Transposable elements are regulated by context-specific patterns of chromatin marks in mouse embryonic stem cells

The majority of mammalian genomes are devoted to transposable elements (TEs). Whilst TEs are increasingly recognized for their important biological functions, they are a potential danger to genomic stability and are carefully regulated by the epigenetic system. However, the full complexity of this regulatory system is not understood. Here, using mouse embryonic stem cells, we show that TEs are suppressed by heterochromatic marks like H3K9me3, and are also labelled by all major types of chromatin modification in complex patterns, including bivalent activatory and repressive marks. We identified 29 epigenetic modifiers that significantly deregulated at least one type of TE. The loss of Setdb1, Ncor2, Rnf2, Kat5, Prmt5, Uhrf1, and Rrp8 caused widespread changes in TE expression and chromatin accessibility. These effects were context-specific, with different chromatin modifiers regulating the expression and chromatin accessibility of specific subsets of TEs. Our work reveals the complex patterns of epigenetic regulation of TEs.
The modification of histones is an elaborate system to regulate gene expression, and provides an epigenetic landscape for the cell-type-specific interpretation of the genome. Yet, the major class of genomic elements in the cellular genome are not genes, but transposable elements (TEs), including endogenous retroviruses (ERVs). TEs were originally thought of as genetic parasites with roles in human disease, but TEs are now understood to contribute to normal biological processes. There are many examples of exapted TEs that have become host cell genes, such as the RAG enzymes which are essential for antibody and T cell receptor recombination, or the Synctin genes which are involved in placental development. TEs can be transcribed to produce RNA, and have contributed to the evolution of long non-coding RNAs, microRNAs, and circular RNAs. In addition, TEs often mimic host cell functions by incorporating cis-regulatory elements, which can recruit transcription factors (TFs) to promote TE activity. This was initially observed for the TE repressor REST, but has since been observed for a wide range of TEs. TEs copy themselves in the genome and are co-opted to form new regulatory elements, and contribute to the rewiring of gene regulatory networks. However, much of this data on exaptation of TEs is derived from genomic data and there is argument over how much is functional.

TEs/ERVs are silenced by a range of molecular mechanisms, including heterochromatin formation, mRNA editing, and DNA methylation. DNA methylation is thought to be the dominant suppressive mechanism in somatic tissues. However, DNA is globally demethylated during early embryonic development, and TEs are released from repression in a controlled, stage-specific manner. The TEs are then free to compete with the epigenetic suppression mechanisms to duplicate themselves and enter the germ line. Consequently, there is a delicate balance between the beneficial effects of TEs, and their deleterious effects on genome integrity.

TEs are suppressed in embryonic cells in a process that is well described for ERVs. Zinc-finger proteins (ZFPs) bind to specific sequences in ERVs and histone H3K9me3 methyltransferase SETDB1 to silence TEs. In addition to H3K9me3, there are other modes of epigenetic suppression of TEs, such as the methylation of H4K20me3 and H3K27me3, the biotinylation and sumoylation of H2A, H3, and H4 histones, and the deposition of the histone variant H3.3. It is clear that the epigenetic system is regulating TEs, however, there are at least 1100 distinct types of TE, comprising millions of genomic copies, for which the epigenetic regulation is unclear.

Here, we reveal that TEs are marked by chromatin modifications in complex patterns. Of the 32 chromatin marks we explored, 22 were enriched on at least one type of TE. We found evidence not only of repressive marks, but widespread marking of TEs by activatory marks, including bivalent marking of TEs by repressive H3K9me3 and activatory H3K27ac chromatin marks. When we knocked down a panel of chromatin regulators, 29 of 41 led to the deregulation of expression of at least one type of TE. The loss of Setdb1, Ncor2, Rnf2, Kat5, Prmt5, Uhrf1, and Rrp8 led to widespread changes in the expression of TEs, and a corresponding change in chromatin accessibility. Finally, we explore the consequences of these observations by showing that the loss of chromatin modifiers induces a gene expression program similar to totipotent 2 cell (2C) embryos, characterized by the upregulation of MERVL TEs and 2C-specific genes. Overall, our data suggests that the chromatin modifying system manages the expression of TEs by context-specific deposition of chromatin marks that regulate the chromatin state and expression level of TEs.

Results

TEs are marked by overlapping patterns of chromatin marks. To explore the chromatin code for the suppression of TEs, we uniformly reanalyzed mouse ESC ChIP/epigenetic-seq data for 32 chromatin marks, including histone methylation, acetylation, ubiquitination, and variant histones, along with DNA 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC), and Assay for Transposase-Accessible Chromatin (ATAC)-seq (Supplementary Data 1). Data was quality controlled by removing poorly correlating biological replicates (See Methods and Supplementary Figure 1a). A concern in the analysis of repeats is the use of multimapped reads. In our analysis, for multimapped reads only the highest scoring alignment was kept, and was randomly assigned to a TE copy. To reduce bias due to random distribution of reads across TE copies we considered TEs as metagenes that describe the entire set of TE copies across the genome. Below, we show some example genomic views, but it should be kept in mind that the analysis of individual TE copies can be ambiguous.

The ChIP-seq data comes from serum + LIF grown ESCs, which have relatively high levels of DNA methylation, particularly at Intracisternal A-particle (IAP) TEs. However, many TEs have demethylated DNA, and ESCs use DNA methylation-independent mechanisms to suppress TEs. Knockout of all three DNA methyltransferases led to upregulation of mostly IAP-family TEs only, and this upregulation is rapidly compensated for by small RNA and polycistrome-mediated repression. Indeed, only 65/679 TE types had >2-fold enrichment of 5mC DNA methylation, and most of the methylated TEs were IAP or IAP LTRs (Long terminal-repeats), and a mixture of other ERVs, for example, BGLLI_Mus (Fig. 1a, b, Supplementary Figure 1b and Supplementary Data 2).

