Supplemental Information

SUPPLEMENTAL FIGURE LEGENDS

Figure S1. TAB-Seq of Specific Loci and 5mC Conversion Rate Test in the Context of Genomic DNA. Relates to Figure 1.

(A) Purified mTet1 catalytic domain used for oxidation of genomic DNA.

(B) Sanger sequencing of M.SssI treated lambda DNA spiked into a genomic DNA background at 0.5% before (- mTet1) and after (+ mTet1) subjecting the DNA to TAB-Seq.

(C) Semiconductor sequencing of M.SssI-treated lambda DNA spiked into a genomic DNA background at 0.5% before (- mTet1) and after (+ mTet1) subjecting the DNA to TAB-Seq. The left y-axis shows the percentage of bases read as C and the right y-axis shows the depth of sequencing at each C position in the targeted amplicon, which is plotted on the x-axis. For reference, a dotted line is plotted at 98% on the left y-axis.

(D) Several loci in mouse cerebellum were tested by both traditional bisulfite sequencing and TAB-Seq. Genuine 5hmC is read as C in both methods (left) while genuine 5mC is read as C in traditional bisulfite sequencing but display as T in TAB-Seq (right).

Figure S2. Relates to Figure 2.

(A) Validation of TAB-Seq signals by semiconductor sequencing. The left y-axis shows the percentage of bases read as C and the right y-axis shows the depth of sequencing at each C position in the targeted amplicon, which is plotted on the x-axis. For reference, TAB-Seq

5hmC calls (orange diamond) and TAB-Seq read depth (clear diamond) have been plotted on the same axes, with all other C signals and read depths being derived from the semiconductor sequencing experiments.

(B) Lambda DNA was used as a spiked-in control as follows: regions 0-10kb were amplified with 5mdCTP, regions 20-30kb used dCTP, and regions 38-48kb used 5hmdCTP. Shown is the percent of bases read as cytosine after high-throughput sequencing, for reads after PCR duplicate removal (green) and after additional removal of reads having greater than 3 cytosine base calls in non-CG context (red). The average percentage of cytosine base calls is indicated below. While the average 5hmC abundance of every cytosine in hmCλ is 84.4%, later analysis shows this rate increases to 87.0% when considering the subset of bases in the similar 5hmC content as mammalian genomes.

(C) A theoretical calculation of the probability of calling a 5hmC for a given sequencing depth and a known abundance of 5hmC (percentages indicated), assuming a binomial distribution with parameter *p* as the 5mC non-conversion rate. Dotted lines indicate the average sequencing depth of H1 and the final p-value cutoff chosen at a false discovery rate of 5%.

(D) The 38-48kb region of lambda DNA (hmCλ), constructed by PCR amplification with 5hmCTP, was used to estimate the 5hmCG protection rate in mammalian cells. Shown is the protection rate of hmCλ as a function of distance to the closest cytosine to the CpG dinucleotide. Data for both strands are shown, and each strand was analyzed independent of the other strand. We observed that hmC protection is most efficient when the closest neighboring hmC residue is 4 bases away (hmCNNNhmC), which we denote as neighborless hmCs. However, even with the extreme case of hmCGhmCG (2 bases away) in hmCλ the analysis shows that each 5hmC still has a ~80% non-conversion rate. To support this conclusion, we prepared a synthetic model DNA (right panel, sequence: 5'-CCTCACCATCTCAACCAATATTATATTAXGXGTATATXGXGTATTTXGXGTTATAATATTGAGGGAGAAGTGGTGA-3', X = 5hmC). After TAB-Seq, cloning and sequencing, we observed at least 78% non-conversion rate for each 5hmC in the two hmCGhmCG sequences presented in the model DNA.

(E) For each 5hmC (top) and 5mC (bottom) in H1, the distance to the nearest 5hmC and 5mC, respectively, was calculated. The distribution of these minimal distances is shown. Data for both strands are shown, and each strand was analyzed independent of the other strand. In H1, >98.8% of 5hmC and 94.4% of 5mC are separated by at least 4 bases to the nearest 5hmC and 5mC, respectively, indicating that the vast majority of 5hmC bases we observe are more efficiently protected.

(F) The number of 5hmCs called for various p-value cutoffs of the binomial distribution, for actual data (black) and randomly sampled 5mCs (grey) (see Extended Experimental Procedures). The final p-value chosen was 3.5E-4, which corresponds to a false discovery rate of 5%.

(G) Sequence context of 5hmC sites in mouse ES cells, compared to the reference mouse genome.

