, 4, and 5 most windows fell closer to TSSs within 50kb of genes (Figure 1.2E), confirming the promoter-based analysis. Furthermore, the GO terms associated with the DMRs of groups 1, 2 and 6 were very few in number and not readily interpretable (data not shown). In contrast, the differentially methylated windows in groups 3, 4, and 5 were associated with many GO terms that corresponded to biological pathways seen for DMRs found in the promoters of differentially expressed genes. For instance: groups 1 DMRs were largely associated with biological pathway terms related to with cell division as before, but also with chromatin organization, and, intriguingly, genetic imprinting; group 4 DMRs were associated with stem cell maintenance, chromatin organization, negative regulation of biosynthetic processes, and cell division, as before; and Group 5 DMRs were once again associated with differentiation and development (data not shown).

In summary, only the 3 of the 6 DMR cluster groups identified were associated with the promoters of genes that exhibited a significant change in expression during the course of reprogramming. Groups 3 and 4, represent regions that lost DNA methylation in the LATE and ESC-specific states, respectively. These regions were associated with genes that showed increased expression and enriched for GO terms of with cell division (group 3) and stem cell maintenance (group 4). On the other hand, genomic regions that gained methylation during
reprogramming (group 5) were associated with down-regulated genes involved in differentiation and developmental.

1.3.5 DNA methylation at reprogramming factor binding sites and its association with histone modifications

Results from a sister study in the Plath laboratory had shown that binding patterns of the four reprogramming factors (O, S, K, and M) change during somatic cell reprogramming. Early in reprogramming, the factors tend to bind somatic sites that are subsequently disengaged later in the process. The majority of ESC cell-specific sites, on the other hand, are engaged only very late in reprogramming. With this in mind, we sought to examine how this genome-wide factor re-organization might be related to DNA methylation, by looking at DNA methylation at and around these sites based on ChIP-seq data sets generated by Kostas Chronis. We first categorized reprogramming factor binding sites based on whether they showed specificity for a particular somatic cell reprogramming stage or stages, or whether their binding was constitutive (Figure 1.3A).

Furthermore, we were also interested in examining whether DNA methylation at the sites of binding was related to the absence of presence of DNA binding motifs. Binding sites were therefore further partitioned into those that contained the reprogramming factor DNA binding motifs within 100bp of the summit and those that did not. The motifs for five factors were considered, namely those for the binding sites of the four reprogramming factors, as well as that for NANOG, regarded as a marker of faithful reprogramming\(^1\). The O, S, K, and M peaks were scanned for the presence of the motif within 100bp of the peak summit using HOMER\(^{33}\).
Since we also had ChIP-seq data for nine histone modifications (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K27me3, H3K36me3, H3K79me2, H3K9me3) and one histone variant (H3.3) across the four cell types under study, we also looked at the enrichment for these marks around the binding sites. These histone marks were selected on the basis of their association with promoters (H3K4me3 and H3K9ac)\(^8,9,34\), enhancers (H3K4me2, H3K4me1 and H3K27ac)\(^8,34-36\), transcription initiation and elongation (H3K79me2, H3K36me3)\(^34,37\), and repression (H3K27me3 and H3K9me3) (Boyer et al., 2006)\(^34,38\). H3.3 is a histone 3 variant that is associated with enhancer activation\(^39\).

We first looked at reprogramming factor binding that was specific to the EARLY stage of reprogramming, which we referred to as the ‘100’ binding pattern for each of the four reprogramming factors (O, S, K, and M) (Figure 1.3A). Looking at mean DNA methylation and histone mark enrichment within 1kb windows around the summits of the partitioned binding sites, we noticed was that DNA methylation and histone mark enrichment values were very similar around OCT4 and SOX2 peaks, which comprised of approximately 60 and 40 thousand peaks, respectively. However, the epigenetic state around KLF4 and MYC peaks was more unique (Figure 1.3B). The DNA around O and S binding sites that were exclusive to the EARLY state had DNA methylation levels that were lower than the average level seen in the population, showing a mean DNA methylation level of 40–50% (Figure 1.3B/Ci). Interestingly, those EARLY-specific O/S sites harboring O, S, or K motifs were 5–10% more methylated on average (Figure 1.3B/Ci). These sites showed similar DNA methylation levels in the pluripotent state, so they appear to maintain their DNA methylation status during the course of somatic cell reprogramming despite the loss of reprogramming factor binding (Figure 1.3B/Ci). The LATE
intermediates showed a 5–10% decrease in DNA methylation at these sites, which is reflective of the globally hypomethylation exhibited by this state compared to the other lines (Figure 1.1C).

Histone modification enrichment values around ‘100’ O and S peaks suggested that these binding sites are found in regions of active, open chromatin, since these sites are both constitutively enriched for H3K4me3 and H3K9ac (marks associated with active promoter regions\(^6\), as well as for H3K4me1, H3K4me2 and H3K27ac (modifications associated with active enhancers\(^6\)) (Figure 1.3B/Ci). Notably, enrichment for H3K4me1 and H3K27ac, is lost during the course of somatic cell reprogramming, suggesting that these are somatic enhancers. Importantly, O/S ‘100’ sites with motifs occur exclusively at more methylated, enhancer-like sites, with these sites showing absolutely no enrichment in H3K4me3 at any time during the reprogramming process (Figure 1.3B/Ci).

Similarly, all ‘100’ KLF4 binding sites, regardless of motif, are exclusively enriched for H3K4me1 and H3K27ac, but not at all for H3K4me3 or H3K9ac, again suggesting a distinct preference for enhancer-like binding sites (Figure 1.3B/Cii). In the same manner to ‘100’ O/S, these sites lose their H3K4me1/H3K27ac enrichment and the DNA methylation does not change in the ESCs, suggesting that these are somatic enhancers.

EARLY-exclusive MYC peaks are fewer in number than for the other three factors while the DNA binding sites are far more demethylated throughout reprogramming. Such MYC binding occurs predominantly in regions enriched for H3Kme3 and H3K9ac, which represent active promoters. Some moderate EARLY H3K4me1 enrichment is also seen but is lost in the LATE and ESC states indicating that any EARLY binding to enhancers regions is transient (Figure 1.3B/Ciii).
We next considered those peaks specific to the pluripotency state, which referred to hereon as ‘001’ peaks (Figure 1.4A). There were significantly fewer ESC-specific OCT4 and SOX2 peaks compared to the EARLY-specific number. It was clear that in the ESC-specific context, O/S DNA binding occurred at sites that become specifically demethylated late in the somatic cell reprogramming process (Figure 1.4B/Ci). Furthermore, this O/S binding was found to occur predominantly at enhancer regions since these sites very highly enriched for H3K4me1 and H3K27ac, but not at all for H3K4me3 or H3K9ac (Figure 1.4B/Ci). Tellingly, enrichment for H3K4me1 appears to precede DNA demethylation, and is already apparent in the LATE intermediates, suggesting that these sites are being primed for activation. Activation, itself, appears to be a LATE stage event based on the fact that enrichment of H3.3, a histone variant associated with enhancer activation\textsuperscript{39}, only becomes prominent in the ESCs (Figure 1.4B). It is also noteworthy that these pluripotency-specific O/S epigenetic profiles are independent of the absence or presence of binding motifs (Figure 1.4B/Ci).

‘001’ KLF4 binding sites show somewhat similar characteristics to those seen for the O and S sites, except that H3K4me1 enrichment at these sites in the pre-iPSCs only occurred when O or S motifs were present (Figure 1.4B/Cii).

Very few ESC-specific MYC binding sites were observed (844), and these occurred at regions showing very broad and very low DNA methylation levels, and a very strong enrichment for H3K4me3, as well as H3K79me2, indicative of active gene promoter regions (Figure 1.4B/Ciii). Intriguingly, a small number of ‘001’ M peaks (37) possessing the OCT4 DNA binding motif showed very late chromatin activation (Figure 1.4B/Ciii).
Apart from EARLY and LATE-specific factor binding, another category of binding pattern of interest was constitutive binding during all three stages of somatic cell reprogramming, which we designated ‘111’ (Figure 1.5A). This category comprised of far fewer peaks than the two previously mentioned categories, but still significant enough to warrant investigation.

DNA methylation levels steadily dropped at sites that were constitutively bound by OCT4, SOX2, or KLF4 during the course of reprogramming (Figure 1.5B/Ci/ii). Furthermore, those constitutive O/S/K sites with O/S/K motifs were 5-10% more methylated on average 48 hours into reprogramming (Figure 1.5B/Ci/Cii).

Constitutive MYC binding sites showed much lower DNA methylation levels than observed for the constitutive sites of the other three factors, and very little H3K4me1 enrichment, other than at sites with O, S, or K motifs (Figure 1.5B/Ciii) consistent with these sites being at promoters.

In summary, from this fine-grained analysis of reprogramming factor binding sites, we see that EARLY-specific and constitutive OCT4 and SOX2 binding sites with O/S motifs occur exclusively at more methylated, enhancer-like sites, whereas ESC-specific O/S binding sites occur exclusively in enhancer-like regions regardless of the absence or presence of motifs. KLF4 binding occurs at enhancer-like states, regardless of motif, at every stage of somatic cell reprogramming. Constitutive O, S and K DNA binding sites appear to gradually lose DNA methylation over the course of reprogramming, but, importantly, these changes are preceded by changes to chromatin state, something that appears to be generally the case. MYC binding sites show little DNA methylation at any stage of reprogramming and stage specificity.
1.3.6 DNA methylation at sites of reprogramming factor co-binding and its association with histone modifications

Up to this point, we had only looked at the binding sites of each individual reprogramming factor and their timing with respect to cell type, but had not yet considered their propensity to co-localize\(^1\). In order to take their combinatorial binding nature into account, we used k-means clustering to find 500bp regions where the summits of the pluripotency factors peaks co-occurred during each of the four stages of reprogramming under study. Specifically, we clustered the OCT4, SOX2, KLF4, and MYC binding peaks from the three stages following induction by doxycycline addition, as well KLF4 and MYC peaks that were observed in MEFs prior to doxycycline-induced up-regulation, to produce 20 transcription factor (TF) clusters representing co-binding events during somatic cell reprogramming (Figure 1.6A). Co-occurrence frequencies were used to hierarchically cluster these TF clusters in turn, to reveal similarity in their binding patterns, as represented by the dendrogram to the left of heat map. This secondary clustering motivated the assignment of the TF clusters to one of five groups (A–E; right of heat map).

The four TF clusters belonging to group A (1–4) all show OSK co-binding in the EARLY intermediates at sites where K was already found to be bound in MEFs (Figure 1.6A/Bi, first column). In general, Group A binding appears to be associated with active promoter in MEFs that largely remain active through the course of reprogramming, since the binding sites for all TF clusters in this group are strongly enriched for H3Kme2/3, H3K9ac, H3K27ac and H3K79me2.

However, Group A can be further partitioned into two subgroups based on whether or not M binding is present 48 hours into reprogramming (Figure 1.6A/Bi, second column). Strikingly, TF clusters 3 and 4 are the only two clusters that show any significant M binding in the EARLY
intermediates, LATE intermediates, or constitutively, and are also unique in exhibiting extremely low DNA methylation levels. This coincides with relatively low H3K4me1, but strong H3K4me2/3, H3K9ac, H3K27ac, H3K79me2, and H3.3 enrichment, suggesting that these M-inclusive co-binding events occur at largely hypomethylated, active gene promoter regions, as already observed for constitutive M-binding alone (Figure 1.5). However, there are differences between the two members of this subgroup: TF cluster 3, which exhibits K binding throughout, showed strong OSKM binding in ESCs and constitutive H3K4me1 enrichment at those few sites with O, S, K, or NANOG motifs present. TF cluster 4, on the other hand, has little K EARLY and shows little to no OSK co-binding in ESCs, and apart from 33 windows possessing the O motif, DNA methylation and H3K4me1 under-enrichment is consistent (Figure 1.6A/Bi).

Those group A members without the presence of M-binding (TF clusters 1 and 2), show moderate/low DNA demethylation levels in MEFs (but not to the same extend seen in TF clusters 3 and 4), with further demethylation during the course of somatic cell reprogramming (Figure 1.6A/Bi). TF clusters 1 and 2 show H3K4me1 enrichment, in addition to enrichment for active transcriptional marks, and somewhat lower H3K79me2 enrichment, suggesting binding to active enhancer as well as active promoter regions. For TF cluster 1, DNA methylation is lowest in the LATE intermediates and ESCs. For TF cluster 2, DNA methylation is lowest in ESCs and H3K4me1 enrichment is strong throughout. Both TF cluster 1 and 2 show strong H3K4me2/3, H3K9ac, and H3K27ac enrichment, except for windows harboring the O motif where H3K4me3 and H3K9ac enrichment is lower, suggesting enhancer specific binding at these sites.

The members of Group B (TF cluster 5, 6, and 7) also show EARLY OSK co-binding, with OS binding persisting in the LATE intermediates (Figure 1.6A/Bii). All three TF clusters in
the group show lower H3K4me3, H3K9ac, and H3K79me2 enrichment than seen for members of Group A, and more specific H3K4me1/2 and H3K27ac enrichment suggesting that the binding sites of this group are more enhancer specific.

Although all three clusters have O binding in ESCs, only TF cluster 7 has OSK binding in the pluripotent state, and this coincides with a pattern of DNA demethylation and H3K27me3 loss during reprogramming and concomitant H3K4me1/2, H3K27ac, and H3.3 enrichment that is not seen for TF clusters 5 and 6, indicating that these regions of co-binding correspond to ESC-specific enhancer sites that seem to be primed for activation in the pluripotent state already in the LATE intermediates. Notably, it appears that histone marks suggesting enhancer activation precede DNA demethylation, as seen previously for ‘001’ O and S binding (Figure 1.4).

TF clusters 5 and 6, show strong particularly strong OS binding in the LATE intermediates and relatively weak O and OS binding in ESCs, respectively. Their binding also appears to be largely enhancer-specific. Consistent with the strong OS binding in the LATE state, they show lowest DNA methylation levels and strongest H3K4me1/2 and K27ac enrichment in LATE intermediates, and much lower H3.3 enrichment in ESCs, suggesting that these co-bound regions are not activated in the pluripotent state (Figure 1.6A/Bii).

TF cluster group C members (8–11) show little to OSKM binding 48 hours into somatic cell reprogramming, with some K and M binding in the LATE intermediates, and no binding in ESCs (Figure 1.6A). In accordance with this, none of the TF clusters in this group show much evidence of an active state in ESCs (Figure 1.7i). As with Group A, all members of this group show H3K4me3, H3K9ac, and H3K79me2 enrichment, suggesting binding at active promoters.
However, this group also shows H3K4me1/2 and H3K27ac enrichment, which indicates binding at enhancers. Only TF cluster 10 exhibits a significant drop in DNA methylation during reprogramming specifically in LATE intermediates and in the pluripotent state. This demethylation corresponds to KLF4 and MYC LATE-specific binding (Figure 1.6Bii).

The members of TF cluster group D (12–15), which only see factor binding in the LATE intermediates and/or ESCs, in contrast to the group C members just discussed, appear to be very enhancer-specific as they show little H3K4me3, H3K9ac, or H3K79me2 enrichment.

Importantly, all members of the group show very pluripotency-specific DNA demethylation and chromatin state activation, with very strong H3.3 enrichment in ESCs at enhancer regions (Figure 1.7ii). TF binding in this group is correspondingly ESC-specific, and these regions of the genome presumably coincide with those of the ‘001’ binding sites discussed above (Figures 1.6A and 1.4). TF cluster 12, which shows strong K binding in the LATE intermediates and ESCs, appears to progressively lose DNA methylation and gain H3K4me1/2, H3K27ac, and H3.3 enrichment during reprogramming, while TF cluster 13, which only sees strong K binding specifically in ESCs, sees these changes occur very specifically in ESCs (Figures 1.6A and 1.7ii). TF clusters 14 and 15 only see factor binding in the pluripotent state, and, even though H3K4me1/2 is already enriched in the LATE intermediates, H3.3 enrichment is ESC-specific (Figures 1.6A and 1.7ii).

In all cases, it appears that H3K4me1/2 and H3K27ac enrichment precedes the fullest extent of DNA demethylation, with H3.3 enrichment closely coinciding with DNA methylation loss, as previously seen with TF cluster 7 and ‘001’ O/S binding (Figures 1.6 and 1.4).
Group E (TF clusters 16–20) comprises all genomic regions that see EARLY-specific factor binding, corresponding to ‘100’ regions for the individual factors discussed previously, with TF clusters 16 and 17 appearing to be more enhancer-specific than the three remaining TF clusters (Figures 1.6A and 1.3). These regions exhibit largely stable DNA methylation values, but particularly notable H3K4me1/2, H3K27ac and H3.3 under-enrichment in ESCs (Figure 1.7iii). This is indicative of the fact that early in reprogramming, the three factors O, S and K can bind enhancers irrespective of the moderate/high levels of DNA methylation present and exert transient effects.

The TF cluster-based analysis presented in this section reveals that regions showing similar patterns of factor binding during somatic cell reprogramming exhibit similar patterns of DNA methylation and histone modification enrichment. Studying the epigenetic state around sites where the reprogramming factors do and do not co-bind, not only confirmed patterns seen in the previous section for the binding sites of the individual factors, but provided additional insight into how these factors influence DNA methylation and chromatin state in a combinatorial fashion.

1.3.7 DNA methylation distributions in different chromatin states

Moving beyond factor binding sites, next we investigated whether there was an association between DNA methylation and histone modifications throughout the genome regardless of factor binding. To do so, we took advantage of a hidden Markov model of chromatin states generated by chromHMM. The model will be described in detail in a separate manuscript (Chronis et al., in preparation), but briefly, it was based on the same ChIP-seq data sets for the nine histone
modifications and one variant introduced earlier in this chapter, as well as a native input library. Using this model, the genomes of each of four cell types for which we had genome-wide DNA methylation data (MEFs, EARLY intermediates, LATE intermediates, and ESCs) were tiled into 200bp windows and classified as belonging to one of 18 chromatin states. Candidate functional annotations were assigned to each of the 18 states based on the prevalence and combination of histone mark present in each state (Figure 1.8A).

We were initially interested in investigating whether DNA methylation levels were different in each of the different chromatin states, and, if so, to what extent. To investigate this, for each of the four cell types, I calculated the mean DNA methylation within the 13.3 million 200bp windows corresponding to those used for the chromatin state model. Only those windows containing at least one CpG with 5X coverage was assigned a value in an effort to ensure high-confidence estimates. I then plotted the distribution of WGBS-based DNA methylation estimates for windows belonging to each chromatin state (Figure 1.8B). The different chromatin states showed characteristic DNA methylation distributions that were similar across the four cell types (Figure 1.8B). For instance, the promoter-associated chromatin states (1 and 2) were comparatively hypomethylated, while several enhancer-related chromatin states (3, 4, 5, and 7) showed a wide spread of DNA methylation levels and an intermediate mean DNA methylation. Most of the other chromatin states showed higher DNA methylation levels (Figure 1.8B).

There were also some cell type-specific differences. Windows belonging to chromatin state 2 (poised promoters) looked more methylated in ESCs on average, while chromatin states 3, 4, 8 and 9 (enhancer-like regions of the genome) looked less methylated in ESCs (Figure 1.8B).
At the CpG-level, the LATE intermediates showed a lower average DNA methylation compared to the other cell types (Figure 1.1C) and, at 500bp resolution, many regions showed specific demethylation (Figure 1.2B, clusters 1 and 2). In accordance with these earlier observations, the genome-wide DNA distributions (Figure 1.8B, blue violin plots) reveal a high density of lowly methylated 200bp windows specifically in the LATE intermediates. However, although average DNA methylation levels are lower in the LATE intermediates, their median methylation level (indicated by the white circles in the violin plots) looks relatively similar to that of the other cell types.

LATE-specific differences are also apparent for certain chromatin states. States 5 and 15 (corresponding to regions with moderate enhancer activity and polycomb repression, respectively) showed more windows with low levels of DNA methylation in the LATE intermediates, than seen in the other three stages of somatic cell reprogramming. Interestingly, the DNA methylation distributions of states 3 and 4 (active enhancers) in the LATE intermediates associated with lower methylation levels that were more similar to those of the ESCs than the other cell types (Figure 1.8B).

Also, many more windows showed low levels of DNA methylation in the LATE intermediates (Figure 1.8B). And since this chromatin state accounted for more than 40% of the genome, it could explain the overall hypomethylated nature of the LATE intermediates.

This analysis showed that DNA methylation distributions differed within each of 18 chromatin states by was largely similar for a given chromatin state across cell types. Exceptions included poised promoters, which appeared to be more methylated in ESCs on average, and enhancer-like regions of the genome, which looked less methylated in ESCs.
1.3.8 Significant changes in DNA methylation with respect to chromatin state

Given that significant changes in DNA methylation were observed to take place during somatic cell reprogramming (section 1.3.2 and 1.3.3, above; Figure 1.2), we were interested in seeing where these changes occurred with respect to chromatin state. To address this, we considered 200bp windows that showed an absolute change in DNA methylation of at least a 50% in each pairwise comparison, and determined the states into which these fell. Only 0.41% of the MEF genome sees this level of significant change in DNA methylation during the first 48 hours of reprogramming, the majority of which occur in the low signal chromatin states 17 and 18 (Figure 1.8Ci). However, these chromatin states are also the most abundant in MEFs in general (Figure 1.8Civ), so the proportions do not represent a significant enrichment over the expected proportion (Figure 1.8Di). Chromatin state 15, representing polycomb repression in MEFs, shows the next highest proportion of windows showing a significant change in DNA methylation between the MEFs and EARLY intermediate cell states, which represents a slight enrichment (Figure 1.8Ci/Di). Each of the chromatin states associated with enhancer regions in MEFs (3–9) show an enrichment for significant differential methylation early in somatic cell reprogramming, whereas the promoter and transcription-associated states (1, 2, 10–14) all exhibit an under-enrichment for such changes (Figure 1.8Ci/Di).

Although a far higher proportion (2.72%) of 200bp windows in the MEF genome show a 50% change in DNA methylation in the LATE intermediates, the distribution of these windows across MEF chromatin states and their enrichment over the expected number is very similar to that seen within the first 48 hours of reprogramming (Figure 1.8Cii/Dii).
A smaller number of windows show a 50% change in DNA methylation going from MEFs to ESCs (2.15%) than seen when going from MEFs to the LATE intermediates (2.72%), but that can probably be explained by the fact that the LATE intermediates possess a high density of very lowly methylated regions relative to the other cell types (Figure 1.8B, blue violin plots), and already discussed in reference to Figures 1.1Ciii and 1.2Cvii. However, the distribution of windows showing that level of DNA methylation change by the pluripotent state across chromatin states is somewhat different from that seen for the EARLY and LATE intermediates, as is the enrichment pattern (Figure 1.8Ciii/Diii). The number of windows showing a 50% change from the level seen in MEFs is very enriched in chromatin states 3 to 11, as well as 14, representing the majority of MEF chromatin states associated with enhancer or transcriptional activity. Although the promoter states (1 and 2) do not show enrichment, they are not as under-enriched in the pluripotent state as seen previously (Figure 1.8Ciii/Diii).

This analysis showed, once again, that few significant changes in DNA methylation (>50%) occur during the first 48 hours of reprogramming. Those changes that do occur are somewhat enriched in chromatin states representing enhancer regions of the genome, as well as polycomb-repressed regions. Many more such changes have occurred by the LATE stage of reprogramming, as well as in the pluripotent state, where most of the MEF chromatin states associated with enhancer or transcriptional activity are enriched for DNA methylation changes of this magnitude. However, although this analysis revealed that changes in DNA methylation are associated with specific chromatin contexts, it did not address the directionality of change.
1.3.9 DNA methylation changes in regions undergoing transitions in chromatin state during somatic cell reprogramming

The previous section showed in which chromatin states significant changes in DNA methylation tend to occur during somatic cell reprogramming, it did not take into account the directionality of change, nor stepwise changes between sequential cell states. Furthermore, more subtle changes (showing an absolute change in DNA methylation below the 50% threshold) were not considered. To address these points, we wanted to intersect DNA methylation changes with changes in histone modifications. To that end, we considered each of the 324 possible chromatin state transitions going from one stage of reprogramming to the next (as described in Figure 1.1A), focusing on the transitions from MEFs to the EARLY intermediates (Figure 1.9Ai), and then subsequently from the EARLY intermediates to ESCs (Figure 1.9Aii). Although many more chromatin state transitions were observed after 48 hours, the chromatin landscape was already relatively dynamic during the first 48 hours.

To see whether these changes in histone modifications were tied to changes in DNA methylation during the same phases of somatic cell reprogramming, we calculated the mean change in DNA methylation in windows belonging to each of the chromatin state transitions for same two steps in reprogramming intervals looked at in Figure 1.9A. To ensure significant changes in DNA methylation, we filtered out those changes associated with transitions consisting of only few windows by requiring a −log10(p-value) ≥ 3 by a Z-test. Mean DNA methylation changes that met this significance requirement were plotted for each of the two reprogramming intervals (Figure 1.9Bi/ii). The only MEF to EARLY chromatin state transition to show a significant change with an absolute magnitude ≥ 10% was that from state 14 to state 2,
representing the transition from a transcriptional to poised promoter chromatin state (Figure 1.9Bi). All the other average changes values shown in the plot were < 2% and were significant simply by virtue of the large number of windows undergoing each of these transitions.

In contrast with the MEF to EARLY transitions, the windows undergoing transitions in chromatin state going from EARLY intermediates to ESCs showed many large and significant changes in mean DNA methylation (Figure 1.9Bii). Those windows transitioning to chromatin states 1 to 9 showed a marked decrease in DNA methylation (indicated by the swath of blue in the heat map), while those transitioning away from states 1 to 5 towards states 9 to 18 showed an increase in DNA methylation (indicated by the red cells; Figure 1.9Bii).

Taken together, although the changes in DNA methylation accompany changes in chromatin state later in somatic cell reprogramming, with those regions transitioning to active states exhibiting DNA demethylation and those transitioning away from active states showing DNA re-methylation, during the first 48 hours, there is little change in DNA methylation even though there are many transitions in chromatin state. This suggests that either DNA methylation is decoupled from chromatin stages early in reprogramming (but not later), or that DNA methylation changes lag changes in chromatin state. The second scenario corresponds to evidence of that DNA methylation changes follow histone modification changes already seen with respect to factor binding sites in section 1.3.5 and co-binding sites in 1.3.6.

1.3.10 Modeling DNA methylation using histone modifications

To better understand what might drive DNA methylation dynamics during somatic cell reprogramming, we compared the DNA methylation profiles from each cell type to a
corresponding set of ChIP-seq profiles for the nine histone modifications and the H3.3 variant introduced earlier in the chapter (section 1.3.5–1.3.8), but also included H3 libraries. We noticed that DNA methylation profiles (based on WGBS data) correlated, or anti-correlated, with certain histone modifications, including H3K4 monomethylation and H3K36 trimethylation (Table 1.6), as has recently been reported\textsuperscript{41}.

Seeing these associations between DNA methylation and histone marks, I examined whether DNA methylation could be modeled using the histone modification data sets as explanatory variables. For each cell type, I used multiple logistic regression to model the log odds of the DNA methylation probability (i.e. the log odds of the DNA methylation level) at each CpG dimer, regressing against the normalized ChIP-seq read counts within 100bp bins around the corresponding CpG dimers. The models for each cell type therefore had the following form, with R's glm() function used to perform the modeling\textsuperscript{42}:

\[
y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{11} x_{i11} (i = 1..n)
\]

where \(y\) represents WGBS DNA methylation estimates per CpG dimer, \(x_{i1}\) to \(x_{i11}\) represent the normalized ChIP-seq read counts within 100bp bins around the corresponding CpG dimers for the 11 histone modifications, and \(n\) represents the number of CpG dimers with data across all 12 data sets used in the model per cell type.

In order to select those histone mark combinations that best predicted DNA methylation levels, I used a stepwise regression procedure, specifically forward selection, which involves starting with no variables in the model, and then adding the variable that improves the model fit the most, repeating this process until all variables are used. For instance since in MEFs, H3K4me2 correlates most strongly with DNA methylation, it was the first variable used in the
model, followed by K36me3 since it improved the fit the most among the remaining marks, and so forth (Figure 1.10Ai, Tables 1.6 and 1.7). The more variables included, the better the fit, although the pseudo $R^2$ plateaus after a time.

The best fit was obtained for ESCs (pseudo $R^2 = 0.39$) with H3K4me dimethylation the strongest predictor of DNA methylation levels in all cell types and H3K36 trimethylation next most influential in all cell types except for ESCs, suggesting that there are differences in DNA methylation’s relationship with chromatin state going from the differentiated to the pluripotent cell state (Table 1.7).

Corresponding DNA methylation estimates based on RRBS libraries were also modeled using the same histone mark data sets as explanatory variables. In the case of the RRBS data, the fits for all four cell types was markedly superior to that based on the whole-genome bisulfite sequencing data with pseudo $R^2$ values greater than 0.5 in all cases, with ESCs once again showed the best fit (pseudo $R^2 = 0.74$; Figure 1.10Bi, Table 1.8).

Figure 1.10B shows genome browser tracks for the extended Esrrb locus illustrate that the model predictions (fitted) in both MEFs (i, blue tracks 1 vs. 2) and ESCs (ii, green tracks 10 vs. 11) correlate with the WGBS DNA methylation estimates (observed) reasonably well ($r = 0.42$ and 0.67, respectively). Intriguingly, model fits were essentially unchanged when only four histone marks (H3K36me3 and the three H3K4 methylation marks) were used as predictors, as opposed to all 11 modifications (tracks 1 vs. 3 and 10 vs. 12).

Distributions of observed (WGBS estimates) versus model fitted DNA methylation levels suggest that, although similar, fitted values lack the variance observed, especially at the lower
end of the range (Figure 1.10Ci). The fact that the ESC fitted DNA methylation distribution (green; right-hand side) shows a higher density of low-end values may help to explain the superior fit and correlation between observed and fitted DNA methylation for ESCs.

Corresponding distributions of RRBS-based fitted values also appear lack some of the low-end values observed using RRBS estimates (Figure 1.10Cii). The bimodal nature of RRBS data is captured quite faithfully, however, and helps to explain the far higher correlation between observed and fitted DNA methylation values ($r = 0.78$ for MEFs and $r = 0.89$ for ESCs; Figure 1.10Cii, Table 1.9).

CpG dimer coverage is a lot lower using RRBS data (4.1% in MEFs and 6.5% in ESCs, Table 1.1) than when using WGBS data (68% for MEFs and 81% for ESCs, Table 1.1), although model-based DNA methylation coverage is also dependent on the coverage of each of the 11 ChIP-seq data sets used to as explanatory variables (Table 1.9).

Models built by predicting RRBS estimates predicted genome-wide DNA methylation levels as well as WGBS-based models (data not shown), suggesting that although these models relied on a much more limited proportion of the genome with a strong CpG bias, similar features were responsible for driving the models in both cases, as seen in the order that the histone modifications were added to the growing models (Figure 1.10A).

Figure 1.10D shows that the distribution of differences between observed and fitted DNA methylation levels based on WGBS data broken up into deciles. The greatest error magnitude and variance is found at the low-end of DNA methylation estimates (deciles in the range of 0–50%), as suggested by difference in the combined distributions in Figure 1.10Ci.
To investigate whether sites where the model failed to accurately predict DNA methylation were associated with particular features, I consider those CpG dimers showing an absolute error in DNA methylation prediction of greater than 20% relative to WGBS estimates and looked at the density of 20 different genomic features in the 100bp around these sites, as listed below the bars in Figure 1.10Ei. In MEFs, the sites of predictions showing a large discrepancy were found to be under-enriched from H3K27ac, H3K36me3, H3K4me3, H3K79me2, H3K9ac, and PolII binding in MEFs (Figure 1.10Ei), suggesting that the model is failing predominantly in regions outside of active promoters and gene bodies. In ESCs, significant predictions errors were found in regions very under-enriched for H3K4me3 and H3K9ac, and slightly enriched for H3K4me1 (Figure 1.10Eii), again suggesting that the model is failing predominantly in regions away from active promoters and potentially at enhancers. In both cases, the majority of these features were highly ranked predictors in the respective models (Ai).

Since we had chromatin state data available (section 1.3.7; Figure 1.8A), I next compared WGBS DNA methylation estimates to model predictions in each of the different 18 chromatin states introduced previously. For each of the four cell types (Figure 1.1A), I calculated the mean DNA methylation within the 13.3 million 200bp windows corresponding to those used for the chromatin state model, with only those windows containing at least one CpG with 5X coverage being assigned a value in an effort to ensure high-confidence estimates. For each cell type, the distributions of the model-fitted DNA methylation values track the chromatin state-specific distribution shapes seen for the WGBS-based estimates reasonably well, as also seen in the way the median values (white circles) track one another (Figure 1.11A). In other words, the common
patterns of hypomethylation in promoter-associated chromatin states (1 and 2), intermediate DNA methylation of the enhancer-related states (3, 4, 5, and 7), and hypermethylation seen for most other states (Figure 1.11A top row of violin plots) were well captured by the distributions of the model-fitted DNA methylation values (bottom row), albeit the variance was lower in the modeled values, as seen in the genome-wide comparisons (blue violin plots and Figure 1.10Ci.

Cell type-specific differences were also captured by the model. For instance, the tendency for windows belonging to chromatin state 2 (corresponding to poised promoters) to be more methylated in ESCs on average is also seen for the model-fitted values suggesting that this is a real difference at the biological level and not just a technical difference in the ESC WGBS data set (Figure 1.11Aiv). However, the slight hypomethylation seen in ESC chromatin states 3, 4, 8 and 9 (corresponding to enhancer-like regions of the genome) do not look more demethylated in the ESC WGBS-based modeled values relative to the other cell types (Figure 1.11Aiv).

Figure 1.11B shows the same distributions shown in Figure 1.11Aiv, but reorganized to allow easier state-by-state comparison, and to introduce the comparison metrics given above the violin plots. ‘% coverage’ gives the percentage of 200bp windows with data per chromatin state alternating between observed and model-fitted DNA methylation estimates, while ‘% pairwise complete’ gives the percentage of 200bp windows with data in both the observed and model-fitted estimates per state. In most cases, the model provides similar coverage to that seen for the WGBS estimates with the exception of chromatin states 1 and 2, where the model-fitted coverage is substantially lower. This is the case for all four cell types, not just he ESCs.

The metrics labeled ‘methyl15’ and ‘methyl25’ give the percentage of 200bp windows where the difference between the observed and model-fitted DNA methylation estimates was at
most 15% and 25%, respectively. Using the more lenient methyl25 comparison, 88% of windows show this level of concordance between estimated (observed) and modeled (fitted) DNA methylation values (Figure 1.11B). The lowest concordance of 69% is seen for chromatin state 7, representing the poised enhancer state, which has a broad spread in DNA methylation estimates around a mean value of ~50% (Figure 1.11B). Other chromatin states that exhibited somewhat intermediate levels of DNA methylation (states 3–6 and 8) also showed relatively low concordance of 70–80% (Figure 1.11B).

In sum, WGBS-based CpG dimer-level DNA methylation estimates could be well modeled using ChIP-seq data for nine histone features using a logistic regression framework, with ESC data producing the best fit. In fact, only four key histone modifications (H3K36me3 and the three H3K4 methylation marks) were sufficient to produce a fit almost as good as that produced by all 10 histone marks combined. RRBS-based estimates produced far superior fits across all cell types, albeit for a far smaller proportion of the genome. The models tended to fail at the lower end of methylation estimates in regions that were under-enriched for histone modifications important to the model. Averaging CpG dimer-level DNA methylation estimates and fitted values within 200bp windows, 10/18 chromatin states showed a concordance within 25% in at least 88% of the comparable windows. The eight states showing lower levels of concordance (69–83%) had more intermediate levels of DNA methylation and exhibited some degree of enhancer-like character (states 2–9).
1.4 Discussion

In this study, we investigated DNA methylation dynamics during the course of somatic cell reprogramming within a mouse model system. We produced whole genome bisulfite sequencing (WGBS) libraries for four cell types representing different stages of the reprogramming process in order to assess their genome-wide DNA methylation levels at the single nucleotide level. An accompanying set of reduced representation bisulfite sequencing (RRBS) libraries were also produced for validation purposes (Figure 1).

With the data, we confirmed that DNA methylation changes do indeed take place during the course of somatic cell reprogramming to pluripotency. DNA methylation profiles diverge progressively (Figure 1.1C/D), with significantly differentially methylated regions (DMRs) mainly being found later in the process (Figure 1.2B). This corresponds to previous findings showing that DNA methylation change, especially loss of methylation, to be critical late in reprogramming\textsuperscript{17,19-21}.

DMRs were clustered into groups reflecting distinct patterns of change that showed strong similarities to those recently observed by Lee et al.\textsuperscript{21} These patterns included both gain and loss of methylation during reprogramming. However, only three of the six cluster groups were associated with the promoters of genes that exhibited a significant change in expression during the course of reprogramming (Figure 1.2B/C, cluster groups 3, 4 and 5).

Genes associated with regions that lost DNA methylation at the LATE and fully reprogrammed stages (cluster groups 3 and 4), showed increased expression during reprogramming (Figure 1D and Table 1.2), and were involved with cell division and stem cell maintenance, respectively (Tables 1.2 and 1.3). In other words, genes associated with pluripotency were demethylated very late in reprogramming (i.e. cluster group 4), corresponding
to previous findings showing\textsuperscript{20}. Interestingly, the DMRs in group 4, representing ESC-specific demethylation, seemed to be found within regions of broader demethylation, which could correspond to DNA methylation valleys (DMVs) found in human ESCs, described by Xie et al.\textsuperscript{30}, or super-enhancers, which have been found to often overlap DMVs\textsuperscript{43-45}. This might also correspond to the hypomethylated nature of enhancer regions in ESCs (Figure 1.8B).

Genes associated with genomic regions that gained methylation during reprogramming (cluster group 5), showed a significant reduction in expression during reprogramming (Figure 1D and Table 1.2), and were involved in differentiation and development (Table 1.4), which again corresponds to previous findings\textsuperscript{20}.

The distribution of the DMRs with respect to genes was different for the different cluster groups with the DMRs for the cluster groups 3, 4, and 5, discussed above, tending to be found in closer proximity to genes (Figure 1.2E).

Looking at reprogramming factor binding sites, MYC binding occurred in exclusively demethylated regions, KLF4 binding predominantly occurring at enhancer sites at all stages, and OCT4 and SOX2 binding sites showing cell types-specific epigenetic patterns and motif dependence (Figures 1.3–1.5). EARLY-specific and constitutive O and SOX2 S binding sites with O/S motifs were found to occur exclusively at more methylated, enhancer-like sites (Figures 1.3 and 1.5), whereas ESC-specific O/S binding sites occurred exclusively in enhancer-like regions regardless of the absence or presence of motifs (Figure 1.4).

Constitutive O, S and K DNA binding sites appeared to lose DNA methylation over the course of reprogramming, but, importantly, these changes appeared to be preceded by changes to
chromatin state, something that appears to be generally the case (Figure 1.4), which agrees with previous findings\textsuperscript{16,20}.

To investigate the epigenetic state around sites where the reprogramming factors do and do not co-bind, we clustered the reprogramming factor binding sites (Figure 1.6A). Sites that saw M binding in MEFs or the EARLY intermediates were the most demethylated of any of the TF clusters binding sites, and appeared to be largely specific to active promoter regions and to be constitutive (Figure 1.6B). O and S often appeared to act in tandem and were usually associated with an active chromatin environment, and in ESCs their binding appeared to be enhancer specific, and motif independent (Figure 1.6A/B/C), as seen previously for ‘001’ O/S binding (Figure 1.4B). K was found to co-localize with O and/or S in many instances, though in some cases strong K binding appeared to compensate where O/S binding was lacking (Figure 1.6A/B/C). Intriguingly, H3K4me1/2 and H3K27ac enrichment were often seen to precede the fullest extent of DNA demethylation, with H3.3 enrichment more closely coinciding with DNA methylation loss (Figure 1.6), again corresponding to what was previously seen for ‘001’ O/S binding (Figure 1.4B).

Investigations into the epigenetic state around TF binding sites during reprogramming have been conducted before\textsuperscript{21}, but our analysis included many additional histone modifications than typically considered and provided interesting differences in the epigenetic state around TF binding sites based on the absence or presence of motifs at the binding sites.

To investigate how DNA methylation distributions might differ within regions of the genome showing different chromatin character beyond factor binding sites, we made used of a chromatin state model (Figure 1.8A). DNA methylation distributions were indeed different
within the different chromatin states, but very similar for a given chromatin state across cell types (Figure 1.8B). Exceptions were states that captured poised promoters, which were more methylated in ESCs on average, and enhancer-like regions of the genome, which were less methylated in ESCs (Figure 1.8B).

As previously observed and discussed above, relatively few significant changes in DNA methylation (>50%) were seen to occur early (within the first 48 hours) in reprogramming (Figure 1.8Ci), but the few changes were slightly enriched in enhancer and polycomb-repressed regions (Figure 1.8Di). There were many more significant changes in DNA methylation changes by the latter stages of reprogramming (Figure 1.8Cii/iii). Differentially methylated regions going to the pluripotent state were very enriched within regions of the genome that possessed an enhancer-like chromatin state character in MEFs (Figure 1.8iii).

Looking at the average change in DNA methylation in regions undergoing chromatin state transitions (Figure 1.9A), although there was essentially no significant change in any of the transitions during the first 48 hours of reprogramming (Figure 1.9Bi), regions transitioning from EARLY intermediates to ESCs showed many large and significant changes in mean DNA methylation, with those regions transitioning to active states showed a marked decrease in DNA methylation, while those transitioning away from active states showed an increase in DNA methylation (Figure 1.9Bii). During the first 48 hours, however, there are many transitions in chromatin state (Figure 1.9A), which suggests that DNA methylation changes are either decoupled from chromatin stages early in reprogramming (but not later), or that they occur subsequent to changes in chromatin state. The latter hypothesis is supported by evidence of that DNA methylation changes lag histone modification changes seen within the context of reprogramming factor binding in sections 1.3.5 and 1.3.6, and has been suggested previously\textsuperscript{16,20}.\textsuperscript{16,20}
We confirmed that a strong relationship existed between DNA methylation and certain histone modifications in all four cell types (Table 1.6), as reported previously\textsuperscript{16,41,46}. Interestingly, the strong positive correlation between H3K36me3 and DNA methylation levels observed for the other three cell lines was not as evident for ESCs, while conversely, (anti-) correlations with other histone modifications (H3K4me1/2/3, K9ac, Ka27ac) became stronger in ESCs (Table 1.6). A recent paper showed that the bodies of transcribed genes, which are enriched for the H3K36 trimethylation mark were preferentially targeted for DNA methylation via DNMT3B\textsuperscript{47}, so the lower correlation between H3K36me3 and DNA methylation observed in ESCs might be due to somatic genes not having been turned on yet. Alternatively, since DNMT3B forms a complex with DNMT3L in ESCs\textsuperscript{48}, and DNMT3L is highly expressed in that cell type (data not shown), perhaps this disrupts DNMT3B’s ability to methylated H3K36.

We used multiple logistic regression to model WGBS and RRBS-based CpG methylation estimates using histone modification data as predictors (Figure 1.10 and Tables 1.7–1.9). Only four key histone modifications (H3K36me3 and the three H3K4 methylation marks) able to produce a fit almost as good as that produced by all 11 histone features combined (Figure 1.10A/B). RRBS-based data resulted in models with better fits across all cell types, but for a far smaller proportion of the genome. For many chromatin states, observed versus fitted DNA methylation estimates fell within 25% of one another in at least 88% of the region. States with intermediate levels of DNA methylation showed lower levels of concordance (69–83%).
Looking ahead, we still have many open questions about the role of DNA methylation during reprogramming. For instance, it is still not known whether the observed DNA demethylation is a passive process that simply occurs in conjunction with cell division, or whether it is an active process involving the ten-eleven translocation (TET) methylcytosine dioxygenases\textsuperscript{14}. However, it is known that the absence of TET3 leads to a failure to demethylate key pluripotency genes such as Pou5f1 and Nanog, so an active process in the latter stages would not be surprising\textsuperscript{49}. One means to try to ascertain this is to utilize TAB-seq to determine levels of 5-hydroxymethylcytosine\textsuperscript{50}, since TETs catalyze the conversion of 5-methylcytosine to 5-hydroxymethylcytosine\textsuperscript{51}.

In fact, since WGBS and RRBS cannot distinguish between 5mC and 5hmC, complementary TAB-seq libraries for 5hmC levels would be informative, as they would reveal the extent to which DNA methylation might be in a state of active transition.

More generally, although DNA methylation is associated with gene repression\textsuperscript{14}, its specific role of is not clear and requires further study.
Material and methods

**Whole-genome bisulfite sequencing (WGBS)**

Genomic DNA from MEFs, induced MEFs (48h OSKM), preiPSCs, and ESCs was isolated using the Blood and tissue DNeasy kit (Qiagen). Isolated DNA was treated with RNaseA for 30 min at 37°C and cleaned up using AMPure XP beads. 5 µg of treated DNA was fragmented to 100–500 bp using a Bioruptor Sonicator. 5 minutes in pulses of 30 sec on, 1 minute off. DNA fragments were visualized on 1% agarose gel, gel extracted and purified using a QIAGEN gel extraction minelute kit. End-repair reactions (50 µl) contained 1x T4 DNA ligase buffer (NEB), ATP, 0.4 mM dNTPs, 15 units T4 DNA polymerase, 5 units Klenow DNA polymerase, 50 units T4 polynucleotide kinase (all NEB) and were incubated for 30 min at 20°C. DNA clean-up was performed using a 2x volume of AMPure XP beads and eluted in 32 µl of dH2O. Adenylation was performed for 30 minutes at 37°C in 50 µl volumes that contained 5 µl 1x Klenow buffer, 0.2 mM dATP and 15 units Klenow exo− (NEB). Adenylated DNA fragments and methylated adapters (Illumina) were ligated for 15 min at 20°C in a 50 µl reaction containing 5,000 units quick ligase (NEB) and 5 µl of adapters. Adaptor-ligated DNA of 200-600 bp, was size-selected on a 2% agarose gel. Bisulfite conversion was performed with an EpiTect Bisulfite Kit (QIAGEN) following the manufactures conditions. Bisulfite converted DNA was amplified for 15 cycles with PfuTurboCx Hotstart DNA polymerase (Agilent technologies). The final library DNA was quantified using a Qubit fluorometer and a Quant-iT dsDNA HS Kit (Invitrogen).

Single-end sequencing for 100bp reads was performed on an Illumina Hiseq 2000.
Reduced Representation Bisulfite sequencing (RRBS)

5 µl of genomic DNA was digested with 50 units of MspI (NEB) in a 100 µl reaction for 6 hours at 37°C. Digested DNA was run on a 3% low-melt agarose gel (Lonza) and fragments of 25 to 300 bp were extracted and purified using a MinElute gel extraction kit (QIAGEN) according to the manufacturers instructions. DNA end-repair and adenylation was as described above with the exception of using a dNTP mix consistent of dATP, dGTP and 5me dCTP. Ligation to methylated adapters and subsequent library construction was performed similarly to the WGBS protocol.

