ARTICLES

Human DNA methylomes at base resolution show widespread epigenomic differences

Ryan Lister[1]*, Mattia Pelizzola[1]*, Robert H. Dowen[1], R. David Hawkins[2], Gary Hon[2], Julian Tonti-Filippini[4], Joseph R. Nery[1], Leonard Lee[2], Zhen Ye[2], Que-Minh Ngo[2], Lee Edsall[2], Jessica Antosiewicz-Bourget[5,6], Ron Stewart[5,6], Victor Ruotti[5,6], A. Harvey Millar[4], James A. Thomson[5,6,7,8], Bing Ren[2,3] & Joseph R. Ecker[1]

DNA cytosine methylation is a central epigenetic modification that has essential roles in cellular processes including genome regulation, development and disease. Here we present the first genome-wide, single-base-resolution maps of methylated cytosines in a mammalian genome, from both human embryonic stem cells and fetal fibroblasts, along with comparative analysis of messenger RNA and small RNA components of the transcriptome, several histone modifications, and sites of DNA–protein interaction for several key regulatory factors. Widespread differences were identified in the composition and patterning of cytosine methylation between the two genomes. Nearly one-quarter of all methylation identified in embryonic stem cells was in a non-CG context, suggesting that embryonic stem cells may use different methylation mechanisms to affect gene regulation. Methylation in non-CG contexts showed enrichment in gene bodies and depletion in protein binding sites and enhancers. Non-CG methylation disappeared upon induced differentiation of the embryonic stem cells, and was restored in induced pluripotent stem cells. We identified hundreds of differentially methylated regions proximal to genes involved in pluripotency and differentiation, and widespread reduced methylation levels in fibroblasts associated with lower transcriptional activity. These reference epigenomes provide a foundation for future studies exploring this key epigenetic modification in human disease and development.

Thirty-four years have passed since it was proposed that cytosine DNA methylation in eukaryotes could act as a stably inherited modification affecting gene regulation and cellular differentiation[1,2]. Since then, intense research effort has expanded our understanding of diverse aspects of DNA methylation in higher eukaryotic organisms. These include elucidation of molecular pathways required for establishing and maintaining DNA methylation, cell-type-specific variation in methylation patterns, and the involvement of methylation in multifarious cellular processes such as gene regulation, DNA–protein interactions, cellular differentiation, suppression of transposable elements, embryogenesis, X-inactivation, genomic imprinting and tumorigenesis[3–9]. DNA methylation, together with covalent modification of histones, is thought to alter chromatin density and accessibility of the DNA to cellular machinery, thereby modulating the transcriptional potential of the underlying DNA sequence[10].

Genome-wide studies of mammalian DNA methylation have previously been conducted, however they have been limited by low resolution[11], sequence-specific bias, or complexity reduction approaches that analyse only a very small fraction of the genome[12–14]. To improve our understanding of the genome-wide patterns of DNA methylation we have generated single-base-resolution DNA methylation maps throughout the majority of the human genome in both embryonic stem cells and fibroblasts. Furthermore, we have profiled several important histone modifications, protein–DNA interaction sites of regulatory factors, and the mRNA and small RNA components of the transcriptome to better understand how changes in DNA methylation patterns and histone modifications may affect readout of the proximal genetic information.

Single-base-resolution maps of DNA methylation for two human cell lines

Single-base DNA methylomes of the flowering plant *Arabidopsis thaliana* were previously achieved using MethylC-Seq[15] or BS-Seq[16]. In this method, genomic DNA is treated with sodium bisulphite (BS) to convert cytosine, but not methylcytosine, to uracil, and subsequent high-throughput sequencing. We performed MethylC-Seq for two human cell lines, H1 human embryonic stem cells[17] and IMR90 fetal lung fibroblasts[18], generating 1.16 and 1.18 billion reads, respectively, that aligned uniquely to the human reference sequence (NCBI build 36/HG18). The total sequence yield was 87.5 and 91.0 gigabases (Gb), with an average read depth of 14.2× and 14.8× per strand for H1 and IMR90, respectively (Supplementary Fig. 1a). In each cell type, over 86% of both strands of the 3.08 Gb human reference sequence are covered by at least one sequence read (Supplementary Fig. 1b), accounting for 94% of the cytosines in the genome.

We detected approximately 62 million and 45 million methylcytosines in H1 and IMR90 cells, respectively (1% false discovery rate (FDR), see Supplementary Information and Fig. 1a), comprising

[1]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, California 92037, USA. [2]Ludwig Institute for Cancer Research, [3]Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, California 92093, USA. [4]ARC Centre of Excellence in Plant Energy Biology, The University of Western Australia, Crawley, Western Australia 6009, Australia. [5]Morgridge Institute for Research, Madison, Wisconsin 53707, USA. [6]Genome Center of Wisconsin, Madison, Wisconsin 53706, USA. [7]Wisconsin National Primate Research Center, University of Wisconsin-Madison, Madison, Wisconsin 53715, USA. [8]Department of Anatomy, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA.
*These authors contributed equally to this work.

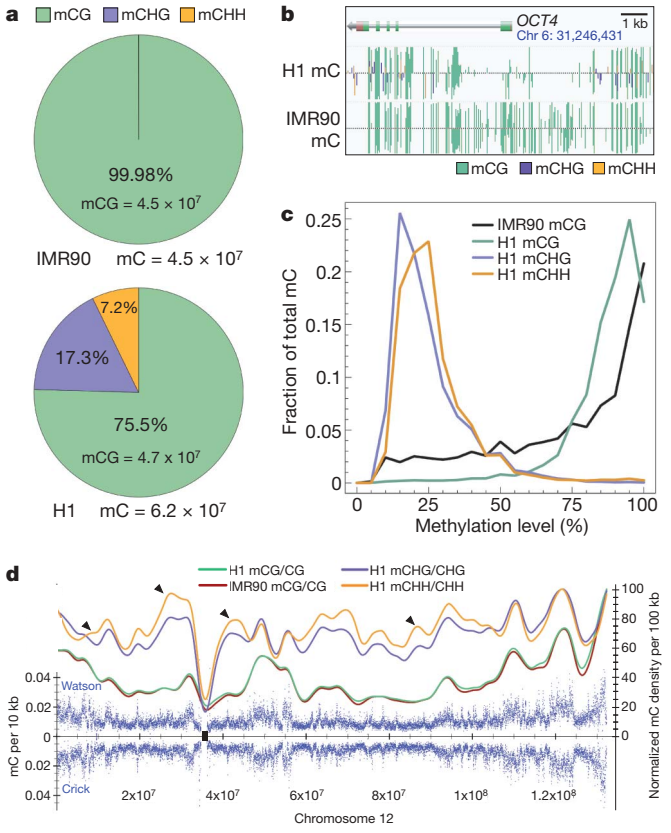


Figure 1 | Global trends of human DNA methylomes. **a**, The percentage of methylcytosines identified for H1 and IMR90 cells in each sequence context. **b**, AnnoJ browser representation of *OCT4*. **c**, Distribution of the methylation level in each sequence context. The *y* axis indicates the fraction of all methylcytosines that display each methylation level (*x* axis), where methylation level is the mC/C ratio at each reference cytosine (at least 10 reads required). **d**, Blue dots indicate methylcytosine density in H1 cells in 10-kb windows throughout chromosome 12 (black rectangle, centromere). Smoothed lines represent the methylcytosine density in each context in H1 and IMR90 cells. Black triangles indicate various regions of contrasting trends in CG and non-CG methylation. mC, methylcytosine.

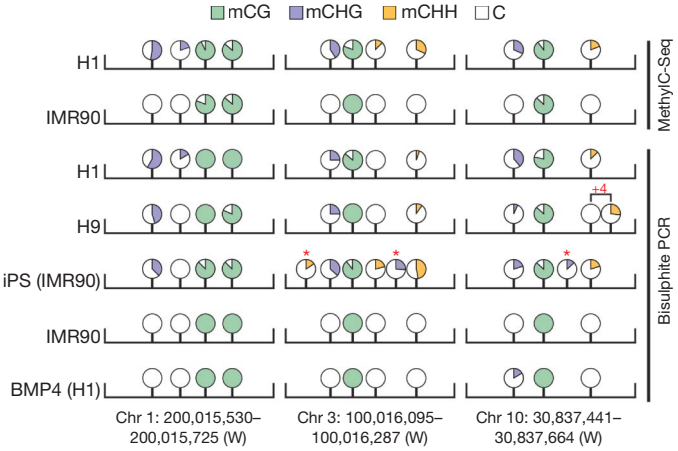


Figure 2 | Bisulphite-PCR validation of non-CG DNA methylation in differentiated and stem cells. DNA methylation sequence context is displayed according to the key and the percentage methylation at each position is represented by the fill of each circle (see Supplementary Table 2 for values). Non-CG methylated positions indicated by an asterisk are unique to that cell type and '+4' indicates a mCHH that is shifted 4 bases downstream in H9 cells. iPS, induced pluripotent stem cell.

5.83% and 4.25% of the cytosines with sequence coverage. Full browsing of the entire data set at single-base resolution can be performed at http://neomorph.salk.edu/human_methylome using the AnnoJ browser (http://www.annoj.org). Of the methylcytosines detected in the IMR90 genome, 99.98% were in the CG context, and the total number of mCG sites was very similar in both cell types. In the H1 stem cells we detected abundant DNA methylation in non-CG contexts (mCHG and mCHH, where H = A, C or T), comprising almost 25% of all cytosines at which DNA methylation is identified, and accounting for most of the difference in total methylcytosine number between the cell types (Fig. 1a). The prevailing assumption is that mammalian DNA methylation is located almost exclusively in the CG context; however, a handful of studies have previously detected non-CG methylation in human cells, and in particular in embryonic stem cells[19,20]. Bisulphite-PCR, cloning and sequencing of selected loci displaying H1 non-CG methylation in several human cell lines revealed that a second embryonic stem cell line, H9[17], displayed mCHG and mCHH at conserved positions, confirming that non-CG methylation is probably a general feature of human embryonic stem cells (Fig. 2, Supplementary Table 2). In addition, like IMR90 cells, BMP4-induced H1 cells lost non-CG methylation at several loci examined whereas methylation in the CG context was maintained, indicating that the pervasive non-CG methylation is lost upon differentiation. Furthermore, analysis of these loci in IMR90 induced pluripotent stem (iPS) cells revealed restored non-CG methylation (Fig. 2). Overall this demonstrates that the CHG and CHH methylation identified in H1 cells and absent in IMR90 cells is not simply due to genetic differences between the two cell types, but rather that the presence of non-CG methylation is characteristic of an embryonic stem-cell state. For each cell type, two biological replicates were performed with cells of different passage number (see Supplementary Information), and comparison of the methylcytosines identified independently in each replicate revealed a high concordance of cytosine methylation status between replicates (Supplementary Fig. 2). For each cell type, the final DNA methylation map presented in this study represents the composite of the two biological replicates. The *OCT4* gene (also called *POU5F1*) exemplifies both cell-specific differential methylation and the presence of non-CG methylation (Fig. 1b), and in addition displayed a ~50-fold reduction in *OCT4* transcript in IMR90 cells (data not shown). The absence of mCHG and mCHH methylation in IMR90 cells coincided with significantly lower transcript abundance of the *de novo* DNA methyltransferases (DNMTs) *DNMT3A* and *DNMT3B* and the associated *DNMT3L* in IMR90 cells (Supplementary Fig. 3), which is supported by a previous study of DNA methylation in embryonic stem cells and somatic cells[19] and by the determined target sequence specificity of these DNMTs[21,22].

Multiple reads covering each methylcytosine can be used as a readout of the fraction of the sequences within the sample that are methylated at that site[16], here referred to as the methylation level of a specific cytosine. Similar to the *Arabidopsis* genome[15], in the H1 genome we observed that 77% of mCG sites were 80–100% methylated, whereas 85% of mCHG and mCHH sites were 10–40% methylated (Fig. 1c), indicating that at sites of non-CG methylation only a fraction of the surveyed genomes in the sample was methylated. Notably, 56% of mCG sites in IMR90 cells were highly methylated (80–100%, Fig. 1c), indicating that although the total number of mCG sites in H1 and IMR90 cells is similar, in general the IMR90 mCG sites were typically less frequently methylated. In support of this, considering all CG site sequencing events, 82.7% and 67.7% were methylated in H1 and IMR90 cells, respectively. A global-scale view of DNA methylation levels revealed that the density of DNA methylation showed large variations throughout each chromosome (Fig. 1d). Sub-telomeric regions of the chromosomes frequently showed higher DNA methylation density (Fig. 1d and Supplementary Fig. 4), which was previously reported as being important for control of telomere length and recombination[23,24]. The smoothed profile of DNA methylation density in 100-kb windows indicated that on the chromosomal level the density profile of mCG in H1

and IMR90 cells was similar. The density profiles of mCHG and mCHH revealed that non-CG methylation was present throughout the entire chromosome. These two non-CG methylation marks showed a moderate correlation and did not always occur together (Pearson correlation 0.5 in 1-kb windows; Supplementary Fig. 2d). Notably, changes in density of the non-CG methylation were distinct from that of mCG in a number of regions.

Pervasive non-CG DNA methylation in embryonic stem cells

To characterize the abundant non-CG methylation in the H1 genome, we compared the average density of methylation relative to the underlying density of all potential sites of methylation in each context (henceforth referred to as the relative methylation density), throughout various genomic features (Fig. 3a and Supplementary Fig. 5). We observed a correlation in the density of mCG and the distance from the transcriptional start site (TSS), with mCG density increasing in the 5′ untranslated region (UTR) to a similar level in exons, introns and the 3′ UTR as to 2 kb upstream of the TSS (Fig. 3a). We generally observed lower relative densities of methylation at CG islands and TSS; however, a subset of these regions did not display this depletion (Supplementary Fig. 6)[13,14,25]. mCHG and mCHH methylation densities also decreased significantly towards the TSS and returned to the same level as 2 kb upstream at the end of the 5′ UTR; however, within exons, introns and 3′ UTRs the non-CG methylation densities were twice as high. Intriguingly, the mCHH density was approximately 15–20% higher in exons than within introns and the 3′ UTR. To identify links between gene activity and non-CG methylation level within the gene body we performed strand-specific RNA-Seq[15] and observed a positive correlation between gene expression and mCHG ($r = 0.60$) or mCHH ($r = 0.58$) density (Fig. 3b), with highly expressed genes containing threefold higher non-CG methylation density than non-expressed genes (Supplementary Fig. 7a). However, no correlation was observed

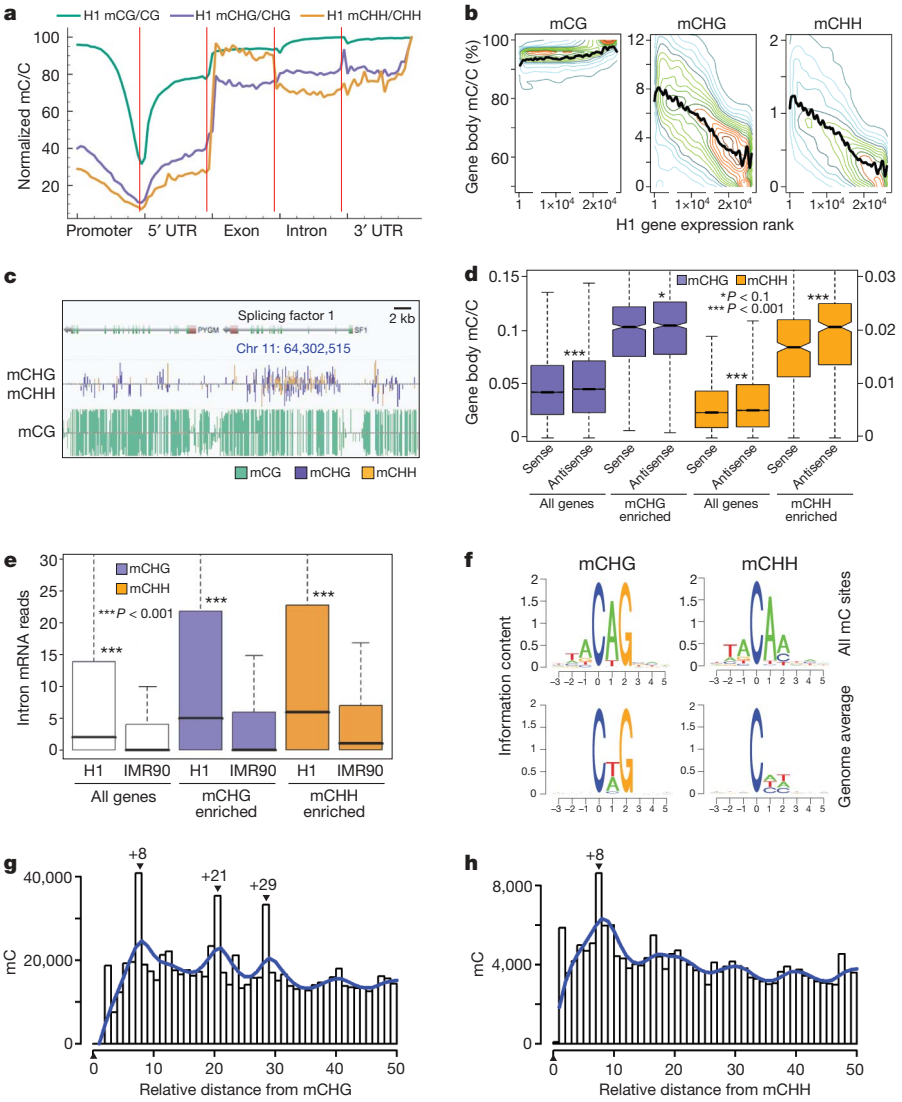


Figure 3 | Non-CG DNA methylation in H1 embryonic stem cells. a, Relative methylation density (the ratio of methylcytosines to reference cytosines) in H1 throughout different gene-associated regions (promoters encompass 2 kb upstream of the transcriptional start site).The mean mC/C profile was normalized to the maximum value. **b**, Relative methylation density within gene bodies (*y* axis) as a function of gene expression (*x* axis), with transcript abundance increasing from right to left. Coloured lines represent data point density and smoothing with cubic splines is displayed in black. **c**, Graphical representation of methylation at a non-CG methylation enriched gene, splicing factor 1. **d**, Average relative methylation densities in each sequence context within gene bodies on the sense or antisense strand relative to gene directionality. *P*-values for differences between sense and antisense densities are indicated. Boxes in **d** and **e** represent the quartiles and whiskers mark the minimum and maximum values. **e**, Number of mRNA intronic reads in all genes or genes associated with non-CG enriched regions, in H1 and IMR90 cells. *P*-values for differences between H1 and IMR90 reads are indicated. **f**, Logo plots of the sequences proximal to sites of non-CG DNA methylation in each sequence context in H1 cells. **g**, **h**, Prevalence of mCHG/mCHH sites (*y* axis) as a function of the number of bases between adjacent mCHG/mCHH sites (*x* axis) based on all non-redundant pairwise distances up to 50 nucleotides in all introns. Blue line represents smoothing with cubic splines.

between CG methylation density and gene expression in the H1 cells (Fig. 3b).

