Supporting Information

Leung et al. 10.1073/pnas.1322273111

SI Materials and Methods

Mouse Embryonic Stell Cell Culture and Treatments. Mouse embryonic stem cells (mESCs) were passaged every 48 h in DMEM supplemented with 15% FBS (90 ml of FBS/600 ml of total media) (HyClone), 20 mM Hepes, 0.1 mM nonessential amino acids, 0.1 mM 2-mercaptoethanol, 100 units/mL penicillin, 0.05 mM streptomycin, leukemia inhibitory factor (1,000 units/mL), and 2 mM glutamine. Trypsinized cells were plated on tissue culture grade polystyrene plates treated with 0.2% gelatin from porcine skin (Sigma) for at least 15 min at room temperature before use. Cells were maintained at 37 °C and 5% CO_2.

Setdb1 conditional knockout (CKO) mESCs carried one null and one floxed allele of Setdb1, which could be deleted upon CRE-mediated excision. Treatment with 4-hydroxytamoxifen (4-OHT) induced the activation of ligand binding domain of estrogen receptor and CRE recombinase fusion protein, leading to conditional deletion of the remaining allele (1). For 4-OHT treatment, ES cells were cultured in ES medium with 800 nM 4-OHT (Sigma) for 4 d, and further cultured without 4-OHT for 2 d. Depletion of Setdb1 was validated by Western blotting with a Setdb1-specific antibody (Millipore).

MethylC-Sequencing and Bisulfite Sanger Sequencing. Genomic DNA was extracted from mESCs and was spiked in with unmethylated lambda DNA (Promega). The DNA was fragmented by sonication. Purified DNA fragments were end repaired. After A-tailing reaction, fragments were ligated to paired-end cytosine-methylated adapters provided by Illumina. Size-selected adapter-ligated DNA was treated with sodium bisulfite using the EZ DNA Methylation-Gold kit (Zymo Research). The resulting DNA molecules were enriched by PCR, purified, and sequenced following standard protocols from Illumina. Libraries were sequenced to a minimum of 7× coverage of the mouse genome. Bisulfite-treated genomic DNA (gDNA) harvested from an independent biological replicate was also amplified twice by seminested PCR as previously described (2). Amplified products were sequenced by Sanger sequencing. Primer sequences can be found on supplemental experimental procedures.

ChIP-Sequencing Library Generation and Sequencing. mESCs were processed according a ChIP protocol as previously described (3). ChIP-sequencing (ChIP-seq) libraries were prepared and sequenced using the Illumina instrument as per manufacturer instructions. Anti-H3K9me3 (Abcam 8898) and anti-Tet1 (Millipore 09-872) antibodies were used.

RNA-Seq Library Generation and Sequencing. Total RNA from wild-type and *Dnmt 3a/3b* double knockout (DKO) mESCs were extracted using TRIzol (Invitrogen) according to manufacturer instructions. Ribosomal RNA was removed using the Ribom-Minus Eukaryote kit (Invitrogen). The mRNA libraries were prepared for strand-specific sequencing as described previously (4).

5-Hydoxylmethylcytosine Detection: Selective Capture of 5-Hydroxymethyl-Cytosine from Genomic DNA. We selectively captured 5-hydroxymethyl-cytosine (5hmC) for genome-wide analyses as previously described (5). In summary, genomic DNA was fragmented by sonication and size selected (~300 bp). The DNA was treated with β-glucosyltransferase (β-GT), which transfers an azide-glucose to 5hmC. The purified products are then biotinylated by click chemistry. This allows the labeled 5hmC containing DNA fragments to be captured by using streptavidin beads. The DNA is purified and sequenced following standard protocols from Illumina.

Hydroxymethyl DNA immunoprecipitation-quantitative PCR (hMeDIP-qPCR) was conducted to validate the findings of genome-wide 5hmC datasets. The hydroxymethylated DNA kit (Diagenode) was used according to the manufacturer's protocol. Captured DNA was analyzed by qPCR with primers targeting Etn/Mus subfamily of ERVs. Primer sequences can be found in Table S3.

Combined Bisulfite Restriction Analysis. Genomic DNA from independent harvests were isolated from WT, *Setdb1* KO, *Dnmt3a/3b* WT, and DKO mESCs. DNA was subjected to bisulfite conversion with the EZ DNA Methylation-Gold kit (Zymo Research), as per manufacturer's protocol. Specific primers (Table S3) were used to amplify the region of interest, which contains a restriction site for either Taq I or Tai I endonuclease, only when CpG is methylated. PCR amplicons were digested overnight to ensure complete digestion. The products were resolved on a 2% agarose gel (2 g of agarose/100 ml of 1X Tris-Borate-EDTA buffer). Bands were quantified with AlphaImager software. The raw signals for both digested and undigested bands are normalized and presented as a proportion of WT methylation (%). Each experiment was repeated at least twice, with error bars reflecting the SD between replicates.

