DNA Methylation and SETDB1/H3K9me3 Regulate Predominantly Distinct Sets of Genes, Retroelements, and Chimeric Transcripts in mESCs

Mohammad M. Karimi, Preeti Goyal, Irina A. Maksakova, Misha Bilenky, Danny Leung, Jie Xin Tang, Yoichi Shinkai, Dixie L. Mager, Steven Jones, Martin Hirst, and Matthew C. Lorincz

INTRODUCTION

DNA methylation and histone H3 lysine 9 trimethylation (H3K9me3) play important roles in silencing of genes and retroelements. However, a comprehensive comparison of genes and repetitive elements repressed by these pathways has not been reported. Here we show that in mouse embryonic stem cells (mESCs), the genes upregulated after deletion of the H3K9 methyltransferase Setdb1 are distinct from those derepressed in mESC deficient in the DNA methyltransferases Dnmt1, Dnmt3a, and Dnmt3b, with the exception of a small number of primarily germline-specific genes. Numerous endogenous retroviruses (ERVs) lose H3K9me3 and are concomitantly derepressed exclusively in SETDB1 knockout mESCs. Strikingly, ~15% of upregulated genes are induced in association with derepression of promoter-proximal ERVs, half in the context of “chimeric” transcripts that initiate within these retroelements and splice to genic exons. Thus, SETDB1 plays a previously unappreciated yet critical role in inhibiting aberrant gene transcription by suppressing the expression of proximal ERVs.

SUMMARY

DNA methylation and histone H3 lysine 9 trimethylation (H3K9me3) play important roles in silencing of genes and retroelements. However, a comprehensive comparison of genes and repetitive elements repressed by these pathways has not been reported. Here we show that in mouse embryonic stem cells (mESCs), the genes upregulated after deletion of the H3K9 methyltransferase Setdb1 are distinct from those derepressed in mESC deficient in the DNA methyltransferases Dnmt1, Dnmt3a, and Dnmt3b, with the exception of a small number of primarily germline-specific genes. Numerous endogenous retroviruses (ERVs) lose H3K9me3 and are concomitantly derepressed exclusively in SETDB1 knockout mESCs. Strikingly, ~15% of upregulated genes are induced in association with derepression of promoter-proximal ERVs, half in the context of “chimeric” transcripts that initiate within these retroelements and splice to genic exons. Thus, SETDB1 plays a previously unappreciated yet critical role in inhibiting aberrant gene transcription by suppressing the expression of proximal ERVs.
Silencing via DNA versus H3K9 Methylation

cells and/or mESCs (Edwards et al., 2010; Yokochi et al., 2009), indicating that H3K9 methylation may generally act independently of DNA methylation to negatively regulate gene expression.

Unlike genic promoter regions, class I and II ERVs are both densely DNA methylated and marked by H3K9me2/3 in mESCs (Dong et al., 2008). Recently, we showed that the H3K9 KMTase Setdb1 (ESET/KMT1E), which plays an important role in stem cell maintenance (Bilodeau et al., 2009; Yuan et al., 2009), is required for H3K9me3 marking and silencing of several ERV subfamilies in mESCs (Matsui et al., 2010). Surprisingly, the overall level of DNA methylation at these ERVs was unchanged or only modestly reduced in Setdb1 conditional knockout (SETDB1 KO) cells. Conversely, H3K9me3 at these elements was not reduced in Dnmt1/Dnmt3a/Dnmt3b triple-knockout (DNMT TKO) mESCs, nor was transcription comparably induced, indicating that SETDB1 functions independently of DNA methylation in these cells. However, a comprehensive genome-wide comparison of the role of SETDB1 versus DNA methylation in transcriptional silencing of genes and ERVs has not been performed.

To identify those genes and/or repetitive elements regulated by DNA methylation and/or SETDB1 genome-wide, and to determine whether SETDB1-mediated deposition of H3K9me3 and associated transcriptional silencing is perturbed in the absence of DNA methylation, we conducted RNA-seq and H3K9me3 native-ChIP (NCHIP)-seq experiments on SETDB1 KO, DNMT TKO, and corresponding wild-type (WT) mESCs. We show that disrupting these two epigenetic pathways results in the derepression of predominantly distinct sets of genes and repetitive elements in mESCs. Furthermore, deletion of Setdb1 leads to widespread reactivation of class I and II ERVs and, unexpectedly, to the aberrant expression of numerous chimeric RNAs that originate in such ERVs and splice to canonical genic exons.