The second major mechanism for the regulation of TEs in ESCs is the formation of heterochromatin, mediated by H3K9me3 and H4K20me3. We found that 87/679 TE types were >2-fold-enriched with H3K9me3, and 72/679 were marked by H4K20me3. H3K9me3 and H4K20me3 tended to co-occur, and they both marked 58/679 TE types (Supplementary Data 2). Examples are shown for the ERVK IAPLTR4_I, and IAPEY-int which were enriched for H3K9me3 and H4K20me3.

Of the remaining TEs, some had simple patterns of chromatin modifications, such as MERVL TEs that were enriched only for H3K56ac (Fig. 1b). In total 41/679 TE types were enriched with a single chromatin mark (Supplementary Data 2). The most common pattern was for multiple chromatin marks on the TEs, as 117/679 had 2 or more chromatin marks, and, at a relaxed enrichment threshold of >1.5 fold, 198/679 TE types were enriched with at least 2 chromatin marks (Supplementary Data 2). Examples include RLRTETN_Mmm, which was enriched for 14 of the 32 chromatin marks we profiled (Fig. 1a, b). Interestingly, TEs were marked by chromatin patterns associated with other genomic features. Many TEs were enriched with chromatin marks indicative of enhancers, such as H3K4me1 which marked 23/679 TE types. For example, a 3.0 fold-enrichment of H3K4me1 was found on the ERVK RLTR9D2 (Supplementary Data 2). The promoter mark H3K4me3 was enriched at 14/679 TEs, including the LINE (long interspersed nuclear element) L1Md_Gf (7.2-fold-enriched), and the ERVK RLRTETN_Mmm (2.9 fold-enriched) (Fig. 1a, b and Supplementary Data 2).

ERVKs/ERVs were enriched with activatory chromatin marks, including histone acetylations, such as H3K27ac (RLTR13G >6.2-fold-enriched) and H2B120ac (MMERGLN_LTR >4.6 fold-enriched) (Fig. 1a, c). In total 75/679 TE types were marked by one or more acetylated histone mark and 42/679 by H3K27ac. To gain a global view of TE chromatin, we clustered the patterns of chromatin marks on TEs...
The full heatmap and table for all TE types is in Supplementary Figure 1b and Supplementary Data 2. Those TE types > 4 fold-enriched are indicated with a cyan asterisk. See Supplementary Data 1 for detailed sample information.

A chromatin mark was considered enriched if it was > 2-fold enriched over background. Fold-enrichment over background was calculated by measuring the RPKM for each TE type, taking the maximum binned read count tag density pileup (in RPKM) of the indicated TE types. The RPKM (reads per kilobase of TE per million library sequence reads) was measured for all copies of that TE type, and each TE type was treated as a metagene. Fold-enrichment over background was calculated by measuring the RPKM for each TE type, taking the maximum binned read count tag density pileup (In RPKM) of the indicated TE types. The RPKMs were scaled to the same size, and the flanking 5 kb regions are shown. n = indicates the number of TEs used in that pileup, for this and all subsequent pileup figures. c,d Native-ChIP-qPCR using antibodies against H3K27ac and H3K9me3 for selected TEs, measured as percent of input. The location of the primer pairs are indicated with red lines in the genomic views in (c). See Supplementary Data 7 for the primer pair sequences. Genomic locations are from the mm10 mouse assembly, for this and all other genomic views. The scale of the profiles ranges from a baseline of 0 up to the indicated value, in normalized arbitrary units, for this and all subsequent genome views. Native-ChIP-qPCR (d) was performed three times in biological replicate each with three technical replicates, the open circles indicate the mean of the technical replicates, the red bar is the mean of the biological replicates, and the error bars are the standard error of the mean. Source data are provided as a Source Data file.

Fig. 1 TEs are marked by TE-type-specific epigenetic signatures. a Heatmap of 32 chromatin marks in normal serum-grown ESCs and their fold-enrichment at the indicated TE types. The RPKM (reads per kilobase of TE per million library sequence reads) was measured for all copies of that TE type, and each TE type was treated as a metagene. Fold-enrichment over background was calculated by measuring the RPKM for each TE type, taking the maximum binned value, and comparing this to size matched, randomly selected genomic regions. A chromatin mark was considered enriched if it was > 2-fold enriched (marked with an asterisk). Those TE types > 4 fold-enriched are indicated with a cyan asterisk. See Supplementary Data 1 for detailed sample information.

( Supplementary Figure 1c). We could identify 6 major clusters of chromatin marks, including: a heterochromatic-cluster (H4K20me3, H3K9me3, H3.3), a H4Rme2-cluster (H4R3me2, 5mC, H1D, H1C), a transcriptional elongation-cluster (H3K36me3, H3K36me2), a variant/histone–polycomb-cluster (H3K27me3, H3K27me2, H3.1 and H3.2), a promoter-cluster (H3K4me3, H3K9ac), and an enhancer-like cluster (H3K4me1, H3K27ac, ATAC, H2B20ac) (Supplementary Figure 1c). Overall, of the 32 chromatin marks we analyzed 22 were enriched >2 fold on at least one TE type, and 27 were enriched >1.5 fold (Supplementary Data 2), indicating broad and complex patterns of chromatin marks on TEs.

LINE elements are a special case as they occupy 18% of the mouse genome, and are increasingly recognized for their important roles in embryogenesis and chromatin...
The analysis of LINE TEs suggests that they are mainly enriched with H4R3me2 (Fig. 1a and Supplementary Data 3). However, LINEs can be divided into longer near-identical elements and shorter truncated versions. We split the LINE1 elements into long (>5 kb) and short (<5 kb), and measured the chromatin mark enrichment at the two classes of LINEs. Whilst both long and short LINEs were enriched for H4R3me2, the long LINEs were enriched for 15/32 chromatin marks (Supplementary Figure 1d). This included chromatin marks representing many genomic elements, including enhancers (H3K4me1, H3K27ac), promoters (H3K3me3), activatory regions (acetylated histones), DNA methylation (5mC and 5hmC), and heterochromatin (H3K9me3, H4K20me3). The chromatin marks tended to be located at the 5' end of the LINEs (Supplementary Figure 1e). The chromatin at LINEs nonetheless remains difficult to analyze as LINEs occupy so much of the mouse genome. They are frequently close to other regulatory elements, and this can bias the chromatin patterns. Nonetheless, the pattern of chromatin marks on LINEs is complex, and highlights their roles in recruiting TFs40, in chromatin organization43, and cell type control44.

