Endogenous retroviruses co-opted into divergently transcribed regulatory elements shape the regulatory landscape of embryonic stem cells

Stylianos Bakoulis[1], Robert Krautz[1], Nicolas Alcaraz[1,2], Marco Salvatore[1], Robin Andersson[1,*]

[1]Department of Biology, University of Copenhagen, Copenhagen, Denmark
[2]Novo Nordisk Foundation Center for Protein Research (CPR), University of Copenhagen, Copenhagen, Denmark
*To whom correspondence should be addressed: robin@binf.ku.dk

Abstract

Transcription factor binding to regulatory elements is the key process underlying gene regulation during cellular differentiation. Although the specific regulation of genes by transcription factors is generally conserved, regulatory elements themselves are associated with high evolutionary turnover, a process that has been attributed to transposable elements. However, it is unclear how frequent co-option of transposable elements into regulatory elements is and to which regulatory programs they contribute. Here, we report an in-depth characterization of the transposon-derived regulatory landscape of mouse embryonic stem cells. We demonstrate that a substantial number of endogenous retroviral elements are divergently transcribed into unstable RNAs, and that these elements contribute to a sizable proportion of active enhancers and gene promoters. We further show that transposon subfamilies contribute to specific regulatory programs through their enrichment of binding sites for transcription factors, shedding light on the formation of regulatory programs and the origins of regulatory elements.

Introduction

Transcriptional regulatory elements are stretches of genomic sequence that exert enhancer and promoter activities essential for the precise spatial and temporal control of gene expression (Haberle and Stark 2018; Beagrie and Pombo 2016; Shlyueva et al. 2014; Andersson and Sandelin 2020). Identifying regulatory elements and deciphering their combinatorial logic have thus emerged as novel disciplines to untangle the mechanisms underlying cell differentiation. A key property of regulatory active elements is that they initiate divergent bidirectional transcription (Andersson et al. 2014b; Scruggs et al. 2015; Andersson et al. 2014a; Koch et al. 2011; Andersson et al. 2015a). This is established at closely spaced pairs of divergently oriented core promoters, resulting in long non-coding enhancer RNA (eRNA) transcripts at regulatory elements with enhancer activity, and pairs of mRNAs and promoter upstream transcripts (PROMPTs) at gene promoters. Their transcriptional output can vary, which correlates with histone modifications of nucleosomes flanking their chromatin-accessible sites (Andersson and Sandelin 2020; Core et al. 2014; Henriques et al. 2018; Rennie et al. 2018).

The cell-type specific activity of a regulatory element is controlled by the specificity of transcription factors (TFs) to bind the element and the density of their binding sites, both of which are in turn dependent on its DNA sequence (Nguyen et al. 2016; Weingarten-Gabbay et al. 2019; Smith et al. 2013; Grossman et al. 2017). TF DNA-sequence preferences (Nitta et al. 2015), gene expression (Chan et al. 2009; Berthelot et al. 2018), and the specific regulation of genes by TFs (Odom et al. 2007; Schmidt et al. 2010) are generally well conserved across eukaryotes. On the contrary, regulatory elements with enhancer activity are generally associated with high evolutionary turnover (Vierstra et al. 2014; Young et al. 2015; Villar et al. 2015).

Fixation of transposable elements (TEs) in the host genome by integrating transposons into germline chromatin is one mode of regulatory innovation. In mammals, the large majority of such endogenous retroviruses (ERVs) are dormant, having lost their ability to replicate and are maintained repressed by H3K9me3, H3K27me3 and DNA methylation (Walter et al. 2016; Maksakova et al. 2013; Matsui et al. 2010; Karimi et al. 2011; Rowe et al. 2013). However, favorable genomic

integration may lead to their co-option into endogenous regulatory elements by utilizing their native retroviral or subsequent evolutionary acquired TF binding sites (Bourque et al. 2008; Chuong et al. 2013; Sundaram et al. 2014; Trizzino et al. 2017; Sundaram et al. 2017; Sun et al. 2018; Cao et al. 2019; Todd et al. 2019; Miao et al. 2020). For example, in embryonic stem cells (ESCs) long terminal repeat (LTR) retrotransposons bound by pluripotency TFs, including OCT4 and NANOG, have been shown to possess enhancer activities (Todd et al. 2019) and initiate transcription of eRNA-like transcripts (Fort et al. 2014). In addition, species-specific TE-derived regulatory elements cause binding differences of pluripotency TFs between human and mouse ESCs (Kunarso et al. 2010). Thus, systematically characterizing the contribution of TEs to transcriptional regulation has the potential to provide insights into the formation of regulatory programs, the origins of regulatory elements and thus the basis for their evolutionary turnover.

