Cell

Resource

Genome-wide Analysis Reveals TET- and TDG-Dependent 5-Methylcytosine Oxidation Dynamics

Li Shen,[1,2,3,7] Hao Wu,[1,2,3,5,7,*] Dinh Diep,[6] Shinpei Yamaguchi,[1,2,3] Ana C. D'Alessio,[1,2,3] Ho-Lim Fung,[6] Kun Zhang,[6] and Yi Zhang[1,2,3,4,*]
[1]Howard Hughes Medical Institute
[2]Program in Cellular and Molecular Medicine
[3]Department of Genetics
[4]Harvard Stem Cell Institute
Harvard Medical School, WAB-149G, 200 Longwood Avenue, Boston, MA 02115, USA
[5]Department of Stem Cell and Regenerative Biology, Harvard University, 7 Divinity Avenue, Cambridge, MA 02138, USA
[6]Department of Bioengineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
[7]These authors contributed equally to this work
*Correspondence: haowu7@gmail.com (H.W.), yzhang@genetics.med.harvard.edu (Y.Z.)
http://dx.doi.org/10.1016/j.cell.2013.04.002

SUMMARY

TET dioxygenases successively oxidize 5-methylcytosine (5mC) in mammalian genomes to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC). 5fC/5caC can be excised and repaired to regenerate unmodified cytosines by thymine-DNA glycosylase (TDG) and base excision repair (BER) pathway, but it is unclear to what extent and at which part of the genome this active demethylation process takes place. Here, we have generated genome-wide distribution maps of 5hmC/5fC/5caC using modification-specific antibodies in wild-type and *Tdg*-deficient mouse embryonic stem cells (ESCs). In wild-type mouse ESCs, 5fC/5caC accumulates to detectable levels at major satellite repeats but not at nonrepetitive loci. In contrast, *Tdg* depletion in mouse ESCs causes marked accumulation of 5fC and 5caC at a large number of proximal and distal gene regulatory elements. Thus, these results reveal the genome-wide view of iterative 5mC oxidation dynamics and indicate that TET/TDG-dependent active DNA demethylation process occurs extensively in the mammalian genome.

INTRODUCTION

Epigenetic modifications of DNA and histones play essential roles in regulating gene expression in development and diseases (Goldberg et al., 2007; Jaenisch and Bird, 2003; Sasaki and Matsui, 2008). The predominant epigenetic modification of DNA is methylation at the fifth position of cytosine (5mC), which is indispensable for normal mammalian embryogenesis and is implicated in a variety of human diseases (Baylin and Jones, 2011; Cedar and Bergman, 2012). The DNA methylation pattern is established and maintained by DNA methyltransferases (DNMTs) and is relatively stable in somatic tissues (Bird, 2002; Jones, 2012). 5mC can be successively oxidized to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC) by the Ten eleven translocation (TET/Tet) family of Fe(II) and 2-oxoglutarate-dependent DNA dioxygenases (He et al., 2011; Ito et al., 2010; Ito et al., 2011; Tahiliani et al., 2009) (Figure S1A available online). Different Tet enzymes (Tet1–3) exhibit distinct expression patterns in vivo and functional analyses of *Tet*-deficient mice indicate that they play important roles in diverse biological processes, including zygotic epigenetic reprogramming, germ cell development, pluripotent stem cell differentiation, and myelopoiesis (Cimmino et al., 2011; Dawlaty et al., 2013; Dawlaty et al., 2011; Gu et al., 2011; Koh et al., 2011; Marcucci et al., 2010; Wu and Zhang, 2011a; Yamaguchi et al., 2012).

The study of biological roles of Tet enzymes has been facilitated by the development of methods to specifically enrich or label 5hmC, a relatively abundant 5mC oxidation derivative detected in many tissues (Globisch et al., 2010; Kriaucionis and Heintz, 2009; Münzel et al., 2010). Immunostaining with antibodies specific for 5hmC reveals that global erasure of paternal DNA methylation is first initiated by Tet3-mediated conversion of 5mC to 5hmC in the male pronucleus, followed by replication-dependent passive loss of 5hmC during preimplantation development (Gu et al., 2011; Inoue and Zhang, 2011; Iqbal et al., 2011; Wossidlo et al., 2011). Similar analysis also suggests a role of Tet1-mediated 5mC oxidation in epigenetic reprogramming during development of primordial germ cells (PGCs) and regulation of parental-origin-specific imprinting (Dawlaty et al., 2013; Hackett et al., 2013; Seisenberger et al., 2012; Yamaguchi et al., 2013; Yamaguchi et al., 2012). Genome-wide 5hmC mapping studies of pluripotent stem cells and differentiated tissues using affinity enrichment-based methods or modified bisulphite sequencing (BS-seq) indicate that 5hmC is enriched in highly transcribed gene bodies, as well as polycomb repression complex bound promoters and distal *cis*-regulatory elements

(Booth et al., 2012; Ficz et al., 2011; Mellén et al., 2012; Pastor et al., 2011; Song et al., 2011; Stroud et al., 2011; Szulwach et al., 2011a; Szulwach et al., 2011b; Williams et al., 2011; Wu et al., 2011a; Wu and Zhang, 2011b; Xu et al., 2011; Yu et al., 2012). Together, these studies not only confirm a functional role of Tet-mediated 5mC oxidation in regulating global DNA demethylation dynamics during specific embryonic stages (one-cell zygotes and developing PGCs) but also suggest that Tet-initiated DNA demethylation process may be more prevalent in the genome than previously anticipated.

In vitro biochemical studies show that DNA repair enzyme thymine-DNA glycosylase (TDG) can excise 5fC and 5caC to generate abasic sites (He et al., 2011; Maiti and Drohat, 2011; Nabel et al., 2012), which are repaired by base excision repair (BER) pathway. These observations suggest a mechanistic paradigm of active DNA demethylation in which Tet proteins first successively oxidize 5mC to 5hmC/5fC/5caC and TDG/BER pathways then excise 5fC/5caC and regenerate unmodified cytosines (Figure S1A). The demonstration that genetic inactivation of *Tdg* in mice causes embryonic lethality (Cortázar et al., 2011; Cortellino et al., 2011), raises the possibility that TET/TDG-mediated active DNA demethylation process may be widespread in mammalian genomes and play an essential role in developmental gene regulation. However, it is currently unclear to what extent and at which part of the genome TDG-dependent 5fC/5caC excision followed by BER contributes to dynamic changes of DNA methylation patterns in vivo.

To directly address this question, we generated genome-wide maps of 5mC and its oxidation derivatives (5hmC/5fC/5caC) in wild-type and *Tdg*-deficient mouse embryonic stem cells (ESCs). We reasoned that depletion of *Tdg* would block the DNA methylation/demethylation cycle and cause accumulation of 5fC and 5caC, which can mark genomic loci actively undergoing TET/TDG-dependent 5mC oxidation dynamics. Our results reveal that TET/TDG-mediated cyclic changes of cytosine modification states occurs at a large cohort of gene regulatory regions and suggest that active DNA demethylation takes place more extensively than previously thought in mammalian cells.

RESULTS

Enrichment of 5fC and 5caC from Genomic DNA by Cytosine Modification-Specific Antibodies

Genome-wide distribution of 5mC and 5hmC can be determined by affinity enrichment or bisulfite conversion-based methods (Song et al., 2012). However, reliable methods are yet to be developed to specifically enrich or label 5fC and 5caC for genome-wide mapping analysis. Antibody-based DNA immunoprecipitation followed by high throughput sequencing (DIP-seq) represents a simple and effective approach for profiling cytosine modifications (especially suitable for detecting regions with multiple modified bases) if a highly specific antibody is available. A strategy for chemical labeling of 5fC with aldehyde-reactive probe (ARP) has previously been suggested (Pfaffeneder et al., 2011), but this approach may also label abasic sites, which are an intermediate product of endogenous DNA repair process and one of the most prevalent lesions in DNA (Nakamura et al., 1998). Thus, proper controls or chemical blocking reactions need to be developed to allow ARP-based chemical-labeling methods to distinguish 5fC from abasic sites (Raiber et al., 2012). More recently, modified BS-seq strategies have been developed to map 5hmC distribution at single-nucleotide resolution (Booth et al., 2012; Yu et al., 2012). However, current base-resolution mapping methods are not compatible for detecting 5fC/5caC and require substantially deeper sequencing depth to reliably detect 5hmC marks (Yu et al., 2012). Given that 5fC/5caC is present in the genome at much lower levels compared to 5hmC, it will be challenging to map 5fC/5caC at a genome-wide scale and at base-resolution.

To better compare various approaches and identify effective methods for genome-wide mapping of 5fC/5caC, we first performed in-depth analysis comparing genome-wide 5hmC mapping results from antibody- or chemical-labeling-based (e.g., GLIB [glucosylation, periodate oxidation and biotinylation]) methods with the base-resolution 5hmC map in mouse ESCs (Pastor et al., 2011; Yu et al., 2012). Among 2.06 million high-confidence 5hmC marks of the base-resolution map, 21.3% (0.44 million) of them are sparsely distributed (single 5hmC mark within $\geq$ 1 kb). This analysis revealed that the performance of the 5hmC antibody-based method was similar to that of the chemical-labeling method (GLIB) in terms of pulling down both clustered (48.1% for 5hmC antibody and 43.1% for GLIB) and sparsely distributed (6.4% for 5hmC antibody versus 5.3% for GLIB) 5hmC marks from in vivo genomic DNA (Figures S1B–S1D). Thus, we focused our efforts on the antibody-based DIP method for detecting 5fC/5caC. The 5fC- and 5caC-specific antibodies were previously used to examine global levels of 5fC/5caC by immunostaining (Inoue et al., 2011). After further confirmation of their specificity by dot blot analysis (Figure 1A), we tested their utility in DIP assays. This analysis indicated that these antibodies could pull-down 5fC- or 5caC-containing oligonucleotides specifically and efficiently, suggesting that they are suitable for DIP assays (Figure 1B).

Quantitative mass spectrometry analysis indicates that 5fC and 5caC levels are approximately 2% or 0.5%, respectively, of the total level of 5hmC in wild-type mouse ESCs (Ito et al., 2011). Given that mouse ESCs possess high levels of Tet enzymatic activities, the relatively low abundance of 5fC/5caC suggests that 5fC and 5caC marks may be rapidly removed by TDG in vivo (He et al., 2011; Maiti and Drohat, 2011; Nabel et al., 2012). Thus, blocking TDG activity may result in accumulation of 5fC and 5caC, which allows the identification of genomic loci targeted by TDG activity. To test this possibility, we generated *Tdg*-deficient mouse ESCs by lentivirus-mediated knockdown (Figure 1C). Mass spectrometry analysis demonstrated that global levels of 5fC and 5caC increased by 5.6-fold and 8.4-fold, respectively, in response to *Tdg* knockdown (Figure 1D). In contrast, neither 5mC nor 5hmC showed significant change upon *Tdg* knockdown (Figure 1D). Consistent with previous results demonstrating that *Tdg* is not required for mouse ESC maintenance (Cortázar et al., 2011), neither the morphology nor the expression levels of pluripotent genes (*Oct4*, *Sox2*, and *Nanog*) or *Tet* genes were altered by *Tdg* knockdown (Figure S2).

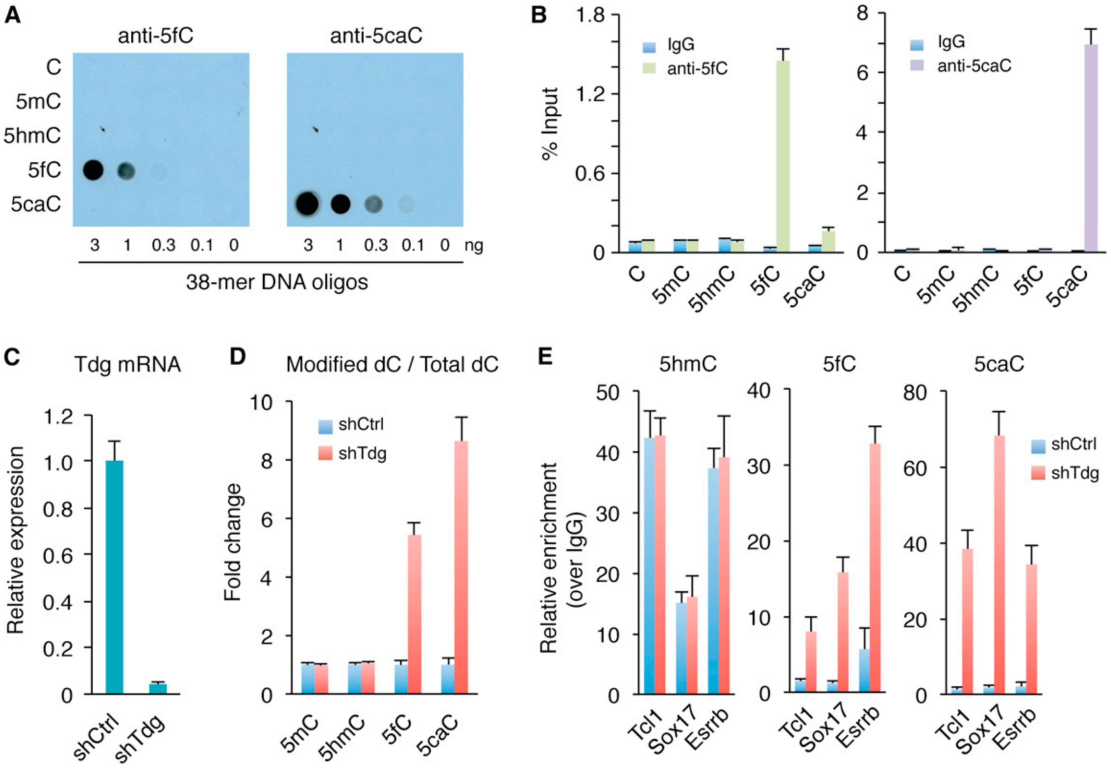

Figure 1. Enrichment of 5fC and 5caC from Genomic DNA by Modification-Specific Antibodies

(A) The 5fC and 5caC antibodies specifically recognize 5fC and 5caC-containing DNA oligos in dot blot assays, respectively. Different amounts of 38-mer DNA oligonucleotides (oligos), where the cytosines in 9 CpGs are either C, 5mC, 5hmC, 5fC, or 5caC, were spotted on membrane and probed with 5fC or 5caC antibodies.