We noticed that many TE types had both an activatory and a repressive chromatin mark. For example, 16/679 TE types had both H3K9me3 and H3K27ac (Supplementary Data 2). One explanation for the overlapping patterns of activatory and repressive chromatin marks is that individual TE copies have discreet chromatin marks. However, genome views and pileup heatmaps suggested that the chromatin marks were simultaneously marking the TE copies (Supplementary Figure 1f, g). These observations come with two important caveats. The first is that the ChIP-seq data was generated from pooled cells, and the marks could separately label different cells. The second caveat is related to uncertainty in mapping reads to specific TE copies due to the repetitive nature to TEs. To address the latter issue, we performed native-ChIP-qPCR for a selection of TEs, using primer pairs that amplified unique regions of the genome close to the TEs (Fig. 1c). As controls, we used primer pairs for IAP-EZ-int, ETnERV-int, IAPLTR1a_Mm and RLTR10B2 (marked only by H3K9me3), and RLTR13E and RLTR13D6 (marked only by H3K27ac). In agreement with the genomic views, the TEs, L1Md_A/T, LTR9E, RLTR1B, RLTRETN_Mm, and LTRIS3, were indeed enriched for both H3K9me3 and H3K27ac, as measured by native-ChIP-qPCR (Fig. 1d). Overall, the native-ChIP-qPCR results closely matched the genome views (Fig. 1c), and indicated that, at least for this limited set of TEs, they are marked by both activatory and repressive chromatin marks.

Chromatin modifiers are recruited to TEs. TEs can function as regulatory elements, recruiting TFs and chromatin modifiers (CMs)39,45. To explore the role of CMs in depositing chromatin marks at TEs we collected 91 publically available ChIP-seq datasets for CMs in ESCs (Supplementary Data 1 and 3). In agreement with the patterns of chromatin marks at TEs, CMs were also bound to TEs in specific patterns (Fig. 2a). All 91 CMs were enriched >2 fold on at least 1 TE type, and 174/691 TE types were enriched with at least 1 CM, and of those, 134/174 were enriched with 2 or more CMs (Supplementary Data 3), for example RLTR46, which was bound by HCFC1, KANSL3, CHD8, and YY1 (Supplementary Figure 2a). TEs often contain DNA-binding motifs for TFs8, which can recruit CMs. To explore the TF binding motifs inside TEs, we performed de novo motif discovery. TEs were rich in DNA-binding motifs (Supplementary Figure 2b), and most types of DNA-binding motif are represented. We noticed that the YY1 and REST motifs were enriched in our analysis. The YY1 motif was enriched at LINE, IAP, and MERVL TEs, as previously reported46, and ChIP-seq data for YY1 showed that YY1 was bound to IAPLTR2b (Supplementary Figure 2c). The REST motif was identified as enriched in RMR21A TEs, and ChIP-seq data indicated REST was binding to RMR21A (Supplementary Figure 2c). One motif that interested us was the NRS2A motif, which was in almost all copies of RLTR13B2 (Supplementary Figure 2d). NRS2A is interesting as it can reprogram primed EpiSCs to naive ESCs47. ChIP-seq data indicated that NRS5A2 was bound to RLTR13B2 TEs (Supplementary Figure 2e), and RLTR13B2 was marked by a pattern of chromatin marks reminiscent of enhancers (H3K27ac, H3K4me1, and other activatory marks) (Supplementary Figure 2f).

In addition to NRS2A binding, there was a range of CMs bound, including mediator complex proteins, MED1, MED12, and NIPBL, and several co-activator/repressor complex proteins, including ASH2L, BRD4, HDAC1/2, KDM1A, P300, and SMARCA4 (Supplementary Figure 2g). Importantly, RNA-seq and ATAC-seq data indicated that NRS5A2 and RLTR13B2 were specifically expressed in naive ESCs, and that the RLTR13B2 TEs were open chromatin only in the naive ESCs (Supplementary Figure 2h, i). This analysis suggests that RLTR13B2 is under the epigenetic control of NRS5A2, and may be acting as a cis-regulatory element for NRS5A2. These results highlight the interactions between TEs, TFs, and CMs.

To interrogate the global arrangement of CMs, chromatin marks and TEs, we generated a relational network. Edges were drawn between the nodes (TEs, CMs, and CMs in each type of CM), if the chromatin mark or CM was enriched >2 fold on the TE type (Fig. 2b). One major group in the network was the IAP-family TEs which were marked by heterochromatic marks, such as H3K9me3, and H4K20me329,30 and the CMs SETDB1, SUV39H1/2, TRIM28, and ZFP809 (Fig. 2b–d). Many of the IAPs also had methylated DNA48,49, and were bound by MBD-family proteins50 (Fig. 2a). The chromatin structural protein CTCF was bound to RLTR46 TEs51, along with its known protein partners RAD21, STAG1, and STAG251 (Fig. 2c). We also observed other CMs enriched at RLTR46 TEs, including HCFC1, KANSL3, CHD8, and YY1 (Supplementary Figure 2a; Supplementary Data 3). The widespread binding of CMs to a TE is similar to the pattern observed for LINE L1s in human cancer cell lines, which are bound by a large number of TFs/CMs, including CTCF40. Interestingly, we did not observe CTCF binding to L1, in agreement with a previous report8, suggesting CTCF binding to L1 is species-specific.