We identified 447 and 226 genes that were proximal to genomic regions highly enriched for mCHG and mCHH, respectively, with 180 genes in common. An example of non-CG methylation enrichment in such a gene, splicing factor 1 (*SF1*), is shown in Fig. 3c. Analysis of gene ontology terms for each set revealed significant enrichment for genes involved in RNA processing, RNA splicing and RNA metabolic processes ($P = 2 \times 10^{-11}$, Supplementary Fig. 7b). Unexpectedly, we found a significant enrichment of non-CG methylation on the antisense strand of gene bodies, for both mCHG and mCHH enriched sets of genes ($P < 0.1$ and $P < 0.001$, respectively, Fig. 3d). The antisense strand serves as the template for RNA polymerization, and further investigation will be required to determine if there are functional repercussions of this non-CG methylation strand bias. We also observed that genes in H1 had significantly more RNA originating from introns than in IMR90, relative to the total number of sequenced reads in each sample, and this discrepancy in intronic read abundance was significantly enhanced in the mCHG and mCHH enriched genes ($P < 0.001$, Fig. 3e). The higher abundance of intronic reads was associated with higher non-CG methylation within gene bodies, rather than differential non-CG methylation of exons versus introns.

In the *Arabidopsis* genome, the methylation state of a cytosine in the CG and CHG contexts is highly correlated with the methylation of the cytosine on the opposite strand within the symmetrical site[15,16]. Whereas we observed that 99% of mCG sites from the human cell lines were methylated on both strands, surprisingly mCHG was highly asymmetrical, with 98% of mCHG sites being methylated on only one strand. This raises an interesting question as to how these sites of DNA methylation are consistently methylated in a considerable fraction of the genomes without two hemi-methylated CHG sites as templates for faithful propagation of the methylation state (Fig. 1c). It is not yet known whether continual, but indiscriminate, *de novo* methyltransferase activity preferentially methylates particular CHG sites after replication, or if a persistent targeting signal is present that drives CHG methylation.

We analysed the genome sequence proximal to sites of non-CG methylation to determine whether enrichment of particular local sequences were evident, as previously reported in the *Arabidopsis* DNA methylomes[15,16]. Whereas no local sequence enrichment was observed for mCG sites, a preference for the TA dinucleotide upstream of non-CG methylation was observed (Fig. 3f and Supplementary Fig. 8). Furthermore, the base following a non-CG methylcytosine was most commonly an A, with a T also observed relatively frequently, a sequence preference observed in previous *in vitro* studies of the mammalian DNMT3 methyltransferases[21,22].

To determine whether there was any preference for the distance between adjacent sites of DNA methylation in the human genome, we analysed the relative distance between methylcytosines in each context within 50 nucleotides in introns. We focused on introns because these are genomic regions enriched in non-CG methylation, but unlike exons, are not constrained by protein coding selective pressures (Fig. 3g, h). Analyses for random genomic sequences and exons are presented in Supplementary Fig. 9, together with mCG spacing patterns. For methylcytosines in all contexts, a periodicity of 8–10 bases was evident (Fig. 3g, h and Supplementary Fig. 9), but interestingly a strong tendency was observed for two pairs of 8-base separated mCHG sites spaced with 13 bases between them. An 8–10 base periodicity was also evident for mCHH sites, corresponding to a single turn of the DNA helix, as previously observed in the *Arabidopsis* genome[16], indicating that the molecular mechanisms governing *de novo* methylation at CHH sites may be common between the plant and animal kingdoms. A structural study of the mammalian *de novo* methyltransferase DNMT3A and its partner protein DNMT3L found that two copies of each form a heterotetramer that contains two active sites separated by the length of 8–10 nucleotides in a DNA helix[26,27]. The consistent 8–10 nucleotide spacing we observed in the human genome suggests that DNMT3A may be responsible for catalysing the methylation at non-CG sites. Notably, the mCHG and mCHH relative spacing patterns were distinct, suggesting that this sub-categorization of the non-CG methylation is appropriate, and that distinct pathways may be responsible for the deposition of mCHG and mCHH in the human genome.

Depleted DNA methylation at DNA–protein interaction sites

Numerous past studies have documented that DNA methylation can alter the ability of some DNA binding proteins to interact with their target sequences[28–32]. To investigate this relationship further we used ChIP-Seq[33] to identify sites of protein–DNA interaction in H1 cells for a set of proteins important for gene expression in the pluripotent state, namely NANOG, SOX2, KLF4 and OCT4, as well as proteins involved in the transcription initiation complex and in enhancers (TAF1 and p300, respectively) (Supplementary Tables 3–8). In general we observed a decrease in the profile of relative methylation density towards the site of interaction, particularly in the non-CG context, independently from proximity to the TSS (Fig. 4a and Supplementary Fig. 10). The IMR90 genome showed higher average density of methylation at H1 SOX2 and p300 interaction sites, but had similar CG methylation densities for the H1 NANOG and OCT4 interaction sites,

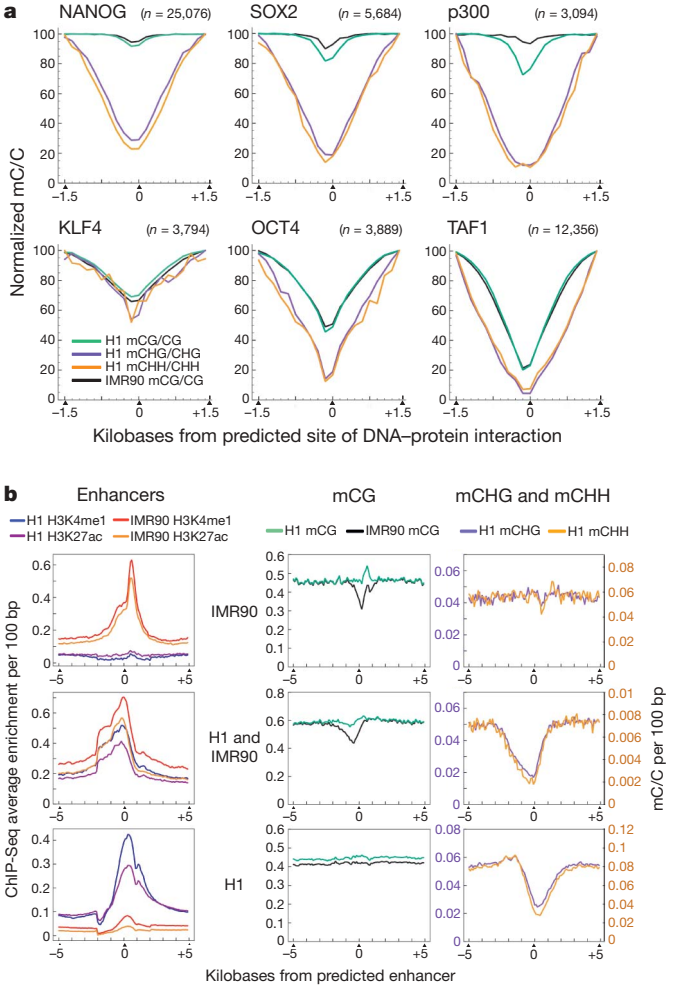


Figure 4 | Density of DNA methylation at sites of DNA–protein interaction. **a**, Average relative DNA methylation densities 1.5 kb upstream and downstream of predicted sites of DNA–protein interaction. **b**, Co-localization of H3K4me1 and H3K27ac ChIP-Seq tag enrichment indicative of enhancer sites that have been grouped into three sets: specific to IMR90 cells (top), H1 cells (bottom), or common to both H1 and IMR90 cells (middle). Average relative DNA methylation densities in each sequence context in 100-bp windows are displayed throughout 5 kb upstream and downstream of the enhancers in each of the sets.

even though the genes encoding these proteins are transcribed at a very low level in IMR90 relative to H1 cells (47–50-fold less mRNA), and are not considered to be functional in fibroblasts. This suggests that these genomic regions are generally maintained in a less methylated state in multiple cell types regardless of the occupancy of these specific DNA binding proteins.

We next analysed the patterns of DNA methylation in sets of enhancers either unique to each cell type or shared. ChIP-Seq was used to detect the location of enhancers throughout the H1 and IMR90 genomes, defined as regions of simultaneous enrichment of the histone modifications H3K4me1 and H3K27ac[34] (Fig. 4b). We examined the average relative DNA methylation density at enhancer sites, as well as the flanking genomic regions, and found a depletion of CG methylation at IMR90-specific enhancers, yet enrichment in mCG density in H1 at the same genomic locations (Fig. 4b). In contrast, at H1-specific enhancers there was no change in mCG density in either the H1 or IMR90 genome, but non-CG methylation density decreased approximately threefold at the enhancer sites, relative to the density 5 kb upstream and downstream. This is in agreement with the depletion of non-CG methylation in the H1 genome at predicted sites of p300 interaction (Fig. 4a), a strong indicator of enhancer activity[34]. The set of enhancer sites present in both H1 and IMR90 cells showed both of these cell-specific patterns: lower mCG density in IMR90 and lower non-CG methylation density in H1. The specific depletion of DNA methylation at active enhancers in each cell type (also recently reported on a limited basis[35]) indicates maintenance of these elements in an unmethylated state, potentially preventing interference in the process of protein–DNA interaction at these sites. Notably, H1 cells had depleted non-CG methylation but not mCG, in contrast to the mCG depletion at IMR90 enhancers. These data might indicate cell-type-specific utilization of different categories of DNA methylation, possibly coupled with novel stem-cell-specific factors that are able to recognize non-CG methylation, akin to the specific binding of the H3K9 histone methyltransferase KRYPTONITE to non-CG methylation sites in *Arabidopsis*[36].

Widespread cell-specific patterns of DNA methylation

The paradigm of DNA methylation controlling aspects of cellular differentiation necessitates that patterns of methylation vary in different cell types. Numerous studies have previously documented differences in DNA methylation between cell types and disease states[7,8,10,37]. With comprehensive maps of DNA methylation throughout the genomes of the two distinct cell types, we next characterized changes in DNA methylation evident between the H1 and

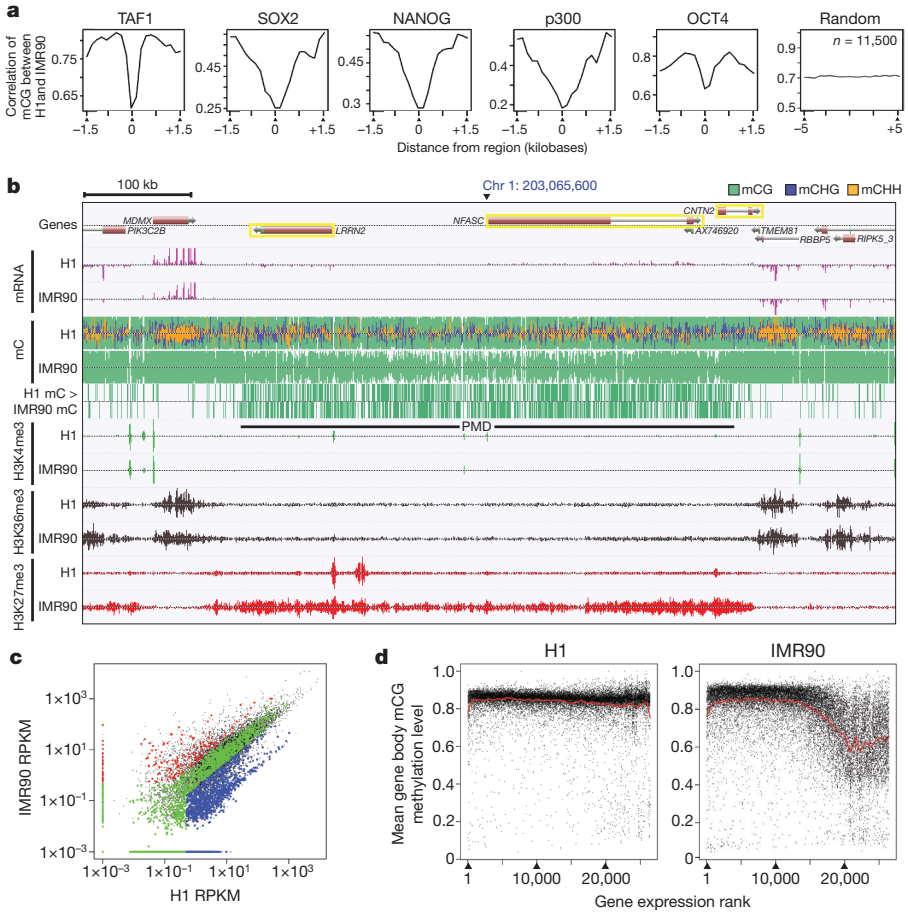


Figure 5 | Cell-type variation in DNA methylation. **a**, Pearson correlation coefficient of mCG methylation density (*y* axis) between H1 and IMR90 at various genomic features. Regions were divided into 20 equally sized bins from 5′ to 3′. Pearson correlation was determined in each bin considering all the H1 and IMR90 occurrences of the given genomic region. **b**, DNA methylation, mRNA and histone modifications in H1 and IMR90 cells associated with a PMD. Vertical lines above and below the dotted central line in DNA methylation tracks indicate methylcytosines on the Watson and Crick strands, respectively. Line vertical height indicates the methylation level. The H1 mC > IMR90 mC track indicates methylcytosines significantly more methylated in H1 than IMR90 at a 5% FDR (Fisher's exact test). Vertical bars in the mRNA and histone modification tracks represent sequence tag enrichment. A yellow box indicates any gene with ≥30-fold higher mRNA abundance in H1 than IMR90. **c**, Comparison of transcript abundance between H1 and IMR90 cells of genes with a transcriptional start site located in or within 10 kb of a PMD. Black dots indicate all genes in the genome; blue, red and green indicate PMD genes expressed ≥3-fold higher in H1, IMR90 or not differentially expressed, respectively. **d**, Mean gene body mCG methylation (at least 10 reads required) as a function of the gene expression rank, 1 being the most expressed. mC, methylcytosine; PMD, partially methylated domain; RPKM, reads per kilobase of transcript per million reads.

IMR90 DNA methylomes, and explored how these changes may relate to the distinctiveness of these cells.

Pairwise comparison of methylation at the same genomic coordinates between H1 and IMR90 is required to reveal cell-specific methylation patterns potentially masked by average profiles. The Pearson correlation coefficient of the mCG methylation state between H1 and IMR90 was calculated for 20 equally sized windows flanking or within various genomic features (Fig. 5a and Supplementary Fig. 11), providing a measure of methylation state conservation at these genomic features between the two cell types, and distinct from the average methylation density profiles presented above (Fig. 4). At the sites of protein–DNA interaction surveyed in Fig. 4a, we observed a decrease in the correlation of methylation compared to the flanking 1.5 kb of the genome (Fig. 5a), except for KLF4 (data not shown). This decrease was most pronounced at the predicted site of protein–DNA interaction, indicating that even though the mCG depletion was a general feature of the surveyed protein binding sites (Fig. 4a), when a pairwise comparison of the methylation status at each cytosine associated with the protein binding site between H1 and IMR90 was performed a significant decrease in the conservation of methylation was observed (Fig. 5a).