(H) Heatmap of the abundance of 5hmC and 5mC for cytosines significantly enriched with 5-hmC in mouse ES cells. 5mC was estimated as the rate from traditional bisulfite sequencing (5hmC + 5mC) minus the measured 5hmC rate.

(I) The distribution of the abundance of 5hmC (red) and 5mC (green) at the 5hmC sites in mouse ES cells. m: median.

(J) In mouse ES cells, the overlap of 5hmCs (black) with 73,173 genomic regions identified as enriched by affinity mapping and 29,794 TET1 peaks identified by ChIP-Seq (Williams et al., 2011; Wu et al., 2011), in comparison to randomly chosen 5mCs (grey) (see Extended Experimental Procedures).

Figure S3. Relates to Figure 3.

(A) The distribution of pair-wise distances between all 5hmCs identified in H1 (red), compared to the same number of randomly selected 5mCs (black).

(B) The distribution of base-level phastCons conservation scores (Siepel et al., 2005) for several tiers of 5hmC abundance.

(C) Total methylation level measured by methylC-Seq (left) and the 5hmC abundance measured by TAB-Seq (right) for DNase I hypersensitive elements ranked by signal strength.

(D) The relative enrichment of H1 5hmCs (black) and random sites (grey) at promoter-distal ChIP-Seq peaks, normalized to the total coverage of the element type. Random consists of 10 random samplings of 5mCs (see Extended Experimental Procedures).

(E) For the subset of H1 cytosines having a ratio of 5hmC to 5mC between 0.9 and 1.10, shown is the relative enrichment of sites (black) and random sites (grey) at genomic elements, normalized to the total coverage of the element type. Random consists of 10 random samplings of 5mCs (see Extended Experimental Procedures).

(F) Overlap of mouse ES cell 5hmCs with genomic elements. Promoter-distal regulatory elements (>5kb from TSS) reflect those experimentally mapped in mouse ES cells from

ChIP-Seq and DNase-Seq experiments. Each 5hmC base is counted once: the overlap of a genomic element excludes all previously overlapped cytosines counterclockwise to the arrow. Green: promoter-proximal; red: promoter-distal regulatory elements; grey: genic regions; white: intergenic regions.

(G) The relative enrichment of mouse ES cell 5hmCs (black) and random 5mCs (grey) at genomic elements, normalized to the total coverage of the element type. Random consists of 5 random samplings of 5mCs (see Extended Experimental Procedures).

(H) The percentage of distal regulatory elements significantly enriched with 5hmCG in mouse ES cells.

(I) H1 promoters were divided into three equally sized groups based on the expression of corresponding genes. Shown is the relative enrichment of 5hmCs (black) and random sites (grey) at these promoters, normalized to the total coverage of each group. Random consists of 10 random samplings of 5mCs (see Extended Experimental Procedures).

(J) Shown is the distribution of total methylation (5mC + 5hmC) and 5hmC abundance at repetitive elements that do not overlap with regulatory elements (promoters, p300/CTCF binding sites, enhancers, DNase I hypersensitive sites). Elements with less than 50 C+T calls for either methylC-Seq or TAB-Seq were excluded.

(K) The percentage of repetitive elements significantly enriched with 5hmCG in H1.

(L) The absolute level of 5hmCG for several classes of repetitive elements significantly enriched with 5hmCG in H1 ($p = 0.01$, binomial).

Figure S4. Relates to Figure 4.

(A) Absolute levels of 5hmCG (red) and 5mCG+5hmCG (black) around distal p300 binding sites. Peaks were identified by MACS (Zhang et al., 2008), and the p300 binding site was estimated as the MACS summit location.

(B) Frequency of 5hmC around distal NANOG binding sites, relative to the NANOG motif (blue bar). The different lines represent the different strands, oriented with respect to the NANOG motif (consensus: GGCCATTAAC). Opp, opposite.

(C) Absolute levels of 5hmCG (red) and 5mCG+5hmCG (black) around distal NANOG binding sites containing an NANOG motif (blue bar, center; consensus: GGCCATTAAC). 5mC (green) was estimated as the rate from traditional bisulfite sequencing rate (5hmC+5mC) minus the measured 5hmC rate. The top half indicates enrichment on the strand containing the motif, with the bottom half indicating the opposite strand.

Figure S5. Relates to Figure 5.

(A) The percentage of promoters and gene bodies having significant strand bias of 5hmCG, relative to the direction of transcription.

