SUPPLEMENTAL INFORMATION

- Supplementary Figures

Figure S1. Comparison of antibody-based and chemical labeling based mapping methods

Figure S2. *Tdg*-depletion does not affect the mouse ESC morphology or the RNA levels of Tet and pluripotency genes

Figure S3. Overall genomic distributions of 5mC/5hmC/5fC/5caC enriched regions in control and *Tdg*-deficient mouse ESCs

Figure S4. *Tdg*-depletion induced 5caC peaks are enriched at distal regulatory regions

Figure S5. *Tdg*-depletion induced 5caC peaks are enriched at active enhancers and pluripotency TF binding sites in mouse ESCs

Figure S6. *Tdg*-depletion induced 5caC peaks are enriched at transcriptionally inactive (bivalent or silent) gene promoters in mouse ESCs

Figure S7. Complex relationship between gene expression and cytosine modification cycling

- Supplementary Tables

Table S1. Summary of sequencing reads (related to Figure 2)

Table S2. 5caC-enriched regions in control mouse ESCs (related to Figure 3)

Table S3. 5caC-enriched regions in *Tdg*-deficient mouse ESCs (related to Figure 3)

Table S4. 5fC-enriched regions in control mouse ESCs (related to Figure 3)

Table S5. 5fC-enriched regions in *Tdg*-deficient mouse ESCs (related to Figure 3)

- Supplementary Figure Legends

- Extended Experimental Procedures

- Supplementary References

SUPPLEMENTARY FIGURE LEGENDS

Figure S1. Comparison of antibody-based and chemical labeling based mapping methods

(A) Diagram illustration of the DNA demethylation process mediated by TET/TDG/Base excision repair (BER).

(B) Comparison of 5hmC signal tracks and enriched regions (peaks) derived from antibody-based (5hmC Ab from this study; CMS Ab (Pastor et al., 2011)) and chemical labeling-based (GLIB (Pastor et al., 2011)) methods at representative loci. The 5hmC base-resolution map in mouse ESCs was also shown as a reference (TAB-seq dataset (Yu et al., 2012)). 5hmC enriched regions identified by two different cutoffs (P< 1e-3 or 1e-5) from 5hmC Ab datasets were shown. 5fC and 5caC signal tracks in control (shCtrl) and *Tdg*-deficient (shTdg) mouse ESCs were also shown. Highlighted were regions that are recovered by both GLIB and 5hmC Ab (region #2, #3 and #4) or only by one method (region #1 by GLIB only; #5 by 5hmC Ab only).

(C) Percentage of total (2.06 million, blue), sparsely distributed (0.44 million, green) and clustered (1.62 million, red) high-confidence 5hmC marks (identified in base-resolution map from Yu et al., 2012) that can be recovered by different 5hmC affinity enrichment methods (antibody-based: 5hmC Ab and CMS Ab; chemical labeling: GLIB). Sparsely distributed 5hmC marks are defined as a single 5hmC within at least 1kb genomic region. Clustered 5hmC marks are defined as two or more 5hmC within 1kb.

(D) Percentage of 5hmC enriched regions identified by different methods (5hmC Ab, CMS and GLIB) that contain one or more high-confidence 5hmC marks of the base-resolution map (Yu et al., 2012).

(E) Correlation matrix (Pearson coefficient) of all sequencing experiments (5mC/5hmC/5fC/5caC/IgG/Input in control (shCONT) or *Tdg*-deficient (shTDG) mouse ESCs). The comparison was based on normalized read density within 2kb non-overlapping genomic intervals. Note that 5caC profiles in shTDG were closely clustered with 5hmC profiles in shCONT or shTDG.

Figure S2. *Tdg*-depletion does not affect the mouse ESC morphology or the RNA levels of Tet and pluripotency genes, Related to Figure 1

(A) Alkaline phosphatase staining of control and *Tdg*-deficient mouse ESCs.

(B) RT-qPCR analyses of Tet and representative pluripotency genes in control and *Tdg*-deficient mouse ESCs. Data are presented as mean ± SEM.

Figure S3. Overall genomic distributions of 5mC/5hmC/5fC/5caC enriched regions in control and *Tdg*-deficient mouse ESCs, Related to Figure 3

(A) Enrichment (log2 ratios of observed over random) of 5mC/5hmC/5fC/5caC in various genomic features. Values represent means of two biological replicates with the ends of the error bars corresponding to the individual data point.

(B) Average signals of 5mC/5hmC/5fC/5caC along concatenated exons/introns of each RefSeq genes. Rp10m, reads per 10 million. The increases of 5fC/5caC signals in *Tdg*-deficient cells compared with control cells are significantly higher in exons than in introns.

(C) Heatmap representation of normalized read density of 5mC/5hmC/5fC/5caC at Tet1 bound regions. The distribution of CpG islands (CGIs) is also shown. The heatmaps are rank-ordered from regions with CGIs of the longest length to those without CGIs in the flanking 5 kb regions. Color scale indicates the DIP-Seq signal in reads per 10 million.

(D) Venn diagram showing the overlap of 5mC-/5hmC-/5fC-/5caC-enriched regions in *Tdg*-deficient mouse ESCs.

(E) Pie charts showing pairwise comparisons between 5fC-/5caC-enriched regions and 5mC-/5hmC-enriched regions in *Tdg*-deficient mouse ESCs. 5fC peaks tend to co-localize with 5mC enriched regions, while 5caC peaks preferentially overlap with 5hmC-enriched regions.

Figure S4. *Tdg*-depletion induced 5caC peaks are enriched at distal regulatory regions, Related to Figure 4

(A) Average conservation PhastCons scores within regions flanking the center of 5mC/5hmC/5fC/5caC peaks (overlapping with promoters or exons) in *Tdg*-deficient

mouse ESCs. The numbers of peaks that are overlapping with exons and proximal promoters (+/- 1kb flanking TSSs) for each cytosine modification are also shown.

(B) Heat maps of 5mC/5hmC/5fC/5caC levels (normalized read density) in control and *Tdg*-deficient cells at centers of annotated genomic features or enriched regions for transcriptional regulators, histone modifications, pluripotency transcription factors (TFs) and distal regulator regions (derived from published datasets in mouse ESCs). The difference in 5hmC and 5caC levels between control and *Tdg*-deficient cells (subtraction of signals in shCtrl from those in shTdg) is also shown for all, proximal (overlapping with +/-1kb flanking TSSs) and distal features. The heatmap was clustered by hierarchical clustering (complete linkage).

Figure S5. *Tdg*-depletion induced 5caC peaks are enriched at active enhancers and pluripotency TF binding sites in mouse ESCs, Related to Figure 5

(A) Heat map of Tet1 binding and DNase I hypersensitivity (normalized read density) in wild-type mouse ESCs at centers of previously identified tissue-specific enhancers (Shen et al., 2012).

(B) Average 5mC/5hmC/5fC/5caC signals in control and *Tdg*-deficient mouse ESCs at center of mouse ESC (mES)-specific and neural progenitor (NP)-specific LMRs (Stadler et al., 2011).

(C) Heat map of 5hmC and 5caC levels (normalized read density) in control and Tdg-deficient cells at centers of binding sites of mouse ESC TFs (Oct4, Nanog and Sox2 and neuronal TFs.

(D) 5mC/5hmC/5fC/5caC levels in control and *Tdg*-deficient mouse ESCs at a representative locus (within the *Vps26a* gene body) of binding sites of pluripotency TFs (Oct4/Nanog/Sox2). Other genomic features (e.g. DNase I hypersensitivity sites, H3K4me1 enriched regions, LMRs and enhancer-related epigenetic regulator LSD1) are also shown.

