Cell

Resource

Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells

Wei Xie,[1] Matthew D. Schultz,[2] Ryan Lister,[2,14] Zhonggang Hou,[3] Nisha Rajagopal,[1] Pradipta Ray,[11] John W. Whitaker,[4] Shulan Tian,[3] R. David Hawkins,[1,15] Danny Leung,[1] Hongbo Yang,[7] Tao Wang,[4] Ah Young Lee,[1] Scott A. Swanson,[3] Jiuchun Zhang,[3,8] Yun Zhu,[4] Audrey Kim,[1] Joseph R. Nery,[2] Mark A. Urich,[2] Samantha Kuan,[1] Chia-an Yen,[1] Sarit Klugman,[1] Pengzhi Yu,[3] Kran Suknuntha,[12] Nicholas E. Propson,[3] Huaming Chen,[2] Lee E. Edsall,[1] Ulrich Wagner,[1] Yan Li,[1] Zhen Ye,[1] Ashwinikumar Kulkarni,[11] Zhenyu Xuan,[11] Wen-Yu Chung,[11,16] Neil C. Chi,[7] Jessica E. Antosiewicz-Bourget,[3] Igor Slukvin,[8,9,12] Ron Stewart,[3] Michael Q. Zhang,[11,13] Wei Wang,[4,6] James A. Thomson,[3,9,10,*] Joseph R. Ecker,[2,*] and Bing Ren[1,5,*]

[1]Ludwig Institute for Cancer Research, La Jolla, CA 92093, USA
[2]Genomic Analysis Laboratory, Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA
[3]Morgridge Institute for Research, Madison, WI 53707, USA
[4]Department of Chemistry and Biochemistry
[5]Department of Cellular and Molecular Medicine, Institute of Genomic Medicine and Moores Cancer Center
[6]Department of Cellular and Molecular Medicine
[7]Department of Medicine, Division of Cardiology
University of California, San Diego, La Jolla, CA 92093, USA
[8]Wisconsin National Primate Research Center
[9]Department of Cell and Regenerative Biology
University of Wisconsin-Madison, Madison, WI 53715, USA
[10]Department of Molecular, Cellular, and Developmental Biology, University of California, Santa Barbara, Santa Barbara, CA 93106, USA
[11]Department of Molecular and Cell Biology, Center for Systems Biology, The University of Texas at Dallas, Richardson, TX 75080, USA
[12]Department of Pathology and Laboratory Medicine, University of Wisconsin Medical School, Madison, WI 53792, USA
[13]Bioinformatics Division, Center for Synthetic and Systems Biology, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China
[14]Present address: Plant Energy Biology (ARC CoE) and Computational Systems Biology (WA CoE), School of Chemistry and Biochemistry, The University of Western Australia, Perth, WA 6009, Australia
[15]Present address: Division of Medical Genetics, Department of Medicine, Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA
[16]Present address: Department of Computer Science and Information Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung 807, Taiwan
*Correspondence: jthomson@morgridgeinstitute.org (J.A.T.), ecker@salk.edu (J.R.E.), biren@ucsd.edu (B.R.)
http://dx.doi.org/10.1016/j.cell.2013.04.022

SUMMARY

Epigenetic mechanisms have been proposed to play crucial roles in mammalian development, but their precise functions are only partially understood. To investigate epigenetic regulation of embryonic development, we differentiated human embryonic stem cells into mesendoderm, neural progenitor cells, trophoblast-like cells, and mesenchymal stem cells and systematically characterized DNA methylation, chromatin modifications, and the transcriptome in each lineage. We found that promoters that are active in early developmental stages tend to be CG rich and mainly engage H3K27me3 upon silencing in nonexpressing lineages. By contrast, promoters for genes expressed preferentially at later stages are often CG poor and primarily employ DNA methylation upon repression. Interestingly, the early developmental regulatory genes are often located in large genomic domains that are generally devoid of DNA methylation in most lineages, which we termed DNA methylation valleys (DMVs). Our results suggest that distinct epigenetic mechanisms regulate early and late stages of ES cell differentiation.

INTRODUCTION

Embryonic development is a complex process that remains to be understood despite knowledge of the complete genome sequences of many species and rapid advances in genomic technologies. A fundamental question is how the unique gene expression pattern in each cell type is established and maintained during embryogenesis. It is well accepted that the gene expression program encoded in the genome is executed by transcription factors that bind to *cis*-regulatory sequences and modulate gene expression in response to environmental cues

(Young, 2011). Growing evidence now shows that maintenance of such cellular memory depends on epigenetic marks such as DNA methylation and chromatin modifications (Bird, 2002; Kouzarides, 2007).

DNA methylation at promoters has been shown to silence gene expression and thus has been proposed to be necessary for lineage-specific expression of developmental regulatory genes, genomic imprinting, and X chromosome inactivation (Bird, 2002). Indeed, the DNA methyltransferases DNMT1 or DNMT3a/3b double-knockout mice exhibit severe defects in embryogenesis and die before midgestation, supporting an essential role for DNA methylation in embryonic development (Li et al., 1992; Okano et al., 1999). On the other hand, mouse embryonic stem cells (mESCs) lacking all three DNMTs can survive and self-renew and can even begin to differentiate to some germ layers (Jackson et al., 2004; Tsumura et al., 2006), raising the possibility that DNA methylation is dispensable for at least initial lineage specification in early embryos. Thus, the role of DNA methylation in animal development needs to be more precisely defined. Like DNA methylation, chromatin modifications have also been shown to play a key role in animal development. Enzymes responsible for methylation of histone H3 at lysine 4, 9, and 27, in particular, are essential for embryogenesis (Kouzarides, 2007; Vastenhouw and Schier, 2012). Additionally, depletion of the histone acetyltransferase p300 or CBP also leads to early embryonic lethality (Yao et al., 1998). Although both DNA methylation and chromatin modifications are critical for mammalian development, the exact role of each epigenetic mark in the maintenance of lineage-specific gene expression patterns remains to be defined.

In humans, studying the epigenetic mechanisms regulating early embryonic development often requires access to embryonic cell types that are currently difficult or impractical to obtain. Human embryonic stem cells (hESCs) (Thomson et al., 1998) can be differentiated into a variety of precursor cell types, providing an in vitro model system for studying early human developmental decisions. We have established protocols for differentiation of hESCs to various cell states, including trophoblast-like cells (TBL) (Xu et al., 2002), mesendoderm (ME) (Yu et al., 2011), neural progenitor cells (NPCs) (Chambers et al., 2009; Chen et al., 2011), and mesenchymal stem cells (MSCs) (Vodyanik et al., 2010). The first three states represent developmental events that mirror critical developmental decisions in the embryo (the decision to become embryonic or extraembryonic, the decision to become mesendoderm or ectoderm, and the decision to become surface ectoderm or neuroectoderm, respectively). MSCs are fibroblastoid cells that are capable of expansion and multilineage differentiation to bone, cartilage, adipose, muscle, and connective tissues (Vodyanik et al., 2010). The specific hESC derivatives chosen thus reflect key lineages in the human embryo and also represent those lineages that currently can be produced in sufficient quantity and purity for epigenomic studies. These lineages will complement other cells from more mature sources, many of which have had their epigenomes well characterized (Hawkins et al., 2010; Lister et al., 2009; Zhu et al., 2013). Importantly, epigenomic analysis of these cell types allows for investigation of chromatin and transcriptional changes that drive the initial developmental fate decisions.

Here, we used high-throughput approaches to examine the differentiation of hESCs into four cell types by generating in-depth maps of transcriptomes, a large panel of histone modifications, and base-resolution maps of DNA methylation for each cell type. Our study provided a full view of the dynamic epigenomic changes accompanying cellular differentiation and lineage specification. As outlined below, an integrative analysis of these data sets provided us with substantial insights into the role of DNA methylation and chromatin modifications in animal development.

RESULTS

Generation of Comprehensive Epigenome Reference Maps for hESCs and Four hESC-Derived Lineages

We differentiated the hESC line H1 to ME, TBL, NPCs, and MSCs (Figure 1A) (Extended Experimental Procedures). ME, TBL, and NPC differentiation occurred quickly (2 days, 5 days, and 7 days, respectively) compared to that of MSC (19–22 days). The expression of various marker genes in these cells was confirmed using immunofluorescence and fluorescence-activated cell sorting (FACS), and the purity of each cell population ranged from 93% to 99% (Figures S1A–S1C available online). ME, NPCs, and MSCs possess further differentiation potentials as shown in Figures S1D and S1E (for ME and NPCs) and our previous study (for MSCs) (Vodyanik et al., 2010). On the other hand, the nature of TBL is still currently under debate (Bernardo et al., 2011; Xu et al., 2002). As a control for terminally differentiated cells, we also cultured and analyzed IMR90, a primary human fetal lung fibroblast cell line. For each cell type, we mapped DNA methylation at base resolution using MethylC-seq (Lister et al., 2009) (20–35× total genome coverage or 10–17.5× coverage per strand). We also mapped the genomic locations of 13–24 chromatin modifications by chromatin immunoprecipitation sequencing (ChIP-seq). Additionally, we performed paired-end (100 bp × 2) RNA-seq experiments, generating more than 150 million uniquely mapped reads for every cell type (Figures 1A and 1B). At least two biological replicates were carried out for each analysis, and the data were publicly released as part of the NIH Roadmap Epigenome Project (http://www.epigenomebrowser.org/). Selected data are also available at http://epigenome.ucsd.edu/differentiation.

Identification of Differentially Expressed Genes in hESC-Derived Cells

We first asked how the genome is differentially transcribed when hESCs are differentiated into each cell type. To do so, we examined the expression of 19,056 RefSeq coding genes (33,797 isoforms), among which 76.6% (14,595) were expressed in at least one cell type (Figure S2A). Using an entropy-based method (Barrera et al., 2008; Schug et al., 2005) (Figure S2B), we identified 2,408 genes that showed cell-type-specific expression (Figures 2A and S2A). For convenience, we use ''lineage-restricted genes'' to reflect both H1-specific and differentiated cell-specific genes. As expected, known lineage markers were highly expressed in the corresponding cell types (Figure 2A). It is worth noting that, in line with a previous report (Yu et al., 2011), the ME cells also express high levels of the hESC regulators

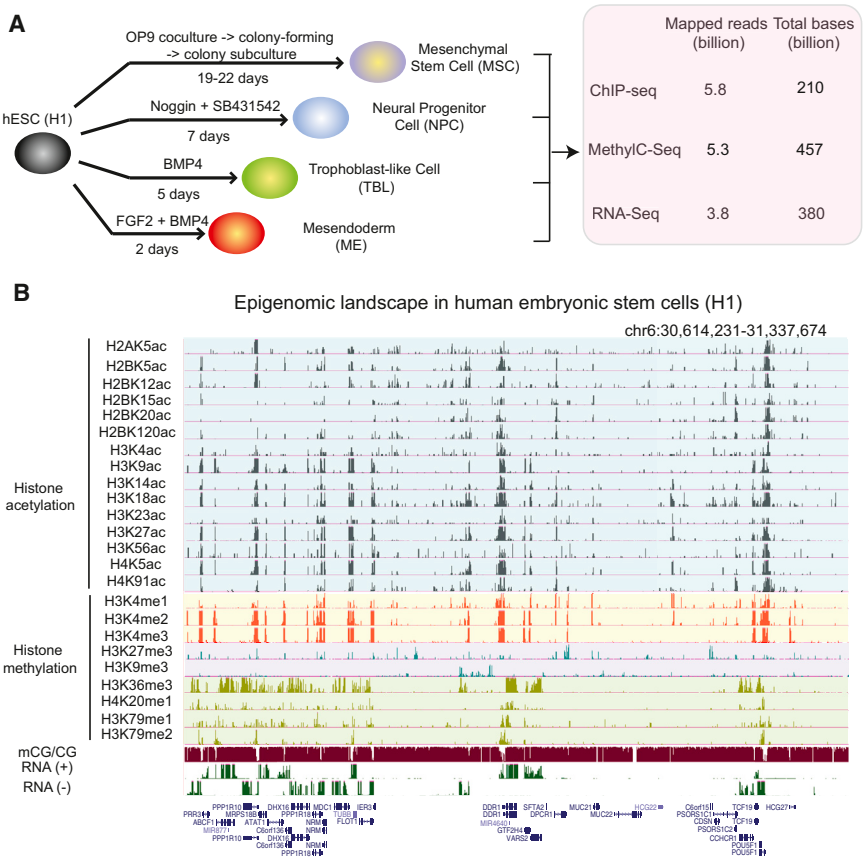


Figure 1. Generation of Comprehensive Epigenome Reference Maps for hESCs and Four hESC-Derived Lineages

(A) Schematic of hESC differentiation procedures and a summary of the epigenomic data sets produced in this study.

(B) A snapshot of the UCSC genome browser shows the DNA methylation level (mCG/CG), RNA-seq reads (+, Watson strand; −, Crick strand), and ChIP-seq reads (RPKM) of 24 chromatin marks in H1.

See also Figure S1.

NANOG, POU5F1, and a reduced but significant level of SOX2. We then investigated a cohort of long noncoding RNA (lncRNA) genes and detected significant levels of transcripts for 2,175 known and 281 unannotated lncRNA genes in at least one cell type (Figures 2A and S2A). Using the same entropy-based approach, we found 930 lncRNA genes defined as lineage restricted (Figure S2C), which constitute 37.9% of total expressed lncRNA genes. By contrast, only 16.5% of expressed coding genes are characterized as lineage restricted (Figure S2D). The above analysis defined a large number of coding and noncoding genes that are differentially expressed in H1 and its derived cells. The lists of all lineage-restricted genes are included in Table S1.

Intriguingly, the promoters of several lncRNA genes highly expressed in H1 overlap with the long terminal repeat (LTR)-containing retrotransposons (Figure 2B). This appears to be a general phenomenon as we observed that significant percentages of transcription start sites (TSSs) of lncRNA genes directly fall into LTRs (Figure 2C). The percentages are notably higher for H1- and ME-enriched lncRNA genes (30% and 31%, respectively), which are in contrast to those of coding genes (<2%). By quantifying the transcription levels of all major classes of mappable repetitive elements, we found that the ERV1 (class I endogenous retrovirus) elements are preferentially expressed in H1 and ME, but not in other cell types (Figure 2D, top). Strikingly, such lineage-specific expression occurs almost exclusively at the ERV1 subfamily HERV-H and its flanking LTR elements LTR7 (Figure 2D, bottom). Together, HERV-H and LTR7 account for more than 43% of LTRs that are present at H1- and ME-specific lncRNA gene promoters. A gene ontology analysis of coding genes near H1-specific HERV-H/LTR7 sites revealed an enrichment of POU5F1-targeted genes (p value = 4×10^{-15}), which is consistent with a previous study showing that NANOG and POU5F1 preferentially bind to repetitive elements (Kunarso et al., 2010). We did not find significant enrichment of LTR subclasses for other lineage-restricted lncRNA genes. Repetitive elements are known to be regulated by DNA methylation and H3K9me3 in ESCs (Leung and Lorincz, 2012). We do not find significant enrichment of H3K9me3 around most HERV-H elements (data not shown). By contrast, a subset of the H1-specific HERV-H elements (n = 70) show hypomethylation in H1 and ME but gain DNA methylation in other H1-derived cells (Figures 2B and 2E). Notably, the overall low level of DNA methylation in IMR90 reflects its globally hypomethylated genome, likely due to the presence of partially methylated domains (PMDs) (Figures S2E and S2F) (Lister et al., 2009). Additionally, by examining published methylomes (Lister et al., 2011), we found that DNA methylation at these regions was depleted upon reprogramming of IMR90 or foreskin fibroblasts to iPSCs and was then reestablished when the fibroblast-derived iPSCs were differentiated to trophoblast-like lineage (Figure 2B). Together, these data suggest that many noncoding RNA genes may be transcriptionally regulated by endogenous retroviral sequences. Of particular interest, the expression of HERV-H/LTR7 is closely correlated with the state of pluripotency and may be regulated by DNA methylation.

Dynamic DNA Methylation and Chromatin Modifications at Promoters of Lineage-Restricted Transcripts

Previous studies have shown that the promoters for somatic-tissue-specific genes are often CG poor and lack CpG islands (CGIs), in contrast to those for housekeeping genes, which are CG rich and predominantly contain CGIs (Barrera et al.,

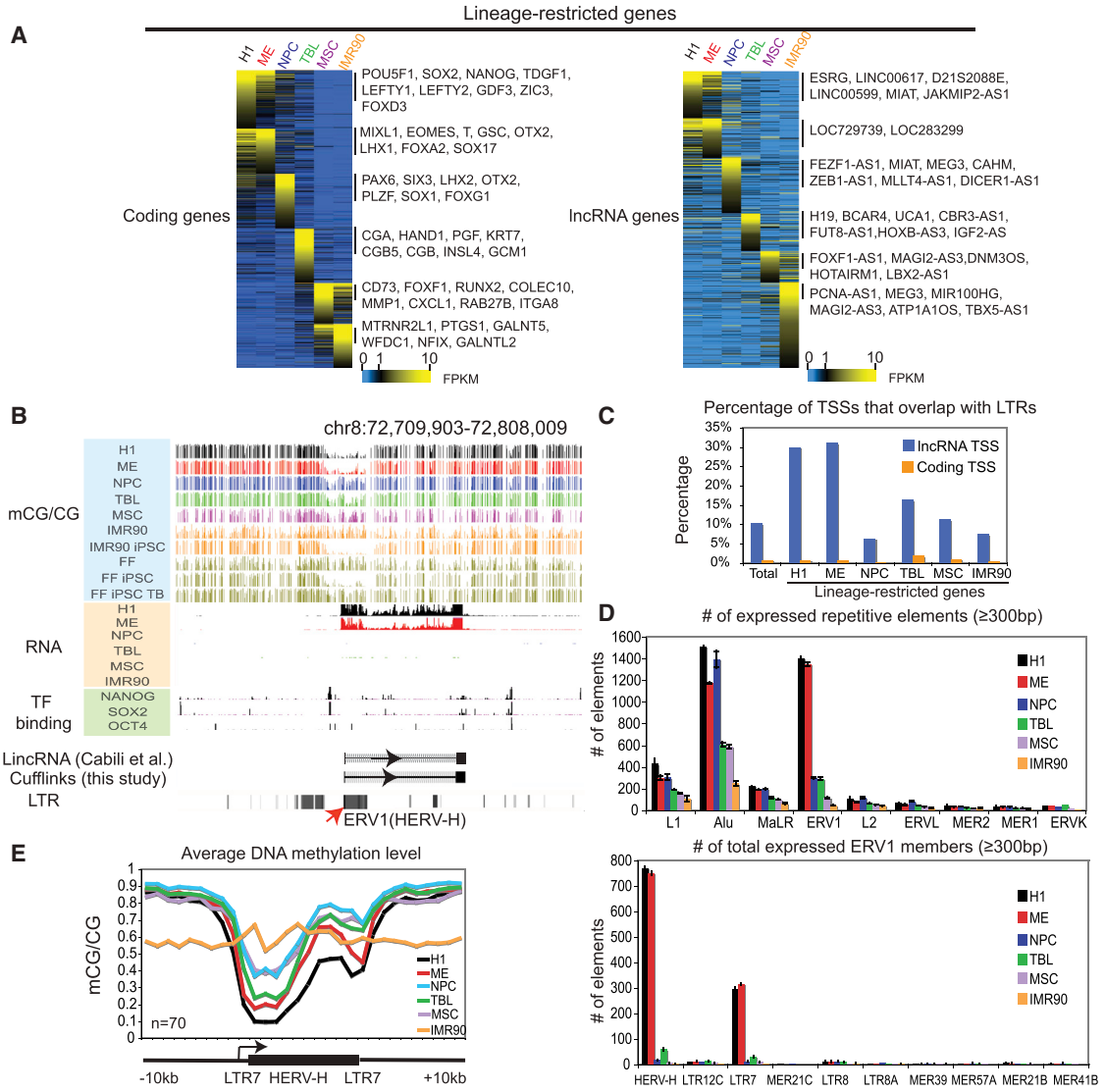


Figure 2. Identification of Lineage-Restricted Transcripts in H1 and H1-Derived Cells

(A) Heatmaps showing the expression levels of lineage-restricted coding genes (left) and lncRNA genes (right). Genes are organized by the lineage in which their expression is enriched. Note that certain genes (such as *SOX2*) can be expressed in more than one cell type.