RESULTS

Genome-wide Profiling of Gene Expression in SETDB1 KO and DNMT TKO mESCs

We isolated mRNA from SETDB1 KO (Matsui et al., 2010) and DNMT TKO (Tsurnura et al., 2006) mESCs and their parent lines TT2 and J1, respectively, and performed RNA-seq as described previously (Morin et al., 2008). More than 20M paired-end reads for each cell line were aligned to mouse genome and transcriptome resources (see Experimental Procedures and Figure S1A available online). Several genes within the MageA and Rhox gene clusters reported previously to be DNA methylated and repressed in mESCs, including MageA5, MageA8, Rhox2, and Rhox4 (Fouse et al., 2008; Oda et al., 2006), as well as the MageA4 and Rhox1 genes, were derepressed in the DNMT TKO line (Figures 1A and 1B). None of these genes were derepressed in the SETDB1 KO line. In contrast, the germline-specific gene Dazl was derepressed in both the DNMT TKO and SETDB1 KO lines, whereas the macrophage-specific gene Mmp12 was derepressed exclusively in the SETDB1 KO line (Figures 1C and 1D). Consistent with these observations, derepression of Dazl in both KO lines was validated by qRT-PCR (data not shown), confirming that for a subset of genes, disruption of either pathway is sufficient for transcriptional activation.

To characterize gene expression patterns in the mutant and WT RNA-seq data sets, we generated reads per kilobase per million mapped reads (RPKM) (Mortazavi et al., 2008) values for all annotated exons of ENSEMBL protein-coding genes (22,848 total). In the SETDB1 KO and DNMT TKO lines, 558 (2.4%) and 239 (1.0%) genes were found to be derepressed, respectively, applying combined thresholds based on Z-score (>1.2) and fold-change (≥2) (Figure 1E; Table S1). Although ~17% of genes reported previously via expression microarray to be upregulated >2-fold in a related DNMT-deficient ESC line (Fouse et al., 2008) were also scored as upregulated in our RNA-seq analysis, the majority were not. This probably reflects the greater specificity of high-throughput sequencing (Marioni et al., 2008) under the stringent threshold applied (Figure S1B).

Strikingly, only 7% (39/558) of genes derepressed in the SETDB1 KO line were also derepressed in the DNMT TKO line (Figure 1F; Table S1), and gene ontology (GO) analysis of the genes up- or downregulated in the DNMT TKO and SETDB1 KO lines revealed that none of the GO terms identified are common to both KO lines (Figures S1C and S1D and Table S2). Furthermore, analysis of a recently published DNA methylation data set revealed that only 7% of the promoter regions of genes upregulated in the SETDB1 KO are DNA methylated in the WT TT2 line (Figure 1G; Myant et al., 2011), indicating that SETDB1 and the DNMTs are required for silencing of predominantly distinct sets of genes.

To determine the genome-wide distribution of H3K9me3, and whether this mark is perturbed in the absence of SETDB1 and/or DNA methylation, we conducted NCHIP-seq (O’Neil and Turner, 2003) on the SETDB1 KO and DNMT TKO lines as well as their parent lines with an antibody specific for H3K9me3 (Figures S2A and S2B). ~255 Mb (13%) or ~215 Mb (11%) of the mappable mouse genome (analyzed in 800 bp bins) is marked by H3K9me3 in the TT2 and J1 parent lines, respectively. Although >50% of H3K9me3-marked regions lost this mark in the SETDB1 KO line, only ~15% did so in the DNMT TKO line (data not shown). Fewer than 1% of the 221 genes downregulated in the SETDB1 KO line are marked by H3K9me3 in the promoter region in WT cells (Figures 1H and 1I), implicating SETDB1 predominantly as a transcriptional repressor, as expected. Surprisingly, however, only 13% of the promoter regions of genes upregulated in the SETDB1 KO line are marked by H3K9me3 in the WT line (Figure 1J), revealing that only a minority of induced genes are direct targets of SETDB1.

To focus specifically on direct genomic targets of this H3K9 KMTase, we realigned a previously reported mESC SETDB1 ChIP-seq data set (Yuan et al., 2009) to the genome and identified 20,177 high confidence SETDB1 binding sites via FindPeaks (Fejes et al., 2008). Of these, 67.3% and 64.8% were marked by H3K9me3 in TT2 and J1 ESCs, respectively, comparable to the 65.3% of these sites marked by H3K9me3 in the original study (Figure S2C; Yuan et al., 2009). A 3-way comparison between the parental mESC lines yielded 87%–93% overlap (11,100 common sites) between H3K9me3-enriched regions at SETDB1 binding sites, revealing that our
ChIP-seq data are highly correlated with those generated by Yuan et al. (2009).

Genome-wide analysis of H3K9me3 enrichment at all SETDB1 bound regions, measured in terms of RPKM, revealed that whereas only 9% (1,097/12,782) of H3K9me3-marked sites are lost in the DNMT TKO line, 78% (8,891/11,346) of sites are lost in the SETDB1 KO (Figure 2A). Similarly, analysis of all SETDB1-bound promoter (±500 bp of the TSS) regions marked by H3K9me3 revealed that 11% and 61% lost this mark in the DNMT and SETDB1 KO lines, respectively (Figures 2A and 2B).

