

Transcriptional and Epigenetic Dynamics during Specification of Human Embryonic Stem Cells

Casey A. Gifford,^{1,2,3,9} Michael J. Ziller,^{1,2,3,9} Hongcang Gu,¹ Cole Trapnell,^{1,3} Julie Donaghey,^{1,2,3} Alexander Tsankov,^{1,2,3} Alex K. Shalek,⁴ David R. Kelley,^{1,3} Alexander A. Shishkin,¹ Robbyn Issner,¹ Xiaolan Zhang,¹ Michael Coyne,¹ Jennifer L. Fostel,¹ Laurie Holmes,¹ Jim Meldrim,¹ Mitchell Guttman,¹ Charles Epstein,¹ Hongkun Park,⁴ Oliver Kohlbacher,⁵ John Rinn,^{1,3,6} Andreas Gnirke,¹ Eric S. Lander,^{1,7} Bradley E. Bernstein,^{1,8} and Alexander Meissner^{1,2,3,*}

¹Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

²Harvard Stem Cell Institute, Cambridge, MA 02138, USA

³Department of Stem Cell and Regenerative Biology

⁴Department of Chemistry

Harvard University, Cambridge, MA 02138, USA

⁵Applied Bioinformatics, Center for Bioinformatics and Quantitative Biology Center, University of Tübingen, 72076 Tübingen, Germany

⁶Beth Israel Deaconess Hospital, Boston, MA 02115, USA

⁷Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

⁸Howard Hughes Medical Institute and Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA

⁹These authors contributed equally to this work

*Correspondence: alexander_meissner@harvard.edu

<http://dx.doi.org/10.1016/j.cell.2013.04.037>

SUMMARY

Differentiation of human embryonic stem cells (hESCs) provides a unique opportunity to study the regulatory mechanisms that facilitate cellular transitions in a human context. To that end, we performed comprehensive transcriptional and epigenetic profiling of populations derived through directed differentiation of hESCs representing each of the three embryonic germ layers. Integration of whole-genome bisulfite sequencing, chromatin immunoprecipitation sequencing, and RNA sequencing reveals unique events associated with specification toward each lineage. Lineage-specific dynamic alterations in DNA methylation and H3K4me1 are evident at putative distal regulatory elements that are frequently bound by pluripotency factors in the undifferentiated hESCs. In addition, we identified germ-layer-specific H3K27me3 enrichment at sites exhibiting high DNA methylation in the undifferentiated state. A better understanding of these initial specification events will facilitate identification of deficiencies in current approaches, leading to more faithful differentiation strategies as well as providing insights into the rewiring of human regulatory programs during cellular transitions.

INTRODUCTION

Coordinated changes to the epigenome are essential for lineage specification and maintenance of cellular identity. DNA methyl-

ation (DNAm) and certain histone modifications critically contribute to epigenetic maintenance of chromatin structures and gene expression programs (Zhou et al., 2011; Smith and Meissner, 2013). Genetic deletion of histone methyltransferases and the catalytically active DNA methyltransferases are embryonic or postnatal lethal (Li, 2002), providing evidence for their essential role in proper execution of developmental programs.

Several groups have reported genome-wide maps of chromatin and DNA methylation in pluripotent and differentiated cell types. From these efforts, a global picture of the architecture and regulatory dynamics is beginning to emerge. For example, active promoters generally contain modifications such as H3K4me3 and H3K27ac, whereas active enhancers are generally enriched for H3K4me1 and H3K27ac (Heintzman et al., 2009; Creyghton et al., 2010; Ernst et al., 2011; Rada-Iglesias et al., 2011). Repressed loci exhibit enrichment for H3K27me3, H3K9me2/3, DNAm, or a combination of the latter two modifications. The enrichment of repressive histone modifications, such as H3K27me3, which is initiated at CpG islands (CGI), is considered a facultative state of repression, whereas DNAm is generally considered a more stable form of epigenetic silencing (Smith and Meissner, 2013).

Recent studies have reported dynamics that suggest epigenetic priming such as the appearance of euchromatic histone modifications prior to gene activation during *in vitro* T cell differentiation (Zhang et al., 2012) and cardiac differentiation (Wamstad et al., 2012). These results are reminiscent of changes that occur during the early stages of reprogramming toward the induced pluripotent state (Koche et al., 2011) and highlight possible similarities between differentiation and dedifferentiation. In parallel to these advances, whole-genome bisulfite sequencing (WGBS) has been used to map DNAm genome

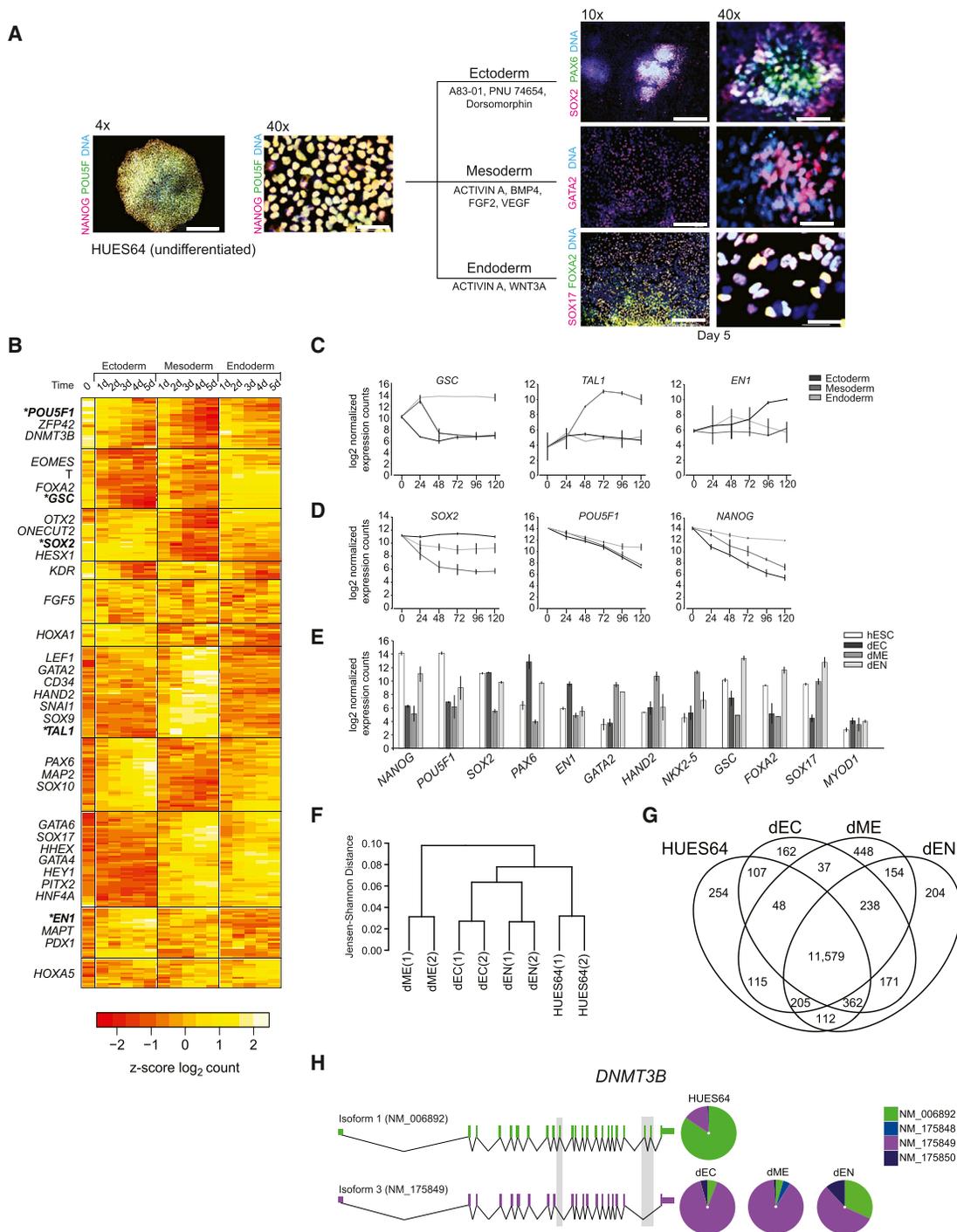


Figure 1. Generation and Characterization of hESCs and hESC-Derived Cell Types

(A) Left: Low (4x) and high (40x) magnification overlaid immunofluorescent images of the undifferentiated human embryonic stem cell (hESC) line HUES64 stained with OCT4 (POU5F1) and NANOG antibodies. Right: Established directed (two-dimensional) differentiation conditions were used to generate representative populations of the three embryonic germ layers: hESC-derived ectoderm, hESC-derived mesoderm, and hESC-derived endoderm. Cells were fixed and stained after 5 days of differentiation with the indicated antibodies. Representative overlaid images at low (10x) and high (40x) magnification are shown. DNA was stained with Hoechst 33342 in all images. Scale bars, 200 μm (4x), 100 μm (10x), and 30 μm (40x).

(B) NanoString nCounter expression data (Z score \log_2 expression value of two biological replicates) for a time course of in vitro differentiation using the conditions shown in (A). 541 genes were profiled, and 268 changing by more than 0.5 are displayed. Selected lineage-specific genes are shown on the left for each category that was identified based on hierarchical clustering (see Table S1 for all). The average \log_2 expression value of two biological replicates is displayed. Error bars represent 1 SD.

(legend continued on next page)

wide. Examination of WGBS data from murine ESCs (mESCs) and neural progenitor cells highlighted lowly methylated regions (LMRs) at distal sites that frequently overlap with DNase I hypersensitive sites (HS) and/or displayed an enhancer signature defined by H3K4me1 and p300 enrichment (Stadler et al., 2011).

Studying the role of epigenetic modifications in the dynamic rewiring of human transcriptional programs in vivo is complicated by numerous technical and ethical limitations. However, models for in vitro differentiation of human embryonic stem cells (hESCs) offer a unique opportunity to explore and characterize critical events that prepare, guide and possibly regulate cell fate decisions (Kriks et al., 2011; Chen et al., 2012; Wei et al., 2012).

To dissect the early transcriptional and epigenetic events during hESC specification, we used two-dimensional, directed differentiation of hESCs to produce early representative populations from the three germ layers, namely ectoderm, mesoderm, and endoderm (Hay et al., 2008; Evseenko et al., 2010; Lee et al., 2010), followed by fluorescence-activated cell sorting (FACS) to enrich for the desired differentiated populations. These three cell types, in addition to undifferentiated hESCs (HUES64), were then subjected to chromatin immunoprecipitation sequencing (ChIP-seq) for six histone marks (H3K4me1, H3K4me3, H3K27me3, H3K27ac, H3K36me3, and H3K9me3), WGBS, and RNA sequencing (RNA-seq). To complement this data, we also performed ChIP-seq for three transcription factors (TFs) (OCT4, SOX2, and NANOG) in the undifferentiated hESCs, as well as ChIP bisulfite sequencing (ChIP-BS-seq) for FOXA2 in the endoderm population. Our data, in combination with the report by Xie et al. (2013), provide a wealth of information, including holistic views of transcriptional and epigenetic dynamics that help further dissect the molecular events during human germ layer specification.

RESULTS

High-Resolution Transcriptional Measurements during Directed Differentiation of hESCs

To better understand the molecular dynamics involved in hESC differentiation, we produced populations representative of each embryonic germ layer, namely ectoderm (Lee et al., 2010), mesoderm (Evseenko et al., 2010), and endoderm (Hay et al., 2008) (see Extended Experimental Procedures). We chose the male hESC line HUES64, a National Institutes of Health (NIH)-approved line that readily differentiates into each of the three

germ layers. These hESCs can be differentiated into a neuroectoderm-like progenitor population positive for SOX2 and PAX6 by inhibition of TGF β , Wingless/integrin1 (WNT), and bone morphogenetic protein (BMP) signaling (Figure 1A, top). Alternatively, canonical mesoderm markers, such as GATA2 (Figure 1A, middle), can be induced using ACTIVIN A, BMP4, VEGF, and FGF2 treatment. Lastly, differentiation toward a definitive endoderm-like fate, positive for markers such as SOX17 and FOXA2 (Figure 1A, bottom), is induced using ACTIVIN A and WNT3A.

We began by measuring the expression of 541 selected genes, including many developmental transcription factors and lineage markers (Bock et al., 2011), at 24 hr intervals during differentiation toward each respective germ layer. We found that 268 of these genes exhibit expression changes (Z score \log_2 expression) during the first 5 days of differentiation (Figure 1B). Genes such as *EOMES*, *T*, *FOXA2*, and *GSC* are upregulated at 24 hr of mesoderm and endoderm induction, but not ectoderm differentiation (Figures 1B and 1C and Table S1 available online). *GSC* expression decreases within 48 hr of differentiation in the mesoderm-like population, whereas the expression level is maintained in the endoderm population (Figures 1B and 1C). *EOMES* and *FOXA2* expression is also maintained in the endoderm population accompanied by upregulation of *GATA6*, *SOX17*, and *HHEX* (Figure 1B). After transient upregulation of mesendodermal markers, activation of mesodermal markers such as *GATA2*, *HAND2*, *SOX9*, and *TAL1* is detected specifically in the mesoderm conditions (Figures 1B and 1C and Table S1). None of these markers are detected during early ectoderm differentiation, which instead upregulates neural markers such as *PAX6*, *SOX10*, and *EN1* (Figures 1B and 1C and Table S1).

We found that *POU5F1* (OCT4), *NANOG*, and, to some extent, *SOX2* expression is maintained in our endoderm population (Figures 1B and 1D and Table S1). This is consistent with prior studies indicating that OCT4 and NANOG expression is detected during the course of early endoderm differentiation and supports NANOG's suggested role in endoderm specification (Teo et al., 2011). *SOX2* expression is downregulated in mesoderm and—to a lesser degree—in endoderm but is maintained at high levels in the ectoderm population (\log_2 expression 10.9) (Figure 1D and Table S1), whereas *ZFP42* (*REX1*) is similarly downregulated in all three lineages (Figure 1B and Table S1). We confirmed that these populations indeed represent a precursor stage for each respective lineage by inducing them to differentiate further, which resulted in upregulation of genes such as *OLIG2* and

(C) The average \log_2 expression values of two biological replicates of lineage-specific genes highlighted in (B) are shown. Error bars represent 1 SD. If no error is evident, SD < 0.5 \log_2 expression units.

(D) The average \log_2 expression values of two biological replicates of pluripotent genes highlighted in (B) are shown. Error bars represent 1 SD. If no error is evident, SD < 0.5 \log_2 expression units.

(E) NanoString nCounter profiling of FACS-isolated ectoderm (dEC), mesoderm (dME), and endoderm (dEN). Expression levels for *MYOD1* (right) are included as a negative control. The average \log_2 expression value of two biological replicates is shown. Error bars represent 1 SD. If no error is evident, SD < 0.5 \log_2 expression units.

(F) Hierarchical clustering of global gene expression profiles as measured by strand-specific RNA-seq for biological replicates of HUES64 and dEC, dME, and dEN is shown as a dendrogram. Pairwise distances between the replicates were measured using the Jensen-Shannon distance metric.

(G) Venn diagram illustrating unique and overlapping genes with expression (FPKM > 1) in HUES64 and the FACS-isolated directed differentiation conditions are shown.

(H) Differential splicing of *DNMT3B* in response to directed differentiation. Relative expression of isoforms 1 (NM_006892, green) and 3 (NM_175849, purple) as measured by RNA-seq are shown on the right.

See also Figure S1.