Global co-correlation of the binding patterns of all CMs and chromatin marks on TEs revealed several clusters of co-occurring CMs/marks (Supplementary Figure 3). We annotated the clusters based on the function of known CM members. There were two supergroups which we designate the Heterochromatin group and the Active/Open group, supplemented by several smaller groups that were distinguished by known regulators/binders of TEs, such as DAXX52, REST7, or CTCF9,10. Amongst the Active/Open super-group were several subgroups which could be distinguished based upon their major members, such as the P300/mediator, or Pol II, along with groups that had patterns resembling enhancers (H3K4me1). One of the subgroups contained CHAF1A, a histone chaperone involved in depositing the variant histone H3.3, both of which have been implicated in TE control24,34. Finally, the Heterochromatin supergroup contained many factors specific for silencing, particularly SETDB1, MBD proteins and the chromatin marks H3K9me3 and H4K20me3. Overall, TEs were bound by a wide-range of CMs associated with both repressive and activatory chromatin (Fig. 2a, c, d and Supplementary Data 3), and these results demonstrate the interconnected binding patterns of the TEs by CMs through chromatin marks. However, although CMs bind promiscuously to TEs in a complex chromatin organization.
Selected subgroups are indicated on the network. Chromatin mark and a TE, or a CM and a TE, if the chromatin mark or CM was enriched at least 2-fold. Node sizes are proportional to the network degree.

Specific for the indicated TEs, chromatin marks and CMs are indicated. Care should be taken in interpreting genomic views due to uncertainty in mapping reads to specific TE copies, this applies to all genome views in the manuscript.

Fig. 2 TEs are marked by type-specific patterns of chromatin modifiers. a The fold-enrichment of CMs bound to TEs was measured, as in Fig. 1a. The data were hierarchically clustered and selected TE types are indicated. The full table is in Supplementary Data 3. See Supplementary Data 1 for details of the samples used in this study. b Relational network for the chromatin marks (purple), the selected CMs (green), and TEs (gray). Edges were drawn between a chromatin mark and a TE, or a CM and a TE, if the chromatin mark or CM was enriched at least 2-fold. Node sizes are proportional to the network degree. Selected subgroups are indicated on the network. c Read count tag density pileups (In RPKM) of the indicated chromatin marks and CMs for the selected TEs. The TEs were scaled to the same size, and the flanking 5 kb regions are shown. The gray line indicates the input/IgG control. d Example genomic views for the indicated TEs, chromatin marks and CMs are indicated. Care should be taken in interpreting genomic views due to uncertainty in mapping reads to specific TE copies, this applies to all genome views in the manuscript.
Chromatin modifiers regulate type-specific TE expression. To explore how CMs control TE expression, we selected a panel of reader, writer, and eraser CMs (Supplementary Figure 4a) to knockdown, based on the clusters we detected in Fig. 2a, b and Supplementary Figure 3. As cells can compensate for the derepression of TEs by activating alternative suppressive mechanisms, we chose to knockdown over a short time period of four days. This decision has two benefits. First, it allowed us to study those CMs that are lethal if knocked down for a prolonged time. Second, it should minimize changes in cell fate which would complicate the results, as the cells reconfigure their chromatin in response to differentiation. In total, we knocked down 41 CMs, and subjected them to RNA-seq. To rule out a change in cell fate we measured the expression of a panel of pluripotency and differentiation marker genes. Their expression indicated that these knockdowns were isogenic with the control samples (Supplementary Figure 4d). Principal component analysis gave similar results, as the knockdown samples did not diverge from the control samples (Supplementary Figure 4e). Analyzing the number of differentially expressed genes came to a similar conclusion, as minimal numbers of genes were differentially expressed in each knockdown (Supplementary Figure 4f). There was, however, an important exception. Knockdown of Mrc1 showed an upregulation of chromatin-related genes and a loss of germ cell character, and Chd4, which showed signs of differentiation towards the immune system (Supplementary Figure 4f). We include Chd4 and Mrc1 in the analysis below, but it should be noted that these two knockdowns are differentiating, and interpretation of their effect on TEs is complicated. For all other knockdowns, there was little evidence of differentiation, and no loss of pluripotency. This data set provides a system to look at the short-term deregulation of TEs when CMs are lost, in the absence of major cell type changes.

We next explored if these CM knockdowns can deregulate TE expression. We aligned the RNA-seq data to LINE, SINE (short interspersed nuclear element), LTR (ERV), and DNA class TEs (See Methods). On average 1–5% of reads would map to TEs (Supplementary Figure 5a), and we observed higher percentages of reads mapping to TEs in the knockdown samples (most >2% versus 1.6% in the shLuc control), suggesting a higher level of TE-derived RNA (Supplementary Figure 5a). As expected, knockdown of Setdb1 caused a robust upregulation of many types of TE (Fig. 3a, b). In addition, 29/41 other CM knockdowns significantly up or down-regulated at least 1 type of TE (Fig. 3a), and 8 knockdowns upregulated at least 1 type of TE >4 fold (Fig. 3b). We used RT-qPCR to validate a selection of TE expression changes in knockdowns with shRNAs against Rnf2, Setdb1, Ncor2, Prmt5, and Ash2l (Supplementary Figure 5b).

We selected eight CM knockdowns (and shNcor1, as a control) that deregulated 20 or more types of TE for further analysis. Western blotting for the pluripotency proteins OCT4 and NANOG confirmed that the eight selected knockdowns were not inducing differentiation (Supplementary Figure 5c). Knockdown of these eight CMs led to a variety of effects on the expression of TEs. Knockdown of Setdb1 deregulated only a few TE types, but it led to the largest change in the magnitude TE expression. Conversely, knockdown of Ncor2 or Rnf2 deregulated more types of TE, but the magnitude of upregulation was reduced (Figs. 3a, b). There was surprisingly little overlap in the TE types that were upregulated, and most TE types were specific to a single knockdown (Supplementary Figure 5d, e).

TEs can span a wide-range of evolutionary time, and some TEs are active, or recently active, whilst others are inactive and ancient. Importantly, young and old TEs are regulated differently. We measured the evolutionary age of TEs, and found that younger TEs tended to have higher enrichment of chromatin marks, although the correlation was modest (Supplementary Figure 1b). Younger TEs were more likely to be deregulated in a CM knockdown, although, again, the correlation was modest (Supplementary Figure 5f). These results suggest that younger TEs are more likely to be under active epigenetic regulation.

