6-7-2023

Co-option of endogenous retroviruses through genetic escape from TRIM28 repression.

Rocio Enriquez-Gasca

Poppy A Gould

Hale Tunbak

Lucia Conde

Javier Herrero

See next page for additional authors

Authors

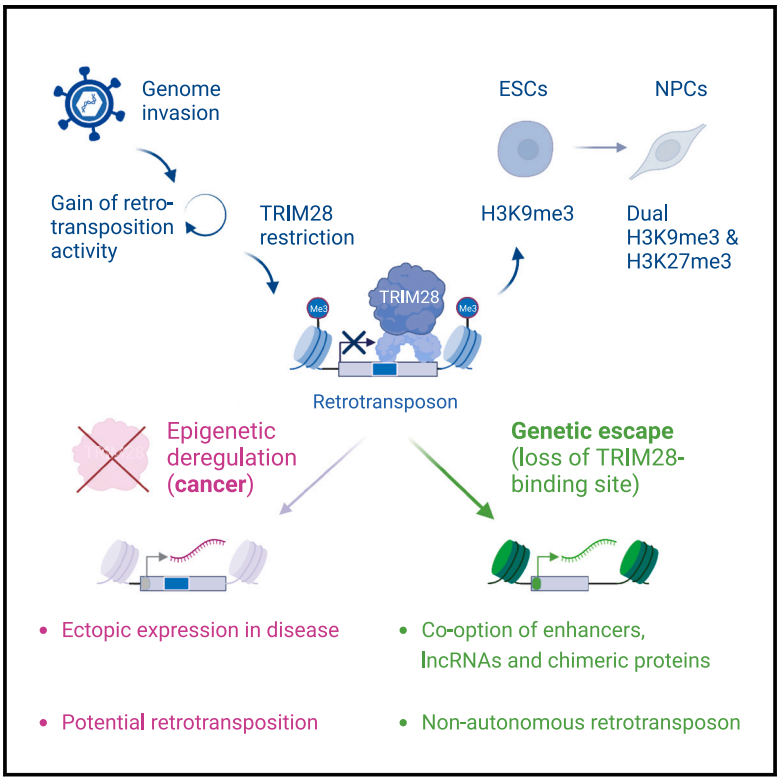
Rocio Enriquez-Gasca, Poppy A Gould, Hale Tunbak, Lucia Conde, Javier Herrero, Alexandra Chittka, Christine R Beck, Robert Gifford, and Helen M Rowe

Cell Reports

Article

Co-option of endogenous retroviruses through genetic escape from TRIM28 repression

Graphical abstract



Highlights

- Tracking the transcriptional fate of ERVs through neural differentiation sheds light on co-option
- Most ERVs succumb to H3K9me3 deposition in ESCs and additionally bear H3K27me3 in NPCs
- ERVs that activate host genes have undergone genetic escape from TRIM28 repression and are expressed
- Evasion of silencing is an evolutionary trade-off that comes with the loss of retrotransposition

Authors

Rocio Enriquez-Gasca, Poppy A. Gould, Hale Tunbak, ..., Christine R. Beck, Robert Gifford, Helen M. Rowe

Correspondence

r.enriquez-gasca@qmul.ac.uk (R.E.-G.), h.rowe@qmul.ac.uk (H.M.R.)

In brief

Potentially harmful endogenous retroviruses can gain beneficial host functions in development and immunity in a process termed co-option. Enriquez-Gasca et al. demonstrate the early steps of this process by showing that endogenous retroviruses undergo genetic changes that allow them to evade transcriptional silencing and activate neural genes.

Enriquez-Gasca et al., 2023, Cell Reports *42*, 112625
June 27, 2023
https://doi.org/10.1016/j.celrep.2023.112625

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

Co-option of endogenous retroviruses through genetic escape from TRIM28 repression

Rocio Enriquez-Gasca,[1,5,*] Poppy A. Gould,[1,5] Hale Tunbak,[1] Lucia Conde,[2] Javier Herrero,[2] Alexandra Chittka,[1] Christine R. Beck,[3] Robert Gifford,[4] and Helen M. Rowe[1,6,*]

[1]Centre for Immunobiology, Blizard Institute, Queen Mary University of London, London E1 2AT, UK
[2]Bill Lyons Informatics Centre, UCL Cancer Institute, London WC1E 6DD, UK
[3]Department of Genetics and Genome Sciences, University of Connecticut Health Center, The Jackson Laboratory for Genomic Medicine, Connecticut, JAX CT, Farmington, CT 06032, USA
[4]MRC-University of Glasgow Centre for Virus Research, Glasgow G611QH, UK
[5]These authors contributed equally
[6]Lead contact
*Correspondence: r.enriquez-gasca@qmul.ac.uk (R.E.-G.), h.rowe@qmul.ac.uk (H.M.R.)
https://doi.org/10.1016/j.celrep.2023.112625

SUMMARY

Endogenous retroviruses (ERVs) have rewired host gene networks. To explore the origins of co-option, we employed an active murine ERV, IAPEz, and an embryonic stem cell (ESC) to neural progenitor cell (NPC) differentiation model. Transcriptional silencing via TRIM28 maps to a 190 bp sequence encoding the intracisternal A-type particle (IAP) signal peptide, which confers retrotransposition activity. A subset of "escapee" IAPs (~15%) exhibits significant genetic divergence from this sequence. Canonical repressed IAPs succumb to a previously undocumented demarcation by H3K9me3 and H3K27me3 in NPCs. Escapee IAPs, in contrast, evade repression in both cell types, resulting in their transcriptional derepression, particularly in NPCs. We validate the enhancer function of a 47 bp sequence within the U3 region of the long terminal repeat (LTR) and show that escapee IAPs convey an activating effect on nearby neural genes. In sum, co-opted ERVs stem from genetic escapees that have lost vital sequences required for both TRIM28 restriction and autonomous retrotransposition.

INTRODUCTION

Mammalian genomes are constantly co-evolving with the abundant transposable element (TE)-derived DNA burden of which they are comprised.[1] TEs possess their own functional sequences, which can be repurposed by the host in a process known as co-option, for example to rewire gene regulatory networks[2] or generate chimeric proteins.[3,4] They also contain sequences that are targeted by TRIM28-mediated transcriptional silencing to protect genome integrity.[5–7] This duality represents an evolutionary dilemma in terms of how and when a TE becomes silenced vs. co-opted to affect a cellular function.

Several prominent examples of co-option of TE-derived sequences in different molecular roles have been described.[4,8,9] Neural lineages represent a fertile ground for innovation with neocortex-specific enhancers conserved in present-day mice that are derived from ancient TEs dating back to amniote genomes.[10] Intriguingly, somatic mosaicism resulting from LINE-1 activity has been documented in the human brain,[11,12] potentially contributing to phenotypic variability. A neuronal protein, *Arc*, is derived from a retroviral *Gag* gene and functions as a viral-like capsid, transferring mRNAs from neuron to neuron.[13]

Despite well-documented cases of TE co-option, however, understanding the early events promoting this process remains a challenge. Here, we set out to pinpoint how co-option events may emerge by interrogating which endogenous retroviruses (ERVs) gain transcriptional activity and why.

Enhancer activity serves as a proxy for subsequent co-option events including rewiring of host genes through *cis*-regulatory elements,[2] as well as the generation of chimeric transcripts[14] and proteins derived from ERVs.[15,16] We employ mouse embryonic stem cells (ESCs) as a developmental model and focus on an actively transposing murine ERV family, the intracisternal A-type particles (IAPEz with long terminal repeats [LTRs] of the LTR1/1a type)[17,18] to map evolutionarily recent gain-of-enhancer events.

We map TRIM28 repression to overlap the signal peptide that targets IAPEz particles to the endoplasmic reticulum and has been shown to have conferred intracisternal A-type particles (IAPs) with the ability to retrotranspose.[19–21] We then measure the histone modifications enriched at IAPEz elements through neural differentiation, which we refer to in this article as their epigenetic fate. Tracking the epigenetic fate of endogenous IAPEz elements reveals that those with

the TRIM28-binding site are laden with both H3K9me3 and H3K27me3 in neural progenitor cells (NPCs). A minority of IAPEz copies exhibit sequence divergence at the TRIM28-binding site, which mirrors their transcriptional derepression in NPCs, and we define these integrants as "escapees." Importantly, we validate an enhancer sequence within the LTR of IAPEzs, and we observe significantly higher expression of escapee-proximal genes compared with genes nearby their repressed counterparts. From this pool of putative enhancers, only those with beneficial effects on adjacent genes would be selected, whereas the rest may reside as neutral events subject to further decay.

Taken together, this work shows that epigenetic silencing depends on TRIM28 targeting of vital sequences needed for retrotransposition. Genetic escape from epigenetic silencing paves the way toward ERV domestication. This in turn explains why the noted downregulation of epigenetic complexes in some cancers[22] can potentially unveil rare retroelement copies intact for retrotransposition.[23]

RESULTS

TRIM28 targets and represses IAPEz elements independent of the YY1 and PBS sites

We explored the early steps of co-option by focusing on IAPEz elements, which adapted to retrotranspose through gain of an endoplasmic reticulum (ER)-targeting signal at the N-terminal of GAG and loss of their retroviral envelope gene (Figure 1A, top panel).[21] We defined full-length IAPEz LTR1/1a integrants in the reference C57BL/6J genome (referred to hereafter as IAPEz) as those that (1) contained two LTRs of the relevant subtype (LTR1/1a), (2) possessed two LTRs in the same orientation, (3) were separated by less than 20 kb, and (4) were annotated to have IAPEz internal sequences. We then cross-referenced these positions with the reported list of deletions in the 129/Ola genome to define a list of 838 full-length IAPEz copies known to be present in both C57BL/6J and 129/Ola mouse strains (Figure 1A, bottom panel; see Table S1).

We employed public chromatin immunoprecipitation sequencing (ChIP-seq) data from De Iaco et al.[24] to visualize the binding intensity of TRIM28 on these 838 IAPEz LTRs and 10 kb flanking regions in ESCs grown in 2i + LIF media. One binding site for TRIM28 corresponded to the LTR and was present in full-length (Figure 1) and solo LTRs (Figure S1A), while a second prominent TRIM28 peak mapped to the IAPEz UTR of full-length elements (Figure 1B). Cloning sequences of an IAPEz integrant (termed IAP575; Figure 1C) located downstream of the gene *Zfp575*, at which we previously detected TRIM28, SETDB1, and repressive H3K9me3 to be enriched,[25] showed that an LTR-UTR/GAG construct, and not the LTR alone, was sufficient to confer repression in a reporter assay in ESCs (Figure 1D), which we verified to express pluripotency markers (Figure S1B). Silencing was detected by day 3 post-transduction (Figure S1C) and was complete by day 8. We verified that reporter repression was not due to lack of vector integration (Figure 1D), and 3T3 cells, which cannot establish heterochromatin, served as an additional control. As expected, both constructs were equally expressed in the latter cells (Figure S1D). This suggested the TRIM28-binding site within the IAPEz UTR/GAG to be a key determinant of epigenetic repression.

The IAPEz LTR harbors a YY1-binding site like LINE-1 elements, which have recently been shown to depend on this site for their epigenetic repression,[12] whereas the retrovirus murine leukemia virus (MLV) is silenced through its PBS[26] and YY1 site.[27] We therefore asked if the YY1 site or the PBS were necessary for repression individually or in combination and found both to be dispensable (Figures 1E and S1E). Subsequent reporter assays mapped repression to the proximal part of the IAPEz UTR (Figures 1F, S1F, and S1G), which was relieved upon short hairpin RNA (*shRNA*)-mediated depletion of TRIM28 (Figures 1G and S1H), as expected.