Surprisingly, we found that a large proportion of the IMR90 genome displayed lower levels of CG methylation than H1 (Fig. 1c). Contiguous regions with an average methylation level less than 70% were identified (mean length = 153 kb), which we termed partially methylated domains (PMDs) (Fig. 5b, Supplementary Fig. 12 and Supplementary Table 9). The PMDs comprised a large proportion of every autosome (average = 38.4%), and 80% of the IMR90 X chromosome (Supplementary Fig. 12), consistent with the lower levels of DNA methylation reported in the inactive X chromosome[38]. As IMR90 cells are derived from a female (XX), it is anticipated that simultaneous sequencing of BS-converted genomic DNA from both the inactive and the active X chromosomes will manifest as PMDs throughout the majority of the X chromosome. However, the widespread prevalence of PMDs on the autosomes was unexpected. We analysed the ratio of methylated to unmethylated CG sites within individual MethylC-Seq reads. The IMR90 reads located within PMDs were more frequently partially methylated or unmethylated compared to all IMR90 reads aligned to the autosomes (Supplementary Fig. 12b). The decrease in PMD methylation manifested similarly in IMR90 autosomes and chromosome X; however, currently we cannot determine whether common pathways are responsible for altering methylation patterns in all chromosomes.

Upon inspection of 5,644 genes with a TSS located in or within 10 kb of a PMD, we found a strong enrichment for these genes to be less expressed in IMR90 ($P = 2 \times 10^{-47}$, Fisher's exact test). Specifically, of all of the genes that were more highly expressed in H1 (H1 $\geq 3\times$ IMR90 transcript abundance), 42% were located within PMDs, compared to only 13% of all more highly expressed genes in IMR90 cells being located in PMDs (Fig. 5b, c and Supplementary Tables 10 and 11). Many of the partially methylated and downregulated genes in IMR90 displayed lower proximal H3K4me3 and H3K36me3 modifications, and higher proximal H3K27me3 levels (Fig. 5b, Supplementary Fig. 13 and R.D.H. *et al.*, submitted). Whereas in IMR90 cells we

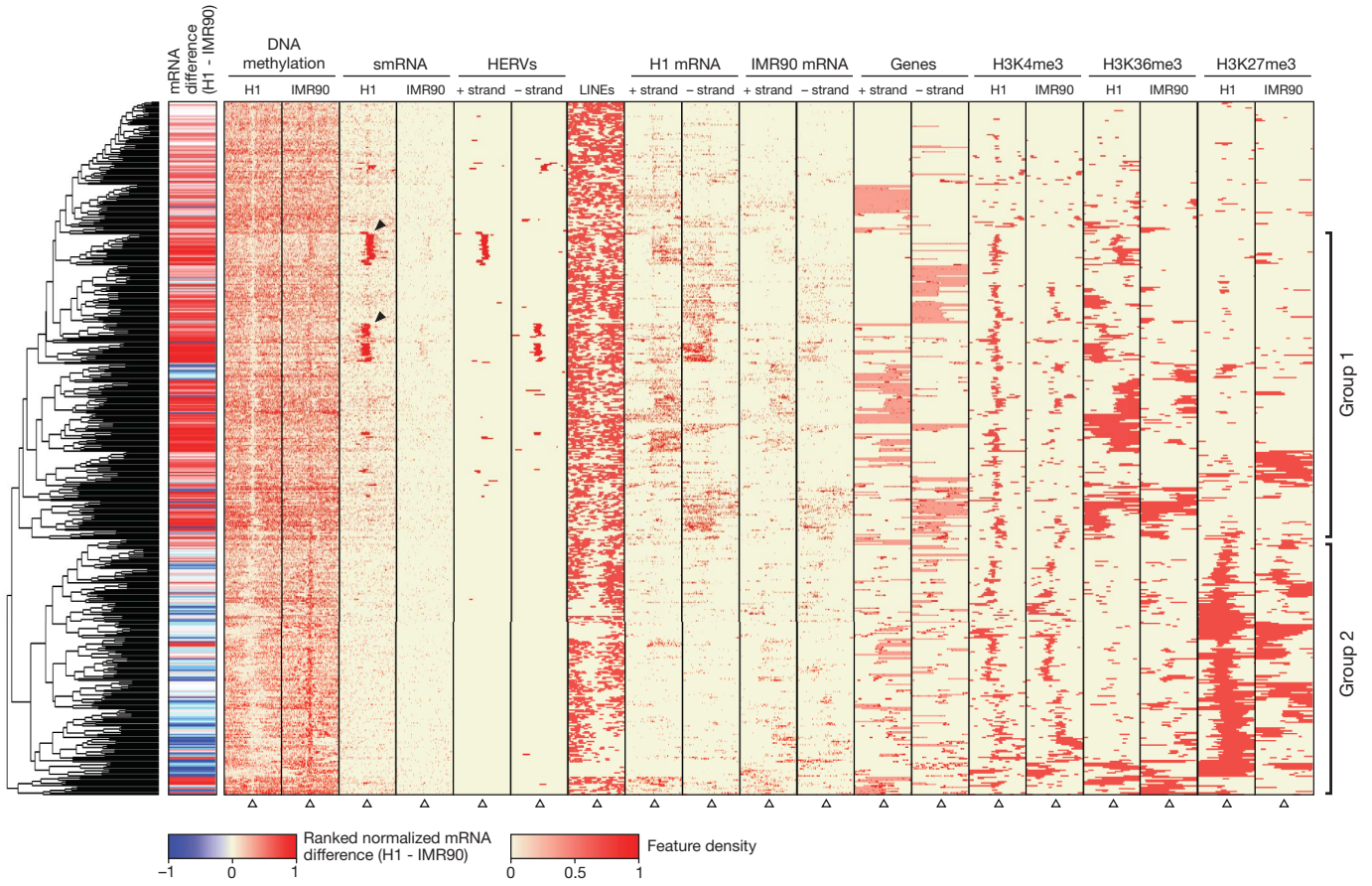


Figure 6 | Clustering of genomic, epigenetic and transcriptional features at differentially methylated regions. The density of DNA methylation, smRNA reads, strand-specific mRNA reads and the presence of domains of H3K4me3, H3k36me3 and H3K27me3 in H1 and IMR90 was profiled through 20 kb upstream and downstream of each of the 491 DMRs where DNA methylation was more prevalent in IMR90 than H1. Open triangles indicate the central point in each window. The side colour bar indicates the difference between H1 and IMR90 mRNA levels. The location of HERVs, LINEs and genes is displayed on each strand, where pink colouring indicates the gene body and dark red boxes represent exons. Black triangles indicate regions enriched for smRNAs that are coincident with HERVs. Group 1 and 2 are discussed in the text. DMRs, differentially methylated regions; HERVs, human endogenous retroviruses.

observed a positive correlation between the mean gene body mCG methylation level and gene expression, no such relationship was discernible in H1 cells (Fig. 5d). Consequently, the positive correlation between gene expression and gene body methylation recently reported[12] could be re-interpreted as a depletion of methylation in repressed genes in differentiated cells.

Stem cell hypomethylated regions

A sliding window approach was used to identify differentially methylated regions (DMRs) enriched for cytosines where IMR90 was more highly methylated than H1 (5% FDR, Fisher's exact test, Supplementary Fig. 14). We identified 491 DMRs, and in a window spanning 20 kb up- and downstream of each DMR we surveyed mCG density, mRNAs, small RNAs (smRNAs), H3K4me3, H3K36me3, H3K27me3, genes and repetitive elements (Fig. 6, Supplementary Table 12 and R.D.H. *et al.*, submitted). The DMRs were associated with 139 and 113 genes more highly expressed in H1 and IMR90, respectively. More than half of these genes were associated with DMRs located within 2 kb upstream of the TSS or the 5′ UTR, which include factors previously defined as having a role in embryonic stem-cell function[39] (Supplementary Fig. 15 and Supplementary Tables 13 and 14).

Complete linkage hierarchical clustering of these data revealed two broad categories of transcriptional activity, histone modifications and DNA methylation proximal to the DMRs (Fig. 6). Group 1 DMRs are associated with high proximal H3K4me3, H3K36me3 and transcriptional activity relative to IMR90, and are unmarked by H3K27me3 in both cell types. Although we did not observe widespread association of small RNA molecules with enrichment of DNA methylation, we found that a subset of group 1 DMRs co-localizes with dense clusters of small RNAs that map to annotated human endogenous retroviruses (HERVs)[40]. Notably, the HERVs were less densely methylated in H1 and frequently associated with high downstream transcriptional activity, in contrast to the more methylated state in IMR90 that was not associated with abundant small RNAs and showed little proximal transcription (Fig. 6 and Supplementary Fig. 16). Accurate targeting of DNA methylation by small RNAs is a well-established process in plants[41]. Although our data did not provide evidence for the existence of an analogous process in the human cells, further experiments may be required to investigate this relationship in greater detail, such as DNA methylation profiling following silencing of components of the RNA interference machinery.

Group 2 DMRs were associated with gene-rich sequences that were more highly expressed in IMR90 cells and generally exhibited a depletion of long interspersed nuclear elements (LINEs) in the flanking sequence, with concomitant H3K27me3 modification and less DNA methylation, as observed in many IMR90 PMDs. Furthermore, group 2 regions in H1 frequently displayed both H3K4me3 and H3K27me3 modifications, characteristic of the bivalent state that is thought to instil a suppressed but poised transcriptional status[42,43]. Many of these regions showed markedly less H3K27me3 in IMR90 cells in addition to more DNA methylation, suggesting that prior repression may have been relieved, and defining a set of genes potentially regulated by DNA methylation and involved in the developmental transition from a pluripotent to differentiated state.

Concluding remarks

We found extensive differences between the DNA methylomes of two human cell types, revealing the highly dynamic nature of this epigenetic modification. The genomic context of the DNA methylation is resolved, here revealing abundant methylation in the non-CG context, which is typically overlooked in alternative methodologies. Profiling of enhancers and different patterning of CG and non-CG methylation in gene bodies and their different correlation with gene expression suggest possible alternative roles for DNA methylation in these two contexts. The exclusivity of non-CG methylation in stem cells, probably maintained by continual *de novo* methyltransferase activity and not observed in differentiated cells, suggests that it may have a key role in the origin and maintenance of this pluripotent state. Essential future studies will need to explore the prevalence of non-CG methylation in diverse cell types, including variation throughout differentiation and its potential re-establishment in induced pluripotent states.

METHODS SUMMARY

Biological materials and sequencing libraries. Human H1, H9, BMP4-induced H1 and IMR90 cells were cultured as described previously[34,44,45]. smRNA-Seq libraries were generated from 10–50-nt small RNAs using the Small RNA Sample Prep v1.5 kit (Illumina), as per the manufacturer's instructions. Strand-specific mRNA-Seq libraries were produced using a modification of a protocol described previously[15]. Unique 5′ and 3′ RNA oligonucleotides were sequentially ligated to the ends of fragments of RNA isolated by depletion of rRNA from total RNA samples. MethylC-Seq libraries were generated by ligation of methylated sequencing adapters to fragmented genomic DNA followed by gel purification, sodium bisulphite conversion and four cycles of PCR amplification. ChIP-Seq libraries were prepared following Illumina protocols with minor modifications (See Supplementary Information). Sequencing was performed using the Illumina Genome Analyser II as per the manufacturer's instructions.

Read processing and alignment. MethylC-Seq sequencing data was processed using the Illumina analysis pipeline and FastQ format reads were aligned to the human reference genome (hg18) using the Bowtie alignment algorithm[46]. The base calls per reference position on each strand were used to identify methylated cytosines at 1% FDR. mRNA-Seq reads were aligned to the human reference and splice junctions of UCSC known genes using the ELAND algorithm (Illumina). smRNA-Seq reads that contained a subset of the 3′ adaptor sequence were selected and this adaptor sequence removed, retaining trimmed reads that were from 16 to 37 nucleotides in length. These processed reads were aligned to the human reference genome (NCBI build 36/HG18) using the Bowtie alignment algorithm, and any read that aligned with no mismatches and to no more than 1,000 locations in the reference genome was retained. Base calling and mapping of Chip-Seq reads were performed using the Illumina pipeline.

Received 19 June; accepted 21 September 2009.
Published online 14 October 2009.

1. Holliday, R. & Pugh, J. E. DNA modification mechanisms and gene activity during development. *Science* **187,** 226–232 (1975).
2. Riggs, A. D. X inactivation, differentiation, and DNA methylation. *Cytogenet. Cell Genet.* **14,** 9–25 (1975).
3. Bestor, T. H. The DNA methyltransferases of mammals. *Hum. Mol. Genet.* **9,** 2395–2402 (2000).
4. Li, E., Bestor, T. H. & Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69,** 915–926 (1992).
5. Lippman, Z. *et al.* Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430,** 471–476 (2004).
6. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for *de novo* methylation and mammalian development. *Cell* **99,** 247–257 (1999).
7. Reik, W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* **447,** 425–432 (2007).
8. Straussman, R. *et al.* Developmental programming of CpG island methylation profiles in the human genome. *Nature Struct. Mol. Biol.* **16,** 564–571 (2009).
9. Weber, M. & Schübeler, D. Genomic patterns of DNA methylation: targets and function of an epigenetic mark. *Curr. Opin. Cell Biol.* **19,** 273–280 (2007).
10. Cedar, H. & Bergman, Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nature Rev. Genet.* **10,** 295–304 (2009).
11. Rauch, T. A., Wu, X., Zhong, X., Riggs, A. D. & Pfeifer, G. P. A human B cell methylome at 100-base pair resolution. *Proc. Natl Acad. Sci. USA* **106,** 671–678 (2009).
12. Ball, M. *et al.* Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature Biotechnol.* **27,** 361–368 (2009).
13. Deng, J. *et al.* Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nature Biotechnol.* **27,** 353–360 (2009).
14. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454,** 766–770 (2008).
15. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in *Arabidopsis. Cell* **133,** 523–536 (2008).
16. Cokus, S. J. *et al.* Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452,** 215–219 (2008).
17. Thomson, J. A. *et al.* Embryonic stem cell lines derived from human blastocysts. *Science* **282,** 1145–1147 (1998).
18. Nichols, W. W. *et al.* Characterization of a new human diploid cell strain, IMR-90. *Science* **196,** 60–63 (1977).
19. Ramsahoye, B. H. *et al.* Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc. Natl Acad. Sci. USA* **97,** 5237–5242 (2000).