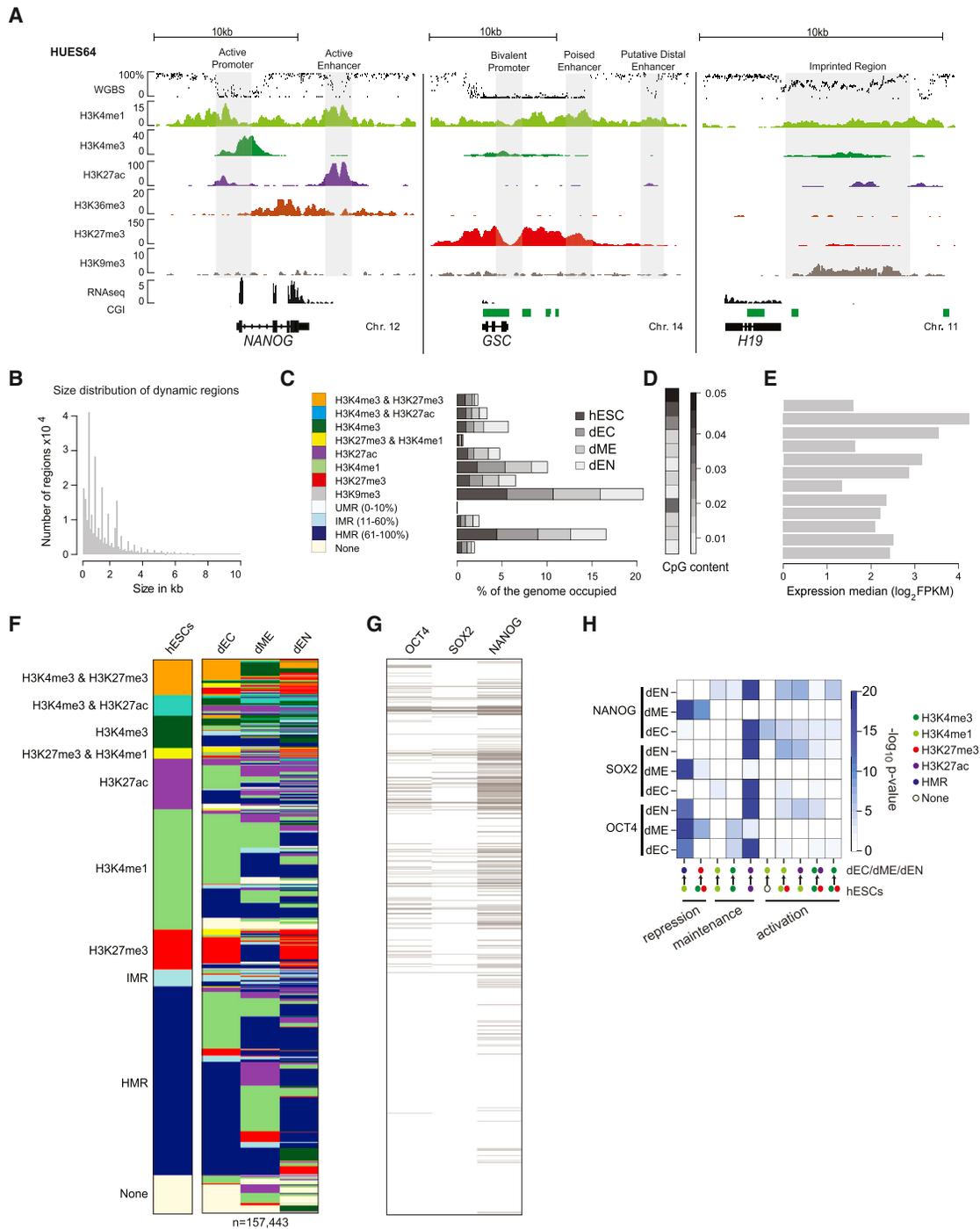


Figure 2. Epigenetic Remodeling Is Lineage Specific during Directed Differentiation

(A) WGBS (% methylation), ChIP-seq (read count normalized to 10 million reads), and RNA-seq (FPKM, read count normalized) for the undifferentiated hESC line HUES64 at three loci: *NANOG* (chr12:7,935,038-7,957,818), *GSC* (chr14:95,230,449-95,250,241), and *H19* (chr11:2,015,282-2,027,359). CGI are indicated in green.

(B) Size distribution of genomic regions enriched for at least one of our six histone modifications in at least one cell type (hESC, dEC, dME, and dEN) and/or classified as UMR or IMR in at least one cell type (n = 297,653).

(C) Definition of epigenetic states used in this study and the genomic space occupied by these in the four cell types under study.

(D) Median CpG content of the genomic regions in distinct epigenetic states defined in (C).

(E) Median expression level of epigenetic states used in this study (C) based on assignment of each region to the nearest RefSeq gene. Median was computed over the states in all four cell types and the corresponding expression profile.

(legend continued on next page)

SST in the ectoderm (Chambers et al., 2012), *TRPV6* in the mesoderm (Evseenko et al., 2010), and *AFP* and *HGF* in the later endoderm populations (Figure S1A) (DeLaForest et al., 2011). Lastly, multidimensional scaling confirmed that, at 24 hr, the mesoderm population is very similar to the endoderm, whereas the ectoderm population has already moved in an alternative direction (Figure S1B). These high-temporal-resolution gene expression signatures suggest that expression programs associated with the three unique cell populations, representing early stages of each germ layer, are established within a similar time frame of hESC differentiation.

Global Transcriptional Dynamics between hESCs and hESC-Derived Cell Types

Based on these results, we selected day 5 as the optimal time point to capture early regulatory events in well-differentiated populations representing all three germ layers. To reduce heterogeneity, we used FACS to enrich populations based on previously reported surface markers (Figure S1C); populations isolated by FACS are referred to as dEC for the ectoderm, dME for the mesoderm, and dEN for the endoderm. Expression analysis of the sorted populations confirms further enrichment for the desired populations (Figure 1E and Table S1).

We next expanded on our selected gene signature profiles by performing strand-specific RNA-seq on poly-A fractions from each day-5-differentiated FACS-isolated population and undifferentiated HUES64 (Table S2). Hierarchical clustering based on the global expression profiles of each cell type reveals that the dME population is the most distantly related cell type and that dEN and dEC are more similar to each other than to dME or hESCs (Figure 1F). This was unexpected, given that the dME and dEN populations are putatively derived through a common mesendoderm precursor stage (Figures 1B and 1C and Table S1), whereas the dEC does not upregulate markers associated with this stage (*EOMES*, *T*, and *GSC*; Table S1). Overall, 14,196 RefSeq-defined coding and noncoding transcripts (~38% of defined transcripts) are expressed (FPKM > 1) in at least one of the populations, with 11,579 (81.6% of the total number of transcripts detected within our cell types) being expressed in all three populations. Examining the overlap of genes expressed (FPKM > 1) in each population reveals that the dME population exhibits expression of the largest number of unique genes ($n = 448$, Figure 1G), such as *RUNX1* (FPKM: 3.4) and *HAND2* (FPKM: 17.8). Examining genes unique to pairs of the differentiated cell types also reveals that dEC and dME have the least in common ($n = 37$; Figure 1G), whereas dEC and dEN have the most number of transcripts in common ($n = 171$; Figure 1G), which is consistent with our clustering analysis. Genes such as *PAX6* (dEC FPKM: 25.9; dEN

FPKM: 5.6) and *NKX6.1* (dEC FPKM: 2.3; dEN FPKM: 3.3), which are each required for both brain (Ericson et al., 1997) and pancreas development (Sander et al., 1997), are expressed in both the dEC and dEN. Canonical markers of embryonic development such as *FOXA2* (FPKM: 12.7) in the dEN and *EN1* (FPKM: 5.8) in the dEC are restricted to their expected germ layers at our early stages (Table S2).

Notably, we also identified 1,296 splicing events (FDR = 5%) as well as alternative promoter usage within our populations (Table S3) (Trapnell et al., 2013). For example, we detected expression of multiple isoforms of *DNMT3B* ($p = 5 \times 10^{-5}$). Expression of *DNMT3B* isoform 1 (NM_006892) was restricted to the undifferentiated hESCs (FPKM: 214.3), whereas the differentiated cell types predominantly express an alternative isoform, *DNMT3B* isoform 3 (NM_175849) (dEC FPKM: 33.9; dME FPKM: 14.2; dEN FPKM: 20.0) (Figure 1H). The presence of this isoform, as well as others, has previously been reported in more advanced stages of embryonic development as well as in normal adult (Robertson et al., 1999) and cancerous tissues (Ostler et al., 2007). Our results suggest that this switch coincides with the exit from the pluripotent state, regardless of the specified lineage. We also identified expression of three *PITX2* isoforms, with differential splicing leading to different isoform expression between the dEN and dME (Table S3). In the chick, *PITX2* is essential for heart looping, and each isoform is responsible for executing distinct functions (Yu et al., 2001). Taken together, this suggests that both transcript levels and isoform expression contribute to cellular identity.

Generation of Comprehensive Reference Epigenome Maps

To gain a more complete picture of the underlying molecular dynamics and to investigate the regulatory events during the specification of the three germ layers, we collected ~12 million cells of the respective dEC, dME, and dEN populations, as well as HUES64. All samples were subjected to ChIP-seq (H3K4me1, H3K4me3, H3K27me3, H3K27ac, H3K36me3, and H3K9me3) and WGBS (Figure 2A), producing a total of 28 data sets with more than 12 billion aligned reads (data are publicly available through the NIH Roadmap Epigenomics Project data repositories: <http://www.roadmapepigenomics.org/>; Table S4).

Integrative Analysis of Epigenetic State Transitions

After completing our basic quality control (see Extended Experimental Procedures and Table S4), we focused our analysis on previously identified informative chromatin states associated with various types of regulatory elements (Ernst et al., 2011;

(F) Epigenetic state map of regions enriched for one of four histone modifications in at least one cell type or classified as UMR/IMR in at least one cell type and changing its epigenetic state upon differentiation in at least one cell type ($n = 157,433$). State definitions are listed in (C).

(G) Regions bound by OCT4, SOX2, and NANOG, as determined by ChIP-seq and organized using the chromatin states in (F).

(H) Enrichment of OCT4, SOX2, and NANOG within various classes of dynamic genomic regions changing upon differentiation of hESC, computed relative to all regions exhibiting the particular epigenetic state change in other cell types. Epigenetic dynamics are categorized into three major classes: repression (loss of H3K4me3 or H3K4me1 and acquisition of H3K27me3 or DNAm), maintenance of open chromatin marks (H3K4me3, H3K4me1, and H3K27ac), and activation of previously repressed states.

See also Figure S2.

Rada-Iglesias et al., 2011), including the following specific combinations: H3K4me3+H3K27me3 (bivalent/poised promoter); H3K4me3+H3K27ac (active promoter); H3K4me3 (initiating promoter); H3K27me3+H3K4me1 (poised developmental enhancer); H3K4me1 (poised enhancer); H3K27ac+H3K4me1 (active enhancer); and H3K27me3 (Polycomb repressed); and H3K9me3 (heterochromatin). In addition, we segmented the WGBS data into three DNAm states: highly methylated regions (HMRs: > 60%), intermediately methylated regions (IMRs: 11%–60%), and unmethylated regions (UMRs: 0%–10%). The latter two differ from the highly methylated background of the genome and likely indicate functional importance as previously suggested (Stadler et al., 2011). We next assigned each genomic region to one of the resulting states (Figure S2A, see Extended Experimental Procedures for details).

Across the four cell populations, we identified 297,653 genomic regions spanning 400 base pairs (bp) to 10 kilobases (kb) (Figure 2B) and covering a total of 512,784,035 bp (21.4% of the human genome) (Figure 2C) that were enriched for at least one of the eight chromatin states and/or classified as an IMR or UMR in at least one of the cell types. Of the identified epigenetic states, H3K27me3-enriched regions and HMRs covered the most base pairs (Figure 2C), and the combination of H3K4me3 and H3K27me3 exhibits the highest CpG content (Figure 2D), which was expected given the close association between H3K27me3 and high CpG density (Ku et al., 2008). The majority of epigenetically dynamic regions are not located near promoters (6.8% +2 kb to –500 bp of the TSS promoters; 48.8% > 50 kb upstream of TSS intergenic; 15.1% > 500 bp downstream of TSS intragenic/gene body) (Figure S2B). As expected, regions of open chromatin exhibit the highest median expression value (Figure 2E). However, overall, we find that many (62%–67%) epigenetic remodeling events are not directly linked to transcriptional changes based on the expression of the nearest gene.

The loss of H3K4 methylation (me1 and me3) is commonly associated with a transition to high DNAm (Figure 2F), which is most prominent in the dEN population and genes involved in neural development (i.e., neural tube development, $q = 9.6 \times 10^{-12}$). We identified 4,639 proximal bivalent domains in hESCs and observed that 3,951 (85.1%) of these domains resolve their bivalent state in at least one hESC-derived cell type (Figures 2F and S2C). When we specifically investigated the promoters of TF-encoding genes, we found that 463 of these promoters are in a bivalent state in hESCs, and 400 of them change in at least one differentiated cell type (Figure S2D). The majority transitions to H3K4me3 only or H3K27me3 only in a lineage-specific manner, as shown for *ISL1* (Figure S2E). In dME, H3K4me3 is gained at the *ISL1* locus while H3K27me3 is lost, leading to expression (FPKM: 14.3). The lineage-specific dynamics in this region are interesting given that this gene has known roles in all three germ layers, although at later time points (Pfaff et al., 1996; Ahlgren et al., 1997; Cai et al., 2003). Notably, in contrast to the limited overall association between many epigenomic dynamics and changes in expression, we found that a large proportion of these bivalent TFs (275) change their expression level during the differentiation (Figure S2D).

Pluripotent TF Binding Is Linked to Chromatin Dynamics during Differentiation

To further explore potential regulators of chromatin dynamics during the exit from pluripotency, we performed ChIP-seq for OCT4, NANOG, and SOX2 in HUES64 (Figures S2F and S2G and Table S8). We found that regions bound by all three factors ($n = 1,556$), by SOX2 only ($n = 955$), or by NANOG only ($n = 14,523$) are frequently associated with inter- and intragenic regions (Figures S2H–S2J, top). In contrast, regions bound by OCT4 only ($n = 8,599$) are more frequently associated with promoter regions (Figure S2H). Examination of regions bound by OCT4, NANOG, and SOX2 in hESCs showed that H3K4me1 regions enriched for OCT4 binding sites frequently become HMRs in all three differentiated cell types, whereas NANOG and SOX2 sites are more prone to change to an HMR state in dME (Figure 2I). In general, many regions associated with open chromatin that are bound by NANOG are more likely to retain this state in dEN compared to dME and dEC (Figure 2I). We also found that regions enriched for H3K27ac in hESCs that maintain this state in dEN or dEC are likely to be bound by SOX2 and NANOG. This is in agreement with the reported role of SOX2 during ectoderm development and differentiation (Wang et al., 2012) but also supports our observation that SOX2 expression is maintained in the dEN. Motif enrichment analysis detected the GATA3 motif in regions bound by OCT4 and SOX2 that transition to an active state in dEC. Furthermore, we found that regions bound by OCT4, NANOG, and SOX2 that gain an active mark in dEC are enriched for the motifs PAX9, p63, and STATs (Table S5). Examining epigenetic dynamics at sites of OCT4, NANOG, and SOX2 binding further supports the observation that some pluripotency-associated TFs are also involved in downstream specification.

Gain of DNAm Occurs at Open Chromatin Enriched for TF Motifs

We next utilized the WGBS data that cover ~26 million CpGs (at ≥ 5 coverage) across all four cell types. Hierarchical clustering analysis of the WGBS data, which included human adult liver and hippocampus for comparison, revealed that the pluripotent hESCs and the hESC-derived cell types form a separate cluster arm with respect to the somatic tissues (Figure 3A).

We defined DMRs as exhibiting a significant ($p \leq 0.05$) minimal difference of CpG methylation level of 0.1 among our four cell types. The majority of all DMRs occur at CpG-poor intergenic regions in line with previous reports (Figure 3B, bottom) (Stadler et al., 2011). The dEN exhibits more than twice the number of regions that gain DNAm compared to dEC and dME (Figure 3B, top). Interestingly, only 65 of the total number of DMRs identified are shared between all three populations. However, reaffirming that our populations are depleted of pluripotent cells, this group of DMRs includes the regulatory region of OCT4 (Figure S3A). In line with the small number of shared regions, more than 60% of regions that gain DNAm are lineage specific (Figure 3C) and include loci such as *SMAD3* (dEC), *CTNNA3* (dME), and *FOXA2* (dEN). *FOXA2* has an upstream CGI that exhibits gain of DNAm (Figure 3D), and transcription in dEN is initiated downstream of this DMR at an alternative TSS, suggesting that TSS usage may be regulated, stabilized, or reflected by DNAm (Maunakea et al., 2010).