Transcriptional repression maps to a 190 bp sequence overlapping the IAPEz signal peptide

In order to more precisely map the sequence required for TRIM28 repression, we drew on previous sequence annotation of the IAPEz UTR/GAG. This region contains two direct repeats (DRs),[28] between which resides an ER-targeting peptide that is upstream of and in frame with GAG and which has been shown to have conferred an infectious IAPEz progenitor with the *de novo* ability to retrotranspose.[21] The remnants of DRs suggests that this signal peptide, which derives from host sequences, could have been captured itself through a retrotransposition event. Deletion of a region containing DR1 (96 bp, of which 86 bp is DR1) relieved most repression, and deletion of a region including DR1 plus a fraction of DR2 (127 bp) was sufficient to completely relieve reporter repression (Figures 2A and S2A). To establish whether DR1 is sufficient to confer reporter repression or DR1 plus DR2 is required, we tested the repressive effect of these sequences upstream of the LTR reporter construct, in either sense (S) or antisense (αS) orientation. This revealed that a 174 bp sequence encompassing DR1 and DR2 in combination was necessary and sufficient to establish reporter repression to the same degree as the whole UTR sequence (Figures 2B and S2B). We then included DR1 plus DR2 and the additional flanking sequence included in the 127 bp mutant to define a 190 bp sequence as a tool to probe the conservation within this region of endogenous IAPEz retrotransposons (Figures 2A–2C).

Having defined a 190 bp repressor region, we determined its conservation across the 838 full-length copies of IAPEz present in the genome of 129/Ola mice. We classified IAPEz copies based on their percent identity to this repressor sequence (190 bp), with those exhibiting 80% or greater identity classified as an IAPEz "with" a complete repressor and those with <80% sequence identity as an IAPEz "without" a repressor, reasoning that these two groups may exhibit differences in their epigenetic fate (Figure 2C). 82% of IAPEz elements contained a full repressor, which we term canonical IAPEzs, while only 18% fell into the without group, in which there was a range of percent divergence from the repressor between individual integrants (Figure 2C). To address whether genetic sequence divergence from the repressor has functional consequences, we compared the TRIM28 binding between the with- and without-repressor IAPEz groups. This analysis unveiled a much stronger TRIM28 signal for the canonical IAPEz copies (Figure 2D), indicating that sequence divergence may alter the epigenetic regulation of these elements.

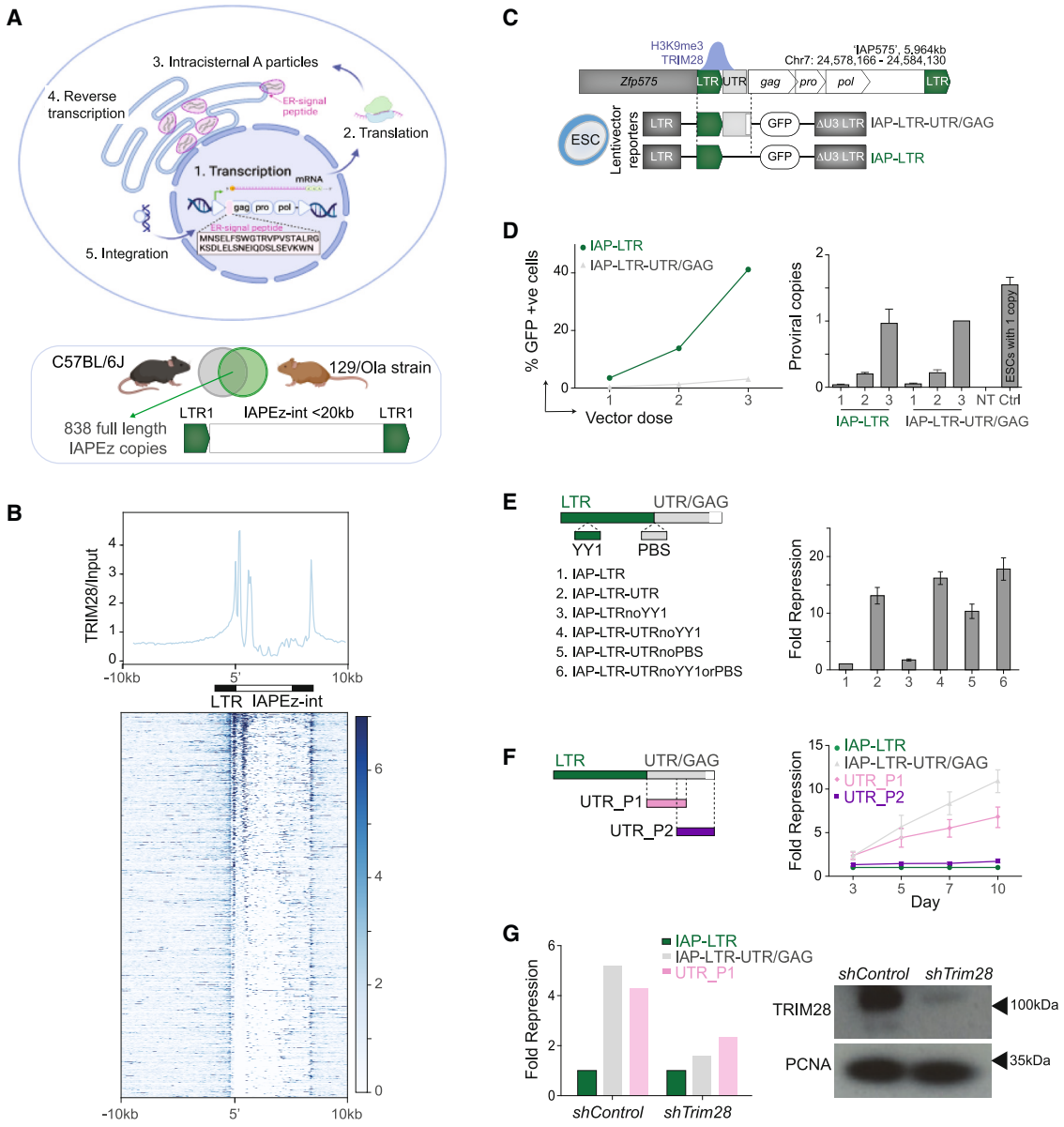


Figure 1. TRIM28 targets and represses IAPEz elements independent of the YY1 and PBS sites

(A) Retrotransposition cycle of IAPEz elements. IAP elements gained an endoplasmic reticulum (ER) signal peptide upstream and in frame with GAG, retargeting particles to the ER (top). Schematic representation of the sequence structure of IAPEz that was used to annotate full elements in this study (bottom).

(B) TRIM28 fold enrichment normalized to total input over full-length IAPLTR1/1a elements, where the 3′ end coordinate of their 5′ LTR was used as the reference point. The fold enrichment is represented as a profile plot (top) or a heatmap sorted by TRIM28-binding intensity (bottom). The scale bar on the heatmap shows the fold enrichment of the read coverage.

(C) Schematic representation of a TRIM28-repressed IAP in chromosome 7 (IAP575) (top) and of the lentivector reporter constructs with the LTR ± its endogenous 5′ UTR and GAG junction (bottom).

(D) Percentage of GFP positive (+ve) cells by vector dose in ESCs transduced with the reporter construct in (C). Results are shown for day 8 post-transduction (left), with one representative experiment shown of five independent biological replicates. Right: relative number of proviral integrants for both constructs. Error bars show standard error of duplicates.

(E) Fold repression of reporters normalized to expression of IAP-LTR in ESCs where constructs harbor a deletion of the YY1-binding site, the PBS, or both (depicted left). Data are shown for day 8 post-transduction (right). Error bars show mean and standard deviation of four independent biological replicates.

(F) Fold repression of the reporter normalized to expression of IAP-LTR in ESCs for constructs comprising part 1 (P1) or part 2 (P2) of the 5′ UTR, as shown left, in addition to constructs shown in (C), over a time course of 3–10 days (right). Error bars show mean and standard deviation for one representative experiment of four independent biological replicates.

(G) Fold repression of indicated constructs in control or *Trim28*-depleted (shRNA) ESCs. Data are from day 3 post-transduction before TRIM28 depletion is lethal (right). One representative experiment is shown of two independent biological replicates. Western blots showing successful knockdown of TRIM28 with PCNA as a loading control. See Tables 1 and 2 for antibodies and shRNA sequences, respectively.

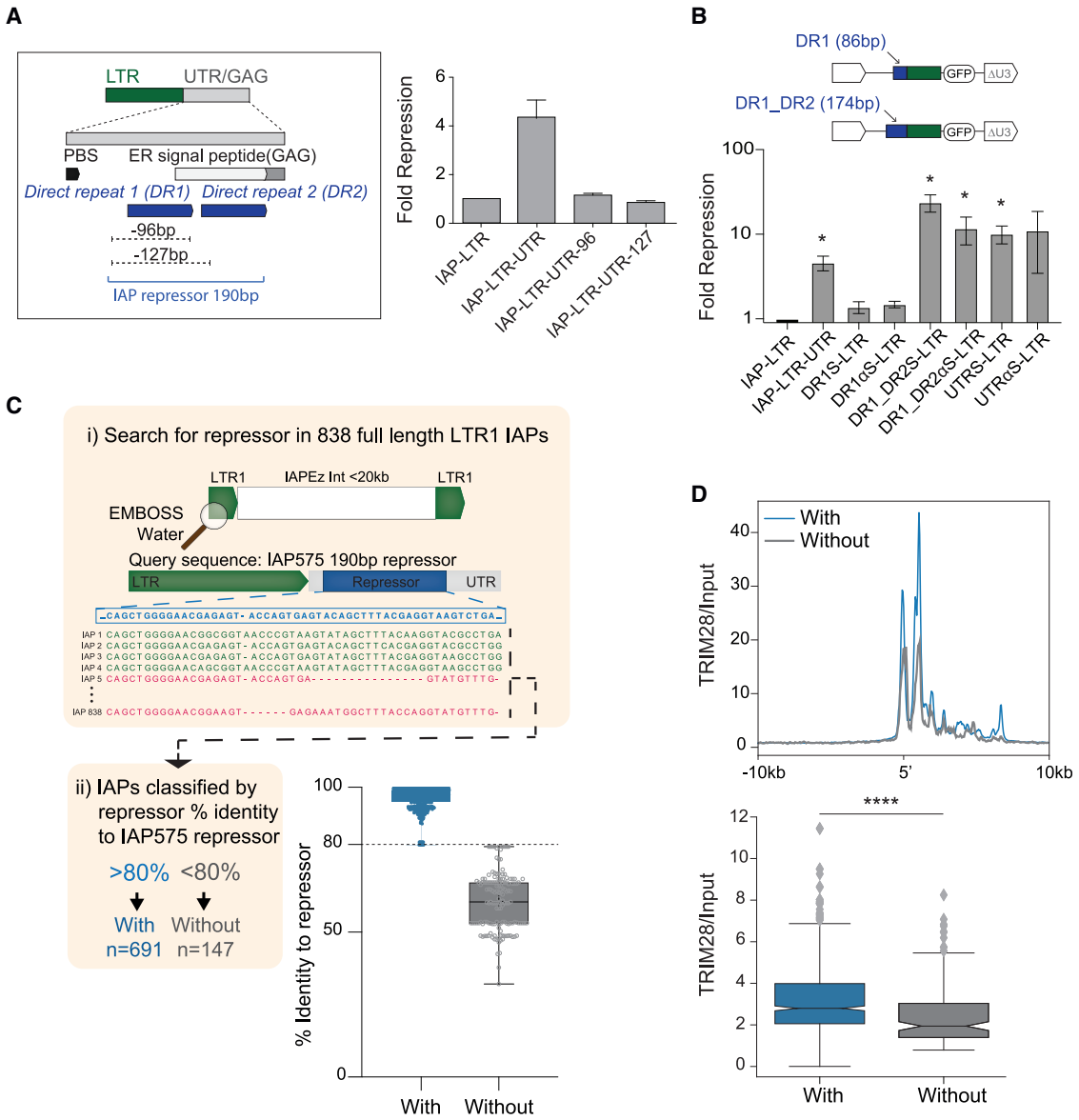


Figure 2. Epigenetic repression maps to a 190 bp sequence overlapping the IAPEz signal peptide

(A) Schematic representation of the first half of the 5′ UTR showing the direct repeats and host-derived ER-targeting signal peptide together with the 96 or 127 bp deletions (left) assayed in ESCs for fold repression as previously (right). Error bars show mean and standard deviation of three independent biological replicates.

(B) Fold repression in ESCs of reporter constructs containing one or both direct repeats upstream of the IAP LTR promoter in either orientation. Error bars show mean and standard deviation of three independent biological replicates. Significance was determined with two-tailed paired t tests, IAP-LTR-UTR p = 0.0162; DR1_DR2S-LTR p = 0.0184; DR1_DR2aS-LTR p = 0.0496; UTRS-LTR p = 0.0226. See Figure S2 for parallel data on vector integration and GFP mRNA.