(B) The levels of DNA methylation and RNA, as well as the binding of NANOG, SOX2, and POU5F1, are shown around an annotated lincRNA gene with the promoter overlapping a HERV-H element.

(C) The percentages of TSSs that overlap with LTRs are shown for coding genes (yellow) and lncRNA genes (blue) for all genes (total) or lineage-restricted genes.

(D) The numbers of expressed (FPKM $\geq$ 1), mappable repetitive elements are shown in each cell type for various repeat classes (top) or subclasses of ERV1 (bottom). Data are represented as mean ± SD based on two replicates of RNA-seq.

(E) The average DNA methylation level in each cell type is shown for a subset of H1-specific HERV-H elements.

See also Figure S2.

2008; Schug et al., 2005). Therefore, we asked whether early lineage-restricted promoters also demonstrate similar features as tissue-specific promoters. We first identified promoters for each lineage-restricted gene and excluded those with ambiguous active promoters (Extended Experimental Procedures). Next, we divided the promoters into three groups based on CG density (high, medium, and low) (Figure S3A). Surprisingly, genes preferentially expressed in early embryonic lineages H1, ME, and NPC tend to be CG rich and contain CGIs (Figure 3A). The percentages of CGI-containing promoters decreased for genes enriched in MSCs and IMR90, which are at relatively late development stages. By contrast, a much lower percentage of promoters (23%) contain CGIs for somatic-tissue-specific genes identified from 18 human tissues (Zhu et al., 2008) (Figure 3A). We further verified this using an independent set of somatic-tissue-specific genes (35%) (Chang et al., 2011). These data

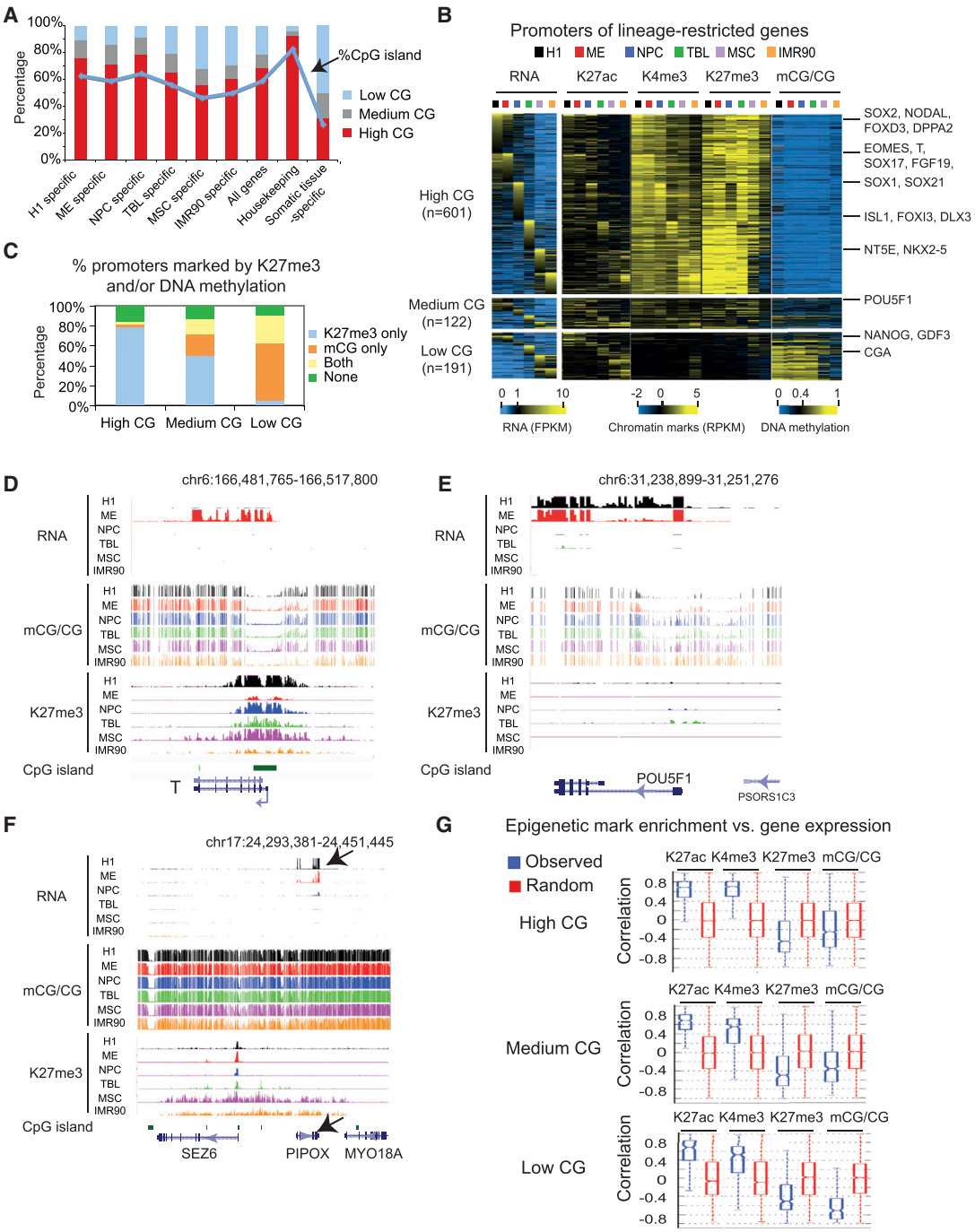


Figure 3. Epigenetic Regulation of Promoters for Lineage-Restricted Genes

(A) Bar graphs showing the percentages of promoters in the high, medium, and low CG classes for genes that are enriched in each cell type, all RefSeq genes, housekeeping genes, and somatic-tissue-specific genes identified in Zhu et al. (2008). The percentages of promoters that contain CGIs are also shown (blue line).

(B) Heatmaps showing the average levels of RNA, H3K27ac, H3K4me3, H3K27me3, and DNA methylation for promoters of lineage-restricted genes. Histone modifications, TSS ± 2 kb; DNA methylation, TSS ± 200 bp; promoter CG density, TSS ± 500 bp.

(C) Bar graphs showing the percentages of promoters that are marked by DNA methylation or K27me3 in at least one cell type.

(D–F) The levels of RNA, DNA methylation, and K27me3 are shown for the locus containing *T* (D), *POU5F1* (E), or *PIPOX* (F). *PIPOX* (black arrow) is a low CG-promoter-containing gene located in a K27me3 domain in MSCs and IMR90, where it is also repressed.

(G) The distribution of Pearson correlation coefficients between gene expression level and the levels of various histone modifications or DNA methylation at promoters.

See also Figure S3.

suggest that the promoters used for lineage specification in early stages of cell differentiation have distinct sequence features compared to those in more mature cell types.

DNA methylation machinery has been shown to be a mechanism of gene silencing during cell differentiation (Bird, 2002). In addition, the Polycomb protein complex, which deposits H3K27me3 at target genes, can also repress developmental genes (Boyer et al., 2006; Lee et al., 2006). We set to determine which promoters are subject to regulation by DNA methylation, H3K27me3, or both. A detailed analysis showed that promoters with high CG density tend to be enriched for H3K27me3, whereas those with low CG density are preferentially marked by DNA methylation (Figures 3B and 3C). This is exemplified by the promoters of the ME marker *T* (high CG, with a CGI) and the hESC marker *POU5F1* (medium CG, no CGIs) (Figures 3D and 3E). Notably, whereas both H3K27me3 and DNA methylation are largely anticorrelated with gene expression, high CG promoters are often marked by reduced but significant enrichment of H3K27me3 even when they are active (Figures 3B and 3D). It has been shown that the PRC2 complex can be directly recruited by CG-rich sequences (Mendenhall et al., 2010). Consistent with this model, our data indicate that the sequence of a promoter could contribute to the epigenetic mechanisms that affect its regulation.

Notably, the majority of developmental regulatory genes, including *SOX2*, *NODAL*, *EOMES*, *T*, *SOX17*, and *SOX1*, belong to the high CG group and are marked by H3K27me3 (Figure 3B). DNA methylation, on the other hand, marks a relatively small number of lineage-restricted genes, including *NANOG* and *POU5F1*. A gene ontology analysis also showed that lineage-restricted genes with high CG promoters are enriched for developmental genes, embryonic morphogenesis, and pattern specification, whereas those with low CG promoters contain genes that function in plasma membrane, disulfite bond, and protein kinase cascade. As controls, somatic-tissue-specific promoters are largely CG poor, often showing high level of DNA methylation; housekeeping gene promoters are predominantly CG rich, showing neither DNA methylation nor H3K27me3 in these cells (Figure S3B). Interestingly, some CG-poor promoters are also marked by low levels of H3K27me3. These promoters are largely observed in the expanded H3K27me3 domains (Figures 3B and 3F, black arrow), a broad pattern of enrichment for H3K27me3 (Hawkins et al., 2010; Zhu et al., 2013) that frequently occurs in MSCs and IMR90, but less so in H1 and other H1-derived cells (Figure S3C and data not shown). These observations suggest that the expansion of H3K27me3 may be a mechanism to lock low CG promoters in a repressed state in later development stages. Consistently, H3K27me3 shows similar negative correlations with gene expression in all three classes (Figure 3G). By contrast, DNA methylation shows the strongest negative correlation with gene expression for low CG genes (see Figure S3D for the analysis of additional histone modifications). Together, our data suggest that, although H3K27me3 may play a widespread role in regulating key factors of cellular differentiation, DNA methylation is involved in modulation of many somatic-tissue-specific genes and a limited number of—albeit critical—developmental regulators.

Dynamic DNA Methylation and Chromatin Modifications at Enhancers Reflect Lineage-Restricted Gene Expression

Enhancers are distal regulatory elements that mediate tissue and developmental-stage-specific gene expression (Ong and Corces, 2011). To examine the potential role of DNA methylation and chromatin modifications at enhancers, we first identified a total of 103,982 putative enhancer sites in the six cell types (Table S2) by using an enhancer prediction method described recently (Rajagopal et al., 2013) (Extended Experimental Procedures). By examining the level of H3K27ac, a marker for active enhancers (Creyghton et al., 2010; Rada-Iglesias et al., 2011), we classified 32,423 enhancers as lineage restricted using the entropy-based analysis (Figure S4A, Table S2, and Extended Experimental Procedures). We validated these enhancers using several approaches by showing that they extensively overlap with the binding sites of transcriptional regulators or DNase I hypersensitive sites (J.A. Stamatoyannopoulos, personal communication) (Figure S4B); they show evolutional conservation in sequences (Figure S4C); they are enriched for motifs of transcription factors known to function in each lineage (Figure S4D and Table S3); and their neighboring genes demonstrate functional enrichment that is related to their lineage identities (Figure S4E). Finally, we constructed eight GFP reporters containing various lineage-specific enhancers and injected them in zebrafish embryos. A high percentage of these enhancers (50%) demonstrated activity in vivo in specific lineages regardless of their positions relative to the reporter gene (Figure S4F). Together, these data suggest that we have identified a set of lineage-restricted enhancers of high quality in hESCs and hESC-derived cells.

We subsequently examined the dynamic epigenetic modifications at lineage-restricted enhancers. As these modifications at intragenic enhancers can be confounded by the activity of their hosting genes, we focused on intergenic lineage-restricted enhancers (n = 6,819) for this analysis (enhancers present in PMDs in IMR90 were also excluded). Most enhancers are CG poor (94%) and appear to be depleted of H3K27me3 (Figure 4A). However, weak enrichment of H3K27me3 is observed at a subset of enhancers in MSCs and IMR90. These enhancers are largely active in H1, ME, NPCs, and TBL, but not in MSCs and IMR90, as indicated by the levels of H3K27ac. A closer examination revealed that these enhancers are preferentially present in the H3K27me3 domains specific to MSCs and IMR90 (see Figure 4B for an example). In IMR90 and MSCs, repressed enhancers are marked by a higher level of H3K27me3 compared to active enhancers (Figure 4C). By contrast, this is less evident for enhancers in H1 and other H1-derived cells. These results are consistent with the mode that the H3K27me3 domains that arise in differentiated cells may function to repress enhancers that are active in other lineages (Hawkins et al., 2010; Zhu et al., 2013).

Our data also showed that the presence of DNA methylation negatively correlates with the activity of enhancers (Figure 4C). Interestingly, although some H1-specific enhancers acquire DNA methylation in MSCs and IMR90, this is less evident in ME, NPCs, and TBL (Figures 4A and 4D). These data are in line with a recent study showing that inactive regulatory elements tend to progressively gain DNA methylation over time during

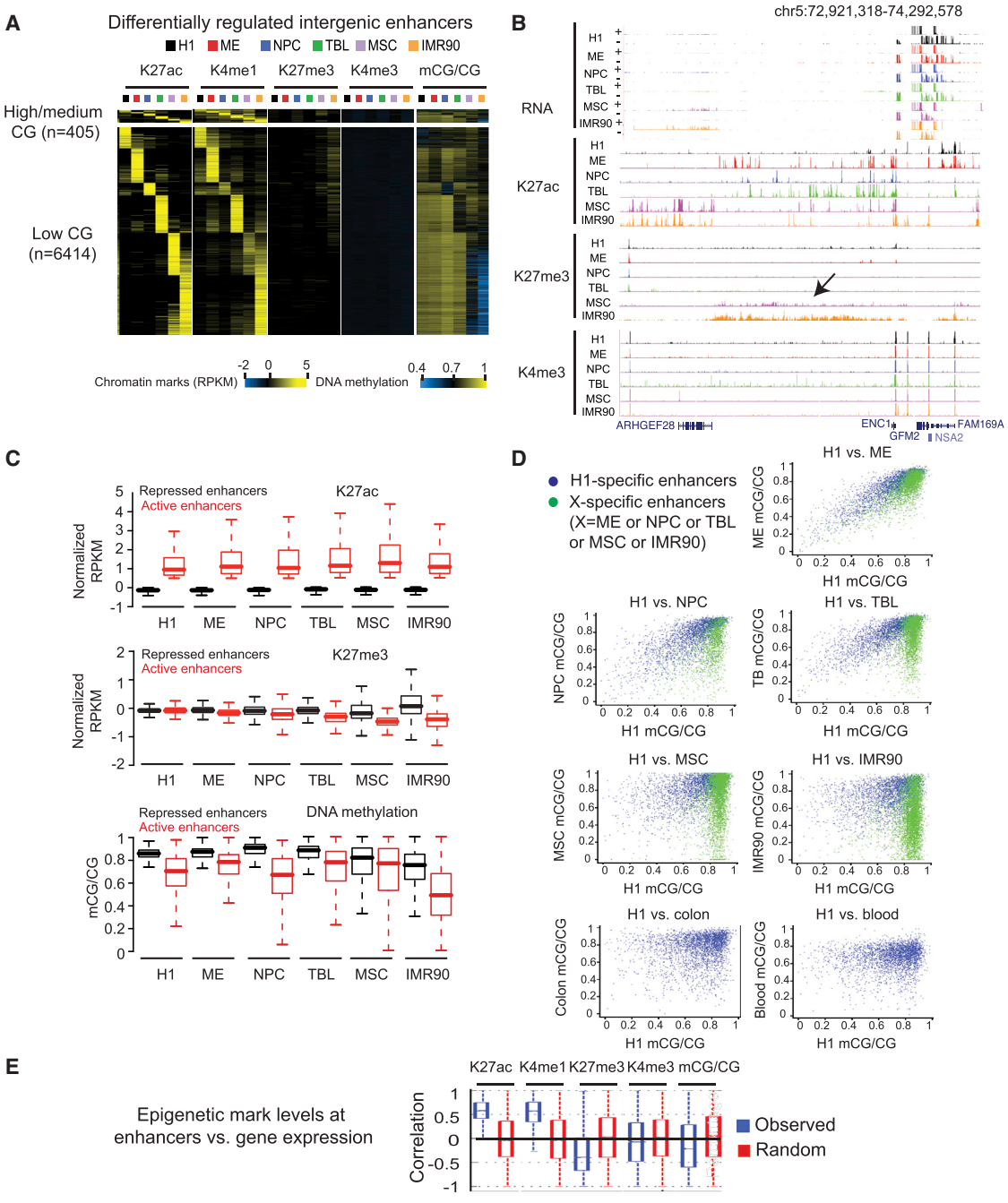


Figure 4. Epigenetic Regulation of Lineage-Restricted Enhancers

(A) Heatmaps showing the average levels of H3K27ac, H3K4me1, H3K4me3, H3K27me3, and DNA methylation around the centers of lineage-restricted enhancers. Histone modifications, enhancer center ± 2 kb; DNA methylation, enhancer center ± 500 bp; CG density, enhancer center ± 500 bp.