Single-end sequencing for 100bp reads was performed on an Illumina Hiseq 2000.

Bisulfite sequencing data analysis

DNA methylation calling was performed using BS-Seeker\textsuperscript{26} using Bowtie 0.12.9\textsuperscript{52} for read alignment on the UCLA Hoffman2 computer cluster. WGBS reads were mapped to the entire mm9 reference genome while RRBS reads were mapped to a reduced reference that was \textit{in silico} digested using the \textit{MspI} recognition sequence and limited to fragments of 20–500bp in length. The 100bp reads were trimmed of adapter sequences and allowed 5 mismatches during mapping. For DNA methylation level calling, only CpG dimers covered by at least 5 reads on both were used in an effort to obtain reliable methylation levels.

All statistical analysis, clustering, and heat map generation were performed using custom R scripts\textsuperscript{53}. Hierarchical clustering of samples was performed using the distance metric 1 – $r$ and the complete linkage agglomeration method.

Native ChIP-seq libraries

All histone modification data used for this study were determined using native ChIP\textsuperscript{54} and will be described in detail in a separate manuscript (Chronis et al., in preparation). Briefly, cells were
resuspended in buffer I (0.3M sucrose, 60mM KCl, 15mM NaCl, 5mM MgCl2, 10mM EGTA, 15mM Tris-HCl, 0.5mM DTT, 0.1% NP40 and protease inhibitor cocktail) and incubated on ice for 10 minutes. Nuclei were generated by centrifugation via a sucrose cushion (1.2M sucrose, 60mM KCL, 15mM NaCl, 5mM MgCl2, 0.1mM Tris-HCl, 0.5mM DTT, and protease inhibitor cocktail). Nuclei were then resuspended in MNase-digestion buffer (0.32M sucrose, 50mM Tris-HCl, 4mM MgCl2, 1mM CaCl2, protease inhibitor cocktail) and digested with 3 units of MNase (Roche) for 10 minutes at 37°C. Soluble chromatin fractions were then incubated with anti-H3K4me3 (abcam; ab8580), anti-H3K4me2 (abcam ab7766), anti-H3K4me1 (ab8895), anti-H3K27me3 (Active motif 39155), anti-H3K27ac (abcam; ab4729), and anti-H3K36me3 (abcam; ab9050). Extracts were washed twice with wash buffer A (50mM Tris-HCl, 10mM EDTA, 75mM NaCl), wash buffer B (50mM Tris-HCl, 10mM EDTA, 125mM NaCl), wash buffer C (50mM Tris-HCl, 10mM EDTA, 250mM NaCl). DNA extraction and library preparation as described54.

Cross-linked ChIP-seq libraries

Transcription factor binding data generated in this study were acquired using cross-linking ChIP and will be described in detail in a separate manuscript (Chronis et al., in preparation). Briefly, cells were grown to a final concentration of 5x10^7 cells for each sequencing experiment. Pre-iPSCs and ESCs were grown in feeders, then were feeder depleted and grown to gelatin overnight prior to each experiment. Cells were chemically cross-linked by the addition of formaldehyde to 1% final concentration for 10 minutes and quenched with 0.125 M final concentration glycine. Cross-linked cells were resuspended in sonication buffer (50mM Hepes, 140mM NaCl, 1mM EDTA, 1% TritonX-100, 0.1% Na-deoxycholate, 0.1% SDS) and sonicated
using a Diagenode Bioruptor for 10 minutes. Nuclear extracts were then incubated overnight at 4°C with the respective antibodies: anti-Esrrb (RnD; H6705), anti-Klf4 (RnD; AF3158), anti-Myc (RnD; AF3696), anti-Nanog (cosmobio), anti-OCT4 (RnD; AF1759), anti-Sox2 (RnD AF2018), anti-p300 (SantaCruz;sc-585) SMC1 (Bethyl labs A300-055A), MED1 (Bethyl labs A300-793A), CTCF (Millipore 07-729), polII (Covance 8WG16), K79me2 (active motif 39143), K9me3 (abcam ab8898 or Millipore 05-1242), H3.3 (abnova). Extracts were washed twice with RIPA, low salt buffer (20mM Tris pH 8.1, 150mM NaCl, 2mM EDTA, 1% Triton X-100, 0.1% SDS), high salt buffer (20mM Tris pH 8.1, 500mM NaCl, 2mM EDTA, 1% Triton X-100, 0.1% SDS), LiCl buffer (10mM Tris pH 8.1, 250mM LiCl, 1mM EDTA, 1% deoxycholate, 1% NP-40), and 1xTE. Reverse cross-linking was performed by overnight incubation at 60°C with 1% SDS and proteinase K. All protocols for Illumina/Solexa sequence preparation, sequencing and quality control are provided by Illumina with a minor modification: PCR amplification step were limited to 10 cycles.

**ChIP-seq data analysis**

Reads were mapped to the mouse genome (mm9) using Bowtie software\textsuperscript{52} and only those reads that aligned to a unique position with no more than two sequence mismatches were retained for further analysis. Multiple reads mapping to the same location in the genome were collapsed to a single read to account for clonal amplification effects.

ChIP-seq peaks were called using MACS software (Version 2)\textsuperscript{55} using a bandwidth parameter of 150 bp. Peaks with an FDR < 0.05 and fold $\geq$ 3-fold were retained.

To define cell type specific peaks we used the BEDTools\textsuperscript{56} suite. Briefly subtractBed was used to identify unique peaks per cell type by subtracting neighboring peak summits within a 400
bp distance. Summits within 100bp of each other were defined as shared summits between cell types.

Motif scanning was performed using HOMER\textsuperscript{33} and the script findMotifsGenome.pl to scan ChIP-seq peaks for motifs within 200bp windows around peak summits in the mm9 genome, using default settings.

ChIP-seq reads counts binned within 100bp bins around mm9 CpG dimers were normalized to 10 million reads and input reads counts (native or cross-linked as appropriate) were subtracted to a minimum of zero reads.

**Transcription factor clusters**

The genome was tiled into 500bp windows and the presence of reprogramming factor peak summits were used to define the (transcription factor) TF clusters. This resulted in a vector of binary data for each TF reflecting its absence or presence within 500bp windows across the genome. The windows represented by these vectors were then clustered using R’s k-means function\textsuperscript{42} (using the Hartigan-Wong method) to obtain groups of windows exhibiting common combinatorial binding patterns across the genome. The combinations of TFs found to co-bind within each window of a cluster are analogous in their combinatorial nature of histone marks defined by the chromatin states (see below). The number of centers (k = 20) was chosen so as to substantially reduce the number of potential combinatorial TF groups ($2^{14} - 1$), while ensuring that each cluster was still represented by a significant number of windows.
Chromatin states

The model will be described in detail in a separate manuscript (Chronis et al., in preparation), but briefly 18 chromatin states in the MEFs, EARLY intermediates, LATE intermediates, and ESCs were identified at a resolution of 200bp using chromHMM as described by Ernst and Kellis\textsuperscript{57} using ChIP-seq data sets for nine histone modification, one histone variant (H3.3), and an input, as listed in Figure 1.8.

RNA-seq

RNA-seq was performed essentially as described in Parkhomchuk et al.\textsuperscript{58}, using 4 ug of total RNA as starting material. Unstranded 50bp single-end read libraries were generated for ESCs and Pre-iPSCs (LATE intermediates). Strand-specific 50bp single-end read libraries were generated in triplicate for both uninduced MEFs and the 48 hour induced counter-parts (EARLY intermediates). And one set of strand-specific 100bp paired-end read libraries were generated for each of the four cell types.

RNA-seq data analysis

Reads were mapped to the mouse genome (mm9) using the UCSC RefSeq annotation using TopHat2 software\textsuperscript{59}. HTSeq\textsuperscript{60} was used to determine gene counts, and DESeq2\textsuperscript{61} for differential analysis. RPKMs were calculated using gene lengths based on concatenated exons.

DNA methylation modeling

Multiple logistic regression was used to model WGBS and RRBS-based DNA methylation estimates (the response vectors) using normalized ChIP-seq read counts (see ‘ChIP-seq data
analysis’) within 100bp windows around corresponding CpG dimers for nine histone modification (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K27me3, H3K36me3, H3K79me2, H3K9me3) and two histone variants (H3.3, and H3) (the predictors) with R’s glm() function42. The four columns of each sub-table in Tables 1.7 and 1.8 are outputs from R’s summary.glm() function42, with the coefficients and standard error transformed from log-odds to percentage change in DNA methylation, using the formula: \( \left( \frac{e^\hat{\beta}}{1 + e^\hat{\beta}} \right) \times 100 \) – 50. Pseudo R² values were calculated using the formula: \( 1 - \left( \text{deviance} / \text{deviance for null model} \right) \).

Data submission

All high-throughput sequencing data will be made available on GEO at the time of publication.
Figure 1.1  Genome-wide DNA methylation during somatic cell reprogramming
**A.** Schematic representation of samples used in the study to represent the four stages of somatic cell reprogramming: 1) mouse embryonic fibroblasts (MEFs; blue) to represent the differentiated state and harboring a ‘stem cell cassette’ that allows for the simultaneous induction of the four reprogramming factors (OCT4 (O), SOX2 (S), KLF4 (K), and MYC (M)) by the addition of doxycycline; 2) EARLY intermediates (yellow) represented by MEFs that had been induced with doxycycline for 48 hours; 3) LATE reprogramming intermediates (red) represented by partially induced pluripotent cells, or pre-iPSCs; and 4) embryonic stem cells (ESCs; green) representing the fully reprogrammed state.

**B.** Bar plots showing the number of CpGs (on either strand) obtained at five different coverage levels (1–20 X) in each of the four whole-genome bisulfite sequencing (WGBS) samples described in A (i–iv).

**C.** Distributions of DNA methylation estimates based for each WGBS sample described in A based on individual CpGs (on either strand) with at least 5X coverage (i–iv). Mean DNA methylation levels shown in top left corner, along with the CpG-level Pearson correlation values for the EARLY intermediates, LATE intermediates, and ESC samples versus the DNA methylation profile for MEFs. Color scheme as in (A).

**D.** Heat maps of the Pearson correlation coefficients ($r$) between the CpG dimer-level DNA methylation profiles of pairs of samples based on (i) WGBS, and (ii) both the WGBS and RRBS samples for all four cell types described in A. Color scales indicate the range of correlation values. Dendrograms are based on hierarchical clustering of the samples using the distance metric $1 - r$ and complete linkage agglomeration method.
Figure 1.2  Differentially methylated regions during somatic cell reprogramming
A. Integrative Genomics Viewer (IGV) browser tracks of i) CpG-level DNA methylation levels, and ii) mean DNA methylation within 500bp windows tiled along the mouse genome around four representative gene loci (Runx1, Olig3, Esrrb, and Dppa2) for each of the four stages of somatic cell reprogramming in the study described in Figure 1.1A, namely MEFs (blue), EARLY (yellow) and LATE (red) reprogramming intermediates, and embryonic stem cells (ESCs; green). Differentially methylated regions highlighted. DNA methylation levels based on whole-genome bisulfite sequencing (WGBS) estimates.

B. 120,744 500bp windows tiled along the mouse genome, representing ~2.3% of the genome, were identified as significantly and consistently differentially methylated regions (DMRs) if they showed an absolute difference in DNA methylation of at least 50% in 3/6 of the pairwise comparisons, based on WGBS estimates. These DMRs were partitioned into six groups by k-means clustering to capture typical patterns of DNA methylation change during reprogramming. The heat map represents the mean DNA methylation level across the windows in each group (given to the right of each cluster group number) per sample (along bottom). The dendrogram was obtained by hierarchically clustering these mean DNA methylation levels across groups using correlation values as the distance metric in order to highlight those groups showing similar patterns of DNA methylation change.

C. Plots of mean DNA methylation per sample within a 20kb region centered around the differentially methylated windows in each of the cluster groups described in B (i–vi), as well as around 20,000 randomly-selected 500bp windows (vii). The color legend identifies cell type.

D. Violin (top) and empirical cumulative distribution function (CDFs, bottom) plots of the expression distributions of UCSC Refseq genes whose promoters (TSS +/- 2,500bp) contain differentially methylated windows in each of the cluster groups described in B (i–vi) for each
sample. The color legend identifies cell type. CDFs included to better highlight differences in the distributions of expression between cell type for genes associated with certain clusters. Asterisks indicate that statistically significant differences exist between at least one of the pairs of distributions (see Table 2).

**E.** Bar plots of the proportion of differentially methylated windows that fall 0–5 kb, 5–50 kb, 50–500 kb, or beyond either upstream or downstream of the TSSs of UCSC known genes for each of the cluster groups described in B (i–vi).
Figure 1.3  DNA methylation levels and histone modification enrichment around EARLY-specific ('100') reprogramming factor binding sites
A. Key for the classification of reprogramming factor peaks according by stage-specific or constitutive binding characteristics in the EARLY, intermediates, LATE intermediates, and ESCs. The heat map shows the binding patterns schematically, illustrating the rationale behind the code used to represent said pattern, which is given in the last column. EARLY-specific binding occurs only in the EARLY samples, so is binding pattern given the code ‘100’.

B. Heat map of DNA methylation levels and enrichment of nine histone modifications and one histone variant (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K27me3, H3K36me3, H3K79me2, H3K9me3, H3.3) within 1kb windows around the ChIP-seq peak summits of pluripotency transcription factor (OCT4, SOX2, KLF4, MYC) binding sites that are exclusive to the EARLY state (‘100’, see panel A), with and without the specified motifs within 100bp of the summit (+/- motif column and accompanying key) for the four samples representing different stages of somatic cell reprogramming described in Figure 1.1A (MEFs (blue), EARLY (yellow), LATE (red), ESCs (green)). The columns labeled ‘# peaks’ and ‘% peaks’ give the counts and percentage of factors peaks with and without motifs, with shading indicative of the relative proportion. The color legends identify cell type, DNA methylation level and histone modification enrichment, respectively. DNA methylation levels based on whole-genome bisulfite sequencing (WGBS) estimates, and histone modification enrichment values on normalized ChIP-seq read counts.

C. Plots of mean DNA methylation (row 1), as well as H3K4me1 and K3K4me3 enrichment (rows 2 and 3) for the four samples representing different stages of somatic cell reprogramming described in Figure 1.1A (MEF (blue), EARLY (yellow), LATE (red), ESC (green)) within 4kb windows around i) all EARLY-specific OCT4 ChIP-seq peak summits, as well as those with and without the OCT4 motif within 100bp of the summit, ii) all EARLY-specific KLF4 ChIP-seq
peak summits with and without the KLF4 motif within 100bp of the summit, and iii) all EARLY-specific MYC ChIP-seq peak summits. DNA methylation levels based on WGBS estimates, and histone modification enrichment values on normalized ChIP-seq read counts. The color legend identifies cell type.
Figure 1.4  DNA methylation levels and histone modification enrichment around ESC-specific ('001') reprogramming factor binding sites

As in Figure 1.3, but for ESC-specific binding ('001').
Figure 1.5  DNA methylation levels and histone modification enrichment around constitutively bound ('111') reprogramming factors
As in Figure 1.3, but for constitutive binding ('111').
Figure 1.6: Clustering of reprogramming factors reveals that co-binding is associated with distinct DNA methylation and histone modification patterns.”
A. Reprogramming factor (OCT4 (O), SOX2 (S), KLF4 (K), MYC (M)) ChIP-seq peaks at each of the three stages of somatic cell reprogramming following induction by doxycycline addition, as well KLF4 and MYC peaks in MEFs prior to doxycycline-induced up-regulation, were clustered using k-means clustering at 500bp resolution to produce 20 transcription factor (TF) clusters. Factor frequency in each TF cluster is represented by the heat map. The color scale denotes the frequency with which a given factor is found at genomic positions corresponding to the respective TF clusters. These frequencies were used to hierarchically cluster the co-binding events in order to reveal similar binding patterns between individual factors (dendrogram above heat map), as well as between the TF clusters (dendrogram left of heat map). TF clusters were assigned to one of five groups (A–E; right of heat map) based on the hierarchical clustering patterns. The color legend identifies cell type.

B. Heat map as in Figure 1.3B, but for the 1kb regions around the midpoints of the 500bp windows belonging to TF clusters 1–7, assigned to cluster groups A and B as in (A) and annotated in the first column. The second column describes particular differences in binding patterns seen for the individual TF clusters within each group. The final two columns describe DNA methylation level and histone modification enrichment patterns observed for each TF cluster, with ESC-specific color coding as described in the legends at the top right.
The DNA methylation and histone modification enrichment patterns associated with transcription factor cluster groups C, D, and E belonging to TF clusters 8–20, assigned to cluster groups C, D, and E.
Figure 1.8  DNA methylation levels with respect to chromatin state

<table>
<thead>
<tr>
<th>State</th>
<th>Genomic %</th>
<th>Input</th>
<th>K27ME3</th>
<th>K27AC</th>
<th>K4ME2</th>
<th>K4ME1</th>
<th>K4ME3</th>
<th>K36ME3</th>
<th>K9AC</th>
<th>K79ME2</th>
<th>H3.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promoter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1_PromA</td>
<td>0.5</td>
<td>3</td>
<td>2</td>
<td>97</td>
<td>91</td>
<td>87</td>
<td>27</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2_PromP</td>
<td>0.2</td>
<td>2</td>
<td>10</td>
<td>47</td>
<td>97</td>
<td>72</td>
<td>23</td>
<td>8</td>
<td>6</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Enhancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1_EnhA</td>
<td>0.9</td>
<td>2</td>
<td>0</td>
<td>64</td>
<td>91</td>
<td>87</td>
<td>27</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>34</td>
</tr>
<tr>
<td>2_EnhP</td>
<td>1.4</td>
<td>3</td>
<td>1</td>
<td>56</td>
<td>95</td>
<td>96</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3_EnhM</td>
<td>1.3</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>74</td>
<td>45</td>
<td>28</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>4_EnhW</td>
<td>2.6</td>
<td>3</td>
<td>1</td>
<td>29</td>
<td>3</td>
<td>54</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Tsscribed enhancer</td>
<td>1.1</td>
<td>22</td>
<td>58</td>
<td>7</td>
<td>39</td>
<td>63</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>8_TxEnhA</td>
<td>1.2</td>
<td>3</td>
<td>0</td>
<td>49</td>
<td>51</td>
<td>86</td>
<td>14</td>
<td>6</td>
<td>30</td>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td>9_TxEnhM</td>
<td>1.0</td>
<td>9</td>
<td>2</td>
<td>37</td>
<td>15</td>
<td>66</td>
<td>1</td>
<td>0</td>
<td>175</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Transcription</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10_Tx</td>
<td>2.9</td>
<td>3</td>
<td>1</td>
<td>8</td>
<td>19</td>
<td>1</td>
<td>0</td>
<td>94</td>
<td>3</td>
<td>51</td>
<td>2</td>
</tr>
<tr>
<td>11_TxS</td>
<td>1.6</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>19</td>
<td>3</td>
<td>0</td>
<td>11</td>
<td>2</td>
<td>31</td>
</tr>
<tr>
<td>12_TxS2</td>
<td>6.7</td>
<td>4</td>
<td>1</td>
<td>50</td>
<td>10</td>
<td>0</td>
<td>84</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>13_TxS22</td>
<td>4.5</td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>14_Tx</td>
<td>0.5</td>
<td>5</td>
<td>3</td>
<td>18</td>
<td>5</td>
<td>16</td>
<td>5</td>
<td>2</td>
<td>34</td>
<td>35</td>
<td>75</td>
</tr>
<tr>
<td>Polycomb</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15_ReprPC</td>
<td>8.5</td>
<td>7</td>
<td>56</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Repeats</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16_Repeats</td>
<td>1.3</td>
<td>4</td>
<td>8</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>64</td>
<td>0</td>
</tr>
<tr>
<td>Low signal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17_Low</td>
<td>22.4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18_LowL</td>
<td>42.2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Enrichment over genome-wide proportions (log2)**

**Figure 1.8** DNA methylation levels with respect to chromatin state
A. Heat map and functional annotation for an 18-level chromatin state model at 200bp resolution built using peak calls made using ChIP-seq data sets for nine histone modifications and one histone variant (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K27me3, H3K36me3, H3K79me2, H3K9me3, H3.3), as well as an input library, for each of the four cell types described in Figure 1.1A: i) MEFs, ii) EARLY intermediates, ii) LATE intermediates, and iv) ESCs. Candidate functional annotations were assigned to each of the 18 chromatin states based on the prevalence and combination of histone mark peaks, which could in turn be classified into the seven categories indicated in the left-hand column. The probability of a window in each state to contain a peak for a given histone modification is given as a percentage in each cell, and visually indicated by the intensity of color in the heat map. The proportion of the concatenated genome (MEFs + EARLY + LATE + ESCs) found in each of the 18 chromatin states is given in the third column.

B. Violin plots of the distributions of whole-genome bisulfite sequencing (WGBS) DNA methylation estimates in each of the 18 chromatin states (described in A), as well as genome-wide (blue), for each of the four cell types under study described in Figure 1.1A: i) MEFs, ii) EARLY intermediates, iii) LATE intermediates, iv) ESCs. The mean DNA methylation level within 200bp windows corresponding to those used for the chromatin state model were used, only considering those windows containing at least one CpG with 5X coverage, in an effort to ensure high-confidence estimates. White circles represent median values.

C. Pie charts showing the proportion of 200bp windows in MEFs showing significant differential DNA methylation at each stage of the three subsequent stages of somatic cell reprogramming (described in Figure 1.1A: i) EARLY intermediates, ii) LATE intermediates, iii) ESCs) found in each of the 18 chromatin states as described in (A). Any window showing an absolute change in
methylated level of at least 50% was considered to be significantly differentially methylated. The distribution of all 13.3 million 200bp windows genome-wide across the 18 chromatin states is shown in iv.

D. Bar plots showing the log2 fold change (observed / expected) number of differentially methylated windows at each stage of the three stages of somatic cell reprogramming as compared to MEFs as in (C).
Figure 1.9  Chromatin state transitions occur throughout somatic cell reprogramming but DNA methylation changes are predominantly occur late in the process

A. Heat map of the number of windows that underwent each of the 324 possible chromatin state transitions going from the MEFs to EARLY intermediates (i), and then to ESCs (ii) during somatic cell reprogramming, with reference to the four stages described in Figure 1.1A and 18 chromatin states described in Figure 1.8A. The number of 200bp windows observed to undergo a
given state transition is shown in each respective cell, and visually indicated by the intensity of color in the heat map (see color key).

**B.** Heat map of the mean change in DNA methylation for the windows that underwent each chromatin state transitions going from the MEFs to EARLY intermediates (i), and then to ESCs (ii) during somatic cell reprogramming, as in (A). Blue indicates a decrease and red an increase in DNA methylation levels for the windows in question on average (see color key). Values for the average change in DNA methylation are provided for those select transitions where the change was determined to be significant by a Z-test (-log10(p-value) >= 3). DNA methylation levels based on whole-genome bisulfite sequencing (WGBS) estimates
Figure 1.10  DNA methylation levels can be modeled using histone modifications
A. Bar plots showing the incremental change in pseudo $R^2$ values as a logistic regression model of DNA methylation levels was built up using stepwise forward selection to obtain an optimal combination of nine histones modifications (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K27me3, H3K36me3, H3K79me2, H3K9me3) and two histone variants (H3.3, and H3) at each step to predict DNA methylation estimated using (i) whole-genome bisulfite sequencing (WGBS) and (ii) reduced representation bisulfite sequencing (RRBS; replicate 1) data for each of the four cell types under study (MEFs, EARLY intermediates, LATE intermediates, and ESCs as described in Figure 1.1A).

B. IGV tracks of observed DNA methylation for MEFs (i, blue) and ESCs (ii, green) in a ~1.5 Mb region around the $Esrrb$ gene locus based on WGBS estimates, and model-fitted values (darker hues) using models built using all 11 and only 4 (H3K4me1, H3K4me2, H3K4me3, and H3K36me3) features, respectively to predict the WGBS estimates. Genome-wide Pearson correlation values between observed and respective fitted DNA methylation levels given in right hand margin. Selected chromatin mark signals as labeled.

C. Violin plots of the distributions of observed DNA methylation for MEFs (blue) and ESCs (green) compared to the distributions of fitted values using models built using 11 histone modifications to predict DNA methylation estimates using (i) WGBS and (ii) RRBS (replicate 1) data. White circles represent median values.

D. Violin plots of the distributions of difference between WGBS DNA methylation estimates (observed) and fitted DNA methylation values using models built using 11 histone modifications to predict DNA methylation estimated using WGBS data for MEFs (blue) and ESCs (green), with observed DNA methylation values partitioned into deciles. White circles represent median values.
E. Bar plots of the log2(observed/expected) feature density within 100bp around CpG dimers exhibiting an absolute difference between WGBS DNA methylation estimates (observed) and fitted DNA methylation of greater than 20% for MEFs (blue) and ESCs (green). Features considered are given below each bar.
Figure 1.11  Modeled DNA methylation levels in different chromatin states
A. Violin plots of the distributions of whole-genome bisulfite sequencing (WGBS) DNA methylation estimates (observed) and fitted DNA methylation values using models built using 11 histone modifications to predict the DNA methylation in each of the 18 chromatin states (described in Figure 1.8A), as well as genome-wide (blue), for each of the four cell types under study described in Figure 1.1A: i) MEFs, ii) EARLY intermediates, iii) LATE intermediates, iv) ESCs. The mean DNA methylation level within 200bp windows corresponding to those used for the chromatin state model were used, only considering those windows containing at least one CpG with 5X coverage, in an effort to ensure high-confidence estimates. White circles represent median values. Note: the observed distributions are the same as those in Figure 1.8B.

B. The same distributions are show for Aiv (ESCs) but re-orientated for easier chromatin state-wise comparisons. ‘% coverage’ gives the percentage of 200bp windows with data per chromatin state alternating between observed and model-fitted DNA methylation estimates. ‘% pairwise complete’ gives the percentage of 200bp windows with data in both the observed and model-fitted estimates per state. The metrics labeled ‘methyl15’ and ‘methyl25’ give the percentage of 200bp windows where the difference between the observed and model-fitted DNA methylation estimates was at most 15% or 25%, respectively.
Tables

<table>
<thead>
<tr>
<th></th>
<th>Total Mapped Reads</th>
<th>Mean CpG coverage depth</th>
<th>CpG dimers with 5X coverage</th>
<th>Percentage of CpG dimers covered</th>
<th>Mean methylation per CpG dimer</th>
</tr>
</thead>
<tbody>
<tr>
<td>WGBS MEF</td>
<td>320,292,576</td>
<td>5.48</td>
<td>14,524,871</td>
<td>68.1%</td>
<td>0.72</td>
</tr>
<tr>
<td>WGBS EARLY</td>
<td>429,374,384</td>
<td>7.80</td>
<td>17,202,917</td>
<td>80.6%</td>
<td>0.71</td>
</tr>
<tr>
<td>WGBS LATE</td>
<td>144,045,811</td>
<td>3.47</td>
<td>9,454,085</td>
<td>44.3%</td>
<td>0.67</td>
</tr>
<tr>
<td>WGBS ESC</td>
<td>391,724,853</td>
<td>7.23</td>
<td>17,022,903</td>
<td>79.8%</td>
<td>0.70</td>
</tr>
<tr>
<td>RRBS MEF</td>
<td>9,925,994</td>
<td>11.62</td>
<td>868,693</td>
<td>4.1%</td>
<td>0.49</td>
</tr>
<tr>
<td>RRBS EARLY</td>
<td>12,826,209</td>
<td>12.49</td>
<td>1,289,663</td>
<td>6.0%</td>
<td>0.48</td>
</tr>
<tr>
<td>RRBS LATE</td>
<td>12,822,123</td>
<td>15.71</td>
<td>1,276,884</td>
<td>6.0%</td>
<td>0.48</td>
</tr>
<tr>
<td>RRBS ESC</td>
<td>18,131,716</td>
<td>18.88</td>
<td>1,358,529</td>
<td>6.4%</td>
<td>0.46</td>
</tr>
</tbody>
</table>

Table 1.1 Bisulfite sequencing library total mapped reads, mean CpG coverage depth, CpG dimer coverage, and mean methylation per cell type

WGBS and RRBS libraries for the four cell types under study (MEFs, EARLY intermediates, LATE intermediates, and ESCs as described in Figure 1A). WGBS reads were mapped the whole genome (mm9) and duplicate reads collapsed. RRBS reads were mapped to an in silico MspI-digested reduced mm9 reference genome. Mean CpG coverage determined for CpGs on either strand. *Note: CpG dimer metrics (last three columns) are based on 21,342,493 CpG dimers in mm9 (excluding chrM).
Table 1.2  Significance values for tests of difference of expression values for genes associated with differential DNA methylation cluster groups between the four cell types

Kolmogorov–Smirnov test significance values for tests of difference of the distributions of expression values (log2(RPKM+1)) of genes associated with the six different DMR cluster groups (Figure 2B) between the four cell types in the study (MEFs, EARLY intermediates, LATE intermediates, and ESCs as described in Figure 1A). A gene was said to be associated with a differential DNA methylation cluster group if its promoter (TSS ± 2,500bp) overlapped a 500bp window belonging to that group. If a gene was associated with more than one cluster group, it was assigned to the group, or groups, with the most associated windows. Lower triangle contains raw p-values, while upper triangle contains adjusted p-values using the Holm method. Intensity of red indicates greater significance.
Table 1.3  Gene ontology enrichment analysis results for genes associated with the differential DNA methylation cluster group 3

752 genes whose promoters (TSS +/- 2,500bp) were associated with the differentially methylated windows in cluster group 3 (Figure 2B) and showed differential expression (DE) between the four cell types in the study (MEFs, EARLY intermediates, LATE intermediates, and ESCs as described in Figure 1A). 499 of the DE genes that had counterparts in the biological pathways category in the DAVID gene ontology (GO) database and were then interrogated for GO term enrichment, the results of which are presented in the table. The column labeled ‘Sample genes’ gives a subset of the genes involved. Only GO terms showing an adjusted p-value < 0.05 (by the Benjamini-Hochberg procedure) and fold enrichment >= 2 are listed, sorted by p-value.

<table>
<thead>
<tr>
<th>GO term ID</th>
<th>GO term Description</th>
<th>GO term ID</th>
<th>GO term Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0007049</td>
<td>cell cycle</td>
<td>GO:0006396</td>
<td>RNA processing</td>
</tr>
<tr>
<td>GO:0016071</td>
<td>mRNA metabolic process</td>
<td>GO:0000087</td>
<td>M phase of mitotic cell cycle</td>
</tr>
<tr>
<td>GO:0022402</td>
<td>cell cycle process</td>
<td>GO:0048285</td>
<td>organelle fission</td>
</tr>
<tr>
<td>GO:0000278</td>
<td>mitotic cell cycle</td>
<td>GO:0051276</td>
<td>chromosome organization</td>
</tr>
<tr>
<td>GO:0034660</td>
<td>ncRNA metabolic process</td>
<td>GO:0000280</td>
<td>nuclear division</td>
</tr>
<tr>
<td>GO:0007067</td>
<td>mitosis</td>
<td>GO:006397</td>
<td>mRNA processing</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GO term ID</th>
<th>GO term Description</th>
<th>GO term ID</th>
<th>GO term Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0007049</td>
<td>cell cycle</td>
<td>GO:0006396</td>
<td>RNA processing</td>
</tr>
<tr>
<td>GO:0016071</td>
<td>mRNA metabolic process</td>
<td>GO:0000087</td>
<td>M phase of mitotic cell cycle</td>
</tr>
<tr>
<td>GO:0022402</td>
<td>cell cycle process</td>
<td>GO:0048285</td>
<td>organelle fission</td>
</tr>
<tr>
<td>GO:0000278</td>
<td>mitotic cell cycle</td>
<td>GO:0051276</td>
<td>chromosome organization</td>
</tr>
<tr>
<td>GO:0034660</td>
<td>ncRNA metabolic process</td>
<td>GO:0000280</td>
<td>nuclear division</td>
</tr>
<tr>
<td>GO:0007067</td>
<td>mitosis</td>
<td>GO:006397</td>
<td>mRNA processing</td>
</tr>
</tbody>
</table>

**Table 1.3**  Gene ontology enrichment analysis results for genes associated with the differential DNA methylation cluster group 3

752 genes whose promoters (TSS +/- 2,500bp) were associated with the differentially methylated windows in cluster group 3 (Figure 2B) and showed differential expression (DE) between the four cell types in the study (MEFs, EARLY intermediates, LATE intermediates, and ESCs as described in Figure 1A). 499 of the DE genes that had counterparts in the biological pathways category in the DAVID gene ontology (GO) database and were then interrogated for GO term enrichment, the results of which are presented in the table. The column labeled ‘Sample genes’ gives a subset of the genes involved. Only GO terms showing an adjusted p-value < 0.05 (by the Benjamini-Hochberg procedure) and fold enrichment >= 2 are listed, sorted by p-value.
<table>
<thead>
<tr>
<th>GO term ID</th>
<th>GO term Description</th>
<th>DE genes assoc. w/ GO term</th>
<th>DE genes %</th>
<th>mm's genes assoc. w/ GO term</th>
<th>mm's gene %</th>
<th>mouse gene %</th>
<th>Fold Enrichment</th>
<th>Modified Fisher EASE score</th>
<th>Adjusted P-value (Benjamini)</th>
<th>Sample of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0051253</td>
<td>negative regulation of RNA metabolic process</td>
<td>56</td>
<td>866</td>
<td>6.47</td>
<td>310</td>
<td>13,588</td>
<td>2.28</td>
<td>2.83</td>
<td>2.87E-12</td>
<td>3.85E-09</td>
</tr>
<tr>
<td>GO:0010558</td>
<td>negative regulation of macromolecule biosynthetic process</td>
<td>67</td>
<td>866</td>
<td>7.74</td>
<td>418</td>
<td>13,588</td>
<td>3.08</td>
<td>2.51</td>
<td>4.67E-12</td>
<td>4.19E-09</td>
</tr>
<tr>
<td>GO:0045892</td>
<td>negative regulation of transcription, DNA-dependent</td>
<td>56</td>
<td>866</td>
<td>6.47</td>
<td>308</td>
<td>13,588</td>
<td>2.27</td>
<td>2.85</td>
<td>2.19E-12</td>
<td>5.90E-09</td>
</tr>
<tr>
<td>GO:0051172</td>
<td>negative regulation of nitrogen compound metabolic process</td>
<td>64</td>
<td>866</td>
<td>7.39</td>
<td>401</td>
<td>13,588</td>
<td>2.95</td>
<td>2.50</td>
<td>1.83E-11</td>
<td>6.16E-09</td>
</tr>
<tr>
<td>GO:0045934</td>
<td>negative regulation of nucleosides, nucleotides, nucleic acids and metabolic process</td>
<td>64</td>
<td>866</td>
<td>7.39</td>
<td>397</td>
<td>13,588</td>
<td>2.92</td>
<td>2.53</td>
<td>1.17E-11</td>
<td>6.30E-09</td>
</tr>
<tr>
<td>GO:0006974</td>
<td>response to DNA damage stimulus</td>
<td>52</td>
<td>866</td>
<td>6.00</td>
<td>287</td>
<td>13,588</td>
<td>2.11</td>
<td>2.84</td>
<td>1.71E-11</td>
<td>6.56E-09</td>
</tr>
<tr>
<td>GO:0010605</td>
<td>negative regulation of macromolecule metabolic process</td>
<td>75</td>
<td>866</td>
<td>8.86</td>
<td>506</td>
<td>13,588</td>
<td>3.72</td>
<td>2.33</td>
<td>3.04E-11</td>
<td>7.02E-09</td>
</tr>
<tr>
<td>GO:0031327</td>
<td>negative regulation of cellular biosynthetic process</td>
<td>67</td>
<td>866</td>
<td>7.74</td>
<td>430</td>
<td>13,588</td>
<td>3.16</td>
<td>2.44</td>
<td>1.70E-11</td>
<td>7.64E-09</td>
</tr>
<tr>
<td>GO:0009890</td>
<td>negative regulation of biosynthetic process</td>
<td>67</td>
<td>866</td>
<td>7.74</td>
<td>434</td>
<td>13,588</td>
<td>3.19</td>
<td>2.42</td>
<td>2.58E-11</td>
<td>7.72E-09</td>
</tr>
<tr>
<td>GO:0010629</td>
<td>negative regulation of gene expression</td>
<td>63</td>
<td>866</td>
<td>7.27</td>
<td>410</td>
<td>13,588</td>
<td>3.02</td>
<td>2.41</td>
<td>1.35E-10</td>
<td>3.64E-08</td>
</tr>
<tr>
<td>GO:0016481</td>
<td>negative regulation of transcription</td>
<td>59</td>
<td>866</td>
<td>6.81</td>
<td>372</td>
<td>13,588</td>
<td>2.74</td>
<td>2.49</td>
<td>1.62E-10</td>
<td>3.95E-08</td>
</tr>
<tr>
<td>GO:0006259</td>
<td>DNA metabolic process</td>
<td>62</td>
<td>866</td>
<td>7.16</td>
<td>421</td>
<td>13,588</td>
<td>3.10</td>
<td>2.31</td>
<td>1.09E-09</td>
<td>2.44E-07</td>
</tr>
<tr>
<td>GO:0033554</td>
<td>cellular response to stress</td>
<td>60</td>
<td>866</td>
<td>6.93</td>
<td>404</td>
<td>13,588</td>
<td>2.97</td>
<td>2.33</td>
<td>1.52E-09</td>
<td>3.15E-07</td>
</tr>
<tr>
<td>GO:0006281</td>
<td>DNA repair</td>
<td>40</td>
<td>866</td>
<td>4.62</td>
<td>222</td>
<td>13,588</td>
<td>1.63</td>
<td>2.83</td>
<td>5.92E-09</td>
<td>1.14E-06</td>
</tr>
<tr>
<td>GO:0000122</td>
<td>negative regulation of DNA polymerase II promoter</td>
<td>39</td>
<td>866</td>
<td>4.50</td>
<td>231</td>
<td>13,588</td>
<td>1.70</td>
<td>2.65</td>
<td>5.77E-08</td>
<td>1.03E-05</td>
</tr>
<tr>
<td>GO:0051276</td>
<td>chromosome organization</td>
<td>54</td>
<td>866</td>
<td>6.24</td>
<td>404</td>
<td>13,588</td>
<td>2.97</td>
<td>2.10</td>
<td>3.63E-07</td>
<td>5.72E-05</td>
</tr>
<tr>
<td>GO:0009792</td>
<td>embryonic development ending in birth or egg hatching</td>
<td>56</td>
<td>866</td>
<td>6.47</td>
<td>425</td>
<td>13,588</td>
<td>3.13</td>
<td>2.07</td>
<td>3.63E-07</td>
<td>5.73E-05</td>
</tr>
<tr>
<td>GO:0043009</td>
<td>chordate embryonic development</td>
<td>55</td>
<td>866</td>
<td>6.35</td>
<td>421</td>
<td>13,588</td>
<td>3.10</td>
<td>2.05</td>
<td>6.12E-07</td>
<td>6.66E-05</td>
</tr>
<tr>
<td>GO:0051301</td>
<td>cell division</td>
<td>41</td>
<td>866</td>
<td>4.73</td>
<td>281</td>
<td>13,588</td>
<td>2.07</td>
<td>2.29</td>
<td>1.35E-06</td>
<td>1.65E-04</td>
</tr>
<tr>
<td>GO:0000279</td>
<td>M phase</td>
<td>39</td>
<td>866</td>
<td>4.50</td>
<td>283</td>
<td>13,588</td>
<td>2.08</td>
<td>2.16</td>
<td>9.99E-06</td>
<td>1.17E-03</td>
</tr>
</tbody>
</table>

**Table 1.4** Gene ontology enrichment analysis results for genes associated with the differential DNA methylation cluster group 4

As for Table 1.3, but for 1,296 DE genes associated with differentially methylated windows in cluster group 4, 866 of which had counterparts in the biological pathways gene set. Only the top 20 GO terms by p-value are listed (32 terms had an adj. p-value < 0.05 and fold change >=2).
### Table 1.5  Gene ontology enrichment analysis results for genes associated with the differential DNA methylation cluster group 5

As for Table 1.3, but for 639 DE genes associated with differentially methylated windows in cluster group 5, of which 442 had counterparts in the biological pathways gene set. Only the top 20 GO terms by p-value are listed (158 terms had an adj. p-value < 0.05 and fold change >=2).
<table>
<thead>
<tr>
<th>Hitone modifications (for each cell type)</th>
<th>MEFs</th>
<th>EARLY</th>
<th>LATE</th>
<th>ESCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3K4ME2</td>
<td>-0.37</td>
<td>-0.45</td>
<td>-0.29</td>
<td>-0.59</td>
</tr>
<tr>
<td>H3K27AC</td>
<td>-0.31</td>
<td>-0.34</td>
<td>-0.20</td>
<td>-0.46</td>
</tr>
<tr>
<td>H3K9AC</td>
<td>-0.27</td>
<td>-0.31</td>
<td>-0.20</td>
<td>-0.44</td>
</tr>
<tr>
<td>H3K4ME3</td>
<td>-0.25</td>
<td>-0.39</td>
<td>-0.17</td>
<td>-0.43</td>
</tr>
<tr>
<td>H3K4ME1</td>
<td>-0.14</td>
<td>-0.13</td>
<td>-0.13</td>
<td>-0.36</td>
</tr>
<tr>
<td>H3.3</td>
<td>-0.12</td>
<td>-0.19</td>
<td>-0.09</td>
<td>-0.20</td>
</tr>
<tr>
<td>H3K27ME3</td>
<td>-0.06</td>
<td>-0.11</td>
<td>-0.07</td>
<td>-0.10</td>
</tr>
<tr>
<td>H3</td>
<td>-0.02</td>
<td>-0.05</td>
<td>-0.09</td>
<td>-0.20</td>
</tr>
<tr>
<td>H3K79ME2</td>
<td>-0.01</td>
<td>-0.08</td>
<td>-0.04</td>
<td>-0.18</td>
</tr>
<tr>
<td>H3K9ME3</td>
<td>0.01</td>
<td>0.01</td>
<td>0.06</td>
<td>0.02</td>
</tr>
<tr>
<td>H3K36ME3</td>
<td>0.22</td>
<td>0.21</td>
<td>0.20</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Table 1.6  CpG dimer-level correlations between WGBS DNA methylation estimates and nine histone modifications and two histone variants in each cell type

Pearson correlation values between WGBS CpG dimer-level DNA methylation estimates and the binned read counts within 100bp windows surrounding each CpG dimer for the specified histone modifications and variants (H3 and H3.3) for the four cell types under study (MEFs, EARLY intermediates, LATE intermediates, and ESCs as described in Figure 1A). Red intensity signifies the strength of a positive correlation, while blue intensity signifies the strength of the anti-correlation.
Table 1.7  DNA methylation model coefficients based on WGBS data

Each of the four sub-tables contains coefficients, associated standard errors, and statistical significance for a logistic regression model of DNA methylation for the four cell types under study (MEFs, EARLY intermediates, LATE intermediates, and ESCs as described in Figure 1A) using normalized ChIP-seq read counts for nine histone modification and two histone variants (H3 and H3.3) within 100bp windows around CpG dimers as predictors. Coefficients and standard error were transformed from log-odds to percentage change in DNA methylation. Pseudo R² values were calculated using the formula: 1 – ( deviance / deviance for null model ).

See text for further details. Red intensity signifies the strength of a positively-related coefficients, while blue intensity signifies the strength of negatively-related coefficients. Coefficients ranked by absolute value.
Table 1.8  DNA methylation model coefficients based on RRBS

As for Table 1.7, but based on RRBS data.
<table>
<thead>
<tr>
<th></th>
<th>Pseudo R-squared</th>
<th>Pearson correlation</th>
<th>CpG dimers fitted</th>
<th>Percentage of CpG dimers fitted</th>
<th>Mean methylation per CpG dimer</th>
</tr>
</thead>
<tbody>
<tr>
<td>WGBS MEF</td>
<td>0.19</td>
<td>0.46</td>
<td>13,934,741</td>
<td>65.3%</td>
<td>0.72</td>
</tr>
<tr>
<td>WGBS EARLY</td>
<td>0.25</td>
<td>0.54</td>
<td>16,512,583</td>
<td>77.4%</td>
<td>0.72</td>
</tr>
<tr>
<td>WGBS LATE</td>
<td>0.13</td>
<td>0.40</td>
<td>9,278,255</td>
<td>43.5%</td>
<td>0.67</td>
</tr>
<tr>
<td>WGBS ESC</td>
<td>0.39</td>
<td>0.67</td>
<td>16,469,438</td>
<td>77.2%</td>
<td>0.70</td>
</tr>
<tr>
<td>RRBS MEF</td>
<td>0.53</td>
<td>0.78</td>
<td>843,484</td>
<td>4.0%</td>
<td>0.49</td>
</tr>
<tr>
<td>RRBS EARLY</td>
<td>0.66</td>
<td>0.85</td>
<td>1,256,680</td>
<td>5.9%</td>
<td>0.48</td>
</tr>
<tr>
<td>RRBS LATE</td>
<td>0.61</td>
<td>0.82</td>
<td>1,246,223</td>
<td>5.8%</td>
<td>0.48</td>
</tr>
<tr>
<td>RRBS ESC</td>
<td>0.74</td>
<td>0.89</td>
<td>1,323,409</td>
<td>6.2%</td>
<td>0.46</td>
</tr>
</tbody>
</table>

Table 1.9 DNA methylation model statistics for all four cell types

Pseudo R-squared, Pearson correlation, coefficients, number and percentage of CpG dimers fitted, and mean fitted methylation estimates per CpG dimer. *Note: Based on 21,342,493 CpG dimers in mm9 (excluding chrM).
References


Chapter 2

Determining differential DNA methylation using methylation-sensitive restriction enzyme bisulfite sequencing
2.1 Abstract

Whole-genome bisulfite sequencing (WGBS) and reduced representation bisulfite sequencing (RRBS) allow methylation levels of cytosines to be determined on a genome-wide scale. However, obtaining sufficient coverage levels for WGBS can be very costly, especially for large genomes. While RRBS greatly reduces these costs, it only interrogates 5–10% of the genome. Here, we introduce methylation-sensitive restriction enzyme bisulfite sequencing (MREBS) as means to determine differential DNA methylation between two samples that offers a compromise between WGBS’s broad coverage and RRBS’s reduced cost. We show that MREBS DNA methylation estimates follow similar trends across different chromatin states to those of estimates based on WGBS and RRBS data. Differential DNA methylation values based on MREBS data correlate well with those based on WGBS and RRBS data at the CpG level. Conversely, differential MREBS read counts within 1kb windows around each CpG anti-correlate with WGBS and RRBS differential DNA methylation values. We build a multiple regression model the combined MREBS differential DNA methylation values and MREBS differential read count data that predicted WGBS-based differential DNA methylation values more accurately than did a model based on RRBS data, albeit for a slightly smaller proportion of the genome (~1.5% vs. 3%). A big advantage of MREBS was that a model based on MREBS differential read data alone to provided differential DNA estimates across close to 60% of the genome.
2.2 Introduction

DNA methylation plays an important role in gene regulation and cell identity, although its specific mechanism of action is still not fully understood, making it a very active area of scientific inquiry\textsuperscript{1-3}. Recent developments in sequencing technologies have fueled this research by enabling the production of DNA methylation mapping on an unprecedented scale\textsuperscript{4-9}.

