Epigenome Analysis of Multilineage Differentiation of Human Embryonic Stem Cells


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 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 mesoderm 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 each 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.epigenomeweb.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 2A). 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.
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 genes. Of particular interest, the expression of HERV-H/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 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., 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.
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 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 ≥ 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
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 hypo-methylation 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 will be 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**


Received: September 29, 2012

Revised: January 7, 2013

Accepted: April 1, 2013

Published: May 9, 2013

**REFERENCES**


Note Added in Proof