Establishment of the Epigenetic Landscape in Mammalian Embryonic Stem Cells

by

Richard Patrick Koche

Submitted to the Harvard-MIT Division of Health Sciences and Technology
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2012

Author ....
Harvard-MIT Division of Health Sciences and Technology
July 13, 2012

Certified by ....
Bradley E. Bernstein, MD, PhD
Associate Professor, Harvard Medical School
Thesis Supervisor

Accepted by ....
Arup Chakraborty, PhD
Director, Institute for Medical Engineering and Sciences
Robert T. Haslam Professor of Chemical Engineering, Chemistry and Biological
Engineering, Massachusetts Institute of Technology

Establishment of the Epigenetic Landscape in Mammalian Embryonic Stem Cells

by

Richard Patrick Koche

Submitted to the Harvard-MIT Division of Health Sciences and Technology
on July 13, 2012, in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Abstract

Temporal and spatial variation of histone methylation is an important factor in mammalian development. Deciphering the details of such epigenetic phenomena has the potential to enrich both stem cell biology and therapeutics, as well as offer insight into various pathologies. While the enzymatic machinery responsible for these transitions is well known, it is their localization to specific genomic regions that controls cell fate, and this has largely remained a mystery. The goal of this thesis was to use an integrative genomics approach to elucidate the role of cis elements in the establishment of repressive chromatin domains. To this effect, we determined the genetic basis for localization of Polycomb repressive complexes (PRCs) in mammalian embryonic stem (ES) cells.

First, by generating genomewide chromatin state maps in mouse and human by high throughput sequencing, we utilized a comparative and motif dictionary approach to computationally identify potential Polycomb recruitment elements. Surprisingly, we found that PRC recruitment is best explained by localization to clusters of unmethylated CpG dinucleotides, elements originally associated with gene activation. Next, in a series of transgenic assays involving human and *E. coli* sequence, we were able to reconstitute the chromatin state of an epigenetic memory element in mouse ES cells. Finally, we found that as somatic identity is reset during induced pluripotent stem (iPS) cell reprogramming, these same elements are central to a coordinated response in which active chromatin domains are established prior to and independently of transcription.

Taken together, these studies highlight the role of a particular cis element in the establishment of both active and repressive chromatin domains. Furthermore, this dynamic underscores how a static genetic element can be utilized to enable the chromatin-based plasticity required of stem cell differentiation and lineage specification.

Thesis Supervisor: Bradley E. Bernstein, MD, PhD
Title: Associate Professor, Harvard Medical School

Acknowledgments

Arthur Schopenhauer said "All truth passes through three stages. First, it is ridiculed. Second, it is violently opposed. Third, it is accepted as self-evident." It is not often that one begins an endeavor and experiences all three of these stages by the time it is finished, but my thesis reflects these three stages to varying extents, and my survival in this process is owed to the courage and imagination of those with whom I worked in graduate school. First, I am deeply indebted to Brad Bernstein, whose enthusiasm, support, and vision (and demand for a diurnal schedule) forever altered my scientific career. Starting as a rotation project and blossoming into a full-fledged thesis, my work with Brad allowed me to gain experience that spanned from computational to technology development to more traditional bench science.

I am also deeply indebted to Alex Meissner, whose creativity and willingness to take risks opened up a whole new chapter in my graduate school life. Having met Alex when he was finishing graduate school himself, and watching him navigate to a faculty position, was inspirational, to say the least.

As for the work itself, I am forever grateful to my colleagues on the ground, in the trenches, the ones whose knowledge and skill allows us to test our ideas in the lab. The work in all phases of this thesis would not have been possible without the expertise, support, encouragement, and never-ending optimism from Manching Ku and Zack Smith, who provided me with a type of yin and yang that influenced me in and out of lab. I am also grateful for Shawn Gillespie, whose dedication and meticulous attention to detail allowed my scrawled diagrams to become working experiments. I am also indebted to Tarjei Mikkelsen, whose philosophy and style of computation will stay with me forever. For the endless discussions in and out of lab that make science so fun, I am thankful for Andrew Chi, Esther Rheinbay, Mario Suva, Mazhar Adli, Eric Mendenhall, Anthony Philippakis, and countless others in the Boston scientific community.

For the parts of graduate school where I was taken well beyond the lab, from stormy surf sessions in Rhode Island to bargaining for boat rides in Indonesia, I owe a lifetime of gratitude to Daniel Brady, Jonathan Gill, Lars Blackmore, and Yulee Newsome.

Finally, I am forever grateful for my family, whose support for me has never waned, and whose generosity, kindness, and dedication I hope is reflected in my own life and work.

Contents

Chapter 1

Introduction

Metazoan development requires cells to both proliferate and differentiate in a hierarchical fashion, with coordinated gene expression changes occurring over successive generations. The differentiation of mammalian embryonic stem (ES) cells poses a particular challenge, in which the plasticity of pluripotency must give rise to the subsequent synchronicity of development, all of which must be established from a more or less static genome. Epigenetic regulation is one important mechanism for coordinating such complex gene expression patterns. In addition to controlling the patterns of development, these same mechanisms have been implicated in the pathogenesis of various diseases. While the enzymatic components of this machinery have been well-studied, their recruitment mechanisms remain largely unknown. This thesis takes advantage of complementary biological and technological advances to uncover the role of cis regulatory elements in the establishment of repressive epigenetic landscapes early in development.

Chromatin modifications and epigenetic memory

Eukaryotic chromatin contains DNA wrapped around nucleosomes, each of which is composed of an octamer of histone proteins. Alterations of chromatin can control accessibility of the underlying DNA, activation or repression of genetic elements, recruitment of proteins, positioning of the genome within the nucleus, and formation of larger structures [1-3]. Such alterations can involve chemical modifications of the

DNA itself or the core histone proteins. Some modifications act through simple biophysical means, such as acetylated lysines neutralizing the positive charge of histones, thereby weakening the interaction with negatively charged DNA and causing a decondensation of chromatin. Other modifications, such as lysine methylation, can give rise to both activated and repressed domains, with each type eliciting a particular response and recruiting a unique set of proteins [4].

Chemical modifications of chromatin by Polycomb (PcG) and Trithorax group (TrxG) proteins allow for the maintenance of a repressive or active transcriptional state, respectively [5]. In particular, histone H3 lysine 27 trimethylation (H3K27me3), mediated by Polycomb repressive complex 2 (PRC2), is associated with transcriptional silencing [6], while histone H3 lysine 4 trimethylation (H3K4me3) promotes gene activation [7]. In *Drosophila*, where these complexes were originally discovered, both PcG and TrxG proteins are recruited to developmental loci early in embryogenesis [5]. Together they serve as a memory module in which a gene is silent but 'poised' for either transcriptional activation or repression. After reaching such a turning point, the transcriptional status is stored in the chromatin and 'remembered' throughout development [8].

While the above enzymatic components are well-studied, the means by which they are recruited and exactly how they elicit transcriptional changes remain largely unknown. However, it is clear that these complexes and their associated histone marks are part of a cascade of signals both up- and downstream. The chain of events must begin with their recruitment to the specific genomic regions at which they act. Given the lack of sequence specific binding factors in many of these complexes, this most likely involves as yet unidentified proteins and their cognate binding sites. The downstream mechanisms must account for the means by which effector proteins are recruited and the chromatin altered. This can include chromatin condensation, inhibition of transcriptional elongation, sequestration to the nuclear periphery, interruption of long-range interactions, and as yet undiscovered mechanisms [9-12]. However, it is the recruitment of these complexes that begins the cascading of signals essential for development and many pathologies, and it is here we focus our efforts.

Steering cell fate and oncogenesis

Modifications to chromatin can control activation, repression, and poising of genomic elements as varied as promoters, enhancers, locus control regions, and imprinted sites [3]. In turn, these elements determine which genes are or can be expressed in a given cell type. Thus, these mechanisms provide a critical means of establishing cellular memory and coordinating cell fate. This is of particular importance in stem and progenitor cells, in which different expression programs are enabled for different cell types. As such, perturbations in chromatin machinery affect events as early as gastrulation and as late as macrophage activation [13,14]. The pluripotent state is of particular interest, as its epigenome must contain the full potential for the activation or repression of genes necessary for the formation of all three germ layers, hence our focus on mammalian ES cells.

For each restriction placed on a cell undergoing lineage specification, this same restriction must be reversed in the processes of reprogramming and transdifferentiation. As such, the activity of chromatin modifiers and their histone marks is thought to play a prominent role. Induced pluripotent stem (iPS) cells are of particular importance for the study of cell state transitions, disease modeling, and regenerative therapeutics, and indeed, alterations of the epigenome are capable of blocking, facilitating, and even accelerating their formation [15-21]. Here, it is again the initiation of the epigenetic signaling cascade that remains largely unknown. Reprogramming through the induction of defined factors may prove to be an ideal closed system in which to explore the cause, effect, and timing of chromatin state transitions. Elucidating such transitions may have important implications for the epigenomic remodeling relevant to development and disease.

Parallels between the proliferation and plasticity of stem and tumor cells have raised the possibility that cancer is a stem cell disease, either by aberrant regulation of stem cells or a de-differentiation of lineage-committed cells [22]. While originally limited to hematologic malignancies, this stem cell model has since been implicated in a variety of solid tumors [23-26]. Importantly, further analyses of multiple cancer types illuminate similarities to ES cell epigenetic state, particularly PcG-mediated repression [27-34]. Reconstitution of an ES cell-like chromatin pattern may help to lock cells into a highly proliferative and de-differentiated state. Deciphering the means by which this epigenetic machinery is localized to a given region may help explain its aberrant recruitment in tumorigenesis. Furthermore, this lends excitement to the possibility of novel therapeutics since, in theory, such epigenetic transitions are reversible [35,36]. Indeed, small molecule inhibitors capable of reversing chromatin state are being actively explored in the clinic, with some initial reports of successful reversion of malignancies [37-39]. Pinpointing the exact nature of these epigenetic transitions, starting with recruitment and initiation, may hint at underlying disease processes which can help guide targeted therapeutics.

Barriers to deciphering Polycomb recruitment

Despite extensive involvement in mammalian development and cancer, the mechanism of recruitment of these chromatin modifying complexes remains elusive. In *Drosophila*, the canonical model involves initial recruitment through Polycomb response elements (PREs), cis-regulatory elements found within or near targeted genes [40]. A functional PRE consists of clusters of short binding sites, which are bound by proteins that then recruit both PcG and TrxG complexes early in development.

Identifying PRE elements in vertebrates has proven difficult for several reasons. First, vertebrates lack close homologs of the majority of DNA binding proteins known to direct PcGs in *Drosophila* [5]. Additionally, the differential localization of PRC1 and PRC2 may lead to the utilization of entirely different recruitment modalities [41]. Also, given the larger genomes of mammals, PREs may be further upstream or downstream from the target genes. Finally, a lack of data systematically characterizing mammalian epigenetic states has prevented the elucidation of putative PREs in the underlying sequences.

The lack of available information has not lead to a shortcoming of hypotheses for recruitment. Proposals hitherto have involved everything from recruitment via a complex histone code to non-coding RNAs (in both cis and trans) to utilization of the main *Drosophila* homolog (YY1), as well as suggestions that the multitude of different

Polycomb sites will involve a multitude of different recruitment mechanisms [42-47]. While some obvious discrepancies and contradictions exist amongst the proposed models, it is possible that one or more are correct. However, before entertaining more complex hypotheses, various genomic elements should be explored as potential recruitment sites, in analogy with the *Drosophila* model. As such, any attempt at defining the mammalian PRE should start with more accurate chromatin state maps and a thorough dissection of motifs within the underlying DNA sequence.

Technology, opportunity, and momentum

Several recent opportunities to discover such sequence elements have presented themselves. Through a combination of chromatin immunoprecipitation (ChIP) and microarrays, overlapping patterns of H3K4me3 and H3K27me3 were discovered in mouse ES cells [48]. These sites were termed 'bivalent domains' and resembled the initial recruitment of PcG and TrxG at PREs in *Drosophila*. Importantly, such sites were markedly enriched for genes controlling development and differentiation, as occurs in *Drosophila*, suggestive of evolutionarily conserved paradigms. Furthermore, recent advances in ChIP followed by high-throughput sequencing (ChIP-seq) have allowed for unprecedented genome-wide mapping of multiple histone modifications and PcG proteins in both mouse and human ES cells [49].

Importantly, these data also revealed an association between PRC2 localization (and that of the associated modification, H3K27me3) with specific genomic features, including highly conserved non-coding elements (HCNEs), CpG islands, and transposon exclusion zones (TEZs) [48]. In addition to these new epigenetic data, there are also more sequenced mammalian genomes, increasing the power of comparative genomics in identifying conserved and potentially functional cis-regulatory elements [50]. Moreover, previously unexplored protein-DNA binding specificities are being documented with recent advances in protein binding microarrays [51]. Finally, computational and functional genomics tools continue to extend our ability to understand complex biological processes.

I aimed to take a computational approach to investigate sequence features that may be responsible for establishing the initial epigenetic state in mouse ES cells. Specifically, the approach took advantage of genome-wide chromatin maps acquired through recently developed ChIP-seq methodology. After the identification of repressed and active chromatin domains, these data sets were used to identify overrepresented motifs that are evolutionarily conserved as well as motif clusters that may serve as templates for the recruitment of chromatin modifiers. Predictions were then subjected to both experimental as well as further in silico validation in the context of pluripotency and lineage specification. More specifically, I utilized a transgenic recruitment system in which both putative and synthetic PREs were tested for their ability to enable de novo recruitment of PcG proteins.

Taken together, several unique intersections of biological and technological advances were integrated to explore the establishment of epigenetic landscapes and, more specifically, to define the sequence-based mechanisms that underlie the recruitment of PcG proteins in mammalian ES cells.

Overview of thesis

Chapter 2: Genomewide Analysis of PRC1 and PRC2 Occupancy Identifies Two Classes of Bivalent Domains

Epigenetic repression is essential to the organization of many developmental and pathological gene expression hierarchies. While all PcG complexes are associated with gene repression, they vary substantially in their subunit composition, histone modifications, and modes of repression. This study aimed to address the structure and function of these discrepancies by mapping the associated histone modifications and core subunits in mouse and human ES cells.

The resultant chromatin state maps and comparative genomic analyses were used to draw two novel conclusions. First, the dominant repressive chromatin state in ES cells - the bivalent domain - was found to actually exist in two discrete PcG states: one with and one without PRC1. The sites containing PRC1 were functionally distinct, as they were more likely to function as repressors of developmental regulators and also showed higher conservation at orthologous loci in the human-mouse comparisons. The second finding provided fundamental insight into what may constitute a mammalian PRE, elements that are key to deciphering how epigenetic repression is initially established early in development. Importantly, the resulting model contained a simplicity that was hitherto lacking in competing proposals: the simplest explanation for PcG localization was an affinity for large, unmethylated CpG islands that lack activating motifs for a given cell type. The model's predictive power was cross-analyzed and confirmed in human ES cells, but the experimental evidence for PcG recruitment was left an open question.

Chapter 3: GC-Rich Sequence Elements Recruit PRC2 in Mammalian ES Cells

The computational model of Chapter 2 hinted at a completely novel mode of PcG recruitment, but was lacking in data to corroborate it. Here we utilized a transgenic bacterial artificial chromosome (BAC) system to test candidate recruitment elements for PcG localization. Initially, several putative PREs already containing PcG enrichment in human were selected via the computational model and subsequently inserted into mouse ES cells. ChIP-qPCR confirmed that the sequences were sufficient to recruit PcG proteins based on sequence alone. Next, the candidate elements were then dissected, demonstrating that the CpG island component was indeed necessary for PcG recruitment and chromatin modification.

Since these elements already contained PcG proteins in human ES cells, a better test of the model involved creating PcG recruitment sites where none had previously existed. This was done in two ways: deletion of motifs in a constitutively active CpG island, and utilization of GC-rich stretches of DNA from the *E. coli* genome. In both cases, the computational model correctly predicted the de novo recruitment of PcG proteins to our 'synthetic' PREs. These findings, in combination with Chapter 2, provide a cohesive model for the establishment of epigenetic repression early in development. Importantly, as this same type of repression is important for carcinogenesis, our model may offer insight into the aberrant silencing of such loci in cancer progression.

Chapter 4: Reprogramming Factor Expression Initiates Widespread Targeted Chromatin Remodeling

The epigenetic activation and repression associated with lineage specification must be reversed as somatic cells reprogram toward a less differentiated, more developmentally potent state. While chromatin state and gene expression dynamics have been described for the process as a whole, the initiation of these events remained a mystery. Here we designed a novel system that allowed for stage-specific cellular states in early factor-induced reprogramming to be tracked in a time- and cell cycle-dependent manner. At each point we collected data for gene expression, histone methylation, and DNA methylation.

Transcriptional dynamics were limited to sites already containing euchromatin, and the few activated genes appeared to be driven via Myc. Instead, unexpectedly, the dominant response upon factor induction was a coordinated, genome-wide increase in the euchromatic histone mark H3K4me2. This preceded transcriptional activation, and occurred regardless of whether the final reprogrammed state was active or repressed. Instead, the initiation of H3K4 methylation appeared to correlate well with the presence of CpG islands, as well as binding sites for Oct4 and Sox2. Sites of DNA methylation were refractory to histone methylation dynamics. This study helped identify an initiating epigenetic event as a means to which cellular identity is reset during reprogramming.

Chapter 5: Conclusions and Perspectives

This chapter places my work in a historical context, and identifies gaps in the literature which are now filled in by the conclusions reached in this thesis. It also attempts to place these studies within the context of a more coherent model of cis element-based epigenetic regulation. Lastly, I outline the implications of this thesis, and how they may pave the road for future studies involving epigenetic transitions in development and disease.

References

1. Bernstein BE, Meissner A, Lander ES (2007) The mammalian epigenome. Cell 128: 669-681.

2. Li G, Reinberg D (2011) Chromatin higher-order structures and gene regulation. Curr Opin Genet Dev 21: 175-186.

3. Zhou VW, Goren A, Bernstein BE (2011) Charting histone modifications and the functional organization of mammalian genomes. Nat Rev Genet 12: 7-18.

4. Dambacher S, Hahn M, Schotta G (2010) Epigenetic regulation of development by histone lysine methylation. Heredity

5. Ringrose L, Paro R (2007) Polycomb/Trithorax response elements and epigenetic memory of cell identity. Development 134: 223-232.

6. Simon JA, Kingston RE (2009) Mechanisms of Polycomb gene silencing: knowns and unknowns. Nat Rev Mol Cell Biol

7. Schuettengruber B, Martinez AM, Iovino N, Cavalli G (2011) Trithorax group proteins: switching genes on and keeping them active. Nat Rev Mol Cell Biol 12: 799-814.

8. Margueron R, Reinberg D (2010) Chromatin structure and the inheritance of epigenetic information. Nat Rev Genet 11: 285-296.

9. Francis NJ, Kingston RE, Woodcock CL (2004) Chromatin compaction by a polycomb group protein complex. Science 306: 1574-1577.

10. Stock JK, Giadrossi S, Casanova M, Brookes E, Vidal M et al. (2007) Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells. Nat Cell Biol 9: 1428-1435.

11. Zullo JM, Demarco IA, Pique-Regi R, Gaffney DJ, Epstein CB et al. (2012)

DNA sequence-dependent compartmentalization and silencing of chromatin at the nuclear lamina. Cell 149: 1474-1487.

12. Tiwari VK, McGarvey KM, Licchesi JD, Ohm JE, Herman JG et al. (2008) PcG proteins, DNA methylation, and gene repression by chromatin looping. PLoS Biol 6: 2911-2927.

13. Meissner A (2010) Epigenetic modifications in pluripotent and differentiated cells. Nat Biotechnol 28: 1079-1088.

14. De Santa F, Totaro MG, Prosperini E, Notarbartolo S, Testa G et al. (2007) The histone H3 lysine-27 demethylase Jmjd3 links inflammation to inhibition of polycomb-mediated gene silencing. Cell 130: 1083-1094.

15. Takahashi K, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126: 663-676.

16. Stadtfeld M, Hochedlinger K (2010) Induced pluripotency: history, mechanisms, and applications. Genes Dev 24: 2239-2263.

17. Plath K, Lowry WE (2011) Progress in understanding reprogramming to the induced pluripotent state. Nat Rev Genet 12: 253-265.

18. Mikkelsen TS, Hanna J, Zhang X, Ku M, Wernig M et al. (2008) Dissecting direct reprogramming through integrative genomic analysis. Nature 454: 49-55.

19. Onder TT, Kara N, Cherry A, Sinha AU, Zhu N et al. (2012) Chromatin-modifying enzymes as modulators of reprogramming. Nature 483: 598-602.

20. Ang YS, Tsai SY, Lee DF, Monk J, Su J et al. (2011) Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network. Cell 145: 183-197.

21. Singhal N, Graumann J, Wu G, Arauzo-Bravo MJ, Han DW et al. (2010) Chromatin-Remodeling Components of the BAF Complex Facilitate Reprogramming. Cell 141: 943-955.

22. Shackleton M, Quintana E, Fearon ER, Morrison SJ (2009) Heterogeneity in cancer: cancer stem cells versus clonal evolution. Cell 138: 822-829.

23. Ben-Porath I, Thomson MW, Carey VJ, Ge R, Bell GW et al. (2008) An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. Nat Genet 40: 499-507.

24. Stingl J, Caldas C (2007) Molecular heterogeneity of breast carcinomas and the cancer stem cell hypothesis. Nat Rev Cancer 7: 791-799.

25. Driessens G, Beck B, Caauwe A, Simons BD, Blanpain C (2012) Defining the mode of tumour growth by clonal analysis. Nature 488: 527-530.

26. Chen J, Li Y, Yu TS, McKay RM, Burns DK et al. (2012) A restricted cell population propagates glioblastoma growth after chemotherapy. Nature 488: 522-526.

27. Widschwendter M, Fiegl H, Egle D, Mueller-Holzner E, Spizzo G et al. (2007) Epigenetic stem cell signature in cancer. Nat Genet 39: 157-158.

28. Schlesinger Y, Straussman R, Keshet I, Farkash S, Hecht M et al. (2007) Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. Nat Genet 39: 232-236.

29. Gal-Yam EN, Egger G, Iniguez L, Holster H, Einarsson S et al. (2008) Frequent switching of Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line. Proc Natl Acad Sci U S A 105: 12979-12984.

30. Mills AA (2010) Throwing the cancer switch: reciprocal roles of polycomb and trithorax proteins. Nat Rev Cancer 10: 669-682.

31. Wilson BG, Wang X, Shen X, McKenna ES, Lemieux ME et al. (2010) Epigenetic antagonism between polycomb and SWI/SNF complexes during oncogenic transformation. Cancer Cell 18: 316-328.

32. Iliopoulos D, Lindahl-Allen M, Polytarchou C, Hirsch HA, Tsichlis PN et al. (2010) Loss of miR-200 inhibition of Suz12 leads to polycomb-mediated repression required for the formation and maintenance of cancer stem cells. Mol Cell 39: 761-772.

33. Sauvageau M, Sauvageau G (2010) Polycomb group proteins: multi-faceted regulators of somatic stem cells and cancer. Cell Stem Cell 7: 299-313.

34. Ryan RJ, Bernstein BE (2012) Molecular biology. Genetic events that shape the cancer epigenome. Science 336: 1513-1514.

35. Mosammaparast N, Shi Y (2010) Reversal of histone methylation: biochemical and molecular mechanisms of histone demethylases. Annu Rev Biochem 79: 155-179.

36. Arrowsmith CH, Bountra C, Fish PV, Lee K, Schapira M (2012) Epigenetic protein families: a new frontier for drug discovery. Nat Rev Drug Discov 11: 384-400.

37. Sharma SV, Lee DY, Li B, Quinlan MP, Takahashi F et al. (2010) A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. Cell 141: 69-80.

38. Filippakopoulos P, Qi J, Picaud S, Shen Y, Smith WB et al. (2010) Selective inhibition of BET bromodomains. Nature 468: 1067-1073.

39. Delmore JE, Issa GC, Lemieux ME, Rahl PB, Shi J et al. (2011) BET bromodomain inhibition as a therapeutic strategy to target c-Myc. Cell 146: 904-917.

40. Ringrose L, Rehmsmeier M, Dura JM, Paro R (2003) Genome-wide prediction of Polycomb/Trithorax response elements in Drosophila melanogaster. Dev Cell 5: 759-771.

41. Schoeftner S, Sengupta AK, Kubicek S, Mechtler K, Spahn L et al. (2006) Recruitment of PRC1 function at the initiation of X inactivation independent of PRC2 and silencing. EMBO J 25: 3110-3122.

42. Xu C, Bian C, Yang W, Galka M, Ouyang H et al. (2010) Binding of different histone marks differentially regulates the activity and specificity of polycomb repressive complex 2 (PRC2). Proc Natl Acad Sci U S A 107: 19266-19271.

43. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X et al. (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129: 1311-1323.

44. Zhao J, Sun BK, Erwin JA, Song JJ, Lee JT (2008) Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science 322: 750-756.

45. Garcia E, Marcos-Gutierrez C, del Mar Lorente M, Moreno JC, Vidal M (1999) RYBP, a new repressor protein that interacts with components of the mammalian Polycomb complex, and with the transcription factor YY1. EMBO J 18: 3404-3418.

46. Kim SY, Paylor SW, Magnuson T, Schumacher A (2006) Juxtaposed Polycomb complexes co-regulate vertebral identity. Development 133: 4957-4968.

47. Ringrose L, Paro R (2004) Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. Annu Rev Genet 38: 413-443.

48. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125: 315-326.

49. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553-560.

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

51. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PWr et al. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol 24: 1429-1435.

Chapter 2

Genomewide Analysis of PRC1 and PRC2 Occupancy Identifies Two Classes of Bivalent Domains

This work was originally published as:

*Ku M, *Koche RP, *Rheinbay E, Mendenhall EM, Endoh M, Mikkelsen TS, Presser A, Nusbaum C, Xie X, Chi AS, Adli M, Kasif S, Ptaszek LM, Cowan CA, Lander ES, Koseki H and Bernstein BE. (2008) Genomewide Analysis of PRC1 and PRC2 Occupancy Identifies Two Classes of Bivalent Domains. PLoS Genet 4(10): e1000242. doi:10.1371/journal.pgen.1000242

Contributions:

Conceived and designed the experiments: MK, RPK, BEB. Performed the experiments: MK, RPK, ER, ME, MA. Analyzed the data: MK, RPK, ER, EMM, TSM, AP, XX, BEB. Contributed reagents/materials/analysis tools: ME, TSM, CN, ASC, SK, LMP, CAC, ESL, HK. Wrote the paper: MK, RPK, ER, BEB.

Abstract

In embryonic stem (ES) cells, bivalent chromatin domains with overlapping repressive (H3 lysine 27 tri-methylation) and activating (H3 lysine 4 tri-methylation) histone modifications mark the promoters of more than 2000 genes. To gain insight into the structure and function of bivalent domains, we mapped key histone modifications and subunits of Polycomb repressive complexes 1 and 2 (PRC1 and PRC2) genomewide in human and mouse ES cells by chromatin immunoprecipitation followed by ultra high-throughput sequencing. We find that bivalent domains can be segregated into two classes: the first occupied by both PRC2 and PRC1 (PRC1-positive) and the second specifically bound by PRC2 (PRC2-only). PRC1-positive bivalent domains appear functionally distinct as they more efficiently retain lysine 27 tri-methylation upon differentiation, show stringent conservation of chromatin state, and associate with an overwhelming number of developmental regulator gene promoters. We also used computational genomics to search for sequence determinants of Polycomb binding. This analysis revealed that the genomewide locations of PRC2 and PRC1 can be largely predicted from the locations, sizes and underlying motif contents of CpG islands. We propose that large CpG islands depleted of activating motifs confer epigenetic memory by recruiting the full repertoire of Polycomb complexes in pluripotent cells.

Introduction

Increasing evidence suggests that Polycomb- (PcG) and trithorax-group (trxG) proteins and associated histone modifications are critical for the plasticity of the pluripotent state, for the dynamic changes in gene expression that accompany ES cell differentiation, and for subsequent maintenance of lineage-specific gene expression programs [1-4]. PcG proteins are transcriptional repressors that function by modulating chromatin structure [2-4]. They reside in two main complexes, termed Polycomb repressive complexes 1 and 2 (PRC1 and PRC2). PRC2 contains Ezh2, which catalyzes histone H3 lysine 27 tri-methylation (H3K27me3), as well as Eed and Suz12. PRC1 contains Ring1, an E3 ubiquitin ligase that mono-ubiquitinylates histone H2A at lysine 119 (H2Aub1) [5,6]. Other PRC1 components include Bmi1, Mel-18, and Cbx family proteins with affinity for H3K27me3 [2,3].

Interplay between PcG complexes and modified histones has been proposed to mediate stable transcriptional repression [2,3]. In the prevailing model, PRC2 is recruited to specific genomic locations where it catalyzes H3K27me3. The modified histones in turn recruit PRC1, which catalyzes H2Aub1 and thereby impedes RNA polymerase II elongation [7,8]. PRC1 may also affect PRC2 function through as yet undefined mechanisms [2,3].

Several groups have combined chromatin immunoprecipitation (ChIP) with microarrays to examine the genomic localizations of individual PcG subunits [9-13]. Lee et al used tiling arrays to map the PRC2 subunit Suz12 in human ES cells, identifying nearly 2000 gene targets. Boyer et al used promoter arrays to identify 512 genes co-occupied by PRC2 and PRC1 components in mouse ES cells. In both studies, the implicated gene sets were highly enriched for developmental transcription factors (TFs), many of which become de-repressed upon ES cell differentiation or in a PRC2-deficient background.

Concurrent studies of histone methylation in ES cells led to the unexpected finding that virtually all sites of PcG activity not only carry the repressive H3K27me3 modification, but are also strongly enriched for the activating, trxG-associated H3 lysine 4 tri-methylation (H3K4me3) mark [14,15]. Genomic regions with the two opposing modifications were termed 'bivalent domains' and proposed to silence developmental regulators while keeping them 'poised' for alternate fates. Upon ES cell differentiation, most bivalent promoters resolve to a 'univalent' state. Induced genes become further enriched for H3K4me3 and lose H3K27me3, while many non-induced genes retain H3K27me3 but lose H3K4me3 [15,16].

Despite this progress, our understanding of PcG regulation and bivalent domains remains limited. In the current study we sought to address two outstanding issues. The first relates to whether all bivalent domains have the same regulatory structure. The recent observation that human and mouse ES cells show overlapping H3K27me3 and H3K4me3 at over 2000 promoters, only a portion of which have developmental functions, suggests that bivalent domains may reflect multiple, distinct regulatory entities [16-18]. The second relates to the mechanisms that underlie the targeting of PcG complexes and the establishment of bivalent domains in ES cells. In *Drosophila*, PcG complexes are recruited to DNA elements termed Polycomb response elements (PREs). However, mammalian equivalents of these elements have yet to be identified [4].

We addressed these outstanding issues through genomewide analysis of PcG complex localization in mouse and human ES cells. We used the newly developed ChIP-Seq method, which leverages ultra high-throughput sequencing to generate uniquely comprehensive maps of protein-DNA interactions [16,19].

The data reveal two classes of bivalent domains with distinct regulatory properties. The first class corresponds to bivalent domains with both PRC2 and PRC1. These PRC1-positive bivalent domains show striking evolutionary conservation, correspond to large H3K27me3 regions in ES cells that are significantly more likely to retain H3K27me3 upon differentiation, and account for a vast majority of implicated developmental regulator genes. By contrast, PRC1-negative bivalent domains, which are exclusively bound by PRC2, are weakly conserved, poorly retain H3K27me3, and largely correspond to membrane proteins or genes with unknown functions. Remarkably, computational genomic analysis of the ChIP-Seq data suggests a simple genomic code in which the locations, sizes and motif contents of CpG islands may predict the genomewide localizations of PRC2, PRC1 and bivalent domains in ES cells. Based on these data, we propose a model in which large CpG islands depleted of activating transcription factor motifs confer epigenetic memory elements through mammalian development by recruiting PRC2 and PRC1 during early embryogenesis.

Results

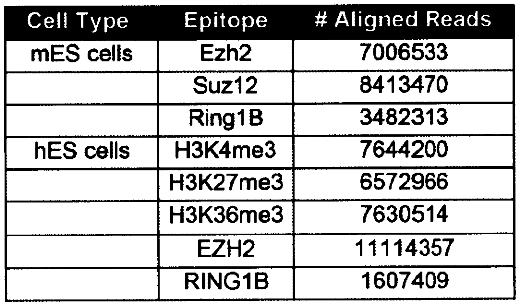
Overview of ChIP-Seq datasets

To gain insight into the structure, function and conservation of bivalent chromatin, we used ChIP-Seq to acquire genomewide maps of PcG complex components and related histone modifications in ES cells (Table 1). Chromatin from mouse v6.5 ES cells or human H9 ES cells was immunoprecipitated using antibodies against Ezh2, Suz12, Ring1B, H3K4me3, H3K27me3 or H3K36me3 (Materials and Methods). We also used biotin-streptavidin interaction (bioChIP) to purify chromatin from a transgenic mouse ES line in which endogenous Ring1B is fused to biotin ligase recognition peptide. DNA isolated in each ChIP experiment was sequenced to high depth using the Illumina Genome Analyzer. Aligned reads were integrated into maps that indicate enrichment of a given epitope as a function of genome position. In total, we created eight genomewide maps that each reflects two to eleven million aligned reads and together represent over 2 Gb of sequence. All data are publicly available at http://www.broad.mit.edu/seq_platform/chip/.

Evolutionary conservation of chromatin state in ES cells

The availability of genomewide data for mouse and human ES cells acquired using identical antibodies and methodologies provides an opportunity to study the conservation of chromatin state in pluripotent cells. We systematically compared chromatin state at 13,200 orthologous promoters, identifying striking similarities at orthologous genomic loci (Figure 1, Figure 2A).

In both mouse and human ES cells, roughly three-quarters of gene promoters are marked by H3K4me3. There is strong correspondence between species as >94% of promoters with H3K4me3 in mouse also carry H3K4me3 in human. Roughly one fifth of H3K4me3 promoters also carry H3K27me3, and thus are bivalent (mouse: n=2978; human: n=2529) (Figure 1C). There is again strong conservation, with more than half of bivalent mouse promoters also carrying bivalent chromatin in human ES cells (Fig 1A and Figure 2B). As shown previously, many bivalent mouse promoters correspond to homeobox TFs or other developmental regulators [14,15]. These gene categories show particularly strong conservation of chromatin state, with roughly 70% correspondence between mouse and human. Still, there are numerous developmental regulators whose chromatin state differs between species (Table 2). Closer inspection of these genes reveals a number of interesting cases that appear to reflect biological differences between the two pluripotency models:



Cell Type	Epitope	# Aligned Reads
mES cells	Ezh2	7006533
	Suz12	8413470
	Ring1B	3482313
hES cells	H3K4me3	7644200
	H3K27me3	6572966
	H3K36me3	7630514
	EZH2	11114357
	RING1B	1607409

Table 1: List of ChIP-Seq datasets and number of aligned reads. mES cells are from genotype 129SVJae x C57BL/6 F1 mice; hES cells are H9.

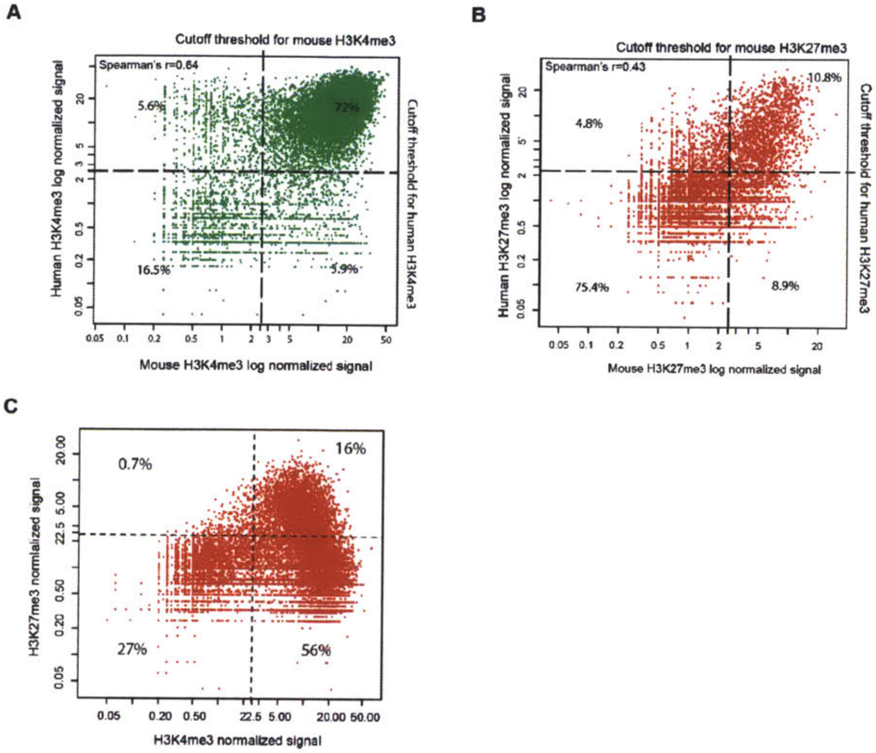


Figure 1: Comparison of chromatin states in mouse and human ES cells. **(A)** Conservation of H3K4me3 for 13,200 transcription start sites between human and mouse. Dashed lines indicate cutoff thresholds used to binarize the data for further analysis. Genes that carry H3K4me3 are likely to be conserved (upper right quadrant), as are those that are not marked (lower left quadrant). Less than 12% of genes are differentially methylated between human and mouse (upper left and lower right quadrants). **(B)** Conservation of H3K27me3 for the same regions used in (A). Most genes in both mouse and human are not marked with H3K27me3 (bottom left quadrant). Only slightly more than half the genes that carry H3K27me3 in mouse do so in human also. (upper and lower right quadrant). **(C)** H3K4me3 vs. H3K27me3 plotted for 17,760 mouse genes reveal three prominent marks in ESC: H3K4me3 only, (lower right quadrant), H3K4me3+H3K27me3/bivalent (upper right quadrant) and "no mark" (lower left quadrant). Very few genes are marked with H3K27me3 only (upper left quadrant).

