Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver

Alayne L. Brunner,1,5 David S. Johnson,1,5,6 Si Wan Kim,1,5 Anton Valouev,2 Timothy E. Reddy,3 Norma F. Neff,1 Elizabeth Anton,1 Catherine Medina,1 Loan Nguyen,1 Eric Chiao,1 Chuba B. Oyolu,1 Gary P. Schroth,4 Devin M. Absher,3 Julie C. Baker,1,8 and Richard M. Myers1,7,8

1Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA; 2Department of Pathology, Stanford, California 94305, USA; 3HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA; 4Illumina, Inc., San Diego, California 92121, USA

To investigate the role of DNA methylation during human development, we developed Methyl-seq, a method that assays DNA methylation at more than 90,000 regions throughout the genome. Performing Methyl-seq on human embryonic stem cells (hESCs), their derivatives, and human tissues allowed us to identify several trends during hESC and in vivo liver differentiation. First, DNA methylation changes at a minimal number of assayed regions, both in vitro and in vivo (2%–11%). Second, in vitro hESC differentiation is characterized by both de novo methylation and demethylation, whereas in vivo fetal liver development is characterized predominantly by demethylation. Third, hESC differentiation is uniquely characterized by methylation changes specifically at H3K27me3-occupied regions, bivalent domains, and low density CpG promoters (LCPs), suggesting that these regions are more likely to be involved in transcriptional regulation during hESC differentiation. Although both H3K27me3-occupied domains and LCPs are also regions of high variability in DNA methylation state during human liver development, these regions become highly unmethylated, which is a distinct trend from that observed in hESCs. Taken together, our results indicate that hESC differentiation has a unique DNA methylation signature that may not be indicative of in vivo differentiation.


During embryonic development, somatic cells are specified through gene expression programs to give rise to all the tissues in the body. Human embryonic stem cells (hESCs) have become an important tool for studying development because they can be differentiated in vitro toward many specific cellular types, which require a precise program of regulation of gene expression (Fritsch and Singer 2008). DNA methylation at CpG dinucleotides has long been considered a key mechanism of transcriptional regulation (Holliday and Pugh 1975; Razin and Cedar 1991; Reik et al. 1999; Siegfried et al. 1999; Fuks et al. 2003; Hashimshony et al. 2005; Zhang et al. 2005; Reik 2007; Esteller 2008; Tiwari et al. 2008). However, recently, regions of methylated DNA have been correlated with the tissue-specific expression of several genes and with active coding regions across the genome (Kumar and Biswas 1988; Ngo et al. 1996; Grunau et al. 2000; Imamura et al. 2001; Krot et al. 2001; Kusui et al. 2001; Futscher et al. 2002; Song et al. 2005; Butta et al. 2006; Da et al. 2006; Fujii et al. 2006; Lavelle et al. 2006; Douet et al. 2007; Hellman and Chess 2007; Kitamura et al. 2007; Shen et al. 2007; Zilberman et al. 2007; Yagi et al. 2008). Additionally, rather than a permanent mark, methylation has been shown to be dynamic, capable of temporally changing at gene promoters (Weiss and Cedar 1997; Kangaspeska et al. 2008; Metivier et al. 2008). This complex range of DNA methylation is known to target the inactive X chromosome, transposons, imprinted loci, promoters, and coding regions (Goll and Bestor 2005; Chang et al. 2006; Rollins et al. 2006; Mandrioli 2007; Weber et al. 2007; Sha 2008; Suzuki and Bird 2008). Because of numerous studies showing it associated with heterochromatin, DNA methylation has gained a reputation as a permanent silencing mark (Jaenisch 1997; Walsh et al. 1998; Bird and Wolfe 1999; Siegfried et al. 1999; Fuks et al. 2003; Hashimshony et al. 2003; Zhang et al. 2005; Reik 2007; Esteller 2008; Tiwari et al. 2008).
targets has raised new questions about its role in early development, particularly in regulating gene expression during tissue differentiation (Suzuki and Bird 2008).

The global distribution of DNA methylation patterns in human tissues is only beginning to be uncovered, and its role during tissue differentiation in early human development remains uncharacterized (Rakyan et al. 2004; Bibikova et al. 2006; Eckhardt et al. 2006; Keshet et al. 2006; Rollins et al. 2006; Schumacher et al. 2006; Hellman and Chess 2007; Ladd-Acosta et al. 2007; Shen et al. 2007; Weber et al. 2007; Bollati et al. 2008; Illingworth et al. 2008; Rauch et al. 2008; Kaminsky et al. 2009). Deconstructing the functional role for DNA methylation in development has remained a challenge because of the large number of potential CpG targets in the genome (Suzuki and Bird 2008). Here, we present a new technique, Methyl-seq, that assays more than 250,000 methyl-sensitive restriction enzyme cleavage sites, which, when grouped, represent more than 90,000 regions in the nonrepetitive human genome; 35,528 of these regions are within annotated CpG islands, while the remaining 55,084 non-CpG island regions are distributed across the genome in promoters, genes, and intergenic regions. We used Methyl-seq to assay the DNA methylation status of these regions in human ES cells, hESC-derived cells, and fetal and adult liver.