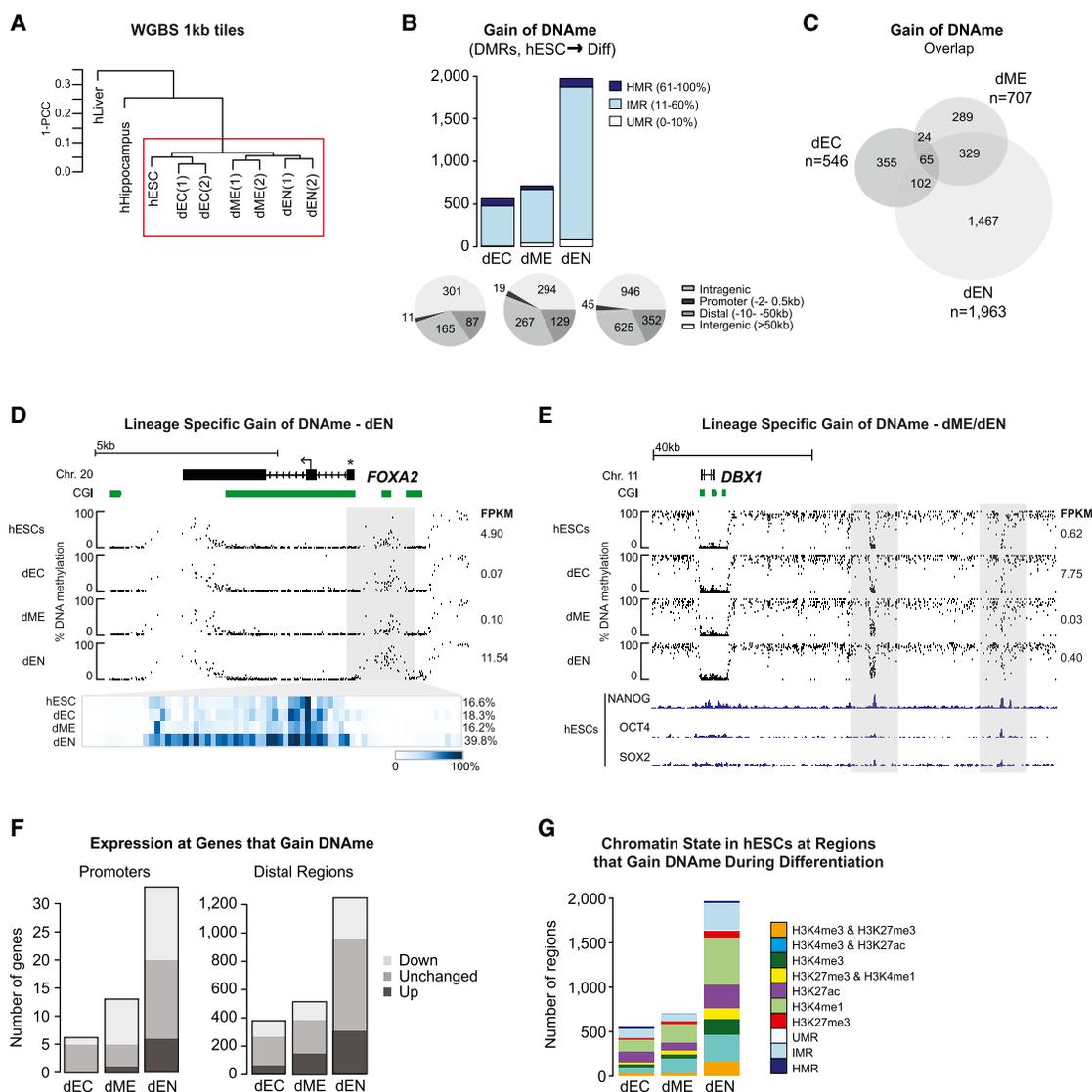


Figure 3. Global DNA Methylation Dynamics—Gain of DNAm

(A) Hierarchical clustering of hESCs, hESC-derived populations (dEC, dME, and dEN), human adult hippocampus, and human adult liver based on mean DNAm levels of 1 kb tiles across the human genome using Pearson Correlation Coefficient (PCC). y axis indicates sample distance in terms of 1 minus PCC. Red box indicates cell types interrogated in this study.

(B) Regions that significantly ($p \leq 0.05$) increase their DNAm levels by at least 0.1 between hESCs and the differentiated cell types. The color code indicates the DNAm state found in hESCs. Bottom: Genomic features associated with DMRs gaining DNAm in each of the differentiated cell types based on RefSeq gene annotation and de novo discovered promoters by RNA-seq.

(C) The overlap of these differentially methylated regions (DMRs) that increase their DNAm level in the three hESC-derived populations.

(D) DNAm levels and RNA-seq expression values of *FOXA2* (chr20:22,559,343-22,571,189) in hESCs and differentiated cell types. The heat map below shows the DNAm values of individual CpGs within the highlighted region. The average DNAm value for the entire highlighted region is shown on the right in red. CGI are shown as green bars. Expression values (FPKM) are displayed on the right. The arrows indicate two known TSSs.

(E) DNAm levels and OCT4, SOX2, and NANOG ChIP-seq at the *DBX1* locus (chr11:20,169,548-20,277,940).

(F) Promoters (left) and distal elements (right) that gain DNAm separated by the changes in FPKM at associated genes.

(G) Chromatin state in hESCs at regions that gain DNAm during differentiation. Regions devoid of any detected chromatin marks are categorized according to their DNA methylation state in hESCs.

See also Figure S3.

We find significant enrichment of various TF motifs as DNAm targets upon differentiation (Table S6), which has some analogy to the gain of methylation observed at myeloid targets in the lymphoid lineage in vivo (Ji et al., 2010; Deaton et al., 2011;

Bock et al., 2012). To extend this observation, we examined the DNAm state at regions bound by SOX2, OCT4, and NANOG in hESCs. For example, two regions 20 kb downstream of *DBX1*, a gene associated with early neural specification, are bound by all

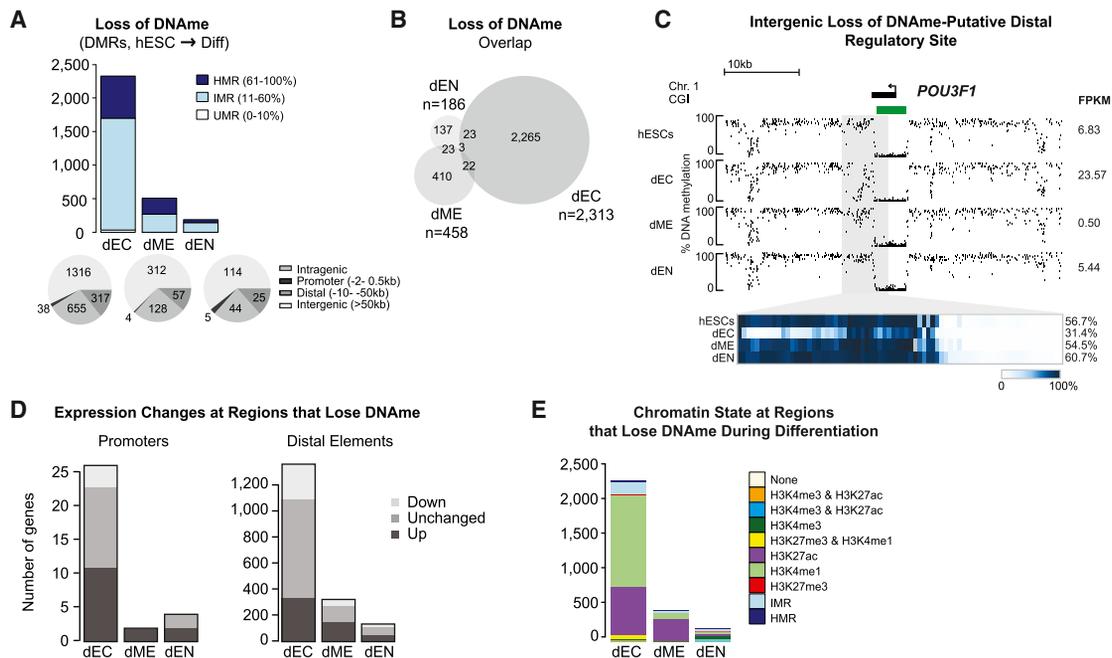


Figure 4. Global DNA Methylation Dynamics—Loss of DNAm

(A) Regions that significantly ($p \leq 0.05$) decrease their DNAm levels by at least 0.1 between hESCs and the differentiated cell types. The color code indicates the DNAm state distribution in the differentiated cell types, revealing that most regions reside in an IMR state after they lost DNAm (left). Genomic features (bottom) associated with DMRs losing DNAm in each of the differentiated cell types based on RefSeq gene annotation and de novo discovered promoters by RNA-seq.

(B) Venn diagram of identified DMRs that decrease their DNA methylation level between the three hESC-derived populations.

(C) DNAm at the *POU3F1* locus (chr1:38,493,152-38,532,618). The heatmap below shows the DNA methylation values of individual CpGs within the gray region. The average DNAm value for the entire highlighted region is shown on the right in red. CGIs are shown as green bars. Expression values (FPKM) are displayed on the right.

(D) Promoters (left) and distal elements (right) that gain DNAm separated by the changes in FPKM at associated genes.

(E) Chromatin state in differentiated cell types at regions that lose DNAm during differentiation.

See also Figure S4.

three TFs and gain DNAm in dME and dEN. In contrast, this region maintains low levels of DNAm in dEC, which has activated transcription of *DBX1* (Figure 3E). We generally find that cobound sites gain DNAm in the dME and dEN, but not dEC (Figure S3B). Further supporting the functional relevance of these dynamics, we find that regions that gain DNAm frequently coincide with DNaseI hypersensitive sites (Figure S3C) (Thurman et al., 2012). Transcriptional silencing was less frequently correlated with gain of DNAm at distal elements than at promoters (Figure 3F).

In examining the chromatin state of regions that gain DNAm during differentiation, we find that most regions exhibited enrichment of one or more histone modifications in hESCs (Figure 3G). These results confirm that, in particular, distal regulatory elements show highly dynamic regulation of DNA methylation during specification.

Loss of DNA Methylation Is Biased toward dEC

Loss of DNAm is asymmetric between the three populations (Figure 4A, top) and occurs in a more lineage-specific fashion than gain (Figure 4B). However, loss also occurs mainly at intergenic regions (Figure 4A, bottom). Notably, the dEC has the most DMRs, and many were associated with neuronal gene categories (for instance: neural tube development, $q = 3.13 \times$

10^{-13}). This includes the ectodermal TF *POU3F1*, which has a bivalent promoter in hESCs, resolves to a H3K4me3-only state (Figure S4A), and exhibits transcriptional activation in dEC (Table S2). Chromatin dynamics and activation at this locus coincide with specific loss of DNAm at a putative regulatory element downstream of the 3' UTR in dEC (Figure 4C). On a global scale, an immediate correspondence between loss of DNAm and expression, such as that observed at *POU3F1*, occurs at about half of the regions (Figure 4D). More than 70% of DMRs that lose DNAm during differentiation are enriched for one of our profiled histone modifications, H3K4me1 or H3K27ac in particular (Figure 4E).

Taken together, our hESC differentiation system reveals several interesting DNAm dynamics, including the lineage-specific silencing of regulatory regions in default or alternative lineages. The asymmetric loss may also explain why our chromatin state analysis revealed fewer regions that gained H3K4me3 in the dEC population.

Gain of H3K27ac Reveals Putative Regulatory Elements

In addition to methylation on H3K4, open chromatin is also demarcated by enrichment of H3K27ac. It has also been suggested that the combination of H3K4me1 and H3K27ac at

distal regions identifies active enhancer elements, whereas H3K4me1 and H3K27me3 correspond to poised enhancer elements (Rada-Iglesias et al., 2011). To extend these observations, we focused specifically on regions that gain H3K27ac during differentiation and found that more than half of the identified regions are HMRs in hESCs (Figure 5A), whereas another large fraction is enriched for H3K4me1 in hESCs (Figure 5A). The majority of regions that gain H3K27ac are intergenic (Figure S5A), as shown for the *RUNX1* locus (Figures 5A and 5B).

We next placed each region into one of three distinct categories (repressed, poised, and open) based on their state in hESCs and subsequently performed gene set enrichment analysis using the GREAT toolbox (Figure 5C) (McLean et al., 2010). This analysis reveals enhancer dynamics in line with the lineage-specific differentiation trajectory for dEC and dME (Figure 5C). In contrast, the dEN population shows an unexpected enrichment for early neuronal genes (e.g., neural tube development, Figure 5C). This observation is consistent with the correlation that we reported between our dEC and dEN RNA-seq data, suggesting that similar networks are induced in the early stages of both our ectoderm and endoderm specification (van Arensbergen et al., 2010).

Moreover, we find strong enrichment of downstream effector genes of the TGF β , VEGF, and BMP pathways in dME, directly reflecting the signaling cascades that were stimulated to induce the respective differentiation. In dEN, we find enrichment of genes involved in WNT/ β -CATENIN and retinoic acid (RA) signaling (Figure 5C). Although we did not use RA, this signaling cascade has previously been implicated in endodermal tissue development, including pharyngeal and pancreatic cell types (Wendling et al., 2000; Oström et al., 2008). Concordantly, we also find high levels of SMAD3 motif enrichment in the repressed dME and dEN, particularly in the poised putative enhancer populations (Figure 5D). Similarly, we observe enrichment of key lineage-specific TF motifs such as the ZIC family proteins in dEC, TBX5 in dME, and SRF in dEN. Interestingly, we also find the FOXA2 motif highly overrepresented in dEN—in which the factor is active, and also dEC, in which the factor is inactive but becomes expressed at a later stage of neural differentiation (Kriks et al., 2011), but not in dME (Figure 5D).

Acquisition of H3K4me1 without Transcriptional Activation Suggests Epigenetic Priming

Many regions that exhibit high DNAm in hESCs and transition to H3K4me1 in one lineage remain HMRs in the two alternative cell types (Figures 2F and 6A). These regions are typically >10 kb from the nearest TSS (Figure 6B). GREAT analysis shows a strong enrichment for categories associated with brain development such as cerebellum morphogenesis in dEC ($q < 10^{-30}$), TGF β pathway targets ($q < 10^{-10}$) in dME, and suppression of EMT in dEN ($q < 0.0001$). To understand whether regions that gain H3K4me1 in our system are associated with somatic identity, we took advantage of published microarray data for 24 human tissues and determined genes upregulated in these tissues with respect to hESCs (termed Tissue Atlas, see [Extended Experimental Procedures](#)). Reaffirming the relevance of our dynamics, we found that regions that gain H3K4me1 in dEC are associated with fetal brain and specific cell types found within the adult brain

(Figure 6C). The dME H3K4me1 pattern was associated with a range of interrogated tissues, such as heart, spinal cord, and stomach, which may be due to heterogeneity of the tissues collected (Figure 6C). The dEN associations were interesting given that, as with the RNA-seq and H3K27ac trends, H3K4me1 was again associated with brain-related categories (Figure 6C).

Overall, less than half of the genes that gain H3K4me1 exhibit immediate transcriptional changes (Figure 6D). *CYP2A6* and *CYP2A7* (Figure 6E) are representative examples that do not show a corresponding change in expression, whereas *LMO2* does (Figure 6F). To investigate these regions in more detail, we carried out motif enrichment analysis and found lineage-specific enrichment of TF motifs near regions that gain H3K4me1. Whereas the FOXA2 motif is enriched in all three cell types, the DBX1 motif is associated with the gain of H3K4me1 in dEC (Figure 6G), which coincides with its transcriptional activation in this cell type (FPKM: 5.36). Conversely, the GLI3, HIC1, and CTF1 motifs are strongly enriched at regions that gain H3K4me1 in dEN (Figure 6G).

To further assess whether this DNAm to H3K4me1 switch acts as a priming event, we differentiated the HUES64 endoderm population for 5 additional days in the presence of BMP4 and FGF2, leading to HNF4 α -positive hepatoblast-like (dHep) cells (Table S2 and Figure S6A). Interestingly, of the motifs enriched in dEN that gain H3K4me1, HIC1, KLF4, and CTF1 (Figure 6G), several of these genes become expressed at the next stage of differentiation (Figure 6H). Lastly, 1,346 of these putatively primed regions are enriched for the active enhancer mark H3K27ac in human liver (Figure 6I).