(B) The epigenetic landscape at an intergenic locus showing a low level of H3K27me3 and absence of H3K27ac in MSC and IMR90.

(C) Box plots showing the levels of H3K27ac (top), H3K27me3 (middle), and DNA methylation (bottom) at active and repressed enhancers in each cell type.

(D) Scatterplots showing the levels of DNA methylation in each cell type at H1-specific enhancers (blue) and differentiated cell-specific enhancers (green). In the last two panels, colon- and blood-specific enhancer information (green dots) is not available in Berman et al. (2012) and Li et al. (2010).

(E) Box plots showing the distribution of Pearson correlation coefficients between the levels of various histone modifications or DNA methylation at enhancers and the expression level of their potential target genes.

See also Figure S4.

cell differentiation (Bock et al., 2012). By contrast, differentiated cell-specific enhancers appear highly methylated in lineages where they are inactive. We do not observe significant differences between H1-specific and differentiated cell-specific enhancers in their proximity to the nearest TSSs (data not shown). Notably, some H1-specific enhancers remain hypomethylated even in MSCs, IMR90, and two human tissues: peripheral blood mononuclear cells (Li et al., 2010) and the colon (mucosa) (Berman et al., 2012) (Figure 4D). The functions of these hypomethylated enhancers remain to be explored. Together, these data indicate that H3K27me3 is preferentially enriched at a subset of enhancers in a later stage of cellular differentiation. By contrast, DNA methylation is widely present at enhancers of all stages and negatively correlates with their activity.

We further examined whether the presence of DNA methylation or H3K27me3 may correlate with the expression of genes that are potentially regulated by enhancers. To do so, we identified candidate target genes of lineage-restricted enhancers using correlative analyses (Ernst et al., 2011) (Table S4 and Extended Experimental Procedures). At enhancers, histone acetylation is generally positively correlated with the expression of enhancer-targeted genes (Figures 4E and S4G). H3K27me3 and DNA methylation, by contrast, show an inverse relationship with gene expression of their potential target genes. The analysis results for expanded histone marks are included in Figure S4G.

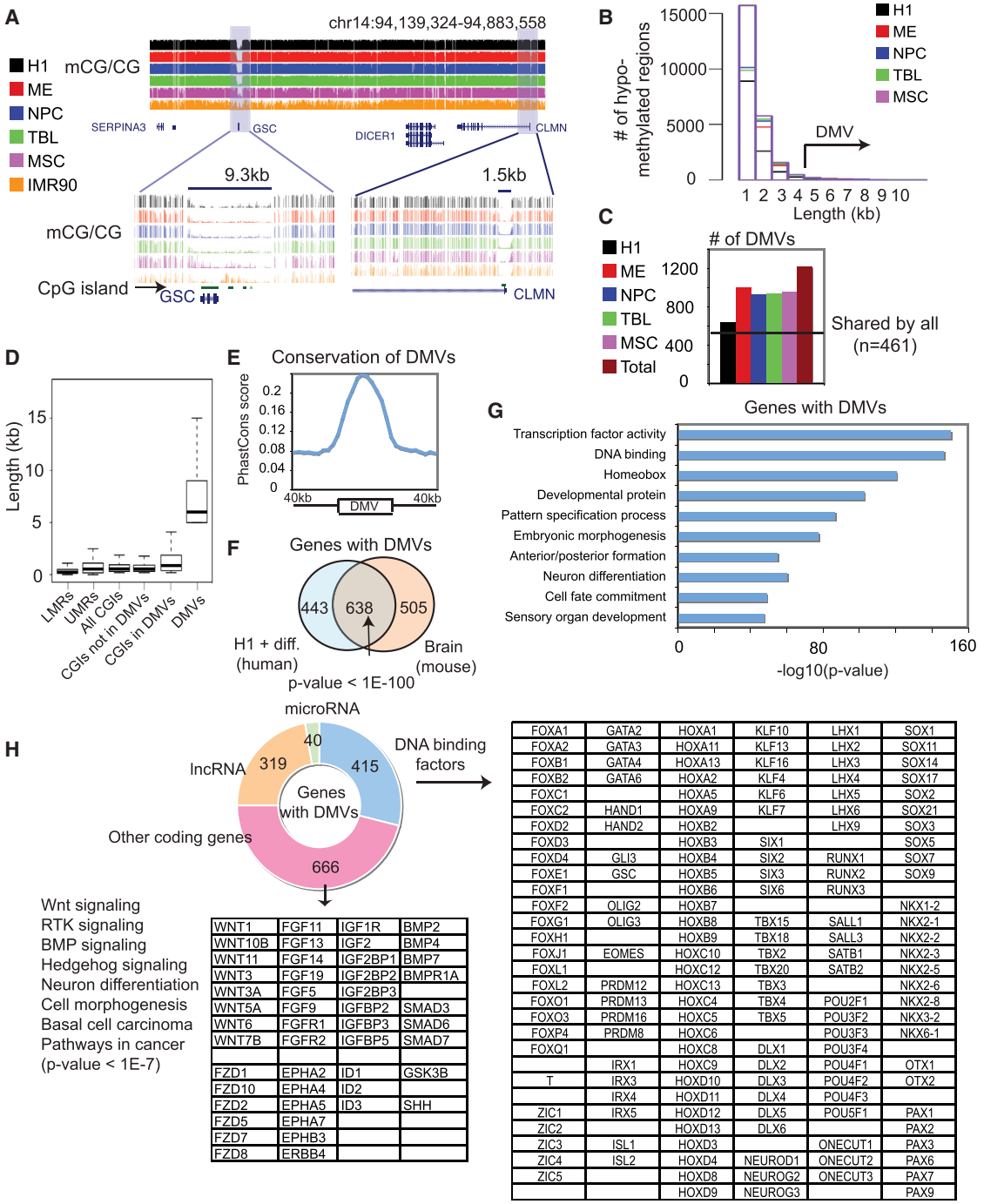
Identification of DNA Methylation Valleys

Previously, low methylation regions (LMRs) and unmethylated regions (UMRs) have been suggested to function as *cis* elements (Stadler et al., 2011). Applying the same approach as Stadler et al. (2011), we defined 5,323 to 31,158 UMRs and 32,744 to 74,541 LMRs in H1 and its derived lineages (Table S5). Indeed, more than 85% of UMRs and 42% of LMRs are present in either enhancers or promoters. Surprisingly, although LMR and UMRs are generally short (median lengths 252 bp and 532 bp, respectively), a number of loci show a much wider depletion of DNA methylation. Interestingly, they often appear near genes for transcription factors and developmental regulators. For example, a 9.3 kb hypomethylated region is observed at *GSC*, a transcription factor specifically expressed in ME (Figure 5A). This unmethylated region covers the entire gene body and regions beyond, which is in contrast to a typical UMR (*CLMN*, Figure 5A). We sought to investigate whether such broad DNA methylation depletion around developmental genes is a general phenomenon. By examining all continuous hypomethylated regions in H1 and the H1-derived cells (Figures S5A and 5B), we identified those that are at least 5 kb long, which constitute less than 3.2% of all hypomethylated regions in any cell type. We named these regions DNA methylation valleys (DMVs). IMR90 was excluded from this part of our study due to the presence of PMDs in these cells (Lister et al., 2009) (Figure S2F), which would confound the analyses. Genome wide, we identified 639, 1,004, 933, 944, and 962 DMVs in H1, ME, NPC, TBL, and MSC, respectively, among which 461 are shared by all cell types (Figure 5C; see Table S6 for the full lists). Together, these regions occupy 1,220 distinct genomic loci. Strikingly, nearly every DMV (99.7%) contains at least one known (89.9%) or putative promoter (9.8%, as indicated by the presence of H3K4me3). The majority of DMVs (93.8%, n = 1,144) contain at least one CGI. Interestingly, whereas 51.8% DMVs contain one or less CGI, 23.7% (289) DMVs contain at least three CGIs (Figure S5B). These DMVs range in size from 5 kb to 68 kb and are much larger than the CGIs in these regions (Figure 5D). About 67% of DMVs contain at least half non-CGI sequences even when we used a much larger CGI list (n = 63,956) (Irizarry et al., 2009) instead of the UCSC CGI list (n = 27,639). We then asked whether DMVs are conserved across species. Indeed, DMVs show high level of sequence conservation (Figure 5E). Additionally, we searched for DMVs in mice using a brain methylome that we recently obtained (Xie et al., 2012). Strikingly, a large number of genes with DMVs in humans (638, or 59%, p value $< 1 \times 10^{-100}$) are also present in DMVs in mice (Figure 5F). Finally, many DMVs (>40%) found in H1 and its derivatives were also observed to be such in adult tissues (Berman et al., 2012; Li et al., 2010) (Figure S5C), suggesting that DMVs are not artifacts of cell culture. The different numbers of DMVs in various cell types may be in part attributed to variations in sequencing depth and methylome coverage of promoters (Figure S5D).

Intriguingly, DMVs contain a unique set of genes. In total, 1,086 coding genes are found in the 1,220 DMVs (Table S7). The majority (91.5%) of their promoters are CG rich (Figure S5E). No significant differences in gene sizes are found for DMV genes with CGIs compared to non-DMV genes with CGIs (data not shown). Strikingly, a gene ontology analysis showed that these genes are strongly enriched for functional groups in transcription factors, homeobox family, developmental protein, and embryonic morphogenesis (Figure 5G). In fact, 38.4% (415) of coding genes in DMVs encode DNA-binding proteins (Figure 5H). These genes include hESC and lineage markers such as *SOX2*, *POU5F1*, *ZIC3* (hESC); *EOMES*, *T*, *GSC* (ME); *GLI3*, *SIX3*, *LHX3*, *PAX6* (NPC); *GATA2*, *GATA6* (TBL); and *RUNX1* (MSC). This list also includes transcription factor families that are located in clusters (such as *HOX*), as well as those that reside in different locations (such as *FOX*, *ZIC*, *GATA*, *KLF*, *SIX*, *TBX*, *LHX*, and *DLX*). In addition, genes in DMVs are strongly enriched for those encoding components of development signaling pathways, including WNT, receptor tyrosine kinase (RTK), BMP, and Hedgehog (Figure 5H). Furthermore, there are 319 lncRNA genes with promoters that overlap with DMVs, including 22 lncRNA genes newly identified in this study (Figure 5H and Table S7). Finally, we found 40 microRNA genes in DMVs (Figure 5H and Table S7), 12 (30%) of which are known to be hESC specific (such as mir-302/367) (Suh et al., 2004) or within 10 kb of lineage-restricted genes that we identified (data not shown). Taken together, our data have revealed a unique class of genomic regions that show wide depletion of DNA methylation and are strongly associated with transcription factor genes and developmental genes.

The Majority of DMVs Remain Largely Unmethylated upon Cell Differentiation

Previously, bivalent genes marked by H3K4me3 and H3K27me3 were shown to be highly enriched for developmental genes (Bernstein et al., 2006). Interestingly, DMV genes appear to be more enriched for transcription factors and developmental genes compared to bivalent genes in hESCs as defined in this



WNT1	FGF11	IGF1R	BMP2
WNT10B	FGF13	IGF2	BMP4
WNT11	FGF14	IGF2BP1	BMP7
WNT3	FGF19	IGF2BP2	BMPR1A
WNT3A	FGF5	IGF2BP3	
WNT5A	FGF9	IGFBP2	SMAD3
WNT6	FGFR1	IGFBP3	SMAD6
WNT7B	FGFR2	IGFBP5	SMAD7
FZD1	EPHA2	ID1	GSK3B
FZD10	EPHA4	ID2	
FZD2	EPHA5	ID3	SHH
FZD5	EPHA7		
FZD7	EPHB3		
FZD8	ERBB4		

FOXA1	GATA2	HOXA1	KLF10	LHX1	SOX1
FOXA2	GATA3	HOXA11	KLF13	LHX2	SOX11
FOXB1	GATA4	HOXA13	KLF16	LHX3	SOX14
FOXB2	GATA6	HOXA2	KLF4	LHX4	SOX17
FOXC1		HOXA5	KLF6	LHX5	SOX2
FOXC2	HAND1	HOXA9	KLF7	LHX6	SOX21
FOXD2	HAND2	HOXB2		LHX9	SOX3
FOXD3		HOXB3	SIX1		SOX5
FOXD4	GLI3	HOXB4	SIX2	RUNX1	SOX7
FOXE1	GSC	HOXB5	SIX3	RUNX2	SOX9
FOXF1		HOXB6	SIX6	RUNX3	
FOXF2	OLIG2	HOXB7			NKX1-2
FOXG1	OLIG3	HOXB8	TBX15	SALL1	NKX2-1
FOXH1		HOXB9	TBX18	SALL3	NKX2-2
FOXJ1	EOMES	HOXC10	TBX2	SATB1	NKX2-3
FOXL1		HOXC12	TBX20	SATB2	NKX2-5
FOXL2	PRDM12	HOXC13	TBX3		NKX2-6
FOXO1	PRDM13	HOXC4	TBX4	POU2F1	NKX2-8
FOXO3	PRDM16	HOXC5	TBX5	POU3F2	NKX3-2
FOXP4	PRDM8	HOXC6		POU3F3	NKX6-1
FOXQ1		HOXC8	DLX1	POU3F4	
	IRX1	HOXC9	DLX2	POU4F1	OTX1
T	IRX3	HOXD10	DLX3	POU4F2	OTX2
	IRX4	HOXD11	DLX4	POU4F3	
ZIC1	IRX5	HOXD12	DLX5	POU5F1	PAX1
ZIC2		HOXD13	DLX6		PAX2
ZIC3	ISL1	HOXD3		ONECUT1	PAX3
ZIC4	ISL2	HOXD4	NEUROD1	ONECUT2	PAX6
ZIC5		HOXD8	NEUROG2	ONECUT3	PAX7
		HOXD9	NEUROG3		PAX9

Figure 5. Genes within DMVs Are Strongly Enriched for Transcription Factors and Developmental Genes

(A) DNA methylation levels for a DMV (*GSC*) and a nearby typical UMR (*CLMN*) are shown.

(B) Histograms showing the distribution of the lengths of hypomethylated regions in various cell types.

(C) The numbers of DMVs found in various cell types. The horizontal line indicates the number of DMVs shared by all cell types.

(D) The distribution of lengths of various genomic elements as indicated.

(E) The average conservation level (PhastCons scores) around DMVs.

(F) A Venn diagram showing the overlap of genes with DMVs in humans (H1 and its derived cells) and in mice (frontal cortex).

(G) Gene ontology analysis results for DMV genes in H1 and the H1-derived cells.

(H) A breakdown of the types of DMV genes in H1 and the H1-derived cells, with examples shown in the tables.

See also Figure S5.

study or previous studies (Pan et al., 2007; Zhao et al., 2007) (Figures 6A and S6A). Additionally, genes in DMVs are not simply genes with long CGIs, high promoter CG density, or CGI clusters (Extended Experimental Procedures) (Figures 6A and S6B). We then asked whether DMVs undergo dynamic epigenetic regulation upon H1 differentiation. We examined the DNA methylation levels in H1, the H1-derived cells, and a panel of published methylomes (see Figure 6D and its legend for the list) (Berman et al., 2012; Li et al., 2010; Lister et al., 2011). Interestingly, most of the promoters in DMVs (89.5%, n = 968) remain hypomethylated in all cell types (Figures 6B and S6C). The other 113 promoters demonstrate methylation level at or above 0.4 in at least one cell type (Figures 6B and S6C), including those at several *HOX* genes as shown previously (Bock et al., 2012; Laurent et al., 2010), and genes that have low CG promoters include *POU5F1* (Figure 3E), *DPPA4* (data not shown), and the hESC-specific microRNA gene cluster *mir-302/367* (Figure 6C). Notably, the expression of the *mir*-302/367 cluster can reprogram somatic cells to pluripotent cells (Anokye-Danso et al., 2011). The activity of *mir*-302/367 may be regulated by DNA methylation as indicated by the hypermethylation of the associated DMV upon differentiation (Figure 6C). Therefore, a small subset of DMVs, including those at the *HOX* genes and a number of CG-poor promoters, shows dynamic DNA methylation during cell differentiation.

Next, we examined DMVs that remain hypomethylated upon cell differentiation. Among all 968 coding genes that are located in these DMVs, 259 are defined as aforementioned lineage-restricted genes. Most promoters of these genes are CG rich and are marked differentially by H3K27me3 in various lineages, while lacking DNA methylation in general (Figure 6D). Additionally, 134 genes are repressed in all six cell types and are also predominantly marked by H3K27me3, including *HOXC5/C12/D3/D4*, *FOXB2/D2/D4/E1*, and *PAX3/5/7* (Figure 6D). We then examined genes with DMVs that are expressed in most lineages ($\geq$4) in the current study, including those that are marked by H3K27me3 in at least one of the six cell types, and those that are not marked by H3K27me3 in any cell types (Figure 6D). The first group shows somewhat weak lineage-restricted expression. The second group is active in all six cell types. Gene ontology analysis shows that this group is not enriched for housekeeping genes but instead is still strongly enriched for transcription regulators, such as *MYC*, *MLL*, *SRF*, and *CBX3*, and several histone demethylase genes, *KDM2A/2B*, *JARID2*, and *JMJD1C*. Together, DMV genes appear to be largely marked by H3K4me3 and/or H3K27me3 (Figures 6D and 6E). Interestingly, this is also true in sperm as we examined data sets from published studies (Hammoud et al., 2009; Molaro et al., 2011) (Figures 6D and 6E). Consistent with the notion that many bivalent developmental genes become monovalent upon cell differentiation (Bernstein et al., 2006), a larger portion of DMVs bear only either H3K4me3 or H3K27me3 in differentiated cells compared to that in sperm or H1 (Figure 6E). Interestingly, the sperm genome contains more DMVs than those in other cell types (n = 4,167), and most DMVs in H1 and the H1-derived cells (82.9%) are also present in sperm (Figure 6F). These observations are exemplified at two loci near *HAND1* (Figure 6G) and *MYC* (Figure 6H). Therefore, we conclude that the majority of genes in DMVs remain hypomethylated upon H1 differentiation and are premarked by H3K27me3 and/or H3K4me3 in sperm.