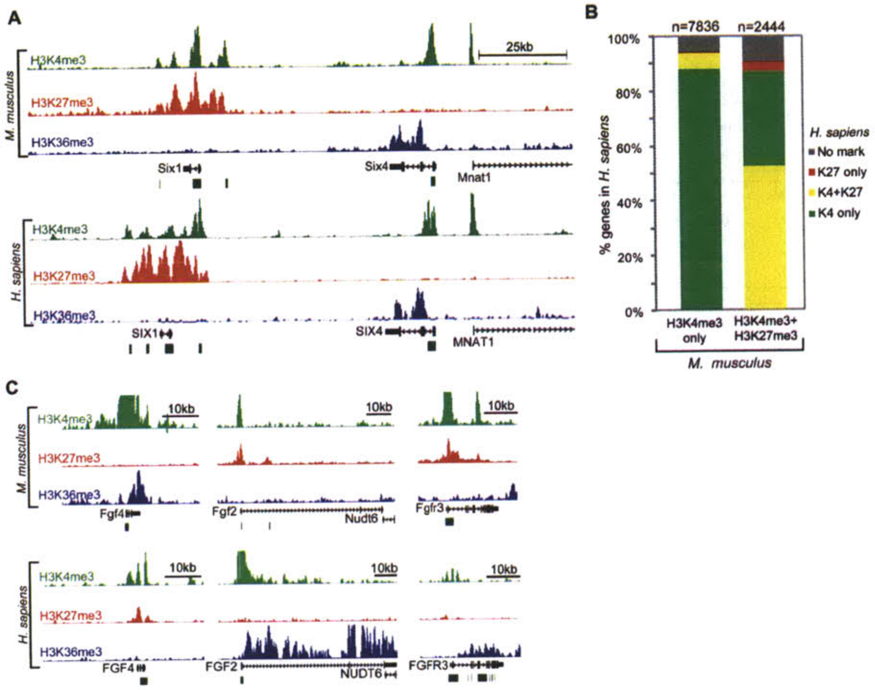
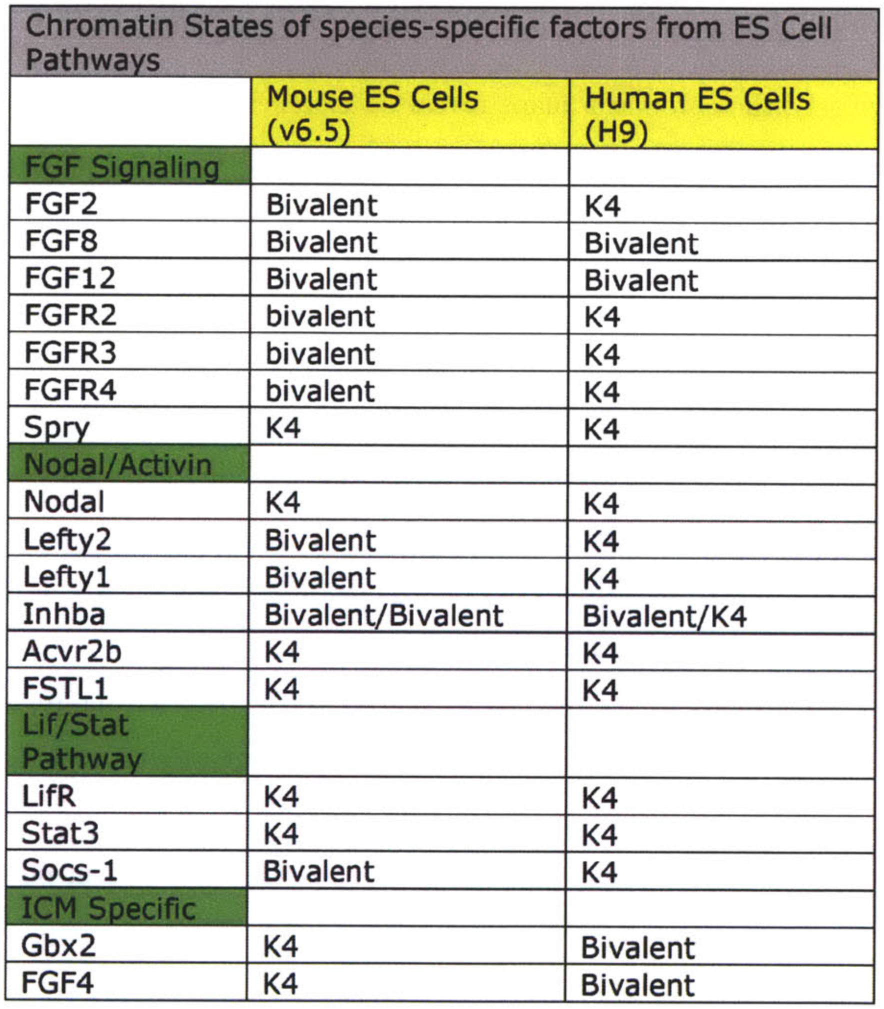


Figure 2: Conservation of chromatin state in mouse and human ES cells. **(A)** ChIP-Seq signals for H3K4me3 (green), H3K27me3 (red) and H3K36me3 (blue) are plotted across 120 kb of orthologous sequence in mouse and human ES cells. **(B)** The proportion of promoters that have a given chromatin state in human ES cells is indicated contingent on their state in mouse ES cells. **(C)** ChIP-Seq signals are shown for developmental regulator loci with divergent chromatin state in mouse and human ES cells. The divergent states correspond to known differences between the two pluripotency models (see text).



Chromatin States of species-specific factors from ES Cell Pathways		
	Mouse ES Cells (v6.5)	Human ES Cells (H9)
FGF Signaling		
FGF2	Bivalent	K4
FGF8	Bivalent	Bivalent
FGF12	Bivalent	Bivalent
FGFR2	bivalent	K4
FGFR3	bivalent	K4
FGFR4	bivalent	K4
Spry	K4	K4
Nodal/Activin		
Nodal	K4	K4
Lefty2	Bivalent	K4
Lefty1	Bivalent	K4
Inhba	Bivalent/Bivalent	Bivalent/K4
Acvr2b	K4	K4
FSTL1	K4	K4
Lif/Stat Pathway		
LifR	K4	K4
Stat3	K4	K4
Socs-1	Bivalent	K4
ICM Specific		
Gbx2	K4	Bivalent
FGF4	K4	Bivalent

Table 2: Divergent chromatin states of species-specific factors in transcription and signaling pathways observed in mouse and human ES cells reflect known distinctive biological functions between the two pluripotency models.

1. The promoters of Fgf2, Fgfr3, Activin A, Lefty1 and Lefty2 are bivalent in mouse ES cells but show active 'H3K4me3 only' states in human (Fig 2C). This is consistent with known expression patterns for these genes, which are associated with the human ES cell-specific Activin/NODAL pathway [20-22]. Another example is SOCS1, an inhibitor of STAT3 signaling that is specifically expressed in human ES cells where it may block response to LIF [23].

2. Conversely, the chromatin maps reveal developmental regulators that are bivalent only in human ES cells, and these may also relate to known physiologic differences between the models (Fig 2C). Examples include Fgf4 and Gbx2, which are associated with the inner cell mass and specifically expressed in mouse ES cells [20,24,25].

Thus, comparative analysis of human and mouse ES cells suggests extensive conservation of the pluripotent chromatin state while also illuminating divergent chromatin regulation associated with signaling pathways and transcriptional programs known to vary between the studied cell models (see also Table 2). The strong conservation of bivalent domains seen here contrasts with the surprisingly weak correspondence observed previously for Oct4 and Nanog targets between mouse and human ES cells [26]. Consistent with prior studies, our data suggest that global patterns of H3K27me3 and H3K4me3 are intimately tied to transcriptional programs and cellular state, and that the bivalent combination is a conserved mark of silent developmental regulators in pluripotent cells.

PcG complex occupancy defines two classes of bivalent domains

PRC2 occupies essentially all bivalent domains: To gain insight into the establishment and function of bivalent domains, we next considered the localization of PcG complexes in mouse ES cells. ChIP-Seq maps for the PRC2-components Ezh2 and Suz12 reveal >3000 sites in the mouse genome significantly enriched for one or both factors. Roughly three-quarters of these PRC2 bound sites correspond to known gene promoters: Ezh2 occupies 2461 promoters, while Suz12 occupies 1944 promoters. There is extensive overlap between these sets of promoters, with more than 89%

of Suz12 targets also having Ezh2 (rphi = 0.77). There is also overwhelming overlap with bivalent promoters: nearly all Suz12 and Ezh2 targets have bivalent histone markings and, conversely, 78% of bivalent promoters have Ezh2 or Suz12 (Figure 3A,C).

Since PRC2 is the only known complex capable of catalyzing H3K27me3 [2], we considered the minority (22%) of bivalent promoters for which PRC2 was not detected by ChIP-Seq. Many of these promoters show relatively low levels of H3K27me3, and we considered whether PRC2 was simply missed due to sensitivity or thresholding issues. Consistent with this possibility, ChIP with quantitative real-time PCR (qPCR) confirmed modest but significant Ezh2 enrichment at each of these promoters (ratios from 2- to 7-fold; Figure 4). This suggests that PRC2 is present at essentially all bivalent promoters. Notably, the correspondence between H3K27me3 and PRC2 is not limited to annotated gene promoters, as near-universal PRC2 binding is also evident at the roughly 1000 sites of bivalent chromatin that do not correspond to known genes (see Materials and Methods).

PRC1 occupies a conserved subset of bivalent domains: We next turned to examine PRC1 localization, focusing on its catalytic component Ring1B. ChIP-Seq maps reveal roughly 1500 significantly enriched genomic sites in mouse ES cells, including 1308 annotated gene promoters. Nearly all (90%) Ring1B targets correspond to bivalent, PRC2-bound genomic regions. However, just 39% of bivalent promoters are enriched for Ring1B (Figure 3B,C). This occupancy rate is roughly half that observed for Ezh2. As an added measure, we created an Ezh2 ChIP-Seq dataset with exactly the same number of reads as the Ring1B dataset (by randomly selecting reads). Analysis of this truncated dataset reveals Ezh2 binding at 74% of bivalent promoters (compare to 75% for the full Ezh2 ChIP-Seq dataset). Hence, sequencing depth does not account for the difference between Ezh2 and Ring1B occupancy.

Thus, ChIP-Seq analysis suggests that while PRC2 is ubiquitously present at bivalent promoters, PRC1 occupies only a distinct subset. Since PRC2 and PRC1 have generally been described at common genes and loci [9,10], we sought to confirm this unexpected result by orthogonal approaches, as follows:

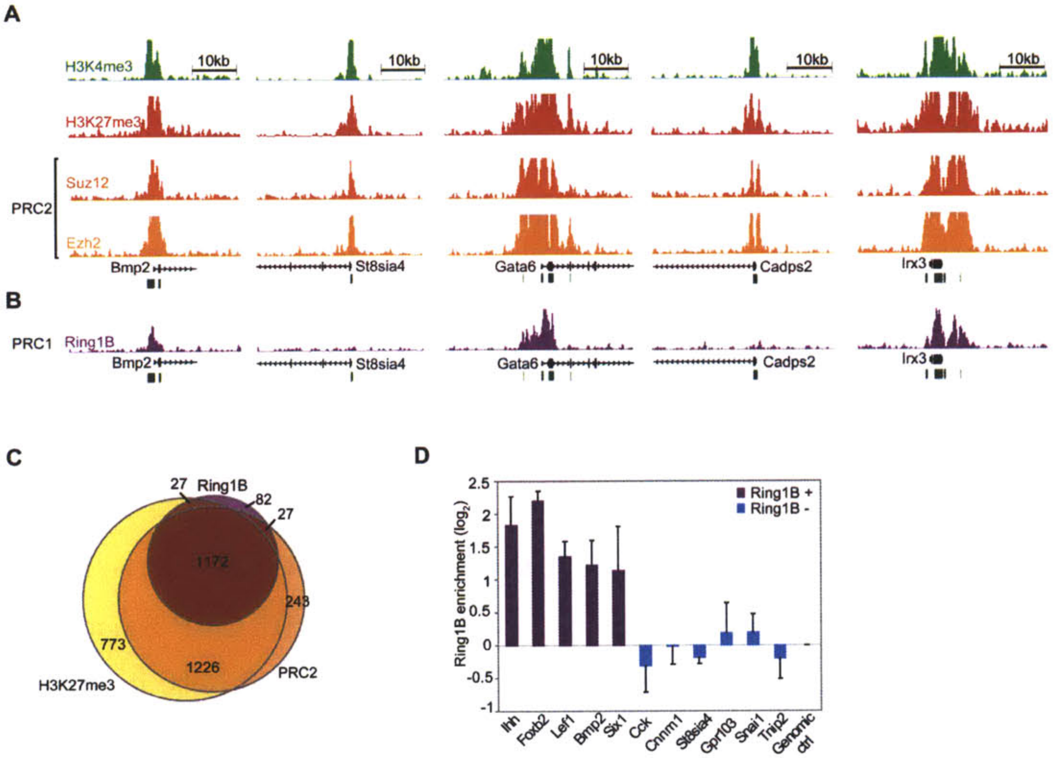


Figure 3: PcG complex occupancy at bivalent domains. **(A)** ChIP-Seq signals are shown for H3K4me3, H3K27me3 and PRC2 subunits, Suz12 and Ezh2, at a representative panel of bivalent gene promoters. **(B)** ChIP-Seq signal for the PRC1 subunit Ring1B at these loci. **(C)** Venn diagram illustrating overlap between promoters marked by H3K27me3, PRC2 and Ring1B. **(D)** ChIP-qPCR data for Ring1B at bivalent promoters classified by ChIP-Seq as Ring1B-positive or Ring1B-negative. Error bars show standard deviation.

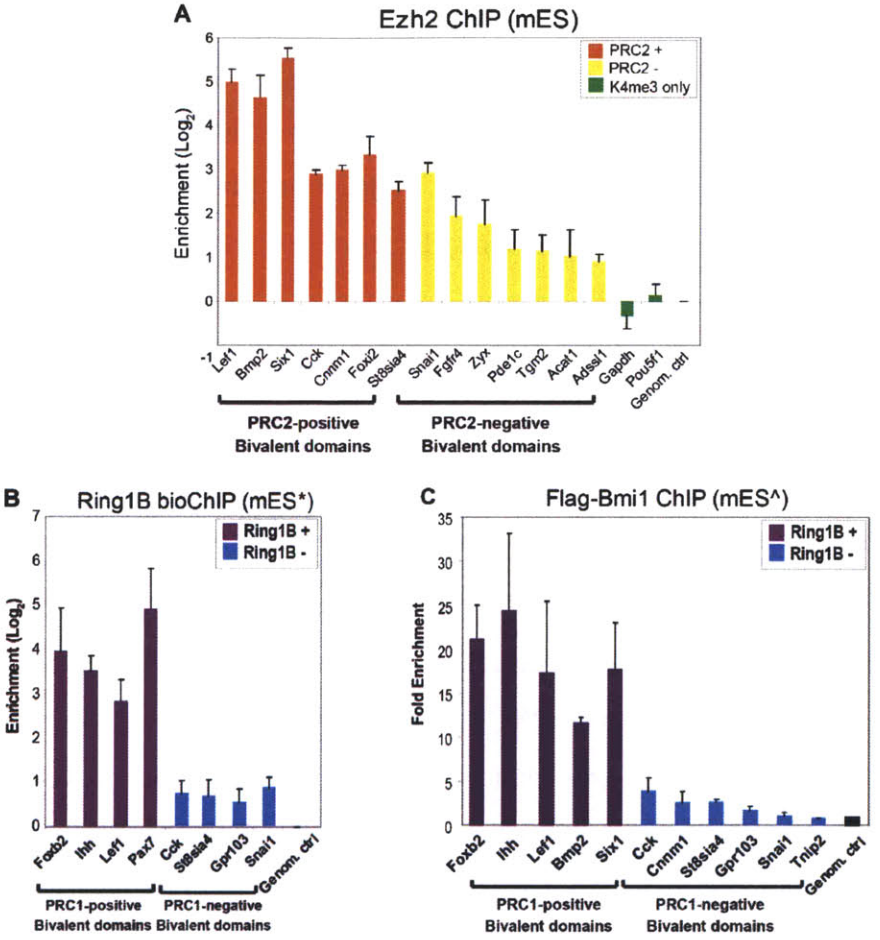


Figure 4: Quantitative PCR enrichment for Ezh2 ChIP, Ring1B bioChIP and Flag-Bmi1 ChIP. **(A)** Plot shows Log2 ChIP-qPCR enrichment of Ezh2 in mouse v6.5 ES cells at bivalent gene promoters. Included are promoters classified as PRC2-bound (orange) or PRC2-unbound (yellow) by ChIP-Seq. **(B)** Plot shows Log2 enrichment of Ring1B bioChIP-qPCR in transgenic mouse ES cells expressing biotin-tagged Ring1B (mES*) at bivalent promoters classified by ChIP-Seq as PRC1-bound (purple) or PRC1-unbound (blue). H3K4me3 only genes are green. **(C)** Plot shows fold enrichment of Flag ChIP-qPCR in transgenic mouse ES cells expressing Flag-tagged Bmi1 (mES$\hat{)}$ at bivalent promoters classified by ChIP-Seq as PRC1-bound (purple) or PRC1-unbound (blue).

(i) First, we used ChIP and qPCR to exclude the possibility that the absence of Ring1B at a subset of bivalent promoters reflected a lack of sensitivity of the ChIP-Seq data. This analysis confirmed that Ring1B-negative bivalent promoters also do not show any enrichment by qPCR (Figure 3D).

(ii) Next, to rule out antibody-related bias, we used bioChIP to purify Ring1B-bound chromatin from transgenic ES cells carrying a fusion between Ring1B and biotin ligase recognition peptide (Figure 4B). Ring1B-positive bivalent promoters again showed consistent enrichment, while Ring1B-negative bivalent promoters showed similar enrichment to background controls.

(iii) Third, to test whether the existence of Ring1B-positive and negative bivalent domains is a conserved phenomenon, we examined Ring1B occupancy in human ES cells by ChIP-Seq. We again found that Ring1B occupies only a subset of bivalent domains. The locations of PRC1 show remarkable cross-species conservation: 60% of Ring1B-positive promoters in human are also Ring1B-positive in mouse.

(iv) Finally, to confirm that Ring1B status is reflective of PRC1 status, we studied the localization of a distinct PRC1 component, Bmi1. Using an epitope tagged construct in ES cells, we showed that Bmi1 specifically localizes to Ring1B-positive bivalent domains (Figure 4C). This suggests that our findings on Ring1B generally apply to the PRC1 complex. Henceforth, the two sets of bivalent domains are notated as PRC1-positive and PRC1-negative.

PRC1-bound bivalent domains are functionally distinct

The identification of a distinct set of bivalent promoters targeted by Ring1B prompted us to investigate the functional significance of PRC1 occupancy. We made several striking observations relevant to chromatin regulation, epigenetic memory, development and differentiation:

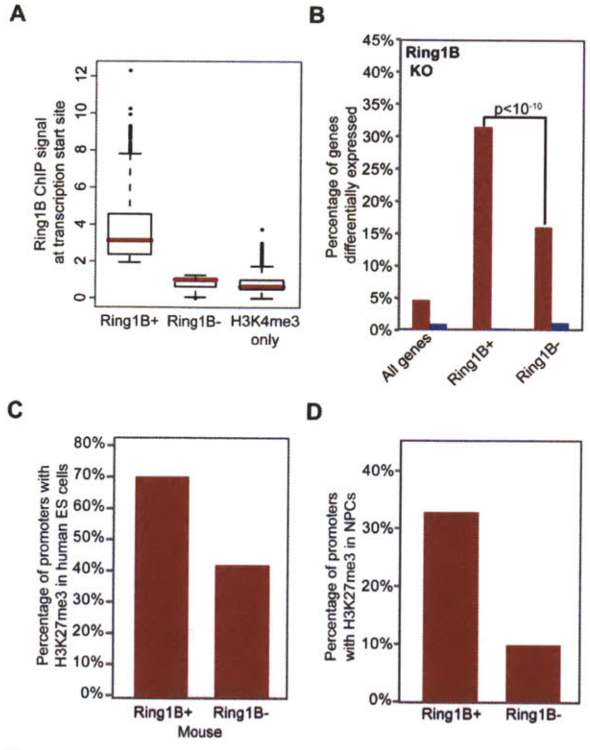
PRC1 occupancy correlates with functional repression: We first considered whether physical targets of PRC1, as defined above, are also regulated by the complex. Since Ring1B and Ring1A are functionally redundant, we employed a conditional Ring1A/B double-knockout ES cell system in which Ring1B depletion is induced by addition of

4-hydroxy tamoxifen (OHT) [13]. We profiled expression changes after 48 hours of OHT treatment, at which time Ring1B protein levels are markedly depleted while Oct4 levels remain essentially unchanged [8,13]. We found that 32% of PRC1-positive bivalent promoters are up-regulated by at least 50%, compared to just 5% of all genes (Figure 5B). A much smaller proportion of PRC1-negative bivalent promoters are up-regulated at this time point (16%). The difference between the two sets is statistically significant ($p < 10^{-10}$), and is not explained by baseline expression levels as bivalent promoters show very low activity, regardless of PRC1 status.

Several factors could contribute to de-repression of this smaller set of PRC1-negative bivalent promoters. The changes may reflect indirect effects as expression is measured after 2 days of OHT treatment. Also, the Ring1 knockout experiment and the location analyses were done in different ES lines, and this could be the basis of some of the discrepancy. Nonetheless, the fact that the PRC1-positive set shows a significantly greater response indicates that PRC1 occupancy correlates with functional repression. As a control, we examined expression changes associated with PRC2 loss. We found that PRC1-positive and PRC1-negative bivalent promoters are de-repressed to roughly equal extents in ES cells lacking the PRC2 component Eed (Figure 6) [13].

PRC1-positive bivalent domains correspond to large and conserved sites of H3K27me3: Next, we asked whether the patterns of histone modification vary between the two sets of bivalent domains. We observed two significant trends. First, PRC1-positive bivalent domains are associated with much larger regions of H3K27me3 than PRC1-negative bivalent domains (median size of 3.2 kb versus 1.0 kb). The large size is consistent with a proposed role for H3K27me3 in PRC1 recruitment [2,3]. Second, PRC1-positive bivalent domains exhibit greater conservation of chromatin state: bivalent mouse promoters with PRC1 have a bivalent human ortholog in 71% of cases, compared to just 43% of bivalent mouse promoters without PRC1 ($p < 10^{-10}$; Figure 5C). Thus, PRC1 occupancy correlates with larger bivalent domains that appear to reflect highly conserved functions.

PRC1-positive bivalent domains correspond to developmental regulator genes:



E

GO term	% of genes	Ratio	P-value
Ring1B+			
developmental process	40.5%	2.4	$p<10^{-20}$
regulation of gene expression	30.4%	2.2	$p<10^{-20}$
cell-cell signaling	7.7%	3.8	$p<10^{-20}$
embryonic morphogenesis	5.4%	5.0	$p<10^{-20}$
Ring1B-			
membrane	50.1%	1.2	$p<10^{-10}$
cell adhesion	7.9%	2.3	$p<10^{-5}$
transporter activity	12.3%	1.8	$p<10^{-5}$

Figure 5: PRC1-positive bivalent domains are functionally distinct. **(A)** Box plot shows 25th, 50th and 75th percentile Ring1B ChIP-Seq signals for Ring1B-positive bivalent promoters, Ring1B-negative bivalent promoters, and for H3K4me3 only promoters. **(B)** Plot illustrates fraction of genes up-regulated (red) or down-regulated (blue) in PRC1-deficient ES cells for the indicated gene sets (see text for details on Ring1A/B dKO ES cell model). De-repression is evident for a significantly greater proportion of PRC1-positive bivalent promoters (p-value by Fisher's exact test). **(C)** The proportion of bivalent mouse promoters for which the human ortholog also carries H3K27me3 is indicated, contingent on Ring1B status in mouse ES cells. **(D)** The proportion of bivalent promoters for which H3K27me3 is retained in ES cell-derived neural progenitors (NPCs), contingent on Ring1B status in mouse ES cells. **(E)** Gene Ontology categories over-represented in PRC1-positive or PRC1-negative bivalent gene sets.

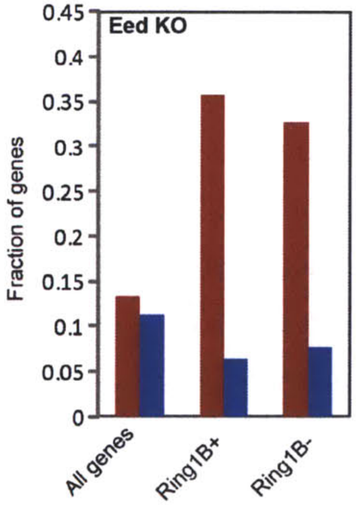


Figure 6: Expression analysis in PRC2 wild-type (WT) and knock-out (KO) mouse ES cells. Expression changes for all genes, Ring1B-positive bivalent and Ring1B-negative bivalent genes in PRC2 knock-out (Eed-/-) mouse ES cells.

Next, we examined the gene targets associated with the different classes of bivalent promoters. The PRC1-positive set contains a dramatic enrichment of genes encoding TFs (30%, $p < 10^{-20}$), including members of the Hox, Sox, Pax and Pou domain families, or cell signaling and morphogenesis molecules, such as Wnts and Fgfs (Table 2). In contrast, the PRC1-negative set of bivalent promoters is instead over-represented for genes that encode membrane proteins (50%; $p < 10^{-10}$). Remarkably, despite the strong correlation of PcG proteins with developmental TFs, this PRC1-negative (PRC2-only) subset of bivalent domains shows statistically significant depletion of TF genes relative to the genome average (4.1% vs 10.2%; $p < 10^{-10}$).

PRC1-positive bivalent domains efficiently maintain repressive chromatin environment: Finally, we compared the behavior of PRC1-positive and PRC1-negative bivalent promoters upon ES cell differentiation. We examined ChIP-Seq data for a population of neural progenitors (NPCs) derived from the same ES cell line [16]. Since PRC1 is implicated in the maintenance of a repressive chromatin state, we reasoned that promoters with PRC1 should more efficiently retain H3K27me3 upon differentiation. Consistent with this hypothesis, we found that 33% of PRC1-positive bivalent promoters retain H3K27me3 in the NPCs, compared to just 10% of PRC1-negative bivalent promoters ($p < 10^{-10}$) (Figure 5D). Many PRC1-positive bivalent promoters that lose the repressive mark upon differentiation do so in association with transcriptional activation as roughly one-fifth are induced at least 5-fold in the NPCs. Thus, PRC1 occupancy is associated with more stable retention of PcG-associated chromatin marks through differentiation.

We conclude that two distinct sets of bivalent domains can be defined based on PcG complex occupancy in ES cells. Bivalent domains that carry both PRC2 and PRC1 are larger, more conserved and more efficiently retained through differentiation. They account for the vast majority of implicated developmental regulators. By contrast, bivalent domains occupied by PRC2 only are poorly maintained, correspond to distinct non-developmental gene sets, and thus may reflect alternate regulatory processes.

Sequence elements and motifs predict PcG complex localization in ES cells

We next studied the chromatin maps to gain insight into another fundamental unanswered question, namely, the mechanisms that underlie the initial recruitment of PcG complexes and the formation of bivalent domains in ES cells. The extensive epigenetic reprogramming that precedes the pluripotent state suggests that elements in the genomic sequence itself must play central roles in this process [1,27,28]. Yet the identity of these PcG-determining sequence elements has remained elusive.

PRC2 associates with CG-rich sequences genomewide: To identify sequence elements that could contribute to PcG recruitment, we applied computational sequence analysis and the new ChIP-Seq data. We focused initially on Ezh2, reasoning that this catalytic PRC2 subunit would most closely reflect the initial recruitment mechanisms. Bivalent domains and PcG target sites have been shown previously to correlate with CG-rich DNA; for example, 50% of Suz12 binding sites in human ES cells correspond to CpG islands [11,16,29]. The ChIP-Seq data for mouse Ezh2 reveal an even higher correspondence, with a full 88% of enriched intervals coinciding with an annotated CpG island. H3K27me3-enriched intervals similarly correlate with CpG islands in 79% of cases. Remarkably, the fraction of Ezh2/H3K27me3 sites that coincide with CpG islands is substantially higher than that of H3K4me3 (68%), which has previously been associated with CpG islands [15]. It is also far greater than that of other chromatin structures (Figure 7), including H3K9me3 (1.1%) and H4K20me3 (0.7%).

When we examined the small minority (12%) of Ezh2 binding sites that do not correspond to an annotated CpG island, we found that three-quarters of these sites overlap highly CG-rich sequences that just fall short of the defined threshold for CpG islands (see Materials and Methods). Including those sites, >97% of Ezh2 binding sites in the ES cell genome correspond to annotated CpG islands or other highly CG-rich sequences. These results suggest that such CG-rich sequences, known to be largely un-methylated at the DNA level in ES cells [27], may contribute to the recruitment of PRC2 and the subsequent establishment of H3K27me3 at bivalent domains.

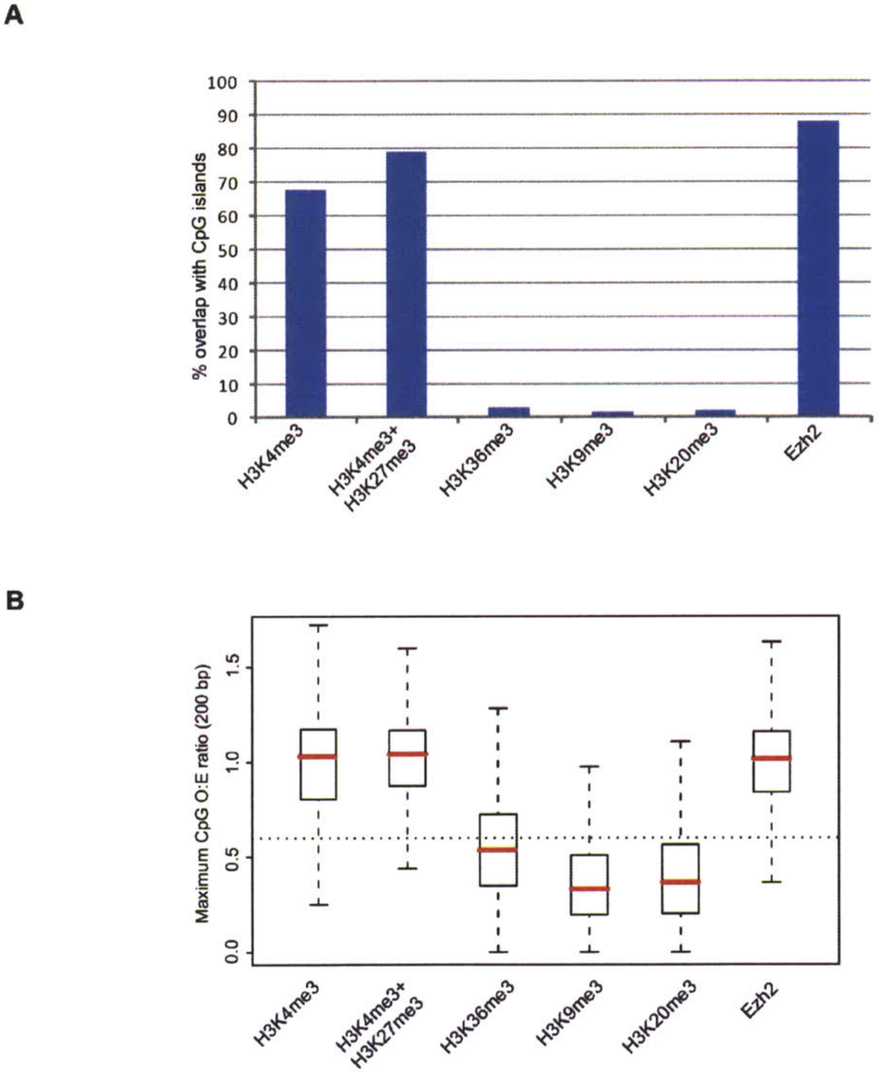


Figure 7: Analysis of the CG-richness of HMM-defined intervals of H3K4me3, H3K27me3, H3K36me3, H3K9me3, H3K20me3, and Ezh2. **(A)**The fraction of intervals that either directly overlap or are within 500 bp of a CpG island. **(B)** The maximum CpG observed-to-expected ratio in any 200 bp window within the interval. The dashed line marks 0.6, one of the criteria used to define a CpG island.

Still, only a minority of CpG islands carries Ezh2 or H3K27me3 in ES cells – that is, are PRC2-positive. Most are enriched for H3K4me3 only and are PRC2-negative (Figure 8A). We thus considered whether additional sequence characteristics distinguish between PRC2-positive and PRC2-negative CpG islands. We collated two sets of CpG islands, one showing clear Ezh2 binding based on ChIP-Seq (n=2608) and the other lacking any Ezh2 signal (n=9097). To maximize the power of our analysis, we excluded a subset of CpG islands showing intermediate levels of Ezh2 enrichment (n=3443).

We considered CpG island length, CG density and the frequency of all possible dinucleotides (Figure 9) as potential characteristics. PRC2-positive CpG islands show a greater median length (721 bp vs 526 bp) and a slightly lower median CpG observed-to-expected ratio (0.88 vs 0.92). However, the overall distributions of length and ratio are largely similar and do not discriminate between PRC2-positive and negative sets.

We also compared the conservation properties of these CpG island sets. Mammalian genomes contain 200 large regions characterized by striking enrichment for highly conserved non-coding elements [30,31] and exceptionally low CpG divergence rates [32]. These loci contain promoters for many developmental genes, most of which are bivalent in ES cells [33]. Although it has been suggested that conserved elements within these loci contribute to PcG recruitment, we find that only 10% of Ezh2 binding sites occur within these regions. Overall, we find that PRC2-positive CpG islands show modestly higher sequence conservation relative to PRC2-negative islands, but with overlapping distributions (Figure 10; Materials and Methods). Thus, conservation analysis does not present an obvious explanation for observed PRC2 binding patterns.