Loss of DNA Methylation and Acquisition of H3K27me3 at Putative Regulatory Elements

More surprisingly, we observe distal/intergenic regions that switch from high DNAm to H3K27me3 ($n = 3,985$ in dEN) (Figure 7A). This transition frequently occurred within CpG-poor regions, which is distinct from the common CpG-island-centric targets of Polycomb Repressive Complex 2 and H3K27me3 (Figure 7B). This switch is highly lineage specific, and DNAm is generally retained in the alternative two cell types (Figures 7C and 2F).

Motif enrichment analysis, combined with the evaluation of publicly available TF binding site (TFBS) data from the ENCODE project, indicated that many regions exhibiting this transition in dEN were near binding sites of the pioneering factor FOXA2. This TF has putative roles in chromatin decompaction, but its distinct functions and limitations remain somewhat unclear (Li et al., 2012). To investigate this association, we performed ChIP-seq for FOXA2 in the endoderm population. This analysis reveals that FOXA2 binding sites frequently overlap with regions that transition from HMR to H3K27me3 (Figure 7D). We also confirmed that gain of H3K27me3 at dEN FOXA2 binding sites occurs predominantly in dEN and not dEC or dME (Figure 7E). A notable example of this transition can be seen at the *ALB* locus, where H3K27me3 is gained at *AFP* and *AFM*, proximal to FOXA2 binding sites (Figure 7F). Many regions that exhibit this transition are required for later stages of development as with *AFP* and *AFM* in the liver (Figure 7F) or *HBB1* in the dME. As expected, the majority of these regions do not yet exhibit significant changes in expression (Figure S7A).

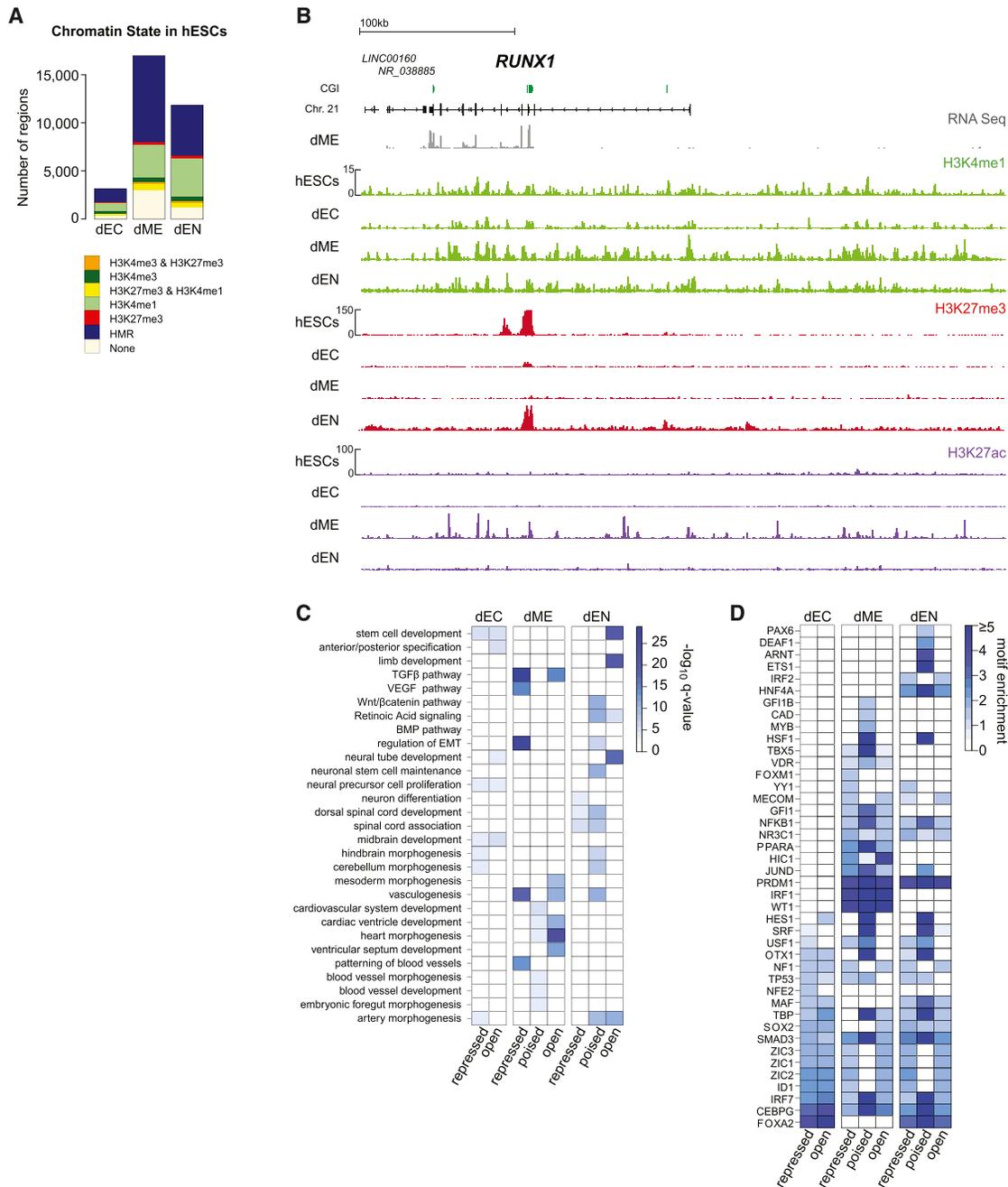


Figure 5. H3K27ac Dynamics at Putative Gene Regulatory Elements

(A) Number of regions and associated epigenetic state distribution in hESCs of regions that are transitioning to H3K27ac in the three populations.

(B) Normalized ChIP-seq tracks (H3K4me1, H3K27me3, and H3K27ac) for the *RUNX1* region (chr21:36,091,108-36,746,447) with corresponding RNA-seq data in dME.

(C) GO categories enriched in regions transitioning to H3K27ac in the cell type indicated on the right compared to hESCs as determined by GREAT analysis. Regions gaining H3K27ac were split up by state of origin in hESC into repressed (none, IMR, HMR, and HK27me3), poised (H3K4me1/H3K27me3), and open (H3K4me3/H3K27me3, H3K4me3, and H3K4me1). Color code indicates multiple testing adjusted q value of category enrichment.

(D) TF motifs enriched in regions changing to H3K27ac in the cell type indicated on the right compared to hESCs. Color code indicates motif enrichment score incorporating total enrichment over background as well as differential expression of the corresponding transcription factor in the respective cell type. Regions were split up by state of origin in hESCs similar to (C). For each region class, the eight highest-ranking motifs are shown.

See also Figure S5.

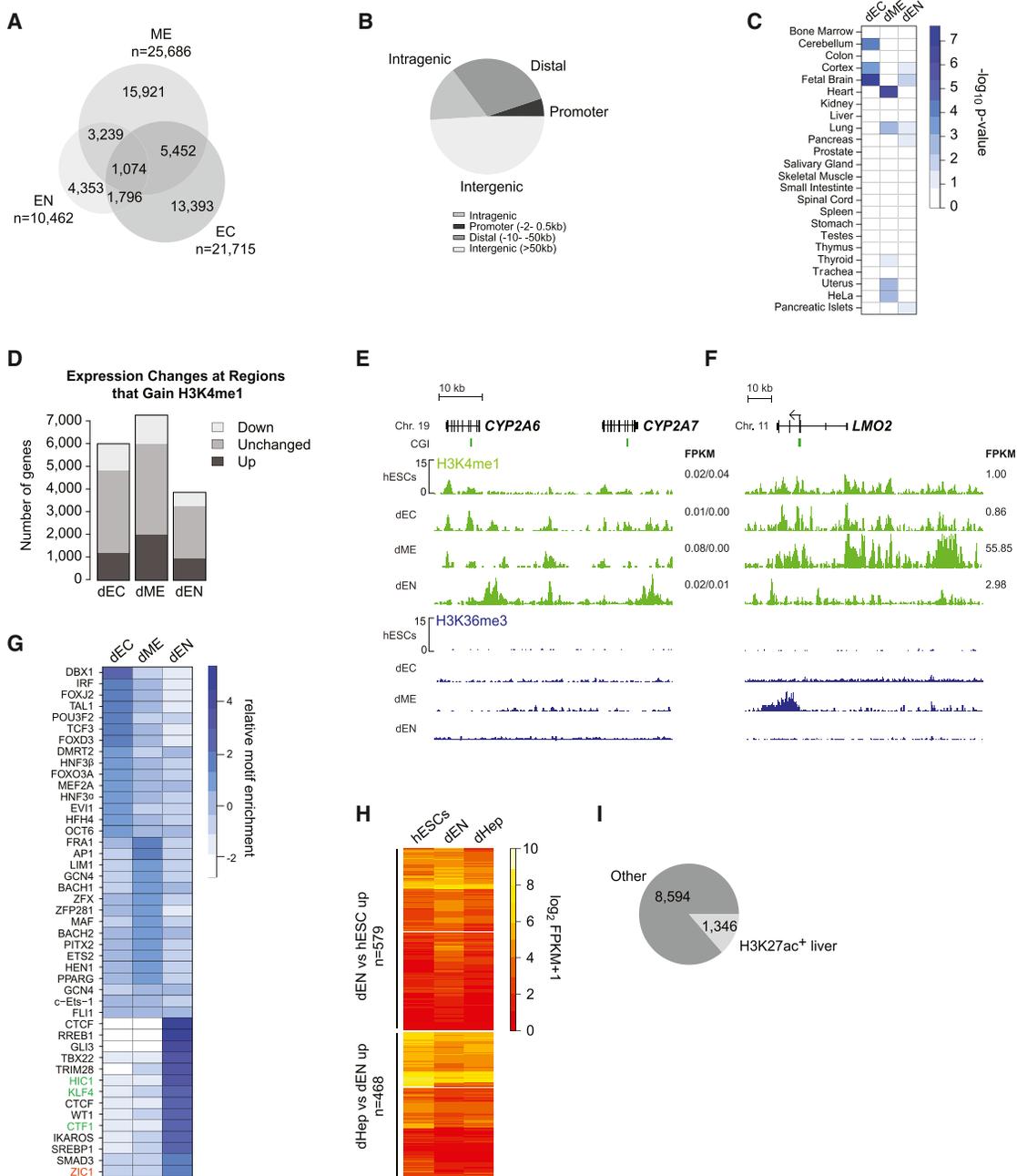


Figure 6. Characterization of H3K4me1 Dynamics at Putative Distal Regulatory Elements

(A) Overlap of regions gaining H3K4me1 in the three differentiated populations relative to hESCs.
 (B) Genomic distribution of all regions gaining H3K4me1 compared to hESCs in at least one of the three differentiated populations.
 (C) Tissue signature enrichment levels of genes assigned to regions specifically gaining H3K4me1 in the differentiated populations indicated on the bottom. For tissue signature definitions, see [Extended Experimental Procedures](#).
 (D) Number and distribution of gene expression changes of genes assigned to regions gaining H3K4me1 in the differentiated populations. Associated genes were classified as either being up/downregulated or unchanged relative to hESCs.
 (E) Normalized ChIP-seq tracks (H3K4me1 and H3K36me3) for the *LMO2* locus (Chr.11:33,865,134-33,977,858). Read counts on y axis are normalized to 10 million reads for each cell type. CGIs are indicated in green.
 (F) Normalized ChIP-seq tracks (H3K4me1 and H3K36me3) for the *CYP2A6/CYP2A7* region (Chr19: 41,347,260-41,395,599). Read counts on y axis are normalized to 10 million reads for each cell type. CGIs are indicated in green.
 (G) Normalized motif enrichment scores for the top 15 motifs enriched in regions specifically transitioning to H3K4me1 in the differentiated cell type indicated on the bottom. Motif highlighted in green corresponds to a TF that is upregulated at the next stage (hepatoblast) of endoderm differentiation, whereas motifs highlighted in red are specifically upregulated in dEN but are downregulated at the dHep stage.

(legend continued on next page)

A previous report found that FOXA1/FOXA2 could bind to regions exhibiting DNAm (Sérandour et al., 2011), which is not a characteristic shared by all TFs (Gifford and Meissner, 2012). Regions bound by these factors subsequently lost DNAm and gained euchromatic histone modifications in our populations. We therefore compared DNAm at FOXA2 binding sites in hESCs to dEN and found a slight reduction specifically in the dEN (Figure 7G). To more directly assess this relationship, we interrogated the DNAm state of regions isolated by FOXA2-ChIP-BS sequencing in dEN (Brinkman et al., 2012). Interestingly, we saw a major depletion of DNAm at sites isolated by FOXA2-ChIP (Figure 7G). To determine whether these regions exhibit transcriptional activation after further differentiation, we examined again our dHep RNA-seq data and found that 50 genes, which were bound by FOXA2 and gained H3K27me3 in dEN, increased their expression at the next stage of differentiation (Figure 7H and Table S2). We also find H3K27ac enrichment at 197 loci in the human liver that had experienced the gain of H3K27me3 in dEN (Figure 7I).

DISCUSSION

Using directed differentiation of hESCs to three distinct, FACS-enriched populations representing early stages of embryonic development, we provide an extensive data set and highlight several insights on the transcriptional and epigenetic dynamics that occur during human *in vitro* lineage specification.

Among other things, we describe two very interesting but distinct lineage-specific dynamics from high DNAm to H3K4me1 or H3K27me3. These transitions occur at many sites that do not significantly change gene expression during our early stages of differentiation. Notably, we made similar observations for H3K4 methylation during the early stages of reprogramming to an iPSC state (Koche et al., 2011), suggesting that this type of epigenetic priming event might be common. At this point, however, it is not clear whether these events reflect a regulatory mechanism to facilitate timely activation upon differentiation or indicate the absence of a critical cofactor necessary for complete transcriptional activation. We also cannot rule out that a subset of the observed priming events are due to heterogeneity in the cell population that are not detected by our RNA-seq. Our observation that high DNAm switches to H3K27me3 enrichment in distal, CpG-poor regions is even more interesting. It remains to be tested whether targeted loss of DNAm at these regions causes a default gain of H3K27me3 in the absence of additional cofactors due to underlying sequence context (Mendenhall et al., 2010) or represents a more active recruitment event and regulatory mechanism. It is also possible that H3K27me3 gain at distal regions is due to genomic conformation changes and reflects H3K27me3 spreading in three dimensions. It was recently reported that the combination of H3K27me3 enrichment and a nearby nucleosome-depleted region creates sites amenable to TF binding (Taberlay et al., 2011). Based on these results, one may speculate that specific TFs, such as FOXA2, exert chromatin decompaction functions resulting in loss of DNAm and leading to gain of H3K27me3, which creates a platform for subsequent binding of other TFs that cannot directly remodel a heterochromatic state but instead function in transcription machinery assembly and transcriptional activation.

In conclusion, examination of our data yields powerful insights on transcriptional and epigenetic dynamics during hESC specification and represents valuable reference maps for many applications, including regenerative biology and the study of human developmental biology.

For full details, see [Extended Experimental Procedures](#).

EXPERIMENTAL PROCEDURES

For full details, see [Extended Experimental Procedures](#).

Cell Culture

Human ESCs were cultured as previously described (Bock et al., 2011). All details regarding the differentiation conditions are in the [Extended Experimental Procedures](#).

Gene Expression Signatures

RNA was hybridized to a custom probe set, processed using the Nanostring prep station, imaged using the Nanostring nCounter, and analyzed as previously described (Bock et al., 2011).

Sequencing Library Construction

For WGBS sequencing libraries, genomic DNA was fragmented using a Covaris S2 sonicator. DNA fragments were cleaned up, end repaired, A-tailed, and ligated with methylated paired-end adapters (purchased from ATDBio).

For ChIP-seq libraries, ChIP-isolated DNA was end repaired, A-tailed, and ligated to barcoded Illumina adaptors, and the library was amplified using PFU Ultra II Hotstart Master Mix (Agilent).

For RNA-seq libraries, polyadenylated RNA was isolated using Oligo dT beads (Invitrogen) and fragmented to 200–600 base pairs and then ligated to RNA adaptors using T4 RNA Ligase (NEB), preserving strand of origin information.

ChIP-BS libraries were constructed as previously described (Brinkman et al., 2012).

Sequencing Data Processing and Analyses

Details regarding the WGBS, ChIP-seq, ChIP-bisulfite seq, and RNA-seq bioinformatics analysis can be found in the [Extended Experimental Procedures](#). All data were aligned to hg19, and the accession numbers for each data set are included in [Table S4](#).