Genes with DMVs Are Hypermethylated in Cancer

As promoters with DMVs are preferentially hypomethylated in most cells that we examined, we sought to examine whether this is also true in cancer. Notably, DMV genes are enriched for genes involved in cancer pathways (Figure 5H), tumor suppressor genes (n = 120, p value = 2×10^{-20}) and oncogenes (n = 72, p value = 5×10^{-14}) (Cancer Gene Database, Memorial Sloan-Kettering Cancer Center) (Table S7). Interestingly, by examining base-resolution methylomes for normal and tumor colon tissues (Berman et al., 2012), we found that promoters in DMVs gain significant levels of DNA methylation in the tumor tissue (Figure 7A). Genome wide, 54.0% of DMVs (n = 659) overlap with the "methylation-prone elements" in colon cancer (Berman et al., 2012). Conversely, 28.9% of methylation-prone elements (n = 1,493) overlap with DMVs. Because the majority of methylation-prone elements (71%) are in nonpromoter regions (Berman et al., 2012), but DMVs are present almost exclusively at promoters, we focused on the promoter regions for the following analysis. Strikingly, promoters that gain most DNA methylation in the tumor sample (ΔmCG/CG $\geq$ 0.4) strongly overlap with DMVs identified in H1 and the H1-derived cells (Figures 7B and 7C). This is true for promoters of both coding genes and lncRNA genes. Similar results were obtained using two additional hypermethylated gene lists in breast cancer and colorectal cancer (Figure S7A). As a control, promoters with DMVs remain hypomethylated in blood cells (Figure 7B). Importantly, most hypermethylated tumor suppressor genes in colon cancer are also DMV genes (16/22, p value = 1×10^{-17}). Unexpectedly, 12 oncogenes are also hypermethylated in colon cancer, among which 9 are DMV genes (p value = 2×10^{-11}). Previously, it was shown that many hypermethylated genes in cancer are Polycomb targets (Bracken and Helin, 2009). Consistently, 87.2% (575/659) of hypermethylated DMVs, compared to 42% (236/561) of nonhypermethylated DMVs, are marked by K27me3 in H1. Taken together, these data suggest that, although DMV promoters are preferentially devoid of DNA methylation in normal cells, they are prone to hypermethylation in cancer.

DISCUSSION

It has long been recognized that epigenetic mechanisms play a critical role in mammalian development, but precisely how DNA methylation and chromatin modifications contribute to development has not yet been clearly elucidated. In this study, we focused on hESCs as a model and generated by far the most comprehensive reference epigenome maps of a multilineage differentiation system in humans. Importantly, we demonstrated that the majority of genes differentially expressed in early progenitors are CG rich and appear to employ H3K27me3-mediated repression in nonexpressing cells. Conversely, genes differentially expressed in later stages are largely CG poor and preferentially show DNA methylation-mediated gene silencing (Figure 7D). Surprisingly, we found more than 1,200 loci, termed DNA methylation valleys, that largely remain unmethylated in most cell types that we examined.

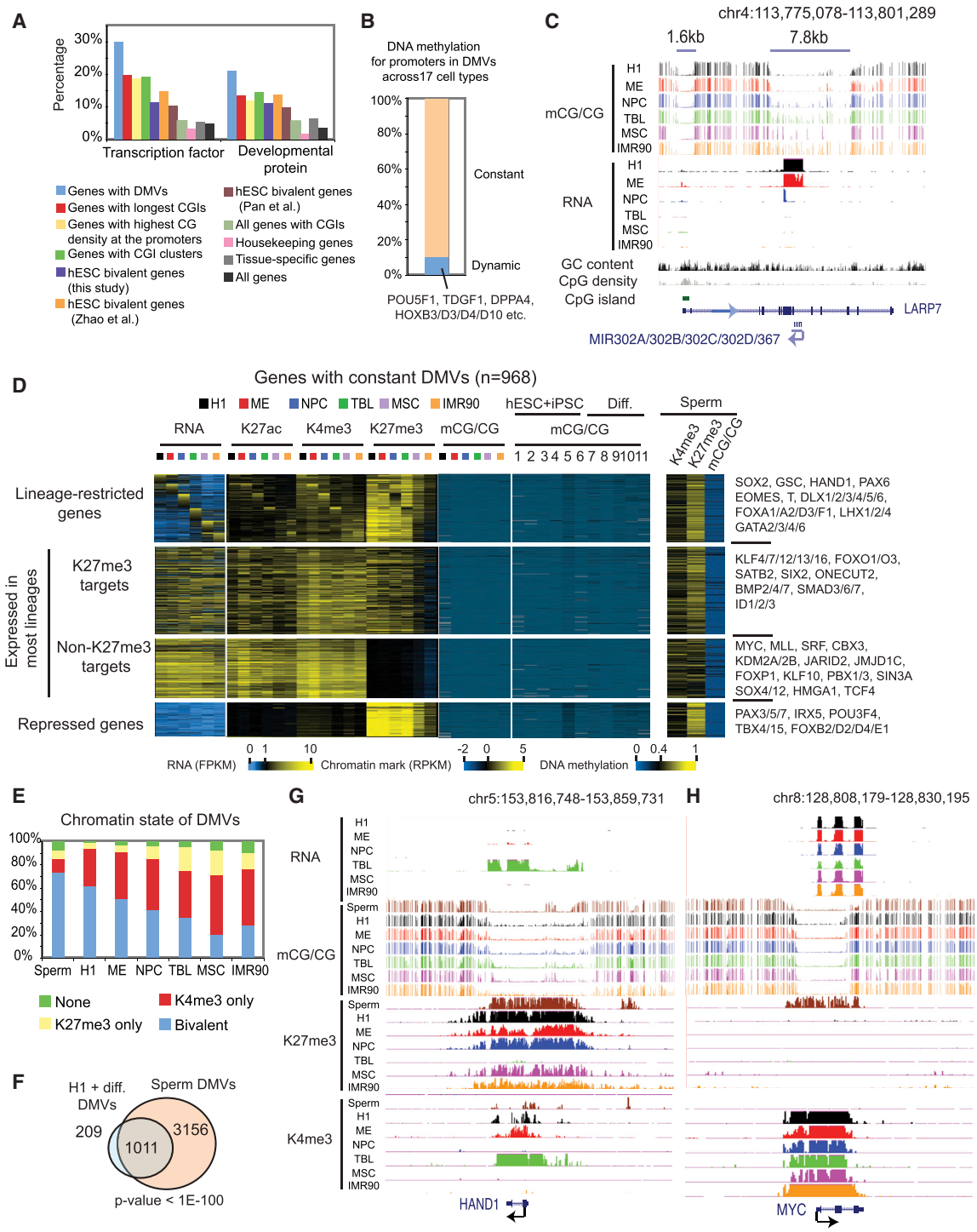


Figure 6. DMVs Largely Remain Hypomethylated in Sperm and Many Terminally Differentiated Cell Types

(A) Percentages of genes that belong to various gene ontology groups are shown as bar graphs for coding genes in DMVs (n = 1,081), genes with longest CGIs (n = 1,081), genes with the highest promoter CG densities (n = 1,081), genes with CGI clusters (n = 1,019), hESC bivalent genes as defined in this study (n = 2,401) or in previous studies (Zhao et al., 2007, n = 1,797 after gene symbol conversion; Pan et al., 2007, n = 3,301 after gene symbol conversion), all RefSeq genes, housekeeping genes (n = 3,140), and somatic-tissue-specific genes (n = 885) as defined in Zhu et al. (2008).

(B) A bar graph showing the percentages of promoters in DMVs that demonstrate dynamic DNA methylation (mCG/CG $\geq$ 0.4 in any cell types) or constant DNA methylation (mCG/CG < 0.4 in any cell types).

(C) The levels of DNA methylation and RNA are shown near *mir-302A/302B/302C/302D/367*. A transcript, likely the hosting transcript for this microRNA gene cluster, is observed mainly in H1 and ME (only − strand RNA reads are shown for simplicity).

(legend continued on next page)

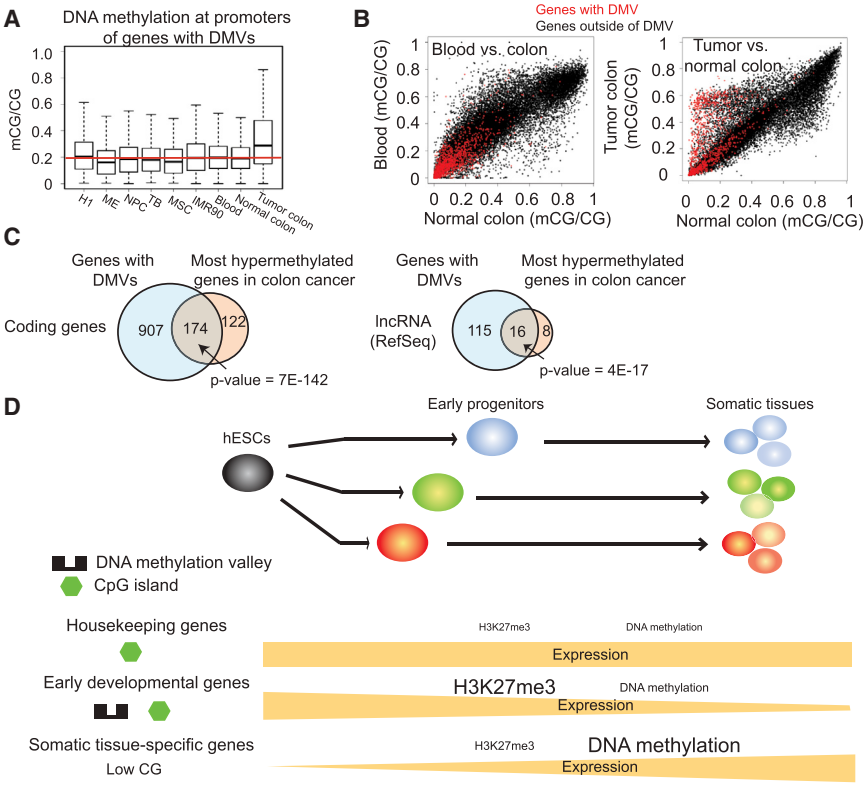


Figure 7. DMVs Are Preferentially Methylated in Cancer

(A) Box plots showing the distribution of the DNA methylation levels at promoters in DMVs for various cell types.

(B) Scatterplots showing the DNA methylation levels at promoters between colon and blood (left) and normal and tumor colon (right). Red, promoters with DMVs; black, all other promoters in the genome.

(C) Venn diagrams showing the overlaps between genes of which the promoters are hypermethylated in colon cancer (ΔmCG $\geq$ 0.4, at least 10 CGs covered) and genes with DMVs for coding genes (left) and lncRNA genes (right).

(D) A model for three classes of promoters with distinct sequence features and epigenetic regulation mechanisms in cell differentiation.

See the Discussion for details and also Figure S7.

These regions are uniquely enriched for transcription factor and developmental regulatory genes. Interestingly, DMVs frequently gain abnormal DNA methylation in cancer, suggesting that alterations in DNA methylation machinery might be an important epigenetic mechanism aiding tumorigenesis. In accordance with an independent study of human ES cells differentiating to cells representative of three germ layers (Gifford et al., 2013 [this issue of *Cell*]), we observed cell-type-specific, dynamic DNA methylation and H3K27me3 at enhancers during ES cell differentiation. Our analysis further demonstrated that dynamic changes of DNA methylation and chromatin marks at enhancers correlate with gene expression, suggesting a potential role of epigenetic modulators in regulating enhancer activities.

Distinct Epigenetic Mechanisms at Lineage-Restricted Genes Expressed at Early and Late Stages of ES Cell Differentiation

Previous studies have shown that somatic-tissue-specific promoters tend to be CG poor (Barrera et al., 2008; Schug et al., 2005). However, we found that a large number of CG-rich promoters appear to drive lineage-specific expression in hESC-derived early precursor cells. In line with previous studies, these CG-rich promoters tend to employ Polycomb, but not DNA methylation, for repression (Meissner et al., 2008; Mendenhall et al., 2010; Mohn et al., 2008). By contrast, dynamic DNA methylation is frequently observed at the late-stage lineage-restricted promoters, which are characterized by CG-poor sequences. Similar results were obtained when we analyzed two published time course data sets for single lineage hESC differentiation to trophoblast (Xu et al., 2002) (Figure S7B) or cardiovascular cells (Paige et al., 2012) (Figure S7C). Together, these data add to the notion that low and high CG promoters are regulated by distinct epigenetic regulatory mechanisms (Meissner et al., 2008) and further suggest a temporal relationship of DNA methylation and Polycomb in regulating cell-type-specific genes.

DMVs Are a Special Class of Genomic Loci Subject to Exquisite Epigenetic Control

Interestingly, many genes encoding for key regulators of embryonic development reside in hypomethylated domains, or DMVs. Importantly, these DMVs are also preferentially hypomethylated in sperm, raising the possibility that these DMVs may be established even earlier. Why are developmental regulatory genes preferentially located in DMVs? One possibility is that DNA methylation at these regions may be incompatible with maintenance of the pluripotency or multipotency of these cells. We

(D) Heatmaps showing RNA, H3K27ac, H3K4me3, H3K27me3, and DNA methylation levels for promoters of genes with DMVs within various categories. The levels of DNA methylation in additional 11 cell types and sperm, as well as the levels of H3K4me3 and H3K27me3 in sperm, are also shown. 1, hESC H9; 2–4, foreskin fibroblast (FF)-derived iPSC lines (19.11,6.9,19.7); 5, adipose-derived stem (ADS) cell iPSCs; 6, FF iPSC-derived trophoblast-like cells; 7, ADS; 8, ADS-derived adipocytes; 9, FF (Lister et al., 2011); 10, PMBC (blood) (Li et al., 2010); 11, colon tissue (Berman et al., 2012).

(E) The chromatin state (presence of H3K4me3 and/or H3K27me3) of DMVs is shown for various cell types.

(F) The overlap of DMVs is shown between those in H1 and its derived cells and those in sperm.

(G and H) The epigenetic landscape is shown for the DMV associated with the gene *HAND1* (G) or *MYC* (H).

See also Figure S6.

noticed that many DMV genes demonstrate a bivalent state (H3K4me3 and H3K27me3), which is linked to poised transcription that may enable developmental genes to be more flexibly modulated (Bernstein et al., 2006). DNA methylation, on the other hand, may be required for more stable silencing of genes in terminally differentiated cells. Another possibility is that the genetic programs regulating embryonic development may actually evolve separately from, or prior to, the evolution of DNA methylation machinery. Supporting this hypothesis, DNA methylation is either absent (such as in *Drosophila* and C. *elegans*) or varies considerably in its pattern relative to gene activity in invertebrates (Feng et al., 2010; Zemach et al., 2010). On the other hand, the Polycomb family of factors regulates key developmental regulatory genes in both invertebrates and vertebrates in a more conserved manner. Several mechanisms of DNA hypomethylation at DMVs can be envisioned. DMVs may be recognized by proteins, such as the Tet family, that actively remove DNA methylation (Wu and Zhang, 2011). Alternatively, DMVs may be associated with histone modifications or histone variants, such as H3K4me3 or H2A.Z, that are incompatible to DNA methylation (Cedar and Bergman, 2009). Future experiments are needed to determine which of the above mechanisms could be responsible for DMV formation in the mammalian genome.

EXPERIMENTAL PROCEDURES

hESC Differentiation

H1 cells were differentiated according to previously established protocols to mesendoderm (Yu et al., 2011), trophoblast-like cells (Xu et al., 2002), neural progenitor cells (Chambers et al., 2009; Chen et al., 2011), and mesenchymal stem cells (Vodyanik et al., 2010). Details of the differentiation methods can be found in Extended Experimental Procedures.

MethylC-Seq Library Generation and Sequencing

Genomic DNA from H1 and the H1-derived cells was extracted and sonicated. Sequencing libraries were constructed using NEBNext DNA Sample Prep Reagent Set 1 (NEB). Methylated adapters were used in place of the standard genomic DNA adapters from Illumina. Ligation products were purified, bisulfite treated, PCR amplified, and sequenced using HiSeq2000 (Illumina).

ChIP-Seq Library Generation and Sequencing

H1 and the H1-derived cells were processed following a ChIP protocol as previously described (Hawkins et al., 2010). ChIP libraries were prepared and sequenced using the Illumina instrument per manufacturer's instructions.

RNA-Seq Library Generation and Sequencing

Total RNA from H1 and the H1-derived cells was extracted and sequencing libraries were constructed using the TruSeq RNA Sample Prep Kit (Illumina) (Poly(A) selected) according to manufacturer's instructions with modifications to confer strand specificity (see Extended Experimental Procedures for details).

ACCESSION NUMBERS

All data have been deposited to the Sequence Read Archive (SRA) under accession number SRP000941.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, seven figures, and seven tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2013.04.022.

ACKNOWLEDGMENTS

This work is supported by the NIH Epigenome Roadmap Project (U01 ES017166) and also in part by NSFC 91019016 and NIH R01 HG001696 (M.Q.Z.), NIH P01 GM081629 (J.A.T.), and CIRM RN2-00905-1 (B.R.). J.R.E. is a Howard Hughes Medical Institute and Gordon and Betty Moore Investigator. N.C.C. is funded by grants from the NIH/NHLBI. H.Y. is funded by grants from the American Heart Association (12POST12050080). We thank Drs. Tomek Swigut and Joanna Wysocka for sharing the zebrafish enhancer reporter vector. We thank Dr. John Stamatoyannopoulos for generating and providing access to the DNase-seq data sets. We also thank members of the Ren lab for helpful comments of the manuscript. B.R., J.A.T., J.R.E., W.W., and M.Q.Z. designed and supervised the research. Z.H., J.Z., P.Y., N.E.P., K.S., J.E.A.-B., and I.S. performed/supervised the H1 differentiation experiments. W.X., R.D.H., D.L., A.Y.L., A.K., S. Kuan, C.Y., and S. Klugman performed ChIP-seq experiments. R.L. and J.R.N. performed MethylC-seq experiments. M.A.U., Y.L., and Y.Z. performed RNA-seq experiments. H.Y. and N.C.C. performed/supervised the enhancer-reporter assay in zebrafish. W.X., M.D.S., N.R., P.R., J.W.W., S.T., T.W., S.A.S., Y.Z., R.L., H.C., L.E.E., U.W., A.K., Z.X., W.Y.C., and R.S. analyzed data. W.X., B.R., and R.S. prepared the manuscript. B.R., J.A.T., J.R.E., W.W., and M.Q.Z. are equally responsible for the analysis results. M.D.S., R.L., Z.H., N.R., P.R., J.W.W., S.T., R.D.H., and D.L. contributed equally to this work.