Taken together, these results indicate that SETDB1-mediated deposition of H3K9me3 is generally not dependent upon the presence of DNA methylation.

Integration of the RNA-seq and ChIP-seq data sets revealed that SETDB1 is bound to the promoter regions of only 21% (117 of 558) of upregulated genes in WT cells (Figure 2C), confirming that the majority of such genes are induced as a result of downstream effects of SETDB1 loss. Surprisingly, of the 231 genes bound by SETDB1 in their promoter regions that lose H3K9me3 in the SETDB1 KO line, 86% are not upregulated.
Analysis of the promoter regions of 194 of these 198 gene promoters for which DNA methylation data are available (Myant et al., 2011) reveals that only 9.8% are DNA methylated in the TT2 line (Figure 3A), similar to the 9.8% (1,498/15,252) of all ENSEMBL protein coding gene promoters that are methylated. Furthermore, analysis of previously published ChIP-seq data (Mikkelsen et al., 2007) reveals that only 22.7% of these promoter regions are marked by H3K27me3 (Figure 3A). Thus, the majority of these H3K9me3-marked genes are not marked concurrently by DNA methylation or H3K27me3 in WT mESCs.

In contrast, of the 33 genes that lose H3K9me3 and are concomitantly derepressed in the SETDB1 KO line, 40.6% are also DNA methylated in TT2 cells (Figure 3B), prompting us to analyze the 39 genes derepressed in both KO lines (see Figure 1F) in greater detail. Strikingly, 20 of the 30 genes derepressed in both KO lines for which gene expression information is available in the BioGPS database (GNF1M Gene Atlas data set) (Wu et al., 2009) are expressed in testis and/or oocytes (Figure 3C). In TT2 cells, 18 of these germline-specific genes are marked by H3K9me3 in their promoter regions, all but one of which lose this mark in the SETDB1 KO, indicating that they are direct SETDB1 targets. Furthermore, the presence of 13 SNPs in the promoter region of the Tuba3a gene allowed us to confirm that both alleles are marked by H3K9me3. Indeed, 18 and 11 reads from the TT2 H3K9me3 data set definitively mapped to the C57BL/6 and CBA alleles, respectively. The majority of these genes are also DNA methylated in TT2 cells (Myant et al., 2011) and many show reduced H3K9me3 in DNMT TKO cells (Figure 3C). Taken together, these results indicate that DNA methylation and H3K9me3 act in cis at a specific set of germline-specific genes in mESCs and play nonredundant roles in silencing of these genes.

Figure 2. The Majority of SETDB1-Bound Promoters Are Depleted of H3K9me3 in SETDB1 KO but Not DNMT TKO Cells

(A) H3K9me3 RPKM values at genomic (light shading) or promoter (heavy shading) regions bound by SETDB1 are plotted for DNMT TKO versus J1 and SETDB1 KO versus TT2 lines, and the number of genomic sites or promoter regions (in parentheses) losing or gaining H3K9me3 in the KO lines is shown. (B) The number and percentage of SETDB1-bound, H3K9me3-marked promoter regions losing, gaining, or showing no change in H3K9me3 in DNMT TKO and SETDB1 KO lines is shown. (C) The percent and number of all genes or genes bound by SETDB1 in their promoter regions that are upregulated are shown for each KO line. (D) The percent and number of genes with SETDB1-bound promoters that lose H3K9me3 and are upregulated in each KO line are shown. See Figures S1 and S2.
ERVs. Strikingly, although 69 ERV subfamilies were derepressed in the SETDB1 KO line, only 5 were derepressed in the DNMT TKO line (Figure 4A) and 4 of the latter were derepressed to a greater extent in the SETDB1 KO line. Analysis of uniquely mapped reads aligning to the annotated internal regions of all ERV subfamilies with >50 copies in the genome revealed that between 4% and 20% of all genomic copies of ten class I or II ERV subfamilies, including RLTR1B, GLN, ERVK10C, ETn, ETnERV, MMTV, ETnERV2/MusD, RLTR45, IAP-d, and RLTR10, were derepressed in the SETDB1 KO line (Table S3). Reactivation of a subset of these ERVs was confirmed by qRT-PCR (Figure S3A). In contrast, no ERV subfamily showed reactivation of >4% of genomic copies in DNMT TKO cells. Summing the total normalized RNA-seq coverage over “intact” ERVs (annotated internal regions flanked by their cognate LTRs) confirmed that the majority of these elements were significantly derepressed exclusively in the SETDB1 KO line (Figure 4B; Figure 4D). In fact, enrichment of this mark is increased at several ERVs in the DNMT TKO line. Strikingly, analysis of the SETDB1 ChIP-seq data set described above (Yuan et al., 2009) revealed that ~40% of the 20,171 SETDB1 binding sites in the mouse genome overlap with, or occur within, 100 bp of an annotated ERV, a significantly greater number than predicted based on random expectation (Figure 4E). This probably significantly underestimates the true overlap, because ChIP-seq reads that map to sites within multicopy ERVs that show no sequence variation are excluded from the analysis.

