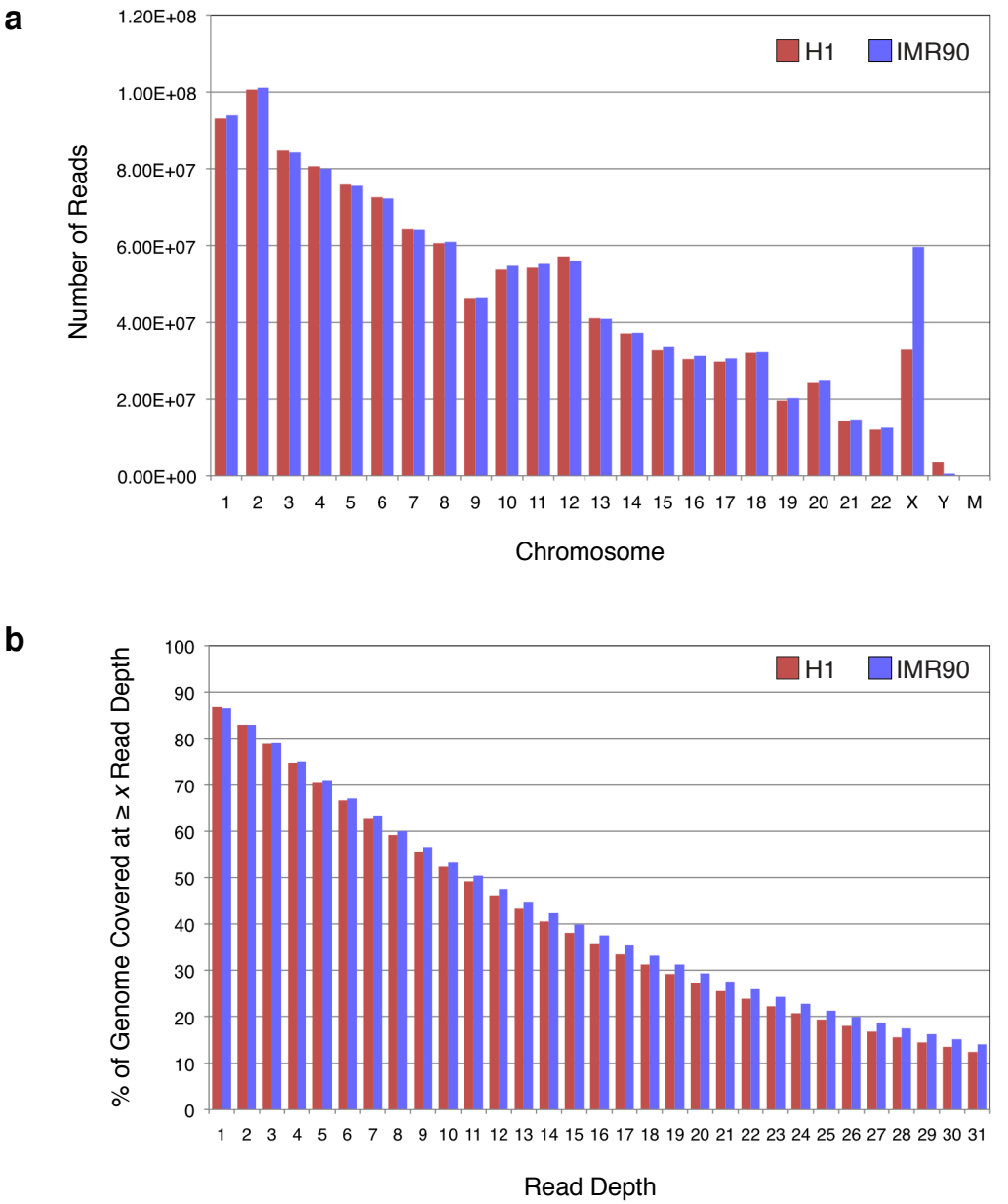
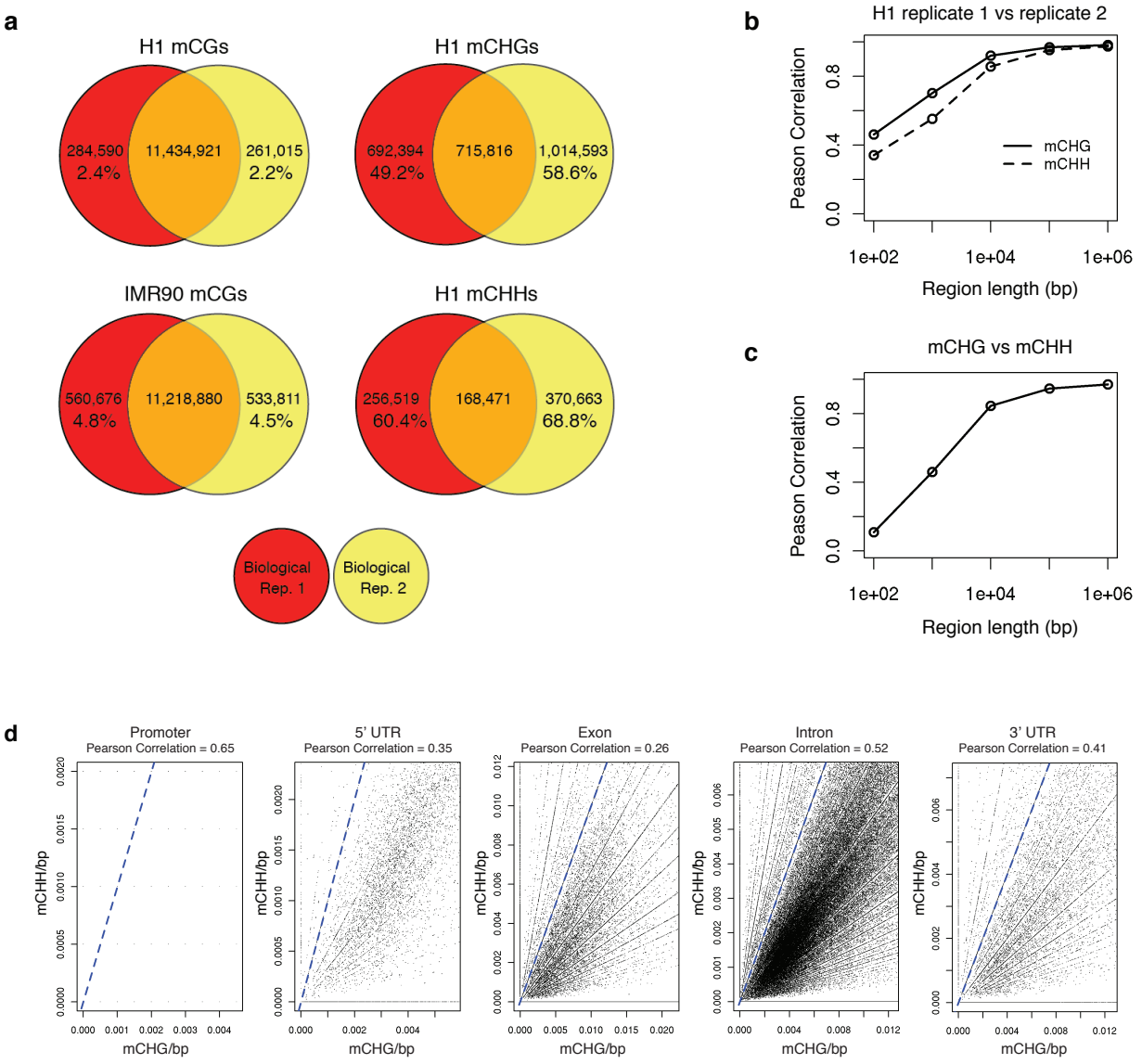
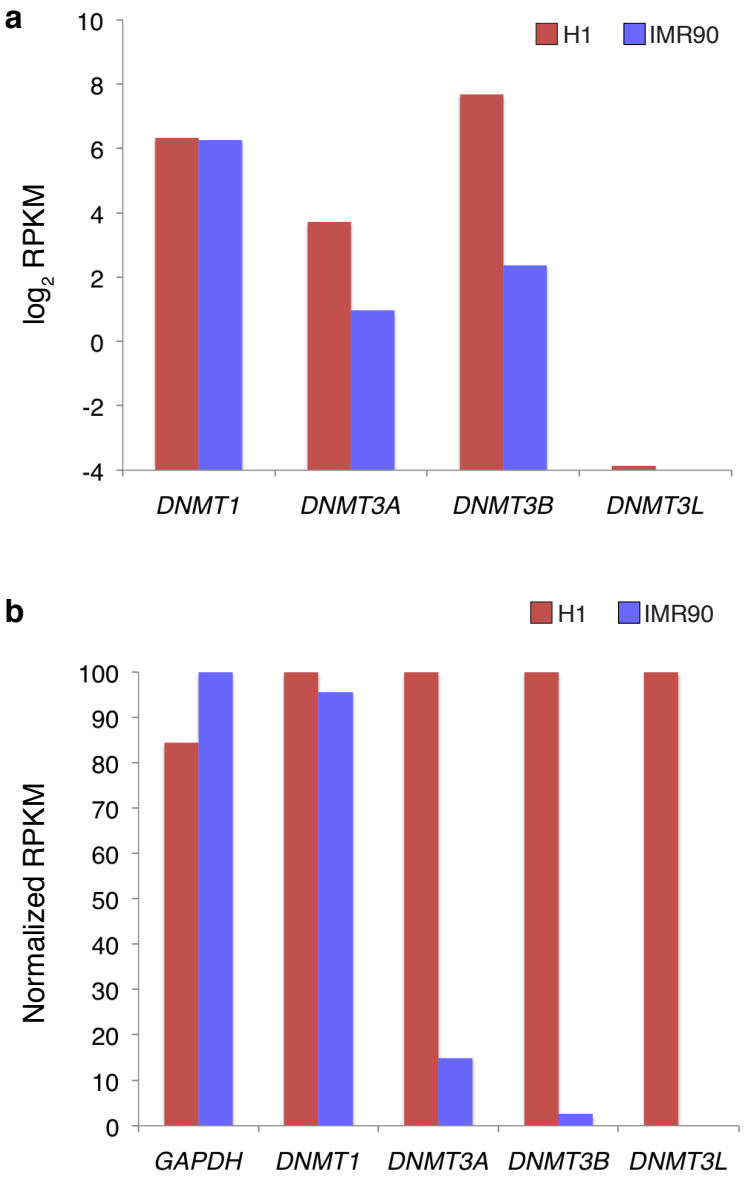


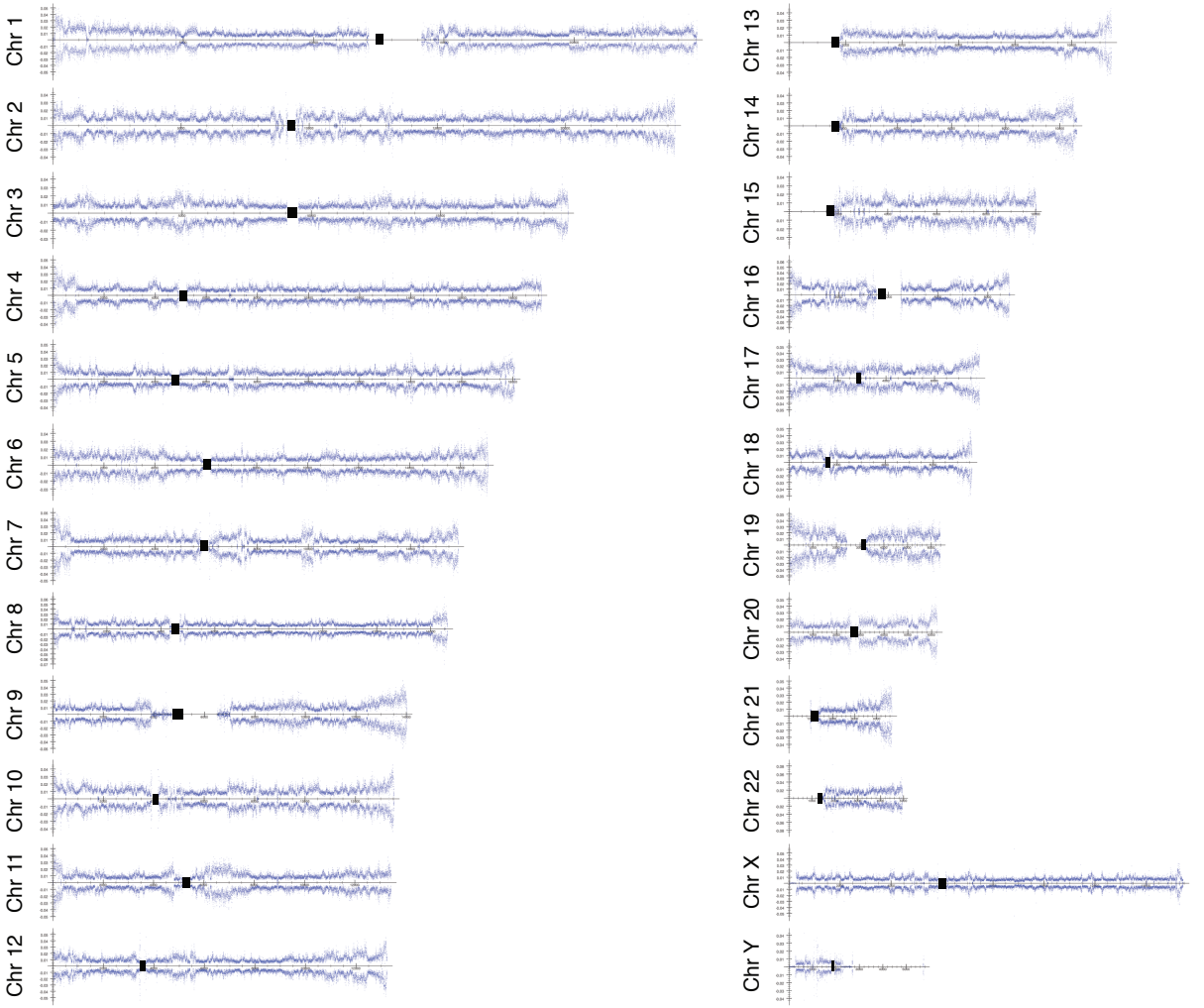
Supplementary Figure 1 | Uniquely Mapped Reads and Coverage for H1 MethylC-Seq. **a,** The number of uniquely mapped