Results

Methyl-seq as a technique for identifying methylation patterns in the genome

Methyl-seq combines DNA digestion by a methyl-sensitive enzyme with next-generation (next-gen) DNA sequencing technology. It is based upon the historically most robust assay for DNA methylation—methylation-sensitive restriction enzymes (Cedar et al. 1979). Here, we used the isoschizomers HpaII and MspI, which recognize the same cleavage site (5'-CCGG-3'); MspI digests at the site regardless of methylation status, while HpaII digests only unmethylated sites. In Methyl-seq, genomic DNA is digested with HpaII or MspI, ligated to next-gen sequencing adapters, and subjected to size selection to isolate small gel fragments (~100–350 bp), including the adapters, corresponding to genomic fragments ~30–260 bp in length), to generate a fragment library for next-gen sequencing (Fig. 1A; Supplemental Fig. 2; Supplemental Tables 1, 2; Methods). The fragment library is sequenced to obtain millions of sequencing tags that are ~25–35 bp in length (~3 × 10^6 tags per HpaII library and 1 × 10^7 tags for the MspI library). Because MspI digests at both methylated and unmethylated sites where HpaII is blocked from digesting methylated sites, sequence tags present in MspI libraries but not in HpaII libraries are derived from methylated regions. Conversely, sequence tags that occur in HpaII libraries come from at least partially unmethylated regions.

Analysis of these sequence tags is described in the Supplemental material, but briefly, aligned sequence reads are matched to methylated CpG sites, and Illumina beta values are calculated (Fig. 1B). Receiver operator characteristic (ROC) curve generated from Methyl-seq and Illumina Infinium data. The area under the curve is 0.944.

Figure 1. The Methyl-seq method and validation. (A) Workflow for the Methyl-seq assay. (B) Comparison of the number of Methyl-seq tags and Illumina beta values at 160 CpGs for four samples: HCT116, H9 hESC, H9 endoderm, and adult liver. HCT116 was used in Methyl-seq development and validation. There is good correlation between methylated calls (zero or one Methyl-seq tags and high beta values) and, likewise, unmethylated calls (low beta values and high numbers of sequencing tags) (for binary comparison, see Supplemental Fig. 7, P-values <10^-8). (C) Receiver operator characteristic (ROC) curve generated from Methyl-seq and Illumina Infinium data. The area under the curve is 0.944.

DNA methylation in differentiated hESCs and fetal liver
islands (55,084 of 90,612). Nevertheless, Methyl-seq is still able to effectively assay 65% of all annotated CpG islands in the human genome (17,945 of 27,639) (Bird 1987; Gardiner-Garden and Frommer 1987). In addition, the assayable regions cover a range of CpG densities, as illustrated with the overlap of Methyl-seq regions with promoters previously defined as high-CpG promoters (HCPs), intermediate-CpG promoters (ICPs), and low-CpG promoters (LCPs) (Table 2; Weber et al. 2007). This suggests that Methyl-seq can assay a range of genomic elements, allowing a broader survey of regions than classic methylation studies limited to CpG islands and promoters.

In each Methyl-seq experiment, we infer the methylation status from the presence or absence of HpaII tags at each digestion site. We called regions with an average tag count of greater than one tag per digestion site as unmethylated and called the remaining regions as methylated. We validated these methylation calls by three separate techniques: Illumina Infinium HumanMethylation27 genotyping, qPCR, and MeDIP (Weber et al. 2005; Gebhard et al. 2006; Zhang et al. 2006). For the sake of brevity, we describe here only the results obtained by the bisulfite genotyping, as such methods are often considered a gold standard for which the nature and extent of epigenetic regulation of lineage differentiation during human development is mostly unknown. However, mouse studies suggest a critical role for DNA methylation during development (Okano et al. 1999). To address how DNA methylation contributes to human somatic differentiation on a genome-wide level, we applied Methyl-seq to developing human tissues and cells (Fig. 2A; Supplemental Table 1). In our initial experiments, we asked how DNA methylation patterns change during the differentiation of hESCs into definitive endoderm and other derivatives. To this end, we generated a homogeneous (98% pure) population of definitive endoderm from H9 hESCs according to previously reported protocols, which include FACs sorting with CXCR4 after a 5-d culture regime in the presence of activin (see Methods) (D’Amour et al. 2005). These sorted cells express high levels of endoderm markers, including SOX17, FOXA2, GATA4, GATA6, and CXCR4 (Fig. 2B). Applying Methyl-seq, we

DNA methylation changes minimally during in vitro differentiation

The nature and extent of epigenetic regulation of lineage differentiation during human development is mostly unknown. However, mouse studies suggest a critical role for DNA methylation during development (Okano et al. 1999). To address how DNA methylation contributes to human somatic differentiation on a genome-wide level, we applied Methyl-seq to developing human tissues and cells (Fig. 2A; Supplemental Table 1). In our initial experiments, we asked how DNA methylation patterns change during the differentiation of hESCs into definitive endoderm and other derivatives. To this end, we generated a homogeneous (98% pure) population of definitive endoderm from H9 hESCs according to previously reported protocols, which include FACs sorting with CXCR4 after a 5-d culture regime in the presence of activin (see Methods) (D’Amour et al. 2005). These sorted cells express high levels of endoderm markers, including SOX17, FOXA2, GATA4, GATA6, and CXCR4 (Fig. 2B). Applying Methyl-seq, we