(C) Depiction of the strategy used to classify sequences of full-length IAPEz elements by their percent identity to the 190 bp functional repressor sequence of IAP575. Boxplots represent first and third quartiles, where the central line corresponds to the median; whiskers are ×1.5 the interquartile range.

(D) TRIM28 signal normalized to total input for IAPEz elements separated by their identity to the functional repressor shown as a profile plot (top) or boxplot where the mean signal across the 20 kb interval depicted above is taken for each element (bottom). Mann-Whitney U test p value = 1.7e−11.

The transcriptional fate of IAPEz ERVs in NPCs is determined by their sequence

To track the epigenetic consequences of the genetic changes we observed at the repressor sequence, we utilized an *in vitro* system of ESC to NPC differentiation, where NPCs are identified by induction of the endogenous SOX1-GFP reporter (Figure S3A). Chromatin profiling and expression were assessed using CUT&RUN and RNA-seq (Figure 3A). While IAPEz elements are largely identical to one another, preventing the detection of uniquely mapping short reads within them, H3K9me3 enrichment at these elements spreads into flanking regions.[29] This phenomenon allows these elements to be mapped at the subfamily level using either multi-mapping (Figure 3B) or uniquely mapping reads (Figure S3B). H3K9me3 enrichment

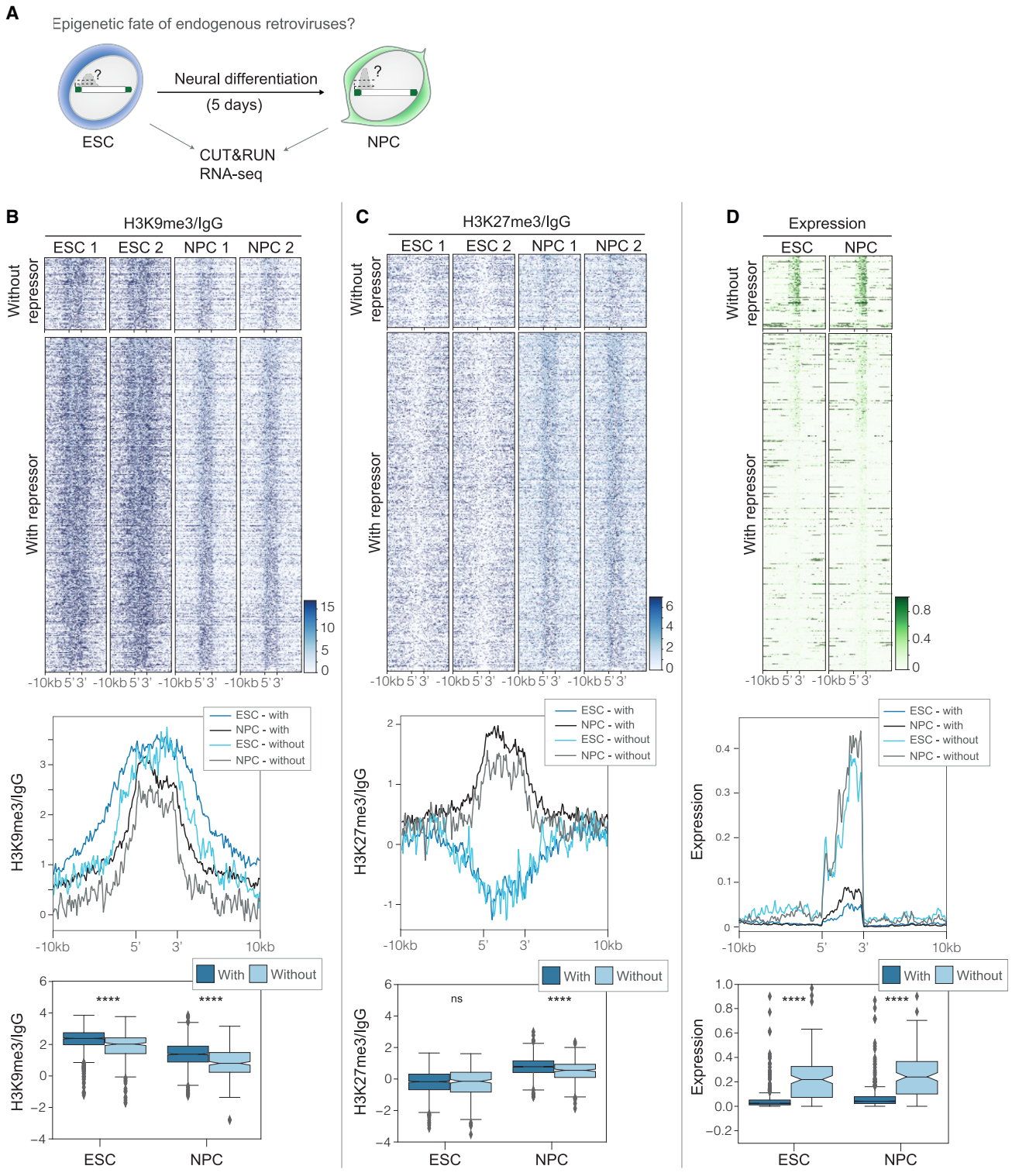


Figure 3. The epigenetic fate of IAPEz ERVs in neural progenitor cells is determined by their sequence

(A) Schematic of ESC to NPC differentiation.

(B) H3K9me3 CUT&RUN signal normalized to immunoglobulin G (IgG) across 838 IAPEz elements and surrounding 10 kb sequences to either end, separated by the presence of the repressor (with and without) and sorted by decreasing levels of intensity as a heatmap for two replicates (top). The scale bar shows the fold enrichment normalized to IgG control. Profile plots depict the trimmed mean across all elements in the indicated category, across 100 bp bins relative to the start

(legend continued on next page)

was associated with integrants containing a functional repressor sequence, consistent with the pattern of TRIM28 binding observed in ESCs. This is in line with the described TRIM28 regulation of ERVs in NPCs.[30] Interestingly, H3K9me3 spreading was prominent in ESCs but not in NPCs, with repressor-less IAPEzs exhibiting the lowest levels of H3K9me3 (Figure 3B). This suggests that IAPEzs can invoke repression on nearby genes in early development, whereas in NPCs, there is less spreading of repression and a more favorable context for epigenetic escape of the repressor-less IAPEz copies (Figure 3B, see summary plots underneath the heatmaps).

The significant loss of H3K9me3 at repressor-less IAPEzs in NPCs prompted us to measure the levels of H3K27me3, which is deposited by the PRC2 complex and is associated with repression of cell-type-specific genes.[31] Strikingly, we saw significant levels of H3K27me3 at IAPEz elements in NPCs, which was absent in ESCs (Figures 3C and S3C). It is unknown, however, whether this mark participates in repressing these elements. While H3K27me3 has been documented to safeguard IAP silencing in germ cells[32] and to repress murine endogenous retrovirus-L (MERVL) elements following induced loss of DNA methylation in ESCs,[33] an epigenetic consolidation from H3K9me3 to H3K27me3 at IAP elements has not been reported during NPC differentiation. However, it has been described at bivalent genes[34] and suggests that IAPEzs may be dynamically regulated in neural development. We next looked at RNA expression of IAPEz elements in both cell types as a proxy for the potential emergence of co-option events. Results showed unequivocally, and independently of mappability issues,[35] that a decrease in silent epigenetic marks at escapee IAPEzs, including less DNA methylation through using available data,[36] correlated with their transcriptional derepression (Figures 3D and S3D–S3G). Escapee IAPEzs exhibited a higher level of transcriptional activity in NPCs than ESCs, in support of co-option occurring in neural lineages, as well as potentially in other lineages not examined here.

The IAP LTR1 U3 region is a potent enhancer and activates nearby neural genes

The transcriptional activation of escapee IAPEz copies posed the exciting possibility that some of these integrants may harbor intact enhancers that could activate host genes. Indeed, as IAPs are derived from retroviruses, they have a conserved putative enhancer in the U3 region.[28,37] We asked if the IAP U3 could act as a classical enhancer by cloning the IAP575 U3 sequence into a reporter construct upstream of a minimal SV40 promoter in sense (S) and antisense (αS) configurations. This illustrated that the U3 sequence and the minimal putative 47 bp enhancer sequence within it function as potent enhancers in a cell-type-independent manner (Figures 4A and S4A). Further, escapee IAPEz elements are marked with H3K27ac in the developing fore- and mid-brain (embryonic day 12.5) compared with repressor IAPEz elements. This is consistent with their capacity to function as enhancers when derepressed (Figures 4B and S4B). Therefore, while the enhancer sequence is present in both repressor and escapee IAPEz elements, its activity is governed by the variable presence of the repressor.

With the aim to investigate whether escapee IAPEzs can act as activators for nearby genes, we first observed that a significant fraction (31%) of IAPEz elements overlap a gene compared with 22% of IAPs with an IAPLTR2-type LTR, which we employed as an older IAP family for comparison. In terms of distance, IAPEz elements were also significantly closer to genes than their older IAPLTR2 counterparts (Figures 4C and S4C), further implicating young IAPEz elements as candidates for co-opted gene regulatory functions. When classifying the IAPEz-proximal genes by gene type, we found their relative proportions to be largely comparable to the whole genome, except for a slight but significant depletion for protein-coding genes (Figure S4D). Of interest, interrogating the function of IAPEz-proximal genes revealed them to be enriched in synapse-associated terms (Figure 4D).

Considering both the enrichment of neural-related terms in IAPEz-proximal genes and the loss of heterochromatin at escapee IAPEzs in NPCs, we asked whether there was an association between changes in gene expression upon NPC differentiation of ESCs (log2 fold change) and the distance to the closest IAPEz depending on the presence or absence of the repressor sequence. This analysis revealed a statistically significant positive correlation between the log2 fold change (NPC/ESCs) and the distance to the closest repressor-less IAPEz. Genes proximal to canonical IAPEzs with the repressor showed the opposite trend, in contrast, potentially pointing to a dampening effect of canonical IAPEzs on the expression of nearby genes (Figure S4E).

We next compared the expression of genes proximal (within 100 kb) to either type of IAPEz in ESCs and in NPCs, showing that genes proximal to escapees are significantly more highly expressed in both cell types than those proximal to IAPEz elements with an intact repressor sequence (Figure 4E, left). We confirmed this observation in a second dataset from Bonev et al.,[39] which also included cortical neurons (Figure 4E, right). We also noted in both datasets that the enhanced expression of escapee-proximal genes was more pronounced in neural cell types, which is consistent with the enrichment for genes associated with synapse-related functional categories (Figure 4D).

The activator effect of escapee IAPEzs is explored by interrogating strain-specific insertions

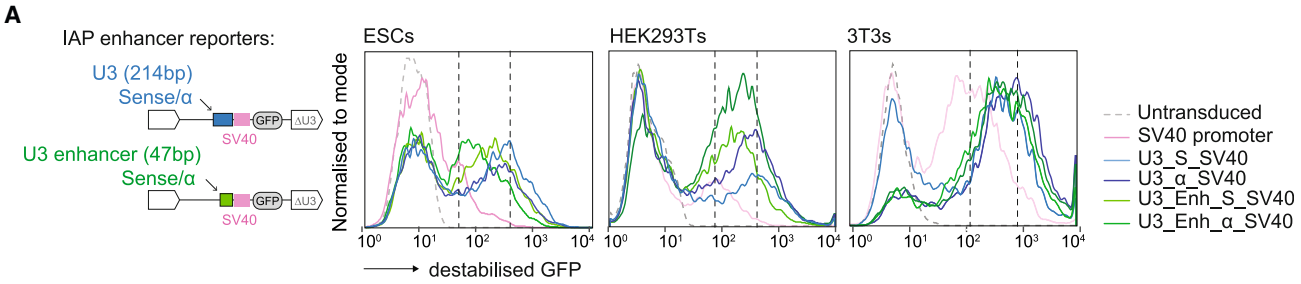
Utilizing the interstrain variability between inbred mouse strains,[40] we defined a list of 176 C57BL/6J escapee IAPEz

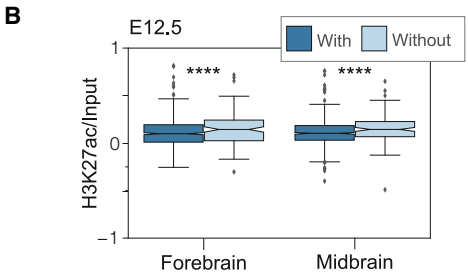
and end coordinates of the element, where each full IAP is depicted in 50 bins (middle), and boxplots show mean signal across the element and flanking regions for each IAPEz. Mann-Whitney U test, false discovery rate (FDR)-corrected p value for ESC p = 2.4e−19; NPC p = 6.4e−21.