Received: September 29, 2012
Revised: January 7, 2013
Accepted: April 1, 2013
Published: May 9, 2013

REFERENCES

Anokye-Danso, F., Trivedi, C.M., Juhr, D., Gupta, M., Cui, Z., Tian, Y., Zhang, Y., Yang, W., Gruber, P.J., Epstein, J.A., and Morrisey, E.E. (2011). Highly efficient miRNA-mediated reprogramming of mouse and human somatic cells to pluripotency. Cell Stem Cell *8*, 376–388.

Barrera, L.O., Li, Z., Smith, A.D., Arden, K.C., Cavenee, W.K., Zhang, M.Q., Green, R.D., and Ren, B. (2008). Genome-wide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs. Genome Res. *18*, 46–59.

Berman, B.P., Weisenberger, D.J., Aman, J.F., Hinoue, T., Ramjan, Z., Liu, Y., Noushmehr, H., Lange, C.P., van Dijk, C.M., Tollenaar, R.A., et al. (2012). Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. Nat. Genet. *44*, 40–46.

Bernardo, A.S., Faial, T., Gardner, L., Niakan, K.K., Ortmann, D., Senner, C.E., Callery, E.M., Trotter, M.W., Hemberger, M., Smith, J.C., et al. (2011). BRACHYURY and CDX2 mediate BMP-induced differentiation of human and mouse pluripotent stem cells into embryonic and extraembryonic lineages. Cell Stem Cell *9*, 144–155.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315–326.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev. *16*, 6–21.

Bock, C., Beerman, I., Lien, W.H., Smith, Z.D., Gu, H., Boyle, P., Gnirke, A., Fuchs, E., Rossi, D.J., and Meissner, A. (2012). DNA methylation dynamics during in vivo differentiation of blood and skin stem cells. Mol. Cell *47*, 633–647.

Boyer, L.A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K., et al. (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature *441*, 349–353.

Bracken, A.P., and Helin, K. (2009). Polycomb group proteins: navigators of lineage pathways led astray in cancer. Nat. Rev. Cancer *9*, 773–784.

Cedar, H., and Bergman, Y. (2009). Linking DNA methylation and histone modification: patterns and paradigms. Nat. Rev. Genet. *10*, 295–304.

Chambers, S.M., Fasano, C.A., Papapetrou, E.P., Tomishima, M., Sadelain, M., and Studer, L. (2009). Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. Nat. Biotechnol. *27*, 275–280.

Chang, C.W., Cheng, W.C., Chen, C.R., Shu, W.Y., Tsai, M.L., Huang, C.L., and Hsu, I.C. (2011). Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. PLoS ONE *6*, e22859.

Chen, G., Gulbranson, D.R., Hou, Z., Bolin, J.M., Ruotti, V., Probasco, M.D., Smuga-Otto, K., Howden, S.E., Diol, N.R., Propson, N.E., et al. (2011). Chemically defined conditions for human iPSC derivation and culture. Nat. Methods *8*, 424–429.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc. Natl. Acad. Sci. USA *107*, 21931–21936.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Feng, S., Cokus, S.J., Zhang, X., Chen, P.Y., Bostick, M., Goll, M.G., Hetzel, J., Jain, J., Strauss, S.H., Halpern, M.E., et al. (2010). Conservation and divergence of methylation patterning in plants and animals. Proc. Natl. Acad. Sci. USA *107*, 8689–8694.

Gifford, C.A., Ziller, M.J., Gu, H., Trapnell, C., Donaghey, J., Tsankov, A., Shalek, A.K., Kelley, D.R., Shishkin, A.A., Issner, R., et al. (2013). Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. Cell *153*. Published online May 9, 2013. http://dx.doi.org/10.1016/j.cell.2013.04.037.

Hammoud, S.S., Nix, D.A., Zhang, H., Purwar, J., Carrell, D.T., and Cairns, B.R. (2009). Distinctive chromatin in human sperm packages genes for embryo development. Nature *460*, 473–478.

Hawkins, R.D., Hon, G.C., Lee, L.K., Ngo, Q., Lister, R., Pelizzola, M., Edsall, L.E., Kuan, S., Luu, Y., Klugman, S., et al. (2010). Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell Stem Cell *6*, 479–491.

Irizarry, R.A., Wu, H., and Feinberg, A.P. (2009). A species-generalized probabilistic model-based definition of CpG islands. Mamm. Genome *20*, 674–680.

Jackson, M., Krassowska, A., Gilbert, N., Chevassut, T., Forrester, L., Ansell, J., and Ramsahoye, B. (2004). Severe global DNA hypomethylation blocks differentiation and induces histone hyperacetylation in embryonic stem cells. Mol. Cell. Biol. *24*, 8862–8871.

Kouzarides, T. (2007). Chromatin modifications and their function. Cell *128*, 693–705.

Kunarso, G., Chia, N.Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.S., Ng, H.H., and Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat. Genet. *42*, 631–634.

Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C.T., Low, H.M., Kin Sung, K.W., Rigoutsos, I., Loring, J., and Wei, C.L. (2010). Dynamic changes in the human methylome during differentiation. Genome Res. *20*, 320–331.

Lee, T.I., Jenner, R.G., Boyer, L.A., Guenther, M.G., Levine, S.S., Kumar, R.M., Chevalier, B., Johnstone, S.E., Cole, M.F., Isono, K., et al. (2006). Control of developmental regulators by Polycomb in human embryonic stem cells. Cell *125*, 301–313.

Leung, D.C., and Lorincz, M.C. (2012). Silencing of endogenous retroviruses: when and why do histone marks predominate? Trends Biochem. Sci. *37*, 127–133.

Li, E., Bestor, T.H., and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell *69*, 915–926.

Li, Y., Zhu, J., Tian, G., Li, N., Li, Q., Ye, M., Zheng, H., Yu, J., Wu, H., Sun, J., et al. (2010). The DNA methylome of human peripheral blood mononuclear cells. PLoS Biol. *8*, e1000533.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315–322.

Lister, R., Pelizzola, M., Kida, Y.S., Hawkins, R.D., Nery, J.R., Hon, G., Antosiewicz-Bourget, J., O'Malley, R., Castanon, R., Klugman, S., et al. (2011). Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature *471*, 68–73.

Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., Jaffe, D.B., et al. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature *454*, 766–770.

Mendenhall, E.M., Koche, R.P., Truong, T., Zhou, V.W., Issac, B., Chi, A.S., Ku, M., and Bernstein, B.E. (2010). GC-rich sequence elements recruit PRC2 in mammalian ES cells. PLoS Genet. *6*, e1001244.

Mohn, F., Weber, M., Rebhan, M., Roloff, T.C., Richter, J., Stadler, M.B., Bibel, M., and Schübeler, D. (2008). Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Mol. Cell *30*, 755–766.

Molaro, A., Hodges, E., Fang, F., Song, Q., McCombie, W.R., Hannon, G.J., and Smith, A.D. (2011). Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. Cell *146*, 1029–1041.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell *99*, 247–257.

Ong, C.T., and Corces, V.G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat. Rev. Genet. *12*, 283–293.

Paige, S.L., Thomas, S., Stoick-Cooper, C.L., Wang, H., Maves, L., Sandstrom, R., Pabon, L., Reinecke, H., Pratt, G., Keller, G., et al. (2012). A temporal chromatin signature in human embryonic stem cells identifies regulators of cardiac development. Cell *151*, 221–232.

Pan, G., Tian, S., Nie, J., Yang, C., Ruotti, V., Wei, H., Jonsdottir, G.A., Stewart, R., and Thomson, J.A. (2007). Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. Cell Stem Cell *1*, 299–312.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279–283.

Rajagopal, N., Xie, W., Li, Y., Wagner, U., Wang, W., Stamatoyannopoulos, J., Ernst, J., Kellis, M., and Ren, B. (2013). RFECS: a Random-Forest based algorithm for Enhancer Identification from Chromatin State. PLoS Comput. Biol. *9*, e1002968.

Schug, J., Schuller, W.P., Kappen, C., Salbaum, J.M., Bucan, M., and Stoeckert, C.J., Jr. (2005). Promoter features related to tissue specificity as measured by Shannon entropy. Genome Biol. *6*, R33.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490–495.

Suh, M.R., Lee, Y., Kim, J.Y., Kim, S.K., Moon, S.H., Lee, J.Y., Cha, K.Y., Chung, H.M., Yoon, H.S., Moon, S.Y., et al. (2004). Human embryonic stem cells express a unique set of microRNAs. Dev. Biol. *270*, 488–498.

Thomson, J.A., Itskovitz-Eldor, J., Shapiro, S.S., Waknitz, M.A., Swiergiel, J.J., Marshall, V.S., and Jones, J.M. (1998). Embryonic stem cell lines derived from human blastocysts. Science *282*, 1145–1147.

Tsumura, A., Hayakawa, T., Kumaki, Y., Takebayashi, S., Sakaue, M., Matsuoka, C., Shimotohno, K., Ishikawa, F., Li, E., Ueda, H.R., et al. (2006). Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. Genes Cells *11*, 805–814.

Vastenhouw, N.L., and Schier, A.F. (2012). Bivalent histone modifications in early embryogenesis. Curr. Opin. Cell Biol. *24*, 374–386.

Vodyanik, M.A., Yu, J., Zhang, X., Tian, S., Stewart, R., Thomson, J.A., and Slukvin, I.I. (2010). A mesoderm-derived precursor for mesenchymal stem and endothelial cells. Cell Stem Cell *7*, 718–729.

Wu, H., and Zhang, Y. (2011). Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation. Genes Dev. *25*, 2436–2452.

Xie, W., Barr, C.L., Kim, A., Yue, F., Lee, A.Y., Eubanks, J., Dempster, E.L., and Ren, B. (2012). Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. Cell *148*, 816–831.

Xu, R.H., Chen, X., Li, D.S., Li, R., Addicks, G.C., Glennon, C., Zwaka, T.P., and Thomson, J.A. (2002). BMP4 initiates human embryonic stem cell differentiation to trophoblast. Nat. Biotechnol. *20*, 1261–1264.

Yao, T.P., Oh, S.P., Fuchs, M., Zhou, N.D., Ch'ng, L.E., Newsome, D., Bronson, R.T., Li, E., Livingston, D.M., and Eckner, R. (1998). Gene dosage-dependent embryonic development and proliferation defects in mice lacking the transcriptional integrator p300. Cell *93*, 361–372.

Young, R.A. (2011). Control of the embryonic stem cell state. Cell *144*, 940–954.

Yu, P., Pan, G., Yu, J., and Thomson, J.A. (2011). FGF2 sustains NANOG and switches the outcome of BMP4-induced human embryonic stem cell differentiation. Cell Stem Cell *8*, 326–334.

Zemach, A., McDaniel, I.E., Silva, P., and Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science *328*, 916–919.

Zhao, X.D., Han, X., Chew, J.L., Liu, J., Chiu, K.P., Choo, A., Orlov, Y.L., Sung, W.K., Shahab, A., Kuznetsov, V.A., 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.

Zhu, J., He, F., Hu, S., and Yu, J. (2008). On the nature of human housekeeping genes. Trends Genet. *24*, 481–484.

Zhu, J., Adli, M., Zou, J.Y., Verstappen, G., Coyne, M., Zhang, X., Durham, T., Miri, M., Deshpande, V., De Jager, P.L., et al. (2013). Genome-wide chromatin state transitions associated with developmental and environmental cues. Cell *152*, 642–654.

Note Added in Proof

While this manuscript was in revision, the two following related papers describing hESC-specific expression of the HERV-H retrotransposable elements were published. Kelley, D.R., and Rinn, J.L. (2012). Transposable elements reveal a stem cell specific class of long noncoding RNAs. Genome Biol. 13, R107. Santoni, F.A., Guerra, J., and Luban, J. (2012). HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. Retrovirology 9, 111.

EXTENDED EXPERIMENTAL PROCEDURES

Cell Culture

All differentiation media is based on the E8 medium (Chen et al., 2011). For mesendoderm differentiation, a protocol derived from Yu et al. (2011) was used. Briefly, undifferentiated H1 cells were maintained in E8 medium (which contains FGF2) on Matrigel coated plates. For differentiation, H1 cells were individualized with TrypLE (Life Technology) and washed once with E8 medium. H1 cells were then seeded onto fresh Matrigel plates at a density of around $1x10^4$ cells/cm^2 and cultured in differentiation medium (E8 with 5ng/ml BMP4 and 25ng/ml Activin A). Media was changed every day. Cells were harvested at the end of day 2. For further differentiation into definitive endoderm cells, mesendoderm cells were further treated with E8 medium supplemented with 100 ng/ml Activin for 2 days.

For trophoblast-like cell differentiation, a protocol modified from Xu et al. (2002) was used. Briefly, undifferentiated H1 cells were maintained in E8 medium on Matrigel coated plates. For differentiation, H1 cells were treated with 0.5mM EDTA in PBS for 3 to 5 min and resuspended in differentiation medium. Resuspended H1 cells were then seeded onto fresh matrigel plates at densities around $4x10^4$ cells/cm^2. Media was changed every day and cells were harvested at the end of day 5. Differentiation medium consists of E8 minus FGF2 with 50ng/ml BMP4.

For neural progenitor differentiation, a modified protocol from Chambers et al. (2009) was employed. Briefly, undifferentiated H1 cells were maintained in E8 medium on Matrigel coated plates. For differentiation, H1 cells were treated with 0.5mM EDTA in PBS for 3 to 5 min and resuspended in NPC differentiation medium. Resuspended H1 cells were then seeded onto fresh matrigel plates at densities around 1-$2x10^4$ cells/cm^2. Media was changed every day. Cells were harvested at the end of day 7. The NPC differentiation medium consists of E8 minus FGF2, minus TGFβ1, with only 5 μg/ml insulin, with 10 μM SB431542 and 100ng/ml Noggin. For further differentiation into neurons, NPCs were cultured in DMEM/F12 medium supplemented with 1x N2, 1x B27, 64 μg/ml vitamin C, 14ng/ml sodium selenite and 5ng/ml FGF2 for additional 25 days.

For mesenchymal stem cell differentiation, a previously described protocol was used without modifications (Vodyanik et al., 2010).

For immunostaining of definitive endoderm cells, cells were fixed in 2% paraformaldehyde in PBS for 15 min at room temperature followed by a wash with PBS. Cells were permeabilized and blocked with 1% BSA in PBS/0.25% Triton X-100 for 1 hr and incubated with primary antibodies overnight at 4°C. For all other stainings, cells were fixed and permeabilized with ice cold 90% methanol in PBS and blocked with 2% FBS in PBS. Cells were then incubated with primary antibody in blocking buffer at 4°C overnight. Cells were then incubated with secondary antibodies for 1 hr at room temperature in the dark. Nuclei were visualized by DAPI (Vector laboratories).

For the FACS analysis, cells were individualized with Accutase (Life Technologies), fixed in 2% paraformaldehyde, and permeabilized with ice cold 90% methanol in PBS. Cells were then incubated with primary antibodies in the FACS buffer (2% FBS in 1xPBS) for 4 hr at RT or 4°C overnight followed by secondary antibody incubation at RT for 1 hr. Stained cells were analyzed using FACS CANTO-II (BD).

Primary antibodies used are as follows: GATA3 (558686, BD Biosciences), EOMES (50-4877-42, eBiosciences), TH (P40101-0, Pel-Freez Biologicals), POU5F1(sc-5279, Santa Cruz). Additionally, NANOG (4903), SOX2 (3579), KRT7 (4465), MAP2 (4542) are from Cell Signaling. T (AF2085), FOXA2(AF2400), CXCR4(MAB172), LHX1(MAB2725), GATA2 (AF2046), bIII-Tubulin (MAB1195) are from R&D Systems. TBR1 (AB10554) and PLZF(OP128) are from Millipore. PAX6 is from Developmental Studies Hybridoma Bank at the University of Iowa.

ChIP-Seq

ChIP was carried out as previously described with 20 μg chromatin and 5 μg antibody with the following antibodies: H3K4me3 (Active Motif, 39159), H3K4me1 (Abcam, ab8895), H3K27Ac (Active Motif, 39133), H3K36me3 (Abcam, ab9050), H3K27me3 (Active Motif, 39155), H3K9me3 (Abcam, ab8898), H3K79me1 (Abcam, ab2886), H2AK5ac (Abcam, ab45152), H2BK120ac (Active Motif, 39119), H2BK5ac (Active Motif, 39123), H3K18ac (Abcam, ab1191), H3K4ac (Millipore, 07-539) and H3K9ac (Active Motif, 39137). ChIP and input library preparation and sequencing procedures were carried out as described previously (Hawkins et al., 2010) according to Illumina protocols with minor modifications (Illumina, San Diego, CA).

MethylC-Seq

Unmethylated Lambda DNA was added to genomic DNA at 1:200 (w/w). Approximately two micrograms of genomic DNA was sonicated to ~100 bp using the Covaris S2 System with the following parameters: cycle number = 6, duty cycle = 20%, intensity = 5, cycles/burst = 200 and time = 60 s. Sonicated DNA was purified using QIAGEN DNeasy minielute columns (QIAGEN). Each sequencing library was constructed using the NEBNext DNA Sample Prep Reagent Set 1 (New England Biolabs, Ipswich, MA) according to the manufacturer's instructions with the following slight modifications. Methylated adapters were used in place of the standard genomic DNA adapters from Illumina (Illumina, San Diego, CA). Ligation products were purified with AMPure XP beads (Beckman, Brea, CA). Adaptor-ligated DNA (450 ng) was bisulfite treated using the MethylCode Kit (Invitrogen, Carlsbad, CA) following the manufacturer's guidelines and then PCR amplified using PfuTurboCx hotstart DNA polymerase (Agilent, Santa Clara, CA) with the following PCR conditions (2 min at 95þC, 4 cycles of 15 s at 98þC, 30 s at 60þC, 4 min at 72þC, then 10 min at 72C). MethylC-Seq libraries were sequenced using the Illumina HiSeq 2000 (Illumina) instrument as per manufacturer's instructions.

Sequencing of libraries was performed up to 101 cycles. Image analysis and base calling were performed with the standard Illumina pipeline version RTA 2.8.0.