Figure S6. *Tdg*-depletion induced 5caC peaks are enriched at transcriptionally inactive (bivalent or silent) gene promoters in mouse ESCs, Related to Figure 6

(A) Average gene expression levels in wild-type mouse ESCs (average of four previously published RNA-seq experiments (Ficz et al., 2011)) are shown for four groups of gene promoters that are associated with distinct chromatin states (active: H3K4me3+/H3K79me2+; initiated: H3K4me3+ only; bivalent: H3K4me3+/H3K27me3+; silent: none).

(B) 5mC/5hmC/5fC/5caC levels in control and *Tdg*-deficient mouse ESCs at representative loci of gene promoters that are associated with different histone modification states. The gene promoters are highlighted by grey bars.

(C) Boxplots of normalized 5hmC, 5fC and 5caC levels (read per million reads and kilo bases, rpkm) in control and *Tdg*-deficient cells within genomic regions enriched for major histone modifications. H3K4me3+H3K79me2 and H3K4me3+H3K27me3 denote regions that are associated with both histone modifications.

Figure S7. Complex relationship between gene expression and cytosine modification cycling, Related to Figure 7

(A) Average signals of 5fC and 5mC within genes expressed at different levels in control (left) and *Tdg*-deficient (right) mouse ESCs.

(B) Average 5fC/5hmC/5mC signals in control (left) and *Tdg*-deficient (right) mouse ESCs at the TSS of down-regulated and up-regulated genes.

EXTENDED EXPERIMENTAL PROCEDURES

Lentiviral vector construction and virus production

Stable knockdown of Tdg was achieved using a lentiviral system obtained from the National Institutes of Health (NIH) AIDS Research and Reference Reagent Program (https://www.aidsreagent.org/). This system allows selection of infected cells with puromycin. The short hairpin RNA (shRNA) targeting *Tdg* (5'-GCAAGGATCTGTCTAGTAA-3') and the control shRNA (5'-GTTCAGATGTGCGGCGAGT-3') were cloned into the BbsI/HindIII sites, where the transcription is under the control of U6 promoter. To generate lentiviruses, the transducing vectors (pTY, pNHP and pHEF1α-VSVG) were co-transfected into 293T cells (He et al., 2008). The supernatant was harvested at 24, 36 and 48 hours after transfection, filtered through 0.45μm membrane and concentrated using a centrifugal filter (EMD Millipore, Amicon Ultra 100k).

5fC and 5caC antibody generation and characterization

To generate 5fC and 5caC antibodies, ribonucleoside forms of 5fC and 5caC were conjugated to KLH before being injected into rabbits (Inoue et al., 2011). For dot blot characterization of the antisera specificities, different amounts of 38-mer DNA oligos (5′-AGCCXGXGCXGXGCXGGTXGAGXGGCXGCTCCXGCAGC-3′, where X is either a C or modified C) were denatured with 0.1 M NaOH and spotted on nitrocellulose membranes (BioRad, 162-0112). The membrane was baked at 80 ˚C and then blocked in 5% non-fat milk in TBS containing 0.1% Tween 20 (TBST) for 1 h at room temperature. The membranes were then incubated with 1:10,000 dilution of anti-5hmC antiserum (Active Motif, 39769), 1:5,000 dilution of anti-5fC antiserum or 1:2,000 dilution of anti-5caC antiserum overnight at 4 ˚C, respectively. After three rounds of washes with TBST, membranes were incubated with 1:2,000 dilution of HRP-conjugated anti-rabbit IgG secondary antibody. The membranes were then washed with TBST and treated with ECL.

Cell culture and lentiviral transduction

E14 mouse ESCs are cultured on the 0.1% gelatin coated plates in DMEM medium (Sigma) supplemented with 100 U/ml penicillin/streptomycin, 15% fetal bovine serum (Sigma), 1x nonessential amino acid, 1x sodium pyruvate, 1x GlutaMax, 1x beta-mercaptoethanol (Invitrogen) and 1000 units/ml leukemia inhibitory factor (ESGRO, EMD Millipore). To generate control and *Tdg*-knockdown cells, $1x10^5$ cells are infected with lentivirus (MOI= 10) in a 24-well plate. 48 hours after infection, puromycin (2 ug/ml) is added to the medium for selecting infected cells. Cells are split when necessary until being harvested after one week.

Alkaline phosphatase staining

Alkaline phosphatase staining was performed using an alkaline phosphatase detection kit (Millipore, SCR004) following manufacturer's instructions.

RT-qPCR analysis

RNA was extracted and purified from cells using QIAshredder (QIAGEN) and RNeasy spin columns (QIAGEN). Total RNA (1 μg) was subjected to reverse transcription using random primers (Promega) and the Superscript II reverse transcriptase (Invitrogen). Real-time qPCR reactions were performed on an Applied Biosystems ViiA7 system using FAST SYBR Green PCR master mix (Applied Biosystems). cDNA levels of target genes were analyzed using the $\Delta\Delta Ct$ method, and the expression of individual genes is normalized to the expression level of GAPDH. Primers for RT-qPCR are listed below.

RT-qPCR primers used in this study

Tdg-F	GCAGGAGAAAATCACAGACG
Tdg-R	CCCAGGGTAGTGATGTCCTT
Oct4-F	CCAATCAGCTTGGGCTAGAG
Oct4-R	CCTGGGAAAGGTGTCCTGTA
Nanog-F	AAGCAGAAGATGCGGACTGT
Nanog-R	ATCTGCTGGAGGCTGAGGTA
Sox2-F	GAACGCCTTCATGGTATGGT
Sox2-R	TTGCTGATCTCCGAGTTGTG

Tet1-F	GAGCCTGTTCCTCGATGTGG
Tet1-R	CAAACCCACCTGAGGCTGTT
Tet2-F	TGTTGTTGTCAGGGTGAGAATC
Tet2-R	TCTTGCTTCTGGCAAACTTACA
Tet3-F	CCGGATTGAGAAGGTCATCTAC
Tet3-R	AAGATAACAATCACGGCGTTCT
GAPDH-F	CATGGCCTTCCGTGTTCCTA
GAPDH-R	GCCTGCTTCACCACCTTCTT

Microarray analysis

Total RNA was extracted and purified from the cells using QIAshredder and RNeasy spin columns (QIAGEN). The RNA processing and hybridization to the Mouse Gene 1.0 ST Array (Affymetrix) were performed by Functional Genomics Core Facility at UNC-Chapel Hill. Three biological independent samples were analyzed for both control and *Tdg*-deficient mouse ESCs. Probe sets with intensity value below the 20th percentile were filtered out, and unpaired *t*-test (asymptontic) was used in data analysis.

Immunocytochemistry of the surface spreads of mouse ESCs

Preparation of the surface spreads was performed as previously described with minor modifications (Yamaguchi et al., 2012). Briefly, ESCs were dissociated by 0.05% trypsin/EDTA, followed by wash and re-suspension into PBS. An equal volume of hypotonic buffer [30 mM Tris-HCl (pH 8.3), 5 mM EDTA, 1.7% sucrose, 0.5% trisodium citrate dehydrate] was added to the suspension. After 7 min of incubation in room temperature (RT), cell suspension was centrifuged and re-suspend in 100 mM sucrose. The cell suspension was spread onto glass slides dipped into fixative solution [1% paraformaldehyde, 0.15% Triton X-100 and 3 mM dithiothreitol, pH 9.2]. The glass slides were kept overnight in a humidified box at 4 °C. The slides were washed in water containing 0.4% Photoflow (Kodak), and completely dried at RT. To stain spreads, dried slide glasses were washed with 0.1% Tween20/PBS (PBST) for 10 min, and incubated with hydrochloric acid solution (4 N hydrochloric acid, 0.1% Triton-X 100 in distilled water) for 20 min at room temperature, followed by washes in PBST. The slides were incubated with blocking buffer (3% BSA, 2% donkey serum/PBST) for 1 h at RT,

and incubated with the primary antibodies at 4 °C overnight, followed by incubation with appropriate secondary antibodies for 1 h at RT.