Data Analysis. ***Processing and alignment of MethylC-Seq read sequences.*** *Dnmt3a/3b* WT and DKO methylomes were mapped with a few modifications as described by Lister et al. (6). Briefly, reads in FastQ format files from the Illumina pipeline were trimmed and all cytosine bases were replaced by thymine. Using Bowtie (v0.12.7) (7) these reads were aligned to two NCBI BUILD 37/MM9 reference sequences, one with thymines replacing cytosines(C) and the complementary sequence with adenines replacing guanines. We analyzed the two replicates for wild type and for *Dnmt3a*$^{-/-}$ *Dnmt3b*$^{-/-}$ (DKO) mESCs independently of each other to ensure reproducibility and finally pooled the respective WT and the KO mESC replicates. For the WT and *Setdb1* KO mESC methylomes, which were produced later than the *Dnmt3a/3b* WT and DKO methylomes, the MethylC-Seq reads were mapped using BSMAP (http://www.biomedcentral.com/1471-2105/10/232). The Lambda virus genome was added throughout the procedure as negative control to estimate the bisulfite sequencing error rate. We analyzed the two replicates for wild-type and for *Dnmt3a*$^{-/-}$ *Dnmt3b*$^{-/-}$ (DKO) mES cells independently of each other to ensure reproducibility and finally pooled the respective WT and the knockout mESC replicates. After the alignment, duplicate reads were removed using the Picard software (http://picard.sourceforge.net) and all output was converted into BAM format using samtools (8). Alignments were merged in a strand specific way, giving rise to distributions of bases at each position using the pileup option of samtools. Individual bases having a PHRED quality score < 20 were filtered out. For each position x in the reference genome with a C the number of reads containing mC (methylated C) were counted and set in 10 relation to the total number of reads containing mC and a C at position ×. The context of the C (CG, CHG or CHH) at a given position x was determined based on the pileup consensus call for positions x+1 and x+2.
Methylation status of genomic regions. Each chromosome was divided up into bins of 100 bp and for each bin total numbers of sequenced mC and C were determined, for each context of the cytosine (CG, CHG or CHH) and for Watson and Crick strand separately. In a similar way to the method presented by Lister et al. (6), a binomial test was used to distinguish whether the ratio of methylated C to all C (mC/all C) in each bin was significantly different from the error rate. As an overall error rate, we estimated the

following numbers: WT, +strand: 0.0055; WT, −strand 0.0057; DKO, +strand: 0.0053; DKO, −strand: 0.0052. Finally, a false discovery rate cutoff of 1% was applied to call the mC ratio of a bin statistically signficantly different from the error rate.

Segmentation and characterization of the CG-methylome in Dnmt3a/3b DKO. For each 100-bp bin of the CG-methylome in DKO, a binomial test was carried out to distinguish between the overall average background level of methylation in the DKO cell line and occurrences of enriched residual levels of methylation. As a conservative measure of methylation level we used a threshold of 10%, which was higher than the actual measured level of 2.7%. A value of 1 was assigned to bins that showed a *P* value of 10^{-7} and lower, a value of 0 was assigned to those bins with higher *P* values. Bins that were not spanned by any reads were ignored in the further segmentation process. The binary data were subjected to segmentation using the R package RHmm (http://cran.rproject.org/web/packages/RHmm/index.html). For each chromosome separately, a two-state hidden Markov model (HMM) was established, corresponding to loci of enriched residual methylation and of loci with depleted methylation. Using RHmm's "discrete" distribution option to model the emission probabilities, parameters were estimated using the Baum–Welch algorithm. Using the Viterbi algorithm, state labels of enriched and depleted methylation were assigned. Bins of enriched state that did not have at least four other bins with the neighboring three bins upstream and three bins downstream were combined with neighboring segments of the depleted state. Larger segments of the enriched state were then aggregated by combining consecutive bins with the same state label. These segments of the enriched state were called enriched residual methylation loci (ERML). These ERML were overlapped with genomic features of the mouse genome National Center for Biotechnology Information (NCBI) BUILD 37/MM9 and H3K9me3 data using the IRanges package in R and the respective annotation for Reference Sequence (RefSeq) transcripts and repeat elements (RepeatMasker) were retrieved from University of California Santa Cruz (UCSC). To produce shuffled ERML, for each chromosome independently, positions of ERML were shuffled within those bins that were overlapped by at least one read, maintaining the number and the size distribution of original ERML found for the respective chromosome.

Determine Hypo- and Hypermethylated Regions in *Setdb1* KO mESCs. The numbers of methylated CGs and unmethylated CGs were determined for 200-bp windows for both WT and *setdb1* KO cells. We only considered those windows that had at least five combined totally CG counts (methylated + unmethylated) in both WT and KO cells. To identify windows that showed significant differences of DNA methylation levels between WT and KO, the numbers of methylated CGs and unmethylated CGs in WT and KO were subjected to Fisher's exact test to calculate the *P* values, which considered both DNA methylation differences and the sequencing coverages for that window. Windows that showed $P \leq 0.01$ were considered as candidate regions that show altered DNA methylation upon ablation of Setdb1. To further eliminate random noises, we identified those windows that had at least one more such candidate window within a 5-kb region nearby, and at least three-quarters of these windows in such a 5-kb region show consistent DNA methylation change directions (either WT > KO or WT < KO).

Analysis of 5hmC Data. The 5hmC data were mapped using Bowtie against the mm9 genome. The reads per kilobases per million of reads (RPKM) were calculated to represent the enrichment levels of 5hmC at each genomic region by normalizing the total number of reads for both WT and *Setdb1* KO cells.

ChIP-Seq Data Analysis. ChIP-seq reads for IPs using antibodies against H3K9me3 and respective input reads in FastQ format files from the Illumina pipeline were aligned to the NCBI BUILD 37/MM9 reference sequences using Bowtie software (v0.12.7) (7, 8). The genome was divided up into bins of 100 bp and for each bin the number of reads in IP and input spanning the respective bin was counted. RPKM values were calculated for IP and input independently Finally, RPKM values were computed by subtracting $\text{RPKM}_{\text{input}}$ from $\text{RPKM}_{\text{H3K9me3}}$ for each bin. Regions of enriched tag densities were identified using ChromaBlocks (3).

RNA-Seq Data Analysis. RNA-seq reads in FastQ format files from the Illumina pipeline were aligned to the NCBI BUILD 37/MM9 reference sequences using TopHat software (v1.4.1) (9). Gene expression was summarized as RPKM values using Cufflinks (9) software based on a gene transfer format file combining Gene Symbol and RefSeq definitions as retrieved from UCSC browser website.