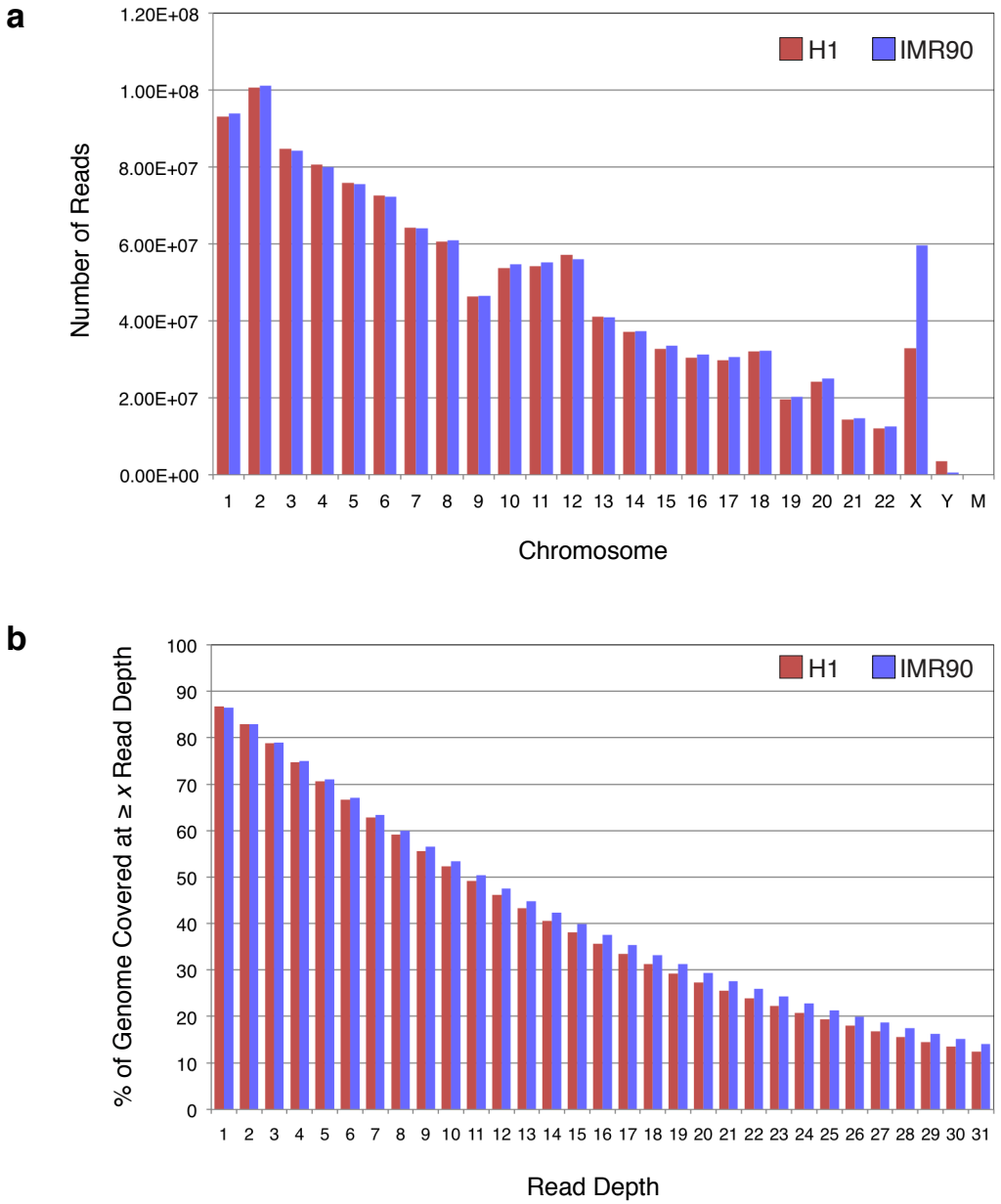
20. Woodcock, D. M., Crowther, P. J. & Diver, W. P. The majority of methylated deoxycytidines in human DNA are not in the CpG dinucleotide. *Biochem. Biophys. Res. Commun.* **145,** 888–894 (1987).
21. Aoki, A. *et al.* Enzymatic properties of de novo-type mouse DNA (cytosine-5) methyltransferases. *Nucleic Acids Res.* **29,** 3506–3512 (2001).
22. Gowher, H. & Jeltsch, A. Enzymatic properties of recombinant Dnmt3a DNA methyltransferase from mouse: the enzyme modifies DNA in a non-processive manner and also methylates non-CpA sites. *J. Mol. Biol.* **309,** 1201–1208 (2001).
23. Gonzalo, S. *et al.* DNA methyltransferases control telomere length and telomere recombination in mammalian cells. *Nature Cell Biol.* **8,** 416–424 (2006).
24. Steinert, S., Shay, J. W. & Wright, W. E. Modification of subtelomeric DNA. *Mol. Cell. Biol.* **24,** 4571–4580 (2004).
25. Brunner, A. L. *et al.* Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. *Genome Res.* **19,** 1044–1056 (2009).
26. Ferguson-Smith, A. C. & Greally, J. Epigenetics: perceptive enzymes. *Nature* **449,** 148–149 (2007).
27. Jia, D., Jurkowska, R. Z., Zhang, X., Jeltsch, A. & Cheng, X. Structure of Dnmt3a bound to Dnmt3L suggests a model for *de novo* DNA methylation. *Nature* **449,** 248–251 (2007).
28. Bell, A. C. & Felsenfeld, G. Methylation of a CTCF-dependent boundary controls imprinted expression of the *Igf2* gene. *Nature* **405,** 482–485 (2000).
29. Clark, S. J., Harrison, J. & Molloy, P. L. Sp1 binding is inhibited by (m)Cp(m)CpG methylation. *Gene* **195,** 67–71 (1997).
30. Hark, A. T. *et al.* CTCF mediates methylation-sensitive enhancer-blocking activity at the *H19/Igf2* locus. *Nature* **405,** 486–489 (2000).
31. Kitazawa, S., Kitazawa, R. & Maeda, S. Transcriptional regulation of rat cyclin D1 gene by CpG methylation status in promoter region. *J. Biol. Chem.* **274,** 28787–28793 (1999).
32. Mancini, D. N., Singh, S. M., Archer, T. K. & Rodenhiser, D. I. Site-specific DNA methylation in the neurofibromatosis (NF1) promoter interferes with binding of CREB and SP1 transcription factors. *Oncogene* **18,** 4108–4119 (1999).
33. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* **316,** 1497–1502 (2007).
34. Heintzman, N. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459,** 108–112 (2009).
35. Schmidl, C. *et al.* Lineage-specific DNA methylation in T cells correlates with histone methylation and enhancer activity. *Genome Res.* **19,** 1165–1174 (2009).
36. Johnson, L. M. *et al.* The SRA methyl-cytosine-binding domain links DNA and histone methylation. *Curr. Biol.* **17,** 379–384 (2007).
37. Jones, P. A. & Baylin, S. B. The epigenomics of cancer. *Cell* **128,** 683–692 (2007).
38. Hellman, A. & Chess, A. Gene body-specific methylation on the active X chromosome. *Science* **315,** 1141–1143 (2007).
39. The International Stem Cell Initiative. Characterization of human embryonic stem cell lines by the International Stem Cell Initiative. *Nature Biotechnol.* **25,** 803–816 (2007).
40. Villesen, P., Aagaard, L., Wiuf, C. & Pedersen, F. S. Identification of endogenous retroviral reading frames in the human genome. *Retrovirology* **1,** 32 (2004).
41. Chan, S. W. Inputs and outputs for chromatin-targeted RNAi. *Trends Plant Sci.* **13,** 383–389 (2008).
42. Azuara, V. *et al.* Chromatin signatures of pluripotent cell lines. *Nature Cell Biol.* **8,** 532–538 (2006).
43. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125,** 315–326 (2006).
44. Ludwig, T. *et al.* Feeder-independent culture of human embryonic stem cells. *Nature Methods* **3,** 637–646 (2006).
45. Ludwig, T. *et al.* Derivation of human embryonic stem cells in defined conditions. *Nature Biotechnol.* **24,** 185–187 (2006).
46. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

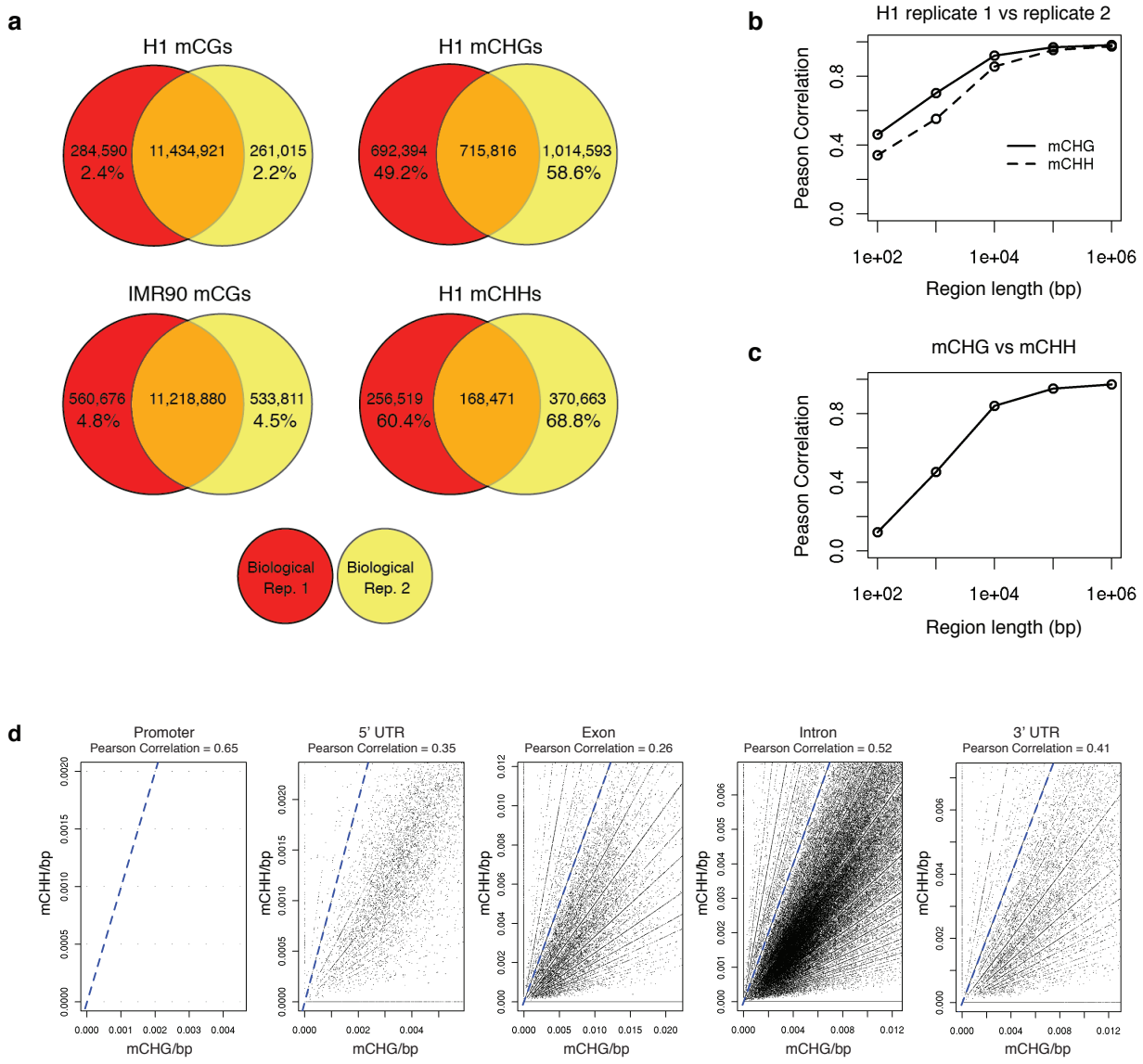
Acknowledgements We thank A. Elwell and A. Hernandez for assistance with sequence library preparation and Illumina sequencing. R.L. is supported by a Human Frontier Science Program Long-term Fellowship. R.D.H. is supported by an American Cancer Society Postdoctoral Fellowship. This work was supported by grants from the following: Mary K. Chapman Foundation, The National Institutes of Health (U01 ES017166 and U01 1U01ES017166-01), the California Institute for Regenerative Medicine (RS1-00292-1), the Australian Research Council Centre of Excellence Program (CE0561495, DP0771156) and Morgridge Institute for Research, Madison, Wisconsin. We thank the NIH Roadmap Reference Epigenome Consortium (http://nihroadmap.nih.gov/epigenomics/referenceepigenomeconsortium.asp) and C. Gunter (Hudson-Alpha Institute) for assistance. This study was carried out as part of the NIH Roadmap Epigenomics Program.

Author Contributions Experiments were designed by J.R.E., B.R., R.L., J.A.T. and R.D.H. Cells were grown by J.A.-B. and Q.-M.N. MethylC-Seq, RNA-Seq and smRNA-Seq experiments were conducted by R.L. and J.R.N. ChIP-Seq experiments were conducted by R.D.H., L.L. and Z.Y. ChIP-Seq data analysis was performed by G.H., R.D.H. and L.E. BS-PCR validation was performed by R.H.D. Sequencing data processing was performed by R.L., J.T.-F., L.E., V.R. and G.H. Bioinformatic and statistical analyses were conducted by M.P., R.L., G.H., J.T.-F., R.H.D., R.S. and A.H.M. AnnoJ development was performed by J.T.F and A.H.M. The manuscript was prepared by R.L., M.P., R.H.D., A.H.M. and J.R.E.

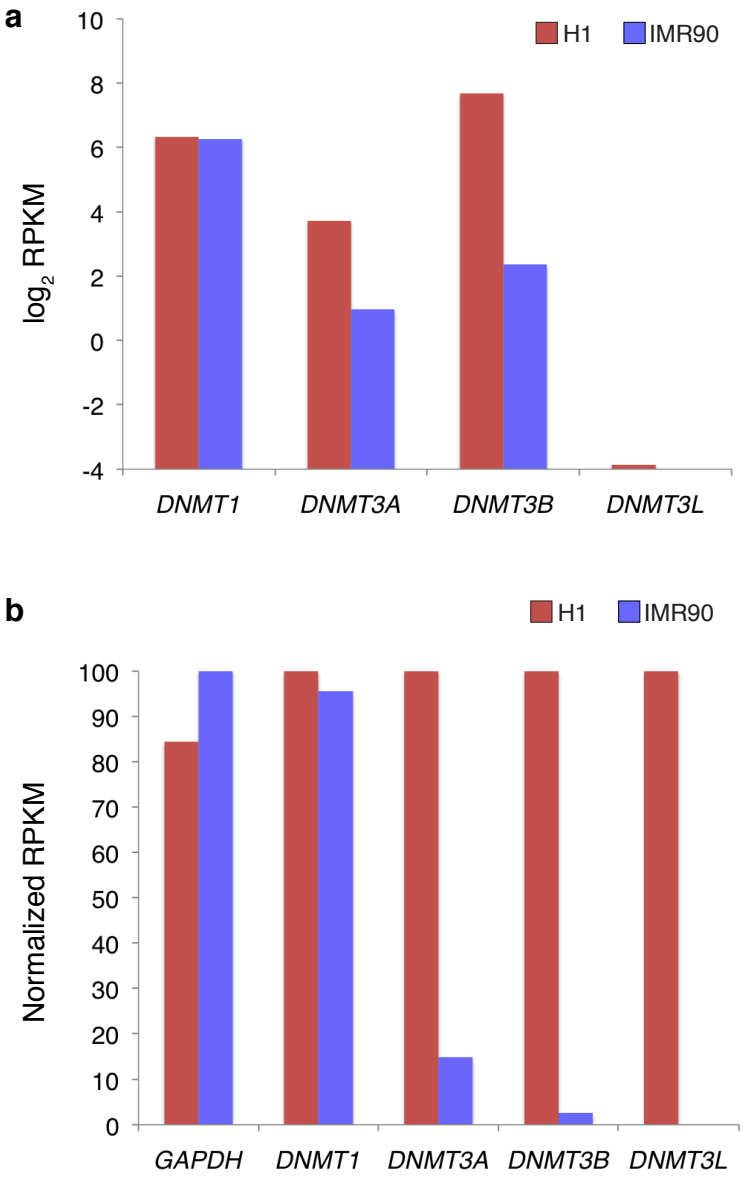
Author Information Sequence data is available under the GEO accessions GSM429321-23, GSM432685-92, GSM438361-64, GSE17917, GSE18292 and GSE16256, and the SRA accessions SRX006782-89, SRX006239-41, SRX007165.1-68.1 and SRP000941. Analysed data sets can be obtained from http://neomorph.salk.edu/human_methylome. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.R.E. (ecker@salk.edu).



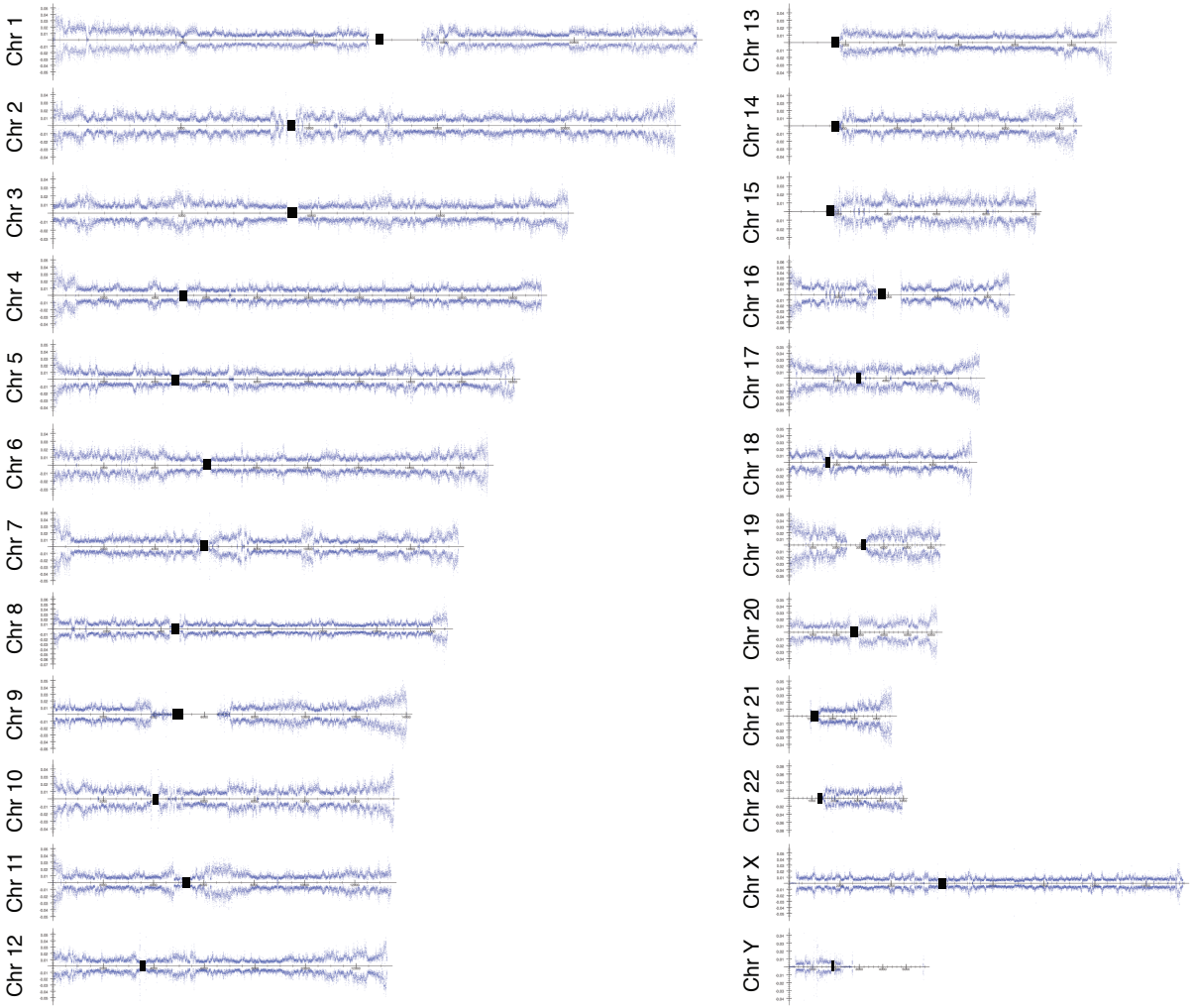
Supplementary Figure 1 | Uniquely Mapped Reads and Coverage for H1 MethylC-Seq. **a,** The number of uniquely mapped MethylC-seq reads for each chromosome of H1 and IMR90. **b,** The percent of the H1 and IMR90 genomes that is covered by differing minimum number of MethylC-seq reads.