ACCESSION NUMBERS

The accession numbers for all of the sequencing data reported in this paper are listed in [Table S4](#).

SUPPLEMENTAL INFORMATION

Supplemental Information includes [Extended Experimental Procedures](#), seven figures, and eight tables and can be found with this article online at <http://dx.doi.org/10.1016/j.cell.2013.04.037>.

(H) Gene expression levels of genes assigned to regions gaining H3K4me1 specifically in dEN compared to hESC and being upregulated in dEN, but not hepatoblast (top). Gene expression levels of genes being upregulated between dEN and dHep (but not between hESC and dEN) and gaining H3K4me1 in dEN are shown on the bottom.

(I) Fraction of regions changing to H3K4me1 in dEN and being enriched for H3K27ac in human liver ($n = 1,346$). See also [Figure S6](#).

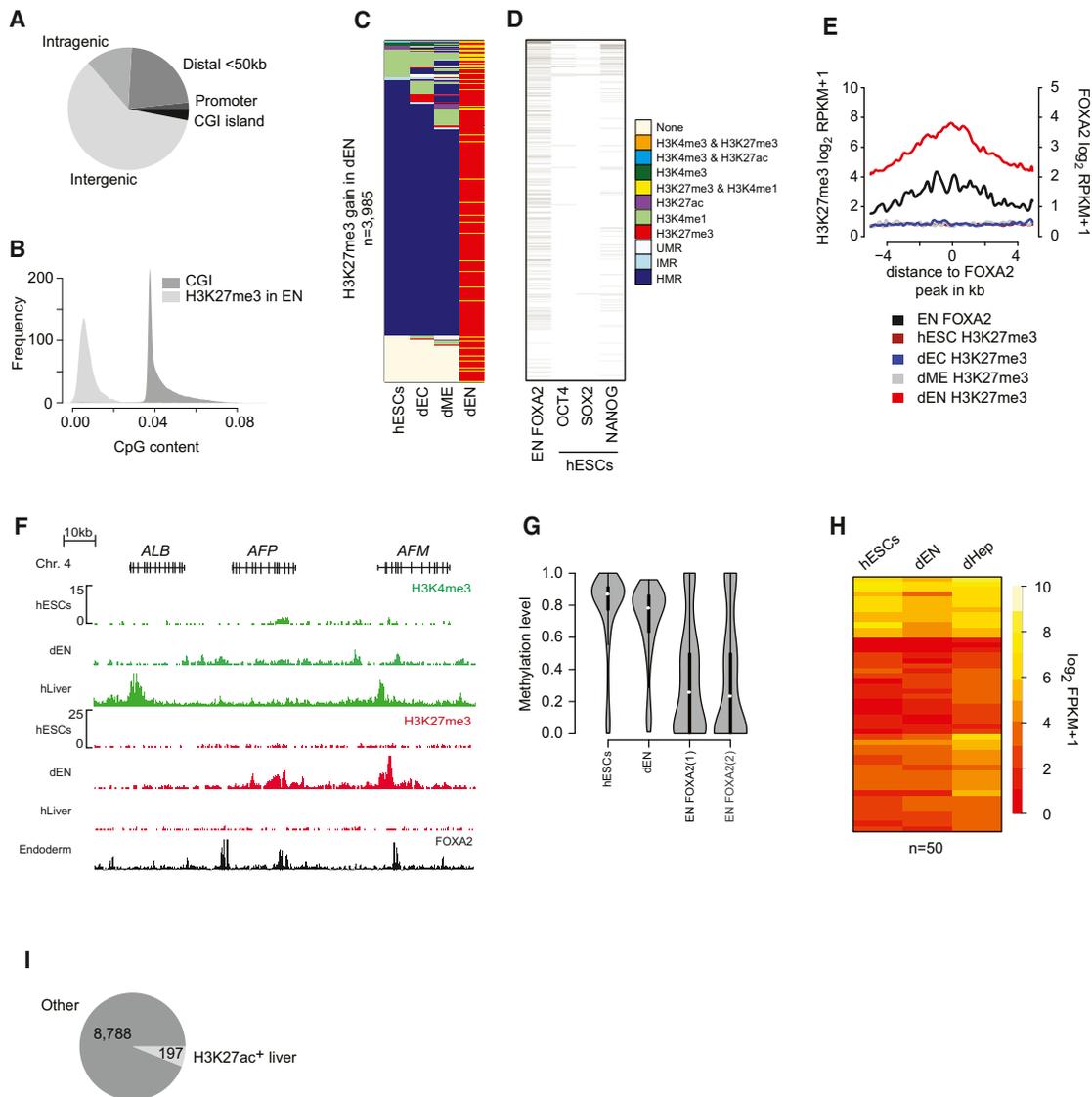


Figure 7. DNase to H3K27me3 Switch and FOXA2 Binding

(A) Distribution of genomic features associated with region gaining H3K27me3 ($n = 22,643$) upon differentiation to any of the three hESC-derived cell types compared to hESC.

(B) CpG content distribution of regions gaining H3K27me3 upon differentiation. For reference, the CpG content distribution of CpG islands is shown.

(C) Epigenetic state distribution in hESC, dEC, and dME of regions that gain H3K27me3 in the dEN population compared to hESC.

(D) Binding profile of FOXA2 in dEN ($n = 357$), OCT4 ($n = 32$), SOX2 ($n = 12$), and NANOG ($n = 124$) in hESC across regions that gain H3K27me3 in dEN upon differentiation.

(E) Composite plot of median normalized tag counts (RPKM) of regions bound by FOXA2 in dEN and gaining H3K27me3 in dEN compared to hESC ($n = 357$).

(F) Normalized H3K27me3 and H3K4me3 ChIP-seq tracks for hESCs, dEN, and human adult liver tissue at the *ALB* locus (chr4:74,257,882-74,377,753). Black bars (bottom) indicate TF binding of OCT4, SOX2, or NANOG in hESCs. Read counts on y axis are normalized to 10 million reads.

(G) Distribution of methylation levels of regions bound by FOXA2 and gaining H3K27me3 in dEN. DNase information is depicted for hESC and dEN WGBS data sets and two biological replicates of FOXA2 CHIP-bisulfite experiments in dEN ($n = 357$).

(H) Gene expression profile of genes upregulated at the hepatoblast stage relative to dEN that are associated with regions bound by FOXA2 and gaining H3K27me3 in dEN ($n = 50$).

(I) Fraction of regions gaining H3K27me3 in dEN and being enriched for H3K27ac in human liver ($n = 197$).

See also Figure S7.

ACKNOWLEDGMENTS

We would like to thank Kendell Clement for support of the WGBS data visualization and Loyal Goff for RNA-seq visualization, as well as Zachary Smith and

Jing Liao for critical reading of the manuscript, and the SCRIB FACS Core for advice regarding FACS analysis. We also thank other members of the Meissner Lab and Epigenomics Platform at the Broad Institute for helpful discussion. A.M. is supported by the Pew Charitable Trusts and is a New York

Stem Cell Foundation (NYSCF) Robertson Investigator. The work was funded by the US National Institutes of Health (NIH) grants (U01ES017155 and P01GM099117) and The New York Stem Cell Foundation.

Received: September 17, 2012

Revised: March 4, 2013

Accepted: April 16, 2013

Published: May 9, 2013

REFERENCES

- Ahlgren, U., Pfaff, S.L., Jessell, T.M., Edlund, T., and Edlund, H. (1997). Independent requirement for ISL1 in formation of pancreatic mesenchyme and islet cells. *Nature* **385**, 257–260.
- Bock, C., Kiskinis, E., Verstappen, G., Gu, H., Boulting, G., Smith, Z.D., Ziller, M., Croft, G.F., Amoroso, M.W., Oakley, D.H., et al. (2011). Reference Maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. *Cell* **144**, 439–452.
- 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.
- Brinkman, A.B., Gu, H., Bartels, S.J., Zhang, Y., Matarese, F., Simmer, F., Marks, H., Bock, C., Gnirke, A., Meissner, A., and Stunnenberg, H.G. (2012). Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. *Genome Res.* **22**, 1128–1138.
- Cai, C.L., Liang, X., Shi, Y., Chu, P.H., Pfaff, S.L., Chen, J., and Evans, S. (2003). Isl1 identifies a cardiac progenitor population that proliferates prior to differentiation and contributes a majority of cells to the heart. *Dev. Cell* **5**, 877–889.
- Chambers, S.M., Qi, Y., Mica, Y., Lee, G., Zhang, X.J., Niu, L., Bisland, J., Cao, L., Stevens, E., Whiting, P., et al. (2012). Combined small-molecule inhibition accelerates developmental timing and converts human pluripotent stem cells into nociceptors. *Nat. Biotechnol.* **30**, 715–720.
- Chen, Y.F., Tseng, C.Y., Wang, H.W., Kuo, H.C., Yang, V.W., and Lee, O.K. (2012). Rapid generation of mature hepatocyte-like cells from human induced pluripotent stem cells by an efficient three-step protocol. *Hepatology* **55**, 1193–1203.
- 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.
- Deaton, A.M., Webb, S., Kerr, A.R., Illingworth, R.S., Guy, J., Andrews, R., and Bird, A. (2011). Cell type-specific DNA methylation at intragenic CpG islands in the immune system. *Genome Res.* **21**, 1074–1086.
- DeLaForest, A., Nagaoka, M., Si-Tayeb, K., Noto, F.K., Konopka, G., Battle, M.A., and Duncan, S.A. (2011). HNF4A is essential for specification of hepatic progenitors from human pluripotent stem cells. *Development* **138**, 4143–4153.
- Ericson, J., Rashbass, P., Schedl, A., Brenner-Morton, S., Kawakami, A., van Heyningen, V., Jessell, T.M., and Briscoe, J. (1997). Pax6 controls progenitor cell identity and neuronal fate in response to graded Shh signaling. *Cell* **90**, 169–180.
- Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shores, 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.
- Evseenko, D., Zhu, Y., Schenke-Layland, K., Kuo, J., Latour, B., Ge, S., Scholes, J., Dravid, G., Li, X., MacLellan, W.R., and Crooks, G.M. (2010). Mapping the first stages of mesoderm commitment during differentiation of human embryonic stem cells. *Proc. Natl. Acad. Sci. USA* **107**, 13742–13747.
- Gifford, C.A., and Meissner, A. (2012). Epigenetic obstacles encountered by transcription factors: reprogramming against all odds. *Curr. Opin. Genet. Dev.* **22**, 409–415.
- Hay, D.C., Fletcher, J., Payne, C., Terrace, J.D., Gallagher, R.C., Snoeys, J., Black, J.R., Wojtacha, D., Samuel, K., Hannoun, Z., et al. (2008). Highly efficient differentiation of hESCs to functional hepatic endoderm requires ActivinA and Wnt3a signaling. *Proc. Natl. Acad. Sci. USA* **105**, 12301–12306.
- Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112.
- Ji, H., Ehrlich, L.I., Seita, J., Murakami, P., Doi, A., Lindau, P., Lee, H., Aryee, M.J., Irizarry, R.A., Kim, K., et al. (2010). Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature* **467**, 338–342.
- Koche, R.P., Smith, Z.D., Adli, M., Gu, H., Ku, M., Gnirke, A., Bernstein, B.E., and Meissner, A. (2011). Reprogramming factor expression initiates widespread targeted chromatin remodeling. *Cell Stem Cell* **8**, 96–105.
- Kriks, S., Shim, J.W., Piao, J., Ganat, Y.M., Wakeman, D.R., Xie, Z., Carrillo-Reid, L., Auyeung, G., Antonacci, C., Buch, A., et al. (2011). Dopamine neurons derived from human ES cells efficiently engraft in animal models of Parkinson's disease. *Nature* **480**, 547–551.
- Ku, M., Koche, R.P., Rheinbay, E., Mendenhall, E.M., Endoh, M., Mikkelsen, T.S., Presser, A., Nusbaum, C., Xie, X., Chi, A.S., et al. (2008). Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet.* **4**, e1000242.
- Lee, G., Chambers, S.M., Tomishima, M.J., and Studer, L. (2010). Derivation of neural crest cells from human pluripotent stem cells. *Nat. Protoc.* **5**, 688–701.
- Li, E. (2002). Chromatin modification and epigenetic reprogramming in mammalian development. *Nat. Rev. Genet.* **3**, 662–673.
- Li, Z., Gadue, P., Chen, K., Jiao, Y., Tuteja, G., Schug, J., Li, W., and Kaestner, K.H. (2012). Foxa2 and H2A.Z mediate nucleosome depletion during embryonic stem cell differentiation. *Cell* **151**, 1608–1616.
- Maunakea, A.K., Nagarajan, R.P., Bilenky, M., Ballinger, T.J., D'Souza, C., Fouse, S.D., Johnson, B.E., Hong, C., Nielsen, C., Zhao, Y., et al. (2010). Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466**, 253–257.
- 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.
- 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.
- Ostler, K.R., Davis, E.M., Payne, S.L., Gosalia, B.B., Expósito-Céspedes, J., Le Beau, M.M., and Godley, L.A. (2007). Cancer cells express aberrant DNMT3B transcripts encoding truncated proteins. *Oncogene* **26**, 5553–5563.
- Oström, M., Löffler, K.A., Edfalk, S., Selander, L., Dahl, U., Ricordi, C., Jeon, J., Correa-Medina, M., Diez, J., and Edlund, H. (2008). Retinoic acid promotes the generation of pancreatic endocrine progenitor cells and their further differentiation into beta-cells. *PLoS ONE* **3**, e2841.
- Pfaff, S.L., Mendelsohn, M., Stewart, C.L., Edlund, T., and Jessell, T.M. (1996). Requirement for LIM homeobox gene Isl1 in motor neuron generation reveals a motor neuron-dependent step in interneuron differentiation. *Cell* **84**, 309–320.
- 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.
- Robertson, K.D., Uzvolgyi, E., Liang, G., Talmadge, C., Sumegi, J., Gonzales, F.A., and Jones, P.A. (1999). The human DNA methyltransferases (DNMTs) 1, 3a and 3b: coordinate mRNA expression in normal tissues and overexpression in tumors. *Nucleic Acids Res.* **27**, 2291–2298.
- Sander, M., Neubüser, A., Kalamaras, J., Ee, H.C., Martin, G.R., and German, M.S. (1997). Genetic analysis reveals that PAX6 is required for normal transcription of pancreatic hormone genes and islet development. *Genes Dev.* **11**, 1662–1673.
- Sérandour, A.A., Avner, S., Percevault, F., Demay, F., Bizot, M., Lucchetti-Miganeh, C., Barloy-Hubler, F., Brown, M., Lupien, M., Métivier, R., et al.