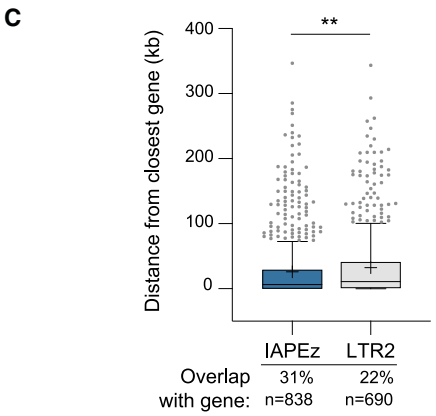
(C) H3K27me3 CUT&RUN signal normalized to IgG for two replicates as in (B), where elements in the heatmap are sorted by the decreasing H3K9me3 signal. The scale bar shows the fold enrichment normalized to IgG control. Mann-Whitney U test, FDR-corrected p value for ESC p = 0.94; NPC p = 1.6e−9. See Figure S3 for DNA methylation data.

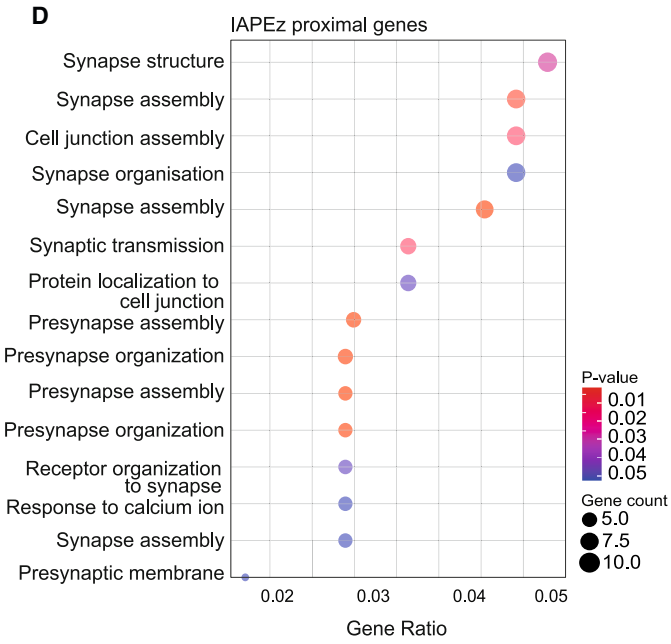
(D) Mean RNA-seq signal across 3 replicates and normalized to number of mapped reads. Rows in the heatmap are sorted by decreasing H3K9me3 signal from (B). The scale bar shows the fold enrichment of normalised expression. Boxplots show mean signal across the element and 1,000 bp flanking regions for each IAPEz. Mann-Whitney U test, FDR-corrected p value for ESC p = 4.6e−08; NPC p = 7.8e−08.

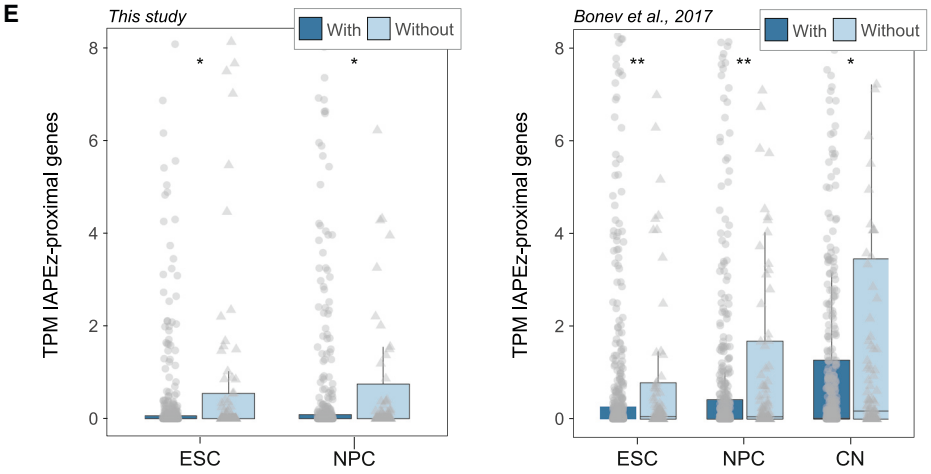
CellPress
OPEN ACCESS











(legend on next page)

elements that are absent from the 129/Ola genome to ask whether these polymorphisms resulted in interstrain gene expression differences. We first verified that there was not an overall global difference in gene expression between the two datasets using randomizations (Figure S4F). We then saw that in C57BL/6J NPCs,[39] the 154 genes proximal to these 176 elements are significantly more highly expressed than in 129/Ola NPCs (this study), indicating a putative strain-specific enhancer effect of escapee IAPEz elements (Figure S4F). On the other hand, using matched data from ESCs derived from either C57BL/6J or 129/Ola[41] showed no difference in gene expression (Figure S4F). One possibility is that a strain-specific enhancer effect of escapee IAPEzs might only be detected in neural lineages.

Significant genetic divergence of the IAPEz signal peptide underpins gain-of-enhancer events

To interrogate the genetic changes permitting escape from repression, we performed multiple sequence alignments of IAPEzs with and without the repressor and the IAP575 sequence; this analysis revealed that the main genetic divergence in the escapee sequences occurs in the ER-targeting peptide region. Performing alignments to the amino acid sequence of the ER-targeting peptide revealed a striking difference in the percent identity separating the escapee and repressed IAPEzs (Figures 5A and S5A). A closer inspection of the multiple sequence alignments of the sequence corresponding to the repressor for all 838 IAPEz elements revealed two subcategories of with-repressor IAPEzs. These were defined by whether the IAPEz, like IAP575, contained either a single copy (IAPEz with 1) or a duplication (IAPEz with 2) of a 33 bp segment of DR2. Interestingly, the escapee IAPEzs could also be subclassified according to iterations of this segment—into those either containing a duplication (IAPEz without 1) or a single copy (IAPEz without 2) or those lacking it entirely (IAPEz without 3) (Figures 5A, S5B, and S5C). Intriguingly, the subcategory with the most severe deletions with respect to the TRIM28-binding site (IAPEz without 3) exhibited the most marked mRNA expression (Figure 5A). This further supports our model in which a gain in ERV transcriptional activity stems from genetic escape from epigenetic silencing.

Finally, we examined the epigenetic signature at IAP elements that still have an intact envelope and which represent events preceding a gain in retrotransposition activity. In line with our model, we find that these copies were less targeted by TRIM28 and H3K9me3 and exhibited higher expression (Figures S5D–S5F). Rather than being subject to silencing, therefore, this subset of ERVs may even function in antiviral defense.[42,43] We propose that the early steps of co-option involve fixation of an endogenous retroviral family in the genome followed by its gain in retrotransposition activity, the molecular basis of which is targeted by TRIM28. Subsequent genetic escape in the repressor sequence enables the ERV regulatory sequences to gain activity while losing/decreasing their retrotransposition ability. Leading on from this, co-option of ERV enhancers, long non-coding RNAs (lncRNAs), and chimeric proteins can emerge (Figure 5B).

DISCUSSION

In this work, we set out to identify the early steps paving the way to co-option of retrotransposons by focusing on IAPEz elements with an LTR1/1a as young and still actively retrotransposing ERVs in the mouse genome. In doing so, our study has highlighted the sequence corresponding to the ER targeting signal as a focal point of conflict between ERVs and their hosts. On the one hand, through previous genetic and biochemical investigation,[21] this sequence has been shown to have been vital to the endogenization of these elements, while on the other hand, here we identify it to represent a genetic vulnerability targeted by TRIM28 in the ongoing evolutionary arms race between TEs and their hosts. These results shed light on why epigenetic perturbations that cause reactivation of repressed retroelements, for example in cancer, may not only lead to ectopic expression of ERV enhancers but also potentially to *de novo* retrotransposition.[44] In the human genome, this would apply to LINE-1 elements, which are the only retroelements intact for autonomous retrotransposition.[23]

We uncover a subset of IAPEz elements exhibiting sequence divergence at the ER targeting signal, illustrating a selective pressure to escape from repression. We were able to identify at least three different versions of diverged sequences, suggesting that this escape has occurred more than once. By tracking the epigenetic state of IAPEz integrants through ESC differentiation to NPCs, we could demonstrate a functional effect of genetic escape: while canonical IAPEzs succumb to a dynamic epigenetic profile involving an initial silencing in ESCs by H3K9me3, and the additional adornment with H3K27me3 in NPCs, repressor-less IAPEzs partially escape this regulation. The functional relevance of the observed H3K27me3 is an outstanding question. We envision that canonical IAPEz elements are likely to be recognized by numerous KRAB-zinc finger proteins

Figure 4. The IAP LTR1 U3 region is a potent enhancer activating nearby neural genes

(A) GFP mean fluorescent intensity (MFI) of the stated cell lines transduced with reporter constructs (shown left) containing either the entire IAP U3 region or the minimal enhancer (47 bp) within the U3, in sense and antisense orientations, upstream of a minimal SV40 promoter reporter that lacked its own enhancer. Experiments were performed three times per cell line, and representative results are shown. See Figure S4 for summary data.

(B) H3K27ac signal at IAPEz elements in E12.5 fore- and mid-brain. Boxplot shows mean signal normalized to input across the element and 2 kb flanking regions for each IAPEz. Mann-Whitney U test, FDR-corrected p for forebrain = 2e−4, mid-brain = 6.15e−06 (left). Data are from He et al.[38]

(C) Boxplot of distances between younger (IAPEz) and older (IAPLTR2/2a/2a2/2b) IAPs and their closest gene in kb. p value calculated using Welch's t test p = 0.0099; percentage of elements overlapping a gene as well as total number in each class is shown below plot.

(D) Gene set enrichment analysis to Gene Ontology annotations for genes closest to IAPEz elements. The GO terms titles are shortened to fit the figure (terms had overlapping names). The scale bars show p values and enrichment scores as bubble sizes as annotated.

(E) Boxplot depicting TPMs of genes proximal to IAPEzs. Expression values are for ESCs and NPCs generated for this study (left) and for ESCs, NPCs, and cortical neurons (CNs) from Bonev et al.[39] p values calculated using a Mann-Whitney test: this study ESC p = 0.02414; NPC p = 0.02709. Bonev et al. ESC p = 0.00696; NPC p = 0.00129; CN p = 0.02511.

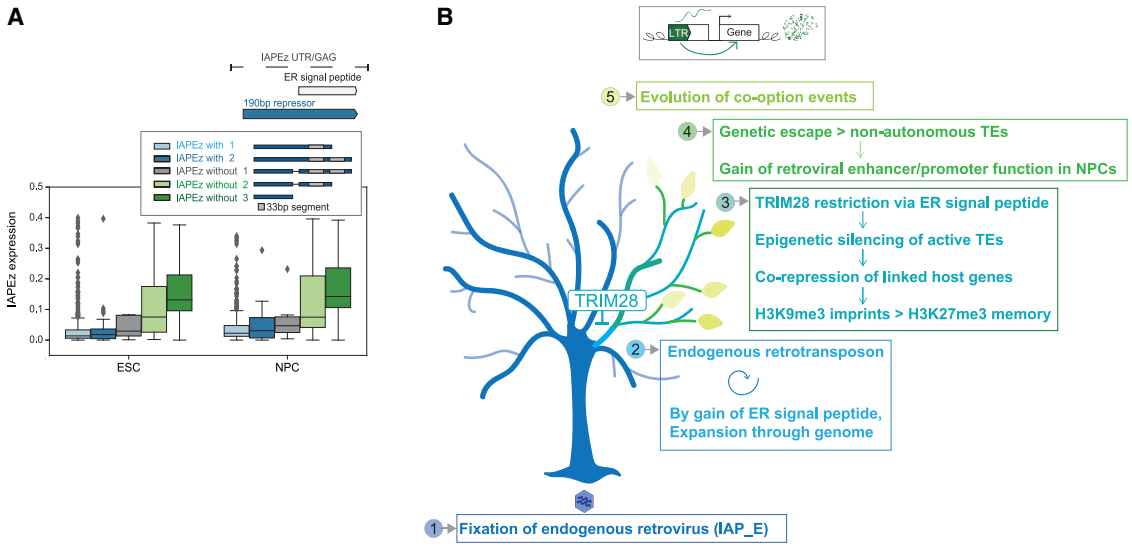


Figure 5. Dissection of genetic escape from TRIM28 repression and summary model of the proposed pathway to co-option

(A) Schematic of IAPEz UTR sequence depicting location of ER targeting peptide relative to 190 bp repressor (top). Boxplots of expression normalized to number of mapped reads, where IAPEz sequences have been further subdivided based on the observed changes in their repressor sequence (depicted in inset box) (below).