(B) DIP-qPCR analysis demonstrates the specificity of the antibodies.

(C) RT-qPCR analysis of Tdg expression levels in control (shCtrl) and *Tdg*-deficient (shTdg) mouse ESCs.

(D) Mass spectrometric quantification of 5mC, 5hmC, 5fC, and 5caC in control and *Tdg*-deficient mouse ESCs.

(E) DIP-qPCR analysis of 5hmC/5fC/5caC at three 5hmC-enriched regions. Data are presented as mean ± SEM.

See also Figures S1 and S2.

We next tested 5fC and 5caC antibodies in immunoprecipitating genomic DNA fragments at three Tet1-bound and 5hmC-enriched regions (*Tcl1*, *Sox17*, and *Esrrb*) (Wu et al., 2011a; Wu et al., 2011b). Consistent with the fact that TDG does not excise 5hmC, 5hmC levels at these loci were not affected by *Tdg* knockdown (Figure 1E). In contrast, 5fC and 5caC levels at these loci were significantly increased in *Tdg*-deficient cells (Figure 1E). Given that significant 5fC- and 5caC-DIP signals are only detected in *Tdg* knockdown ESCs, we conclude that 5fC and 5caC antibodies are highly specific and are potentially suitable for genome-wide 5fC/5caC-DIP analysis.

Preferential Enrichment of 5fC/5caC at Pericentric Heterochromatin in Mouse ESCs

To determine genome-wide distribution of 5fC/5caC, we performed 5fC and 5caC DIP-seq experiments in replicates using genomic DNA of wild-type and *Tdg*-deficient mouse ESCs (Figure S1E and Table S1). We also performed 5mC, 5hmC, and mock IgG DIP-seq experiments using the same batches of genomic DNA (Figure S1E). Sequencing reads mapped to multiple genomic regions (multihit reads) generally represent repetitive sequences in the genome. Indeed, 89%–98% of multihit reads were found to be overlapped with the UCSC Repeat-Masker (RMSK) track (Dreszer et al., 2012), whereas only 20%–31% of the uniquely mapped reads overlapped with RMSK (Figure 2A). To evaluate the potential enrichment of 5fC/5caC at repeats, multihit reads were retained in the initial analysis. In wild-type mouse ESCs, 48% 5fC reads and 41% 5caC reads are multihit reads, which are higher percentages than that of 5hmC (25%) but lower than that of 5mC (70%). Thus, these results not only confirm that 5mC and 5hmC are relatively enriched and depleted from repetitive sequences, respectively (Ficz et al., 2011; Williams et al., 2011; Yoder et al., 1997), but also suggest that 5fC and 5caC may accumulate to detectable levels at repetitive sequences in wild-type mouse ESCs.

To determine the types of repeats at which 5fC and 5caC are relatively enriched, we classified all sequencing reads on the basis of RMSK annotation and calculated the number of reads in each repeat class. After correcting the relative percentage of various classes of repeats, 5mC, 5fC, and, to a lesser extent, 5caC are found to be preferentially enriched at major satellite repeats, whereas 5hmC tends to accumulate at short interspersed

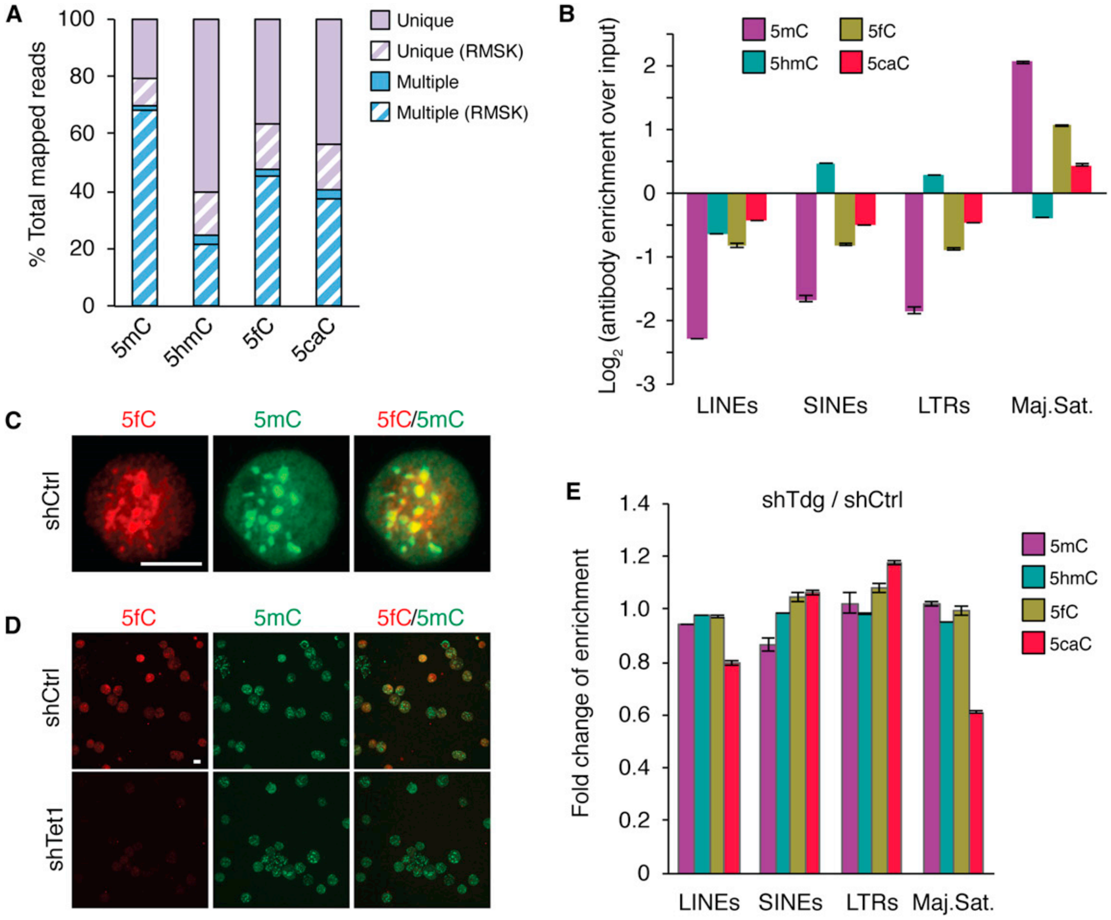

Figure 2. Preferential Accumulation of 5fC and 5caC at Major Satellite Repeats in Wild-Type Mouse ESCs

(A) Percentages of uniquely mapped and multihit reads in total mapped reads. Reads that overlap with the UCSC RepeatMasker (RMSK) track are highlighted by forward slash.

(B) Relative enrichment (log2 ratio of IP over input) for each cytosine modification at major classes of repetitive sequences in mouse ESCs. Values represent means of two biological replicates with ends of error bars corresponding to individual data points.

(C and D) Representative images of mouse ESC surface spreads costained with 5fC and 5mC antibodies. The same exposure time was used for comparing control (shCtrl) and *Tet1* knockdown (shTet1) mouse ESCs in (D). Scale bar, 100 μm.

(E) Bar graph presentation of the fold change of the enrichment in each class of repetitive sequences upon *Tdg* knockdown.

See also Table S1.

nuclear elements (SINEs) and long-terminal repeats (LTRs) (Figures 2B). Furthermore, costaining of surface spreads of mouse ESCs with 5fC and 5mC antibodies revealed significant overlap of 5fC signals with 5mC (Figure 2C), which is known to be enriched at pericentric heterochromatin (i.e., major satellite repeats). Immunostaining analysis further showed marked reduction of 5fC signals in the *Tet1* knockdown cells (Figure 2D), validating the specificity of 5fC signals at pericentric regions. *Tdg* knockdown does not significantly alter the relative enrichment ratio of 5fC at major satellite repeats, but it significantly reduced that of 5caC at the major satellite repeats (Figure 2E), probably due to the increase of 5caC at other genomic regions. Collectively, these results indicate that 5fC (and to a lesser extent 5caC) tends to accumulate at major satellite repeats in wild-type mouse ESC, and TDG may not efficiently excise 5fC/5caC at pericentric heterochromatin.

Tdg-Depletion Induces Accumulation of 5fC and 5caC in Nonrepetitive Regions

To further investigate *Tdg*-deficiency-induced changes of 5fC/5caC signals at nonrepetitive regions, we analyzed uniquely mapped reads. To identify genomic loci enriched for high-confidence 5fC/5caC signals, we first identified peak candidates using input genomic DNA as a negative control and then quantitatively filtered out peaks with relatively high levels of signals in IgG controls. In wild-type mouse ESCs, we identified 1,673 regions enriched for 5caC (Figure 3A and Table S2). Upon *Tdg* knockdown, a marked increase in the number of 5caC peaks (n = 89,503) was observed (Figures 3A and 3B and Table S3). Many newly appeared 5caC peaks colocalize with 5hmC peaks, which are largely unaffected by *Tdg* knockdown (Figure 3A and 3B). *Tdg* depletion also leads to a less pronounced increase in the number of 5fC peaks (Figure 3A and 3B and Tables S4 and

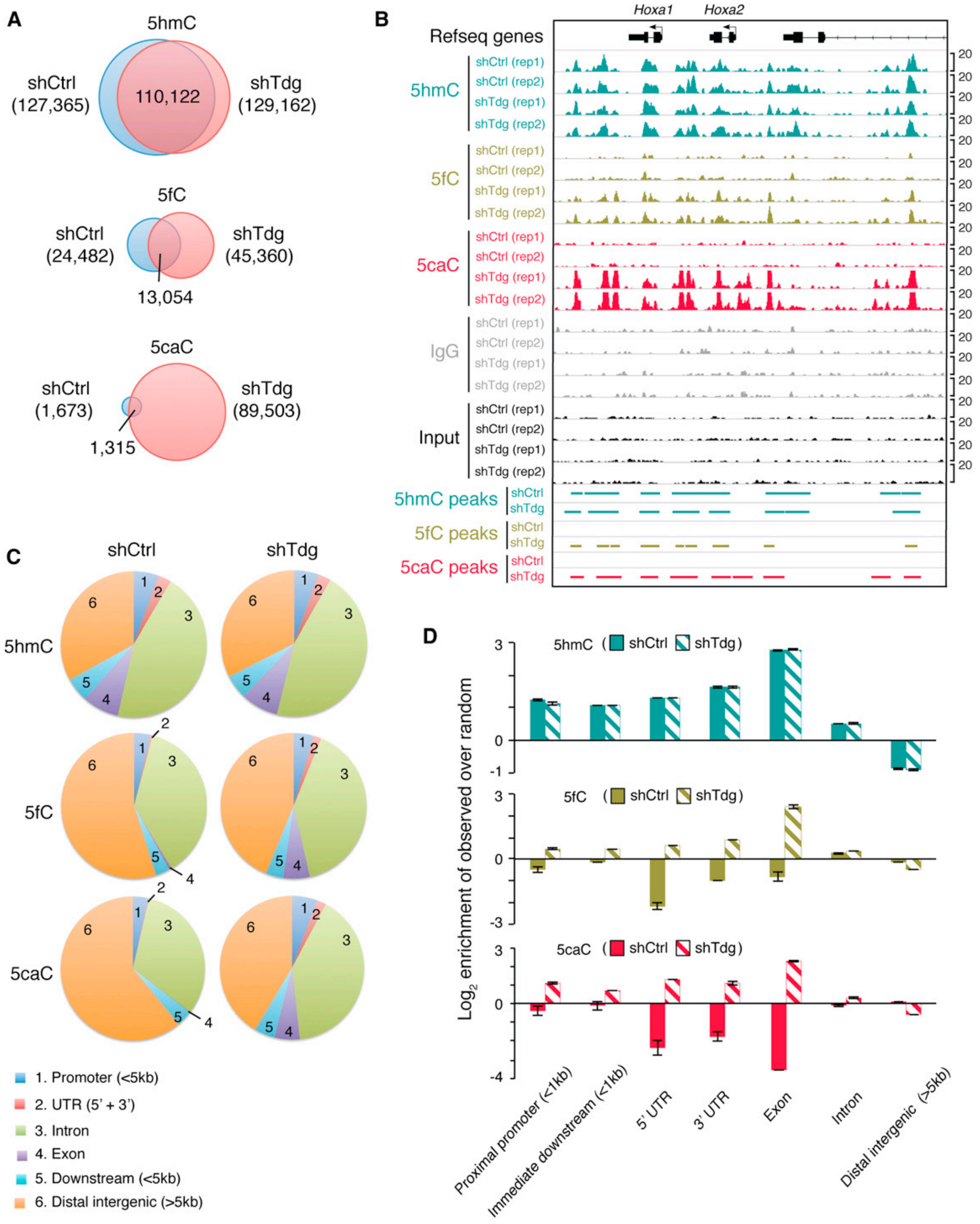

Figure 3. Accumulation of 5fC and 5caC in Nonrepetitive Regions in *Tdg*-Deficient Mouse ESCs

(A) Venn diagrams showing the overlap of 5hmC, 5fC, or 5caC peaks in control and *Tdg*-deficient mouse ESCs.

(B) Representative genomic loci (*Hoxa1* and *Hoxa2*) showing 5hmC/5fC/5caC peaks in control and *Tdg*-deficient mouse ESCs.

