Published in final edited form as:
Genomics. 2012 November ; 100(5): 320–326. doi:10.1016/j.ygeno.2012.07.006.

CpG Island Structure and Trithorax/Polycomb Chromatin Domains in Human Cells

David A. Orlando[1,*], **Matthew G. Guenther**[1,*], **Garrett M. Frampton**[1,2,*], and **Richard A. Young**[1,2,†]

[1]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA

[2]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Abstract

TrxG and PcG complexes play key roles in the epigenetic regulation of development through H3K4me3 and H3K27me3 modification at specific sites throughout the human genome, but how these sites are selected is poorly understood. We find that in pluripotent cells, clustered CpG-islands at genes predict occupancy of H3K4me3 and H3K27me3, and these "bivalent" chromatin domains precisely span the boundaries of CpG-island clusters. These relationships are specific to pluripotent stem cells and are not retained at H3K4me3 and H3K27me3 sites unique to differentiated cells. We show that putative transcripts from clustered CpG-islands predict stem-loop structures characteristic of those bound by PcG complexes, consistent with the possibility that RNA facilitates PcG recruitment or maintenance at these sites. These studies suggest that CpG-island structure plays a fundamental role in establishing developmentally important chromatin structures in the pluripotent genome, and a subordinate role in establishing TrxG/PcG chromatin structure at sites unique to differentiated cells.

Keywords

Polycomb; trithorax; stem cell; bivalent; H3K4me3; H3K27me3; stem-loop

1. Introduction

Trithorax group (TrxG) and Polycomb group (PcG) genes were discovered in *Drosophila melanogaster* as activators and repressors of Hox transcription factor genes, which specify cell identity along the anteroposterior axis of segmented animals [1; 2]. TrxG proteins catalyze trimethylation of histone H3 lysine 4 (H3K4me3) at the promoters of active genes and facilitate maintenance of active gene states during development. PcG proteins catalyze trimethylation of histone H3 lysine 27 (H3K27me3) and function to silence genes encoding key regulators of development. TrxG and PcG proteins have been implicated in control of cell identity, proliferation, X inactivation, genomic imprinting and cancer [3; 4; 5; 6; 7].

[†]Corresponding Author: Richard A. Young, Ph.D., Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, Phone: (617) 258-5218, Fax: (617) 258-0367, young@wi.mit.edu.
[*]These authors contributed equally

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

How TrxG and PcG protein complexes are recruited to their sites of action in mammalian cells is not fully understood. Nucleosomes with H3K4me3 are found immediately downstream of transcription initiation sites [8] consistent with proposals that TrxG complexes are recruited to active promoter regions by transcription factors and/or the transcription initiation apparatus [9; 10; 11]. In embryonic stem (ES) cells and induced pluripotent stem (iPS) cells, regions that are occupied by PcG proteins contain nucleosomes with both H3K4me3 and H3K27me3, and this “bivalent” structure frequently occurs at the promoters of genes encoding key developmental regulators [12; 13]. There is evidence that ncRNAs and DNA binding cofactors have roles in targeting PcG protein complexes to some of these sites [14; 15; 16; 17; 18; 19; 20; 21; 22; 23; 24; 25; 26; 27; 28]. Several models have been proposed to explain the recruitment of PcG proteins genome-wide [29; 30; 31], some of which suggest that transcription of CG-rich domains contributes to PcG binding.

CpG islands are small genomic elements, ~1kb in length, which have been “protected” from the loss of CG dinucleotide content that is characteristic of the rest of the genome [32]. Two features of CpG islands led us to investigate the potential relationship between them and the chromatin structure catalyzed by TrxG and PcG proteins. First, the majority of CpG islands span the transcription start sites (TSS) of genes, and this is coincident with the genomic locations of nucleosomes with H3K4me3 and H3K27me3. Second, ncRNA species that are known to recruit PcG complexes have characteristic GC-rich stem loop structures that are required for their function [16; 19; 20; 21; 22; 33], suggesting that transcription of CpG islands might generally be involved in PcG complex recruitment.

2. Results and Discussion

TrxG/PcG chromatin structure and local CpG content

We compared the genome-wide location of CpG islands to the occupancy of H3K4me3 and H3K27me3 modified nucleosomes for a collection of cells using data from both published and unpublished ChIP-Seq experiments (Supplemental Information, Table S1). These cells included human ES and iPS cells, primary CD4+ T cells, IMR90 fetal lung fibroblasts, T-cell lymphoma cells, CD24+ and CD44+ mammary cells, and HCC1954 breast cancer cells. In all these cells, the presence of H3K4me3 nucleosomes throughout the genome was highly correlated with the location of CpG islands (Table S2), as shown in a correlation plot for human ES and iPS cells in Figure 1. The majority of H3K4me3 bound regions were associated with CpG islands in all cell types examined (57%–88%, Figure S1). H3K27me3 modified nucleosomes also occurred at CpG islands, albeit at a smaller set of these islands than H3K4me3 nucleosomes (Figure S1). In ES and iPS cells, approximately 65% of H3K27me3 bound regions occur at CpG islands.

To further explore the relationship between genes, CpG islands, and H3K4me3 and H3K27me3 histone modifications, we first classified the complete set of human genes by the number of CpG islands in their promoter regions (Figure 2A). We found that ~30% of genes do not have a CpG island at their TSS (Class I genes, 7301 genes), ~60% of genes have a single CpG island at their TSS (Class II genes, 13101 genes), and ~10% of genes have two or more CpG islands in the vicinity of their TSS (Class III genes, 3928 genes)(Table S3).