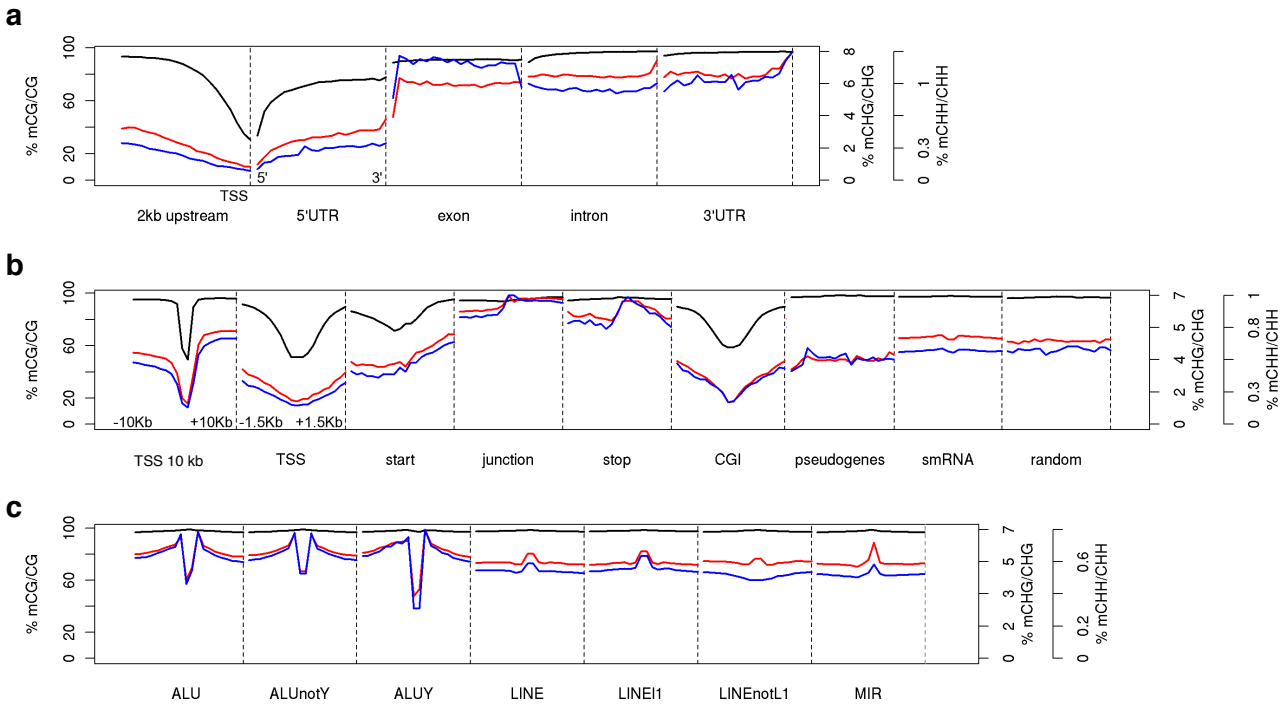
Supplementary Figure 2 | Direct Overlap in Methylcytosines Between the H1 and IMR90 Cell Types, and Regional Correlation of non-CG Methylation Between Biological Replicates and mCHG/mCHH. a, Methylcytosines with similar sequencing depth were compared and classified as unique to biological replicate 1 (red), unique to replicate 2 (yellow) or common to both replicates (orange). The number of methylcytosines in each category is listed, as well as the percent methylcytosines unique within each biological replicate. **b,** Pearson correlation of the density of non-CG methylation sites within adjacent regions of chromosome 1 of varying length between the two H1 biological replicates. The correlation was determined independently for mCHG and mCHH. **c,** Pearson correlation was computed as in panel **b**, comparing mCHG to mCHH density from methylcytosine sites identified in the composite of the two biological replicates. **d,** Scatter plot of mCHG and mCHH density for each promoter, 5' UTR, exon, intron and 3'UTR occurrence. A blue dashed line with slope 1 along regions with equal mCHG and CHH density is displayed. Pearson correlation is reported in the plot title.



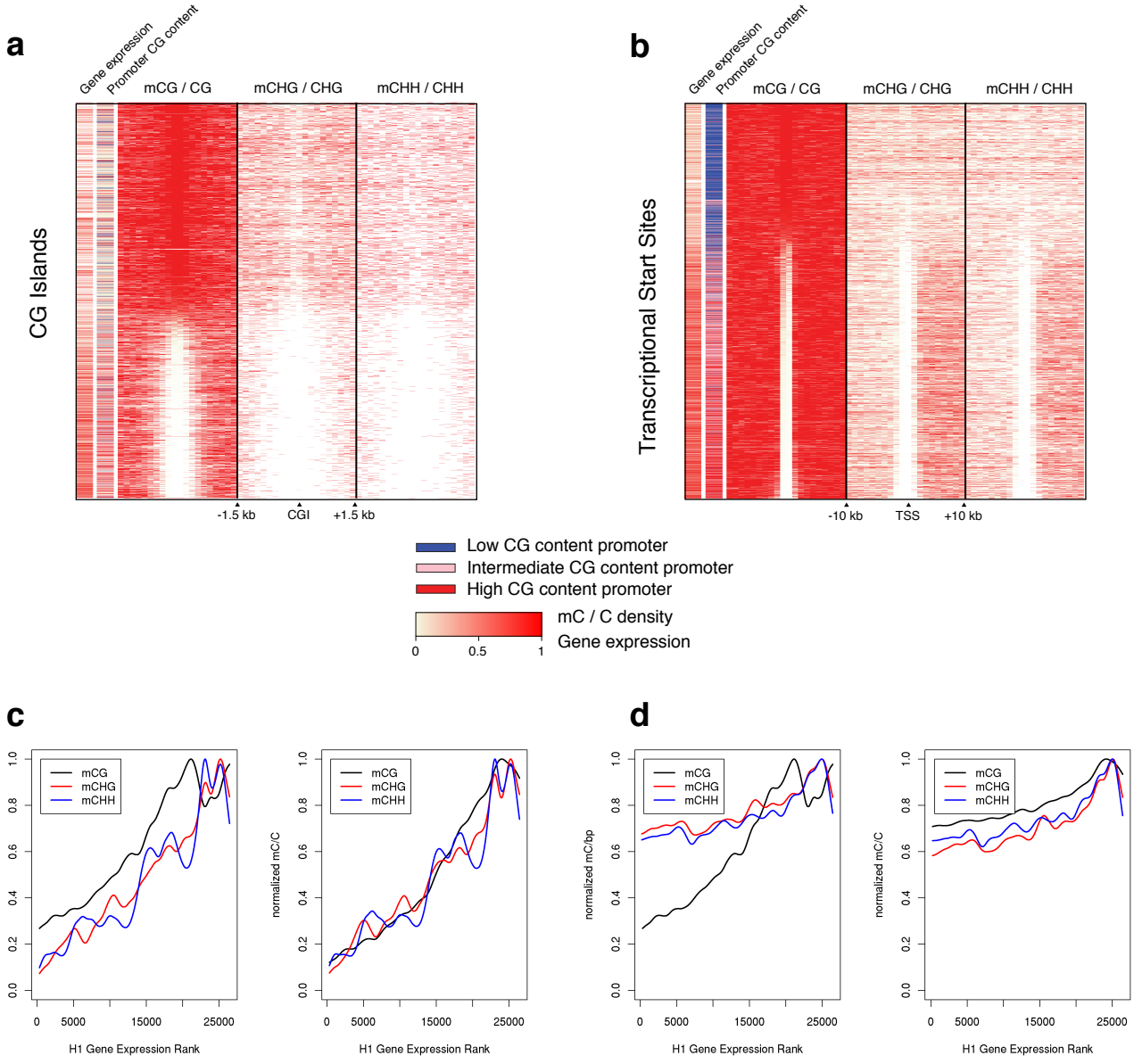
Supplementary Figure 3 | Differentually Expression of *DNMT* Genes in H1 and IMR90. **a,** log2RPKM and **b,** Maximum normalized RPKM measurements of transcript abundance for *DNMT1, DNMT3a, DNMT3b, DNMT3L* and *GAPDH* from RNA-seq. Abbreviations: RPKM, reads per kilobase of exon model per million mapped reads.



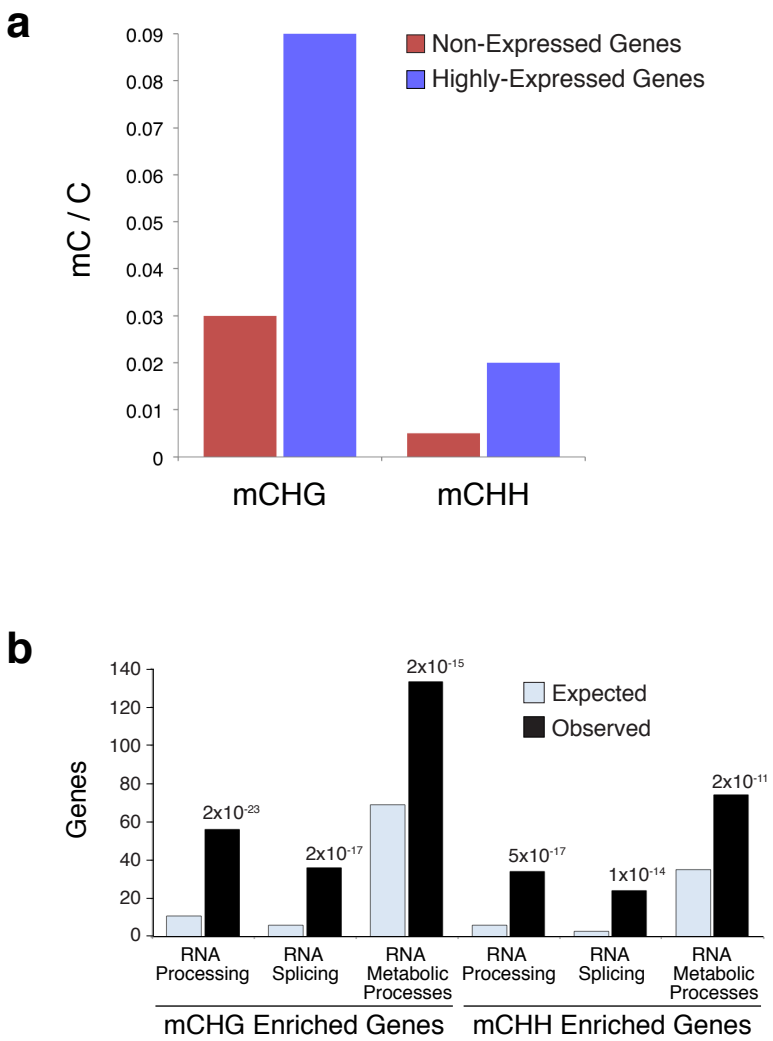
Supplementary Figure 4 | The Density of Methylcytosines Identified in All Chromosomes in H1 Cells. Blue dots indicate the density of all methylcytosines in 10 kb windows. Black rectangles indicate approximate centromere positions.



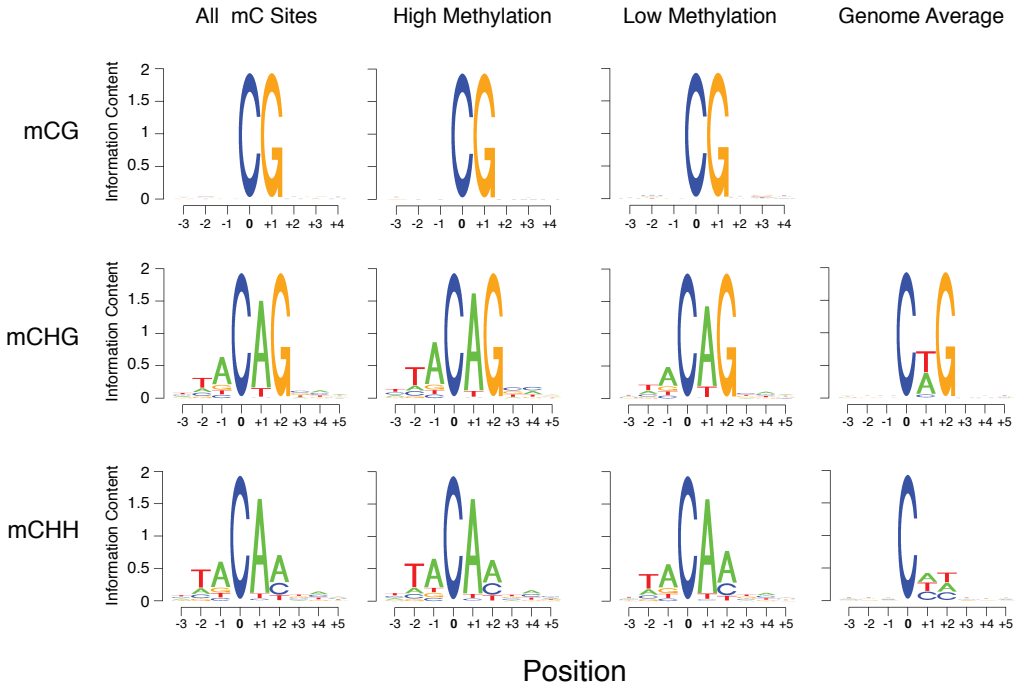
Supplementary Figure 5 | Mean mC/C Profiles Over Genomic Regions. a, gene body regions were divided in 20 bins from 5' to 3' end, and the mean mC/C level within each bin for each methylation type was determined (mCG/CG black, mCHG/CHG red, CHH/CHH blue). Mean over 3Kb regions centered at **b,** regulatory and **c,** transposable repeated genomic regions are displayed.