PRC2-positive CpG islands can be distinguished based on motif content: Because the distinction between PRC2-positive and PRC2-negative CpG islands is not explained by simple sequence composition, we next considered more complex sequence motifs. In D. melanogaster, PcG recruitment is mediated by combinations of motifs recognized by specific TFs [4]. We thus explored whether TF motifs could predict PRC2 localization in mammalian ES cells. Since the motifs and TFs implicated in

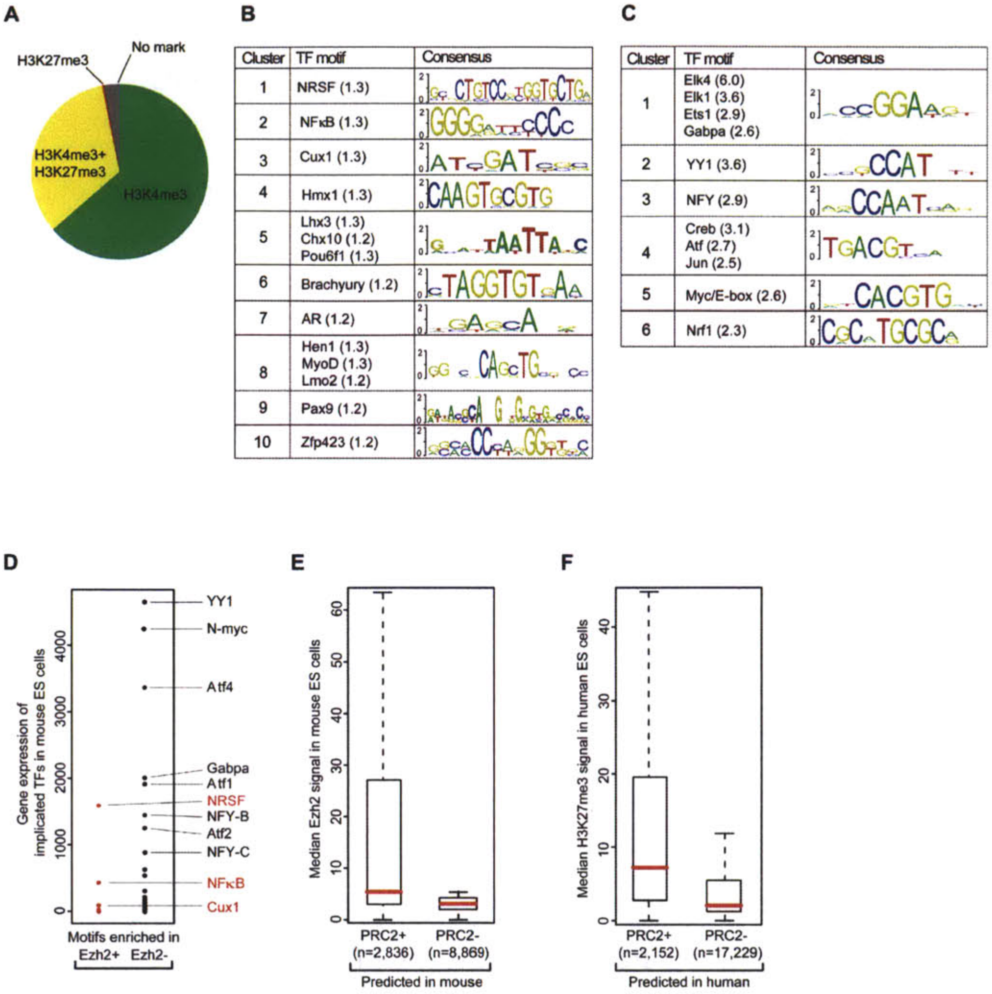


Figure 8: CG-density and DNA motif occurrences predict genomewide PcG complex localization. **(A)** Proportion of CpG islands with a given chromatin state in mouse ES cells. More than 97% of Ezh2 sites in mouse ES cells correspond to CpG islands or other highly CG-rich sequences. A systematic screen reveals sets of DNA motifs over-represented in **(B)** Ezh2-positive CpG islands or **(C)** Ezh2-negative CpG islands (enrichment in parentheses). **(D)** Expression levels of implicated TFs in mouse ES cells. Motifs enriched in Ezh2-positive CpG islands correspond to repressors or to TFs that are not expressed. Motifs enriched in Ezh2-negative CpG islands correspond to highly expressed activators. **(E)** Ezh2 ChIP-Seq signals for CpG islands predicted as PRC2-positive or PRC2-negative based on motif occurrences. **(F)** H3K27me3 ChIP-Seq signals for human ES cells for CpG islands predicted to be PRC2-positive or PRC2-negative based on occurrences of the motifs originally identified in mouse.

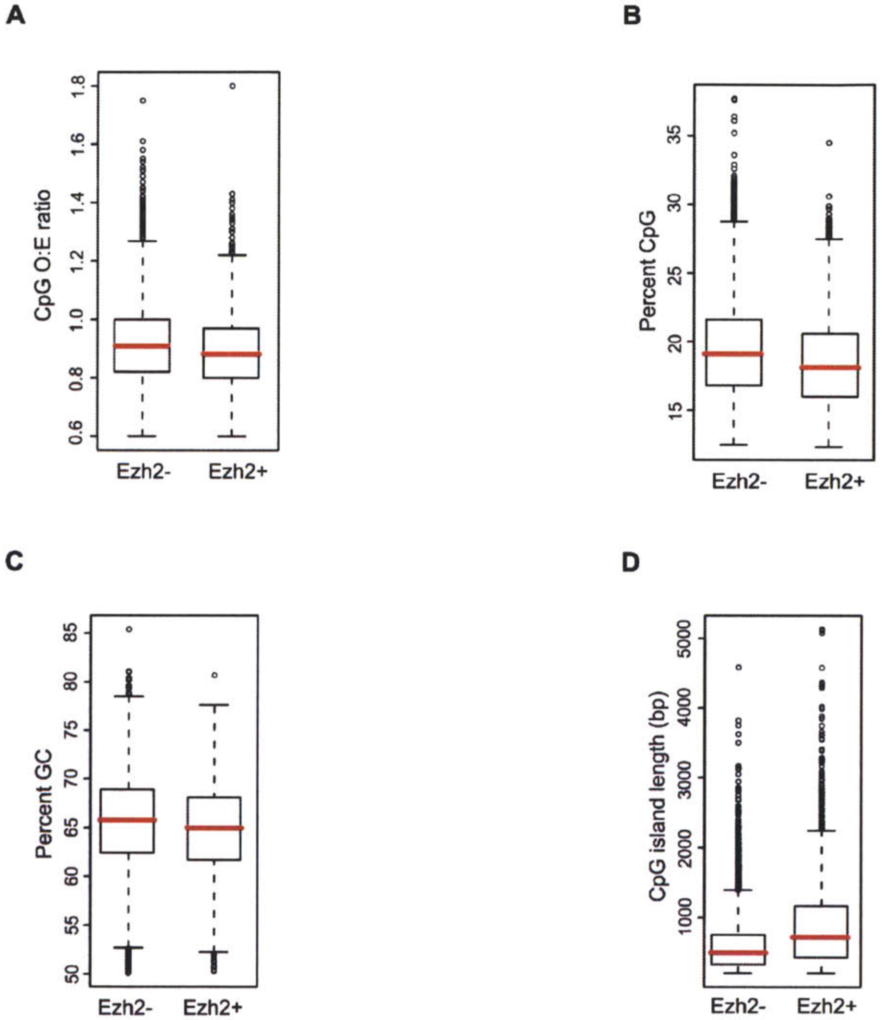


Figure 9: Comparison of Ezh2-positive and Ezh2-negative CpG islands. No marked difference was observed in CpG observed-to-expected ratio **(A)**, percent CpG **(B)**, or percent GC **(C)**, whereas Ezh2-positive CpG islands tend to be longer (median 721 bp vs 526 bp; **D**).

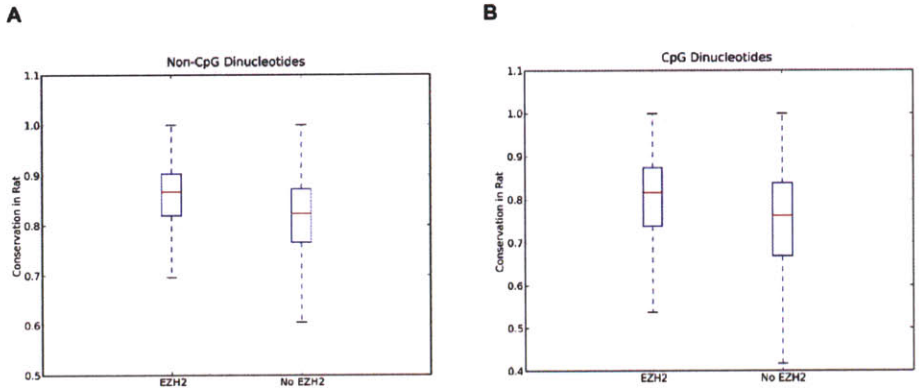


Figure 10: Conservation of Ezh2-bound and Ezh2-unbound dinucleotides between rat and mouse. Aligning regions in rat (rn4) for both classes of CpG island were identified, and a dinucleotide level comparison was performed on the conservation between the two species. Both non-CpG **(A)** and CpG **(B)** dinucleotides were conserved at slightly higher levels in the Ezh2-bound CpG islands than in those islands that did not bind Ezh2.

fly show little or no conservation in vertebrates, we broadened our analysis to include all 668 vertebrate DNA binding motifs annotated in the TRANSFAC and Jaspar databases [34,35].

We used the MAST algorithm [36] and position weight matrices (PWMs) from these databases to identify motifs. Taking an unbiased approach, we searched for motifs over-represented in either Ezh2-positive or Ezh2-negative CpG islands. Over-represented motifs were ranked by enrichment ratio, and their significance was confirmed using Fisher's exact test. We also excluded the possibility that enriched motifs simply reflected differences in underlying nucleotide content by repeating each survey with scrambled PWMs. Finally, since there is redundancy among factors and PWMs in the TRANSFAC and Jaspar databases, a clustering algorithm was used to collapse highly similar PWMs to a single representative motif. This analysis yielded a total of 14 motifs enriched between 1.2 and 1.3-fold in the Ezh2-positive CpG islands, and these fall into 10 motif clusters. It also revealed 11 motifs enriched between 2.3 and 6.0-fold in the Ezh2-negative CpG islands, falling into 6 clusters (Figure 8B,C, Figure 11).

We initially focused on the motifs associated with Ezh2-positive CpG islands as these could potentially mediate PRC2 recruitment. Although the enrichment ratios were relatively low, it is conceivable that combinations of factors might be required, as in *Drosophila*. However, most of the corresponding TFs are not actually expressed in ES cells, but rather are expressed in differentiated cells. These include developmental regulators induced along specific differentiation pathways, such as MyoD (myogenesis), Lmo2 (hematopoiesis), Brachyury (paraxial mesoderm) and Pou6F1 (neurogenesis) [37-40]. PRC2 targets include many developmental genes with complex expression patterns which may explain why they are enriched for lineage-specifying TF motifs. Hence, it is unlikely that these non-expressed TFs contribute to PRC2 localization in ES cells.

However, three of the factors identified in the Ezh2-positive islands are expressed in ES cells, and these cases are illustrative (Figure 8D). The most highly-expressed is neuron-restrictive silencing factor (NRSF/REST), a potent transcriptional repressor

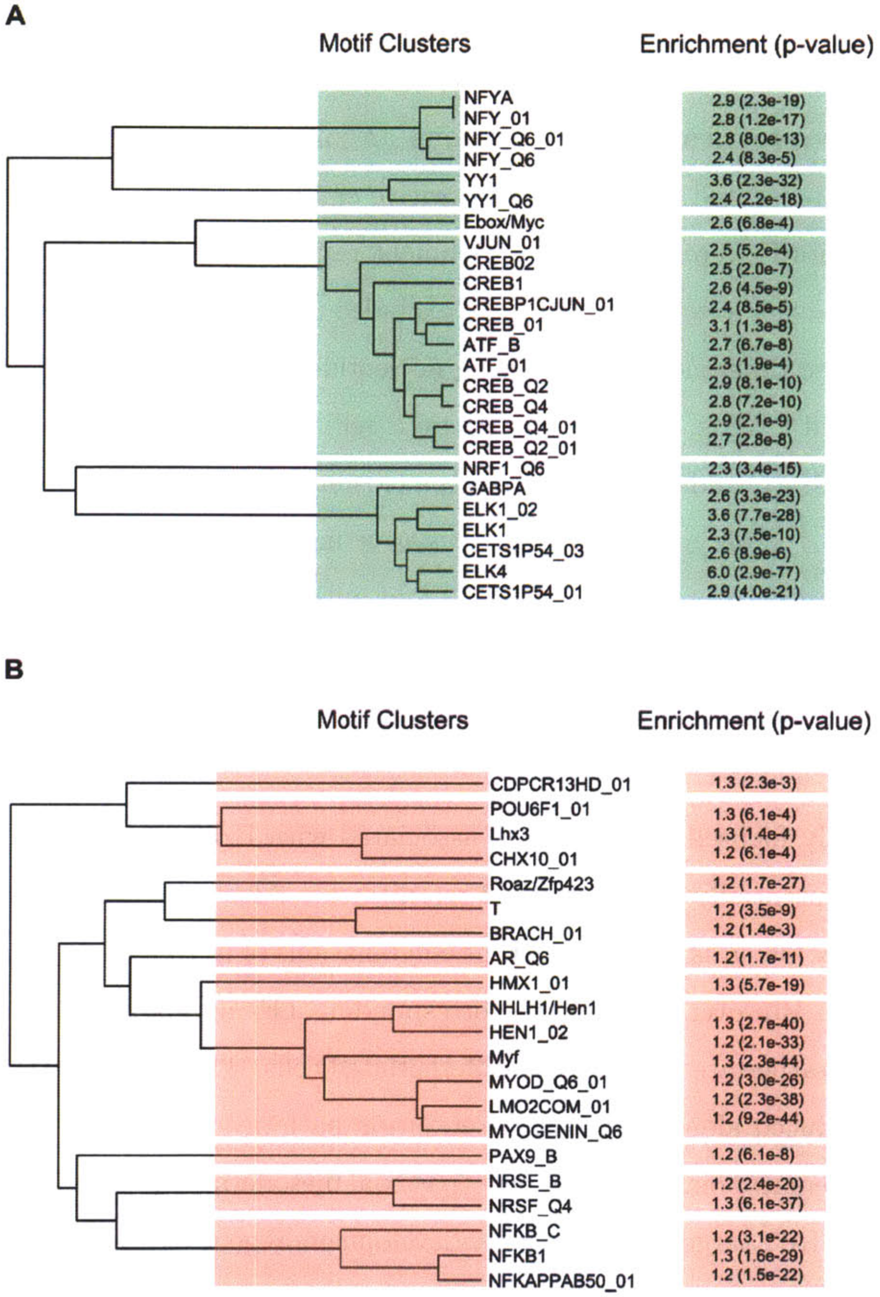


Figure 11: Motif clusters and their respective enrichment p-values for Ezh2-positive and Ezh2-negative CpG islands. The top ranking motifs (and their Bonferroni-corrected p-values from Fisher's exact test) for Ezh2-negative **(A)** and positive **(B)** CpG islands. The motifs were clustered and collapsed to reduce redundancy.

essential for ES cell pluripotency [41]. Notably, the NRSF motif is among the best characterized and highly predictive binding elements in mammalian genomes [42]. A second expressed factor is Cux1, which also functions as a transcriptional repressor [43]. The third expressed factor is NFϰB, a widely studied transcriptional regulator with diverse functions related to immunity, inflammation and differentiation [44]. Although NFϰB is clearly expressed, its activity is strongly inhibited in ES cells by the pluripotency factor Nanog [45]. Thus, motifs enriched in Ezh2-positive CpG islands are recognized either by repressors or by TFs that are inactive in ES cells.

Next, we turned to examine motifs enriched in the Ezh2-negative CpG islands. We were immediately struck that these motifs are recognized by several well-characterized classes of transcriptional activators that are highly expressed in ES cells (Figure 8C,D). Some of the implicated factors have key functions in the ES cell regulatory network (e.g., NFY, Myc) while others are constitutive activators with general housekeeping functions (e.g., Ets1) [46-48]. The magnitudes of enrichment observed for these activating motifs are much greater than those observed for motifs identified in Ezh2-positive sequences above. Thus, the strongest sequence correlate of Ezh2 binding at a CpG island appears to be the *absence* of motifs capable of conferring transcriptional activity.

A simple count of the motif occurrences within a CpG island allows accurate prediction of roughly two-thirds of Ezh2 binding sites (see Materials and Methods; Figure 8E). This compares favorably with the Polycomb response elements predicted in *Drosophila*, which are present at 6 to 27% of experimentally-determined PcG binding sites [4,49-51]. Notably, the motif occurrences we identified in mouse also have considerable predictive value for identifying PcG targets in human ES cells (Figure 8F).

In sum, we find that PRC2-positive CpG islands are characterized by an over-representation of repressor motifs and a strong depletion of transcriptional activator motifs. While it is possible that the implicated repressors directly mediate PRC2 recruitment, each has been well-studied and linked to distinct biological processes. Rather, we favor the view that the paucity of activating motifs and, to a lesser extent, the presence of repressive motifs dictate a transcriptionally inactive state in ES cells that is permissive to PRC2 binding. We suggest that CpG islands play a central role in PRC2 recruitment and, in the absence of transcriptional activity, assume a bivalent chromatin state by default in ES cells (see Discussion).

PRC1 occupies large PRC2-positive CpG islands: Lastly, we considered whether PRC1 association can also be predicted from genome sequence. PRC1 occupies roughly half of all PRC2 sites in ES cells, and is essentially never observed in the absence of this second PcG complex. We collated and compared two sets of Ezh2-positive CpG islands, one with Ring1B (n = 1036) and the other without Ring1B (n = 981) (see Methods). We found no significant differences in nucleotide content (CG-density, dinucleotide frequencies) or in the occurrences of the motifs discussed above.

Rather, the best predictor appears to be the length of CG-rich DNA. PRC1-positive CpG islands are roughly twice as large as those that carry only PRC2 (Figure 12). They are also much more likely to reside in close proximity to other bivalent CpG islands. Consideration of CpG island size and proximity to other bivalent islands enables accurate prediction of PRC1 status for >70% of PRC2-positive CpG islands (see Materials and Methods). Thus, our findings suggest that the genomewide localization of the two main PcG complexes in ES cells may be largely predicted from the location, size and underlying motif content of CpG islands.

Discussion

We have applied ChIP-Seq and computational genomic analysis to study the genomewide distributions of key histone modifications and PcG subunits in mouse and human ES cells, thereby gaining insight into the structure, function and establishment of bivalent domains.

The ChIP-Seq data reveal two distinct sets of bivalent domains in ES cells. One set, defined based on co-occupancy by both PRC1 and PRC2, shows special epigenetic properties, including higher evolutionary conservation of chromatin state and robust retention of repressive chromatin through differentiation. This set is exquisitely en-

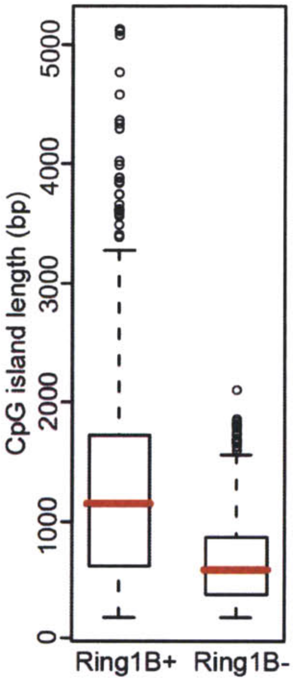


Figure 12: Length of CpG islands in Ring1B-positive and Ring1B-negative bivalent promoters. Ring1B-positive bivalent CpG islands are larger than bivalent CpG islands that are only bound by PRC2.

riched for developmental targets in that over one third of the corresponding genes encode TFs, morphogens or cytokines. In striking contrast, a second set of bivalent domains, occupied by PRC2 only, is actually under-represented for TF genes relative to the genome average, and shows weak conservation and retention of the PcG-associated chromatin marks. We suggest that the complete repertoire of PcG machinery is needed for full functionality of bivalent domains and associated chromatin in the epigenetic regulation of key developmental genes.

The data also suggest a potential model for understanding the initial recruitment of PcG complexes for the coordinated establishment of bivalent chromatin. In particular, we find that PRC2 association in ES cells is entirely restricted to sequences with high CpG content, the vast majority being annotated CpG islands. The status of a given CpG island – whether it carries PRC2 and bivalent H3K4me3/H3K27me3 chromatin or only H3K4me3 – correlates with underlying motif content. CpG islands with PRC2 show a striking depletion of transcriptional activator motifs and a modest enrichment of repressor motifs. Thus, PRC2 appears to localize to CpG islands that are transcriptionally silent in ES cells because they lack activating DNA sequence motifs.

CpG islands have been extensively correlated with trxG complexes and H3K4me3; recruitment of the former likely involves CXXC proteins with affinity for un-methylated CpG dinucleotides [15,52,53]. We propose that CpG islands by default similarly mediate PcG recruitment and catalysis of H3K27me3 in mammalian ES cells, except when the default is over-ridden by transcriptional activity. In this model, the extent of PcG/H3K27me3 and trxG/H3K4me3 at any given CpG island is determined by its baseline transcriptional status which is dictated by underlying motif content. The view that transcriptional status is upstream of PcG status in ES cells is consistent with the subtle transcriptional changes evident in PcG-deficient ES cells [9,54]. Although our analyses do not shed light on the underlying mechanisms, PRC2 recruitment may also involve proteins with affinity for un-methylated CpGs or may be mediated indirectly through recognition of other histone modifications such as H3K4me3. In either case, active transcription within a locus would preclude stable PRC2 association and thereby restrict it to inactive CpG islands.

Large PRC2-positive CpG islands tend to also carry PRC1. The expansive regions of H3K27me3 associated with these islands may contribute to PRC1 recruitment via chromodomain proteins [2,3]. As discussed above, bivalent domains that carry both PRC2 and PRC1 appear to have unique epigenetic regulatory properties. We therefore propose that large CpG islands depleted of activating motifs confer epigenetic regulation by recruiting both key PcG complexes in pluripotent cells. Such islands may thereby reflect mammalian memory elements analogous to Polycomb response elements in flies.

The tight correspondence between DNA sequence and PcG localization may have implications for important cellular processes, such as development and epigenetic reprogramming. Induced pluripotent stem (iPS) cells and ES cells exhibit nearly identical chromatin patterns, including the locations of bivalent domains [55,56]. The sequences described above may function as templates for the robust assembly and appropriate positioning of PcG complexes and bivalent domains during pre-implantation development or the artificial reprogramming of somatic cells to iPS cells [1,28].

What then might be the purpose of an initial chromatin state fully encoded by genetic sequence and an associated transcriptional program? Based on existing evidence, we suggest that PcG complexes and associated chromatin buffer the pluripotent ground state by reinforcing the repression of factors that induce differentiation. The initial chromatin architecture also appears poised for the dynamic expression changes that accompany differentiation and for the subsequent engagement of epigenetic controls to maintain lineage-specific transcriptional programs. Our analysis suggests that such epigenetic functions mainly apply to large bivalent CpG islands that also carry PRC1. It remains to be seen whether small PRC1-negative bivalent domains have distinct regulatory functions or are simply byproducts of the mechanisms that have evolved for establishment of the former.

Further studies are needed to determine the precise DNA elements and protein interactions that mediate PcG recruitment. As discussed above, the proposed central role for CG-rich sequences implies the involvement of CXXC domains or other proteins that recognize CG dinucleotides. However, several factors complicate the interpretation of our genomic findings. In particular, CpG islands are at least partly a consequence of reduced CpG deamination rates in regions that lack DNA methylation in the germ line [27]. PcG-occupied regions are largely un-methylated at the DNA level, at least in ES cells [57], and this could favor retention of CG-rich sequences. Thus, it remains possible that evolutionary dynamics and/or the generally high CpG content of target regions are masking other key sequence features.

Finally, it should be emphasized that our findings on the relationships among PRC2 and PRC1 and the sequences that underlie their genomic localizations pertain specifically to ES cells. PcG complexes show remarkable tissue-specificities in terms of their expression levels, stoichiometry and localization [2,3,11,12]. Further study is needed to understand how the genomic localizations and regulatory functions of PcG complexes vary with differentiation, lineage specification, environment, and disease.

Materials and methods

Cell culture

Mouse v6.5 (genotype 129SvJae x C57BL6, male, passages 10-15) ES cells were cultured on fibroblast feeders in DMEM (Sigma) with 15% fetal bovine serum (Hyclone), GlutaMax (Invitrogen), MEM non-essential amino acids (Invitrogen), pen/strep (Invitrogen), ESGRO (Chemicon) and 2-mercaptoethanol (Sigma), incubating at 37°C, 5% CO2 [16]. Prior to harvest, these cells were passaged 2-3 times on feeder-free gelatinized tissue culture plates. A transgenic ES cell line expressing a fusion between Ring1B and biotin ligase recognition peptide from the endogenous Ring1B locus and the BirA biotin ligase from the Rosa26 locus (H.K., unpublished) was cultured as described above. Human H9 (female, passage 45) ES cells were cultured as described [58] and at http://www.WiCell.org. Briefly, the human ES cells were cultivated on irradiated MEFs (strain DR4) in Knockout DMEM (Invitrogen) containing 10% Knockout Serum Replacement (Invitrogen), 10% Plasmanate (Bayer Healthcare), GlutaMax (2mM), pen/strep, MEM non-essential amino acids (0.1mM), 10ng/ml Îš-FGF (Invitrogen) and 2-mercaptoethanol. Cells were incubated at 37°C, 5% CO2. MEF-free ES cells were used for analysis. MEF-free culture was prepared in the following manner: First, MEFs were depleted at the time of trypsin passaging through brief transfer (thirty minutes) of hES cells onto gelatin-coated plates. MEF-subtracted ES cells were then propagated on plates coated with Matrigel (Invitrogen). ES cells grown on Matrigel were supported with the aforementioned human ES cell medium that had first been conditioned on MEFs for 24 hours. Fresh beta-FGF was added to the conditioned medium immediately prior to use.

Generation of Flag-Bmi1 mES cells

Doxycyclin-inducible Flag-Bmi1 transgenic ES cell line was generated by PCR amplifying a 1X flag tagged Bmi1 ORF (Addgene) with primers that incorporate a 3X flag tag as well as EcoRI and XbaI restriction enzyme sites. This was cloned into the pLox vector (pPGK-loxP-neoEGFP) and incorporated into Ainv15 mouse ES cells using a cre recombinase expression vector as previously described [59]. Flag-Bmi1 ES cells were cultured similarly to wild-type mES cells as described above. Prior to harvest, Flag-Bmi1 expression was induced by incubating with 1 ug/ml of Doxycycline for two days on gelatinized culture plates.

Chromatin immunoprecipitation and antibodies

ChIP experiments for H3K4me3, H3K27me3 and H3K36me3, Ring1B and Flag-Bmi1 were carried out as described [15,16]. ES cells were crosslinked in 1% formaldehyde, lysed and sonicated with either a Branson 250 Sonifier (mouse ES cells) or a Diagenode bioruptor (human ES cells) to obtain chromatin fragments in a size range between 200 and 700 bp. Solubilized chromatin (whole cell lysate or WCE) was diluted in ChIP dilution buffer (1:10) and incubated with antibody overnight at 4°C. Protein A sepharose beads (Sigma) were used to capture the antibody-chromatin complex and washed with low salt, LiCl, as well as TE (pH 8.0) wash buffers. Enriched chromatin fragments were eluted at 65°C for 10 min, subjected to crosslink reversal at 65°C for 5 hrs, and treated with Proteinase K (1mg/ml), before being extracted by phenol-chloroform-isoamyl alcohol, and ethanol precipitated. ChIP DNA was then quantified by Quant-iT Picogreen dsDNA Assay kit (Invitrogen).

ChIP experiments for Ezh2 and Suz12 were carried out on nuclear preps. Crosslinked ES cells were incubated in swelling buffer (0.1M Tris pH7.6, 10mM KOAc, 15mM MgOAc, 1% NP40), on ice for twenty minutes, passed through a 16G needle 20 times and centrifuged to collect nuclei [60]. Isolated nuclei were then lysed, sonicated and immunoprecipitated as described above.

BioChIP assays were carried out using transgenic Ring1B-Biotin ligase recognition peptide ES cells (above). Nuclei were isolated, lysed and sonicated as described above. Dynabeads M-280 Streptavidin (Invitrogen 112.05D) were used to capture biotinylated Ring1B-DNA complex. Beads were washed with a 2% SDS buffer and a high salt buffer (50mM HEPES, pH7.5, 1mM EDTA, 500mM NaCl, 1% Triton X-100, 0.1% Deoxycholate), in addition to the regular washes. Elution and cross-link reversal were done simultaneously by incubating Dynabeads in 300mM NaCl at 65oC overnight [46]. DNA was isolated as described above.

Antibodies used in this study include anti-H3K4me3 (Abcam ab8580), anti-H3K27-me3 (Upstate 07-449), anti-H3K36me3 (Abcam ab9050), anti-Ezh2 (Active Motif 39103), anti-Suz12 (Abcam ab12073), anti-Ring1B [61] and anti-Flag (M2) (Sigma F1804).

Sequencing library preparation and Illumina/Solexa sequencing

Library preparation and ultra high-throughput sequencing were carried out as described [16]. Briefly, one to ten nanograms (ng) of ChIP DNA were end-repaired and 5'phosphorylated using END-It DNA End-Repair Kit (Epicentre). We then followed steps four through seven of Illumina standard sample prep protocol (v1.8) using Genomic DNA Sample Prep Kit (Illumina) with minor modifications. A single Adenine was added to 3'ends by Klenow (3'→ 5'exo), and double-stranded Illumina Adapters were ligated to the ends of the ChIP fragments. Adapter-ligated ChIP DNA fragments between 275 bp to 700 bp were gel-purified and subjected to 18 cycles of PCR. Prepared libraries were quantified using PicoGreen and sequenced on the Illumina Genome Analyzer per standard operating procedures.

Read alignment and generation of density maps and modified intervals

Sequence reads (36 bases) from each ChIP experiment were compiled, post-processed and aligned to the appropriate reference genome using a general purpose computational pipeline as described previously [16]. Aligned reads are used to estimate the number of end-sequenced ChIP fragments that overlap any given genomic position (at 25-bp resolution). For each position, we counted the number of reads that are oriented towards it and closer than the average length of a library fragment (~300 bp). The result is a high-resolution density map that can be viewed through the UCSC Genome Browser [62] and is used for downstream analyses. Prior comparisons to microarray analysis and quantitative real-time PCR have shown that ChIP-Seq density maps accurately reflect enrichment [16]. ChIP-Seq data can be accessed at http://www.broad.mit.edu/seq_platform/chip/.

We used a Hidden Markov Model (HMM) to demarcate chromosomal segments likely to be enriched for a given chromatin modification or PcG protein [16]. In order to model ChIP-Seq read density variations along the genome, we define four observed states: masked, low density, medium density, and high density. This discretization of the data into the four states was based on the signal intensity in known modified regions versus known unmodified regions as determined in prior ChIP-Seq, microarray and ChIP-PCR analyses [15,16], and adjusted for each sample. The model was then used to discriminate enriched and unenriched intervals genome wide. In order to more properly classify enriched regions containing several short interspersed peaks and facilitate subsequent analyses intervals within 2 kb were merged.

Promoter classification and definition of gene and transcript intervals

We defined 17760 mouse and 18522 human promoters for 17442 and 17383 genes, respectively, as the sequences between -0.5 kb and +2.0 kb of the annotated transcription start site, using the mouse mm8 and human hg18 genome builds. Transcripts were defined for these genes as the range from transcription start to end [62]. To identify regions enriched for histone marks or chromatin-associated proteins, we generated a null-hypothesis background model by dividing the alignable parts of each chromosome into 200 bp bins and randomly redistributing the reads aligned on this chromosome. Based on a histogram of the cumulative distribution of reads per bin, a cutoff threshold was determined. Stability of the calculated background cutoff threshold was confirmed through 1000 independent simulations for each ChIP-Seq track and showed remarkable invariance. For promoters, a 200 bp sliding window was moved across the 2.5 kb promoter region and the ratio of median read density over background was calculated. The maximum enrichment achieved in any window at this promoter site was then used for further analysis. Maximum enrichment cutoff thresholds were determined empirically for all tracks, and promoters were then classified based on the maximum enrichment for the various histone marks and PcG proteins. The same procedure was applied to a pan-H3 (modification-insensitive) ChIP-Seq dataset as control where virtually no significant enrichment over background was found. Ring1B-positive bivalent promoters were defined based on normalized ChIP-Seq signal and comprise 40% of all bivalent promoters. A set of Ring1B-negative bivalent promoters was also defined based on absence of ChIP-Seq enrichment, and includes another 40% of all bivalent promoters. The remaining bivalent promoters (20%) with indeterminate Ring1B ChIP-Seq signals were excluded from this analysis.

For conservation analyses of human and mouse promoter states, we used NCBI HomoloGene (build 58) gene clusters to assign orthologous human promoters and transcripts to the 17442 mouse promoters and transcripts, yielding a set of 13200 orthologous promoters and 13625 orthologous transcripts for which human and mouse chromatin state could be compared (ftp://ftp.ncbi.nih.gov/pub/HomoloGene/). Genes with multiple start sites were excluded from this analysis. Promoters were associated with CpG states as described previously [16].

For comparison of Ezh2 and Ring1B occupancy at target genes, a reduced Ezh2 read set was generated by randomly selecting the same number of reads that were available for Ring1B from the full Ezh2 read pool (~3.5 million). Read mapping to the mouse genome and analysis of promoter state were performed as described above.

Real-time PCR

PCR primer pairs were designed to amplify designated genomic regions using

Primer3 (http://fokker.wi.mit.edu/primer3/input.htm). Real-time PCR assays were carried out on ABI 7000 or 7500 detection systems. We used Quantitect SYBR green PCR mix (Qiagen) with 0.1 ng ChIP or 0.1 ng un-enriched input DNA (WCE) as template. Log2 enrichment was calculated from geometric means obtained from three independent ChIP experiments, each evaluated by duplicate PCR assays. Background was subtracted by normalizing over negative genomic control.

Gene expression analysis

Gene expression data for Ring1A/B-dKO (Ring1A -/-; Ring1B fl/fl; Rosa26::CreERT2) ES cells (2 days post-tamoxifen treatment and no-treatment control, H. Koseki unpublished data) and Eed KO ES cells (Eed -/- and control Eed+/+ ES) [13], acquired with Affymetrix Mouse Genome 430 2.0 Arrays, were normalized using the Genepattern expression data analysis package (http://www.broad.mit.edu/cancer/software/genepattern). CEL files were processed with RMA, quantile normalization and background correction [63]. For a given comparison (Ring1A/B-dKO vs control; or Eed -/- vs +/+), we only considered probes in which at least one of the experiments had a “P” significance call. Fold changes were calculated for each passing probe. Genes with multiple corresponding probes were assigned the geometric average fold change value. Gene expression data for mouse v6.5 mES and NPCs were obtained from previously published Affymetrix mRNA profiles [16].

Gene class enrichment analysis

Gene ontology (GO) functional annotation for the Ring1B positive and negative sets was done using DAVID analysis tool (http://david.abcc.ncifcrf.gov/home.jsp). P-values were adjusted for multiple hypothesis testing using Bonferroni correction.

CG content and motif enrichment analysis

The HMM described above was used to define enriched intervals for each modification or chromatin protein from the mouse ES cell ChIP-Seq data. We determined the extent to which Ezh2 intervals (and those for other epitopes) overlap with CG-rich sequences. CpG island coordinates were obtained from the UCSC Genome Browser

[62]. We identified all Ezh2 intervals that overlap these CpG island coordinates within 500 bp. Next, the EMBOSS analysis package [64] was used to determine the portion of remaining Ezh2 intervals overlapping a 'mini' CpG island defined as a 100 bp window with at least 50% GC content and an O:E ratio >0.6 (instead of the standard CpG island window of 200 bp).

We next classified CpG islands according to their chromatin state (e.g., Ezh2-positive v. Ezh2-negative, H3K4me3 v. bivalent). This was done by computing the median ChIP-Seq read density across each defined CpG island, and setting thresholds using a null background model of randomized reads. For these analyses we excluded CpG islands that fall within unalignable regions, typically due to low complexity sequence, and thus could not be evaluated by ChIP-Seq (<7% of all CpG islands). To maximize discriminatory power, we excluded intermediate CpG islands with sub-threshold Ezh2 signal. We computed median values and distributions for length, CG density and observed-to-expected ratio for the different CpG island sets, and also evaluated nucleotide content by calculating the frequencies of all 16 dinucleotide combinations. Conservation scores were determined for each CpG island by aligning the regions between mouse and rat, and performing a dinucleotides level comparison of the conservation between the two species. Both CpG and non-CpG dinucleotides were conserved at slightly higher levels in the Ezh2-bound CpG islands (Figure 10).

We next screened the CpG island sets for TF motif occurrences. 668 position weight matrices (PWMs) were obtained from the Jaspar (Release 3.0 [34]) and TRANSFAC (Release 9.4; [35]) databases, excluding any non-vertebrate factors. We prepared sets of Ezh2-positive and Ezh2-negative sequences by extracting each CpG island along with flanking sequence equal to 50% of its length. The MAST algorithm [36] was then used to search for significant PWM matches ($p < 5x10^{-5}$) in the Ezh2-positive and negative sets. Occurrences were length-normalized and used to calculate ratios that reflect the enrichment in the Ezh2-positive set relative to the Ezh2-negative set, or vice versa. We identified significantly over-represented motifs using Fisher's exact test with Bonferroni-adjusted p-values. These candidate motifs were then scrambled, re-scored, and excluded if any enrichment was observed in the scramble.

We used a clustering algorithm to collapse similar motifs identified as enriched in one of the sets to a single consensus sequence [65]. This was necessary due to high motif redundancy in the databases. After clustering, all intra-cluster motif occurrences overlapping by more than 50% were counted as a single instance. Expression values for corresponding DNA binding proteins were determined from previously published Affymetrix mRNA profiles for v6.5 ES cells [16].

A simple count-based model was used to determine the extent to which motif occurrences are predictive of Ezh2 status. The motif content which allowed for maximum discrimination in mouse is as follows: a CpG island was predicted to be Ezh2-positive if it either (i) contained $>$ 8 'Ezh2-positive' motifs or (ii) contained $>$ 4 'Ezh2-positive' motifs and $<$ 2 'Ezh2-negative' motifs. Ezh2 status in human was predicted using the motifs identified in mouse but with the following metric: a CpG island was predicted to be Ezh2-positive if it contained $>$ 15 'Ezh2-positive' motifs and $<$ 2 'Ezh2-negative' motifs.

In order to quantify Ring1B presence in CpG islands, we considered the distribution of ChIP-Seq reads in control regions. We specifically used all alignable, H3K4me3-only CpG islands as our null hypothesis background model. The distribution of Ring1B ChIP-Seq read densities across these islands was calculated and a threshold was set to minimize the false positive detection rate. We then calculated Ring1B ChIP-Seq read density in sliding 200 bp windows in all Ezh2-positive CpG islands, with a CpG island assigned the maximum enrichment in any of its 200 bp windows. For maximum discriminatory power, we excluded 20% of CpG islands with sub-threshold Ring1B signal. Ring1B status was predicted using the length of CpG-richness in PRC2-positive CpG islands. Islands were predicted to be Ring1B-positive if they were either $>$ 1200 bp or within 2 kb of another CpG island.