MethylC-seq reads for each chromosome of H1 and IMR90. **b,** The percent of the H1 and IMR90 genomes that is covered by differing minimum number of MethylC-seq reads.



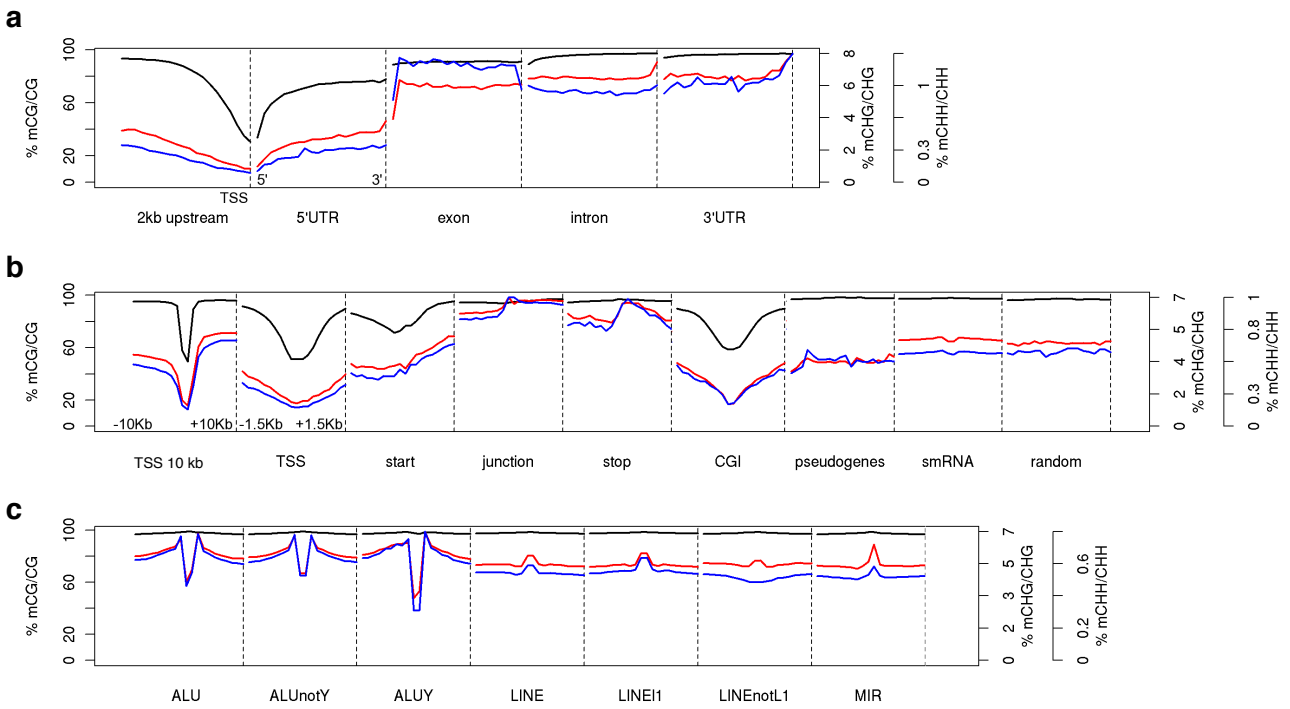
Supplementary Figure 2 | Direct Overlap in Methylcytosines Between the H1 and IMR90 Cell Types, and Regional Correlation of non-CG Methylation Between Biological Replicates and mCHG/mCHH. a, Methylcytosines with similar sequencing depth were compared and classified as unique to biological replicate 1 (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.