Visualization of ChIP-Seq and mC/All C Ratio Data in Given Loci. To display ChIP-seq data and mC ratios in given loci in form of a heat map, data were extracted for each locus including 5 kb upstream of the start and 5 kb downstream of the end of the locus. To standardize the different lengths of the loci, the data for each locus were divided up into 100 bins of identical size and the respective values within one bin were averaged. These values were color coded and displayed using TreeView (10) with the order of the genes depending, for example, on gene expression or *k*-means clustering (using Cluster 3.0) (11) giving rise to heat maps. Similarly, to create intensity profiles in exons and introns of all RefSeq genes, for each exon and intron, respectively, 100-bp bin data were extracted from 5 kb upstream of the start to 5 kb downstream of the end, taking into account the orientation of the respective gene. Within an exon or an intron, data were divided up into 100 bins of identical size and the respective values within one bin were averaged. Finally, for any of the 200 bin positions, within exon/intron or outside, values were averaged over all exons and introns, respectively. For the creation of averaged intensity profiles around transcription start sites (TSSs) of all RefSeq genes, 100-bp bin data for 5 kb up and 5 kb downstream of the TSSs were extracted and for any of the 100-bin positions, values were averaged over all genes.

Gene Ontology Analysis. Gene Ontology analysis was carried out using DAVID Bioinformatics Resources (12). For the GO analysis of differentially expressed genes, the list of background genes was reduced to those genes that showed RPKM > 0 in at least one condition, WT or DKO.

1. Matsui T, et al. (2010) Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. *Nature* 464(7290):927–931.
2. Dong KB, et al. (2008) DNA methylation in ES cells requires the lysine methyltransferase G9a but not its catalytic activity. *EMBO J* 27(20):2691–2701.
3. Hawkins RD, et al. (2010) Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* 6(5):479–491.
4. Parkhomchuk D, et al. (2009) Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res* 37(18):e123.
5. Li Y, Song CX, He C, Jin P (2012) Selective capture of 5-hydroxymethylcytosine from genomic DNA. *J Vis Exp* (68).
6. Lister R, et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462(7271):315–322.
7. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25.
8. Li H, et al.; 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.
9. Trapnell C, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28(5):511–515.
10. Saldanha AJ (2004) Java Treeview—extensible visualization of microarray data. *Bioinformatics* 20(17):3246–3248.
11. de Hoon MJ, Imoto S, Nolan J, Miyano S (2004) Open source clustering software. *Bioinformatics* 20(9):1453–1454.
12. Huang W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4(1):44–57.