Acknowledgments

We thank the staff of the Broad Institute Genome Sequencing Platform for assistance with reagents and data generation. We thank Manolis Kellis, Pouya Kheradpour and Alex Meissner for helpful discussions. We acknowledge L. Zagachin and the MGH RT-PCR core for assistance with quantitative PCR. We thank Miguel Vidal for Ring1A KO cells and reagents for generating the Ring1B conditional knock-down and transgenic biotin-Ring1B ES cells. The plox (pPGK-loxP-neoEGFP) plasmid was a kind gift from G. Daley.

References

1. Jaenisch R, Young R (2008) Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell 132: 567-582.

2. Schuettengruber B, Chourrout D, Vervoort M, Leblanc B, Cavalli G (2007) Genome regulation by polycomb and trithorax proteins. Cell 128: 735-745.

3. Sparmann A, van Lohuizen M (2006) Polycomb silencers control cell fate, development and cancer. Nat Rev Cancer 6: 846-856.

4. Ringrose L, Paro R (2007) Polycomb/Trithorax response elements and epigenetic mcmory of cell identity. Development 134: 223-232.

5. de Napoles M, Mermoud JE, Wakao R, Tang YA, Endoh M, et al. (2004) Polycomb group proteins Ring1A/B link ubiquitylation of histone H2A to heritable gene silencing and X inactivation. Dev Cell 7: 663-676.

6. Wang H, Wang L, Erdjument-Bromage H, Vidal M, Tempst P, et al. (2004) Role of histone H2A ubiquitination in Polycomb silencing. Nature 431: 873-878.

7. Zhou W, Zhu P, Wang J, Pascual G, Ohgi KA, et al. (2008) Histone H2A monoubiquitination represses transcription by inhibiting RNA polymerase II transcriptional elongation. Mol Cell 29: 69-80.

8. Stock JK, Giadrossi S, Casanova M, Brookes E, Vidal M, et al. (2007) Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells. Nat Cell Biol 9: 1428-1435.

9. Boyer LA, Plath K, Zeitlinger J, Brambrink T, Medeiros LA, et al. (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature 441: 349-353.

10. Bracken AP, Dietrich N, Pasini D, Hansen KH, Helin K (2006) Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. Genes

Dev 20: 1123-1136.

11. Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, et al. (2006) Control of developmental regulators by Polycomb in human embryonic stem cells. Cell 125: 301-313.

12. Squazzo SL, O'Geen H, Komashko VM, Krig SR, Jin VX, et al. (2006) Suz12 binds to silenced regions of the genome in a cell-type-specific manner. Genome Res 16: 890-900.

13. Endoh M, Endo TA, Endoh T, Fujimura Y, Ohara O, et al. (2008) Polycomb group proteins Ring1A/B are functionally linked to the core transcriptional regulatory circuitry to maintain ES cell identity. Development 135: 1513-1524.

14. Azuara V, Perry P, Sauer S, Spivakov M, Jorgensen HF, et al. (2006) Chromatin signatures of pluripotent cell lines. Nat Cell Biol 8: 532-538.

15. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125: 315-326.

16. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553-560.

17. Pan G, Tian S, Nie J, Yang C, Ruotti V, et al. (2007) Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. Cell Stem Cell 1: 299-312.

18. Zhao XD, Han X, Chew JL, Liu J, Chiu KP, et al. (2007) Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. Cell Stem Cell 1: 286-298.

19. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, et al. (2007)

High-resolution profiling of histone methylations in the human genome. Cell 129: 823-837.

20. Wei CL, Miura T, Robson P, Lim SK, Xu XQ, et al. (2005) Transcriptome profiling of human and murine ESCs identifies divergent paths required to maintain the stem cell state. Stem Cells 23: 166-185.

21. Besser D (2004) Expression of nodal, lefty-a, and lefty-B in undifferentiated human embryonic stem cells requires activation of Smad2/3. J Biol Chem 279: 45076-45084.

22. Xu RH, Peck RM, Li DS, Feng X, Ludwig T, et al. (2005) Basic FGF and suppression of BMP signaling sustain undifferentiated proliferation of human ES cells. Nat Methods 2: 185-190.

23. Schuringa JJ, van der Schaaf S, Vellenga E, Eggen BJ, Kruijer W (2002) LIF-induced STAT3 signaling in murine versus human embryonal carcinoma (EC) cells. Exp Cell Res 274: 119-129.

24. Tesar PJ, Chenoweth JG, Brook FA, Davies TJ, Evans EP, et al. (2007) New cell lines from mouse epiblast share defining features with human embryonic stem cells. Nature 448: 196-199.

25. Goldin SN, Papaioannou VE (2003) Paracrine action of FGF4 during periimplantation development maintains trophectoderm and primitive endoderm. Genesis 36: 40-47.

26. Loh YH, Wu Q, Chew JL, Vega VB, Zhang W, et al. (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat Genet 38: 431-440.

27. Bernstein BE, Meissner A, Lander ES (2007) The mammalian epigenome. Cell 128: 669-681.

28. Surani MA, Hayashi K, Hajkova P (2007) Genetic and epigenetic regulators of pluripotency. Cell 128: 747-762.

29. Mohn F, Weber M, Rebhan M, Roloff TC, Richter J, et al. (2008) Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Mol Cell 30: 755-766.

30. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, et al. (2005) Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol 3: e7.

31. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438: 803-819.

32. Tanay A, O'Donnell AH, Damelin M, Bestor TH (2007) Hyperconserved CpG domains underlie Polycomb-binding sites. Proc Natl Acad Sci U S A 104: 5521-5526.

33. Bernstein E, Duncan EM, Masui O, Gil J, Heard E, et al. (2006) Mouse polycomb proteins bind differentially to methylated histone H3 and RNA and are enriched in facultative heterochromatin. Mol Cell Biol 26: 2560-2569.

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

35. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, et al. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31: 374-378.

36. Bailey TL, Gribskov M (1998) Combining evidence using p-values: application to sequence homology searches. Bioinformatics 14: 48-54.

37. Weintraub H, Davis R, Tapscott S, Thayer M, Krause M, et al. (1991) The myoD gene family: nodal point during specification of the muscle cell lineage. Science 251: 761-766.

38. Yamada Y, Warren AJ, Dobson C, Forster A, Pannell R, et al. (1998) The T cell leukemia LIM protein Lmo2 is necessary for adult mouse hematopoiesis. Proc Natl Acad Sci U S A 95: 3890-3895.

39. Donahue LM, Reinhart AJ (1998) POU domain genes are differentially expressed in the early stages after lineage commitment of the PNS-derived stem cell line, RT4-AC. Brain Res Dev Brain Res 106: 1-12.

40. Yamaguchi TP, Takada S, Yoshikawa Y, Wu N, McMahon AP (1999) T (Brachyury) is a direct target of Wnt3a during paraxial mesoderm specification. Genes Dev 13: 3185-3190.

41. Singh SK, Kagalwala MN, Parker-Thornburg J, Adams H, Majumder S (2008) REST maintains self-renewal and pluripotency of embryonic stem cells. Nature 453: 223-227.

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

43. Ellis T, Gambardella L, Horcher M, Tschanz S, Capol J, et al. (2001) The transcriptional repressor CDP (Cutl1) is essential for epithelial cell differentiation of the lung and the hair follicle. Genes Dev 15: 2307-2319.

44. Hayden MS, Ghosh S (2004) Signaling to NF-kappaB. Genes Dev 18: 2195-2224.

45. Torres J, Watt FM (2008) Nanog maintains pluripotency of mouse embryonic stem cells by inhibiting NFkappaB and cooperating with Stat3. Nat Cell Biol 10: 194-201.

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

47. Grskovic M, Chaivorapol C, Gaspar-Maia A, Li H, Ramalho-Santos M (2007) Systematic identification of cis-regulatory sequences active in mouse and human embryonic stem cells. PLoS Genet 3: e145.

48. Hollenhorst PC, Shah AA, Hopkins C, Graves BJ (2007) Genome-wide analyses reveal properties of redundant and specific promoter occupancy within the ETS gene family. Genes Dev 21: 1882-1894.

49. Negre N, Hennetin J, Sun LV, Lavrov S, Bellis M, et al. (2006) Chromosomal distribution of PcG proteins during Drosophila development. PLoS Biol 4: e170.

50. Schwartz YB, Kahn TG, Nix DA, Li XY, Bourgon R, et al. (2006) Genome-wide analysis of Polycomb targets in Drosophila melanogaster. Nat Genet 38: 700-705.

51. Tolhuis B, de Wit E, Muijrers I, Teunissen H, Talhout W, et al. (2006) Genome-wide profiling of PRC1 and PRC2 Polycomb chromatin binding in Drosophila melanogaster. Nat Genet 38: 694-699.

52. Voo KS, Carlone DL, Jacobsen BM, Flodin A, Skalnik DG (2000) Cloning of a mammalian transcriptional activator that binds unmethylated CpG motifs and shares a CXXC domain with DNA methyltransferase, human trithorax, and methyl-CpG binding domain protein 1. Mol Cell Biol 20: 2108-2121.

53. Birke M, Schreiner S, Garcia-Cuellar MP, Mahr K, Titgemeyer F, et al. (2002) The MT domain of the proto-oncoprotein MLL binds to CpG-containing DNA and discriminates against methylation. Nucleic Acids Res 30: 958-965.

54. Pasini D, Bracken AP, Hansen JB, Capillo M, Helin K (2007) The polycomb group protein Suz12 is required for embryonic stem cell differentiation. Mol Cell Biol 27: 3769-3779.

55. Wernig M, Meissner A, Foreman R, Brambrink T, Ku M, et al. (2007) In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. Nature.

56. Maherali N, Sridharan R, Xie W, Utikal J, Eminli S, et al. (2007) Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. Cell Stem Cell 1: 55-70.

57. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, et al. (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 454: 766-770.

58. Thomson JA, Itskovitz-Eldor J, Shapiro SS, Waknitz MA, Swiergiel JJ, et al. (1998) Embryonic stem cell lines derived from human blastocysts. Science 282: 1145-1147.

59. Kyba M, Perlingeiro RC, Daley GQ (2002) HoxB4 confers definitive lymphoid-myeloid engraftment potential on embryonic stem cell and yolk sac hematopoietic progenitors. Cell 109: 29-37.

60. Weinmann AS, Bartley SM, Zhang T, Zhang MQ, Farnham PJ (2001) Use of chromatin immunoprecipitation to clone novel E2F target promoters. Mol Cell Biol 21: 6820-6832.

61. Atsuta T, Fujimura S, Moriya H, Vidal M, Akasaka T, et al. (2001) Production of monoclonal antibodies against mammalian Ring1B proteins. Hybridoma 20: 43-46.

62. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome Res 12: 996-1006.

63. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, et al. (2006) GenePattern 2.0. Nat Genet 38: 500-501.

64. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology

Open Software Suite. Trends Genet 16: 276-277.

65. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature 434: 338-345.

Chapter 3

GC-Rich Sequence Elements Recruit PRC2 in Mammalian ES Cells

This work was originally published as:

*Mendenhall EM, *Koche RP, Truong T, Zhou VW, Issac B, Chi AS, Ku M and Bernstein BE. (2010) GC-Rich Sequence Elements Recruit PRC2 in Mammalian ES Cells. PLoS Genet 6(12): e1001244. doi:10.1371/journal.pgen.1001244

Contributions:

Conceived and designed the experiments: EMM, RPK, BEB. Performed the experiments: EMM, RPK, TT, VWZ, BI, ASC, MK. Analyzed the data: EMM, RPK. Wrote the paper: EMM, RPK, BE.

Abstract

Polycomb proteins are epigenetic regulators that localize to developmental loci in the early embryo where they mediate lineage-specific gene repression. In *Drosophila*, these repressors are recruited to sequence elements by DNA binding proteins associated with Polycomb repressive complex 2 (PRC2). However, the sequences that recruit PRC2 in mammalian cells have remained obscure. To address this, we integrated a series of engineered bacterial artificial chromosomes into embryonic stem (ES) cells and examined their chromatin. We found that a 44 kb region corresponding to the Zfpm2 locus initiates de novo recruitment of PRC2. We then pinpointed a CpG island within this locus as both necessary and sufficient for PRC2 recruitment. Based on this causal demonstration and prior genomic analyses, we hypothesized that large GC-rich elements depleted of activating transcription factor motifs mediate PRC2 recruitment in mammals. We validated this model in two ways. First, we showed that a constitutively active CpG island is able to recruit PRC2 after excision of a cluster of activating motifs. Second, we showed that two 1 kb sequence intervals from the *Escherichia coli* genome with GC-contents comparable to a mammalian CpG island are both capable of recruiting PRC2 when integrated into the ES cell genome. Our findings demonstrate a causal role for GC-rich sequences in PRC2 recruitment and implicate a specific subset of CpG islands depleted of activating motifs as instrumental for the initial localization of this key regulator in mammalian genomes.

Introduction

Polycomb proteins are epigenetic regulators required for proper gene expression patterning in metazoans. The proteins reside in two main complexes, termed Polycomb repressive complex 1 and 2 (PRC1 and PRC2). PRC2 catalyzes histone H3 lysine 27 tri-methylation (K27me3), while PRC1 catalyzes histone H2A ubiquitination and mediates chromatin compaction [1,2]. PRC1 and PRC2 are initially recruited to target loci in the early embryo where they subsequently mediate lineage-specific gene repression. In embryonic stem (ES) cells, the complexes localize to thousands of genomic sites, including many developmental loci [3-5]. These target loci are not yet stably repressed, but instead maintain a "bivalent" chromatin state, with their chromatin enriched for the activating histone mark, H3 lysine 4 tri-methylation (K4me3), together with the repressive K27me3 [6,7]. In the absence of transcriptional induction, PRC1 and PRC2 remain at target loci and mediate repression through differentiation. The mechanisms that underlie stable association of the complexes remain poorly understood, but likely involve interactions with the modified histones [8-12].

Proper localization of PRC1 and PRC2 in the pluripotent genome is central to the complex developmental regulation orchestrated by these factors. However, the sequence determinants that underlie this initial landscape remain obscure. Polycomb recruitment is best understood in *Drosophila*, where sequence elements termed Polycomb response elements (PREs) are able to direct these repressors to exogenous locations [13]. PREs contain clusters of motifs recognized by DNA binding proteins such as Pho, Zeste and GAGA, which in turn recruit PRC2 [14-17]. Despite extensive study, neither PRE sequence motifs nor binding profiles of PRC2-associated DNA binding proteins are sufficient to fully predict PRC2 localization in the *Drosophila* genome [1,16,18,19].

While protein homologs of PRC1 and PRC2 are conserved in mammals, DNA sequence homologs of *Drosophila* PREs appear to be lacking in mammalian genomes [13]. Moreover, it remains controversial whether the DNA binding proteins associated with PRC2 in *Drosophila* have functional homologs in mammals. The most compelling candidate has been YY1, a Pho homolog that rescues gene silencing when introduced into Pho-deficient *Drosophila* embryos [20]. YY1 has been implicated in PRC2-dependent silencing of tumor suppressor genes in human cancer cells [21]. However, this transcription factor has also been linked to numerous other functions, including imprinting, DNA methylation, B-cell development and ribosomal protein gene transcription [22-26].

Recently, researchers identified two DNA sequence elements able to confer Polycomb repression in mammalian cells. Sing and colleagues identified a murine PRE-like element that regulates the MafB gene during neural development [27]. These investigators defined a critical 1.5 kb sequence element that is able to recruit PRC1, but not PRC2 in a transgenic cell assay. Woo and colleagues identified a 1.8 kb region of the human HoxD cluster that recruits both PRC1 and PRC2 and represses a reporter construct in mesenchymal tissues [28]. Both groups note that their respective PRE regions contain YY1 motifs. Mutation of the YY1 sites in the HoxD PRE resulted in loss of PRC1 binding and partial loss of repression, while comparatively, deletion of a separate highly conserved region from this element completely abrogated PRC1 and PRC2 binding as well as repression [28].

In addition to these locus-specific investigations, genomic studies have sought to define PRC2 targets and determinants in a systematic fashion. The Ezh2 and Suz12 subunits have been mapped in mouse and human ES cells by chromatin immunoprecipitation and microarrays (ChIP-chip) or high-throughput sequencing (ChIP-Seq) [3-5,29]. Such studies have highlighted global correlations between PRC2 targets and CpG islands [5,30] as well as highly-conserved genomic loci [4,7,31]. Recently, Jarid2 has been shown to associate with PRC2 and to be required for proper genome-wide localization of the complex [32-35]. Intriguingly, Jarid2 contains an ARID and a Zinc-finger DNA-binding domain. However, it is unclear how Jarid2 could account for PRC2 targeting given the lack of sequence specificity and the low affinity of its DNA binding domains [33,36]. In summary, a variety of sequence elements including CpG islands, conserved elements and YY1 motifs have been implicated in Polycomb targeting in mammalian cells. Causality has only been demonstrated in two specific instances and a unifying view of the determinants of Polycomb recruitment remains elusive.

Here we present the identification of multiple sequence elements capable of recruiting PRC2 in mammalian ES cells. This was achieved through an experimental approach in which engineered bacterial artificial chromosomes (BACs) were stably integrated into the ES cell genome. Evaluation of a series of modified BACs specifically identified a 1.7 kb DNA fragment that is both necessary and sufficient for PRC2 recruitment. The fragment does not share sequence characteristics of *Drosophila* PREs and lacks YY1 binding sites, but rather corresponds to an annotated CpG island. Based on this result and a genome-wide analysis of PRC2 target sequences we hypothesized that large GC-rich sequence elements lacking transcriptional activation signals represent general PRC2 recruitment elements. We tested this model by assaying the following DNA sequences: (i) a 'housekeeping' CpG island which was re-engineered by removal of a cluster of activating motifs; and (ii) two large GC-rich intervals from the *E. coli* genome that satisfy the criteria of mammalian CpG islands. We found that all three GC-rich elements robustly recruit PRC2 in ES cells. We propose that a class of CpG islands distinguished by a lack of activating motifs play causal roles in the initial localization of PRC2 and the subsequent coordination of epigenetic controls during mammalian development.

Results

Recruitment of Polycomb repressors to a bacterial artificial chromosome integrated into ES cells

To identify DNA sequences capable of recruiting Polycomb repressors in mammalian cells, we engineered human BACs that correspond to genomic regions bound by these proteins in human ES cells.

We initially targeted a region of the human Zfpm2 (hZfpm2) locus, which encodes a developmental transcription factor involved in heart and gonad development [37]. In ES cells, the endogenous locus recruits PRC1 and PRC2, and is enriched for the bivalent histone modifications, K4me3 and K27me3 (Figure 1A). We used recombineering to engineer a 44 kb BAC containing this locus and a neomycin selection marker. The modified BAC was electroporated into mouse ES cells, and individual transgenic ES cell colonies containing the full length BAC were expanded (Figure 2). Fluorescent in situ hybridization (FISH) confirmed integration at a single genomic location (Figure 3).

We used ChIP and quantitative PCR (ChIP-qPCR) with human specific primers to examine the chromatin state of the newly incorporated hZfpm2 locus. This analysis revealed strong enrichment for K27me3 and K4me3 (Figure 1B). In addition, we explicitly tested for direct binding of the Polycomb repressive complexes using antibody against the PRC1 subunit, Ring1B, or the PRC2 subunit, Ezh2. We detected

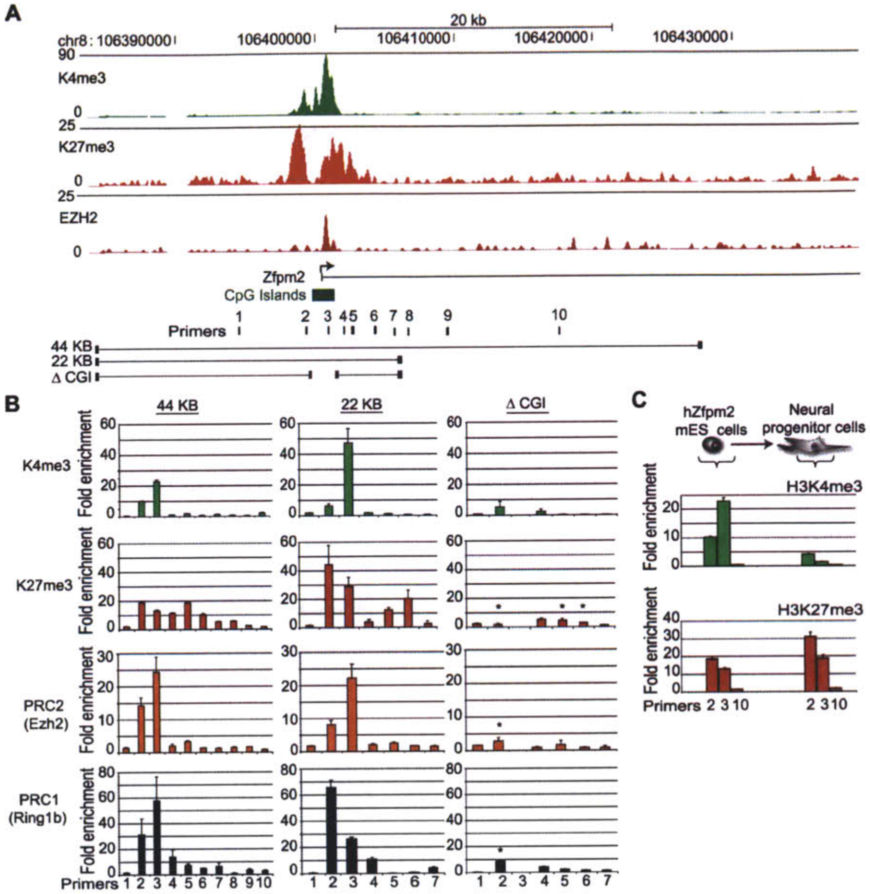


Figure 1: Recruitment of Polycomb repressors to a BAC integrated into ES cells. **(A)** ChIP-Seq tracks depict enrichment of K27me3 (the modification catalyzed by PRC2), Ezh2 (the enzymatic component of PRC2), and K4me3 across the endogenous hZfpm2 locus in human ES cells. Primers and constructs used in this study are indicated below the gene track. **(B)** BAC constructs from (A) containing the hZfpm2 locus were stably integrated into mouse ES cells. ChIP-qPCR enrichments are shown for K4me3, K27me3, Ezh2, and the PRC1 component Ring1b across the locus. The integrated locus adopts a bivalent chromatin state with K27me3 and K4me3 in all constructs except the ΔCGI BAC. The locations of PCR amplicons are designated on the horizontal axis. **(C)** Transgenic ES cells differentiated along a neural lineage show enrichment for K27me3 but not K4me3 in NP cells. Error bars show standard error of the mean (SEM) for n = 3 (44 kb) or n = 2 (22 kb; ΔCGI) biological replicates.

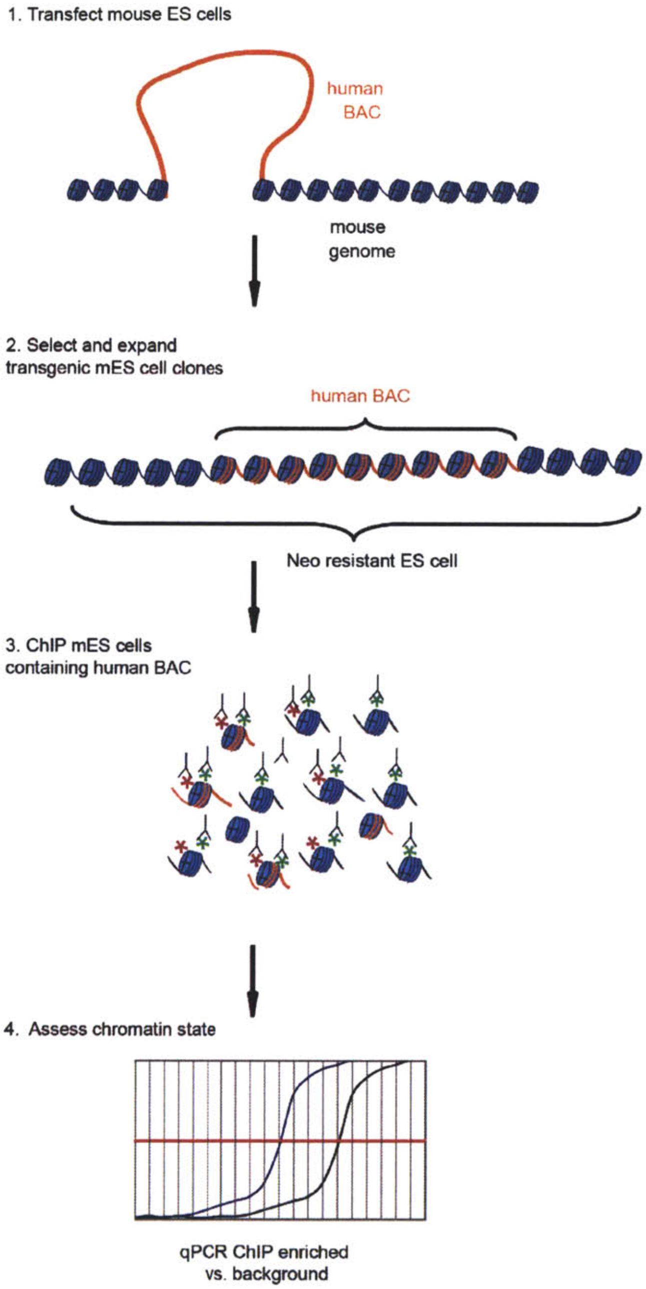


Figure 2: A schematic of the transgenic chromatin assay that was used to examine the role of DNA sequence in determining histone modification patterns in embryonic stem cells.

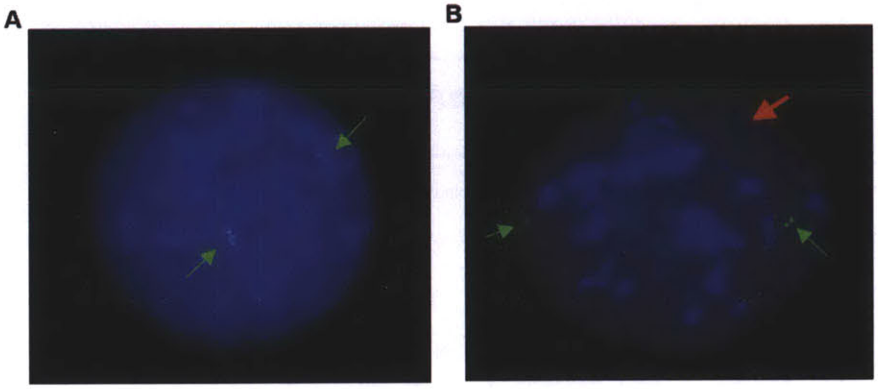


Figure 3: Transgenic mouse ES cells and associated mouse feeder cells were probed by FISH using Human BAC CTD331719L (hZFPM2), labeled with Cy3-dUTP (red), and a control mouse probe BAC (RP23-442F1, located on mouse chromosome 15), labeled with FITC-dUTP (green) along with DNA stained with DAPI (blue). A MEF feeder cell **(A)** shows two copies of the mouse probe (green arrows), and lacks a copy of hZfpm2. A transgenic ES cell **(B)** shows two copies of the mouse probe (green arrows) and one copy of hZFPM2 probe (red arrow).

robust enrichment for both complexes in the vicinity of the hZfpm2 gene promoter (Figure 1B). To confirm this result and eliminate the possibility of integration site effects, we tested two additional transgenic hZfpm2 ES cell clones with unique integration sites as well as a fourth transgenic ES cell line containing a distinct Polycomb target locus, Pax5. In each case, we observed a bivalent chromatin state analogous to the endogenous loci (Figure 4). Similar to endogenous bivalent CpG islands, we found the Zfpm2 CpG island was DNA hypomethylated (Figure 5). These results suggest that DNA sequence is sufficient to initiate de novo recruitment of Polycomb in ES cells.

The Zfpm2 BAC maintains K27me3 through ES cell differentiation

A key function of Polycomb repressors is to maintain a repressive chromatin state through cellular differentiation. To determine if the integrated BAC is capable of maintaining K27me3, the hZfpm2 transgenic ES cells were differentiated to neural progenitor (NP) cells in vitro [38]. ChIP-qPCR analysis revealed continued enrichment of K27me3 but loss of K4me3 (Figure 1C), a pattern frequently observed at endogenous loci that are not activated during differentiation [39]. This indicates that DNA sequence at the hZfpm2 locus is sufficient to initiate K27me3 chromatin modifications in ES cells, and maintain the repressive chromatin state through neural differentiation.

Distinguishing Polycomb recruiting sequences in the Zfpm2 BAC

We next sought to define the sequences within the hZfpm2 BAC required for recruitment of Polycomb repressors. First, we re-engineered the 44 kb hZfpm2 BAC to remove 20 kb of flanking sequences that contained distal non-coding conserved sequence elements (Figure 1A). When we integrated the resulting 22 kb construct into ES cells we found that it robustly enriches for PRC1, PRC2, K4me3 and K27me3 (Figure 1B). Hence, these particular distal elements do not appear to be required for the recruitment of the complexes. Next, we considered the necessity of the CpG island which corresponds to the peak of Ezh2 enrichment in ChIP-Seq profiles (Figure 1A). We excised a 1.7 kb fragment containing the CpG island, and integrated the resulting

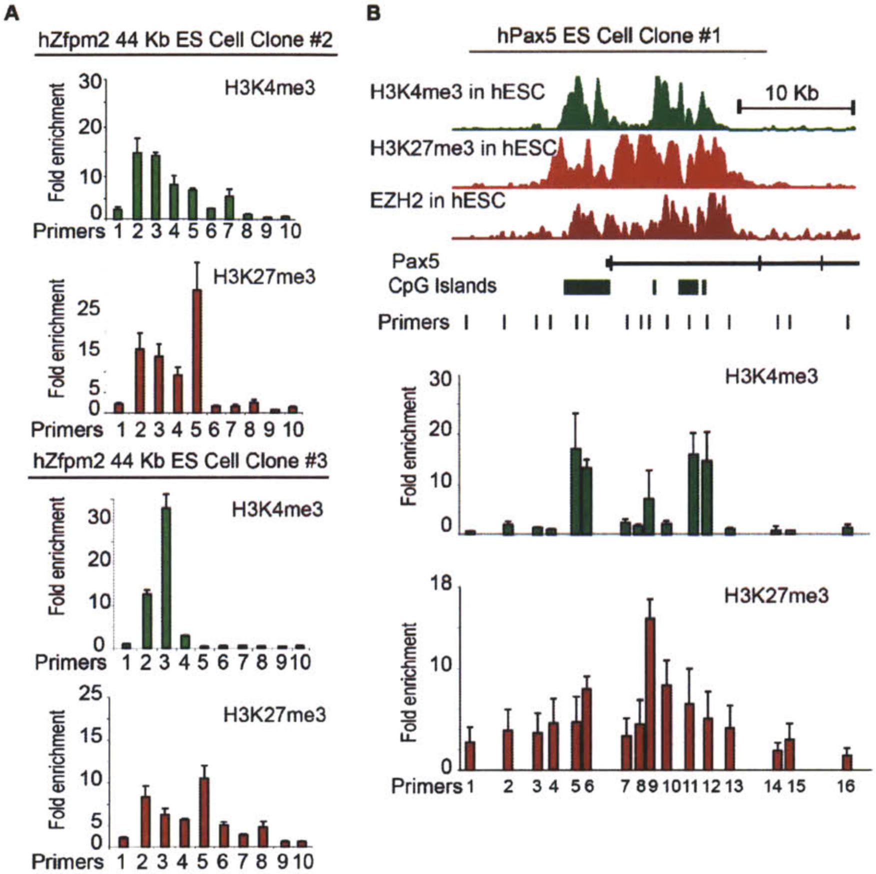


Figure 4: Confirmation of transgenic BAC chromatin state. **(A)** Two additional mES cell clones containing the 44 kb hZfpm2 locus were examined using ChIP-qPCR similar to Figure 1C. Both show enrichment of H3K4me3 and H3K27me3 across the gene promoter. **(B)** ChIP-seq map of the human Pax5 locus in human ES cells show broad regions of H3K4me3 and H3K27me3 enrichment. Bottom panel shows ChIP-qPCR of transgenic mouse ES cells carrying a 50 kb region of the hPax5 locus showing a similar enrichment of H3K4me3 and H3K27me3 across the region. (Error bars represent SEM, n = 3).

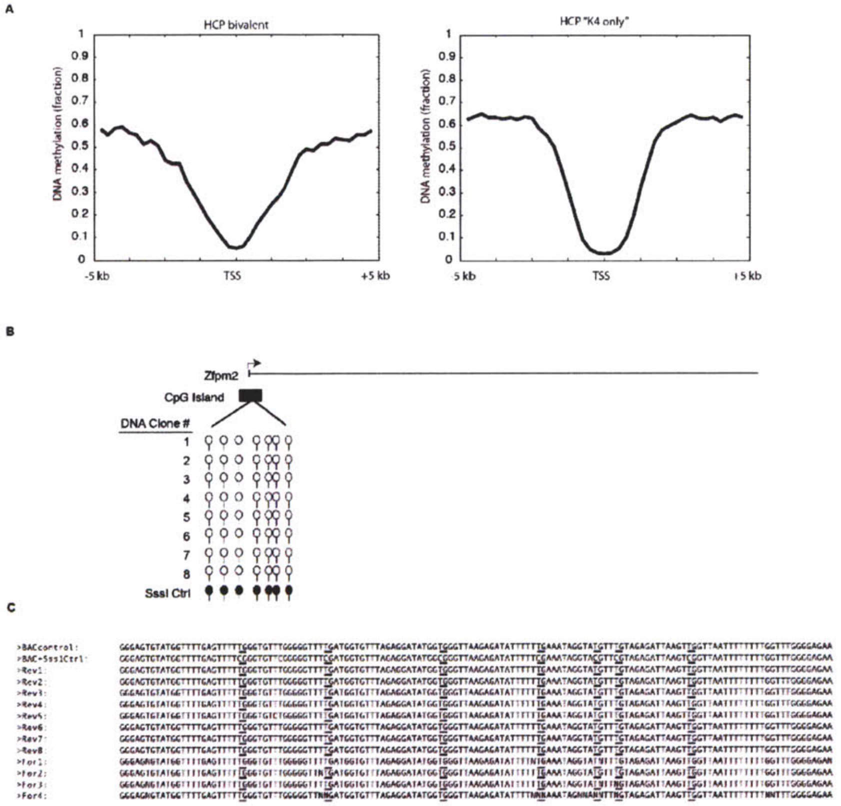


Figure 5: The BAC CpG island remains hypomethylated. **(A)** Composite plots showing the lack of DNA methylation at both bivalent and K4me3 only promoters in mouse ES cells. **(B)** Schematic showing the CpG island of the Zfpm2 BAC remains free of DNA methylation upon integration into mouse ES cells. **(C)** The raw data used to create (B) shows aligned sequencing reads of Zfpm2 ES cell genomic DNA that was bisulfite treated (see Methods). Unmethylated and in vitro methylated BAC DNA are shown as controls. The underlined bases indicate sites of CG dinucletides.

BAC (ΔCGI) into ES cells. The ΔCGI BAC failed to recruit PRC1 or PRC2, and showed significantly reduced K27me3 levels relative to the other constructs (Figure 1B). This suggests that the CpG island is essential for recruitment of Polycomb proteins to the hZfpm2 locus.

A 1.7 kb CpG island is sufficient to recruit PRC2 to an exogenous locus

We next asked whether the hZfpm2 CpG island is sufficient to recruit Polycomb repressors to an exogenous locus. To test this, we selected an unremarkable gene desert region on human chromosome 1 that shows no enrichment for PRC1, PRC2 or K27me3 in ES cells (Figure 6A). We also verified that the gene desert BAC alone does not show any enrichment for K27me3 or Ezh2 when integrated into ES cells (Figure 6B). Using recombineering, we inserted the 1.7 kb sequence that corresponds to the hZfpm2 CpG island into the gene desert BAC. The resulting construct was integrated into mouse ES cells and three independent clones were evaluated. ChIP-qPCR analysis revealed strong enrichment for K27me3, K4me3 and PRC2 over the inserted CpG island (Figure 6C, Figure 7). In contrast, we observed relatively little enrichment for the PRC1 subunit Ring1B (Figure 6C). We confirmed the specificity of these enrichments with primers that span the boundary between the insertion and adjacent gene desert sequence. Notably, K27me3 enrichment was detected across the gene desert locus up to 2.5 kb from the inserted CpG island (Figure 6C). This indicates that the localized CpG island can initiate K27me3 that then spreads into adjacent sequence. Lastly we found no YY1 enrichment across the CpG island by ChIP-qPCR (Figure 7). Together, these data suggest that the hZfpm2 CpG island contains the necessary signals for PRC2 recruitment but is insufficient to confer robust PRC1 association.