Although the process of regulatory innovation by TEs is well established, its extent and how TE-derived regulatory elements compare to non-TE associated regulatory elements with regards to divergent transcription initiation, RNA metabolism, TF binding, regulatory function and associated chromatin signatures remain unclear. This is due to their repetitive nature in their host genomes, which makes mapping of short sequencing reads to such genomic locations complicated. Encouraged by the possibility to study TE-associated regulatory elements through transcription initiation mapping (Fort et al. 2014), we here systematically investigate how the wide repertoire of mouse TEs contribute to the transcription initiation landscape of mouse ESCs (mESCs). We find that many dormant TEs, in particular ERV TE family elements, are divergently transcribed into unstable RNAs that are targeted by the 3'-5' nuclear exosome for degradation. Furthermore, a large fraction of open chromatin regions carry transcribed ERVs with enhancer potential or contribute to mRNA gene promoters, together reinforcing the process of TE co-option into regulatory elements. Finally, we characterize the regulatory potential of ERV subfamilies, including those primarily serving as binding sites for pluripotency TFs, and demonstrate that TF binding specificity maintained by different ERV subfamilies contributes to a diverse set of regulatory programs.

Results

Transcription at transposable elements is divergent and results in unstable RNAs

To investigate the prevalence and co-option of TEs as regulatory elements, we first characterized the association of TEs with divergent transcription initiation, a key property of regulatory elements with either enhancer or promoter activities (Andersson et al. 2014a, 2015b; Andersson and Sandelin 2020; Rennie et al. 2018; Mikhaylichenko et al. 2018; Henriques et al. 2018). To this end, we analyzed the genomic location of mapped sequencing reads of capped RNA 5' ends (CAGE: Cap Analysis of Gene Expression (Takahashi et al. 2012)) from mESCs depleted for the ribonucleolytic RNA exosome core component Rrp40 (Lloret-Llinares et al. 2018). Transcription start sites (TSSs) identified by this data thus also include those whose RNAs are targeted by the exosome for degradation.

Of all mouse TE insertion events (3,677,522), only 1.3% (46,424) were associated with measurable RNA polymerase II transcripts in mESCs (Methods). Among these, endogenous retroviral TE family elements (ERV1, ERVK, ERVL, ERVL-MaLR) were most prevalent ($p<2e-16$, Fisher's exact test; Supplementary Fig. 1A). However, expressed TEs were associated with ~4% of all identified TSSs in mESCs and ~2% of uniquely mapped CAGE reads. Thus, only a small fraction of annotated TEs exhibit transcription initiation and, likewise, only a small fraction of CAGE reads originate in TEs.

Since repeats are inherently hard to map uniquely with short sequencing reads, we investigated the feasibility of probabilistically assigning multi-mapping reads by the locations of TSSs and their expression levels inferred from uniquely mapped CAGE reads (Faulkner et al. 2008, 2009; Hashimoto et al. 2009). Employing the MuMRescueLite multi-mapping rescue approach (Hashimoto et al. 2009) led to an increase in the number of detected expressed TEs (the number of detected TEs increased by 77.5% to 82,383 TEs) and an overall higher expression level (Supplementary Figs. 1B, 2, 3). Importantly, the inferred expression levels of TE-derived RNA by this approach largely agreed with an alternative quantification strategy (TELocal) (Jin et al. 2015) based on maximum likelihood alignments to

annotated repeats (Spearman's rho=0.752, $p<0.01$; Supplementary Fig. 4). However, the benefit of being able to characterize and quantify the expression levels of TSSs genome-wide and not having to rely on custom annotations of TE insertions made us opt for the probabilistic multi-mapping rescue strategy. We further investigated how rescuing multi-mapping CAGE reads affects expression level quantification of genes with TE-associated promoters in HeLa cells. Using generalized linear Poisson regression, we observed that the multi-mapping rescue approach improved the agreement between promoter-derived gene expression levels inferred from HeLa CAGE (Andersson et al. 2014b) with those from HeLa RNA-seq data (Andersen et al. 2013) quantified from exonic reads ($p=0.0112$ and $p<2e-16$ for uniquely mapped CAGE data versus data also including rescued reads, respectively, F-test). This shows that quantification of TE-derived transcripts with CAGE may underestimate their abundances and that rescuing multi-mapping reads can alleviate some of these challenges.