We then investigated the relationship between the number of CpG islands at genes and occupancy by H3K4me3 and H3K27me3 modified histones. In ES and iPS cells, genes with zero CpG islands were rarely occupied by H3K4me3 (15–17%) or H3K27me3 (9–19%) modified nucleosomes (Figure 2B, Figure S2). Genes with one or more CpG island were almost always occupied by H3K4me3 modified nucleosomes (~94–96%; Figure 2B). In contrast, the genes with one or more CpG islands were occupied by H3K27me3 nucleosomes in a manner that was positively correlated with the number of CpG islands

(Figure 2B). Thus, for genes containing CpG islands, occupancy with H3K4me3 modified nucleosomes is highly likely with just a single CpG island, whereas occupancy by H3K27me3 modified nucleosomes is increasingly likely with increasing numbers of CpG islands. These observations suggest that the number of CpG islands that occur at a gene's promoter is predictive of the TrxG/PcG chromatin structure at those genes in pluripotent stem cells.

We examined the genes with bivalent chromatin and multiple CpG islands in pluripotent cells in more detail and observed that the H3K4me3/H3K27me3 modified nucleosomes and CpG island clusters spanned the same genomic regions and that the peaks of H3K4me3 occupancy in these regions were aligned with the individual CpG islands, as shown in Figure 2A for class III genes. This phenomenon, the spread of bivalent nucleosomes across the span of clustered CpG islands, was most prominent in the four Hox gene clusters, which each contain ~40 CpG islands (Figure S3A–D, Table S4A). Similar bivalent domain spreading was also evident at approximately 1,000 genes encoding developmental regulators and cellular signaling components (Table S4B), including DLX and IRX family members (Figure S3E–H). These results indicate that the TrxG/PcG chromatin structure at bivalent genes is highly aligned with the local CpG island structure (Additional file S1–S4).

Chromatin and CpG islands in differentiated cells

We next investigated whether the relationships observed between CpG island structure and H3K4me3 and H3K27me3 occupancy in pluripotent cells were preserved in a spectrum of differentiated human cell types (Primary CD4+ T cells, IMR90 fetal lung fibroblasts, T-cell lymphoma cells, CD24+ and CD44+ mammary cells, and HCC1954 breast cancer cells) (Figure 3). Initial inspection of the data suggested that these relationships were similar in the pluripotent and differentiated cell types. Indeed, there is a large population of genes that are similarly occupied by H3K4me3 and/or H3K27me3 in both pluripotent and differentiated cell types (Figure 3A, B, left panels). However, genes occupied by H3K4me3 and/or H3K27me3 exclusively in differentiated cells showed striking differences in the relationships between modified nucleosomes and CpG islands. H3K4me3 occupied genes specific to differentiated cells were rarely associated with CpG islands (8–17%)(Figure 3A, right panel, Table S3A) and H3K27me3 occupied genes specific to differentiated cells were modestly associated with CpG islands (36–58%; Figure 3B, right panel; Table S3B). These results show that there is less of a correlation between CpG islands and occupancy by H3K4me3 and H3K27me3 nucleosomes at genes that are occupied by these modified nucleosomes exclusively in differentiated cells.

Why do nucleosomes with histone H3K4me3 occur at the majority of genes containing CpG islands in human cells (Figure 1, 2)? Previous studies have indicated that transcription initiation occurs at most, if not all, of these sites [8] and TrxG proteins can be recruited via the transcription apparatus [34]. It is also possible that proteins that bind to CG dinucleotides, such as Cfp1, recruit TrxG-containing COMPASS (Complex Proteins Associated with Set1) complexes to all these sites [11; 35; 36; 37].

RNA stem-loop structure and CpG islands

Why is there a nearly linear relationship between occupancy by H3K27me3 nucleosomes and the number of CpG islands in pluripotent cells (Figure 2B)? RNA species containing GC-rich stem loop structures can contribute to PcG complex recruitment [16; 21; 33; 38]. It has been proposed that transcripts from GC-rich promoter regions frequently contain these structures, suggesting a general model for establishing PcG domains that involves transcripts from these domains [16; 31]. We investigated the possibility that CpG islands are generally transcribed in human ES cells and that these RNA species are likely to form the GC-rich stem loop structures known to recruit PcG proteins. Previous studies of transcripts in ES cells have noted that most protein-coding genes experience some level of transcription initiation [8; 39] and that bidirectional transcription can occur in CpG islands [40; 41]. Small ncRNAs are pervasively transcribed from CpG islands and some have been shown to form CG-rich hairpin structures that can directly bind PcG complex proteins Suz12 and EZH2 [21; 33]. Using the known examples of small ncRNAs that bind to PcG proteins as a guide (Figure 4A) [21; 33], we examined the promoter sequence of every gene and searched for sequence elements that would be likely to form the characteristic GC-rich stem loop structure (Supplemental Information). We found that the promoter regions with two or more CpG islands have a much higher probability of forming the GC-rich RNA structures that bind PcG protein complexes than those that contain zero or one CpG island ($p < 10^{-100}$, Figure 4B; Table S3C, See Supplemental Information for details regarding calculation of p-value). Furthermore, we found that among genes with one CpG island and among genes with more than one CpG island, those that were occupied by H3K27me3 had a much higher probability of forming the characteristic stem loops than those that were not occupied by H3K27me3 ($p < 10^{-61}$; Figure 4C).