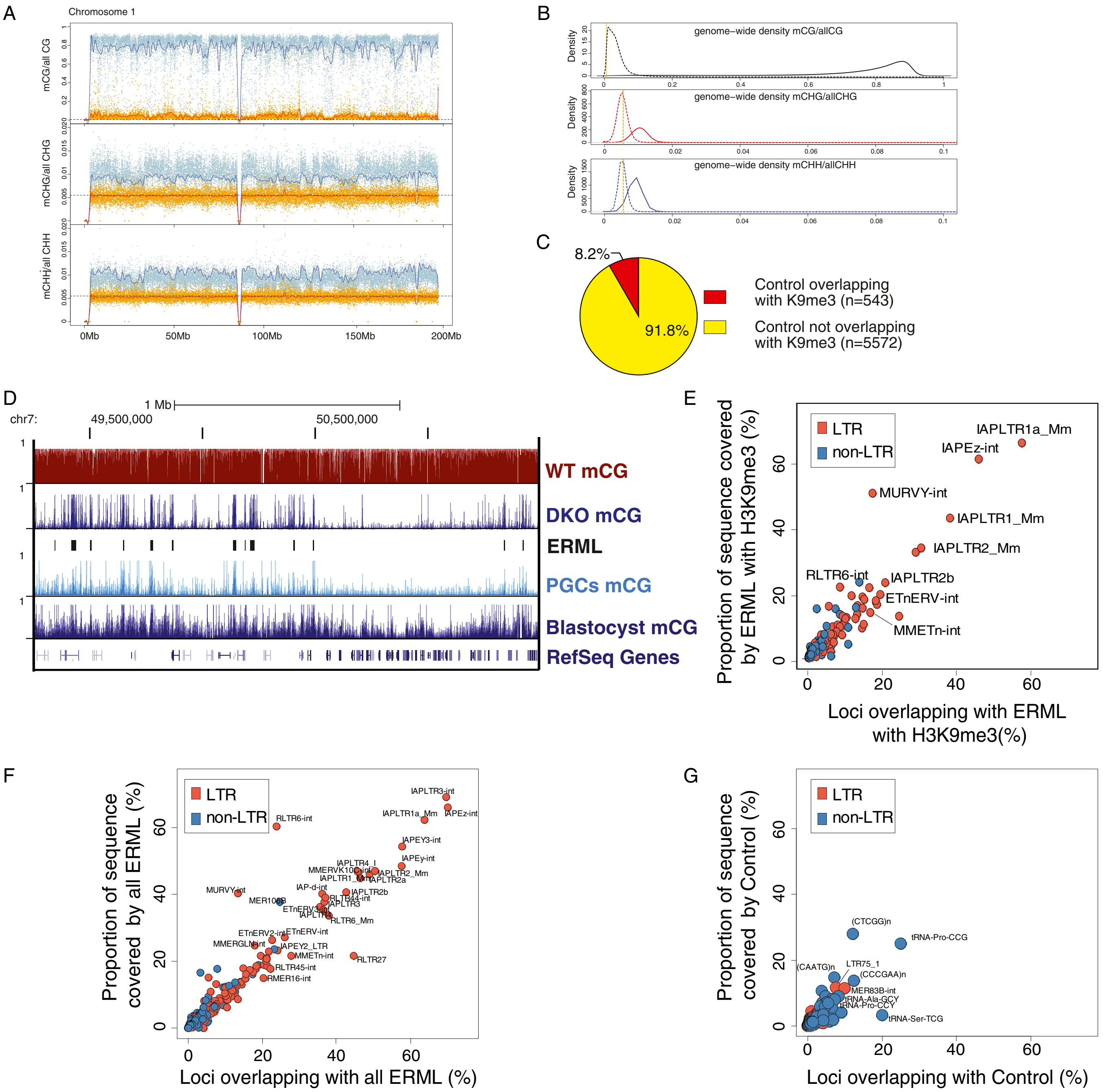

Fig. S1. MethlyC-seq in wild-type and *Dnmt* DKO mESCs (relating to Fig. 1). (*A*) Distribution of mC/all C ratios throughout chromosome 1 (bin size 10 kb) for CG, CHG, and CHH (where H is A, T, or C) context in WT (blue dots) and DKO (orange dots). Dark blue and red lines represent smoothed average values for WT and DKO, respectively. Black dotted lines represent the estimated error rate. (*B*) Genome-wide distribution of mC/all C in the CG, CHG, and CHH contexts for all 10-kb bins of WT mESC (continuous line) and of DKO mESC (dotted line). Vertical orange lines represent the estimated error rate. (*C*) Pie chart demonstrating the overlap between regions enriched for H3K9me3 in both WT and DKO mESCs and shuffled control regions as described in *SI Materials and Methods*. (*D*) UCSC genome browser screen captures illustrating the DNA methylation profiles for WT, DKO mESCs, primordial germ cells (PGCs) and blastocysts across large regions on chromosome 7. DNA methylation at enriched residual methylation loci (ERML) in DKO mESCs, PGCs, and blastocysts appear higher than neighboring regions. (*E*) Scatterplot showing for each repeat subfamily, the percentage of respective occurrences in the genome overlapping with ERML marked by H3K9me3 versus the percentage of base pairs of each repeat subfamily overlapping with ERML marked by H3K9me3 (*n* = 2,451). (*F*) Scatterplot showing for each repeat subfamily, the percentage of respective occurrences in the genome overlapping with all ERMLs versus the proportion of the total sequence in base pairs of each repeat subfamily overlapping with all ERML (*n* = 6,115). (*G*) Scatterplot showing for each repeat subfamily the percentage of respective occurrences in the genome overlapping with all control regions versus the proportion of the total sequence in base pairs of each repeat subfamily overlapping with all control regions (*n* = 2,451).

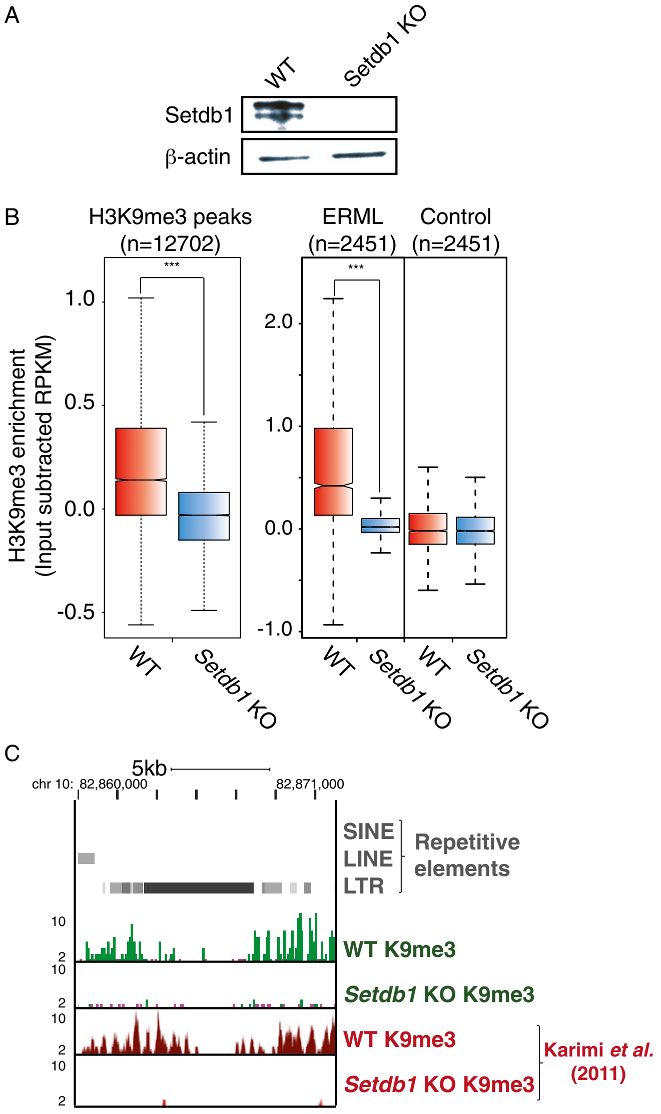

Fig. S2. H3K9me3 profile of *Setdb1* KO mESCs (relating to Fig. 2). (*A*) Western blot confirming efficient depletion of Setdb1 upon 4-hydroxytamoxifen (4-OHT) treatment. (*B*) Boxplot showing median of H3K9me3 enrichment (input subtracted RPKM) in WT and *Setdb1* KO mESCs for H3K9me3 peaks identified in WT mESCs (*Left*), ERML marked with H3K9me3 and control regions (*Right*). (*C*) UCSC genome browser screen capture illustrates that the H3K9me3 ChIP-seq patterns in WT and *Setdb1* KO mESCs generated in this study are consistent with previously published datasets (1).