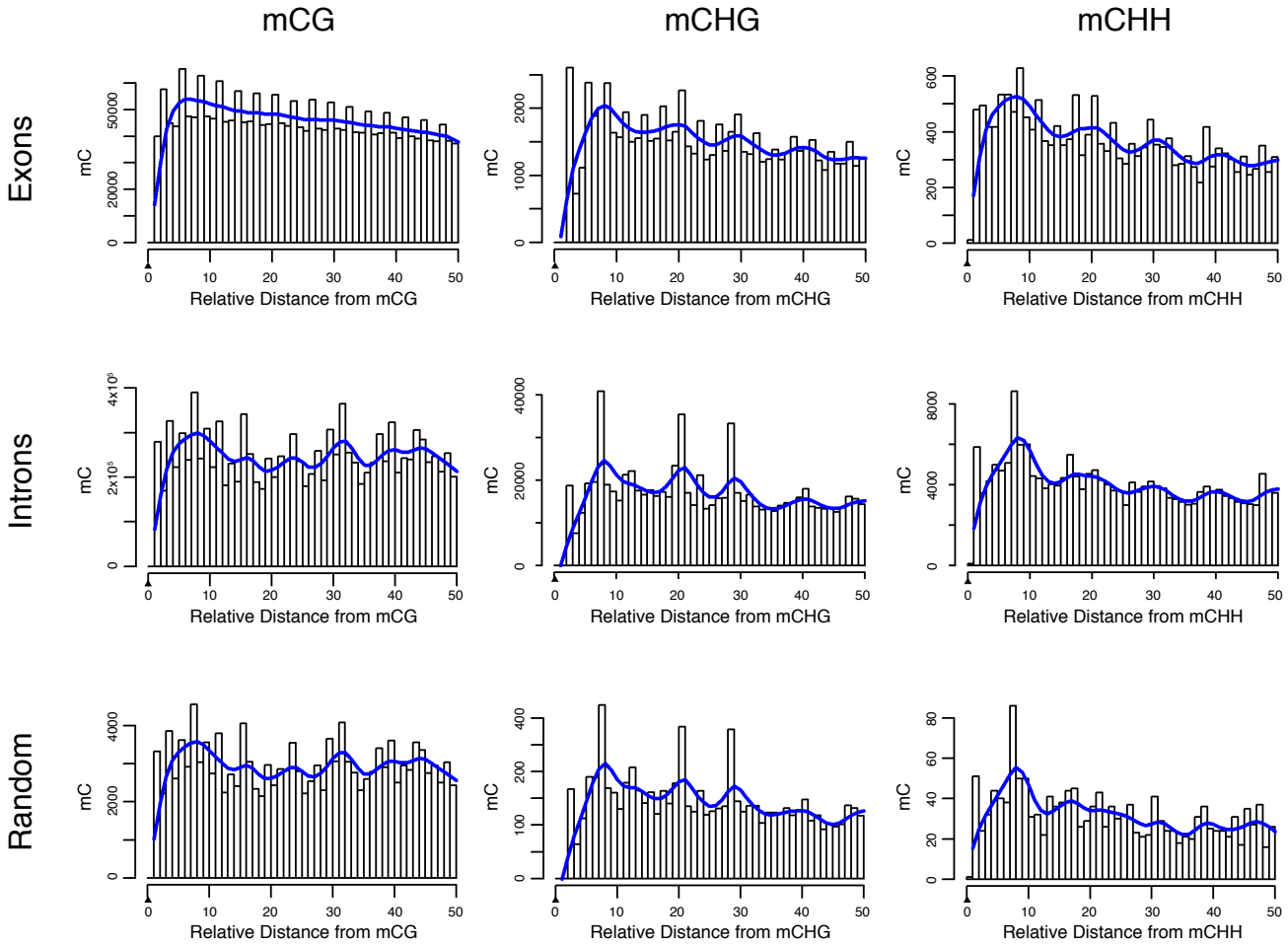
Supplementary Figure 6 | DNA Methylation at CG Islands, Transcriptional Start Sites and Promoters. Relative DNA methylation density at **a,** CG islands (1.5 kb upstream/downstream) and **b,** transcriptional start sites (10 kb upstream/downstream) is displayed with downstream gene expression and promoter CG content. Each CG island was assigned to the closest gene whose TSS is within 10Kb. As expected, low CG content promoters are highly methylated, or close to highly methyalted CG islands, and close to low expressed genes. High CG content promoters are poorly methylated and usually close to highly expressed genes. CG and non-CG methylation density was profiled upstream of the transcriptional start site (TSS) and have compared this to the expression of the downstream gene, for all genes. For both proximal TSS (**c,** defined as -150 bp to +150 bp across TSS) and promoter (**d,** defined as the region 1.5 kb upstream of the TSS) there is a clear anti-correlation of gene expression in respect to both the absolute and relative mC content (mC/bp and mC/C, respectively). This trend is more evident for the region proximal the TSS. Abbrevitations: CGI, CG island. mC, methylcytosine. TSS, transcriptional start site.



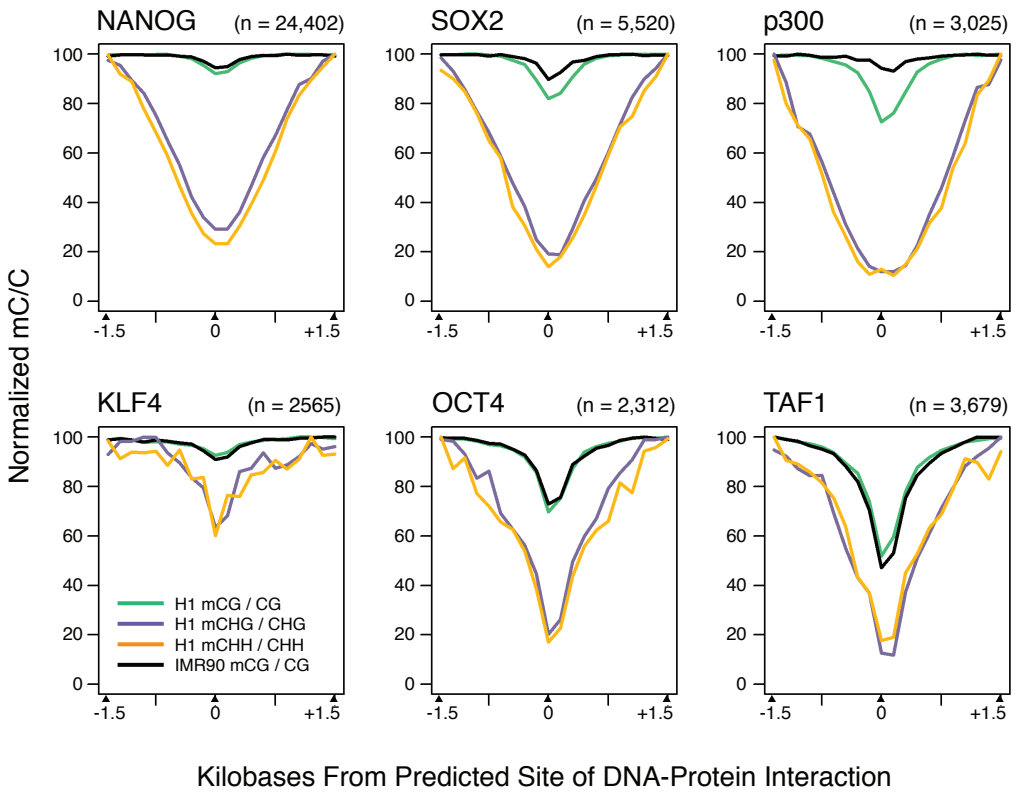
Supplementary Figure 7 | a, Enrichment of non-CG methylation in non-expressed and highly-expressed genes in H1. **b,** Over-representation of GO terms of genes within 20 kb of genomic regions displaying the highest enrichment of CHG and CHH methylation. The enrichment P-value is shown for each GO term.



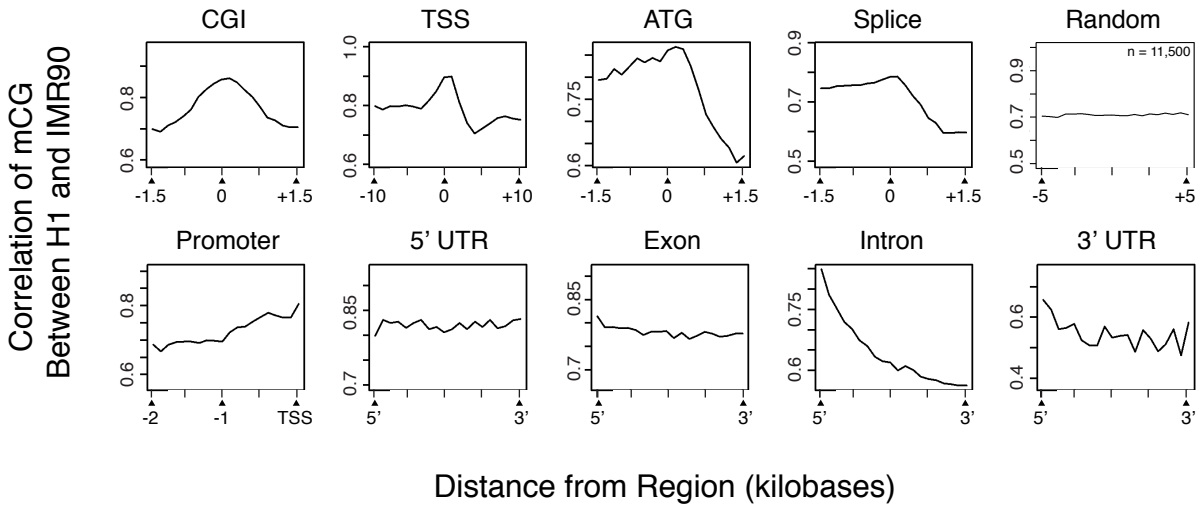
Supplementary Figure 8 | Logo Plots of the Sequences Proximal to Sites of DNA Methylation in Each Sequence Context in H1 Cells. Logo plots are presented for all methylcytosines, and methylcytosines that display a high methylation level (CG ≥75% methylated, non-CG ≥25% methylated), and low methylation level (CG <75% methylated, non-CG <25% methylated). Three bases flanking every site of methylation were analysed to identify local sequence preferences. The information content of each base represents the level of sequence enrichment. Local sequence enrichments were not evident when all cytosines were analyzed, regardless of their methylation status, and the level of methylation at a non-CG methylation site did not appear to influence the local sequence enrichment.



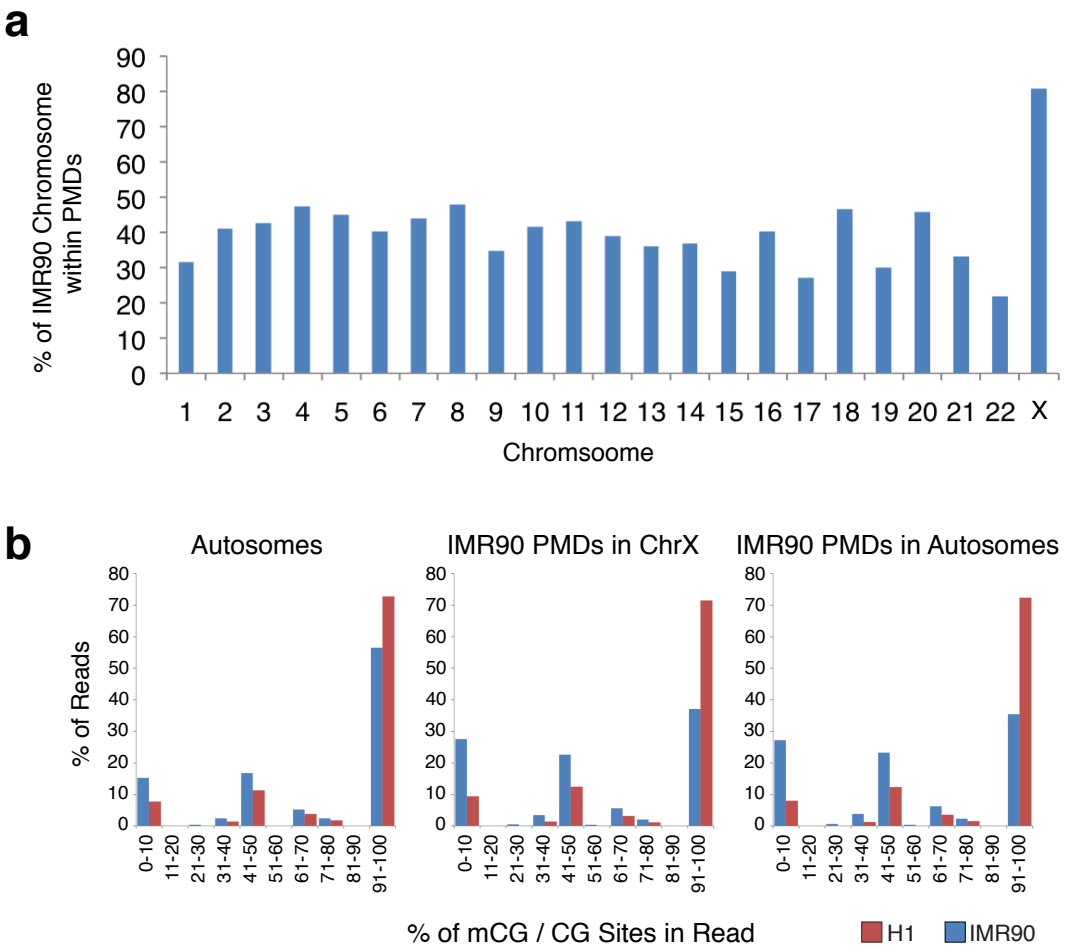
Supplementary Figure 9 | Spacing of Adjacent Methylcytosines in Different Contexts. Prevalence of mCHG/mCHH sites (y-axis) as a function of the number of bases between adjacent mCHG/mCHH sites (x-axis) based on all non redundant pair-wise distances up to 50 nt in exons, introns and random sequences. The blue line represents smoothing by cubic splines.



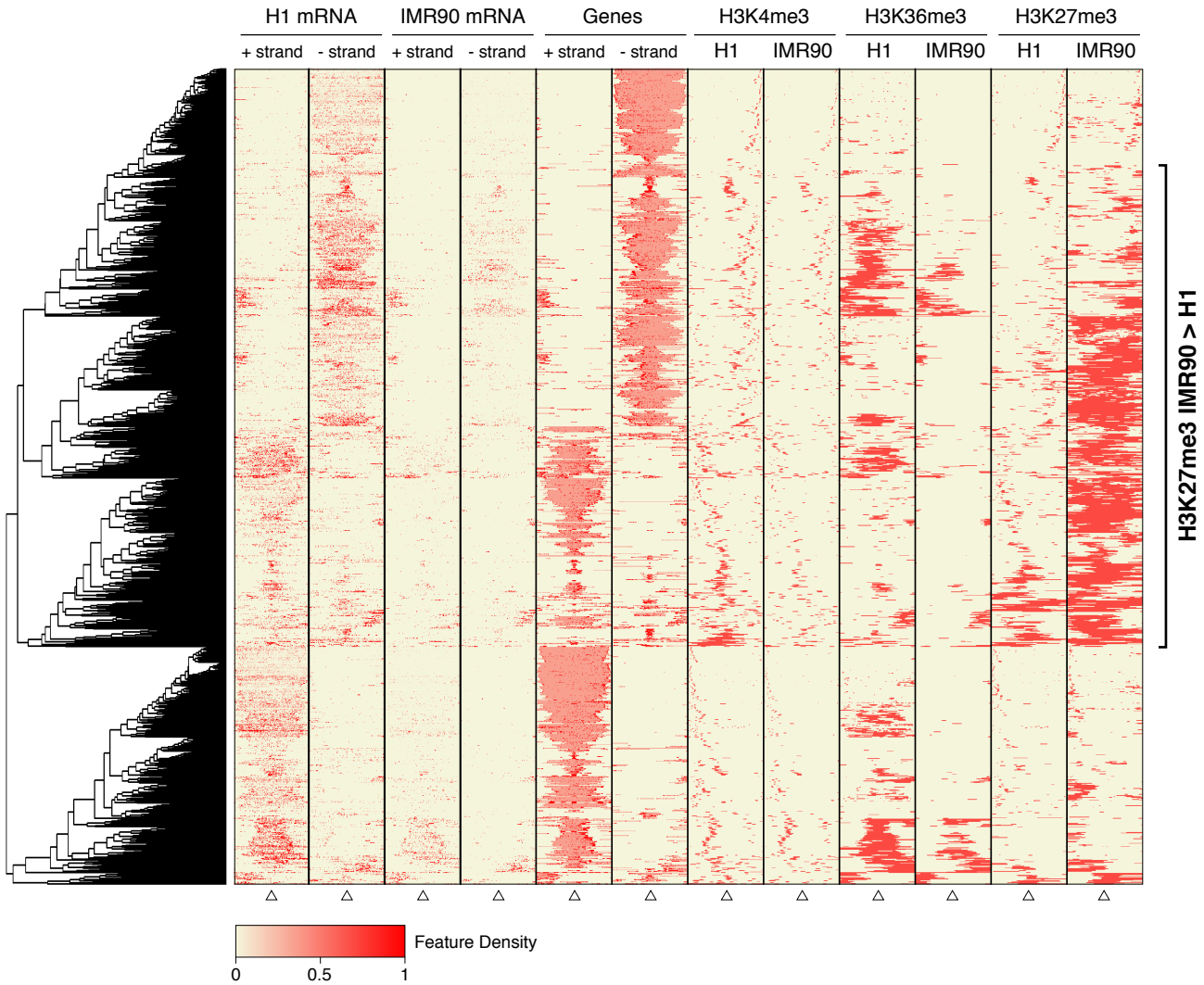
Supplementary Figure 10 | DNA Methylation at Sites of Protein-DNA Interaction. The average relative DNA methylation densities in each sequence context are shown from 1.5 kb upstream to 1.5 kb downstream of the predicted sites of DNA-protein interaction identified by ChIP-seq that were at least 1.5 kb from the closest transcriptional start site.