We next analyzed the expression and H3K9me3 states of individual full-length ERVs, considering only uniquely aligned reads (Table S4). Analysis of a subset of the class I and II ERV subfamilies derepressed in the SETDB1 KO line revealed a significant increase in expression and loss of H3K9me3 in all cases.

Figure 3. Genes Depleted of Promoter H3K9me3 in the SETDB1 KO Are Generally Not Marked by DNA Methylation or H3K27me3

(A and B) The DNA methylation (Myant et al., 2011) and H3K27me3 (Mikkelsen et al., 2007) states (in WT cells) of genes depleted of H3K9me3 in their promoter regions (TSS ± 500 bp) in the SETDB1 KO line showing no increase (A) or increased (B) expression are shown.

(C) The tissue specificity of genes represented in the BioGPS database that are derepressed in both the SETDB1 KO and DNMT TKO lines (30 of 39 total) is shown, along with the SETDB1 binding (Yuan et al., 2009) (in WT cells), DNA methylation (Myant et al., 2011) (in WT cells), and H3K9me3 states in the promoter regions of these genes (see Figure S2, Tables S1 and S2). Genes highlighted in yellow are expressed in the germline. NA, promoters of MGI gene not represented in the DNA methylation data set.
Plotting expression versus H3K9me3 levels of the parental and KO mESC lines for each of these ERV subfamilies revealed a strong correlation between loss of H3K9me3 and induction of ERV expression (Figure 5B; Figure S5), although not all of the ERVs depleted of H3K9me3 showed increased expression, perhaps because a number of these elements are transcriptionally inert. In contrast, representative class III ERVs and non-LTR LINE1 elements were generally not marked by H3K9me3 (consistent with a previous report by Mikkelsen et al., 2007) nor derepressed in either KO line (Figure 5C; Figure S5).

Because DNA methylation may play a role in maintaining a subset of these elements in a silent state in the absence of H3K9me3, we determined whether simultaneous depletion of DNA methylation and SETDB1 leads to a higher level of ERV reactivation than depletion of SETDB1 alone. Dnmt1 and Setdb1 were targeted via RNAi either alone or in combination, and expression of several ERV subfamilies was monitored via qRT-PCR (Figure 5D). While knockdown (KD) of Setdb1 induced expression of GLN, RLTR4/MLV, ERVK10C, IAPE-z, and in particular ETnERV2/MusD ERVs, KD of Dnmt1 had a relatively modest effect on expression of these proviruses. For each of these subfamilies, simultaneous KD of Setdb1 and Dnmt1 did not increase the level of expression over that observed upon KD of Setdb1 alone, with the exception of the young IAPE-z subfamily, for which the double KD behaves synergistically. Taken together, these data reveal that although SETDB1 plays a role in maintaining a subset of these elements in a silent state in the absence of H3K9me3, DNA methylation may play a role in maintaining a subset of these elements in a silent state in the absence of H3K9me3, we determined whether simultaneous depletion of DNA methylation and SETDB1 leads to a higher level of ERV reactivation than depletion of SETDB1 alone. Dnmt1 and Setdb1 were targeted via RNAi either alone or in combination, and expression of several ERV subfamilies was monitored via qRT-PCR (Figure 5D). While knockdown (KD) of Setdb1 induced expression of GLN, RLTR4/MLV, ERVK10C, IAPE-z, and in particular ETnERV2/MusD ERVs, KD of Dnmt1 had a relatively modest effect on expression of these proviruses. For each of these subfamilies, simultaneous KD of Setdb1 and Dnmt1 did not increase the level of expression over that observed upon KD of Setdb1 alone, with the exception of the young IAPE-z subfamily, for which the double KD behaves synergistically. Taken together, these data reveal that although SETDB1 plays a role in maintaining a subset of these elements in a silent state in the absence of H3K9me3, DNA methylation may play a role in maintaining a subset of these elements in a silent state in the absence of H3K9me3, we determined whether simultaneous depletion of DNA methylation and SETDB1 leads to a higher level of ERV reactivation than depletion of SETDB1 alone. Dnmt1 and Setdb1 were targeted via RNAi either alone or in combination, and expression of several ERV subfamilies was monitored via qRT-PCR (Figure 5D). While knockdown (KD) of Setdb1 induced expression of GLN, RLTR4/MLV, ERVK10C, IAPE-z, and in particular ETnERV2/MusD ERVs, KD of Dnmt1 had a relatively modest effect on expression of these proviruses. For each of these subfamilies, simultaneous KD of Setdb1 and Dnmt1 did not increase the level of expression over that observed upon KD of Setdb1 alone, with the exception of the young IAPE-z subfamily, for which the double KD behaves synergistically. Taken together, these data reveal that although SETDB1 plays...
a dominant role in silencing of class I and II ERVs, for a subset of these elements, DNA methylation provides an additional layer of silencing in the absence of H3K9me3.