(C) Pie chart presentation of the overall genomic distribution of 5hmC/5fC/5caC-enriched regions.

(D) Enrichment (log2 ratios of observed over random) of 5hmC/5fC/5caC in control and *Tdg*-deficient cells at various genomic features. Values represent means of two biological replicates with ends of error bars corresponding to individual data points.

See also Figures S1 and S3 and Tables S2, S3, S4, and S5.

S5). Notably, significantly more 5fC peaks (n = 24,482) are detected in wild-type mouse ESCs relative to 5caC peaks (n = 1,673) (Figure 3A), which is consistent with previous findings that 5fC is more abundant than 5caC in mouse ESCs (Ito et al., 2011; Pfaffeneder et al., 2011). Importantly, the observation that a large number of 5fC/5caC peaks are specifically detected in *Tdg*-knockdown cells further validates the specificity of the 5fC/5caC DIP-seq method.

Next, we sought to determine where ectopic 5fC and 5caC peaks are preferentially located in the genome. Compared to random control regions, the relative percentage of 5fC and 5caC peaks in genic regions (promoters, group 1; introns, group 3; and exons, group 4 in Figure 3C) is increased in *Tdg*-deficient cells compared to that in control cells (Figures 3D and S3A). Moreover, 5fC and 5caC signals are increased preferentially within exons (solid lines in Figure S3B) relative to introns (dash lines in Figure S3B) in response to Tdg knockdown. This finding is consistent with previous studies demonstrating the enrichment of 5hmC at exon/intron boundaries (Khare et al., 2012), suggesting that a potential role of TET/TDG-mediated generation and excision of 5fC/5caC in regulating transcriptional elongation and/or splicing.

5fC and 5caC Exhibit Common and Unique Distributions

Both Tet1 and Tet2 are highly expressed in mouse ESCs (Ito et al., 2010; Koh et al., 2011), and 5fC and 5caC levels are significantly reduced upon *Tet1* depletion (Ito et al., 2011). To test whether Tet1 occupancy correlates with 5fC/5caC generation in vivo, we examined 5fC and 5caC signals at regions enriched for Tet1. In *Tdg*-deficient cells, Tet1 bound regions with medium-to-low CpG density are preferentially enriched for 5fC and 5caC signals (Figure S3C). In contrast, Tet1 bound regions with high CpG density tend to be depleted of 5fC/5caC in both control and *Tdg*-deficient mouse ESCs (Figure S3C), which is in agreement with the finding that CpG-rich regions are generally depleted of 5mC and 5hmC (Szulwach et al., 2011a; Wu et al., 2011a; Yu et al., 2012). Notably, in *Tdg*-deficient mouse ESCs, 5fC peaks tend to colocalize with 5mC-enriched regions, while 5caC peaks preferentially overlap with 5hmC-enriched regions (Figures S3D and S3E). Thus, four cytosine modifications have both common and unique distributions in the genome, suggesting that processivity of Tet proteins may be regulated by local sequence context and/or chromatin structure.

TDG-Mediated 5fC/5caC Excision Occurs Extensively at Distal Regulatory Elements

Further analysis of 5fC and 5caC peaks in *Tdg*-deficient mouse ESCs indicates that the majority of *Tdg*-depletion-induced 5caC peaks (76.3% of all 5caC peaks, n = 68,326) are located outside promoter or exonic regions. To investigate whether these distal regions are of functional relevance, we calculated the sequence conservation scores within these ectopic 5fC/5caC peaks. As expected, peaks overlapping with exons and proximal promoters show strong evolutionary conservation (Figure S4A). Interestingly, sequences overlapping with *Tdg*-depletion-induced nonexonic and nonpromoter 5caC peaks (to a less extent for ectopic 5fC peaks) are also relatively conserved compared to flanking regions (Figure 4A). Furthermore, 5caC peaks in *Tdg*-deficient mouse ESCs frequently overlap with low-methylated regions (LMRs) (Figure 4B), a unique group of genomic regions that display features of distal regulatory regions and are generally associated with intermediate (~30%, measured by BS-seq) DNA methylation level (Stadler et al., 2011). These results suggest that both 5mC oxidation and 5fC/5caC excision activity may be preferentially recruited to a large cohort of distal regulatory elements.

To further characterize regions where TDG actively excises 5fC/5caC, we calculated the averaged signals of 5mC oxidation derivatives within genomic features derived from published genome-wide mapping data sets for a number of DNA-binding factors and major histone modifications (Figure 4C and S4B). This analysis indicates that in *Tdg*-deficient mouse ESCs, 5caC accumulates at binding sites of pluripotency transcription factors (TFs) such as Oct4, Nanog, Sox2, and Esrrb (Figure 4C, yellow color group) (Chen et al., 2008; Marson et al., 2008). For instance, 5caC peaks in *Tdg*-deficient cells show a significant overlap (observed/expected = 16.0, $p < 2.2 \times 10^{-16}$, Fisher's exact test) with Oct4-binding regions (±100 bp flanking peak summits) compared to that expected by chance. Ectopic 5caC signals are also preferentially detected at peaks of factors (e.g., p300) and histone marks (e.g., H3K4me1 and H3K27ac) that are associated with active/poised enhancers (Figure 4C, yellow color group) (Creyghton et al., 2010; Shen et al., 2012). Notably, the presence of *Tdg*-depletion-induced 5caC signals at these distal elements is not simply due to a higher level of 5hmC because Smad1-binding sites are not enriched for high levels of 5hmC but are frequently associated with ectopic 5caC signals (Figure 4C). Furthermore, ectopic 5caC signals are frequently detected at binding sites of the cohesion complex and mediator proteins, both of which are implicated in regulating interactions between promoter and enhancers (Figure 4C, pink color group) (Kagey et al., 2010). Many regions enriched for repressor complexes (e.g., LSD1, Hdac1/2) that are involved in decommissioning active ESC enhancers during differentiation also frequently overlap with ectopic 5caC peaks (Figure 4C, red color group) (Whyte et al., 2012). By contrast, regions associated with basal transcriptional machineries (e.g., RNA Pol2, TBP in green color group), insulators (CTCF in pink color group), and topological domain boundaries (in pink color group) are not enriched for ectopic 5caC peaks (Figure 4C) (Dixon et al., 2012; Kagey et al., 2010; Rahl et al., 2010). For most features analyzed in Figure 4C, distally located elements are preferentially associated with *Tdg*-depletion-induced 5fC/5caC signals relative to proximally located ones (within ±1 kb regions flanking transcriptional start sites [TSSs]) (Figures 4C and S4B), suggesting that TET/TDG may be more active at or preferentially recruited to regions outside proximal promoters. Taken together, these results indicate that TET/TDG-mediated 5mC oxidation and 5fC/5caC excision actively occur at a large cohort of distal *cis*-regulatory elements.

TDG-Mediated 5fC/5caC Excision Occurs Preferentially at Active Enhancers in Mouse ESCs

To further analyze distal *cis*-elements targeted by TDG in mouse ESCs, we calculated averaged signals of 5mC oxidation

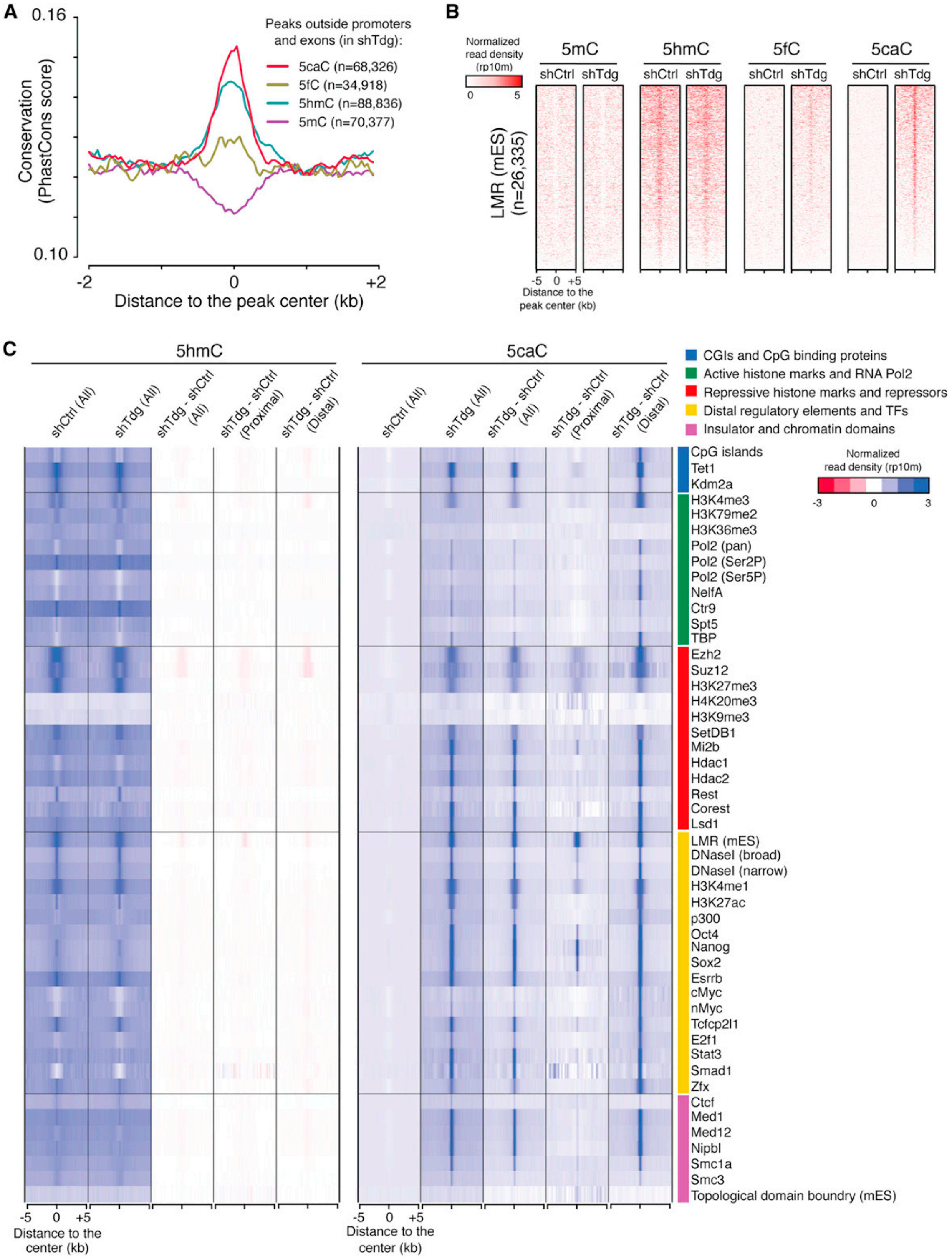

Figure 4. ***Tdg*****-Depletion-Induced Ectopic 5fC and 5caC Accumulate at Distal Regulatory Regions**

(A) Average conservation (phastCons) scores within regions flanking the center of 5mC/5hmC/5fC/5caC peaks (nonoverlapping with exons or promoters) in *Tdg*-deficient mouse ESCs. The number of peaks that are located outside exons and proximal promoters (±1 kb flanking TSSs) for each cytosine modification is also shown.

(B) Heatmaps of 5mC/5hmC/5fC/5caC levels (normalized read density) at LMRs identified in mouse ESCs. The heatmaps are ranked by the mean of 5caC signals in *Tdg*-deficient cells (top, highest; bottom, lowest).

(legend continued on next page)

derivatives in wild-type and *Tdg*-deficient mouse ESCs at tissue-specific enhancers identified by mouse ENCODE project (Shen et al., 2012). Interestingly, enhancers that are specifically active in mouse ESCs are associated with the highest level of ectopic 5caC signals (Figure 5A). These mouse ESC-specific enhancers are also preferentially bound by Tet1 and overlapped with DNase I hypersensitivity sites (Figure S5A), suggesting that cytosines within or surrounding active enhancer regions tend to undergo TET/TDG-mediated 5mC oxidation dynamics. Similarly, mouse ESC-specific LMRs are associated with higher levels of *Tdg*-depletion-induced 5fC/5caC signals than are neural progenitor (NP)-specific LMRs (Stadler et al., 2011) (Figure S5B). In addition, analysis comparing binding sites of pluripotency TFs and neuronal TFs indicates that pluripotency TF-bound regions in mouse ESCs are preferentially marked by *Tdg*-depletion-induced 5fC/5caC signals (Kim et al., 2010; Marson et al., 2008) (Figures 5B and S5C). As exemplified in Figures 5C and S5D, a large cohort of distal regions bound by pluripotency TFs (Oct4/Sox2/Nanog) are associated with ectopic 5caC peaks in *Tdg*-deficient mouse ESCs regardless of the presence of stable Tet1 occupancy. Together, these results suggest that TET/TDG-mediated 5mC oxidation dynamics in mouse ESCs may contribute to the regulation of active enhancer activity.

TDG-Mediated 5fC/5caC Excision Occurs Preferentially at Transcriptionally Inactive Gene Promoters in Mouse ESCs

Although only a small portion of ectopic 5fC/5caC peaks overlap with proximal promoters, we frequently observed 5caC accumulation at regions flanking gene promoters or 3′ gene bodies of transcribed genes in *Tdg*-deficient mouse ESCs. Previous studies suggest that distinct genic regions are associated with specific histone lysine methylation patterns (Barski et al., 2007; Bernstein et al., 2006; Mikkelsen et al., 2007; Whyte et al., 2012), which may in turn contribute to regulation of gene expression (Figure S6A).