The two most popular methods for assessing DNA methylation levels on a genome-wide scale are whole-genome bisulfite sequencing (WGBS, also known as BS-seq, methyl-seq, or methylC-seq)\textsuperscript{10-12} and reduced representation bisulfite sequencing (RRBS)\textsuperscript{13,14}. As implied by their names, both are bisulfite-based protocols, where the methylation status of cytosines can be estimated as the proportion protected from conversion to uracils by the bisulfite treatment. WGBS is far more comprehensive and can theoretically assess the methylation status of every single cytosine in the mappable genome, but requires a tremendously deep level of sequencing, and, hence, can be very costly, especially if one is working with an organism that has a large genome. RRBS interrogates a smaller, CpG-rich portion of the genome, thereby significantly reducing the amount of sequencing required. The RRBS protocol achieves this by introducing a step where the genomic DNA is first digested. Typically the methylation insensitive restriction endonuclease MspI is used with the recognition sequence C\textsuperscript{\textregistered}CGG. By enriching for CpG-rich regions (5-10\% of the genome), one arguably includes most regions that are of regulatory interest, such as promoters and enhancers\textsuperscript{14}, but this represents a rather narrow view of the entire epigenomic landscape.

In order to address the respective deficiencies of WGBS and RRBS, we developed a new method that seeks a compromise between the two, which we call methylation-sensitive
restriction enzyme bisulfite sequencing (MREBS) as it adds a bisulfite step to an existing protocol (MRE-seq\textsuperscript{15}). MRE-seq typically utilizes three methylation-sensitive restriction endonucleases, HpaII (C\textsuperscript{CGG}), HinP1I (G\textsuperscript{CGC}), and AciI (C\textsuperscript{CGC}), in parallel to digest DNA, and is therefore also known as triple-MRE-seq (Table 2.1). Areas exhibiting a high density of MRE-seq aligned reads should represent demethylated or lowly methylated regions of the genome within the population of cells comprising the sample. More specifically, the cut sites of the DNA fragments would necessarily be demethylated due to the sensitivity of the endonucleases. However, subsequent bases may not necessarily be methylated to the same extent and bisulfite treatment could provide additional information regarding the DNA methylation levels downstream of these cut sites, which was the motivation for the additional step in MREBS.

MREBS reads from would be expected to show a strong bias for demethylated regions, and not provide a fair indication of absolute DNA methylation levels. Therefore, we envisaged that the data would be best for determining differentially methylated regions (DMRs) between a pair of samples. The big advantage afforded by MREBS is that the use of three methylation sensitive enzymes to digest the genome would result in higher coverage levels, and one could determine differential DNA methylation in two different ways: 1) Based on read depth alone, regardless of bisulfite treatment-based results, because due to the use of methylation sensitive digestion, read counts around CpGs should anti-correlate with DNA methylation levels, allowing one to determine differential methylation between two samples by comparing the densities of their mapped reads, as can be done with traditional MRE-seq. 2) Additionally, for CpGs with
sufficient read coverage one could obtain estimates of differential DNA methylation directly with high confidence, just as one would typically do when using WGBS and RRBS data.

To test our approach, we first compared MREBS-based methylation estimates and coverage to those based on WGBS and RRBS data, using two mouse cell types between which we expected to see substantial differential methylation. We found that MREBS-based DNA methylation estimates showed similar trends to WGBS- and RRBS-based values across chromatin states and that the proportion of CpG dimers with read coverage high enough to obtain relatively high confidence in methylation estimates (5X) was comparable to the proportion of CpGs using RRBS. Additionally, nearly 60% of all CpG dimers had a minimum of two reads falling within the surrounding 1kb window in at least one of the two cell types. We employed a multiple regression model using both MREBS differential DNA methylation and differential read count data to predict differential DNA methylation values for ~3% of CpGs and produced estimates that compared favorably to those obtained from models built using RRBS methylation data alone. Additionally, differential DNA methylation estimates for far greater proportion of CpGs (~60%) could be obtained using a model that only used MREBS differential read data with surrounding 1kb windows, and that corresponded well to WGBS-based differential values.
2.3 Results

2.3.1 Study design and data sets

Recognizing that MREBS was best suited for determining differential DNA methylation, we required a pair of samples that we knew to be differentially methylated to effectively evaluate the approach. Since we had produced WGBS libraries for four cell types that represented different stages during the somatic cell reprogramming process, as discussed in the preceding chapter, we selected two stages that we had shown to exhibit differential DNA methylation, namely 1) MEFs that were induced to ectopically express the so-called Yamanaka reprogramming factors, namely OCT4 (O), SOX2 (S), KLF4 (K), and MYC (M; also known as ‘c-MYC’)\textsuperscript{16} for 48 hours, representing early somatic cell reprogramming (EARLY), and 2) mouse embryonic stem cells (ESCs) representing the fully reprogrammed state, since they are more readily available and abundant than iPSCs, but essentially equivalent (Figure 2.1A).

WGBS libraries for the two cell types were sequenced deeply enough to ensure average coverage of over 7 reads per CpG after mapping to the mm9 genome using BS-Seeker\textsuperscript{2}\textsuperscript{17} (Figure 2.1B, Table 2.2). The comparable sequencing depth and CpG coverage level of the pair of samples was another reason for choosing these two WGBS samples for testing and comparison purposes (Figure 2.1Bi vs. ii, Tables 2.2 and 2.3).

RRBS and MREBS libraries for the same cell lines were generated, with the MREBS libraries in duplicate. The RRBS libraries were mapped to an \textit{in silico} MspI-digested reduced reference mm9 genome using BS-Seeker2 (Figure 2.1C). The MREBS reads were first filtered for those with the expected 5’ cut sites and mapped to the whole genome, also using BS-Seeker2.
(Figure 2.1D). The total mapped reads for WGBS were easily an order of magnitude greater than that seen for the MREBS libraries, for a comparable coverage depth per CpG (Figure 2.1, Table 2.2).

2.3.2 MREBS DNA methylation estimates follow the same trends was WGBS and RRBS in different chromatin states

To compare DNA methylation estimates not only on the genome-wide scale, but in different genomic contexts, we took advantage of a hidden Markov model of chromatin states generated by chromHMM$^{18}$. The model will be described in detail in a separate manuscript (Chronis et al., in preparation), but briefly, it was on based on ChIP-seq data sets for the 10 histone modifications, including the native input library for each of the three cell types described in Figure 2.1A, namely MEFs (i), EARLY intermediates (ii), and ESCs (iii), as well as partially induced pluripotent cells (pre-iPSCs), a late reprogramming intermediates otherwise not used in this study. Using this model, the genomes of the two cell types for which we had DNA methylation estimates (EARLY intermediates and ESCs) were tiled into 200bp windows and classified as belonging to one of 18 chromatin states. Candidate functional annotations were assigned to each of the 18 states based on the prevalence and combination of histone mark peaks (Figure 2.2A).

The mean DNA methylation values within the 13.3 million 200bp windows corresponding to those used for the chromatin state model were calculated based on the CpG-level estimates for each of the three bisulfite sequencing methods, for each cell type. Only those windows containing at least one CpG with 5X coverage being assigned a value in an effort to
ensure high-confidence estimates. Distributions of the mean DNA methylation values for all the windows belonging to each chromatin state, as well as genome-wide, were then plotted for each cell type (EARLY intermediates (Figure 2.2B) and ESCs (Figure 2.2C)) and for each bisulfite sequencing method (WGBS (i), RRBS (ii), and MREBS (iii)). The different chromatin states showed characteristic DNA methylation distributions that were relatively similar in both cell types (Figure 2.2B/C). For instance, the promoter-associated chromatin states (1 and 2) were comparatively hypomethylated, while several of the enhancer-related chromatin states (3, 4, 5, and 7) showed a wide spread of DNA methylation levels and an intermediate mean DNA methylation. Most of the other chromatin states were largely hypermethylated (Figure 2.2B/C).

There are observable differences between the cell type, but the differences between sequencing methods are perhaps more obvious. Most particularly, the distributions of the MREBS-based DNA methylation estimates are systematically lower in all chromatin states (Figure 2.2B/Ciii). Most notably the genome-wide DNA methylation levels based on MREBS estimates are very low, close to that of the more demethylated states (blue violin plots). This implies that the MREBS DNA methylation calls are largely obtained from these lowly methylated regions of the genome.

To confirm this, Figure 2.3 shows the proportion of 200bp windows with DNA methylation estimates found in each of the 18 chromatin states using i) WGBS, RRBS), and iii) MREBS data for the EARLY intermediate (A) and ESC (B) samples. The accompanying bar plots (v-vii) show that the MREBS samples are particularly enriched for chromatin states 1–5, 7, and 15, which also happen to show the lowest DNA methylation levels (Figure 2.3).
This analysis showed that although MREBS DNA methylation estimates are systematically lower than those obtained using WGBS and RRBS data, they follow similar trends across different chromatin states. These estimates could therefore be utilized to determine differential methylation levels between two samples, if scaled appropriately, or incorporated into a model to predict differential DNA methylation.

2.3.3 MREBS read counts provide additional information for determining differential DNA methylation

The proportion of the 21.3 million CpGs in the mouse genome covered with a minimum of 5X coverage was far higher for the WGBS samples (~80%), than for either the RRBS (6%) or the MREBS samples (4–5%), as expected (Table 2.3). However, 64–69% of CpG dimers had at least one MREBS read falling within a surrounding 1kb window, 42–48% had two or more reads, 22–25% had five or more, and 15-17% had ten or more (Table 2.3).

Since we envisioned that MREBS would be best utilized for determining DMRs between a pair of samples, we also looked at the proportion of CpGs exhibiting differential DNA methylation or differential read coverage (Table 2.4). In the case of DNA methylation, this meant the proportion of CpGs in both samples with at least 5X coverage, and from the point of view of counts, this was the proportion of CpGs with two or more reads in the surrounding 1kb window in at least one sample. As for the individual samples, the WGBS samples showed far higher pairwise 5X coverage (75.5%) than for either the RRBS (5.6%) or the MREBS samples (~3%) (Table 2.4). However, nearly 60% of CpG dimers had at least two reads falling within the
surrounding 1kb window in at least one of the two cell types, and this was the case for both pairs of replicates (Table 2.4).

Since MREBS utilizes methylation sensitive digestion, read counts around CpGs should anti-correlate with their methylation levels. Therefore, MREBS read counts within windows around CpGs can be used to assign differential methylation estimates to the CpGs in question, thereby providing broader coverage than one would get relying on high confidence DNA methylation calls in both samples.

### 2.3.4 MREBS DNA methylation estimates correlate and read counts anti-correlate with WGBS and RRBS DNA methylation estimates

We next looked to confirm that MREBS DNA methylation estimates did indeed correlate and MREBS binned reads counts would anti-correlate with WGBS and RRBS DNA methylation estimates.

The DNA methylation levels of both EARLY MREBS replicates correlated more closely with the EARLY WGBS sample \( (r = 0.77 \text{ and } r = 0.78, \text{ respectively}) \) than they did with the ESC WGBS sample \( (r = 0.53 \text{ in both cases}) \) (Table 2.5). And EARLY MREBS replicates also correlated more closely with the EARLY RRBS sample \( (r = 0.82 \text{ in both cases}) \) than they did with the ESC RRBS sample \( (r = 0.66 \text{ in both cases}) \) (Table 2.5). Conversely, the DNA methylation levels of the both the ESC MREBS replicates both correlated more closely with the ESC WGBS sample \( (r = 0.70 \text{ and } r = 0.71, \text{ respectively}) \) than they did with the EARLY WGBS sample \( (r = 0.53 \text{ in both cases}) \) (Table 2.5). And the ESC MREBS replicates also correlated more
closely with the ESC RRBS sample ($r = 0.80$ in both cases) than they did with the EARLY RRBS sample ($r = 0.64$ in both cases) (Table 2.5).

Furthermore, as expected, MREBS read counts anti-correlated with the DNA methylation levels based on WGBS-, RRBS-, and MREBS data, though somewhat surprisingly not in a particularly cell type-specific manner (Table 2.5).

However, MREBS differential read counts (EARLY - ESC) anti-correlated more strongly with the differential DNA methylation values based on WGBS ($r = -0.34$ for both replicate pairs), RRBS ($r = -0.51$ and $r = -0.52$ for the two respective replicate pairs), and MREBS ($r = -0.34$ and $r = -0.35$ for the two respective replicate pairs) (Table 2.6). MREBS differential DNA methylation correlated positively with those of WGBS ($r = 0.55$ and $r = 0.56$ for the two respective replicate pairs) and RRBS ($r = 0.64$ for both replicate pairs) (Table 2.6).

These results indicated that (i) the MREBS differential data correlated with the differential WGBS and RRBS differential DNA methylation in the expected directions, and (ii) that the MREBS differential DNA methylation and MREBS differential read counts could be used together to better estimate differential DNA methylation.

### 2.3.5 CpG-level DNA methylation can be modeled using MREBS data

The strong correlations discussed in the previous section suggested that WGBS differential DNA methylation values might be modeled using the MREBS data. To investigate this possibility, I built the following four different linear regression models using R's `lm()` function$^{19}$, coefficients for which are provided in Table 2.7:
1) \[ y_i = \beta_0 + \beta r_i \ (i = 1..n) \]
where \( y \) represents WGBS differential DNA methylation, and \( r \) represents RRBS differential DNA methylation, both at CpG dimer level \((n = 666,214)\).

2) \[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \ (i = 1..n) \]
where \( y \) represents WGBS differential DNA methylation per CpG dimer, \( x_1 \) represents MREBS differential DNA methylation per CpG dimer, and \( x_2 \) represents MREBS differential read count within the 1kb window around each CpG dimer \((n = 318,400 \) for MREBS replicate 1 and \( n = 322,431 \) for the second MREBS replicate).

3) \[ y_i = \beta_0 + \beta_1 x_{i1} \ (i = 1..n) \]
where \( y \) represents WGBS differential DNA methylation per CpG dimer, and \( x_1 \) represents MREBS differential DNA methylation per CpG dimer \((n = 319,304 \) for MREBS replicate 1 and \( n = 323,431 \) for MREBS replicate 2).

4) \[ y_i = \beta_0 + \beta_2 x_{i2} \ (i = 1..n) \]
where \( y \) represents WGBS differential DNA methylation per CpG dimer, and \( x_2 \) represents MREBS differential read count within the 1kb window around each CpG dimer \((n = 9,485,471 \) for MREBS replicate 1 and \( n = 9,670,440 \) for MREBS replicate 2).

Model 1 was used for comparison purposes and shows that WGBS-based differential DNA methylation values (EARLY - ESC) can be modeled using RRBS data for \( \sim 3\% \) of the CpG dimers (666,214) in the mouse genome, which represents sites with sufficient coverage (5X) in
both the WGBS (75.5%) and RRBS (5.6%) samples. The model had an $R^2 = 0.39$, implying a correlation between the WGBS estimated differential DNA methylation values and the model-fitted one of $r = 0.63$ (Table 2.8). The root-mean-square error (RMSE) between the observed and fitted values was 20.4% and the mean absolute error (MAE) was 15.2%.

The metrics ‘methyl15’ and ‘methyl25’ give the percentage of CpG dimers where the difference between the WGBS differential DNA methylation estimate and that of the model was at most 15% and 25%, respectively. Using the more lenient methyl25 comparison, the RRBS-based model-fitted DNA methylation values show 80% concordance with the WGBS-based estimates (Table 2.8).

Model 2 is a multiple regression model using both MREBS differential DNA methylation (EARLY - ESC) and MREBS differential read count data to predict WGBS-based differential DNA methylation values. A model was built for each replicate pair. The fits for both replicates are similar ($R^2 = 0.35$ and $R^2 = 0.36$) and similar to that obtained using the RRBS data (Table 2.8). Interestingly, both RMSE ($\sim 18\%$ for both replicates) and MAE ($\sim 12.5\%$ for both replicates) values are better than that seen for the RRBS-based model. The concordance metrics (methyl15 = $\sim 73\%$ and methyl25 = $\sim 86\%$, for both replicates) were also better than seen for the RRBS-based model (Table 2.8).

However, the differential DNA methylation values of only 1.5% ($n = \sim 320$ thousand) CpG dimers genome-wide could be predicted, half that using the RRBS data ($n = 666,214$ or $\sim 3\%$ of CpG dimers), although this determined to large extent by sequencing depth.
Model 3 and 4 are simple linear models using each of the two independent variables from model 2, respectively. The fit for model 3, using only MREBS differential DNA methylation (EARLY - ESC), somewhat worse without the additional count data (R² = 0.30 and R² = 0.31 for the replicate pairs), but, interestingly, the RMSE (~18.5% for both replicates) and MAE (~12.5% for both replicates) is still superior to that of the RRBS based model, as are the concordance metrics (methyl15 = ~73% and methyl25 = ~85%, for both replicates) (Table 2.8).

Model 4 is based only on MREBS differential read counts within 1kb window around each CpG (EARLY - ESC). Only those CpGs with at least two reads in the surrounding +/- 500bp in at least one sample were considered, amounting to 12.5 (58.8%) and 12.8 (59.7%) million CpG dimers for MREBS replicate 1 and 2, respectively (Table 2.8). The represents ~10X more CpG dimers than available for use in the RRBS-based model 1 (1.2 million CpG dimers), and ~20X more CpG dimers than available for use in models 2 and 3 based on MREBS sites with 5X coverage in both samples (~648–665 thousand CpG dimers).

Even though the fits for model 4 are worse than that based on the MREBS DNA methylation estimates (R² = 0.11 for both replicate pairs), the RMSE (~24–25%) and MAE (~18.5%) are not that much worse than seen for the RRBS based model, nor are the concordance metrics (methyl15 = ~53% and methyl25 = ~73%) (Table 2.8).

However, the differential DNA methylation values for 44–45% of CpGs were be predicted using this model, representing the overlap of those CpG dimers with 5X WGBS coverage (75.5%) and with at least 2 MREBS reads within the surrounding 1kb window (58.8 and 59.7% for the respective MREBS replicates).
To get the benefit of both the extended coverage of model 4 and the improved accuracy of model 2, I combined their results, updating the model 4 estimates with those of model 2 where available. This marginally improved all the applicable metrics discussed previously (Table 2.8). Figure 2.4 shows how these combined modeled differential DNA estimates (iii, green tracks, two replicates) compared to WGBS (i, dark blue tracks) and RBBS (ii, light blue tracks) at different length scales: 611kb (A), 19kb (B), and an extended locus partitioned in three 18kb panels (C). Below the modeled estimates are tracks showing the MREBS-based differential DNA methylation (iv, orange, two replicates) and MREBS-based differential read counts (v, red, two replicates), which were used in the models. While the MREBS differential DNA methylation coverage is comparable to that of the RRBS data (cf. tracks iv and ii, Table 2.8), the MREBS differential read count coverage is more comparable to that of the WGBS data (cf. tracks v and i, Table 2.8), and it is clear that the majority of the differential DNA estimates are driven by the read count data (model 4). Nonetheless, the modeled differential DNA methylation appears to track WGBS-based estimates reasonably faithfully for regions that are more methylated in the EARLY intermediates (Figures 2.4B/C and 2.5A), as well as regions that are more methylated in the ESCs (Figures 2.4A and 2.5B).
2.4 Discussion

Whole-genome bisulfite sequencing (WGBS)\textsuperscript{11} and reduced representation bisulfite sequencing (RRBS)\textsuperscript{14} are two popular methods for assessing DNA methylation levels on a genome-wide scale. WGBS can potentially determine the methylation status of every single cytosine, but the amount of sequencing required to obtain sufficient coverage to do so can be prohibitively expensive. Sequencing costs can be significantly lowered by RRBS, but one incurs an 80–90\% loss in cytosine coverage. In order to address these respective shortcomings, we introduce methylation-sensitive restriction enzyme bisulfite sequencing (MREBS), which adds a bisulfite conversion step to the existing protocol, methylation-sensitive restriction enzyme digestion followed by high-throughput sequencing (MRE-seq)\textsuperscript{15}.

As expected, due to the methods reliance on methylation sensitive endonucleases, the distributions of the MREBS-based DNA methylation estimates were systematically lower than those obtained using WGBS or RRBS data. However, the MREBS-based estimates followed similar trends across all chromatin states (Figure 2.2). Equally unsurprisingly, high-confidence MREBS methylation estimates (CpGs with 5X coverage) were particularly enriched in chromatin states with the lowest DNA methylation levels (Figure 2.3). Since these chromatin states are known to be highly important with respect to gene regulation, their enrichment at lower sequencing cost is beneficial.

For the MREBS libraries, 4.3–4.6\% of CpG dimers possessed the 5X read coverage required to make relatively confident DNA methylation level calls, which was comparable to the coverage obtained from RRBS libraries at a similar level of sequencing. However, close on 60\%
of CpG dimers had two or more reads falling within the surrounding 1kb window in at least one of the two cell types (Table 2.4). These data can potentially be used for estimating differential DNA methylation of a high proportion of CpGs, since MREBS utilizes methylation sensitive digestion and therefore read counts around CpGs should anti-correlate with their methylation levels, which we showed to be the case (Tables 2.5 and 2.6).

To obtain estimates of differential DNA methylation based on MREBS data, we built a multiple regression model that incorporate both MREBS differential DNA methylation and MREBS differential read count data to predict differential DNA methylation values for ~3% of CpGs. The fits for both replicates were similar and correspondence metrics comparing the model-fitted values to WGBS estimates were superior to those obtained from models built using MREBS or RRBS methylation data alone (Table 2.8). Furthermore, differential DNA methylation estimates for a far greater proportion of CpGs (~60%) could be obtained using a model that used only MREBS differential read data with surrounding 1kb windows. Although the fit was not as good, estimates still corresponded to WGBS-based differential values (Table 2.8 and Figures 2.4 and 2.5).
Methylation-sensitive restriction enzyme bisulfite sequencing (MREBS)

Three enzymatic digestions were performed on 1 µg of purified genomic DNA using 10 U of each one of the MRE restriction enzymes (*HpaII*, *Hin6* and *AciI* - Fermentas) in a 50 µl final volume with TANGO buffer. 2.5 µl of RNase cocktail mix (Ambion) were added and the reaction was incubated overnight at 37°C. After the digestion, the three reactions were pooled and the DNA was purified using AMPure XP beads (Beckman Coulter). Subsequent reactions of DNA End Repair, A-tailing and Adapter Ligation were performed using Illumina TruSeq reagents, following manufacturer’s instructions and the DNA was size selected between 200 and 500 bp using AMPure XP beads. Size selected DNA was then treated with bisulfite using the EpiTect kit (QIAGEN) according to the protocol suggested from the manufacturer, except that the conversion step was performed twice, for a total time of 10 h. For each bisulfite-converted sample, two parallel PCR reactions were set up in a final volume of 50 µl using MyTaq HS Mix (Bioline) and 2.5 µl of Illumina TruSeq PCR Cocktail Primers. The amplification cycles were as follows: 98°C – 2 min; 12 cycles of: 98°C – 15 sec, 60°C – 30 sec, 72°C – 30 sec; 72°C – 5 min. The final PCR products were purified using AMPure XP beads and the final concentration of the libraries was measured using Qubit DNA BR Assay (Life Technologies).

Single-end sequencing for 100bp reads was performed on an Illumina Hiseq 2000.

Whole-genome bisulfite sequencing (WGBS)

Genomic DNA from induced MEFs (48h OSKM) and ESCs was isolated using the Blood and tissue DNeasy kit (Qiagen). Isolated DNA was treated with RNAseA for 30 min at 37°C and cleaned up using AMPure XP beads. 5 µg of treated DNA was fragmented to 100–500 bp using a
Bioruptor Sonicator. 5 minutes in pulses of 30 sec on, 1 minute off. DNA fragments were visualized on 1% agarose gel, gel extracted and purified using a QIAGEN gel extraction minelute kit. End-repair reactions (50 µl) contained 1x T4 DNA ligase buffer (NEB), ATP, 0.4 mM dNTPs, 15 units T4 DNA polymerase, 5 units Klenow DNA polymerase, 50 units T4 polynucleotide kinase (all NEB) and were incubated for 30 min at 20°C. DNA clean-up was performed using a 2x volume of AMPure XP beads and eluted in 32 µl of dH2O. Adenylation was performed for 30 minutes at 37°C in 50 µl volumes that contained 5 µl 1x Klenow buffer, 0.2 mM dATP and 15 units Klenow exo− (NEB). Adenylated DNA fragments and methylated adapters (Illumina) were ligated for 15 min at 20°C in a 50 µl reaction containing 5,000 units quick ligase (NEB) and 5 µl of adapters. Adapter-ligated DNA of 200-600 bp, was size-selected on a 2% agarose gel. Bisulfite conversion was performed with an EpiTect Bisulfite Kit (QIAGEN) following the manufactures conditions. Bisulfite converted DNA was amplified for 15 cycles with PfuTurboCx Hotstart DNA polymerase (Agilent technologies). The final library DNA was quantified using a Qubit fluorometer and a Quant-iT dsDNA HS Kit (Invitrogen).

Single-end sequencing for 100bp reads was performed on an Illumina Hiseq 2000.

**Reduced Representation Bisulfite sequencing (RRBS)**

5 µl of genomic DNA was digested with 50 units of MspI (NEB) in a 100 µl reaction for 6 hours at 37°C. Digested DNA was run on a 3% low-melt agarose gel (Lonza) and fragments of 25 to 300 bp were extracted and purified using a MinElute gel extraction kit (QIAGEN) according to the manufacturers instructions. DNA end-repair and adenylation was as described above with the exception of using a dNTP mix consistent of dATP, dGTP and 5meC dCTP. Ligation to methylated adapters and subsequent library construction was performed similarly to the WGBS protocol.
Single-end sequencing for 100bp reads was performed on an Illumina Hiseq 2000.

Bisulfite sequencing data analysis
DNA methylation calling was performed using BS-Seeker2\textsuperscript{17} using Bowtie 0.12.9\textsuperscript{20} for read alignment on the UCLA Hoffman2 computer cluster. WGBS and MREBS reads were mapped to the entire mm9 reference genome while RRBS reads were mapped to a reduced reference that was \textit{in silico} digested using the \textit{MspI} recognition sequence and limited to fragments of 20–500bp in length. The 100bp reads were trimmed of adapter sequences and allowed 5 mismatches during mapping. MREBS reads were first filtered so that only with the expected 5’ trimers (CGG and CGC; Table 2.1) were retained. For DNA methylation level calling, only CpG dimers covered by at least 5 reads on both were used in an effort to obtain reliable methylation levels.

All statistical analysis were performed using custom R scripts\textsuperscript{21}.

ChIP-seq libraries
All histone modification data used for this study were determined using native ChIP\textsuperscript{22} and will be described in detail in a separate manuscript (Chronis et al., in preparation). Briefly, cells were resuspended in buffer I (0.3M sucrose, 60mM KCl, 15mM NaCl, 5mM MgCl\textsubscript{2} 10mM EGTA, 15mM Tris-HCl, 0.5mM DTT, 0.1% NP40 and protease inhibitor cocktail) and incubated on ice for 10 minutes. Nuclei were generated by centrifugation via a sucrose cushion (1.2M sucrose, 60mM KCL, 15mM NaCl, 5mM MgCl\textsubscript{2}, 0.1mM Tris-HCl, 0.5mM DTT, and protease inhibitor cocktail). Nuclei were then resuspended in MNase-digestion buffer (0.32M sucrose, 50mM Tris-HCl, 4mM MgCl\textsubscript{2}, 1mM CaCl\textsubscript{2}, protease inhibitor cocktail) and digested with 3 units of MNase (Roche) for 10 minutes at 37\textdegree C. Soluble chromatin fractions were then incubated with anti-H3K4me3 (abcam; ab8580), anti-H3K4me2 (abcam ab7766), anti-H3K4me1 (ab8895), anti-
H3K27me3 (Active motif 39155), anti-H3K27ac (abcam; ab4729), and anti-H3K36me3 (abcam; ab9050). Extracts were washed twice with wash buffer A (50mM Tris-HCl, 10mM EDTA, 75mM NaCl), wash buffer B (50mM Tris-HCl, 10mM EDTA, 125mM NaCl), wash buffer C (50mM Tris-HCl, 10mM EDTA, 250mM NaCl). DNA extraction and library preparation as described\(^2\).

**ChIP-seq data analysis**

Reads were mapped to the mouse genome (mm9) using Bowtie software\(^2\) and only those reads that aligned to a unique position with no more than two sequence mismatches were retained for further analysis. Multiple reads mapping to the same location in the genome were collapsed to a single read to account for clonal amplification effects.

**Chromatin states**

The model will be described in detail in a separate manuscript (Chronis et al., in preparation), but briefly 18 chromatin states in the MEFs, EARLY intermediates, LATE intermediates, and ESCs were identified at a resolution of 200bp using chromHMM as described by Ernst and Kellis\(^2\) using ChIP-seq data sets for nine histone modification, one histone variant (H3.3), and an input, as listed in Figure 2.2.

**Differential DNA methylation modeling**

Linear regression was used to model differential CpG dimer methylation estimates based on WGBS (the response vectors) using differential methylation estimates based on RRBS and MREBS, as well as differential read counts within 1kb windows based on MREBS data around
corresponding CpG dimers with R's \texttt{lm()} function\textsuperscript{19}. The coefficients in Table 2.7 are outputs from R's \texttt{summary.lm()} function\textsuperscript{19}.

**Data submission**

All high-throughput sequencing data will be made available on GEO at the time of publication.
Figure 2.1  WGBS, RRBS, and MREBS for samples representing two stages of somatic cell reprogramming
A. Schematic representation of the two cell types used in the study. Mouse embryonic fibroblasts (MEFs; blue), modified to harbor a ‘stem cell cassette’ allowing for the simultaneous induction of the four pluripotency factors (OCT4 (O), SOX2 (S), KLF4 (K), and MYC (M)) by the addition of doxycycline, were induced for 48 hours. These EARLY somatic cell reprogramming intermediates (yellow) were the first of the two cell types sampled, with embryonic stem cells (ESCs; green), representing the fully reprogrammed state, being the second.

B. Bar plots showing the number of CpGs obtained at five different coverage levels (1–20 X) in each of the two whole-genome bisulfite sequencing (WGBS) samples: i) EARLY intermediates and ii) ESCs.

C. As in (B), but for reduced representation bisulfite sequencing (RRBS) samples.

D. As in (B), but for duplicate samples produced using methylation-sensitive restriction enzyme bisulfite sequencing (MREBS).
Figure 2.2 DNA methylation estimates based on WGBS, RRBS and MREBS data in different chromatin states
A. Heat map and functional annotation for an 18-level chromatin state model at 200bp resolution built using peak calls made using ChIP-seq data sets for 10 histone modifications, as well as an input library, for each of the three cell types described in Figure 2.1A: i) MEFs, ii) EARLY intermediates, and iii) ESCs, as well as a late reprogramming intermediate (LATE), partially induced pluripotent cells, or pre-iPSCs, not otherwise used in the study. Candidate functional annotations were assigned to each of the 18 chromatins states based on the prevalence and combination of histone mark peaks, which could in turn be classified into the seven categories indicated in the left-hand column. The probability of a window in each state to be contain a peak for a given histone modification is given as a percentage in each cell, and visually indicated by the intensity of color in the heat map. The proportion of the concatenated genome (MEFs + EARLY + LATE + ESCs) found in each of the 18 chromatin states is given in the third column.

B. Violin plots of the distributions of the DNA methylation estimates in each of the 18 chromatin states (described in A), as well as genome-wide (blue), for EARLY intermediates using i) whole-genome bisulfite sequencing (WGBS), reduced representation bisulfite sequencing (RRBS), and iii) methylation-sensitive restriction enzyme bisulfite sequencing (MREBS). The mean DNA methylation estimates within 200bp windows corresponding to those used for the chromatin state model were used, only considering those windows containing at least one CpG with 5X coverage, in an effort to ensure high-confidence estimates. White circles represent median values.

C. As in (B), but for ESCs.
Figure 2.3  Chromatin state coverage by DNA methylation estimates by WGBS, RRBS, and MREBS
A. Pie charts show the proportion of 200bp windows with DNA methylation estimates found in each of the 18 chromatin states (as described in Figure 2.2A) using i) whole-genome bisulfite sequencing (WGBS), reduced representation bisulfite sequencing (RRBS), and iii) methylation-sensitive restriction enzyme bisulfite sequencing (MREBS) EARLY intermediate samples, as compared to the proportion of chromatin states in genome for all 13.3 million windows (iv). Bar plots show the log2 fold change (observed / expected) number of windows with estimates per method: i) WGBS), RRBS, and iii) MREBS. The mean DNA methylation estimates within 200bp windows corresponding to those used for the chromatin state model were used, only considering those windows containing at least one CpG with 5X coverage, in an effort to ensure high-confidence estimates.

B. As in (A), but for ESCs.
Figure 2.4  Differential DNA methylation levels modeled using MREBS data
A. IGV tracks of differential DNA methylation estimates between EARLY intermediates and ESCs based on WGBS data (i, dark blue), RRBS data (ii, light blue), modeled data based on combined model 4 and 2 (iii, green, two replicates), MREBS DNA methylation-based data (iv, orange, two replicates), and MREBS read count-based data (v, red, two replicates), within a 611 kb region of chr17. Bottom two tracks show CpG dimer and Refseq gene locations. Gray background reflects regions (CpG dimers) where data was not available.

B. As in (A), but for a 19kb region around the Gata2 gene.

C. As in (A), but for the extend Olig1/2 gene locus, divided into three 18 kb panels.
Figure 2.5  Examples of modeled differential DNA methylation around gene loci

A. As in Figure 2.4A, but for 13 genes up-regulated in EARLY intermediates relative to ESCs.

B. As in Figure 2.4A, but for 10 genes up-regulated in ESCs relative to EARLY intermediates.
Table 2.1  
MRE endonuclease recognition sequence frequency within the mm9 genome

Showing their position within the ranked frequencies for all the 4mer CpG, including chrM.

*Note: HpaII has the same recognition sequence as MspI (the endonuclease typically used for RRBS libraries), albeit HpaII is methylation sensitive, as are AciI and Hin6I.
<table>
<thead>
<tr>
<th>Sample Type</th>
<th>Total Mapped Reads</th>
<th>Mean CpG coverage depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>WGBS EARLY</td>
<td>429,374,384</td>
<td>7.80</td>
</tr>
<tr>
<td>WGBS ESC</td>
<td>391,724,853</td>
<td>7.23</td>
</tr>
<tr>
<td>RRBS EARLY</td>
<td>12,826,209</td>
<td>12.49</td>
</tr>
<tr>
<td>RRBS ESC</td>
<td>18,131,716</td>
<td>18.88</td>
</tr>
<tr>
<td>MREBS EARLY Rep1</td>
<td>11,963,716</td>
<td>5.72</td>
</tr>
<tr>
<td>MREBS EARLY Rep2</td>
<td>12,400,629</td>
<td>5.78</td>
</tr>
<tr>
<td>MREBS ESC Rep1</td>
<td>11,835,343</td>
<td>6.35</td>
</tr>
<tr>
<td>MREBS ESC Rep2</td>
<td>12,222,192</td>
<td>6.28</td>
</tr>
</tbody>
</table>

Table 2.2  Bisulfite sequencing library mapped reads and mean CpG coverage depth

WGBS and MREBS reads were mapped the whole genome (mm9). Mean CpG coverage determined for CpGs on either strand. Mean CpG coverage determined for CpGs on either strand. RRBS reads were mapped to an in silico MspI digested reduced reference genome. MREBS reads were filtered in silico to have the expected 5’ start sites.
### Table 2.3  CpG dimer coverage per bisulfite sequencing library

The percentage of CpG dimers with at least 5X coverage for each bisulfite sequencing library, as well as the percentage of CpG dimers with the specified number of MREBS reads within a surrounding 1kb window. *Note: This is based on 21,342,493 CpG dimers in mm9, excluding chrM.

<table>
<thead>
<tr>
<th>Library Type</th>
<th>CpG dimers*</th>
<th>% Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>WGBS EARLY DNAme</td>
<td>17,202,917</td>
<td>80.6%</td>
</tr>
<tr>
<td>WGBS ESC DNAme</td>
<td>17,022,903</td>
<td>79.8%</td>
</tr>
<tr>
<td>RRBS EARLY DNAme</td>
<td>1,289,663</td>
<td>6.0%</td>
</tr>
<tr>
<td>RRBS ESC DNAme</td>
<td>1,358,529</td>
<td>6.4%</td>
</tr>
<tr>
<td>MREBS EARLY Rep1 DNAme</td>
<td>962,559</td>
<td>4.5%</td>
</tr>
<tr>
<td>MREBS EARLY Rep2 DNAme</td>
<td>924,551</td>
<td>4.3%</td>
</tr>
<tr>
<td>MREBS ESC Rep1 DNAme</td>
<td>918,508</td>
<td>4.3%</td>
</tr>
<tr>
<td>MREBS ESC Rep2 DNAme</td>
<td>973,441</td>
<td>4.6%</td>
</tr>
<tr>
<td>MREBS EARLY Rep1 Counts&gt;0</td>
<td>14,695,688</td>
<td>68.9%</td>
</tr>
<tr>
<td>MREBS EARLY Rep2 Counts&gt;0</td>
<td>14,718,855</td>
<td>69.0%</td>
</tr>
<tr>
<td>MREBS ESC Rep1 Counts&gt;0</td>
<td>13,602,796</td>
<td>63.7%</td>
</tr>
<tr>
<td>MREBS ESC Rep2 Counts&gt;0</td>
<td>13,947,144</td>
<td>65.3%</td>
</tr>
<tr>
<td>MREBS EARLY Rep1 Counts=&gt;2</td>
<td>10,250,065</td>
<td>48.0%</td>
</tr>
<tr>
<td>MREBS EARLY Rep2 Counts=&gt;2</td>
<td>10,315,026</td>
<td>48.3%</td>
</tr>
<tr>
<td>MREBS ESC Rep1 Counts=&gt;2</td>
<td>8,987,596</td>
<td>42.1%</td>
</tr>
<tr>
<td>MREBS ESC Rep2 Counts=&gt;2</td>
<td>9,227,587</td>
<td>43.2%</td>
</tr>
<tr>
<td>MREBS EARLY Rep1 Counts=&gt;5</td>
<td>5,317,311</td>
<td>24.9%</td>
</tr>
<tr>
<td>MREBS EARLY Rep2 Counts=&gt;5</td>
<td>5,387,683</td>
<td>25.2%</td>
</tr>
<tr>
<td>MREBS ESC Rep1 Counts=&gt;5</td>
<td>4,696,304</td>
<td>22.0%</td>
</tr>
<tr>
<td>MREBS ESC Rep2 Counts=&gt;5</td>
<td>4,659,571</td>
<td>21.8%</td>
</tr>
<tr>
<td>MREBS EARLY Rep1 Counts=&gt;10</td>
<td>3,557,311</td>
<td>16.7%</td>
</tr>
<tr>
<td>MREBS EARLY Rep2 Counts=&gt;10</td>
<td>3,610,365</td>
<td>16.9%</td>
</tr>
<tr>
<td>MREBS ESC Rep1 Counts=&gt;10</td>
<td>3,389,630</td>
<td>15.9%</td>
</tr>
<tr>
<td>MREBS ESC Rep2 Counts=&gt;10</td>
<td>3,352,308</td>
<td>15.7%</td>
</tr>
</tbody>
</table>
Table 2.4  
**CpG dimer coverage for differential analysis per bisulfite sequencing library**

The percentage of CpG dimers with at least 5X coverage in both the EARLY intermediate and ESC samples for each bisulfite sequencing library, as well as the percentage of CpG dimers with at least two MREBS reads within a surrounding 1kb window. *Note: This is based on 21,342,493 CpG dimers in mm9, excluding chrM.*

<table>
<thead>
<tr>
<th></th>
<th>CpG dimers</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>WGBS EARLY-ESC</td>
<td>16,113,172</td>
<td>75.5%</td>
</tr>
<tr>
<td>differential meCpG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RRBS EARLY-ESC</td>
<td>1,204,249</td>
<td>5.6%</td>
</tr>
<tr>
<td>differential meCpG Rep1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MREBS EARLY-ESC</td>
<td>649,614</td>
<td>3.0%</td>
</tr>
<tr>
<td>differential meCpG Rep1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MREBS EARLY-ESC</td>
<td>665,431</td>
<td>3.1%</td>
</tr>
<tr>
<td>differential meCpG Rep2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MREBS EARLY-ESC</td>
<td>12,542,720</td>
<td>58.8%</td>
</tr>
<tr>
<td>differential Counts Rep1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MREBS EARLY-ESC</td>
<td>12,746,080</td>
<td>59.7%</td>
</tr>
<tr>
<td>differential Counts Rep2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 2.5  
CpG dimer-level correlations between bisulfite sequencing libraries

Pearson correlation values between WGBS, RRBS, and MREBS CpG dimer-level methylation estimates, as well as binned read counts within 1kb windows around CpG dimers, for those with at least two MREBS reads. Red intensity signifies the strength of a positive correlation, while blue intensity signifies the strength of the anti-correlation.
Table 2.6  
CpG dimer-level correlations between differential values for all bisulfite sequencing library pairs

Pearson correlation values between WGBS, RRBS, and MREBS differential CpG dimer-level methylation estimates (EARLY – ESC), as well as EARLY – ESC differential read counts between all CpG dimers with at least two MREBS reads within a surrounding 1kb window. Red intensity signifies the strength of a positive correlation, while blue intensity signifies the strength of the anti-correlation.

<table>
<thead>
<tr>
<th></th>
<th>delta WGBS DNAme</th>
<th>delta RRBS DNAme</th>
<th>delta MREBS DNAme</th>
<th>delta MREBS readcounts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rep1</td>
<td>Rep2</td>
<td>Rep1</td>
<td>Rep2</td>
</tr>
<tr>
<td>delta WGBS DNAme</td>
<td>1.00</td>
<td>0.63</td>
<td>0.55</td>
<td>0.56</td>
</tr>
<tr>
<td>delta RRBS DNAme</td>
<td>0.63</td>
<td>1.00</td>
<td>0.64</td>
<td>0.64</td>
</tr>
<tr>
<td>delta MREBS DNAme</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rep1</td>
<td>0.55</td>
<td>0.64</td>
<td>1.00</td>
<td>0.64</td>
</tr>
<tr>
<td>Rep2</td>
<td>0.56</td>
<td>0.64</td>
<td>0.64</td>
<td>1.00</td>
</tr>
<tr>
<td>delta MREBS read counts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rep1</td>
<td>-0.34</td>
<td>-0.51</td>
<td>-0.35</td>
<td>-0.34</td>
</tr>
<tr>
<td>Rep2</td>
<td>-0.34</td>
<td>-0.52</td>
<td>-0.35</td>
<td>-0.34</td>
</tr>
<tr>
<td>Model</td>
<td>RRBS-seq differential DNA methylation</td>
<td>MREBS differential DNA me + counts</td>
<td>MREBS differential DNA me only</td>
<td>MREBS differential counts only</td>
</tr>
<tr>
<td>-------</td>
<td>--------------------------------------</td>
<td>-----------------------------------</td>
<td>-------------------------------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rep1</td>
<td>Rep2</td>
<td>Rep1</td>
</tr>
<tr>
<td>B0 (intcept)</td>
<td></td>
<td>5.29</td>
<td>5.09</td>
<td>4.92</td>
</tr>
<tr>
<td>B1</td>
<td></td>
<td>0.75</td>
<td>0.53</td>
<td>0.54</td>
</tr>
<tr>
<td>B2</td>
<td>N/A</td>
<td>0.90</td>
<td>-0.09</td>
<td>-0.09</td>
</tr>
</tbody>
</table>

Table 2.7  Differential DNA methylation model coefficients

Coefficient values (first column) for four different models (top row and described in the text).
Table 2.8   Differential DNA methylation model metrics

The table gives comparison metrics (first column) for four different models, as well as a combined model, (top row and described in the text). The column labeled ‘Total / optimal value’ gives the maximum or best value achievable for each metric. The column labeled ‘WGBS differential DNA methylation’ provides coverage information for comparison purposes. *Note: In the case methylation levels, only CpG dimers with 5X coverage in both EARLY intermediates and ESCs were considered. With respect to counts, CpG dimers with 2+ reads in the surrounding 1Kb bin, in at least one sample, were considered. Red intensity signifies how close the metrics are to the optimal values.
References


Chapter 3

A high-throughput screen of inactive X chromosome reactivation identifies the enhancement of DNA demethylation by 5-aza-2’-dC upon inhibition of ribonucleotide reductase
3.1 Abstract

DNA methylation is important for maintenance of the silent state of genes on the inactive X chromosome (Xi). Here, we screened for siRNAs and chemicals that reactivate an Xi-linked reporter in the presence of 5-aza-2’-deoxycytidine (5-aza-2’-dC), an inhibitor of DNA methyltransferase 1, at a concentration that, on its own, is not sufficient for Xi-reactivation. We found that inhibition of ribonucleotide reductase (RNR) induced expression of the reporter. RNR inhibition potentiated the effect of 5-aza-2’-dC by enhancing its DNA incorporation, thereby decreasing genome-wide DNA methylation levels. Since both 5-aza-2’-dC and RNR-inhibitors are used in the treatment of hematological malignancies, we treated myeloid leukemia cell lines with 5-aza-2’-dC and the RNR inhibitor hydroxyurea, and observed synergistic inhibition of cell growth and decreases in genome-wide DNA methylation. Taken together, our study identifies a drug combination that enhances DNA demethylation by altering nucleotide metabolism.
3.2 Introduction

X chromosome inactivation (XCI) is a program of transcriptional gene silencing that occurs on one of two X chromosomes in female mammalian cells to equalize gene dosage of X-linked genes to male cells. The inactive X chromosome (Xi) is a striking example of developmentally regulated heterochromatin formation in mammals. XCI has served as paradigm for understanding factors with generalized roles in genome-wide gene silencing such as DNA methylation and Polycomb protein-mediated histone methylation\(^1\)\(^-\)\(^3\). The Xi is established early in female embryonic development through a series of stepwise molecular changes that cooperate to ensure stable chromosome-wide gene silencing. Once established, the Xi is inherited through all somatic cell divisions and adult life\(^1\)\(^-\)\(^3\). Early in embryonic development, XCI is initiated by the upregulation of the long noncoding RNA \(Xist\) from the maternal or paternal X chromosome\(^1\)\(^-\)\(^3\). \(Xist\) coats the X chromosome from which it is expressed and initiates a cascade of events including exclusion of RNA polymerase II, changes in histone marks, and recruitment of structural chromosome proteins\(^1\)\(^-\)\(^3\). Accumulation of the histone variant macroH2A1 and gain of CpG island methylation characterize the transition to the maintenance phase of XCI, which is marked by resistance to X chromosome reactivation (XCR) upon deletion of \(Xist\)\(^4\)\(^-\)\(^8\). Thus, initially \(Xist\) is absolutely required for the initiation of XCI, but later is largely dispensable for the maintenance of the Xi, due to the presence of various other repressive chromatin marks.