The ability to infer the TSSs and expression levels of TE-derived RNAs from CAGE allowed us to characterize the transcription initiation patterns across TE families (Fig. 1A). LTRs, in particular endogenous retroviral families (ERV1, ERVK, ERVL and ERVL_MaLR), and L1 LINE elements showed preferential TSS locations close to their repeat body boundaries (Fig. 1B). TSS location preferences for these families could not be explained by varying mappability over repeat bodies and flanks (Supplementary Fig. 6). We note, however, that lack of detected expression in DNA transposons could be due to low mappability, suggesting that expression quantification and TSS mapping of some repeats may require alternative approaches or longer sequencing reads.

The average profiles of TSS locations for ERVs and L1 LINEs (Fig. 1A,B; Supplementary Fig. 6) indicated divergent transcription initiation, reminiscent of that of gene promoters and gene-distal enhancers (Kim et al. 2010; Andersson et al. 2014a, 2014b; Chen et al. 2016). To investigate divergent transcription in individual genomic TE instances, we quantified transcriptional directionality in TE-associated open chromatin loci, as measured by DNase I hypersensitive sites (DHSs). The large majority of TE-associated CAGE-derived TSSs were proximal to DHSs (Supplementary Fig. 7) and the majority of TE-associated DHSs displayed balanced

bidirectional transcription initiation (Fig. 1C), a hallmark of gene-distal regulatory elements with enhancer activity (Andersson et al. 2014a). This suggests that some of the investigated TEs act as enhancers.