(B) There are 16 pairs of neighborless 5hmCGs in hmCλ, and shown in red is the asymmetry score (median absolute difference in 5hmCG abundance between pairs). The background distribution was computed as the asymmetry score of 100,000 randomly sampled sets of 16 neightborless CGs from each strand. The data indicates no asymmetry of 5hmCG in the control lambda DNA. Thus, our observations of asymmetry in H1 are not a result of the assay itself being biased.

(C) HPLC chromatogram (at 260 nm) of the nucleosides derived from a fully-hydroxymethylated double-stranded DNA before and after βGT catalyzed glucosylation. The peak of 5hmC decreased dramatically after glucosylation which indicates that over 90% of 5hmC is protected. The 5gmC elutes within the peak of dG in the chromatograph; however, formation of 5gmC was independently confirmed by mass spectrometry analysis of the product DNA as shown in Figure 1C.

Figure S6. Relates to Figure 6.

(A) The percentage of guanine bases found in the 75-bp region [-25, +50] around 5hmCG sites, compared to randomly selected 5-methylcytosines (see Extended Experimental Procedures).

(B) In hmCλ, the abundance of 5hmC as a function of guanine content was plotted for all neighborless CpGs. The dotted line indicates the median 5hmC abundance. The data indicates that the guanine content around these bases does not significantly correlate with 5hmC abundance in hmCλ ($R^2 = 0.018$, $p = 0.035$), indicating that our observations in H1 cells is not a result of TAB-Seq being biased.

(C) In H1, the sequence context ±10bp around 5hmCG sites that are on the Watson strand, for sites in various genomic elements. Similar results are observed on the Crick strand.

(D) In mouse ES cells, the sequence context ±10bp around 5hmCG sites that are on the Watson strand, for sites in various genomic elements. Similar results are observed on the Crick strand.

(E) In mouse ES cells, the sequence context ±150bp around all 5hmCG sites. Shown sequences are on the same strand as the 5hmC base. Positive coordinates indicate the 3’ direction.

(F) In mouse ES cells, the sequence context ±150bp around the subset of 5hmCG sites at enhancers. Shown sequences are on the same strand as the 5hmC base. Positive coordinates indicate the 3’ direction.

Figure S7. Relates to Figure 7.

(A) Heatmap of total methylation ±250bp from TSSs, as a function of CpG density.

(B) Heatmap of percent 5hmCG ±250bp from distal p300 binding sites, as a function of CpG density.

(C) Heatmap as in (B), but for the subset of binding sites with DNase I hypersensitivity.

(D) Heatmap of total methylation ±250bp from DNase I hypersensitive sites lacking H3K4me1 and H3K27ac, as a function of CpG density.

(E) Heatmap of total methylation ±250bp from DNase I hypersensitive sites having H3K4me1 but not H3K27ac, as a function of CpG density.

(F) Heatmap of total methylation ±250bp from DNase I hypersensitive sites having both H3K4me1 and H3K27ac, as a function of CpG density.

(G) The distribution of CpG content (top) and GC content (bottom) for 5hmC-enriched (red) and 5hmC-unenriched (green) bivalent promoters (left) and H3K4me3-only promoters (right).

(H) The density of 5hmC at promoters classified as having low (LCP), intermediate (ICP), and high (HCP) CpG content, normalized by the number of CpG dinucleotides in these promoters.

EXTENDED EXPERIMENTAL PROCEDURES

Cell Culture

E14 (E14Tg2A) ES cell lines were cultured in feeder-free gelatin-coated plates in Dulbecco's Modified Eagle Medium (DMEM) (Invitrogen Cat. No. 11995) supplemented with 15 % FBS (GIBCO), 2 mM L-glutamine (GIBCO), 0.1 mM 2-mercaptoethanol (Sigma), 1x non-essential amino acids (GIBCO), 1,000 units/ml LIF (Millipore Cat. No. ESG1107), 1x pen/strep (GIBCO), 3 µM CHIR99021 (Stemgent) and 1 µM PD0325901 (Stemgent). The culture was passaged every 2 days.

Expression and Purification of Recombinant mTET1

The catalytic domain (amino acids 1367-2039) of Mouse *TET1* (GU079948) gene was cloned into BssH1 and NotI sites of N-terminal Flag-tagged pFastBac Dual vector (Invitrogen, cat:10712024) and then expressed in Bac-to-Bac baculovirus insect cell expression system. The recombinant protein was first purified with the anti-Flag M2 antibody agarose affinity gel (Sigma-Aldrich) as reported (Ito et al., 2010) and then loaded onto a Superdex 200 (GE Healthcare) gel-filtration column equilibrated with 20 mM HEPES (pH 8.0), 150 mM NaCl and 1 mM DTT.