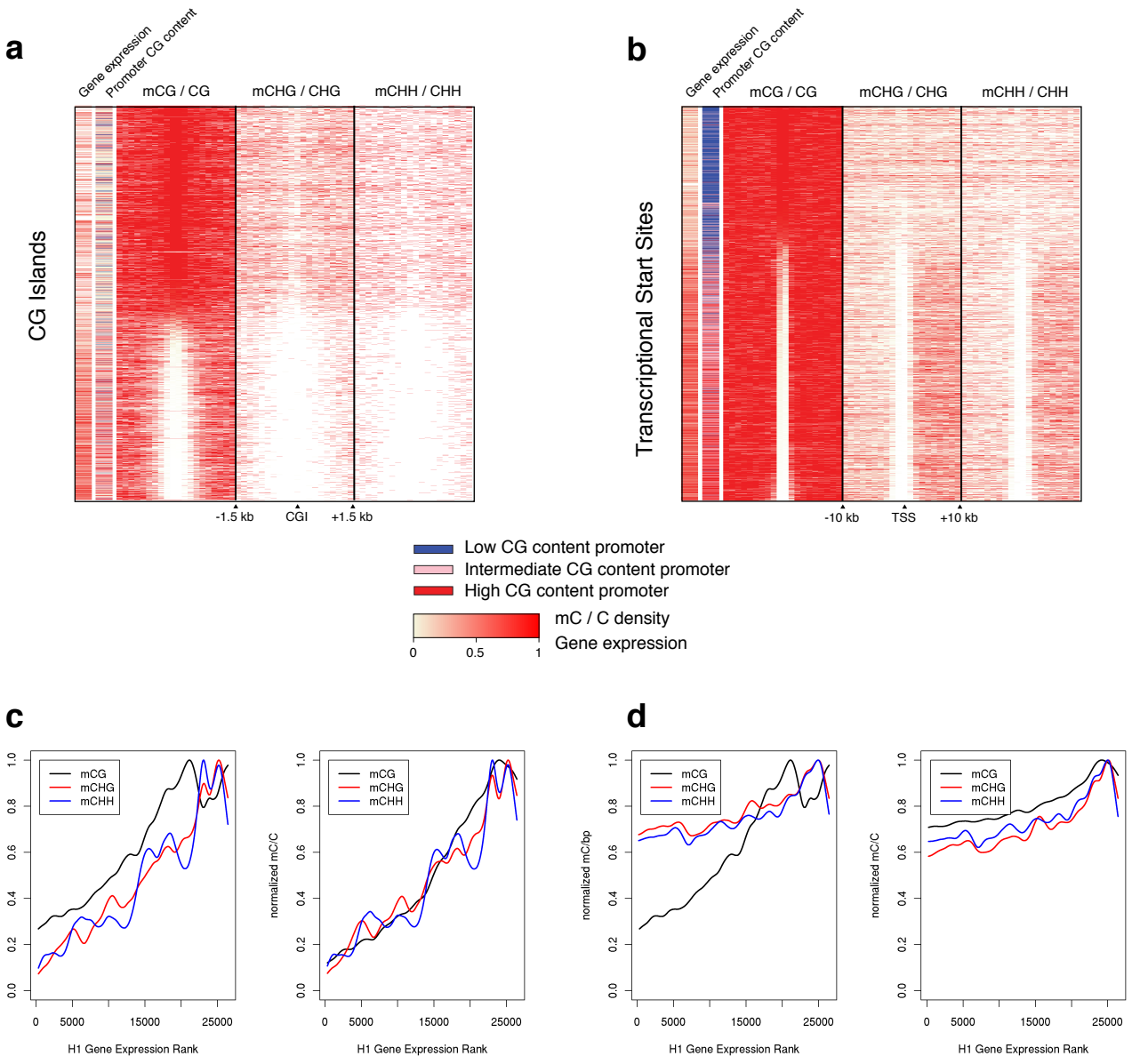
Supplementary Figure 3 | Differentually Expression of *DNMT* Genes in H1 and IMR90. **a,** log2RPKM and **b,** Maximum normalized RPKM measurements of transcript abundance for *DNMT1, DNMT3a, DNMT3b, DNMT3L* and *GAPDH* from RNA-seq. Abbreviations: RPKM, reads per kilobase of exon model per million mapped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



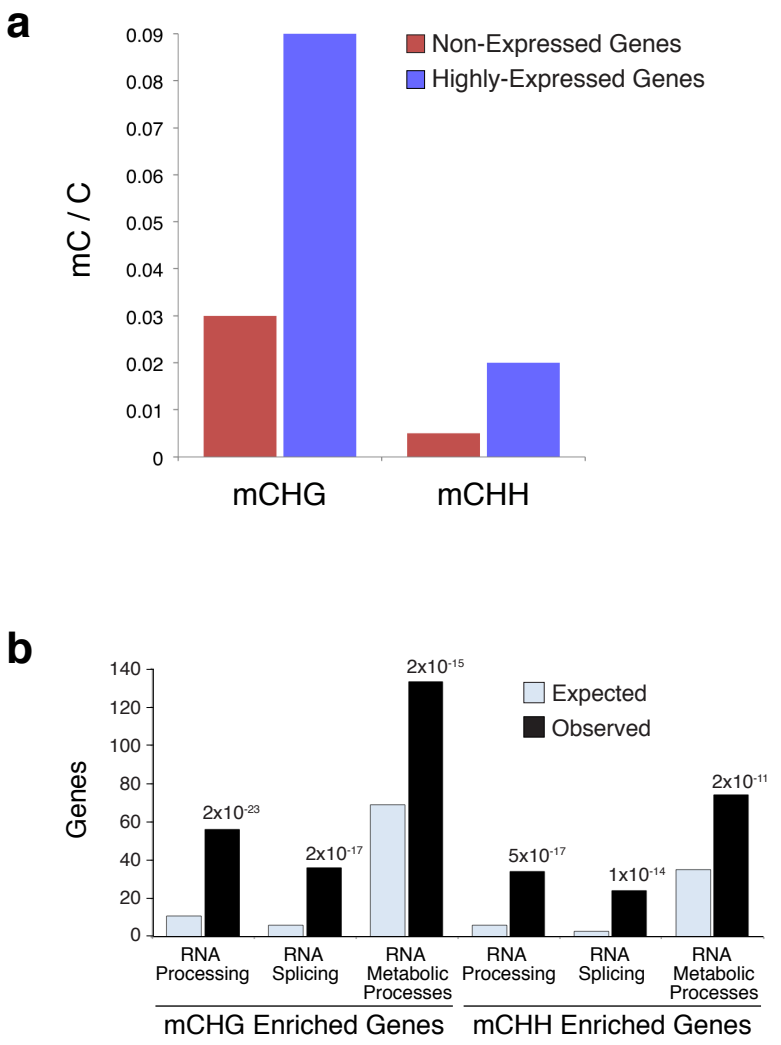
Supplementary Figure 4 | The Density of Methylcytosines Identified in All Chromosomes in H1 Cells. Blue dots indicate the density of all methylcytosines in 10 kb windows. Black rectangles indicate approximate centromere positions.