### Table 1. Distribution of the 90,612 Methyl-seq regions by chromosome (hg18)

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Regions</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>7946</td>
</tr>
<tr>
<td>chr2</td>
<td>5923</td>
</tr>
<tr>
<td>chr3</td>
<td>3665</td>
</tr>
<tr>
<td>chr4</td>
<td>3212</td>
</tr>
<tr>
<td>chr5</td>
<td>3715</td>
</tr>
<tr>
<td>chr6</td>
<td>3765</td>
</tr>
<tr>
<td>chr7</td>
<td>4914</td>
</tr>
<tr>
<td>chr8</td>
<td>3794</td>
</tr>
<tr>
<td>chr9</td>
<td>4331</td>
</tr>
<tr>
<td>chr10</td>
<td>4315</td>
</tr>
<tr>
<td>chr11</td>
<td>5060</td>
</tr>
<tr>
<td>chr12</td>
<td>3847</td>
</tr>
<tr>
<td>chr13</td>
<td>1911</td>
</tr>
<tr>
<td>chr14</td>
<td>2813</td>
</tr>
<tr>
<td>chr15</td>
<td>2536</td>
</tr>
<tr>
<td>chr16</td>
<td>5167</td>
</tr>
<tr>
<td>chr17</td>
<td>5778</td>
</tr>
<tr>
<td>chr18</td>
<td>1814</td>
</tr>
<tr>
<td>chr19</td>
<td>6674</td>
</tr>
<tr>
<td>chr20</td>
<td>3129</td>
</tr>
<tr>
<td>chr21</td>
<td>1505</td>
</tr>
<tr>
<td>chr22</td>
<td>3055</td>
</tr>
<tr>
<td>chrX</td>
<td>1717</td>
</tr>
<tr>
<td>chrY</td>
<td>26</td>
</tr>
</tbody>
</table>

### Table 2. The breakdown of the 90,612 Methyl-seq regions into specific annotated genomic elements and binding locations

<table>
<thead>
<tr>
<th>Category</th>
<th>Regions</th>
</tr>
</thead>
<tbody>
<tr>
<td>CpG island</td>
<td>35,528</td>
</tr>
<tr>
<td>Non-CpG island</td>
<td>55,084</td>
</tr>
<tr>
<td>Promoter (~1000 bp to TSS)</td>
<td>13,575</td>
</tr>
<tr>
<td>5’ UTR</td>
<td>4366</td>
</tr>
<tr>
<td>Coding exon</td>
<td>11,986</td>
</tr>
<tr>
<td>Intron</td>
<td>28,035</td>
</tr>
<tr>
<td>3’ UTR</td>
<td>1653</td>
</tr>
<tr>
<td>Intergenic region</td>
<td>30,997</td>
</tr>
<tr>
<td>Bivalent</td>
<td>5200</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>27,537</td>
</tr>
<tr>
<td>H3K27me3</td>
<td>2366</td>
</tr>
<tr>
<td>Neither</td>
<td>55,509</td>
</tr>
<tr>
<td>HCP</td>
<td>14,330</td>
</tr>
<tr>
<td>ICP</td>
<td>1254</td>
</tr>
<tr>
<td>LCP</td>
<td>385</td>
</tr>
<tr>
<td>7× Regulatory potential</td>
<td>75,297</td>
</tr>
</tbody>
</table>