Mass spectrometric analysis

Genomic DNA was extracted and purified from cells using DNeasy Blood & Tissue Kit (QIAGEN). For mass spectrometric quantification, 30 ug of genomic DNA was heat-denatured, hydrolyzed with 600U of nuclease S1 (Sigma) in Buffer 1 (0.5 mM $ZnSO_4$, 14 mM sodium acetate, pH 5.2) at 37 ˚C for at least 1 hour (total volume is 250 µl), followed by the addition of 30 µl 10xBuffer 2 (560 mM Tris-HCl, 30 mM NaCl, 10 mM $MgCl_2$, pH 8.3), 5 µg of phosphodiesterase I (Worthington) and 20 U of CIAP for an additional 1 hour (final volume is 300 µl), and filtered with Nanosep3K (PALL). The hydrolyzed samples were subjected to HPLC fraction collection and then mass spectrometric analysis as previously described (Ito et al., 2011; Shen and Zhang, 2012). The amounts of 5mC/5hmC/5fC/5caC in each sample were normalized to the amount of total cytosine.

DNA immunoprecipitation (DIP)-qPCR analysis

To test the utility of 5fC/5caC specific antibodies in DIP assays, the 38-mer double-stranded DNA oligos containing 5mC, 5hmC, 5fC, 5caC or unmodified C, which were also used in dot blot assays, were ligated to a PCR adaptor (Forward: 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3', Reverse: 5'-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'). These oligos (5 pg) were denatured and used as the input in the presence of 10 ug sonicated salmon sperm DNA (ssDNA). For each DIP assay, 1 ul (anti-5fC) or 0.3 ul (anti-5caC) of antiserum were used, alternatively, 1 ul of mouse and rabbit pre-immune serum mixture was used as IgG control. DNA and antibodies were incubated at 4 ˚C overnight in a final volume of 500 ul DIP buffer (10 mM sodium phosphate (pH 7.0), 140 mM NaCl, 0.05% Triton X-100) as previously described (Wu et al., 2011a). After the DNA-antibody incubation, 30 ul protein G Dynabeads (Invitrogen) were added to the tube and incubated with the DNA-antibody mixture for 2 h at 4 ˚C. The beads were washed three times with 1 mL of DIP buffer, then treated with proteinase K at 55 ˚C for 3 hours, and the immunoprecipitated DNA was purified by phenol-chloroform extraction followed by ethanol precipitation.

Immunoprecipitated DNA was then analyzed by qPCR (Forward Primer: 5-ACACTCTTTCCCTACACGACGC-3', Reverse Primer: 5'-CTCGGCATTCCTGCTGAAC-3'), and data were calculated as percentage of input before further analysis.

To perform DIP-qPCR assays with genomic DNA, purified genomic DNA was sonicated to ~500 bp using Bioruptor (Diagenode), heat-denatured and used as the input (10 ug for each DIP assay). The DIP procedures are the same as the steps in the oligo DIP assay described above. The qPCR primers are: 5'-AGCCAGTATGGCGTACATCTGTGT-3' and 5'-TGTGAAGAGTGGCTCACGGACAAA-3' for *Esrrb*, 5'-TTTTCTCTGTCTTCCCTGTCTTGG-3' and 5'-CGGGCTTTCTTTCTAACCACTTTC-3' for *Sox17*, 5'-CAAAATTGGAATATCTTTAAGGTAGC-3' and 5'-TTTGGCTTTACAAGTGGAACA-3' for *Tcl1*.

Genome-wide 5mC/5hmC/5fC/5caC sequencing (DIP-Seq)

To prepare sequencing libraries, purified genomic DNA from either control or Tdg knockdown mouse ESCs was first sonicated to ~250 bp using Bioruptor (Diagenode), then end-repaired and ligated to Illumina PE adaptors using NEBNext DNA Library Prep Master Mix Set (NEB). In DIP assays, 10 ug of the adaptor ligated genomic DNA was used as input, and 5 ul of 5mC antibody (Eurogentec, BI-MECY-0500), 5 ul of 5hmC antibody (Active Motif, 39791), 1 ul of 5fC anti-serum or 0.3 ul of 5caC anti-serum was added to immunoprecipitate modified DNA. For the IgG control, 1 ul of mouse and rabbit preimmune serum mixture was used. The DIP procedures are the same as the steps in the oligo DIP assay described above. Each of the immunoprecipitated DNA, as well as the input, was amplified with forward primer PCR_F (5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC -3') and indexed reverse primers (5'-CAAGCAGAAGACGGCATACGAGATXXXXXXCTCGGCATTCCTGCTGAACCG CTCTT-3', where XXXXXX is the 6 bp barcode) in a 50 µl PCR reaction with 0.2X Sybr-Green I and 1X Phusion High-fidelity PCR Master Mix (NEB). PCR reactions were terminated before entering the non-exponential plateau phase. The amplified libraries of the expected average size range (400 bp) were purified using Qiaquick PCR

Amplification columns (Qiagen) and quantified using Qubit fluorometer (Invitrogen). For each batch of the experiment, the 12 libraries were then separated into two batches of six libraries, resulting in two "indexed libraries". The two libraries were sequenced on the Illumina GAIIx sequencer, each occupying 4 lanes of an 8-lane single-end flow-cell. Barcodes were sequenced after the standard first read using N2IndSeq primer (5'-AAGAGCGGTTCAGCAGGAATGCCGAG-3'). After base-calling, each sequence was decoded using an in-house Perl script which required at least five out of six matching positions in the barcode sequence. All sequencing reads were trimmed from 5' end to obtain final 37 bases in length for further analysis.

Repetitive sequence analysis

We first mapped all reads to the mouse genome (mm9) using bowtie (v0.12.7), with the options "-v 2 -m 1 --best" to determine the total numbers of unmappable, multi-hit, and uniquely mapped reads. And then we separately mapped those multi-hit and uniquely mapped reads to the UCSC RepeatMasker track (RMSK) sequences, to determine the percentages of RMSK reads in multi-hit and uniquely mapped reads, respectively. To determine which class of repetitive sequences these RMSK reads overlap with, we created separate Bowtie indices for each class of repetitive sequences separately using RMSK annotation, and mapped all RMSK reads to each using the same options as above. Only those reads that uniquely belong to one class of repetitive sequences were counted.

Identification of 5mC/5hmC/5fC/5caC enriched regions

We mapped DIP-seq reads using Bowtie (v0.12.7) to obtain only those reads that are mapped uniquely to mouse genome (mm9) with at most 3 mismatches. The parameters used in Bowtie were "-v 3 -m 1 -y --best --strata". To accurately identify regions enriched for each cytosine modification, we adapted a computational pipeline as previously described (Shen et al., 2012). Specifically, we first identified peak candidates with MACS (v1.4.2) (Zhang et al., 2008) using input as the control dataset and parameters allowing at most two reads at the same genomic position ("--gsize=mm --pvalue=1e-5 --keep-dup 2 --nomodel"). To obtain high quality peaks for downstream analysis, we further filtered out peak candidates that have high signals in IgG mock DIP experiments.

To this end, we computed normalized read density values (reads per million reads per kilo bases, rpkm) within each peak in both DIP-seq and IgG-seq data. To remove poor quality peaks, we applied following procedures: DIP_rpkm >= 2 * IgG_rpkm and (DIP_rpkm - IgG_rpkm) > 1.

Generation of wig tracks for peak visualization and downstream analysis.