In the maintenance phase of XCI, interference with DNA methylation has thus far shown the largest effect on eliciting XCR\(^9\)\(^-\)\(^11\). Currently, the only manipulations known to produce complete XCR do so by reversing cellular identity to the embryonic state rather than by directly interfering with silencing mechanisms\(^12\),\(^13\). Despite the observation that many factors are implicated in Xi establishment, DNA methylation may uniquely ‘lock-in’ the silenced state and
execute a greater influence on the robust nature of Xi maintenance than other repressive regulatory mechanisms. DNA methylation concentrates on CpG islands in the course of XCI with redistribution away from intragenic and intronic CpGs relative to the active X chromosome\(^5,7,14,15\). CpG island methylation on the Xi is established by the \textit{de novo} methyltransferase DNMT3B upon initiation and is subsequently propagated by the maintenance methyltransferase DNMT1\(^5,11,16\). Interference with DNA methylation by deletion of \textit{Dnmt1} or treatment with 5-aza-2’-deoxycytidine (5-aza-2’-dC, also called decitabine) has been shown to induce the reactivation of an Xi-linked reporter gene in a low proportion of female somatic cells\(^10\). 5-aza-2’-dC is a deoxycytidine analog that upon phosphorylation is incorporated into DNA and irreversibly inhibits DNMT1\(^17\). Subsequent rounds of DNA replication in the absence of DNMT1 activity lead to passive DNA demethylation\(^18\).

In the field of cancer biology, there is growing appreciation that abnormalities in these epigenetic pathways can drive tumorigenesis across many cancer types and there is promise for improved therapies aimed at reversal of gene silencing\(^19\). In this study, we bridge the study of the Xi with the development of strategies to reactivate silenced genes. 5-aza-2’-dC is used clinically in the setting of hematologic malignancy with the rationale of reactivating silenced genes\(^18\). The drug is currently approved for the treatment of myelodysplastic syndrome (MDS) and acute myeloid leukemia (AML)\(^19\). Several studies have confirmed that 5-aza-2’-dC at low doses elicits genome-wide DNA demethylation in AML patient samples\(^20-22\). One approach to increase the epigenetic activity of 5-aza-2’-dC in myeloid malignancy is to use it in combination with other agents known to elicit reactivation of silenced genes, such as histone deacetylase inhibitors\(^19\). Notably, for the Xi, such co-treatment approaches increase the rate of Xi-reactivation\(^10\). For instance, it has been shown that the combination of histone deacetylase inhibitors with 5-aza-2’-
dC increases rates of XCR\textsuperscript{10}. The similar efficacy of 5-aza-2’-dC alone or in combination with other chromatin-modifying agents in Xi-linked genes and in myeloid leukemia supports the translation of findings from X chromosome inactivation to epigenetic cancer therapies.

Here, we set out to find additional pathways that, in combination with 5-aza-2’-dC, elicit Xi-reactivation. Specifically, we applied high-throughput siRNA and chemical screening to identify factors that could reactivate a silent reporter transgene on the Xi. Our screen employed treatment with a low dose of 5-aza-2’-dC to sensitize somatic cells for DNA demethylation and X chromosome reactivation. We identified that inhibition of the ribonucleotide reductase protein complex significantly enhances DNA demethylation action of 5-aza-2’-dC and hence, activity of the Xi-reporter. We characterize the mechanism of action as increasing DNA incorporation of 5-aza-2’-dC and thus its demethylating activity. While our approach initially centered on the Xi, we found a pathway that altered DNA methylation levels genome-wide. Our study therefore demonstrates that assays of XCR can be adapted to optimize the epigenetic activity of a DNA demethylating drug combination.
3.3 Results

3.3.1 An siRNA screen for XCR in the presence of a low 5-aza-2′-dC dose identifies the ribonucleotide reductase pathway

Previous work from our lab has shown that an Xi-linked, CAG promoter-driven luciferase transgene in the Hprt locus on the X chromosome (Xi-luciferase) is a sensitive reporter of Xi silencing when tested in primary mouse embryonic fibroblasts (MEFs)\textsuperscript{23}. Our Xi-luciferase MEFs faithfully inactivate the luciferase-bearing X chromosome in female embryonic development rather than undergoing random XCI because an Xist deletion on the other X chromosome forces XCI on the chromosome carrying the wildtype Xist allele\textsuperscript{24} (Figure 3.1A). The Xi-luciferase gene body and promoter are highly methylated at the DNA level and Xi-luciferase reporter MEFs increase luciferase activity in a dose-dependent fashion in response to 5-aza-2′-dC treatment\textsuperscript{23}. Here, we used the Xi-luciferase reporter MEFs to screen for gene knockdown or chemicals that could elicit XCR.

In order to perform a high-throughput screen for XCR, we established an siRNA knockdown assay in 384-well format, with each individual siRNA tested in a single well. As positive control, we chose knockdown of Dnmt1 since interference with Dnmt1 either by knockout or 5-aza-2′-dC treatment has previously been described to elicit XCR in MEFs\textsuperscript{10,23}. Initially, we tested increasing concentrations of 5-aza-2′-dC in combination with Dnmt1 knockdown to determine a condition for which the depletion of Dnmt1 by siRNAs yielded robust XCR rates. Our titration experiment demonstrated that the combination of 5-aza-2′-dC at a concentration ranging from 0.10 to 0.20 uM, along with siDnmt1 treatment, enhanced luciferase activity in the 384-well format. 5-aza-2′-dC treatment or knockdown of Dnmt1 alone did not
induce a significant difference in luciferase signal compared to untreated wells (Figure 3.2A). The requirement for 5-aza-2’-dC co-treatment with \( \text{Dnmt1} \) knockdown to detect XCR likely reflects that \( \text{Dnmt1} \) knockdown alone does not lead to sufficient levels of XCR to be detected by this assay in 384-well format. By comparison, a higher dose of 5-aza-2’-dC (1 uM) elicited strong reactivation of the Xi-linked luciferase reporter that was not as dramatically enhanced by \( \text{siDnmt1} \) treatment. Thus, a low dose of 5-aza-2’-dC has a sensitizing effect on eliciting XCR by \( \text{Dnmt1} \) knockdown. The interaction of 5-aza-2’-dC with other Xi maintenance factors indicates a similar sensitizing effect with respect to XCR. For instance, the knockdown of the candidate Xi-maintenance factor \( \text{Atf7ip} \) or deletion of \( \text{Xist} \) produces a low rate of XCR that is significantly boosted by the addition of 5-aza-2’-dC\(^{10,23}\). Therefore, we extended low concentration 5-aza-2’-dC treatment to the entire genome-wide siRNA screen with the rationale that knockdown of other chromatin-modifying factors may require concurrent DNA demethylation to produce detectable Xi-luciferase reporter reactivation.

We performed a genome-wide mouse siRNA screen with 51,150 siRNAs against 21,114 genes on 153 384-well plates (see Methods section for details on the library used) in duplicate using the female Xi-linked luciferase reporter MEFs in the presence of 0.2 uM 5-aza-2’-dC. We measured luciferase levels 72 hrs after siRNA transfection (Figure 3.1B). To eliminate batch effects, we normalized luminescence data by 384-well plate, and then analyzed the data by prioritizing gene hits with multiple active siRNAs by redundant siRNA activity (RSA) analysis\(^{25}\) (Figures 3.1C and 3.3). Notably, \( \text{Dnmt1} \) was the top hit in our genome-wide screen, which provided internal validation of the method (Figure 3.1C). Further support came from another hit, identified as \( \text{Atf7ip} \), which our group recently reported as a maintenance factor in XCI\(^{23}\). As with
other previously described maintenance factors, we found that the Xi-luciferase signal in response to knockdown of *Atf7ip* was greatly increased by 5-aza-2’-dC (0.2 uM) co-treatment\(^{21}\). Identification of *Atf7ip* in the screen supports the strategy of 5-aza-2’-dC co-treatment to unmask functional contribution of Xi-maintenance factors.

To select novel hits, we chose the top 54 genes from the RSA analysis with at least two unique active siRNAs inducing an increase in luciferase levels in the 384-well screen, omitting genes we deemed irrelevant such as those for olfactory receptors, and retested the active siRNAs sequences from the library (data not shown). Several of these siRNAs showed reproducible increases in luciferase activity in the validation assay (Figure 3.4). Here, we decided to focus on *Rrm2* as a hit since one siRNA against it had produced the next highest level of luciferase activity after the siRNAs targeting *Atf7ip*. Follow-up assays with a greater number of starting cells demonstrated an increase in luciferase activity for all three siRNAs against the ribonucleotide reductase (RNR) M2 subunit gene (*Rrm2*) (Figure 3.1D).

The luminescence generated with si*Rrm2* treatment was in proportion to individual extent of *Rrm2* knockdown, suggesting specificity of *Rrm2* targeting for the XCR effect (Figure 3.1D). As part of the RNR enzyme complex, RRM2 catalyzes the conversion of ribonucleoside 5’-disphosphates to their 2’-deoxyribonucleoside form in the rate-limiting step of *de novo* dNTP biosynthesis\(^{26}\). The RRM2 subunit, which was identified in this siRNA screen, is specifically up-regulated at S phase of cell cycle and is necessary for the activity of the RNR complex\(^{26}\). Since we identified si*Rrm2* in combination with 5-aza-2’-dC (0.2 uM) in the genome-wide screen, we next asked if knockdown of *Rrm2* could elicit XCR in the absence of 5-aza-2’-dC, since interference with *Atf7ip* or *Xist* produces a low rate of XCR that is significantly boosted by the addition of 5-aza-2’-dC\(^{10,23}\). However, unlike previously described maintenance factors, we did
not find that siRrm2 produced measurable levels of luciferase activity in the absence of 5-aza-2’-dC (Figure 3.1D). We conclude that low doses of 5-aza-2’-dC are necessary for the XCR effect of siRrm2 identified by our genome-wide screen for factors involved in the maintenance of Xi silencing.

3.3.2 Inhibitors of ribonucleotide reductase elicit XCR with 5-aza-2’-dC present

We used a complimentary approach to further probe the pathways contributing to Xi maintenance by performing a companion screen analogous to the siRNA screen but instead using a collection of annotated chemicals (Figure 3.1B and 3.5). In the screen, we found that Resveratrol, a chemical agent known for mimicking cellular effects of caloric restriction, demonstrated the potential to activate the Xi-luciferase reporter (Figure 3.1E). The bell-shaped dose-response activity of Resveratrol in combination with a fixed (low) concentration of 5-aza-2’-dC (0.2 uM) suggested a maximal XCR activity at a concentration of 20 uM27 (Figure 3.1F).

In order to confirm an XCR-specific rather than an off-target luciferase effect, we tested whether the combination of Resveratrol with 5-aza-2’-dC could re-activate a different Xi-linked reporter, encoding H2B-citrine, in the same locus as the luciferase reporter23. Indeed, Resveratrol and 5-aza-2’-dC (0.2 uM) together could reactivate the Xi-linked fluorescent reporter (data not shown).

Resveratrol is described to mediate its metabolic effects through direct and indirect activation of the histone deacetylase SIRT1 though no specific role in reversal of chromatin silencing or XCI has been characterized28-30. We did not find that knockdown of Sirt1 attenuated the ability of Resveratrol with 5-aza-2’-dC to elicit Xi-luciferase reactivation (data not shown).
further search for the cellular target of Resveratrol in XCR led us to a study that described Resveratrol as an inhibitor of RNR, the same enzyme complex that we identified as a hit in the genome-wide siRNA screen for XCR described above\textsuperscript{31}. This link between the complimentary screening approaches pointed to Resveratrol’s role in XCR in the presence of 5-aza-2’-dC by means of RNR inhibition.

In order to further investigate whether RNR is the target of Resveratrol in eliciting XCR, we tested another well-characterized inhibitor of RNR, hydroxyurea, and found that it also increased Xi-luciferase activity in the presence of a low dose of 5-aza-2’-dC (0.2 uM)\textsuperscript{31} (Figure 3.1G). We reasoned that if Resveratrol and siRrm2 converge on inhibition of RNR, that the XCR effect of Resveratrol should require the co-treatment with 5-aza-2’-dC as seen for the Rrm2 knockdown. Indeed, we found that all three forms of RNR inhibition-mediated XCR (siRrm2, Resveratrol, and hydroxyurea) demonstrated a complete dependence on low levels of 5-aza-2’-dC to elicit XCR, and that RNR inhibition alone by these various means did not increase Xi-luciferase activity (Figures 3.1D/F/G and 3.6).

We further investigated the relationship between 5-aza-2’-dC and RNR inhibition in XCR by querying whether 5-aza-2’-dC can be replaced by knockdown of Dnmt1. In previous studies where 5-aza-2’-dC had a sensitizing effect towards XCR, the effect is attributable to interference with Dnmt1\textsuperscript{23}. For instance, 5-aza-2’-dC treatment can be substituted by knockdown of Dnmt1 to elicit synergistic XCR by Atf7ip knockdown\textsuperscript{23}. Contrary to these prior findings, Dnmt1 depletion did not replace the contribution of 5-aza-2’-dC to XCR induced by RNR-inhibition (Figure 3.1H). These data suggest a mechanism of action whereby RNR inhibition specifically affects the action of the cytidine analogue 5-aza-2’-dC.
3.3.3 RNR inhibition increases incorporation of 5-aza-2’-dC into DNA

Next, we sought to understand how RRM2 inhibition interacts with low amounts of 5-aza-2’-dC to elicit XCR. The pool of dNTPs in the nucleus is tightly regulated and studies have speculated that RNR inhibition can increase the likelihood of nucleoside analog DNA incorporation by reducing the pools of endogenous nucleotide concentrations\textsuperscript{26,32}. Accordingly, we postulated that RRM2 inhibition may increase 5-aza-2’-dCTP concentration in the nucleus relative to the endogenous dCTP pool, leading to more 5-aza-2’-dCTP DNA incorporation (Figure 3.6A). Higher rates of 5-aza-2’-dC incorporation into DNA subsequently would lead to greater DNA demethylation and XCR (Figure 3.6A).

Consistent with this model, we observed that knockdown of \textit{Rrm2} or Resveratrol treatment reproducibly increased the amount of tritiated 5-aza-2’-dC incorporated into DNA approximately by two-fold (Figure 3.6B). We further tested the role of the ratio of 5-aza-2’-dCTP to endogenous dCTP by the converse manipulation of increasing dCTP relative to 5-aza-2’-dC. This experiment was performed by adding increasing concentrations of deoxycytidine (dC) into media, which is metabolized to dCTP within the cell, in the presence of 5-aza-2’-dC with \textit{Rrm2} knockdown or Resveratrol treatment (Figures 2C/D). Importantly, dC does not require the action of RNR for DNA incorporation. Our expectation was that an increase in dCTP levels in the cell would reduce the incorporation of 5-aza-2’-dC into the DNA, and therefore reduce the reactivation of the Xi-linked luciferase reporter. As expected, the luciferase signal decreased in a dose-dependent fashion when exogenous deoxycytidine was supplied in the media (Figure 3.6C/D). The loss of the Xi-reporter reactivation is consistent with the notion that the relative nuclear concentration of 5-aza-2’-dCTP to dCTP could be shifted by addition of exogenous nucleotide substrate to reduce the effective concentration of the 5-aza-2’-dC analog.
An alternate explanation for the observed decrease in luciferase signal upon addition of dC is a reduction in viable cell number. To rule out possible nucleotide treatment-dependent cell growth effects, we confirmed that protein concentrations in lysates were similar for the various treatment conditions (data not shown). Furthermore, we used uridine as a control because it is a nontoxic precursor of pyrimidine synthesis that, like deoxycytidine, can be taken up by cells and used as a substrate via the nucleoside salvage synthetic pathway\textsuperscript{33} (Figure 3.6C/D). Unlike deoxycytidine, uridine requires reduction by RNR in order to contribute to dNTP pools\textsuperscript{33}. Increasing levels of uridine did not alter Xi-luciferase levels and thereby XCR in the presence of 5-aza-2’-dC with \textit{Rrm2} knockdown and Resveratrol treatment, respectively, compared to control (Figure 3.6C/D). This result supports the role of deoxycytidine in reversing the XCR effect downstream of RNR.

In sum, RRM2/RNR inhibition was identified in the XCR screen because it augmented 5-aza-2’-dC DNA incorporation. This mechanism is consistent with the observation that RNR inhibition alone, i.e. in the absence of 5-aza-2’-dC, did not produce measurable Xi-reporter reactivation in prior assays (Figure 3.1D/F/G).

3.3.4 \textit{Rrm2} inhibition enhances demethylation caused by 5-aza-2’-dC

If \textit{Rrm2} inhibition potentiates low dose 5-aza-2’-dC action to increase XCR by increasing the incorporation of 5-aza-2’-dC, then methylation distributions in cells treated with a low dose of 5-aza-2’-dC with RRM2 inhibition should approximate those of cells treated with a high dose of 5-
aza-2’-dC. We investigated DNA methylation patterns at genome-scale by reduced representation bisulfite sequencing (RRBS). Specifically, MEFs were treated with siRrm2 or Resveratrol alone, low or high doses of 5-aza-2’-dC, and combinations of siRrm2 or Resveratrol with a low dose of 5-aza-2’-dC (Figures 3.7 and 3.8). As expected, compared to control treatments, the treatment of MEFs with a low dose of 5-aza-2’-dC (0.2 uM) induced only a slight reduction in the level of genome-wide DNA methylation, while high dose of 5-aza-2’-dC (10.0 uM) resulted in marked demethylation (Figure 3.7A/B/C). Notably, the combination of Rrm2 knockdown or Resveratrol with the low dose 5-aza-2’-dC reduced global methylation to a similar extent as the high dose 5-aza-2’-dC treatment, while the methylation profile remained largely unchanged when Rrm2 knockdown and Resveratrol addition, respectively, were applied without 5-aza-2’-dC (Figure 3.7A/B/C). The effect on the methylation profile and hierarchical clustering patterns of the autosomes and X chromosome was similar, as well as the effect on promoters and CpG islands (Figures 3.7B and 3.8A/B). These findings are consistent with a genome-wide effect on DNA methylation rather than an Xi-specific mechanism, owing to increased DNA incorporation of 5-aza-2’-dC under RRM2 inhibition conditions. We observed that CpGs with the highest levels of methylation in the control samples showed the most dramatic 5-aza-2’-dC-induced demethylation, both on the X chromosome and on autosomes (Figures 3.7D and 3.8C). For CpGs with lower methylation levels in the untreated conditions, demethylation due to 5-aza-2’-dC incorporation is still visible but less extensive.

We believe that the greater apparent effect in highly methylated regions does not represent a predilection of 5-aza-2’-dC for highly methylated regions, as has been previously suggested, but rather that the random incorporation of 5-aza-2’-dC disproportionately affects the methylation estimates of highly methylated sites20.
We also extracted the available methylation data for the Xi-linked luciferase reporter gene to determine whether the methylation levels correlated with the extent of Xi-luciferase reactivation in the various conditions. We found that CpG sites within the luciferase reporter gene followed the genome-wide methylation changes, and that the low 5-aza-2’-dC treatment together with RNR inhibition, by either Rrm2 knockdown or Resveratrol, induced similar demethylation as the high dose of 5-aza-2’-dC (Figure 3.9). The similar behavior of the luciferase reporter CpG sites supports that the augmentation of DNA incorporation describes the Rrm2 result in our Xi-reporter reactivation screen.

Taken together, our genome-wide methylation analysis for high dose 5-aza-2’-dC and low dose 5-aza-2’-dC with RRM2 inhibition supports the idea that RRM2 inhibition increases the effective concentration of 5-aza-2’-dC and thereby its DNA incorporation, leading to global DNA demethylation.

### 3.3.5 Hydroxyurea and 5-aza-2’-dC synergistically inhibit myeloid leukemia cell line proliferation in a dose-dependent fashion

Given that RRM2 inhibition increases DNA incorporation of 5-aza-2’-dC, we next applied the combination of RRM2 inhibition and 5-aza-2’-dC to a disease model in which 5-aza-2’-dC has therapeutic relevance. 5-aza-2’-dC is an FDA-approved drug and commonly used off-label in the setting of acute myeloid leukemia (AML). Therefore, we tested the drug combination in four myeloid leukemia cell lines (THP1, U937, K562, HL60) (Figure 3.10). We hypothesized that, since RRM2 inhibition increased DNA incorporation of 5-aza-2’-dC, the combination of RRM2
inhibition with 5-aza-2’-dC could improve the therapeutic index of 5-aza-2’-dC, allowing lower doses to maximize demethylation activity with fewer cytotoxic off-target effects. We chose to use hydroxyurea (HU) as the form of RRM2 inhibition because it also is an FDA-approved agent commonly used off-label for cyto-reductive purposes, also in the setting of AML.}

To assess the effect of combining HU and 5-aza-2’-dC on myeloid leukemia cell line proliferation, we applied a luminescence-based cell viability assay that linearly scales with cell number, and titered HU and 5-aza-2’-dC individually to determine IC50 values for each cell line. For 5-aza-2’-dC, IC50s were difficult to approximate given a plateau in cell proliferation changes at higher concentrations (data not shown). Thus we chose the 5-aza-2’-dC concentration corresponding to halfway to the point of plateau effect.

We then combined HU and 5-aza-2’-dC at fixed ratios, empirically determined for each of the four myeloid leukemia cell lines (Figure 3.10A/B). In each of the four cell lines tested, the combination treatment inhibited cell proliferation more than either treatment alone. In order to make a quantitative determination of the drug interaction we performed fixed drug ratio treatments and calculated Chou-Talalay Combination Indices (CI) where CI <1, =1, >1 indicate synergism, additive effect, and antagonism, respectively. The combination of HU and 5-aza-2’-dC demonstrated evidence of drug synergism across a range of drug concentrations in the four cell lines tested (Figure 3.10A/B). We repeated the drug treatments with K562 cells in a soft agar assay to confirm the synergistic effect of HU and 5-aza-2’-dC on clonal cell expansion. Consistent with the proliferation studies, the combination HU and 5-aza-2’-dC reduced colony formation compared to either treatment alone. Together, these results demonstrate a synergistic interaction between HU and 5-aza-2’-dC (data not shown).
We next assessed whether DNA demethylation related to the synergistic drug effect observed. Specifically, we determined the DNA methylation profile of K562 cells treated at a low, mid, and high concentration of 5-aza-2’-dC and HU at a fixed ratio by RRBS (Figure 3.10C). The low average genome-wide CpG methylation levels of approximately 35% in K562 cells with few highly methylated CpGs is consistent with a prior studies reporting overall global hypomethylation inherent to K562 cells36 (Figure 3.11, DMSO-treated control conditions). Nonetheless, treatment with a fixed ratio of lower HU and 5-aza-2’-dC concentrations, that induced a synergistic effect on cell growth (Figure 3.10A, low condition), reduced DNA methylation compared to the respective low 5-aza-2’-dC treatment alone (Figure 3.10C). As expected, HU treatment alone did not alter DNA methylation levels (Figure 3.10C). As with MEFs, filtering by CpGs that are highly methylated in control conditions best displayed the enhancing effect of HU to low 5-aza-2’-dC (Figure 3.10C).

Unexpectedly, methylation levels did not appreciably decrease and even increased with the higher dose combinations of HU and 5-aza-2’-dC (mid and high treatment combinations) (Figure 3.10A/C). Particularly at the high concentration combination, HU addition almost completely blunted the effect of 5-aza-2’-dC (Figure 3.10C). We hypothesized that the differing effects of the low and high concentration combinations may be due to interference of cell cycle progression with increasing concentrations of HU, which in turn interferes with the incorporation of 5-aza-2’-dC into DNA during DNA replication. Accordingly, flow cytometry analysis revealed a significant cell-cycle arrest of K562 cells at the high HU concentration, but not at the low concentration (Figure 3.10D).
Our data suggest that at lower concentrations, HU and 5-aza-2’-dC act synergistically on cell growth, at least partially via DNA demethylation, while at higher concentrations, direct effects on cell cycle progression without increased DNA demethylation inhibit cell growth. In summary, these data indicate that the combination of HU and 5-aza-2’-dC synergistically decreases cell proliferation of the four myeloid leukemia cell lines tested. Moreover, it appears that the mechanism of action of this synergistic drug combination changes in a dose-dependent fashion.
3.4 Discussion

Using an Xi-linked luciferase reporter sensitized to reactivate by low concentration 5-aza-2’-dC treatment, we screened genome-wide siRNA and chemical libraries for reactivation activity. We found that inhibition of the RRM2 subunit of the ribonucleotide reductase enzyme increases rates of Xi-linked reporter reactivation. We attribute the effect of RRM2 inhibition on the Xi in MEFs to augmentation of 5-aza-2’-dC incorporation into DNA, which in turn induces increased genome-wide DNA demethylation in a pattern similar to a high dose 5-aza-2’-dC treatment alone. Moreover, treatment of myeloid leukemia cells with 5-aza-2’-dC and the RRM2 inhibitor hydroxyurea together synergistically inhibited cell proliferation and altered DNA methylation levels in these cancer cell lines in a dose-dependent manner.

Our screen utilized a single copy Xi-linked reporter to identify the effect of Rrm2 inhibition, which was then characterized as a genome-wide effect of augmenting 5-aza-2’-dC-mediated demethylation. The extension of our findings from a single gene reporter on the Xi to a genome-wide effect indicates that the Xi can be used as a model system for identifying and targeting general mechanisms of gene silencing. The robust nature of Xi silencing in differentiated cells, however, contributes to one of the challenges of high-throughput screening with this model: XCR is partial and occurs at low rates thus XCR assays must be optimized in sensitivity. Despite the use of the sensitive luciferase assay method for detecting XCR events and co-treatment with a low dose of 5-aza-2’-dC to sensitize for Xi-reactivation, our siRNA and chemical screens produced only a short candidate gene list. We believe that further modification of the Xi-luciferase screening approach with the use of other co-treatments besides 5-aza-2’-dC to increase rates of XCR would lead to identification of more Xi maintenance factors.
Regardless, the adoption of 5-aza-2’-dC in the optimization of this screen in order to sensitize for DNA demethylation ultimately led to identification of a 5-aza-2’-dC-interacting pathway with therapeutic relevance. From the standpoint of optimizing epigenetically acting drugs, monitoring gene reactivation from the Xi can therefore provide a readout of chromatin reprogramming with immediate effects on gene expression.

We used cell proliferation assays and genome-wide methylation level estimates in myeloid leukemia cell lines to gauge the activity of 5-aza-2’-dC. Our data suggest that at a low concentration of 5-aza-2’-dC, the addition of low dose hydroxyurea, increases the fraction of 5-aza-2’-dC that is incorporated into DNA and available to inhibit DNMT1. This DNA incorporation augmentation effect has the potential to represent a therapeutic advantage. The clinical use of 5-aza-2’-dC is hampered by incomplete disease response in AML and MDS and by high rates of adverse effects17,37,38. Its mechanism of action in patients is most likely due to a combination of demethylating and direct cytotoxic actions that differ in their relative contribution according to disease context and 5-aza-2’-dC concentration. At higher doses, 5-aza-2’-dC is thought to form DNA adducts leading to DNA synthesis arrest, which inhibits its DNA incorporation19,39. Higher doses therefore contribute to higher rates of adverse reactions including hematologic toxicities19. Accordingly, lower doses have been favored in more recent clinical trials and have shown greater likelihood in eliciting gene expression changes as well as producing clinic responses in AML and even solid tumors19,21,39. Thus, increasing DNA incorporation of 5-aza-2’-dC at low doses is a promising strategy to increase its therapeutic index by biasing its activity profile towards DNA demethylation.

In this study, we observed synergistic anti-proliferative effect of 5-aza-2’-dC in combination with hydroxyurea, however, did not capture genome-wide methylation changes at all
concentrations to explain this effect. The anti-proliferative effect in the absence of global DNA demethylation changes is possibly secondary to cytotoxic effects such as DNA adduct formation and DNA synthesis arrest. Alternatively, it is possible that differentially methylated loci are preferentially demethylated by 5-aza-2’-dC at lower concentrations and expression of these genes drives the phenotypic effects of inhibiting proliferation, even when mean global methylation levels are not affected.

Previous studies have reported that 5-aza-2’-dC and hydroxyurea drug combination is antagonistic to DNA methylation based on bisulfite sequencing analysis of three loci in two other cancer cell lines\textsuperscript{40}. Our data supports these findings at high concentration hydroxyurea with 5-aza-2’-dC in K562 cells but shows a synergistic effect on DNA demethylation at lower doses of hydroxyurea and with RNR inhibition. The extent of RNR inhibition is likely critical for a synergistic interaction with 5-aza-2’-dC as too little RNR inhibition will not increase DNA incorporation of 5-aza-2’-dC and too much RNR inhibition with lead to S-phase arrest and interfere with 5-aza-2’-dC-mediated passive DNA demethylation (see model Figure 3.6A).

Another relevant disease model to test a potential therapeutic benefit of the combination of 5-aza-2’-dC and hydroxyurea is sickle cell anemia. Current therapies to treat the genetic defect in adult hemoglobin are aimed at reactivating the fetal hemoglobin gene\textsuperscript{41}. Hydroxyurea is a standard therapy that when administered at cytotoxic doses to patients severely affected with sickle cell anemia increases HbF, but only in a subset of patients for unknown reasons\textsuperscript{41}. As opposed to myeloid leukemia, where the efficacy of 5-aza-2’dC is partially attributable to demethylation, in sickle-cell anemia clinical responses to 5-aza-2’-dC do correlate with demethylation of the fetal hemoglobin gene and increases in hemoglobin levels\textsuperscript{41-43}. Thus it is appealing to explore modified dosing schedules of hydroxyurea and 5-aza-2’-dC for sickle cell
patients already receiving these therapies in order to potentially exploit some synergistic effect of combination therapy for raising hemoglobin levels.
Materials and methods

Genome-wide siRNA library plate preparation

The Silencer Mouse Druggable siRNA Library V3 and Extension set V3 (Ambion) were provided as 250 pmol of lyophilized powder in a total of 153 384-well source plates, containing one siRNA per well except in columns 23 and 24, which were reserved for controls. Each of 21,114 genes is represented by mostly 3 unique (some 2 unique) siRNAs on different 384-well plates. Plates were centrifuged at 1700x g, 50 ul of nuclease-free water was added to each well, sealed and briefly vortexed to resuspend the siRNAs in individual wells. RNA concentrations were confirmed by measuring 1 ul of siRNA solution from 14 randomly chosen wells by NanoDrop spectrophotometer (Thermo Scientific). 2 ul of siRNA diluted to 0.5 pmol/ul from each source plate was stamped in duplicate onto Matrix white opaque 384-well tissue culture-treated plates (Thermo Scientific) by BenchCel 4X system with a PlateLoc plate sealer, Vcode Barcode Printer, and Vprep pipettor fitted with a 96 LT head (all from Agilent Technologies) and stored in -80°.

Derivation of MEFs

Xi reporter MEFs were derived from a cross between transgenic male mice bearing a CAG promoter-driven luciferase or H2B-Citrine allele in the Hprt locus and transgenic female mice heterozygous for an Xist knockout allele\(^\text{23}\). MEFs were derived at embryonic day 14.5 and cultured in MEF media (DMEM supplemented with 10% FBS, nonessential amino acids, L-glutamine, penicillin-streptomycin, β-mercaptoethanol) following standard procedures. The reporter MEFs with genotypes Xi\(^{\text{CAG-Luciferase}}\) Xa\(^{\text{AXist}}\) and Xi\(^{\text{CAG-H2BCitrine}}\) Xa\(^{\text{AXist}}\) were obtained at expected Mendelian ratios of 1 out of 4 embryos and identified by PCR genotyping for presence
of an *Xist* knockout allele, presence of a FLP-Frt recombination production in the *Hprt* locus, and lack of Y chromosome gene *Zfy*.

**High-throughput screening siRNA and chemical screening assays**

The screening assay was optimized to maximize the Z-factor statistical measure of signal-to-noise ratio between the positive control of *Dnmt1* knockdown and negative control or no siRNA mock-transfected cells. Pilot experiments sequentially tested individual variables of the assay such as incubation times and reagent types to increase the Z-factor of the assay. The 5-aza-2’-dC concentration of 0.2 uM used in the screen was determined in this empiric fashion, by titrating a range of 5-aza-2’-dC concentrations to determine which would maximally increase the signal separation between *Dnmt1* knockdown and control samples, calculated as the Z-factor of the assay. The Z-factor of the finalized screening assay was 0.11 (Figure 3.2B). Screening data analysis was performed by first normalizing raw luminescence values by robust z-score which is the number of median absolute deviations for a given well luminescence value from the plate median luminescence value.

Primary MEFs from four female Xi-luciferase reporter embryos were thawed in 15 cm² plates, passaged twice at a 1:6 split, pooled to ensure a homogeneous cell population, and then frozen into 144 vials for use in screening and hit validation. For the large-scale screen, for each batch of 30 plates carrying the genome-wide siRNA library, 2 vials of cells were thawed in MEF media. After one day in culture, adherent cells were trypsinized, live cells excluding Trypan blue were counted using a hemocytometer and brought up in suspension with MEF media agitated by a stir bar.
Meanwhile, a batch of 30 plates including duplicates from 15 source plates of 384-well siRNA library were thawed at room temperature, centrifuged, and cleaned with RNAse-reducing solution (Life Technologies). A positive control siRNA targeting Dnmt1 (Ambion AM161526) was stamped by BenchCel 4X system with an 8 channel LT head (Agilent Technologies) into 16 wells of column 24 of each library plate by adding 4ul of nuclease-free water containing 1 picomole of siDnmt1 to each well. The 16 wells of the column 23 were reserved as negative control and contained no siRNA. Transfection was initiated by adding 20ul of Opti-MEM (Life Technologies) and 0.05ul RNAimax (Life Technologies) per well by Multidrop 384 (Thermo Scientific) and incubating for 20 mins to 1 hour. 20 ul of cell suspension containing 2,000 cells with 5-aza-2’-dC (0.4 uM, Sigma) was added to the transfection mix, bringing the final 5-aza-2’-dC concentration to 0.2 uM. Cells were incubated for 3 days in a humidified 37°C incubator at 5% CO₂. 20 ul of media was then aspirated off using an ELx 405 plate washer (BioTek Instruments) and 20 ul of One-Glo luciferase assay reagent (Promega) was added using the Multidrop 384 and incubated for 20 mins. As luminescence data were collected on an Acquest reader (Molecular Devices), quality control for each plate was performed by visual inspection of positive and negative controls on the heat map during data collection.

Chemical screening was performed analogously with several exceptions: 384-well plates were not pre-treated. Rather, 50ul of cell suspension with 2,000 MEFs and 5-aza-2’-dC (0.2 uM) were plated in 15 384-well plates. A positive control mixture was distributed to a row of wells on each plate by mixing 50 ul of cell suspension with 2000 cells per well in 1x MEF media with high concentration 5-aza-2’-dC (10.0 uM). The screening compounds were added to all but positive control wells as 0.5 ul of 1 mM stock in DMSO by Biomek FX (Beckman Coulter).
After 72 hours incubation, 30 ul of media were aspirated off, and the luciferase assay was performed as described for the siRNA screen. Libraries screened include 4,266 compounds from Microsource (2,000), Biomol enzyme inhibitor (337) and bioactive lipid libraries (203), Prestwick chemical library (1,120), and NIH clinical collections (606) at the UCLA MSSR\textsuperscript{46}. The 30 chemicals producing highest luciferase values were chosen for subsequent validation.

High throughput siRNA screening analysis

Genome-wide siRNA screen hits were identified by Redundant siRNA Activity (RSA) analysis using robust z-scores as the input values\textsuperscript{25}. The R script provided by Konig et al. was used with minor modifications to adapt it for our workflow (http://carrier.gnf.org/publications/RSA). RSA works by ranking hits in order of activity then assigning P values for genes based on whether their siRNAs rank higher than would be expected by chance. We obtained two activity measurements for each siRNA since the siRNA library was screened in duplicate, and treated these data points as independent measurements with regard to the analysis. Therefore, most genes were represented by 6 data points (and some with 4 data points) in the RSA analysis.

Cell culture and treatment methods

For subsequent Xi-reactivation/validation assays, MEFs at passage 1 or 2 post-derivation were seeded at a density of 6.0×10\textsuperscript{4} cells per 12-well well and chemicals in MEF media and/or siRNAs in Opti-MEM media (Gibco) were added and incubated for 72 hours. For 5-aza-2’-dC (Sigma), which was resuspended in DMSO and stored at -80°, final DMSO concentration on the cells was kept below 0.1%. Total volumes of MEF and/or Opti-MEM media were normalized across samples when different treatments were used. Hydroxyurea, Resveratrol, Uridine, and Deoxycytidine (Sigma) were resuspended in water and stored at -20°. K562, HL60, U937, and
THP1 cells were purchased from ATCC. K562, U937, and THP1 cells were cultured in RPMI media (Gibco) with 10% FBS and HL60 cells were cultured in IMDM (Gibco) with 20% FBS. ATCC culture method suggestions were followed for expanding the cells.

The soft agar assay was performed by mixing of 1.2% nobel agar (Sigma) in water with 2X RPMI to achieve final concentration of 0.6% agar for the bottom layer. After this solidified in 6-well plates, top soft agar was prepared at final 0.3% nobel agar concentration containing K562 cells to achieve $4.0 \times 10^4$ cells per well. DMSO or 5-aza-2’-dC (0.05 uM) and/or HU (50 uM) were added to both bottom and top agar layers. This 1000:1 ratio of HU to 5-aza-2’-dC was determined to be optimal for the soft agar assay, which is different from the 4000:1 optimal ratio used in CellTiter Glo assay. Small colonies started appearing four days after plating. On day 8, colonies were stained with 0.01% crystal violet for 1 hour, washed with PBS, and the plates were scanned to obtain images.

**Luciferase assay**

For each luciferase assay, MEF Xi-luciferase reporter treatments were performed in triplicate 12-well wells for 72 hours and lysed with 200 ul passive lysis buffer (PLB, Promega) for 20 mins at room temperature on an orbital shaker. Lysates were cleared by 30 seconds of centrifugation and 20 ul were assayed for luciferase activity with 50 ul of LARI reagent (Promega) on a GloMax microplate luminometer (Promega). Protein concentration measurements were performed on corresponding PLB lysates by Quick Start™ Bradford Protein Assay Kit (Bio-Rad) and analyzed by interpolating to standard curve according to Manufacturer’s Instruction. For the proliferation assays of leukemia cell lines, 100 ul of well-suspended cells were mixed with 100 ul of CellTiter
Glo® reagent (Promega), incubated at room temperature for 20 mins, and luciferase units were measured using a GloMax microplate luminometer (Promega).

**RT-qPCR analysis**

Cells were harvested from a 6-well format in Trizol (Invitrogen) and RNA purification was performed with the RNeasy kit (Qiagen) according to manufacturer’s instructions with on-column DNAse treatment. cDNA was prepared using SuperScript III (Invitrogen) with random hexamers and RT-qPCR was performed using a Mx3000 thermocycler (Stratagene) with primers for *Rrm2* (F-GCACTGGGAAGCTCTGAAAC, R-GGCAATTTGGAAGCCCATAGA), *Dnmt1* (F-CATGAAATTCTGCAAACAGAA, R-TTGACTTTAGCCAGGTAGCC), or *Gapdh* (F-GGCCTTCCGTGTTCC, R-GCCTGCTTCACCACCTTCT). Results were normalized to *Gapdh* by the ΔCt method.

**Knockdowns in follow-up experiments**

Knockdowns by siRNA were performed by reverse transfection at 25 nM final concentration of siRNA. Briefly, a cell suspension was added to a pre-incubated mixture of Lipofectamine RNAimax, 100 ul of reduced serum Opti-MEM media, and siRNA. The siRNAs used were *Rrm2* (Ambion, 150659 (A), 64497 (B), 150661 (C)), *Dnmt1* (Ambion, 161526), and, as negative controls, *Scramble* (Ambion, 4636), *Luciferase* (Dharmacon, D-001210-02), *Aurkb* (Dharmacon, D-063793-01), and *GFP* (Dharmacon, P-002048-01). For *Rrm2* knockdown where the siRNA is not specified, siRNA 66497 was used.

**Flow cytometry**

Flow cytometry for measuring the reactivation of the Xi-linked H2B Citrine reporter was performed as described previously. For the cell cycle measurement with K526 cells, $5.0 \times 10^6$
cells (determined by trypan blue exclusion assay) were taken from each treatment condition, washed once with PBS, and stained with propidium iodide buffer (3mM EDTA pH 8.0, 0.05% NP40, 50ug/ml PI, 1mg/ml RNaseA in PBS) for 30 min at room temperature. Stained cells were passed through a strainer and analyzed by FACSDiva (BD Biosciences) with FlowJo software (Tree Star, Inc.).

3H Decitabine Incorporation

This assay was analogous to the reactivation treatment assays with a few modifications: assays were scaled 2.5-fold to 6-well format, 1 ul (1 uCi) of tritiated 5-aza-2’-dC (3H-Decitabine, Moravek Biochemicals Inc.) was added instead of cold 5-aza-2’-dC, and samples were harvested after 48 hrs of incubation. Cells were trypsinized, genomic DNA isolated using the Quick-gDNA MinPrep kit (Zymo Research), and measured by QuBit fluorometer (Life technologies). Tritium content of 25 ul of genomic DNA was measured using a scintillation counter and normalized to the measured DNA concentration.

Reduced Representation Bisulfite Sequencing

Primary Xi-reporter MEFs were subjected to the same chemical treatment as used for the luciferase assays, but in 6-well format, and a fraction of the cells was taken to confirm appropriate luciferase reporter activity. Genomic DNA was isolated using the Blood and Cell Culture Mini Kit (Qiagen) with RNase A treatment (Life Technologies). The RRBS libraries were generated at previously described by Orozco et al. with minor modifications. DNA purifications for each enzymatic reaction was carried out using AMPure XP beads (Beckman Coulter). Bisulfite conversion was performed using the Epitect kit (Qiagen) twice compared to manufacturer’s instruction to optimize the efficiency. Bisulfite-converted libraries were
amplified using MyTaq Mix (Bioline) with the following program: (98° for 15 sec, 60° for 30 sec, 72° for 30 sec) 12 cycles, 72° for 5 min, 4°C storage.

RRBS data analysis

DNA Methylation calling was performed using BS-Seeker2 using Bowtie 0.12.9 for read alignment on the UCLA Hoffman2 computer cluster\textsuperscript{48}. CpG islands (CGIs) were obtained from UCSC (http://genome.ucsc.edu) and CGI tracks were based on methods by Gardiner-Garner and Frommer\textsuperscript{49}. Promoters were defined as the region transcription start site (TSS) minus 1kb to TSS for all UCSC genes. Only sites covered by at least 5 reads across all samples under consideration were used in an effort to obtain reliable methylation levels. The methylation levels of samples were hierarchically clustered using complete linkage and the Euclidean distance metric. Statistical analysis, clustering, and heat map generation were performed using custom R scripts\textsuperscript{50}.

Data submission

All RRBS data will be made available on GEO at the time of publication.
Figure 3.1  High-throughput siRNA and chemical screens identify RRM2 depletion and Resveratrol as mediators of XCR
A. Schematic of the X chromosomes in female reporter MEFs carrying the luciferase reporter transgene in the *Hprt* locus specifically on the Xi. The *Xist* deletion on one of the chromosomes skews X-inactivation to the wild-type *Xist*-bearing X chromosome.

B. Diagram of the screening workflow. siRNAs from the mouse genome-wide library and selected chemical libraries were assayed in 384-well plates containing a column of positive and negative controls. Xi-luciferase reporter MEFs were added and incubated for 72 hours in the presence of 5-aza-2’-dC (0.2 uM) prior to luciferase assay.

C. Gene activity distribution plot ranked by the –log of the p-value obtained with the redundant siRNA activity (RSA) assay from duplicate genome-wide siRNAs screens following transformation of the luminescence activity values into robust z-scores. The top validated hits, *Dnmt1*, *Atf7ip*, and *Rrm2*, are labeled.

D. (i) Graph depicting Xi-luciferase reporter reactivation upon knockdown of *Rrm2* with the three siRNAs (A, B, C) present in the genome-wide library in the presence or absence of 5-aza-2’-dC (0.2 uM). Luminescence was measured 72 hours after the start of the treatment. Error bars indicate standard deviation of luminescence unit values from three individual wells with a given treatment in one experiment. (ii) RT-qPCR for RNA levels of *Rrm2* normalized to siGFP control and *Gapdh* expression. RNA was harvested in parallel to luciferase assays shown in (i). Error bars indicate standard deviation from three measurements in one experiment.

E. Activity of chemicals in the chemical screen in the presence of 5-aza-2’-dC (0.2 uM), ranked by luminescence unit with the value corresponding to Resveratrol designated.

F. Xi-luciferase reporter assay as described in (Di) titrating the Resveratrol concentration with or without 5-aza-2’-dC (0.2 uM).
G. Xi-luciferase reporter assay as in (Di) titrating hydroxyurea with or without (untreated) 5-aza-2’-dC (0.2 uM). Luminescence activity was negligible for untreated samples. The result for Resveratrol treatment in the same experiment is given for comparison.

H. (i) Xi-luciferase reporter assay as in (Di) comparing the consequences of 0.2 uM 5-aza-2’-dC treatment and siRNA-mediated knockdown of Dnmt1 to elicit reporter reactivation by 20 uM Resveratrol. Luminescence activity was negligible for untreated samples. (ii) RT-qPCR for Dnmt1 RNA levels normalized to siGFP control and Gapdh expression in the same experiment as (i).
**Figure 3.2**  Optimization of 5-aza-2’-dC concentration for genome-wide siRNA screen

A. Bar chart illustrating luciferase activity from Xi-reporter MEFs upon knockdown of *Dnmt1* and treatment with varying concentrations of 5-aza-2’-dC in 384-well format for 72 hours. Error bars indicate standard deviation from eight measurements in one experiment. Asterisks indicate p<0.01 by Student’s T-test.

B. Scatterplot of luminescence values from the optimized Xi-reactivation screening assay in 384-well format in the presence of 5-aza-2’-dC (0.2 uM) with *siDnmt1* (red) or negative control *siAurkb* (Aurora kinase B, blue). The Z-factor, a measure of separation between positive and negative control populations used in the assessment of high-throughput assays, is shown.⁴⁵
Figure 3.3  Batch effects in siRNA screen and robust z-score normalization

A. Box plot of all raw luciferase measurements distributions per individual 384-well plate from one of the duplicates of the siRNA screen. These plates were prepared and assayed in 30-plate batches according to their numerical order in the source library plates, keeping duplicate plates together.