CpG island coordinates were obtained from the UCSC genome browser where they were calculated for hg18 by using the definition from Gardiner-Garden and Frommer (1987), where islands have a GC content of 50% or greater, a length greater than 200 bp, and a ratio greater than 0.6 of observed number of CG dinucleotides to the expected number on the basis of the number of Cs and Cs in the segment (1987). Coordinates for promoters and other genomic elements were calculated using the transcription start sites and gene annotations from the Known Genes file, compiled by the UCSC Genome Browser. Coordinates for strong CpG islands or high-CpG promoters (HCPs), weak CpG islands or intermediate-CpG promoters (ICPs), and sequences with no local enrichment of CpGs or low-CpG promoters (LCPs) were from Weber et al. (2007). Coordinates for regions of 7× regulatory potential, or regions (average score >0.5) conserved in human, chimpanzee, macaque, mouse, rat, dog, and cow, were obtained from King et al. (2005). Histone ChIP-chip binding data was from Pan et al. (2007), where “bivalent” regions were regions that overlapped with both H3K4me3 and H3K27me3 peaks. Methyl-seq regions overlapping with H3K4me3 peaks only were called H3K4me3; regions overlapping with H3K27me3 peaks only were called H3K27me3; and all unassigned regions were defined as “neither” (Pan et al. 2007).
compared the DNA methylation profiles of the undifferentiated H9 hESC line and CXCR4-sorted definitive endoderm cells (Supplemental Table 2; Libraries 5, 6, respectively). Although these hESCs and definitive endoderm populations have markedly different genome-wide transcript profiles (Fig. 2C), we found that, upon surveying a total of 90,612 regions, only a small fraction of the regions change in their methylation state following differentiation of H9 hESC to definitive endoderm: 1017 (1.1%) become methylated and 682 (0.8%) become demethylated (see Supplemental Table 5C). These results suggest that significant methylation changes are not needed during endoderm formation to control gene expression, which is consistent with previous evidence showing that embryos lacking de novo methyltransferases are viable during gastrulation (Okano et al. 1999).
Based on these results, we hypothesized that a cell type representing a later stage of development might exhibit more differences in DNA methylation than we observed between endoderm and hESCs. To test this notion, we derived several different cell types from hESCs, including alpha-fetoprotein (AFP)-positive-sorted cells with hepatocyte-like characteristics (Chiao et al. 2008), AFP-negative-sorted cells with highly heterogeneous mixture of cell types, embryoid bodies (EBs), and heterogeneous populations of cells derived from EBs after 20 d of plating (EB-derived) (Fig. 2A). By using Methyl-seq, we found that DNA methylation changed at only a small percentage (2%–5%) of regions during hESC differentiation, regardless of heterogeneity (see Fig. 3 for example of regional methylation data). In the most homogeneous population (the AFP positive), we found 2628 differences (2.9%) compared with the naïve hESC state (Supplemental Table 5E). Similar numbers were observed in EBs, EB-derived, as well as in the AFP-negative populations. More than half of the changes (1732) are shared between the derived cell types, even though they are comprised of unique cell types (Fig. 4B; Supplemental Table 5F,H). These same changes occur in endoderm cells and EBs (see Fig. 4A,C). Furthermore, of the 1011 changes common to both AFP differentiation and endoderm differentiation, only 182 are absent in the differentiation toward AFP-negative cells, indicating that a similar group of changes is occurring in hESC differentiation regardless of the derived cell type (Supplemental Table 5D,J). To test whether these changes are specific to the particular hESC line used, H9, or represent a general hESC trend, we also tested several of these differentiation regimes on another hESC line, BG02. We found that many of the changes occurring upon differentiation are common between samples derived from either hESC line (Supplemental Table 5C–W). The high frequency of overlapping methylation changes in the derived heterogeneous populations, which are frequently composed of beating cardiomyocytes and other tissue types, strongly suggests that similar changes in methylation occur during hESC differentiation regardless of induced cell type. Only a small fraction (<5%) of the regions that we assayed by Methyl-seq showed changes in DNA methylation during in vitro differentiation. We asked whether this was unique to cells derived from hESCs or whether it was also true for in vivo human fetal differentiation. Thus, we measured human 11- and 24-wk fetal, as well as adult liver samples for their DNA methylation patterns by Methyl-seq (Supplemental Table 5C–W; Libraries 15–17, respectively). Because the fetal liver also plays a role in hematopoiesis throughout gestation, it is a heterogeneous combination of erythrocytes and hepatocytes (MacSween et al. 2007). Conversely, the adult liver is composed of ~78% hepatocytes, making it one of the more homogeneous adult organ systems (Blouin et al. 1977; MacSween et al. 2007). Although we were concerned about the heterogeneity of the fetal liver, we were not able to separate these populations due to the limiting nature of the tissue (Dan et al. 2006; Tosh and Strain 2005).
However, between 11 and 24 wk, the relative proportions of cell types remain similar, so we can still assess changes occurring between these two stages of liver development. When we applied Methyl-seq to these tissues, we found that 6034 regions changed their methylation status between 11- and 24-wk fetal liver, 10,237 regions changed between 11-wk fetal and adult liver, and 9359 changed between 24-wk fetal and adult liver (Fig. 3; Supplemental Table S5A–AI). We also observed 12,934 regions with methylation differences between adult liver and hESCs, of which 5173 were common to fetal liver at 11 wk and 7215 were common to fetal liver at 24 wk (Supplemental Table S5Z,AB,AC). These results suggest that although these tissues are heterogeneous, the methylation patterns seen for all stages of liver have significant overlap and indicate progression of methylation changes over time in vivo.

Genomic context guides DNA methylation regardless of tissue type

We compared the DNA methylation status results of the 90,612 assayable regions in the distinct cells and tissues tested in our study. By comparing the relative percentage of methylation in these regions between samples, we found that hESCs and their derivatives have a larger fraction of methylated regions than do the in vivo fetal and adult tissues. The fraction of DNA regions methylated in hESCs and their derivatives (54%–59%) is significantly higher than in that of fetal liver, adult liver, and pancreas (40%–49%), \( P = 5.5 \times 10^{-4} \), Mann-Whitney U-test; Supplemental Table 4A; data not shown). However, the relative percentage of methylation in the hESC derivatives does not appear to change substantially when compared with the initial hESC methylation levels, regardless of time in culture or heterogeneity of the population (see Supplemental Table 4A). Conversely, the global methylation percentage appears to decrease during liver development, with the 11-wk liver having the highest percentage of DNA methylation (49%) and the adult liver the lowest (43%), suggesting that in vivo liver development is biased toward demethylation events (Supplemental Table 4A).

We assigned the 90,612 assayable regions into distinct categories of genomic elements and then analyzed the Methyl-seq data in each sample according to each type of functional element. The annotated genomic elements are (1) predicted CpG islands (which overlap extensively with other classifications); (2) promoters and 5’-UTRs; (3) exons; (4) introns; (5) intergenic regions; (6) high density CpG promoters (HCPs); (7) intermediate density CpG promoters (ICPs); (8) low density CpG promoters (LCPs); (9) regions with predicted regulatory potential (7× RP); and (10) regions previously shown to be bound in hESCs by specific modified histone H3 proteins, including H3K4me3 (referred to as univalent regions), H3K27me3, and H3K4me3/H3K27me3 (referred to as bivalent regions) (Table 2; King et al. 2005; Bernstein et al. 2006; Pan et al. 2007; Weber et al. 2007). In Supplemental Table 4, B through N, we present the degree of DNA methylation for each genomic element in each sample. We found that, in every sample analyzed, regions that include CpG islands, promoters and both univalent and bivalent H3K4me3 domains are more highly unmethylated (less than 15% of regions methylated), whereas regions outside of CpG islands, including those occupied by H3K27me3; exons; introns; 3’-UTRs; and intergenic regions are more highly methylated (ranging from 39% to more than 78% of the sites are methylated; Fig. 5). Thus, different genomic functional elements vary in their degree of DNA methylation.