Expression and Purification of β-Glucosyltransferase Protein (βGT)

The βGT protein was expressed and purified following the previous protocol (Song et al., 2011).

Oligonucleotide Synthesis

9-mer oligonucleotides containing modified cytosine (5mC or 5hmC) were prepared by using Applied Biosystems 392 DNA synthesizer with 5-Me-dC-CE or 5-hydroxymethyl-dC-CE phosphoramidite (Glen Research). All synthetic oligonucleotides were then purified by denaturing PAGE. The complementary 11-mer oligonucleotide without modified bases was purchased from Operon. 11-mer and 13-mer 5hmC containing oligonucleotides for HPLC analysis were prepared in the same way.

Preparation of 76-mer Double-Stranded DNA with 5mC or 5hmC Modification

The 76-mer dsDNA with one 5mC or 5hmC on one strand (as shown in Figure 1B) were generated using PCR reaction with 5-methyl-2'-deoxycytidine 5'-triphosphate (5mdCTP) (Fermentas) or 5-hydroxymethyl-2'-deoxycytidine 5'-triphosphate (5hmdCTP) (Bioline) in place of dCTP and RED Taq polymerase (Sigma-Aldrich). To remove the unmodified template from product, two rounds of PCR were applied with 22 cycles in the first round and 30 cycles in the second round as described (Jin et al., 2010). The PCR products were then purified using PCR purification kits (Qiagen) (Forward primer: 5'-CCTCACCATCTCAACCAATA-3'; Reverse primer: 5'-TCACCACTTCTCCCTCAAT-3').

TAB-Seq of 76-mer dsDNA

The glucosylation reactions were performed in a 20 µl solution containing 50 mM HEPES buffer (pH 8.0), 25 mM $MgCl_2$, 100 ng/µl model DNA, 200 µM UDP-Glc, and 1 µM βGT. The reactions were incubated at 37 ^{o}C for 1 h. After the reaction, the DNA was purified by QIAquick Nucleotide Removal Kit (Qiagen). The oxidation reactions were performed in a 20 µl solution containing 50 mM HEPES buffer (pH 8.0), 100 µM ammonium iron (II) sulfate, 1 mM α-ketoglutarate, 2 mM ascorbic acid, 2.5 mM DTT, 100 mM NaCl, 1.2 mM ATP, 15 ng/µl glucosylated DNA and 3 µM recombinant mTet1. The reactions were incubated at 37 ^{o}C for 1.5 h. After proteinase K treatment, the DNA was purifited with QIAquick Nucleotide Removal Kit (Qiagen) and then applied to EpiTect Bisulfite Kit (Qiagen) following the supplier's instruction. After PCR amplification with Hotstar Taq polymerase (Qiagen) (Forward primer: 5'-CCCTTT TATTATTTTAATTAATATTATATT-3'; Reverse primer: 5'-CTCCGACATTATCACTACCATCAACCACCCATCCTACCTGGACTACATTCTTATTCAG TATTCACCACTTCTCCCTCAAT-3'), the PCR product was purified using PCR purification kits (Qiagen) and sent for sequencing.

HPLC Analysis of βGT Catalyzed Glucosylation

The glucosylation reactions were performed in a 120 µl solution containing 50 mM HEPES buffer (pH 8.0), 25 mM $MgCl_2$, 10 µM fully-hydroxymethylated dsDNA, 200 µM UDP-Glc, and 1 µM βGT. The reactions were incubated at 37 ^{o}C for 1 h. After the reaction, both the substrate DNA (1.2 nmol) and glucosylated DNA were digested by two unit Nuclease P1 (Sigma) in 0.01 M NH_4Ac (pH 5.3) at 45 ^{o}C for 2 h and then two unit of Alkaline Phosphatase (Sigma) in 0.1 M fresh NH_4HCO_3 at 37 ^{o}C overnight. The digested DNA was analyzed by HPLC with a C18 reverse-phase column equilibrated with buffer A (50mM ammonium acetate) and buffer B (50mM ammonium acetate, 0.1% TFA, 60% CH_3CN).