RNA-Seq

Approximately four micrograms of total RNA was used as input. Each sequencing library was constructed using the TruSeq RNA Sample Prep Kit (Illumina, San Diego, CA) (Poly(A) selected) according to manufacturer's instructions with the following modifications to confer strand-specificity. The RNA was incubated in the Elute, Prime, Fragment Mix at 94þC for 4 min. After first strand synthesis the cDNA:RNA hybrid was purified using RNAClean XP beads (Beckman, Brea, CA). The second strand synthesis was performed using a dNTP mix containing dUTPs (10mM dATPs, 10mM dGTPs, 10mM dCTPs, and 20mM dUTPs) and DNA Polymerase I (*E. coli*) (New England Biolabs, Ipswich, MA). The purified ligation products were incubated with Uracil DNA Glycosylase (Fermentas) before PCR amplification. The completed library was then gel size selected to approximately 350-450 bp using the QIAquick Gel Extraction Kit (QIAGEN). RNA-seq libraries were sequenced using the Illumina HiSeq 2000 (Illumina) instrument as per manufacturer's instructions. Sequencing of libraries was performed up to 2 × 101 cycles. Image analysis and base calling were performed with the standard Illumina pipeline version RTA 2.8.0.

Data Analyses

Sequencing Data Mapping

For H1 and the H1 derived cells, to account for the genetic differences between the H1 genome and the hg18 reference genome, we called SNPs using the Bioscope software package with data available from SRA (SRA049981) and incorporated the homozygous variants into the hg18 reference (i.e., replaced the reference base with the called variant). This reference genome, which we refer to as the H1-modified reference genome, was used for all mapping of the H1 and H1 derived cells. Data for IMR90 were obtained from previous publications without modifications (Hawkins et al., 2010; Lister et al., 2009). To be consistent with the RNA-Seq data for H1 and the H1-derived cells, we reproduced the IMR90 RNA-Seq data using the Illumina sequencing platform, which were then mapped to the hg18 reference genome.

ChIP-Seq Data Processing

For H1 and the H1 derived cells, ChIP-seq reads were aligned to the H1-modified reference genome with Bowtie (version 0.12). We used the first 25 bp for the alignment and only reads with less than two mismatches were accepted. To generate the ChIP-seq signals for each histone modification shown in the UCSC genome browser, we normalized the read counts for both the ChIP and the input samples by computing the number of reads per kilobase of bin per million reads sequenced (RPKM). The RPKM values for the ChIP signal were then subtracted by those for the input signal as described previously (Hawkins et al., 2010) and were shown as the UCSC genome browser tracks. For the downstream data analyses, RPKM values were averaged for each bin between replicates. To minimize the batch and cell type variation, the RPKM values were further normalized through Z-score transformation, by subtracting the mean of RPKM across the genome and divided by the standard deviation of RPKM across the genome.

MethylC-Seq Data Processing

For H1 and the H1 derived cells, MethylC-Seq reads were aligned to the H1-modified reference genome with a pipeline that was previously established (Lister et al., 2009).

RNA-Seq Data Processing

For H1 and the H1 derived cells, the RNA-Seq reads were mapped to the H1-modified reference with TopHat (version 1.20). The mapped reads were further analyzed by Cufflinks (Trapnell et al., 2012) and the expression levels for each transcript were quantified as Fragments Per Kilobase of transcript per Million mapped reads (FPKM). For coding genes, we used the well-curated RefSeq database (Pruitt et al., 2012) and selected those RefSeq IDs starting with "NM." For long non-coding genes, we examined several databases including RefSeq, a collected lincRNA database (Cabili et al., 2011), and the NONCODE database (Bu et al., 2012). As these databases contain overlapping annotations, we examined the databases in the following order: RefSeq, lincRNA (Cabili et al., 2011) and NONCODE. Genes that have 80% or more annotated regions overlapping with prior databases were removed. For RefSeq non-coding genes, we first collected all genes starting with "NR" (n = 7445) and then curated a total of 2,868 lncRNA genes from all NR genes, by removing small RNA genes and pseudogenes. For all potential lncRNA genes we required the minimal length to be 200bp. For genes in the NONCODE database we also required genes to have more than one exon to increase the probability that they are true lncRNA genes. For genes with multiple isoforms, the FPKM values were summed across all isoforms as the expression values for the genes.

Analysis of Repetitive Elements

The RepeatMasker annotation file was downloaded from the UCSC genome browser, and Cufflinks was used to measure the transcription levels for each mappable repetitive element. To define expressed repetitive elements in each cell type, we used the following criteria: 1) elements do not overlap with any of the RefSeq annotations; 2) FPKM > = 1; and 3) length > = 300 bp.

De Novo lncRNA Identification

To identify potential novel lncRNA genes, all mapped reads were analyzed and compared to a panel of expanded annotation databases. For this purpose, we first combined the annotations from multiple databases using GFFRead (as part of the Cufflinks package). We used the following databases to create a combined reference transcriptome: RefSeq, NONCODE, Ensembl (Hubbard et al., 2002), UCSC Known Gene database (Hsu et al., 2006), together consisting of a total of 134,377 known transcripts. We then

assembled the transcripts with Cufflinks, Cuffmerge and Cuffdiff based on the standard protocol (Trapnell et al., 2012) against the combined gene annotation. The resulting assembled transcripts were further compared to the latest released Gencode gene annotation (Derrien et al., 2012). Transcripts that do not overlap with any existing annotations, or with less than 80% total length overlapping with existing lncRNA genes, were selected as candidate novel lncRNA genes. We then employed a method similar to that used by Cabili et al. (2011) to identify a stringent set of novel transcripts that have a greater potential of being functional lncRNAs. First, we only selected multi-exonic transcripts to minimize the chances of including transcriptional "noise" (Guttman and Rinn, 2012). Second, we selected those with the expression level of above FPKM 0.5 (and both replicates should be above 0.25) in at least one cell type. Third, we calculated the coding potential score of each transcript as well as its list of BLAST hits in known and predicted mammalian protein database. The coding potential score for each transcript is indicative of its chances of being part of a coding protein isoform, which takes into account factors such as the presence of Open Reading Frames (ORFs), presence of stop codon, evolutionary statistics of codon usage, and homology with known proteins. We used Coding Potential Calculator which is robust also for low quality transcripts (Kong et al., 2007). We removed those transcripts with coding potential scores greater than 0 and subjected the rest to an additional BLAST search (Altschul et al., 1997) in the Uniref90 database (Apweiler et al., 2004) to eliminate ancient or partial coding transcripts (Kong et al., 2007). Those without any BLAST hits were reported as the final novel lncRNA genes.

Identification of Lineage-Restricted Genes

To identify lineage restricted genes, we used a strategy described previously based on the Shannon entropy to compute a cell type-specificity index to each gene (Barrera et al., 2008; Schug et al., 2005; Shen et al., 2012). Specifically, for each gene, we defined its relative expression in a cell type i as Ri = Ei / Σ E, where Ei is the FPKM value for gene in the cell type i; Σ E is the sum of FPKM values in all cell types; N is the total number of cell types. Then the entropy score for this gene across cell types can be defined as H = −1*Σ Ri * logRi($1 \leq i \leq N$), where the value of H ranges between 0 to log2(N). An entropy score close to zero indicates the expression of this gene is highly cell type-specific, while an entropy score close to log2(N) indicates that this gene is expressed ubiquitously. Based on an examination of the entropy distribution (Figure S2B), genes with entropy less than 2 were selected as the candidate lineage restricted genes. Among these genes, we selected candidates of lineage-restricted genes for each cell type based on the following criteria: the gene is highly expressed in this lineage (FPKM > = 1, which doubles the threshold that we used for calling a gene to be expressed), and such high expression cannot be observed in more than two additional cell types. These genes were then reported in the final lineage restricted gene lists.

Lists of somatic tissue-specific genes and housekeeping genes were obtained from (Zhu et al., 2008) and were available from http://www.wikicell.org/index.php/HK_Gene and http://www.wikicell.org/index.php/TS_Gene per the authors' instruction. Independent tissue-specific and housekeeping gene lists were downloaded from (Chang et al., 2011).

Identification of Active Promoter for Lineage-Restricted Genes

Our initial analysis of RNA-Seq data showed that many genes contain multiple promoters, including those that were not annotated yet. Although Cufflinks allows quantification of transcription levels for each isoform with different promoters, a visual inspection indicated that it may be difficult to accurately assign each sequencing read to different isoforms for each gene. To precisely identify promoters for genes that are expressed in a lineage-restricted manner, we examined all possible promoters for each lineage-restricted gene from the Cufflink output, which include all possible isoforms for each gene, with either annotated promoters or de novo assembled promoters. To reliably identify the promoters that are actively utilized in each lineage, we only selected promoters that have a transcription start site within 1kb of a DNase I hypersensitive site in the same cell type. To examine the epigenetic landscape at promoters across H1 and the H1 differentiated cells, we only examined genes that have a single active promoter, or genes that contain multiple active promoters but within 1kb of each other.

Identification of High CG, Medium CG, and Low CG Promoters

For the promoter of every RefSeq gene, we examined the sequence immediate around the TSS (+/− 500bp) and counted the number of CG dinucleotides per 100bp. An examination of the distribution of the CG density for all promoters revealed two distinct promoter populations (Figure S3A). Therefore, we empirically chose thresholds that separate the promoters into three classes: high CG class (CG density > = 4CGs per 100bp), low CG class (CG density < 2CGs per 100bp), and medium CG class for those with CG density in between.

For analyses of DNA methylation involving low CG promoters, we used a relatively smaller window (+/− 200bp) as we found that many low-CG promoters show hypo-methylation only for very short regions around TSSs.

Identification of Enhancers

We recently developed a random-forest based algorithm, RFECS (Random Forest for Enhancer Identification using Chromatin States), for the purpose of enhancer prediction (Rajagopal et al., 2013). Briefly, the enhancer identification procedure was as follows. We used histone modification profiles at p300 binding sites in H1 to train a random-forest for enhancer prediction. We constructed the forest using a selected set of histone modifications that provide largely non-redundant information, including H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, and H3K9me3. The enrichment of these modifications was determined and used in 100 bp bins from −1 to +1kb along the p300-binding sites or selected non-p300 background sites to train the RFECS classifier. Using this classifier, we predicted enhancers genome wide in the 6 cell-types. In order to compare enhancers across cell-types, it is preferable to have enhancer predictions with the same level of confidence. To determine the appropriate cutoff for multiple cell-types, we calculated a false discovery rate by randomly permuting 100 bp bins across the genome and computing the ratio of enhancers predicted in permuted data versus enhancers predicted in observed data for various cutoffs of voting percentages. We selected a

FDR cutoff of 2.5% for the 6 cell-types analyzed in this study. Enhancers predicted in each cell type were then examined, and those that directly overlap with H3K4me3 peaks (identified by MACS [Zhang et al., 2008]) or within 2.5kb of the H3K4me3 peak centers or TSSs of RefSeq genes were removed. The resulting lists of enhancers were used for downstream analyses. Additionally, for a joint non-redundant list of enhancers existing in all cell types, we merged enhancers from all cell types that are located close to each other (<2 kb) and used the midpoint as the center of the "new" enhancer.

Identification of Lineage-Restricted Enhancers

To identify lineage restricted enhancers, we used H3K27ac (K27ac) as a marker for active enhancers (Rada-Iglesias et al., 2011; Creyghton et al., 2010), and employed a strategy described previously based on the Shannon entropy to compute a cell type-specificity index to each enhancer (Barrera et al., 2008; Schug et al., 2005; Shen et al., 2012). We first selected lineage restricted enhancer candidates in each one of the 6 cell types by the following criteria: the enhancer is significantly enriched by K27ac (Z-score transformed RPKM > = 0.5), and such significant enrichment is observed in no more than two additional cell types. Then for each enhancer in the candidate enhancer set of a particular cell type, we computed the entropy to assess the cell type specificity of its K27ac enrichment. Specifically, for each enhancer, we defined its relative K27ac in a cell type i as Ri = Ei / Σ E, where Ei is the normalized K27ac RPKM value in the cell type i; Σ E is the sum of K27ac RPKM values in all cell types; N is the total number of cell types. Then the entropy score for this element across cell types can be defined as H = $-1^{*}\Sigma$ Ri * logRi($1 \leq i \leq N$), where the value of H ranges between 0 to log2(N). An entropy score close to zero indicates the activity of this enhancer is highly cell type-specific, while an entropy score close to log2(N) indicates that this enhancer is active/silenced ubiquitously. We selected those with entropy less than 1 as lineage restricted enhancers. Inactive enhancers were defined as those that show no H3K27ac enrichment (Z-score transformed RPKM $\leq$ 0).

GREAT Analysis for Lineage-Restricted Enhancers

The functional enrichment for genes that are near lineage specific enhancers was analyzed using the GREAT tool (McLean et al., 2010). The following parameters were used: Basal plus extension, proximal 5kb upstream and 1 kb downstream, plus distal up to 100 kb or 50 kb.

Motif Analysis for Lineage-Restricted Enhancers

To identify a broad set of de novo motifs, we used two programs: (i) HOMER (Heinz et al., 2010), and (ii) our own method, Epigram. Epigram works by calculating the enrichment of 9-mers within the peaks in comparison to two backgrounds: (i) the entire genome, and (ii) the shuffled peak sequences. Those with fold-enrichment above 1.5 are classified as enriched, and the topmost is taken as a seed for motif construction. All enriched 9-mers having exactly 1 or 2 mismatches to the seed are grouped with the seed to create an initial ungapped alignment. A motif is produced from the alignment by weighting each of the 9-mers' contributions by their corresponding enrichment scores. The motif's enrichment is calculated by comparing the motif to all the non-neutral 9-mers, those with relative fold enrichment greater than 1.5 in the peaks, or the backgrounds.

The alignment is then widened by adding each enriched 9-mer that can be aligned, with an offset of one and with at most one mismatch, to at least one 9-mer already in the alignment. During the widening step an expansion is rejected if it causes the enrichment score to decrease. The initial alignment may be extended for a maximum for four bases in either direction, to produce a motif of at most 17 bases.

Once a motif is completed the process repeats, using only the set of enriched 9-mers that have not yet been included in a previous motif. The process continues until all enriched 9-mers have been used to initialize a motif or have been included in a motif.

The final set of motifs is scanned against the peaks, and the shuffled peak, sequences and best score for each motif, in each set of sequences, is recorded. These scores are then used to calculate the area under the curve (AUC), which represents how well the motif can differentiate the two sets of sequences. Motif with AUCs < 0.55 are excluded from subsequent analysis. Finally, to allow the motifs to be ranked alongside the motifs produced by HOMER, which ranks by *P*-value, the hypergeometric distribution is used to calculate enrichment *P*-values for each of the motifs.

To identify which of the de novo motifs matched known motifs, Tomtom was run with an *E*-value cut-off of < 0.05 (Gupta et al., 2007; Tanaka et al., 2011). When running Tomtom, a library of known motif was constructed from the following four databases: (i) TRANSFAC (Matys et al., 2003), (ii) JASPAR (Bryne et al., 2008), (iii) Uniprobe (Robasky and Bulyk, 2011) and (iv) hPDI (Xie et al., 2010). JASPAR motifs with IDs starting with "LM" or "PF" were excluded, as the interacting partners of these motifs are unknown. Any motifs that did not significantly match any known motifs were ignored from subsequent analysis.

To produce the final motif table the motif lists went through two rounds of manual curation. During the first round of curation, the origin and the quality of the known motifs were checked. If the known motif was not produced using a mammalian transcription factor, or if it was a mammalian transcription factor not present in humans, then the motif was ignored. Furthermore, if a de novo motif significantly matched multiple motifs from the same family of transcription factors then it was summarized as being the family. This approach is necessary as both the de novo and known motif making processes are noisy and slight variations that make a de novo motif more similar to a single member of a family cannot be taken as strong evidence that a particular family member is enriched. This curation step resulted in the motif table that is shown in Table S3. Next, this motif table was further refined to establish the final list shown in Figure S4D. During the final curation, several factors were taken into consideration. When the identified motif matches common regulatory elements (e.g., the CREB-element) that are recognized by many transcriptions factors, it is impossible to accurately identify transcription factors that are likely responsible for the enrichment; in such cases the motif was excluded. Motifs that were identified in multiple sets of enhancers were excluded, as they might well reflect

ubiquitously expressed transcription factors. The reason for deciding to include, or exclude, each motif from the final table is given in Table S3.

Zebrafish Reporter Assays

Identified human enhancers were evaluated for their ability to regulate tissue specific expression using a Tol2 transposon-mediated zebrafish transgenesis approach as previously described (Rada-Iglesias et al., 2011). Selected human enhancers were PCR amplified and subcloned either upstream or downstream of the hsp70 promoter/eGFP cassette in the pT2HE vector (gift from Dr. Joanna Wysocka). Tol2 transposase was in vitro transcribed from a linearized pCS2+ vector using the mMessage mMachine Sp6 kit (Ambion), and mixed with reporter plasmid DNA containing corresponding human enhancer sequences. This cocktail was subsequently injected in one-cell-stage AB wild-type zebrafish embryos. Zebrafish imaging was performed on these injected embryos with a Leica M205 FA fluorescent stereoscope, and eGFP expression patterns were monitored at 24-28 hr postfertilization (hpf). Expression patterns were considered as representative for a given enhancer if expression was consistently displayed by at least 30%–40% embryos (at least 100 surviving zebrafish embryos were analyzed per enhancer). Candidate enhancers were further validated by a second set of injections where at least 50 additional surviving injected zebrafish embryos were analyzed for recapitulation of expression patterns.

Correlation Analysis for Epigenetic Mark Enrichment at Promoters and Gene Expression

For each promoter, we averaged normalized RPKM values for each histone modification within −2 to +2 kb of the TSS and percentage of CG methylation within −200 to +200bp of the TSS, and computed the Pearson correlation coefficient between the enrichment of each epigenetic mark and the corresponding RNA-seq expression value across the 6 cell-types. In this analysis, we used a small window for DNA methylation as we found that many low-CG promoters show hypo-methylation only for very short regions around TSSs. We also performed a similar correlation analysis by permuting gene expression values across the 6 cell-types for each promoter 10 times, to assess the ''random'' background correlation levels.

Correlations Analysis for Epigenetic Mark Enrichment at Enhancers and Expression for Potential Enhancer-Targeted Genes

To identify the potential targeted genes for each enhancer, we averaged the normalized H3K27ac RPKM values within −1 to +1kb of lineage-specific enhancers, and examined the correlation between the H3K27ac level at enhancers and the expression level of genes within 200kb of this enhancer across the 6 cell types. As a control, we permuted genes within 200kb 100 times to create random enhancer-gene pairs and computed the correlation similarly. We then selected all enhancer-gene pairs that showed a significant positive correlation of H3K27ac with gene expression compared to randomly permutated pairs, based on the Wilcoxon signed-rank test. Next, for each of these predicted enhancer-gene pairs, we computed the correlations between the enrichment of each epigenetic mark (averaged RPKM values for histone marks within −1 to +1 kb, and CG methylation within −500 to +500bp) at lineage-restricted enhancers and the expression level of the associated genes. A similar analysis was done for those enhancer-gene pairs using randomly permutated gene sets to assess the ''random'' correlation levels.