Consideration of sequence determinants of PRC2 recruitment

The functionality of a CpG island in PRC2 recruitment is consistent with prior observations that a majority of PRC2 sites in ES cells correspond to CpG islands [4,5] and with the striking correlation between intensity of PRC2 binding and the GC-richness of the underlying sequence (Figure 6D). We therefore considered whether

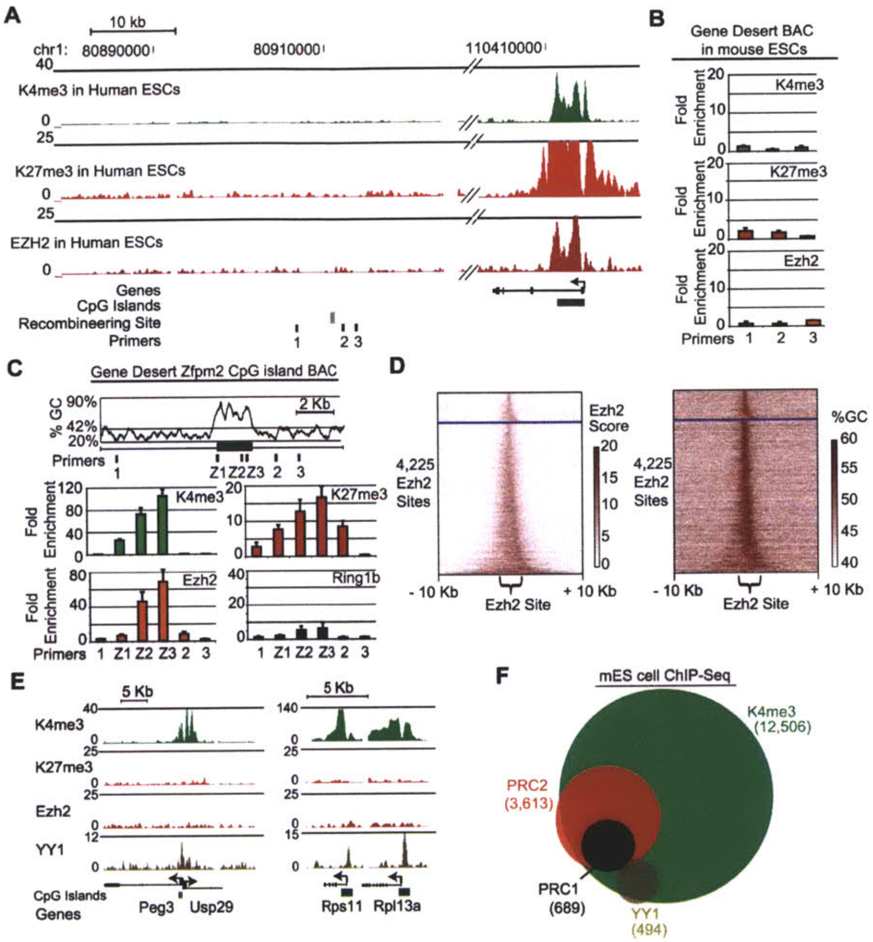


Figure 6: A 1.7 kb GC-rich sequence element is sufficient to recruit PRC2. **(A)** ChIP-Seq tracks show no enrichment for K4me3, K27me3 or Ezh2 in human ES cells across the gene desert region. For comparison a nearby locus is shown. The recombineering site and primers used in this study are indicated below the tracks. **(B)** The gene desert BAC shows no enrichment of K4me3, K27me3 or PRC2 upon integration in mouse ES cells. **(C)** The hZfpm2 CpG island is depicted at the site of insertion into the gene desert BAC, along with the corresponding GC percentage (42% indicates genome average) and primers used for qPCR. Underlying plots represent ChIP-qPCR enrichment of K4me3, K27me3, PRC2 (Ezh2), and PRC1 (Ring1b) at the indicated sites (n = 2 biological replicates). **(D)** Heat maps show Ezh2 ChIP-Seq signal (left panel) or GC-percentage (right panel) for all Ezh2-bound regions in ES cells. Each row depicts a 20 kb region centered on the Ezh2 signal. Rows are separated into two groups based on whether the site overlaps a CpG island (below the blue line) and are then sorted based on the width of Ezh2 enrichment (see Methods). **(E)** ChIP-Seq was used to profile the mammalian Pho homolog YY1 in mouse ES cells. Genome browser views show ChIP-Seq enrichment signals for K4me3, K27me3, Ezh2 and YY1 for YY1 target loci. **(F)** Venn diagram shows overlap of K4me3, Ezh2, Ring1b, and YY1 at promoters in mES cells.

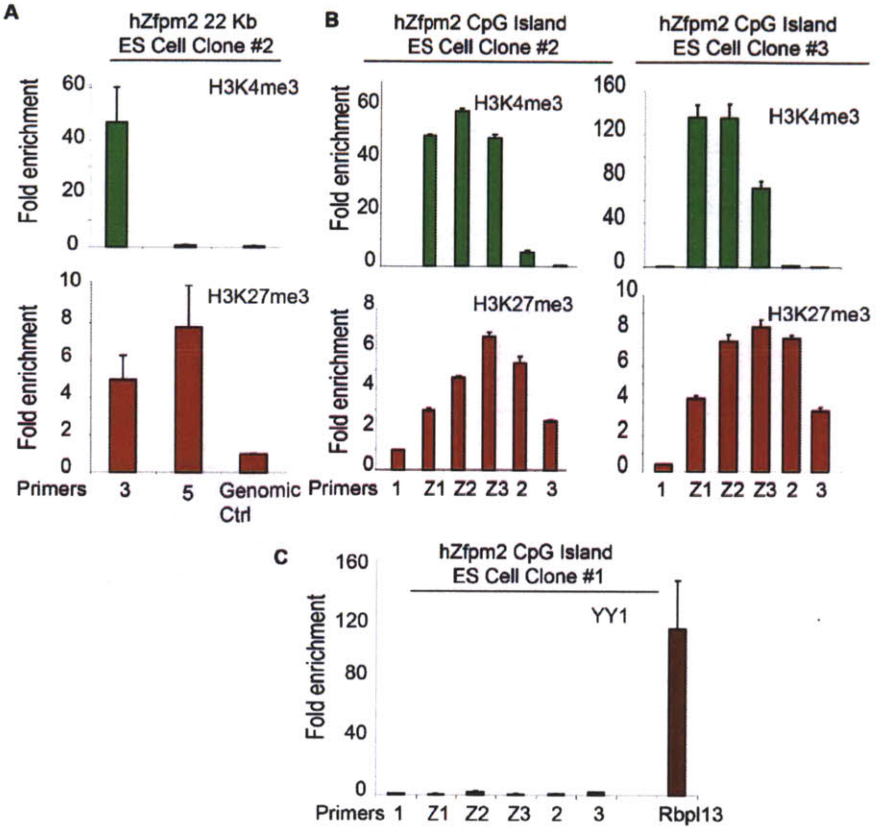


Figure 7: Independent validation of different BAC clones. **(A)** One additional mES cell clone containing 22 kb of the hZfpm2 locus was examined using ChIP-qPCR. As seen with the first clone (Figure 1B) this clone also shows enrichment of H3K4me3 and H3K27me3 at the gene promoter. **(B)** Additional clones of transgenic ES cells containing the Gene Desert BAC with the hZfpm2 CpG island inserted show enrichment of H3K4me3 and H3K27me3 as seen with clone 1 (Figure 6C). **(C)** The Zfpm2 Gene Desert BAC shows no enrichment of YY1, in contrast to the promoter of Rpl13a. Error bars equal to SEM (n = 2). Genomic Ctrl = mouse neg genomic control.

specific signals within the Zfpm2 CpG island might underlie its capacity to recruit PRC2.

First, we searched for sequence motifs analogous to the PREs that recruit PRC2 in *Drosophila*. We focused on motifs recognized by YY1, the nearest mammalian homolog of the *Drosophila* recruitment proteins. Notably, both of the recently described mammalian PREs contain YY1 motifs [27,28]. The 44 kb hZfpm2 BAC contains 11 instances of the consensus YY1 motif. However, none of these reside within the CpG island (Figure 8) (see Methods). We also examined YY1 binding directly in ES cells and NS cells using ChIP-Seq. Consistent with prior reports, YY1 binding is evident at the 5'ends of many highly expressed genes, including those encoding ribosomal proteins, and is also seen at the imprinted Peg3 locus (Figure 6E) [26]. However, no YY1 enrichment is evident at the Zfpm2 locus. Moreover, at a global level, YY1 shows almost no overlap with PRC2 or PRC1, but instead co-localizes with genomic sites marked exclusively by K4me3 (Figure 6F, Figure 8). Thus, although YY1 may contribute to Polycomb-mediated repression through distal interactions or in trans, it does not appear to be directly involved in PRC2 recruitment in ES cells.

We previously reported that CpG islands bound by PRC2 in ES cells could be predicted based on a relative absence of activating transcription factor motifs (AMs) in their DNA sequence [5]. We reasoned that transcriptional inactivity afforded by this absence of AMs is a requisite for PRC2 association [40,41]. This could explain why PRC2 is absent from a majority of CpG islands, many of which are found at highly active promoters. Consistent with this model, when we examined a recently published RNA-Seq dataset for poly-adenylated transcripts in ES cells, we found that virtually all of the high-CpG promoters (HCPs) lacking Ezh2 are detectably transcribed (Figure 9). The small proportion of HCPs that are neither Ezh2-bound nor transcribed may reflect false-negatives in the ChIP-Seq or RNA-Seq data. Alternatively, these HCPs tend to correspond to CpG islands with relatively low GC-contents and lengths and may therefore have insufficient GC-richness to promote PRC2 binding (Figure 9). Thus, correlative analyses implicate large GC-rich elements that lack transcriptional activation signals as general PRC2 recruitment elements in mammals.

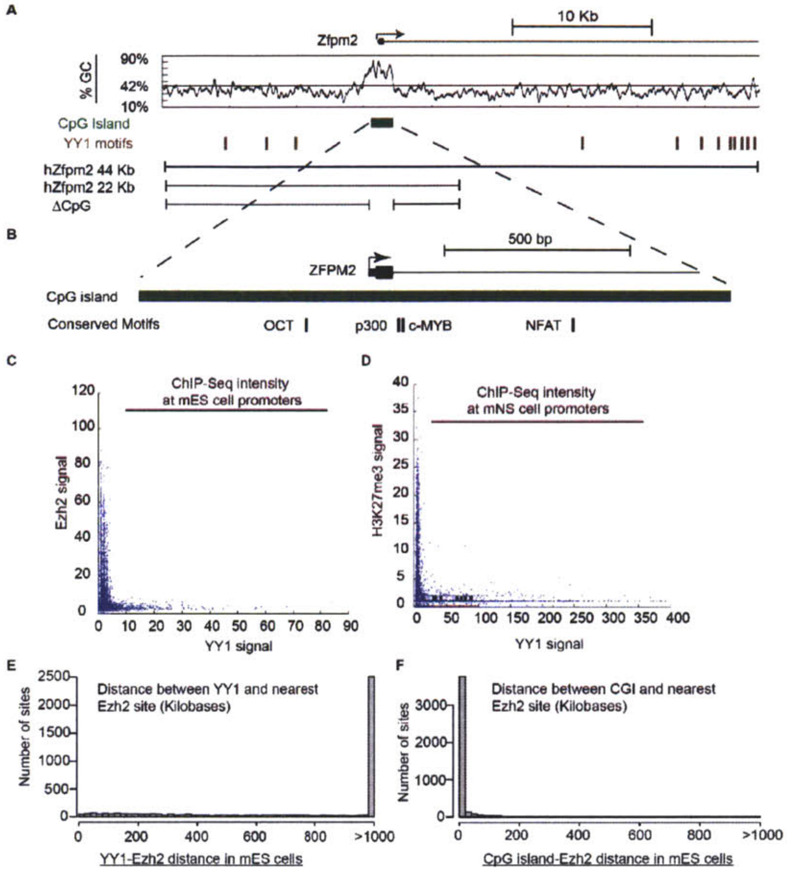


Figure 8: Motif content and YY1 binding relative to CpG island observations. **(A)** The GC-richness and locations of YY1 motifs for the Zfpm2 locus are shown. **(B)** The 1.7 kb CpG island contains 4 conserved motifs (see Methods). **(C)** PRC2 signal is inversely correlated with YY1 signal at 17,761 promoters in mouse ES cells. **(D)** PRC2 activity as measured by K27me3 also shows an inverse correlation with YY1 in mouse neural stem (NS) cells. **(E)** Genome-wide binding profiles show YY1 is predominantly over 1 mb away from the nearest Ezh2 site. By comparison CpG islands **(F)** show close proximity to Ezh2 sites in ES cells.

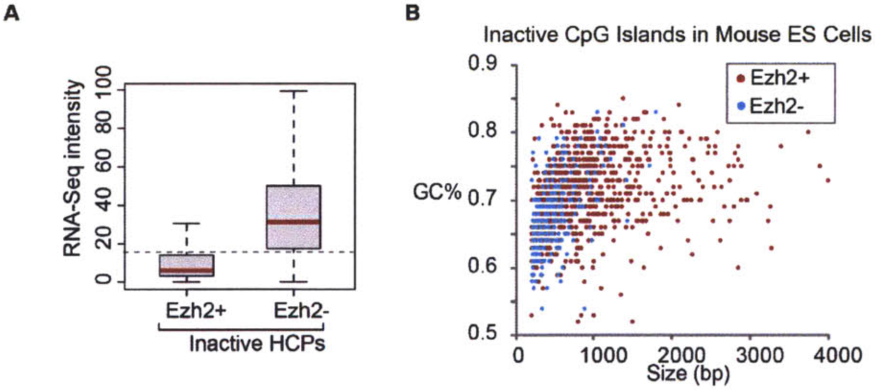


Figure 9: Characteristics of active versus inactive CpG islands. **(A)** Analysis of gene promoters with high CpG content (HCPs) shows Ezh2 positive promoters have significantly lower RNA-Seq scores compared to Ezh2 negative promoters. The dashed line represents the highest expression seen at LCPs. All transcriptionally inactive HCPs containing a single CpG island were scored for Ezh2 enrichment (see text and Methods). **(B)** The scatter plot indicates length and %GC for Ezh2-positive and Ezh2-negative CpG islands with low RNA-Seq scores in mouse ES cells.

Sufficiency of GC-rich sequences for PRC2 recruitment

To obtain direct experimental support for the general sufficiency of large GC-rich elements lacking AMs in PRC2 recruitment, we carried out the following experiments. First, we tested whether a K4me3-only CpG island could be turned into a PRC2 recruitment element by removing activating motifs. We targeted a 1.3 kb CpG island that overlaps the promoters of two ubiquitously expressed genes – Arl3 and Sfxn2. Neither gene carries K27me3 in ES cells, or in any other cell type tested (Figure 10, and data not shown). This CpG island was selected as it has many conserved AMs clustered in one half of the island (Figure 11A). We hypothesized that the portion of the Arl3/Sfxn2 CpG island lacking AMs would, in isolation, lack active transcription and recruit PRC2. In contrast, we predicted that the half containing multiple AMs would lack Polycomb. To test this, we generated two additional BAC constructs containing the respective portions of the Arl3/Sfxn2 CpG island positioned within the gene desert, and integrated these constructs into ES cells (Figure 11A). ChIP-qPCR shows that the portion of the CpG island lacking AMs is able to recruit PRC2 and becomes enriched for K27me3 (Figure 11B). In contrast, the AM-containing portion shows no enrichment for K27me3 or Ezh2, but is instead marked exclusively by K4me3, similar to the endogenous human locus (Figure 11C, Figure 10). Thus, a GC-rich sequence element with no known requirement for Polycomb regulation can recruit PRC2 when isolated from activating sequence features.

Next, we tested whether even more generic GC-rich elements might also be capable of recruiting PRC2 in ES cells. Here, we focused on sequences derived from the genome of *E. coli*, reasoning that there would be no selection for PRC2 recruiting elements in this prokaryote given the complete lack of chromatin regulators. We arbitrarily selected three 1 kb segments of the *E. coli* genome. Two with GC contents above the threshold for a mammalian CpG island but that each contained few AMs, and one AT rich segment as a control. We recombined each segment into the gene desert BAC and integrated the resulting constructs into ES cells. ChIP-qPCR confirmed that both GC-rich *E. coli* segments recruit Ezh2 and form a bivalent chromatin state (Figure 12A,B, Figure 13). Notably, the GC-rich segment also enriches for

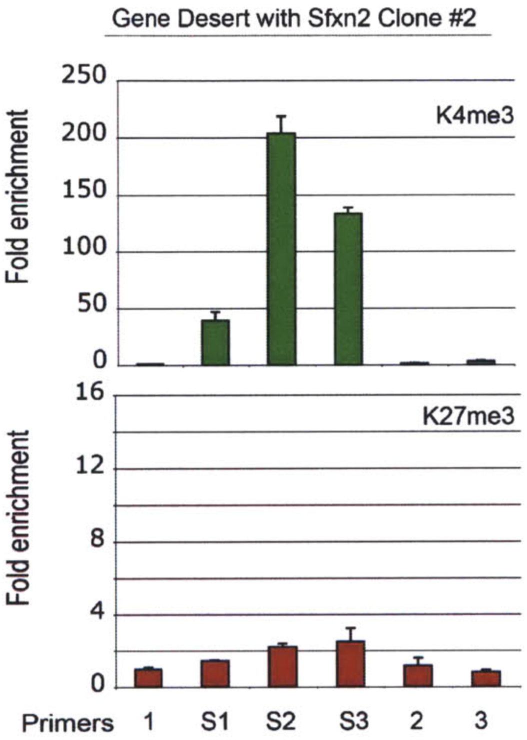


Figure 10: One additional mES cell clone containing gene desert BAC with the Sfxn2 CpG island was examined using ChIP-qPCR. As seen with the first clone (Figure 11B) this clone also shows significant enrichment of H3K4me3 but not H3K27me3 at the CpG island. Error Bars represent SEM (n = 2).

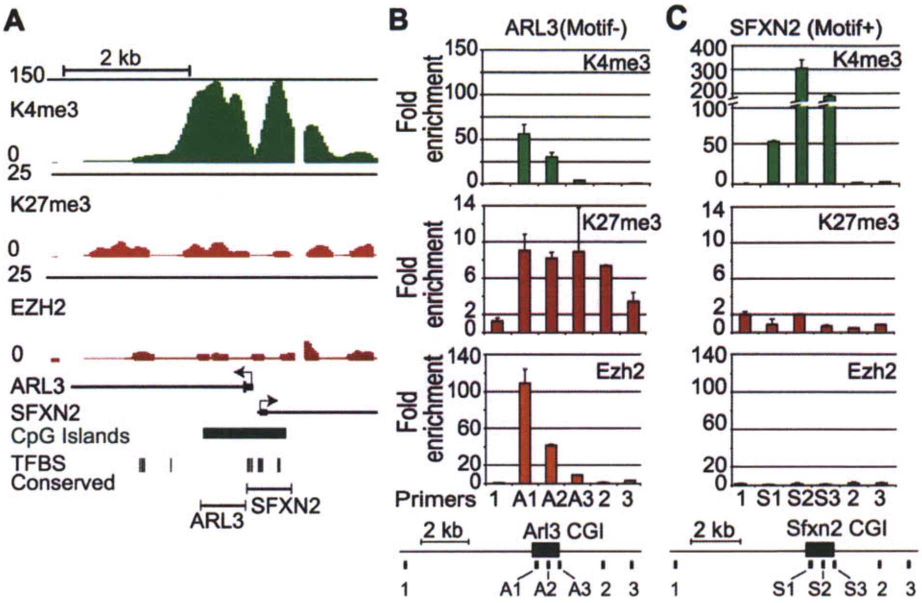


Figure 11: Removal of activating transcription factor motifs initiates PRC2 recruitment. **(A)** Genome browser views shows a locus containing the promoters for the housekeeping genes Arl3 and Sfxn2 with ChIP-Seq enrichment signals for K4me3, K27me3, and Ezh2 in mouse ES cells. This region contains a 1.8 kb CpG island that has the transcription factor motifs clustered on one side. Below shows the regions used for integration into the gene desert BAC. **(B)** After integration into mouse ES cells, ChIP-qPCR was conducted using three primers from the CpG island inserts and 3 primers in the flanking gene desert sequence. The motif devoid Arl3 section shows de novo PRC2 (Ezh2) recruitment and K4me3 and K27me3 enrichment. **(C)** The motif containing Sfxn2 half shows no enrichment for K27me3 but significant enrichment for K4me3, similar to the endogenous locus shown in (A) (n = 2 biological replicates).

Jarid2, a PRC2 component with DNA binding activity (Figure S10). In contrast, the AT-rich segment did not recruit Ezh2 or enrich for either K4me3 or K27me3 (Figure 12C, Figure 13). Together, our findings suggest that GC-rich sequence elements that lack signals for transcriptional activation have an innate capacity to recruit PRC2 in mammalian ES cells.

Discussion

Several lines of evidence suggest that the initial landscape of Polycomb complex binding is critical for proper patterning of gene expression in metazoan development [1,2,13]. Failure of these factors to engage their target loci in embryogenesis has been linked to a loss of epigenetic repression at later stages. Accordingly, the determinants that localize Polycomb complexes at the pluripotent stage are almost certainly essential to the global functions of these repressors through development.

We find that DNA sequence is sufficient for proper localization of Polycomb repressive complexes in ES cells, and specifically identify a CpG island within the Zfpm2 locus as being critical for recruitment. We provide evidence that GC-rich elements lacking activating signals suffice in general to recruit PRC2. This includes demonstrations (i) that a motif devoid segment of an active 'housekeeping' CpG island can recruit PRC2; and (ii) that arbitrarily selected GC-rich elements from the *E. coli* genome can themselves mediate PRC2 recruitment when integrated into the ES cell genome.

Several possible mechanistic models could explain the causality of GC-rich DNA elements in PRC2 recruitment (Figure 15). First, we note that CpG islands have been shown to destabilize nucleosomes in mammalian cells [42]. At transcriptionally inactive loci, this property could increase their accessibility to PRC2-associated proteins with DNA affinity but low sequence specificity, such as Jarid2 or AEBP2 [32-35,43] (Figure 14). Although this association would be abrogated by transcriptional activity at most CpG islands, those lacking activation signals would remain permissive to PRC2 association (Figure 15). In support of this model, PRC2 targets in ES cells are also enriched for H2A.Z and H3.3, histone variants linked to nucleosome exchange

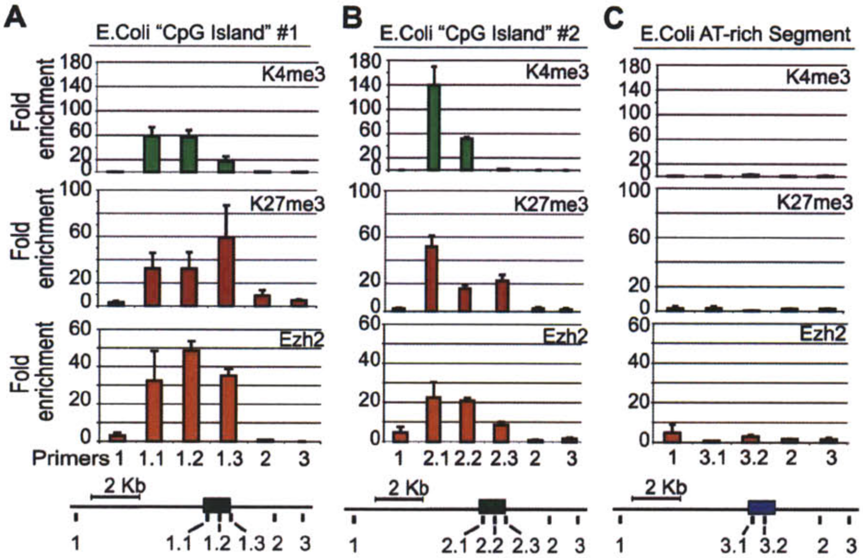


Figure 12: PRC2 is recruited to *E.coli* GC-rich sequences in mouse ES cells. The *E. coli* genome was scanned for 1 kb regions that met the criteria for a mammalian CpG island and had few motifs for mammalian transcription factors (see Methods). **(A,B)** Both GC-rich segments adopt a bivalent chromatin state with K27me3 and K4me3 and recruit PRC2 (Ezh2) upon integration in mouse ES cells. **(C)** A non-CG rich region of the *E. coli* genome failed to recruit Ezh2 and lacked K4me3 and K27me3 (n = 2 biological replicates).

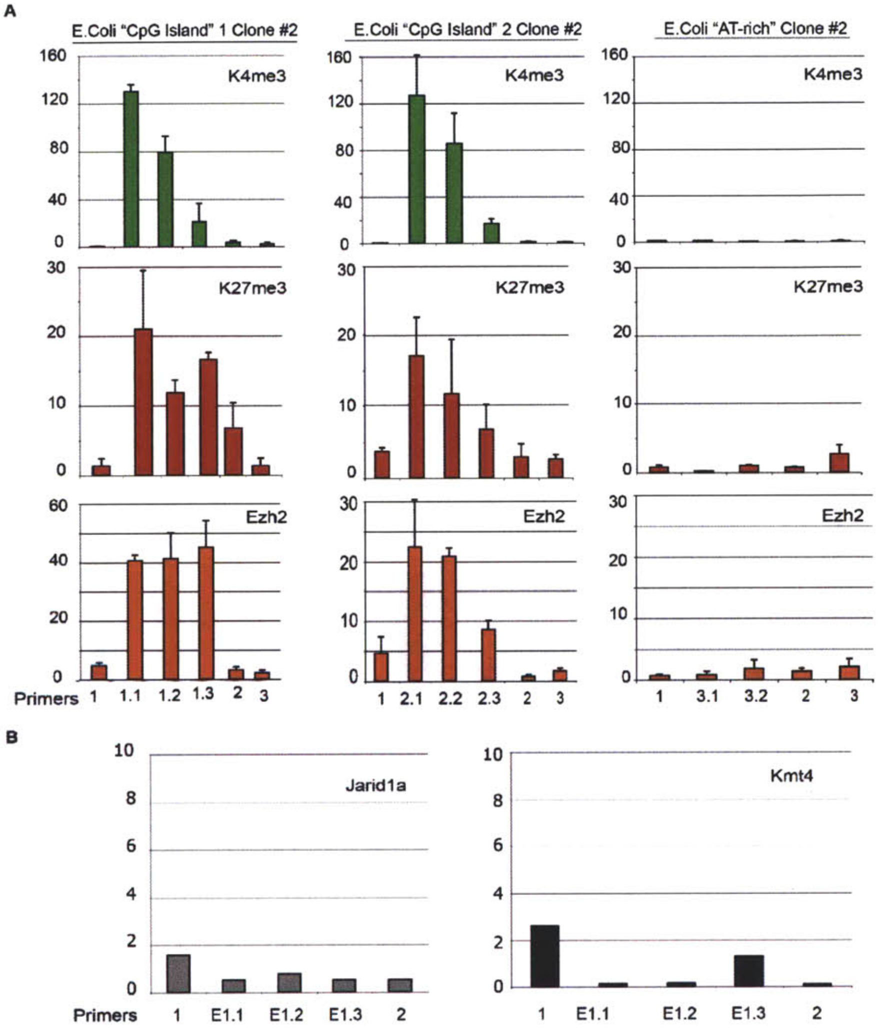


Figure 13: Confirmation of *E. coli* PREs. **(A)** One additional mES cell clone for each *E. coli* DNA construct was analyzed by ChIP-qPCR. As seen with the first clones (Figure 12A-C) the CpG island clones show significant enrichment of K4me3, K27me3 and Ezh2 at the gene promoter. Error Bars represent SEM (n = 2) **(B)** As a negative control, *E. coli* CpG island 1 was also tested for the chromatin modifiers Jarid1a and Kmt4, which showed no enrichment.

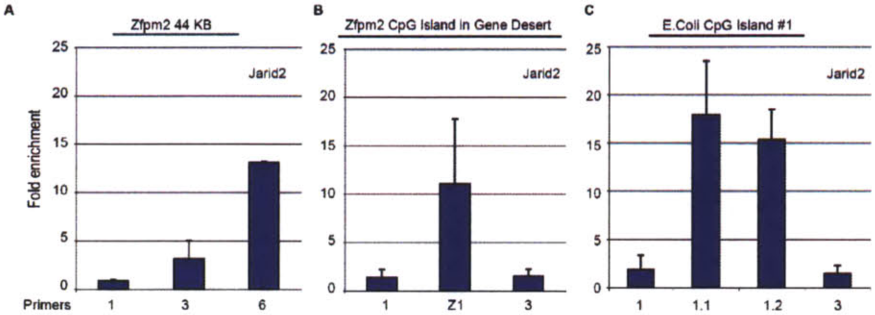


Figure 14: ChIP-qPCR shows Jarid2 enrichment signal at the CpG island (primer 6) of the 44 kb BAC **(A)**, the Zfpm2 CpG island (primer Z1) within the Gene Desert BAC **(B)** and the GC-rich element (primers 1.1, 1.2) from *E. coli* **(C)**. Error Bars represent SEM (n = 2).

dynamics [44,45]. Alternatively or in addition, targeting could be supported by DNA binding proteins with affinity for low complexity GC-rich motifs or CpG dinucleotides, such as CXXC domain proteins [46]. Localization may also be promoted or stabilized by long and short non-coding RNAs [47-50] as well as by the demonstrated affinity of PRC2 for its product, H3K27me3 [11,12]. Notably, PRC2 recruitment in ES cells appears distinct from that in *Drosophila*, as we do not find evidence for involvement of PRE-like sequence motifs or mammalian homologues such as YY1.

It should be emphasized that PRC2 localization does not necessarily equate with epigenetic repression. Indeed virtually all PRC2 bound sites in ES cells, and all CpG islands tested here, are also enriched for K4me3, and presumably poised for activation upon differentiation. Epigenetic repression during differentiation may require PRC1 and thus depend on additional binding determinants. YY1 remains an intriguing candidate in this regard, given prior evidence for physical and genetic interactions with PRC1 [51,52]. YY1 consensus motifs are present in the Polycomb-dependent silencing elements recently identified in the MafB and HoxD loci. Interestingly, the HoxD element combines a CpG island with a cluster of conserved YY1 motifs. Mutation of the motifs abrogated PRC1 binding but left PRC2 binding intact. Still, the fact that only a small fraction of documented PRC2 and PRC1 sites have YY1 motifs or binding suggests that this transcription factor may act indirectly and/or explain only a subset of cases. Nonetheless, it is likely that a fully functional epigenetic silencer would require a combination of features, including a GC-rich PRC2 element as well as appropriate elements to recruit PRC1. Further study is needed to expand the rules for PRC2 binding to include a global definition of PRC1 determinants and ultimately, to understand how the initial landscape facilitates the maintenance of gene expression programs in the developing organism.

Methods

BAC construct design

BAC constructs CTD331719L ('Zfpm2 44'), CTD-2535J16 ('Pax5') and CTD-3219L19 ('Gene Desert') were obtained from Open Biosystems. Recombineering was

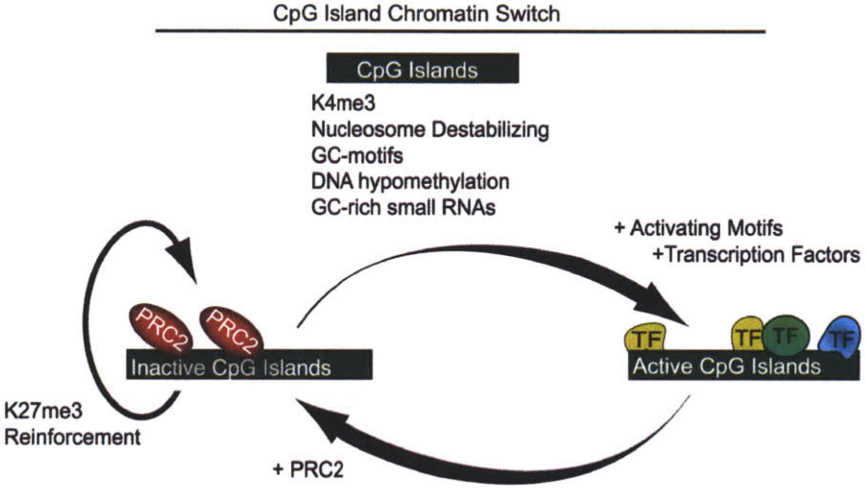


Figure 15: A model showing CpG islands as a chromatin switch. Features common to both active and inactive CpG islands include destabalization of nucleosomes, simple GC-motifs, K4me3 and lack of DNA methylation. Additionally, many CpG island transcribe small non-coding GC-rich RNAs. Active CpG islands contain motifs associated with numerous activating transcription factors and transcriptional machinery, which likely prevent PRC2 from binding. In contrast, CpG islands lacking activating motifs are bound by PRC2 which, through a positive feedback loop with K27me3, maintains an inactive state.

done using the RedET system (Open Biosystems) in DH10B cells. Homology arms 200-500 bp in length were PCR amplified and cloned into a PGK; Neomycin cassette (Gene Bridges). This cassette was used to recombineer all BACs to enable selection in mammalian cells. The 22 kb hZfpm2 BAC was created by restricting the hZfpm2 BAC at two sites using ClaI, and re-ligating the BAC lacking the intervening sequence. The CpG island was excised from the 22 kb hZfpm2 BAC by amplification of flanking homology arms, and cloned into a construct containing an adjacent ampicillin cassette (Frt-amp-Frt; Gene Bridges). After recombination, the ampicillin cassette was removed using Flp-recombinase and selection for clones that lost ampicillin resistance (Flp-706; Gene Bridges). PCR across the region confirmed excision of the CpG island. For the Gene Desert BACs, the Zfpm2, Arl3, Sfxn2 and *E. coli* CpG islands were amplified with primers containing XhoI sites and cloned into the Frt-amp-Frt vector that contains homology arms from the Gene Desert region. The final constructs were confirmed by sequencing across recombination junctions.

Transgenic ES cell and ChIP experiments

ES cells (V6.5) were maintained in ES cell medium (DMEM; Dulbecco's modified Eagle's medium) supplemented with 15% fetal calf serum (Hyclone), 0.1 mM beta-mercaptoethanol (Sigma), 2 mM Glutamax, 0.1 mM non-essential amino acid (NEAA; Gibco) and 1000U/ml recombinant leukemia inhibitory factor (ESGRO; Chemicon). Roughly 50 ug of linearized BAC was nucleofected using the mouse ES cell nucleofector kit (Lonza) into 10^6 mouse ES cells, and selected 7-10 days with 150 ug/ml Geneticin (Invitrogen) on Neomycin resistant MEFs (Millipore). Individual resistant colonies were picked, expanded and tested for integration of the full length BAC by PCR. Differentiation of hZfpm2 ES cell clone 1 into a population of neural progenitor (NP) cells was done as previously described [53]. FISH analysis was done as described previously [54]. DNA methylation analysis was done as previously described [55].

For each construct, between one and three ES cell clones were expanded and subjected to ChIP using antibody against K4me3 (Abcam ab8580 or Upstate/Millipore 07-473), K27me3 (Upstate/Millipore 07-449), Ezh2 (Active Motif 39103 or 39639), or

Ring1B (MBL International d139-3) as described previously [5,7,39]. ChIP DNA was quantified by Quant-iT Picogreen dsDNA Assay Kit (Invitrogen). ChIP enrichments were assessed by quantitative PCR analysis on an ABI 7500 with 0.25 ng ChIP DNA and an equal mass of un-enriched input DNA. Enrichments were calculated from 2 or 3 biologically independent ChIP experiments. For K27me3, and Ezh2 enrichment, background was subtracted by normalizing over a negative genomic control. Error bars represent standard error of the mean (SEM). We confirmed that the human specific primers do not non-specifically amplify mouse genomic DNA.

Genomic and computational analysis

Genomewide maps of YY1 binding sites were determined by ChIP-Seq as described previously [39]. Briefly, ChIP was carried out on $6x10^7$ cells using antibody against YY1 (Santa Cruz Biotechnology sc-1703). ChIP DNA was used to prepare libraries which were sequenced on the Illumina Genome Analyzer. Density profiles were generated as described [39]. Promoters (RefSeq; http://genome.ucsc.edu) were classified as positive for YY1, H3K4me3 or H3K27me3 if the read density was significantly enriched ($p<10^{-3}$) over a background distribution based on randomized reads generated separately for each dataset to account for the varying degrees of sequencing depth. ChIP-Seq data for YY1 are deposited to the NCBI GEO database under the following accession number GSE25197. Sites of Ezh2 enrichment ($p<10^{-3}$) were calculated genomewide using sliding 1 kb windows, and enriched windows within 1 kb were merged. DNA methylation levels were calculated using previously published Reduced Representation Bisulphite Sequenced (RRBS) libraries [55]. Composite plots represent the mean methylation level in sliding 200 bp windows in the the 10 kb surrounding the TSSs of the indicated gene sets.

YY1 motifs were identified using the MAST algorithm [56] where a match to the consensus motif was defined at significance level $5x10^{-5}$. Candidate CpG islands for TF motif analysis were identified by scanning annotated CpG islands for asymmetric clustering of motifs related to transcriptional activation in ES cells [5]. Motifs shown in Figure 11A and Figure 8 are from UCSCs TFBS conserved track. GC-rich elements from the *E. coli* K12 genome were selected by calculating %GC and CpG O/E in sliding 1 kb windows. Sequences matching the criteria for mammalian CpG islands while simultaneously being depleted of motifs related to transcriptional activation [5] were chosen for insertion into mouse ES cells. Transcriptionally inactive HCPs were selected based on a lack of transcript enrichment by both expression arrays [39] and RNA-Seq data [57]. In the case of RNA-Seq, each gene was assigned the maximum read density within any 1 kb window of exonic sequence. To ease analysis of promoter CpG island statistics, only HCPs containing a single CpG island were considered.