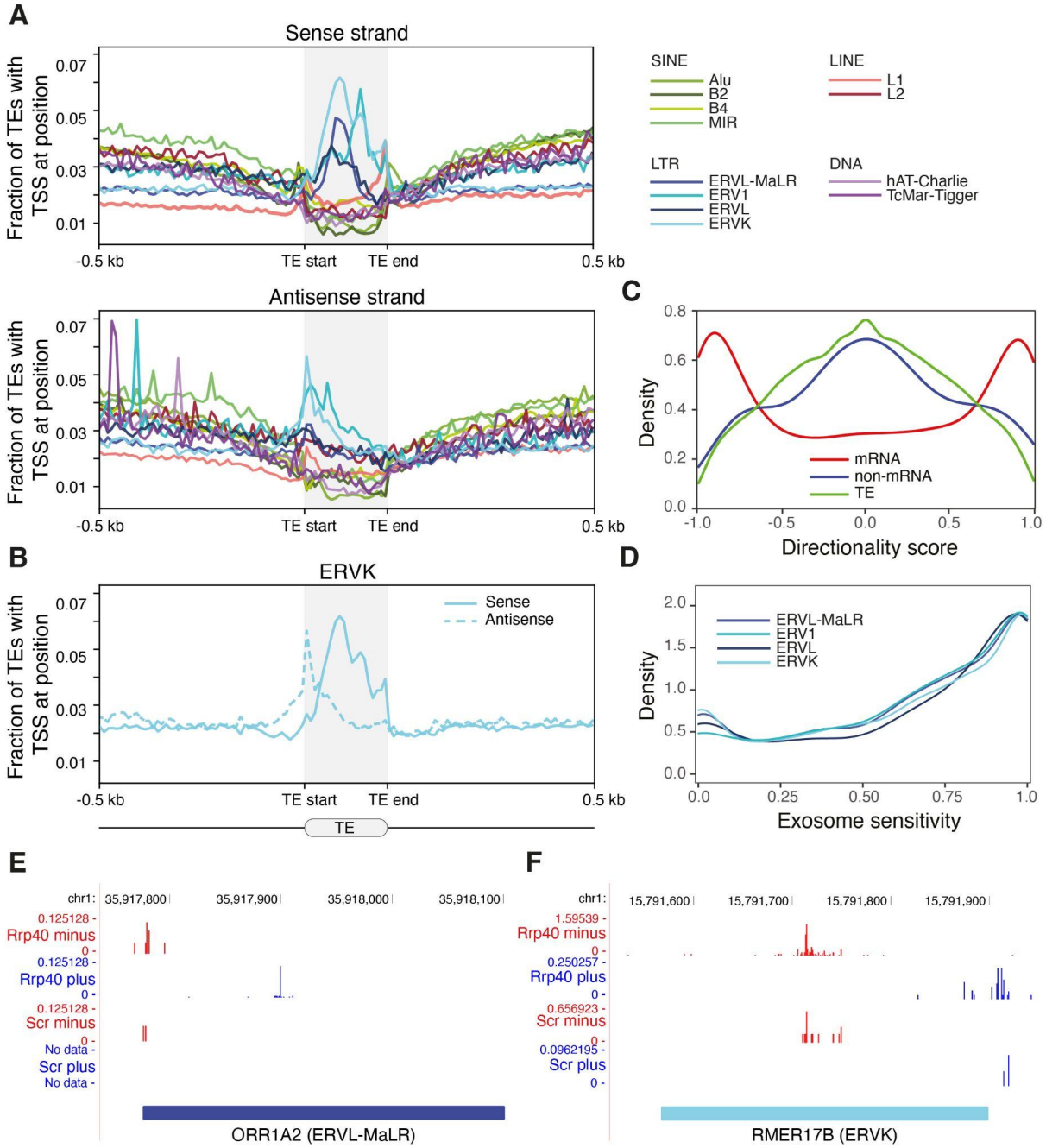


***Figure 1**: **TEs are divergently transcribed into unstable RNAs. A:** Average distribution of CAGE-inferred TSS locations (vertical axis; expression agnostic) +/- 500 bp upstream/downstream and across the body of major TE families (horizontal axis). TSS locations are visualized separately for the sense (upper panel) and antisense (middle panel) strands. **B:** Average distribution of CAGE-inferred TSS locations for the ERVK family. **C:** Transcriptional directionality score, describing the strand bias in expression levels (ranges between -1 for 100% minus strand expression and +1 for*

100% plus strand expression), for mRNA, non-mRNA (non-protein-coding GENCODE transcripts) and TE-associated RNAs (regardless of annotation). ***D:*** *Exosome sensitivity, measuring the relative amount of exosome degraded RNAs (ranges between 0 for RNAs unaffected by the exosome and 1 for 100% unstable RNAs), for transcripts associated with LTR families ERV1, ERVK, ERVL, and ERVL-MaLR.* ***E-F:*** *Genome browser tracks for two loci of unannotated transcripts with characteristic divergent expression patterns falling on TE insertions of ORR1A2 (ERVL-MaLR;* ***E****) and RMER17B (ERVK;* ***F****) subfamilies. Pooled replicate CAGE expression levels in control (Scr) and after exosome depletion (Rrp40) split by plus (blue) and minus (red) strands are shown. For visibility reasons, the scales of CAGE signals differ between strands and conditions.*

Divergent transcripts from enhancers and gene promoters are frequently associated with nuclear decay (Preker et al. 2008; Ntini et al. 2013; Andersson et al. 2014b). Comparing CAGE data from exosome-depleted mESCs with wildtype mESCs (scrambled shRNA control) (Lloret-Llinares et al. 2018) confirmed that TE-derived RNAs are also degraded by the exosome (Fig. 1D; Supplementary Figs. 1-3,5; Supplementary Table 1), as exemplified by CAGE data at genomic loci containing insertion sites for ORR1A2 (ERVL-MaLR) and RMER17B (ERVK) (Fig. 1E-F). Across TE families, we generally observed more TE-associated TSSs in exosome-depleted mESCs compared to wild type mESCs (73,246 versus 28,215). Overall, a higher fraction of TEs with transcription initiation was detected in exosome-depleted (1.75%) versus wildtype (<1%) mESCs (26,079 in wildtype mESCs; 64,299 in exosome-depleted mESCs; 82,383 in either exosome-depleted or wildtype mESCs, pooled data). Exosome-depleted cells further displayed an increased expression level of TE-derived RNAs (Supplementary Fig. 1; Supplementary Table 1). Together, these results indicate that, although a prominent number of TEs are transcribed, the majority of derived RNAs are degraded.

To assess how generalizable these results are beyond mESCs, we next analyzed TE-associated transcription initiation events using CAGE data from exosome-depleted HeLa cells (Andersson et al. 2014b). We observed a similar increase in TE expression levels upon exosome depletion, and that probabilistically rescued multi-mapping reads increased the number of detected transcribed TEs (Supplementary Fig. 8).

Taken together, our results demonstrate that transcribed mammalian TEs are associated with divergent transcription initiation of RNAs that are substrates for endonucleolytic decay, which implies the potential for TEs to function as enhancers.

Transcription of dormant ERVs reveals co-opted regulatory elements

The characteristics of TE-derived transcripts and their similarities with eRNAs and PROMPTs led us to investigate whether transcribed TEs carry further similarities with regulatory elements. Genomic annotation of transcribed TEs in mESCs revealed that a substantial fraction was either located in gene-distal intergenic regions or overlapped with gene promoters (Fig. 2A). Across all transcribed TE families, ERV- and L1-families contained the biggest fractions of TEs in intergenic regions localized at least 10 kb from the nearest gene, indicating their preference for gene-distal regulatory elements (50.6% of ERVs on average across ERV families and 59.8% of L1 LINEs). In contrast, other TE families displayed an elevated proportion of expressed TEs in genic regions. Overall, transcribed TEs generally overlapped with open chromatin regions as defined by DHSs (62,514, ~65%, of transcribed TEs overlapped with 7,431 DHSs). This suggests that a prominent fraction of transcribed TEs may act, alone (34% of DHSs overlap one transcribed TE) or in combination with other transcribed TEs (66% of DHSs overlap multiple transcribed TEs), as *bona fide* gene regulatory elements (Fig. 3).