**Aberrant ERV Transcription in the SETDB1 KO Line**

The widespread derepression of ERVs in the SETDB1 KO line prompted us to explore the possibility that a subset of the genes showing ectopic transcription were induced as a consequence of derepression of proximal ERVs. We classified all genes based on the absence or presence of an annotated ERV ±5 kb from the annotated TSS and further subdivided the latter on the basis of RNA-seq coverage over these ERVs in the TT2 WT and/or SETDB1 KO lines. Intriguingly, genes 3' of ERVs transcribed in both lines (RNA-seq RPKM > 1) were generally expressed at higher levels than genes lacking an ERV within 5 kb of the TSS or genes in which an ERV is present but not transcribed (coverage <1 RPKM) in either line (Figure 6A). Deletion of Setdb1 had little effect on these relationships. Strikingly, however, 56 of the 261 genes with a promoter-proximal ERV showing a ≥10-fold increase in transcription (and a minimum expression level of RPKM > 1) in the SETDB1 KO line are themselves concomitantly upregulated, representing ~10% of the 558 upregulated genes (shown in Figure 1F and listed in Table S1) in this line. This is significantly greater than the 2.4% of all genes showing an increase in expression in this line (p value < 10^-15), indicating that constitutively expressed ERVs positively influence the expression of proximal genes and that aberrant activation of ERVs may alter the expression of neighboring genes.
Silencing via DNA versus H3K9 Methylation

Figure 6. Increased Genic Expression in SETDB1 KO mESCs Is Associated with Increased Expression of Promoter-Proximal ERVs

(A) Protein coding genes were grouped according to the presence of an annotated ERV within 5 kb of the annotated TSS(s) and then classified solely on the basis of the presence or absence of RNA-seq reads over these promoter-proximal ERVs in the TT2 and/or SETDB1 KO lines. The distribution of RNA-seq coverage (normalized exonic RPKM) is shown in box plot form (median, 25th, and 75th percentiles) for genes with no proximal ERV, along with genes harboring promoter-proximal ERVs that are (1) repressed in both lines (RNA-seq coverage < 1.0 aRPKM); (2) expressed in both lines (RNA-seq coverage 1.0 aRPKM/TT2 aRPKM between .75 and 1.3; or (3) expressed predominantly in the SETDB1 KO line (RNA-seq coverage ≥ 1.0 aRPKM and SETDB1 KO aRPKM/TT2 aRPKM ≥ 10). The number of genes in each category is also shown.

(B) UCSC genome-browser screen shot of the 5' end of the Akr1c21 gene, showing H3K9me3 ChIP-seq and RNA-seq tracks, alignment of the split-paired-end RNA-seq reads in the locus, and ERVs 5' of the gene. See Figure S6.

Figure S7B). To identify chimeric transcripts induced as a result of Setdb1 deletion, we screened for genes showing a ≥ 4-fold increase in such reads in the SETDB1 KO (and an arbitrary minimum of three chimeric reads). Strikingly, we identified 84 such genes, 63 of which show 3 or more chimeric reads in the SETDB1 KO but none in the TT2 line (Table S5). Interestingly, none of these genes showed a ≥ 4-fold increase in chimeric reads in the DNMT TKO line (Table S5). Furthermore, in contrast to the genes associated with constitutive chimeric transcripts, 38 of these genes, representing 6.8% of all upregulated genes in the SETDB1 KO, intersect with the list of genes showing increased expression (as measured by total exonic RNA-seq coverage; Table S1) in this line (Figure S7A). Thirteen of these chimeric genes are among the 56 upregulated genes associated with a derepressed promoter-proximal ERV (identified in Figure 6A), yielding a total of 81 genes upregulated in association with derepression of a nearby ERV. Strikingly, 4 of the top 10 and 17 of the top 100 genes ranked in terms of fold-increase in expression in the SETDB1 KO line are included in this list (Table S1), indicating that genes associated with ERV-initiated chimeric transcripts can be transcribed at very high levels. The annotated ERVs associated with chimeric transcripts are generally truncated elements, indicating that transcription is more likely to extend into flanking genomic sequence when the splice acceptor and/or poly(A) sites of the ERV are deleted. Taken together, these results indicate that transcription from promoter-proximal ERVs can increase mRNA levels of associated downstream genes, frequently in association with the generation of chimeric transcripts.