References

1. Schuettengruber B, Ganapathi M, Leblanc B, Portoso M, Jaschek R, et al. (2009) Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos. PLoS Biol 7: e1000013. doi:10.1371/journal.pbio.1000013.

2. Schwartz YB, Pirrotta V (2007) Polycomb silencing mechanisms and the management of genomic programmes. Nat Rev Genet 8: 9-22.

3. Boyer LA, Plath K, Zeitlinger J, Brambrink T, Medeiros LA, et al. (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature 441: 349-353.

4. Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, et al. (2006) Control of developmental regulators by Polycomb in human embryonic stem cells. Cell 125: 301-313.

5. Ku M, Koche RP, Rheinbay E, Mendenhall EM, Endoh M, et al. (2008) Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. PLoS Genet 4: e1000242. doi:10.1371/journal.pgen.1000242.

6. Azuara V, Perry P, Sauer S, Spivakov M, Jorgensen HF, et al. (2006) Chromatin signatures of pluripotent cell lines. Nat Cell Biol 8: 532-538.

7. Bernstein B, Mikkelsen T, Xie X, Kamal K, Huebert D, et al. (2006) A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell 125: 315-326.

8. Cao R, Wang L, Wang H, Xia L, Erdjument-Bromage H, et al. (2002) Role of histone H3 lysine 27 methylation in Polycomb-group silencing. Science 298: 1039-1043.

9. Czermin B, Melfi R, McCabe D, Seitz V, Imhof A, et al. (2002) Drosophila enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that

marks chromosomal Polycomb sites. Cell 111: 185-196.

10. Kuzmichev A, Nishioka K, Erdjument-Bromage H, Tempst P, Reinberg D (2002) Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. Genes Dev 16: 2893-2905.

11. Hansen KH, Bracken AP, Pasini D, Dietrich N, Gehani SS, et al. (2008) A model for transmission of the H3K27me3 epigenetic mark. Nat Cell Biol 10: 1291-1300.

12. Margueron R, Justin N, Ohno K, Sharpe ML, Son J, et al. (2009) Role of the polycomb protein EED in the propagation of repressive histone marks. Nature 461: 762-767.

13. Ringrose L, Paro R (2007) Polycomb/Trithorax response elements and epigenetic memory of cell identity. Development 134: 223-232.

14. Simon J, Chiang A, Bender W, Shimell MJ, O'Connor M (1993) Elements of the Drosophila bithorax complex that mediate repression by Polycomb group products. Dev Biol 158: 131-144.

15. Dejardin J, Rappailles A, Cuvier O, Grimaud C, Decoville M, et al. (2005) Recruitment of Drosophila Polycomb group proteins to chromatin by DSP1. Nature 434: 533-538.

16. Tolhuis B, de Wit E, Muijrers I, Teunissen H, Talhout W, et al. (2006) Genome-wide profiling of PRC1 and PRC2 Polycomb chromatin binding in Drosophila melanogaster. Nat Genet 38: 694-699.

17. Wang L, Brown JL, Cao R, Zhang Y, Kassis JA, et al. (2004) Hierarchical recruitment of polycomb group silencing complexes. Mol Cell 14: 637-646.

18. Schwartz YB, Kahn TG, Nix DA, Li XY, Bourgon R, et al. (2006) Genome-wide analysis of Polycomb targets in Drosophila melanogaster. Nat Genet 38: 700-705.

19. Negre N, Hennetin J, Sun LV, Lavrov S, Bellis M, et al. (2006) Chromosomal distribution of PcG proteins during Drosophila development. PLoS Biol 4: e170. doi:10.1371/journal.pbio.0040170.

20. Atchison L, Ghias A, Wilkinson F, Bonini N, Atchison ML (2003) Transcription factor YY1 functions as a PcG protein in vivo. Embo J 22: 1347-1358.

21. Ko CY, Hsu HC, Shen MR, Chang WC, Wang JM (2008) Epigenetic silencing of CCAAT/enhancer-binding protein delta activity by YY1/polycomb group/DNA methyltransferase complex. J Biol Chem 283: 30919-30932.

22. Sui G, Affar el B, Shi Y, Brignone C, Wall NR, et al. (2004) Yin Yang 1 is a negative regulator of p53. Cell 117: 859-872.

23. Yue R, Kang J, Zhao C, Hu W, Tang Y, et al. (2009) Beta-arrestin1 regulates zebrafish hematopoiesis through binding to YY1 and relieving polycomb group repression. Cell 139: 535-546.

24. Liu H, Schmidt-Supprian M, Shi Y, Hobeika E, Barteneva N, et al. (2007) Yin Yang 1 is a critical regulator of B-cell development. Genes Dev 21: 1179-1189.

25. Xi H, Yu Y, Fu Y, Foley J, Halees A, et al. (2007) Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. Genome Res 17: 798-806.

26. Kim JD, Kang K, Kim J (2009) YY1's role in DNA methylation of Peg3 and Xist. Nucleic Acids Res 37: 5656-5664.

27. Sing A, Pannell D, Karaiskakis A, Sturgeon K, Djabali M, et al. (2009) A vertebrate Polycomb response element governs segmentation of the posterior hindbrain. Cell 138: 885-897.

28. Woo CJ, Kharchenko PV, Daheron L, Park PJ, Kingston REA region of the human HOXD cluster that confers polycomb-group responsiveness. Cell 140:

99-110.

29. Bracken AP, Dietrich N, Pasini D, Hansen KH, Helin K (2006) Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. Genes Dev 20: 1123-1136.

30. Mohn F, Weber M, Rebhan M, Roloff TC, Richter J, et al. (2008) Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Mol Cell 30: 755-766.

31. Tanay A, O'Donnell AH, Damelin M, Bestor TH (2007) Hyperconserved CpG domains underlie Polycomb-binding sites. Proc Natl Acad Sci U S A 104: 5521-5526.

32. Pasini D, Cloos PA, Walfridsson J, Olsson L, Bukowski JP, et al. JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. Nature 464: 306-310.

33. Li G, Margueron R, Ku M, Chambon P, Bernstein BE, et al. Jarid2 and PRC2, partners in regulating gene expression. Genes Dev 24: 368-380.

34. Peng JC, Valouev A, Swigut T, Zhang J, Zhao Y, et al. (2009) Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. Cell 139: 1290-1302.

35. Shen X, Kim W, Fujiwara Y, Simon MD, Liu Y, et al. (2009) Jumonji modulates polycomb activity and self-renewal versus differentiation of stem cells. Cell 139: 1303-1314.

36. Kim TG, Kraus JC, Chen J, Lee Y (2003) JUMONJI, a critical factor for cardiac development, functions as a transcriptional repressor. J Biol Chem 278: 42247-42255.

37. Tevosian SG, Albrecht KH, Crispino JD, Fujiwara Y, Eicher EM, et al. (2002)

Gonadal differentiation, sex determination and normal Sry expression in mice require direct interaction between transcription partners GATA4 and FOG2. Development 129: 4627-4634.

38. Conti L, Pollard SM, Gorba T, Reitano E, Toselli M, et al. (2005) Niche-independent symmetrical self-renewal of a mammalian tissue stem cell. PLoS Biol 3: e283. doi:10.1371/journal.pbio.0030283.

39. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553-560.

40. Poux S, McCabe D, Pirrotta V (2001) Recruitment of components of Polycomb Group chromatin complexes in Drosophila. Development 128: 75-85.

41. Schmitt S, Prestel M, Paro R (2005) Intergenic transcription through a polycomb group response element counteracts silencing. Genes Dev 19: 697-708.

42. Ramirez-Carrozzi VR, Braas D, Bhatt DM, Cheng CS, Hong C, et al. (2009) A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. Cell 138: 114-128.

43. Kim H, Kang K, Kim J (2009) AEBP2 as a potential targeting protein for Polycomb Repression Complex PRC2. Nucleic Acids Res 37: 2940-2950.

44. Creyghton MP, Markoulaki S, Levine SS, Hanna J, Lodato MA, et al. (2008) H2AZ is enriched at polycomb complex target genes in ES cells and is necessary for lineage commitment. Cell 135: 649-661.

45. Goldberg AD, Banaszynski LA, Noh KM, Lewis PW, Elsaesser SJ, et al. Distinct factors control histone variant H3.3 localization at specific genomic regions. Cell 140: 678-691.

46. Tate CM, Lee JH, Skalnik DG (2009) CXXC Finger Protein 1 Contains

Redundant Functional Domains That Support Embryonic Stem Cell Cytosine Methylation, Histone Methylation, and Differentiation. Mol Cell Biol.

47. Zhao J, Sun BK, Erwin JA, Song JJ, Lee JT (2008) Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science 322: 750-756.

48. Kanhere A, Viiri K, Araujo CC, Rasaiyaah J, Bouwman RD, et al. Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. Mol Cell 38: 675-688.

49. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, et al. (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129: 1311-1323.

50. Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, et al. Long noncoding RNA as modular scaffold of histone modification complexes. Science 329: 689-693.

51. Lorente M, Perez C, Sanchez C, Donohoe M, Shi Y, et al. (2006) Homeotic transformations of the axial skeleton of YY1 mutant mice and genetic interaction with the Polycomb group gene Ring1/Ring1A. Mech Dev 123: 312-320.

52. Garcia E, Marcos-Gutierrez C, del Mar Lorente M, Moreno JC, Vidal M (1999) RYBP, a new repressor protein that interacts with components of the mammalian Polycomb complex, and with the transcription factor YY1. Embo J 18: 3404-3418.

53. Pollard SM, Benchoua A, Lowell S (2006) Neural stem cells, neurons, and glia. Methods Enzymol 418: 151-169.

54. Mrak RE, Yasargil MG, Mohapatra G, Earel J Jr, Louis DN (2004) Atypical extraventricular neurocytoma with oligodendroglioma-like spread and an unusual pattern of chromosome 1p and 19q loss. Hum Pathol 35: 1156-1159.

55. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, et al. (2008)

Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 454: 766-770.

56. Bailey TL, Gribskov M (1998) Combining evidence using p-values: application to sequence homology searches. Bioinformatics 14: 48-54.

57. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, et al. (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 5: 613-619.

Chapter 4

Reprogramming Factor Expression Initiates Widespread Targeted Chromatin Remodeling

This work was originally published as:

*Koche RP, *Smith ZD, Adli M, Gu H, Ku M, Gnirke A, Bernstein BE, Meissner A. (2011) Reprogramming factor expression initiates widespread targeted chromatin remodeling. Cell Stem Cell 8: 96-105.

Contributions:

Conceived and designed the experiments: RPK, ZDS, AM. Performed the experiments: RPK, ZDS, MA, HG, MK, AG. Analyzed the data: RPK, ZDS. Wrote the paper: RPK, ZDS, AM.

Abstract

Despite rapid progress in characterizing transcription factor-driven reprogramming of somatic cells to an induced pluripotent stem cell (iPSC) state, many mechanistic questions still remain. To gain insight into the earliest events in the reprogramming process, we systematically analyzed the transcriptional and epigenetic changes that occur during early factor induction after discrete numbers of divisions. We observed rapid, genome-wide changes in the euchromatic histone modification, H3K4me2, at more than a thousand loci including large subsets of pluripotency-related or developmentally regulated gene promoters and enhancers. In contrast, patterns of the repressive H3K27me3 modification remained largely unchanged except for focused depletion specifically at positions where H3K4 methylation is gained. These chromatin regulatory events precede transcriptional changes within the corresponding loci. Our data provide evidence for an early, organized, and population-wide epigenetic response to ectopic reprogramming factors that clarify the temporal order through which somatic identity is reset during reprogramming.

Introduction

Exposure to ectopic transcription factors has been established as a robust way to shift somatic cells toward alternative somatic states and to pluripotency [1]. Ectopic expression of four transcription factors, Oct4, Sox2, Klf4, and c-Myc (OSKM), is capable of directing cells from any tissue toward the formation of induced pluripotent stem cells (iPSCs) in mouse and human [2]. Fully reprogrammed iPSCs can contribute to all germ layers and can form complete, fertile mice by tetraploid embryo complementation [2]. Moreover, iPSCs are similar to their embryo-derived counterparts on a molecular level, indicating a genome-wide cascade of transcriptional and epigenetic changes that lead to a stable, newly acquired state [3].

Despite the remarkable fidelity that governs the transition to pluripotency, the overall frequency in which it occurs within induced populations is low and requires an extended latency of one or several weeks [4]. Previous studies and the general reprogramming timeline suggest a requirement for secondary or stochastic events through which certain cells acquire unique advantages that permit transition to pluripotency [4-7]. Therefore, the ectopic expression of the current set of embryonic factors appears insufficient to completely reset the somatic nucleus alone and the mechanism of action probably includes the activation of additional yet unidentified downstream effectors.

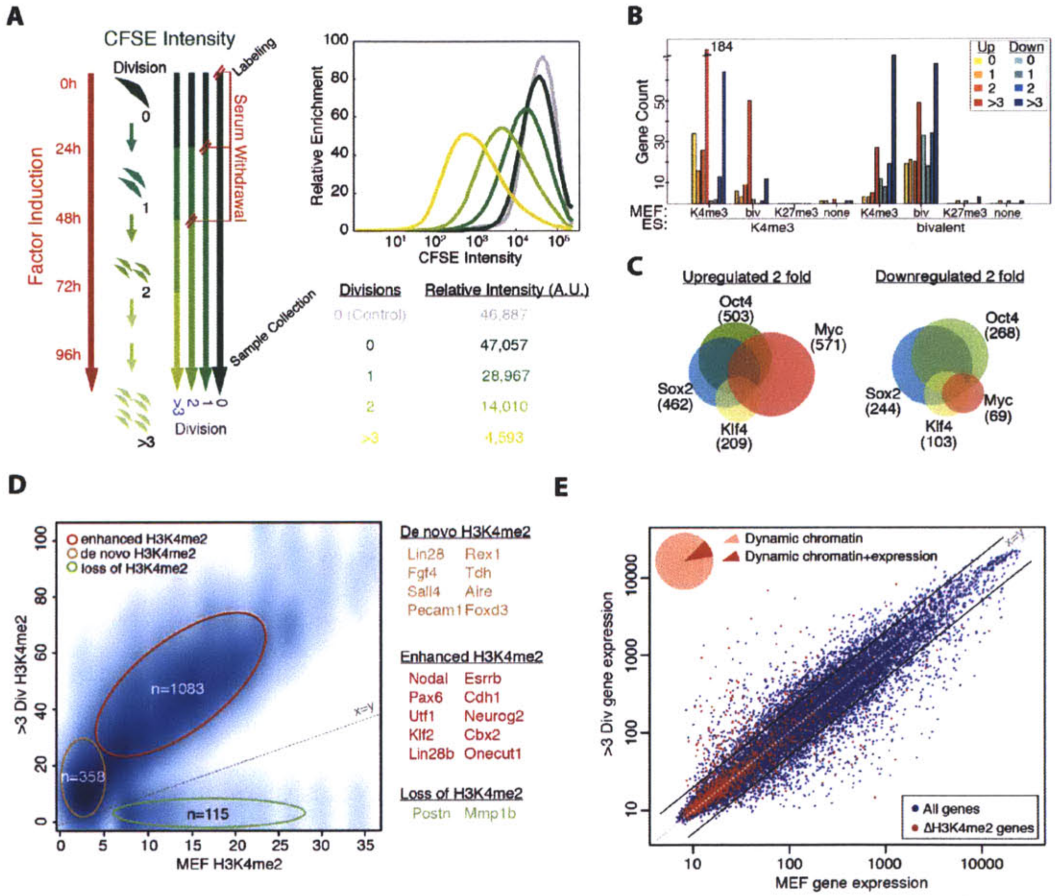
Recent evidence suggests that certain phases of the reprogramming process may be more coordinated than previously assumed. This includes live imaging analysis that demonstrates conserved transitions within reprogramming populations [8]. Transcriptional profiling and RNAi screening in clonally reprogramming populations have demonstrated that robust silencing of somatic transcription factors and effectors as well as activation of critical epithelial markers, govern the most immediate definitive transition from fibroblast toward a "primed" or reprogramming amenable state; the output of somatic factor repression or intermediate stabilizing signaling factors have demonstrated improved iPSC colony generation that suggests that this phase is an essential early step [9]. Despite recent progress, the global nature and scale of these early events as well as their impact on transcriptional and epigenetic landscapes remain unknown.

To gain more insight into the early events during reprogramming, we assayed global gene expression, chromatin state, and DNA methylation in populations of induced fibroblasts that have undergone a discrete number of divisions. We find that dynamic transcription within the reprogramming population is limited and restricted to promoters with pre-existing euchromatin. In contrast to the relative rarity of transcription changes, we found that euchromatin-associated H3K4 methylation is a predominant global early activating response and occurs in the absence of transcriptional activation at corresponding loci. Interestingly, these targets include the promoters of many essential pluripotency-related and developmentally regulated genes and describe a coherent shift in cellular identity. We observe highly localized, coordinated depletion of repressive chromatin (H3K27me3) exclusively at promoters where H3K4 methylation is gained. Finally, this targeted remodeling extends to enhancers across the genome, which transition dramatically from the somatic state, and represents an additional level of cell state transition. Taken together, our results suggest that early transcriptional dynamics are largely dependent on pre-existing, accessible chromatin and that ectopic factor induction initiates a concerted change in target chromatin through which pluripotent targets are primed for subsequent activation.

Results

CFSE labeling enables enrichment of cells that have undergone discrete numbers of cell divisions

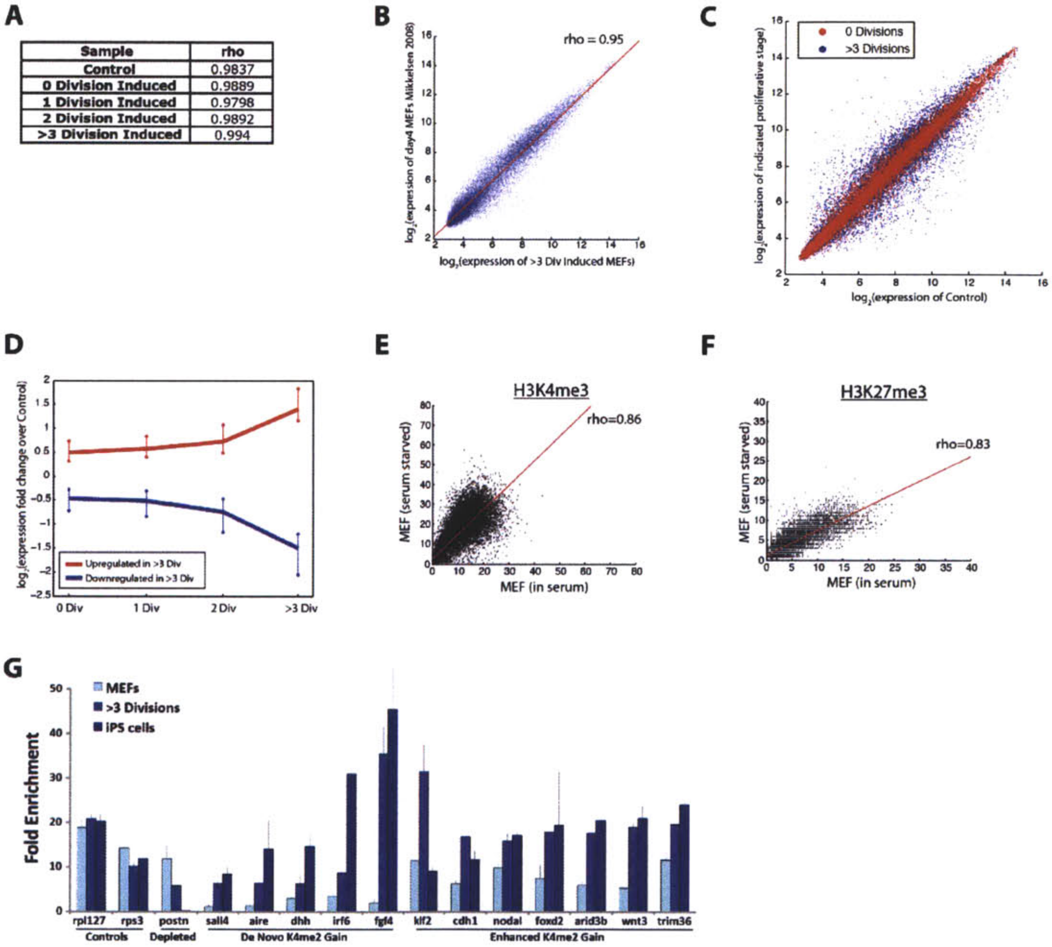
To further elucidate critical early steps in the reprogramming process, we investigated responses to reprogramming factor expression in cells that had undergone no cell division and cells that had divided 1, 2, or more than 3 times. By using inducible (OSKM) secondary mouse embryonic fibroblasts (MEFs), we could ensure rapid and homogenous induction of the four factors as described previously [3,10]. We isolated doxycycline-induced cells that had undergone a defined number of cell divisions by combining the live stain CFSE (carboxyfluorescein succinimidyl ester) and a serum pulsing protocol. Four distinct fractions were enriched based upon their mean proliferative number in a manner that ensures that proliferation is the predominant experimental variable (Figure 1A). All cells were collected in an arrested (serum-starved) state except the final sample, which was allowed to divide continuously under factor induction. We confirmed that the relative fluorescence intensity remains unchanged in the serum-starved control compared to a serum-starved, doxycycline-induced population that remains exposed to the reprogramming factors for 96 hr and experiences minimal or no cell division (Figure 1A). Importantly, CFSE-labeled cells that proliferated continuously for 96 hr (with a fluorescence reduction indicating three or more divisions) show highly similar global transcriptional attributes to populations that had not undergone CFSE labeling or serum withdrawal, demonstrating that this protocol does not interfere with the general reprogramming process (Figure 2A,B).



Divisions	Relative Intensity (A.U.)
0 (Control)	46,887
0	47,057
1	28,967
2	14,010
>3	4,593

Figure 1

Figure 1: Global Transcriptional and Epigenetic Dynamics during Early Induction of Reprogramming Factors. **(A)** Schematic for enrichment of distinct proliferative cohorts by means of the live dye CFSE and serum pulsing under constant factor induction and time. After 96 hr of continued culture in doxycycline-supplemented medium, samples were scored via flow cytometry. Median fluorophore intensity was assessed as a relative metric for proliferative number and is shown on the right. Relative intensity is displayed in arbitrary units (A.U.). **(B)** mRNA expression dynamics conditional on MEF/ES chromatin state progressing across cell division number (shown color coded in the inset) for up- and downregulated genes. ESC H3K4me3-only loci and their respective states in MEFs are shown on the left, and ESC bivalent (H3K4me3/H3K27me3) loci are shown on the right. **(C)** Enrichment for Oct4, Sox2, Klf4, and c-Myc (OSKM) binding in promoter elements of dynamically regulated genes shows an asymmetric bias toward gene activation within targets of the myc oncogene. Transcription factor binding taken from genome-scale profiling of embryonic stem cells [22,23]. **(D)** Density plot of genes with dynamic H3K4me2 in reprogramming populations compared to control MEFs. Promoters exhibiting a dynamic shift in H3K4me2 (n$\sim$1500) fall into three distinct classes: de novo (beige), enhanced (red), and loss (green). Representative genes from all three classes are highlighted on the right. **(E)** Expression data between starting state (control) and the >3 divisions induced population with dynamic H3K4me2 genes highlighted in red. Pie chart shows the representation of genes that exhibit only H3K4me2 changes (pink) or both H3K4me2 and gene expression changes (red; n$\sim$10%).



Sample	rho
Control	0.9837
0 Division Induced	0.9889
1 Division Induced	0.9798
2 Division Induced	0.9892
>3 Division Induced	0.994

Figure 2

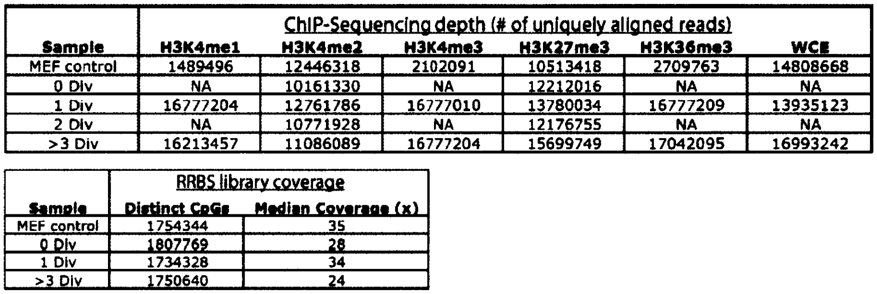
Figure 2: Fidelity of reprogramming system and molecular assays across samples and to previously published controls. **(A)** Pearson correlation for biological replicates purified using our CFSE serum pulsing protocol present highly similar expression data in biological duplicate that confirms the reproducibility of our assay. **(B)** Scatterplot of expression values from a preceding data set using our inducible MEF system show that neither CFSE labeling nor transient serum starvation inhibits or hinders the normal response to reprogramming factor induction. **(C)** Continuity of expression dynamics across proliferative samples: scatterplot superimposing our two extreme experimental samples, 0 Division and >3 Division over the uninduced control shows a progressive divergence away from the somatic state. **(D)** When differentially regulated genes (>2 fold expression) in our terminal >3 Division sample are mapped across earlier divisions, they exhibit continuous trends. **(E)** Comparison of global H3K4me3 levels in serum arrested MEFs to pre-existing data for MEFs grown in serum (rho=0.86, $p < 10^{-16}$). **(F)** Comparison of global H3K27me3 levels in serum arrested MEFs to pre-existing data for MEFs grown in serum (rho=0.83, $p < 10^{-16}$) **(G)** ChIP-qPCR validation of H3K4me2 dynamics within identified classes confirm genome-wide observations. Gene names are highlighted and are organized into specific classes: including positive controls, depleted, de novo, and enhanced promoters (error bars are standard deviation for n=3 replicate experiments).

Transcriptional dynamics of early reprogramming populations are limited to sites with pre-existing H3K4 trimethylation

We next used our discrete cell populations to investigate the early gene expression and chromatin dynamics induced by the four factors (Table 1). Global mRNA expression profiles revealed continuous trends across populations and a primary response to factor induction that operates almost exclusively within accessible H3K4me3 chromatin (Figure 1B, 97%, Fisher's exact test $p < 10^{-16}$). Upregulated (2-fold, t test $p < 0.05$) targets are predominantly associated with promoter histone H3K4me3 in MEFs prior to induction, and moreover are enriched 2.2-fold for loci that are H3K4me3 within ESCs (Figure 1B). Repressed genes (2-fold, t test $p < 0.05$) were enriched for H3K4me3 only or H3K4me3/H3K27me3 (bivalent) promoters in MEFs, but enriched 2.8-fold for the bivalent state in pluripotent cells (Figure 1B). Both activated and repressed gene sets exhibited preferential promoter binding for the induced factors, with an asymmetric bias for enhanced expression among c-Myc-regulated targets (9.5-fold increased likelihood, Fisher's exact text $p < 10^{-16}$), consistent with its function in the transition to transcriptional elongation as opposed to PolII recruitment/initiation (Figure 1C; [11]). These observations indicate that early expression changes mediated by factor induction are in large part constrained by pre-existing chromatin and may operate only at promoters that are already in an open and accessible state. Moreover, these changes occur immediately and gradually increase with additional cell divisions (Figure 2C,D). These data suggest that in the earliest phase of reprogramming, fibroblast identity is predominantly perturbed by transcriptional silencing of somatic targets and not the activation of pluripotency-associated targets of the reprogramming factors.

Activating chromatin marks are targeted to promoters prior to transcriptional activation

Next we investigated the consequences of ectopic factor activity at the chromatin level by comparing the dynamics of functional epigenetic markers to the more limited observations that could be made when measuring transcriptional output alone. We



Sample	ChIP-Sequencing depth (# of uniquely aligned reads)					
	H3K4me1	H3K4me2	H3K4me3	H3K27me3	H3K36me3	WCE
MEF control	1489496	12446318	2102091	10513418	2709763	14808668
0 Div	NA	10161330	NA	12212016	NA	NA
1 Div	16777204	12761786	16777010	13780034	16777209	13935123
2 Div	NA	10771928	NA	12176755	NA	NA
>3 Div	16213457	11086089	16777204	15699749	17042095	16993242

Sample	RRBS library coverage	
	Distinct CpGs	Median Coverage (x)
MEF control	1754344	35
0 Div	1807769	28
1 Div	1734328	34
>3 Div	1750640	24

Table 1: Sequencing depth of ChIP-Seq libraries for all histone marks analyzed as well as corresponding data for methylation profiling using RRBS

generated genome-wide chromatin maps for the three methylation marks on H3K4 (mono-, di-, and trimethylation) as well as for H3K27 trimethylation and H3K36 trimethylation across the isolated populations via ChIP-Seq [12]. We then focused our initial query on H3K4me2, because it is a general marker of both promoter and enhancer regions and is broadly amenable to genome-wide analysis (as opposed to trimethylation that is exclusive to promoters) [13,14]. H3K27me3 was chosen as a marker associated with transcriptional silencing, in particular of developmental transcription factors [12,15,16]. Comparison with previously published data sets confirms that our serum-starvation protocol does not induce significant chromatin changes in the MEFs (Figure 2E,F), and ChIP followed by quantitative PCR for representative loci confirms the trends observed in our ChIP-Seq results (Figure 2G).

Surprisingly, H3K4me2 peaks exhibit dramatic changes at more than 1500 genes and continuously increase with successive cell divisions (Figure 1D). The results highlight two striking findings. First, H3K4me2 target loci do not correspond to observed changes in gene expression (Figure 1E, chi square test $p > 0.1$). Furthermore, changes in H3K4me2 are apparent even in populations that have not yet divided based on CFSE intensity (Mann-Whitney U test $p < 10^{-16}$). Notably, these regions are strongly enriched for pluripotency and developmentally regulated targets, such as Sall4, Lin28, and Fgf4, which will not become transcriptionally active until later stages of iPSC formation. These results provide insights into the reprogramming process and describe an unexpected chromatin-remodeling response to the reprogramming factors that precedes transcriptional activation of ESC-exclusive genes (Figure 3A). We confirmed this observation with the transcriptionally associated histone mark H3K36me3, which exhibits no enrichment at identified loci across the early reprogramming phase or outside of pluripotent cell types, and by RNA PolII occupancy at representative promoters, which did not yield apparent enrichment when compared to established iPSC lines (Figures 3B,C). This suggests that complete chromatin remodeling to transcriptional initiation is either unstable or not yet established during this early phase.

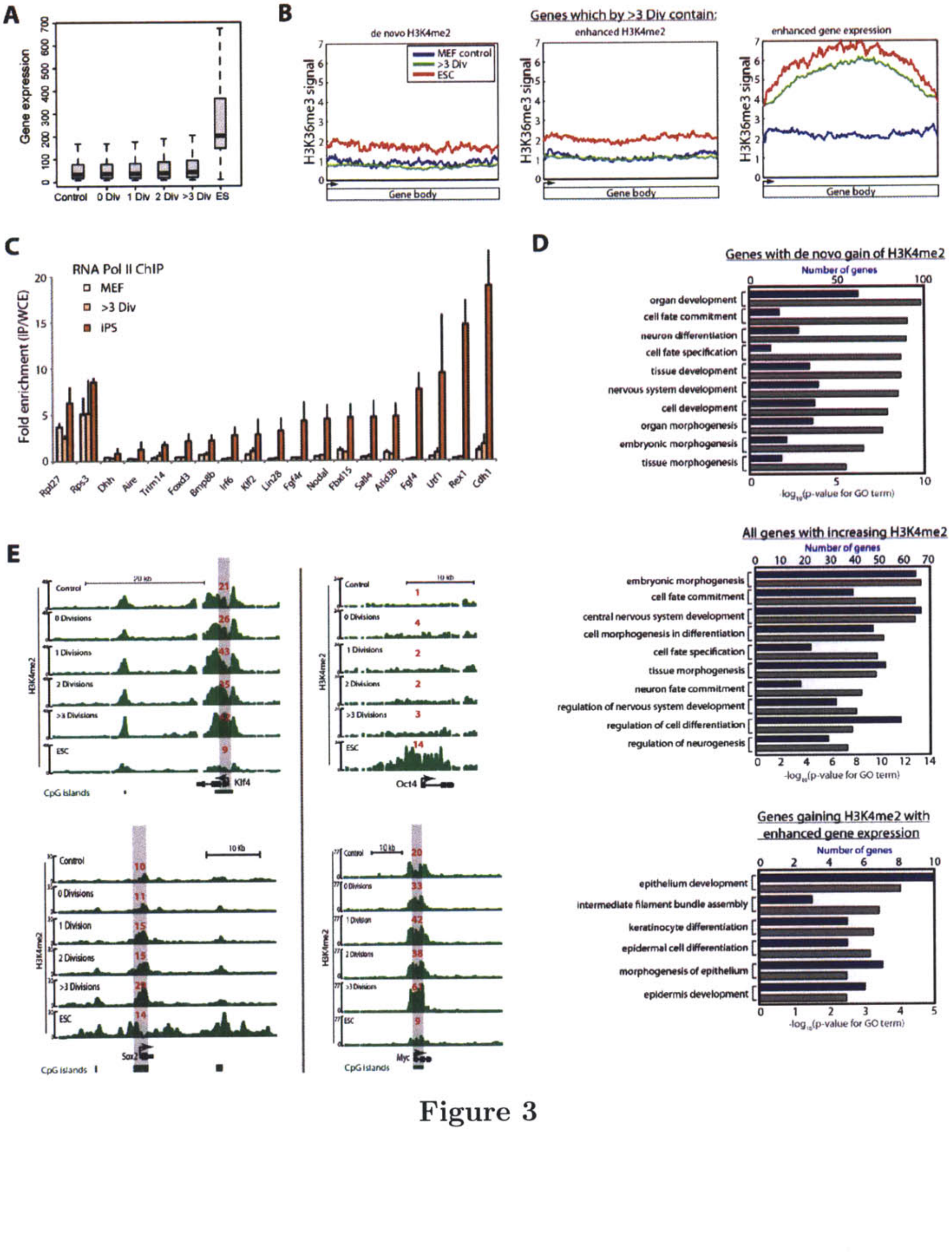
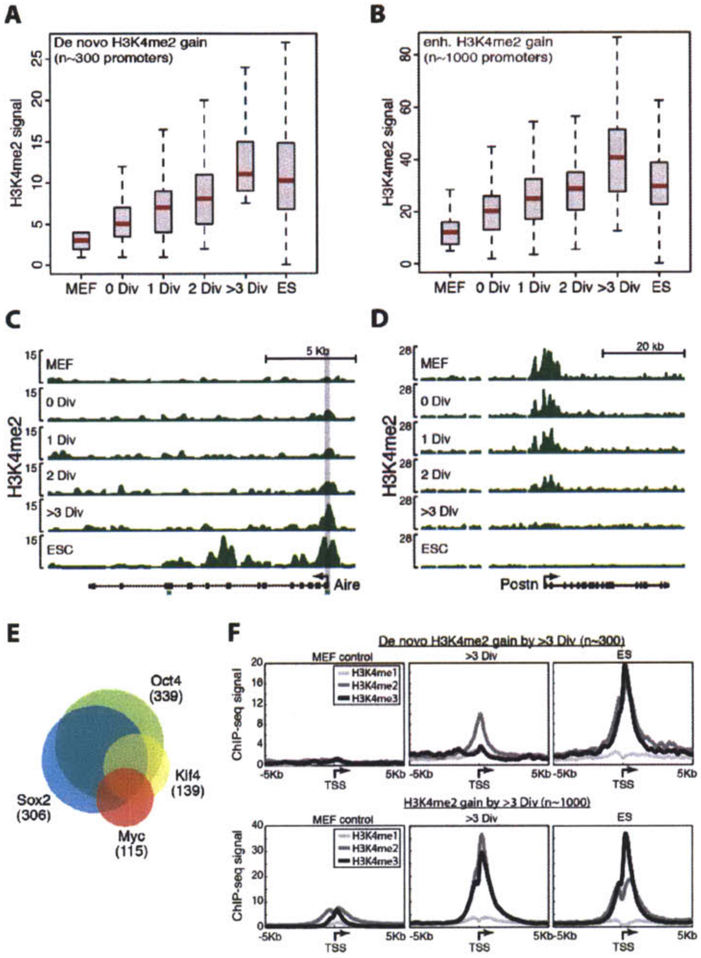

A
Gene expression
0
100
200
300
400
500
600
700
Control
0 Div
1 Div
2 Div
>3 Div
ES
B
Genes which by >3 Div contain:
de novo H3K4me2
enhanced H3K4me2
enhanced gene expression
H3K36me3 signal
0
1
2
3
4
5
6
7
MEF control
>3 Div
ESC
Gene body
C
RNA Pol II ChIP
MEF
>3 Div
iPS
Fold enrichment (IP/WCE)
0
5
10
15
20
Rpl27
Rps3
Dhh
Aire
Trim14
Foxd3
Bmp8b
Irf6
Klf2
Lin28
Fgf4r
Nodal
Fbxl15
Sall4
Arid3b
Fgf4
Utf1
Rex1
Cdh1
D
Genes with de novo gain of H3K4me2
Number of genes
0
50
100
organ development
cell fate commitment
neuron differentiation
cell fate specification
tissue development
nervous system development
cell development
organ morphogenesis
embryonic morphogenesis
tissue morphogenesis
0
5
10
-log10(p-value for GO term)
All genes with increasing H3K4me2
Number of genes
0
10
20
30
40
50
60
70
embryonic morphogenesis
cell fate commitment
central nervous system development
cell morphogenesis in differentiation
cell fate specification
tissue morphogenesis
neuron fate commitment
regulation of nervous system development
regulation of cell differentiation
regulation of neurogenesis
0
2
4
6
8
10
12
14
-log10(p-value for GO term)
Genes gaining H3K4me2 with enhanced gene expression
Number of genes
0
2
4
6
8
10
epithelium development
intermediate filament bundle assembly
keratinocyte differentiation
epidermal cell differentiation
morphogenesis of epithelium
epidermis development
0
1
2
3
4
5
-log10(p-value for GO term)
E
20 kb
Control
21
0 Divisions
26
1 Divisions
43
2 Divisions
35
>3 Divisions
42
ESC
9
H3K4me2
Klf4
CpG islands
10 kb
Control
1
0 Divisions
4
1 Divisions
2
2 Divisions
2
>3 Divisions
3
ESC
14
H3K4me2
Oct4
10 Kb
Control
10
0 Divisions
11
1 Division
15
2 Divisions
15
>3 Divisions
28
ESC
14
H3K4me2
Sox2
CpG islands
10 kb
Control
20
0 Divisions
33
1 Division
42
2 Divisions
38
>3 Divisions
61
ESC
9
H3K4me2
Myc
CpG islands

Figure 3

Figure 3: Chromatin modifications are enriched at developmental genes and for pluripotency associated targets without changes in transcription. **(A)** Box plots of expression for ES cell genes display minimal and insignificant transcriptional changes across the time series (t test $p > 0.3$) while demonstrating significant changes in H3K4 methylation status (Figure 4A,B). **(B)** H3K36me3 status across gene bodies for identified subsets: de novo H3K4me2 (n$\sim$300), enhanced H3K4me2 (n$\sim$1200), as well as for a positive control set (genes demonstrating enhanced expression n$\sim$150). Signal is assayed within three cell states: our MEF control, >3 divisions post factor induction and within mES cells. No observable or significant H3K36me3 occurs within the gene subsets for which expression is not observed. **(C)** PolII enrichment at the Transcription Start Site (TSS) for 17 loci identified showing increased promoter H3K4me2 enrichment as well as for Rpl27 and Rps3, which serve as positive, housekeeping controls. No appreciable changes in PolII recruitment are observed after >3 divisions of reprogramming factor induction compared to starting fibroblasts. Data is averaged over 3 biological replicates for each timepoint and normalized over Whole Cell Extract with SEM highlighted. **(D)** Gene set enrichment analysis of all sites exhibiting de novo H3K4me2 (n$\sim$300) enhanced H3K4me2 (n$\sim$1200) and enhanced H3K4me2 with co-occurring increase in gene expression (n$\sim$167). Enhanced H3K4me2 peaks demonstrating transcriptional activity are highly enriched for keratinization components as a likely artifact of somatic Klf4 activity. Blue bars highlight the number of genes found against the scaling present on the top of each plot; Grey bars represent the Log10 P values of these enrichments and are scaled at the bottom. **(E)** H3K4 methylation status of Oct4, Sox2, Klf4, and c-Myc loci during early phase reprogramming. By >3 divisions, Sox2, Klf4 and c-Myc have enhanced H3K4me2 levels gained at their respective CpG island promoters. Note that reads mapping outside the coding regions reconstitute these trends and are distinct from any potential ambiguities mapping to the transgenes. The Oct4 locus, which is not CpG dense and is DNA methylated, does not change its basal promoter H3K4me2 levels.