(B) Model. (1) An infectious retrovirus invades the germline and becomes fixed. (2) Evolution of an efficient retrotransposon through gain of a GAG ER-targeting signal derived from the host and loss of the envelope gene.[21] (3) Emergence of sequence-specific KZFP and TRIM28 restriction in early development for epigenetic repression of active retrotransposons through recognition of the ER-targeting signal. This involves co-repression of nearby genes through heterochromatin spreading of H3K9me3, which is consolidated with H3K27me3 following NPC differentiation. (4) Retrotransposon integrants that gain enhancer/promoter activity have undergone genetic escape through loss of the TRIM28-binding site, rendering them inactive TEs. Importantly, these retroviral enhancers/promoters are not active in ESCs, only following differentiation into NPCs. (5) Natural selection can then operate on "escapee" retrotransposons and co-opt them for host gene expression. Escapee IAPEz retrotransposons represent a snapshot of evolution in action. Gain-of-enhancer function in NPCs suggests that the brain represents a hotbed for retroviral co-option.

(KZFPs), targeting the UTR to initiate epigenetic repression via TRIM28, in a redundant manner.[45,46] Epigenetic derepression and the associated striking transcriptional activation are hallmarks of genetic escape that we identify as modifiers of the expression of nearby genes. We postulate that some of these escapees would contribute to future co-option events through natural selection.

Table 1. List of antibodies used in this article, related to key resources table in the STAR Methods

Protein	Protocol	Species	Cat. no.	Manufacturer
OCT4	FACS	rat	12-5841	eBioscience
SSEA1	FACS	mouse	eBioMC-480	eBioMC
PCNA	WB	mouse	NA03	Calbiochem
OCT4	WB	mouse	sc-5279	Santa Cruz Biotech
TRIM28	WB	rabbit	Ab10483	Abcam
NANOG	WB	rabbit	Ab80892	Abcam
H3K9me3	C + R	rabbit	Ab8898	Abcam
H3K27me3	C + R	rabbit	Ab195477	Abcam
IgG control	C + R	rabbit	12-370	Millipore
IgG control	FACS	rat	12-4321	eBioscience

Related to the key resources table in the STAR Methods. FACS, fluorescence-activated cell sorting; WB, western blotting; C + R, CUT&RUN.

As we have focused on the early stages of co-option, most escapee events that we have documented here may be neutral or deleterious, rather than functional, enhancers. Expression of ERVs even if they are not full length can lead to collateral damage. For example, ERV regulatory elements can unduly affect gene expression[47] and cell fate,[48] and retroelement-derived nucleic acids can mimic viral replication intermediates and drive interferon responses and inflammation.[22,49] Importantly though, these changes are expected to be selected against and therefore not perpetuated, while beneficial acquired activity of ERVs will be co-opted and retained. Polymorphic IAPEzs have been documented to reside as metastable epialleles with variable DNA methylation levels,[50] and it is likely that the few integrants on their way to being co-opted to regulate host genes are those that are best able to resist DNA methylation. Of note, although not focused on here, repetitive elements can also be co-opted as repressors/poised enhancers.[51]

The proximity of IAPEz elements to genes enriched in neural functions is curious. Reasons for this could relate to the longer than average length of neural-related genes[52,53] or because this lineage is permissive to some degree of perturbation.[54] In addition, and perhaps as a cause or consequence of the previous points, we have seen a gain of H3K27me3 at IAPEzs specifically in NPCs. The enrichment of this histone modification, which has been proposed to function as a placeholder of sequences to be activated later in neural development,[31] may

Table 2. shRNA and primer sequences used in this manuscript, related to key resources table in STAR Methods

	Gene		Sequence (5'-3')
shRNA	*Trim28*	sh*Trim28*	5' CCGGGCTCTCTAAGAAGCTGATCTACTCGAGTA GATCAGCTTCTTAGAGAGCTTTTTG 3'
qRT-PCR	*GFP*	forward	5' CTGCTGCCCGACAACCAC 3'
		reverse	5' ACCATGTGATCGCGCTTCTC 3'
	Cox6a1	forward	5' CTCTTCCACAACCCTCATGT 3'
		reverse	5' GAGGCCAGGTTCTCTTTAC 3'

Related to the key resources table in the STAR Methods.

point to another layer of gene regulation associated with IAPEz elements in NPCs. Future work will be needed to understand the complexity and role of epigenetic marks at IAPEz elements in late development and adult tissues. Although the strain-specific effects we have found suggest that some of these activator effects already have a functional impact on the expression of host genes, the history of IAP co-option in the mouse genome is still being written. Still, IAPEz elements appear to be remarkably well poised, both in terms of their genomic context and their epigenetic regulation, for co-option in neural tissues, as well as potentially other lineages. Looking back at ancient co-option events that have been preserved throughout millions of years indeed reveals that ERVs have been notably coerced into co-option in neural lineages.[13]

The emerging picture from our work, here on an active retrotransposon, highlights the ongoing battle between selective forces driving both the expression and further transposition of these sequences and the host genome's struggle to keep these genomic invaders in check. We envision that in the face of this evolutionary arms race, an ultimate compromise is likely to be struck whereby these invading sequences can be repurposed for the benefit of the host and thus earn the genomic space that they have colonized.

Limitations of the study

This work is focused on the first steps of co-option of endogenous retroviruses in terms of their escape from epigenetic repression and does not explain how and why individual proviral integrants become subsequently selected for by the host for gene regulatory or other beneficial host roles. Furthermore, while we document co-enrichment of H3K27me3 and H3K9me3 at repressed ERVs in NPCs, the functional significance of these dual epigenetic marks is an outstanding question.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Cell culture and reagents
- METHOD DETAILS
  - Plasmids and lentiviral vectors
  - Intracellular POU5F1 staining/SSEA1 staining
  - RNA extraction and quantification and DNA quantification
  - Western blotting
  - NPC differentiation
  - CUT&RUN
  - CUT&RUN data analysis
  - ChIP-seq data analysis
  - RNA sequencing
  - Functional analysis of genes
  - Sequence analyses
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2023.112625.

ACKNOWLEDGMENTS

We thank Austin Smith for the *Sox1*-GFP reporter mouse ESCs (mESCs) and James Briscoe for advice on NPC differentiation. We thank Connor Husovsky for technical help. We thank Miguel Branco and Pradeepa Madapura for helpful comments on the manuscript. We thank labs within QMUL Epigenetics Hub for reagents and advice and labs in the Blizard Center for Immunobiology for advice. We thank members of the Rowe Lab, Liane P. Fernandes and James Holt, for helpful discussions. Some figures were made using BioRender. This work was funded through a European Research Council starting grant (678350, TransposonsReprogram) to H.M.R., supporting R.E.-G. and P.A.G., and a Sir Henry Dale Fellowship through the Wellcome Trust and Royal Society (grant number 101200/Z/13/Z) awarded to H.M.R., which supported H.T. H.M.R. is funded by a Barts Charity Lectureship (MMBG1R).

AUTHOR CONTRIBUTIONS

Conceptualization, R.E.-G., P.A.G., and H.M.R.; methodology, R.E.-G., P.A.G., L.C., J.H., R.G., A.C., and H.M.R.; investigation, R.E.-G., P.A.G., H.T., L.C., and H.M.R.; writing, review, & editing, R.E.-G., P.A.G., and H.M.R.; funding acquisition, H.M.R.; resources, A.C., C.R.B., and H.M.R.; supervision, R.E.-G. and H.M.R.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in science. One or more of the authors of this paper self-identifies as a member of the LGBTQ+ community.

Received: July 6, 2022
Revised: April 4, 2023
Accepted: May 23, 2023
Published: June 7, 2023