Identification of DMVs

As DMVs are larger than CpG islands and typical LMRs and UMRs, we employed a window-based approach to identify DMVs. To identify each DMVs in a cell type, the genome was first divided in 1kb bins and the DNA methylation level was averaged within each bin. Then a sliding 5kb window (with 1kb step) was used to identify regions that have an averaged methylation level less than 0.15 in a 5kb window. Continuous regions resulting from this analysis were then merged to form DMVs.

PMD and FMR/LMR/UMR Analysis

PMDs were identified as described previously (Lister et al., 2009). Only those that are no less than 50 kb were included. For the FMR/LMR/UMR analysis, the methylomes of H1 and the H1-derived cell types were segmented using the Hidden Markov Model parameters and methodology described in (Stadler et al., 2011). Unmethylated regions (UMRs), lowly methylated regions (LMRs), and fully methylated regions (FMRs) were identified accordingly.

CpG Island Cluster Analysis

A list of CpG islands (CGIs) was downloaded from the UCSC Genome Browser or from a previous study (Irizarry et al., 2009). CGI clusters were identified as following: a CGI cluster contains at least three CGIs, with the maximal distance between two adjacent CGIs no more than 5kb. Genes with TSSs overlapping CGI cluster loci were subsequently identified.

Identification of Histone Modification Peaks and Domains

For H3K4me3, MACS (Zhang et al., 2008) was used to identify its peaks using the default setting. For H3K27me3 and H3K9me3 which typically show broad enrichment, HOMER (Heinz et al., 2010) was used to identify their enriched regions, using the findPeaks script with the module designed for broad histone modification peak calling. For even larger H3K27me3 and H3K9me3 domains (such as those that are preferentially present in MSCs and IMR90) (Hawkins et al., 2010), a different approach was employed using a sliding window based approach modified from a method described previously (Hon et al., 2012). Briefly, the genome was divided into 10kb bins and the RPKM values of K27me3 and K9me3 enrichment were averaged within each bin. A bin was counted as an ''enriched bin'' (+1) if the average value is no less than 0.95 quantile of all bins in the genome. If a bin shows an average FPKM value below 0.8 quantile of all bins in the genome, a penalty was given (−1). Next, for K27me3, the percentage of enriched bins within a 10-bin sliding window was counted (including penalty if any) and was compared to the expected percentage of enriched bins in the entire genome. A binomial test was then employed to compute the p-value to assess the likelihood that the observed number of enriched bins occurs by chance. When the p-value is less than 0.001, the sliding window is identified as an enriched window. To avoid including boundaries that are only marginally enriched by K27me3, only the 2.5kb-7.5kb region within the 0-10kb window was labeled as the final domain

regions. Such analysis was repeated for each sliding window with a step of 10kb. The contiguous domain regions were then merged. The K9me3 domains were identified similarly but with a larger sliding window (25bins), as we observed that the K9me3 domains are typically larger than the K27me3 domains.

Enrichment of DMV Genes for Oncogenes and Tumor Suppressor Genes

A total of 496 oncogenes and 874 tumor suppressor genes were obtained from the Cancer Gene database, Memorial Sloan-Kettering Cancer Center (http://cbio.mskcc.org/CancerGenes/Select.action). The significance of the overlap between DMV genes and oncogenes or DMV genes and tumor suppressor genes was assessed by the hypergeometric test.

Data Analyses

Details of bioinformatic analyses can be found in Supplemental Information.

SUPPLEMENTAL REFERENCES

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. (2004). UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. *32*(Databaseissue), D115–D119.

Bryne, J.C., Valen, E., Tang, M.H., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B., and Sandelin, A. (2008). JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res. *36*(Databaseissue), D102–D106.

Bu, D., Yu, K., Sun, S., Xie, C., Skogerbø, G., Miao, R., Xiao, H., Liao, Q., Luo, H., Zhao, G., et al. (2012). NONCODE v3.0: integrative annotation of long noncoding RNAs. Nucleic Acids Res. *40*(Databaseissue), D210–D215.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. *25*, 1915–1927.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789.

Fang, F., Turcan, S., Rimner, A., Kaufman, A., Giri, D., Morris, L.G., Shen, R., Seshan, V., Mo, Q., Heguy, A., et al. (2011). Breast cancer methylomes establish an epigenomic foundation for metastasis. Sci. Transl. Med. *3*, 75ra25.

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007). Quantifying similarity between motifs. Genome Biol. *8*, R24.

Guttman, M., and Rinn, J.L. (2012). Modular regulatory principles of large non-coding RNAs. Nature *482*, 339–346.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Hon, G.C., Hawkins, R.D., Caballero, O.L., Lo, C., Lister, R., Pelizzola, M., Valsesia, A., Ye, Z., Kuan, S., Edsall, L.E., et al. (2012). Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. Genome Res. *22*, 246–258.

Hsu, F., Kent, W.J., Clawson, H., Kuhn, R.M., Diekhans, M., and Haussler, D. (2006). The UCSC Known Genes. Bioinformatics *22*, 1036–1046.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. (2002). The Ensembl genome database project. Nucleic Acids Res. *30*, 38–41.

Kong, L., Zhang, Y., Ye, Z.Q., Liu, X.Q., Zhao, S.Q., Wei, L., and Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. *35*(WebServer issue), W345–W349.

Matys, V., Fricke, E., Geffers, R., Gössling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., et al. (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. *31*, 374–378.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. *28*, 495–501.

Pruitt, K.D., Tatusova, T., Brown, G.R., and Maglott, D.R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. *40*(Databaseissue), D130–D135.

Robasky, K., and Bulyk, M.L. (2011). UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. Nucleic Acids Res. *39*(Databaseissue), D124–D128.

Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V., and Ren, B. (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116–120.

Tanaka, E., Bailey, T., Grant, C.E., Noble, W.S., and Keich, U. (2011). Improved similarity scores for comparing motifs. Bioinformatics *27*, 1603–1609.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562–578.

Weisenberger, D.J., Siegmund, K.D., Campan, M., Young, J., Long, T.I., Faasse, M.A., Kang, G.H., Widschwendter, M., Weener, D., Buchanan, D., et al. (2006). CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. Nat. Genet. *38*, 787–793.

Xie, Z., Hu, S., Blackshaw, S., Zhu, H., and Qian, J. (2010). hPDI: a database of experimental human protein-DNA interactions. Bioinformatics *26*, 287–289.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

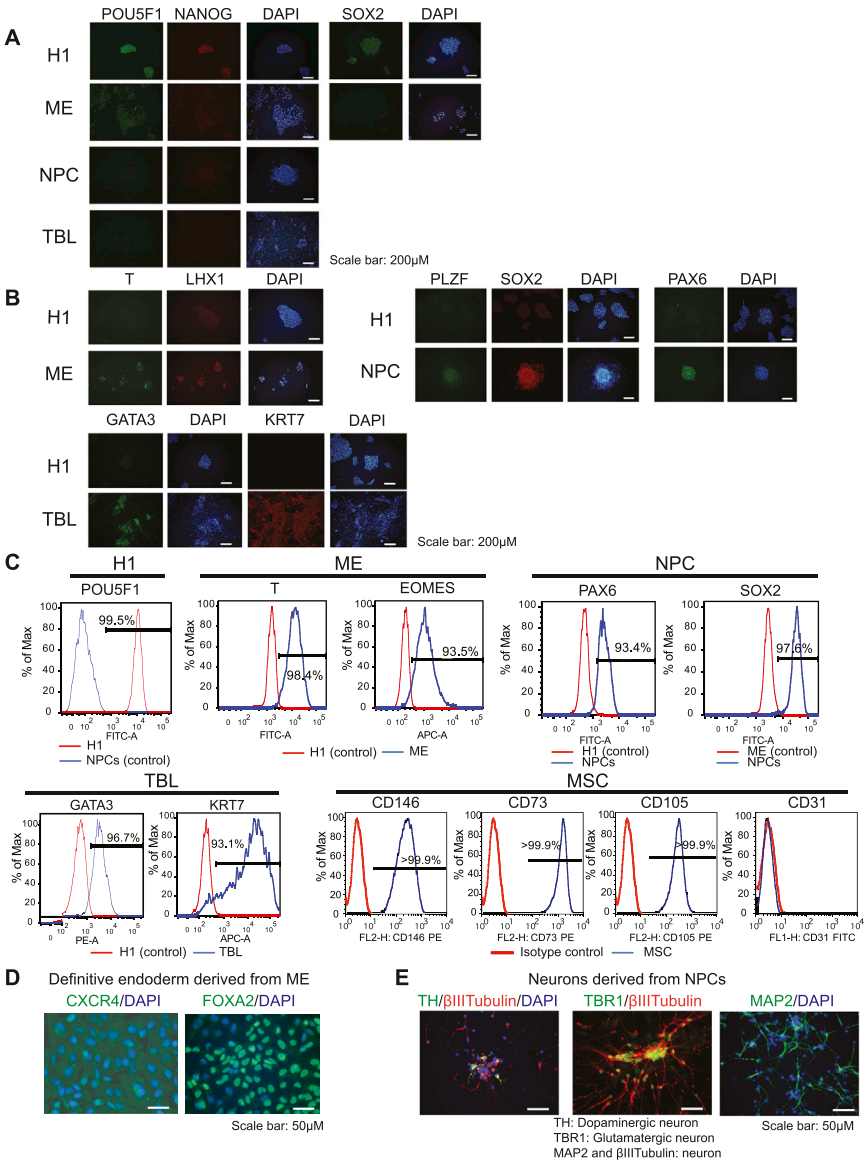


Figure S1. Differentiation of hESCs (H1) to ME, NPCs, TBL, and MSCs, Related to Figure 1

(A and B) H1 and the H1-derived cells were stained with various (A) hESC-specific markers or (B) lineage specific markers using immunofluorescence.

(C) H1 and the H1-derived cells were analyzed by FACS for various lineage specific markers. The numbers indicate the percentages of cells stained positive for the specific marker compared to the negative controls (differentiation efficiency). For H1, ME, NPC, and TBL, a second cell type in which the marker is not expressed was used as a negative control as indicated. For MSCs, an isotype control was used. CD31 is also a negative control marker for MSCs.

(D) Definitive endoderm cells differentiated from H1-derived ME were stained for the definitive endoderm markers CXCR4 and FOXA2 using immunofluorescence.

(E) Neurons differentiated from H1-derived NPCs were stained for various neuronal markers TH, TBR1, MAP2, and βIII Tubulin using immunofluorescence.

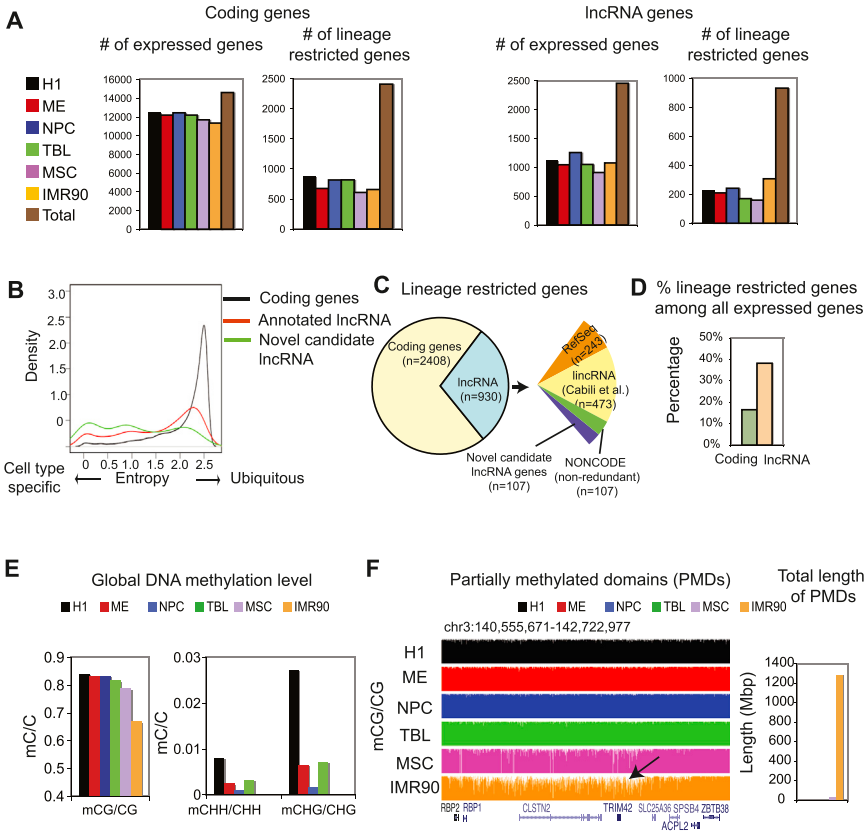


Figure S2. Identification of Lineage-Restricted Transcripts in H1 and H1-Derived Cells, Related to Figure 2

(A) Bar graphs showing the numbers of expressed and lineage-restricted genes identified in each cell type used in this study for coding genes (RefSeq) (left) or long non-coding RNA genes (lncRNA) (right). The total numbers of non-redundant genes are also shown.

(B) Distributions of the entropy of expression values (FPKM) across H1, H1 derivatives and IMR90 are shown as probability density curves for coding genes (black), known lncRNA genes (red) and novel lncRNA genes (green).

(C) A pie chart showing the distribution of lineage-restricted coding genes and lncRNA genes. For known lncRNA genes, annotations from RefSeq (Pruitt et al., 2012), a long intergenic non-coding RNA (lincRNA) gene database (Cabili et al., 2011) and NONCODE (Bu et al., 2012) were used.

(D) A bar chart showing the number of lineage-restricted coding or lncRNA genes as a percentage of total number of expressed coding or lncRNA genes (expressed in at least one cell type).

(E) The average levels of DNA methylation are shown as bar graphs for both CG (left) and non-CG (right) sites in each of the 6 cell types.

(F) CG methylation levels are shown for a region where a partially methylated domain (PMD) is present in IMR90 (left, black arrow). The total lengths of PMDs present in each of the 6 cell types are shown as bar graphs (right).

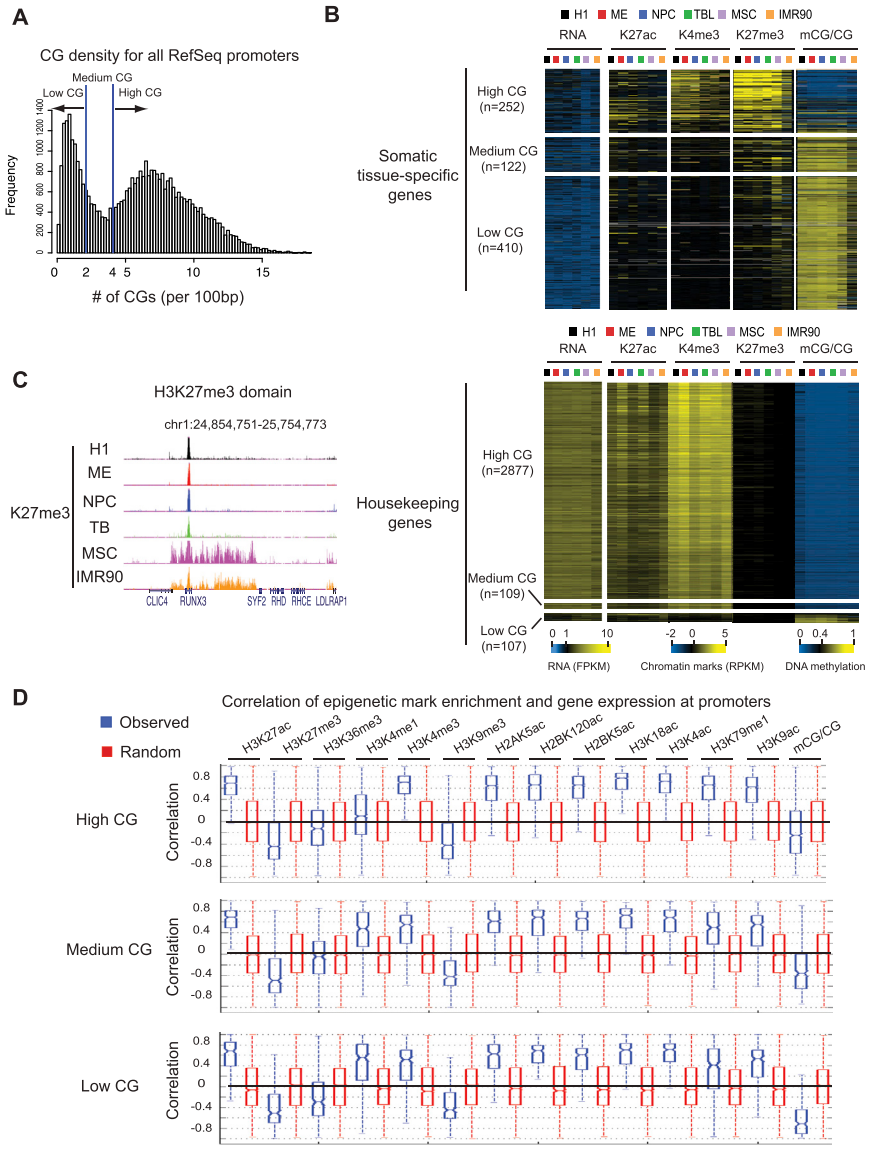


Figure S3. Epigenetic Regulation of Lineage-Restricted Promoters, Related to Figure 3

(A) A histogram showing the distribution of CG densities at all RefSeq promoters. The thresholds used to separate high-, medium-, and low-CG promoters are also shown.

(B) RNA, K27ac, K4me3, K27me3 and DNA methylation levels are shown for promoters of either somatic tissue-specific genes (top) or housekeeping genes (bottom), identified previously using a panel of gene expression data from 18 human tissues (Zhu et al., 2008). Genes were classified based on their promoter CG density.

(C) The levels of H3K27me3 enrichment for all 6 cell types are shown at a locus where the expanded H3K27me3 domains are observed in MSCs and IMR90.