To visualize peaks in the genome browser, we generated wig track files for each dataset with MACS (v1.4.2) by extending the uniquely mapped reads (keeping at most two read at the same genomic position) to 200bp toward the 3' end and binning the read count to 50bp intervals. We further normalized tag counts in each bin to the total number of uniquely mapped reads (reads per 10 million reads, rp10m). To remove non-specific signals, IgG samples were processed similarly and their rp10m values were subtracted from DIP-seq wig files. Thus, the final wig track for each DIP-seq sample is computed as: DIP_rp10m – IgG_rp10m. The wig tracks are available for download at http://labs.idi.harvard.edu/zhang/data.htm.

Conservation analysis

To calculate the conservation at 5hmC/5fC/5caC peak centers (Figure 4A and S4A), we downloaded the PhastCons11way (mm9) track from UCSC (http://genome.ucsc.edu/), which include all euarchontoglires. This track contains pre-calculated conservation score for each base in the mouse genome based on multi-species alignments between mouse and all euarchontoglires (Siepel et al., 2005).

Promoter classification by chromatin states

Gene promoters were classified into four groups (active, initiated, bivalent and silent) by histone modifications as previously described (Whyte et al., 2012). Specifically, 1) active promoters are associated with H3K4me3 (+/- 2kb flanking TSS) and H3K79me2 (5kb downstream of TSS). 2) initiated promoters are associated with H3K4me3 (+/- 2kb flanking TSS) only. 3) bivalent promoters are associated with H3K4me3 (+/- 2kb flanking TSS) and H3K27me3 (+/- 5kb flanking TSS). 4) silent promoters are not associated with H3K4me3/H3K79me2/H3K27me3.

Published datasets

Annotations of known Refseq transcripts and CpG islands were obtained from UCSC (downloaded on Jan, 2012). To calculate Figure S6A and 7A, previously published RNA-seq datasets of wild-type mouse ESCs (Ficz et al., 2011) were processed using TopHat (v1.2.0) and Cufflinks (v1.0.1) as previously described (Wu and Zhang, 2011). Specifically, normalized expression levels (measured by Fragments Per Kilobase per Million mapped fragments, FPKM) associated with distinct isoforms from the same transcript unit were summed to estimate the gene-level expression. To calculate heat map in Figure 4C, we used following published datasets: Tet1 (Wu et al., 2011b), Kdm2a (Blackledge et al., 2010), H3K4me3, H3K36me3, H3K27me3, H3K9me3, H4K20me3 (Mikkelsen et al., 2007), H3K4me1 (Meissner et al., 2008), Pol2 (pan), Pol2 (Ser2), Pol2 (Ser5), NelfA, Ctr9, Spt5 (Rahl et al., 2010), Ezh2, Suz12 (Ku et al., 2008), Oct4, Nanog, Sox2, H3K79me2 (Marson et al., 2008), Med1, Med12, Nipbl, Smc1a, Smc3, TBP (Kagey et al., 2010), LSD1, Mi2b, Hdac1, Hdac2, Rest, Corest (Whyte et al., 2012), LMR (Stadler et al., 2011), H3K27ac, p300 (Creyghton et al., 2010), Esrrb, cMyc, nMyc, Tcfcp2l1, E2f1, Stat3, Smad1, Zfx (Chen et al., 2008), CTCF (Shen et al., 2012), DNase I hypersensitive sites (ENCODE project at genome.ucsc.edu), and Topological domain boundary (Dixon et al., 2012). To calculate heat map in Figure 5A, we used the list of tissue-specific enhancers that were identified previously by the modENCODE project (Shen et al., 2012). To calculate Figure 5B and S5A, we used a previously identified list of neuronal TF binding sites (Kim et al., 2010).

SUPPLEMENTAL REFERENCES

Blackledge, N.P., Zhou, J.C., Tolstorukov, M.Y., Farcas, A.M., Park, P.J., and Klose, R.J. (2010). CpG islands recruit a histone H3 lysine 36 demethylase. Mol Cell *38*, 179-190.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J.*, et al.* (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell *133*, 1106-1117.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A.*, et al.* (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A *107*, 21931-21936.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376-380.

Ficz, G., Branco, M.R., Seisenberger, S., Santos, F., Krueger, F., Hore, T.A., Marques, C.J., Andrews, S., and Reik, W. (2011). Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. Nature *473*, 398-402.

He, J., Kallin, E.M., Tsukada, Y., and Zhang, Y. (2008). The H3K36 demethylase Jhdm1b/Kdm2b regulates cell proliferation and senescence through p15(Ink4b). Nat Struct Mol Biol *15*, 1169-1175.

Inoue, A., Shen, L., Dai, Q., He, C., and Zhang, Y. (2011). Generation and replication-dependent dilution of 5fC and 5caC during mouse preimplantation development. Cell Res *21*, 1670-1676.

Ito, S., Shen, L., Dai, Q., Wu, S.C., Collins, L.B., Swenberg, J.A., He, C., and Zhang, Y. (2011). Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science *333*, 1300-1303.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S.*, et al.* (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430-435.

Kim, T.K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S.*, et al.* (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182-187.

Ku, M., Koche, R.P., Rheinbay, E., Mendenhall, E.M., Endoh, M., Mikkelsen, T.S., Presser, A., Nusbaum, C., Xie, X., Chi, A.S.*, et al.* (2008). Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. PLoS Genet *4*, e1000242.

Marson, A., Levine, S.S., Cole, M.F., Frampton, G.M., Brambrink, T., Johnstone, S., Guenther, M.G., Johnston, W.K., Wernig, M., Newman, J.*, et al.* (2008). Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell *134*, 521-533.

Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., Jaffe, D.B.*, et al.* (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature *454*, 766-770.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P.*, et al.* (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553-560.

Pastor, W.A., Pape, U.J., Huang, Y., Henderson, H.R., Lister, R., Ko, M., McLoughlin, E.M., Brudno, Y., Mahapatra, S., Kapranov, P.*, et al.* (2011). Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. Nature *473*, 394-397.

Rahl, P.B., Lin, C.Y., Seila, A.C., Flynn, R.A., McCuine, S., Burge, C.B., Sharp, P.A., and Young, R.A. (2010). c-Myc regulates transcriptional pause release. Cell *141*, 432-445.

Shen, L., and Zhang, Y. (2012). Enzymatic analysis of tet proteins: key enzymes in the metabolism of DNA methylation. Methods Enzymol *512*, 93-105.

Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V.*, et al.* (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116-120.

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 Res *15*, 1034-1050.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Scholer, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D.*, et al.* (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490-495.

Whyte, W.A., Bilodeau, S., Orlando, D.A., Hoke, H.A., Frampton, G.M., Foster, C.T., Cowley, S.M., and Young, R.A. (2012). Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. Nature *482*, 221-225.

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

Wu, H., D'Alessio, A.C., Ito, S., Xia, K., Wang, Z., Cui, K., Zhao, K., Eve Sun, Y., and Zhang, Y. (2011b). Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. Nature *473*, 389-393.

Wu, H., and Zhang, Y. (2011). Tet1 and 5-hydroxymethylation: A genome-wide view in mouse embryonic stem cells. Cell Cycle *10*.

Yamaguchi, S., Hong, K., Liu, R., Shen, L., Inoue, A., Diep, D., Zhang, K., and Zhang, Y. (2012). Tet1 controls meiosis by regulating meiotic gene expression. Nature.

Yu, M., Hon, G.C., Szulwach, K.E., Song, C.X., Zhang, L., Kim, A., Li, X., Dai, Q., Shen, Y., Park, B.*, et al.* (2012). Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. Cell *149*, 1368-1380.

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 Biol *9*, R137.