To further characterize the positive correlation between the number of chimeric paired-end reads detected and the read coverage (across all exons) of associated genes, we analyzed the top 20 genes ordered in terms of the number of chimeric transcripts in the SETDB1 KO line in greater detail (Figure 7A). Intriguingly, although the majority of cognate genic promoters are not marked by H3K9me3 or bound by SETDB1, 16 of the 20 ERVs in which transcription apparently initiates are marked by H3K9me3. Furthermore, many of the ERVs in which transcription of these chimeric mRNAs initiate, such as IP, ETn, and RLR1B elements, are in the same subfamilies of ERVs that are broadly derepressed in SETDB1 KO cells (see Figure 4B; Figure S3).

To validate the existence of these SETDB1 KO-dependent chimeric transcripts, RT-PCR was conducted with primers of elements upstream of the TSS and the other to the 5’ end of the second annotated exon (Figure 6B). Similar observations were made for the Angpt6 and Cyp2b23 loci (Figure S6). To identify additional chimeric mRNAs, we surveyed paired-end RNA-seq reads for the presence of individual transcripts with one of the mate-pair reads aligning to an ERV and the other to an annotated genic exon. Numerous genes with such chimeric reads were found in all four cell lines (Figure S7A; Table S5). Analysis of the 117 genes associated with such constitutive chimeric transcripts in the TT2 and SETDB1 KO lines revealed that the genic expression levels (RPKM over annotated exons) were similar in most cases, with only five of these genes showing increased expression in the SETDB1 KO line. Furthermore, the ERVs identified were generally distinct from those derepressed in the SETDB1 KO line (compare Figure 4 and Figure S3 with Figure 4B).

Surprisingly, inspection of paired-end read alignments at several such genes revealed the presence of numerous “chimeric” transcripts (Peaston et al., 2004; van de Lagemaat et al., 2003), with one mate pair read mapping within the promoter-proximal ERV and the other within an annotated genic exon. For example, 20 paired-end reads in the Akr1c21 locus show one mate-pair read mapping to an ERV within a cluster...
Silencing via DNA versus H3K9 Methylation

Figure 7. Chimeric Transcripts Initiating in LTR Elements 5' of Genic TSSs and Splicing to Canonical Genic Exons Detected Exclusively in the SETDB1 KO Line

(A) Genes with one paired-end read mapping to an annotated ERV and the other to a genic exon were identified. The top 20 genes in the SETDB1 KO, in terms of the number of chimeric reads identified, are shown, along with RNA-seq coverage over genic exons. Annotation of the ERV in which transcription initiates, the orientation of the ERV in relation to the gene, and the presence of SETDB1 or H3K9me3 in the ERV or at the 5' end of the gene are also shown. Stars indicate those subclASSES of ERVs that are broadly reactivated.

(B) The presence of chimeric transcripts of the Akr1c21, Angptl6, Gm1110, Mep1b, and Cyp2b23 genes exclusively in the SETDB1 KO line was validated by RT-PCR with primers (arrows) designed within the 50 bp regions to which the chimeric paired-end reads aligned. β-actin was used as a control.

(C) Amplicons were cloned and sequenced and the structure of the chimeric paired-end reads of five of these genes (Figure 7B), including Akr1c21, Angptl6, Gm1110, Mep1b, and Cyp2b23. As expected, PCR products were observed only in the SETDB1 KO. The upstream sequences of these transcripts, including any cryptic splice site junctions, were subsequently determined by Sanger sequencing, confirming that they frequently splice from cryptic splice donor sites embedded in an ERV itself or in 5' flanking genomic DNA, to 5' genic splice acceptor sites (Figure 7C). Analysis of the coding potential of these transcripts reveals that the complete native ORF of only the Angptl6 gene is retained, but a cryptic upstream ORF is also encoded which probably precludes the expression of the Angptl6 protein. To establish the coding potential of the remaining chimeric genes, we carried out ab initio transcript assembly via Cufflinks (Trapnell et al., 2010). Of the 38 induced chimeric genes, transcript modeling revealed that 20 initiate in an ERV and extend to a genic exon, 13 of which are associated with genes upregulated in the SETDB1 KO (Table S5). Although nine of these chimeric transcripts encode the native genic ORF, only three, CD209c, 2810474O19Rik, and 2010005H15Rik, do not also encode a cryptic upstream ORF. Thus, paradoxically, although the level of transcript over genic exons is dramatically upregulated for a number of the chimeric transcripts identified, translation of the native ORF is likely to be reduced for the majority of constitutively expressed genes associated with chimeric transcripts, because of the presence of cryptic upstream ORFs.