REFERENCES

1. de Koning, A.P.J., Gu, W., Castoe, T.A., Batzer, M.A., and Pollock, D.D. (2011). Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. *7*, e1002384. https://doi.org/10.1371/journal.pgen.1002384.
2. Chuong, E.B., Elde, N.C., and Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science *351*, 1083–1087. https://doi.org/10.1126/science.aad5497.
3. Cornelis, G., Funk, M., Vernochet, C., Leal, F., Tarazona, O.A., Meurice, G., Heidmann, O., Dupressoir, A., Miralles, A., Ramirez-Pinilla, M.P., and Heidmann, T. (2017). An endogenous retroviral envelope syncytin and its cognate receptor identified in the viviparous placental Mabuya lizard. Proc. Natl. Acad. Sci. USA *114*, E10991–E11000. https://doi.org/10.1073/pnas.1714590114.
4. Mi, S., Lee, X., Li, X., Veldman, G.M., Finnerty, H., Racie, L., LaVallie, E., Tang, X.Y., Edouard, P., Howes, S., et al. (2000). Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature *403*, 785–789. https://doi.org/10.1038/35001608.
5. Jacobs, F.M.J., Greenberg, D., Nguyen, N., Haeussler, M., Ewing, A.D., Katzman, S., Paten, B., Salama, S.R., and Haussler, D. (2014). An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature *516*, 242–245. https://doi.org/10.1038/nature13760.
6. Rowe, H.M., Friedli, M., Offner, S., Verp, S., Mesnard, D., Marquis, J., Aktas, T., and Trono, D. (2013). De novo DNA methylation of endogenous retroviruses is shaped by KRAB-ZFPs/KAP1 and ESET. Development *140*, 519–529. https://doi.org/10.1242/dev.087585.
7. Wolf, G., Yang, P., Füchtbauer, A.C., Füchtbauer, E.M., Silva, A.M., Park, C., Wu, W., Nielsen, A.L., Pedersen, F.S., and Macfarlan, T.S. (2015). The KRAB zinc finger protein ZFP809 is required to initiate epigenetic silencing of endogenous retroviruses. Genes Dev. *29*, 538–554. https://doi.org/10.1101/gad.252767.114.
8. Cosby, R.L., Chang, N.C., and Feschotte, C. (2019). Host-transposon interactions: conflict, cooperation, and cooption. Genes Dev. *33*, 1098–1116. https://doi.org/10.1101/gad.327312.119.
9. Enriquez-Gasca, R., Gould, P.A., and Rowe, H.M. (2020). Host gene regulation by transposable elements: the new, the old and the ugly. Viruses *12*, 1089. https://doi.org/10.3390/v12101089.
10. Notwell, J.H., Chung, T., Heavner, W., and Bejerano, G. (2015). A family of transposable elements co-opted into developmental enhancers in the mouse neocortex. Nat. Commun. *6*, 6644. https://doi.org/10.1038/ncomms7644.
11. Muotri, A.R., Chu, V.T., Marchetto, M.C.N., Deng, W., Moran, J.V., and Gage, F.H. (2005). Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature *435*, 903–910. https://doi.org/10.1038/nature03663.
12. Sanchez-Luque, F.J., Kempen, M.J.H.C., Gerdes, P., Vargas-Landin, D.B., Richardson, S.R., Troskie, R.L., Jesuadian, J.S., Cheetham, S.W., Carreira, P.E., Salvador-Palomeque, C., et al. (2019). LINE-1 evasion of epigenetic repression in humans. Mol. Cell *75*, 590–604.e12. https://doi.org/10.1016/j.molcel.2019.05.024.
13. Pastuzyn, E.D., Day, C.E., Kearns, R.B., Kyrke-Smith, M., Taibi, A.V., McCormick, J., Yoder, N., Belnap, D.M., Erlendsson, S., Morado, D.R., et al. (2018). The neuronal gene arc encodes a repurposed retrotransposon Gag protein that mediates intercellular RNA transfer. Cell *172*, 275–288.e18. https://doi.org/10.1016/j.cell.2017.12.024.
14. Lu, X., Sachs, F., Ramsay, L., Jacques, P.É., Göke, J., Bourque, G., and Ng, H.H. (2014). The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. Nat. Struct. Mol. Biol. *21*, 423–425. https://doi.org/10.1038/nsmb.2799.
15. Grow, E.J., Flynn, R.A., Chavez, S.L., Bayless, N.L., Wossidlo, M., Wesche, D.J., Martin, L., Ware, C.B., Blish, C.A., Chang, H.Y., et al. (2015). Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. Nature *522*, 221–225. https://doi.org/10.1038/nature14308.
16. Ng, K.W., Attig, J., Young, G.R., Ottina, E., Papamichos, S.I., Kotsianidis, I., and Kassiotis, G. (2019). Soluble PD-L1 generated by endogenous retroelement exaptation is a receptor antagonist. Elife *8*, e50256. https://doi.org/10.7554/eLife.50256.
17. Qin, C., Wang, Z., Shang, J., Bekkari, K., Liu, R., Pacchione, S., McNulty, K.A., Ng, A., Barnum, J.E., and Storer, R.D. (2010). Intracisternal A particle genes: distribution in the mouse genome, active subtypes, and potential roles as species-specific mediators of susceptibility to cancer. Mol. Carcinog. *49*, 54–67. https://doi.org/10.1002/mc.20576.
18. Rebollo, R., Galvao-Ferrarini, M., Gagnier, L., Zhang, Y., Ferraj, A., Beck, C.R., Lorincz, M.C., and Mager, D.L. (2020). Inter-strain epigenomic profiling reveals a candidate IAP master copy in C3H mice. Viruses *12*, 783. https://doi.org/10.3390/v12070783.
19. Fehrmann, F., Jung, M., Zimmermann, R., and Kräusslich, H.G. (2003). Transport of the intracisternal A-type particle Gag polyprotein to the endoplasmic reticulum is mediated by the signal recognition particle. J. Virol. *77*, 6293–6304. https://doi.org/10.1128/jvi.77.11.6293-6304.2003.
20. Magiorkinis, G., Gifford, R.J., Katzourakis, A., De Ranter, J., and Belshaw, R. (2012). Env-less endogenous retroviruses are genomic superspreaders. Proc. Natl. Acad. Sci. USA *109*, 7385–7390. https://doi.org/10.1073/pnas.1200913109.
21. Ribet, D., Harper, F., Dupressoir, A., Dewannieux, M., Pierron, G., and Heidmann, T. (2008). An infectious progenitor for the murine IAP retrotransposon: emergence of an intracellular genetic parasite from an ancient retrovirus. Genome Res. *18*, 597–609. https://doi.org/10.1101/gr.073486.107.
22. Tunbak, H., Enriquez-Gasca, R., Tie, C.H.C., Gould, P.A., Mlcochova, P., Gupta, R.K., Fernandes, L., Holt, J., Van der Veen, A.G., Giampazolias, E., et al. (2020). The HUSH complex is a gatekeeper of type I interferon through epigenetic regulation of LINE-1s. Nat. Commun. *11*, 5387.
23. Rodriguez-Martin, B., Alvarez, E.G., Baez-Ortega, A., Zamora, J., Supek, F., Demeulemeester, J., Santamarina, M., Ju, Y.S., Temes, J., Garcia-Souto, D., et al. (2020). Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. Nat. Genet. *52*, 306–319. https://doi.org/10.1038/s41588-019-0562-0.
24. De Iaco, A., Planet, E., Coluccio, A., Verp, S., Duc, J., and Trono, D. (2017). DUX-family transcription factors regulate zygotic genome activation in placental mammals. Nat. Genet. *49*, 941–945. https://doi.org/10.1038/ng.3858.
25. Rowe, H.M., Kapopoulou, A., Corsinotti, A., Fasching, L., Macfarlan, T.S., Tarabay, Y., Viville, S., Jakobsson, J., Pfaff, S.L., and Trono, D. (2013). TRIM28 repression of retrotransposon-based enhancers is necessary to preserve transcriptional dynamics in embryonic stem cells. Genome Res. *23*, 452–461. https://doi.org/10.1101/gr.147678.112.
26. Wolf, D., and Goff, S.P. (2007). TRIM28 mediates primer binding site-targeted silencing of murine leukemia virus in embryonic cells. Cell *131*, 46–57. https://doi.org/10.1016/j.cell.2007.07.026.
27. Schlesinger, S., Lee, A.H., Wang, G.Z., Green, L., and Goff, S.P. (2013). Proviral silencing in embryonic cells is regulated by Yin Yang 1. Cell Rep. *4*, 50–58. https://doi.org/10.1016/j.celrep.2013.06.003.

28. Mietz, J.A., Grossman, Z., Lueders, K.K., and Kuff, E.L. (1987). Nucleotide sequence of a complete mouse intracisternal A-particle genome: relationship to known aspects of particle assembly and function. J. Virol. *61*, 3020–3029.

29. Rebollo, R., Karimi, M.M., Bilenky, M., Gagnier, L., Miceli-Royer, K., Zhang, Y., Goyal, P., Keane, T.M., Jones, S., Hirst, M., et al. (2011). Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms. PLoS Genet. *7*, e1002301. https://doi.org/10.1371/journal.pgen.1002301.

30. Fasching, L., Kapopoulou, A., Sachdeva, R., Petri, R., Jönsson, M.E., Männe, C., Turelli, P., Jern, P., Cammas, F., Trono, D., and Jakobsson, J. (2015). TRIM28 represses transcription of endogenous retroviruses in neural progenitor cells. Cell Rep. *10*, 20–28. https://doi.org/10.1016/j.celrep.2014.12.004.

31. Mohn, F., Weber, M., Rebhan, M., Roloff, T.C., Richter, J., Stadler, M.B., Bibel, M., and Schübeler, D. (2008). Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Mol. Cell *30*, 755–766. https://doi.org/10.1016/j.molcel.2008.05.007.

32. Huang, T.C., Wang, Y.F., Vazquez-Ferrer, E., Theofel, I., Requena, C.E., Hanna, C.W., Kelsey, G., and Hajkova, P. (2021). Sex-specific chromatin remodelling safeguards transcription in germ cells. Nature *600*, 737–742. https://doi.org/10.1038/s41586-021-04208-5.

33. Walter, M., Teissandier, A., Pérez-Palacios, R., and Bourc'his, D. (2016). An epigenetic switch ensures transposon repression upon dynamic loss of DNA methylation in embryonic stem cells. Elife *5*, e11418. https://doi.org/10.7554/eLife.11418.

34. Bilodeau, S., Kagey, M.H., Frampton, G.M., Rahl, P.B., and Young, R.A. (2009). SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. Genes Dev. *23*, 2484–2489. https://doi.org/10.1101/gad.1837309.

35. Huang, W., Li, L., Myers, J.R., and Marth, G.T. (2012). ART: a next-generation sequencing read simulator. Bioinformatics *28*, 593–594. https://doi.org/10.1093/bioinformatics/btr708.

36. Haggerty, C., Kretzmer, H., Riemenschneider, C., Kumar, A.S., Mattei, A.L., Bailly, N., Gottfreund, J., Giesselmann, P., Weigert, R., Brändl, B., et al. (2021). Dnmt1 has de novo activity targeted to transposable elements. Nat. Struct. Mol. Biol. *28*, 594–603. https://doi.org/10.1038/s41594-021-00603-8.

37. Zierler, M., Christy, R.J., and Huang, R.C. (1992). Nuclear protein binding to the 5' enhancer region of the intracisternal A particle long terminal repeat. J. Biol. Chem. *267*, 21200–21206.

38. He, Y., Hariharan, M., Gorkin, D.U., Dickel, D.E., Luo, C., Castanon, R.G., Nery, J.R., Lee, A.Y., Zhao, Y., Huang, H., et al. (2020). Spatiotemporal DNA methylome dynamics of the developing mouse fetus. Nature *583*, 752–759. https://doi.org/10.1038/s41586-020-2119-x.

39. Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G.L., Lubling, Y., Xu, X., Lv, X., Hugnot, J.P., Tanay, A., and Cavalli, G. (2017). Multiscale 3D genome rewiring during mouse neural development. Cell *171*, 557–572.e24. https://doi.org/10.1016/j.cell.2017.09.043.

40. Lilue, J., Doran, A.G., Fiddes, I.T., Abrudan, M., Armstrong, J., Bennett, R., Chow, W., Collins, J., Collins, S., Czechanski, A., et al. (2018). Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. Nat. Genet. *50*, 1574–1583. https://doi.org/10.1038/s41588-018-0223-8.

41. Ferraj, A., Audano, P.A., Balachandran, P., Czechanski, A., Flores, J.L., Varun Mosur, A.A.R., Gordon, D.S., Walawalkar, I.A., Eichler, E.E., Reinholdt, L.G., and Beck, C.R. (2022). Resolution of structural variation in diverse mouse genomes reveals chromatin remodeling due to transposable elements. Preprint at bioRxiv. https://doi.org/10.1101/2022.09.26.509577.

42. Frank, J.A., Singh, M., Cullen, H.B., Kirou, R.A., Benkaddour-Boumzaouad, M., Cortes, J.L., Garcia Pérez, J., Coyne, C.B., and Feschotte, C. (2022). Evolution and antiviral activity of a human protein of retroviral origin. Science *378*, 422–428. https://doi.org/10.1126/science.abq7871.

43. Yap, M.W., Young, G.R., Varnaite, R., Morand, S., and Stoye, J.P. (2020). Duplication and divergence of the retrovirus restriction gene Fv1 in Mus caroli allows protection from multiple retroviruses. PLoS Genet. *16*, e1008471. https://doi.org/10.1371/journal.pgen.1008471.

44. Gu, Z., Liu, Y., Zhang, Y., Cao, H., Lyu, J., Wang, X., Wylie, A., Newkirk, S.J., Jones, A.E., Lee, M., Jr., et al. (2021). Silencing of LINE-1 retrotransposons is a selective dependency of myeloid leukemia. Nat. Genet. *53*, 672–682. https://doi.org/10.1038/s41588-021-00829-8.

45. Imbeault, M., Helleboid, P.Y., and Trono, D. (2017). KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. Nature *543*, 550–554. https://doi.org/10.1038/nature21683.

46. Wolf, G., de Iaco, A., Sun, M.A., Bruno, M., Tinkham, M., Hoang, D., Mitra, A., Ralls, S., Trono, D., and Macfarlan, T.S. (2020). KRAB-zinc finger protein gene expansion in response to active retrotransposons in the murine lineage. Elife *9*, e56337. https://doi.org/10.7554/eLife.56337.

47. Morgan, H.D., Sutherland, H.G., Martin, D.I., and Whitelaw, E. (1999). Epigenetic inheritance at the agouti locus in the mouse. Nat. Genet. *23*, 314–318. https://doi.org/10.1038/15490.

48. Macfarlan, T.S., Gifford, W.D., Driscoll, S., Lettieri, K., Rowe, H.M., Bonanomi, D., Firth, A., Singer, O., Trono, D., and Pfaff, S.L. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature *487*, 57–63. https://doi.org/10.1038/nature11244.

49. Ahmad, S., Mu, X., Yang, F., Greenwald, E., Park, J.W., Jacob, E., Zhang, C.Z., and Hur, S. (2018). Breaching self-tolerance to alu duplex RNA underlies MDA5-mediated inflammation. Cell *172*, 797–810.e13. https://doi.org/10.1016/j.cell.2017.12.016.

50. Kazachenka, A., Bertozzi, T.M., Sjoberg-Herrera, M.K., Walker, N., Gardner, J., Gunning, R., Pahita, E., Adams, S., Adams, D., and Ferguson-Smith, A.C. (2018). Identification, characterization, and heritability of murine metastable epialleles: implications for non-genetic inheritance. Cell *175*, 1259–1271.e13. https://doi.org/10.1016/j.cell.2018.09.043.

51. Fernandes, L.P., Enriquez-Gasca, R., Gould, P.A., Holt, J.H., Conde, L., Ecco, G., Herrero, J., Gifford, R., Trono, D., Kassiotis, G., and Rowe, H.M. (2022). A satellite DNA array barcodes chromosome 7 and regulates totipotency via ZFP819. Sci. Adv. *8*, eabp8085. https://doi.org/10.1126/sciadv.abp8085.

