Transcriptional and Epigenetic Dynamics during Specification of Human Embryonic Stem Cells

Casey A. Gifford,1,2,3,9 Michael J. Ziller,1,2,3 Hongcang Gu,1 Cole Trapnell,1,3 Julie Donaghey,1,2,3 Alexander Tsankov,1,2,3 Alex K. Shaalek,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 Gmitko,1 Eric S. Lander,1,7 Bradley E. Bernstein,1,8 and Alexander Meissner1,2,3,*

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 methylation (DNAme) 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, DNAme, 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 DNAme 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 DNAme 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 log2 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 log2 expression value of two biological replicates is displayed. Error bars represent 1 SD.
To dissect the early transcriptional and epigenetic events during hESC specification, we used two-dimensional, directed differentiation of HUES64, a National Institutes of Health (NIH)-approved cell 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 TGFβ, Wingless/integrase1 (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 FGFR2 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 log2 expression) during the first 5 days of differentiation (Figure 1B). Genes such as EOMES, 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). FOXA2 is upregulated during early ectoderm differentiation, which instead upregulates neural markers such as PAX6, SOX10, and EN1 (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 (log2 expression 10.9) (Figure 1D and Table S1), whereas ZFP4 (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 OLG2 and
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 × 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;...
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 2). 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 2). 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 differentiation.

Gain of DNAme 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 ≤ 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 DNAme 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 DNAme are lineage specific (Figure 3C) and include loci such as SMAD3 (dEC), CTNNB3 (dME), and FOXA2 (dEN). FOXA2 has an upstream CGI that exhibits gain of DNAme (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 DNAme (Maunakea et al., 2010).
We find significant enrichment of various TF motifs as DNAme 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 DNAme 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 three factors. The DNAme state at these regions is shown in Figure 3C. We also examined the DNAme state at regions bound by SOX2, OCT4, and NANOG ChIP-seq at the DBX1 locus (Deaton et al., 2011; Bock et al., 2012). To examine the DNAme state at regions bound by SOX2, OCT4, and NANOG in hESCs, we used whole-genome bisulfite sequencing (WGBS) to examine the DNAme state at regions bound by SOX2, OCT4, and NANOG ChIP-seq at the DBX1 locus (Deaton et al., 2011; Bock et al., 2012).
three TFs and gain DNAme in dME and dEN. In contrast, this region maintains low levels of DNAme in dEC, which has activated transcription of DBX1 (Figure 3E). We generally find that cobound sites gain DNAme in the dME and dEN, but not dEC (Figure S3B). Further supporting the functional relevance of these dynamics, we find that regions that gain DNAme frequently coincide with DNaseI hypersensitive sites (Figure S3C) (Thurman et al., 2012). Transcriptional silencing was less frequently correlated with gain of DNAme at distal elements than at promoters (Figure 3F).

In examining the chromatin state of regions that gain DNAme 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 DNAme 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 x 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 DNAme at a putative regulatory element downstream of the 3’ UTR in dEC (Figure 4C). On a global scale, an immediate correspondence between loss of DNAme and expression, such as that observed at POU3F1, occurs at about half of the regions (Figure 4D). More than 70% of DMRs that lose DNAme 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 DNAme 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 TGFβ, 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 DNAme 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^{-53}), TGFβ pathway targets (q < 10^{-18}) 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 DNAme 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-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 DNAme 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 DNAme 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 red 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.
A previous report found that FOXA1/FOXA2 could bind to regions exhibiting DNAme (Sérandour et al., 2011), which is not a characteristic shared by all TFs (Gifford and Meissner, 2012). Regions bound by these factors subsequently lost DNAme and gained euchromatic histone modifications in our populations. We therefore compared DNAme 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 DNAme state of regions isolated by FOXA2-ChIP-BS sequencing in dEN (Brinkman et al., 2012). Interestingly, we saw a major depletion of DNAme 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 DNAme 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 in the human liver, suggesting that these transitions are due to genomic conformation changes and reflect H3K27me3 enrichment in dEN (Figure 7I). To more directly assess this relationship, we interrogated the DNAme state of regions isolated by FOXA2-ChIP-BS sequencing in dEN (Brinkman et al., 2012). Interestingly, we saw a major depletion of DNAme 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).

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.
Figure 7. DNAme 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. DNAme information is depicted for hESC and dEN WGBS datasets 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 SCRB 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...
REFERENCES


Epigenetic switch involved in activation of pioneer factor FOXA1-dependent enhancers. Genome Res. 21, 555–565.


**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 (10 ng/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 (100 ng/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 by 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 μM CHIR99021 (TOCRIS), 10 μM SU5402 (TOCRIS), and 10 μM 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, ab8895, Lot 659352), H3K9me3 (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 lyses at 55°C overnight in 300–600 μl 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/μl DNase-free RNase (Roche) and 1 μg/μl proteinase K (NEB). After extraction with an equal volume of phenol:chloroform:isopropanol alcohol (25:24:1; Invitrogen) and addition of 0.5 μl (20 μg/μl) 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 μl 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 μg) 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 μl) 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 for the end-repair to avoid artificially unmethylated sites. Adenylation was performed for 30 min. at 37°C in 50 μl 1x Klenow buffer containing 0.2 mM dATP and 15 units Klenow exonuclease (NEB). Adenylated DNA fragments and methylated paired-end adapters (purchased from ATDBio) were incubated overnight at 16°C in a 50 μl reaction containing 5,000 Ci ¹⁴C and washed with 70% ethanol. DNA was re-suspended in 30–100 μl 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).
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 ≥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 ≤ 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), 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 ≤ 0.05, Fisher’s exact test) different DNA methylation levels of at least 0.1.

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
and computed the enrichment over the WCE. Only windows enriched at a significance level below \( p < 10^{-3} \) as well as the WCE. Next, we used the Poisson model proposed in Mikkelsen et al. (2010) to determine nominal p-value of enrichment insert size extended sequence tags whose midpoint was located within the window of interest for the ChIP-Seq track of interest K3K4me1 using 400bp windows. To compute the enrichment statistics on the window level, we determined the number of unique

This analysis was conducted separately for two groups, 1. H3K27ac, H3K4me3 using 200bp windows and H3K27me3, H3K9me3, retaining all H3K4me3 & H3K27ac regions but merging or splitting enriched H3K4me1 regions.

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 \( \mu \)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 Library Construction

DNA was first subjected to end-repair in a 30-\( \mu \)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-\( \mu \)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 \( \mu \)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, K3K4me1 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 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 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 \( \geq 3 \) 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 \( \geq 2.5 \) fold at a significance level of \( \rho \leq 0.01 \) 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 (Kunarso 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) 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) 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 \( m \)
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 $\geq 1.2$.

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 $\geq 1.2$. To that end we multiplied the motif enrichment score for the cell type of interest with the log2 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 5G.

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 ranked ordered the motifs. For each cell type, we then report the top 20 enriched motifs in Figure 6G.

SUPPLEMENTAL REFERENCES


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

(B) Composite plots of DNAme 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 DNAme 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.
Figure S6. HUES64-Derived Hepatoblast-like Population, Related to Figure 6
Immunofluorescent HNF4α stain of hepatoblast-like population after 10 days of differentiation (10x, scale bar = 100 μm).
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.