B. As in (A) except each measurement was normalized by the robust z-score\textsuperscript{45} (median absolute deviations from the plate median).
Figure 3.4 Validation of gene hits identified by genome-wide siRNA screening

The chart displays the luminescence for the Xi-luciferase assay in 24-well format with knockdown by the indicated siRNAs, chosen as top hits of the genome-wide screen, in combination with 5-aza-2’-dC (0.2uM) for 72 hours. For each gene hit, siRNAs were re-ordered to match the sequences of the 2 or 3 active siRNA identified by RSA activity analysis of the genome-wide siRNA screen. Error bars indicate one standard deviation from duplicate wells. *siDnmt1* positive control is shown in red.
Figure 3.5  Chemical screen results and validation

A. Box plot of all raw luciferase measurements from the chemical screen by individual 384-well plate, demonstrating lack of obvious batch effect. Chemical library plates were prepared and assayed as one batch of 15 plates.

B. Chart displaying results from the Xi-luciferase assay in the 24-well format upon treatment with various chemicals (at 10 uM) in the presence of 5-aza-2’-dC (0.2uM) for 72 hours. Error bars indicate one standard deviation from duplicate wells except for negative control 5-aza-2’-dC (0.2 uM) alone (n=16) and positive control 5-aza-2’-dC (10.0uM) alone (n=16). Resveratrol is indicated with an asterisk.
Figure 3.6  Inhibition of RNR enhances DNA incorporation of 5-aza-2’-dC to elicit XCR

A. Illustration of model in which (1) inhibition of ribonucleotide reductase (RNR) by various means leads to (2) increased relative dCTP utilization for DNA synthesis from salvage pathways which are supplemented with exogenous 5-aza-2’-dC. (3) DNMT1 inhibition occurs upon binding to DNA-incorporated 5-aza-2’-dC leading to (4) increased loss of DNA methylation with successive cell divisions.

B. Quantification of 3H-5-aza-2’-dC (3H-Decitabine) incorporation into genomic DNA upon either Resveratrol treatment or Rrm2 knockdown for 48 hours. Genomic DNA was isolated and an equal volume measured for 3H-Decitabine incorporation (disintegrations per minute, DPM), then normalized to the amount of DNA loaded (ug). Error bars indicate standard deviation from three independent treatment wells. Asterisks indicate p<0.01 by Student’s T-test.

C. Xi-luciferase reactivation assay as in Figure 3.1Di in the presence of 0.2uM 5-aza-2’-dC and 20 uM Resveratrol and increasing concentrations of deoxycytidine (dC) or uridine.

D. As in (C) except with siRrm2 in the place of Resveratrol.
Figure 3.7  RNR inhibition increases 5-aza-2’-dC-mediated DNA methylation in MEFs

A. Bar chart displaying average genome-wide CpG methylation levels for the indicated 72 hrs treatments filtered for CpGs with at least 5X sequencing coverage by RRBS across samples. Label color reflects the various treatment groups. Subscripts (A and B) indicate duplicates where applicable. Treatment concentrations include LowAza (5-aza-2’-dC 0.2 uM), HighAza (5-aza-2’-dC 10.0 uM), and Resv (Resveratrol 20 uM).
B. Heat map of unsupervised hierarchical clustering of autosomal CpG methylation levels assayed by RRBS in MEFs treated with the indicated chemicals for 72 hours as in (A) with at least 5X sequencing coverage by RRBS across samples. A methylation level of 1 indicates 100% methylation, while 0 represents complete absence of methylation.

C. Heat map as in (B) but for subsets of autosomal CpG sites partitioned into four groups representing different DNA methylation levels in the untreated control sample as (i) 0.75 - 1.0, (ii) 0.50 - 0.75, (iii) 0.25 - 0.50, and (iv) 0 - 0.25. In each case, the combination of RNR inhibition with 0.2 uM 5-aza-2’-dC clusters away from all other samples, but together with the high dose of 5-aza-2’-dC, and appears more demethylated.

D. Histograms display whole-genome CpG methylation level distribution for the indicated treatments, based on CpGs with at least 5X sequencing coverage by RRBS across samples.
Figure 3.8  Analysis of DNA methylation in MEFs treated with combinations of RNR inhibition and 5-aza-2'-dC
A. Heat map of the unsupervised hierarchical clustering of CpG methylation levels in MEFs as in Figure 3.7B with at least 5X sequencing coverage by RRBS across samples, except that the data for X chromosome CpG sites are shown.

B. (i) Heat maps of the unsupervised hierarchical clustering as in Figure 3.7B but only for CpG sites within CpG islands. Genomic locations of CpG islands were obtained from UCSC Genome browser (see methods). (ii) Heat maps of the unsupervised hierarchical clustering as in Figure 3.7B but only for CpG sites within promoters. Promoters were defined as the region TSS -1 Kb to TSS for all UCSC genes.

C. Heat maps of the unsupervised hierarchical clustering as in Figure 3.7C except CpG sites are filtered by the methylation level in the untreated sample of either (i) 0.75-1.0, (ii) 0.50-.75, (iii) 0.25-0.50, or (iv) 0-0.25.
Figure 3.9  DNA methylation status of the luciferase transgene in MEFs treated with combinations of RNR inhibition and 5-aza-2’-dC
A. Bar chart displaying average genome-wide CpG methylation levels for the indicated treatments filtered by CpGs with at least 5X sequencing coverage by RRBS across all samples as in Figure 3.7A, but only considering the CpGs in the luciferase reporter gene/promoter.

B. Heat map of unsupervised hierarchical clustering of CpG methylation levels as in Figure 3.6B except for CpG sites in the luciferase reporter gene/promoter.

C. As in (B), except for CpG sites within the luciferase reporter with a methylation level greater than 0.75 in both the untreated sample.

D. Histograms showing the distribution of CpG methylation levels within the luciferase reporter gene.
Figure 3.10  Hydroxyurea and 5-aza-2’-dC treatment of myeloid leukemia cell lines

A. Dose-response curves measuring viable K562 cells using the Cell-Titer Glow Assay (Promega). Chemical treatments were performed with hydroxyurea (HU) or 5-aza-2’-dC (Aza) alone or in combination at a fixed concentration ratio of 4000:1 HU:Aza (based on individual IC50 values). Red lines indicate concentrations assayed in the subsequent RRBS analysis displayed in (C). The table depicts the Chow-Talalay analysis of Combination Index (CI) at the given treatment combination concentrations. CI values < 1.0 indicates synergy, 1 = additive effect, and > 1.0 antagonism of the combination drug effect.

B. As in (A), except dose-response curves for HL60 cells using a fixed concentration ratio of 150:1 HU:Aza in the combined treatments.

C. Heat map showing an unsupervised hierarchical clustering of X chromosome CpG
methylation levels in K562 cells treated with the indicated chemicals for 72 hours. CpGs with greater than 0.75 methylation level in the DMSO-treated samples are displayed, as these are the CpGs most dramatically affected by 5-aza-2’-dC treatment. CpGs are filtered for at least 10X sequencing coverage across all samples. The combined treatment conditions used (Low, Mid, High) are indicated with red lines in (A). In addition, we treated cells with either 5-aza-2’-dC or HU at the respective concentrations (Low HU (0.16 mM), Mid HU (0.4 mM), High HU (0.80 mM), and Low Aza (0.04uM), Mid Aza (0.1 uM), High Aza (0.2 uM)). Aza is resuspended in DMSO thus DMSO only controls are matched to volume of Aza added in corresponding Low and High Aza samples.

D. Flow cytometry analysis of propidium iodide-stained K562 cells treated with low and high HU concentrations as described in (C) compared to low or high DMSO control treatment.
Figure 3.11  Extended data on the methylation analysis of K562 cells

Histograms showing CpG methylation distribution on the X chromosome in K562 cells, for CpGs with at least 10X sequence coverage across all samples as determined by RRBS. Chemical treatments are as shown in Figure 3.10A/C.
References


Chapter 4

A genome-wide chronological study of gene expression and two histone modifications, H3K4me3 and H3K9ac, during developmental leaf senescence
4.1 Abstract

The genome-wide abundance of two histone modifications, histone 3 lysine 4 trimethylation (H3K4me3) and histone 3 lysine 9 acetylation (H3K9ac), both associated with actively expressed genes, was monitored in *Arabidopsis thaliana* leaves at different time points during developmental senescence, along with expression in the form of RNA-seq data. H3K9ac and H3K4me3 marks were highly convergent at all stages of leaf aging, but H3K4me3 marks covered nearly twice the gene area as H3K9ac marks. Genes with the greatest fold-change in expression displayed the largest positively-correlated percent change in coverage for both marks. Most senescence up-regulated genes were pre-marked by H3K4me3 and H3K9ac, but at levels below the whole-genome average, and for these genes, gene expression increased without a significant increase in either histone mark. However, for a subset of genes showing increased or decreased expression, the respective gain or loss of H3K4me3 marks were found to closely match the temporal changes in mRNA abundance. 22% of genes that increased expression during senescence showed accompanying changes in H3K4me3 modification, and they include numerous regulatory genes, which may act as primary response genes.
4.2 Introduction

Leaf senescence is the final nutritive stage of leaf development in which the organ that provided sugars to the growing plant undergoes a controlled degradation process that recycles many of the nutrients located in the protein-rich photosynthetic apparatus. Leaf senescence is regulated by developmental age and can be accelerated by adverse environmental conditions such as drought, nutrient deprivation and heat stress. Ethylene, jasmonic acid and salicylic acid promote, while cytokinin prevents leaf senescence\(^1\). Numerous microarray studies have catalogued major changes in gene expression during leaf senescence and a large overlap with defense responses has been noted\(^2-5\).

Towards understanding the mechanisms that regulate gene expression during senescence, numerous studies have demonstrated that transcription factors such as WRKY53\(^6-8\), ORE1/NAC2\(^9,10\), and EIN3\(^11\) promote leaf senescence, as well as many others listed in the Leaf Senescence Database\(^12\). In addition, mutants that affect chromatin structure have been shown to affect senescence, however phenotypes are often pleiotropic\(^13-16\).

In a previous study, changes in two euchromatin histone modifications, histone 3 lysine 4 trimethylation (H3K4me3)\(^17\) and histone three lysine 27 trimethylation (H3K27me3)\(^18\), were measured on a genome-wide scale by ChIP-seq in young and old leaves\(^19\). This study did not have accompanying gene expression data and published microarray data\(^4\) were used to compare gene expression and changes in histone methylation status. Even with these limitations, a correlation between the H3K4me3 mark and gene expression was apparent. Genes that displayed a gain in the H3K4me3 mark were most highly represented by genes expressed to a greater extent in older leaves and the opposite distribution was observed for genes that lost the H3K4me3 mark. Genes that lost the H3K27me3 mark were expressed in older leaf tissue, but
overall gene numbers were low and a gain of the H3K27me3 mark for down-regulated genes was rarely observed.

H3K4me3 marks are localized to actively transcribed genes, but they become associated with chromatin after RNA polymerase II binding\textsuperscript{20,21} and are thus considered to be marks for transcribed genes. However, interaction between the H3K4me3 mark and TAF3, a TFIID subunit of the pre-initiation complex, has been observed in human cells\textsuperscript{22}, suggesting the H3K4me3 mark plays a role during the initiation of transcription. In addition, the Arabidopsis trithorax1 (ATX1) methyltransferase was found to be important for recruiting TATA-binding protein and RNA polymerase II to promoters. Interestingly this activity was independent of the ATX1 methyltransferase activity, which was necessary for elongation of transcription\textsuperscript{23}. The extent of H3K4me3 modifications was recently shown to be greatest for cell identity genes, and in young leaves, genes with the greatest breadth of H3K4me3 coverage were enriched for photosynthesis\textsuperscript{24}.

A second well-studied histone mark is histone 3 lysine 9 acetylation (H3K9ac)\textsuperscript{25}. Changes in H3K9 acetylation have been shown to correlate with changes in gene expression during de-etiolation and after UV-B treatment\textsuperscript{26,27}. The timing of H3K9ac gain or loss closely paralleled gene expression changes for the circadian clock components and response to drought stress\textsuperscript{21,28}, or occurred after genes had been activated, in the case of flowering\textsuperscript{29}. Thus the exact role of H3K9ac in gene activation is not yet defined, and it likely plays numerous roles.

In this study, both of these well-established histone marks, which positively correlate to gene expression – H3K4me3 and H3K9ac – have been measured on a genome-wide scale using chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq), along with
accompanying RNA sequencing (RNA-seq) data at different stages of leaf aging. Most (78%) senescence up-regulated genes were pre-marked with H3K4me3 and H3K9ac and most (85%) of down-regulated genes retained both marks even when gene expression had dropped to low levels. However, the breadth of the H3K4me3 and H3K9ac marks were positively related to gene expression. For a subset of genes (22% of up-regulated genes and 15% of down-regulated genes), a strong relationship between temporal changes in gene expression and gain/loss of the H3K4me3 mark was observed during leaf aging. By identifying genes that show combined changes in gene expression and histone marks, we have produced a list of genes that include many known to play important roles in senescence, as well as potentially novel players in this important biological process.
4.3 Results

4.3.1 RNA-seq gene expression analysis

RNA-seq was performed on mature, fully-expanded rosette leaf tissue at four time points: 29d, 35d, 42d and 57d. These time points were chosen to encompass the steps of leaf senescence, which commences soon after the development of the inflorescence\(^2\). Three independently sampled replicates were subject to RNA-seq using multiplex single-end Illumina sequencing (Table 4.1). Replicates sampled at each day show a high degree of similarity with the 29d samples being the most distinct (Figure 4.1A), allowing us to use the replicates to assess variation for confident differential expression analysis, as well as to subsequently pool the replicate read count for the calculation of RPKMs (reads per kilobase transcript length per million reads). The initial data analysis identified genes that showed significant changes in expression between adjacent time points (≥ 2-fold change in expression, p-value ≤ 0.05): 29-35d, 35-42d, and 42-57d. Confirming the distinct separation of the 29d replicates and those of the other time points seen in the dendrogram of the clustering results (Figure 4.1A), the largest number of genes was up- (Figure 4.1B) or down- (Figure 4.1C) regulated during the first time interval, indicating that major changes in gene expression occur prior to the visual manifestations of leaf aging: the loss of chlorophyll. The greatest overlap between time intervals was between 35-42d and 42-57d for up-regulated genes and 29-35d and 35-42d for down-regulated genes (Figure 4.1B/C).

A hypergeometric test revealed a significant overlap between all intervals for up-regulated genes (p-value ≤ 5\(\times\)04) indicating that although most genes are fully up-regulated by 35d, a significant proportion continue to be more gradually up-regulated across time points.
The down-regulated genes only showed a significant overlap across the first two intervals (p-value < 1.1e-81) indicating that a largely distinct set of genes decreases from 42 to 57d (Figure 4.1C).

We sought to generate a high-confidence set of senescence up-regulated genes (SURGs) and senescence down-regulated genes (SDRGs), by requiring that they showed significant changes in expression (≥2-fold, p-value ≤ 0.05) in two of six pairwise comparisons (29-35d, 29-42d, 29-57d, 35-42d, 35-57d and 42-57d). This analysis permits the inclusion of genes with significant changes in expression in just one interval; for example the distinct group of genes that are down-regulated between 42-57d (Figure 4.1C) will be listed in the 29-57d, 35-57d and 42-57d pairwise comparisons. To remove genes with low expression we also required that they have RPKM values (after merging replicate data sets) above the median RPKM value (0.764 for 29d, 0.911 for 35d, 0.790 for 42d and 0.752 for 57d) at the time of higher expression. Figure 4.1D/E shows the robust up- and down-regulation of expression in the SURGs and SDRGs, respectively. As was generally the case, the biggest changes in expression occurred between 29d and 35d, but the respective upward and downward trends persisted for the duration of the time course.

In contrast to the SURGs and SDRGs, the expression distributions of all other genes show no trend indicating the classification was reasonable (Figure 4.1F). Figure 4.1G shows that setting the threshold at two of six pairwise comparisons resulted in a fair estimate of gene expression changes that represented a good compromise between overly stringent and more lenient criteria. WRKY transcription factor genes, usually associated with senescence and thus representing a likely false positive, were observed in the down-regulated category when only one of six pairwise comparisons was the threshold; conversely SAUR genes, down-regulated in
senescence and representing a likely false positive, were seen at increasing numbers in the up-regulated category when only one of six pairwise comparisons was the threshold. Genes encoding bHLH transcription factors showed no preference for up- or down-regulation during leaf aging and the numbers of these genes became more plentiful as the threshold decreased.

The selection procedure described above resulted in 1432 SURGs (plus 11 pseudogenes and 6 transposable element genes) and 964 SDRGs (plus 4 pseudogenes and 5 transposable element). Small numbers of transposable element genes, mostly retroposons, and pseudogenes were both up- and down-regulated during leaf senescence. This contrasts to animal cells where a general increase in expression and mobility of transposable element genes has been observed in older somatic cells. A Gene Ontology (GO) analysis was performed on this list of SURGs and SDRGs and enriched Biological Processes with false discovery rate (FDR) below 1% are shown in Table 4.2. Genes related to defense, jasmonic acid and transport were enriched in SURGs as expected. In addition, enrichment for indole glucosinolate synthesis genes suggested a role for these secondary metabolites during senescence. SDRGs were enriched for photosynthesis and growth-related processes such as response to auxin stimulus, response to light stimulus, response to gibberellin, lipid biosynthesis, and cell wall organization.

4.3.2 ChIP-seq analysis for H3K4me3 and H3K9ac

Nuclei were prepared from the same tissue used in RNA-seq, and ChIP-seq was performed using an antibody that recognized H3K4me3. A second set of tissue was grown and harvested at 30, 34, and 42d for H3K9ac ChIP libraries. Although the days were not identical, the developmental stages were similar to the first three time points of the first experiment. ChIP and input sequencing reads are summarized in Table 4.3. Figure 2A shows two typical gene-rich regions
on chromosome 1 and chromosome 5 with H3K4me3 and H3K9ac read counts. The peaks of read counts are most commonly located in the 500 bp region downstream of the TSS (Figure 2B).

Regions of significant H3K4me3 and H3K9ac ChIP enrichment were determined (see Methods section). Approximately 1.82 Mb of the 115 Mb genome was found to be significantly enriched for one or the other of the two histone marks relative to input, with the marks coinciding across 1.42 Mb (78%; Figure 2C). A cooperative interaction between H3K4me3 and H3K9ac modifications is suggested by the high degree of colocalization, as has been noted previously.

4.3.3 Changes in histone marks during leaf senescence

To determine if the levels of the two histone marks changed significantly during leaf senescence, six pairwise comparisons were made between ChIP read counts at the four different time points for the H3K4me3 mark and three pairwise comparisons were made for the three different time points for H3K9ac read counts. Based on these pairwise comparisons, a region was defined as displaying a consistent and significant gain or loss of histone marks during the course of the experiment, if at least three of the six pairwise comparisons for H3K4me3 showed significant (–log(p-value) ≥ 6) changes, or at least two of three comparisons for H3K9ac, thereby ruling out spurious changes. We used these more stringent thresholds for the histone marks relative to gene expression (2 of 6 pairwise comparisons and –log(p-value) ≥ 1.3) because the lower requirement for leaf material allowed us to produce replicate RNA-seq libraries (n=3; Figure 4.1A), which in turn allowed for greater confidence in the differential expression results. Representative regions displaying a gain in H3K4me3 or H3K9ac are shown in Figure 4.3.
Panel 3A shows At5g13080 that encodes WRKY75, a TF known to be important in regulating leaf senescence\(^{34}\). ChIP and input read counts in progressively older leaves are displayed at different time points, and the increase in H3K4me3 is clearly seen. The region that shows significant gain is indicated in the “K4_GAIN 3 of 6” track, and three of the six comparisons (29-35d, 29-42d and 29-57d) show $-\log(p \text{ values}) \geq 6$. H3K9ac is enriched at the *WRKY75* locus, but does not show a change in abundance. In panel 3B, the gain in H3K9ac marks can be seen for At1g66760, which encodes a member of the MATE efflux transporter family. The “K9_GAIN 2 of 3” track displays regions with $-\log(p \text{ values}) \geq 6$ in two of three comparisons (30-34d and 30-42d). This gene has consistent levels of the H3K4me3 mark.

Overall, 564 genes showed a significant gain of H3K4me3 within the region from the TSS to 500 bp downstream during leaf aging. 222 genes were associated with a significant loss of H3K4me3 over the same period. Of these genes, 56% (315/564) that gained the H3K4me3 mark were SURGs and 63% (139/222) that lost the H3K4me3 mark were SDRGs. This demonstrates a high positive correlation between significant changes in these histone modifications and gene expression. For H3K9ac marks, the number of genes associated with a consistent significant gain (128) and loss (150) was lower than for H3K4me3 marks, in part due to one less time point and reduced read counts as compared to H3K4me3. Furthermore, the correlations with gene expression were not as high: 20% (26/128) for gain of the acetylation mark and up-regulation and 17% (26/150) for loss of the acetylation mark and down-regulation. In fact, of the 128 genes that significantly gained the H3K9ac mark, 51 (40%) showed no upward or downward trend in expression during the course of leaf senescence, while 59 of the 150 genes (39%) that significantly lost H3K9ac had unchanged gene expression.
Considering SURGs as a whole, 22% (315/1432) gained H3K4me3, while only 2% (26/1432) gained the H3K9ac mark. Similarly, 14% (139/964) of SDRGs displayed a loss of the H3K4me3 mark, with 3% (26/964) showing H3K9ac loss. Intrigued by these relatively low proportions, we first studied the 1117 SURGs (1432–315 genes) and 825 SDRGs (964–139 genes) that were not associated with H3K4me3 acquisition or loss. Average H3K4me3 and H3K9ac read counts ± 2500 bp from the TSS for these genes are displayed in Figure 4.4. These SURGs were marked with H3K4me3 at the earliest time point, and a slight increase in marks was apparent, however counts remained below the whole-genome time point average (Figure 4.4A). Interestingly H3K9ac counts increased with gene expression and were above the whole-genome time point average at 42d (Figure 4.4B). Increasing gene expression RPKM values are shown in Figure 4.4C. These two histone modifications were present before the first harvest and prior to the significant increase in expression captured by the RNA-seq time course experiments, yet for these genes, only the H3K9ac mark more closely reflected gene expression. For this subset of SDRGs, there is a decrease in H3K4me3 modification levels that takes place in parallel with decreased mRNA abundance, however there was only a moderate enrichment for this mark (approximately the same as on average) and the losses were not significant by our criteria, with a peak still evident at 57d, when RPKM levels were low (Figure 4.4D, F). For H3K9ac, the modification levels hover below the whole genome time point average and do not correspond to changes in gene expression (Figure 4.4E). The decreasing gene expression profiles for the SDRGs not associated with a significant decrease in histone mark levels are shown in Figure 4.4F. Thus, for SDRGs, a decrease in H3K4me3, albeit not a significant one, still accompanies reduced gene expression, yet both marks are retained even in older tissue.
Next, we considered those genes for which significant and consistently changing histone modifications and gene expression were identified. The 315 genes with increases in both H3K4me3 and mRNA levels (K4-SURGs) include NAC- and WRKY-domain transcription factors, numerous classes of receptor-like protein kinases, late embryogenesis abundant dehydration factors, MATE efflux transporters and RING/U-box proteins. The number of genes with increased H3K9ac and mRNA levels (K9-SURGs) is much smaller (26) and these include the WRKY41 transcription factor known to play a role in defense. Abundant gene classes within the 139 K4_SDRGs (genes with a decrease in both H3K4me3 and mRNA) include tubulin, gibberellin response and cell wall extension. Three YUCCA (auxin biosynthesis) and 13 SAUR-like (auxin response) genes show decreased expression accompanied by a loss in H3K4me3 marks. The 26 K9_SDRGs include the PSAK gene encoding a Photosystem I subunit as well as one SAUR-like gene. The intersection of K4 & K9-SURGs was only 9 genes while that for K4 & K9-SDRGs was 11 genes (data not shown).

4.3.4 Temporal patterns in histone mark acquisition and loss

Our time course allowed a comparative analysis of temporal changes in gene expression and histone modifications. Regions that showed significant gain or loss for each histone modification were subject to k-means clustering to produce three groups with different temporal patterns. Figure 4.5 shows the clustering for H3K4me3 acquisition (panel A) and how gene expression for the genes in each cluster coincides with the trend seen in the average histone mark profiles (panel B). The average H3K4me3 profiles for each cluster of genes at each time point are shown for the region -2500 to +2500 bp in relation to the transcription start site (TSS) and the mean levels for all genes at each time point are denoted with the dashed lines. In all cases, H3K4me3 peaks are
centered approximately 400 bp downstream from the TSS, as seen before. A rapid increase in H3K4 trimethylation is observed between 29 and 35 days for cluster #1, and this is reflected in the accompanying gene expression box plot (Figure 4.5B, cluster #1). The more gradual increase in H3K4me3 marks for cluster #2 is also mirrored in a more gradual increase in gene expression for genes in cluster #2, and the more subtle increases in H3K4me3 marks are accompanied by a smaller fold-increase in gene expression for cluster #3 (Figure 4.5B, clusters #2 and #3). The p-values for all pairwise comparisons (after correction for multiple testing) can be viewed in Table 4.4, and in all but one case, differences in gene expression are significant. The one case in which gene expression did not differ significantly, 42-57d for cluster #1, showed a concomitant negligible gain in H3K4 trimethylation. K-means clustering was also used to partition regions exhibiting a loss of H3K4me3 into three groups reflecting different temporal patterns of diminution (Figure 4.6A).

As with the gain of H3K4me3 marks, the magnitude and timing of loss in histone marks and decreases in gene expression mirror one another for all three clusters (Figure 4.6B). For all clusters, the decrease in gene expression was not significant between 42 and 57d (Table 4.4), and H3K4me3 marks are at nearly identical levels. These correlations support an active and finely-tuned role for the H3K4me3 mark in regulating gene expression during leaf aging for a subset of genes.

A corresponding analysis was performed for peaks that gained the H3K9ac histone mark, which revealed less distinct temporal trends. Of the 33 genes comprising the H3K9ac gain cluster #3, a significant increase in gene expression was observed during leaf aging, between 29–35d and 29–42d (Figure 4.7B, cluster #3 and Table 4.4). The remaining two k-means clusters
showed an upward trend when gene expression was compared to histone modifications, however changes in gene expression were not significant (Figure 4.7B, clusters #1 and 2 and Table 4.4). Gene clusters #1 and #2 showed increased acetylation approximately 1.5 kb upstream of the TSS, which was most apparent at 34 and 42 days, and co-localized with the H3K4me3 mark (Figure 4.8).

Peaks that lost the H3K9ac mark were also subject to k-means clustering (Figure 4.9A) and cluster-associated gene expression showed a similar downward trend for all three clusters, but again, changes in gene expression were significant for only one pairwise comparison, 29–42d, cluster #2 (Figure 4.9B, Table 4.4). In the case of this one significant interval, the mean H3K9ac counts at 30d and 42d were almost identical (Figure 4.9B, cluster #2). These data suggest that H3K9ac marks do not dynamically correspond to gene expression during leaf senescence as do the H3K4me3 marks, most strikingly for down-regulated genes. This is well illustrated by the H3K9ac profiles for the 314 K4-SURGs (Figure 4.10). K4-SURGs significantly increased expression during leaf senescence, and the H3K4me3 profiles start well below the time point average and consistently increase during leaf senescence, while the H3K9ac profiles are on a par with the time point average at 30d and 34d, and only show an increase at 42d.

GO enrichment analysis was performed for the k-means gene clusters. H3K4me3_Gain cluster #2 showed a dramatic increase in gene expression and enrichment for “multidrug transport” and “indole glucosinolate biosynthesis” was observed, while the genes in H3K4me3 gain cluster #3 with a smaller increase in expression were enriched for “immune response” and “secondary metabolic process”. For H3K4me3 loss cluster #2, which showed a dramatic decrease in expression, “response to hormone stimulus” and “plant-type cell wall modification”
were enriched while the less steep decline in expression observed in H3K4me3 loss cluster #3 was also enriched for “response to hormone stimulus” as well as “response to light stimulus” and “developmental growth involved in morphogenesis”. The remaining clusters showed no significant enriched GO biological process clusters.

4.3.5 Histone mark gene coverage correlates with gene expression during leaf senescence

A recent study by Benayoun et al. showed that the extent of H3K4me3 marks was greatest for genes that impart cell identity. In leaves, for instance, a high proportion of genes with the broadest H3K4me3 coverage were assigned the photosynthesis GO term. With this in mind, the extent to which each gene was covered by the two respective histone modifications was determined. Using 100 bp bins that exhibited a significant enrichment in ChIP-seq read counts relative to input (-log(p-value ≥ 4), we calculated coverage both in terms of absolute base pairs and as a proportion of the gene to normalize for length (TSS to TTS +/- 500bp to allow for potential spreading of the marks beyond the confines of the gene body). Considering the absolute extent of significant H3K4me3 enrichment at 42d, the top 5% of genes had at least 1500 bp of coverage by the mark (Figure 4.11A). As a proportion of the gene length, the top 5% of genes with the highest coverage had at least 50% of their lengths significantly enriched for H3K4me3 (Figure 4.11B). In contrast, genes in the 95th percentile of H3K9ac coverage at 42d showed significant ChIP-to-input enrichment over at least 900 bp (Figure 4.11C). Normalizing for gene length, the top 5% of genes with the broadest H3K9ac enrichment had at minimum 32% of their gene bodies covered (Figure 4.11D). Similar overall coverage distributions were observed for the respective marks at all leaf ages (data not shown), with H3K9ac observed to cover only 60-65% of the region covered by H3K4me3.
Enrichment for Biological Process GO terms was performed for genes with the top 5% broadest coverage for both modifications. “Photosynthesis” was enriched for H3K4me3 breadth for all time points, but was not enriched at any time for genes with the broadest H3K9ac coverage. The 57d H3K4me3 time point showed enrichment for “response to bacterium”, defense genes expressed during senescence. “Regulation of transcription” was enriched in 30d and 34d H3K9ac samples, but not enriched in any of the H3K4me3 samples, suggesting that H3K9ac coverage breadth may control regulatory loci.

To better illustrate the association between the breadth of the histone marks and the expression levels of genes involved in leaf senescence, genes were placed into different bins based on their log$_2$ fold change in expression across the four time points used for the H3K4me3 analysis and the three time points for the H3K9ac analysis. The change in the proportion of histone mark coverage was determined across the corresponding time points for each bin, and their distributions are plotted in Figure 4.11E (H3K4me3) and Figure 4.11F (H3K9ac). A positive correlation between fold change in gene expression and breadth of histone marks was apparent, with the correlation being higher for H3K4me3 marks ($r = 0.47$) than that for H3K9ac marks ($r = 0.20$). The strength of the relationship between gene expression and H3K4me3 gene coverage was further illustrated by the dendrograms for correlation matrices made for the RNA-seq data and for the H3K4me3 coverage data (Figure 4.12).
4.4 Discussion

The abundance of H3K4me3 and H3K9ac histone modifications change at specific loci during numerous developmental transitions in plants\textsuperscript{19,21,26-29,31,35-37}, and thus it is likely that the final stage of leaf development, leaf senescence, would be accompanied by changes in these two well-characterized histone modifications that positively correlate with gene expression. Leaves were harvested from progressively more senescent plants and subject to RNA-seq and ChIP-seq so that a direct correlation between gene expression and abundance and breadth of histone modifications could be performed. Our data showed a strong correlation between the breadth of both histone marks and gene expression. Differing amounts of H3K4me3 marks correlated well to gene expression, however a weaker trend was observed between gene expression and the abundance of H3K9ac marks.

Our harvest protocol maximized material to allow for both nuclei and RNA extraction, and used multiple as well as larger leaves (approximately leaves 12-15) than leaf 7 which was used in an extensive transcriptome analysis of developmental senescence\textsuperscript{2}.. Genes from cluster U1 were expressed early in both experiments, genes from clusters U2 and U3 mostly showed the greatest fold-induction in the first time interval (29-35d), while genes from the U5, U6 and U7 clusters showed highest expression at 57d, the final time point for the RNA-seq data. Although the current study did not have the temporal refinement of the extensive transcriptome work, the intervals are generally equivalent and trends should translate among studies.

The signal strength for the H3K4me3 peaks was higher than that of the H3K9ac peaks. A recent analysis of multiple chromatin marks shows weaker peaks for numerous acetylation marks compared to H3K4me3 marks, suggesting that acetylation marks are less pronounced than
H3K4me3 modifications in Arabidopsis\textsuperscript{38}. The high coincidence of H3K4me3 and H3K9ac marks (Figure 2) has been noted previously\textsuperscript{20,21,28,31,33,39} and supports the validity of the H3K9ac datasets. Although the coincidence of these two modifications has been noted before, our observation that the extent of H3K9ac coverage per gene was about 60\% of the H3K4me3 coverage level is novel.

For most genes, modifications centered +400 bp from the TSS, but for genes identified in the H3K9ac\_GAIN\_clusters #1 and #2, modifications that peaked at ~ -1200 bp from the TSS were also observed for both marks (Figure 4.8). The combination of the two marks in the promoter region might contribute to an upward trend in gene expression, however neither of these H3K9ac\_GAIN clusters exhibited significantly increased gene expression (Table 4.4). Promoter methylation for H3K4me3\_GAIN genes did increase with plant age, but no clear peak was identified and the magnitude of the change was much less compared to the increase that centered at +400 bp from the TSS. Thus, for these genes, the role of promoter histone modifications is unclear.

The breadth of histone mark coverage was found to correlate well with gene expression, especially for H3K4me3 (Figures 4.11 and 4.12). Genes with the greatest fold-change in expression had the largest corresponding changes in gene coverage by the mark. These data support the fact that the extent of gene coverage, as well as the peak intensity, of these histone marks is important for gene regulation. Since this study is based on an ensemble of cells, the increase in the total histone mark signal at active genes could be due to the fact that an increasing number of cells up-regulate the genes involved in the senescence process over time.
A similar experiment coupling expression to the H3K4me3 mark was performed on rice subject to drought stress\textsuperscript{40}. Results of this study corresponded to our work noting that genes with a significant change in the H3K4me3 mark and in gene expression (609 rice genes) were directionally correlated such that gain of the H3K4me3 mark and up-regulation occurred in 89% of cases and loss of the H3K4me3 mark and down-regulation occurred in 90% of cases. Identification of significantly changed H3K4me3 marks was less precise because only two conditions, control and drought, were studied, and thus the significant change in three of six (or two of three) pairwise comparisons that was done for this work, was not possible. The observation that of the 4,387 rice genes with differential H3K4me3 modification levels, many were not expressed (45%) or had no change in expression during drought (40%) might be due to an over-estimation of genes with differential H3K4me3 marks.

The relatively few genes that showed significant changes in H3K4me3 (786 genes) and H3K9ac (278 genes) marks is surprising. Our work showed that only a small proportion of SURGs (22%) and SDRGs (15%) had changes in the H3K4me3 mark, and an even smaller percentage (2% for SURGs and 3% for SDRGs) for the H3K9ac mark. This can be partially explained by the stringent criteria used to call regions of significant change, but raising the p-value resulted in the inclusion of genes that did not show convincing gains/losses when raw data were viewed. Additionally, many SURGs already had both histone marks prior to the first leaf harvest, however levels were below the whole-genome average for H3K4me3 and did not change significantly (Figure 4.4). For these genes, H3K9ac levels showed an increasing trend that peaked above whole-genome time point averages, indicating that small changes in this mark do accompany changes in gene activation. SDRGs that did not show a significant loss in H3K4 trimethylation over the time points studied did display a decreasing trend in H3K4me3 marks,
however they remained above the genome-wide average, and H3K9ac marks for these genes were low during the entire time course. Thus, for the bulk of up-regulated genes, low levels of H3K4me3 marks were present prior to the first time point and H3K9ac showed an increasing trend. For the bulk of down-regulated genes, the opposite was observed: low levels of H3K9ac were present while H3K4me3 levels showed a decreasing trend. This reciprocal relationship between the two marks has not been noted previously, and is worth further study.

Pathogen infiltration of Arabidopsis leaves was found to result in increased expression of WRKY70, PR1 and TH12.1. WRKY70 was found to have differential H3K4me3 and to be a direct target of the ATX1 histone methyltransferase while PR1 and TH12.1 had constant levels of H3K4me3 marks and were targets of WRKY70\(^{41}\). The pathogen infiltration study suggested that changing H3K4me3 may occur for primary-responsive regulatory genes, but not for secondary response genes. In agreement with this observation, our list of SURGs that displayed a gain of H3K4me3 (K4-SURGs) includes 4 WRKY and 6 NAC TFs. Of these WRKY and NAC genes, mutant analysis has shown that WRKY75\(^{34}\), NAC016\(^{42}\) and NAC019\(^{43}\) are positive regulators of senescence while WRKY54\(^{44}\) is a negative regulator of senescence. Thus numerous important primary regulators of senescence do display a concomitant change in H3K4me3 marks and gene expression. A small number of regulatory genes displayed changes in H3K9ac as leaf senescence progressed. K9-SURGs include WRKY41, which plays a protective role during defense\(^{45}\) and NDR1/HIN1-Like10, a MAPK gene that is SA-responsive during senescence\(^{46}\).

Of the 175 K4-SDRGs that displayed a loss of H3K4me3 marks and decreased gene expression, 13 encoded YUCCA and SAUR-like genes respectively involved in auxin biosynthesis\(^{47-49}\) and response\(^{50,51}\). Auxin synthesis sets in motion major changes in gene expression and physiology\(^{52}\) and the down-regulation of auxin synthesis and signaling can be
considered a primary event in the progression of senescence. Although reports of the role of auxin in leaf senescence conflict\textsuperscript{53-56}, our molecular data suggest auxin action is down-regulated during the progression of leaf senescence.

A well-characterized senescence regulator, \textit{WRKY53}, showed a constant high level of H3K4me3 and H3K9ac marks in our study (Figure 4.13A), thus indicating that not all important regulators were identified as K4-SURGs. RPKM levels for \textit{WRKY53} increased from 42.6 to 548.1 between 29d and 35d indicating that a large change in expression (13-fold) did occur during the time points included in this study. \textit{WRKY53} is one example of the many genes that are marked prior to significant up-regulation of mRNA levels. \textit{WRKY53} expression is down-regulated by Whirly1\textsuperscript{57}, which may explain the coincidence of low transcript levels and high levels of H3K4me3 marks. Examples of post-transcriptional regulation mediated by small RNAs have been identified during leaf senescence\textsuperscript{11,58,59}, and may also explain some of the inconsistencies between H3K4me3 marks and gene expression. \textit{SAG12}, a molecular marker for senescence was up-regulated 3300-fold between 29 and 57d (0.022–72.5 RPKM) and showed increased levels of H3K4me3 marks 250- 650 bp upstream of the TSS, but no clear H3K9ac marks (Figure 4.13B), suggesting that H3K4me3 modifications in the promoter can be associated with changes in gene expression. This differed from our previous study in which \textit{SAG12} was not marked by H3K4me3\textsuperscript{19}. The other gene that was reported to be strongly up-regulated and that did not have H3K4me3 marks, At1g73220, was devoid of both marks in the current study (Figure 4.13C), however expression levels were low for this gene, and only reached median RPKM levels at 57d. As four time points were included in this more comprehensive work, these data have more certainty than our previous data.
During leaf senescence, changes in H3K4me3 marks reflected changes in gene expression to a much greater extent than did H3K9ac marks. This is, in part, due to the slightly different ages of the H3K9ac ChIP-seq data set compared to the RNA-seq data set. However, the criteria for a significant change required changes that covered two of three intervals for H3K9ac, thus the differences in leaf harvest dates (29d vs. 30d and 35d vs. 34d) for the two marks were minimized by the nature of the analysis. A small number of genes that were specifically up- or down-marked between 29–35d, but not between the shorter interval, 30-34d, would be missed though. In drought stress, H3K9ac and H3K4me3 marks increased in parallel for three drought-responsive genes. During rehydration, expression of the drought-responsive genes was down-regulated, and the H3K9ac marks were rapidly lost while the H3K4me3 marks were retained and decreased at a rate similar to mRNA levels. The drought study was done on the time scale of hours while this developmental senescence study was performed on a time scale of days, yet in both H3K4me3 marks paralleled gene expression. It should be noted, however, that H3K4me3 marks can change rapidly, as has been observed for clock genes in which both H3K9ac and H3K4me3 marks cycle each 24 hours. Sampling was always performed at the same time each harvest day, so diurnal cycles would not interfere with our study.

In this study, the 387 K4-SURG genes were noted to have average H3K9ac levels at the earliest time points that increased by 42d, unlike the H3K4me3 marks which at first were substantially below average and greatly increased by the end of the time course (Figure 4.10). These data suggest that H3K9ac modifications are present prior to H3K4 trimethylation for the K4-SURGs. In animal systems, evidence supporting K4 trimethylation of acetylated H3 has been
published\textsuperscript{60-62}, and it is possible that the H3K9 acetylation serves as a template for the gain of H3K4me3 marks during leaf senescence.

Interestingly, for K4\_SURGs which display the largest gain in H3K4me3 marks between the first two time points (29-35d), additional H3K9ac acquisition, above average levels, only occurs at a subsequent stage between 34 and 42d (Figure 4.10). Furthermore, those genes that gain H3K9ac (K9\_SURGs) already show elevated levels of H3K4me3 and show no further increase in trimethylation (Figure 4.8), indicating that H3K4me3 acquisition occurred sometime prior to that of H3K9ac for these genes. Non-K4-SU/DRGs showed a reciprocal relationship with one mark remaining at a basal low level while the other showed an increasing/decreasing trend (Figure 4.4).
Materials and methods

Plant growth and nuclei isolation

Plants were grown as described\textsuperscript{19} with the exception that the light intensity was 30 µmols photons m\textsuperscript{-2} sec\textsuperscript{-1}, at 22°C, and the diurnal cycle was 20 hours of light and 4 hours of darkness. The low light intensity was chosen because the leaves of older plants grown at standard light intensities (120-150 µmols photons m\textsuperscript{-2} sec\textsuperscript{-1} for 16h is recommended by the Arabidopsis Biological Resource Center) became purple and stressed; to compensate for the low light intensity, a longer light period, 20h, was utilized. Mature fully-expanded rosette leaves were harvested. Newly emerging leaves at the center of the rosette were not harvested and petioles were trimmed such that harvested tissue contained leaf blades, only. No inflorescence tissue was harvested. The first harvest was performed after the vegetative to reproductive transition when incipient bolts (0.5 to 3 cm) had developed (29 day-old plants, 29d). Mature rosette leaf tissue was then harvested at 35d when bolts had lengthened and secondary bolts were forming, at 42d when siliques were present on primary and secondary bolts and at 57d when about 50% of siliques were brown and dry and chlorophyll loss was evident throughout the rosette leaves.

RNA-seq and ChIP-seq library preparation

Isolation of nuclei and chromatin immunoprecipitation were performed as previously described with the exception that the final pellet was resuspended in 1 ml cold nuclei lysis buffer and 2 ml of cold ChIP dilution buffer. The H3K9ac antibody (07-352) and H3K4me3 antibody (17-678) were purchased from Millipore. The H3K4me3 ChIP library was prepared for sequencing as described previously\textsuperscript{19} and the H3K9ac library was prepared using the Illumina TruSeq\textsuperscript{TM} ChIP Sample Preparation Kit – Set A.
RNA was isolated using Trizol reagent (Life Technologies) and RNA was prepared for sequencing using the Illumina TruSeq™ RNA Sample Prep Kit v2-SetA.

Sequencing was performed at UCLA using an Illumina HiSeq2000 to produce 50bp single reads from both the ChIP-seq and RNA-seq libraries.

RNA-seq data processing

RNA-seq reads were mapped to the TAIR10 reference genome and gene annotation using tophat-2.0.4calling bowtie-0.12.8 using the following parameters:

Tophat --solexa1.3-quals --no-coverage-search -g 1 –G <TAIR10 annotation>

HTSeq was used to perform gene counts and DESeq2 for differential analysis. Intersection and union analyses of gene lists were performed using the Virtual Plant web site.

ChIP-seq data processing

ChIP-seq reads were mapped to the TAIR10 reference genome using bowtie-0.12.8 using the following parameters:

bowtie -m 1 -v 2 -a --best --strata

Regions of significant H3K4me3 and H3K9ac ChIP enrichment were determined by comparing the number of ChIP to input reads within 100bp bins across the genome. Bins showing –log(p-value) ≥ 6 by the Poisson test were deemed to be significantly enriched. To ensure an even higher level of confidence in these peaks, we required that at least two sequential 100bp bins showed significant ChIP enrichment when comparing the locations of the two marks (Figure 4.2). Pearson correlations and significance values were calculated using R's 'cor.test' function.
**ChIP-seq differential analysis**

Pairwise comparisons were made between ChIP signals at each time point for each histone mark, using a Poisson test to determine regions of significant difference (\(-\log(p\text{-value}) \geq 6\)). Based on these pairwise comparisons, a region was defined as displaying a consistent gain or loss of histone marks during the course of the experiment, if changes were observed in at least three of the six pairwise comparisons for H3K4me3 and two of the three pairwise comparisons for H3K9ac.

**K-means clustering**

For each mark, only regions displaying a consistent significant gain or loss of the mark during the course of the experiment were considered (see ChIP-seq differential analysis section above). In each of the four cases (significant K4me3 gain, K4me loss, K9ac gain, K9ac loss), \(-\log(p\text{-value})\) results from the ChIP vs. input Poisson test (see above) were grouped by k-means clustering (k=3) using the ‘kmeans’ function in the R package\(^\text{68}\). Genes were assigned to these peaks, if the peaks overlapped the TSS to TSS + 500bp region of the genes. In some cases, multiple regions from a single gene were assigned to different clusters.

**Gene coverage**

For each mark, the coverage of each gene was determined as the proportion of 100bp bins from TSS - 500bp to TSS + 500bp that exhibited a significant enrichment in ChIP-seq read counts relative to input (\(-\log(p\text{-value}) \geq 4\)). A lower threshold was chosen in this case as we were comparing IP to input as opposed to IP vs IP, which has the potential of containing more false positives, and because this helped to better capture the continuity in the breadth of coverage of
significant histone mark deposition along the gene body. The top 5% of genes with the highest proportion of coverage at each time point were analyzed for absolute length of coverage and percent gene coverage.

Data submission

All high-throughput sequencing data are available at GEO under GSE67778.
Figures

Figure 4.1    Gene expression differences during leaf senescence

A. Pearson correlation matrix of gene expression data (log2(read counts + 1)) from all RNA-seq libraries. Darker red indicates a higher correlation. Dendrograms generated by hierarchically clustering samples based on correlation values transformed into distance values (1-r).

B. Genes with significant (≥ 2-fold, p-value ≤ 0.05) increases in expression between adjacent ages are shown in the Venn diagram.

C. Genes with significant (≥ 2-fold, p-value ≤ 0.05) decreases in expression between adjacent ages are shown in the Venn diagram.
D. Box plots for SURGs. RNA-seq RPKM data (log₂ scale) for genes are shown. Boxes represent first to third quartiles, or interquartile range (IQR). Whiskers extend to the most extreme data points, but no more than +/-1.5 times the IQR from the box, beyond which outliers are plotted individually. Notches extend to +/-1.58 IQR/√n ⁶⁹.