We next investigated the relationship between TE expression changes and chromatin marks. There was no global alteration in the levels of histone modifications, with the exception of the Rnf2 knockdown, which reduced H2AK119ub and H3K27ac (Fig. 3c). This indicated that the changes in histone modifications were context-specific, as we saw previously with the overexpression of HDAC3 in ESCs which did not alter global H3K27ac levels, but did alter context-specific H3K27ac sites. There was a direct relationship between CMs and chromatin mark. For example, knockdown of Setdb1 led to significant upregulation of TEs high in H3K9me3, whilst knockdown of the DNA methylation cofactor, Uhrf1 led to upregulation of TEs high in 5mC, particularly IAPs (Fig. 3d, e, and Supplementary Figure 5g). Knockdown of Prmt5 has previously been demonstrated to upregulate LINE L1 and IAP elements in primordial germ cells, and ground state 2i + LIF ESCs, and we observe the same in serum + LIF grown ESCs (Fig. 3f, g, h). The LINE and IAP TEs upregulated in the Prmt5 knockdown were enriched with H4R3me2, the catalytic target for PRMT5, indicating a link between PRMT5 and H4R3me2 at TEs (Figs. 1a, 3g, and Supplementary Figure 6e). These results show that changes in accessibility are a suitable proxy for chromatin marks.

**Chromatin modifiers alter chromatin accessibility at TEs.** The loss of CMs leads to context-specific deregulation of TEs, which in many cases was associated with specific chromatin marks. However, whilst many CMs are specific epigenetic writers/erasers (e.g., SETDB1 writes H3K9me3, RNF2 writes H2AK119ub), they are also members of large multi-protein complexes involved in multiple regulations (e.g., RNF2 is part of polycomb repressor complex 1), or there is substantial redundancy (e.g., SETDB1 and SUV39H1 both target H3K9me3). Consequently, the number of potential chromatin mark changes caused by the loss of any single CM can be large. To gain an overview of changes in chromatin in response to the knockdown of a CM, we used ATAC-seq to probe TE chromatin accessibility after knockdown. Of the 142,119 non-redundant peaks in the ATAC-seq data, we could detect changes in 15,076 loci (Supplementary Figure 6a, b). Chromatin that became accessible in each shRNA knockdown showed matching upregulation of genes within 10 kb of the ATAC-seq peak (Supplementary Figure 6c, d). The changes in chromatin accessibility were also reflecting chromatin marks. For example, ATAC peaks that became accessible when Prmt5 was knocked down were enriched for H4R3me2. A similar pattern was observed for Setdb1 knockdown and H3K9me3 (Supplementary Figure 6e). These results show that changes in accessibility are a suitable proxy for chromatin changes.
We measured the chromatin accessibility at TEs. 153 TE types showed an increase in chromatin accessibility (Fig. 4a, Supplementary Data 5). These changes were mainly specific to a single knockdown, as 108 TE types were uniquely upregulated, whilst 42 TE types were upregulated by more than one shRNA knockdown (Fig. 4a, b, and Supplementary Data 5). Importantly, opening of chromatin at TEs was matched by significant upregulation of the corresponding RNA expression (Fig. 4c). This pattern extended to individual types of TE, for example, the chromatin opened and regulation of expression by PRMT5 through modification of H3K9Me3 and expression was upregulated when Prmt5 was knocked down (Fig. 4d). Similarly, at LINE1 L1Md_F2 TEs, chromatin opened and expression was upregulated when Prmt5 was knocked down (Fig. 4d, e). This suggests a direct relationship between L1Md_F2, and regulation of expression by PRMT5 through modification of H3K9Me3.
**Fig. 4** Loss of CMs opens chromatin at specific TEs. 

a. Heatmap of the fold-changes in chromatin accessibility (measured by ATAC-seq) at TE types. The list of TE types are those that have 20% more or 20% less open or closed chromatin. Selected TE types are indicated, and the full table of all TE chromatin changes is in Supplementary Data 5. 

b. Bar chart of the number of TE types that acquire open chromatin in a unique shRNA knockdown (left), or in more than one shRNA knockdown (right). 

c. Boxplots showing the relationship between chromatin (ATAC-seq, top rows) and expression (RNA-seq, bottom rows) for selected TE families for the indicated CM knockdown, for the indicated type of TE. 

P values were from a Mann–Whitney U test. 

d. Boxplots of the levels of ATAC-seq (top rows, dark gray) and RNA-seq (bottom rows, light gray) for selected TE families for the indicated CM knockdown, for the indicated type of TE. 

P values were from a Mann–Whitney U test. 

**Supplementary Data 5** contains the list of TEs acquired open chromatin and expression in the indicated shRNA knockdown. Examples include the LINE1 elements (L1Md_T, L1Md_A), ERVL (MERVL-int), MT2_Mm, ERVKs (RLTR13B1, ETnERV2-int, and MMERVK10C-int). 

f. Enrichment of chromatin marks at TE types that become open in response to the indicated shRNA knockdown. Significance is derived from a Fisher exact test, P < 0.01 and is indicated with a asterisk. For the boxplots, the midline indicates the median, boxes indicate the upper and lower quartiles and the whiskers indicate 1.5x interquartile range.
H4R3me2 (Supplementary Figure 1d). A similar relationship was observed for the H3K9me3 enriched ETnERV2-int which acquired open chromatin and RNA expression when Setdb1 was knocked down (Fig. 4d, e). To explore these relationships, we measured the correlation between the chromatin marks and changes in accessibility (Fig. 4l). The clearest relationships were between shSetdb1 and heterochromatin marks H3K9me3 and H4K20me3, and other repressive marks, H3.3 and 5mc (Fig. 4l). Knockdown of Rnf2 also correlated with repressive epigenetic marks (Fig. 4f). Overall, these results demonstrate that the loss of CMs leads to the alteration of chromatin at TEs, and a matching upregulation of the expression of TEs. There is a relationship between the CMs target, and the chromatin modifications marking the TEs, although due to the nature of CMs as members of multi-protein complexes, a direct relationship was not always observable.

Knockdown of CMs activates early embryonic gene expression. TEs are expressed and their chromatin is regulated in a phase-specific pattern in the early embryo, TE expression has become a useful marker for early embryonic gene expression programs in ESCs, as MERVL expression marks a small sub-population of 2C-like cells, that have totipotent properties, a capability ESCs lack. Knockdown of CMs can increase the fraction of these cells but the relationship between CM, chromatin, MERVLs, and the 2C gene expression program remains unclear.