H3K4me3 and H3K4me3/H3K27me3 binding correlates highly with unmethylated promoters

Transcriptional promoters are critical components of gene regulation. To understand how DNA methylation is regulated at these elements and because H3K4me3 and H3K27me3 binding are correlated with distinct methylation status in hESCs (i.e., unmethylated and methylated, respectively), we focused on a comparison of DNA methylation and these binding events in promoter regions. Examining HCPs and LCPs with the occupancy of H3K4me3 and H3K27me3, we found that HCPs are highly enriched (82%) for H3K3me3 binding. Of these HCP univalent regions, 98% are unmethylated. When we examined the small fraction of these HCP regions that are methylated, we observed that most of them are located on the edge of unmethylated CpG islands. Therefore, most of these promoters show some unmethylated portion but are not fully unmethylated across their entirety. Although univalent domains are much less frequent in LCPs (13%), these few LCP univalent regions are also highly unmethylated (80%). This is in stark contrast to the LCP regions lacking both H3K4me3 and H3K27me3, where only 9% are unmethylated. Bivalent domains in both HCPs and LCPs show a similar trend; HCP bivalent regions are 95% unmethylated, and LCP bivalent regions are 71% unmethylated. Thus, univalent and bivalent domains are both strong indicators of regions of unmethylated DNA and are particularly informative at predicting unmethylated LCPs. This is not a hESC-specific phenomenon, as we observed the same trend in 11-wk fetal liver and in all hESC-derived cells, and a similar result was recently reported for mouse tissues (Meissner et al. 2008).
hESC differentiation is defined by significant changes in methylation status at LCPs and H3K27me3 and H3K4me3 / H3K27me3-occupied regions

We next analyzed the DNA methylation changes that occur during differentiation in vitro and in vivo. We divided the 90,612 assayable Methyl-seq regions into those that are different and those that are the same between all samples. Seventy-five percent of the regions analyzed (68,116 of 90,612) showed no changes in DNA methylation in any of the cell types or tissues that we tested; 49% of these unchanged regions were always methylated, while 51% were always unmethylated, in all samples (Supplemental Table 5A). In agreement with the trends described above with other work, the regions that are always methylated in all samples tended to be located in non-CpG islands, exons, introns, intergenic regions, and regions not bound by either H3K4me3 or H3K27me3 (Fig. 5). In addition, the regions that are always unmethylated tended to be located in CpG islands, promoters (including HCPs and ICPs), and H3K4me3-bound regions. The remaining 25% (22,496) of the assayable regions show at least one difference in methylation status between all samples (Supplemental Table SB).

In Supplemental Table 5C through AI, we provide details of the relative changes of these regions between hESCs, hESC-derived cells, and tissues. We found that, during hESC differentiation, only 120 regions change uniquely in all hESC-derived samples (Supplemental Table SW). Seventy-seven (64%) of these regions become methylated, and 43 (36%) become unmethylated. Seven of these 120 regions are transcriptional promoters; interestingly, all of these promoters become methylated during hESC differentiation, in contrast to the nonmethylated state of most promoters in the genome (see Supplemental Table 5A).

We next identified regions in the genome that changed in DNA methylation state during differentiation. We observed an average of 1.3% of the regions undergoing de novo methylation and 1.6% undergoing demethylation (see Fig. 6A, total). Dividing these changes into different genomic contexts, we found that this trend is seen in all categories except for LCPs, H3K27me3, and bivalent domains. These elements are highly variable (Fig. 6A). H3K27me3-occupied regions and bivalent domains become substantially de novo methylated in the majority of differentiated profiles (1.4 × 10^−27 < P < 1.2 × 10^−6 and 1.8 × 10^−23 < P < 6.5 × 10^−11, respectively; hypergeometric tests). LCPs are also extremely variable in DNA methylation status during hESC differentiation, but unlike bivalent and H3K27me3-occupied regions, the changes are significantly enriched in both directions (1.5 × 10^−8 < P < 1.8 × 10^−3; Fig. 6A). Therefore, within the regions whose DNA methylation state changes during hESC differentiation, LCPs, bivalent domains, and H3K27me3 marks are significantly more represented than any other genomic region, indicating that these regions are “hotspots” for either de novo methylation or demethylation in hESC differentiation. Although it is unknown if these differentially methylated regions are functionally important, these regions may be prime targets for the regulation of unique cell fates.

Fetal liver differentiation characterized by demethylation

While hESC changes are defined by a combination of de novo methylation and demethylation enriched at particular genomic elements, fetal liver differentiation is characterized by substantial demethylation (Fig. 6B–D). For example, there are 6034 DNA methylation differences between 11- and 24-wk human fetal liver; 76% of these become demethylated by 24 wk (Fig. 6B; Supplemental Table 5AG). This finding is further supported by the analysis of regions that change uniquely within the developing and adult liver, but not within hESC or hESC derivatives (Fig. 6D; Supplemental Table 5AF). Of these 3050 genomic regions, almost all (99.7%) become demethylated. Similar to our findings with
hESC differentiation, significant methylation changes occur in LCPs, bivalent regions, and H3K27me3-bound regions. However, liver development consists of significantly fewer methylation changes in bivalent regions, suggesting these regions are far less variable during in vivo differentiation compared to in vitro differentiation ($1.5 \times 10^{-11} < P < 3.8 \times 10^{-66}$, hypergeometric test).