- (2011). Epigenetic switch involved in activation of pioneer factor FOXA1-dependent enhancers. *Genome Res.* *21*, 555–565.
- Smith, Z.D., and Meissner, A. (2013). DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* *14*, 204–220.
- 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.
- Taberlay, P.C., Kelly, T.K., Liu, C.C., You, J.S., De Carvalho, D.D., Miranda, T.B., Zhou, X.J., Liang, G., and Jones, P.A. (2011). Polycomb-repressed genes have permissive enhancers that initiate reprogramming. *Cell* *147*, 1283–1294.
- Teo, A.K., Arnold, S.J., Trotter, M.W., Brown, S., Ang, L.T., Chng, Z., Robertson, E.J., Dunn, N.R., and Vallier, L. (2011). Pluripotency factors regulate definitive endoderm specification through eomesodermin. *Genes Dev.* *25*, 238–250.
- Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. *Nature* *489*, 75–82.
- Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* *31*, 46–53.
- van Arensbergen, J., García-Hurtado, J., Moran, I., Maestro, M.A., Xu, X., Van de Casteele, M., Skoudy, A.L., Palassini, M., Heimberg, H., and Ferrer, J. (2010). Derepression of Polycomb targets during pancreatic organogenesis allows insulin-producing beta-cells to adopt a neural gene activity program. *Genome Res.* *20*, 722–732.
- Wamstad, J.A., Alexander, J.M., Truty, R.M., Shrikumar, A., Li, F., Eilertson, K.E., Ding, H., Wylie, J.N., Pico, A.R., Capra, J.A., et al. (2012). Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell* *151*, 206–220.
- Wang, Z., Oron, E., Nelson, B., Razis, S., and Ivanova, N. (2012). Distinct lineage specification roles for NANOG, OCT4, and SOX2 in human embryonic stem cells. *Cell Stem Cell* *10*, 440–454.
- Wei, H., Tan, G., Manasi, Qiu, S., Kong, G., Yong, P., Koh, C., Ooi, T.H., Lim, S.Y., Wong, P., et al. (2012). One-step derivation of cardiomyocytes and mesenchymal stem cells from human pluripotent stem cells. *Stem Cell Res. (Amst.)* *9*, 87–100.
- Wendling, O., Dennefeld, C., Chambon, P., and Mark, M. (2000). Retinoid signaling is essential for patterning the endoderm of the third and fourth pharyngeal arches. *Development* *127*, 1553–1562.
- Xie, W., Schultz, M.D., Lister, R., Hou, Z., Rajagopal, N., Ray, P., Whitaker, J.W., Tian, S., Hawkins, R.D., Leung, D., et al. (2013). Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* *153*. Published online May 9, 2013. <http://dx.doi.org/10.1016/j.cell.2013.04.022>.
- Yu, X., St Amand, T.R., Wang, S., Li, G., Zhang, Y., Hu, Y.P., Nguyen, L., Qiu, M.S., and Chen, Y.P. (2001). Differential expression and functional analysis of Pitx2 isoforms in regulation of heart looping in the chick. *Development* *128*, 1005–1013.
- Zhang, J.A., Mortazavi, A., Williams, B.A., Wold, B.J., and Rothenberg, E.V. (2012). Dynamic transformations of genome-wide epigenetic marking and transcriptional control establish T cell identity. *Cell* *149*, 467–482.
- Zhou, V.W., Goren, A., and Bernstein, B.E. (2011). Charting histone modifications and the functional organization of mammalian genomes. *Nat. Rev. Genet.* *12*, 7–18.

EXTENDED EXPERIMENTAL PROCEDURES

Cell Culture

All in vitro derived cell types were derived from HUES64 (Chen et al., 2009). Human embryonic stem cells were expanded on murine embryonic fibroblasts (Global Stem) in KO-DMEM (Life Technologies) containing 20% Knockout serum replace (Life Technologies) and FGF2 (10 ng/ml) (Millipore). Cultures were passaged by enzymatic dissociation using Collagenase IV (1 mg/ml) (Life Technologies). Prior to differentiation, cells were plated on matrigel-coated plates (BD Biosciences) and cultured in mTeSR1 (Stem Cell Technologies) for 3 to 4 days. Endoderm differentiation was induced in Advanced RPMI (Invitrogen), 0.5% FBS (Hyclone), Activin A (100 ng/ml) (R&D) and WNT3A (50 ng/ml) (R&D). HUES64-derived hepatoblasts (dHep) were induced by culturing day 5 endoderm in RPMI media containing B27 (1X), FGF2 (10ng/ml) (Millipore) and BMP4 (20 ng/ml) (R&D) for five days, and collected after 10 days total of differentiation. Hepatocyte-like cells were derived by culturing the HUES64-derived hepatoblasts in Lonza hepatocyte culture media containing 10ng/ml of HGF (R&D) for 5 additional days, or 15 days total. Mesoderm differentiation was induced by the addition of media consisting of in DMEM/F12 (Life Technologies), 0.5% FBS (Hyclone), Activin A (100ng/ml) (R&D) (for the first 24 hr only), BMP4 (100 ng/ml) (R&D), VEGF (100 ng/ml) (R&D) and FGF2 (20 ng/ml) (Millipore). To induce osteoblast differentiation, the day 5 mesoderm population was dissociated with accutase and replated on matrigel coated plates (BD) in EGM-2 media (Lonza) for 7 days, or 12 days total. Ectoderm differentiation was induced using A83-01 (2um) (Tocris), PNU 74654 (2um) (Tocris) and Dorsomorphin (2um) (Tocris), DMEM/F12 (Life Technologies) containing 15% Knock serum replacer (Life Technologies). Neurectoderm differentiation was induced by switching the day 5 ectoderm population to media containing 3 μ M CHIR99021 (TOCRIS), 10 μ M SU5402 (TOCRIS), and 10 μ M DAPT (TOCRIS), and collected after 6 more days, or 11 days total. N2-supplement (Life Technologies) was added to cells in 25% increments every other day beginning four days after the initiation of ectoderm differentiation. For all cell types, media was changed daily.

Antibodies

ChIP was performed using the following antibodies: H3K4me3 (Millipore, 07-473, Lot DAM1623866), H3K27ac (Abcam, ab4729, Lot 509313), H3K27me3 (Millipore, 07-449, Lot DAM1514011), H3K36me3 (Abcam, ab9050, Lot 499302), H3K4me1 (Abcam, ab8895, Lot 659352), H3K9me3 (Abcam, ab8898, Lot 484088), POU5F1 (Abcam, ab19857), SOX2 (Santa Cruz, sc-17320X), NANOG (R&D, AF1997) and FOXA2 (R&D, AF2400).

For live cell FACS isolation, cells were stained for 30 min on ice with the following antibodies directed toward extracellular surface proteins: CD326-PerCP-Cy5 (clone EBA1) (BD Biosciences), CD56-PE (clone NCAM16.2) (BD Biosciences), and CD184-PE-Cy5 (clone 12G5) (BD Biosciences).

Immunostaining was done with the following primary antibodies: FOXA2 (R&D, AF2400), GATA2 (Santa Cruz, sc-16044) SOX17 (R&D, AF1924), PAX6 (Covance, PRB-278P) and HNF4 α (Abcam, ab41989). Cells were fixed in 4% Formaldehyde, incubated in primary antibody overnight at 4°C, and then incubated in secondary antibody for 1 hr at room temperature. DNA was detected using Hoechst 33342 trihydrochloride trihydrate (Invitrogen).

FACS Analysis

FACS was done on a BD FACSAria II using linear FSC and SSC scaling, followed by height and width-based doublet discrimination. The viability of the populations was assessed by Propidium Iodide staining, with the positively stained populations being excluded from the sorting gates. Compensation was calculated using FACS Diva autocompensation algorithms, and supplemented by manual compensation to correct for autofluorescence.

Genomic DNA Isolation

Flash-frozen human tissues or cell pellets were lysed at 55°C overnight in 300–600 μ l lysis buffer (10 mM Tris-HCl pH 8.0, 10 mM EDTA, 10 mM NaCl and 0.5% wt/vol SDS) supplemented with 50 ng/ μ l DNase-free RNase (Roche) and 1 μ g/ μ l proteinase K (NEB). After extraction with an equal volume of phenol:chloroform:isopropanol alcohol (25:24:1; Invitrogen) and addition of 0.5 μ l (20 μ g/ μ l) glycogen (Roche) and 1/20 vol 5 M NaCl, DNA was precipitated with 2.5 vol ethanol, spun down (30 min/16,000 g) at 4°C and washed with 70% ethanol. DNA was re-suspended in 30–100 μ l of TE buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA) and quantified using a Qubit fluorometer and a dsDNA BR Assay Kit (Life Technologies).

WGBS Library Construction

Genomic DNA (1–5 μ g) was fragmented to 100–500 bp using a Covaris S2 sonicator 9 times for 60 s at duty cycle 20%, intensity 5 and 200 cycles per burst. DNA fragments were cleaned up using a QIAGEN PCR purification kit. End-repair reactions (100 μ l) contained 1x T4 DNA ligase buffer (NEB), ATP, 0.4 mM dNTPs, 15 units T4 DNA polymerase, 5 units Klenow DNA polymerase, 50 units T4 polynucleotide kinase (all NEB) and were incubated for 30 min. at 19°C and 15 min. For some libraries we used a dCTP-free dNTP mix instead of all four dNTPs during for the end-repair to avoid artificially unmethylated sites. Adenylation was performed for 30 min. at 37°C in 50 μ l 1x Klenow buffer containing 0.2 mM dATP and 15 units Klenow exo⁻ (NEB). Adenylated DNA fragments and methylated paired-end adapters (purchased from ATDBio) were incubated overnight at 16°C in a 50 μ l reaction containing 5,000

units concentrated T4 DNA ligase (NEB) and 3 μM of adapters. Each enzymatic reaction was terminated and cleaned-up by phenol/chloroform extraction and ethanol precipitation as described above.

To determine unmethylated cytosine conversion rates and methylated cytosine over-conversion rates by sodium bisulfite treatment, adaptor-ligated fully methylated and fully unmethylated internal control DNA fragments (Table S7), were spiked into WGBS library preparation at a molar ratio (spike-in to WGBS library) of 1:16,000 each. Adaptor-ligated DNA of 270-370 bp, corresponding to DNA insert sizes of 150-250 bp, was size-selected on a 2.5% Nusieve (3:1) agarose gel (Lonza). Two consecutive bisulfite conversions were performed with an EpiTect Bisulfite Kit (QIAGEN) following the protocol specified for DNA isolated from FFPE tissue samples. One of 40 μl bisulfite-converted DNA was used in each of four 10- μl reactions to determine the minimal PCR cycle number for library amplification. PCR reactions contained 0.5 U of *PfuTurboCx* Hotstart DNA polymerase (Agilent technologies), 1 μl of 10x PCR buffer, 250 μM dNTPs, 1.5 μM of Primer 1.0 and 2.0 (Illumina). The thermocycling profile was 2 min. at 95°C followed by 5-15 cycles of 30 s at 95°C, 30 s at 65°C, 1 min. at 72°C, and a final 7 min. extension at 72°C. Preparative library amplification using the empirically determined number of PCR cycles was performed in eight 25- μl aliquots, each containing 3 μl of bisulfite-converted DNA, 1.25 U of *PfuTurboCx* Hotstart DNA polymerase, 2.5 μl of 10x PCR buffer, 250 μM of dNTP, 1.5 μM of Primer 1.0 and 2.0. PCR products were pooled and purified twice using Agencourt *AMPure XP SPRI Beads* (Beckman Coulter) as per the manufacturer's instructions. The final library DNA was quantified using a Qubit fluorometer and a Quant-iT dsDNA HS Kit (Invitrogen). The insert size was checked on a 4%–20% non-denaturing polyacrylamide gel (Bio-Rad). Paired-end sequencing with 100 base reads was performed on an Illumina HiSeq 2000 followed the manufacturer's guidelines.

WGBS Data Processing and Analysis

WGBS raw sequencing reads were aligned using *maq* in bisulfite mode against human genome version hg19/GRCh37, discarding duplicate reads. DNA methylation calling was performed based on an extended custom software pipeline published previously for RRBS (Bock et al., 2011). To ensure comparability of region DNA methylation levels across all samples, only CpGs covered by $\geq 5x$ in 85% of the samples qualified for the computation of region DNA methylation levels. To assess the DNA methylation state of various genomic regions, we resorted to our previously published protocol estimating a genomic region's methylation state as the coverage weighted average across all CpGs within each region. Subsequently, we averaged a region's DNA methylation level over replicates. Differentially methylated regions (DMRs) were defined as exhibiting significantly ($p \leq 0.05$, Fisher's exact test) different DNA methylation levels of at least 0.1.

Many gene regulatory elements (GREs) are marked by spatially highly constrained reduced DNA methylation levels. It has recently been suggested that besides CpG islands, which are mostly unmethylated (UMR) a second class of GRE is marked by low to intermediate DNA methylation (IMR) (Stadler et al., 2011). We reasoned that these regions might be of particular regulatory importance in our system and might be missed by looking at histone modification enrichments alone. Therefore we adopted a similar Hidden Markov model approach as proposed in Stadler et al. (2011) to identify regions of reduced DNA methylation level. Briefly, we utilized a three state Hidden Markov Model operating on the methylation levels of each CpG in the human genome. Each state's emission probabilities for the DNA methylation levels were modeled by a normal distribution. The model was trained on all CpGs of chromosome 19 in the HUES64 data set using an adaption of the well known Baum Welch algorithm to incorporate the normal distribution (Press, 2007). After initial parameter estimation, we utilized the approach reported by Stadler et al. (2011) to determine the FDR for IMR regions and adapted the initial parameter estimates for the IMR and HMR states to finally 0.01(UMR), 28.8 (IMR), 81.6 (highly methylated, HMR), yielding an FDR of 2%. This parameter set was subsequently used to segment all WGBS data sets. Finally, we used the Viterbi algorithm to compute the most probable path through each chromosome separately and assigned the CpG states accordingly to either unmethylated, intermediate or highly methylated. Subsequently, we merged neighboring CpGs residing in the same state and being less than 200bp apart into unmethylated, intermediate, or highly methylated regions. Only regions harboring more than 3 CpGs were retained for subsequent analysis. The resulting region set is more likely to pick up DMRs due to the highly spatially constrained nature of the marked GRE (often 200–400bp) which easily gets masked by a coarse grained tiling based approach. The HMM inference framework was implemented as custom software in python (<http://python.org/>) and extended to incorporate other state distribution types. To determine differentially methylated regions between two samples, we followed our previously established protocol (Bock et al., 2011).

ChIP and ChIP-Seq Library Production

Cells collected by FACS were crosslinked in 1% formaldehyde for 15 min at room temperature, with constant agitation, followed by quenching with 125mM Glycine for 5 min at room temperature with constant agitation. Nuclei were isolated and chromatin was sheared using Branson sonifier until the majority of DNA was in the range of 200–700 base pairs. Chromatin was incubated with antibody overnight at 4°C, with constant agitation.

Co-immunoprecipitation of antibody-protein complexes was completed using Protein A or Protein G Dynabeads for 1 hr 4°C, with constant agitation. ChIPs were completed using previously reported methods (Mikkelsen et al., 2010). Sequencing library production details can be found in the [Extended Experimental Procedures](#). Sequencing libraries were submitted for sequencing on the Illumina HiSeq 2000.

Immunoprecipitated DNA was end repaired using the End-It DNA End-Repair Kit (Epicenter), extended using a Klenow fragment (3'-5' exo) (NEB), and ligated to sequencing adaptor oligos (Illumina). Each library was then PCR-amplified using PFU Ultra II Hotstart

Master Mix (Agilent), and a size range of 300–600 was selected for sequencing. We confirmed binding of OCT4, NANOG and SOX2 at the NANOG promoter using qPCR (Table S8).

ChIP-BS-Seq Library Construction

DNA was first subjected to end-repair in a 30- μ l reaction containing 6 units T4 DNA polymerase, 2.5 units DNA Polymerase I (Large Klenow Fragment), 20 units T4 Polynucleotide Kinase (all New England Biolabs), dATP, dCTP, dGTP, and dTTP (0.125 mM each), and 1 \times T4 Ligase buffer with ATP for 30 min at 20°C. DNA was then adenylated in a 20- μ l reaction containing 10 units Klenow Fragment (3' \rightarrow 5' exo-) (New England Biolabs), 0.5 mM dATP and 1 \times NEB buffer 2 for 30 min at 37°C. DNA was then ligated to preannealed Illumina genomic DNA adapters containing 5-methylcytosine instead of cytosine (ATDBio) using T4 DNA ligase (New England Biolabs).

Adaptor-ligated DNA fragments were subsequently purified by phenol extraction and ethanol precipitation and size-selected on gel. 50 ng sheared and dephosphorylated *Escherichia coli* K12 genomic DNA was added to adaptor-ligated DNA as carrier during size-selection and bisulfite conversion. DNA was run on 2.5% Nusieve 3:1 Agarose (Lonza) gels. Lanes containing marker (50 bp ladder; New England Biolabs) were stained with SYBR Green (Invitrogen), and size regions to be excised were marked with toothpicks and adaptor-ligated DNA fragments from 200–400 and 400–550 bp were excised. DNA was isolated from gel using the MinElute Gel Extraction kit (QIAGEN). The low and high libraries were kept separate in subsequent steps.

Adaptor-ligated and size-selected DNA was subjected to two subsequent 5-h bisulfite treatments using the EpiTect Bisulfite kit (QIAGEN) following the manufacturer's protocol for DNA isolated from FFPE tissue samples. PCR amplification was done with 1.25 units Pfu Turbo Cx Hotstart DNA Polymerase (Stratagene), primer LPX 1.1 and 2.1 (0.3 μ M each), dNTPs (0.25 mM each), 1 \times Turbo Cx buffer. Amplified libraries were purified with the MinElute PCR Purification kit (QIAGEN) and subsequently purified from gel essentially as described above; whole gels were stained with SYBR Green, and no carrier DNA was added. Final libraries were analyzed on analytical 4%–20% TBE Criterion precast gels (BioRad), and measured by Quant-iT dsDNA HS Assays (Invitrogen) (protocol adapted from Brinkman et al. (2012)).