To generally assess whether transcribed dormant TEs have been co-opted into transcriptional regulatory elements, we first evaluated their association with FANTOM enhancers (Andersson et al. 2014a; Arner et al. 2015; Dalby et al. 2018), an extensively validated set of regulatory elements with predicted enhancer function from bidirectional transcription initiation across a large number of cell types and tissues. Notably, 1,979 out of the 8,942 FANTOM mouse enhancers with detectable expression in mESCs (22.1%) overlapped with 1,504 out of the 7,431 TE-associated DHSs in mESCs (20.2%; p=1e-5, Fisher's exact test). This indicates that TEs contribute to a sizable fraction of regulatory elements with enhancer activity. Moreover, half of this set is composed of DHSs with LTR elements, in particular ERVKs and ERVL-MaLRs (Fig. 2B; Supplementary Fig. 9A), consistent with previously reported bidirectional transcription of ERVs (Fort et al. 2014). Of note, given that FANTOM enhancers were neither identified from exosome depleted

mESCs nor multi-mapping rescued reads, it is conceivable that some TE-associated enhancers are not present in the FANTOM enhancer set.

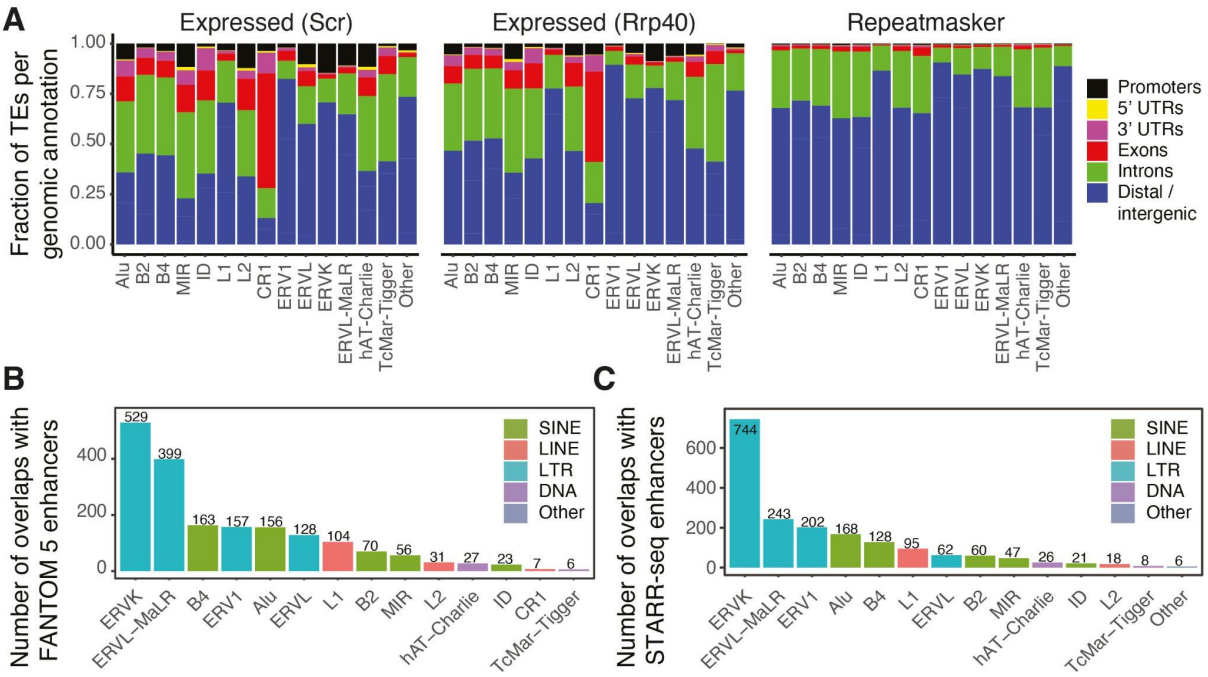


***Figure 2**: **Transcription of transposable elements reveals co-opted regulatory elements. A:** Fraction of expressed TEs in control (Scr) and exosome KD (Rrp40) mESCs as well as all TEs annotated in Repeatmasker (expression agnostic) per genomic annotation group for each TE family. **B:** The number of transcribed TEs overlapping FANTOM 5 mouse enhancers at the TE family level. TE subfamily counts are displayed in Supplementary Figure 9A. **C:** The number of transcribed TEs overlapping STARR-seq mESC enhancers at the TE family level. TE subfamily counts are displayed in Supplementary Figure 9B.*