To shed light on the regulatory network controlling 2C-like cells, we interrogated our shRNA knockdowns for activation of a 2C-like gene expression program. Downregulation of Rnf2, Brd7, and Hdac5 led to the upregulation of MERVLs, and a corresponding increase in 2C-stage embryonic genes, such as Zscan4c (Fig. 5a, b). To assess the type of embryonic gene program being activated we measured the expression of the significantly upregulated genes in each of the knockdowns, against RNA-seq data from a range of embryonic stages. Genes upregulated in the Rnf2, Hdac5, and Brd7 knockdowns were significantly highly expressed in 2C embryos, but not in other embryonic stages (Fig. 5c). Knockdown of Setdb1 did not have this effect (Fig. 5c). This pattern could be confirmed by measuring the expression level of the top 200 upregulated genes (versus ESCs) in 2C embryos (Fig. 5d). A characteristic of the 2C stage is

**Fig. 5** Knockdown of chromatin modifiers activates a 2C-like gene expression program. a RNA-seq expression level of the 2C-related gene Zscan4c in the indicated knockdowns (left), and a boxplot showing the level of MERVL TEs in the same knockdown RNA-seq data. Bar chart values indicate the mean of the two biological replicates (shHdac5 or shRnf2), or a biological singlicate (shBrd7), relative to shLuc control knockdown. The boxplot shows the Log2 normalized read count mapping to the total set of MERVL TEs, from the same RNA-seq data. b Example genome view showing the RNA-seq pileup data and activation of MERVL-int and MT2_Mm TE expression in Brd7, Hdac5, and Rnf2 shRNA knockdowns. c Boxplot of RNA-seq expression for genes upregulated in the indicated knockdowns. For the sets of genes significantly upregulated in the indicated CM knockdowns, their expression was measured in the indicated embryonic stages and in ESCs (RNA-seq data was taken from). P values are from a Mann–Whitney U test. The x-axis indicates the overall level of expression of the knockdown-specific genes, in those cell types. d Upregulation of a 2C-related gene signature by the indicated CM knockdowns. A 2C-related gene signature was generated by taking the fold-change of all 2C genes versus ESCs and taking the top 200 fold-change upregulated genes. The expression of those top 200 2C genes was then measured in the indicated knockdowns, relative to shLuc control. e MA-scatter plots showing the expression of all genes and TEs in the shRnf2, shHdac5, and shBrd7 knockdowns. Labels and red spots indicate the expression of the shRNA target gene, the MERVL-int TE, and selected 2C-related genes (Zscan4a/b/c/d/e/cluster, Dux3, Hsp90aa1, Hsp90ab1, and Gata2). Note that the effect is not related to deregulation of Trim28 or Kdm1a, two known MERVL regulators. For the boxplots in this figure, the midline indicates the median, boxes indicate the upper and lower quartiles and the whiskers indicate 1.5*interquartile range
the splicing of MERVL TEs into genes on the same strand orientation, pattern we observed only in the Brd7, Hdac5, and Rnf2 knockdowns (Supplementary Figure 7a, b). These results showed that the 2C gene expression program was only seen in the shBrd7, shHdac5 and shRnf2 knockdowns (Fig. 5d).

We next sought to understand how these knockdowns were activating a 2C gene expression program. Our analysis indicates that MERVLs are enriched for the activatory H3K36ac chromatin mark, but are not enriched for H2AK119ub (Fig. 1a and Supplementary Data 2). This is interesting as RNF2 catalyses the deposition of H2AK119ub, but downregulation of H2AK119ub is a requirement for 2C-like cells52. The TF GATA2 has been implicated in driving cells to a 2C-fate61, and in our data only knockdown of Rnf2 could activate Gata2 (Fig. 5e). Interrogating ChIP-seq data for Rnf2 and H2AK119ub53, indicated that they are both enriched at the Gata2 locus (Supplementary Figure 7c). These results suggest that 2C-like cells can emerge by the manipulation of two pathways, one pathway involving H3K56ac, HDAC5, and BRD7, and an indirect pathway involving H2AK119ub, RNF2, and Gata2. These pathways complement other factors known to activate the 2C-like state, including Chaf1a/b, Chd5, Rif1, Kdm1a, Ehmt2, Yy1, Rypb, and miR-34a15,26,28,51,64. These results suggest that there are multiple convergent pathways that can drive cells to a 2C state, although whether all of these factors are converging on the same pathway or act independently remains unclear.

Discussion

TEs are the single most prevalent genomic element, more numerous than coding genes or other regulatory units. Although potentially dangerous for genome integrity they are nonetheless a pool of potential regulatory sequences, chromatin control elements, and protein coding genes, and they can function as a substrate for evolutionary processes3,9,21. During early embryonic development there is extensive chromatin remodeling58, and the activation of TE expression18. Consequently, a careful balance is maintained between the potential beneﬁts of TEs, versus their deleterious aspects, and this control is carried out by the epigenetic regulatory system.

Heterochromatin formation is a major mechanism for the suppression of TEs, and we ﬁnd that many TEs are marked by heterochromatic histone modiﬁcations. However, data suggests that, at least for LINEs and IAP-family TEs, TEs are repressed by the loss of activatory marks, rather than through the gain of repressive heterochromatin19. Our data indicate that TEs are marked by many activatory marks, like histone acetylations. We also observed bivalent marking of TE chromatin by H3K9me3 and H3K27ac, which may help to explain why we only observed a limited number of downregulated TEs in response to CM knockdown. This was surprising, as ESCs express over 200 TE types at levels comparable to many genes. Bivalent marking may mean that these TEs are kept in partially activated and repressed states.