ChIP-Seq Data Processing and Analysis

ChIP-Seq data were aligned to the hg19/GRCh37 reference genome using bwa version 0.5.7 (Li and Durbin, 2009) with default parameter settings. Subsequently, reads were filtered for duplicates and extended by 200bp. Visualization of read count data was performed by converting raw bam files to .tdf files using IGV tools (Thorvaldsdóttir et al., 2013) and normalizing to 1 million reads.

In order to identify regions enriched for chromatin modifications we employed a two step approach, first identifying all regions enriched for any chromatin modification. Next, using this comparatively small region set, we determined the quantitative enrichment level as well as significance of enrichment using a Poisson background model based on the whole cell extract (WCE). Finally, we utilize conservative enrichment and significance cutoffs to binarize our enrichment signal in order to increase robustness and simplify subsequent analysis.

First, we segmented the genome into non-overlapping windows and classified each window into either enriched or no enriched. This analysis was conducted separately for two groups, 1. H3K27ac, H3K4me3 using 200bp windows and H3K27me3, H3K9me3, H3K4me1 using 400bp windows. To compute the enrichment statistics on the window level, we determined the number of unique insert size extended sequence tags whose midpoint was located within the window of interest for the ChIP-Seq track of interest as well as the WCE. Next, we used the poisson model proposed in Mikkelsen et al. (2010) to determine nominal p-value of enrichment and computed the enrichment over the WCE. Only windows enriched at a significance level below $p < 10^{-5}$ (in case replicates were used both had to fulfill this criterion) and an enrichment above background greater than 3 (in case replicates were used, the average enrichment had to be greater than 2.5) were retained. However, for most enrichment analysis we employed only the replicate with the strongest signal.

Next, enriched windows within a distance of 850bp were merged into larger regions. Regions smaller than 400bp (600bp for broad marks) after merging were discarded as due to noise and regions greater than 10kb were split. This procedure was carried out for three groups of histone ChIP-Seq tracks separately: H3K4me3 & H3K27ac, H3K4me1 and H3K27me3 & H3K9me3 across all 4 cell types. The resulting three lists of enriched regions were then merged in a hierarchical fashion: first regions identified based on H3K4me3&H3K27ac and H3K4me1, retaining all H3K4me3&H3K27ac regions but merging or splitting enriched H3K4me1 regions.

After completion of this initial processing step, regions were again filtered for minimal size discarding regions smaller than 400bp. Next, the same procedure was repeated for the new H3K4me3, H3K27ac, H3K4me1 region set and the H3K27me3, H3K9me3 region list. Finally, the resulting list was merged with the regions classified as UMRs and IMRs, adding only regions not overlapping with any region identified so far. This procedure gave rise to the region catalog used in subsequent analysis.

In the second processing step, comparative analysis of ChIP-Seq experiments and assignment of chromatin states was carried out, excluding regions enriched for H3K9me3 only. First, for each region in the region catalog the significance and enrichment over WCE was determined using Poisson statistics (Mikkelsen et al., 2010) applied to the duplicate filtered and insert size extended sequencing tag counts overlapping each identified region. Regions with tag counts deviating at a significance level of $p \leq 0.001$ from the WCE and exhibiting enrichment over WCE ≥ 3 were classified as enriched. We chose these moderately stringent thresholds in order also pick up chromatin state changes that occur only in a subset of the investigated cell population and therefore have lower signal. However, this comes at the expense of a higher false positive rate. Next, we compared the enrichment levels for all four cell

types (hESC, dEC, dME, dEN) for each epitope separately. To that end we used the Poisson model based approach proposed in Mikkelsen et al. (2010) and defined regions deviating by ≥ 2.5 fold at a significance level of $p \leq 0.01$ as being different. Next, we reconciled these differential enrichment calls with our enrichment over background classification. Since in our setting we were mostly concerned with incorrectly called differences between cell states (false positives) due to heterogeneity in the distinct populations and varying ChIP-Seq library complexity, we redefined regions that were classified as enriched in hESC and not enriched in one of the differentiated cell types but exhibiting no significant difference according to our differential analysis as being enriched in the differentiated cell type under study. This approach yields a lower false positive rate in terms of dynamics at the expense of a higher false negative rate. However, at this point it still remains to be determined what magnitudes of differences in chromatin modifications are actually meaningful. In this sense, our binary classification approach is rather conservative and relies on previously established observations. Subsequently, we classified each genomic region identified in this way into one of 11 epigenetic states based on the binary classification of enrichment levels for the various modifications. DNA methylation levels were not taken into account when histone modification based states were assigned. Only states devoid of significant enrichment for one of the histone modifications were classified based on DNA methylation levels. Genomic regions were associated with their nearest RefSeq gene using the R package ChIPpeakAnno (Zhu et al., 2010) and classified into promoter, intragenic, distal (<50kb from TSS and not promoter) and intergenic.

TF ChIP-Seq Analysis

For OCT4, SOX2, NANOG and FOXA2 aligned read files were processed with macs version 1.4 (Zhang et al., 2008) using the following parameters: $-g 2.7e9$ $-tsize = 36$ $-pvalue = 1e-5$ $-keep-dup = 1$ and the HUES64 WCE as input control. All other parameters were left at their default setting. For our 25bp libraries, $tsize$ was set to 25. FDR was calculated using macs built in function essentially comparing the original read count distribution with a randomly shuffled distribution. Following this initial peak calling, only peaks significant at an FDR of 0.05 and present in both replicates were retained. As a second replicate for our OCT4 ChIP-Seq experiment we took advantage of publically available OCT4 data (Kunarse et al., 2010).

ChIP BS-Seq Analysis

For the FOXA2 ChIP-bisulfite sequencing experiment, the bisulfite treated ChIP library was processed similarly to the WGBS processing described above and subsequently overlaid with the peak calling results from the FOXA2-ChIP-Seq library that was not bisulfite treated.

RNA-Seq Data Processing and Differential Expression Analysis

Strand specific libraries were constructed as described in the main text using a strand specific method (Levin et al., 2010). Reads were mapped to the human genome (hg19) using TopHat v2.0.6 (Trapnell et al., 2009) (<http://tophat.cbcb.umd.edu>) with the following options: “ $-library-type firststrand$ ” and “ $-transcriptome-index$ ” with a TopHat transcript index built from RefSeq. Transcript expression was estimated with an improved version of Cuffdiff 2 (Trapnell et al., 2013) (<http://cufflinks.cbcb.umd.edu>). Cuffdiff was run with the following options: “ $-min-reps-for-js-test 2$ $-dispersion-method per-condition$ ” against the UCSC iGenomes GTF file from Illumina (available at <http://cufflinks.cbcb.umd.edu/igenomes.html>). The workflow used to analyze the data is described in detail in Trapnell et al. (2012) (alternate protocol B).

To identify a gene or transcript as DE, Cuffdiff 2 tests the observed log-fold-change in its expression against the null hypothesis of no change (i.e., the true log-fold-change is zero). Because of measurement error, technical variability, and cross-replicate biological variability might result in an observed log-fold-change that is nonzero, Cuffdiff assesses significance using a model of variability in the log-fold-change under the null hypothesis. This model is described in detail in Trapnell et al. (2013). Briefly, Cuffdiff two constructs, for each condition, a table that predicts how much variance there is in the number of reads originating from a gene or transcript. The table is keyed by the average reads across replicates, so to look up the variance for a transcript using the table, Cuffdiff estimates the reads originating from that transcript, and then queries the table to retrieve the variance for that number of reads. Cuffdiff 2 then accounts for read mapping and assignment uncertainty by simulating probabilistic assignment of the reads mapping to a locus to the splice isoforms for that locus. At the end of the estimation procedure, Cuffdiff 2 obtains an estimate of the number of reads that originated from each gene and transcript, along with variances in those estimates. The read counts are reported along with FPKM values and their variances. Change in expression is reported as the log fold change in FPKM, and the FPKM variances allow the program to estimate the variance in the log-fold-change itself. Naturally, a gene that has highly variable expression will have a highly variable log-fold-change between two conditions.

The modifications made to Cuffdiff 2 improve sensitivity in calling differentially expressed (DE) genes and transcripts while maintaining a low false positive rate. They stem from the method used to calculate the variability in the log fold change in expression. In Trapnell et al., Cuffdiff 2 used the “delta method” to estimate the variance of the log fold change estimate for a gene or transcript. This method yields a simple equation that takes as input the mean and variance of the transcript’s expression in two conditions and produces a variance for the log fold change. However, the equation contains no explicit accounting for the number of replicates used to produce those estimates – they are assumed to be perfectly accurate.

The improved version of Cuffdiff 2 more accurately estimates the variance in the log-fold-change using simulated draws from the model of variance in expression for each of the two conditions. Imagine an experiment that has n replicates in condition A and m

replicates in condition B. To estimate the distribution of the log-fold-change in expression for a gene G under the null hypothesis, Cuffdiff first draws n times from the distribution of expression of G according to the algorithm's model of expression. Cuffdiff then takes the average of the n draws to obtain an expression "measurement." Then, Cuffdiff draws m from the same distribution and again takes their average. Cuffdiff then takes the log ratio of these averages, places this value in a list, and then repeats the procedure until there are thousands of such log-fold-change samples in the list. The software then makes a similar list, this time using the expression model for condition B – the null hypothesis assumes both sets of replicates originate from the same condition, but we do not know whether A or B is the better representative of that condition, so we must draw samples from both and combine them. To calculate a p -value of observing the real log-fold-change under this null model, we simply sort all the samples and count how many of them are more extreme than the log fold change we actually saw in the real data. This number divided by the total number of draws is our estimate for the p -value.

Cuffdiff 2 reports not only genes and transcripts that are significantly differentially expressed between conditions, but also *groups* of transcripts (i.e., the isoforms of a gene) that show significant changes in expression relative to one another. The test for this is similar to what is described in Trapnell et al., but comparably modified along the lines described above for single genes or transcripts. Draws of expression are made for each transcript in a group according to the number of replicates in the experiment. These are averaged, and the shift in relative transcript abundance for the draw is made using the Jensen-Shannon metric. These draws are added to a list and used to calculate p -values for significance of observed shifts in relative abundance under the null hypothesis.

Clustering of gene expression profiles was achieved with the `csDendro()` function from `CummeRbund` (<http://compbio.mit.edu/cummeRbund/>). This function first transforms the FPKMs of all genes in each sample by adding 1 and then takes the logarithm. Next, it converts each genes transformed expression into a fraction of the total transformed expression. The distances between these transformed expression profiles are then measured by the Jensen-Shannon metric. The distances are then used to build a dendrogram via complete linkage hierarchical clustering using the R function `hclust()`.

Motif Analysis

Predefined sets of genomic regions were scanned for occurrences of motifs contained in the Transfac professional database (2009) using the FIMO program from the MEME suite (Grant et al., 2011). Only motifs with at least one known associated human transcription factor and detected at a significance level of $p \leq 10^{-5}$ were used for further analysis. Next, the total number of occurrences was calculated for each motif. To correct for sequence composition, we trained a Hidden Markov Model on each set of input sequence sets and generated 10 sets of number and size matched region sets using the inferred probabilities as controls. Subsequently, these sequence sets were also subjected to the same motif identification procedure and motif enrichment results were averaged over the 10 control runs. We defined the final motif enrichment score as the fraction of total motif occurrences in the region set of interest and the total number of motif occurrences in the averaged control region set. To determine differentially enriched motifs between region sets from different hESC-derived cell types, we calculated the fraction of motif scores between the two conditions, retaining only motifs with a differential enrichment ≥ 1.2 .

For the H3K27ac motif analysis, we computed overall motif enrichment scores for each region class separately as described above. Next, we correlated the motif enrichment scores only focusing on those motifs with scores ≥ 1.2 . To that end we multiplied the motif enrichment score for the cell type of interest with the \log_2 fold change of the associated transcription factor in that cell type, giving rise to a new combined motif score. If multiple TFs mapped to one motif, we took the average motif score. For each cell type we rank ordered the motifs according to their enrichment scores and report the top 20 motifs with their raw motif score in Figure 5D.

For the H3K4me1 analysis, we wanted to focus on all potential TFBS gaining H3K4me1 and not only those that also become expressed as in the H3K27ac analysis. First, we again determined the motif enrichment scores over background. To focus on motifs differentially enriched between the different cell types, we subtracted the mean motif enrichment across hESd cell types for each motif separately from the enrichment level and rank ordered the motifs. For each cell type, we then report the top 20 enriched motifs in Figure 6G.

SUPPLEMENTAL REFERENCES

- Chen, A.E., Egli, D., Niakan, K., Deng, J., Akutsu, H., Yamaki, M., Cowan, C., Fitz-Gerald, C., Zhang, K., Melton, D.A., and Eggan, K. (2009). Optimal timing of inner cell mass isolation increases the efficiency of human embryonic stem cell derivation and allows generation of sibling cell lines. *Cell Stem Cell* 4, 103–106.
- Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27, 1017–1018.
- 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.
- Levin, J.Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D.A., Friedman, N., Gnirke, A., and Regev, A. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* 7, 709–715.
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760.
- Mikkelsen, T.S., Xu, Z., Zhang, X., Wang, L., Gimble, J.M., Lander, E.S., and Rosen, E.D. (2010). Comparative epigenomic analysis of murine and human adipogenesis. *Cell* 143, 156–169.
- Press, W.H. (2007). *Numerical Recipes: the Art of Scientific Computing*, Third Edition (Cambridge, UK: Cambridge University Press).

- Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* *14*, 178–192.
- Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* *25*, 1105–1111.
- 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.
- 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.
- Zhu, L.J., Gazin, C., Lawson, N.D., Pagès, H., Lin, S.M., Lapointe, D.S., and Green, M.R. (2010). ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* *11*, 237.

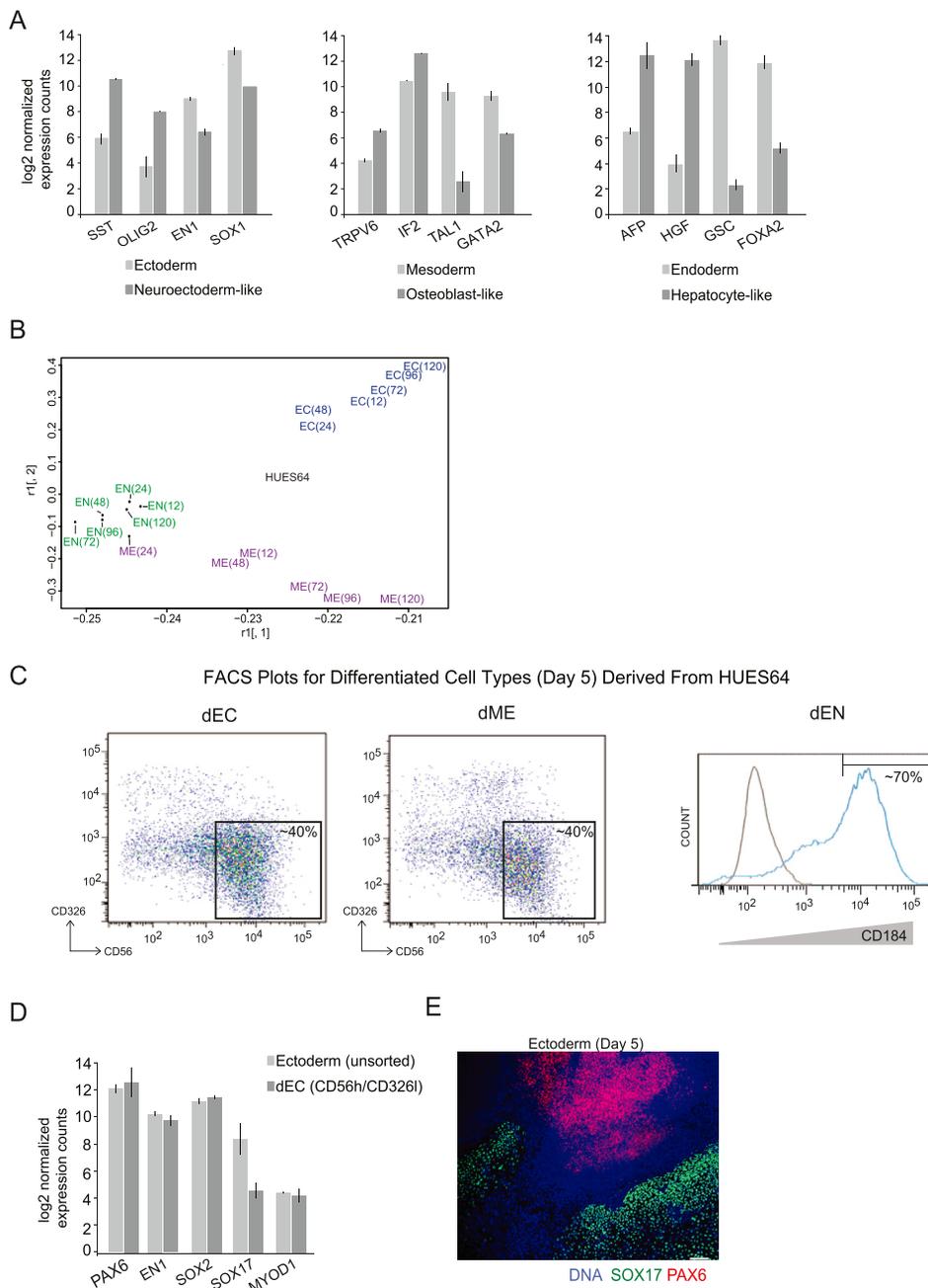


Figure S1. Additional Characterization of the Differentiated Populations, Related to Figure 1

(A) Median Nanostring expression values (log₂) of populations derived from dEC, dME and dEN.