Furthermore, whereas hESC differentiation showed significant numbers of de novo methylation events at H3K27me3-bound regions, the in vivo differentiation shows significant numbers of demethylation events occurring in these regions ($1.2 \times 10^{-11} < P < 8.1 \times 10^{-7}$, hypergeometric test). We obtained similar results for LCPs; these elements were significantly variable for methylation changes in both directions during hESC development. However, during fetal liver development, LCP regions are significantly enriched for demethylation ($P = 1.7 \times 10^{-13}$, hypergeometric test). Thus, while hESCs show variable methylation patterns (including both de novo methylation and demethylation) during differentiation, liver development is specifically characterized by demethylation during differentiation. However, the changes observed for both in vitro and in vivo differentiation occur in similar genomic contexts, including LCPs, bivalent regions, and H3K27me3-bound regions. This suggests that the DNA methylation changes during hESC differentiation are not an accurate reflection of the changes in DNA methylation in vivo, at least in liver development.

**BG02 and H9 hESC lines share 49 specific DNA methylation regions**

hESC lines exhibit different phenotypic responses to differentiation, and each line has a long history of being cultured under distinct conditions. In this study, we used two different lines, BG02 and H9, and the observations described above are consistent in both of these lines. Comparison of the assayable DNA methylation regions between BG02 and H9 revealed a larger number of differences between these two lines (2913) than occur between the naive hESCs and their derivatives (see Supplemental Table S5 vs. Supplemental Table SC–SU). Most striking is that 78% of the differences between the two hESC lines occur where H9 is methylated and BG02 is not. These changes do not appear to influence gene expression, as the hESC expression profiles are highly similar. We suggest that these changes are the result of exposure to different culturing conditions and derivation strategies. Comparison

**Figure 6.** Percent of regions showing methylation changes by genomic feature. (A) Mean percentage of regions that gain or lose methylation in in vitro-differentiated cells compared with hESCs ($y$-axis), grouped by genomic features ($x$-axis). Categories are total changes, changes in CpG islands (CGI), promoters and 5' UTRs (promoter), exons, introns, 3' UTRs, intergenic regions (intergen), HCPs, ICPs, LCPs, 7× regulatory potential (7× RP), H3K4me3/H3K27me3-occupied regions (bivalent), H3K4me3-occupied regions (H3K4), and H3K27me3-occupied regions (H3K27). Error bars, maximum and minimum values across five differentiated sample comparisons: H9 hESC vs. H9 Endoderm (5–6), H9 hESC vs. H9 AF+ hESC-derived cells (5–7), H9 hESC vs. H9 Embryoid bodies (5–9), H9 hESC vs. H9 EB-derived cells (5–11), BG02 hESC vs. BG02 EB-derived cells rep1 and rep2 (12–13,14). (B) Percentage of regions that gain or lose methylation in 11-wk fetal liver compared with 24-wk fetal liver ($y$-axis), grouped by genomic features ($x$-axis). Data are from the sample comparison between 11-wk fetal liver and 24-wk fetal liver (15–16). (C) Mean percentage of regions that gain or lose methylation during tissue differentiation ($y$-axis), grouped by genomic features ($x$-axis). Error bars, maximum and minimum values across three tissue sample comparisons: 11-wk fetal liver vs. 24-wk fetal liver (15–16), 11-wk fetal liver vs. adult liver (15–17), 24-wk fetal liver vs. adult liver (16–17). (D) Mean percentage of regions that gain or lose methylation in tissues compared with hESCs ($y$-axis), grouped by genomic features ($x$-axis). Error bars, maximum and minimum values across three tissue sample comparisons: H9 hESC and BG02 hESC vs. 11-wk fetal liver (5,12–15), H9 hESC and BG02 hESC vs. 24-wk fetal liver (5,12–16), H9 hESC and BG02 hESC vs. adult liver (5,12–17). Library numbers are in parentheses.
of both BG02 with H9 to all other hESC-derived cells and in vivo liver samples revealed 49 hESC-specific regions that changed between all samples but were similar between BG02 and H9 (Supplemental Table 5A). Twenty-three of these regions unique to BG02 and H9 surround genes, in coding regions, promoters, or introns; only one lies within a CpG island. The remaining 26 regions are in intergenic sequences and are mostly unmethylated (70%) in the hESCs. Several of the genes flanked by these regions are of notable developmental importance, including NEAT2 and IGF2. Thus, although there are many differences in the DNA methylation signatures between BG02 and H9, there are only 49 regions showing shared methylation patterns between these cell lines. These regions may be useful biomarkers for the undifferentiated state.

Discussion

During mouse embryogenesis, a genome-wide loss of DNA methylation is observed in the preimplantation embryo (Monk et al. 1987). At the blastocyst stage, de novo methylation occurs within the inner cell mass, but not the trophoderm (Santos et al. 2002). By implantation, DNA methylation in the genome is greatly elevated compared with preimplantation levels and is thought to be maintained throughout subsequent differentiation in somatic lineages (Jaenisch and Jahnner 1984; Razin and Cedar 1991; Kafri et al. 1992; Jaenisch 1997). This logic has largely been extrapolated to include human development; although due to ethical restraints and tissue procurement difficulties, it remains unknown whether a similar pattern of DNA methylation occurs in the early human embryo. Furthermore, how DNA methylation changes during both human and mouse development in somatic lineages is not understood (Reik et al. 2001; Latham et al. 2008; Suzuki and Bird 2008). Several studies have shown that human and mouse somatic tissues have distinct DNA methylation patterns, which is consistent with DNA methylation guiding the repression or activation of tissue-specific loci. However, many of these studies measured only small numbers of loci and did not study the changes in the differentiating cells through developmental time (Yeivin and Razin 1993; Shiota et al. 2002; Kremenskoy et al. 2003; Shiota 2004). Recently, several groups have applied genomic techniques toward understanding DNA methylation in ES cells and differentiating tissues. Fouse et al. (2008) reported genome-wide mapping of DNA methylation patterns in proximal promoter regions in mouse embryonic stem cells (mESCs). They found that binding of Pou5f1 (Oct4), Nanog, Polycomb group proteins (PcG), and H3K27 trimethylation is correlated with DNA methylation of promoters in mESC cells (Fouse et al. 2008). Furthermore, Meissner et al. (2008) recently developed a method based on bisulfite treatment and next-gen sequencing (RRBS) and applied it to measure DNA methylation in mESCs, neural progenitor cells derived from mESCs, and several mouse tissues. Their report provided insight into how DNA methylation becomes modified during the differentiation of mESCs into neural tissue, and demonstrated that CpGs are dynamic marks that undergo specific changes during differentiation. Furthermore, they showed that the DNA methylation status can be predicted based upon association of histone methylation patterns (Meissner et al. 2008). Based upon these studies and others, a clearer picture is emerging of how developmental processes are influenced by DNA methylation. However, there remain many questions that need to be addressed, particularly in human ESCs and within specific developmental lineages, which are likely to have unique DNA methylation signatures (Weiss and Cedar 1997; Shiota 2004; Suzuki and Bird 2008).