TEs are increasingly implicated in a wide-range of biological functions in embryonic cell types. It has long been known that there is reverse transcriptase activity (indicative of TE activity), in developing mouse embryos65, and HERV-derived viral-like particles have been observed in normal human embryos66. TEs are expressed in a stage-speciﬁc pattern in both human and mouse embryos18,19, and their expression in ESCs is associated with alternative processes. For example, MERVL expression marks the 2C stage of embryonic development as well as 2C-like cells in ESCs26, and the inhibition of MERVLs, or MERVL-containing lincRNAs, impairs the transition from the zygote to a 4C embryo stage57,68. LINE1 RNA is transiently required for 2C stage embryos to exit to the 4C stage, and in ESCs LINE1 is required for self-renewal, and chromatin remodeling53,64. It is increasingly clear that TEs are involved in normal biological processes.

In conclusion, our study reveals the complexity of epigenetic regulation of TEs in ESCs. TEs are marked not only by repressive chromatins marks, but are also widely marked by activatory chromatins marks, including patterns of simultaneous marking by activatory and repressive marks. We show that the loss of CMs causes speciﬁc changes in TE expression and chromatin accessibility, and these effects are not limited to the formation of heterochromatin or DNA methylation. Rather, TEs can be activated by a large fraction of known chromatins modiﬁcations. This work suggests that the regulation of TEs is a function of overlapping epigenetic pathways, and that a major task of the epigenetic system is to manage the expression of TEs.

Methods

Cell lines and cell culture medium. All experiments were performed in mouse E14 ESCs (RRID:CVCL_C320), to avoid any differences in genetic background that may influence the number of CMs. Mouse ESCs were grown in serum + LIF medium: DMEM-HIGH GLUCOSE (SH30022.01, HyClone), 15% (v/v) fetal bovine serum (S1580–500, Bioreuss), 1% (v/v) L-Glutamine (35500-061, Gibcco), 1% (v/v) sodium pyruvate (25-000-CIR, Corning), 1% (v/v) non-essential amino acids (25-025-CL, Corning), 0.5% (v/v) penicillin-streptomycin solution (SV30010, HyClone), 0.1 mM 2-mercaptoethanol (21985-023, Gibcco), and 1000 U/ml mLIF (E651107, Millipore).

Chromatin modifier knockdowns. CMs were knocked down in serum + LIF medium using pLKO.puro lentiviral system as previously described27,41. Briefly, pLKO.puro lentiviruses were added to ESCs and the medium was changed after 8 h. Cells were selected with 2 μg/ml puromycin (Gibco) for 2 days, and harvested after a further 2 days. Sequences for the shRNAs are in Supplementary Data 6.

Western blot. Histone proteins were extracted by acid precipitation. Briefly, cells from a 6-well dish were scraped, washed with PBS and resuspended in extraction buffer (PBS, 0.5% Triton-X-100 (9002-93-1, Sigma-Aldrich), 5 mM sodium butyrate (B5887, Sigma-Aldrich), and allowed to extract for 20 mins. The extract was centrifuged at 14,000g for 20 mins and the supernatant was discarded. The pellet was resuspended in 0.2 N HCl and left at 4°C overnight. The mixture was vortexed, and the pH was neutralized with 1 M Tris-HCl pH 8.0. Western blots were performed using typical laboratory procedures with the antibodies: anti-H3K36me3 (1:1000; ab8988, abcam), anti-H3K4me3 (1:1000; ab8580, abcam), anti-H3K36acme3 (1:1000; 07-449, Millipore), anti-H3ac (1:1000; 06-599, Millipore), anti-H4K12ac (1:1000; ab64983, abcam), anti-H4ac (1:1000; 06-866, Millipore), anti-H3K27ac (1:1000; ab4729, abcam), anti-H3K27me3 (1:1000; 07-027, Millipore), and anti-H4K20me1 (1:1000; ab1791, abcam), anti-STAT3 (1:1000; Santa Cruz), anti-NANOG (1:1000; A300-3979A, Bethyl), and anti-GAPDH (1:1000; MAB374, Millipore). Uncropped raw images of the Western blot membranes are in Supplementary Figure 8.

RNA-seq and RT-qPCR experimental protocols. RNA-seq was performed as described49. Briefly, a 6-well dish of ESCs was scraped and lysed in TRIzol (RN190- 200, Invitrogen). RNA was extracted according to the manufacturer’s instructions, and 1 μg was submitted for sequencing. All RNA-seq was sequenced to a depth of 200 million reads per sample, and 30 million paired-end.reads. For RT-qPCR RNA was extracted using TRIzol (RN190-200, Invitrogen), cDNA was prepared with reverse transcriptase kit (RNR6B, Takara), and RT-qPCR was performed with SYBR Green Master kit (RR820R, Takara). RT-qPCR primers are described in Supplementary Data 7.

ATAC-seq and native-ChIP-qPCR experimental protocols. ATAC-seq was performed as described27. A total of ~50,000 cells were washed once with 50 μl cold PBS and resuspended in 50 μl lysis buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.2% (v/v) IGEPAI CA-630). The suspension of nuclei was then centrifuged for 10 min at 500 × g at 4°C, followed by the addition of 50 μl
transposition reaction mix (25 μl TD Buffer, 2.5 μl Entranspose and 22.5 μl nuclease-free H2O) of Nextera DNA library Preparation Kit (96 samples) (FC-121-1031, Eppendorf). Samples were then PCR amplified and incubated at 72 °C for 30 min. DNA was isolated using a MinElute Kit (Qiagen). ATAC-seq libraries were first subjected to five cycles of pre-amplification. To determine the suitable number of cycles required for the second round of PCR, the library was assessed by quantitative PCR[25], and the library was then PCR amplified for the appropriate number of PCR cycles. Libraries were purified with a Qiaquick PCR (Qiagen) column and were quantified using a KAPA Library Quantification kit (KK4824, KAPA Biosystems) according to the manufacturer’s instructions. Library integrity was checked by gel electrophoresis. Finally, the ATAC library was sequenced on a NextSeq 500 High Output Kit v2 (150 cycles) (FC-400-2001, Illumina) according to the manufacturer’s instructions.