Previous studies have identified some aspects of the phenomena we describe here, but have not revealed the extent of the relationship between CpG islands and H3K4me3 and H3K27me3 nucleosomes in human pluripotent stem cells or addressed how polycomb recruitment might occur at bivalent domains. Previous work [11; 42] has shown that, in ES cells, H3K4me3 occupancy is correlated with CpG islands. We have found that H3K4me3 nucleosomes and CpG islands are almost entirely co-incident across gene loci in pluripotent cells and that this relationship does not extend to H3K4me3 nucleosomes that are specific to differentiated cell types. Other studies have noted that PcG occupancy can be predicted from the locations, sizes, and motif contents of CpG islands [43; 44; 45], by conservation properties [46] or by DNA sequence motifs [47; 48], but did not reveal that in pluripotent stem cells, there is a positive relationship between the number of CpG islands at genes, occupancy of these genes by H3K27me3 nucleosomes and the likelihood that CpG island transcripts can form polycomb recruiting structures. One of the most striking insights from this analysis is that the positive relationships between CpG islands and nucleosomes with H3K4me3 or H3K27me3 are specific to pluripotent cells, and do not extend to genes uniquely occupied by these modified nucleosomes in differentiated cells. While CpG transcription may still play a role in the deposition of H3K27me3 and H3K4me3 in differentiated cells at those genes which are already marked in ES cells, we suggest that at genes uniquely occupied in differentiated cells, mechanisms of H3K27me3 and H3K4me3 deposition that are independent of CpG density may prevail.

3. Conclusions

We have identified several striking relationships between the histone modifications catalyzed by TrxG and PcG proteins and CpG island structure in human gene promoters. These relationships are specific to pluripotent stem cells, and are not retained at H3K4me3 and H3K27me3 sites unique to differentiated cells. First, genes that do not have a CpG island at their start site are rarely occupied by H3K4me3 or H3K27me3. Second, H3K4me3 modified nucleosomes and CpG islands are coincident genome-wide. Third, genes with increasing numbers of CpG islands are increasingly likely to be occupied by H3K4me3 and H3K27me3, and increasingly likely to produce transcripts with structures known to recruit polycomb. Finally, where they are found, these H3K4me3 and H3K27me3 bivalent chromatin domains precisely span the CpG island clusters. We conclude that CpG island structure plays a fundamental role in defining TrxG/PcG chromatin structure in human pluripotent stem cells and suggest that CpG islands may play a subordinate role in establishing TrxG/PcG chromatin structure at sites unique to differentiated cells. It will be valuable to understand the mechanisms that cause differentiated cells to behave differently than pluripotent cells in the acquisition of TrxG/PcG chromatin structure.

4. Materials and Methods

4.1 Cells and cell culture

The human induced pluripotent stem (iPS) cell line M2^{3F} was derived as described previously [49] and was maintained on mitomycin C-inactivated mouse embryonic fibroblast (MEF) feeder layers in hESC medium (DMEM/F12 [Invitrogen] supplemented with 15% FBS [Hyclone], 5% KnockOut Serum Replacement [Invitrogen], 1 mM glutamine [Invitrogen], 1% nonessential amino acids [Invitrogen], 0.1 mM β-mercaptoethanol [Sigma], and 4 ng/ml FGF2 [R&D Systems]). Cultures were passaged every 5 to 7 days either manually or enzymatically with collagenase type IV (Invitrogen; 1.5 mg/ml). hiPS cell lines were passaged 15–25 times prior to ChIP-Seq analysis. Information about cells and cell culture for the human ES cell lines WIBR1, WIBR2, and WIBR7, the human iPS cell lines iPS PDB^{1lox}-17puro-5, and iPS PDB^{1lox}-21puro-26, were described previously [12; 50] (GEO accession number GSE23455; Table S1). CUTLL, CD4+, CD24+, CD44+, HCC1954, and IMR90 data were taken from GSE29600, GSE15735, GSE26137, GSE26137, GSE29118, and GSE16256 respectively.

4.2 Chromatin immunoprecipitation

Protocols describing chromatin immunoprecipitation (ChIP) materials and methods can be downloaded from http://web.wi.mit.edu/young/hES_PRC and have previously been described in detail [51; 52].

Briefly, cells were grown to a final count of ~5×10^7 cells to obtain starting material for six chromatin immunoprecipitations. Cells were chemically cross-linked and sonicated to solubilize and shear cross-linked DNA. Whole cell extract was incubated overnight at 4 degrees C with 10μl of Dynal Protein G magnetic beads that had been pre-incubated with approximately 3 μg of the appropriate antibody. Each individual immunoprecipitation used 1/6 of the 3ml total, or ~8 × 10^6 cells per IP. The immunoprecipitation was allowed to proceed overnight. Beads were washed three times (3 × 1.5ml) with RIPA buffer and one time (1 × 1.5ml) with TE containing 50 mM NaCl. Bound complexes were eluted from the beads by heating at 65 degrees C with occasional vortexing and cross-linking was reversed by overnight incubation at 65 degrees C. Immunoprecipitated DNA and whole cell extract DNA were then purified by treatment with RNAse A, proteinase K and two phenol:chloroform:isoamyl alcohol extractions prior to solexa sample preparation. All protocols for Solexa sample preparation and sequencing are provided by Illumina (http://www.illumina.com/). The ChIP antibodies used were ab8580 (Abcam) for H3K4me3 and ab6002 (Abcam) for H3K27me3. These antibodies were reported to be selective for the H3K4me3 and H3K27me3 modifications, respectively (Supplemental Information).

4.3 ChIP-Seq sample preparation and Solexa sequencing

All protocols for Solexa sample preparation and sequencing are provided by Illumina (http://www.illumina.com/).

4.4 ChIP-Seq density calculation

The genome was divided into bins 25 base pairs in width, beginning at the first base of each chromosome. Each ChIP-Seq read was shifted 100 bp from its mapped genomic position and strand to the approximate middle of the sequenced DNA fragment. Subsequently, the ChIP-Seq density within each genomic bin was calculated as the number of ChIP-Seq reads mapping within a 1kb window (+/− 500bp) surrounding the middle of that genomic bin.

4.5 Identification of genes occupied by H3K4me3 and H3K27me3

The genomic coordinates of the full set of transcripts from the RefSeq database (http://www.ncbi.nlm.nih.gov/RefSeq/) from the March 2006 version of the human genome sequence (NCBI Build 36.1, hg18) was downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTables) on September 1, 2010.