1. Karimi MM, et al. (2011) DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. *Cell Stem Cell* 8(6): 676–687.

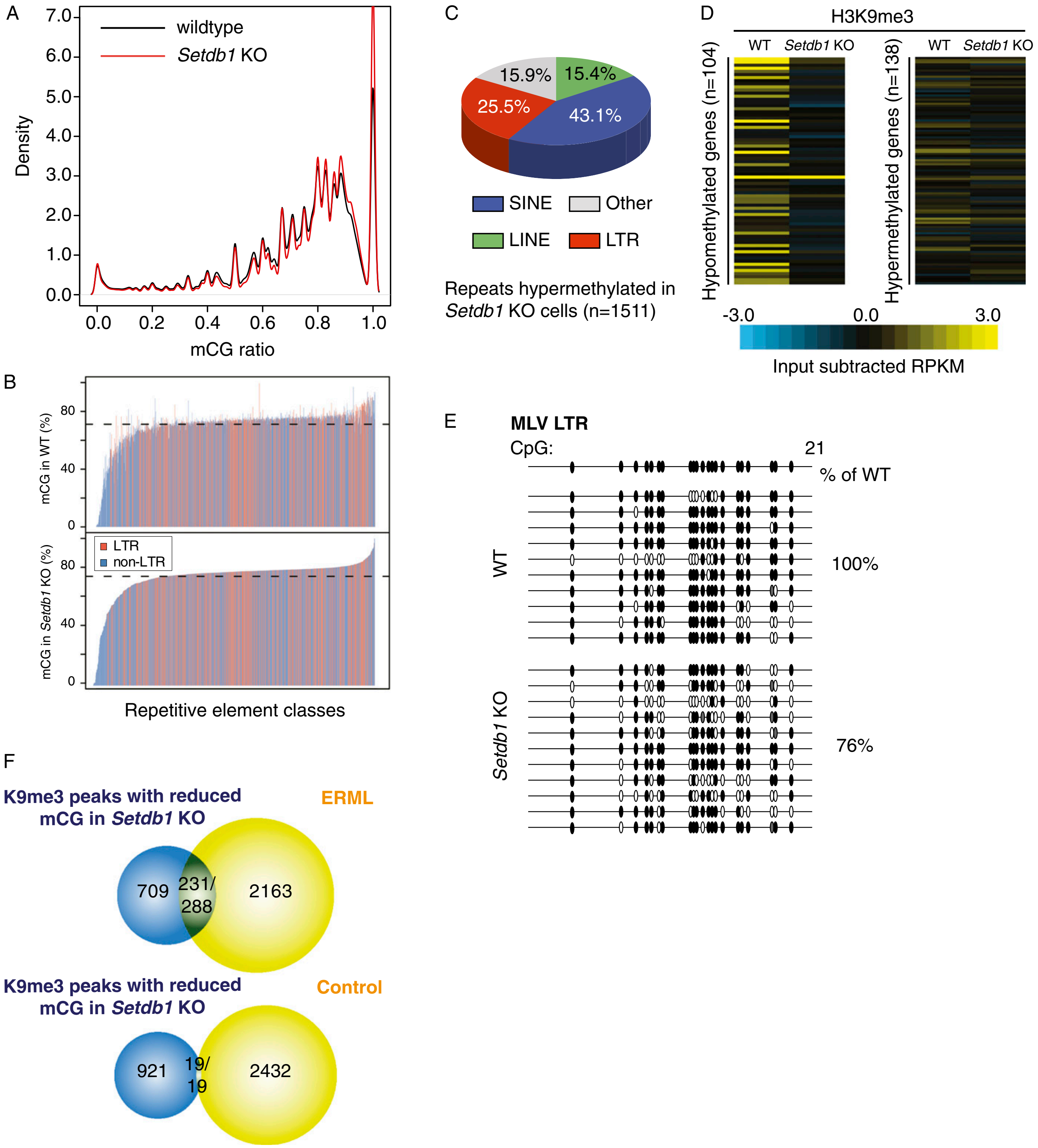

Fig. S3. DNA methylation profile of *Setdb1* KO mESCs (relating to Fig. 2). (*A*) Densityplot shows the distribution of mCG in WT (black) and *Setdb1* KO (red) mESCs. A modest increase of fully methylated regions is found in *Setdb1*-deleted cells. (*B*) Barplot of average mCG levels of mappable LTR (red) and non-LTR (blue) repeat families in WT (*Upper*) and *Setdb1* KO (*Lower*) mESCs. Subsets of repetitive element families show mean hypomethylation, whereas others exhibit mean hypermethylation. (*C*) Pie chart demonstrating the distribution of repetitive element classes among significantly hypermethylated elements (n = 1,511) in *Setdb1* KO mESCs. (*D*) Heat maps demonstrating the enrichment of H3K9me3 of WT and *Setdb1* KO mESCs at genes that become hypomethylated (*Left*) or hypermethylated (*Right*) in the absence of Setdb1. (*E*) Sanger sequencing of bisulfite converted genomic DNA from an independent harvest validates a reduction of mCG at murine leukemia virus (MLV) 5′ LTR upon *Setdb1* deletion. Black, white, and gray ovals represents methylated CpG, unmethylated CpGs, and mutated CpG, respectively. Each horizontal line represents one sequenced molecule. The mean methylation level relative to WT (%) is shown on the *Right*. (*F*) Venn diagrams showing overlap between significantly hypomethylated regions in *Setdb1* KO mESCs and ERML (*Upper*) and control shuffled regions (*Lower*).