Although several recent studies have examined DNA methylation in individual human tissues by using different genomic technologies that assess promoters and CpG islands (Eckhardt et al. 2006; Ladd-Acosta et al. 2007; Shen et al. 2007; Weber et al. 2007; Illingworth et al. 2008; Rauch et al. 2008), our study provides the most comprehensive portrait to date of DNA methylation changes at a wide array of genomic elements during human cellular differentiation, including in hESCs and in vivo fetal tissues. It demonstrates that in vivo liver differentiation is characterized by demethylation through time; a theory that has been generally proposed for somatic differentiation but that has little supportive evidence (Weiss and Cedar 1997). Furthermore, these demethylation events are particularly enriched at H3K27me3-bound regions and LCPs, suggesting that these regions are key mediators in tissue-specific gene expression. However, it must also be noted that the methylation differences observed between the liver tissues of different developmental stages were from different individuals, and it therefore remains a possibility that the differences observed were due to individual methylation variation and not a consistent developmental regime. More samples will need to be profiled in the future to address this issue. Our work also characterizes the methylation changes occurring in several hESC differentiation regimes, and using this extensive data set, we propose that the changes are composed of both de novo methylation and demethylation at particular genomic elements, including H3K27me3-bound regions and LCPs. We also found conclusively that hESCs are more methylated than other tissues, especially upon differentiation. This finding is consistent with the previous report that cultured cells show increased DNA methylation relative to primary tissues (Antequera et al. 1990), but contradicts reports that naive hESCs and other mammalian cell lines capable of in vitro differentiation are hypomethylated (Razin et al. 1984, 1988; Young and Tilghman 1984; Jost and Jost 1994; Zvetkova et al. 2005; Bibikova et al. 2006).

The approach we describe here, Methyl-seq, offers a new tool for measuring the DNA methylation status of a very large number of regions of the human genome and any other vertebrate genome whose DNA sequence is known. Several other methods have been recently described, and all, including Methyl-seq, have strengths and weaknesses (Kremenskoy et al. 2003; Weber et al. 2005, 2007; Keshet et al. 2006; Zhang et al. 2006; Shen et al. 2007; Illingworth et al. 2008; Meissner et al. 2008). Methyl-seq allows measurement of the DNA methylation status of more than 90,000 regions, including a significant number that are not covered by any other single technique. The method is sensitive, highly specific with very low background, reproducible, and simple to execute. It is relatively inexpensive, requiring fewer reads on next-gen sequencers than RRBS (Meissner et al. 2008), and is likely to become less expensive as the sequencing technologies become more efficient and widespread. Despite these advantages, Methyl-seq has some key limitations that are important to keep in mind. First, Methyl-seq assays only the CpGs in a specific subset of HpaII restriction enzyme cleavage sites (those that appear within 35–75 bp of each other in the human genome). This significantly limits the number of CpGs assayed but nevertheless provides a snapshot of regional methylation patterns. Additionally, because Methyl-seq is sensitive in detecting small numbers of HpaII tags, this creates a problem measuring methylation quantitatively. While this may likely be improved in the future, we limited the analysis in this current study to binary calls of "methylated" and "unmethylated."
DNA methylation in differentiated hESCs and fetal liver

Because Methyl-seq currently uses “unmethylated” to describe any region showing HpaII fragments in the collection of next-gen sequencing reads, the quantitative methylation state of any individual DNA molecule is lost. This particularly limits the ability of Methyl-seq to detect partial methylation or imprinted loci. This is not likely to be a problem for most CpGs because numerous studies have shown that most CpGs are mostly methylated (>80%) or mostly unmethylated (<20%) (Meissner et al. 2008), and Methyl-seq is able to distinguish these quantitative differences. However, a more accurate interpretation of Methyl-seq regions showing HpaII tags may be “not fully methylated” as opposed to “unmethylated” to take into account that Methyl-seq calls are sensitive to any demethylation. Lastly, because the current Methyl-seq protocol relies on small HpaII fragments (grouped HpaII sites) as opposed to single cut sites, Methyl-seq does not assay most of the very low density CpG regions of the genome, including a number of low density CpG promoters. Because most of these promoters are assayed by Illumina’s Infinium HumanMethylation27 platform, combining the methylation results from Methyl-seq and Illumina’s Infinium HumanMethylation27 beadchips will provide a broad, comprehensive survey of methylation targets in the genome. Thus, when deep genomic bisulfite sequencing is not an option, Methyl-seq is a valuable addition to the mixture of emerging methods used to assay the methylation status of CpG dinucleotides in the human genome.