Since FANTOM enhancers were predicted based on the same property that we had observed for transcribed TEs, namely balanced transcriptional directionality, alternative experimental data is required to evaluate the true enhancer potential of TEs (Halfon 2019). To this end, we considered the *in vitro* enhancer potential of transcribed TEs using genome-wide mESC STARR-seq data (Peng et al. 2020). 1,404 out of the 7,431 transcribed TE-associated DHSs (18.9%) overlapped with 1,949 out of 7,078 identified open chromatin-associated STARR-seq enhancers (26.2%; Fig. 2C; Supplementary Fig. 9B). Again, LTRs, in particular ERVKs, constitute a sizable fraction of these enhancers, in agreement with the FANTOM enhancer overlap. In agreement with the strong association between transcription initiation and *in vitro* enhancer potential (Andersson et al. 2014a; Wu et al. 2014;

Rennie et al. 2018), we observed that non-transcribed TE-associated DHSs were to a lesser degree validated by STARR-seq enhancers (p<2e-16 for ERVKs and ERVL-MaLR, Fisher's exact test). Combined, the STARR-seq and FANTOM sets indicate that 2,378 (~32%) transcribed TE-associated DHSs in mESCs are likely enhancers. Example loci are displayed in Figure 3, which clearly illustrate the association between divergent transcription initiation and enhancer activity from either individual TEs (Fig. 3A; RLTR41 insertion) or pairs of TEs (Fig. 3B; RLTR41 in pairs with MYSERV-int and RMER10B insertions).