For each RefSeq gene the peak ChIP-Seq density in the region +/−1 kb around the transcription start site (TSS) for H3K4me3 and +/− 25kb around the TSS for H3K27me3 was examined. A gene was considered to be occupied by H3K4me3 or H3K27me3 if there was a statistically significant region of ChIP-Seq density in this region. See Supplemental Information for a more complete description of identification of statistically significant binding. Due to differences in data quality it was necessary to select different cutoffs for each dataset. See Table S1 for the statistical cutoffs used, and number of enriched regions in each dataset. Genes were considered differentially occupied by H3K4me3 or H3K27me3 between pluripotent and differentiated experiments if a gene was called occupied in at least one cell-type of one group and not called occupied in any of the cell types of the other. A summary of the genes occupied by H3K4me3 and H3K27me3 in each cell line is provided (Table S3).

4.6 Identification of CpG islands and assignment to genes

The base composition criteria that were originally used to define CpG islands [53] were created long before the complete sequence of the genome was known. Since then, new definitions of CpG islands have been developed, which offer greatly enhanced sensitivity and specificity in the genome-wide identification of CpG islands [54; 55; 56; 57; 58].

We used a modified version of these methods optimized to provide the greatest sensitivity and specificity in comparing CpG island positions to ChIP-Seq datasets. We tabulated the local CG dinucleotide frequency in a 1 kb window (+/− 500bp) at every position in the genome. CG dinucleotides in protein coding regions of the genome were excluded. Scanning the genome in 25 bp bins, using the CG frequency at the middle position of each bin. We identified bins in which the CG frequency was greater than or equal to 4.6%. Adjoining bins were collapsed into regions and regions that were less than 300 bin in length were excluded. This method identifies 42,371 CpG islands in the NCBI build 36.1 (hg18) of the human genome (Table S2).

In order to assign CpG islands to genes, the following method was used. The genome was scanned and CpG islands within 4,000 bp of one another were merged into CpG island clusters. Most of these clusters consisted of only one CpG island, but there were several thousand clusters of multiple CpG islands. The region at the transcription start site of each gene (+/− 1kb) was overlapped with the CpG island clusters and if a CpG island cluster overlapped with the TSS all CpG islands from that cluster were assigned to that gene.

4.7 Gene ontology analysis

For gene ontology analysis ChIP-Seq results from the human ES cell line WIBR2 were used. Gene ontology analysis was performed using the online tool DAVID (http://david.abcc.ncifcrf.gov/). A summary of the genes annotated as encoding regulators of development and homeobox transcription factors is provided (Table S4).

4.8 Analysis of RNA secondary structure

For each gene, the sequence +/− 5kb around the TSS was analyzed for sequences which may form the characteristic CG rich, PcG recruiting, hairpin structure [21; 33]. A 28bp window was slid across each sequence in 1bp increments and for each window the minimum free energy of that sequence folding into the structure shown in Figure 4A was calculated using rnaEval [59]. In order to account for the possibility of the hairpin forming transcript being generated from transcription of either strand and/or in either direction, the free energy was calculated for four different sequence/structure pairs: the 28bp sequence and its reverse complement were compared to the structure in Figure 4A as well as its mirror image (with the smaller loop on the left). The minimum free energies of those 4 combinations was used as the free energy for that window. Windows with a CG content of at least 50% and a minimum free energy at or below −5 kcal/mol were counted as a potential PcG recruiting hairpin sequence. A summary of the number of hairpin hits for each gene is provided (Table S3).

The fold enrichment shown in Figure 4B and 4C, is a measure of the enrichment of genes with a particular range of potential hairpin hits versus the expected number of genes. As an example of how this is calculated, consider the calculation for the fold enrichment of genes with 15–20 hairpin hits in the multiple CpG class. First we calculate the percentage of multiple CpG genes with between 15 and 20 possible hairpins, ~15%. Then, we calculate the percentage of all genes with this range of hairpin hits, ~8%. The fold enrichment for 15–20 hairpin hits in the multiple CpG class is then simply the observed percentage divided by the expected percentage, or 15/8, for a fold enrichment of ~1.9. This calculation was done for each RNA hairpin range in each gene class. The reported p-values of enrichment were calculated using a one-sided t-test on the distributions of number of potential hairpin hits between the relevant gene classes. Additional supplemental information is available on the *Genomics* website.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank T. Volkert, J. Love, S. Gupta, J. Kwon and V. Dhanapal for assistance with ChIP-Seq experiments as well as T. Lee, L. Lawton, F. Soldner, and J. Reddy for helpful discussions about the manuscript

Abbreviations

ES cell	embryonic stem cell
iPS cell	induced pluripotent stem cell
CpG	cytosine phosphate guanidine
ncRNA	noncoding RNA
PcG	polycomb group
TrxG	trithorax group
H3K4me3	histone H3-lysine-4-trimethyl
H3K27me3	histone H3-lysine-27-trimethyl

References

1. Ringrose L, Paro R. Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. Annu Rev Genet. 2004; 38:413–43. [PubMed: 15568982]
2. Schuettengruber B, Chourrout D, Vervoort M, Leblanc B, Cavalli G. Genome regulation by polycomb and trithorax proteins. Cell. 2007; 128:735–45. [PubMed: 17320510]

3. Schwartz YB, Pirrotta V. Polycomb silencing mechanisms and the management of genomic programmes. Nat Rev Genet. 2007; 8:9–22. [PubMed: 17173055]

4. Simon JA, Kingston RE. Mechanisms of polycomb gene silencing: knowns and unknowns. Nat Rev Mol Cell Biol. 2009; 10:697–708. [PubMed: 19738629]

5. Surface LE, Thornton SR, Boyer LA. Polycomb group proteins set the stage for early lineage commitment. Cell Stem Cell. 2010; 7:288–98. [PubMed: 20804966]