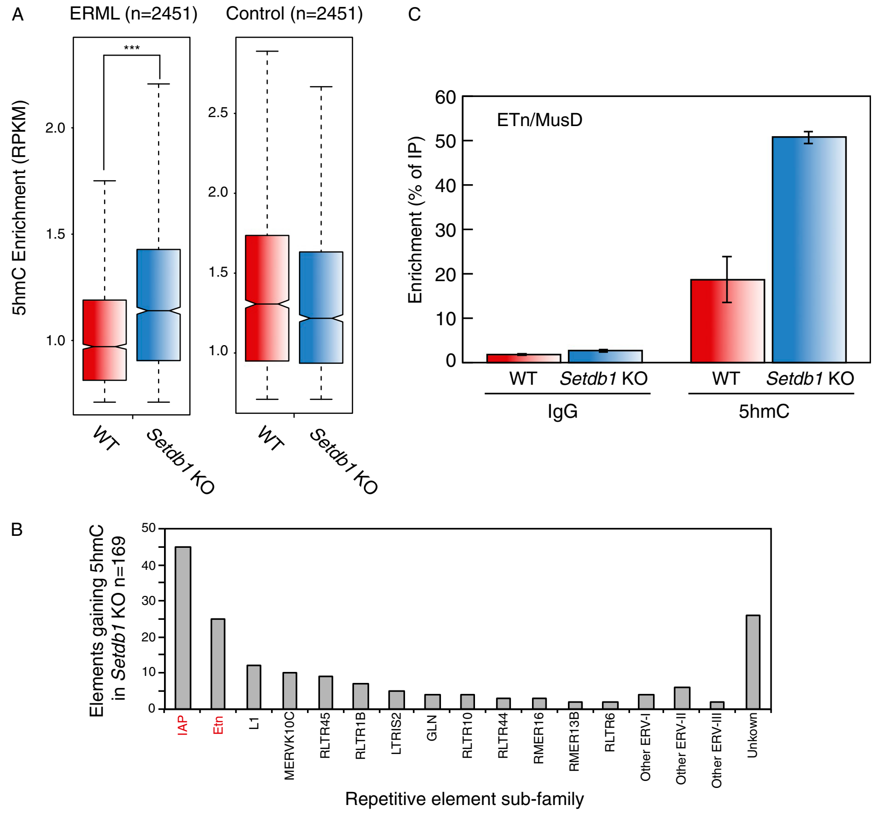

Fig. S4. Description of 5hmC enrichment at ERVs in Setdb1-depleted mESCs (relating to Fig. 3) (*A*) Boxplot showing the 5hmC enrichment (RPKM) at ERML (n = 2,451) and control regions (n = 2,451) in WT and *Setdb1* KO mESCs (***P < 0.001 as comparing WT versus *Setdb1* KO with the Wilcoxon test). (*B*) Barplot showing the distribution of various repetitive element subfamilies among the regions exhibiting the greatest 5hmC increase upon Setdb1 deletion (n = 169). Details on the filtering threshold are described in *SI Materials and Methods*. The same thresholds applied to shuffled 5hmC data yielded fewer than five regions. (*C*) hMeDIP followed by qPCR analyses validate an increase in 5hmC enrichment at early mouse transposon (ETn)/Mus musculus type D retrovirus (MusD) LTR regions in *Setdb1* KO compared with WT mESCs. IgG IP is included to show the background levels of immunoprecipitation. Error bars represent technical replicates.