E. As in (D), but for SDRGs.

F. As in (D), but for non-SURG or SDRG gene expression.

G. The abundance of three gene families for different thresholds used to classify senescence up-regulated genes (SURGs) and senescence down-regulated genes (SDRGs) is shown.
Figure 4.2  Coincidence of H3K4me3 and H3K9ac marks

A. ChIP-seq reads from two gene-rich regions of the Arabidopsis genome, the top from chromosome 1 and the bottom from chromosome 5, are displayed using Integrated Genome Viewer (IGV). Data from 29d (H3K4me3) and 30d (H3K9ac) are displayed in pairs of ChIP and
input reads. The scale for H3K4me3 ChIP and input is 0 to 120 while the scale for H3K9ac ChIP and input is 0 to 60. Gene tracks with exons and introns are shown in blue below the read tracks.

**B.** At5g45340 (the boxed gene in panel A) is shown zoomed-in. The peaks for both histone modifications are co-localized to the 5’ ends of the gene. At5g45340 is transcribed from right to left in the figure and encodes an ABA hydroxylase.

**C.** ChIP reads were compared to input and peaks were called as described in the Methods section for each histone modification. The Venn diagram shows the overlap between peaks comprising at least two sequential 100 bp bins exhibiting significant ChIP counts for each histone modification across all time points.
Figure 4.3  Histone modifications in leaves from differently aged plants

A. IGV tracks for ChIP and input reads for the WRKY75 gene (At5g13080) which shows a significant increase in H3K4me3 marks (K4_GAIN 3 of 6 track). The read count tracks show ChIP and input reads for the numbers of days of plant growth for each histone modification as indicated. Data from younger leaves are darker green while data from older leaves are more yellow. The read range is 0–120 for H3K4me3 reads, 0–60 for H3K9ac reads. The criteria for significance are $-\log(p\text{-value}) \geq 6$ in three of six pairwise comparisons, and p-values for the six pairwise comparisons are shown below (K4_GAIN 29-35 to K4_GAIN 42-57, p-value scale is 0 to 10).
B. IGV tracks as in A, but for one member of the MATE Efflux Antiporter family (At1g66760) which shows a significant increase in H3K9ac marks. Significant changes in acetylation were identified as those that showed $-\log(p\text{-value}) \geq 6$ in two of three comparisons (K9_GAIN 2 of 3 track). K9_GAIN 30-34 to K9_GAIN 34-42 tracks show $-\log(p\text{-values})$ for the three pairwise comparisons on a 0–10 scale.
Figure 4.4  H3K4me3 and H3K9ac histone marks in SURGs and SDRGs not identified as K4-SURGs or K4-SDRGs
**A.** Histone modification profiles for 1117 SURGs are shown as solid lines for H3K4me3 excluding the 315 K4-SURGs that had significant and consistent changes in H3K4me3 marks. The average counts per 100 bp bins for all genes at each time point are shown by the dashed lines.

**B.** As in (A), but for H3K9ac.

**C.** The gene expression values are shown for the group of 1117 SURGs below the histone modification profiles. All pairwise gene expression differences are significant: 1.33e-62*** for 29-35d, 2.54e-106*** for 29-42d, 2.53e-134*** for 29-57d, 2.92e-04*** for 35-42d, 2.64e-09*** for 35-57d and 0.009** for 42-57d. Box plot representations are described in the legend for Figure 4.1E.

**D.** Histone profiles for the 825 SDRGs shown as solid lines for D) H3K4me3 excluding the 139 K4-SDRGs.

**E.** As in (D), but for H3K9ac.

**F.** The gene expression values are shown for the group of 825 SRDGs below the histone modification profiles. The p-values for pairwise comparisons of gene expression are: 5.49e-57*** for 29-35d, 3.47e-79*** for 29-42d, 3.99e-89*** for 29-57d, 3.81e-07*** for 35-42d, 2.46e-10*** for 35-57d and 0.34 for 42-57d. Box plot representations are described in the legend for Figure 4.1E.
Figure 4.5  K-means clustering for H3K4me3 gain peaks and correlation to gene expression

A. Peaks with a significant gain in H3K4me3 marks were subjected to k-means clustering to generate three cluster groups with different temporal trends of acquisition.
**B.** The mean counts per 100 bp bin were calculated for genes associated with peaks in each cluster (solid lines), as well as for all genes at each time point (dashed lines), for clusters #1, #2 and #3. Data are shown for the region comprising -2500 to +2500 bp in relation to the Transcription Start Site (TSS). RNA-seq data for genes that coincided with each peak are shown in the box plots for each cluster. Box plot representations are described in the legend for Figure 4.1E.
Figure 4.6  K-means clustering for H3K4me3 loss peaks and correlation to gene expression

A. As in Figure 4.5A, but for regions exhibiting a significant loss of H3K4me3 marks and gene expression data are displayed as described in Figure 4.5B.

B. As in (A), but for H3K4me3.
Figure 4.7  K-means clustering for H3K9ac gain peaks and correlation to gene expression

A. As in Figure 4.5A, but for regions exhibiting a significant gain of H3K9ac marks.

B. H3K9ac marks and gene expression data are displayed as described in Figure 4.5B.
Figure 4.8 Average H3K4me3 and H3K9ac read count profiles for genes associated with H3K9ac_GAIN Clusters #1 and #2

Histone modification profiles as described in Figure 4.5B but for two clusters of genes associated with H3K9ac acquisition. H3K4me3 and H3K9ac marks were found to coincide within the promoter regions of two of the three H3K9ac_GAIN clusters identified in Figure 4.7.
Figure 4.9  K-means clustering for H3K9ac loss peaks and correlation to gene expression

A. As in Figure 4.5A, but for regions exhibiting a consistent loss of H3K9ac.

B. H3K9ac marks and gene expression data are displayed as described in Figure 4.5B.
Figure 4.10  Average H3K4me3 and H3K9ac read count profiles for the 387 K4-SURGs

Histone modification profiles as described in Figure 4.5B for the 315 K4-SURGs. Both histone marks are shown for K4-SURGs. H3K4me3 marks increase consistently throughout leaf aging, starting at a low level while H3K9ac marks start at the whole-genome time point average and only increase at 42d.
Figure 4.11  Breadth of histone modifications during leaf senescence
A. The breadth of H3K4me3 modifications for 42d samples are shown in terms of coverage of gene lengths in basepairs (bp). Red dashed vertical line delineate the top 95\textsuperscript{th} percentile.

B. As in (A), but coverage given in terms of percentage of gene length.

C. As in (A), but for H3K9ac.

D. As in (C), but coverage given in terms of percentage of gene length.

E. The percent change in H3K4me3 gene coverage from 29d to 57d is plotted for gene expression bins. Genes are placed in bins according to the fold change in gene expression over the same time interval. Gene counts per bin are given in parentheses. For example, the bin labeled 0.5 includes 2712 genes that display a log\textsubscript{2} increase in gene expression greater than 0.5 to less than or equal to 1.0. The box plot representations are as described for Figure 4.1E. The overall Pearson correlation between percent change in gene coverage and fold-change in gene expression was 0.47 (top left corner) with a 95\% confidence interval of [0.463; 0.481] (p-value < 2.2e-16).

F. As in (E), but for H3K9ac gene coverage from 30d to 42d. The overall Pearson correlation in this case was 0.20 with a 95\% confidence interval of [0.185; 0.207] (p-value < 2.2e-16).
Figure 4.12  Correlation matrices and dendrograms for RNA-seq RPKMs and H3K4me3 gene coverage

Dendrograms represent the hierarchical clustering of Spearman rank correlation values transformed into distance measures \((1-\rho)\) for pooled RNA-seq samples and the percent gene coverage by H3K4me3. The Spearman correlation distance scale is shown to the left of each dendrogram. Corresponding correlation matrices are shown below the dendrograms.
Figure 4.13  H3K4me3 and H3K9ac marks for three genes

A. *WRKY53* (At4g23810) encodes a positive regulator of leaf senescence and has high levels of both histone modifications at all of the time points in this study.

B. *SAG12* (At5g45890) is a standard marker of leaf senescence and shows an increase in H3K4me3 marks in the promoter region upstream of the TSS.

C. At1g73220 encoding a putative transporter is up-regulated to median RPKM levels only at the last time point, but not marked by H3K4me3 or H3K9ac at any time point.
### Sequencing and alignment summary for RNA-seq libraries

This table summarizes the RNA-seq replicate samples for all time points.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Total reads</th>
<th>Aligned reads</th>
<th>% Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>29d-1</td>
<td>50,585,318</td>
<td>47,809,911</td>
<td>95%</td>
</tr>
<tr>
<td>29d-2</td>
<td>56,317,560</td>
<td>52,160,493</td>
<td>93%</td>
</tr>
<tr>
<td>29d-3</td>
<td>74,709,300</td>
<td>70,289,777</td>
<td>94%</td>
</tr>
<tr>
<td>35d-1</td>
<td>61,640,834</td>
<td>56,589,865</td>
<td>92%</td>
</tr>
<tr>
<td>35d-2</td>
<td>59,127,761</td>
<td>56,810,411</td>
<td>96%</td>
</tr>
<tr>
<td>35d-3</td>
<td>74,084,899</td>
<td>69,345,945</td>
<td>94%</td>
</tr>
<tr>
<td>42d-1</td>
<td>49,651,744</td>
<td>47,583,338</td>
<td>96%</td>
</tr>
<tr>
<td>42d-2</td>
<td>64,098,111</td>
<td>59,183,060</td>
<td>92%</td>
</tr>
<tr>
<td>42d-3</td>
<td>65,807,180</td>
<td>60,144,918</td>
<td>91%</td>
</tr>
<tr>
<td>57d-1</td>
<td>68,197,891</td>
<td>61,300,159</td>
<td>90%</td>
</tr>
<tr>
<td>57d-2</td>
<td>59,962,671</td>
<td>57,633,192</td>
<td>96%</td>
</tr>
<tr>
<td>57d-3</td>
<td>55,578,330</td>
<td>53,576,333</td>
<td>96%</td>
</tr>
<tr>
<td>SURG GO Biological Process</td>
<td>Gene Count</td>
<td>FDR (p-value)</td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------</td>
<td>------------</td>
<td>---------------</td>
<td></td>
</tr>
<tr>
<td>Defense response</td>
<td>146</td>
<td>1.5E-20</td>
<td></td>
</tr>
<tr>
<td>Protein amino acid phosphorylation</td>
<td>147</td>
<td>1.5E-19</td>
<td></td>
</tr>
<tr>
<td>Response to organic substance</td>
<td>128</td>
<td>2.3E-08</td>
<td></td>
</tr>
<tr>
<td>Glucosinolate metabolic process</td>
<td>19</td>
<td>2.8E-08</td>
<td></td>
</tr>
<tr>
<td>Jasmonic acid metabolic process</td>
<td>12</td>
<td>1.7E-04</td>
<td></td>
</tr>
<tr>
<td>Indole derivative metabolic process</td>
<td>14</td>
<td>8.8E-04</td>
<td></td>
</tr>
<tr>
<td>Multidrug transport</td>
<td>17</td>
<td>0.010</td>
<td></td>
</tr>
<tr>
<td>Aromatic compound biosynthetic process</td>
<td>28</td>
<td>0.307</td>
<td></td>
</tr>
<tr>
<td>Response to oxidative stress</td>
<td>33</td>
<td>0.741</td>
<td></td>
</tr>
<tr>
<td>Chitin catabolic process</td>
<td>8</td>
<td>0.767</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SDRG GO Biological Process</th>
<th>Gene Count</th>
<th>FDR (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response to auxin stimulus</td>
<td>49</td>
<td>3.8E-12</td>
</tr>
<tr>
<td>Cell wall organization</td>
<td>34</td>
<td>1.1E-05</td>
</tr>
<tr>
<td>Response to light stimulus</td>
<td>43</td>
<td>3.5E-04</td>
</tr>
<tr>
<td>Regulation of cellular component size</td>
<td>24</td>
<td>0.068</td>
</tr>
<tr>
<td>Lipid biosynthetic process</td>
<td>34</td>
<td>0.093</td>
</tr>
<tr>
<td>Photosynthesis</td>
<td>19</td>
<td>0.207</td>
</tr>
<tr>
<td>Microtubule-based movement</td>
<td>12</td>
<td>0.217</td>
</tr>
<tr>
<td>Response to gibberellin</td>
<td>15</td>
<td>0.280</td>
</tr>
</tbody>
</table>

Table 4.2  Gene Ontology enrichment for Senescence Up-Regulated Genes (SURGs) and Senescence Down-Regulated Genes (SDRGs)

Enriched GO Biological Processes and p values for false discovery rate (FDR) are reported.
<table>
<thead>
<tr>
<th>H3K4me3</th>
<th>Total Reads</th>
<th>Aligned reads</th>
<th>% Aligned</th>
<th>Unique Reads</th>
<th>% Unique</th>
</tr>
</thead>
<tbody>
<tr>
<td>29d_ChIP</td>
<td>72,651,120</td>
<td>56,515,347</td>
<td>78%</td>
<td>12,471,170</td>
<td>22%</td>
</tr>
<tr>
<td>29d_input</td>
<td>241,859,534</td>
<td>129,130,261</td>
<td>60%</td>
<td>68,163,423</td>
<td>53%</td>
</tr>
<tr>
<td>35d_ChIP</td>
<td>94,549,362</td>
<td>65,914,502</td>
<td>70%</td>
<td>15,235,440</td>
<td>23%</td>
</tr>
<tr>
<td>35d_input</td>
<td>256,832,423</td>
<td>136,813,852</td>
<td>59%</td>
<td>67,497,500</td>
<td>49%</td>
</tr>
<tr>
<td>42d_ChIP</td>
<td>79,935,864</td>
<td>54,513,086</td>
<td>69%</td>
<td>13,123,594</td>
<td>24%</td>
</tr>
<tr>
<td>42d_input</td>
<td>219,636,473</td>
<td>127,612,919</td>
<td>62%</td>
<td>66,732,566</td>
<td>52%</td>
</tr>
<tr>
<td>57d_ChIP</td>
<td>150,759,372</td>
<td>83,983,773</td>
<td>56%</td>
<td>19,116,262</td>
<td>23%</td>
</tr>
<tr>
<td>57d_input</td>
<td>232,064,371</td>
<td>131,234,324</td>
<td>62%</td>
<td>66,787,863</td>
<td>51%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>H3K9ac</th>
<th>Total Reads</th>
<th>Aligned reads</th>
<th>% Aligned</th>
<th>Unique Reads</th>
<th>% Unique</th>
</tr>
</thead>
<tbody>
<tr>
<td>30d_ChIP</td>
<td>72,033,414</td>
<td>52,750,857</td>
<td>80%</td>
<td>11,019,515</td>
<td>21%</td>
</tr>
<tr>
<td>30d_input</td>
<td>88,614,150</td>
<td>60,738,145</td>
<td>78%</td>
<td>16,030,599</td>
<td>26%</td>
</tr>
<tr>
<td>34d_ChIP</td>
<td>72,232,723</td>
<td>54,470,143</td>
<td>81%</td>
<td>10,605,940</td>
<td>19%</td>
</tr>
<tr>
<td>34d_input</td>
<td>92,834,798</td>
<td>61,939,766</td>
<td>77%</td>
<td>26,812,313</td>
<td>43%</td>
</tr>
<tr>
<td>42d_ChIP</td>
<td>81,732,114</td>
<td>62,777,923</td>
<td>82%</td>
<td>6,623,315</td>
<td>11%</td>
</tr>
<tr>
<td>42d_input</td>
<td>88,141,553</td>
<td>61,854,129</td>
<td>78%</td>
<td>6,449,022</td>
<td>10%</td>
</tr>
</tbody>
</table>

Table 4.3  Sequencing and alignment summary for ChIP-seq libraries

This table summarizes the ChIP-seq read output for both H3K4me3 and H3K9ac at all time points.
### Table 4.4  Significance of pairwise comparisons in gene expression (p-values)

Pairwise comparisons are listed at the top of each table. Accompanying H3K4me3 GAIN cluster data are shown in Figure 4.5 and LOSS cluster data are shown in Figure 4.6. Accompanying H3K9ac GAIN cluster data are shown in Figure 4.7 and LOSS cluster data are shown in Figure 4.9. * p ≤ 0.05, ** p ≤ 0.01, *** p ≤ 0.001
References


<table>
<thead>
<tr>
<th></th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>Benayoun, B. <em>et al.</em> H3K4me3 breadth is linked to cell identity and transcriptional consistency. <em>Cell</em> 158, 673-688 (2014).</td>
</tr>
</tbody>
</table>


Chapter 5

A mechanistic link between gene regulation and genome architecture in mammalian development
A mechanistic link between gene regulation and genome architecture in mammalian development
Giancarlo Bonora, Kathrin Plath and Matthew Denholtz

The organization of chromatin within the nucleus and the regulation of transcription are tightly linked. Recently, mechanisms underlying this relationship have been uncovered. By defining the organizational hierarchy of the genome, determining changes in chromatin organization associated with changes in cell identity, and describing chromatin organization within the context of linear genomic features (such as chromatin modifications and transcription factor binding) and architectural proteins (including Cohesin, CTCF, and Mediator), a new paradigm in genome biology was established wherein genomes are organized around gene regulatory factors that govern cell identity. As such, chromatin organization plays a central role in establishing and maintaining cell state during development, with gene regulation and genome organization being mutually dependent effectors of cell identity.

Introduction
Gene regulatory processes that govern the establishment and maintenance of cell identity during development occur within the three-dimensional (3D) space of the nucleus. Following the pioneering work of Job Dekker and colleagues in 2002 [1], elucidation of 3D chromosome folding has been greatly spurred by an expanding suite of chromosome conformation capture (3C)-based techniques, including those leveraging the power of high-throughput sequencing [2,3] (summarized in Table 1). These methods jointly rely on cross-linking of spatially juxtaposed chromatin, fragmentation of cross-linked chromatin with restriction endonucleases or sonication, ligation of proximal DNA fragments, and amplification of ligation pairs via PCR, with or without sequencing, allowing for the identification of physically interacting chromatin fragments, with more frequently interacting fragments showing a higher prevalence in the resulting PCR-amplified libraries.

The recent explosion of 3C-based genome organization studies, in combination with widespread mapping of linear genomic features (such as transcription factor binding sites, chromatin modifications, and transcription) in cell types of varying developmental stages and across numerous species, has made it clear that genome organization is an important and dynamic contributor to nuclear processes [2–8]. In particular, the discovery of various cell type-specific and cell type-invariant organizational features of the mammalian genome and their correlation with transcriptional regulators has offered insights into causal relationships between chromatin organization and gene regulation that we will discuss in this review. Briefly, at the largest scale, these findings include the spatial segmentation of the nucleus into open, transcriptionally permissive and closed, transcriptionally inert compartments [9]. Developmentally regulated switches of chromatin segments from the open to the closed compartments allow for the sequestration of transcriptionally repressed developmental genes at the nuclear lamina, ensuring their stable silencing [10,11]. Cell type-specific long-range interactions between distal genomic regions may megabases (Mb) away on the same chromosome (in cis), or on different chromosomes (in trans), have been identified and occur between genomic regions residing in the same compartment (open or closed) [12*,13*]. Genomic regions interacting over long distances often exhibit enrichment for common gene regulatory factors, such as chromatin regulators or transcription factors [13*,14,15*,16], and appear to occur between megabase-scale self-associating genomic regions termed topologically associating domains (TADs) [17–19]. Notably, although their long-range interactions can be developmentally regulated, the linear position of TADs have been argued to be largely cell type-invariant and evolutionarily conserved [17,18], and function to restrict the distance over which enhancer–promoter interactions can occur [20]. Within TADs, however, enhancer–promoter interactions can change in scope, relevance, and dynamics. Finally, recent work has demonstrated that various architectural proteins, including Cohesin, CTCF, and the Mediator complex, are important for the establishment and maintenance of a variety of cell type-specific and -invariant
Table 1

<table>
<thead>
<tr>
<th>Method</th>
<th>Acronym</th>
<th>Range</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromosome conformation capture</td>
<td>3C</td>
<td>One-to-few</td>
<td>The first step of 3C-based methods is to covalently cross-link spatially adjacent chromatin segments. Restriction endonuclease digestion and ligation of cross-linked chromatin produces chimeric DNA fragments. PCR primer pairs are designed to amplify chimeric DNA fragments consisting of hypothesized interacting regions. As such, this method requires a priori hypotheses about potential interacting chromatin fragments within a population of cells [1].</td>
</tr>
<tr>
<td>Circular chromosome conformation capture</td>
<td>4C</td>
<td>One-to-all</td>
<td>Captures the genome-wide interaction profile ('interactome') of a single locus (‘bait’ or ‘viewpoint’). Following 3C library production, a second round of restriction endonuclease digestion and ligation results in circularized, chimeric DNA products. Inverse PCR primers based on the selected bait fragment are designed to amplify intervening interacting sequences, obviating the need to hypothesize interaction regions [57–59]. An ensemble version of 3C that produces a matrix of interaction frequencies ('contact map') within specified regions of interest, by tiling high-throughput-sequencing amenable PCR primer pairs across a number of given regions, allowing for the identification of interactions between any two primer pairs [60].</td>
</tr>
<tr>
<td>Chromosome conformation capture carbon copy</td>
<td>5C</td>
<td>Many-to-many</td>
<td>An ensemble version of 3C that produces a matrix of interaction frequencies ('contact map') within specified regions of interest, by tiling high-throughput-sequencing amenable PCR primer pairs across a number of given regions, allowing for the identification of interactions between any two primer pairs [60].</td>
</tr>
<tr>
<td>Genome-wide chromosome conformation capture</td>
<td>Hi-C</td>
<td>All-to-all</td>
<td>Allows interactions between any two genomic regions to be interrogated simultaneously to produce genome-wide contact maps. Biotinylated nucleotides are incorporated into ligation junctions during 3C library production. Ligated chromatin is then sonicated and isolated with streptavidin beads for identification of interacting fragments via paired-end sequencing [9,61].</td>
</tr>
<tr>
<td>Tethered genome-wide chromosome conformation capture</td>
<td>TCC</td>
<td>All-to-all</td>
<td>A Hi-C variant wherein proteins are biotinylated in the initial cross-linked complex and tethered to streptavidin-coated beads. Subsequent Hi-C library generation steps can therefore be performed on immobilized chromatin fragments reducing the possibility of spurious ligation between free-floating chromatin fragments [62].</td>
</tr>
<tr>
<td>Chromatin interaction analysis by paired-end tag sequencing</td>
<td>ChIA-PET</td>
<td>All-to-all interactions of chromatin fragments that are associated with a protein of interest</td>
<td>A Hi-C variant incorporating a chromatin immunoprecipitation (ChIP) step to capture only interactions between chromatin fragments associated with a protein of interest [63].</td>
</tr>
</tbody>
</table>

gene organizational features, including enhancer–promoter contacts and long-range inter-TAD chromatin contacts [14,21,22], as well as TAD boundaries [17,23,24].

In this review, we focus on the latest findings from 3C-based studies conducted in mouse and human cells that have begun to establish causal links between gene regulation and nuclear architecture, and have demonstrated the importance of this coupling to mammalian development. We will pay particular attention to the mounting evidence for the role of developmentally regulated linear chromatin features in organizing the genome in 3D. Importantly, these recent findings suggest that chromatin organization contributes to the maintenance and establishment of cell identity in differentiation and reprogramming processes, making the identification of mechanistic links between chromatin organization and the linear genomic features that determine cell type a vitally important task for future work.

The segregated nucleus: compartmentalization of nuclear function

The mammalian genome is highly organized within the nucleus. Microscopy-based approaches demonstrated that each chromosome resides within a discrete volume of space known as a chromosome territory (CT), with individual CTs exhibiting minimal overlap [25–27]. More recently, 3C-based methods have demonstrated a further spatial segregation of the genome between transcriptionally permissive, euchromatic regions, and transcriptionally inert regions enriched for features of constitutive heterochromatin and nuclear lamina association, defined as the A and B, or open and closed compartments, respectively [9]. Chromatin segments residing in specific compartments can interact with each other, and typically eschew interactions with segments in the alternative compartment [4,9]. 3C-based approaches have also identified self-associating chromatin domains of approximately 1 Mb in size, termed topologically associating domains (TADs), that appear to be very stable across cell types and species, and are composed of complex networks of enhancer–promoter interactions that are restricted by the domains’ boundaries [17,18]. These TADs appear to be the fundamental modular unit of chromatin organization.

Thus, the genome is structured in a hierarchical manner with promoter–enhancer interactions occurring within TADs, chromosomes being subdivided into many TADs,
and co-localization between TADs composed of similarly transcriptionally permissive or inert chromatin, respectively, in cis and in trans, leading to the establishment of A and B compartments, and, at the highest level, chromosomes residing in discrete, minimally overlapping CTs. This organizational hierarchy is conserved across mammalian species and Drosophila [17,19], which attests to its importance in nuclear biology. Although the necessary and sufficient components of mammalian TAD boundaries are yet to be identified, highly expressed genes are enriched at these boundaries [17]. Notably, this finding is echoed even in prokaryotes, where the insertion of a highly expressed gene into the Caulobacter crescentus genome was sufficient to demarcate a TAD-analogous ‘chromosomal interaction domain’ despite the absence of a nucleosome-based chromatin structure [28**].

As described above, TADs look to be the fundamental building blocks of high-order chromosome organization. However, the position of a given TAD within the 3D space of the nucleus with respect to other TADs, or nuclear structures such as the transcriptionally repressive nuclear lamina, can change during development, supporting a role for TAD localization in cell type specification. Mirroring and expanding microscopy- and genomics-based findings that demonstrated a sequestration of lineage-specific loci to the transcriptionally repressive nuclear lamina [10,29,30], Lin et al. mapped global chromatin organization during differentiation of pre-pro-B cells to the pro-B stage. Various genes associated with the nuclear lamina in pre-pro-B cells relocate away from the nuclear periphery to the center of the nucleus, switching from the B to the A compartment, concurrent with differentiation to pro-B cells [11]. Similarly, during the course of mammalian X-chromosome inactivation in early embryonic development, entire TADs on the X-chromosome relocate to the nuclear lamina [18]. These reports suggest that TAD sequestration to the nuclear lamina-associated B compartment is an important genome organization-based mechanism for the establishment or maintenance of lineage restricted gene expression during development [11,18]. The developmentally regulated switch of TADs between the active and inactive compartments is an extreme example of the modular nature of TAD localization. Across cell types, long-range interactions between TADs (inter-TAD interactions), in both cis and trans, also change within the A and B compartments, respectively [11,12*,13*].

**Long-distance relationships: cell type-specific inter-TAD interactions point to a role for gene regulatory factors in higher order genome organization**

Several recent 4C-based studies interrogated changes in genome organization upon differentiation of embryonic stem cells (ESCs) and during reprogramming of somatic cells to induced pluripotent stem cells (iPSCs), mediated by the expression of the Yamanaka reprogramming factors Oct4, Sox2, Klf4, and cMyc [31]. These reports revealed a large-scale re-organization of long-range, inter-TAD chromatin contacts of pluripotency loci including the Nanog [14,15*], Dppa2/4 [13*,32], Oct4 [13*,22], and Sox2 [15*] genes during differentiation, and demonstrated that the ESC-specific organization of the genome is re-established upon reprogramming to iPSCs [13*,14,15*]. This pluripotency-specific organization of the mammalian genome suggests a role for pluripotency-associated gene regulatory networks in the organization of long-range chromatin contacts in ESCs and iPSCs. In support of this idea, genomic regions bound by the master pluripotency transcription factors Oct4, Sox2, and Nanog were found to interact with each other over large distances in the ESC nucleus [13*,14,15*,21,22] (Figure 1).

Similarly, extended genomic regions enriched for binding by the transcriptionally repressive Polycomb repressive complex 2 (PRC2), which mediates methylation of histone H3 at lysine 27, also co-localize in ESCs, albeit separately from the pluripotency transcription factors [13*] (Figure 1). Both pluripotency factor and Polycomb-enriched genomic region interactions occur within the context of the A compartment in pluripotent cells [13*]. Specific gene regulatory network-based inter-TAD interactions have also been described within the transcriptionally repressive B compartment in mouse olfactory neurons, within which monogenic olfactory receptor (OR) expression is ensured in part through the formation of OR-exclusive heterochromatic foci formed by aggregation of OR clusters from multiple chromosomes [12*]. Together these results argue for a cell type-specific segregation of genomic compartments based on transcriptionally permissive or inert chromatin, within which specific inter-TAD interactions form between distal regions enriched for similar transcriptional networks (regulators). This in turn begs the question of whether these transcriptional regulators are critical for the formation of these long-range chromatin interactions.

Testing the model wherein particular transcriptional networks drive specific inter-TAD interactions, we found that disruption of the Polycomb/H3K27me3 network by genetic ablation of Eed, a core subunit of PRC2, specifically abolished contacts between genomic regions highly enriched for Polycomb proteins and H3K27me3 in wild-type cells, while not effecting overall chromosome conformation [13*] (Figure 1). Notably, it was previously shown that the TAD structure within the X chromosomes inactivation center is not affected by the Eed knockout [18], indicating that different regulatory mechanisms function at different scales of genome organization. The demonstration of Polycomb-dependent chromatin co-localization in mammalian cells echoes findings in Drosophila [8,33], suggesting an evolutionarily conserved
mechanism of Polycomb-mediated gene silencing and genome organization [8].

Supporting a causative relationship between cell type-specific gene regulatory networks and genome organization, loss of Klf4 [22], Nanog [14,15*], or Oct4 [15*] disrupted pluripotency-specific long-range chromatin contacts in pluripotent cells (Figure 1). Furthermore, ectopic recruitment of Nanog to chromatin was sufficient to induce chromatin interactions between the targeted locus and other Nanog-bound regions [15*]. Although these functional studies have made it clear that gene-regulatory factors play causal roles in the establishment and maintenance of chromatin organization, in future studies it will be important to discern between the direct effects of these factors on genome organization and secondary effects due to changes in transcription or chromatin environment upon loss or gain of these factors.

Reprogramming of somatic cells to pluripotency is a useful tool for defining the temporal relationship between the establishment of pluripotency-specific genome organization, pluripotency factor binding, and pluripotency-specific transcription. Analysis of pre-iPSCs, which represent a late reprogramming intermediate, showed that pluripotency-specific long-range chromatin interactions are not yet established for pluripotency genes, especially not for those that remain inactive and unbound by pluripotency transcription factors in this late intermediate stage, such as Dppa2 and Zfp42 [13*,32]. Another line of experimentation found that pluripotency factor binding at pluripotency genes early during reprogramming is insufficient for induction of gene
expression in the absence of intra-chromosomal loops to
bring their enhancer and promoters into close proximity
[34]. Interestingly, genomic regions that interact with the
Nanog locus in reprogramming intermediates are
enriched for the open chromatin mark H3K4me3 and
bound by the reprogramming factor Klf4, but, only about
half of all genes associated with newly formed 3D-con-
tacts show an increase in expression, either in the inter-
mediate or subsequent fully reprogrammed cells [14].
Surprisingly, Nanog, itself is not up-regulated in a repro-
gramming intermediate despite its promoter being
looped towards an enhancer already enriched for binding
by reprogramming factors at this stage [14]. Together,
these data show that regulatory factor binding and the
establishment of distal chromatin interactions correlate
with the re-establishment of pluripotency and expres-
sion. However, the data also argue that neither binding
by key pluripotency factors nor looping alone is always
sufficient for the induction of gene expression, indicat-
ing the requirement of additional mechanisms for
the establishment of the pluripotency transcription
program.

The studies introduced thus far suggest a causative
relationship between gene regulatory factors and the
establishment of 3D chromatin organization, however
the requirement of specific inter-TAD chromatin con-
tacts for the induction of gene expression is very difficult
to show unequivocally. To this end, Fanucchi and col-
leagues demonstrated a hierarchy of gene expression
among distally located genes [35**] known to co-localize
upon TNF-alpha stimulation [16]. Among the genes
analyzed, SLC6A5 expression is rarely detected without
TNFAIP2 and SMAD4A expression, while TNFAIP2
expression is rarely detected without SMAD4A expres-
sion, arguing that, for their own expression, genes at the
bottom of the hierarchy (SLC6A5) show a strong reliance
on expression of genes above them in the hierarchy
(SMAD4A) [35**]. Remarkably, disruption of the
SMAD4A chromatin loop by TALEN-directed double
strand DNA break abrogated the expression of both
genes lower in the hierarchy, arguing that chromatin
loops and co-localization of genes over long distances in
cis and in trans are required for gene expression [35**].
Similar approaches applied to different interaction
scenarios will show how general the requirement for
co-localization is for the expression of co-regulated
genes.

In summary, the co-localization of distal chromatin frag-
ments bound by members of the same transcriptional
network within the 3D space of the nucleus appears to be
an important aspect of transcriptional regulation, perhaps
due to the resulting increase in the concentration of
specific gene regulatory factors at specialized transcrip-
tion factories [36] or Polycomb bodies [33]. This model
also explains how changes in cell identity lead to changes
in chromatin organization, as different transcriptional
networks bring about the co-localization of different
genomic regions during the course of development.
How these distal sites find each other and avoid co-
localizing with genes regulated by disparate transcription
networks within the nuclear volume remains unclear.
Another interesting observation is that specific 3D-inter-
actions could be essential for the function of long-non-
coding (lnc) RNAs. For instance, we speculated that the
interactions observed between Hox clusters could provide
the 3D conformation necessary for HOTAIR, a lncRNA
transcribed from the HoxC cluster, to find target genes
located within the HoxD cluster on a different chromo-
some, using a mechanism analogous to that employed
by another lncRNA, Xist, during X-chromosome inactivation
[137,37,38].

The logic behind enhancer–promoter–exon
looping
Apart from guiding global chromatin organization through
the establishment of long-range chromatin contacts, cell
type-specific gene regulatory factors also govern short-
range enhancer–promoter contacts, forming the founda-
tion for tissue-specific regulation of transcription. Exam-
ining promoter interactions in 1% of the genome across
three human cell lines (GM12878, K562 and HeLa-S3),
the ENCODE consortium demonstrated a surprising
promiscuity of enhancer–promoter interactions, showing
that many promoters in a given cell are contacted by
multiple enhancers, and vice versa, and that gene expres-
sion driven from a given promoter positively correlates
with the number of enhancers contacting it in a cell
population [39].

As the primary driver of cell type-specific gene expres-
sion, enhancer usage is dynamic during the course of
development. Correlation between the chromatin state at
enhancers and RNA polymerase II (RNAPII) occupancy
at promoters across numerous cell lines allowed for the
identification of co-regulated promoters and enhancers
[20]. These enhancer–promoter pairs showed a propen-
sity to cluster linearly in the genome, often falling within
TADs, and supporting the model that functional promo-
ter–enhancer interactions are delimited by TAD bound-
aries [20]. Genes at TAD boundaries, however, appear to
be able to switch their interactions between different
TADs. For instance, genes lying at the interface of two
TADs within the HoxD cluster switch the set of enhan-
cers with which they interact between contiguous
TADs, allowing for co-linear gene expression of the
HoxD cluster during the course of mouse limb develop-
ment [40**]. The co-regulation of enhancer chromatin
state and RNAPII occupancy, as well as development-
tally regulated changes in enhancer usage argue for a
role of developmental stage- and cell type-specific
transcription factors in the orchestration of enhancer-
promoter contacts.
Within the context of B-cell development, the cell type-specific transcription factors E2A or PU.1, as well as the histone acetyltransferase p300 (indicative of enhancers), are enriched at sites of both intra- and inter-TAD interactions that vary with developmental progression, suggesting that at least some interactions involving enhancer elements can cross TAD boundaries [11]. In line with these findings, a study by Phillips-Cremins and colleagues showed that Mediator and Cohesin, architectural proteins that are thought to facilitate 3D-chromatin interactions, act together within TAD boundaries to support enhancer–promoter interactions, but are also associated with longer-range promoter-enhancer interactions [21]. In the context of stimulus response, enhancers adjacent to 17β-oestradiol-upregulated genes in a human breast cancer cell line exhibited an increase in enhancer–promoter looping upon stimulation, supporting the importance of enhancer–promoter looping in control of gene expression [41]. Together these findings demonstrate that developmentally and stimulus-driven transcription programs are governed at the level of enhancer–promoter networks within TADs, with rare enhancer–promoter interactions crossing TAD boundaries.

Kieffer-Kwon et al. utilized ChiA-PET to identify 3D-chromatin interactions involving the pre-initiation transcriptional complex at promoters and found differential enhancer utilization across two cell types, not only for tissue-specific genes, but, surprisingly, also for constitutively expressed genes [42*], implying that highly dynamic enhancer–promoter interactions govern both cell type-specific and cell-type invariant transcriptional programs. A similar approach found that intragenic looping between promoters and exons facilitates alternative splicing in a cell-type-specific manner by bringing promoters and specific exons into close spatial proximity while looping out intronic sequences [43*]. Together, these results suggest that chromatin looping can occur between a variety of genetic elements within a given cell type, linking local genome organization to cis-regulation of both gene expression and alternative splicing.

Remarkably, despite the apparent role for transcription factor-driven enhancer–promoter loops and gene transcription, TNF-α-responsive enhancers are in contact with their target promoters prior to the induction of signaling genome-wide [44**]. This suggests that the 3D chromatin landscape is stable in a given cell type in the absence of signaling activation and that signaling networks act on pre-existing networks of enhancer–promoter contacts. Importantly, this finding also indicates that enhancer–promoter co-localization can be insufficient to initiate transcription. A similar case has been made for anti-pause enhancers that regulate promoter-proximal pause release. Binding by the histone demethylase JMJD6 and the bromodomain-containing protein Brd4 appears to occur at pre-established enhancer–promoter contacts which are not disrupted by loss of either of these two factors (Figure 2) [45**]. This suggests that enhancer–promoter contacts can be established without initiating gene expression, and that JMJD6 and Brd4-mediated pause release is an independently regulated event downstream of enhancer–promoter looping. The mechanism of establishment and maintenance of enhancer–promoter contacts in the absence of transcription may rely on the Mediator complex (see below), whose depletion leads to loss of enhancer–promoter looping at anti-pause enhancers [45**]. The function of enhancer–promoter contacts with regards to the initiation of transcription, and additional factors required to initiate transcription from an enhancer-contacted promoter will be important areas for future study.

### The linchpins of looping: architectural proteins and chromatin contacts

The establishment and maintenance of both inter- and intra-TAD chromatin interactions is thought to occur via recruitment of Cohesin, a protein complex that is known for its role in sister chromatid cohesion during mitosis [46]. Recruitment of Cohesin can occur through transcription factor-mediated recruitment of the Mediator complex and the Cohesin loading factor Nipbl [47], allowing for cell type-specific chromatin organization associated with gene-regulatory networks. Cohesin can also be recruited by the insulator protein CTCF [48–50], which governs cell type-invariant features of genome organization [11,21] and is required for proper Cohesin localization to CTCF-enriched sites [51]. As such, CTCF, Cohesin, and Mediator act as the ‘architectural’ proteins of the nucleus (Figure 2). In mouse ESCs and neural progenitor cells, CTCF, Cohesin, and Mediator are found at more than 80% of chromatin interactions, as defined by 5C, further supporting the notion that the three proteins play a central role in organizing chromatin [21].

Consistent with their role as effectors of cell type invariant features of chromatin organization, TAD boundaries are enriched for CTCF and Cohesin binding [17,18]. Genes found within chromatin loops anchored by CTCF binding sites often share similar chromatin modifications [52], in agreement with the co-regulated nature of genes located within a single TAD [20], supporting the idea that gene regulation often acts at the scale of TADs. TAD boundaries are well conserved across mammalian species and cell types [17,18], and insulator-binding proteins also serve to delimit distinct chromatin domains in Drosophila [53,54] arguing that insulator accumulation at TAD boundaries is an evolutionarily conserved aspect of genome organization. Despite an enrichment at TAD boundaries, CTCF/Cohesin-bound sites are not sufficient to block chromatin interactions [18,39], and CTCF binding is not sufficient to demar-
Architectural proteins act combinatorially to organize chromatin at different length-scales. (a) TAD boundaries are enriched for CTCF and Cohesin, but these proteins can also act in combination with other factors, such as Mediator to partition these large Mb-scale TADs into smaller sub-TADs and facilitate enhancer–promoter interactions. (b) A gene regulatory event involving a constitutive promoter–enhancer interaction. Mediator establishes a loop of a ‘anti-pause’ enhancer to a target gene promoter. Recruitment of the jumonji C-domain–containing protein 6 (JMJD6) and bromodomain–containing protein 4 (Brd4) complex leads to erasure of H4R3me2 and concomitant decapping/demethylation of 7SK snRNA, ensuring the release of the 7SK snRNA/HEXIM complex, which inhibits elongation factor P-TEFb, thereby permitting pause release and transcriptional elongation.

cate TAD boundaries, as only ~15% of all CTCF binding sites are found at TAD boundaries [17]. Similarly, insulator-binding proteins do not always block chromatin interactions in *Drosophila* [54]. Interestingly, CTCF/Cohesin co-occupancy within TADs form chromatin loops at length scales of a few hundred kilobases, leading to the concept of ‘sub-TADs’, which often form constitutive interactions around developmentally regulated, tissue-specific genes [21] (Figure 2). Together, these results suggest that architectural proteins can serve as boundaries of interactions of different strength, blocking certain interactions while allowing others dependent on the context.

Knockdown of CTCF not only reduces intra-TAD interactions, but also increases inter-TAD interactions, implying that CTCF depletion results in less well-defined TAD boundaries and more promiscuous short-range chromatin interactions, which are accompanied by alterations in gene expression [55]. Conversely, disruption of the Cohesin complex via proteolytic cleavage of the Rad21 protein leads to a diminution of intra-TAD interactions, but the TADs themselves remained intact [55], demonstrating a role for Cohesin in the maintenance of intra-TAD interactions. In line with this finding, knockdown of a Cohesin subunit in ESCs disrupted an interaction between the Pou5f1 promoter and a neighboring enhancer, causing the loss of self-renewal in pluripotent cells [34]. Extending the functional requirement for Cohesin to inter-TAD interactions, Apostolou and colleagues demonstrated the necessity of the Cohesin and Mediator complexes in the re-establishment and maintenance of pluripotency-specific long-range contacts of the Nanog locus upon reprogramming [14]. Similarly, depletion of Klf4 in ESCs leads to loss of Cohesin loading at the Pou5f1 enhancer, and loss of inter-TAD chromatin
contacts that are specific for the pluripotent state [22]. Supporting a combinatorial role for Cohesin and Mediator in facilitating tissue-specific contacts, Phillips-Creemans and colleagues showed that these two factors act together to facilitate interactions between enhancers and core promoters, mainly within TADs, but also at long-range between TADs [21]. Altered chromatin conformations and gene expression profiles upon loss of Cohesin do not appear to be due to mitotic defects, as genetic ablation of Cohesin in post-mitotic astrocytes caused decreased intra- and inter-TAD contacts, resulting in profound global architectural changes and extensive misregulation of gene expression [23]. Cohesin deletion did not ablate TAD boundaries, arguing that although Cohesin is required for proper chromatin organization and gene expression, it is not necessary for TAD boundary formation [23].

Together, the emerging data suggest that architectural protein-mediated inter-and intra-TAD chromatin contacts constitute a key mechanism for ensuring the stability of both cell type-specific and cell type-invariant features of mammalian genome architecture and global gene regulation, and for facilitating changes in genome architecture associated with differentiation (Figure 2).

Completing the loop and looping ahead: future directions
Recent cutting-edge cytological and 3C-based genome-scale research has helped to provide a deeper understanding of the complicated relationship between gene regulation and nuclear architecture in mammalian development. This work has made clear that the linear genomic features that control transcription help to shape the 3D space of the nucleus, and that the 3D organization of chromatin in turn plays a vital role in the regulation of gene expression, and, by extension, in the maintenance and establishment of cell identity.

Given the strong propensity of genomic regions bound by similar gene regulatory factors to co-localize, it will be important to determine how specific genomic regions locate each other within the space of the nucleus. Complementary work on the mechanisms used to avoid contacts with regions bound by different regulatory factors will also be important. Similarly, defining the molecular events that follow enhancer–promoter contacts and precede initiation of transcription will be important to properly define enhancer action and the relevance of promoter–enhancer interactions to gene expression.

A limitation of 3C-based studies is the requirement for a large population of cells during library preparation, meaning the resulting data represent the average chromatin contacts across the entire ensemble of cells, making it difficult to gauge the relevance and frequency of individual chromatin interactions. Single-cell, genome-wide chromatin contact maps recently recapitulated the domain structure characterized using population-based Hi-C, and showed that inter-TAD and inter-chromosomal contacts are highly variable between individual cells and that active domains were generally found at CT boundaries [56]. In future studies, it will be important to compare the variability observed for chromosomal interactions with that of gene expression at the single cell level. Matching this work with genome editing approaches able to disrupt and induce specific chromatin interactions [35**], single cell studies will go a long way towards resolving the direct effect of chromatin organization on gene expression.

Acknowledgements
K.P. is supported by the NIH (P01 GM099134), CIIRM (RB3-05880 and RB4-06133), the Jonsson Comprehensive Cancer Center, and the Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research at UCLA; G.B. by the Whitcome Pre-doctoral Training Program; and M.D. by pre-doctoral fellowships from the Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research at UCLA, CIIRM, and the UCLA Graduate Division.

References and recommended reading
Papers of particular interest, published within the period of review, have been highlighted as:
• of special interest
•• of outstanding interest


15. de Wit E et al.: The pluripotent genome in three dimensions is shaped around pluripotency factors. Nature 2013, 501:227-231 http://dx.doi.org/10.1038/nature12420. Demonstrated that ectopic recruitment of the pluripotency transcription factor Nanog to chromatin was sufficient to induce long-range, inter-TAD chromatin interactions with other genomic regions enriched for binding by Nanog.


Used Hi-C in the prokaryote Caulobacter crescentus to demonstrate the existence of TAD-analogous chromosomal interaction domains (CIDa) in this organism and applied chemical and genetic approaches to demonstrate that CIDa are a biophysical property of the prokaryotic genome, dependent upon transcription.
loops with their respective promoters. Exons with these features are enriched for alternative splicing events.


Used ultra-deep sequencing of Hi-C data to generate genome-wide contact maps in human cells with enhancer/promoter contact resolution. Surprisingly, they found that many NF-κB-responsive gene promoters were already in contact with their respective enhancers prior to induction of signaling, and that these pre-existing contacts were predictive of transcriptional activation upon signaling activation.


Demonstrated a mechanism by which paused polymerases could be released from promoters with pre-established promoter-enhancer looping.


52. Handoko L et al.: CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet* 2011, 43:630-638 [dx.doi.org/10.1038/ng.857](http://dx.doi.org/10.1038/ng.857).


Developed single-cell Hi-C, using it to demonstrate that the TAD-based structure of the genome is conserved at the single cell level, and that active domains tend to cluster at the surface of chromosome territories.

57. Simonis M et al.: Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* 2006, 38:1348-1354 [dx.doi.org/10.1038/ng1896](http://dx.doi.org/10.1038/ng1896).