Figure S1

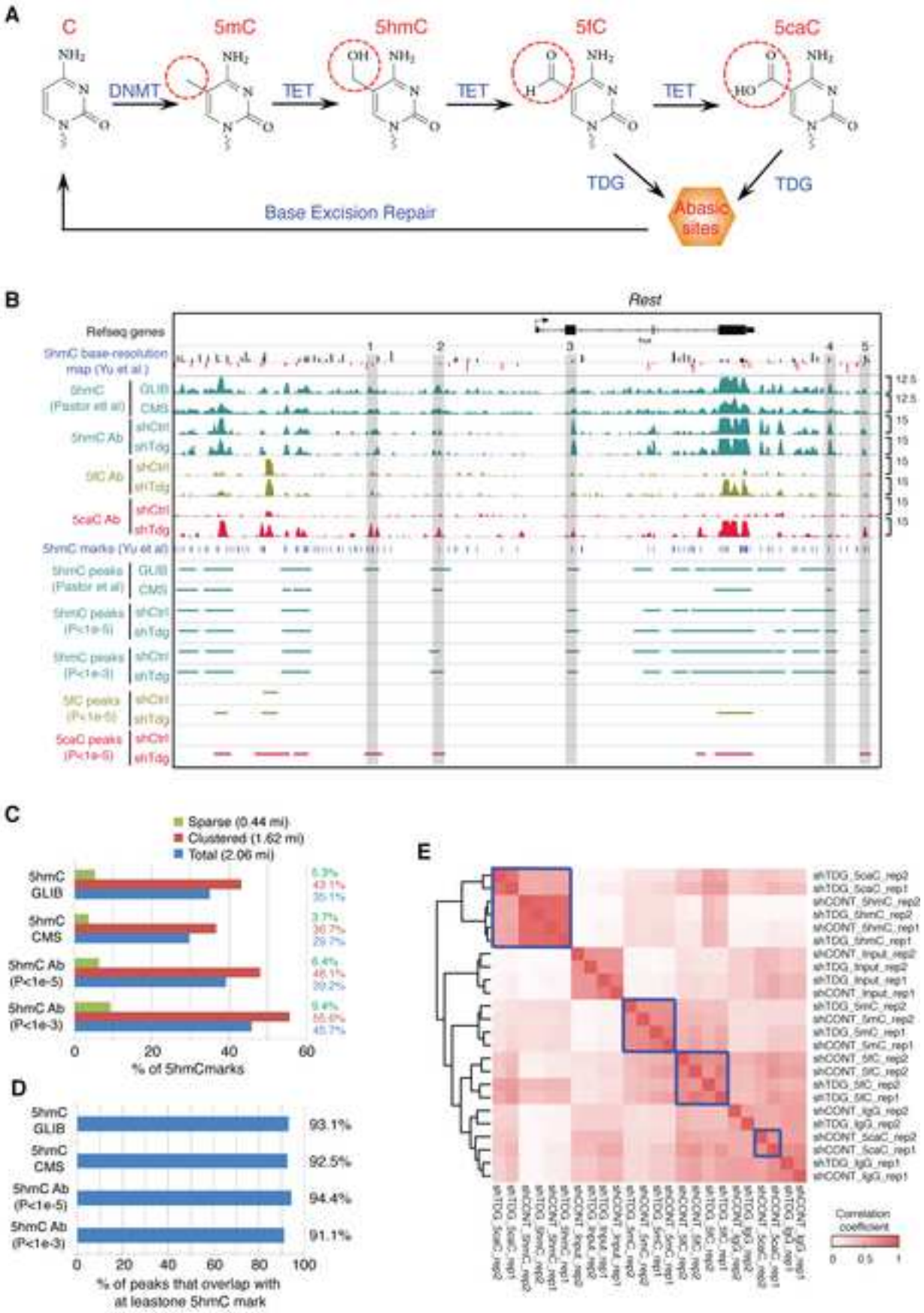
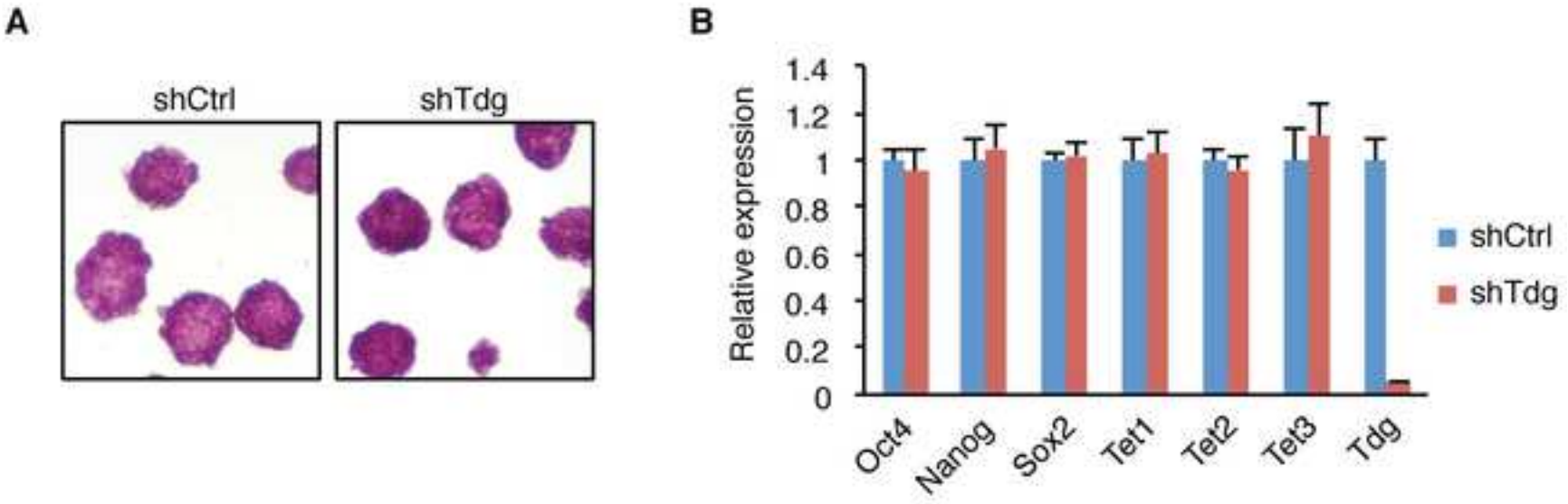

A
C
5mC
5hmC
5fC
5caC
DNMT
TET
TET
TET
TDG
TDG
Base Excision Repair
Abasic sites
B
Rest
Refseq genes
5hmC base-resolution map (Yu et al.)
5hmC (Pastor et al)
GLIB
CMS
5hmC Ab
shCtrl
shTdg
5fC Ab
5caC Ab
5hmC marks (Yu et al)
5hmC peaks (Pastor et al)
5hmC peaks (P<1e-5)
5hmC peaks (P<1e-3)
5fC peaks (P<1e-5)
5caC peaks (P<1e-5)
C
Sparse (0.44 mi)
Clustered (1.62 mi)
Total (2.06 mi)
5hmC GLIB
5hmC CMS
5hmC Ab (P<1e-5)
5hmC Ab (P<1e-3)
6.3%
43.1%
35.1%
3.7%
30.7%
28.7%
6.4%
46.1%
39.2%
9.4%
55.6%
45.7%
0
20
40
60
% of 5hmCmarks
D
5hmC GLIB
5hmC CMS
5hmC Ab (P<1e-5)
5hmC Ab (P<1e-3)
93.1%
92.5%
94.4%
91.1%
0
20
40
60
80
100
% of peaks that overlap with at leastone 5hmC mark
E
Correlation coefficient
0
0.5
1

Figure S2


A
shCtrl
shTdg
B
Relative expression
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Oct4
Nanog
Sox2
Tet1
Tet2
Tet3
Tdg
shCtrl
shTdg