To explore the possibility that distinct chromatin states may influence the generation and excision of 5fC/5caC by TET/TDG, we compared average signal profiles of 5mC/5hmC/5fC/5caC at four groups of extended gene promoters (±5 kb relative to TSSs): (1) "Active," characterized by the presence of high levels of H3K4me3 (active histone mark) at proximal promoters and H3K79me2/3 (indicative of elongation) at 5′ of gene bodies; (2) "Initiated," only associated with promoter H3K4me3; (3) "Bivalent," associated with both H3K27me3 (repressive histone mark) and medium-to-low levels of promoter H3K4me3; (4) "Silent," lack of promoter H3K4me3. This analysis revealed that 5fC/5caC levels are largely comparable between control and *Tdg*-deficient mouse ESCs at gene promoters that are associated with active transcription (green in Figure 6A, exemplified by *Rest* in Figure 6B and P2 promoter of *Dnmt1* in Figure S6B) or transcription initiation (gray in Figure 6A). These observations suggest that TET/TDG-mediated 5mC oxidation dynamics is generally absent at these transcriptionally active/permissive promoters. By contrast, a substantial increase of 5fC/5cac levels was detected at genomic regions flanking bivalent promoters in the absence of Tdg (exemplified by *Tbx5* in Figure 6B and *HoxA cluster* in Figure S6B). Considering that 5hmC is also enriched at bivalent domains (Figure 6A), these results suggest that bivalent domains are targeted by relatively high levels of TET/TDG activities in mouse ESCs. Silent promoters (blue in Figure 6A) were also associated with relatively high levels of ectopic 5fC/5caC signals (exemplified by *Spink2* in Figure 6B and P1 promoter of *Dnmt1* in Figure S6B). Because 5mC is also enriched at silent gene promoters (Figure 6A), it seems that silent gene promoters in mouse ESCs are simultaneously targeted by activities of DNMT/TET/TDG. Collectively, these results indicate that transcriptionally inactive (silent or bivalent/poised) gene promoters are preferentially regulated by TET/TDG activity and tend to undergo active DNA demethylation in mouse ESCs.

H3K27me3 within bivalent domains are deposited by polycomb repression complex 2 (PRC2) (Cao et al., 2002), and PRC2 binding to chromatin is antagonized by the presence of 5mC (Bartke et al., 2010; Wu et al., 2010). To further evaluate the relationship between PRC2 binding and TET/TDG-mediated 5mC oxidation dynamics, we examined ectopic 5fC/5caC levels within regions enriched for two core PRC2 subunits, Ezh2, and Suz12 (Ku et al., 2008). In *Tdg*-deficient cells, 5fC/5caC accumulates to a significant level within Ezh2- and Suz12-bound regions (Figures 6C, 6D and S6C), suggesting a potential role of TET/TDG proteins in regulating PRC2 activity or targeting.

TDG-Mediated 5fC/5caC Excision and Gene Expression

Next, we examined the relationship between 5fC and 5caC distribution and the global gene expression profile. Using published RNA-seq data sets of wild-type mouse ESCs (Ficz et al., 2011), ectopic 5fC/5caC signals are found to be depleted in the promoters of highly expressed genes but were relatively enriched in the intragenic regions (especially 3′ end) of highly and moderately expressed genes (Figure 7A). In support of the notion that silent or repressed/poised promoters tend to be targeted by TET/TDG activity, promoters of genes with low-to-medium expression levels are enriched for ectopic 5fC/5caC signals (Figure 7A and S7A). Thus, TET/TDG-dependent cytosine modification dynamics may play a complex role in transcriptional regulation, depending on their genomic location.

To further study the potential role of cytosine modification cycling in regulating gene expression, we performed microarray analysis comparing gene expression in control and *Tdg*-deficient mouse ESCs. Consistent with the grossly

(C) Heatmaps of 5hmC and 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 regulatory regions (derived from published data sets in mouse ESCs). The difference in 5hmC and 5caC levels between control and *Tdg*-deficient cells (shTdg minus shCtrl) is also shown for all, proximal (overlapping with ±1 kb flanking TSSs), and distal features.
See also Figure S4.

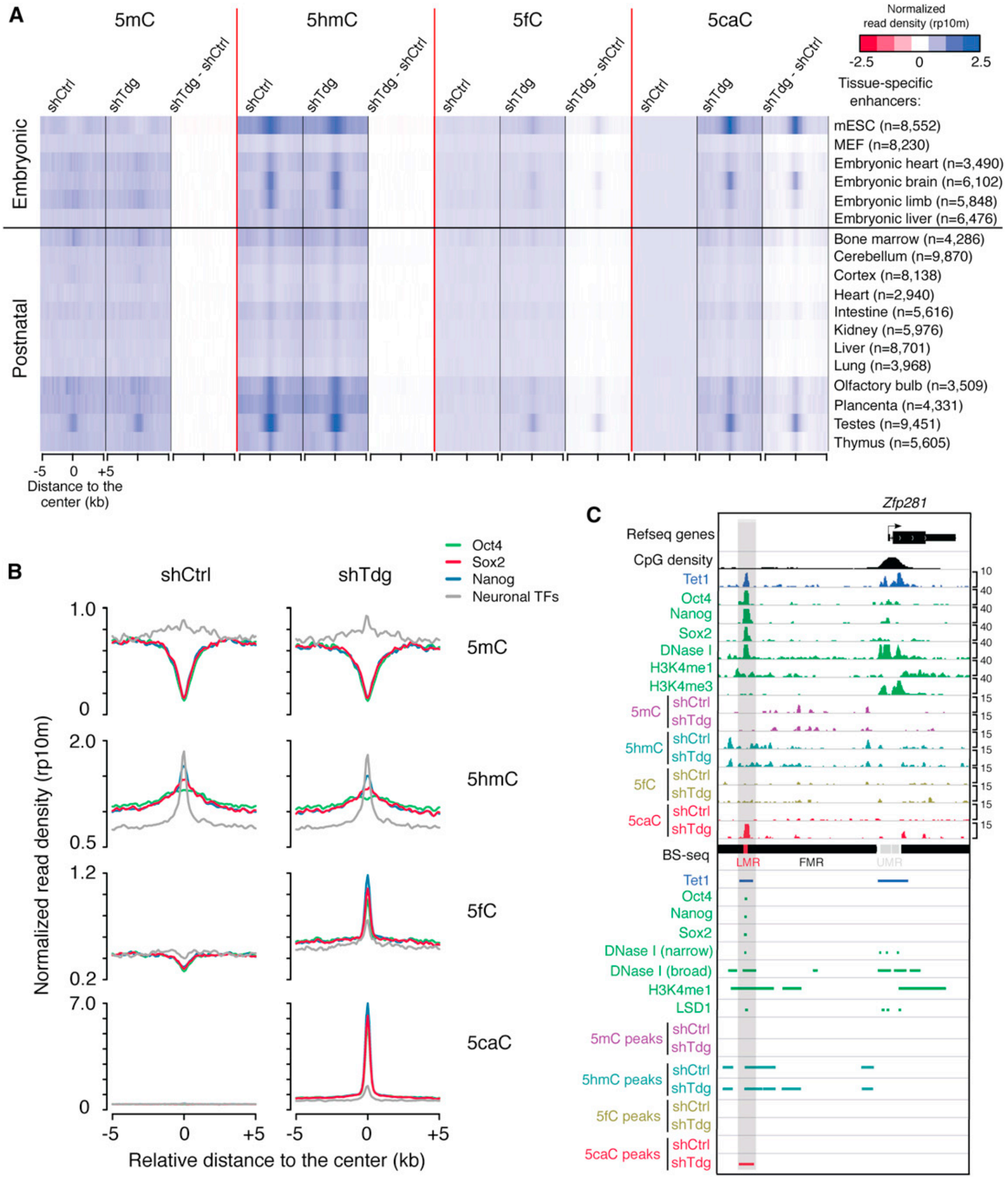

Figure 5. TET/TDG Activities Are Preferentially Recruited to Active Enhancers and Distal Pluripotency TF-Binding Sites in Mouse ESCs

(A) Heatmaps of 5mC/5hmC/5fC/5caC levels (normalized read density) in control and *Tdg*-deficient mouse ESCs at tissue-specific enhancers.

(B) Average 5mC/5hmC/5fC/5caC signals in control and *Tdg*-deficient mouse ESCs at the center of binding sites for pluripotency TFs (Oct4/Nanog/Sox2) and neuronal TFs.

(C) Shown are 5mC/5hmC/5fC/5caC distributions in control and *Tdg*-deficient mouse ESCs at a representative locus (upstream of the *Zfp281* gene) bound by pluripotency TFs (Oct4/Nanog/Sox2). Other genomic features (e.g., DNase I hypersensitivity sites, H3K4me1-enriched regions, and enhancer-related epigenetic regulator LSD1) are also shown. Fully methylated regions (FMRs, in black), low methylated regions (LMRs, in red), and unmethylated regions (UMRs, in gray) were derived from previously published BS-seq data sets (Stadler et al., 2011).

See also Figure S5.

normal phenotype of *Tdg*-deficient mouse ESCs (Figure S2), gene expression changes upon *Tdg* knockdown are minor, with only 99 genes showing relatively marked expression change ($p < 0.01$ and fold change > 1.5). More genes (n = 1,192) exhibited relatively subtle change in expression ($p < 0.01$) in response to *Tdg* depletion (Figure 7B). We then compared

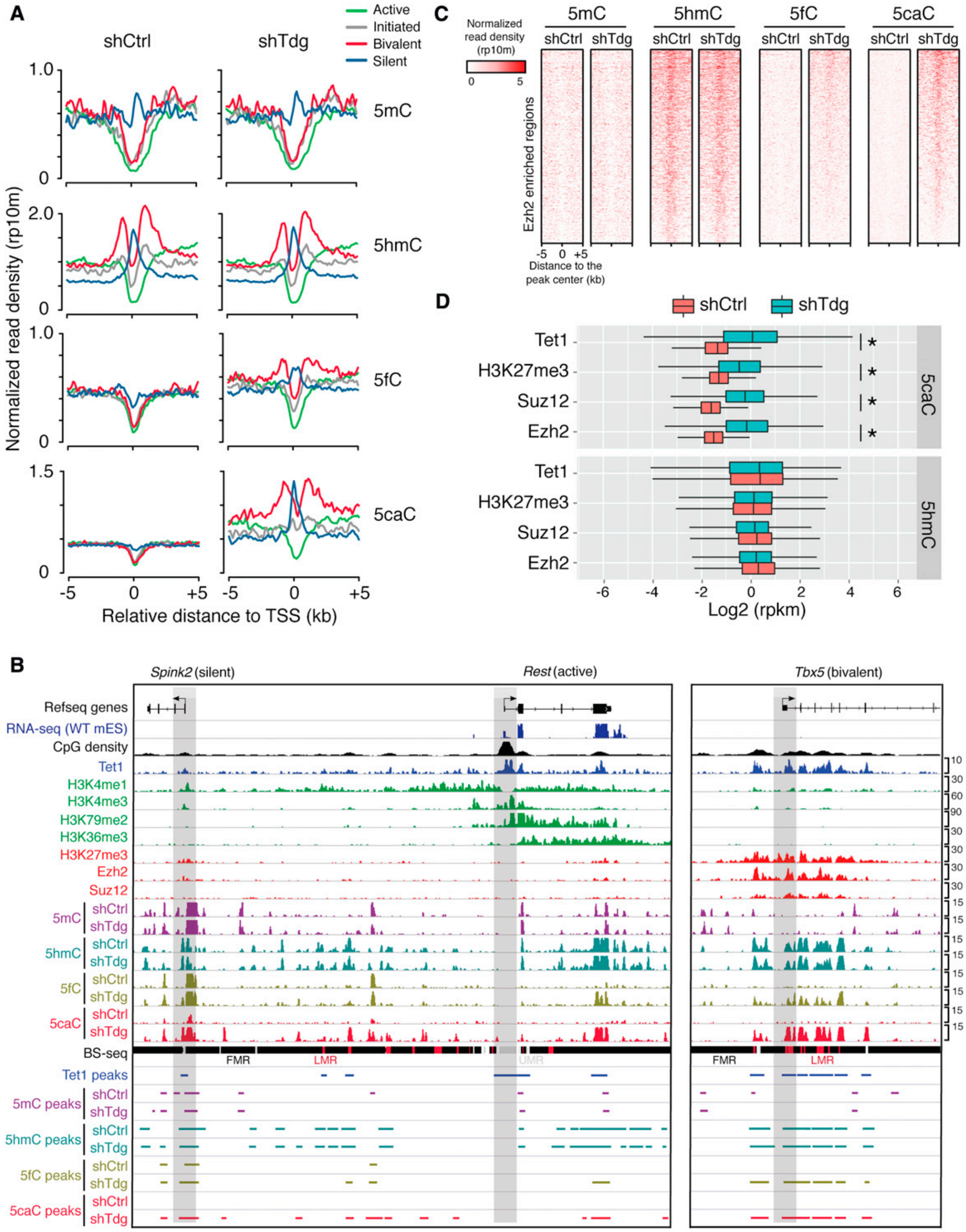

Figure 6. *Tdg*-Depletion-Induced 5fC/5caC Signals Are Enriched at Bivalent and Transcriptionally Silent Gene Promoters in Mouse ESCs

(A) Average 5mC/5hmC/5fC/5caC signals in control and *Tdg*-deficient mouse ESCs at TSSs of four groups of gene promoters that are associated with distinct chromatin states (active, H3K4me3+/H3K79me2+; initiated, H3K4me3+ only; bivalent, H3K4me3+/H3K27me3+; silent, none).

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

(C) Heatmaps of 5mC/5hmC/5fC/5caC levels (normalized read density) in control and *Tdg*-deficient mouse ESCs at centers of Ezh2-binding sites.

(D) Box-plots of normalized 5hmC and 5caC levels (read per million reads and kilo bases [rpkm]) in control and Tdg-deficient cells within genomic regions enriched for Tet1, Ezh2, Suz12 and H3K27me3. *$p < 2.2 \times 10^{-16}$ (p values were calculated by two-tailed t test).

See also Figure S6.