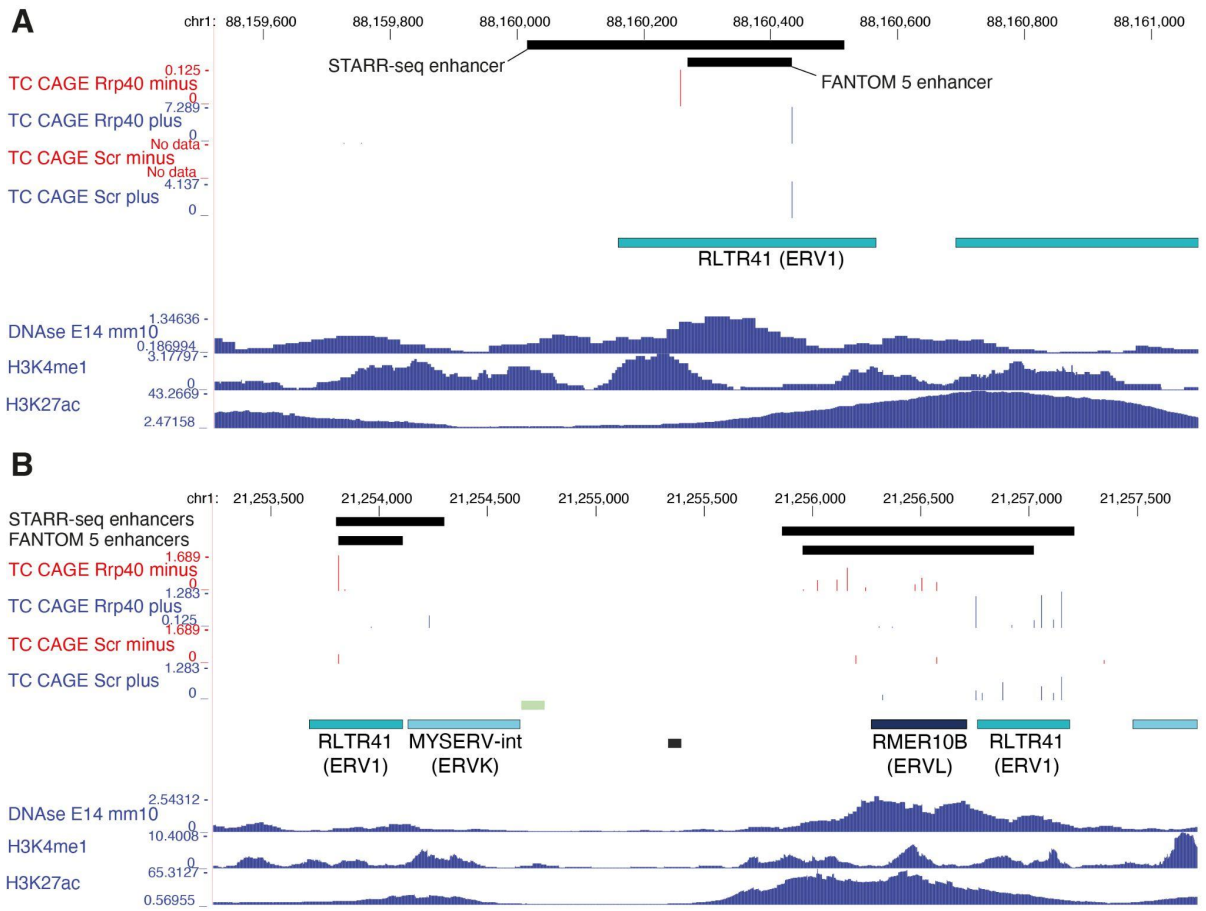


***Figure 3**: **TE insertions co-opted into divergently transcribed enhancers. A-B:** Genome browser tracks for two intergenic loci showing TPM-normalized CAGE data pooled across replicates and split by plus (blue) and minus (red) strands. Shown are also the locations of FANTOM5 mouse enhancers and STARR-seq mESC enhancers and signal tracks for ENCODE DNase-seq data and H3K4me1 and H3K27ac ChIP-seq data for E14 mESCs. The CAGE signals identify divergent transcription initiation from ERV1 RLTR41 insertions, alone (**A**) and in pairs (**B**) with TE insertions of RMER10B and MYSERV-int.*

The strong overlap of ERVK elements with FANTOM and STARR-seq enhancer sets implies differences in the regulatory potential among TE families (Fig. 2B,C). To validate these differences further, we assessed the chromatin states at transcribed TEs. For this, ChIP-seq signal for histone modifications was aggregated around CAGE-inferred TSSs in DHSs (The ENCODE Project Consortium 2012; Yue et al. 2014). Hierarchical clustering of these aggregated signals revealed six major chromatin states (Fig. 4A).

The majority of loci (clusters 1, 3, 4) carried histone modification signal associated with active regulatory elements (Creyghton et al. 2010; Heintzman et al. 2007; Robertson et al. 2008): mono- or tri-methylation of histone H3 lysine 4 residue (H3K4me1/3) and/or acetylation of histone H3 lysine 27 residue (H3K27ac) (Fig. 4A; Supplementary Fig. 10). The elements in cluster 3 and 4 were mainly intronic or gene-distal (>80%; Fig. 4B) and displayed H3K4me1 signal, associated with weak transcription (Core et al. 2014; Andersson and Sandelin 2020), in combination with H3K27ac. We observed a strong association with LTRs (Fig. 4B,C), in particular ERVs (Fig. 4B,D), in these two clusters, reinforcing the association statistics with STARR-seq and FANTOM5 enhancers. Although cluster 3 loci were associated with many TE families, ERVLs, ERVL-MaLRs and B4s were observed in high numbers (ERVL: 871, 33.9% of clustered ERVLs; ERVL-MaLR: 4,106, 37.4% of clustered ERVL-MaLRs; B4: 1,597, 35.5% of clustered B4s). ERV1s and ERVLs were enriched among cluster 4 TEs (Fig. 4D).

A considerable fraction of TE insertions (cluster 1) displayed strong H3K4me3 signal, associated with promoter activity of highly expressed transcription units, e.g. mRNA genes (Core et al. 2014; Andersson and Sandelin 2020). The majority of these fell close to GENCODE annotated TSSs and were enriched with SINE (Alu) and LTR (ERVK) elements (Fig. 4B-D). Hence, TEs are not only evolutionary co-opted into distal regulatory elements, but also gene promoters. Indeed, out of 7,431 transcribed TE-associated DHSs, 1,012 (13,6%) were associated with GENCODE annotated mRNA gene TSSs and 558 (7,5%) with GENCODE long non-coding RNA (lncRNA) gene TSSs. Complementary chromatin state segmentation analysis confirmed the strong bias of transcribed TEs having histone modifications indicative of active regulatory elements (Supplementary Fig. 11).