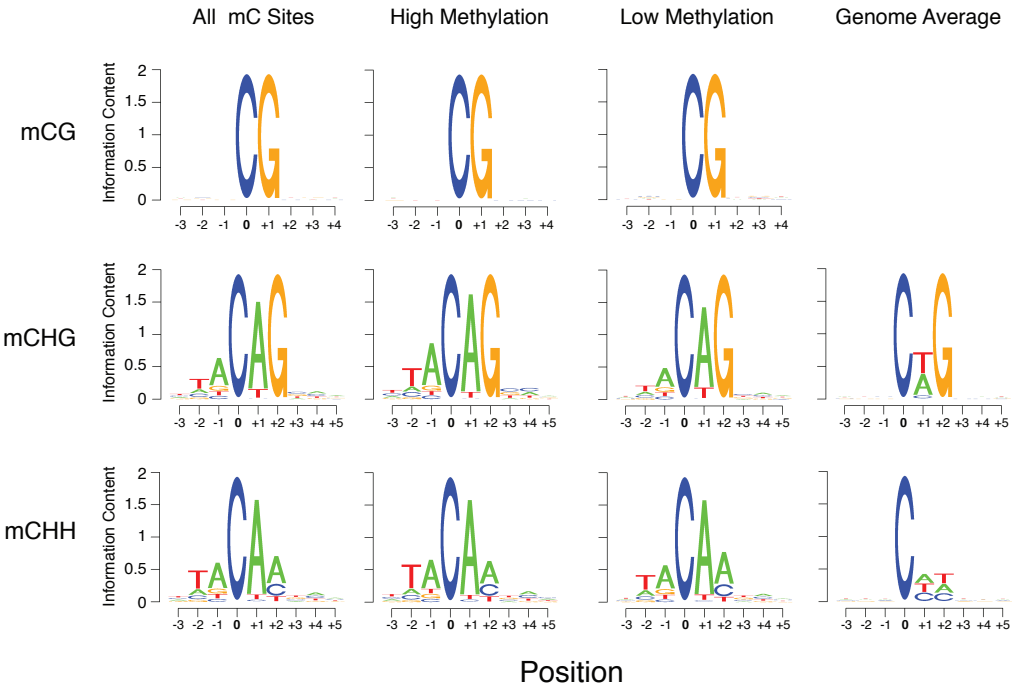
Supplementary Figure 5 l Mean mC/C Profiles Over Genomic Regions. a, gene body regions were divided in 20 bins from 5' to 3' end, and the mean mC/C level within each bin for each methylation type was determined (mCG/CG black, mCHG/CHG red, CHH/CHH blue). Mean over 3Kb regions centered at **b,** regulatory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



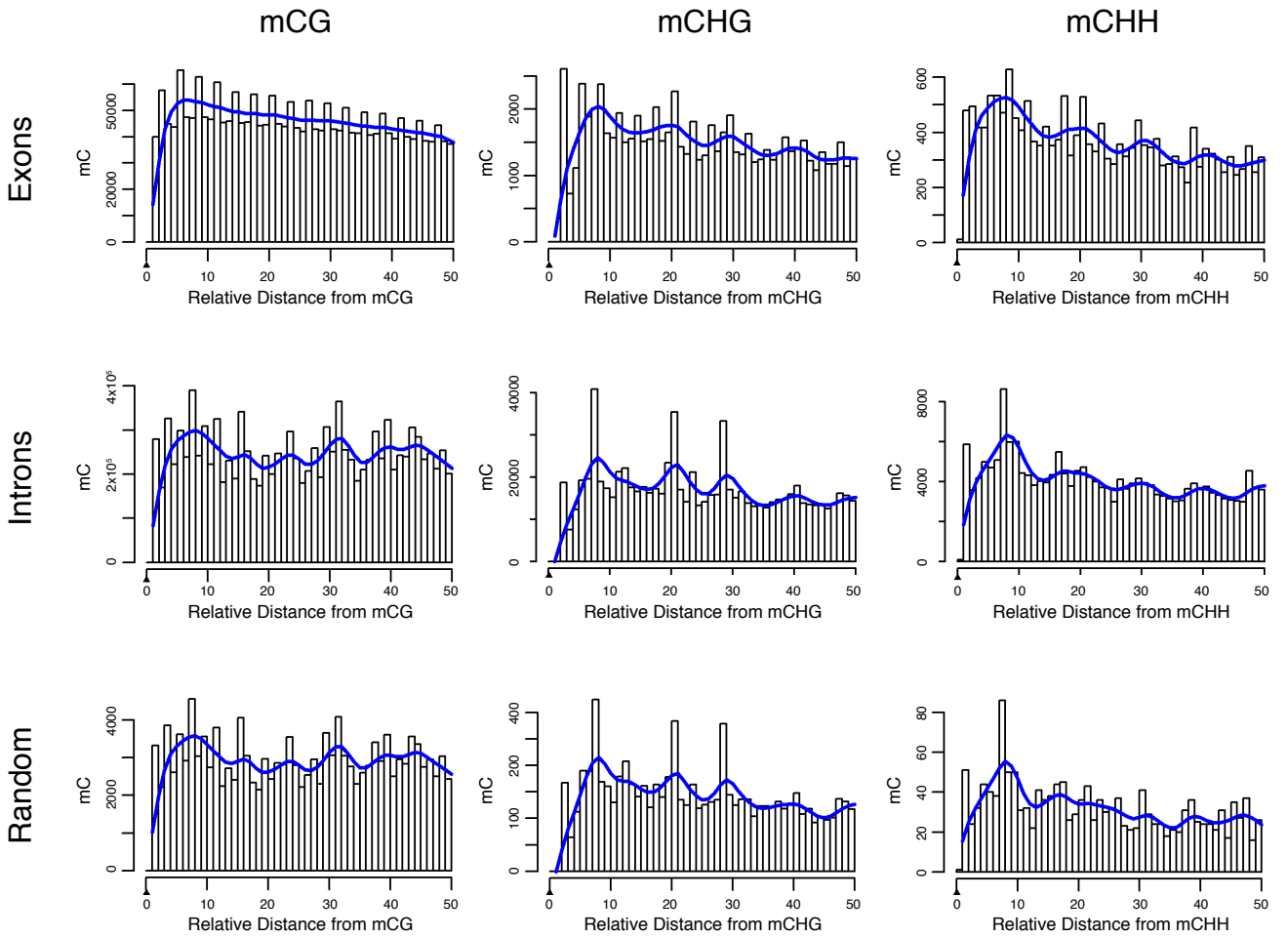
Supplementary Figure 6 | DNA Methylation at CG Islands, Transcriptional Start Sites and Promoters. Relative DNA methylation density at **a,** CG islands (1.5 kb upstream/downstream) and **b,** transcriptional start sites (10 kb upstream/downstream) is displayed with downstream gene expression and promoter 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 promoter (**c,** defined as the region 1.5 kb upstream of the TSS) and proximal TSS (**d,** defined as -150 bp to +150 bp across 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.