Dot Blot Assay

βGT-treated and βGT/mTet1-treated mouse ES genomic DNA was generated as described above. 2 μg of DNA was denatured in 0.4 M NaOH, 10 mM EDTA at 95 ^{o}C for 10 min, and then neutralized by adding an equal volume of cold 2 M ammonium acetate (pH 7.0). 150 ng denatured DNA samples were spotted on nitrocellulose membrane (GE Healthcare). The membrane was then blocked with 5% non-fat milk and incubated with 5mC antibody (1:500) (Epigentek), 5hmC antibody (1:10000) (Active Motif), 5fC antibody (1:5000) (Active Motif) or 5caC antibody (1:2000) (Active Motif). Binding of an HRP-conjugated secondary antibody (1:1000) was visualized by enhanced chemiluminescence.

Semiconductor Sequencing

E14Tg2a genomic DNA was spiked with 0.5% M.SssI treated DNA and subjected to TAB-Seq treatment as described above or used directly in sodium bisulfite conversion. After MethylCode bisulfite conversion of 50 ng, 1 μL of converted DNA was PCR amplified as follows in a 50 μL final reaction volume: 2.5U PfuTurbo Cx Hotstart DNA polymerase, 5 μL 10X PfuTurbo Cx reaction buffer, 1 μL 10 mM dNTPs, 1 μL 10 μM FW primer (5'-CCATCTCATCCCTGCGTGT CTCCGACTCAGAATTTGGTGGTGAGTAATGGTTTTA), 1 μL 10 μM RV primer (5'-CCTCTCTATGGGCAGTCGGTGATAACCTACCCCAACACCTATTTAAAT). Cycling parameters: 95ºC 2 min, 35 cycles of 95ºC 30 sec, 55ºC 30 sec, 72ºC 1 min, followed by 72ºC 5 min. Fusion PCR primers were designed to incorporate sequences at their 5’ ends that are compatible with Ion Torrent template generation. PCR products were purified on Qiagen MinElute columns and quantified on an Agilent 2100 Bioanalyzer High Sensitivity DNA Chip. Sequencing template was generated using and Ion Torrent OneTouch System and Ion OneTouch System Template Kits (Life Technologies). Sequencing reactions were performed for 100 cycles on an Ion PGM semiconductor instrument using an Ion 314 chip and Ion Sequencing Kit (Life Technologies). Sequences were aligned to an index built from only the targeted amplicon using Bowtie in an analogous way to that used for genome-wide sequencing, except without preprocessing reads and requiring full-length perfect matches. Validation of TAB-Seq calls was done in the same manner using an independent mTet1 oxidation of H1 genomic DNA and testing two separate loci that were not previously identified as enriched with 5hmC (Szulwach et al., 2011). These loci included a total of 57 cytosines (11 CpG dincleotides) and 9 same strand 5hmC calls. The hg18 genomic coordinates for the amplicons were chr4:182,423,188-182,423,312 and chr11:45,723,245-45,723,393. The corresponding fusion primer sequences, respectively, were (FW-5'-CCATCTCATCCCTGCGTGTCTCCGACTCAGTAGAAGTAAA GGAAGTAAAGGAAGTATG; RV-5'-CCTCTCTATGGGCAGTCGGTGATAAACCTAAAT AATAACAAACACACC) and (FW-5'-CCATCTCATCCCTGCGTGTCTCCGACTCAG GAAGTTGTATAAAATTTTTGGATGTG; RV-5'-CCTCTCTATGGGCAGTCGGTGAT CCTCTCCTATCTCCCTTAACTACTC)

TAB-Seq of Specific Loci in Mouse Cerebellum

500 ng-1 μg untreated or βGT/mTet1-treated (the same procedure as mouse ES/H1 cell) mouse cerebellum sample was applied to EpiTect Bisulfite Kit (Qiagen) following the supplier's

instruction. After PCR amplification with RED Taq polymerase (Sigma-Aldrich) or Hotstar Taq polymerase (Qiagen) (for 5hmC site, Forward primer: 5’-TTTGATTTTTGTGTTTAGTAGTT TTGTG-3’; Reverse primer: 5’-CCTCCTCAATTTTAAAATCTATTCC-3’; for 5mC site, Forward primer: 5’-TTTAGGAATTGATAGGTAGTTGTAG-3’; Reverse primer: 5’-AAACACAAACAATCTTACAAAAAAAAA-3’), the PCR product was purified using PCR purification kits (Qiagen) and sent for sequencing.