Native-ChIP-qPCR was performed as described[27]. Ten million cells were centrifuged and resuspended in Buffer 1 (0.32 M Sucrose, 60 mM KCl, 15 mM NaCl, 5 mM MgCl2, 0.1 mM EDTA, 15 mM Tris-HCl pH 7.5, 0.5 mM DTT, 0.1 mM PMSF, 1:1000 protease inhibitor cocktail (Sigma-Aldrich)), along with 2 ml of Buffer 1 + 6% IGEPAL. The resulting 4 ml of nuclei were layered on top of 8 ml of Buffer 3 (Same as Buffer 1, but with 1.2 M Sucrose) and centrifuged at 10,000 × g for 20 min at 4 °C unbraked. Nuclei were resuspended in Buffer A (0.34 M sucrose, 15 mM Hepes [pH 7.4], 15 mM NaCl, 60 mM KCl, 4 mM MgCl2, 1 mM DTT, 0.1 mM PMSF, 1:1000 protease inhibitor cocktail (Sigma-Aldrich)) and digested for 10 mins at 37 °C with MNoase (Affymetrix) in buffer A supplemented with 3 mM CaCl2. The reaction was stopped with 5 mM MgATP, centrifuged at 13,500 × g for 10 min, and chromatin resuspended in (10 mM EDTA [pH 8.0], 1 mM PMSF, 1:1000 protease inhibitor cocktail) and rotated at 4 °C for 2-4 h. The mixture was adjusted to 500 μM NaCl, allowed to rotate for another 45 min and then centrifuged at 13,500 × g for 10 min. Chromatin supernatant was diluted to 100 ng/μl with buffer B (20 mM Tris [pH 8.0], 5 mM EDTA, 500 mM NaCl, 0.2% Tween 20) and incubated for 20 mins at 4 °C with 60 μl protein G beads (GE Healthcare). Antibody was added and rotated overnight at 4 °C. A volume of 100 μl of Protein G beads were added and the sample was rotated for a further 4 h at 4 °C. The beads were washed three times with buffer B, and once with buffer B without Tween. The DNA was eluted with 300 μl of elution buffer (20 mM Tris [pH 7.5], 20 mM EDTA, 0.5% SDS, 500 μg/ml Protease K) and incubated for 4 h at 56 °C. The resulting samples were extracted with phenol-chloroform followed by purification using a QiaGen MinElute columns, according to the manufacturer’s instructions. Antibodies used: anti-H3K9me3 (9 μg ab8989, abcam), anti-H3K27ac (5 μg ab4729, abcam). qPCR was performed with SYBR Green Master kit (RR820B, Takara). Primer pairs are described in Supplementary Data 7.
12. Simonti, C. N., Pavlicev, M. & Capra, J. A. Transposable element exaptation
13. Leung, D. C. et al. Lysine methyltransferase G9a is required for de novo DNA methylation and the establishment, but not the maintenance, of proviral silencing. *Nat. Rev. Genet.* **14**, 631–634 (2010).
10. Sun, X. et al. Transcription factor pro
12. Johnson, R. et al. Identification of the REST region reveals extensive transposable element-mediated binding site duplication. *Nucleic Acids Res.* **34**, 3862–3877 (2006).
8. Bourque, G. et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.* **18**, 1752–1762 (2008).
1. Kouzarides, T. G. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* **42**, 631–634 (2010).
11. Sun, X. et al. Transcription factor profiling reveals molecular choreography and key regulators of human retrotransposon expression. *Proc. Natl Acad. Sci. USA* **115**, E5526 (2018).
12. Jacques, P. E., Jeyakani, J. & Bourque, G. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genet.* **9**, e1003504 (2013).
38. Walter, M., Teissandier, A., Perez-Palacios, R. & Bourc...
66. Grow, E. J. et al. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522**, 221–225 (2015).

67. Kigami, D., Minami, N., Takahashi, Y. & Imai, H. MuERV-L is one of the earliest transcribed genes in mouse one-cell embryos. *Biol. Reprod.* **68**, 651–654 (2003).

68. Wang, J. et al. A novel long intergenic noncoding RNA indispensable for the cleavage of mouse two-cell embryos. *EMBO Rep.* **17**, 1452 (2016).

69. Hutchins, A. P., Takahashi, Y. & Miranda-Saavedra, D. Genomic analysis of LPS-stimulated myeloid cells identifies a common pro-inflammatory response but divergent IL-10 anti-inflammatory responses. *Sci. Rep.* **5**, 9100 (2015).

70. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **109**, 21–29 (2015).

71. Hasson, D. et al. The octamer is the major form of CENP-A nucleosomes at human centromeres. *Nat. Struct. Mol. Biol.* **20**, 687–695 (2013).

72. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nucleic Acids Res.* **39**, 2–22 (2011).

73. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

74. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-content normalization for RNA-Seq data. *BMC Bioinformatics* **12**, 480 (2011).

75. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

76. Jin, Y., Tam, O. H., Paniagua, E. & Hammell, M. TETranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31**, 3593–3599 (2015).

77. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* **11**, R14 (2010).

78. Hutchins, A. P., Jauch, R., Dyla, M. & Miranda-Saavedra, D. gbase: a framework for combining, analyzing and displaying heterogeneous genomic and high-throughput sequencing data. *Cell Regen.* **3**, 1 (2014).

79. Ramirez, F., Dundar, F., Diehl, S., Gruning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).

80. Bailey, T. L. et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).

Acknowledgements

We thank Ralf Jauch for discussions. This work was supported by the National Natural Science Foundation of China (31801217, 31850410463, 31850410486, 31671537, 31501185), Shenzhen Peacock plan, the Shenzhen Science and Technology Innovation Committee general program (JCYJ20170307110638600), and the Science and Technology Planning Project of Guangdong Province (2017B030314056).

Author contributions

J.H. performed the bioinformatic analysis, and prepared the manuscript. X.F. did the ATAC-seq, RNA-seq, M.Z., W.L., M.M.A., I.Z. performed the knockdown panel RNA-seq. H.L., W.K., F.H., C.C. performed experiments. L.S., Y.L., I.A.B., helped with the bioinformatics. A.P.H. conceived the project, supervised and funded the study, and prepared the manuscript with assistance from M.A.E, C.J, Q.Z., I.A.B, and Y.-H.L.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-08006-y.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.