The epigenetic state of undifferentiated stem cells has been extensively studied in recent years, but clearly the next frontier is to understand how this undifferentiated state is modified into specific tissue types (Bernstein et al. 2006; Boyer et al. 2006; Lee et al. 2006; Mikkelsen et al. 2007). Because DNA methylation is key to embryonic survival and differentiation, gaining information about how this process is modified in vivo and in vitro could be central to the future engineering of hESCs for therapeutic purposes (Keller 2005). Future studies involving hESCs should address the importance of the endogenous epigenetic state for tissue function. Further investigation will be required to determine whether DNA methylation changes are similar between somatic tissue function. Further investigation will be required to determine whether DNA methylation changes are similar between somatic tissue function.

Methods

Samples

Twenty samples (Supplemental Table 1) were used to obtain genomic DNA; we also purified RNA from 14 of these (see Supplemental Methods). Eight of the hESC in vitro samples were sorted by FACS prior to nucleic acid extraction.

In vitro differentiation

Definitive endoderm precursors were generated from hESCs by using activin A as previously described (D’Amour et al. 2005). EB-derived cells were obtained by plating clumps of undifferentiated hESCs in suspension to promote the formation of EBs. After 8 d, the EBs were plated onto tissue culture dishes coated with gelatin type A, and further differentiation was continued for 2 wk until a peak level of the hepatic protein AFP was detected. To generate AFP-positive fetal hepatocytes, we developed a hESC line expressing a reporter green fluorescent protein (GFP) from the AFP promoter. After 30 d of differentiation in the presence of FGF, AFP-positive cells were isolated by FACS (Chiao et al. 2008).

Methyl-seq and MeDIPSeq library construction

Methyl-seq and MeDIPSeq libraries were constructed by using the procedures outlined in Supplemental Protocols 1 and 2, respectively. Briefly, in Methyl-seq, 5 μg genomic DNA was digested with HpaII or MspI. Fragments were made into next-gen sequencing libraries by using adapters and reagents from Illumina’s Library Construction Kit. After an initial round of PCR, DNA bands (fragments with adapters) of 100–350 bp were isolated by gel extraction (Supplemental Fig. 2). Size-selected libraries were then PCR-amplified prior to sequencing. For MeDIPSeq, 5 μg genomic DNA was digested with MspI and immunoprecipitated with anti-methylcytosine antibody (Calbiochem); fragments were made into Illumina libraries as described above.

Illumina sequencing and primary analysis

Libraries were sequenced on an Illumina 1G Genome Analyzer sequencing machine, and reads were aligned to the genome (UCSC hg18) by using the standard Illumina sequence analysis pipeline (25 nucleotides, two mismatches). We normalized all HpaII Methyl-seq libraries to contain between 3 million and 3.5 million usable reads (those that align to the genome and possess the CGG HpaII-cut signature on their 5’ ends; Supplemental Table 2; Supplemental Methods). The MspI library that we used contained 10 million usable reads.

Analysis for calling restriction enzyme digestion sites and region methylation status

Usable sequence reads were mapped to CCGG sites predicted in silico. Sites with four or more MspI tags occurring in either the forward or reverse direction were retained in the analysis. These “assayable” sites were then grouped with neighboring sites that were within 35–75 bp. Regions therefore represent any number of digestion sites (between two and 18) that have neighboring sites within 35–75 bp of each other. Methylation calls were made by using HpaII tag data from all assayable cut sites. The larger of either the forward read count or reverse read count for each site was averaged across each region. Regions that had an average of zero or one reads per digestion site were called methylated, and regions with more than one sequence read per site were considered unmethylated.

Illumina Infinium validation

We used Infinium HumanMethylation27 BeadChips (Illumina) to validate the methylation calls obtained from Methyl-seq. Genomic DNA from seven samples corresponding to four Methyl-seq libraries was treated with bisulfite and genotyped. Comparison of Infinium data and Methyl-seq tag counts was performed at 160 CpGs common to both methods. The ROC curve was generated with the ROC package for R. Illumina’s Infinium HumanMethylation27 data were used as the “truth” by dichotomizing the beta value with a threshold of 0.6 (i.e., >0.6 was methylated and <0.6 was unmethylated). Additionally, when the dichotomized Infinium beta values were compared with Methyl-seq data (zero to one tag methylated, >1 tags unmethylated), Fisher exact test on each sample produced P-values <10⁻⁸.
Gene expression

We used HumanRef-8 v2 Expression BeadChips (Illumina) for mRNA expression analysis. Array intensities were rank-invariant normalized, log-transformed, clustered using Cluster 3.0, and visualized by using TreeView (Eisen et al. 1998; Saldanha 2004). Genes were filtered to exclude those found in <30% of the samples.

Acknowledgments

We thank Hunter Shain for assistance with the tissue culture work, Larisa Tsvaler at the Stanford Human Genome Center for performing the mRNA expression experiments, Anning Xiong and Jeffrey Glenn for providing the fetal liver samples, Arend Sidow for helping analyze the data, and Rami Rauch for helping organize and submit the genomic data. This work was supported by a grant from the California Institute of Regenerative Medicine to J.C.B., NIH grant 7U54HG004576 to R.M.M., and funds from the HudsonAlpha Institute for Biotechnology to R.M.M. A.L.B. was funded by a NIH grant to the Stanford Genetics and Developmental Biology Training Program.

References


Received December 1, 2008; accepted in revised form March 6, 2009.