DISCUSSION

DNA methylation and posttranslational histone modifications are highly dynamic epigenetic marks, particularly early in development, when transcriptional networks undergo reprogramming associated with differentiation. A recent microarray study revealed that the majority of genes derepressed in the absence of DNA methylation are not marked by H3K27me3, suggesting that DNA methylation acts independent of H3K27me3 to maintain genes in a silent state (Fouse et al., 2008). Here, we show that the majority of genes derepressed in the absence of DNA methylation are not derepressed in the absence of SETDB1/H3K9me3, and vice versa. Genes in the MageA and Rhox clusters that are reactivated in the DNMT TKO line, for example, are not marked by H3K9me3 in WT cells nor reactivated in the SETDB1 KO line, while genes that are reactivated in the SETDB1 KO line are generally not DNA methylated in WT cells, or reactivated in the DNMT TKO line. In contrast, a relatively small number of predominantly germine-specific genes are DNA methylated and marked by H3K9me3 in their promoter regions and derepressed in both KO lines. Why H3K9me3 and DNA methylation marks are required to repress these germline-specific genes remains unknown but may reflect the expression of multiple transcriptional activators that can act independently to promote transcription, each of which must be inhibited by one or the other of these pathways to maintain their promoter regions in an inaccessible state.

Genome-wide reactivation of class I and class II ERVs in mESCs lacking SETDB1 but not DNA methylation confirms our previous qRT-PCR and northern blotting-based observations...
(Matsui et al., 2010) and is consistent with recent reports showing that many of the same ERV families are derepressed in mESCs and blastocysts deficient in the SETDB1 binding partner KAP1 (Rowe et al., 2010) but not in two independently derived DNMT TKO lines (Hutnick et al., 2010; Matsui et al., 2010; Tsumura et al., 2006). Taken together, these data clearly show that while DNA methylation may be critical for silencing of these ERVs in somatic cells and at specific stages in germline development (Walsh et al., 1998), an alternative silencing pathway maintains these elements in a silent state in mESCs and early in embryonic development. The relatively high turnover of DNA methylation in primordial germ cells and in the early embryo (Morgan et al., 2005) may reduce the efficacy of this pathway at these stages. Regardless, given that newly retrotransposed ETn and IAP ERVs are responsible for a significant number of mouse germline mutations (Maksakova et al., 2006), at least some of these elements are clearly capable of evading the host silencing machinery in the germline or early in embryonic development.

Exogenous (Bushman et al., 2005; Lewinski et al., 2006) and young endogenous (Medstrand et al., 2002) viruses generally integrate within or near genes. However, given their propensity to interfere with gene expression, ERVs are generally excluded from genes and adjacent regions by natural selection (Medstrand et al., 2002). Nevertheless, perhaps because of their high transcriptional activity, a number of ERVs have been domesticated to provide new regulatory elements for tissue- or cell-specific expression of developmentally regulated genes (van de Lagemaat et al., 2003). MT and MuERV-L class III ERVs, for example, are highly expressed in oocytes and 2-cell embryos and drive expression of chimeric transcripts that comprise 14% and 3% of all ESTs at these stages, respectively (Peaston et al., 2004). Moreover, a recent genome-wide analysis of cap-selected mouse and human transcripts from different tissues and developmental stages revealed that up to 30% of transcripts initiate within repetitive elements, many of them tissue specific (Faulkner et al., 2009).

Our genome-wide analyses revealed a number of chimeric mRNAs expressed predominantly in SETDB1-deficient cells that are initiated primarily by the same subclasses of class I and II ERVs that are broadly derepressed in these cells. The majority of ERVs in which these chimeric transcripts initiate (15 of 21) are in the sense orientation, consistent with a previous report showing that promoter-proximal LTR elements are more likely to be used as gene promoters when in the sense orientation (Dunn et al., 2005). While these chimeric transcripts are not detected in WT mESCs because of SETDB1-mediated silencing, 11 of the top 20 chimeric transcripts identified in the SETDB1 KO line, including the Cyp2b23, Mmp12, Angptl6, and Mep1b chimeras, are expressed in a subset of normal and/or tumor tissues, according to the AceView cDNA database (Table S5; Thierry-Mieg and Thierry-Mieg, 2006), indicating that silencing of a number of the ERV-initiated chimeric transcripts, while robust in mESCs, is relaxed in other tissues. Intriguingly, aberrant expression of several ERV-initiated proto-oncogenes is linked to transformation in mice (Howard et al., 2008; Lee et al., 1999) and humans (Lamprecht et al., 2010).