52. Gabel, H.W., Kinde, B., Stroud, H., Gilbert, C.S., Harmin, D.A., Kastan, N.R., Hemberg, M., Ebert, D.H., and Greenberg, M.E. (2015). Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. Nature *522*, 89–93. https://doi.org/10.1038/nature14319.

53. Sibley, C.R., Emmett, W., Blazquez, L., Faro, A., Haberman, N., Briese, M., Trabzuni, D., Ryten, M., Weale, M.E., Hardy, J., et al. (2015). Recursive splicing in long vertebrate genes. Nature *521*, 371–375. https://doi.org/10.1038/nature14466.

54. Linker, S.B., Marchetto, M.C., Narvaiza, I., Denli, A.M., and Gage, F.H. (2017). Examining non-LTR retrotransposons in the context of the evolving primate brain. BMC Biol. *15*, 68. https://doi.org/10.1186/s12915-017-0409-z.

55. Ying, Q.L., Stavridis, M., Griffiths, D., Li, M., and Smith, A. (2003). Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture. Nat. Biotechnol. *21*, 183–186. https://doi.org/10.1038/nbt780.

56. Tie, C.H., Fernandes, L., Conde, L., Robbez-Masson, L., Sumner, R.P., Peacock, T., Rodriguez-Plata, M.T., Mickute, G., Gifford, R., Towers, G.J., et al. (2018). KAP1 regulates endogenous retroviruses in adult human cells and contributes to innate immune control. EMBO Rep. *19*, e45000. https://doi.org/10.15252/embr.201745000.

57. Robbez-Masson, L., Tie, C.H.C., Conde, L., Tunbak, H., Husovsky, C., Tchasovnikarova, I.A., Timms, R.T., Herrero, J., Lehner, P.J., and Rowe, H.M. (2018). The HUSH complex cooperates with TRIM28 to repress

young retrotransposons and new genes. Genome Res. *28*, 836–845. https://doi.org/10.1101/gr.228171.117.

58. Rowe, H.M., Jakobsson, J., Mesnard, D., Rougemont, J., Reynard, S., Aktas, T., Maillard, P.V., Layard-Liesching, H., Verp, S., Marquis, J., et al. (2010). KAP1 controls endogenous retroviruses in embryonic stem cells. Nature *463*, 237–240. https://doi.org/10.1038/nature08674.
59. Gouti, M., Tsakiridis, A., Wymeersch, F.J., Huang, Y., Kleinjung, J., Wilson, V., and Briscoe, J. (2014). In vitro generation of neuromesodermal progenitors reveals distinct roles for wnt signalling in the specification of spinal cord and paraxial mesoderm identity. PLoS Biol. *12*, e1001937. https://doi.org/10.1371/journal.pbio.1001937.
60. Krueger, F. Trim Galore. Available at: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
61. Anders, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
62. Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169. https://doi.org/10.1093/bioinformatics/btu638.
63. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.
64. Zhu, A., Ibrahim, J.G., and Love, M.I. (2019). Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. Bioinformatics *35*, 2084–2092. https://doi.org/10.1093/bioinformatics/bty895.
65. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021). clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innovation *2*, 100141. https://doi.org/10.1016/j.xinn.2021.100141.
66. Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191. https://doi.org/10.1038/nprot.2009.97.
67. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797. https://doi.org/10.1093/nar/gkh340.
68. Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189–1191. https://doi.org/10.1093/bioinformatics/btp033.
69. Eddy, S.R. (2011). Accelerated profile HMM searches. PLoS Comput. Biol. *7*, e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rat anti-OCT4	eBioscience; see Table 1	Cat# 12-5841; RRID:AB_914368
Mouse anti-SSEA1	eBioscience; see Table 1	Cat# eBioMC-480 (MC-480); RRID:AB_11217476
Mouse anti-PCNA	Calbiochem; see Table 1	Cat# NA03; RRID:AB_2160355
Mouse anti-OCT4	Santa Cruz Biotech; see Table 1	Cat# sc-5279; RRID:AB_628051
Rabbit anti-TRIM28	Abcam; see Table 1	Cat# Ab10483; RRID:AB_297222
Rabbit anti-NANOG	Abcam; see Table 1	Cat# Ab80892; RRID:AB_2150114
Rabbit anti-H3K9me3	Abcam; see Table 1	Cat# Ab8898; RRID:AB_306848
Rabbit anti-H3K27me3	Abcam; see Table 1	Cat# Ab195477; RRID:AB_2819023
Rabbit anti-IgG control	Millipore; see Table 1	Cat# 12-370; RRID:AB_145841
Rat anti-IgG control	eBioscience; see Table 1	Cat# 12-4321; RRID:AB_1518773
Bacteria and virus strains		
One Shot™ TOP10 Chemically Competent*E. coli*	Invitrogen	Cat# C404010
Chemicals, peptides, and recombinant proteins		
2-Mercaptoethanol (50mM)	Life Technologies	Cat# 31350-010
Bovine Serum Albumin	Merck, Sigma Aldrich	Cat# a9418
Penicillin-Streptomycin (10,000 U/mL)	ThermoFisher Scientific	Cat# 15140122
Mouse LIF	Sigma-Aldrich	Cat# ESG1107
PD0325901	Merck, Sigma Aldrich	Cat# 444966
CHIR99021	Merck, Sigma Aldrich	Cat# 361571
StemPro™ Accutase™	Gibco, Thermo Fisher Scientific	Cat# 11599686
DMEM	Gibco, Thermo Fisher Scientific	Cat# 11995040
Gelatine	Sigma-Aldrich	Cat# G9391
Trypsin-EDTA (0.25%)	Gibco, Thermo Fisher Scientific	Cat# 25200056
ESC FBS	Gibco, Thermo Fisher Scientific	Cat# 10439024
DMEM/F12	Gibco, Thermo Fisher Scientific	Cat# 11320033
Neurobasal	Gibco, Thermo Fisher Scientific	Cat# 21103049
N2 (100x)	Gibco, Thermo Fisher Scientific	Cat# 17502048
B27 (50x)	Gibco, Thermo Fisher Scientific	Cat# 17504044
Puromycin	Sigma-Aldrich	Cat# P8833
Fugene6	Promega	Cat# E2691
Dnase	Ambrio	Cat# AM1907
Sodium chloride solution 5M	Merck, Sigma Aldrich	Cat# 1386-1L
Triton X-100	Merck, Sigma Aldrich	Cat# X100
sodium deoxycholate	Merck, SAFC	Cat# S1827
SDS	Merck, Sigma Aldrich	Cat# 5030
Tris Buffer, 1.0 M, pH 8.0	merck, millipore	Cat# 648314
cOmplete™, Mini, EDTA-free	Roche	Cat# 11836170001
TBS	Merck, Sigma Aldrich	Cat# t5912
Tween 20	Merck, Sigma Aldrich	Cat# P1379
Laminin Mouse Protein	Gibco, Thermo Fisher Scientific	Cat# 23017015
N2 Supplement-B	StemCell Technologies	Cat# 7156
Recombinant Mouse FGF basic	R&Dsystems	Cat# 3139-FB
HEPES buffer solution, 1M	Merck, Sigma Aldrich	Cat# 83264

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Spermidine, 99%	ThermoFisher Scientific	Cat# A19096.14
Concanavalin A Conjugated Paramagnetic Beads	CUTANA™, Epicypher	Cat# 21-1401
Digitonin	Sigma-Aldrich	Cat# D141
EDTA, 0.1M	Merck, Supelco	Cat# EX0546A
pAG-MNase	CUTANA™, Epicypher	Cat# 15-1016
EDTA	Sigma-Aldrich	Cat# E3889
RNase A	ThermoFisher Scientific	Cat# R1253
AMPure XP Reagent	Beckman Coulter	Cat# A63882
Critical commercial assays		
eBioscience intracellular staining buffer kit	eBioscience	Cat# 88-8824-00
RNeasy micro kit	Qiagen	Cat# 74004
SuperScript II Reverse Transciptase kit	ThermoFisher Scientific	Cat# 18064022
SYBR green Fast PCR mastermix	Applied Biosystems	Cat# 4385612
Amersham ECL prime	Cytiva	Cat# GERPN2232
Amersham ECL select	Cytiva	Cat# GERPN2235
MinElute PCR Purification Kit	Qiagen	Cat# 28004
NEBNext® Ultra™ II DNA Library Prep Kit for Illumina®	New England BioLabs	Cat# E7645
NEBNext® Multiplex Oligos for Illumina® Index Primers Set 1	New England BioLabs	Cat# E7335
NEBNext® Multiplex Oligos for Illumina® Index Primers Set 2	New England BioLabs	Cat# E7500
Deposited data		
H3K9me3 CUT&RUN sequencing in ESCs and NPCs	This study, GEO: GSE207184	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE207184
H3K27me3 CUT&RUN sequencing in ESCs and NPCs	This study GEO: GSE207184	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE207184
RNA-sequencing in ESCs and NPCs	This study GEO: GSE207184	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE207184
Experimental models: Cell lines		
ES3 mouse embryonic stem cells	C57BL/6J mouse embryo (male), derived by Trono group (Rowe et al.[6])	NA
46C mouse embryonic stem cells	E14Tg2a.IV ES cells from 129/Ola mouse embryo (male); derived by Smith Lab, Ying et al.[55]	NA
3T3 cells	Gift from Trono Lab	NA
HEK293T cells	Gift from Rehwinkel Lab	NA
Oligonucleotides		
sh*Trim28* hairpins, see Table 2	This study	NA
GFP RT-qPCR primers, see Table 2	Tie et al., 2018[56]	NA
Cox6a1 RT-qPCR primers, see Table 2	Robbez-Mason et al., 2018[57]	NA
Recombinant DNA		
Lentiviral MND vector	Trono Lab, Rowe et al.[6]	NA
MISSION® pLKO.1 lentiviral vector	Sigma-Aldrich	SHC001
pMD2.G	Trono Lab	Addgene plasmid # 12259
p8.91	Trono Lab	NA
Software and algorithms		
BioRender	BioRender	https://www.biorender.com/
GraphPad Prism v9.5.1	GraphPad	https://www.graphpad.com/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
FlowJo	Tree Star	https://www.flowjo.com
TrimGalore v0.4.1	Babraham Bioinformatics	https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
FastQC v0.11.8	Babraham Bioinformatics	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
STAR v2.7.0f	https://doi.org/10.1093/bioinformatics/bts635	https://github.com/alexdobin/STAR
Bedtools v2.27.1	https://doi.org/10.1093/bioinformatics/btq033	https://github.com/arq5x/bedtools2
Samtools v1.10	https://doi.org/10.1093/bioinformatics/btp352	http://samtools.sourceforge.net/
Python v3.8.5	Python Software Foundation	https://www.python.org/
pyBigWig	bioconda	https://github.com/deeptools/pyBigWig
matplotlib	https://doi.org/10.1109/MCSE.2007.55	https://matplotlib.org/
R v4.1.1	R Core	https://www.r-project.org/
DESeq2	https://doi.org/10.1186/s13059-014-0550-8	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
clusterProfiler	https://doi.org/10.18129/B9.bioc.clusterProfiler	https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html
HTSeq-Count	https://doi.org/10.1093/bioinformatics/btu638	https://pypi.org/project/HTSeq/
WATER	EMBOSS	https://www.ebi.ac.uk/Tools/psa/emboss_water/
ART	https://doi.org/10.1093/bioinformatics/btr708	https://www.niehs.nih.gov/research/resources/software/biostatistics/art/index.cfm
MUSCLE v3.8.31	https://doi.org/10.1186/1471-2105-5-113	http://www.drive5.com/muscle/
HMMER v3.1b2	https://doi.org/10.1371/journal.pcbi.1002195	http://hmmer.org/
Jalview	https://doi.org/10.1093/bioinformatics/btp033	https://www.jalview.org/
Other		
Mappability tracks	Umap mm10 k100	https://bismap.hoffmanlab.org/
Nanopore Methylation rates	https://doi.org/10.1038/s41594-021-00603-8	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM5468444
TRIM28 binding data	https://doi.org/10.1038/ng.3858	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE94325
In vivo H3K27ac profiling in fore and midbrain	https://doi.org/10.1038/s41586-020-2119-x	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE100685
Ex vivo NPC expression by RNA-Seq	https://doi.org/10.1016/j.cell.2017.09.043	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE96107
ESC and NPC expression by RNA-Seq	https://doi.org/10.1101/2022.09.26.509577	https://www.biorxiv.org/content/10.1101/2022.09.26.509577v1

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Helen Rowe (h.rowe@qmul.ac.uk).