6. Margueron R, Reinberg D. The Polycomb complex PRC2 and its mark in life. Nature. 2011; 469:343–9. [PubMed: 21248841]

7. Rodriguez-Paredes M, Esteller M. Cancer epigenetics reaches mainstream oncology. Nat Med. 2011; 17:330–9. [PubMed: 21386836]

8. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. A chromatin landmark and transcription initiation at most promoters in human cells. Cell. 2007; 130:77–88. [PubMed: 17632057]

9. Li B, Carey M, Workman JL. The role of chromatin during transcription. Cell. 2007; 128:707–19. [PubMed: 17320508]

10. Eissenberg JC, Shilatifard A. Histone H3 lysine 4 (H3K4) methylation in development and differentiation. Dev Biol. 2010; 339:240–9. [PubMed: 19703438]

11. Thomson JSP, Selfridge J, Clouaire T, Guy J, Webb S, Kerr A, Deaton A, Andrews R, James KD, Turner DJ, Illingworth R, Bird A. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature. 2010; 464:1082–1086. [PubMed: 20393567]

12. Guenther MG, Frampton GM, Soldner F, Hockemeyer D, Mitalipova M, Jaenisch R, Young RA. Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells. Cell Stem Cell. 2010; 7:249–57. [PubMed: 20682450]

13. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell. 2006; 125:315–26. [PubMed: 16630819]

14. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL, Wang Y, Brzoska P, Kong B, Li R, West RB, van de Vijver MJ, Sukumar S, Chang HY. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature. 2010; 464:1071–6. [PubMed: 20393566]

15. Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A, Regev A, Lander ES, Rinn JL. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci U S A. 2009; 106:11667–72. [PubMed: 19571010]

16. Kanhere A, Viiri K, Araujo C, Rasaiyaah J, Bouwman R, Whyte W, Pereira C, Brookes E, Walker K, Bell G, Ponbo A, Fischer A, Young R, Jenner R. Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. Mol Cell. 2010 in press.

17. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, Chang HY. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell. 2007; 129:1311–23. [PubMed: 17604720]

18. Pandey RR, Mondal T, Mohammad F, Enroth S, Redrup L, Komorowski J, Nagano T, Mancini-Dinardo D, Kanduri C. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Molecular Cell. 2008; 32:232–46. [PubMed: 18951091]

19. Yap KL, Li S, Munoz-Cabello AM, Raguz S, Zeng L, Mujtaba S, Gil J, Walsh MJ, Zhou MM. Molecular Interplay of the Noncoding RNA ANRIL and Methylated Histone H3 Lysine 27 by Polycomb CBX7 in Transcriptional Silencing of INK4a. Molecular Cell. 2010; 38:662–674. [PubMed: 20541999]

20. Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E, Chang HY. Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. Science. 2010

21. Zhao J, Sun BK, Erwin JA, Song JJ, Lee JT. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science. 2008; 322:750–6. [PubMed: 18974356]

22. Zhao J, Ohsumi TK, Kung JT, Ogawa Y, Grau DJ, Sarma K, Song JJ, Kingston RE, Borowsky M, Lee JT. Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol Cell. 2010; 40:939–53. [PubMed: 21172659]

23. Shen X, Kim W, Fujiwara Y, Simon MD, Liu Y, Mysliwiec MR, Yuan GC, Lee Y, Orkin SH. Jumonji modulates polycomb activity and self-renewal versus differentiation of stem cells. Cell. 2009; 139:1303–14. [PubMed: 20064376]

24. Peng JC, Valouev A, Swigut T, Zhang J, Zhao Y, Sidow A, Wysocka J. Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. Cell. 2009; 139:1290–302. [PubMed: 20064375]

25. Li G, Margueron R, Ku M, Chambon P, Bernstein BE, Reinberg D. Jarid2 and PRC2, partners in regulating gene expression. Genes Dev. 2010; 24:368–80. [PubMed: 20123894]

26. Pasini D, Cloos PA, Walfridsson J, Olsson L, Bukowski JP, Johansen JV, Bak M, Tommerup N, Rappsilber J, Helin K. JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. Nature. 2010; 464:306–10. [PubMed: 20075857]

27. Kaneko S, Li G, Son J, Xu CF, Margueron R, Neubert TA, Reinberg D. Phosphorylation of the PRC2 component Ezh2 is cell cycle-regulated and up-regulates its binding to ncRNA. Genes Dev. 2010; 24:2615–20. [PubMed: 21123648]

28. Yang L, Lin C, Liu W, Zhang J, Ohgi KA, Grinstein JD, Dorrestein PC, Rosenfeld MG. ncRNA- and Pc2 methylation-dependent gene relocation between nuclear structures mediates gene activation programs. Cell. 2011; 147:773–88. [PubMed: 22078878]

29. Mendenhall EM, Bernstein BE. Chromatin state maps: new technologies, new insights. Curr Opin Genet Dev. 2008; 18:109–15. [PubMed: 18339538]

30. Margueron R, Reinberg D. Chromatin structure and the inheritance of epigenetic information. Nat Rev Genet. 2010; 11:285–96. [PubMed: 20300089]

31. Guenther MG, Young RA. Repressive transcription. Science. 2010; 329:150–1. [PubMed: 20616255]

32. Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes Dev. 2011; 25:1010–22. [PubMed: 21576262]

33. Wutz A, Rasmussen TP, Jaenisch R. Chromosomal silencing and localization are mediated by different domains of Xist RNA. Nat Genet. 2002; 30:167–74. [PubMed: 11780141]

34. Eissenberg JC, Shilatifard A. Histone H3 lysine 4 (H3K4) methylation in development and differentiation. Dev Biol. 2009; 339:240–9. [PubMed: 19703438]