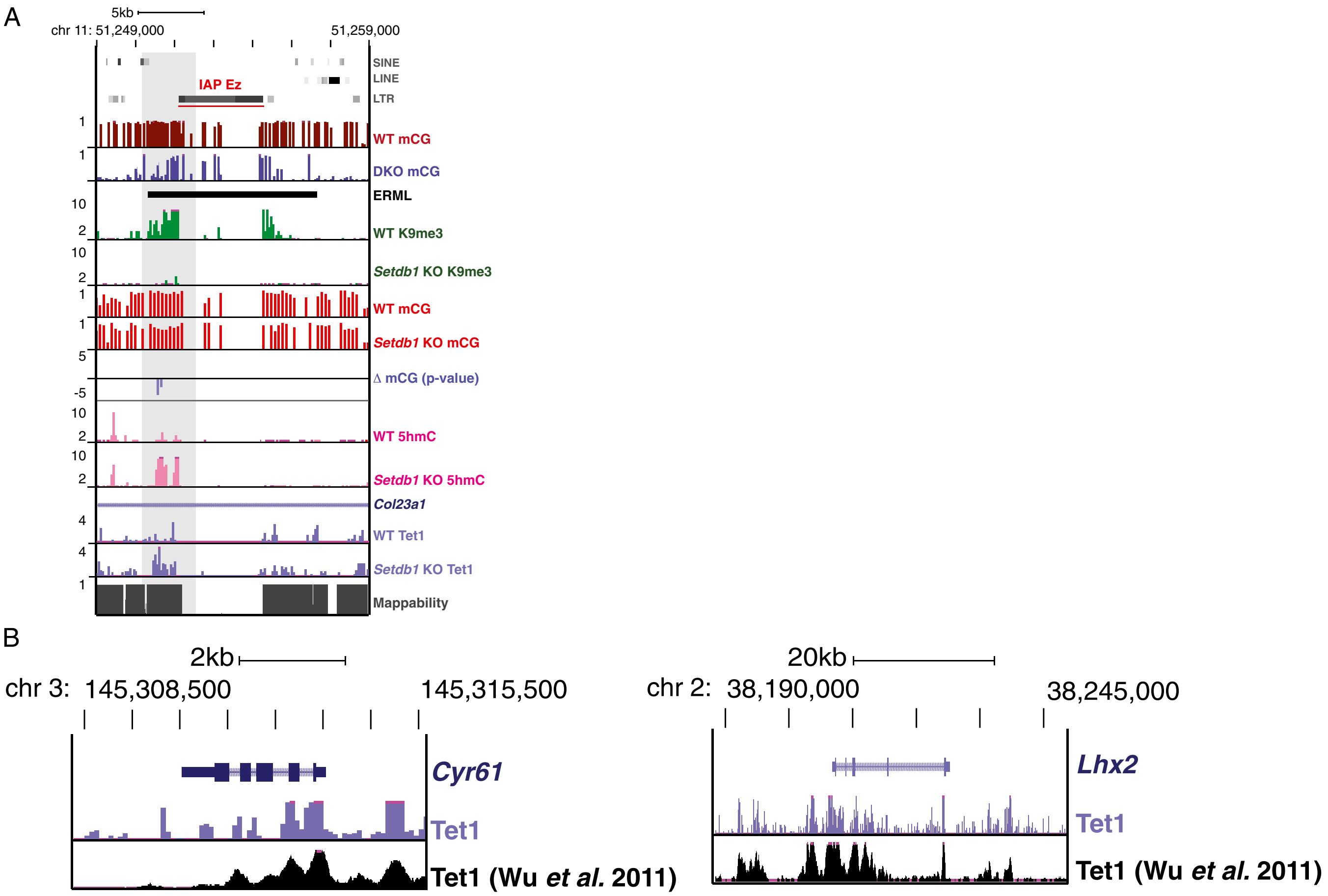

Fig. S5. Description of 5hmC and Tet1 enrichment at endogenous retroviruses (ERVs) in Setdb1 depleted mESCs (relating to Fig. 3) (*A*) Screen capture of a region on chromosome 11. 5hmC and Tet1 profiles are shown as RPKM values between the scales of 2–10 and 0–4, respectively. The example illustrates 5hmC increase (marked by gray shading) at the flanks of an intracisternal A-particle subfamily Ez (IAP-Ez) repeat. Mappability or uniqueness of reference genome from the ENCODE project for 50-bp segments is shown with the score between 0 and 1. (*B*) Screen capture shows similar patterns of Tet1 enrichment at known targets as shown by ChIP-seq in WT mESCs, generated in this study compared with previously published dataset (1).

1. Wu H, et al. (2011) Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature* 473(7347):389–393.

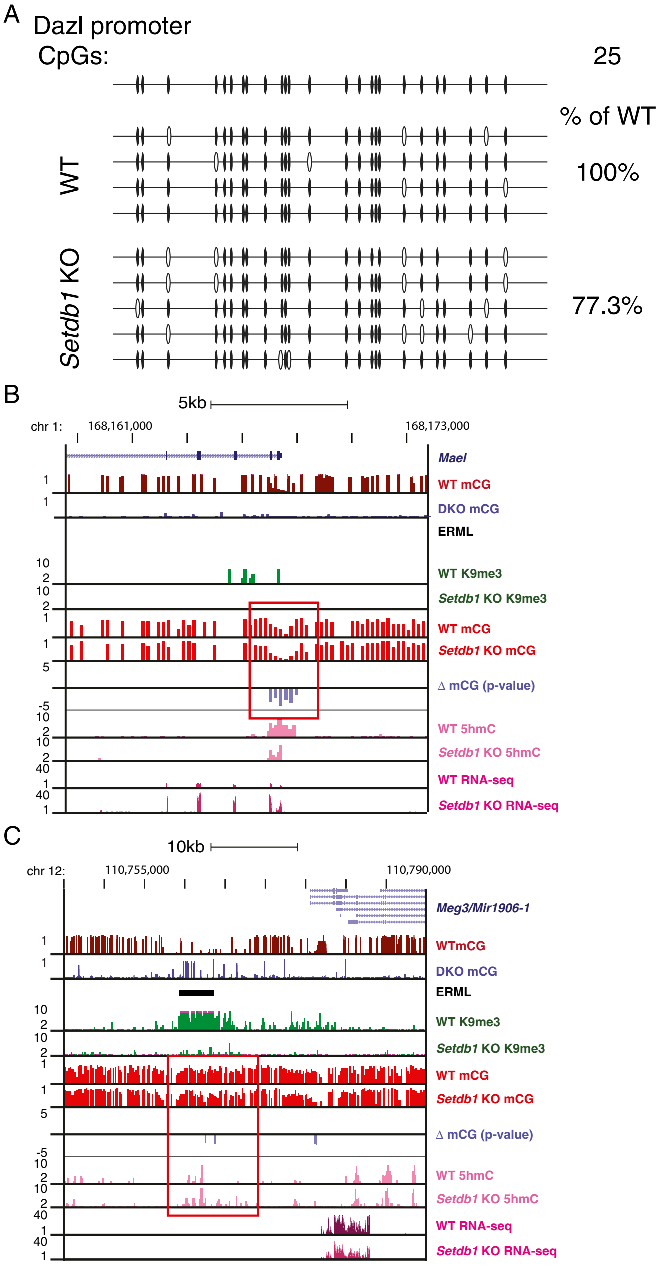

Fig. S6. Setdb1 depletion also resulted in hypomethylation of single copy genes (relating to Fig. 4) (*A*) Sanger sequencing of bisulfite converted genomic DNA showing a modest reduction of mCG at the promoter of the *Dazl* gene in *Setdb1* KO mESCs. (*B*) UCSC genome browser screen capture showing a reduction of mCG (red box) at the promoter of Mael, a testes-specific gene, concomitant with transcriptional up-regulation. (*C*) Screen capture demonstrating hypomethylation of differentially methylated region at *Meg3*, and imprinted gene, in *Setdb1* KO mESCs. This decrease is coupled with an increase in 5hmC (red box).