A
De novo H3K4me2 gain
(n~300 promoters)
H3K4me2 signal
0
5
10
15
20
25
MEF
0 Div
1 Div
2 Div
>3 Div
ES
B
enh. H3K4me2 gain
(n~1000 promoters)
H3K4me2 signal
0
20
40
60
80
MEF
0 Div
1 Div
2 Div
>3 Div
ES
C
5 Kb
H3K4me2
15
MEF
0 Div
1 Div
2 Div
>3 Div
ESC
Aire
D
20 kb
H3K4me2
28
MEF
0 Div
1 Div
2 Div
>3 Div
ESC
Postn
E
Oct4
(339)
Klf4
(139)
Sox2
(306)
Myc
(115)
F
De novo H3K4me2 gain by >3 Div (n~300)
MEF control
>3 Div
ES
H3K4me1
H3K4me2
H3K4me3
ChIP-seq signal
0
4
8
12
16
20
-5Kb
TSS
5Kb
H3K4me2 gain by >3 Div (n~1000)
MEF control
>3 Div
ES
H3K4me1
H3K4me2
H3K4me3
ChIP-seq signal
0
10
20
30
40
-5Kb
TSS
5Kb

Figure 4

Figure 4: H3K4 Dimethylation Increases at Pluripotency-Related Genes and Is Lost in Repressed Somatic Targets. **(A)** De novo H3K4me2 acquisition is continuous across cohorts and already visible before a single division (n$\sim$300). Red line indicates median. Whiskers represent 2.5 and 97.5 percentile. **(B)** Enhanced H3K4me2 at a subset of $\sim$1000 promoters over proliferative cohorts exhibit similar trends and approach expected ESC levels in dividing populations of reprogramming cells. Red line indicates median. Whiskers represent 2.5 and 97.5 percentile. **(C)** ChIP-Seq tracks showing de novo H3K4me2 at the endogenous promoter of Aire as part of an orchestrated enrichment that is preferential for Oct4- and Sox2-regulated promoters. Green bars on the bottom indicate CpG islands. Gray bar highlights the putative nucleosome-depleted region that is flanked by H3K4me2 within ESCs. **(D)** H3K4me2 ChiP-seq map of the Postn locus, which is expressed in MEFs and silenced by >3 divisions, shows a loss of H3K4me2 levels at its promoter region to ESC-like levels. The Postn locus represents 115 promoters for which H3K4me2 is lost during reprogramming factor induction. **(E)** ESC transcription factor occupancy of genes demonstrating H3K4me2 enrichment show a predominance of Oct4 and Sox2 binding. **(F)** Composite plots of H3K4 mono-, di-, and trimethylation distribution at de novo and enhanced promoter classes in control MEFs, after three divisions, and within ESCs.

For further analysis, we subdivided loci that gain H3K4me2 during early reprogramming into two classes: a set of "de novo" H3K4me2 loci that have essentially undetectable H3K4me2 levels in MEFs and a set of "enhanced" H3K4me2 loci whose H3K4me2 signals increase by a minimum of 2.5-fold relative to the MEF control (Figure 4A,B). In both cases, the chromatin changes are reproducible across the target loci and increase in magnitude with cell divisions, suggestive of a progressive and coordinated process (Figure 4C). A third class of promoters was less represented but exhibited a loss of promoter H3K4me2 that correlates with transcriptionally silenced somatic determinants such as Postn (Figure 4D, 1.75-fold decrease in expression, $n \sim 110$ genes, Mann-Whitney U test $p < 0.02$). Overall, the changes in promoter H3K4me2 occur rapidly and are primarily targeted to a set of loci that function in early development or as active mediators of pluripotency, including epigenetic reprogramming of the endogenous Sox2, Klf4, and c-Myc promoters themselves (Figures S2D and S2E). Moreover, promoters gaining H3K4me2 are significantly enriched for targets of Oct4 and Sox2 (Figure 4E, Fisher's exact test $p < 0.0009$ and 0.00039 for Oct4 and Sox2, respectively).

We next investigated the positioning of the related histone marks H3K4me1 and H3K4me3 to explore potential overlaps with H3K4me2. Surprisingly, we find that H3K4me2 is exclusive within the de novo promoter set, which is devoid of all forms of H3K4 methylation in MEF controls and does not gain H3K4me1 or H3K4me3 concurrently with H3K4me2 (Figure 4F). Alternatively, the "enhanced" promoter set, which exhibits both H3K4me2 and H3K4me3 within control populations, coordinately increases both marks as induced populations continue to proliferate (Figure 4F). These data emphasize the value of H3K4me2 as a dynamic mark across promoters because it detects nascent histone modification at de novo promoters, which are under-enriched for these marks in MEFs, as well as increased representation of pre-existing chromatin modifications within enhanced promoters that are augmented by ectopic factor activity. Additionally, within pluripotent cells, H3K4me3 is enriched at the vast majority of genes that gain H3K4me2 within the early reprogramming phase. These H3K4me2-exclusive promoters may therefore imply a decoupled and transiently stable epigenetic mechanism that precedes complete remodeling and gene activation.

The dynamic gain of H3K4 methylation occurs without promoter-wide changes in somatically defined, repressive H3K27me3 when inspected across the entirety of target promoters (Figure 5A; Kolmogorov-Smirnov test $p > 0.1$). The retention of somatic heterochromatin at the same promoters highlights a possible barrier that prevents gene activation and suggests that repressive modifications might be less dynamic than H3K4me2.

Repressive H3K27me3 is lost specifically at sites where H3K4 methylation is gained

We next investigated the positional context of H3K4me2 to explore possible epigenetic or genetic determinants of the early response to ectopic factor induction. Enhanced H3K4me2 peaks occur directly at transcription start sites (TSS) in two distinct promoter classes: those that will ultimately be activated at the iPS cell stage and those that are not activated but are rather reset to a poised bivalent state (Figure 5B, Figure 6A). The positional gain of H3K4me2 is targeted to the TSS and does not display the bimodality seen in ESCs/iPSCs that is associated with nucleosome depletion at the site of initiation (Figure 6B, shaded region). We also examined chromatin changes at the subset of promoters with H3K27me3 in MEFs. Here, we found that positional gain of H3K4me2 is accompanied by a corresponding depletion of H3K27me3 (Figure 6C, Student's t test $p < 0.01$). Remarkably, this H3K27me3 reduction is present only within the punctate boundaries of a sharply gained H3K4me2 peak and does not spread to the surrounding regions, which retain somatic levels of facultative, inhibitory heterochromatin as in the starting state.

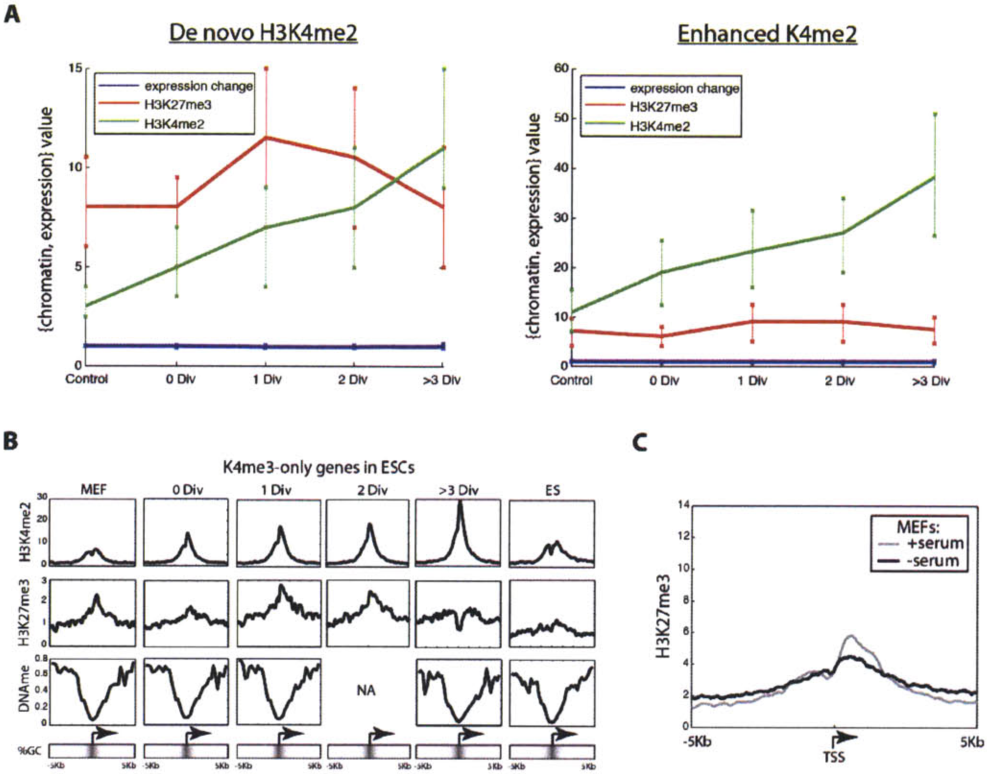


Figure 5: Promoter wide and conditional relationships of inhibitory and euchromatic chromatin marks. **(A)** Continuously increasing H3K4 methylation at pluripotency-associated promoters: De novo (left) and enhanced (right) H3K4me2 levels across promoters exhibit a progressive increase as cells divide that does not dramatically alter promoter H3K27me3 levels and is not associated with detectable expression changes. Blue line is normalized median expression for the included gene sets. Vertical lines represent the 25th and 75th percentile. **(B)** General trends of epigenetic reprogramming events at ES cell H3K4 methylated promoters (n=192) within induced populations: Upper Panel: Composite plots at active ES cell promoters compared against somatic and ES cell controls. Gain of H3K4me2 occurs at the transcription start site. Middle Panel: Composite plot of H3K27me3 levels are generally low but display the same concurrent depletion at the site of H3K4 methylation by >3 divisions. Lower Panel: CpG methylation values at regions of enhanced K4me2 gain are predominantly hypomethylated CpG density across the promoters analyzed is highlighted and demonstrates the boundary of the dynamic changes in chromatin state. Scale ranges between 40% (white) and 80% (black) GC content. **(C)** Composite plot for all ES bivalent genes demonstrating increasing K4me2 across the reprogramming timeline as in Figure 3B for MEFs in the presence or absence of serum.

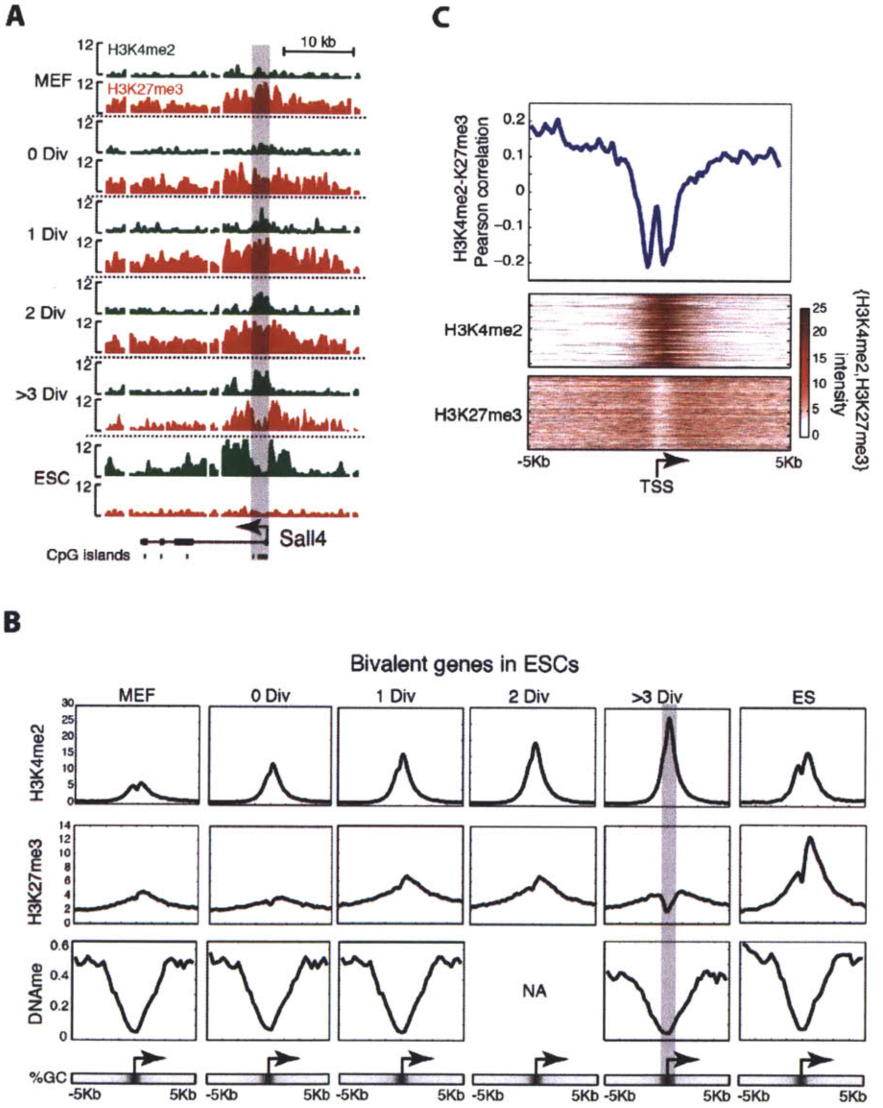

A
10 kb
H3K4me2
H3K27me3
12
MEF
0 Div
1 Div
2 Div
>3 Div
ESC
Sall4
CpG islands
C
H3K4me2-K27me3
Pearson correlation
0.2
0.1
0
-0.1
-0.2
H3K4me2
H3K27me3
-5Kb
TSS
5Kb
{H3K4me2,H3K27me3}
intensity
25
20
15
10
5
0
B
Bivalent genes in ESCs
MEF
0 Div
1 Div
2 Div
>3 Div
ES
H3K4me2
30
25
20
15
10
5
0
H3K27me3
14
12
10
8
6
4
2
0
DNAme
0.6
0.4
0.2
0
NA
%GC
-5Kb
5Kb

Figure 6

Figure 6: Chromatin Remodeling and Genetic Determinants Define the Early Reprogramming Phase **(A)** The Sall4 locus exhibits a de novo gain of H3K4 methylation at two CpG islands (green bars). Gain of H3K4me2 corresponds to a targeted depletion of H3K27 methylation within cycling cells that is limited to the site of H3K4 methylation. Highlighted region displays the CpG island and the site of ESC-specific nucleosome depletion. **(B)** General trends of epigenetic reprogramming events at ESC bivalent promoters (n = 688) within induced populations. Top: Composite plots of H3K4me2 gain within ESC bivalent promoters compared against somatic and ESC controls. Middle: Composite plot of H3K27me3 levels stay constant except in the most proliferative cohort (>3 divisions) where levels are inversely proportional to the gain in H3K4me2 and are subsequently depleted. Bottom: CpG methylation values at regions of enhanced H3K4me2 gain are predominantly hypomethylated across states as expected given the high CpG density of this promoter set (82% CpG islands). CpG density across the promoters analyzed is highlighted and demonstrates the boundary of the dynamic changes in chromatin state. Scale ranges between 40% (white) and 80% (black) GC content. **(C)** Pearson correlation between H3K4me2 and H3K27me3 levels in 200 base pair sliding windows. Negative correlation between the two marks reaches significance within 500 bp from the TSS. Histone mark enrichments for the promoter set are included as heat maps and emphasize this inverse relationship.

We also generated genome-wide DNA methylation data from the 0, 1, and >3 division populations and compared them to control and ESC promoters. As expected, the majority of regions exhibiting dramatic H3K4me2 gain displayed promoter hypomethylation in all states (Figure 6B). Moreover, promoters with the most dramatic shifts in chromatin state generally exhibit higher CpG density and preferentially enrich for CpG islands (82%, Fisher's exact test $p < 10^{-33}$). DNA methylation data confirmed that these regions were consistently hypomethylated across populations, including in the starting fibroblast state, an expected epigenetic landscape that is generally characteristic of CpG islands. Additionally, it is interesting to note that regions with depletion of H3K4me2 were frequently associated with transcriptional repression and a vast majority (95%, Fisher's exact test $p < 10^{-41}$) corresponded to non-CpG island promoters at which H3K4 methylation status is often predictive of transcriptional activity. Taken together, these data suggest that the plasticity of somatic chromatin to changes by reprogramming factors is most amenable within certain boundaries in part governed by genetic determinants, such as CpG density and the targeting sequences for the reprogramming factors themselves.

Enhancer Signatures Are Driven from a Somatic toward an ESC-like State

The activity of reprogramming factors on target chromatin is not restricted to the promoter regions and operates similarly within intergenic regions (Figure 7A, Figure 8A). Nonpromoter intervals enriched for H3K4me2 have been correlated to functional enhancers genome-wide, the patterns of which are remarkably variable across cell type and have been used as a high information content signature of a given cell state [14]. We thus reasoned that nonpromoter H3K4me2 elements that differ between MEFs and iPSCs could provide further insight into the early dynamics of reprogramming. Unlike promoter elements, which predominantly gain H3K4me2, epigenetic signatures of enhancers are gained and lost as reprogramming populations shift away from the somatic state (Figure 7B). Moreover, enhancer dynamics are shifted rapidly; a majority of intergenic H3K4me2 dynamics occur on or before a single cell division (54%

gained, 66% lost) and progress continuously with division number (Figure 8B). Of the 11,228 H3K4me2 enhancers identified in the reprogramming populations, 46% are shared with ESCs and 8,407 somatic exclusive enhancer regions are depleted (Figure 7B). Intergenic analysis of additional H3K4 methylation marks confirm the canonical architecture of enhancer elements, with strong overlap of H3K4me1 and H3K4me2 and relative lack of promoter-exclusive H3K4me3 (Figure 7C). Moreover, reprogramming induced enhancer signatures appear to acquire stable H3K4 methylation sequentially, first gaining H3K4me1 (Figure 7C, middle) followed by H3K4me2 (Figure 7C, right). From this context, examination of the epigenetic changes within intergenic regions provide a unique opportunity to model enhancer dynamics; moreover, genome-wide characterization of H3K4me2 confirms its value as a highly informative epigenetic mark, being present in disparate promoter and intergenic contexts where H3K4me1 or H3K4me3 are mutually exclusive (Figure 8D). Intergenic shifts in H3K4me2 enrichment thus serve as a unique barcode for cellular identity and sensitively measure the epigenetic changes caused by reprogramming factor induction.

We incorporated genome-scale DNA methylation maps of ESCs and MEFs [17] with those generated for our induced populations for use in our analysis of intergenic H3K4me2. Genomic intervals that display rapid gain of H3K4me2 tended to exhibit relatively lower DNA methylation levels in MEFs (Figure 7D, left). In contrast, ESC enhancer elements that are not activated after 96 hr of factor induction have significantly higher DNA methylation levels in MEFs (Figure 7D, right, Student's t test $p < 10^{-32}$). Interestingly, the MEF-exclusive enhancers that are lost during reprogramming display complete hypermethylation within ESCs, but not within induced populations (Figure 8C). This suggests that ESC-like DNA methylation patterns are not fully established until later stages of reprogramming. The failure to re-establish DNA methylation at somatic intergenic H3K4me2 enhancers may, in part, account for the instability/elasticity of reprogramming populations, which may traverse back toward a fibroblast-like state upon premature removal of ectopic factor expression [9].

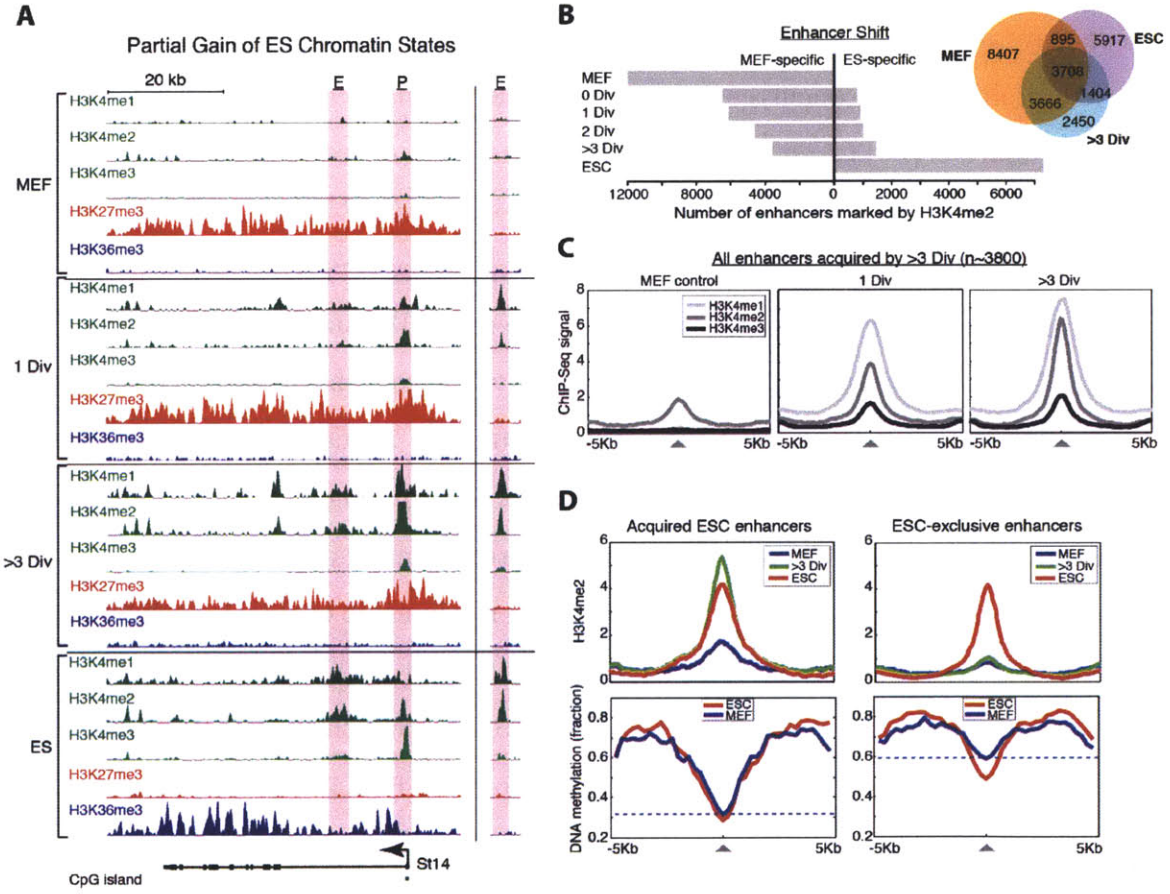
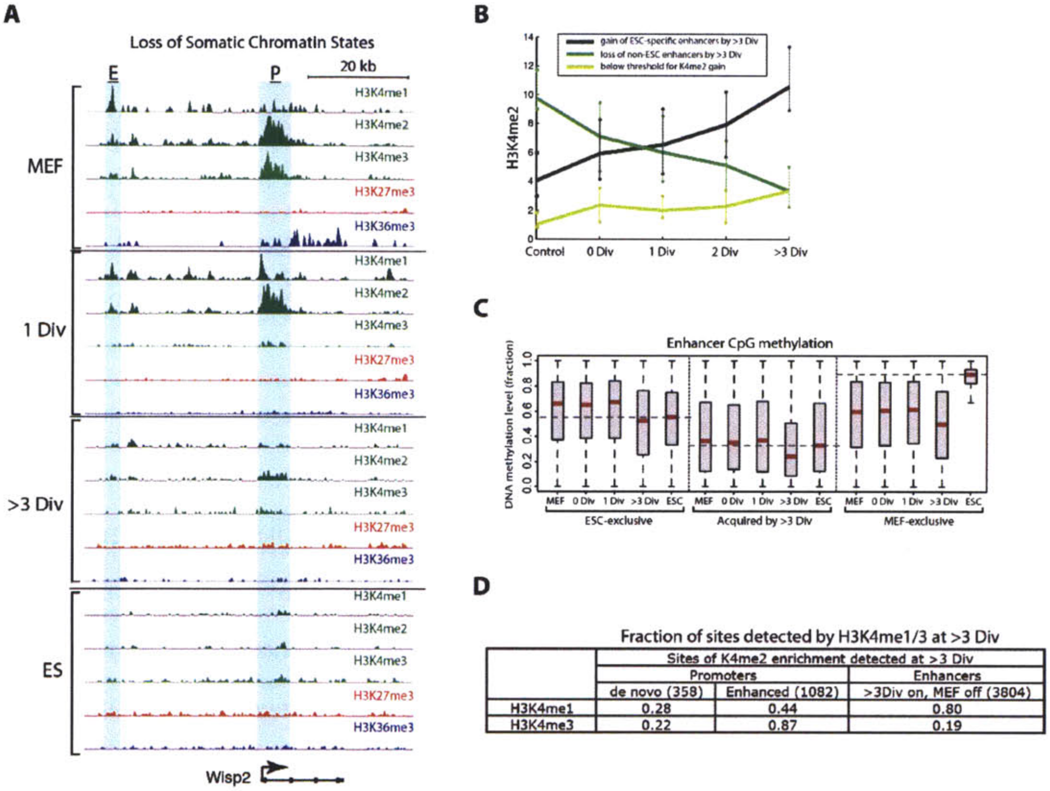

A
Partial Gain of ES Chromatin States
20 kb
E
P
E
MEF
H3K4me1
H3K4me2
H3K4me3
H3K27me3
H3K36me3
1 Div
H3K4me1
H3K4me2
H3K4me3
H3K27me3
H3K36me3
>3 Div
H3K4me1
H3K4me2
H3K4me3
H3K27me3
H3K36me3
ES
H3K4me1
H3K4me2
H3K4me3
H3K27me3
H3K36me3
St14
CpG island
B
Enhancer Shift
MEF-specific
ES-specific
MEF
0 Div
1 Div
2 Div
>3 Div
ESC
12000
8000
4000
0
2000
4000
6000
Number of enhancers marked by H3K4me2
MEF
8407
895
5917
ESC
3708
3666
1404
2450
>3 Div
C
All enhancers acquired by >3 Div (n~3800)
MEF control
1 Div
>3 Div
H3K4me1
H3K4me2
H3K4me3
ChIP-Seq signal
8
6
4
2
0
-5Kb
5Kb
-5Kb
5Kb
-5Kb
5Kb
D
Acquired ESC enhancers
ESC-exclusive enhancers
H3K4me2
6
4
2
0
MEF
>3 Div
ESC
MEF
>3 Div
ESC
ESC
MEF
ESC
MEF
DNA methylation (fraction)
0.8
0.6
0.4
0.2
-5Kb
5Kb
-5Kb
5Kb

Figure 7

Figure 7: Global Epigenetic Dynamics during the Early Stage of Reprogramming Factor Induction Extends beyond Target Promoter Regions to Putative Enhancers **(A)** The CpG island promoter (P) (pink highlight) of the ESC-expressed St14 gene displays minimal H3K4 methylation in the somatic state and increases in H3K4me2 with proliferation, concurrent with punctate loss of H3K27me3 at the CpG island (see also Figure 6A). The de novo K4me2 gain is accompanied by gain of an intronic enhancer signature (E) (pink highlight). Expression levels for St14 are not detected until complete remodeling at later stages. Intergenic enhancers (E) (pink highlight, right) are also gained and are progressively enriched for H3K4me1 and me2. **(B)** Number of MEF-exclusive or ESC-exclusive putative enhancers that are gained or lost across division. The "ESC-specific" enhancer set does not include the 3708 enhancers that are shared between MEF, ESCs, and all reprogramming populations. Inset: Venn diagram of represented enhancers within reprogramming cells against the starting somatic state and ESCs. **(C)** Architecture and relationship of H3K4 methylation marks gained at newly acquired enhancer signatures called after >3 divisions as in (B). Enhancers gain significant H3K4me1 in early proliferative cohorts followed by subsequent H3K4me2 enrichment. **(D)** Composite plot of ESC H3K4me2 enhancer peaks gained in reprogramming populations demonstrate an equivalent CpG hypomethylation in somatic stem cells and ESCs. Alternatively, ESC-specific enhancers that are not acquired after 96 hr of factor induction demonstrate differential and higher mean CpG methylation. Dashed lines highlight somatic CpG methylation in the acquired versus ESC-exclusive sets.



Fraction of sites detected by H3K4me1/3 at >3 Div

	Sites of K4me2 enrichment detected at >3 Div		
	Promoters		Enhancers
	de novo (358)	Enhanced (1082)	>3Div on, MEF off (3804)
H3K4me1	0.28	0.44	0.80
H3K4me3	0.22	0.87	0.19

Figure 8: Dynamics of epigenetic enhancer signatures within reprogramming populations **A** Representative tracks displaying the major shifts in chromatin state upon the clonal induction of reprogramming factors. At the low CpG density promoter (P, blue highlight) of the somatically expressed wisp2, K4me2/me3 enrichment is specific to the promoter and lost with increasing division; this loss accompanies loss of gene body H3K36me3 and expression. A nearby intergenic enhancer (E, blue highlight) is H3K4me1/2 positive in MEFs and is also lost as cells divide. **(B)** Enhancer levels as categorized in Figure 4B exhibit continuous trends across division number. This plot includes an additional 1,235 ES cell specific enhancers (20%) that gain a significant 2-fold increase in H3K4me2 levels but do not reach a suitable threshold for confident scoring (Mann-Whitney U test $p < 10^{-16}$). **(C)** CpG methylation of H3K4me2 enriched enhancer elements: Box plots convey the methylation status of enhancer elements categorized into ES cell exclusive, ES cell/reprogramming shared, and MEF exclusive subsets. Red bars indicate medians and whiskers represent 2.5th and 97.5th percentiles. **(D)** Fraction of promoters and enhancers identified within the >3 division reprogramming populations that are also identified via H3K4me1 or H3K4me3 enrichment.

The sensitivity of H3K4me2 enhancement to DNA methylation is consistent with a model where DNA methylation and associated repressive chromatin structures limit the accessibility of these elements to nuclear reprogramming [3]. Newly activated enhancers that are covered by genome-scale CpG methylation assays exhibit lower methylation levels at the site of H3K4me2 gain and are generally hypomethylated in starting fibroblasts (Figure 7D). These data corroborate changes in promoter histone methylation, where H3K4me2 gain is restricted to sites of high CpG density, which are generally hypomethylated [17] and uniquely amenable to rapid epigenetic reconfiguration [18].

Discussion

To further advance our understanding of the transcription factor-mediated reprogramming process, we isolated clonally induced cells that had undergone defined cell divisions for genomic characterization. Our data demonstrate a robust trend within the early reprogramming population toward a primed epigenetic state that clearly precedes transcriptional activation and complete reprogramming. In addition to suggesting an early coordinated response, our data highlight transcriptional measurement as an incomplete descriptor of the cellular response to reprogramming factor induction. Importantly, gain of H3K4 methylation includes a broader array of notable targets such as key pluripotency and early development genes. As we report, these are particularly enriched for CpG island-containing promoters. Moreover, at sites where H3K4me2 is dynamic, somatic heterochromatin (marked by H3K27me3) is depleted exclusively within the CpG island context but continues to be present in the periphery. Re-establishment of H3K27me3 at bivalent promoters is not observed and must pertain to a later phase of iPSC generation [19].

Our results provide a sensitive measurement of the somatic response to transcription factor activity, which displays a greater trend toward promoter-associated H3K4 methylated euchromatin and may represent a critical step toward transcriptional activation. The continuous behavior of this trend as populations divide clearly demonstrates unique underlying activity that is likely to utilize the endogenous epigenetic machinery. The unexpected genome-wide extent of these events appears mostly limited by sequence context and is most likely to occur within CpG islands in which reprogramming factor regulatory motifs are present. The scope through which promoters and enhancers are modified supports a deterministic model for the initial reprogramming response, because the global events are at expected targets and occur at a detectable frequency similar to what is observed within pluripotent populations. This is further consistent with more recent image-based data [8] and provides an interpretation for the epigenetic response to factor induction, in which genome-wide remodeling occurs within the majority of cells in the induced population, as opposed to selectively within an exclusive subpopulation that will contribute iPSC progeny [7]. The immediate and progressive accumulation of euchromatin-associated marks at ESC-specific promoters and enhancers suggests that a detectable majority of cells in which the factors are induced undergo a certain level of epigenetic reprogramming even in the absence of cell division; these events are immeasurable by expression profiling alone and have to date been largely overlooked.

Moreover, because these events precede detectable transcription, it is likely that the chromatin dynamics observed at the endogenous loci are a critical initial step in the transition to molecular pluripotency. It is intriguing that the promoter dynamics observed are initially restricted to areas of high CpG density and especially CpG islands, whereas peripheral chromatin retains its original, somatic pattern. CpG islands are noted for their plasticity and responsiveness to transcription factor activity [20]. The periphery of these regions behave inversely–they are less CpG rich and more susceptible to DNA methylation and/or extended H3K27me3 spreading, marks that may stably maintain heterochromatin domains in restricted cell types and may require transcriptional activation to be completely depleted. Notably, it is in these regions where somatic epigenetic artifacts might be observed in iPSC characterization studies and a likely explanation could be that these regions are generally less responsive to chromatin remodeling. In our model, the type of mark, the developmental history of its acquisition, and its distribution along target promoter elements all contribute to the response observed. At CpG-dense, hypomethylated transcription start sites, factor expression is sufficient to induce the rapid redistribution of H3K4me2 marks at the promoter that may signal or prime that locus for transcriptional activation. This principle is recapitulated at enhancer sites, where H3K4me2 gain is restricted to somatically hypomethylated regions. As discussed earlier, factor induction alone is not sufficient for complete reprogramming. Instead, the process probably depends on the presence of further chromatin remodeling complexes or transcriptional recruitment elements that may be unavailable in somatic cells.