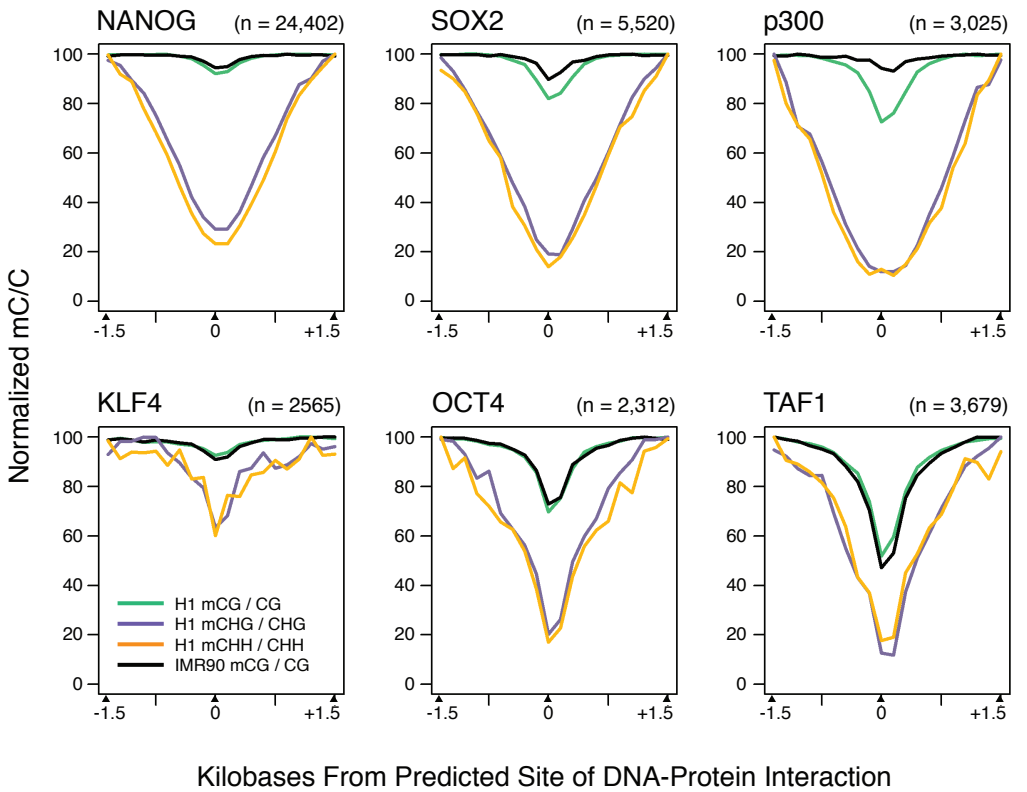
Supplementary Figure 7 I a, Enrichment of non-CG methylation in non-expressed and highly-expressed genes in H1. **b,** Over-representation of GO terms of genes within 20 kb of genomic regions displaying the highest enrichment of CHG and CHH methylation. The enrichment P-value is shown for each GO term.



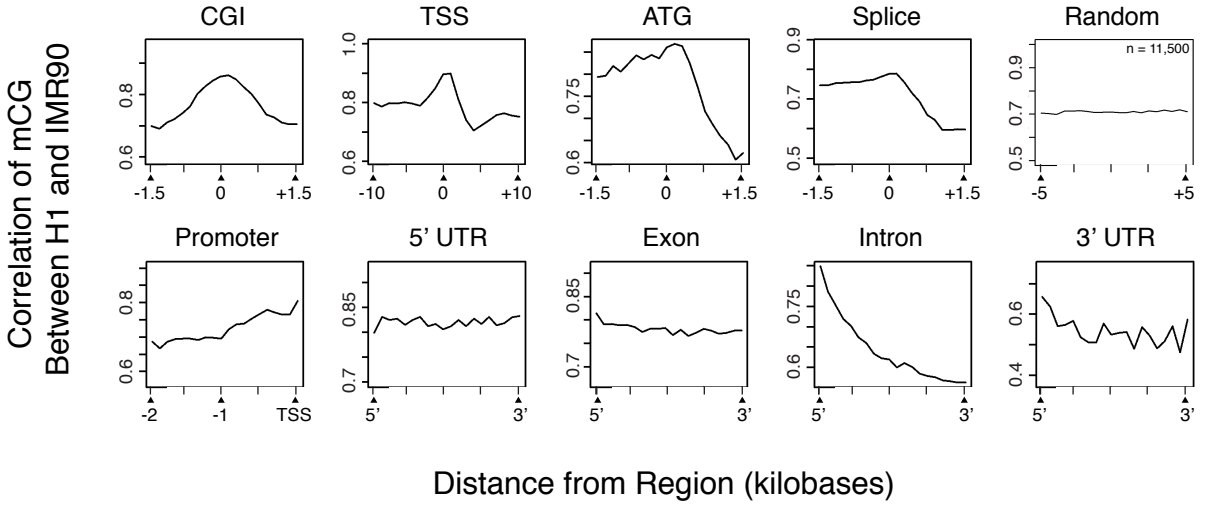
Supplementary Figure 8 | Logo Plots of the Sequences Proximal to Sites of DNA Methylation in Each Sequence Context in H1 Cells. Logo plots are presented for all methylcytosines, and methylcytosines that display a high methylation level (CG ≥75% methylated, non-CG ≥25% methylated), and low methylation 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.