58. Zhao Z et al.: Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet* 2006, 38:1341-1347 [dx.doi.org/10.1038/ng1891](http://dx.doi.org/10.1038/ng1891).


61. van Berkum NL et al.: Hi-C: a method to study the three-dimensional architecture of genomes. *J Vis Exp* 2010 [dx.doi.org/10.3791/1869](http://dx.doi.org/10.3791/1869).


Chapter 6

Long-range chromatin contacts in embryonic stem cells reveal a role for the pluripotency factors and Polycomb proteins in genome organization
Long-Range Chromatin Contacts in Embryonic Stem Cells Reveal a Role for Pluripotency Factors and Polycomb Proteins in Genome Organization

Matthew Denholtz,1,4 Giancarlo Bonora,1,4 Constantinos Chronis,1 Erik Splinter,3 Wouter de Laat,3 Jason Ernst,1 Matteo Pellegrini,5,6 and Kathrin Plath1,4
1Department of Biological Chemistry at the David Geffen School of Medicine
2Department of Molecular, Cell, and Developmental Biology
University of California, Los Angeles, CA, 90095, USA
3Hubrecht Institute-KNAW & University Medical Center, Utrecht, 3584 CT, The Netherlands
4These authors contributed equally to this work
*Correspondence: matteop@mcdb.ucla.edu (M.P.), kplath@mednet.ucla.edu (K.P.)
http://dx.doi.org/10.1016/j.stem.2013.08.013

SUMMARY
The relationship between 3D organization of the genome and gene-regulatory networks is poorly understood. Here, we examined long-range chromatin interactions genome-wide in mouse embryonic stem cells (ESCs), iPSCs, and fibroblasts and uncovered a pluripotency-specific genome organization that is gradually reestablished during reprogramming. Our data confirm that long-range chromatin interactions are primarily associated with the spatial segregation of open and closed chromatin, defining overall chromosome conformation. Additionally, we identified two further levels of genome organization in ESCs characterized by colocalization of regions with high pluripotency factor occupancy and strong enrichment for Polycomb proteins/H3K27me3, respectively. Underlining the independence of these networks and their functional relevance for genome organization, loss of the Polycomb protein Eed diminishes interactions between Polycomb-regulated regions without altering overarching chromosome conformation. Together, our data highlight a pluripotency-specific genome organization in which pluripotency factors such as Nanog and H3K27me3 occupy distinct nuclear spaces and reveal a role for cell-type-specific gene-regulatory networks in genome organization.

INTRODUCTION
Chromosome conformation capture (3C)-based technologies (de Wit and de Laat, 2012; Dekker et al., 2002) have led to a new paradigm wherein gene regulation can be studied in the context of the three-dimensional (3D) organization of the genome (Bickmore and van Steensel, 2013). Recent work has demonstrated an organizational hierarchy to metazoan genome structure (Gibcus and Dekker, 2013). At the smallest scale, up to a few hundred kilobases (kb) of linear DNA, enhancers and promoters come into physical contact to establish cell-type-specific expression programs (Sanyal et al., 2012; Shen et al., 2012; Smallwood and Ren, 2013). These interactions are maintained by the Cohesin complex, which can be recruited to interphase chromatin via the Mediator complex (Kagey et al., 2010) and cell-type-specific transcription factors (TFs) (Denholtz and Plath, 2012; Wei et al., 2013). In mammals, promoter-enhancer interactions are confined to topologically associating domains (TADs), which typically represent ~1 megabase (Mb) cell-type-invariant, self-associating genomic regions whose boundaries are enriched for the insulator protein CTCF (Dixon et al., 2012; Nora et al., 2012).

As a second level of the organizational hierarchy, TADs appear to function as the fundamental modular unit of gene regulation and genome organization, with changes in gene expression and nuclear lamina association during differentiation often occurring in a TAD-wide manner (Dixon et al., 2012; Nora et al., 2012; Shen et al., 2012). A third level in the organizational hierarchy occurs as a result of the preferential colocalization of specific TADs. These interactions can be identified as long-range chromatin contacts between genomic regions many Mb apart on the same chromosome (cis or intrachromosomal) or on different chromosomes (trans or interchromosomal) (Hakim et al., 2011, 2013; Noordermeer et al., 2011; Osborne et al., 2004; Schoenfelder et al., 2010; Simonis et al., 2006). Although spatially colocalizing distal genomic regions have been reported to be enriched for the sequence motifs of specific TFs (Schoenfelder et al., 2010), coexpressed genes (Osborne et al., 2004), or coregulated genes (Hakim et al., 2013; Noordermeer et al., 2011), the relationship between long-range chromatin interactions and the regulatory features enriched in the colocalizing regions is poorly understood. Superimposed upon the milieu of specific short- and long-range chromatin interactions is a general preference for open, accessible chromatin to colocalize with itself and segregate away from closed, inaccessible chromatin (Lieberman-Aiden et al., 2009).

In this study, we explored long-range (distal) chromatin contacts in mouse embryonic stem cells (ESCs) using 4C-seq to understand the relationship between chromatin contacts and gene-regulatory networks that govern cell identity. We define the distal chromatin interactions made by a variety of genomic “bait” regions representing diverse chromatin and pluripotency
TF binding profiles and extend our findings genome-wide. We further examine how genome organization changes in the absence of a critical chromatin regulator and upon differentiation and transcription-factor-induced reprogramming of somatic cells into induced pluripotent stem cells (iPSCs). Together, our data define a previously unappreciated hierarchy in the organization of long-range chromatin contacts and reveal that distal genomic regions sharing common gene-regulatory features co-localize within the 3D space of the nucleus.

RESULTS

Experimental Approach to Studying Chromatin Contacts

To investigate long-range chromatin interactions between genomic regions Mb away on the same or different chromosomes in ESCs, we performed 3C coupled to high-throughput sequencing (4C-seq) (Splinter et al., 2012) for 16 bait regions (Table S1 available online). 4C-seq allows one to identify any mappable genomic region in close physical proximity to a specific genomic (bait) region within a population of cells at the moment of fixation by means of proximity-based ligation of juxtaposed DNA fragments. The product is a library of chimeric DNA fragments containing the bait region and its interacting DNA partner or partners ligated at a restriction site, in our case HindIII, that can be identified by high-throughput sequencing (Experimental Procedures).

Our 4C-seq data were highly reproducible across biological and technical replicates utilizing distinct primer pairs for amplification of ligated fragments and different HindIII fragments within bait regions as anchor points and passed stringent quality control requirements (Figures S1A, S1D, S1E, and S2A available online, Table S2 and Table S3, and Experimental Procedures). As a result, replicate data sets for each bait locus were pooled for downstream analysis (Experimental Procedures). To obtain a semiquantitative measure of interactions, we calculated an average hit probability, referred to as “hit percentage,” within 200 kb windows along the genome (Experimental Procedures). For all of our baits, the hit percentage was higher in cis than in trans, as exemplified by the 4C-seq interactome of the Pou5f1-encoding region (Figure 1A). Furthermore, we expected that our results have a previously published genome-wide ESC interactome (Lieberman-Aiden et al., 2009) would decays as a function of genomic distance to the bait along the linear DNA (Figure 1Aii).

Significantly interacting regions were identified as those 200 kb windows that showed a markedly higher hit probability than expected based on a binomial test (Figures 1Aii–1Av, 1B, Figure S3A, Table S1, Table S4, Experimental Procedures). To identify distal interactions in cis, we empirically modeled the expected background hit probability as a function of distance from the bait locus (Figure 1Aiii). Since trans interactions show no such positional biases, the average hit probability across each trans-chromosome was used for the expected background level of these interactions. False discovery rates (FDRs) were estimated using simulated data (Table S4, Experimental Procedures).

To test the reliability of our analysis and 4C-seq data, we partitioned our Pou5f1 4C-seq replicates into two equal subsets, pooled the libraries within each subset, and ran each pooled subset through our analysis pipeline. We obtained significantly overlapping intrachromosomal and interchromosomal interactomes from the partitioned data sets (Figure 1B), confirming the quality of our data and the robust nature of our analysis.

Several additional approaches were employed to further validate our 4C-seq-defined chromatin interactions and rule out technical biases affecting data generated by 4C-seq. These include 3C proof of ligation products (Figure S1F); 3D fluorescence in situ hybridization (FISH) demonstrating that an interacting region 52 Mb away from the Pou5f1 bait region is significantly closer to the bait in 3D space of the ESC nucleus than a noninteracting region located 35 Mb away (Figure 1C); reciprocal 4C-seq (Figure 1D); control 4C-seq experiments with unixed cells and genomic DNA showing no significant interactions (Figures 1A and 1C); the demonstration of comparable mappability inside and outside our interacting regions (Figure S1G); and the comparison of our 4C-seq data with a previously published genome-wide ESC interactome based on Hi-C (Dixon et al., 2012) showing strong correlation (Figure 3H, Table S5, Experimental Procedures).

A Pluripotency-Specific Organization of the Mouse Genome

The Pou5f1 gene encodes the TF Oct4 that is essential for the establishment and maintenance of mouse ESCs and is specifically expressed in the pluripotent state (Nichols et al., 1998). The initial analysis of our Pou5f1 4C-seq data revealed an extensive intrachromosomal interaction network of this bait region (Figure 1A), raising the question of how the spatial interactions of this important locus are regulated. As a first step toward understanding this question, we determined whether the chromatin interactions of this genomic region change upon differentiation. 4C-seq in mouse embryonic fibroblasts (MEFs) demonstrated that the Pou5f1 interactome in MEFs is distinct from that of ESCs, both in terms of hit probability (Figure 2A and Figure S2A) and interacting domains (Figure 2B) across the cis chromosome, indicating that changes in expression and chromatin state that are associated with the silencing of the Pou5f1 locus during differentiation (Feldman et al., 2009) coincide with changes to that locus’ interactions in 3D space. To expand the purview of this analysis, we determined the interactomes of nine additional baits in MEFs for which we had already generated ESC profiles by 4C-seq and found that the interactomes of all these baits differed significantly between ESCs and MEFs (Figures 2E and 2F, Figure S2A, Table S1). We noted that the chromatin contacts of the Dppa2 bait region are much more different between MEFs and ESCs than those of the other examined bait regions (Figures 2C and 2D, Figure S2A). Consistent with this result, it has been shown that the Dppa2 region repositions toward the nuclear periphery and switches cis-interaction preferences from regions of early DNA replication to regions of late DNA replication during ESC differentiation (Hiratani et al., 2010; Takebayashi et al., 2012). Together, these data highlight an ESC-specific organization of the genome and indicate that changes in 3D chromatin interactions during differentiation are regulated at a locus-specific level.

We next tested whether the 3D organization of the MEF genome can be reset to an ESC-like state by transcription-factor-mediated reprogramming to iPSCs (Takahashi and Yamanaka, 2006). To this end, we determined the interactomes of eight of our bait regions, including the Pou5f1 locus, for two
additional cell types that represent distinct stages of the reprogramming process (Table S1). First, we performed 4C-seq on faithfully reprogrammed, MEF-derived iPSCs and found that the long-range chromatin contacts in iPSCs are highly similar to those of ESCs, in terms of both hit probability across the cis chromosome (Figures 2A, 2C, and 2E, Figure S2A) and interacting domains in cis and in trans to the bait region (Figures 2B, 2D, and 2F, Figure S2B). Second, we examined chromatin interactions in pre-iPSCs that represent a late reprogramming stage at which the pluripotency expression and chromatin program is not yet fully induced (Sridharan et al., 2009). The long-range chromatin interactomes in these cells are typically distinct from those in ESCs and iPSCs, as well as those in MEFs (Figure 2, Figure S2), indicating that the reorganization of chromatin contacts is not complete at this late stage of reprogramming, in line with previous findings detailing the long-range interactions of the Nanog locus (Apostolou et al., 2013). Together, these data show that the large-scale changes in genome organization that
The finding that changes in distal chromatin contacts occur during differentiation and reprogramming suggested a link between genome organization and genomic features that establish and maintain cell identity. To investigate such an association in the context of mouse ESCs, we compiled a compendium of genomic features available for this cell type (Table S6). Specifically, we considered the binding profiles of the basic transcriptional machinery (TBp, RNA polymerase II), coactivators (Mediator and p300), and architectural proteins (Cohesin and CTCF). We also included transcriptional regulators of three important gene-regulatory networks: the pluripotency TF network (Oct4, Sox2, Nanog, and Klf4), TFs that cooperate with cMyc (cMyc, Max, E2F4), and the repressive Polycomb protein network (Ring1b and Polycomb Repressive Complex 2 [PRC2]) (Young, 2011). Since regulatory genomic regions exhibit extensive co-occupancy by these factors, we grouped them into 11 groups (clusters) based on co-binding at a 1 kb resolution (Figure 3A, Experimental Procedures). We also took into account histone modifications and their combinatorial nature by summarizing the relationship of six histone marks in terms of four functionally distinct chromatin states (ChromHMM states) (Ernst and Kellis, 2012) that are associated with Polycomb repression, transcriptional elongation, enhancers, and promoters. Regions lacking these histone modifications were assigned to a “low signal” state (Figure 3B, Experimental Procedures). In addition, DNA replication timing, DNasel hypersensitivity, gene density (in terms of transcriptional start sites [TSSs]), RNA-seq-based expression data, and LaminB association were considered.

Upon binning of the resulting 22 linear genomic data sets into 200 kb windows across the genome, principal components analysis (PCA) was used to reduce their high dimensionality (Experimental Procedures). The first three principal components (PCs) were retained for downstream analyses (Figure 3C). To explore the relationship between genomic features and genome organization, we next compared the linear genomic character represented by the PCs to the 4C interactome data for each of the 16 bait regions analyzed in ESCs (Figure 3A, Figure 1B, Table S1).

The first PC (PC1) captures 51% of the variance across all features (Figure 3C) and distinguishes open, accessible chromatin and closed, inaccessible chromatin (Figure 3D). Specifically, regions of the genome with positive PC1 scores are characterized by high gene density, DNasel hypersensitivity, binding of Cohesin, the basic transcriptional machinery, and TFs, as well as active or Polycomb-repressed chromatin states within 200 kb windows. In contrast, regions of the genome with negative PC1 scores are strongly depleted for these features and instead are LaminB associated and replicate their DNA late in S phase (Figure 3D). Notably, the 1 Mb regions surrounding each bait locus had widely differing open/closed chromatin character as defined by their PC1 scores (Figure 3Ei, top panel). Strikingly, we found that the mean PC1 scores for the 1 Mb bait regions correlated strongly with the mean PC1 scores of the
Figure 3. Interactions between Regions with Similar Open/Closed Chromatin Character Are an Intrinsic Aspect of Chromosome Conformation in Mouse ESCs

(A) Transcription factor (TF) clusters defined using k-means clustering at 1 kb resolution for noted mouse ESC data sets, annotated based on feature frequency as represented by the heatmap. Gray scale denotes the frequency with which a given factor is found at genomic positions corresponding to the cluster. The color legend identifies specific gene-regulatory networks.

(legend continued on next page)
corresponding interactomes in both cis and trans (Figure 3Ei, Spearman’s rho = 0.75 and 0.85, respectively). That is, baits with highly open and accessible chromatin preferentially interact with regions of the genome with similarly high PC1 scores. By contrast, the interactomes of closed chromatin baits, i.e., those with negative PC1 scores, preferentially colocalize with genomic regions of similarly low PC1 scores in cis and trans. These findings were confirmed by a 4C-seq analysis for a partially overlapping set of bait regions in a second, independent ESC line (Figure 3Eii).

To test whether this trend holds genome-wide, we analyzed a previously published Hi-C data set for genome-wide chromatin interactions in mouse ESCs (Dixon et al., 2012) in a similar manner (Experimental Procedures). For the Hi-C-based analysis, each 200 kb region of the genome was treated as a “pseudo-bait” and its intrachromosomal interactome was extracted (Figure 3F). Genome-wide, we found a striking positive correlation between the mean PC1 scores of the extended 1 Mb pseudo-bait regions and the mean PC1 scores of their most frequently interacting intrachromosomal sites (Figure 3G), corroborating and extending our 4C-seq-based findings. A comparison of the profiles of the most and least likely interacting regions further demonstrated that genomic regions with very high, open PC1 character, for instance the Pou5f1 region, interact extensively with most other genomic regions with similarly high PC1 scores on the chromosome in cis and do not interact with genomic regions of negative PC1 scores (Figures 3G and 3H, Figure S3Ci). In contrast, genomic regions of closed chromatin character, defined by negative PC1 scores, avoid regions of high PC1 character and appear to interact selectively with only a subset of PC1-negative genomic regions (Figure S3Ci). This may reflect a tethering of closed chromatin regions to the nuclear lamina, thus limiting their sampling of distal interactions. The positive correlation between bait and interactome character in terms of their PC1 scores persisted when the 10 Mb surrounding the bait region were excluded (Figure S3Dii), demonstrating that interacting regions far away from the bait in cis also show an association with chromatin of similar open/closed character.

Taken together, we conclude that genomic loci with similar PC1 characters preferentially interact (or colocalize) within the 3D space of the ESC nucleus both in cis and in trans. The Hi-C-based results indicate that these associations are a general, genome-wide feature of long-range chromatin interactions. Our findings suggest that the strong interaction preferences between regions of similar PC1 character are an intrinsic aspect of overall chromosome conformation, in line with previous findings (Imakaev et al., 2012; Lieberman-Aiden et al., 2009; Simonis et al., 2006). The data further demonstrate that interaction preferences in ESCs are not accurately described by a binary model of spatial segregation between open and closed chromatin states into two genome-wide compartments, where genomic regions with open chromatin character may colocalize with any other open chromatin region and vice versa (Lieberman-Aiden et al., 2009). Instead, the ESC interactome follows a more graduated model, where highly open chromatin regions predominantly interact with regions of similarly high PC1 scores, mid PC1 regions predominantly interact with other mid to low PC1 regions, and negative PC1 regions predominantly interact with other closed chromatin regions, supporting findings in other cell types (Imakaev et al., 2012). Notably, PC1 scores across the ESC genome are continuous (Figure S3B) and correlate with the continuum of ESC contact frequencies.

Genomic Regions Enriched for Oct4/Sox2/Nanog and Polycomb Proteins Frequently Colocalize in ESCs

Next, we wanted to explore the extent to which long-range chromatin interactions are associated with specific transcriptional networks, beyond their association with the large-scale open/closed chromatin properties demonstrated by the correlation of interactomes with the genomic PC1 character. Since PC1 positive regions reflect enrichment for multiple features with diverse functionalities, we reasoned that an interactome’s PC1 character may not reflect any specific mechanistic role of PC1-enriched features, but is most likely a consequence of the overarching chromosome conformation framework (Figures S4A–S4D). Therefore, we considered the second and third principal components (PC2 and PC3), which account for 7% and 6% of

(B) Chromatin states were determined based on the six indicated histone modifications in ESCs by a multivariate hidden Markov model, at 200 bp resolution (Ernst and Kellis, 2012). Gray scale denotes the frequency with which a given histone mark is found at genomic positions corresponding to the chromatin state.

(C) PCA was performed on ESC chromatin states and TF clusters from (A) and (B), RNA-seq expression data, DNaseI hypersensitivity (HS), LaminB binding, early and late DNA replication timing (Rep. timing), and density of transcriptional state sites (TSSs) upon binning of the linear genomic data into 200 kb windows across the genome. Proportion of total variance in genomic features described by each principal component is shown.

(D) PC1 eigenvector ranked by genomic feature contribution.

(E) (Ei) Top to bottom: mean PC1 score within the 1 Mb bait region centered on each listed bait’s locus; interacting regions in cis; noninteracting regions in cis; interacting regions in trans; and noninteracting regions in trans. Spearman’s rho values give the rank correlation between the mean PC1 score of the 1 Mb bait regions and their interactomes in both cis and trans. (Eii) Similar analysis to (Ei) with a partially overlapping set of baits, for an independently derived ESC line, which is discussed in Figure 7 as EedESC line.

(F) Schematic of genome-wide, Hi-C-based, pseudo–4C analysis. (Fi) Each extracted row of the Hi-C contact matrix, adapted from (Dixon et al., 2012), represents the interactome of one pseudo-bait, at 200 kb resolution. (Fii) Plot of the PC1 character of the same chromosome. For each 200 kb pseudo-bait, the mean PC1 score within the extended 1 Mb pseudo-bait region (Pseudo-bait PC1 character) and the mean PC1 score within the pseudo-bait’s interactome (top 5% of 200 kb windows ranked by reads and excluding the 1 Mb pseudo-bait region) were obtained and plotted as a red point in the scatterplot shown in (G). Pseudo-bait regions corresponding to genomic regions that we used as baits in our 4C-seq analysis in (Ei) are plotted in yellow (4C-bait loci).

(G) Result of the analysis described in (F). 4C-bait loci show a similar trend when analyzed based on Hi-C data as in (Ei) based on 4C-seq. Correlations between bait and interactome PC1 scores are noted. The Hi-C data are also summarized by the regression line in black, and the mean bait and interactome PC1 scores are demarcated by vertical and horizontal gray lines, respectively. Contour lines represent data density.

(H) Top to bottom: comparison of the Pou5f1 4C-seq-based cis-interactome (Hi1%), the Pou5f1 cis-interactome defined by Hi-C read counts, PC1 scores, DNaseI HS, and late DNA replication timing along chr17. Specific correlation values are specified. See also Figures S3 and S4.
Figure 4. Regions of Shared Transcriptional Network Occupancy Preferentially Interact
(A) PC2 eigenvector with individual feature contributions. TF clusters and chromatin states are as in Figure 3. (B) Integrative Genomics Viewer tracks of a representative genomic region with (top to bottom) PC1 and PC2 scores; Sox2, Oct4, and Nanog occupancy; and enhancer density. (legend continued on next page)
the variance in genome features, respectively (Figure 3C). Importantly, these components capture regions of the genome enriched for previously described gene-regulatory networks in ESCs (Figure 4A and 4E). Specifically, genomic regions with positive PC2 scores are exceptionally enriched for Mediator and Cohesin binding, as captured by TF clusters 5, 6, and 11, but also late-replicating and strongly LaminB-bound across 200 kb, while regions with negative PC2 scores are enriched for binding of factors belonging to the pluripotency network, including Nanog, Oct4, and Sox2, as well as p300, which colocalize as TF clusters 7 and 9 (Figure 4A). Figure 4B displays the relationship between PC1 and negative PC2 scores in terms of these features: whereas genomic regions of open chromatin (defined by positive PC1 values) generally have a higher level of pluripotency factor binding than those with negative PC1 scores, genomic regions with negative PC2 scores have an increased density of pluripotency factor co-occupancy over and beyond what is explained by their PC1 state.

Conversely, regions of the genome with positive PC3 values are highly enriched for occupancy by the Polycomb protein complexes PRC1 (Ring1b) and PRC2, as well as histone H3K27me3. Negative PC3 regions, on the other hand, seemingly capture the recently described super-enhancers associated with highly transcribed genes in pluripotent cells (Whyte et al., 2013) and represent a strong enrichment for active transcriptional elongation along with dense occupation by Mediator, Cohesin, the cMyc complex, and the pluripotency TFs Oct4, Sox2, and Nanog (Figure 4E, Figures S4E and S4F). Genomic regions with positive or negative PC3 scores have an elevated density of their characteristic features above what is explained by their PC1 scores (Figure 4F, data not shown). We conclude that PC2 and PC3 describe a finer layer of chromatin structure associated with specific TF- and chromatin-regulatory networks that is not captured by the open and closed chromatin character defined by PC1.

To examine whether the genomic states captured by PC2 and PC3 are associated with long-range chromatin contacts, we examined the PC2 and PC3 scores within our 16 ESC bait regions and their respective interactomes in an analogous manner to our analysis of PC1 in Figure 3. Importantly, our bait regions exhibited widely different PC2 characters, and the mean PC2 score of each bait’s intrachromosomal interactome showed a strong concordance with the mean PC2 score of the 1 Mb bait region itself (Figure 4Ci). This observation was confirmed for additional bait regions in a second ESC line (Figure 4Gi) and, remarkably, also for cis interactions genome-wide based on Hi-C data (Figure 4D, Figures S3Cii and S3Dii). The PC2-based analysis of genomic regions with particularly strong Mediator, Cohesin, and Lamin binding preferentially colocalize in cis as do regions extremely highly enriched for Oct4, Sox2, and Nanog binding without a concomitant extreme accumulation of Mediator and Cohesin (captured by TF clusters 7 and 9). Notably, however, the colocalization of pluripotency-factor-enriched, PC2-negative genomic regions was not evident for trans interactions (Figure 4C).

The comparison of bait and interactome PC3 scores demonstrated a preferential colocalization of Polycomb/H3K27me3-enriched, PC3-positive genomic regions (captured by TF clusters 6, 8, and 10 and the Polycomb Repression ChromHMM state), both in cis and in trans (Figure 4Gi). For example, Hox loci, which belong to the most strongly Polycomb-occupied and H3K27me3-enriched (PC3-positive) regions in the ESC genome, were found to interact with other regions characterized by high H3K27me3/Polycomb occupancy in cis and trans (Figure 4G). The colocalization of distal genomic regions extensively occupied by Polycomb proteins and H3K27me3 may be analogous to the frequent association of Polycomb response elements in Drosophila (Bantignies and Cavalli, 2011) and may therefore represent an evolutionarily conserved feature of genome organization linked to gene regulation. Conversely, bait loci with negative PC3 values that are strongly enriched for ESC super enhancers colocalize with genomic regions of similar PC3-negative character in both cis and trans (Figure 4Gi). These trends were confirmed in a second ESC line (Figure 4Gii) and extended to Hi-C data (Figure 4H, Figures S3Ciii and S3Diii), showing them to be a genome-wide phenomena.

Together, these results argue that regions of the genome enriched for specific gene-regulatory features preferentially colocalize within the 3D space of the nucleus, raising the possibility that specific transcriptional and chromatin-regulatory networks are involved in mediating long-range chromatin contacts in ESCs.

**Spatial Segregation of Nanog and H3K27me3 in the ESC Nucleus**

The preferential interactions of genomic regions with positive and negative PC3 scores (i.e., Polycomb protein versus super enhancer/Oct4/Sox2/Nanog-enriched regions) suggested a segregation of genomic regions with opposing PC3 character into distinct compartments in the nucleus. To test this hypothesis, we examined the colocalization of Nanog (PC3-negative), RNA polymerase II (PC3-negative), and H3K27me3 (PC3-positive) in the ESC nucleus by immunostaining. Image analysis showed that Nanog and RNA polymerase II have a localization pattern distinct from that of H3K27me3,
with sites strongly enriched for H3K27me3 displaying weak Nanog and RNA polymerase II accumulation and vice versa (Figures 5A and 5B). By contrast, RNA polymerase II and Nanog overlap more extensively, albeit not perfectly (Figure 5C), consistent with their differential contribution to PC2 and PC3 scores. For instance, Nanog occupancy is strongly captured by TF cluster 7 and RNA polymerase binding by TF cluster 11, which have similar contributions to PC3, but opposing contributions to PC2 (Figures 4A and 4E). Overall, the immunofluorescence localization patterns support the spatial segregation of functionally distinct TF- and chromatin-regulatory networks in the ESC nucleus and are consistent with the colocalization of distinct gene-regulatory modules detected by our 4C-seq-based analysis.
Changes in Open/Closed Chromatin Character Mirror Changes in Genome Organization during Differentiation

Our analysis revealed a close relationship between chromatin character and spatial interactions, with the combination of linear genomic features summarized by PC1 (open/closed chromatin) showing the strongest association with the fundamental organization of chromatin interactions. Based on these results, we predicted that dramatic changes in open/closed chromatin character that occur during differentiation should coincide with strong changes in interactome character.

To test this, we examined differences in chromatin interactions between ESCs and MEFs with respect to changes in their open/closed chromatin character. To this end, we performed PCA on concatenated ESC and MEF genomes using linear genomic feature data sets that were available for both cell types but did not include cell-type-specific TFs (Figure 6A, Table S6, See also Figure S5.).
Experimental Procedures). This resulted in a new PC1 eigenvector that allowed the comparison of PC1 scores for ten bait regions and their interactomes across both cell types (Figures S5A–S5D). We found that, as in ESCs, regions with similar PC1 character in MEFs preferentially colocalize in both cis and trans. (Figure 6C). Furthermore, changes in bait character between MEFs and ESCs are generally associated with similar changes in their respective interactomes (Figures 6B–6D). For instance, the Dppa2 locus participates in extensive interactions with genomic regions of positive PC1 character in ESCs, in accordance with its early replicating and highly transcribed state in pluripotent cells (Takebayashi et al., 2012). In MEFs, the Dppa2 bait region displays a PC1-negative, repressed, late replicating state, and the MEF-specific interactions likewise exhibit negative PC1 scores, both in cis and in trans (Figures 6B–6D, Figure S5E). In contrast, the Rhbdc1 bait region transitions from negative to positive PC1 scores from ESCs to MEFs, and its MEF-specific interactions have significantly higher PC1 scores than the ESC-specific interactions (Figures 6C and 6D). In addition, bait regions without a change in PC1 character do not change interaction preferences with regards to PC1 scores (Figures 6C and 6D, Hoxa10).

Together, these data support a model where open/closed chromatin character is the strongest predictor of interaction preferences between distal genomic loci, and long-range chromatin interaction preferences are subject to change during differentiation in concordance with changes to the PC1 nature of the regions in question (Figure 6E). Of note, although the Pou5f1 gene itself becomes repressed during differentiation (Feldman et al., 2006), the extended 1 Mb Pou5f1 bait region is strongly positive in PC1 character even in MEFs (Figure 6C), likely explaining the more limited difference in spatial interactions between ESCs and MEFs for this bait region compared to the Dppa2 bait (Figure 2), which shows a more dramatic change in PC1 character upon differentiation.

The Preferential Colocalization of Polycomb-Enriched Genomic Regions Is Eed Dependent

Our data demonstrated that genomic regions are more likely to contact each other when they share strong enrichment of similar regulatory proteins (Figure 4). Therefore, we considered testing the functional importance of a specific gene-regulatory network for long-range chromatin interactions. To this end, we determined the long-range chromatin contacts in ESCs lacking the protein Eed, a subunit of Polycomb complex PRC2 that is required for all genomic H3K27me3 (Montgomery et al., 2005). We examined particularly the interactions of bait regions with positive PC3 scores, which capture high occupancy by PRC2 and H3K27me3 (Table S1). Importantly, despite the complete loss of H3K27me3 (Figure S6A), Eed+/− ESCs continue to express pluripotency-specific TFs including Oct4, Sox2, and Nanog at normal levels, maintain their ability to self-renew, and do not spontaneously differentiate when cultured appropriately (Chamberlain et al., 2008), allowing us to test the role of PRC2 in genome organization without a change in cell identity.

Notably, 4C-seq analysis showed that the intrachromosomal long-range interactions of the Hoxd cluster, a Polycomb-targeted genomic region with highly positive PC3 scores, correlated strongly between the Eed+/− and the Eed−/− ESC lines (Figure 7A). However, despite this overall similarity, numerous intrachromosomal interactions present in Eed+/− ESCs are lost or have less significant interactions by p value, indicative of a reduced interaction frequency, in Eed−/− ESCs (Figure 7A, yellow and orange highlights, respectively). A visual inspection of chromatin contacts indicated that these losses and reductions appear to occur at regions of high Polycomb enrichment in wild-type ESCs, as defined by high PC3 scores (Figure 7A). In

Figure 7. Eed Is Required for the Colocalization of Polycomb-Occupied Genomic Regions

(A) Integrative Genomics Viewer tracks showing the Hoxd12 interactome in cis in terms of Hit% for Eed+/− (blue) and Eed−/− (red) ESCs, chromosome-wide PC1 scores (black) overlaid with PC3 scores (green, positive values shown only), binomial test – log(p values), and interacting domains. Regions that lose significant interactions with the Hoxd12 bait upon Eed ablation are marked with yellow triangles and shading; those that do not lose interactions with the Hoxd12 locus upon Eed ablation but show a decrease in interaction strength are marked with orange triangles and shading. The Spearman’s ρ value shows the rank correlation of the Hit% between Eed+/− and Eed−/− ESCs, (B) Hoxd12 4C-seq Hit% tracks in Eed+/− and Eed−/− ESCs from (A) were subtracted and the 200 kb windows with the top and bottom 5% of resulting values were used to define regions of the cis chromosome that showed stronger interactions in Eed+/− (WT > MT) and Eed−/− (MT > WT) ESCs, respectively. The PC3 score distributions of these genomic regions and of the entire chromosome are shown. Box and whisker demarcation are as in Figure 1C, with notches = 95% confidence interval around medians. ’’p < 0.001, Wilcoxon rank-sum test.

(C–E) As in (B), but for the Hoxa10, Hoxb3, and Tbx5 4C-seq cis interactions. ’p < 0.05. (F–H) Trans interactions between the indicated (PC3-positive) Hox loci in Eed+/− and Eed−/− ESCs, displayed as in (A). Hoxa refers to results of the Hoxa10 bait locus; Hoxb, to Hoxb3; Hoxc, to Hoxc4; and Hoxd, to Hoxd12.

(I) DNA FISH analysis of the trans interactions between Hox clusters. (II) Cumulative frequency distribution plots of colocalization frequencies between Hoxb3 (chr1) and the other three Hox loci (Hoxa10-chr6, Hoxc4-chr15, Hoxd12-chr2) (left), as well as between Hoxb3 and the Sox2 (chr3) locus (right), with colocalization distances noted on the x axis, measured in Eed+/− (blue) and Eed−/− (red) ESCs. (III) Colocalization frequencies at 1 μm for Hoxb3 and the other Hox loci (left), as well as for Hoxb3 and Sox2 (right), derived from (II), n = FISH signal pairs analyzed in both (II) and (III); p value from two-tailed Fisher exact test. (J) The cis interactomes of the six PC3-positive (Polycomb/H3K27me3 enriched) bait loci (Hoxa10, Hoxb3, Hoxc4, Hoxd12, Pcdh19, and Tbx5; see Figure 4GII) were ranked by –log(p value) for both Eed+/− (blue) and Eed−/− (red) ESCs, and the 500 top genomic sites were plotted against their average PC3 scores in wild-type ESCs. Loess regression was used for curve generation. KS, Kolmogorov-Smirnov test to determine the probability that the two underlying probability distributions differ (D = KS – test D statistic).

(K) As in (J), but for the trans interactomes.

(L) As in (J), except for PC1 scores.

(M) As in (K), except for PC1 scores.

(N) Chromatin interaction model wherein in the absence of Eed, the frequency of interactions between regions with high PC3 scores is reduced, but large-scale chromosome conformation is largely conserved. Gray scale reflects the regions of the genome that are more open/accessible (light) versus more closed (dark). See also Figure S6.
agreement with this, long-range chromatin interactions of the Hoxd cluster that specifically occur in Eed+/− ESCs have a significantly more positive PC3 character than those that are specific for Eed−/− ESCs (Figure 7B). This result extends to other Polycomb-regulated regions such as the Hoxa and Hoxb cluster and the Tbx5 locus (Figures 7C–7E, Figure S6D).

A comparison of chromatin interactions in trans between Eed+/− and Eed−/− ESCs revealed that the Hox clusters interact with each other as well as with other regions of high Polycomb/H3K27me3 enrichment that encode developmental regulators in wild-type ESCs (Figures 7F–7H). Importantly, many of these interchromosomal contacts are also lost or reduced in the absence of Eed (Figures 7F–7H). For instance, interactions of the Hoxb cluster with the Hoxa, Hoxc, and Hoxd clusters are observed in wild-type ESCs and are diminished in knockout cells (Figures 7F–7H). We also found that the colocalization frequency between the Hoxb cluster and the Hoxa, Hoxc, and Hoxd clusters in wild-type ESCs was significantly higher than in knockout ESCs when measured by FISH (Figure 7I, Figures S6B and SBC), which is consistent with our 4C-seq results.

To explore whether the absence of Eed specifically affects chromatin contacts that occur between genomic regions characterized by positive PC3 scores, we examined the chromatin character of the most significantly interactingdistal regions across six PC3-positive baits (Table S1). We found that in Eed+/− ESCs, the highest-ranking (and likely most frequent) interactions fall within genomic regions with highly positive PC3 and PC1 scores and with less strongly positive PC2 scores both in cis and in trans (Figures 7J–7M, Figure S6E). In the absence of Eed, the strongest chromatin contacts no longer occur with regions that are highly positive for PC3 in wild-type ESCs (Figures 7J and 7K), although they still take place between distal genomic regions of similarly positive PC1 and PC2 scores both in cis and in trans (Figures 7L and 7M, Figure S6E). The corollary to this finding is demonstrated by two bait regions that are not enriched for Polycomb binding (Pou5f1 and Pltpg): they do not show a similarly dramatic difference in interaction preferences between Eed+/− and Eed−/− ESCs with regards to any of the three principal components (Figures S6F–S6I). These data indicate that loss of PRC2 and H3K27me3 specifically alters the coassociation of PC3-positive genomic regions but does not dramatically affect spatial interactions associated with PC1 and PC2 character.

Based on these data we conclude that Eed is required for the establishment and/or stable maintenance of interchromosomal and intrachromosomal chromatin interactions between Polycomb-occupied, PC3-positive regions in ESCs. Our data also suggest that the overall chromosome topology does not dramatically change upon loss of Eed (Figure 7A). The results indicate that regions that are spatially interacting and Polycomb protein-bound in wild-type ESCs remain confined by a similar chromosome topology in the absence of Eed, but their interaction frequency, i.e., their proximity, is dramatically reduced (Figures 7F–7H, model in Figure 7N).

**DISCUSSION**

Our work describes a pluripotency-specific organization of the mouse genome and suggests that distal regions of the genome bound by similar regulatory proteins colocalize within the 3D space of the ESC nucleus. Based on our data, we propose a model with two layers of regulation for long-range chromatin contacts in ESCs. (1) We posit that, at the largest scale, the open/closed chromatin character (described by the PC1 character of the genome) defines the regions of the genome that have the potential to come into close spatial proximity with one another, both intrachromosomally and interchromosomally, which is intricately linked to the overall folding of the chromosome and in agreement with other recent findings (Imakaev et al., 2012; Lieberman-Aiden et al., 2009). (2) Our data also suggest that on a finer scale, and within the constraints established by the open/closed chromatin architecture, genomic regions are more likely to contact each other when they share strong enrichment of similar regulatory proteins, such as binding by the pluripotency TFs Oct4, Sox2, and Nanog with Mediator and Cohesin (represented by negative PC3 scores) or binding of the repressive Polycomb complex (captured by positive PC3 scores).

Our results demonstrate that the depletion of a single gene-regulatory network in ESCs specifically affects long-range interactions of genomic regions particularly strongly enriched for occupancy by this network (in our case PRC2 and H3K27me3) without altering the global interaction network associated with open/closed chromatin character. A potential explanation for the limited effect on overall chromatin interactions may be that the chromosomal conformation chassis is maintained by many combinatorially acting regulatory factors that probably involve numerous interactions mediated by Cohesin (Apostolou et al., 2013; Phillips-Cremins et al., 2013). Notably, in addition to our description of PRC2 as a critical regulator of specific long-range chromatin interactions (Figure 7I), a functional requirement for the TF Klf4 in the maintenance of long-range chromatin contacts in ESCs has recently been reported (Wei et al., 2013), extending previous findings that demonstrated a requirement for Oct4 in the organization of short-range chromatin interactions within the Nanog locus (Levasseur et al., 2008).

The role of cell-type-specific gene-regulatory networks in defining specific long-range chromatin interactions potentially allows TADs enriched for specific gene-regulatory features to colocalize in the 3D space of the nucleus. Interestingly, while our results demonstrate that Polycomb complexes are important for long-range chromatin contacts between Polycomb targets in mammals, it has recently been shown that the TAD structure within a specific locus, the X chromosome inactivation center, is not affected by the Eed knockout (Nora et al., 2012), indicating different regulatory mechanisms at the different hierarchies of genome organization. It is also interesting to speculate that the interaction between the Hoxd and Hoxc clusters may provide a mechanism for how the noncoding RNA HOTAIR, encoded within the Hoxc cluster, finds its target genes within the Hoxd cluster located on a different chromosome (Rinn et al., 2007): by exploiting 3D conformation of the genome in a manner similar to that employed by the long-noncoding RNA Xist (Engreitz et al., 2013).

Based on the data presented here and on other reports (Hakim et al., 2011; Noordermeer et al., 2011), we suggest that cell-type-specific gene-regulatory networks generally guide specific spatial interactions within the context of a cell-type-specific chromosome folding pattern that relates to the open/closed...
chromatin state. A preferential colocalization of distal genomic regions with similar regulatory networks tens of Mb apart on the same chromosome and in trans suggests a previously unappreciated role for transcriptional networks in influencing the 3D positional preferences of chromatin in mammalian cells, which may represent an evolutionarily conserved interaction between eukaryotic genome organization and gene regulation (Sexton et al., 2012; Tanizawa et al., 2010). We speculate that this organizational hierarchy facilitates the recruitment of regulatory proteins and potentially noncoding RNAs to their genomic target sites and the establishment of chromatin environments, which are both critical for the efficient regulation of gene expression.

EXPERIMENTAL PROCEDURES

4C-seq and 3C
4C-seq libraries listed in Table S1 were prepared as described (Splinter et al., 2012) using the primers given in Table S2 and sequenced on Illumina machines. Reads that aligned to unique HindIII sites in the genome (build mm9) with at most two mismatches were retained. Read distribution statistics for all libraries are given in Table S3. 3C libraries were prepared in an analogous manner to 4C libraries, ending with the first ligation step and amplified with primers listed in Table S2.

Data Analysis
For each 4C-seq library, read counts at each unique HindIII site were collapsed to a hit to reduce effects of clonal amplification. Replica libraries that passed stringent quality control were pooled for downstream analysis by calculating the probability of a hit at each HindIII site across all replicates. Next, we determined the average hit probability within 200 kb windows tiled along each chromosome for each pooled data set, referring to this as the hit percentage. A binomial test with different background models for cis and trans interactions was used to identify significantly interacting regions of each bait locus.

For PCA, 31 linear genomic features (summarized in Table S6) were transformed into five chromatin states and 11 TF clusters (Figures 3A and 3B), which were in turn aggregated within 200 kb windows across the genome, together with six unclustered features, to obtain a feature density matrix. To describe the linear genomic feature state of ESCs and MEFs, vectors containing feature density within 200 kb windows for each cell type were concatenated, allowing PCA to be conducted on the combined feature matrix (Table S6). To obtain a PC score enrichment value for each 4C bait, the mean PC score within the 200 kb bait window and the four flanking windows was calculated (i.e., five 200 kb bait windows = 1 Mb bait region). The PC score enrichment within the bait’s interactome was calculated as the mean PC score within 200 kb windows that overlapped 4C positive domains (as determined by the binomial test) by at least 25%, excluding the five bait windows. The rankings of the bait and interactome PC score enrichment values were correlated using Spearman’s rho statistic.

For the Hi-C data comparison, normalized, mouse ESC Hi-C interaction matrices based on 40 kb bins were downloaded from the Ren Lab website (Dixon et al., 2012) and rebinned into 200 kb bins to match the resolution of our 4C and feature data. Each 200 kb window of rebinned Hi-C data was treated as pseudo-bait and its intrachromosomal interactome was extracted from the chromosome-wide contact matrix. Bait and interactome PC scores were calculated as described above for 4C-seq interactomes, except that the interactome was defined as those 5% of 200 kb windows that had the highest read count, excluding the 1 Mb bait region.

FISH and Immunostaining Analysis
3D FISH and immunostaining were done following standard methods.

ACCESSION NUMBERS
Our 4C-seq data are available at GEO under GSE50029.
gene reprogramming events occur in the same spatial clusters of distal regulatory elements. Genome Res. 21, 697–706.


Supplemental Information

Long-range chromatin contacts in embryonic stem cells reveal a role for pluripotency factors and Polycomb proteins in genome organization

Matthew Denholtz *, Giancarlo Bonora*, Constantinos Chronis, Erik Splinter, Wouter de Laat, Jason Ernst, Matteo Pellegrini, and Kathrin Plath

Inventory

Supplemental Figures

Supplemental Figure 1
This figure is related to Figure 1 and provides further evidence of the reproducibility of 4C-seq data sets, quality control of 4C-seq data, and 3C validation of 4C-seq defined chromatin contacts.

Supplemental Figure 2
This figure is related to Figure 2 and shows that individual 4C-seq replicate data sets per cell type correlate strongly, revealing pluripotency-specific chromatin contacts in both cis and trans.

Supplemental Figure 3
This figure is related to Figures 3 and 4 and depicts the interacting domains for the bait regions analyzed in ESCs by 4C-seq in addition to the Pou5f1 bait, principal component (PC) score distributions, and additional Hi-C-based analysis.

Supplemental Figure 4
This figure is related to Figures 3 and 4 and provides a comparison of average PC1 score and average feature density within 4C-seq bait regions and their corresponding interactomes in ESCs, as well as the enrichment of genomic features and PC scores within ESC super enhancers.

Supplemental Figure 5
This figure is related to Figure 6 and shows individual feature contributions to the first principal component of the ESC+MEF PCA, as well as the 3C validation of ESC-specific chromatin interactions of the Dppa2 bait region detected by 4C-seq.

Supplemental Figure 6
This figure is related to Figure 7 and provides additional validation and characterization of interaction preferences of PC3-positive, Polycomb protein-enriched, and PC3-negative bait regions in ESCs with and without Eed.
Supplemental Tables

Supplemental Table 1
This table summarizes the bait regions by cell type for all 4C-seq libraries in this study (worksheet one). It also provides the genomic positions of all significantly interacting domains for all bait regions and cell types (worksheet two). This table relates to data in figures 1-4, 6-7, and S1-S6.

Supplemental Table 2
This table lists the 4C-seq and 3C primers used in the study (worksheets one and two, respectively). This table relates to data in figures 1-4, 6-7, and S1-S6 for 4C primers, and figures S1 and S5 for 3C primers.

Supplemental Table 3
This table contains read distribution and quality control statistics for all 4C-seq libraries that passed quality control measures and were used in this study. This table relates to data in figures 1-4, 6-7, and S1-S6.

Supplemental Table 4
This table contains binomial test results for all baits, including the number of replicates/simulations, -log_{10}(p-value) threshold, number of significant HindIII sites (true positives), the number of false positives, and the false discovery rate (FDR), in both cis and trans. This table relates to data in figures 1-4, 6-7, and S1-S6.

Supplemental Table 5
This table contains Spearman’s rho statistic values for the rank correlation of the 4C hit percentages determined in this study and published Hi-C read counts corresponding to each 4C bait (Dixon et al., 2012) shown in figures 3 and 4.

Supplemental Table 6
This table provides a summary of the linear genomic feature data sets used for data analysis shown in figures 3, 4, 6, 7, and S3-S6, as well as source information for each data set.

Supplemental Table 7
This table lists the empirical background model parameters for each cell type, including the intercept and slope for the segmented linear (0 – 1 Mb around the bait) and log-log (1 – 8 Mb from the bait) regression models (Supplemental Experimental Procedures) used to identify interacting domains and generate binomial test p-values shown in figures 1-4, 6, 7, and S1-S6.
Additional Supplemental Information

Supplemental Experimental Procedures
Detailed descriptions of all experimental methods and data analyses.