Supplemental Figure3

Figure S3

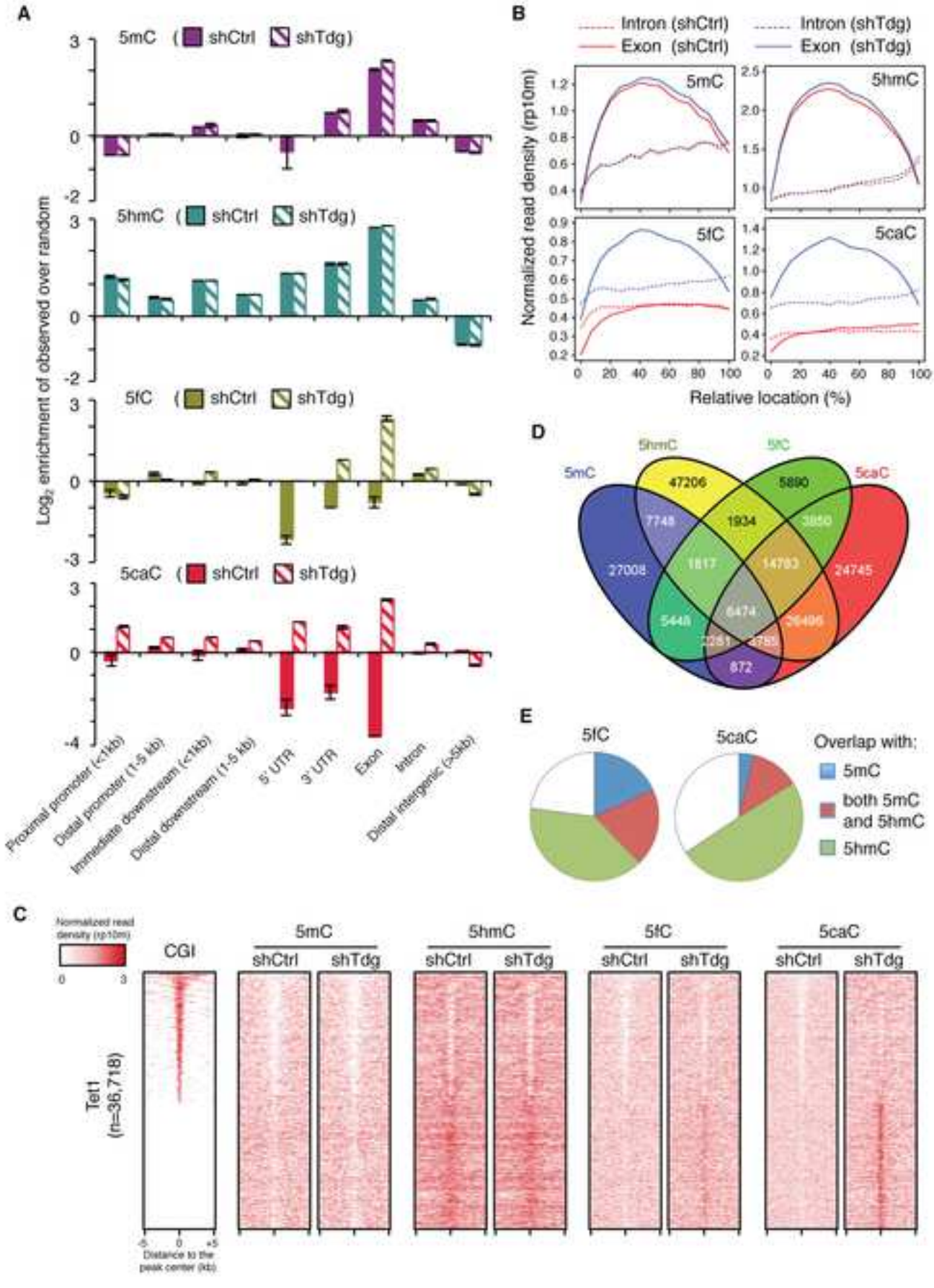


Figure S4

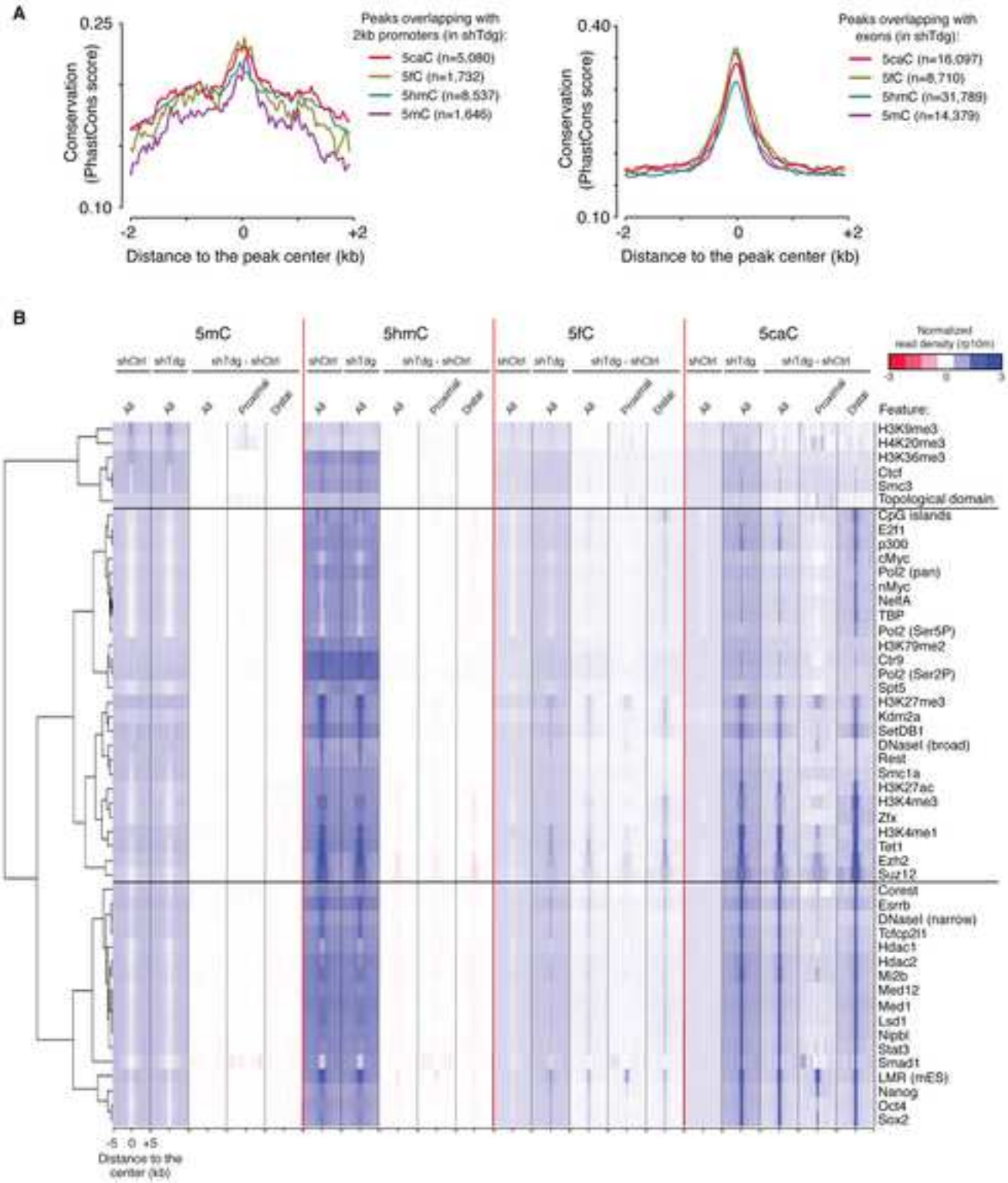
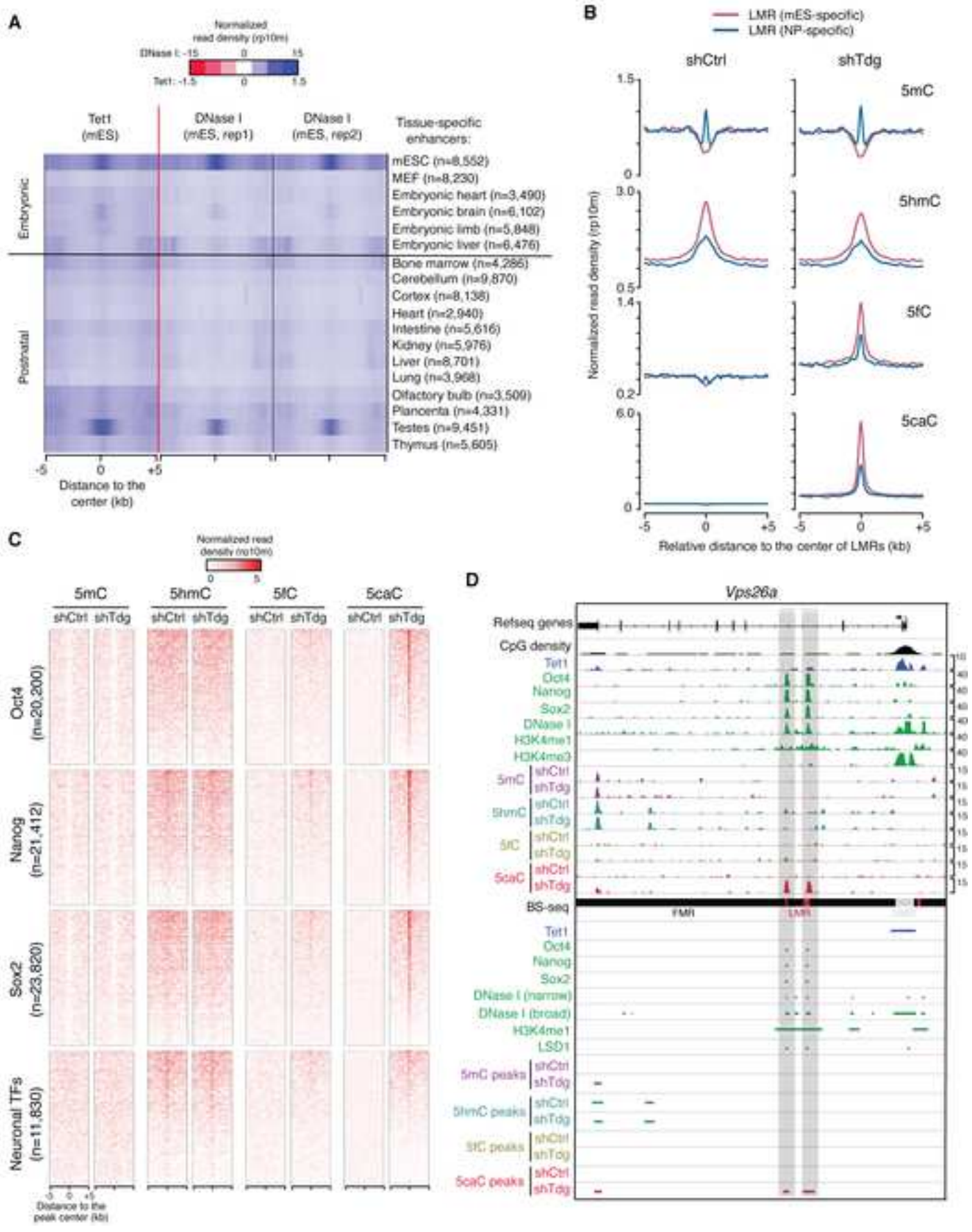

A
Conservation (PhastCons score)
0.25
0.10
-2
0
+2
Distance to the peak center (kb)
Peaks overlapping with 2kb promoters (in shTdg):
5caC (n=5,080)
5fC (n=1,732)
5hmC (n=8,537)
5mC (n=1,646)
0.40
0.10
Peaks overlapping with exons (in shTdg):
5caC (n=16,097)
5fC (n=8,710)
5hmC (n=31,789)
5mC (n=14,379)
B
5mC
5hmC
5fC
5caC
shCtrl
shTdg
shTdg - shCtrl
Normalized read density (rpl0m)
Feature:
H3K9me3
H4K20me3
H3K36me3
Ctcf
Smc3
Topological domain
CpG islands
E2f1
p300
cMyc
Pol2 (pan)
nMyc
NelfA
TBP
Pol2 (Ser5P)
H3K79me2
Ctr9
Pol2 (Ser2P)
Spt5
H3K27me3
Kdm2a
SetDB1
DNaseI (broad)
Rest
Smc1a
H3K27ac
H3K4me3
Zfx
H3K4me1
Tet1
Ezh2
Suz12
Corest
Esrrb
DNaseI (narrow)
Tcfcp2l1
Hdac1
Hdac2
Mi2b
Med12
Med1
Lsd1
Nipbl
Stat3
Smad1
LMR (mES)
Nanog
Oct4
Sox2
-5 0 +5
Distance to the center (kb)

Figure S5


A
Normalized read density (rp10m)
DNase I: -15 0 15