(D) The levels of various epigenetic mark enrichment at the promoters were compared to gene expression across 6 cell types, and the Pearson correlation coefficients are shown as boxplots. The analyses were done before (blue) and after (red) permutation of gene expression values across cell types, for high-CG (top), medium-CG (middle) and low-CG promoters (bottom).

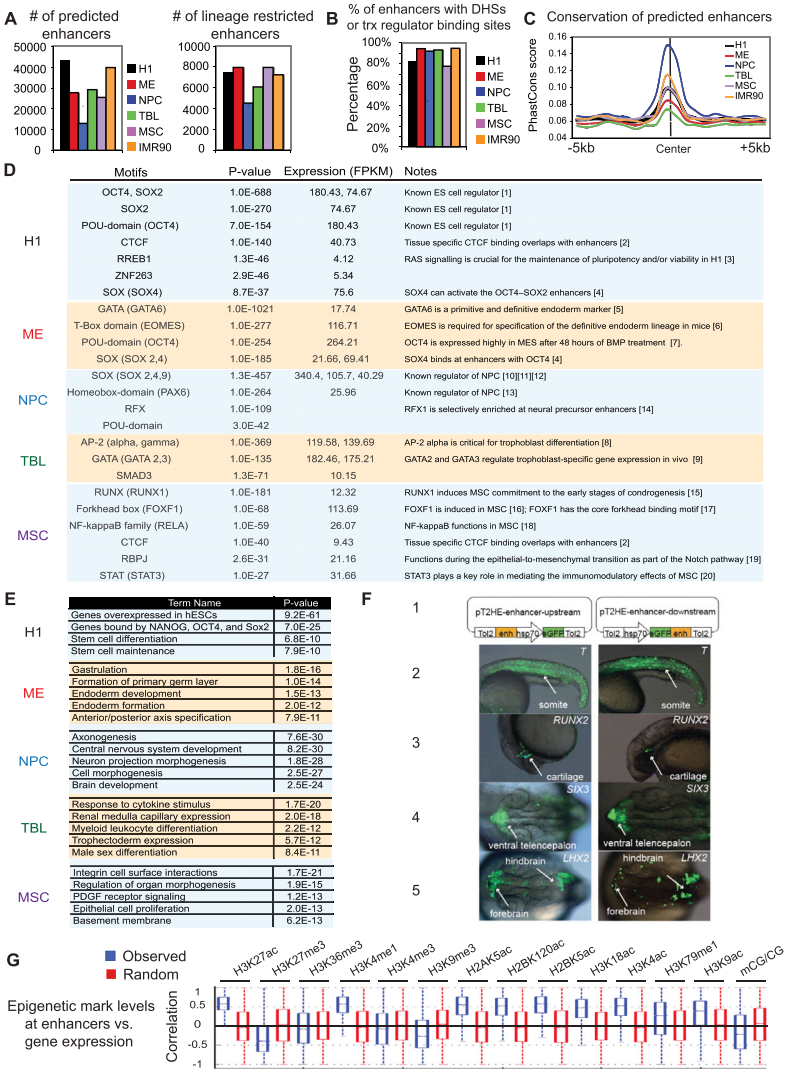


Figure S4. Epigenetic Regulation of Lineage-Restricted Enhancers, Related to Figure 4

(A) Bar graphs showing the numbers of all (left) and lineage-restricted (right) enhancers predicted in each cell type used in this study.

(B) Bar graphs showing the percentages of enhancers that overlap with DNase I hypersensitive sites (DHSs) in each cell type. For H1, a combined set of genomic loci for DHSs, the binding sites of p300, NANOG, SOX2 and OCT4 were used.

(C) The PhastCons scores are shown for +/− 5kb regions around the center of lineage specific enhancers in each of the 6 cell types.

(D) Motifs identified in lineage-restricted enhancers using existing and our in-house developed motif identification algorithms are shown for each cell type, together with their P-values, expression levels of corresponding transcription factors in that lineage, and notes and references supporting the roles of these transcription factors in that lineage (Extended Experimental Procedures). The full list of references used in the table is included in Table S3.

(E) Gene ontology terms are shown for genes near lineage-restricted enhancers as analyzed by the GREAT tool (McLean et al., 2010).

(F) Predicted enhancer elements have developmental enhancer activity in vivo. 1). Schematic representation of the enhancer reporter vectors used in the zebrafish reporter assay. Selected human enhancers (orange) were PCR amplified and subcloned either upstream (left) or downstream (right) of the hsp70 promoter/eGFP cassette (green) in the pT2HE vector. The resulting constructs and transposase RNA were co-injected into one-cell stage wild-type AB zebrafish embryos. 2-5) Fluorescence and brightfield microscopy imaging of individual injected embryos were merged to show representative GFP expression for each reporter construct in 24 hr post fertilization (hpf) zebrafish embryos. 2) The identified human *T* enhancer regulates GFP expression in the somites (white arrow). Lateral view of body/tail. 3) The identified human *RUNX2* enhancer regulates GFP expression in the cartilage (white arrow). Lateral view of body/head. 4) The identified human *SIX3* enhancer regulates GFP expression in the ventral telencephalon (white arrow). Dorsal view of body/head. 5) The identified human *LHX2* enhancer regulates GFP expression in the forebrain and hindbrain (white arrow). Dorsal view of body/head. Sequences of enhancers were retrieved from the following genomic loci: *T*, chr6:166,506,973-166,507,828; *RUNX2*, chr6:45,513,329-45,514,246; *SIX3*, chr2:45,053,109-45,054,078; *LHX2*, chr9:125,835,444-125,836,299.

(G) The levels of various epigenetic marks at enhancers were compared to the expression level of predicted enhancer-targeted genes, and Pearson correlation coefficients are shown before (blue) and after (red) permuting all genes within 200kb of enhancers. Associated enhancer-promoter pairs were identified when the level of H3K27ac at an enhancer shows strong correlation with the expression of a gene within 200kb of enhancers (Extended Experimental Procedures).

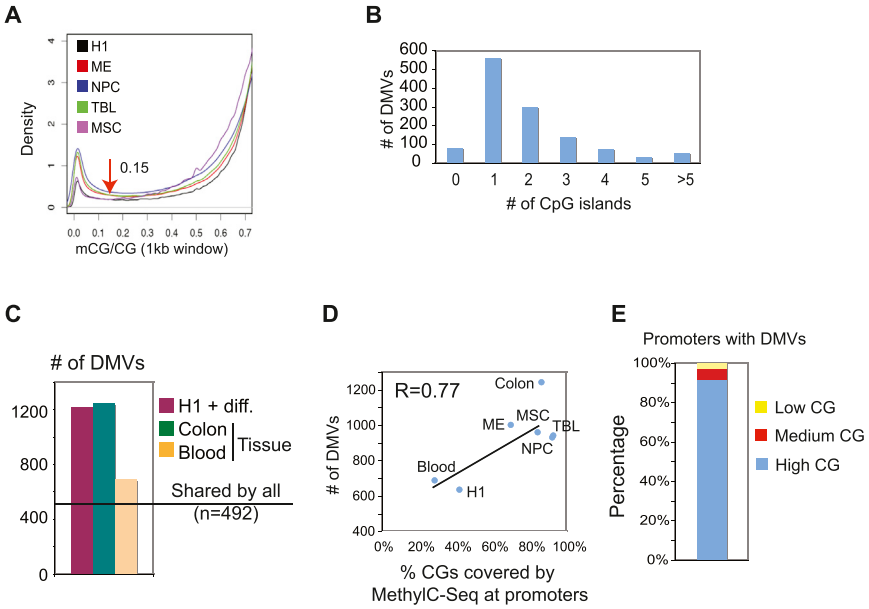


Figure S5. Identification and Characterization of DMVs, Related to Figure 5

(A) Distributions of DNA methylation levels across the genome in H1 and the H1-derived cells (1kb window) are shown as density plots. The threshold used to identify hypo-methylated windows in the genome is indicated (red arrow).

(B) The numbers of CpG islands in DMVs are shown as a histogram plot.

(C) Bar graphs showing the numbers of DMVs identified in H1, the H1-derived cells, as well as those identified in two human tissues colon (Berman et al., 2012) and blood (Li et al., 2010). The number of DMVs shared by all is indicated by the horizontal line.

(D) A scatter plot showing the relationship of numbers of DMVs identified and the percentages of CGs at promoters (TSS +/− 2.5kb) covered by high quality MethylC-Seq data (as defined by Berman et al. [2012] for colon or at least 10 reads for all other cell types) in each cell type. A Pearson correlation coefficient is also shown.

(E) A bar graph showing the percentages of promoters in DMVs that are in the classes of high CG, medium CG and low CG.

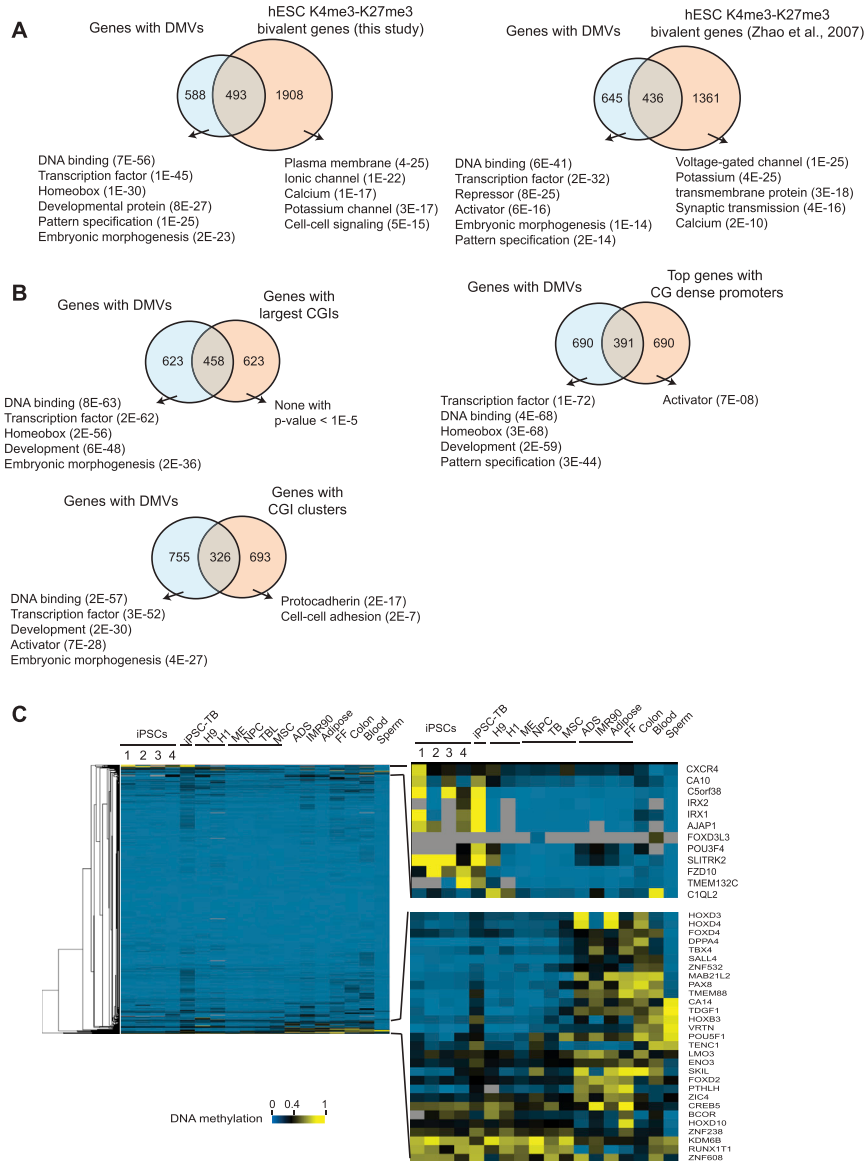


Figure S6. Regulation of DMVs in Cell Differentiation, Related to Figure 6

(A) Venn diagrams showing the overlaps between coding genes in DMVs (n = 1,081) and bivalent genes (marked by both H3K4me3 and H3K27me3) in hESCs as defined in this study (n = 2,401) (left) or (Zhao et al., 2007) (n = 1,797 after gene symbol conversion) (right). A similar result was also obtained using bivalent genes defined in Pan et al., (2007) (data not shown). GO terms for genes that are unique to each gene set are shown with p-values indicated in parentheses.

(B) Venn diagrams showing the overlaps between coding genes in DMVs (n = 1,081) and various gene classes, including genes with the longest CpG islands (n = 1,081) (top left), or genes with the highest promoter CG densities (n = 1,081) (top right), or genes with CpG island clusters (n = 1,019) (bottom). GO terms for genes that are unique to each gene set are shown with p-values indicated in parentheses.

(C) A heatmap showing DNA methylation levels for all promoters in DMVs (average of +1/-1 kb of TSSs) in 17 cell types for which base-resolution maps are available. The following cell types are shown (from left to right): 1-3, foreskin fibroblast (FF)-derived iPSC lines (19.11,6.9,19.7); 4, adipose-derived stem (ADS) cell iPSCs; FF iPSC-derived trophoblasts (iPSC-TB); H9 hESC line; H1, ME, NPC, TBL, MSC, adipose-derived stem (ADS) cells; IMR90, ADS-derived adipocytes; foreskin fibroblasts (FF); the colon tissue (Berman et al., 2012); peripheral blood mononuclear cells (PBMC) (Li et al., 2010); and sperm (Hammoud et al., 2009). Methylome data for cells other than H1 and the H1-derived cells were obtained from studies as indicated or Lister et al., 2011 otherwise. Notably, several genes in DMVs are hypermethylated only in iPSCs as reported previously (Lister et al., 2011), indicating aberrant epigenetic reprogramming at these genes.

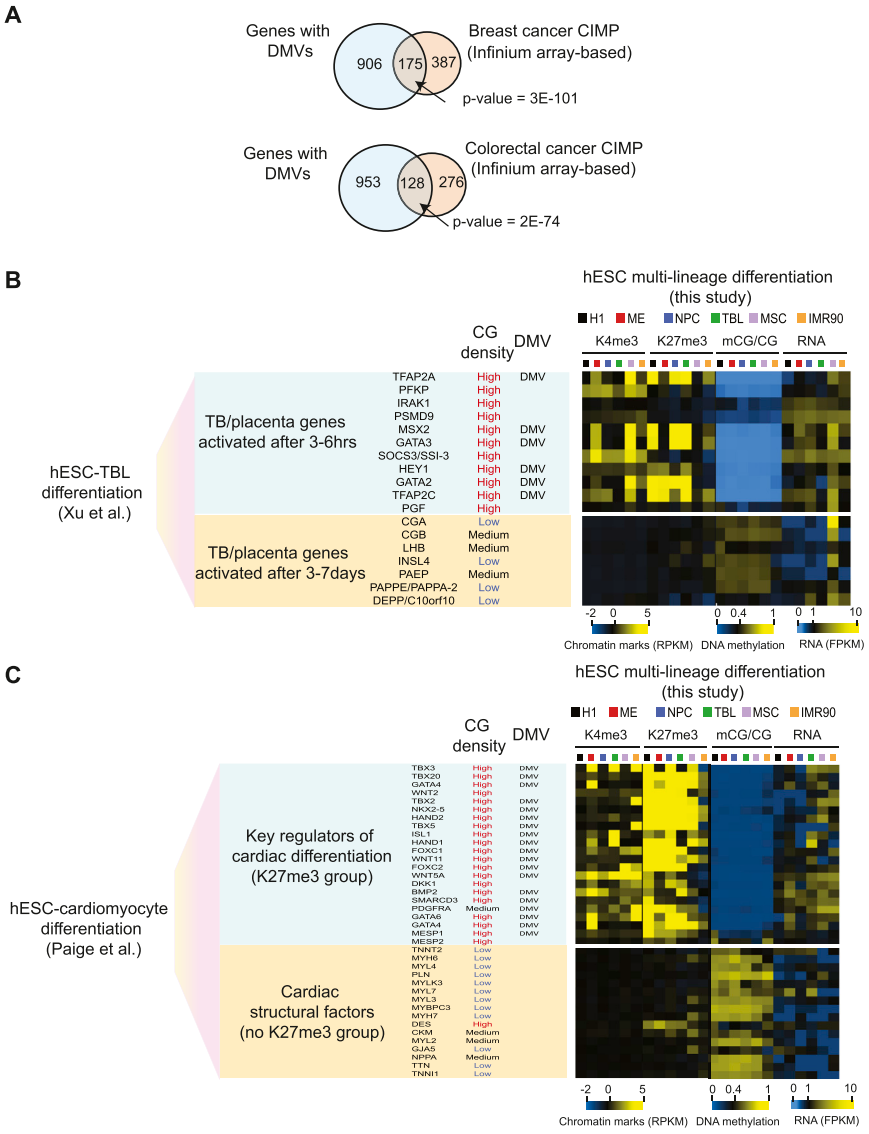


Figure S7. Analyses of Published hESC-TBL and hESC-Cardiomyocyte Differentiation Data, Related to Figure 7

(A) The overlap of genes between those in DMVs and those that are associated with CpG island methylator phenotype (CIMP) in breast cancer (top) or colorectal cancer (bottom) identified using the Illumina Infinium arrays (Fang et al., 2011; Weisenberger et al., 2006).

(B) Trophoblast (TB) and placenta related genes that were activated early (3-6hrs) or late (3-7 days) during hESC-TBL differentiation (Xu et al., 2002) (left) are shown for their promoter CG density and the presence of DMVs (middle). A heatmap also shows the levels of H3K4me3, H3K27me3, DNA methylation at promoters and the RNA level for these genes in H1, the H1-derived cells, and IMR90 (right).

(C) Key regulator genes of cardiac differentiation (left), many of which are known to be enriched for H3K27me3 before activation during cardiac differentiation (K27me3 group) (Paige et al., 2012), are shown for their promoter CG density and the presence of DMVs (middle). A heatmap also shows the levels of H3K4me3, H3K27me3, DNA methylation at promoters and the RNA level for these genes in H1, the H1-derived cells, and IMR90 (right). A similar plot is shown for cardiac structural factor genes, many of which are known to lack H3K27me3 during cardiac differentiation (no K27me3 group) (Paige et al., 2012). Notably, the key regulators of cardiac development are generally activated early during differentiation, as evidenced by the advanced enrichment of active chromatin marks, compared to cardiac structural factors (Paige et al., 2012). A cardiac regulator gene (*MEF2C*) was removed from the list as it contains multiple promoters that show different categories of CG density.