Generation of 5mC Spike in Conversion Controls for Mouse ES Cells

For the E14Tg2a mouse ES samples, unmethylated Lambda cI857 DNA (Promega) was treated with M.SssI to fully-methylate all CpG cytosines. CpG methylation was confirmed by spiking M.SssI treated Lambda DNA into genomic DNA at 0.5% followed by standard bisulfite conversion, PCR amplification, TOPO cloning, and Sanger sequencing. After MethylCode bisulfite conversion of 50 ng, 1 μL of converted DNA was PCR amplified as follows in a 50 μL final reaction volume: 2.5U PfuTurbo Cx Hotstart DNA polymerase, 5 μL 10X PfuTurbo Cx reaction buffer, 1 μL 10 mM dNTPs, 1 μL 10 μM FW primer (5'-TTTGGGTTATGTAAGTTGATTTTATG), 1 μL 10 μM RV primer (5'-CACCCTACTTACTAAAATTTACACC). Cycling parameters: 95 ºC 2min, 35 cycles of 95 ºC 30 sec, 57 ºC 30 sec, 72 ºC 1 min, followed by 72ºC 5 min. 1 μL PCR product was TOPO cloned using the Zero-blunt TOPO cloning kit (Invitrogen) and individual clones were subjected to Sanger sequencing using the an SP6 priming site. 5mC conversion after βGT glucosylation and Tet oxidation was assessed in the same way.

Generation of Spiked-in Conversion Controls for H1 Cells

Several spiked-in controls were generated and tested. Spike-in control A consisted of a 1:1 mixture of unmethylated lambda DNA (Promega Cat. No. D1521) with M.SssI-converted pUC19 DNA (NEB, Cat. No. M0226S). To generate the spiked-in control B, unmethylated lambda DNA (Promega Cat. No. D1521) was PCR amplified and purified by gel electrophoresis in non-overlapping 2-kb amplicons, with a cocktail of dATP/dGTP/dTTP and either: d5mCTP (Zymo Research, Cat. No. D1035) at genomic positions 0-10kb, dCTP at genomic positions 20-30kb, and d5hmCTP (Zymo Research, Cat. No. D1045) at genomic positions 38-48kb. Amplicons with d5mCTP/d5hmCTP and dCTP were amplified by ZymoTaq DNA polymerase (Zymo Research, Cat. No. E2001) and Phusion HF DNA polymerase (NEB, Cat. No. M0530S), respectively, as per manufacturers' instructions. Spiked-in DNA was added to H1 genomic DNA to a final concentration of 0.5% (control A for replicate 1, control B for replicate 2), and sonicated to a range of 300-500bp with a Biorupter 300 (high power, 15s on, 15s off, 20 cycles).

Assessing 5hmC Protection Rate in H1

To estimate the 5hmC protection rate in H1, we have performed further analysis of our spiked-in lambda control. The 38-48kb region of lambda, which we designate hmCλ, was constructed by PCR amplification with 5hmCTP. Thus we assume that every cytosine sequenced in hmCλ exists as 100% 5hmC. As the structure of the glucose moiety of 5gmC suggests steric hindrance with neighboring 5gmC residues, we observe that 5hmC protection is most efficient when the closest neighboring 5hmC residue is at least 4 bases away (hmCNNNhmC), which we denote as neighborless 5hmCs (Figure S2D, left). In H1, >98.8% of 5hmC and 94.4% of 5mC are separated by at least 4 bases to the nearest 5hmC and 5mC, respectively (Figure S2E), indicating that the vast majority of 5hmC bases we observe are more efficiently protected. In hmCλ, these neighborless 5hmCs in CG context are protected at a median level of 87.0% (Figure S6B). In addition, even with an hmCGhmCG sequence (99.89% 5hmCs exist in CG context in H1) our analyses of both hmCλ and a model DNA indicate over 78% protection of each hmC (Figure S2D)

Library Generation

500 ng-1 μg treated genomic DNA was end-repaired, adenylated, and ligated to methylated (5mC) adapters (Illumina TruSeq Genomic DNA adapters) according to standard Illumina protocols for genomic DNA library construction, maintaining the proper molar ratios of adapter to insert. Adapter ligated fragments with 200-600 bp inserts were gel purified by 2% agarose gel electrophoresis and sodium-bisulfite treated using the MethylCode kit (Invitrogen). Bisulfite treated adapter-ligated DNA was amplified by PCR with PfuTurbo Cx Hotstart DNA polymerase. The number of PCR cycles used was determined by quantification of bisulfite treated adapter-ligated DNA by qPCR (KAPABiosystems library quant kit for Illumina libraries) such that the final library concentration obtained was approximately 20 nM. Final sequencing libraries were purified with AMPure XP beads or 2% agarose gel electrophoresis and quantified by qPCR (KAPABiosystems library quant kit for Illumina libraries). Up to 3 separate PCR reactions were performed per sample.