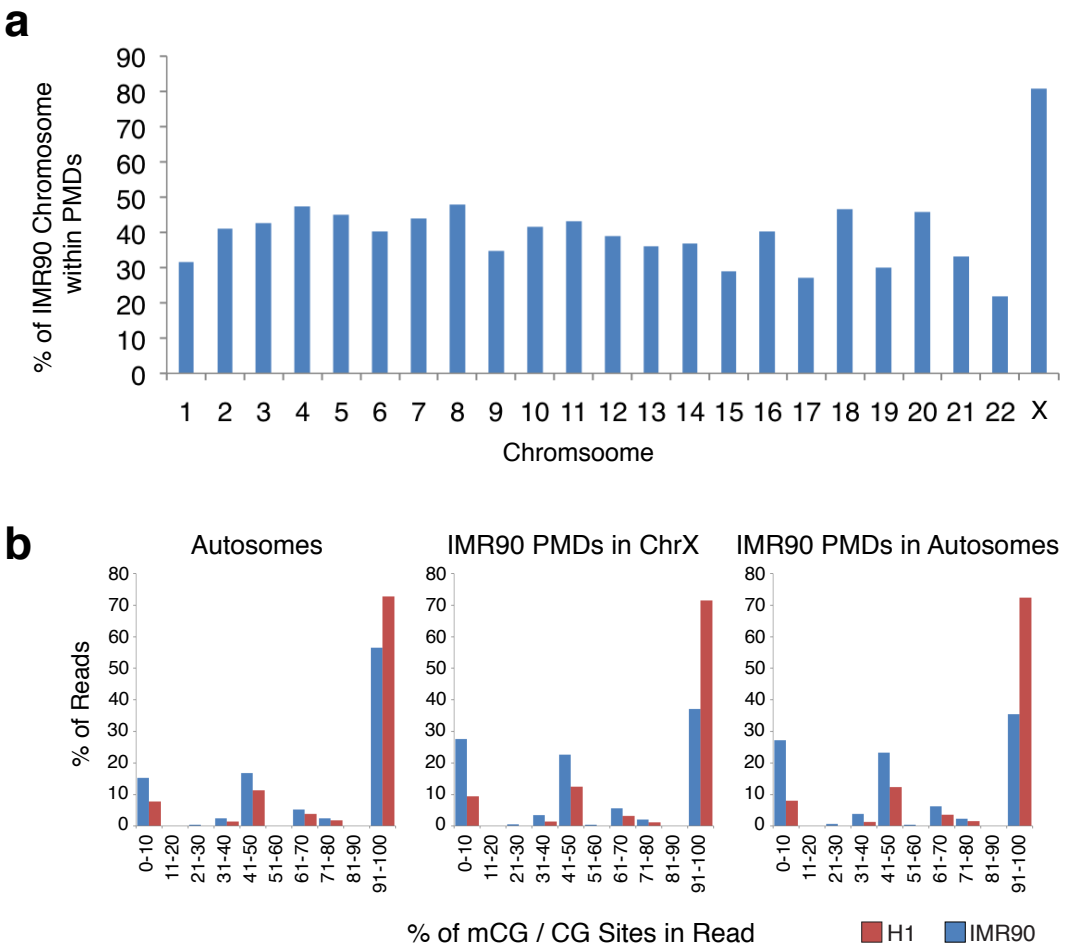
Supplementary Figure 9 | Spacing of Adjacent Methylcytosines in Different Contexts. Prevalence of mCHG/mCHH sites (y-axis) as a function of the number of bases between adjacent mCHG/mCHH sites (x-axis) based on all non redundant pair-wise distances up to 50 nt in exons, introns and random sequences.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



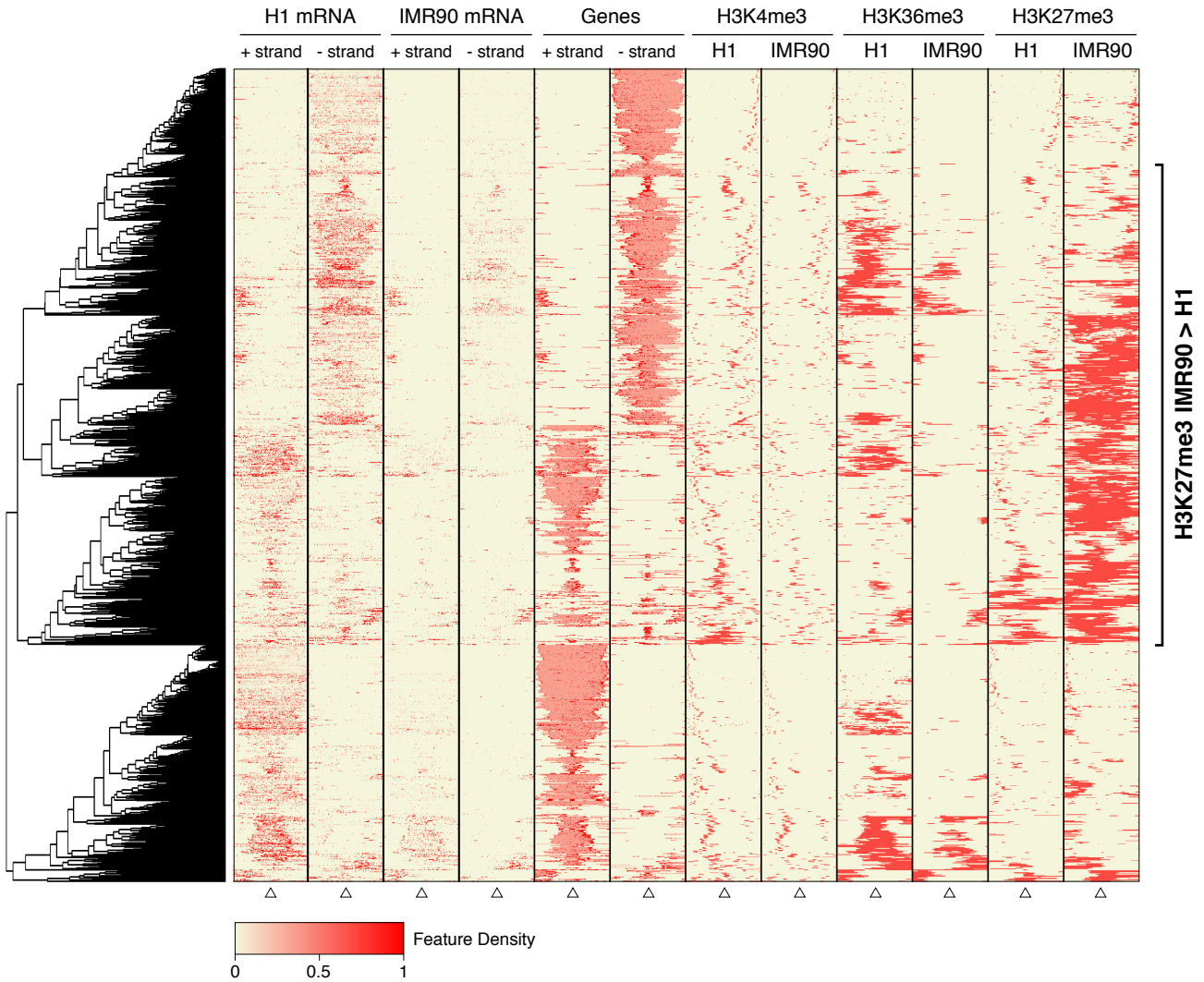
Supplementary Figure 10 | DNA Methylation at Sites of Protein-DNA Interaction. The average relative DNA methylation densities in each sequence context are shown from 1.5 kb upstream to 1.5 kb downstream of the predicted sites of DNA-protein interaction identified by ChIP-seq that were at least 1.5 kb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

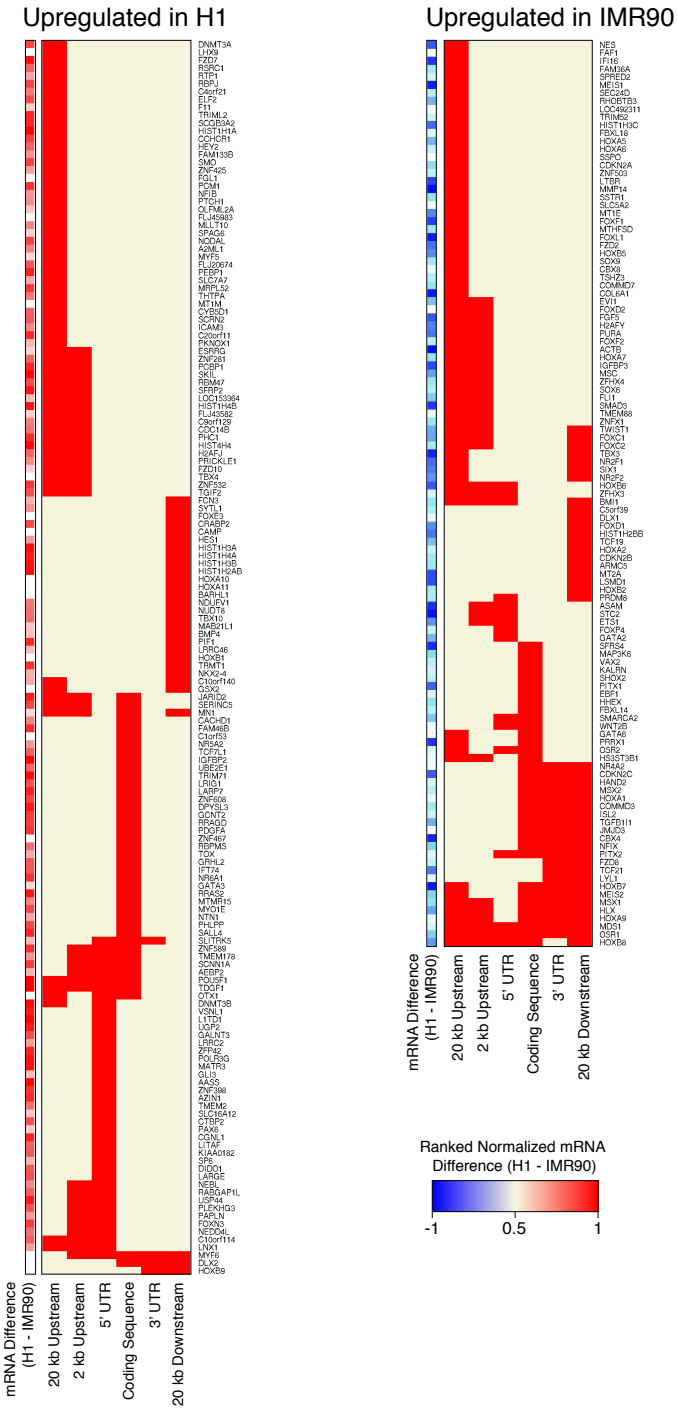