Materials availability

Plasmids generated in this study are available on request by contacting the lead author.

Data and code availability

- Original RNA-Seq and CUT&RUN sequencing data can be accessed from GEO: GSE207184.
- This paper does not report any original code.
- This paper analyses published data. The accession numbers for the datasets are GEO: GSE96107 (Bonev et al., 2017) and GEO: GSE94323 (De Iaco et al., 2017). Any additional information required to reanalyse the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell culture and reagents

ES3 mouse embryonic stem cells were derived from C57BL/6J mouse embryo (male) and were a kind gift from Prof. Didier Trono.[58] 46C ESCs[55] are a Sox1-GFP reporter cell line generated by gene targeting of E14Tg2a.IV ES cells which are derived from a 129/Ola mouse embryo (male) and were a gift from Prof. Austin Smith (University of Cambridge, UK). Mouse embryonic stem cells were cultured in N2B27 media: DMEM/F12 (Gibco, Thermo Fisher Scientific), Neurobasal (Gibco, Thermo Fisher), N2 (Gibco, Thermo Fisher), B27 (Gibco, Thermo Fisher), 0.1mM 2-Mercaptoethanol (Life Technologies) and supplemented with 0.08% BSA and 100 U/mL penicillin/streptomycin (Pen-Strep, Life Technologies), under 2i+LIF culture conditions: 1,000units/mL Leukemia inhibitory factor (LIF, Chemicon), 1μM PD0325901 (Merck, Sigma Aldrich) and 3μM CHIR99021 (Merck, Sigma Aldrich). Cells were grown at 5% CO2 at 37°C and split 1:4 every 2 days with Accutase. 3T3 and HEK293T cells were cultured in Dulbecco's Modified Eagle's Medium (DMEM; Gibco, Thermo Fisher Scientific) with 10% FBS and 100 U/mL penicillin/streptomycin, grown at 5% CO2 and split every 1:5 every 2 days with trypsin.

METHOD DETAILS

Plasmids and lentiviral vectors

Locus-specific cloning of the IAP sequences described in the text into a lentiviral MND vector[6] with a GFP reporter was carried out as follows: The IAPEz regulatory regions adjacent to the gene *Zfp575* were PCR amplified from C57BL/6 mESCs and cloned in place of the MND promoter using *XhoI* and *Bam HI*. The coordinates of the IAPEz element are chr7: 24,578,166–24,584,130. Cloning was verified by sequencing. Flow cytometry was used to measure GFP expression for the reporter assays. For RNAi, *Trim28* hairpins were designed (bioinfo.clontech.com/rnaidesigner/siRNA-SequenceDesignInit.do, Table 2) and cloned into pLKO.1 (Dharmacon) dual promoter lentviral vector with a second promoter driving puromycin resistance; the empty backbone (shControl) was used as a control. Cells were selected with puromycin for two days (or until control cells had all died) before collection and analysis. Lentiviral vectors were produced by Fugene6 co-transfection of HEK293T cells with 1.5μg of plasmid, 1μg p8.91 and 1μg pMDG2 encoding VSV-G. Ultracentrifugation of the supernatant (20,000g for 2h at 4°C) was carried out 2 days post transfection.

Intracellular POU5F1 staining/SSEA1 staining

ESCs were fixed and permeabilized using the eBioscience intracellular staining buffer kit (eBioscience 88-8824-00) and stained with POU5F1-PE, SSEA1-PerCP or isotype control and analyzed by flow cytometry. See Table 1 for antibody information.

RNA extraction and quantification and DNA quantification

Total RNA was extracted using an RNeasy micro kit (Qiagen), treated with DNase (Ambrio, AM1907). cDNA was synthesised from 500ng of RNA with SuperScript II Reverse Transciptase kit (ThermoFisher Scientific) using random primers. RT-qPCR was carried out using SYBR green Fast PCR mastermix (Life Technologies) on an ABI 7500 Real-Time PCR System (Applied Biosystems). CT values were normalised to *Cox6a1* and fold change was calculated using the -DDCt method. See Table 2 for primer sequences. To compare the relative proviral integration between reporter vectors, DNA was extracted using the Qiagen kit and DNA input normalized for the PCR and quantified by TaqMan PCR using primer and probe sets specific to GFP and to titin or albumin. For comparison, cell lines were quantified in parallel on the same qPCR plate that were known to habour one copy of GFP lentiviral vector per cell.

Western blotting

Cells were collected by trypsinisation, washed in PBS and lysed in cold RIPA buffer (150 mM NaCl; 1% Triton X-100; 0.5% sodium deoxycholate; 0.1% SDS and 50 mM Tris, pH 8.0, and protease inhibitor cocktail (cOmplete, Mini, EDTA-free, Roche)), lysates were quantified for normalisation (BCA Protein Assay kit, Millipore) and loaded on 10% denaturing SDS-polyacrylamide gels. Wet transfers were carried out onto PVDF membranes, blocked in 5% milk in TBS-T (TBS, 0.1% Tween 20 (Sigma)) and incubated with antibodies. Membranes were visualised using Amersham ECL kits. Antibodies used were: anti-PCNA, anti-POU5F1, anti-TRIM28, anti-Nanog. See Table 1 for antibody information.

NPC differentiation

46C ESCs were maintained in 2i/LIF conditions as described above and then cultured for two passages without LIF. NPCs were generated from these ESCs as follows using the protocol from,[59] with some modifications: ESCs were plated on laminin coated 6 well plates at a density of 65000 cells per well in N2B27 media (as above but with N2 Supplement-B from StemCell Technologies) supplemented with bFGF (10ng/uL bFGF (R&D)) and 1 μg/mL laminin. Cells were cultured for 5 days with daily media changes (day 1–2 with bFGF, day 3–5 without bFGF) at 7% CO2 and at day 5, cells were collected and analyzed by flow cytometry (ACEA Novocyte 3000) to measure GFP expression and used for downstream analysis.

CUT&RUN

Cut&Run was carried out according to the EpiCypher CUTANATM CUT&RUN Protocol (v1.6) (https://www.epicypher.com/resources/protocols/cutana-cut-and-run-protocol/). 100,000 2i+LIF-cultured ESCs or day 5 NPCs were collected per sample/antibody and washed twice with EpiCypher Wash buffer (20mM HEPES pH 7.5, 150mM NaCl, 0.5mM spermidine) plus protease inhibitors (cOmplete, Mini, EDTAfree, Roche) before attachment to activated Concanavalin A coated magnetic beads. Beads and cells were resuspended in Antibody Buffer (20mM HEPES pH 7.5, 150mM NaCl, 0.5mM Spermidine, 1x protease inhibitors, 0.01% w/v digitonin, 2mM EDTA) with antibodies (1ug of anti-H3K9me3, anti-H3K27me3 or IgG control; Table 1) and incubated overnight at 4°C with gentle rocking. The next day beads were washed twice in cold Digitonin Buffer (20 mM HEPES pH 7.5, 150mM NaCl, 0.5mM Spermidine, 1x Roche complete protease inhibitors, 0.01% digitonin) and then incubated with Digitonin Buffer plus 2.5μL pAG-MNase (CUTANA, Epicypher) before addition of 2mM $CaCl_2$ to activate cleavage at 4°C for 2h. The reaction was quenched by addition of 33μL Stop Buffer (340mM NaCl, 20mM EDTA, 4mM EGTA, 50 μg/mL glycogen, 50 μg/mL RNase A), vortexted and incubated at 37°C for 10minutes to enable the release of DNA fragments. Sample was cleaned using a magnetic rack and the supernatant containing DNA was purified with a MinElute PCR Purification Kit (Qiagen). Libraries were prepared according to manufacturer's instructions using the following kits: NEBNext Ultra II DNA Library Prep Kit for Illumina and NEBNext Multiplex Oligos for Illumina (New England BioLabs) and pooled in equimolar quantities. Sequencing was carried out on a Novaseq6000 with 150bp PE reads.

CUT&RUN data analysis

Reads were trimmed and adapters removed using TrimGalore v0.4.1[60]; quality was checked using FastQC v.0.11.8.[61] Alignment to the GRCm38 mouse genome was performed using STAR (Dobin et al., 2013) with parameters adjusted to generate one random location for multimapping reads *[–outFilterMultimapNmax 5000 –outSAMmultNmax 1 –outFilterMismatchNmax 999*]. The genomecov tool from bedtools was used to generate BedGraph files of genome coverage, scaled by library size, and converted to BigWig files. The pybigwig library in python3 was used to retrieve coverage over 100bp windows. CUT&RUN signal was normalised to IgG signal, this data was represented at heatmaps, profile plots and boxplots using the matplotlib and seaborn libraries. Alignments of uniquely mapping reads were also performed with STAR with the *–outFilterMultimapNmax 1* option and processed as above.

ChIP-seq data analysis

Data from[24](GSE94323), was downloaded using the SRA toolkit. STAR alignments of reads were performed as above, allowing for one random location for multimapping reads. Alignments were processed similarly as CUT&RUN data to generate genomewide coverage in BigWig format for IP and input samples. IP signal was normalised to input in python3 using the pybigwig library and depicted was profile plots or heatmaps as indicated.

RNA sequencing

RNA was extracted as described above and RNA quality and quantity was assessed using a Spectrophotometer UV5 (Mettler Toledo). Preparation of mRNA libraries was carried out by Novogene Co. Ltd and sequenced on the Illumina NovaSeq6000 with 150bp PE reads. Data was demultiplexed and fastq files generated using the bcl2fastq software from Illumina. TrimGalore v0.4.1 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) was used for read trimming and adapter removal and FastQC v.0.11.8[61] was used for quality checking. Reads were aligned to the GRCm38 mouse genome using STAR *[–outFilterMultimapNmax 5000 –outSAMmultNmax 1 –outFilterMismatchNmax 999*]. Read counts per gene were obtained with HTSeq-Count[62] and differential expression analysis was performed with DESeq2[63] under R v4.1.1 was used to call differential expression analysis for genes, and the Approximate Posterior Estimation method[64] was used to shrink the logarithmic fold change. TPM values were calculated in R. For depiction of expression signal, genome coverage scaled by library sized was calculated with the genomecov tool of bedtools and converted to bigwig to be processed in python3 as above. Data from[39] (GSE96107) were downloaded using the SRA toolkit and the reads mapped with STAR and TPMs were calculated as above.

Functional analysis of genes

bedtools closest tool (v2.27.1) was used to call the closest gene and distance to each IAP. GO analysis was performed on the list of closest genes using the Bioconductor clusterprofiler tool[65] in R. Genes were classified according to their transcript biotype where this information was collected using the Bioconductor biomaRt tool[66] in R.

Sequence analyses

The intersect and closest tools from bedtools were used to obtain full-length IAPEz elements which complied with the following: contained two IAPLTR1 or IAPLTR1a, in the same orientation and were separated by less than 20kb. We then verified that elements which fulfilled the previous also overlapped a 'IAPEz' internal sequence. The sequences corresponding to these elements were extracted with the getfasta tool from bedtools. Matches to the 190bp repressor sequence of IAP575 were calculated with the water tool from EMBOSS *[-gapopen 10 -gapextend 0.5*] and the %identity was used to classify elements based on their repressor sequences. Multiple sequence alignments were generated with muscle[67] and visualized using Jalview.[68] Alignments were manually curated to subcategorise repressor types and consensus sequences were generated with HMMER[69] using the hmmbuild and hmmemit tools.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data shown in this study are shown with error bars representing standard deviation. Where shown statistical significance was assessed with two tailed, paired Student's t tests or as described in the figure legends using GraphPad Prism or R v4.1.1. Biological replicates are denoted in the figure legends. For flow cytometry 10,000 events were recorded. p-values of <0.05 were considered significant (****$p < 0.0001$, ***$p < 0.001$, **$p < 0.01$ and *$p < 0.05$) and p-values are shown in the figure or legends. Error bars show the standard deviation.