Tet1: -1.5 0 1.5
Tet1 (mES)
DNase I (mES, rep1)
DNase I (mES, rep2)
Tissue-specific enhancers:
Embryonic
Postnatal
mESC (n=8,552)
MEF (n=8,230)
Embryonic heart (n=3,490)
Embryonic brain (n=6,102)
Embryonic limb (n=5,848)
Embryonic liver (n=6,476)
Bone marrow (n=4,286)
Cerebellum (n=9,870)
Cortex (n=8,138)
Heart (n=2,940)
Intestine (n=5,616)
Kidney (n=5,976)
Liver (n=8,701)
Lung (n=3,968)
Olfactory bulb (n=3,509)
Placenta (n=4,331)
Testes (n=9,451)
Thymus (n=5,605)
-5 0 +5
Distance to the center (kb)
B
LMR (mES-specific)
LMR (NP-specific)
shCtrl
shTdg
5mC
5hmC
5fC
5caC
Normalized read density (rp10m)
Relative distance to the center of LMRs (kb)
C
Normalized read density (rp10m)
5mC
5hmC
5fC
5caC
shCtrl shTdg
Oct4 (n=20,200)
Nanog (n=21,412)
Sox2 (n=23,820)
Neuronal TFs (n=11,830)
Distance to the peak center (kb)
D
Vps26a
Refseq genes
CpG density
Tet1
Oct4
Nanog
Sox2
DNase I
H3K4me1
H3K4me3
5mC
5hmC
5fC
5caC
shCtrl
shTdg
BS-seq
FMR
LMR
DNase I (narrow)
DNase I (broad)
LSD1
5mC peaks
5hmC peaks
5fC peaks
5caC peaks