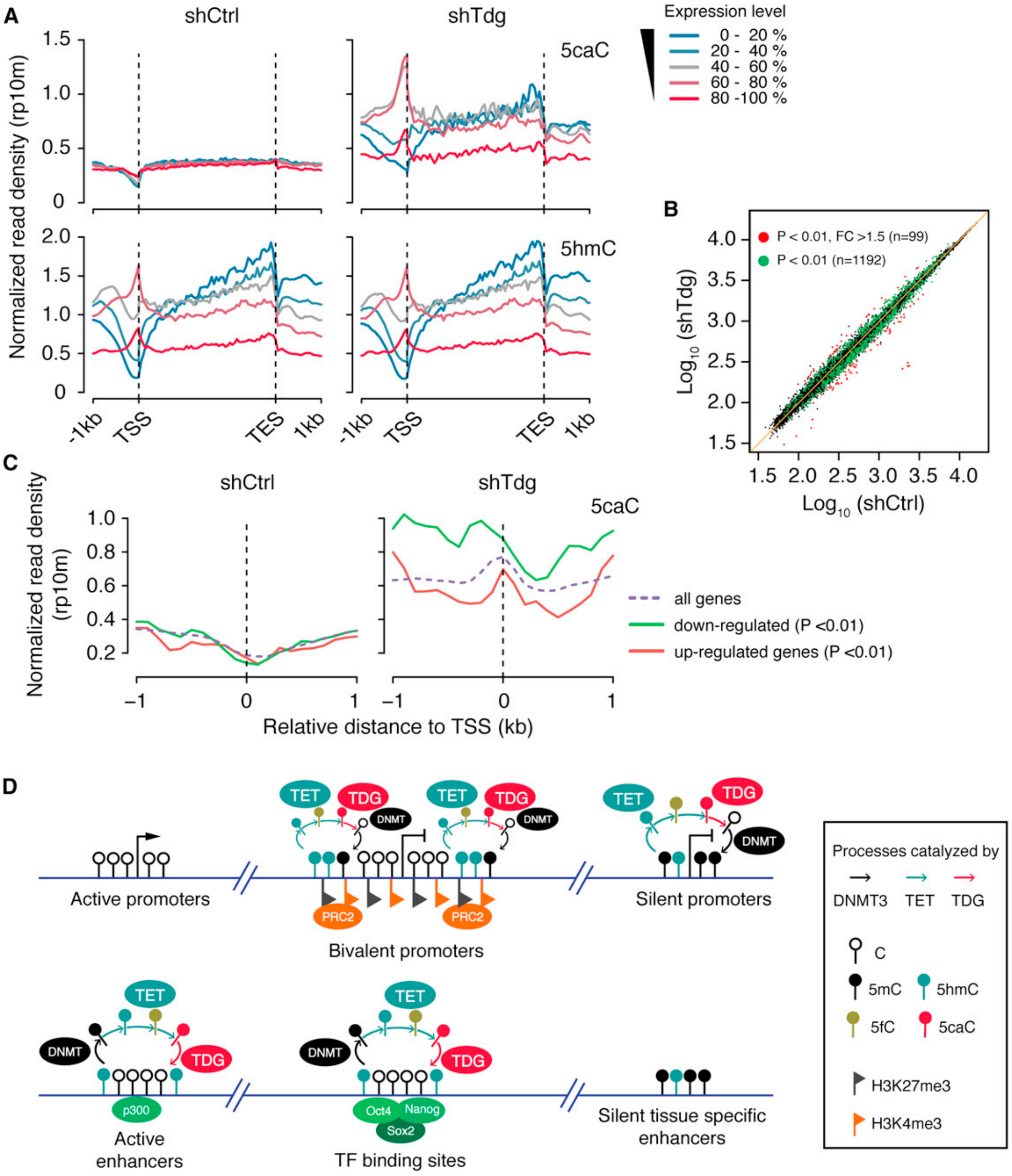

Figure 7. Complex Relationship between Gene Expression and TET/TDG-Mediated 5mC Oxidation Dynamics

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

(B) Scatter plots comparing gene expression profiles of control and *Tdg*-deficient mouse ESCs. Green and red dots indicate differentially expressed genes at $p < 0.01$ and $p < 0.01$, $FC > 1.5$, respectively.

(C) Average 5caC signals in control (left) and *Tdg*-deficient (right) mouse ESCs at the TSS of downregulated and upregulated genes ($p < 0.01$).

(D) Schematic diagram illustrating the relationship between transcriptional activity and DNMT/TET/TDG-mediated cytosine modification cycling at promoters and distal regulatory regions in mouse ESCs. Dynamic cyclic changes of cytosine modifications preferentially occur within bivalent and silent promoters, as well as active enhancers and pluripotency TF-binding sites.

See also Figure S7.

average signals for all cytosine modifications at the TSS of upregulated (n = 413) and downregulated genes (n = 636). This analysis indicated that 5fC/5caC signals at proximal promoters of downregulated genes tend to increase more dramatically when compared to those of upregulated genes in response to *Tdg* depletion (Figures 7C and S7B), suggesting a transcriptional inhibitory role of 5fC/5caC at proximal promoters.

DISCUSSION

TET/TDG-Mediated 5mC Removal Occurs Extensively in the Mammalian Genome

In conjunction with DNMTs, the step-wise process of active DNA demethylation entailed by TET/TDG/BER in principle permits cyclic changes of modification state at all cytosine bases (predominantly in the context of CpG) in the mammalian genome. In contrast to the readily detectable 5hmC, 5fC and 5caC are present at much lower levels in mammalian cells. Several nonmutually exclusive mechanisms may be responsible for the observed scarcity of 5fC/5caC. First, oxidation of 5hmC to 5fC/5caC by TET proteins may be tightly regulated and is less efficient compared to conversion of 5mC to 5hmC. Second, 5hmC can be passively removed by replication-dependent dilution in proliferating cells or converted to other form of modifications (e.g., 5-hydroxymethyluracil or direct conversion of 5hmC to C) before 5hmC is further oxidized by Tet proteins (Chen et al., 2012; Guo et al., 2011; Inoue and Zhang, 2011). Third, TDG-mediated excision of 5fC/5caC is highly efficient, thus 5fC and 5caC are short-lived (Globisch et al., 2010; Ito et al., 2011). To better understand the relative contribution of TET/TDG-mediated active demethylation pathway (5mC oxidation and excision of 5fC/5caC) to DNA methylation dynamics, we applied antibody-based DIP-seq analysis to mouse ESCs and generated genome-wide maps of all 5mC oxidation derivatives. Comparative analysis of 5hmC/5fC/5caC distributions in wild-type and *Tdg*-deficient mouse ESCs has revealed that a large number of genomic loci are targeted by TET/TDG activities, suggesting that relatively large-scale 5mC oxidation dynamics may not be a unique feature for developing zygotes and PGCs, but rather a prevalent event that may takes place in the genome of diverse cell types. Because mouse ESCs are highly proliferative and 5hmC/5fC/5caC can also be removed by the replication-dependent dilution mechanism, future studies of TET/TDG activity in terminally differentiated and postmitotic cells will facilitate understanding of the functional role of TET/TDG-dependent active DNA demethylation pathway in gene regulation and development.

Steady-State Accumulation of 5fC and 5caC at Specific Class of Repetitive Sequences

We observed that, much like 5mC, on a population average, 5fC and 5caC (to a lesser extent) are relatively enriched at repetitive sequences, particularly at major satellite repeats. The accumulation of 5fC and potentially 5caC in major satellite repeats can be explained, at least in part, by two nonmutually exclusive mechanisms: (1) TET proteins tend to oxidize their substrates with a higher processivity within major satellite repeats due to the unique sequence context (e.g., CpG density), local chromatin states, or DNMT activity; and (2) Tdg is less efficient in removing 5fC/5caC within major satellite repeats (located in pericentric heterochromatin) relative to other genomic regions. In support of the second possibility, previous studies have shown that TDG is unable to associate with heterochromatized promoters (Cortázar et al., 2011).

TET/TDG-Mediated 5mC Oxidation Dynamics at Bivalent and Transcriptionally Silent Gene Promoters

Tet1 tend to be enriched at CpG-rich gene promoters through its CXXC domain (Tahiliani et al., 2009; Xu et al., 2011). However, both affinity enrichment-based and base-resolution mapping (e.g., TAB-seq) analyses of 5hmC distribution indicate that 5hmC tends to be enriched at promoters with medium-to-low levels of CpG density but depleted from CpG-rich promoters (Wu et al., 2011a; Yu et al., 2012). The discrepancy between Tet1 occupancy and the 5hmC level suggests that at CpG-rich promoters, either 5hmC is not efficiently generated by Tet1 due to lack of 5mC or 5hmC is rapidly oxidized to 5fC/5caC followed by TDG-mediated 5fC/5caC excision. The fact that, in *Tdg*-deficient cells, 5fC/5caC are not accumulated at CpG-rich, actively transcribed gene promoters suggests that these CpG-rich, Tet1-bound promoters are generally not associated with the active demethylation process. In contrast, a marked increase in 5fC/5caC levels is detected at Tet1-bound bivalent domains flanking transcriptionally repressed/poised promoters (Figure 7D, upper). These bivalent promoters generally encode developmental regulators and lineage-specific transcription factors, suggesting that TET/TDG-mediated active demethylation process may be required to maintain a transcriptionally poised state at these promoters. Interestingly, previous studies have suggested that bivalent promoters show a tendency of being DNA methylated in cancer cells (Baylin and Jones, 2011), so the dysregulation of active demethylation process in tumors may contribute to the observed hypermethylation status at bivalent promoters.

TET/TDG-Mediated 5mC Oxidation Dynamics at Distal-Regulatory Regions

The ability to determine the genome-wide distribution of all 5mC oxidation derivatives offered a unique opportunity to assess the TET/TDG-mediated 5mC oxidation dynamics at various genomic features and regulatory elements. Unlike widespread distribution of 5mC and 5hmC, 5fC and 5caC in wild-type mouse ESCs are hardly detectable at nonrepetitive regions. Upon depletion of *Tdg*, many ectopic 5fC and 5caC peaks appeared at distal, but not proximal, regulatory elements. This observation agrees with recent findings from base-resolution mapping of 5mC and 5hmC in the mouse ESCs (Stadler et al., 2011; Yu et al., 2012) and suggests that TET/TDG-mediated active DNA demethylation may occur extensively at a large cohort of distal regulatory regions (Figure 7D, lower). Further studies are warranted to elucidate the function of cyclic change of cytosine modifications at distal regulatory elements. Moreover, developing base-resolution mapping methods of 5fC and 5caC (e.g., modified BS-seq strategies) is needed to overcome the resolution limit of affinity enrichment method used in the current study and may provide insights into the mechanism by which TET/TDG selectively regulate a subset of cytosines within or near distal *cis*-regulatory elements.

In summary, we have developed an affinity enrichment-based approach to determine genome-wide distribution of 5fC and 5caC and have generated 5fC and 5caC maps in both wild-type and *Tdg*-deficient mouse ESCs. Analysis of these data sets suggests that dynamic cytosine methylation/demethylation cycle occurs at an unexpectedly large number of genomic loci across the genome. Genome-wide mapping of all 5mC oxidation derivatives described in this study sets the stage to systematically study the function of DNA methylation and demethylation dynamics in development and diseases.

EXPERIMENTAL PROCEDURES

Cell Culture and Lentiviral Knockdown of *Tdg*

Mouse ESCs (E14Tg2A) were cultured under feeder-free conditions. For *Tdg* knockdown, mouse ESCs were infected with lentiviruses expressing both the puromycin N-acetyl-tranferase and the short hairpin RNA (shRNA) targeting *Tdg* (5′-GCAAGGATCTGTCTAGTAA-3′). Infected cells were selected by puromycin for 1 week before being harvested for further experiments. Detailed procedures can be found in Extended Experimental Procedures.

5mC/5hmC/5fC/5caC DIP and High-Throughput Sequencing Analysis

For genome-wide analysis of 5mC/5hmC/5fC/5caC distributions, 10 μg of sonicated, adaptor-ligated genomic DNA from control or *Tdg* knockdown mouse ESCs was used as input, and 5 μl of 5mC antibody (Eurogentec, BI-MECY-0500), 5 μl of 5hmC antibody (Active Motif, 39791), 1 μl of 5fC antiserum, or 0.3 μl of 5caC antiserum was used to immunoprecipitate modified DNA as previously described (Wu et al., 2011a). Immunoprecipitated DNA was amplified for Illumina sequencing. Detailed DIP-seq procedures can be found in Extended Experimental Procedures.

Repetitive Sequence Analysis and Identification of 5mC/5hmC/5fC/5caC-Enriched Regions

All reads were aligned to the mouse genome (mm9) using bowtie (v0.12.7) to determine the total numbers of unmapped, multihit, and uniquely mapped reads. To determine which class of repetitive sequences these RMSK reads overlap with, we first constructed separate Bowtie indices for each class of repetitive sequences based on RMSK annotation. Only reads that uniquely belong to one class of repetitive sequences were counted. For identifying genomic regions enriched for 5mC/5hmC/5fC/5caC signals, we adapted a two-step computational procedure as previously described (Shen et al., 2012). Briefly, we first identified peak candidates with MACS (v1.4.2) using input as the control data set. Then, we quantitatively filtered out peak candidates that have high signals in IgG mock DIP experiments. Detailed description of data analysis is presented in Extended Experimental Procedures.

RT-qPCR and Gene Expression Microarray Analysis

Purified total RNA was reverse transcribed and analyzed by quantitative PCR. Primers for RT-qRCR are listed in Table S6. Transcriptome analysis was performed with the Mouse Gene 1.0 ST Array (Affymetrix). Three biological independent samples were analyzed for both control and *Tdg*-deficient mouse ESCs. Detailed procedures of gene expression analysis can be found in Extended Experimental Procedures.