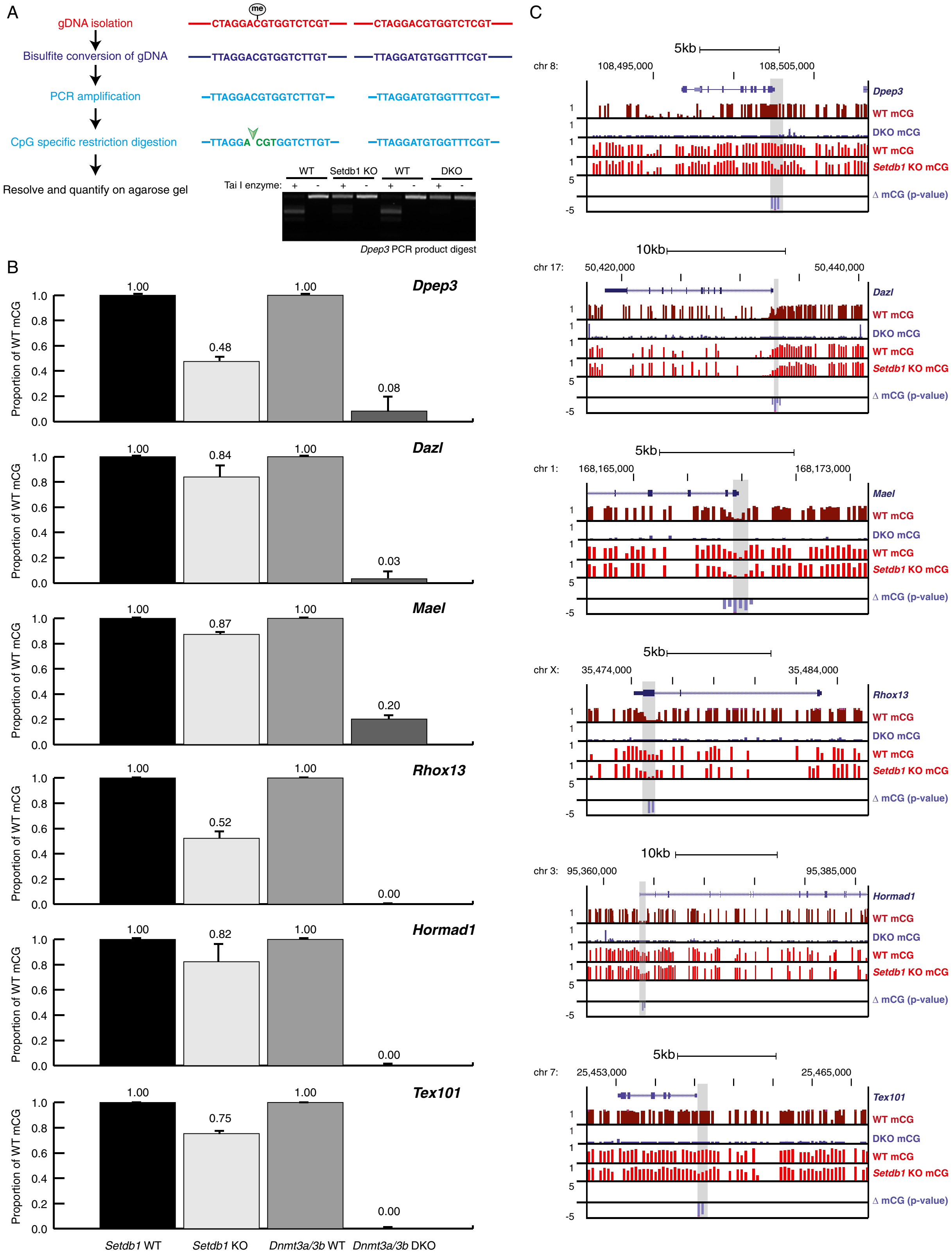

Fig. S7. Validation of single copy genes hypomethylation in Setdb1 KO mESCs by combined bisulfite restriction analysis (COBRA). (*A*) Summary of the procedures in COBRA. Independent harvests of gDNA were bisulfite converted and amplified by PCR. The products would contain a recognition sequence (highlighted in green), which would only be preserved if a give CpG was methylated. The digested products were resolved on agarose gels and quantified. The example image is the digestion of *Dpep3* PCR products by the Tai I endonuclease. (*B*) Quantitation of methylation of single copy developmental or germ-line genes. The raw signals were normalized and presented as a proportion of WT methylation. Error bars reflect SD between replicate experiments. (*C*) Genome browser screen captures showing the DNA methylation of corresponding regions integrated in *B*. Gray shading indicates region targeted by PCR primers.

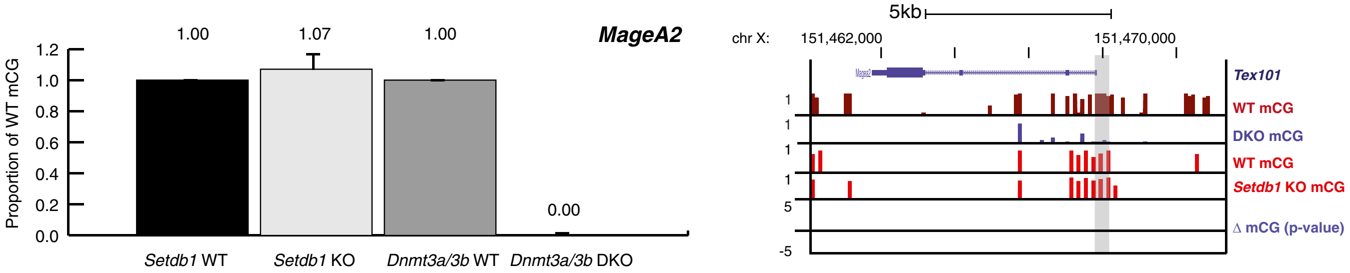

Fig. S8. Validation of no DNA hypomethylation in Setdb1 KO mESCs at negative control gene by COBRA. The *MageA2* gene promoter, which is not marked by K9me3 or shown to be hypomethylated by methylC-seq in *Setdb1* KO cells, is targeted by COBRA as a negative control region. Error bars reflect SD between replicate experiments.

Table S1. Top 20 mappable repeat subfamilies overlapping with ERML

Table S1

Twenty repeat subfamilies with the most significant overlap with ERML are listed. For each subfamily, the percentage of repeat incidences and the percentage of base pairs (bp) overlapping with ERML are shown. In addition, the percentage of ERML incidences overlapping with elements of each repeat subfamily is included.

Table S2. All mappable hypomethylated repeats in *Setdb1* KO mESC

Table S2

The chromosomal location and repeat subclass and subfamily of all mappable repeats that become hypomethylated in Setdb1 KO mESCs (n = 2,395).

Table S3. Sequences and applications of all primers used in this study

Table S3