TAB-Seq Library Sequencing

TAB-Seq libraries were sequenced using the Illumina HiSeq2000 platform. Cluster generation was performed with Illumina TruSeq-PE cluster kit v3-cBot-HS. 2X 101-bp sequencing was completed with Illumina TruSeq SBS kit v3-HS. A dedicated PhiX control lane, as well as 1% PhiX spike in all other lanes, was used for automated matrix and phasing calculations. Image analysis and base calling were performed with the standard Illumina pipeline.

Data Processing

Reads were processed as previously reported (Hon et al., 2012; Lister et al., 2009). Briefly, raw reads were trimmed for low quality bases and adapter sequences. Then, cytosine bases were computationally replaced with thymines, mapped with the Bowtie program (Langmead et al., 2009) against computationally converted copies of hg18 or mm9, and mapped reads were resorted to their pre-computationally-converted bases. PCR duplicates were removed for each PCR amplification reaction using the Picard program (http://picard.sourceforge.net). To eliminate reads not bisulfite converted, reads having more than 3 base calls in non-CG context were removed, as previously (Lister et al., 2009). All libraries were then merged and indexed by the SAMtools suite (Li et al., 2009).

Calling 5-Hydroxymethylcytosines

Since traditional bisulfite sequencing identifies both 5mC and 5hmC, we restricted our search space for 5-hydroxymethylcytosines to the subset of cytosines previously called as methylated by methylC-Seq/BS-Seq. For each such base, we counted the number of “C” bases from TAB-Seq reads as hydroxymethylated (denoted N_C) and the number of “T” bases as not hydroxymethylated (denoted N_T). Then, we used the binomial distribution having parameters N as the sequencing depth (N_C+N_T) and p as the 5mC non-conversion rate (2.22% for H1), to assess the probability of observing NC or greater cytosines by chance.

Assessing False Discovery Rate of 5hmC in H1

To estimate the false discovery rate of calling hydroxymethylated cytosines, we repeated the steps above on randomly sampled methylcytosines. First, for each (chromosome *chr*, strand *str*, context *con*) combination, we counted the number of cytosine base calls having Phred score ≥ 20 spanned by every read (denoted $C_{chr,str,con}$). Then, using calls of methylcytosines from methylC-Seq, we randomly sampled $C_{chr,str,con}$ methylcytosines spanned by TAB-Seq reads on chromosome *chr*, strand *str*, and context *con*, with probability proportional to sequencing depth at each cytosine. This sampling method guarantees an equal chromosomal, strand, and context distribution as the original data, and normalizes for sequencing depth. Thus, the false discovery rate for a given p-value cutoff of the binomial distribution is the average number of hydroxymethylcytosines called in 10 random samplings divided by the number observed in the original data.

Quantifying Enrichment of 5hmC Bases at Genomic Elements

To calculate the enrichment of hydroxymethylcytosine calls at a set of genomic loci, we counted the number of overlapping 5hmCs and divided by the average of 10 random samplings of hydroxymethylcytosine calls, as performed above. Finally, we normalized this enrichment value by the genomic span of the corresponding set of genomic elements.

Generation of E14Tg2a 5hmC Enrichment Profiles

5hmC enrichment from E14Tg2a genomic DNA was done as previously described (Song et al., 2011) utilizing a 5hmC specific chemical labeling and capture approach. Sequence reads were generated and analyzed in the same manner as previously reported for H1 hES cells (Szulwach et al., 2011). Enriched regions were identified by MACS (Zhang et al., 2008) analysis with a p-value threshold of 1e-8 against a matched unenriched input genomic DNA library prepared and sequenced in parallel with the 5hmC enriched DNA.

ChIP-Seq Correlation at Distal Elements

To correlate %5hmCG with histone modifications measured by ChIP-Seq (Hawkins et al., 2010), we calculated the enrichment of histone modifications at each DNase I hypersensitive site as log2 (ChIP RPKM / input RPKM), using a pseudocount as previously (Hon et al., 2012).