ACCESSION NUMBERS

The DIP-seq and expression microarray datasets have been deposited in Gene Expression Omnibus (GEO) under the accession number GSE42250.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, seven figures, and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2013.04.002.

ACKNOWLEDGMENTS

We thank Drs. Jin He and Falong Lu for their helpful discussions. This project is supported by NIH grant U01DK089565 (to Y.Z.) and R01GM097253 (to K.Z.). H.W. is supported by a postdoctoral fellowship (Merck Fellow) from Jane Coffin Childs Funds for Medical Research. D.D. is a CIRM predoctoral fellow. Y.Z. is an investigator of the Howard Hughes Medical Institute.

Received: November 9, 2012
Revised: January 18, 2013
Accepted: March 26, 2013
Published: April 18, 2013

REFERENCES

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823–837.

Bartke, T., Vermeulen, M., Xhemalce, B., Robson, S.C., Mann, M., and Kouzarides, T. (2010). Nucleosome-interacting proteins regulated by DNA and histone methylation. Cell *143*, 470–484.

Baylin, S.B., and Jones, P.A. (2011). A decade of exploring the cancer epigenome - biological and translational implications. Nat. Rev. Cancer *11*, 726–734.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315–326.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev. *16*, 6–21.

Booth, M.J., Branco, M.R., Ficz, G., Oxley, D., Krueger, F., Reik, W., and Balasubramanian, S. (2012). Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. Science *336*, 934–937.

Cao, R., Wang, L., Wang, H., Xia, L., Erdjument-Bromage, H., Tempst, P., Jones, R.S., and Zhang, Y. (2002). Role of histone H3 lysine 27 methylation in Polycomb-group silencing. Science *298*, 1039–1043.

Cedar, H., and Bergman, Y. (2012). Programming of DNA methylation patterns. Annu. Rev. Biochem. *81*, 97–117.

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.

Chen, C.C., Wang, K.Y., and Shen, C.K. (2012). The mammalian de novo DNA methyltransferases DNMT3A and DNMT3B are also DNA 5-hydroxymethylcytosine dehydroxymethylases. J. Biol. Chem. *287*, 33116–33121.

Cimmino, L., Abdel-Wahab, O., Levine, R.L., and Aifantis, I. (2011). TET family proteins and their role in stem cell differentiation and transformation. Cell Stem Cell *9*, 193–204.

Cortázar, D., Kunz, C., Selfridge, J., Lettieri, T., Saito, Y., MacDougall, E., Wirz, A., Schuermann, D., Jacobs, A.L., Siegrist, F., et al. (2011). Embryonic lethal phenotype reveals a function of TDG in maintaining epigenetic stability. Nature *470*, 419–423.

Cortellino, S., Xu, J., Sannai, M., Moore, R., Caretti, E., Cigliano, A., Le Coz, M., Devarajan, K., Wessels, A., Soprano, D., et al. (2011). Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. Cell *146*, 67–79.

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. USA *107*, 21931–21936.

Dawlaty, M.M., Ganz, K., Powell, B.E., Hu, Y.C., Markoulaki, S., Cheng, A.W., Gao, Q., Kim, J., Choi, S.W., Page, D.C., and Jaenisch, R. (2011). Tet1 is dispensable for maintaining pluripotency and its loss is compatible with embryonic and postnatal development. Cell Stem Cell *9*, 166–175.

Dawlaty, M.M., Breiling, A., Le, T., Raddatz, G., Barrasa, M.I., Cheng, A.W., Gao, Q., Powell, B.E., Li, Z., Xu, M., et al. (2013). Combined deficiency of

Tet1 and Tet2 causes epigenetic abnormalities but is compatible with postnatal development. Dev. Cell *24*, 310–323.

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.

Dreszer, T.R., Karolchik, D., Zweig, A.S., Hinrichs, A.S., Raney, B.J., Kuhn, R.M., Meyer, L.R., Wong, M., Sloan, C.A., Rosenbloom, K.R., et al. (2012). The UCSC Genome Browser database: extensions and updates 2011. Nucleic Acids Res. *40*(Databaseissue), D918–D923.

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.

Globisch, D., Münzel, M., Müller, M., Michalakis, S., Wagner, M., Koch, S., Brückl, T., Biel, M., and Carell, T. (2010). Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. PLoS ONE *5*, e15367.

Goldberg, A.D., Allis, C.D., and Bernstein, E. (2007). Epigenetics: a landscape takes shape. Cell *128*, 635–638.

Gu, T.P., Guo, F., Yang, H., Wu, H.P., Xu, G.F., Liu, W., Xie, Z.G., Shi, L., He, X., Jin, S.G., et al. (2011). The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. Nature *477*, 606–610.

Guo, J.U., Su, Y., Zhong, C., Ming, G.L., and Song, H. (2011). Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. Cell *145*, 423–434.

Hackett, J.A., Sengupta, R., Zylicz, J.J., Murakami, K., Lee, C., Down, T.A., and Surani, M.A. (2013). Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine. Science *339*, 448–452.

He, Y.F., Li, B.Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L., et al. (2011). Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science *333*, 1303–1307.

Inoue, A., and Zhang, Y. (2011). Replication-dependent loss of 5-hydroxymethylcytosine in mouse preimplantation embryos. Science *334*, 194.

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.

Iqbal, K., Jin, S.G., Pfeifer, G.P., and Szabó, P.E. (2011). Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. Proc. Natl. Acad. Sci. USA *108*, 3642–3647.

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.

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.

Jaenisch, R., and Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat. Genet. *33*(*Suppl*), 245–254.

Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. *13*, 484–492.

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.

Khare, T., Pai, S., Koncevicius, K., Pal, M., Kriukiene, E., Liutkeviciute, Z., Irimia, M., Jia, P., Ptak, C., Xia, M., et al. (2012). 5-hmC in the brain is abundant in synaptic genes and shows differences at the exon-intron boundary. Nat. Struct. Mol. Biol. *19*, 1037–1043.

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.

Koh, K.P., Yabuuchi, A., Rao, S., Huang, Y., Cunniff, K., Nardone, J., Laiho, A., Tahiliani, M., Sommer, C.A., Mostoslavsky, G., et al. (2011). Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. Cell Stem Cell *8*, 200–213.

Kriaucionis, S., and Heintz, N. (2009). The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. Science *324*, 929–930.

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.

Maiti, A., and Drohat, A.C. (2011). Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. J. Biol. Chem. *286*, 35334–35338.

Marcucci, G., Maharry, K., Wu, Y.Z., Radmacher, M.D., Mrózek, K., Margeson, D., Holland, K.B., Whitman, S.P., Becker, H., Schwind, S., et al. (2010). IDH1 and IDH2 gene mutations identify novel molecular subsets within de novo cytogenetically normal acute myeloid leukemia: a Cancer and Leukemia Group B study. J. Clin. Oncol. *28*, 2348–2355.

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.

Mellén, M., Ayata, P., Dewell, S., Kriaucionis, S., and Heintz, N. (2012). MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. Cell *151*, 1417–1430.

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.

Münzel, M., Globisch, D., Brückl, T., Wagner, M., Welzmiller, V., Michalakis, S., Müller, M., Biel, M., and Carell, T. (2010). Quantification of the sixth DNA base hydroxymethylcytosine in the brain. Angew. Chem. Int. Ed. Engl. *49*, 5375–5377.

Nabel, C.S., Jia, H., Ye, Y., Shen, L., Goldschmidt, H.L., Stivers, J.T., Zhang, Y., and Kohli, R.M. (2012). AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation. Nat. Chem. Biol. *8*, 751–758.

Nakamura, J., Walker, V.E., Upton, P.B., Chiang, S.Y., Kow, Y.W., and Swenberg, J.A. (1998). Highly sensitive apurinic/apyrimidinic site assay can detect spontaneous and chemically induced depurination under physiological conditions. Cancer Res. *58*, 222–225.

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.

Pfaffeneder, T., Hackner, B., Truss, M., Münzel, M., Müller, M., Deiml, C.A., Hagemeier, C., and Carell, T. (2011). The discovery of 5-formylcytosine in embryonic stem cell DNA. Angew. Chem. Int. Ed. Engl. *50*, 7008–7012.

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.

Raiber, E.A., Beraldi, D., Ficz, G., Burgess, H.E., Branco, M.R., Murat, P., Oxley, D., Booth, M.J., Reik, W., and Balasubramanian, S. (2012). Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase. Genome Biol. *13*, R69.

Sasaki, H., and Matsui, Y. (2008). Epigenetic events in mammalian germ-cell development: reprogramming and beyond. Nat. Rev. Genet. *9*, 129–140.

Seisenberger, S., Andrews, S., Krueger, F., Arand, J., Walter, J., Santos, F., Popp, C., Thienpont, B., Dean, W., and Reik, W. (2012). The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. Mol. Cell *48*, 849–862.

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

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. Nat. Biotechnol. *29*, 68–72.

Song, C.X., Yi, C., and He, C. (2012). Mapping recently identified nucleotide variants in the genome and transcriptome. Nat. Biotechnol. *30*, 1107–1116.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, 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.

Stroud, H., Feng, S., Morey Kinney, S., Pradhan, S., and Jacobsen, S.E. (2011). 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. Genome Biol. *12*, R54.

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. (2011a). Integrating 5-hydroxymethylcytosine into the epigenomic landscape of human embryonic stem cells. PLoS Genet. *7*, e1002154.

Szulwach, K.E., Li, X., Li, Y., Song, C.X., Wu, H., Dai, Q., Irier, H., Upadhyay, A.K., Gearing, M., Levey, A.I., et al. (2011b). 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. Nat. Neurosci. *14*, 1607–1616.

Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L., and Rao, A. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science *324*, 930–935.

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.

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.

Wossidlo, M., Nakamura, T., Lepikhov, K., Marques, C.J., Zakhartchenko, V., Boiani, M., Arand, J., Nakano, T., Reik, W., and Walter, J. (2011). 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. Nat Commun. *2*, 241.

Wu, H., and Zhang, Y. (2011a). Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation. Genes Dev. *25*, 2436–2452.

Wu, H., and Zhang, Y. (2011b). Tet1 and 5-hydroxymethylation: a genome-wide view in mouse embryonic stem cells. Cell Cycle *10*, 2428–2436.

Wu, H., Coskun, V., Tao, J., Xie, W., Ge, W., Yoshikawa, K., Li, E., Zhang, Y., and Sun, Y.E. (2010). Dnmt3a-dependent nonpromoter DNA methylation facilitates transcription of neurogenic genes. Science *329*, 444–448.

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., Sun, Y.E., and Zhang, Y. (2011b). Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. Nature *473*, 389–393.

Xu, Y., Wu, F., Tan, L., Kong, L., Xiong, L., Deng, J., Barbera, A.J., Zheng, L., Zhang, H., Huang, S., et al. (2011). Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells. Mol. Cell *42*, 451–464.

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 *492*, 443–447.

Yamaguchi, S., Hong, K., Liu, R., Inoue, A., Shen, L., Zhang, K., and Zhang, Y. (2013). Dynamics of 5-methylcytosine and 5-hydroxymethylcytosine during germ cell reprogramming. Cell Res. *23*, 329–339.

Yoder, J.A., Walsh, C.P., and Bestor, T.H. (1997). Cytosine methylation and the ecology of intragenomic parasites. Trends Genet. *13*, 335–340.

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.

EXTENDED EXPERIMENTAL PROCEDURES

Lentiviral Vector Construction and Virus Production

Stable *Tdg* knockdown 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′-GTTCAGATGTGCGGC GAGT-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 cotransfected into 293T cells (He et al., 2008). The supernatant was harvested at 24, 36 and 48 hr 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% nonfat milk in TBS containing 0.1% Tween 20 (TBST) for 1 hr 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), 1× nonessential amino acid, 1× sodium pyruvate, 1× GlutaMax, 1× beta-mercaptoethanol (Invitrogen) and 1,000 units/ml leukemia inhibitory factor (ESGRO, EMD Millipore). To generate control and *Tdg*-knockdown cells, 1 × 10^5 cells are infected with lentivirus (MOI = 10) in a 24-well plate. 48 hr after infection, puromycin (2 μg/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 ΔΔ*Ct* method, and the expression of individual genes is normalized to the expression level of GAPDH. Primers for RT-qPCR are listed in Table S6.

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 at room temperature, cell suspension was centrifuged and resuspended in 100 mM sucrose. The cell suspension was spread onto glass slides dipped into fixative solution (1% paraformaldehyde, 0.15% Triton X-100 and 3cmM dithiothreitol, [pH 9.2]). The glass slides were kept overnight in a humidified box at 4°C. The slides were washed with 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 (4N 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 hr at room temperature, and incubated with the primary antibodies at 4°C overnight, followed by incubation with appropriate secondary antibodies for 1 hr at RT.

Mass Spectrometric Analysis

Genomic DNA was extracted and purified from cells using DNeasy Blood & Tissue Kit (QIAGEN). For mass spectrometric quantification, 30 μg 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 hr (total volume is 250 μl), followed by the addition of 30 μl 10× buffer 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 hr (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.

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 an adaptor for PCR analysis (Forward: 5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′, Reverse: 5′-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3′). These oligos (5 pg) were denatured and used as the input in the presence of 10 μg sonicated salmon sperm DNA (ssDNA). For each DIP assay, 1 μl (anti-5fC) or 0.3 μl (anti-5caC) of antiserum were used, alternatively, 1 μl of mouse and rabbit preimmune serum mixture was used as IgG control. DNA and antibodies were incubated at 4°C overnight in a final volume of 500 μl 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 μl protein G Dynabeads (Invitrogen) were added to the tube and incubated with the DNA-antibody mixture for 2 hr 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 hr, 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′).