Supplementary Figure 12 | PMDs and the Distribution of Unmethylated, Partial, and Completely Methylated Reads. a, The percent of each IMR90 chromosome that is within PMDs. **b,** For MethylC-seq reads located within a set of genomic regions, the percentage of CG sites within each read that were methylated 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.

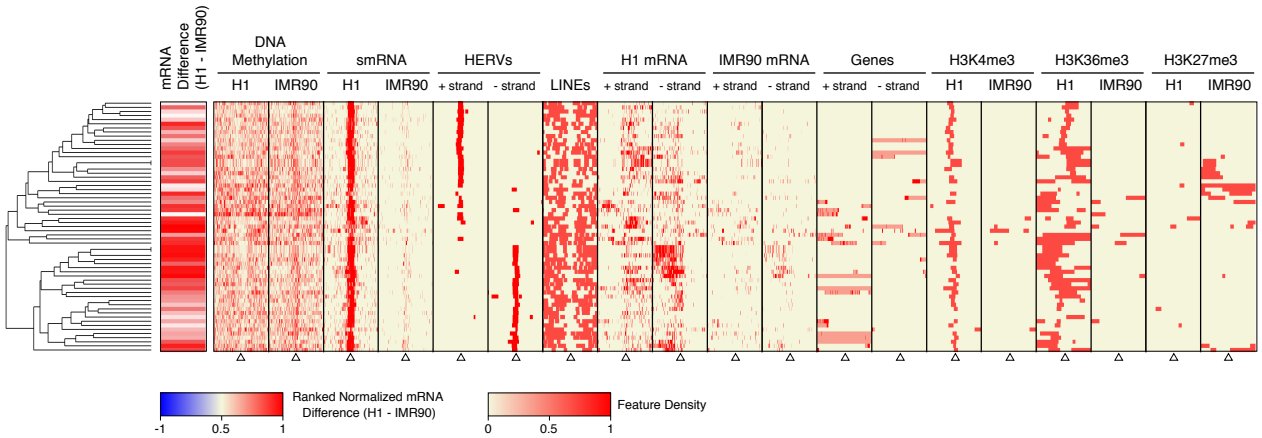


Supplementary Figure 13 | Transcriptional Activity and Epigenetic Modifications at Partially Methylated Domains. The density of strand-specific mRNA reads, and the presence of domains of H3K4me3, H3k36me3 and H3K27me3 in H1 and IMR90 was profiled 20 kb upstream to 20 kb downstream of each gene located 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 | Differentually 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 generartion. 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 methylation 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. The total number of sites was again normalized by the bin size. 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).