35. Demers C, Chaturvedi CP, Ranish JA, Juban G, Lai P, Morle F, Aebersold R, Dilworth FJ, Groudine M, Brand M. Activator-mediated recruitment of the MLL2 methyltransferase complex to the beta-globin locus. Mol Cell. 2007; 27:573–84. [PubMed: 17707229]

36. Sarvan S, Avdic V, Tremblay V, Chaturvedi CP, Zhang P, Lanouette S, Blais A, Brunzelle JS, Brand M, Couture JF. Crystal structure of the trithorax group protein ASH2L reveals a forkhead-like DNA binding domain. Nat Struct Mol Biol. 2011; 18:857–9. [PubMed: 21642971]

37. Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, Lajoie BR, Protacio A, Flynn RA, Gupta RA, Wysocka J, Lei M, Dekker J, Helms JA, Chang HY. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature. 2011; 472:120–4. [PubMed: 21423168]

38. Zhao J, Ohsumi TK, Kung JT, Ogawa Y, Grau DJ, Sarma K, Song JJ, Kingston RE, Borowsky M, Lee JT. Genome-wide identification of polycomb-associated RNAs by RIP-seq. Molecular Cell. 2010; 40:939–53. [PubMed: 21172659]

39. Rahl PB, Lin CY, Seila AC, Flynn RA, McCuine S, Burge CB, Sharp PA, Young RA. c-Myc regulates transcriptional pause release. Cell. 2010; 141:432–45. [PubMed: 20434984]

40. Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, Sharp PA. Divergent transcription from active promoters. Science. 2008; 322:1849–51. [PubMed: 19056940]

41. Flynn RA, Almada AE, Zamudio JR, Sharp PA. Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. Proc Natl Acad Sci U S A. 2011; 108:10460–5. [PubMed: 21670248]

42. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A,

Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature. 2007; 448:553–60. [PubMed: 17603471]

43. 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, Bernstein BE. Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. PLoS Genet. 2008; 4:e1000242. [PubMed: 18974828]

44. Mendenhall EM, Koche RP, Truong T, Zhou VW, Issac B, Chi AS, Ku M, Bernstein BE. GC-rich sequence elements recruit PRC2 in mammalian ES cells. PLoS Genet. 2010; 6:e1001244. [PubMed: 21170310]

45. Lynch MD, Smith AJ, De Gobbi M, Flenley M, Hughes JR, Vernimmen D, Ayyub H, Sharpe JA, Sloane-Stanley JA, Sutherland L, Meek S, Burdon T, Gibbons RJ, Garrick D, Higgs DR. An interspecies analysis reveals a key role for unmethylated CpG dinucleotides in vertebrate Polycomb complex recruitment. EMBO J. 2011; 31:317–29. [PubMed: 22056776]

46. Tanay A, O'Donnell AH, Damelin M, Bestor TH. Hyperconserved CpG domains underlie Polycomb-binding sites. Proc Natl Acad Sci U S A. 2007; 104:5521–6. [PubMed: 17376869]

47. Liu Y, Shao Z, Yuan GC. Prediction of Polycomb target genes in mouse embryonic stem cells. Genomics. 2010; 96:17–26. [PubMed: 20353814]

48. Woo CJ, Kharchenko PV, Daheron L, Park PJ, Kingston RE. A region of the human HOXD cluster that confers polycomb-group responsiveness. Cell. 2010; 140:99–110. [PubMed: 20085705]

49. Soldner F, Hockemeyer D, Beard C, Gao Q, Bell GW, Cook EG, Hargus G, Blak A, Cooper O, Mitalipova M, Isacson O, Jaenisch R. Parkinson's disease patient-derived induced pluripotent stem cells free of viral reprogramming factors. Cell. 2009; 136:964–77. [PubMed: 19269371]

50. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. High-resolution profiling of histone methylations in the human genome. Cell. 2007; 129:823–37. [PubMed: 17512414]

51. Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, Kumar RM, Chevalier B, Johnstone SE, Cole MF, Isono K, Koseki H, Fuchikami T, Abe K, Murray HL, Zucker JP, Yuan B, Bell GW, Herbolsheimer E, Hannett NM, Sun K, Odom DT, Otte AP, Volkert TL, Bartel DP, Melton DA, Gifford DK, Jaenisch R, Young RA. Control of developmental regulators by Polycomb in human embryonic stem cells. Cell. 2006; 125:301–13. [PubMed: 16630818]

52. Lee TI, Johnstone SE, Young RA. Chromatin immunoprecipitation and microarray-based analysis of protein location. Nat Protoc. 2006; 1:729–48. [PubMed: 17406303]

53. Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. J Mol Biol. 1987; 196:261–82. [PubMed: 3656447]

54. Takai D, Jones PA. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A. 2002; 99:3740–5. [PubMed: 11891299]

55. Ponger L, Mouchiroud D. CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. Bioinformatics. 2002; 18:631–3. [PubMed: 12016061]

56. Hackenberg M, Previti C, Luque-Escamilla PL, Carpena P, Martinez-Aroza J, Oliver JL. CpGcluster: a distance-based algorithm for CpG-island detection. BMC Bioinformatics. 2006; 7:446. [PubMed: 17038168]

57. Wu H, Caffo B, Jaffee HA, Irizarry RA, Feinberg AP. Redefining CpG islands using hidden Markov models. Biostatistics. 2010; 11:499–514. [PubMed: 20212320]

58. Glass JL, Thompson RF, Khulan B, Figueroa ME, Olivier EN, Oakley EJ, Van Zant G, Bouhassira EE, Melnick A, Golden A, Fazzari MJ, Greally JM. CG dinucleotide clustering is a species-specific property of the genome. Nucleic Acids Res. 2007; 35:6798–807. [PubMed: 17932072]

59. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P. Fast Folding and Comparison of RNA Secondary Structures. Monatshefte f Chemie. 1994; 125:167–188.