In conclusion, our data argue for an orchestrated response that yields an epigenetically definable intermediate state in the earliest stages of the reprogramming timeline. However, it cannot as of yet be ascertained if the continuation to full pluripotency is predetermined by existing effectors within a select subpopulation or by stochastic activation of these players in iPSC-forming lineages. It is also likely that these epigenetic reprogramming events describe the limiting effect of the four factors (OSKM) themselves as they act within a population where only a select subset will progress to endogenous target activation; transition through this phase toward complete reprogramming probably involves additional factors. Regardless, continued dissection of the reprogramming process promises for a comprehensive identification of a sufficient factor set for complete and safe somatic to pluripotent reprogramming.

Experimental Procedures

CFSE labeling and enrichment for proliferative cohorts

Mouse E13.5 fibroblasts were generated by blastocyst injection with doxycycline-inducible Oct4, Sox2, Klf4, and c-Myc primary iPSCs as previously described. Cells were passaged several times and serum starved with 0.5% FBS-containing medium for 18 hr before CFSE labeling. Cells were labeled with CFSE in 5x10^6 cell batches with 5 ÎijM cellTrace CFSE (Invitrogen) in PBS according to the manufacturer's protocol and plated at 1x10^6 cells per 10 cm dish in 0.5% FBS for an additional 12 hr before the induction of OSKM-reprogramming factors. Factors were induced with 2 Îijg/ml doxycycline-supplemented medium in either 0.5% or 15% FBS to control the relative number of proliferation for 96 hr (see Figure 1A). In brief: our "no division" cohort was cultured exclusively in 0.5% FBS-containing medium and each successive proliferative cohort was cultured in 15% FBS-containing medium containing doxycycline medium for 24 hr, 48 hr, and 96 hr. After serum pulsing, cells were switched back into 0.5% FBS medium to quell further division; all samples were cultured in doxycycline-supplemented medium for the entire 96 hr. The relative proliferative number for each cohort was ascertained with a BD LSR II fluorescent cytometer against an uninduced, serum-starved control. RNA was collected with TRIzol (Invitrogen) and cells were crosslinked with 1% formaldehyde.

ChIP-seq library preparation and RRBS

After necessary treatments, approximately 500K MEF cells were crosslinked with 1% formaldehyde for 10 minutes at 37 C. After quenching with glycine for 5 min, the cells were washed twice with ice cold PBS with 10% serum. Cell pellets were re-suspended in 100 ml of lysis buffer (1% SDS, 10mM EDTA, 50mM Tris-HCl, pH 8.1) and incubated on ice for 10 min. The lysate was then diluted with 400 ml of ChIP dilution buffer containing (0.01% SDS, 1.1% Triton X-100, 1.2mM EDTA, 16.7mM Tris-HCl, pH 8.1)). Chromatin was sonicated for 3.5 min using a Branson 250 at 40% power amplitude (pulses: 0.7 second "on", and 1.3 second "off"). The fragmented chromatin was then immunoprecipitated overnight in a total volume of 1 ml ChIP Dilution buffer containing protease inhibitor cocktails (Roche), using: 1 mg/ml K4me1 (Abcam ab8895 lot #151302), 1 mg/ml K4me2 (Abcam ab7766 lot #56293), 1 mg/ml K4me3 (Millipore 07473 lot #DAM1623866), 2 mg/ml K27me3 (Millipore 07449 lot #AM15140) or 1 mg/ml K36me3 (Abcam ab9050 lot #761748) antibody. Next, the samples were incubated with 10 ml of pre-washed Protein A-Sepharose beads at 4 deg C for 2 hours. We then collected the beads by brief centrifugation at 1,000 x g, keeping the unbound fraction to check chromatin fragmentation. Then, the beads were washed twice with 700 ml of each of the following buffers at 4 deg C: Low Salt Immune Complex Wash Buffer (0.1% SDS, 1% Triton X-100, 2mM EDTA, 20mM Tris-HCl, pH 8.1, 150mM NaCl); LiCL wash buffer (0.25M LiCl, 1% NP40, 1% deoxycholate, 1mM EDTA, 10mM Tris-HCl, pH 8.1); and TE (10mM Tris-HCl, 1mM

EDTA, pH 8.0). We used filter columns (Costar 8160) in order to minimize the beads and sample loss during washes. DNA was then eluted from the beads twice in 125 ml of Chip Elution Buffer (0.2% SDS, 0.1 M NaHCO3 supplemented with fresh 5 mM DTT) by incubation at 65 deg C for 10 min. The eluted chromatin and the "input" sample were then incubated at 65 deg C for 5 hrs and Proteinase K digested at 37 deg C for 2 hours. The ChIP DNA was recovered by phenol-chloroform extraction and ethanol precipitation. After validating the ChIP enrichments in the precipitated DNA, ChIP DNA was processed into Illumina sequencing libraries, as described before. Enrichment was confirmed on independently generated ChIP samples via qPCR using the Applied Biosystems 7500 Fast Real Time PCR SystemÂő and Quantitect Sybr Green Master Mix (Qiagen). PolII ChIP was performed identically using a Pan PolII antibody raised against the N-terminal domain (Santa Cruz, sc899 lot #H0510).

Gene expression profiling

Gene expression profiles were acquired with Affymetrix Mouse Genome 430 2.0 Arrays and Robust Multi-Array (RMA)-normalized with GenePattern. ChIP libraries were sequenced with the Illumina Genome Analyzer and mapped to the mouse mm8 genome as previously described [12]. Description of enrichment calculations, statistical analyses, and normalizations are available as Supplemental Information. OSKM factor enrichment was performed with previously published data and analysis [22] [23].

Analysis of genome wide libraries

Enrichment was scored in sliding 1Kb windows and significance (threshold $p < 10^{-3}$) was quantified using an Extreme Value background distribution based on the total number of uniquely aligned reads for a given sample. Such a computational background model assumes a uniform, randomized distribution of reads and is insufficient for complete analysis, as the mapping of reads in a control input sample often deviates from random. As such, the ChIP signal was compared with the sequencing of a matched whole cell extract (WCE) sample in order to decrease false positives resulting from biased sequencing of particular genomic loci (Figure 9). Our analysis utilized the WCE in two ways. First, any windows genome-wide enriching significantly for WCE ($p < 10^{-3}$) were eliminated from subsequent analyses. These sites appear as undocumented repeat-like elements not covered by RepeatMasker, and while some groups report success in analyzing repetitive elements in ChIP-Seq datasets, we discarded them in an attempt to remove potential ambiguities. Second, the ChIP signal in all significant 1Kb windows was required to be at least three-fold enriched over the WCE in that region. In order to compare ChIP-Seq signal intensity across samples of varying sequencing depths, an adjusted score was calculated as read density per ten million aligned reads.

Accession numbers

The data sets are available in the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/gds) under the accession number GSE26100.

Acknowledgments

We would like to thank Tarjei Mikkelsen for critical reading of the manuscript. We would like to apologize to authors whose primary work we didn't cite because of space restrictions. B.E.B. is an early career scientist of the HHMI. A.M. is a New Investigator of the Massachusetts Life Science Center (MLSC) and Pew Scholar. This work was funded by the MLSC and Pew Charitable Trusts.

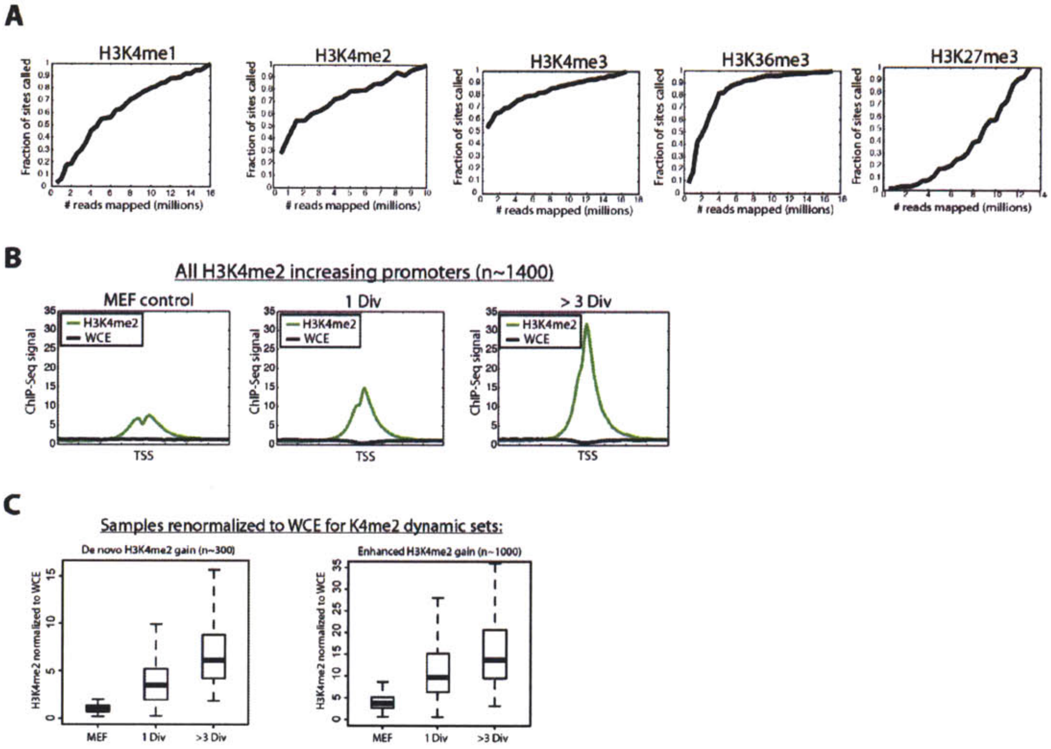


Figure 9: Library sequencing depth and analysis relative to whole cell extract **(A)** Saturation analysis for representative sample (>3 Division Cohort) for histone marks assayed by ChIP-seq in this study. **(B)** Composite plots of H3K4me2 and whole cell extract (WCE) around the transcription start site (TSS) of all H3K4me2 increasing promoters. **(C)** The H3K4me2 signal normalized to WCE for two dynamic gene sets.

References

1. Graf T, Enver T (2009) Forcing cells to change lineages. Nature 462: 587-594.

2. Hanna JH, Saha K, Jaenisch R (2010) Pluripotency and cellular reprogramming: facts, hypotheses, unresolved issues. Cell 143: 508-525.

3. Mikkelsen TS, Hanna J, Zhang X, Ku M, Wernig M et al. (2008) Dissecting direct reprogramming through integrative genomic analysis. Nature 454: 49-55.

4. Jaenisch R, Young R (2008) Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell 132: 567-582.

5. Hanna J, Saha K, Pando B, van Zon J, Lengner CJ et al. (2009) Direct cell reprogramming is a stochastic process amenable to acceleration. Nature 462: 595-601.

6. Meissner A, Wernig M, Jaenisch R (2007) Direct reprogramming of genetically unmodified fibroblasts into pluripotent stem cells. Nat Biotechnol 25: 1177-1181.

7. Yamanaka S (2009) Elite and stochastic models for induced pluripotent stem cell generation. Nature 460: 49-52.

8. Smith ZD, Nachman I, Regev A, Meissner A (2010) Dynamic single-cell imaging of direct reprogramming reveals an early specifying event. Nat Biotechnol

9. Samavarchi-Tehrani P, Golipour A, David L, Sung HK, Beyer TA et al. (2010) Functional genomics reveals a BMP-driven mesenchymal-to-epithelial transition in the initiation of somatic cell reprogramming. Cell Stem Cell 7: 64-77.

10. Wernig M, Lengner CJ, Hanna J, Lodato MA, Steine E et al. (2008) A drug-inducible transgenic system for direct reprogramming of multiple somatic cell types. Nat Biotechnol 26: 916-924.

11. Rahl PB, Lin CY, Seila AC, Flynn RA, McCuine S et al. (2010) c-Myc

regulates transcriptional pause release. Cell 141: 432-445.

12. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553-560.

13. Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK et al. (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. Cell 120: 169-181.

14. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW et al. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39: 311-318.

15. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125: 315-326.

16. Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS et al. (2006) Control of developmental regulators by Polycomb in human embryonic stem cells. Cell 125: 301-313.

17. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J et al. (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 454: 766-770.

18. Xu J, Watts JA, Pope SD, Gadue P, Kamps M et al. (2009) Transcriptional competence and the active marking of tissue-specific enhancers by defined transcription factors in embryonic and induced pluripotent stem cells. Genes Dev 23: 2824-2838.

19. Pereira CF, Piccolo FM, Tsubouchi T, Sauer S, Ryan NK et al. (2010) ESCs require PRC2 to direct the successful reprogramming of differentiated cells toward

pluripotency. Cell Stem Cell 6: 547-556.

20. Ramirez-Carrozzi VR, Braas D, Bhatt DM, Cheng CS, Hong C et al. (2009) A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. Cell 138: 114-128.

21. Gu H, Bock C, Mikkelsen TS, Jager N, Smith ZD et al. (2010) Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. Nat Methods

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

23. Marson A, Levine SS, Cole MF, Frampton GM, Brambrink T et al. (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell 134: 521-533.

Chapter 5

Conclusions and Perspectives

The goal of this thesis was to use an integrative genomics approach to elucidate the role of cis elements in the establishment of repressive chromatin domains. To this effect, I focused on mammalian embryonic stem (ES) cells, reasoning that such a developmentally potent state might better serve to illuminate the points of initial recruitment of chromatin regulators, upstream of further silencing and spreading events associated with lineage specification [1,2]. Indeed, it is difficult to reach a cohesive narrative if one starts with a more fully differentiated cell type, as modifications often cover megabases of genome and encompass a range of regulatory elements. This is most likely due to the folding and sequestering of large regions of genome, for example within macro-scale chromatin structures or to the nuclear periphery, rather than being reflective of a diversity of recruitment elements [3].

At the onset of this thesis, it was clear that the establishment of a transcriptionally repressive chromatin environment in mammals was analogous to that in *Drosophila*, at least with regard to several core components of the Polycomb group (PcG) proteins. However, with mammals lacking homologs for the majority of sequence-specific PcG recruitment proteins found in *Drosophila*, the means of PcG localization remained obscure. Hence I set forth to find the heretofore elusive mammalian Polycomb recruitment element (PRE), and in the process discovered that large unmethylated CpG islands have the innate ability to recruit Polycomb repressive complex 2 (PRC2; Chapters 2 and 3). A separate study (Chapter 4) found that as somatic identity is reset during induced pluripotent stem (iPS) cell reprogramming, CpG islands are central to a coordinated response in which chromatin modifications occur in sequential order and precede transcriptional activation.

Taken together, these studies highlight the role of a particular cis element in the establishment of both active and repressive chromatin domains, and lend insight into a long-standing question in our field as well as suggest several future directions for studying the interplay between DNA sequence and chromatin state.

CpG islands: anomalies in gene regulation

Islands of mammalian DNA enriched with CpG dinucleotides above background were noted for their anomalies long before the sequencing of mammalian genomes or epigenomes. Based on pioneering work by Adrian Bird and colleagues in the 1980s, these elements were found to be mostly free of DNA methylation and associated with highly expressed 'housekeeping' genes [4]. A small subset of islands were DNA methylated and transcriptionally silenced, fitting the logic of repression by DNA methylation. However, confusion arose early on from the discovery of several genes with large CpG island promoters that were DNA methylation free yet also transcriptionally silenced. Observations of human globin genes regulation illuminated this difference: the beta globin cluster was CpG-poor and contained DNA methylation when silenced, whereas the alpha globin cluster contained several CpG islands that remained DNA methylation free even when the locus was silenced in non-erythroid lineages [5]. This was a first hint that this gene class was subject to a different type of transcriptional regulation.

We are finally able to address some of these discrepancies, by considering CpG islands as all or part of the mammalian PRE. While several studies found a strong correlation between CpG islands and PRC2 localization, proposals for the PRE ranged from clusters of motifs analogous to *Drosophila*, to highly conserved non-coding elements, to transposon exclusion zones [1,6,7]. Through both the computational and experimental work of this thesis, I conclude that PRC2 is recruited to large unmethylated CpG islands that lack transcriptional activator motifs for a given cell type. This remains the simplest explanation for the simultaneous localization of PRC2 to thousands of sites with little sequence similarity outside of CpG enrichment, and is supported by data showing a housekeeping CpG island without activator motifs as well as *E. coli* sequences rich in CpGs can recruit PRC2 in mouse ES cells (Chapter 3). In addition, studies published by several other groups have both corroborated and extended our conclusions (see below).

While a complete picture of PcG recruitment has yet to emerge, evidence of CpG island involvement continues to accumulate. One important study by Woo et al in 2010 provided a comprehensive analysis of a functional PRE in the mouse HoxD cluster [8]. Though the data also implicated YY1 motifs and a conserved element in

PRE function, the authors noted a large overlap of these sites with a CpG island. A more recent study demonstrated the ability of a large CpG island-containing human repeat element to recruit PRC2 when placed in a bacterial artificial chromosome (BAC) and introduced into CHO cells [9].

Finally, two studies from the lab of Douglas Higgs and colleagues bring our generalized model of PRC2 recruitment back to the original specific case of globin gene regulation. First, they demonstrate that in non-erythroid cell types, the CpG islands of the human alpha globin genes do indeed serve as recruitment points for PRC2 [10]. In an extensive follow up study, they use a comparative analysis and several transgenic assays between the CpG-rich human locus and the CpG-poor mouse locus to conclude that a CpG island without activator motifs is sufficient for de novo recruitment of PRC2 to the alpha globin genes [11]. They also fragment this region and observe that PRC2 recruitment is encoded redundantly, i.e. each section of the CpG island is capable of functioning as a PRE. Lastly, they used Dnmt3A/B double knockout cell lines to reveal that novel PcG recruitment sites are created when hypermethylated CpG islands lose their DNA methylation.

While a consensus has yet to be reached on the exact definition of the mammalian PRE, the work in this thesis suggests a simple solution to a complex problem. Specifically, it indicates that the innate ability of CpG islands in an inactive state to recruit PRC2 endows them with the capacity to mediate epigenetic regulation through development. Our work and complementary studies by other colleagues continues to provide a voice to CpG islands in the ongoing conversations on PcG recruitment. Exactly how our model may fit into trans recruitment models, or alternative theories, remains to be seen.

Expanding upon the mammalian PRE

As with *Drosophila*, there are most likely recruitment factors which serve as intermediaries between DNA sequence and the core PRC2 components. Given evidence for the role of CpG islands in recruitment, this should inform future studies of trans recruitment models. However, challenges quickly arise when attempting what amounts to a "reverse ChIP," i.e. starting with a nucleic acid sequence and probing for protein interactions. Nonetheless, several recent studies have successfully used this approach to probe proteins interacting with telomeric repeats, TF binding sites, methylated CpGs, modified DNA/chromatin domains, and non-coding RNA [12-16].

A direct line of inquiry might involve using the CpG islands themselves as bait to look for interacting partners in a quantitative proteomics screen. More specifically, one could use stable isotope labeling by amino acids in cell culture (SILAC) to get enrichment of proteins bound to a biotinylated CpG island relative to another sequence, either an AT-rich region or a DNA methylated version of that same CpG island. The result is a direct quantification of proteins bound to different sequence types using peptide isotope ratios [13]. Once potential recruitment factors are identified, they can be genetically fused to a Gal4 DNA binding domain and tested for their ability to silence a reporter gene downstream of Gal4 binding sites, as well as for their potential to recruit PRC2 components to such a synthetic locus. Knockdown studies should also ensue, although this is more complicated, given the ability of PRC2 components to self-propagate once their H3K27me3 mark is present [17,18], nicely reviewed in [19]. That is, elimination of the recruitment protein may have no effect beyond the initial recruitment, and this may be one reason why such recruitment proteins have remained elusive: their effects must be tested in a dynamic system that involves de novo PRC2 recruitment.

A more sophisticated approach to query potential recruitment proteins would involve quantitative mass spectrometry of an endogenous, PRC2-positive locus. While this has been a long sought after technology, it was only achieved relatively recently in a method termed proteomics of isolated chromatin segments (PICh), in which nucleic acid probes are used to isolate and purify a genomic region of interest along with associated proteins [12]. In the first test case, it was able to identify both known and unknown proteins that interacted with human telomeric DNA, which is present at close to 100 copies. While this technique holds promise for future studies, the authors note it would need to be modified to maintain the signal-to-noise ratio for a single copy sequence, such as a CpG island. Alternatively, it is exciting to consider what one might find if a generic CpG-rich probe set could be designed to pull down all or most CpG islands, which at over ten thousand copies should provide enough bound protein as starting material. Even if PRC2-repressed CpG islands could not be isolated specifically, perhaps a picture would emerge for factors that are found at all or most CpG islands (e.g. Kdm2A or Cfp1, respectively) [20-22]. Such a study could be key to deciphering the structure surrounding a cis element known to be subject to a different set of rules regarding histone modifications as well as nucleosome remodeling [23,24].

Implications for trans recruitment models

Even without proteomics-based screening, use of discerning logic combined with a literature search already yields potential targeting candidates: proteins or protein domains which have an affinity for unmethylated CpG-rich DNA, such as CXXC domains and some ARID domains [25,26]. Indeed, several reports have demonstrated an interaction between Jarid2 and PRC2 [27-30]. However, the lack of sequence specificity renders it less likely as the candidate responsible for localization to CpG islands. The converse problem was encountered when the CXXC domain-containing Tet1 was explored as a potential PRC2 recruitment protein: it bound tightly to nearly 90% of CpG islands and its depletion decreased Ezh2 binding, but no interaction was found between the two proteins [31,32]. Thus the effect on PRC2 binding is likely due to an alteration of CpG islands themselves, perhaps resulting from an increase in DNA methylation in the absence of Tet1.

One candidate recruitment factor stands apart as one of the few *Drosophila* PcG targeting proteins conserved between flies and mammals. YY1 is the mammalian homolog of the *Drosophila* protein PHO, a key component of PcG recruitment, and it has held particular appeal because it has been shown to function as both an activator and repressor [33]. However, the data in mammals remains confusing at best. Early reports on YY1 demonstrated a stable interaction with PRC2, through one of its own protein domains as well as through RYBP, which may bridge YY1 and PRC1 [34,35]. However, more recent studies have not been able to replicate the YY1-

PRC2 interaction data, and genome-wide maps of YY1 localization show little overlap with PRC2 targets [28,36]. Nonetheless, at least one study found that both YY1 binding sites as well as RYBP are necessary for full repression by a mammalian PRE in a heterologous context [8]. Yet another paper demonstrated that YY1 can act as a newfound intermediary player by binding both DNA and RNA, opening up possibilities not yet considered [37].

The past decade has seen an explosion in data and theories regarding the roles of long noncoding RNAs (lncRNAs) in gene regulation. Importantly, a 2002 study showing localization of PRC2 with Xist opened up the possibility of lncRNA-based recruitment of PcG in X inactivation [38]. Expanding upon this possibility for targeting in cis, Zhao and colleagues have since isolated an RNA domain responsible for PRC2 interaction, as well as expanded the mechanism to potentially thousands of sites [39,40]. A separate study of short ncRNAs found CpG-rich RNA at PRC2 enriched sites and proposed a cis-based model for recruitment [41]. Transcription from and tethering to a CpG island by ncRNAs provides an appealing model that accounts for both localization and specificity in PRC2 recruitment, though more data is needed to address specific discrepancies between this and trans recruitment. A seminal paper in 2007 demonstrated PRC2 recruitment to the HoxD locus via a novel lncRNA transcribed from the HoxC locus, implicating lncRNAs for both a wider role in development as well as recruitment in trans [42]. Evidence continues to accumulate for lncRNA involvement in gene activation, repression, and molecular scaffolding (nicely reviewed in [43]). It is not yet clear if epigenetic repression by lncRNAs occurs mostly through scaffolding functions or via direct recruitment, and while it is worth noting that many lncRNAs contain a statistically significant GC bias, this may simply be related to constraints in sequence content required for secondary structure formation.

More data are needed to clarify the role of the above recruitment candidates, and is anxiously awaited, as it should shed light on targeting mechanisms as well as further refine the specific sequence characteristics within CpG islands that allow for PRC2 localization.

A coordinated response at CpG islands in reprogramming

While the area of iPS cell reprogramming continues to find new avenues for disease modeling and potential therapeutics, my interest was in utilizing this system to learn fundamental principles of chromatin dynamics and transcriptional regulation. A study into iPS cell reprogramming was initiated for two main reasons. First, a previous study found that in partially reprogrammed cell lines, PRC2 was aberrantly localized to CpG islands, which hinted to us that this might be a useful tool to study the regulation and misregulation of PcG recruitment in light of our PRE model [44]. Second, another study highlighted macroscopic transitions that occurred at the onset of the induction of the "Yamanaka" factors, and it was thought that this outward transition was reflective of an underlying transition in cellular state [45]. Thus we set out to track changes in histone modifications, DNA methylation, and transcription in the first days of reprogramming.

Though we did not observe the expected intermediary PRC2-associated histone modifications at CpG islands in early reprogramming, we did note a highly coordinated upregulation in H3K4me2 at thousands of loci genomewide. Approximatey 10% of these showed a concomitant increase in gene expression, while 90% did not. Strikingly, the sites with accompanying expression changes were CpG poor, while the sites with the chromatin dynamic alone were CpG island promoters. Thus it appeared that, independent of RNA Pol II, which was not detected at these promoters, chromatin regulators were recruited to this particular element to facilitate previously unforeseen epigenetic transition upstream of a cellular transition. This phenomenon appeared to be only dependent on the underlying cis element, in that this H3K4me2 gain occurred regardless of whether the gene was to be activated or repressed in the final iPS cell state. These findings serve to highlight yet another version of chromatin-based plasticity at CpG islands.

Caveats and extensions for the CpG island PRE model

The CpG island-based recruitment model for PRC2 is not without alternatives, both in place of and in addition to our current hypothesis. First and foremost, it may seem odd that the element I propose to mediate PRC2 repression is the same element that is present at constitutively active genes. However, when one considers that in *Drosophila* a PRE is the same as a Trithorax response element (TRE), this begins to fit in perfectly with an element which can be as powerful a player in bolstering gene expression as it is in repressing it [46]. Indeed, a more recent study in *Drosophila* noted that a surprising number of PcG recruitment proteins associated with transcriptionally active rather than repressive loci [47]. One uncertainty in our findings is how important CpG dinucleotides themselves are in the islands themselves, versus general GC-richness. Ideally a synthetic "GpC" island would be added into the BAC system and tested for de novo PRC2 recruitment. Also, there is the possibility that PRC2 is recruited to as yet undiscovered motifs within a CpG island, and it is the footprint of these proteins that has allowed these particular CpGs to remain unmethylated in embryonic development and the germline, not the other way around. While this is possible, the capacity of GC-rich sequence from "E. coli," which could not be conceived to evolve such motifs, to mediate PRC2 recruitment provide an argue against this.

Another open question is how applicable our findings are for PRC2 recruitment outside of ES cells, though several examples cited so far follow the CpG island paradigm, to varying degrees [8,10,11]. However, one important exception is a study that identified a functional PRE in mouse neural development which does not overlap a CpG island [48]. Notably, the 1.5 kb element was capable of recruiting PRC1 but not PRC2 in a transgenic system. This discrepancy highlights a puzzle that has baffled the PcG community for years: what drives differential recruitment of PRC1 and PRC2. While in the canonical model, PRC2 binds first and PRC1 is later recruited through chromodomains, more recent findings challenge this hierarchy. For example, recruitment of PRC1 in the absence of PRC2 has been documented to occur through REST, ZRF1, Runx1, and noncoding RNAs [49-53]. The interplay between PRC1 and PRC2 and the possibility that PRC1 is recruited by a multitude of different targeting complexes will have to be addressed in the near future.

The bulk of my work has focused on the generation and integrative analysis of genetic and epigenetic data. Through intersection with traditional biology, it has helped advance our understanding of Polycomb recruitment mechanisms, a long-standing question in our field. While many challenges remain, the findings presented here should offer CpG islands a place in the ongoing dialogue on Polycomb recruitment and chromatin dynamics.

References

1. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553-560.

2. Hawkins RD, Hon GC, Lee LK, Ngo Q, Lister R et al. (2010) Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell Stem Cell 6: 479-491.

3. Wang J, Kumar RM, Biggs VJ, Lee H, Chen Y et al. (2011) The Msx1 Homeoprotein Recruits Polycomb to the Nuclear Periphery during Development. Dev Cell

4. Bird A, Taggart M, Frommer M, Miller OJ, Macleod D (1985) A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell 40: 91-99.

5. Bird AP, Taggart MH, Nicholls RD, Higgs DR (1987) Non-methylated CpG-rich islands at the human alpha-globin locus: implications for evolution of the alpha-globin pseudogene. EMBO J 6: 999-1004.

6. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125: 315-326.

7. Mohn F, Weber M, Rebhan M, Roloff TC, Richter J et al. (2008) Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Mol Cell 30: 755-766.

8. Woo CJ, Kharchenko PV, Daheron L, Park PJ, Kingston RE (2010) A region of the human HOXD cluster that confers polycomb-group responsiveness. Cell 140: 99-110.

9. Cabianca DS, Casa V, Bodega B, Xynos A, Ginelli E et al. (2012) A long ncRNA links copy number variation to a polycomb/trithorax epigenetic switch in FSHD muscular dystrophy. Cell 149: 819-831.

10. Garrick D, De Gobbi M, Samara V, Rugless M, Holland M et al. (2008) The role of the polycomb complex in silencing alpha-globin gene expression in nonerythroid cells. Blood 112: 3889-3899.

11. Lynch MD, Smith AJ, De Gobbi M, Flenley M, Hughes JR et al. (2012) An interspecies analysis reveals a key role for unmethylated CpG dinucleotides in vertebrate Polycomb complex recruitment. EMBO J 31: 317-329.

12. Dejardin J, Kingston RE (2009) Purification of proteins associated with specific genomic Loci. Cell 136: 175-186.

13. Mittler G, Butter F, Mann M (2009) A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. Genome Res 19: 284-293.

14. Bartels SJ, Spruijt CG, Brinkman AB, Jansen PW, Vermeulen M et al. (2011) A SILAC-based screen for Methyl-CpG binding proteins identifies RBP-J as a DNA methylation and sequence-specific binding protein. PLoS One 6: e25884.

15. Bartke T, Vermeulen M, Xhemalce B, Robson SC, Mann M et al. (2010) Nucleosome-interacting proteins regulated by DNA and histone methylation. Cell 143: 470-484.

16. Butter F, Scheibe M, Morl M, Mann M (2009) Unbiased RNA-protein interaction screen by quantitative proteomics. Proc Natl Acad Sci U S A 106: 10626-10631.

17. Hansen KH, Bracken AP, Pasini D, Dietrich N, Gehani SS et al. (2008) A model for transmission of the H3K27me3 epigenetic mark. Nat Cell Biol 10: 1291-1300.

18. Margueron R, Justin N, Ohno K, Sharpe ML, Son J et al. (2009) Role of the polycomb protein EED in the propagation of repressive histone marks. Nature

19. Margueron R, Reinberg D (2011) The Polycomb complex PRC2 and its mark in life. Nature 469: 343-349.

20. Blackledge NP, Zhou JC, Tolstorukov MY, Farcas AM, Park PJ et al. (2010) CpG islands recruit a histone H3 lysine 36 demethylase. Mol Cell 38: 179-190.

21. Zhou JC, Blackledge NP, Farcas AM, Klose RJ (2012) Recognition of CpG island chromatin by KDM2A requires direct and specific interaction with linker DNA. Mol Cell Biol 32: 479-489.

22. Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J et al. (2010) CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature 464: 1082-1086.

23. Ramirez-Carrozzi VR, Braas D, Bhatt DM, Cheng CS, Hong C et al. (2009) A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. Cell 138: 114-128.

24. Hargreaves DC, Horng T, Medzhitov R (2009) Control of inducible gene expression by signal-dependent transcriptional elongation. Cell 138: 129-145.

25. Voo KS, Carlone DL, Jacobsen BM, Flodin A, Skalnik DG (2000) Cloning of a mammalian transcriptional activator that binds unmethylated CpG motifs and shares a CXXC domain with DNA methyltransferase, human trithorax, and methyl-CpG binding domain protein 1. Mol Cell Biol 20: 2108-2121.

26. Tu S, Teng YC, Yuan C, Wu YT, Chan MY et al. (2008) The ARID domain of the H3K4 demethylase RBP2 binds to a DNA CCGCCC motif. Nat Struct Mol Biol 15: 419-421.

27. Pasini D, Cloos PA, Walfridsson J, Olsson L, Bukowski JP et al. (2010)

JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. Nature 464: 306-310.

28. Li G, Margueron R, Ku M, Chambon P, Bernstein BE et al. (2010) Jarid2 and PRC2, partners in regulating gene expression. Genes Dev 24: 368-380.

29. Peng JC, Valouev A, Swigut T, Zhang J, Zhao Y et al. (2009) Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. Cell 139: 1290-1302.

30. Shen X, Kim W, Fujiwara Y, Simon MD, Liu Y et al. (2009) Jumonji modulates polycomb activity and self-renewal versus differentiation of stem cells. Cell 139: 1303-1314.

31. Wu H, D'Alessio AC, Ito S, Xia K, Wang Z et al. (2011) Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. Nature 473: 389-393.

32. Williams K, Christensen J, Pedersen MT, Johansen JV, Cloos PA et al. (2011) TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. Nature 473: 343-348.

33. Park K, Atchison ML (1991) Isolation of a candidate repressor/activator, NF-E1 (YY-1, delta), that binds to the immunoglobulin kappa 3' enhancer and the immunoglobulin heavy-chain mu E1 site. Proc Natl Acad Sci U S A 88: 9804-9808.

34. Atchison L, Ghias A, Wilkinson F, Bonini N, Atchison ML (2003) Transcription factor YY1 functions as a PcG protein in vivo. EMBO J 22: 1347-1358.

35. Garcia E, Marcos-Gutierrez C, del Mar Lorente M, Moreno JC, Vidal M (1999) RYBP, a new repressor protein that interacts with components of the mammalian Polycomb complex, and with the transcription factor YY1. EMBO J 18: 3404-3418.

36. Mendenhall EM, Koche RP, Truong T, Zhou VW, Issac B et al. (2010) GC-rich

sequence elements recruit PRC2 in mammalian ES cells. PLoS Genet 6: e1001244.

37. Jeon Y, Lee JT (2011) YY1 Tethers Xist RNA to the Inactive X Nucleation Center. Cell 146: 119-133.

38. Mak W, Baxter J, Silva J, Newall AE, Otte AP et al. (2002) Mitotically stable association of polycomb group proteins eed and enx1 with the inactive x chromosome in trophoblast stem cells. Curr Biol 12: 1016-1020.

39. Zhao J, Sun BK, Erwin JA, Song JJ, Lee JT (2008) Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science 322: 750-756.

40. Zhao J, Ohsumi TK, Kung JT, Ogawa Y, Grau DJ et al. (2010) Genome-wide Identification of Polycomb-Associated RNAs by RIP-seq. Mol Cell 40: 939-953.

41. Kanhere A, Viiri K, Araujo CC, Rasaiyaah J, Bouwman RD et al. (2010) Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. Mol Cell 38: 675-688.

42. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X et al. (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129: 1311-1323.

43. Rinn JL, Chang HY (2012) Genome Regulation by Long Noncoding RNAs. Annu Rev Biochem 81: 145-166.

44. Mikkelsen TS, Hanna J, Zhang X, Ku M, Wernig M et al. (2008) Dissecting direct reprogramming through integrative genomic analysis. Nature 454: 49-55.

45. Smith ZD, Nachman I, Regev A, Meissner A (2010) Dynamic single-cell imaging of direct reprogramming reveals an early specifying event. Nat Biotechnol

46. Ringrose L, Paro R (2004) Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. Annu Rev Genet 38: 413-443.

47. Schuettengruber B, Ganapathi M, Leblanc B, Portoso M, Jaschek R et al. (2009) Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos. PLoS Biol 7: e13.

48. Sing A, Pannell D, Karaiskakis A, Sturgeon K, Djabali M et al. (2009) A vertebrate Polycomb response element governs segmentation of the posterior hindbrain. Cell 138: 885-897.

49. Dietrich N, Lerdrup M, Landt E, Agrawal-Singh S, Bak M et al. (2012) REST-mediated recruitment of polycomb repressor complexes in mammalian cells. PLoS Genet 8: e1002494.

50. Richly H, Rocha-Viegas L, Ribeiro JD, Demajo S, Gundem G et al. (2010) Transcriptional activation of polycomb-repressed genes by ZRF1. Nature 468: 1124-1128.

51. Yu M, Mazor T, Huang H, Huang HT, Kathrein KL et al. (2012) Direct recruitment of polycomb repressive complex 1 to chromatin by core binding transcription factors. Mol Cell 45: 330-343.

52. Schoeftner S, Sengupta AK, Kubicek S, Mechtler K, Spahn L et al. (2006) Recruitment of PRC1 function at the initiation of X inactivation independent of PRC2 and silencing. EMBO J 25: 3110-3122.

53. Yap KL, Li S, Munoz-Cabello AM, Raguz S, Zeng L et al. (2010) Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. Mol Cell 38: 662-674.