Supplemental References
Includes the references cited in the Supplemental Information.
Supplemental Figures and Legends
**Figure S1. Reproducibility, validation and quality control of 4C-seq data** (Related to Figure 1)

**A,** Integrative Genomics Viewer tracks demonstrating the reproducibility of *Pou5f1* 4C-seq hit percentage data for four biological and three technical replicate data sets. Biological replicates are designated with numbers, technical replicates with letters. The interacting domains, as identified by the binomial test for the pooled 4C-seq *Pou5f1* data, are given below the hit percentage tracks. 4C-seq experiments with genomic DNA and control libraries from unfixed ESCs follow, demonstrating a lack of significant interactions. 

**B,** To determine the reproducibility of our data in both *cis* and *trans,* *Pou5f1* replicates were pooled (blue) or partitioned into two groups (A and B, red and green, respectively) and interacting regions were determined for each set and are shown on the genome-wide plot. Yellow regions are those that were called as interacting domains in both A and B partitions, and blue regions were called as interacting domains in only the pooled data set. Jaccard similarity coefficients are noted for the overlap of significant domains between A and B, *cis* only, and *trans* only, and the significance of the overlap was determined by the hypergeometric test. The darker chromosome represents the *cis* chromosome.

**C,** 4C-seq experiments on pooled control libraries, here shown for the *Pou5f1* locus (*Pou5f1* Control) in ESCs, display no significant interactions genome-wide. The red mark denotes the bait locus, and the darker chromosome represents the *cis* chromosome.

**D,** Quality control (QC) metrics. All 4C-seq data sets used in this study were required to have at least 20% of HindIII sites covered by reads within 2Mb of the bait locus (x-axis), as well as at least 20% of their reads occurring in *cis* to the bait locus (y-axis), among other quality control metrics (see Experimental Procedures). Only libraries that passed these QC metrics were analyzed further and included in this study.

**E,** Read distribution as a function of distance from the bait locus for all data sets that passed QC. This analysis excludes the self-ligated and undigested products. Boxplots show the median value (line within box), and the first and third quartiles (lower and upper edges of boxes, respectively). Whiskers demarcate +/- 1.5 times the interquartile range.

**F,** 3C validation of *Pouf51* interactions in ESCs. (i) Schematic representation of the 3C experimental design (performed as described in (Miele et al., 2006)) to confirm interactions between the *Pou5f1* locus (primer A) and distal regions of the same chromosome (in *cis*), including the four interacting fragments (based on 4C-seq) (primers B-E) and an intervening non-interacting fragment (primer F). (ii) 3C PCR results for the setup described in (i) confirming the presence of ligation products resulting from the juxtaposition of the genomic regions (B-E) with A, and an absence of ligation products between A and the intervening genomic region (F). A H₂O control and a BAC-generated positive control are shown for each PCR product. The asterisks mark the respective PCR products.

**G,** UCSC ENCODE mm9 36mer-based mappability scores (ftp://encodeftp.cse.ucsc.edu/pipeline/mm9/wgEncodeMapabilitywgEncodeCrgMapabilityAlign36mer.bigWig) were compared between the interacting domains (determined based on our 4C-seq analysis pipeline) and the scores outside these domains. The graph displays the difference in the mean 36-mer-based mappability scores inside and outside of the intrachromosomal interaction domains for the indicated baits and cell types, as well as the average across all baits and cell types (labeled ‘MEAN’), demonstrating that our 4C-seq results are not biased by mappability.
### A

<table>
<thead>
<tr>
<th>Gene</th>
<th>POU5F1</th>
<th>DPPA2</th>
<th>DPPA3</th>
<th>PRSS22</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hoxa10</td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
</tr>
<tr>
<td>Krt13</td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
</tr>
<tr>
<td>Maoa</td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
</tr>
<tr>
<td>Rhbdd1</td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
</tr>
<tr>
<td>Nfla</td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
</tr>
<tr>
<td>Zfp42</td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
</tr>
<tr>
<td>17006P10Rik</td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
</tr>
<tr>
<td>Stk35</td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
</tr>
<tr>
<td>Ukgc</td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
</tr>
<tr>
<td>Pcdh19</td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
</tr>
<tr>
<td>Tbx5</td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
</tr>
<tr>
<td>Vegfc</td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
<td><img src="image" alt="Heatmap" /></td>
</tr>
</tbody>
</table>

**Correlation:** Spearman rho

### B

#### i

<table>
<thead>
<tr>
<th>All baits Hit%, trans</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Heatmap" /></td>
</tr>
</tbody>
</table>

#### ii

<table>
<thead>
<tr>
<th>All baits Jaccard similarity, trans</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Heatmap" /></td>
</tr>
</tbody>
</table>

---

Denholtz, Bonora et al. Figure S2
Figure S2. 4C-seq replicate data sets cluster by cell type, revealing pluripotency-specific chromatin contacts (Related to Figure 2)

A, Spearman rho correlation matrices based on the percentage of HindIII sites hit within 1000kb windows in cis for individual 4C replicates in wildtype ESCs (V6.5 line), iPSCs, pre-iPSCs, and MEFs for the noted baits. Note the preferential clustering of individual replicates by cell type, and the high correlations of the data between ESCs and iPSCs. Correlation values are indicated by the key. Numbers listed next to cell types correspond to biological replicates, while technical replicates are designated with letters.

B, (i) Unsupervised hierarchical clustering of Spearman rank correlation values of the hit percentages within 200kb windows along the trans chromosomes of eight different bait loci (Pou5f1, Stk35, 1700067P10Rik, Nfia, Dppa3, Rhbddd1, Hoxa10, and Dppa2) in wildtype ESCs (V6.5 line), MEFs, iPSCs, and pre-iPSCs, demonstrating the pluripotency-specific organization of interchromosomal chromatin contacts within the mouse genome. Color key defines Spearman rho values. (ii) Unsupervised hierarchical clustering of Jaccard similarity coefficients for the overlapping interacting domains in trans in ESCs, iPSCs, pre-iPSCs, and MEFs, for the same eight bait loci as (A). Color key defines Jaccard similarity values.
Figure S3. Genome-wide analysis of interactomes in ESCs (Related to Figures 3 and 4)

A, We obtained 4C-seq data for a total of 16 baits in the V6.5 ESC line. The interacting domains for the Pou5f1 4C-seq data are shown in Figure 1B, and those identified for the remaining 15 baits are depicted here, genome-wide, with their relative chromosomal locations. Blue marks represent significantly interacting windows, red marks the bait locus, and the darker chromosome denotes the cis chromosome.

B, Demonstration of the continuous nature of PC scores. (i) Box plot of genome-wide PC1 scores overlaid with a violin plot. (ii) As in (i), but for PC2. (iii) As in (i), but for PC3. Boxplots show the median value (line within box), and the first and third quartiles (lower and upper edges of boxes, respectively). Whiskers demarcate +/- 1.5 times the interquartile range.

C, Hi-C data analysis demonstrating that the preferential co-localization of genomic regions with similar genomic features identified by 4C-seq is a genome-wide phenomenon for interactions in cis. Pseudo-4C analysis was performed on Hi-C data as described in Figures 3F/G. (i) (left) The mean PC1 score within the top 5% of 200kb
windows (ranked by read count based on Hi-C data) along the cis-chromosome, excluding the 1Mb extended bait region, and the mean PC1 score within the respective 1Mb pseudo-bait region were determined, and plotted as red point on the scatterplot. If the bait region is one of the our baits analyzed by 4C-seq in Figure 3E-i, the data point was plotted in yellow (4C-bait loci). Correlations between bait and interactome PC1 scores are noted. The data are also summarized by the regression line in black, and the mean bait and interactome PC1 scores are demarcated by vertical and horizontal gray lines, respectively, and contour lines represent data density. The same data are shown in Figure 3G. (i) (right) Similar to the left plot, but an analysis was performed for the 5% of windows that are the least likely to interact based on Hi-C read count (bottom 5%) and their mean PC1 score was plotted in blue against the mean score of the 1Mb bait regions. Again, if the bait region is one of the our baits analyzed by 4C-seq in Figure 3E-i, the data point was plotted in yellow (4C-bait loci). (ii) As in (i), except that the analysis was performed for PC2 scores, and the red (left) scatterplot is repeated from Figure 4D. (iii) As in (i), except that the analysis was performed for PC3 scores, and the red (left) scatterplot is repeated from Figure 4H.

D, As in (C), except that all 200kb windows within the 10Mb region around the bait loci were excluded from the determination of the mean PC scores within the top and bottom 5% of 200kb windows along the cis-chromosome. Importantly, the correlation values between interactome and bait character, with regards to PC scores, were not significantly changed upon exclusion of the 10Mb region around the bait loci, or of increasingly larger proximal interactions around the bait (up to 25Mb on either side of the bait region, data not shown), indicating that even distal interactions follow the logic that regions with similar PC character interact.
Figure S4. Relationship between PC1 score and individual feature enrichment (Related to Figures 3 and 4)

A-D, Comparison of PC1 and feature scores between bait regions and their interactomes: A, Top to bottom: Mean PC1 score for the 1Mb region centered on each bait locus in ESCs; mean PC1 score within interacting regions in cis for each bait locus; mean PC1 score within non-interacting regions in cis for each bait locus; mean PC1 score within interacting regions in trans for each bait locus; mean PC1 score within non-interacting regions in trans for each bait locus. Spearman rho’s give the correlation between PC1 bait character and interactome character across all analyzed baits in cis and trans. These results are duplicated from Figure 3E-i for illustration. B, As in (A), but showing the average, standardized Oct4 enrichment within the bait regions, and their interacting and non-interacting regions. C, As in (B), but for H3K27me3 (which is
captured by ChromHMM state 1). **D**, As in (B), but for LaminB enrichment. The data shown in (A-D) suggest that, due to the strong enrichment of transcription factors, chromatin regulators, basic transcriptional machinery, Cohesin, and active and Polycomb-repressed chromatin states in genomic regions with high PC1 scores and their preferential co-localization in 3D space, their interactomes are also highly enriched for any of the features positively correlating with PC1. Indeed, the extent of Oct4 and H3K27me3 enrichment in a bait’s interacting regions mirrors its PC1 character. Conversely, LaminB occupancy, which correlates with closed, PC1-negative chromatin character, shows the opposite relationship. As such, the enrichment of features positively correlating with PC1 in a given bait’s interactome may not be indicative of specific mechanistic roles for each enriched feature but may be instead a consequence of the overarching nature of chromosomal conformation. **E**, Relationship of ESC super enhancers to the genomic features tested in our study. Enrichment of chromatin states and transcription factor clusters within and outside of ESC super-enhancers. **F**, Mean PC1, PC2, and PC3 scores within and outside of ESC super-enhancers.
Figure S5. Comparison of the spatial interactomes between ESCs and MEFs (Related to Figure 6)

A, PCA was performed on genomic features across ESCs and MEFs. Proportion of total variance in genomic features described by each principal component for the ESC+MEF PCA.

B, Genomic feature contribution to the ESC + MEF feature PC1 eigenvector.

C, Correlation of ESC genomic feature density with PC1 scores.

D, Correlation of MEF genomic feature density with PC1 scores.

E, (i) Schematic representation of the 3C experimental design to confirm the cell type-specificity of interactions with the Dppa2 locus. At the top, a portion of chromosome 16 harboring the Dppa2 locus is depicted, and a subset of interacting domains in ESCs and MEFs as defined by our 4C-seq analysis is indicated. Three ESC-specific interaction sites were examined (primers W-Y) by 3C in both MEFs and ESCs. A positive 3C control
was designed to amplify a ligation product between the *Dppa2* locus (primer Z) and a proximal restriction site (primer V), which should be detectable in both ESCs and MEFs. (ii) 3C PCR results confirming the presence and cell type-specificity of the ESC-specific ligation products Z/W, Z/X, and Z/Y. Note the presence of the positive control PCR product (Z+V) in both the ESC and MEF 3C experiments.
Figure S6. Additional validation and characterization of interaction preferences in Eed\textsuperscript{+/+} and Eed\textsuperscript{-/-} ESCs (Related to Figure 7)

A, Immunostaining for H3K27me3 (green) in wildtype (Eed\textsuperscript{+/+}) and Eed\textsuperscript{-/-} ESCs, grown on irradiated MEFs. DAPI staining in blue marks the nuclei. Note that Eed ablation leads to the loss of H3K27me3. Wildtype MEFs still stain positively for H3K27me3.

B. DNA FISH provides an independent confirmation of 4C-seq-defined differences in chromatin interactions between Eed\textsuperscript{+/+} and Eed\textsuperscript{-/-} ESCs. Cumulative distribution plots for interaction frequencies (y-axis) at different distances (x-axis) for the trans interactions measured between the Hoxb region and the individual Hox loci indicated in the figure (the sum of these individual plots is shown in Figure 7I-i), and, for comparison, between the Hoxb region and the Sox2 region, which does not interact with Hoxb based on our 4C-seq data (same FISH analysis plot as shown in Figure 7I-i). The data for wildtype ESCs are shown in blue, and for mutant ESCs in red.

C, Co-localization frequencies at 1um for each pair of interactions listed in (B). The aggregate of these individual plots between Hox genes is shown in Figure 7I-ii, and the Hoxb-Sox2 analysis plot is the same as that shown in Figure 7I-ii for comparison. n = FISH signal pairs analyzed for each cell type.

D, Top: Overlap of interacting domains of the Hoxb3 bait region between Eed\textsuperscript{+/+} and Eed\textsuperscript{-/-} ESCs genome-wide. Bottom: PC3 score distribution for the Eed\textsuperscript{+/+} and the Eed\textsuperscript{-/-} ESC-specific interacting domains. These data demonstrate that interacting domains that are specific for wildtype ESCs have significantly higher PC3 scores genome-wide than those that are specific for Eed\textsuperscript{-/-} ESCs. Boxplots show the median value (line within box), and the first and third quartiles (lower and upper edges of boxes, respectively). Whiskers demarcate +/- 1.5 times the interquartile range. Notches \( \approx \) 95% confidence interval around medians. ** p-value <= 0.01.

E, The cis (top) and trans (bottom) interactomes of the six PC3-positive (Polycomb/H3K27me3 enriched) target bait loci Hoxa10, Hoxb3, Hoxc4, Hox12, Pcdhb19, and Tbx5 were ranked by -log(p-value) for both Eed\textsuperscript{+/+} and Eed\textsuperscript{-/-} ESCs, and, in each case, the 500 top ranked sites were plotted against their average PC2 scores in wildtype ESCs, indicating minor changes in interaction preferences with regards to PC2 scores upon Eed ablation.

F, As in (D), except for the non-Polycomb, PC3-negative bait Pou5f1. Eed\textsuperscript{+/+} and Eed\textsuperscript{-/-} ESC specific interactions do not show significant differences in PC3 scores.

G-I, The cis and trans interactomes of the two PC3-negative target bait loci Pou5f1 and Ptprg were ranked by -log(p-value) for both Eed\textsuperscript{+/+} and Eed\textsuperscript{-/-} ESCs, and the 500 top ranked sites in cis (top) or trans (bottom) plotted against their average (G) PC3; (H) PC1; and (I) PC2 scores in wildtype ESCs. These data show that non-Polycomb targets do not show major changes in interaction preferences with regards to the PC’s. Kolmogorov-Smirnov test (KS) p-values shown with KS-test D value.
Supplemental Experimental Procedures

4C Library Preparation
4C libraries were prepared from mouse ESCs (V6.5 line), MEFs (wildtype of 129SvJae background), an iPSC line (described in Chin, Plath et al., manuscript in preparation), the pre-iPSCs 12-1 line (Sridharan et al., 2013), and the Eed mutant ESC line (17Rn5-3354SB) and a sibling wild-type ESC line (Morin-Kensicki et al., 2001), essentially as described (Splinter et al., 2012). Specifically, 10⁷ cells were trypsinized and filtered to single cell suspensions with 40um cell strainers. Following a PBS wash, cells were cross-linked in 1xPBS with 10% fetal bovine serum (FBS) and 2% formaldehyde for 10 minutes at 25°C while rocking. Ice-cold 1M glycine was added to the cells on ice to a final concentration of 0.13M to quench the crosslinking reaction. Cells were spun at 500g for 5 minutes at 4°C, resuspended in ice-cold lysis wash buffer (10mM Tris-HCl pH7.5, 10mM NaCl, 5mM MgCl₂, 0.1mM EGTA, and 1x protease inhibitors (Roche)), and re-pelleted at 4°C. Pellets were subsequently resuspended in 1ml ice-cold lysis buffer (lysis wash buffer with 0.2% IGEPAL (Sigma)), cells were lysed on ice for 30 minutes, and dounced using a tight piston for 10 strokes to isolate nuclei. Nuclei were spun down at 200g for 7 minutes at 4°C, washed with 1.2x buffer B (Roche), resuspended in 1ml ice-cold 1.2x buffer B, and transferred to non-stick tubes. Cells were brought to 37°C and 20ul 15% SDS were added to each tube, which were then incubated for 1 hour at 37°C while rotating end over end. 150ul 20% Triton X-100 was added and tubes were allowed to incubate for another hour at 37°C. 800U of high concentration HindIII (Roche, cat# 10798983001) were added to each tube and the restriction digest reaction was run overnight at 37°C while tubes rotated. The restriction enzyme was inactivated at 65°C for 25 minutes and digest efficiency was determined as described (Splinter et al., 2012). Digested samples were transferred to 50ml falcon tubes, and 5.3ml H₂O and 700ul 10x ligation buffer (660mM Tris-HCl pH 7.5, 50mM MgCl₂, 50mM DTT, 10mM ATP) were added to each sample. 100U T4 ligase (Roche cat#10799009001) were added and samples ligated overnight at 16°C. Ligation efficiency was checked as described (Splinter et al., 2012). 30ul 10mg/ml proteinase K were added to efficiently ligated samples and the samples were incubated overnight at 65°C. Subsequently, 30ul RNaseA (Invitrogen, cat# 12091-021) were added to each sample, and samples were incubated for a additional 45 minutes at 37°C. DNA was phenol-chloroform extracted, and precipitated by the addition of 7ml H₂O, 1ml 3M Na-acetate pH 5.6, 7ul glycogen
Precipitated DNA was spun down at 8800g for 45 minutes at 4°C, washed with ice-cold 70% ethanol, and re-spun at 3000g for 15 minutes at 4°C. Upon drying, DNA was resuspended in 150ul 10mM Tris pH 7.5, 300ul H₂O, 50ul 10x DpnII restriction buffer (NEB), and 50U high concentration DpnII (NEB, cat# R0543M) were added to each tube and the DNA was digested overnight at 37°C. DNA was phenol-chloroform purified, precipitated via addition of 50ul 3M Na-Acetate and 1ml ethanol, and re-dissolved in 100ul H₂O. The DpnII-digested DNA was then transferred to a falcon tube to which 12.5ml H₂O, 1.4ml 10x ligation buffer, and 200U high concentration T4 ligase were added for ligation overnight at 16°C. Following phenol-chloroform purification, samples were precipitated via the addition of 14ul glycogen and 35ml 100% ethanol at -80°C. The precipitated DNA was pelleted and washed as above, and resuspended in 150ul 10mM Tris-HCl pH 7.5. Residual salt was removed via Qiagen PCR purification columns.

4C library PCR amplification and Illumina high-throughput sequencing
Inverse PCR primers (Table S2) were designed to anneal to a bait locus HindIII/DpnII restriction fragment (selected with the criteria that it be longer than 300bp and be within 50kb of the indicated gene) and to amplify the unknown portion of the chimeric DNA circle generated during 4C library preparation. The resulting DNA circles consist of the bait locus restriction fragment and its interacting partner’s restriction fragments. The six 3’ nucleotides of the primers annealing to the HindIII side of the restriction fragment contained the HindIII restriction site when possible, or were generally within 4bp of the start of the HindIII site, to avoid uninformative reads upon sequencing. Primers on the DpnII side were allowed more positional flexibility, as sequencing data were not produced from the DpnII end of the restriction fragment.

200ng of DNA from the 4C library were used as template for the PCR amplification using the Expand Long Range PCR system (Roche). 5uM each of forward and reverse primer lacking Solexa sequencing adaptors were applied to amplify the interactome of interest in a 25ul reaction volume under the following PCR conditions: 1 cycle at 92°C for 2 minutes; (92°C 30 seconds; 58°C 1 minute; 68°C 1 minute) x 10 cycles; 1 cycle of 68°C 7 minutes. PCR products were run on an agarose gel, and amplicons between 100-500bp were isolated by gel extraction (Qiagen Gel purification system), and used as template for a second PCR reaction utilizing the same primers.
with the addition of the Solexa adaptors in a 50ul volume as follows: 1 cycle of 92°C for 2 minutes; (92°C 30 seconds; 58°C 1 minute; 68°C 1 minute) x 10 cycles; (92°C 30 seconds; 68°C 1 minute +20 seconds/additional cycle; 68°C 1 minute) x 15 cycles; and 1 cycle at 68°C for 7 minutes. The PCR-amplified library was purified over GFX PCR DNA purification kit columns (GE Healthcare) to remove primer dimers, followed by a second purification with Qiagen MinElute Reaction Cleanup kit (Qiagen) to remove residual salt. Samples were quantified using the Quant-iT dsDNA BR assay kit quantification system (Invitrogen) with a Qubit fluorometer. Purified, PCR-amplified 4C-seq libraries were pooled in EB (Qiagen), 0.1% Tween-20 for multiplexed sequencing as primer distinctiveness allowed, and sequenced at the Broad Stem Cell Research Center at UCLA.

4C sequencing and read mapping
Two to seven 4C libraries were multiplexed and sequenced using the Illumina Genome Analyzer II to obtain 76 base pair (bp) reads or Illumina Hi-seq-2000 to obtain 100 bp reads. Reads were parsed based on a unique, non-annealing, two base pair bar code and/or unique bait-specific primer sequences. The resultant reads were mapped to the mouse genome (build mm9) using Bowtie software (Langmead et al., 2009). Only reads that aligned to a unique position in the genome with no more than two sequence mismatches were retained for further analysis. Reads that were successfully aligned to the genome were then remapped to the 736,199 unique HindIII sites along the genome by matching their respective loci. Because we were concerned about any potential biases created by differential mappability, we excluded all HindIII sites that do not precede a unique 50bp along both DNA strands. In other words, we only mapped reads to HindIII sites that are unique in the mm9 genome with respect to the 100bp centered on the hexamers that comprise the sites, as illustrated below:

(+) AAGCTT--- 50bp ---
(−) --- 50bp ---TTCGAA

By only considering unique HindIII sites, we restricted our 4C analysis to highly mappable regions of the genome

Table S1 provides a summary of all the bait regions for which 4C-seq libraries were generated for all cell types discussed in this study. Read distributions and statistics
for all individual data sets that passed quality control steps (see below) are contained in Table S3.

4C hit determination
To reduce potential clonal amplification effects inherent to PCR-based genomic approaches, we collapsed the raw 4C-seq read count at each HindIII site down to a 'hit' if a read count threshold was met. The threshold was chosen so that at least 80% of all hits were intrachromosomal, i.e. at most 20% of our hits fell in trans (Figure S1E). For each library, the threshold and the number of sites that passed this threshold (i.e. the number of hits) are provided in Table S3. For the majority of libraries the read count threshold for calling hits was a single read.

4C library quality control (QC)
We used three criteria adapted from van de Werken et al. (2012) to estimate the quality of our individual 4C-seq libraries, as well as an additional two criteria of our own. First, we checked whether the library under consideration was comprised of at least 500,000 reads in total. Second, the cis/genome read count ratio (the proportion of total mapped reads in cis) had to be at least 20%. Third, at least 20% of HindIII sites within the 2Mb region around the bait had to be covered by at least one read. If a library passed all three of these de Laat group-inspired criteria, then it received one credit. Libraries received additional credit for passing each of the following two tests: 1) having a cis:trans hit ratio that was at least 4:1 (i.e. a maximum of 20% of thresholded hits could fall in trans for the library to be credited); and 2) having at least 1.5% of all sites along the cis chromosome covered by a hit. Hence, each library could achieve a maximum score of 3/3. Libraries that received a total score of 3/3 automatically passed QC, while those with 2/3 were subject to further scrutiny and only passed if they exhibited strong metrics for their two passing criteria. The libraries scoring less than 2/3 did not pass QC. Excluding three control libraries, 198/242 (84%) of the 4C-seq libraries passed the QC and were kept for further analysis (Figure S1D). Those libraries that passed QC are given in Table S3.

To ensure that those genomic regions within closed chromatin environments (negative PC1 scoring regions) were digested as efficiently as those regions in open chromatin surroundings, we required two things: 1) Our 4C-seq libraries had to exhibit at
least 20% HindIII site coverage within the 2Mb regions around the bait, as shown in Table S3 and Figure S1D. Notably for baits with negative PC1 scores, this demonstrates that closed chromatin is subject to proper digestion. 2) Regions proximal to the bait of high and low PC1 character had to show similar average hit probability (data not shown).

**Pooling of replicate 4C-seq libraries**

Replicate 4C-seq libraries for a single bait locus in a given cell type that passed QC (described above) were pooled by calculating the average number of times each site was called a hit (by attaining the read count threshold described above) in all replicates. In essence, we determined the probability of a hit at each HindIII site along the genome for each bait and cell type. Table S1 summarizes the bait regions for which pooled 4C-seq libraries were produced in each cell type, while Table S3 lists all the replicate libraries that were pooled, as well as statistics pertaining to hit probability. In this study we considered 66 pooled 4C-seq data sets (Table S1).

**Definition and calculation of 4C hit percentage**

In order to obtain a smoother continuous signal at a scale that was compatible with our genomic feature and PCA data (see below), we determined the average hit probability within 200 kilobase (kb) windows tiled along each chromosome, referring to this as the ‘hit percentage.’ We observed a strong correlation between hit percentage and binned Hi-C read counts using equivalent window sizes (Table S5).

**Binomial test analysis**

To demarcate positively interacting regions of a bait locus along each chromosome, we sought to identify statistically significant clusters of HindIII sites that exhibited a high probability of being hit across replicates. We used R’s binomial test function (Team, 2011) to calculate the probability of seeing the observed proportion of hits to HindIII sites, or ‘hit percentage’, within a 200kb window around each HindIII site along each chromosome, relative to the expected background hit percentage. In cis, the background hit proportion was obtained by modeling the average hit percentage as a function of distance from the bait locus across all data sets in a given cell line (see below for details on the background modeling), while in trans the average hit percentage across each respective trans chromosome was used as the expected hit percentage. Using the observed hit probability within the 200Kb surrounding each site, we determined the
number of hits that this represented, given the number of sites within the window: i.e. number of hits = hit probability * number of sites with 200Kb window. We used the resulting number of hits as the value for the binomial test parameter representing the number of successes, with the number of HindIII sites being the number of trials, and the hit percentage obtained from the empirical background model (for a locus at the given distance from the bait locus in cis), or the average hit percentage in trans, being the hypothesized probability of success. Only sites centered within windows containing at least ten HindIII sites were considered. A threshold of –log$_{10}$(p-value) >= 1.8 was used to determine HindIII sites centered within windows showing significant clusters of interaction. This threshold was used as it resulted in a small false discovery rate (FDR; see determination below) in cis, while allowing us to pick up significant trans interactions. The binomial test results are given in Table S4. Adjacent and overlapping positive windows were concatenated into 4C-positive domains. Table S1 catalogs these significantly-interacting domains (“interactomes”) for pooled libraries as determined by our 4C-seq analysis pipeline.

In order to determine an FDR for each of our 4C-seq interactomes, we generated corresponding data sets of simulated hit probabilities. For each bait, we generated as many data sets of simulated hits as we had experimental replicates for that bait. To simulate intrachromosomal hits we used the hit percentage specified by the empirical background model (described below) as the probability of sampling a hit at each site with respect to its distance from the bait. To simulate interchromosomal hits, the average hit probability for each chromosome in the experimental pooled data set was used as the probability of sampling a hit at each site along the chromosome. We pooled the resulting simulated replicates in the manner described above for our experimental replicates. Table S4 lists the number of significant windows and FDR in cis and trans for each of our 4C-seq data sets.

A 200kb window-size for the binomial test was chosen after having tested various window sizes and generated a multi-scale representation of the results (domainogram) to confirm the consistency of the binomial test p-values across window sizes (data not shown). This window size contained a sufficient number of HindIII sites (trials) to produce a robust binomial test result, while providing sub-megabase resolution.
To confirm that there was no strong bias for genomic regions with higher mappability within our libraries, illustrating that our selection of unique HindIII sites (as described above) had the desired effect of only interrogating highly mappable regions of the genome, we compared the average mappability within our 4C-positive regions (4C domains) to that within 4C-negative regions, for intra-chromosomal interactions. We downloaded the UCSC ENCODE mm9 mappability scores based on 36bp alignment (ftp://encodeftp.cse.ucsc.edu/pipeline/mm9/wgEncodeMapability/wgEncodeCrgMapabilityAlign36mer.bigWig), using the 36mer mappability data because it reflects the typical length of our 4C-seq reads after trimming of the primer prefixes. The 36mer-based mappability scores are also the most stringent of the available mappability scores. All intra-chromosomal interactomes produced in our study are shown in the bar plot of the average difference in mappability scores (mean score of 4C domains (inside) - mean score of outside 4C domains (outside)) for all 66 of our baits (Figure S1G). We found that the average difference in mappability score for all baits (labeled ‘MEAN’) is small, supporting the conclusion that our results are not significantly biased by mappability. A similar lack of mappability bias was observed when genome-wide interactions were examined (data not shown). This result is consistent with the fact that we only considered data at unique HindIII sites in the genome.

Intrachromosomal empirical background model

To obtain an estimate of the intrachromosomal expected probability of success for the binomial test, we used the average hit probability within 200kb windows from the bait locus across all baits to build a regression model of hit probability as a function of distance to the bait for each cell type. Adapting the approach used by Lieberman-Aiden et al. (Lieberman-Aiden et al., 2009) for their Hi-C data, we used the 1 – 8 Mb region proximal to the bait locus to fit a log-log regression model. The data in this region produced a much better fit than using the entire range of data points available (Table S7). We also noticed that within 1 Mb of the bait, the data deviated sharply from the power law scaling observed over the 1 – 8 Mb region, and that a linear regression model was more appropriate for this immediately proximal region around the bait locus. We therefore performed segmented regression analysis by partitioning the distance from the bait locus into the 0 – 1 Mb region and the 1 – 8 Mb region, using linear and log-log regression on these two regions, respectively. Model parameters and R-squared values
are provided in Table S7. Figure 1A gives an example of the resulting empirical background model generated for the ESC Pou5f1 pooled data set.

Data set correlation, overlap determination, and clustering
To compare interactomes across cell types, we performed Spearman’s rho correlations on hit probabilities within 200kb windows tiled across each chromosome, and calculated Jaccard similarity coefficients using binary vectors representing 200kb tiled windows that overlapped 4C-positive domains (Figures 2 and S2). Dendrograms were obtained by converting the Spearman’s rho statistics and Jaccard coefficients into distance measures and performing unsupervised hierarchical clustering using R’s hclust function (Team, 2011) using the Ward agglomeration method.

RNA-seq
Strand-specific RNA-seq from V6.5 ECSs and wildtype 129SvJae MEFs (Table S6) was performed essentially as described in Parkhomchuk et al (Parkhomchuk et al., 2009), using 4 ug of total RNA as starting material. Reads were mapped to the mouse genome (mm9) using TopHat software (Trapnell et al., 2009) and only those reads that aligned with no more than two sequence mismatches were retained.

ChIP-seq
All histone modification data used for this study (Table S6) were determined using native ChIP (Wagschal et al., 2007) and will be described in detail in a separate manuscript (Chronis et al., in preparation). Briefly, nuclei were isolated from non-crosslinked V6.5 ESCs and 129SvJae MEFs by centrifugation through a sucrose cushion (1.2M sucrose, 60mM KCL,15mM NaCl, 5mM MgCl₂, 0.1mM Tris-HCl, 0.5mM DTT and protease inhibitor cocktail). Nuclei were then resuspended in Mnase-digestion buffer (0.32M sucrose, 50mM Tris-HCl, 4mM MgCl₂, 1mM CaCl₂, and protease inhibitor cocktail) and digested with 3 units of MNase (Roche) for 10 minutes at 37°C. Soluble chromatin fractions were incubated with anti-H3K4me3 (Abcam; ab8580), anti-H3K4me2 (Abcam ab7766), anti-H3K4me1 (Abcam; ab8895), anti-H3K27me3 (Active Motif; 39155), anti-H3K27ac (Abcam; ab4729), and anti-H3K36me3 (Abcam; ab9050), respectively. Extracts were washed twice with wash buffer A (50mM Tris-HCl, 10mM EDTA, 75mM NaCl), wash buffer B (50mM Tris-HCl, 10mM EDTA, 125mM NaCl), wash buffer C
DNA extraction and library preparation as described (Wagschal et al., 2007).

Transcription factor binding data generated in this study (Table S6) were acquired using cross-linking ChIP and will be described in detail in a separate manuscript (Chronis et al., in preparation). V6.5 ESCs and 129SvJae MEFs were grown to a final concentration of 5x10^7 cells for each sequencing experiment. Cells were chemically cross-linked by the addition of formaldehyde to 1% final concentration for 10 minutes and quenched with glycine at a final concentration of 0.125 M. Cells were then resuspended in buffer I (0.3M sucrose, 60mM KCl, 15mM NaCl, 5mM MgCl_2, 10mM EGTA, 15mM Tris-HCl, 0.5mM DTT, 0.2% NP-40, and protease inhibitor cocktail), and incubated on ice for 10 minutes. Nuclei were generated by centrifugation in a sucrose cushion (1.2M sucrose, 60mM KCL, 15mM NaCl, 5mM MgCl_2, 0.1mM Tris-HCl, 0.5mM DTT, and protease inhibitor cocktail). Isolated nuclei were resuspended in sonication buffer (50mM Hepes, 140mM NaCl, 1mM EDTA, 1% TritonX-100, 0.1% Na-deoxycholate, 0.1% SDS), and sonicated using a Diagenode Bioruptor. Subsequently, nuclear extracts were incubated overnight at 4°C with one of the following antibodies: anti-Klf4 (R&D; AF3158), anti-Myc (R&D; AF3696), anti-Nanog (Cosmobio), anti-Oct4 (R&D; AF1759), anti-Sox2 (R&D AF2018), anti-p300 (SantaCruz; sc-585). Extracts were washed twice with RIPA, low salt buffer (20mM Tris pH 8.1, 150mM NaCl, 2mM EDTA, 1% Triton X-100, 0.1% SDS), high salt buffer (20mM Tris pH 8.1, 500mM NaCl, 2mM EDTA, 1% Triton X-100, 0.1% SDS), LiCl buffer (10mM Tris pH 8.1, 250mM LiCl, 1mM EDTA, 1% deoxycholate, 1% NP-40), and 1xTE. Reverse cross-linking was performed by overnight incubation at 65°C with 1% SDS and proteinase K. All protocols for Illumina/Solexa sequencing library preparation, sequencing, and quality control were performed as recommended by Illumina, with the minor modification of limiting the PCR amplification step to 10 cycles.

Reads were mapped to the mouse genome (mm9) using Bowtie software (Langmead et al., 2009) and only those reads that aligned to a unique position with no more than two sequence mismatches were retained for further analysis. Multiple reads mapping to the same location in the genome were collapsed to a single read to account for clonal amplification effects. ChIP-seq peaks were called using MACS software (Version 1.4.2) (Zhang et al., 2008) using a bandwidth parameter of 150 bp.
Chromatin states
Five chromatin states in ESCs and MEFs were identified at a resolution of 200bp as described by Ernst and Kellis (Ernst and Kellis, 2012) using the six histone modification ChIP-seq data sets listed in Table S6, including a ChIP input dataset.

Transcription factor clusters
The genome was tiled into 1kb windows and the presence of transcription factor (TF) peaks from sixteen in-house and previously published ChIP-seq data sets for ESCs were used to define the TF clusters used in this analysis (Figure 3A and Table S6). For published data sets, we used peaks determined by the authors of the respective studies. The “Cohesin” data set represents the merging of peaks from the Smc1 and Smc3 data sets, and the “PRC2” data set represents the merging of peaks from the Eed, Ezh2, and Suz12 datasets (Table S6). This procedure resulted in a vector of binary data for each TF reflecting its absence or presence within 1kb windows across the genome. The windows represented by these vectors were then clustered using R’s k-means function using the Hartigan-Wong method (Team, 2011) to obtain groups of windows exhibiting common combinatorial binding patterns across the genome. The number of centers (k=11) was chosen so as to substantially reduce the number of potential combinatorial TF groups ($2^{16}$-1), while ensuring that each cluster was represented by a significant number of windows (>7000 windows or ~0.25% of the genome). The eleven combinations of TFs found to co-bind within each window of a cluster are analogous in their combinatorial nature to the five chromatin states described above.

Principal Component Analysis
To compare our 4C interactome data to linear genomic features, including gene density, gene expression, replication timing, chromatin states, and transcription factor combinations in ESCs, we used Principal Component Analysis (PCA) to reduce the dimensionality of the 31 linear ESC genomic feature data sets (Table S6; discussed above and below). This allowed us to focus on weighted combinations of features, or principal components (PCs), that best characterized the genomic landscape of a cell type. Chromatin states and transcription factor clusters were used in order to capture the biologically important combinatorial nature of these features. The five described 200bp-based chromatin states and eleven 1 kb-resolution transcription factor cluster data were
binned into 200kb windows resulting in a semi-quantitative profile of feature density across the genome. Density profiles for DNase and LaminB were obtained in a similar manner by tallying the number of times they were present (at 1kb resolution) within 200kb windows. DNA replication timing data (Hiratani et al., 2010) were used to designate 1kb windows across the genome as either early (> 0.2) or late replicating (< - 0.2) (Table S6). Again, the number of 1kb windows positive for either of these two DNA replication timing states was tallied within larger 200kb windows to obtain vectors representing the density of early and late replicating regions of the genome, respectively. RNA-seq reads were binned in 200kb windows along the genome and the resulting read count totals were log-transformed to obtain a log-normal. For gene density profiles, counts of unique transcription start sites from the UCSC mm9 refGene table within 200kb windows across the genome were obtained. This preprocessing step resulted in a 200 kb-resolution, 13,283 x 22 (windows x features) ESC feature matrix. This matrix of genomic feature data was passed to R’s prcomp function (Team, 2011). Each column of data was scaled prior to performing the PCA. We used the top three PCs that best characterized the genomic landscape of ESCs for further analysis (PC1/2/3) (Figures 3 and 4).

We investigated whether mappability was particularly associated with any of the principal components considered in the study by including the average mappability within 200 Kb windows in the input matrix for our PCA of genomic features. We found that mappability does not contribute significantly to any of the top three principal components discussed in our study (data not shown).

To describe the linear genomic state of ESCs and MEFs, features available for both cell types were selected for PCA (Table S6). Vectors containing feature counts within 200kb windows for each cell type were concatenated, allowing PCA to be conducted on the combined genomic features. The resulting 26,566 x 15 (concatenated windows x common features) ESC+MEF feature matrix was passed to R’s prcomp function to conduct the PCA. Only PC1 was considered for the ES+MEF features (Figure 6A, S5A-D).

Bait versus interactome comparisons based on PC score enrichment
To obtain a PC score enrichment value for each 4C bait, the mean PC score within the 200kb bait window and the four flanking windows (for a total of five 200kb bait windows = 1Mb bait region) was calculated. The PC enrichment within the bait’s interactome was calculated in two ways: 1) as a Spearman’s rho statistic by correlating the vector of 4C hit percentages (average hit probabilities) within 200kb windows across the chromosome to the corresponding vector of PC scores; and 2) by calculating the mean PC score within 200kb windows that overlapped 4C positive domains (as determined by the binomial test) by at least 25%. The five bait windows were excluded from the interactome enrichment. The rankings of the bait and interactome PC score enrichment values were then correlated to obtain the Spearman rho statistics shown in Figures 3E, 4C, and 4G.

Curve fitting
We used R’s loess (Team, 2011) function to perform local fitting of a curve to PC scores as a function of -log$_{10}$(p-value) rank (Figure 7J-M, S6E, G-I), and subsequently R’s predict.loess function to predict a loess fit and estimated standard error for each predicted value. An estimated 95% confidence interval was obtained by drawing a band +/- 2 s.e. on either side of the fitted curve.

Hi-C comparisons
Mouse ESC, normalized, Hi-C interaction matrices based on 40kb bins were downloaded from the Ren Lab website (http://chromosome.sdsc.edu/mouse/hi-c/download.html) (Dixon et al., 2012) and re-binned into 200kb bins to match the resolution of our 4C data. The row/columns within the Hi-C interaction matrix were extracted.

The rebinned Hi-C data (as described above) were also used for conducting “pseudo-4C”, the Hi-C equivalent of the 4C-seq-based bait-interactome Spearman rank correlations in Figures 3E, 4C and 4G. Specifically, the row/columns within the Hi-C interaction matrix corresponding to each of the sixteen 4C baits analyzed in our ESC study were extracted. Bait PC score enrichments were calculated as described above for the 4C data sets by taking the mean PC score within the bait window and four flanking windows (five 200kb bait windows = 1Mb bait region). The mean PC score within the top and bottom 5% of windows based on read count in each row of the rebinned Hi-C
matrices were similarly calculated. To ensure that the interactome enrichment was not driven by the strong contacts centered on the bait locus, we excluded regions around the bait locus ranging in size from 1 to 50Mb to show that the enrichment was robust across the length of the chromosome (Figure S3C/D, and data not shown). Furthermore, the vector of Hi-C read counts rebinned within the 200kb windows, besides the 5 bait windows, was correlated to the PC scores across the respective chromosome to obtain an enrichment value in terms of the Spearman rank rho statistic (Figure 3H, Table S5).

For the genome-wide bait-interactome PC score enrichment correlations (Figure 3G, 4D/H, S3C/D), each 200kb window of rebinned Hi-C data was treated as a pseudo-bait. Bait and interactome PC score enrichment values were calculated as described above for specific baits.

**Fluorescence in situ hybridization (FISH)**

Cells were grown on glass coverslips, washed for 30 seconds with ice cold cytoskeletal (CSK) buffer (100mM NaCl, 300mM sucrose, 3mM MgCl₂, 10mM PIPES, pH6.8), 30 seconds with CSK buffer containing 0.5% Triton X-100, and again for 30 seconds with CSK buffer, fixed in 4% PFA in 1xPBS for 10 minutes at room temperature, transferred to 70% ethanol, and stored at 4°C. Cells were dehydrated through a series 5 minute incubations in ice-cold 85%, 95%, and 100% ethanol, rehydrated in 2xSSC for 5 minutes, incubated in 2xSSC with 100ug/ml RNaseA (Invitrogen) at 37°C for 30 minutes, and washed three times for 5 minutes each in 2xSSC. DNA was denatured for 20 minutes at 80°C in 2xSSC with 70% deionized formamide (CalBiochem), followed by immediate quenching in ice-cold 70% ethanol and a second dehydration series (performed as above). DNA FISH probes were denatured at 95°C for 5 minutes and allowed to pre-hybridize for 1 hour at 37°C before being added to dry slides. Probes were then allowed to hybridize with cellular DNA at 37°C for 16-48 hours in a humid chamber containing 50% formamide in 2x SSC. Following hybridization, cells were washed three times in each of the following solutions, pre-warmed to 42°C: 2x SSC/50% formamide, 2x SSC, and 1x SSC. The second 1x SSC wash contained 100 ng/ml DAPI to visualize nuclei. Slides were mounted in aquapolymount (Polysciences) and allowed to set overnight. FISH probes were generated from bacterial artificial chromosome (BAC) DNA (Figure 1c: Pou5f1 locus – RP23-213M12; A – RP23-98F21; B – RP23-106C23; C – RP23-85E24; D – RP23-2B8; Figure S6c/d: Hoxa - RP24-283F1; Hoxb - RP23-290I2; Hoxc -
RP23-473J19; Hoxd - RP24-398B4; Sox2 - RP23-2B8) by incorporation of fluorescently labeled nucleotides (Cy3-dCTP, Perkin Elmer and Alexa 488-dUTP, Invitrogen) via biopriming (Invitrogen).

3D-FISH image acquisition and analysis
3D images were constructed from a series 0.2um z-stacks through selected individual ESCs or ESC colonies. 3D-distance measurements between FISH signal centers were acquired using the Smart FISH3D plugin for ImageJ (Gue et al., 2005; Schneider et al., 2012 ). Distance distribution statistics were calculated in R (Team, 2011).

Immunofluorescence and image analysis
V6.5 mouse ESCs were grown on glass coverslips, washed for 30 seconds with ice cold cytoskeletal (CSK) buffer (100mM NaCl, 300mM sucrose, 3mM MgCl₂, 10mM PIPES, pH6.8), incubated for 1 hour with CSK buffer containing 0.7% Triton X-100, and washed again for 30 seconds with CSK buffer, fixed with 1xPBS/4% PFA for 10 minutes at room temperature, washed for 5 minutes in PBS/0.2% tween, incubated in blocking buffer (1x PBS, 5% goat serum, 0.2% fish skin gelatin, 0.2% tween) for 30 minutes, and incubated overnight in primary antibody diluted in blocking buffer at 4°C (anti-Nanog [eBioscience 14-5761-80]; anti-H3K27me3 [Active Motif 39155]; anti-RNAPII-S5 [Millipore 05-623]). Following primary antibody incubation cells were washed three times in PBS/0.2% tween and incubated in secondary antibody in blocking buffer at 25°C for 30 minutes. Cells were then washed three additional times in PBS/0.2% tween, with the second wash containing 100 ng/mL DAPI, and mounted in Polyaquamount.

Image acquisition was performed on a spinning disc confocal microscope. Line intensities in Figure 5 were determined using Slidebook software from 3i. Images were exported as tif files, and subsequent analysis was performed with customized R software (Team, 2011). Nuclear signal was identified via the removal of blue pixels whose intensities were below the 20th percentile of all blue pixel intensities for a given capture. Red and green signal outside of these preliminarily defined ‘nuclear’ signals were discarded. To remove nucleoli from the analysis ‘nuclear’ pixels with 0 values for either red or green pixels were similarly dropped from analysis. The remaining red and green pixels were quantile normalized, and the relationship between the remaining normalized red and green pixel positions and intensities were analyzed as shown in Figure 5.
3C

3C libraries were prepared in an analogous manner to 4C libraries, ending with the first ligation step. 3C primers are listed in Table S2, and PCRs were run with the following parameters: Primers A+B/C/D/E/F 95C 2 minutes, (95C 30 seconds, 60C 45 seconds, 72C 1 minute) x30, 72C 2 minutes, 1:5 dilution used as template for nested PCR with same parameters; Primers Z+V/X/Y/W - 95C 2 minutes, (95C 30 seconds, 68C 45 seconds, 72C 1 minute) x35, 72C 2 minutes.

Supplemental References