Assessing Potential Biases in TAB-Seq

In H1, >98.8% of hmCs are separated by at least 4 bases to the nearest 5hmC (Figure S2E). We analyzed these neighborless 5hmC bases (hmCNNNhmC) within hmCλ. There are 16 pairs of neighborless 5hmCG's in hmCλ, which we assume to be symmetrically modified by 5hmC. As a measure of asymmetry, we computed the median absolute difference in 5hmCG abundance between pairs to be 4.96%. To get a background distribution for this asymmetry score, we computed the same score for 100000 randomly sampled sets of 16 neighborless CGs from each strand. We find that the observed asymmetry score is not significantly different from that expected by chance (Figure S5B). In the remaining 1.2% hmCs in H1, even with an hmCGhmCG sequence (99.89% 5hmCs exist in CG context in H1) our analyses of both hmCλ and model DNA indicate over 78% protection of each hmC (Figure S2D). Therefore, we conclude that there is no asymmetry of 5hmCG in lambda DNA, and that our observations of asymmetry in H1 are not a result of the assay itself being biased. We find that that guanine content is a predictor of 5hmC in H1 cells. To assess if this observation is a result of TAB-Seq being biased by sequence content, we focused on neighborless 5hmC bases within hmCλ. Since hmCλ is assumed to be fully modified, we expect no correlation between sequence content and 5hmC. We find that guanine content around these bases does not significantly correlate with 5hmC abundance in hmCλ (R^2 = 0.018, p = 0.035) (Figure S6B), suggesting that our observations of the opposite to be true in H1 cells is not a result of TAB-Seq being biased.

External Data

CTCF ChIP-Seq peaks and DNase I hypersensitive sites for H1 ES cells were downloaded from the UCSC Genome Browser (Kent et al., 2002) and produced by the ENCODE Project Consortium (Myers et al., 2011). Distal regulatory elements are defined as those that are at least 5-kb from a transcription start site. Mouse Tet1 binding sites were derived from (Williams et al., 2011; Wu et al., 2011). Raw Tet1 ChIP-Seq sequence reads from both studies (SRA accessions:

SRR070927, SRR070925, SRR096330, SRR096331) were aligned and monoclonal reads from each were combined into a single set. Peaks were identified against the combined set of IgG control monoclonal reads (SRA accessions: SRR070931, SRR096334, SRR096335), as well as monoclonal reads from the E14Tg2a input genomic DNA sample sequenced as part of this study, using a standard MACS analysis (Zhang et al., 2008).

SUPPLEMENTAL REFERENCES

Hawkins, R.D., Hon, G.C., Lee, L.K., Ngo, Q., Lister, R., Pelizzola, M., Edsall, L.E., Kuan, S., Luu, Y., Klugman, S., *et al.* (2010). Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell stem cell *6*, 479-491.

Hon, G.C., Hawkins, R.D., Caballero, O.L., Lo, C., Lister, R., Pelizzola, M., Valsesia, A., Ye, Z., Kuan, S., Edsall, L.E., *et al.* (2012). Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. Genome research *22*, 246-258.

Ito, S., D'Alessio, A.C., Taranova, O.V., Hong, K., Sowers, L.C., and Zhang, Y. (2010). Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature *466*, 1129-1133.

Jin, S.G., Kadam, S., and Pfeifer, G.P. (2010). Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. Nucleic acids research *38*, e125.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome research *12*, 996-1006.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology *10*, R25.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., *et al.* (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315-322.

Myers, R.M., Stamatoyannopoulos, J., Snyder, M., Dunham, I., Hardison, R.C., Bernstein, B.E., Gingeras, T.R., Kent, W.J., Birney, E., Wold, B., *et al.* (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS biology *9*, e1001046.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome research *15*, 1034-1050.

Song, C.X., Szulwach, K.E., Fu, Y., Dai, Q., Yi, C., Li, X., Li, Y., Chen, C.H., Zhang, W., Jian, X., *et al.* (2011). Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. Nature biotechnology *29*, 68-72.

Szulwach, K.E., Li, X., Li, Y., Song, C.X., Han, J.W., Kim, S., Namburi, S., Hermetz, K., Kim, J.J., Rudd, M.K., *et al.* (2011). Integrating 5-hydroxymethylcytosine into the epigenomic landscape of human embryonic stem cells. PLoS Genet *7*, e1002154.

Williams, K., Christensen, J., Pedersen, M.T., Johansen, J.V., Cloos, P.A., Rappsilber, J., and Helin, K. (2011). TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. Nature *473*, 343-348.

Wu, H., D'Alessio, A.C., Ito, S., Wang, Z., Cui, K., Zhao, K., Sun, Y.E., and Zhang, Y. (2011). Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. Genes & development *25*, 679-684.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., *et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome biology *9*, R137.