Further evidence for ERV-mediated perturbation of gene expression comes from studies of mouse mutants harboring novel ERV insertions (Druker et al., 2004; Duhl et al., 1994; Maksakova et al., 2006; Vasicek et al., 1997). The most well-known example is the Aα epiallele, an epimutation resulting from the insertion of an IAP element in a pseudoxon upstream of the Agouti gene (Waterland and Jirtle, 2003). A cryptic promoter in the IAP element promotes constitutive ectopic expression of a chimeric transcript consisting of a novel IAP 5’LTR-encoded exon spliced to the canonical splice acceptor site of exon 2 of the Agouti gene (Duhl et al., 1994). This chimeric mRNA encodes a functional Agouti protein, the aberrant expression of which leads to yellow fur, obesity, and tumorigenesis in the Aα mouse at non-Mendelian ratios. Another example involves a distinct IAP insertion in the Pcdax v8 gene, which results in reduced expression of this gene in brain tissue due to DNA methylation of the IAP element. Strikingly, Pcdax v8 expression is induced more than 100-fold in neuroblastoma cell lines in association with upregulation of this IAP element (Sugino et al., 2004).

In summary, we find that the widespread reactivation of class I and class II ERVs triggered exclusively by Setdb1 deletion is accompanied by the expression of novel ERV-initiated genic transcripts. Although many of these chimeric transcripts encode novel ORFs upstream of the canonical genic ORF that probably preclude expression of the native protein, the regulatory elements within these ERVs may represent a reservoir of alternative promoters that have the potential to be domesticated, should they confer a selective advantage. Regardless, the results presented here clearly reveal that SETDB1 not only is required for silencing of a subset of genes in mESCs, but also plays a critical role in protecting the integrity of the transcriptome in these cells by inhibiting the aberrant expression of ERVs and ERV-initiated transcripts that splice to genic exons.

**EXPERIMENTAL PROCEDURES**

**Cell Culture, RNA Isolation, qRT-PCR, and RNA-seq**

J1 and TT2 mESCs were passaged as described (Matsui et al., 2010). RNA was isolated with the GenElute Kit (Sigma-Aldridge). For RT analysis, RNA was reverse transcribed with SuperScript III (Invitrogen) as per the manufacturer’s instructions, and qRT-PCR was carried out with SsoFAST EvaGreen Supermix (BioRad) on StepOne Software v2.1 (Applied Biosystems). RNA-seq libraries were constructed from mRNA as described in Morin et al. (2008) from 10 μg of DNasel-treated total RNA and paired-end sequencing performed on an Illumina Genome Analyzer, according to the recommended protocol (Illumina Inc., Hayward, CA). Sequence reads were aligned to the mouse reference genome (mm9) using MAQ v0.7.1 (Li et al., 2008) with Smith-Waterman alignment disabled and annotated exon-exon junctions compiled from Ensemble (Flicek et al., 2010), RefSeq (Prutt and Maglott, 2001), and UCSC (Rhead et al., 2010) (downloaded from http://genome.ucsc.edu on 03/17/09). Oligonucleotide sequences used in RNAi and PCR experiments are listed in Supplemental Experimental Procedures.

**NChIP, Data Normalization, RPKM, and Z-Score**

NChIP was conducted as described in the Supplemental Information. To compare expression and H3K9me3 coverage levels across samples, we calculated RPKM values in regions of interest for both RNA-seq and NChIP-seq samples (Mortazavi et al., 2008), as described in detail in the Supplemental Information. For pairwise sample comparisons, a Z-score was calculated assuming that the distribution of read coverage for each sample follows a Poisson model, as described in detail in the Supplemental Information.
SetDB1 contributes to repression of genes encoding developmental genes in mouse embryonic stem cells. We thank M. Okano for Dmnt TKO mESCs and T. Jenuwein for Suv39h DKO mESCs. We are also grateful to Jim Bone, Bryeanna Villani, Nina Thiessen, and to members of the GSC Functional Genomics and Sequence Production groups and M.C.L. lab for technical support. We thank Keji Zhao, Ting Wang, Irina Stancheva, and Dirk Schübeler for helpful discussions. This work was supported by grants from the Genome Network Project from MEXT of Japan and supported by grants from the Genome Network Project from MEXT of Japan to Y.S. and CIHR grants 77805 to M.C.L. and 92093 to M.C.L. and M.H. M.C.L. is an MSFHR Scholar and a CIHR New Investigator.

Received: October 6, 2010
Revised: February 18, 2011
Accepted: April 4, 2011
Published: June 2, 2011

REFERENCES


Cell Stem Cell

Silencing via DNA versus H3K9 Methylation


**Note Added in Proof**

While this manuscript was under review, Macfarlan et al. (Macfarlan, T.S., Gifford, W.D., Agarwal, S., Driscoll, S., Lettieri, K., Wang, J., Andrews, S.E., Franco, L., Rosenfeld, M.G., Ren, B., and Pfaff, S.L. (2011). Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. Genes Dev. 25, 594–607. Published online February 28, 2011. 10.1101/gad.2008511) reported that the H3K4 demethylase LSD1 represses expression of class III ERVs and neighboring genes. Taken together with our observations, these results reveal that distinct chromatin-based pathways are responsible for silencing of class I and II versus class III ERVs.