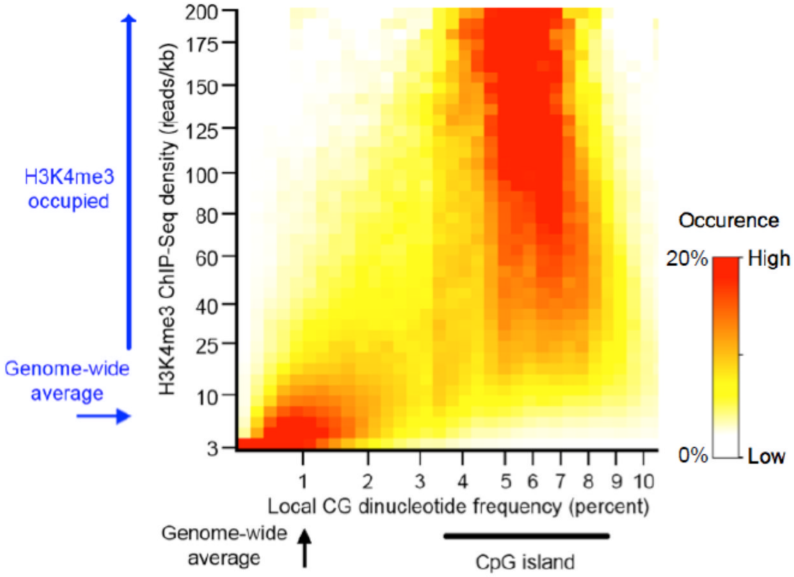

Figure 1. Relationship between CpG islands and H3K4me3 modified nucleosomes in pluripotent stem cells

Co-occurrence plot showing H3K4me3 occupancy is highly coincident with CpG islands in human ES cell line WIBR2. The local CG dinucleotide density and H3K4me3 ChIP-Seq density were tabulated across the genome and are presented in a heatmap where each spot in Figure 1 shows the percent of the genome that has a H3K4me3 occupancy level (y-axis) for a given local CG dinucleotide density (x-axis) (Described in Supplemental Information). The top of the scale is 20% and the bottom of the scale is 0%.

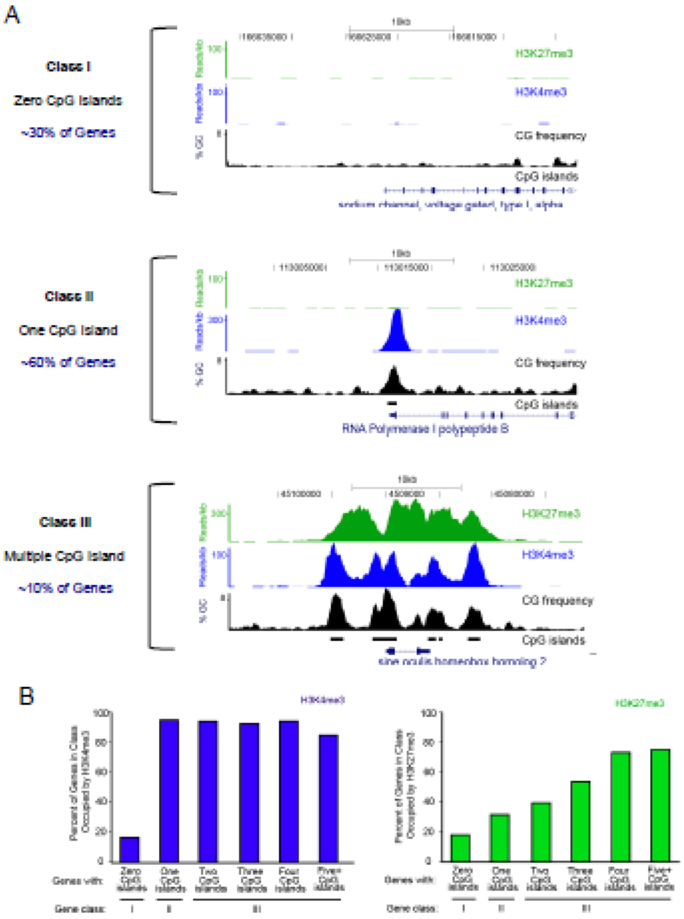

Figure 2. CpG island structure predicts the genomic occupancy of H3K4me3 and H3K27me3 modified nucleosomes in pluripotent stem cells

A. Genes were categorized by the number of CpG islands associated with their transcription start site. The genes encoding the homeobox transcription factor sine oculis homeobox homolog 2 (SIX2), RNA polymerase I polypeptide B (POLR1B), and sodium channel, voltage-gated, type I alpha (SCN1A), all located on chromosome 2 are shown as examples. H3K27me3 and H3K4me3 ChIP-Seq density in the hES line WIBR2, local CG dinucleotide density, and CpG islands are shown.

B. The portion of genes with zero, one, two, three, four, or more than four (five+) CpG islands that are occupied by H3K4me3 (left) and H3K27me3 (right) modified nucleosomes in WIBR2 pluripotent stem cells is shown.