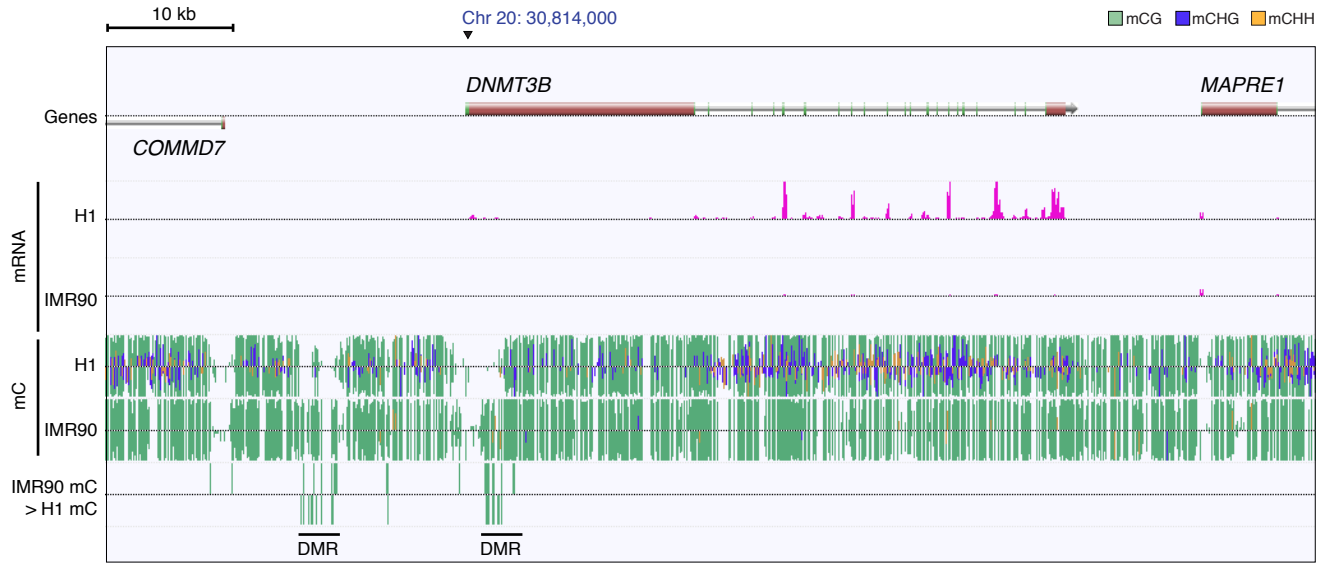
Supplementary Figure 11 | Correlation of DNA Methylation Between IMR90 and H1 at Different Genomic Features. The Pearson correlation coefficient of mCG methylation density (y-axis) between H1 and IMR90 at various genomic features. Regions were divided in 20 equally sized bins from 5' to 3' or based on the distance from the localization of the genomic feature as indicated. Pearson correlation was determined in each bin considering all the H1 and IMR90 occurrences of the given genomic region. An increased and high mCG correlation level was observed in correspondence to genomic regions expected to display a more constitutive epigenetic state, such as CG islands and TSS. We also observed a greater correlation at translational start sites and splice junctions. Gene promoters displayed an increase in correlation as the distance from the TSS decreased. We observed that the correlation in introns is highest toward the 5' exon-intron junction and decreased throughout the length of the introns. Abbreviations: CGI, CG islands. mC, methylcytosine. TSS, transcriptional start site.



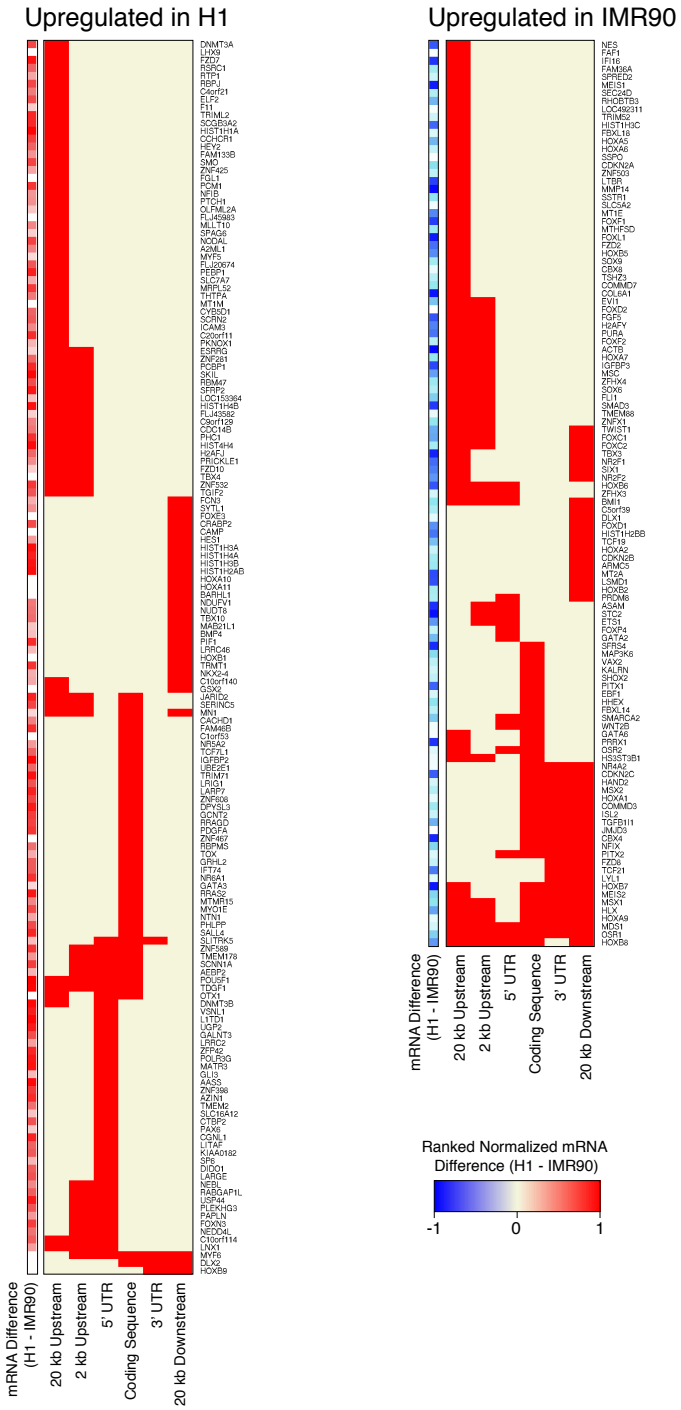
Supplementary Figure 12 | PMDs and the Distribution of Unmethylated, Partial, and Completely Methylated Reads. a, The percent of each IMR90 chromosome that is within PMDs. **b,** For MethylC-seq reads located within a set of genomic regions, the percentage of CG sites within each read that were methylated was calculated, and the percent of all reads within the regions (y-axis) that were methylated at given percentages (x-axis) is displayed. This is presented for H1 and IMR90 MethylC-seq reads in autosomes, in IMR90 partially methylated domains on chromosome X, and IMR90 partially methylated domains in autosomes. Abbreviations: mC, methylcytosine. PMD, partially methylated domain.



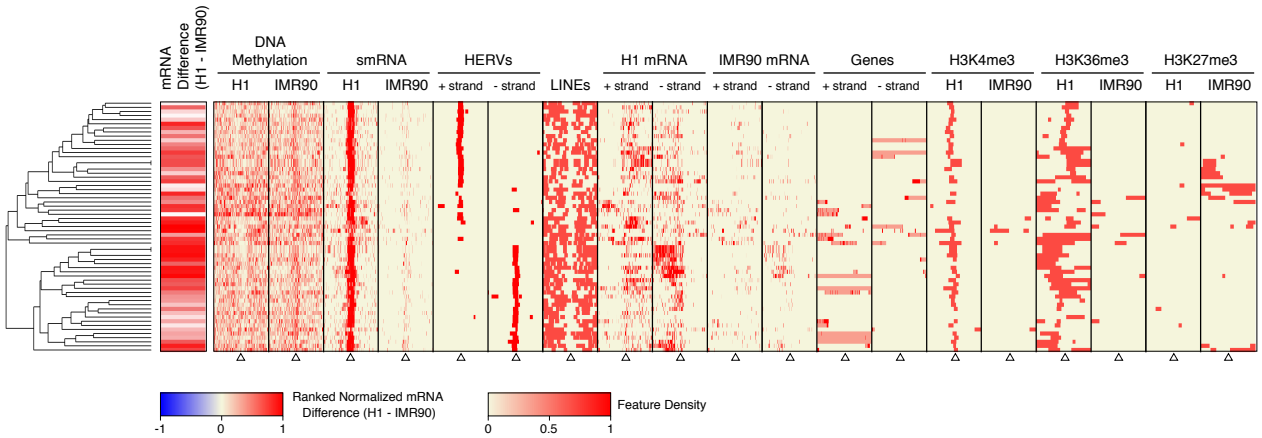
Supplementary Figure 13 | Transcriptional Activity and Epigenetic Modifications at Partially Methylated Domains. The density of strand-specific mRNA reads, and the presence of domains of H3K4me3, H3k36me3 and H3K27me3 in H1 and IMR90 was profiled 20 kb upstream to 20 kb downstream of each gene located in an IMR90 PMD. Open triangles indicate the central point in each 40 kb window. Also displayed is the presence within the Human reference sequence of genes on each strand, where pink coloring indicates the gene body and dark red boxes represent exons. The complete linkage hierarchical clustering of the regions based on these data is presented. Abbreviations: mC, methylcytosine. PMD, partially methylated domain.



Supplementary Figure 14 | Differentially Methylated Regions proximal to *DNMT3B*. AnnoJ genome browser display of DNA methylation and mRNA at two DMRs upstream of *DNMT3B*. For DNA methylation tracks, vertical lines above and below the dotted central line indicate the presence of methylcytosines on the Watson and Crick strands, respectively. The color represents the context of DNA methylation, as indicated, and the vertical height of the line indicates the methylation level of each methylcytosine. The IMR90 > H1 mC track indicates methylcytosines that are significantly more methylated in IMR90 relative to H1 at a 5% FDR (Fisher's Exact Test), and the color represents the context of DNA methylation. Abbreviations: mC, methylcytosine. DMR, differentially methylated region.



Supplementary Figure 15 | Genes Within 20 kb of IMR90 Hypermethylated Regions Being More Expressed in H1 or IMR90. Gene regions bearing differential methylation are indicated in red. Side colorbar displays normalized differential expression (red and blue for genes upregulated in H1 and IMR90, respectively).



Supplementary Figure 16 | Clustering of Genomic, Epigenetic and Transcriptional Features at Differentially Methylated HERVs. The density of DNA methylation, smRNA reads, strand-specific mRNA reads, and the presence of domains of H3K4me3, H3k36me3 and H3K27me3 in H1 and IMR90 was profiled 20 kb upstream to 20 kb downstream of each of the 61 smRNA clusters that co-localize with DMRs. Abbreviations: DMRs, Differentially Methylated Regions. HERVs, Human Endogenous Retroviruses.

SUPPLEMENTARY MATERIALS

METHODS

Cell culture. IMR90 cells were obtained from ATCC and cultured under recommended conditions, during which replicate 1 and 2 cells underwent 4 and 5 passages, respectively. H1 and H9 cells were grown in 10cm^2 dishes (approximately 1 x 10^7 cells / dish) in feeder free conditions on Matrigel (BD Biosciences, San Jose, CA) using quality controlled mTeSR1 media for several passages as described previously[44,45], with/without 200 ng/μl BMP4 for 6 days (RND systems, Minneapolis, MN). The cells used for H1 replicate 1 and 2 cells were passage 25 and 27, including the 9 and 5 passages in mTeSR1 media, respectively. H9 cells were passage 42 including several passages in mTeSR1. IMR90 iPS cells were passage 65, with 33 passages in mTeSR1, and prior to cell harvest aliquots of cells were assayed for *Oct4* expression by flow cytometry as described previously[44,45]. Cells were submitted to the WiCell Cytogenetics Laboratory to confirm normal karyotype.

MethylC-Seq library generation Five μg of genomic DNA was extracted from frozen cell pellets using the DNeasy Mini Kit (Qiagen, Valencia, CA) and spiked with 25 ng unmethylated *c*I857 *Sam7* Lambda DNA (Promega, Madison, WI). The DNA was fragmented by sonication to 50-500 bp with a Bioruptor (Diagenode, Sparta, NJ), followed by end repair with a nucleotide triphosphate mix free of dCTP. Cytosine-methylated adapters provided by Illumina (Illumina, San Diego, CA) were ligated to the sonicated DNA as per manufacturer's instructions for genomic DNA library construction. Adapter-ligated DNA of 140-210 bp was isolated by 2% agarose gel electrophoresis, and sodium bisulfite conversion performed on it using the MethylEasy *Xceed* kit (Human Genetic Signatures, NSW, Australia) as per manufacturer's instructions. One third of the bisulfite-converted, adapter-ligated DNA molecules were enriched by 4 cycles of PCR with the following reaction composition: 2.5 U of uracil-insensitive *PfuTurboC$_x$* Hotstart

DNA polymerase (Stratagene), 5 μl 10X *PfuTurbo* reaction buffer, 25 μM dNTPs, 1 μl Primer 1.1, 1 μl Primer 2.1 (50 μl final). The thermocycling parameters were: 95˚C 2 min, 98˚C 30 sec, then 4 cycles of 98˚C 15 sec, 60˚C 30 sec and 72˚C 4 min, ending with one 72˚C 10 min step. The reaction products were purified using the MinElute PCR purification kit (Qiagen, Valencia, CA) then separated by 2% agarose gel electrophoresis and the amplified product purified from the gel using the MinElute gel purification kit (Qiagen, Valencia, CA). Up to three separate PCR reactions were performed on subsets of the adapter-ligated, bisulfite-converted DNA, yielding up to three independent libraries from the same biological sample. We obtained the final sequence coverage by sequencing all libraries for a sample separately, thus reducing the incidence of "clonal" reads which share the same alignment position and likely originate from the same template molecule in each PCR. Quantitative PCR was used to measure the concentration of viable sequencing template molecules in the library prior to sequencing. The sodium bisulfite non-conversion rate was calculated as the percentage of cytosines sequenced at cytosine reference positions in the Lambda genome.

Small RNA library generation. RNA fractions enriched for small RNAs were isolated from cell pellets treated with RNAlater (Life Technologies, Carlsbad, CA) using the mirVana miRNA isolation kit (Life Technologies, Carlsbad, CA) and treated with DNaseI (Qiagen, Valencia, CA) for 30 min at room temperature. Following ethanol precipitation, the small RNAs were separated by electrophoresis on a 15% TBE-urea gel and RNA molecules between approximately 10 and 50 nt were excised and eluted from the gel fragments. Following ethanol precipitation, smRNA-Seq libraries were produced using the Small RNA Sample Prep v1.5 kit (Illumina, San Diego, CA) as per manufacturer's instructions.

Directional RNA-Seq library generation. Total RNA was isolated from cell pellets treated with RNAlater using the mirVana miRNA isolation kit and treated with DNaseI (Qiagen, Valencia, CA) for 30 min at room temperature. Following ethanol precipitation,

biotinylated LNA oligonucleotide rRNA probes complementary to the 5S, 5.8S, 12S, 18S and 28S ribosomal RNAs were used to deplete the rRNA from 20 μg of total RNA in two sequential RiboMinus reactions (Life Technologies, Carlsbad, CA) as per manufacturer's instructions. Two hundred ng of the remaining RNA was fragmented by metal hydrolysis in 1X fragmentation buffer (Life Technologies, Carlsbad, CA) for 15 min at 70˚C, stopping the reaction by addition of 2 μl fragmentation stop solution (Life Technologies, Carlsbad, CA). Fragmented RNA was treated with 5 U Antarctic phosphatase (New England Biolabs, Ipswich, MA) for 40 min at 37˚C in the presence of 40 U RNaseOut (Life Technologies, Carlsbad, CA), followed by phosphatase heat inactivation at 65˚C for 5 min. Phosphorylation was performed by addition of 10 U PNK (New England Biolabs, Ipswich, MA), 1 mM ATP, and 20 U RNaseOut and incubation at 37˚C for 1 h. The RNA was purified using 66 μl SPRI beads (Agencourt, Beverly, MA) and eluted in 11 μl 10 mM Tris buffer pH 8.0. One μl of 1:10 diluted adenylated 3' RNA adapter oligonucleotide (5'-UCGUAUGCCGUCUUCUGCUUGidT-3') was added to the phosphorylated RNA and incubated at 70˚C for 2 min followed by placement on ice. The 3' RNA adapter ligation reaction was performed by addition of 2 μl 10x T4 RNA ligase 2 truncated ligation buffer, 1.6 μl 100 mM $MgCl_2$, 20 U RNaseOut and 300 U T4 RNA ligase 2 truncated (New England Biolabs, Ipswich, MA) and incubation at 22˚C for 1 h. Ligation of the 5' RNA adapter was performed by addition to the 3' adapter ligated reaction of 1 μl 1:1 diluted, heat denatured (70˚C 2 min) 5' RNA adapter oligonucleotide (5'-GUUCAGAGUUCUACAGUCCGACGAUC-3'), 1 μl 10 mM ATP, and 10 U T4 RNA ligase (Promega, Madison, WI), and incubation at 20˚C for 1 h. The RNA was purified using 66 μl SPRI beads (Agencourt, Beverly, MA) and eluted in 10 μl 10 mM Tris buffer pH 8.0. To the RNA ligation products, 2 μl 1:5 diluted RT primer (5'-CAAGCAGAAGACGGCATACGA-3') was added and heat denatured (70˚C 2 min), followed by incubation on ice. Added to the denatured RNA/primer solution was 4 μl 5x first strand buffer, 1 μl 12.5 mM dNTPs, 2 μl 100 mM DTT, and 40 U RNaseOut, followed by incubation at 48˚C for 1 min. To this, 200 U Superscript II reverse transcriptase (Life Technologies, Carlsbad, CA) was added, followed by incubation at 44˚C for 1 h. The RT reaction was used in a PCR enrichment containing 0.25 μM GEX1

(5'-AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA-3') and 0.25 μM GEX2 (5'-CAAGCAGAAGACGGCATACGA-3') primers, 0.25 mM dNTPs, 1x Phusion polymerase buffer and 4 U Phusion hot-start high fidelity DNA polymerase (New England Biolabs, Cambridge, MA) in a 100 μl reaction using the following thermocycling parameters: 98˚C 30 sec, then 15 cycles of 98˚C 10 sec, 60˚C 30 sec and 72˚C 15 sec, ending with one 72˚C 10 min step. The PCR products were purified in two steps, first by purification using 180 μl SPRI beads and elution in 30 μl 10 mM Tris buffer pH 8.0, followed by purification with 39 μl SPRI beads and elution in 10 μl 10 mM Tris buffer pH 8.0. All oligonucleotides were obtained from Illumina (San Diego, CA). Quantitative PCR was used to measure the concentration of viable sequencing template molecules in the library prior to sequencing.