(B) Multidimensional scaling of populations included in differentiation time course.

(C) Representative FACS plots used to isolate differentiated populations. Square boxes (left and middle panels) and line (in right panel) indicate population collected for further analysis. Approximate percent of population collected is given.

(D) Average Nanostring expression values (log₂) of unsorted ectoderm versus CD56high/CD326low sorted dEC cells.

(E) Immunofluorescent staining of SOX17 and PAX6 in day 5 ectoderm population (40x, scale bar equals 200um).

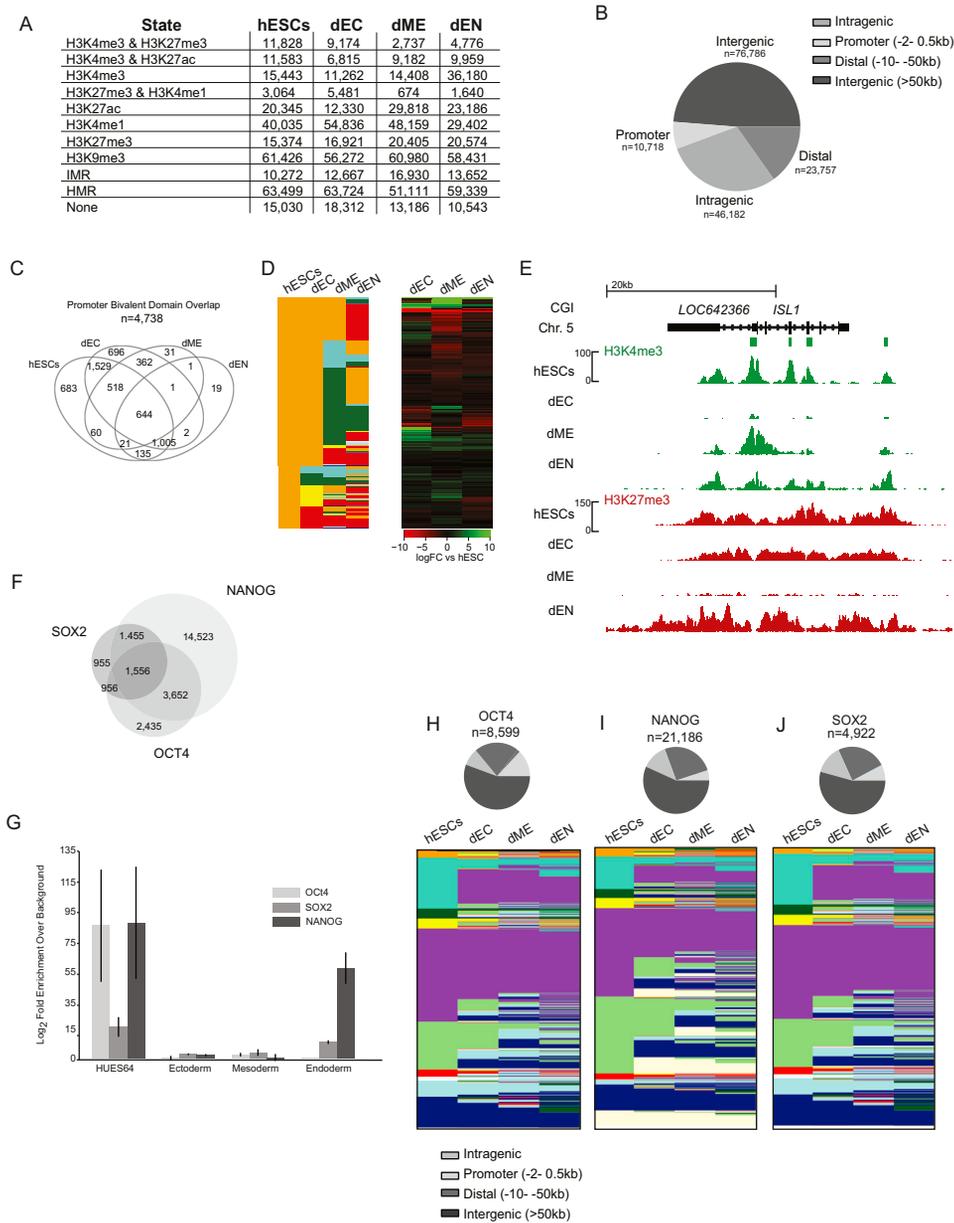


Figure S2. Overall Chromatin Dynamics and TFS Binding, Including OCT4, SOX2, and NANOG, Related to Figure 2

(A) Total number of regions found in each state, in each population.
 (B) Genomic features associated with all regions that change their epigenetic state in at least one cell type.
 (C) Venn diagram showing the overlap of identified proximal bivalent domains.
 (D) Left: Chromatin state map for all TFs that are bivalent in hESCs and change their epigenetic state in at least one cell type (n = 400). Right: Hierarchical clustering ordered heatmap of TF expression (logFC relative to hESCs).
 (E) Normalized ChIP-seq tracks of H3K4me3 and H3K27me3 at the *ISL1* locus (chr5:50,661,163-50,703,879) indicating H3K27me3 is selectively maintained at high levels in dEC but not dME, in contrast to dEN, where H3K4me3 increases while H3K27me3 is lost, promoting active transcription. Read counts on y axis are normalized to 10 million reads and CpG islands (CGI) are indicated in green.
 (F) Venn diagram of the overlap between OCT4, NANOG and SOX2 binding sites identified in hESCs (total overlap = 1,556).
 (G) Fold enrichment of OCT4, NANOG and SOX2 binding at the NANOG locus in hESCs, and each differentiated population on day 5 of differentiation.
 (H) Genomic features of OCT4 binding sites (top) and the associated epigenetic states (bottom) (n = 8,599).
 (I) Genomic features of NANOG binding sites (top) and the associated epigenetic states (bottom) (n = 21,186).
 (J) Genomic features of SOX2 binding sites (top) and the associated epigenetic states (bottom) (n = 4,922).

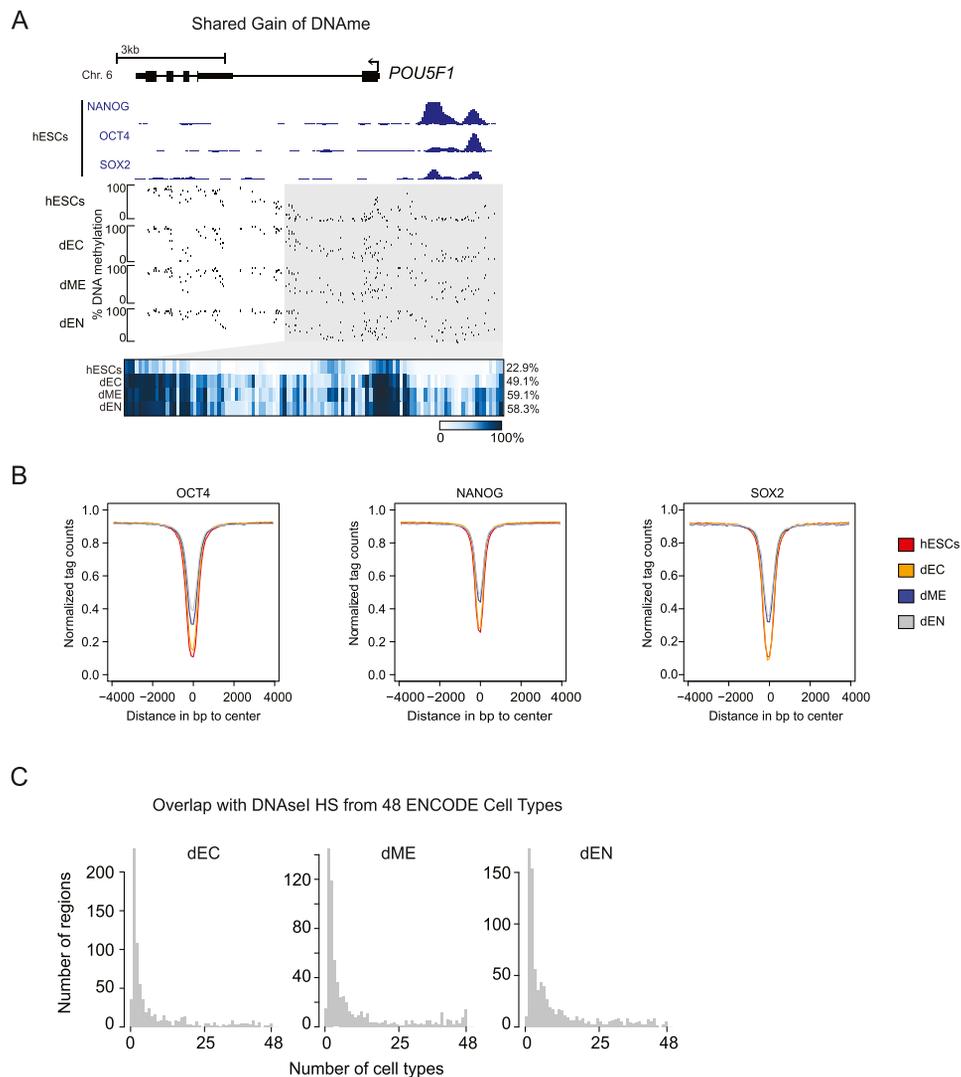


Figure S3. DNA Methylation Dynamics at *POU5F1* and Regions Associated with TF Binding, Related to Figure 3

(A) Gain of DNAm at the *POU5F1* locus (chr6:31,135,410-31,141,237). NANOG, OCT4, SOX2 ChIP-seq tracks (hESCs only) and DNAm levels in hESCs and differentiated cell types. Individual CpG methylation values across the locus are displayed using the IGV. The heatmap below shows the DNAm values of individual CpGs within the gray region. The average DNAm value for the entire highlighted region is shown on the right in red. The TSS is indicated by the arrow. Gain of DNAm is seen at the distal enhancer, as well as over the TSS, in all three differentiated cell types.

(B) Composite plots of DNAm levels in hESC and differentiated populations across SOX2/OCT4/NANOG binding sites in hESCs. Average CpG methylation levels were computed for 100bp tiles across an 8kb region centered at the middle of each transcription factor binding site.

(C) Frequency distribution of overlapping DMRs gaining DNAm in the differentiated populations with DNase I hypersensitive sites across 48 ENCODE cell types (Thurman et al., 2012).

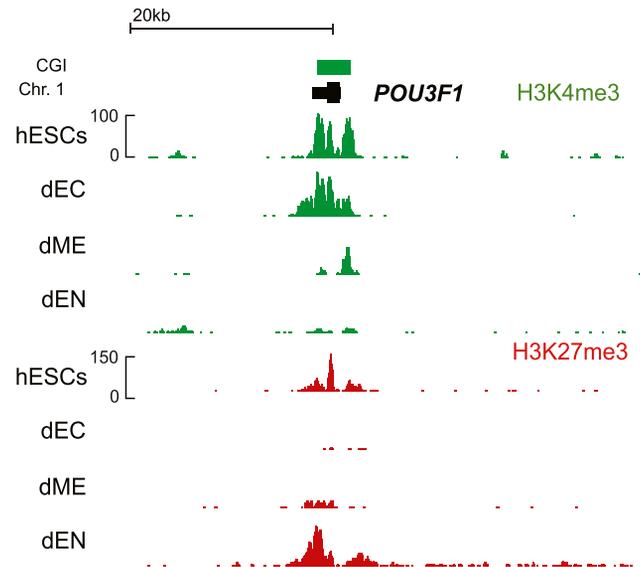


Figure S4. H3K4me3 and H3K27me3 Enrichment at *POU3F1*, Related to Figure 4

H3K4me3 and H3K27me3 tracks at the *POU3F1* locus (chr1:38,493,152-38,532,618) show lineage-specific resolution of the bivalent domain.

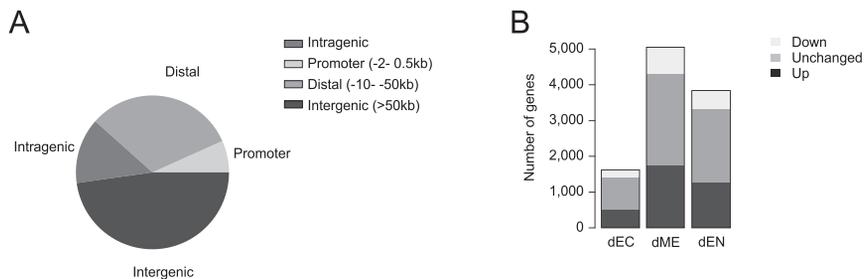


Figure S5. Genomic Features and Expression Changes at Regions that Gain H3K27ac, Related to Figure 5

(A) Genomic features associated with gain of H3K27ac during differentiation. See key on the right.

(B) Classification of gene expression associated with regions gaining H3K27ac in each germ layer into either upregulated (FDR < 0.05), downregulated (FDR < 0.05), or unchanged.

Phase HNF4a

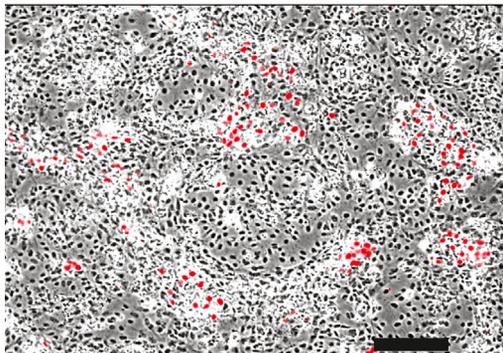


Figure S6. HUES64-Derived Hepatoblast-like Population, Related to Figure 6

Immunofluorescent HNF4 α stain of hepatoblast-like population after 10 days of differentiation (10x, scale bar = 100 μ m).

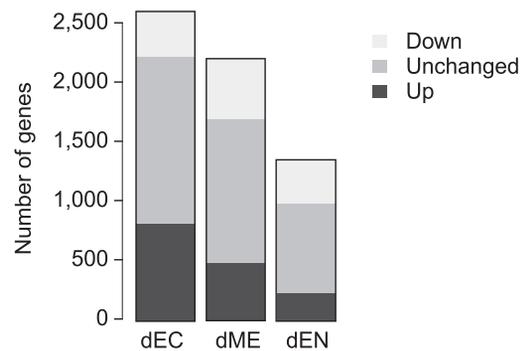


Figure S7. Gene Expression Changes at Genes that Gain H3K27me3, Related to Figure 7

Classification of gene expression associated with regions gaining H3K27me3 in each germ layer into either upregulated (FDR < 0.05), downregulated (FDR < 0.05), or unchanged.