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 μg 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

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 μg of the adaptor-ligated genomic DNA was used as input, and 5 μl of 5mC antibody (Eurogentec, BI-MECY-0500), 5 μl of 5hmC antibody (Active Motif, 39791), 1 μl of 5fC antiserum or 0.3 μl of 5caC antiserum was added to immunoprecipitate modified DNA. For the IgG control, 1 μl 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′-CAAGCAGAAGACGGCATACGAGATXXXXXXCTCGGCATTCCTGCTGAACCGCTCTT-3′, where XXXXXX is the 6 bp barcode) in a 50 μl PCR reaction with 0.2× Sybr-Green I and 1× Phusion High-fidelity PCR Master Mix (NEB). PCR reactions were terminated before entering the nonexponential 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, multihit, and uniquely mapped reads. And then we separately mapped those multihit and uniquely mapped reads to the UCSC RepeatMasker track (RMSK) sequences, to determine the percentages of RMSK reads in multihit 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 data set 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 $\geq$ 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 data set with MACS (v1.4.2) by extending the uniquely mapped reads (keeping at most two read at the same genomic position) to 200 bp 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 nonspecific 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 precalculated conservation score for each base in the mouse genome based on multispecies 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 (±2 kb flanking TSS) and H3K79me2 (5 kb downstream of TSS). (2) initiated promoters are associated with H3K4me3 (±2 kb flanking TSS) only. (3) bivalent promoters are associated with H3K4me3 (± 2 kb flanking TSS) and H3K27me3 (±5 kb flanking TSS). (4) silent promoters are not associated with H3K4me3/H3K79me2/H3K27me3.

Published Data Sets

Annotations of known Refseq transcripts and CpG islands were obtained from UCSC (downloaded on January, 2012). To calculate Figures S6A and 7A, previously published RNA-seq data sets 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 heatmap in Figure 4C, we used following published data sets: 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 heatmap 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.

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.

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.

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

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.

Wu, H., and Zhang, Y. (2011). Tet1 and 5-hydroxymethylation: a genome-wide view in mouse embryonic stem cells. Cell Cycle *10*, 2428–2436.

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

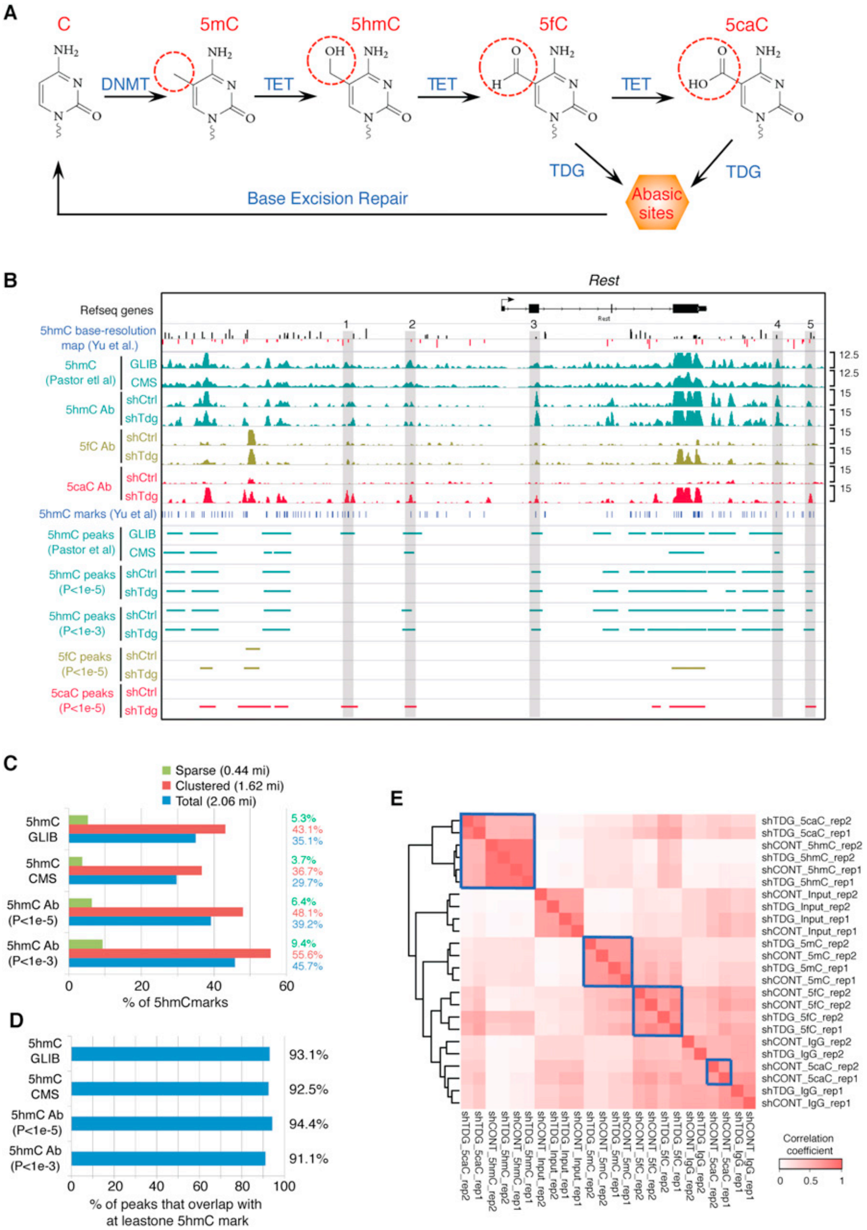

Figure S1. Comparison of Antibody-Based and Chemical-Labeling-Based Mapping Methods, Related to Figure 1

(A) Schematic diagram of the active 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 [Yu et al., 2012]). 5hmC-enriched regions identified by two different cutoffs ($p < 1 \times 10^{-3}$ or 1×10^{-5}) from 5hmC Ab data sets 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 the 5hmC 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 those without any neighboring 5hmC within 1 kb of their flanking regions. Clustered 5hmC marks are defined as those with at least one neighboring 5hmC within 1 kb of their flanking regions.

(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 2 kb nonoverlapping genomic intervals. Note that 5caC profiles in shTDG were closely clustered with 5hmC profiles in shCONT or shTDG.

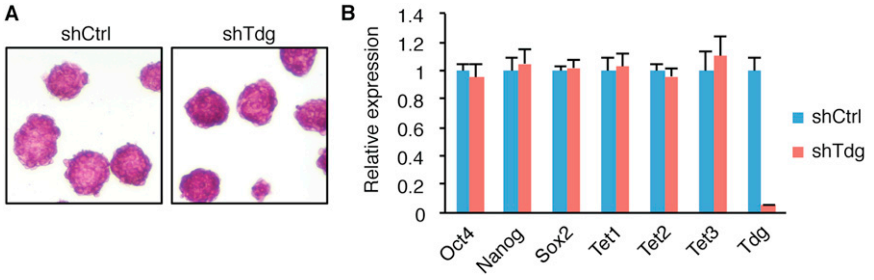

Figure S2. *Tdg*-Depletion Does Not Affect the Mouse ESC Morphology or the RNA Levels of Tet and Pluripotency Genes, Related to Figure 1
(A) Alkaline phosphatase staining of control and *Tdg*-deficient mouse ESCs.
(B) RT-qPCR analyses of Tet and representative pluripotency genes in control and *Tdg*-deficient mouse ESCs. Data are presented as mean ± SEM.

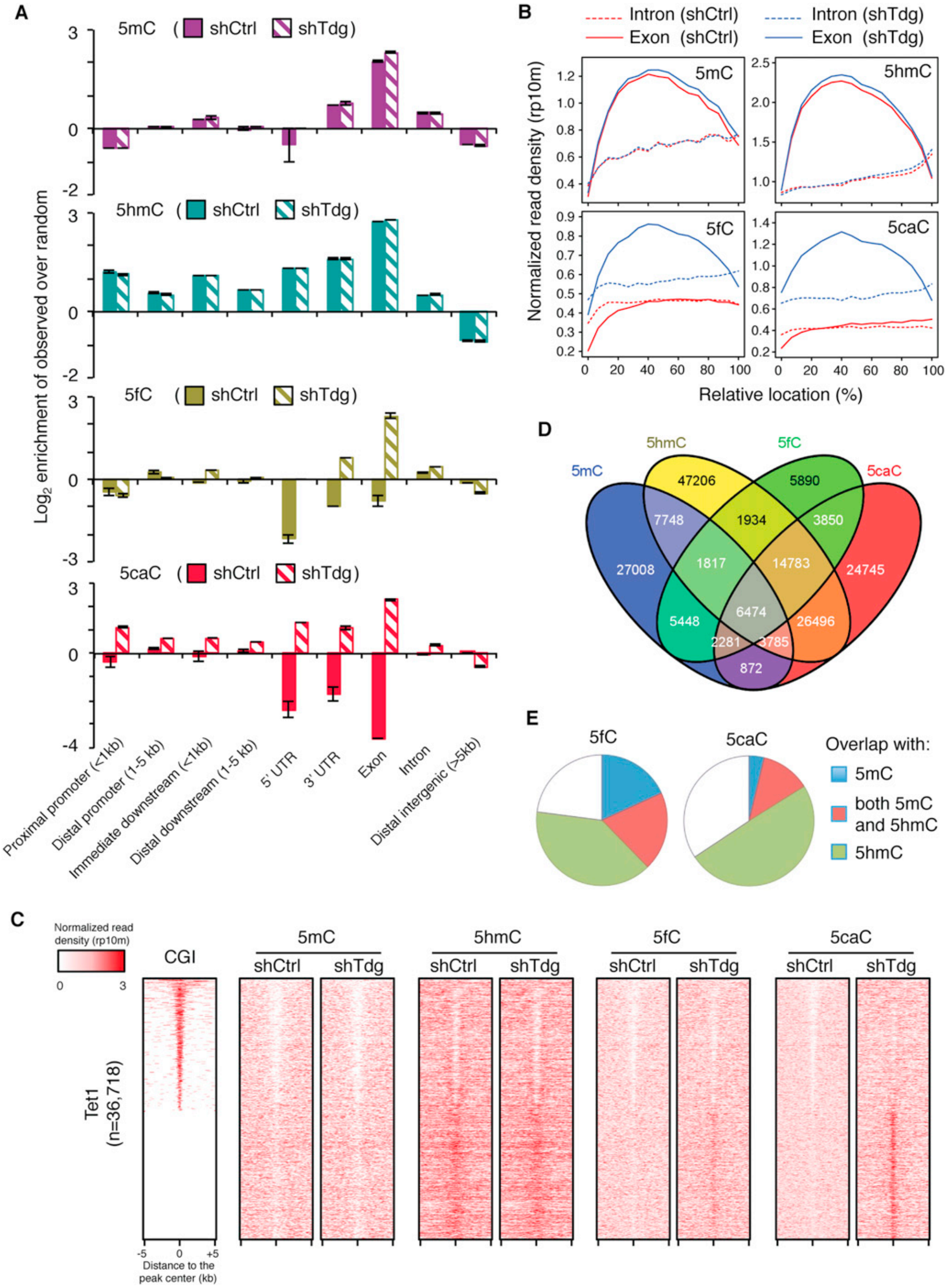

Figure S3. Overall Genomic Distributions of 5mC/5hmC/5fC/5caC-Enriched Regions in Control and *Tdg*-Deficient Mouse ESCs, Related to Figure 3

(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 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 increase of 5fC/5caC signals in *Tdg*-deficient cells compared with control cells is more apparent in exons than in introns.

(C) Heatmap representation of normalized read density of 5mC/5hmC/5fC/5caC at Tet1-bound regions in mouse ESCs (Wu et al., 2011b). 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 within 5 kb of their flanking 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.

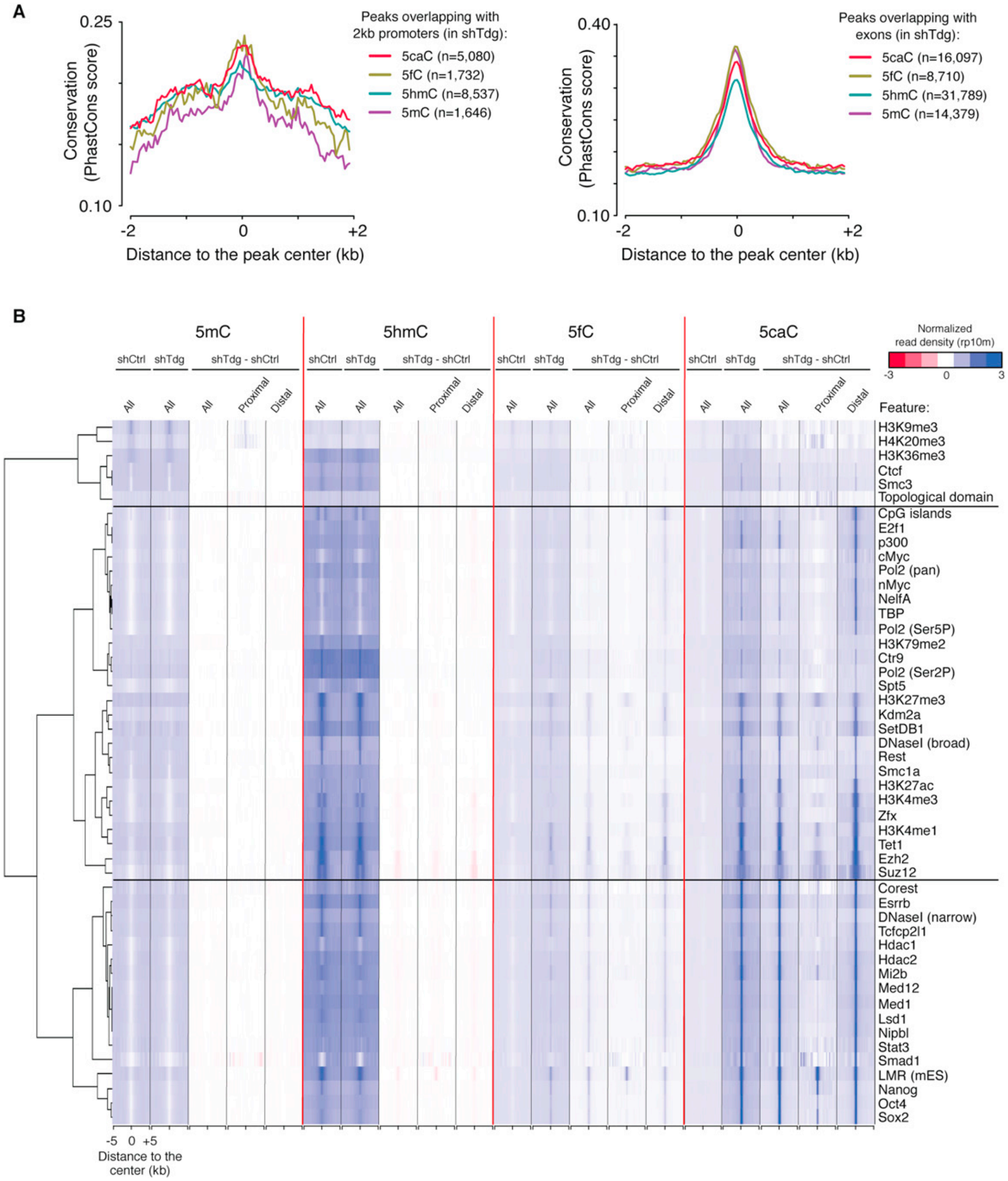

Figure S4. *Tdg*-Depletion-Induced 5caC Peaks Are Enriched at Distal Regulatory Regions, Related to Figure 4

(A) Average conservation PhastCons scores at 5mC/5hmC/5fC/5caC enriched regions (overlapping with promoters or exons) in *Tdg*-deficient mouse ESCs. The numbers of peaks that are overlapping with exons and proximal promoters (±1 kb flanking TSSs) for each cytosine modification are also shown.