Chromatin immunoprecipitation and ChIP-Seq library generation. Chromatin immunoprecipitation (ChIP) for SOX2 (R&D Systems, #AF2018; 5ug) and NANOG (R&D Systems, #AF1997, 5ug) was performed as recently described (Hawkins *et al*., submitted). ChIP for OCT4 (Santa Cruz, #sc8626, 2ug; Santa Cruz, #sc9081, 2ug; R&D Systems, #AF17566, 2ug), p300 (Santa Cruz, #sc585, 5ug), KLF4 (Abcam, #ab21949, 10ug) TAFIIp250/TAF1 (Santa Cruz, #sc735, 5ug) were carried out as previously described with 500ug chromatin and 2-10ug antibody[47,48]. ChIP libraries for sequencing were prepared following standard protocols from Illumina (San Diego, CA) with the following minor modifications. Following linker ligation, libraries were run on an 8% acrylamide gel and size selected for 175 - 250bp. This was repeated following PCR amplification. After each size selection, acrylamide was shredded and incubated with 300ul EB buffer (Qiagen, Valencia, CA) overnight at 4ºC or 50 ºC for 20 mins with shaking. DNA was eluted using Nanosep MF filter tubes (Pall, East Hills, NY). The experimental detail and in depth data analysis of the histone modifications will be described separately (Hawkins *et al*., Submitted).

High-throughput sequencing. MethylC-Seq and RNA-Seq libraries were sequenced using the Illumina Genome Analyzer II (GA II) as per manufacturer's instructions.

Sequencing of MethylC-Seq libraries was performed up to 87 cycles to yield longer sequences that are more amenable for unambiguous mapping to the human genome reference sequence. Image analysis and base calling were performed with the standard Illumina pipeline (Firecrest v1.3.4 and Bustard v1.3.4), performing automated matrix and phasing calculations on the PhiX control that was run in the eighth lane of each flowcell.

Validation of bisulfite sequencing results. Primers were designed to amplify a limited number of specific regions of the genome following bisulfite conversion. Genomic DNA was isolated from H1, BMP-treated H1, H9, IMR90 and IMR90 iPS cells, fragmented by sonication and 1 μg of genomic DNA from each sample was bisulfite converted according to the procedures described above. For each cell type, approximately one tenth of the converted sample was used in 3 distinct PCR reactions (Master*Taq* Kit, 5 Prime, Gaithersburg, MD), each containing a different pair of primers designed to amplify a distinct genomic region (Supplementary Table 2). Amplified products were separated by gel electrophoresis, gel purified, and cloned using the Zero Blunt TOPO PCR cloning kit (Life Technologies, Carlsbad, CA). Sanger sequencing of multiple clones for each cell type and amplicon was performed to identify the methylation status of cytosines within each region.

DATA ANALYSIS

Processing and alignment of MethylC-Seq read sequences. Read sequences produced by the Illumina pipeline in FastQ format were first pre-processed in three steps. Firstly, reads were trimmed to before the first occurrence of a low quality base (PHRED score ≤ 2). Secondly, as a subset of reads contained all or part of the 3' adapter oligonucleotide sequence, every read was searched for the adapter sequence, and if detected the read was trimmed to the preceding base. If the full adapter sequence was not detected, iterative searching of the *k* 3' terminal bases of the read for the *k* 5' bases of the adapter was performed, and if detected the read was trimmed to the

preceding base. Thirdly, any cytosine base in a read was replaced with thymine. Following pre-processing, reads were sequentially aligned using the Bowtie algorithm[46] (v0.9.9.1) to two computationally converted NCBI BUILD 36/HG18 reference sequences, the first in which cytosines were replaced with thymines, and the second in which guanines were replaced with adenines. The 48,502 bp Lambda genome was included in the reference sequence as an extra chromosome so that reads originating form the unmethylated control DNA could be aligned. As all cytosines in the reads were replaced with thymines, the methylation status of a particular genomic sequence has no bearing on its ability to map to the reference. Sequences originating from the Watson strand of the genome aligned to the cytosine-free reference sequence, whereas sequences originating from the Crick strand (complement) of the genome aligned to the guanine-free reference sequence after reverse complementation. The following parameters were used in the Bowtie alignment process: --solexa-quals -e 140 -l 20 -n 0 -k 10 --best --nomaqround. For each read, up to 10 of the most highly scoring alignment positions in the reference sequences were returned, tolerating a maximum sum quality score of 140 at mismatch positions. All results of aligning a read to both the Watson and Crick converted genome sequences were combined, and if more than one alignment position existed for a read it was categorized as ambiguously aligned and disregarded. For each cell line, the reads from two biological replicates were pooled to provide greater coverage for identification of the methylcytosines that are presented in this study. Additionally, parallel analysis was performed on each biological replicate to analyse the variability of DNA methylation. Whole lanes of aligned read sequences were combined in a manner based on the experimental setup. As up to three independent libraries from each biological replicate were sequenced, we first removed reads that shared the same 5' alignment position within each library, referred to as "clonal" reads, leaving the read at that position that had the highest sum quality score. Subsequently, the reads from all libraries of a particular sample were combined. All unambiguous, or "unique", read alignments were then subjected to post-processing, which consisted of 3 steps. Firstly, if a read contained more than 3 mismatches compared to the reference sequence, it was trimmed to the base preceding the fourth mismatch. Secondly, the

cytosines that were originally removed from the read sequences prior to alignment were incorporated back into the aligned reads. Thirdly, to remove reads that were likely not bisulfite converted, reads that contained more than 3 cytosines in a non-CG context were discarded. Finally, the number of calls for each base at every reference sequence position and on each strand was calculated. Read number for each replicate before and after removal of clonal reads and post-processing is detailed in Table S1.

Identification of methylated cytosines. At each reference cytosine the binomial distribution was used to identify whether at least s subset of the genomes within the sample were methylated, using a 0.01 FDR corrected P-value. Each context of methylation was considered independently: CG, CHG or CHH (where H = A, C or T). We identified methylcytosines while keeping the number of false positives methylcytosine calls below 1% of the total number of methylcytosines we identified. The probability p in the binomial distribution B(n,p) was estimated from the number of cytosine bases sequenced in reference cytosine positions in the unmethylated Lambda genome (referred to as the error rate: non-conversion plus sequencing error frequency). The bisulfite conversion rates for all samples were over 99%, and the error rates were as follows: H1 replicate 1, 0.007; H1 replicate 2, 0.004; H1 combined replicates, 0.0050; IMR90 replicate 1, 0.002; IMR90 replicate 2, 0.003; IMR90 combined replicates, 0.0024. We interrogated the sequenced bases at each reference cytosine position one at a time, where read depth refers to the number of reads covering that position. For each position, the number of trials (n) in the binomial distribution was the read depth. For each possible value of n we calculated the number of cytosines sequenced (k) at which the probability of sequencing k cytosines out of n trials with an error rate of p was less than the value M, where M * (number of unmethylated cytosines) < 0.01 * (number of methylated cytosines). In this way, we established the minimum threshold number of cytosines sequenced at each reference cytosine position at which the position could be called as methylated, so that out of all methylcytosines identified no more than 1% would be due to the error rate.

Correction of DNA methylation context calls proximal to SNPs. As the cell lines studied have distinct genotypes compared to the Human reference sequence, the sequencing data downstream of every site of non-CG methylation was interrogated to determine whether the cytosine in the H1 and IMR90 cell lines was truly in the non-CG context. If the consensus call at the base downstream (+1) of a non-CG methylcytosine was a guanine, the methylcytosine context was corrected to mCG. Furthermore, the context of any methylcytosine that had been identified on the opposite strand to the +1 guanine was subsequently corrected to mCG. At positions where +1 bases were potentially heterozygous for a SNP, two conditional tests were performed on the surrounding sequence to test for any evidence that the site represented a CG dinucleotide. Firstly, when there was sequence coverage on the opposite strand, if the +1 position displayed at least 20% guanine and on the opposite strand displayed at least 20% cytosine, the methylcytosine context was corrected to mCG. Furthermore, a methylcytosine was added on the opposite strand at this site if the base calls at the position passed the binomial test to the same significance threshold as used in the initial methylcytosine calling. Secondly, if the strand opposite the +1 position had no sequence coverage and the +1 position displayed a similar number of guanine base calls as the cytosine calls at the methylcytosine, the methylation context was corrected to mCG.

Identification of differentially methylated cytosines. For each cell type the DNA methylation data is comprised of the combination of MethylC-Seq performed on the two biological replicates of different passage number. To compare the mCG overlap between the two biological replicates for H1 and IMR90 cells, the mCG from the binomial distribution analysis from each replicate were selected and the read coverage for each replicate was determined at each position. To compare only mCG that possess similar sequence read coverage, a ratio of the coverage between replicates was calculated and only positions with a depth ratio between 0.8 and 1.2 were considered for the Venn diagram analysis. The mCHG and mCHH for the H1 biological replicates were compared in an identical fashion.

A two-tailed Fisher's Exact Test was used to identify cytosines that are differentially methylated between the H1 and IMR90 cell types. Only mCG determined using the binomial distribution analysis in at least one cell type and those mCG covered by at least 3 reads in at least one cell type were considered for testing. P-value thresholds were selected such that the number of false positives is less than 5% of all mCG positions called as significantly different (5% FDR). A total of 6,023,738 mCG where identified as more highly methylated in H1 cells (p-value < 0.007433) and 124,161 mCG where identified as more highly methylated in IMR90 cells (p-value < 0.000153).

mCHG and mCHH enriched genes. Density of methylated or all occurrences of CHG and CHH in 10Kb regions throughout the genome was determined. The hypergeometric distribution was used to determine the enrichment of methylated occurrences in comparison to the total number of sites in a given window, taking into account the total number of methylated and total occurrences across the whole chromosome. Windows with over-representation P-value less then 1e-20 were considered and Ref Seq whose TSS is within 10Kb from the centre of each window were selected.

Genome annotation. Genomic regions were defined based on NCBI BUILD 36/HG18 coordinates downloaded from UCSC web site. Promoters are arbitrarily defined as regions 2Kb upstream the TSS for each RefSeq transcript. According to the UCSC annotation many Ref Seq transcripts can be associated with a given gene, and they can have the same or alternative TSS. Gene bodies are defined as the transcribed regions, from the start to the end of transcription sites for each RefSeq. In case of genomic regions with strand information, those on the reverse strand were reversed. Consequently, mean profiles over all the occurrences of a genomic region on the genome are oriented from 5' to 3'.

mC and mC/C methylation profiles. Genomic regions were divided in 20 uniformly sized bins. In particular, for genomic regions in genes, the 20 bins span from the 5' to

the 3' end. Rather, for genomic regions centered at annotated genomic elements or obtained by ChIP-Seq experiments, an arbitrarily sized window was centered at the centre of each genomic element or ChIP-Seq peak, as indicated in the figures or figure legends. All occurrences of genomic regions were checked for having sufficient coverage in H1 and IMR90 methylomes. Regions with more than one quarter insufficiently covered (less than a total of 3 reads in both strands) are masked. For regions centered at annotated features the same criteria were applied to check the coverage in the central 10% of the region. Masked genomic regions were not used in the determination of the mean profile.

Absolute (mC) methylation content was determined for each bin based on the number of calls of a given methylation type (mCG, mCHG or mCHH) divided by the bin size. For the symmetric mCG, sites where methylation is observed in at least one strand were counted, while for mCHG and mCHH this measure is determined as the sum of methylation calls of a given type on both strands. Relative methylation content (mC/C) was determined as the absolute methylation content divided by the total number of sites of the same type on the genome independently from their methylation level. In particular, for mCG the total number of CG sites was determined only for one strand, as there is a correspondent number if the same sites on the opposite strand. Rather, for mCHG and mCHH, the total number of CHG or CHH occurrences on the genome was determined. Analysis of NCBI BUILD 36/HG18 genome reference sequence was performed using R and Bioconductor tools and annotation libraries (www.r-project.org, www.bioconductor.org)[49].

Identification of differentially methylated regions (DMRs). A sliding window approach was used to find regions of the genome enriched for sites of higher levels of DNA methylation in IMR90 relative to H1, as identified by Fisher's Exact Test. A window size of 1 kb was used, progressing 100 bp per iteration. When a 1 kb window containing at least 4 differential mCG was identified, the region was extended in 1 kb increments until a 1 kb increment was reached that contained less than 4 differential mCG. After

extension termination, a region containing at least 10 differential mCG and at least 2 kb in length were reported as a DMRs.

Identification of partially methylated domains (PMDs). A sliding window approach was used to find regions of the genome in IMR90 that were partially methylated, based on the measurements of the level of methylation at each mCG. A window size of 10 kb was used, progressing 10 kb per iteration. When a 10 kb window was identified that contained at least 10 mCG, each covered by at least 5 MethylC-Seq reads, for which the average methylation level of these mCG was less than 70%, the region was extended in 10 kb increments. Extension was terminated when a 10 kb increment was reached that had an average methylation level of greater than 70% or less than 10 mCG, and the region was reported as a PMD.

Mapping smRNA-Seq reads. smRNA sequence reads in FastQ format were produced by the Illumina analysis pipeline. smRNA-Seq reads that contained at least 5 bases of the 3' adapter sequence were selected and this adapter sequence removed, retaining the trimmed reads that were from 16 to 37 nt in length. These processed reads in FastQ format were aligned to the human reference genome (NCBI BUILD 36/HG18) with the Bowtie alignment algorithm using the following parameters: --solexa-quals -e 1 -l 20 -n 0 -a -m 1000 --best --nomaqround. Consequently, any read that aligned with no mismatches to the and to no more than 1000 locations in the NCBI BUILD 36/HG18 reference genome sequence was retained for downstream analysis.

Identification of smRNA clusters. A sliding window approach was used to find regions of the genome in that displayed dense clusters of smRNAs. A window size of 1 kb was used, progressing 200 bp per iteration. When a 1 kb window was identified that contained more than 10 non-redundant smRNA reads the region was extended in 500 bp increments until a 500 bp increment was reached that contained less than 3 non-redundant smRNA reads. After extension termination, a region containing at least 100 smRNA and at least 3 kb in length were reported as a smRNA cluster.

Mapping RNA-Seq reads. Read sequences produced by the Illumina analysis pipeline were aligned with the ELAND algorithm to the NCBI BUILD 36/HG18 reference sequence and a set of splice junction sequences generated from known splice junctions in the UCSC Known Genes. Reads that aligned to multiple positions were discarded. Reads per kilobase of transcript per million reads (RPKM) were calculated with the CASAVA software package.

Mapping and enrichment analysis of ChIP-Seq reads. Following sequencing cluster imaging, base calling and mapping were conducted using the Illumina pipeline. Clonal reads were removed from the total mapped tags, retaining only the monoclonal unique tags that mapped to one location in the genome, where each sequence is represented once. Regions of tag enrichment were identified as recently described (Hawkins *et al*., submitted).

Data visualization in the AnnoJ browser. MethylC-Seq, RNA-Seq, ChIP-Seq and smRNA-Seq sequencing reads and positions of methylcytosines with respect to the NCBI BUILD 36/HG18 reference sequence, gene models and functional genomic elements were visualized in the AnnoJ 2.0 browser, as described previously[15]. The data mentioned above can be viewed in the AnnoJ browser at: http://neomorph.salk.edu/human_methylome.

SUPPLEMENTARY NOTES

Supplementary References

15. Lister, R. et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**, 523-536 (2008).

44. Ludwig, T. et al. Feeder-independent culture of human embryonic stem cells. *Nature Methods* **3**, 637-646 (2006).
45. Ludwig, T. et al. Derivation of human embryonic stem cells in defined conditions. *Nat Biotechnol* **24**, 185-187 (2006).
46. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
47. Heintzman, N. D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311-318 (2007).
48. Kim, T. H. et al. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128**, 1231-1245 (2007).
49. Gentleman, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).