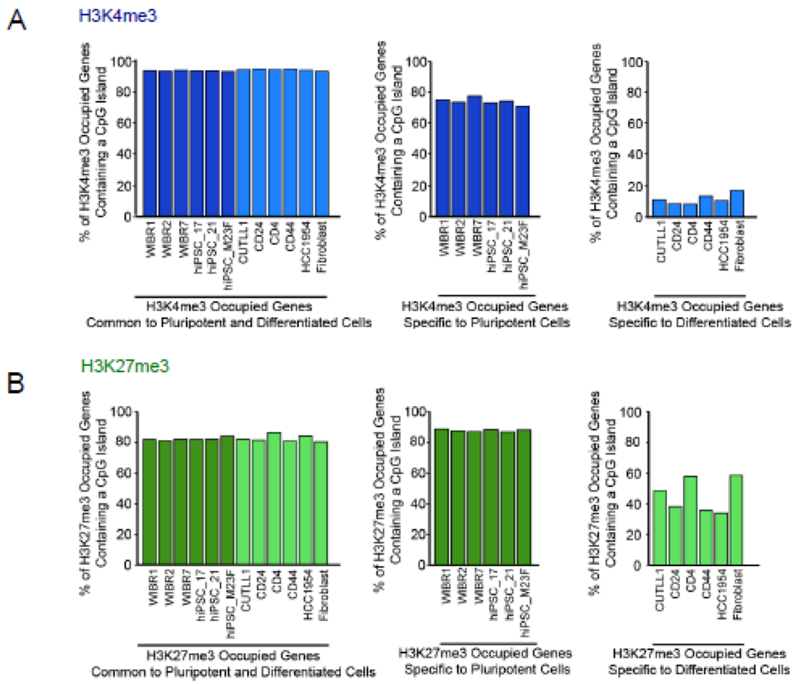

Figure 3. Relationships between H3K4me3 and H3K27me3 modified nucleosomes and CpG islands in differentiated cells

A. (Left) For the set of H3K4me3 occupied genes that are common to pluripotent and differentiated cells, the percentage of genes that are associated with at least one CpG island was calculated. Pluripotent cell lines including ES (WIBR1, WIBR2, WIBR7) and iPS cells (hiPSC_17, hiPSC_21, hiPSC M23F), and differentiated cell lines, Primary CD4+ T cells, IMR90 fetal lung fibroblasts, T-cell lymphoma cells, CD24+ and CD44+ mammary cells, and HCC1954 breast cancer cells, are indicated at bottom. (Middle) For the set of H3K4me3 occupied genes that are specific to pluripotent cells, the percentage of genes that are associated with at least one CpG island was calculated. (Right) For the set of H3K4me3 occupied genes that are specific to differentiated cells, the percentage of genes that are associated with at least one CpG island was calculated.

B. (Left) For the set of H3K27me3 occupied genes that are common to pluripotent and differentiated cells, the percentage of genes that are associated with at least one CpG island was calculated. Cell lines are indicated as in (A). (Middle) For the set of H3K27me3 occupied genes that are specific to pluripotent cells, the percentage of genes that are associated with at least one CpG island was calculated. (Right) For the set of H3K27me3 occupied genes that are specific to differentiated cells, the percentage of genes that are associated with at least one CpG island was calculated. See Supplemental information and Table S3 for genes bound in each cell type.

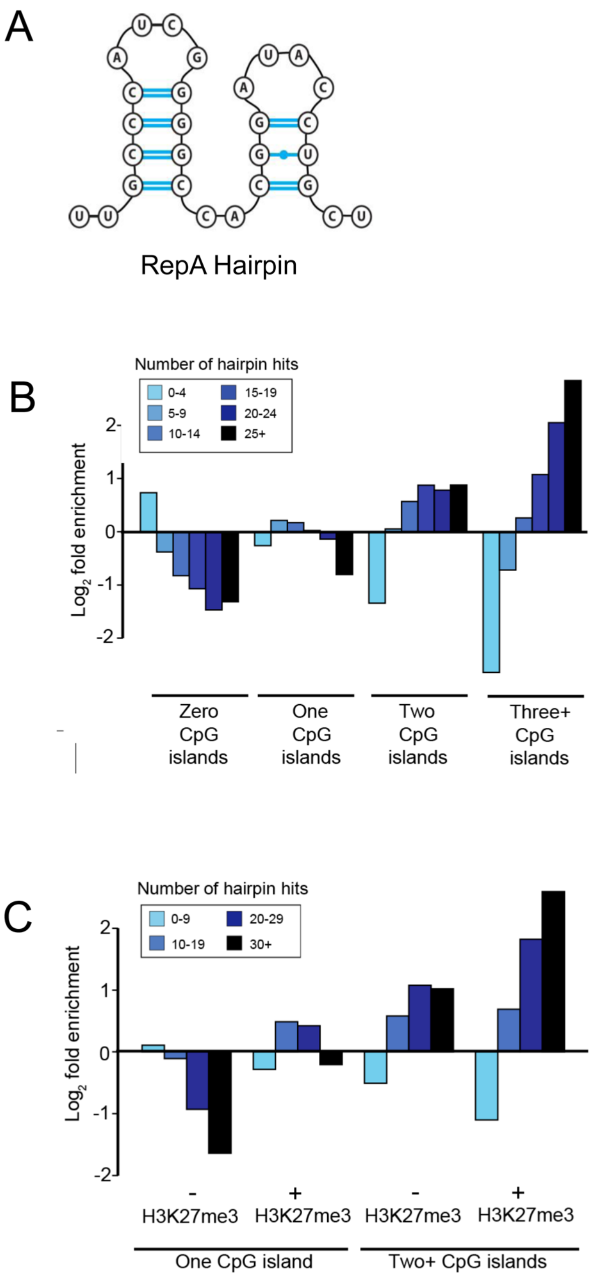

Figure 4. RNA transcripts from H3K27me3 domains have a high probability of GC hairpin structures

A. The dual RepA stem-loop RNA hairpin that is known to bind PRC2 is shown [21; 33].

B. The number of RNA hairpin hits around the TSS (+/−5kb) for genes within zero, one, two, and three or more CpG islands shows a bias for more hits in genes with multiple CpG islands (see Supplemental Information for the definition of an RNA hairpin hit). The relative fold enrichment versus all genes is shown for each CpG class.

C. The number of RNA hairpin hits for genes with one CpG island with or without H3K27me3 occupancy and for genes with more than one CpG island with or without H3K27me3 occupancy is shown. Data are represented as in (b).