Supplemental Figure6

Figure S6

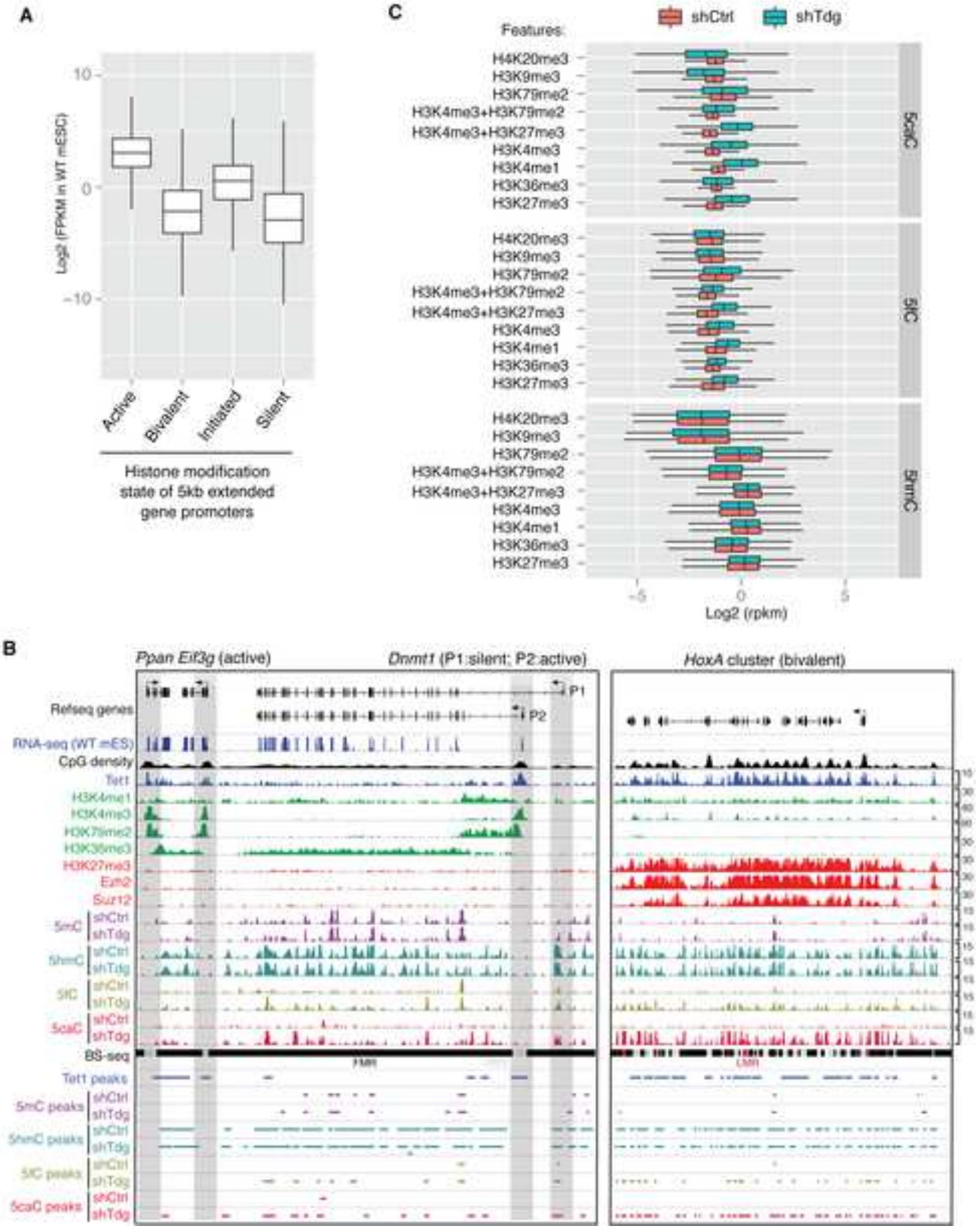
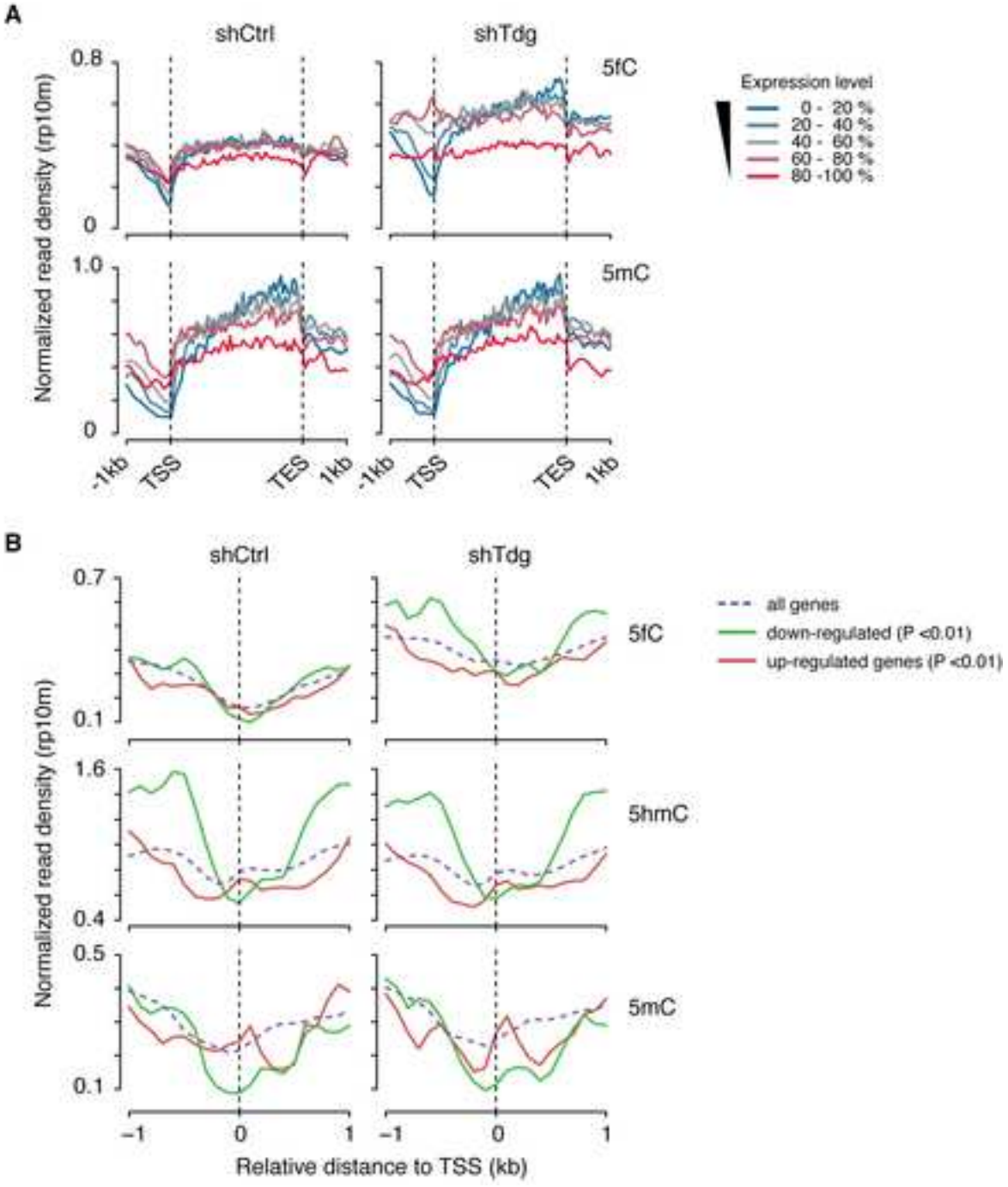


Figure S7


A
shCtrl
shTdg
5fC
5mC
Normalized read density (rp10m)
0.8
0
1.0
0
-1kb
TSS
TES
1kb
Expression level
0 - 20 %
20 - 40 %
40 - 60 %
60 - 80 %
80 -100 %
B
shCtrl
shTdg
5fC
5hmC
5mC
Normalized read density (rp10m)
0.7
0.1
1.6
0.4
0.5
0.1
−1
0
1
Relative distance to TSS (kb)
all genes
down-regulated (P <0.01)
up-regulated genes (P <0.01)