(B) Heatmaps of 5mC/5hmC/5fC/5caC levels (normalized read density) in control and *Tdg*-deficient cells at annotated genomic features or enriched regions for transcriptional regulators, histone modifications, pluripotency transcription factors (TFs) and distal regulator regions (derived from published data sets in mouse ESCs). The difference in 5mC/5hmC/5fC/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 ±1 kb flanking TSSs) and distal features. The heatmap was clustered by hierarchical clustering (complete linkage).

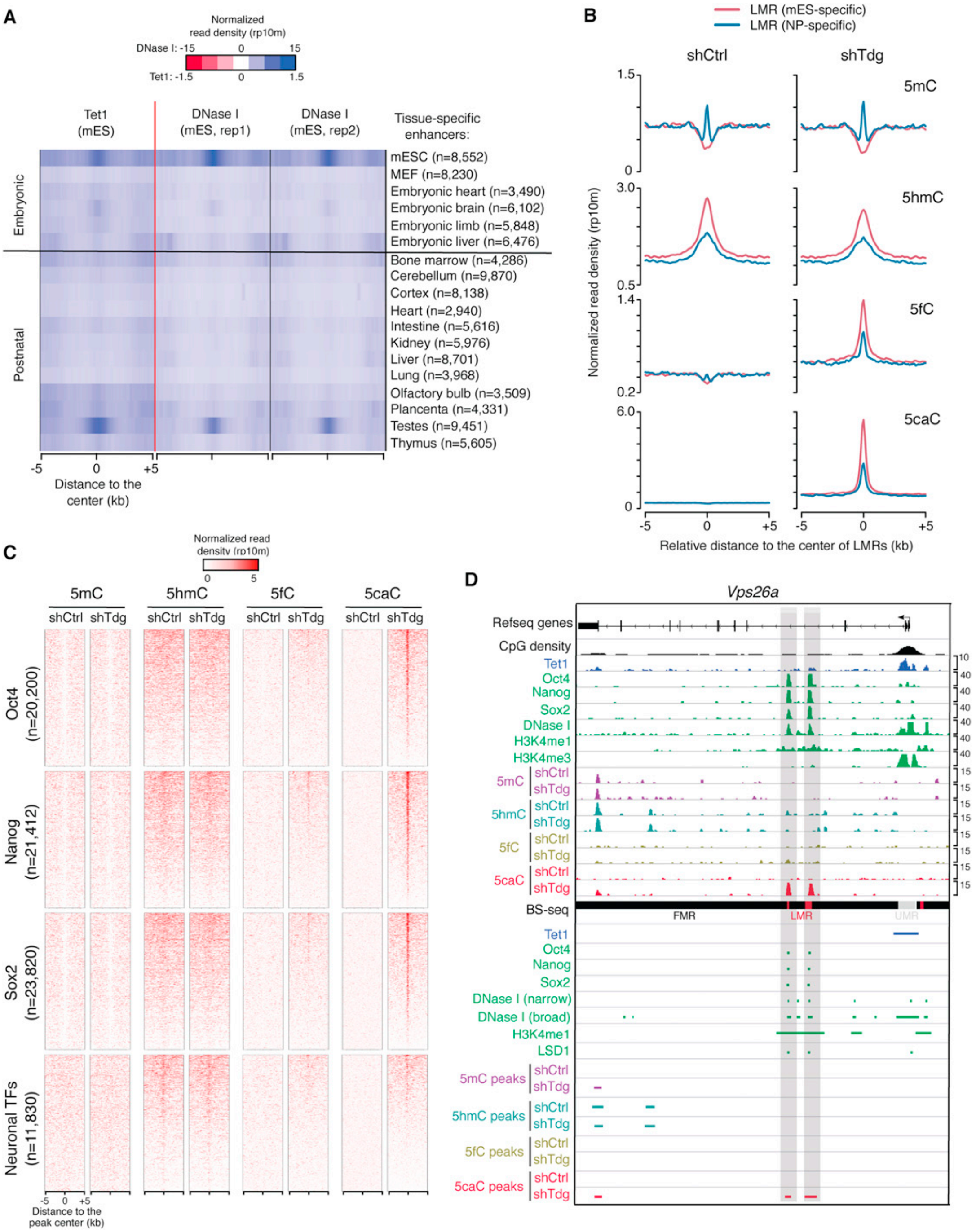

Figure S5. ***Tdg*-Depletion-Induced 5caC Peaks Are Enriched at Active Enhancers and Pluripotency TF Binding Sites in Mouse ESCs, Related to Figure 5**

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

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

(C) Heatmap of 5mC/5hmC/5fC/5caC levels (normalized read density) in control and *Tdg*-deficient mouse ESCs at binding sites of mouse pluripotency TFs (Oct4, Nanog and Sox2) and neuronal TFs.

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

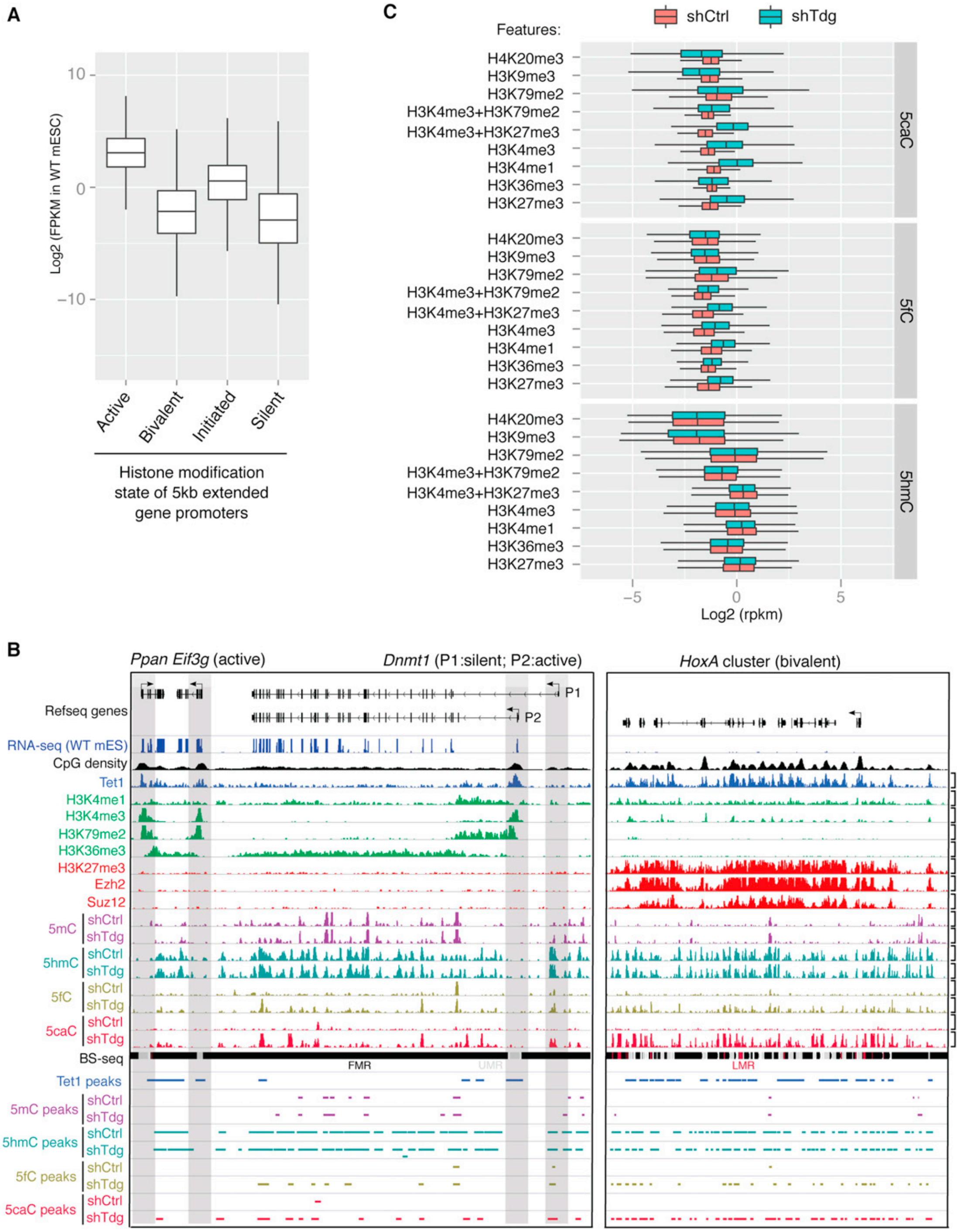

Figure S6. *Tdg*-Depletion-Induced 5caC Peaks Are Enriched at Bivalent and Transcriptionally Inactive Silent Gene Promoters in Mouse ESCs, Related to Figure 6

(A) Average gene expression levels in wild-type mouse ESCs (derived from 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) Shown are 5mC/5hmC/5fC/5caC distributions in control and *Tdg*-deficient mouse ESCs at representative loci that are associated with different histone modification states. The gene promoters are highlighted by gray 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.

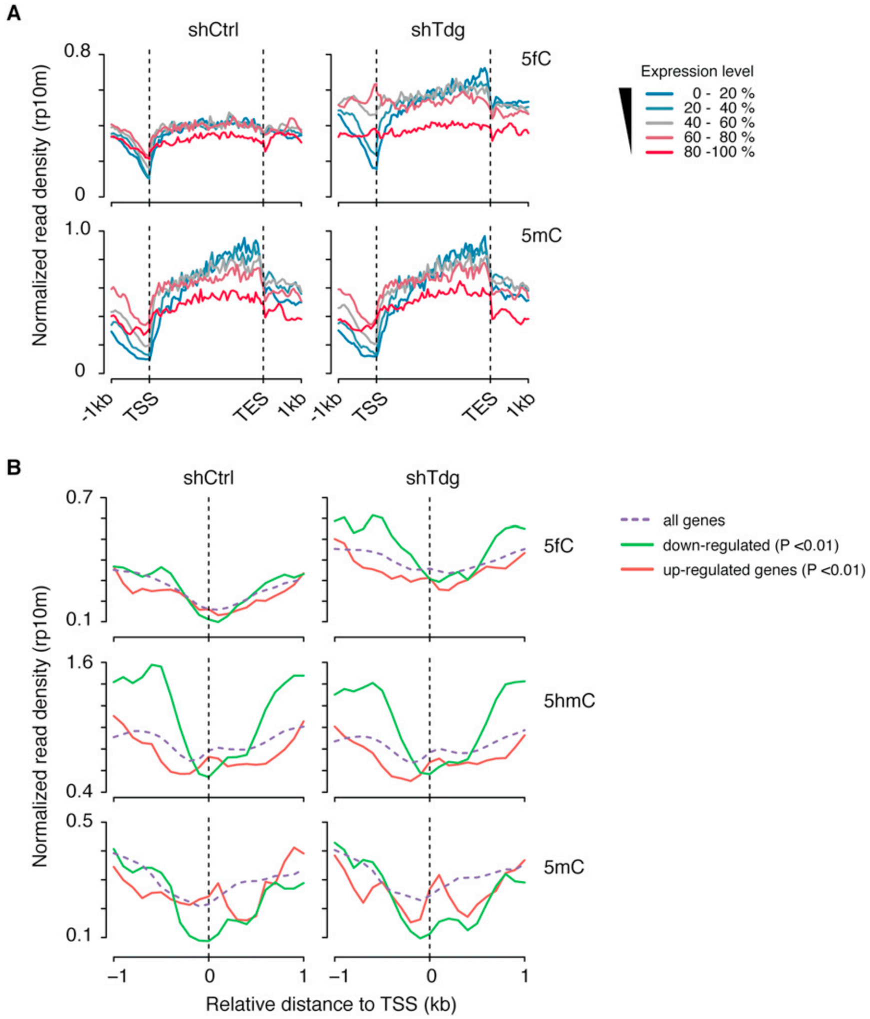

Figure S7. Complex Relationship between Gene Expression and Cytosine Modification Cycling, Related to Figure 7

(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 downregulated and upregulated genes.