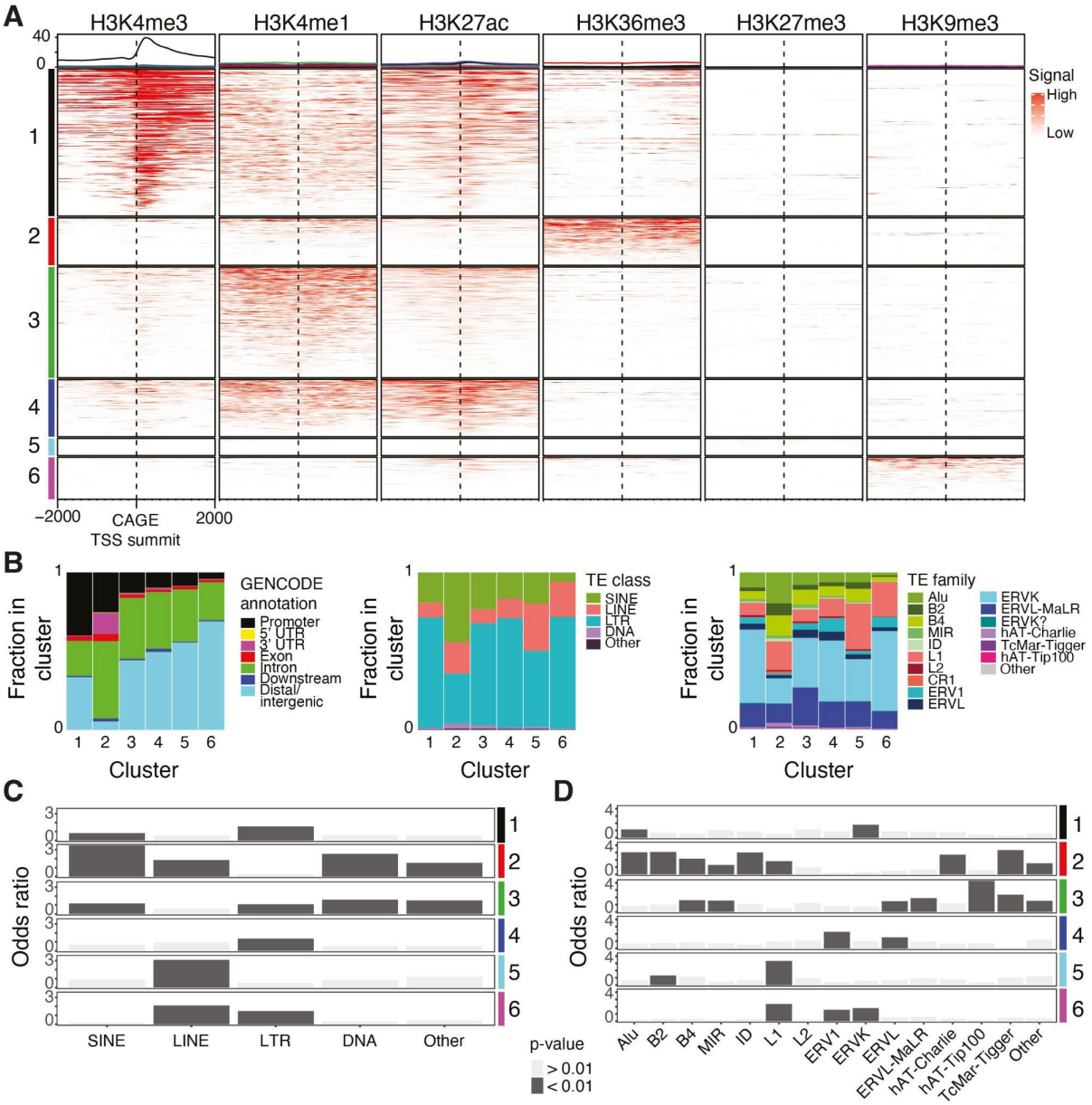


***Figure 4**: **Transcribed TEs exhibit chromatin features of regulatory elements. A:** Hierarchical clustering of histone modification (ChIP-seq) signals +/- 2,000 bp around the summits of TE-associated clusters of CAGE-inferred TSSs (CAGE tag clusters, Methods). The ChIP-seq signal is shown as fold-change over input control. Clusters are represented in rows (color coded in left legend) and histone modifications in columns. Average distributions of ChIP-seq signals for each cluster are shown (top panel). **B:** Annotations of TE-associated CAGE tag clusters based on GENCODE and RepeatMasker TE classes and families. **C,D:** Barplots of odds ratios of enrichments (Fisher's exact test) of TE classes (**C**) and TE families (**D**) in each cluster.*

Interestingly, ERV1s and ERVKs, in addition to L1 LINEs, were also found to be associated with both H3K27ac and tri-methylation of histone H3 lysine 9 residue (H3K9me3; cluster 6). The latter is present at repressed TEs in heterochromatin

(Rowe et al. 2013; Matsui et al. 2010; Maksakova et al. 2013; Karimi et al. 2011) and the bivalent combination of both repressing (H3K9me3) and activating (H3K27ac) histone modifications has been suggested to keep these TE-associated regulatory elements in a partially activated state (He et al. 2019). We also identified a group of transcribed TE loci (cluster 5) that was highly enriched with L1s and associated with low signal of tri-methylation of histone H3 lysine 27 residue (H3K27me3), indicative of polycomb-mediated repression. TE loci of cluster 5 and 6 thus likely represent bivalent, repressed regulatory elements with low transcriptional activity, which may facilitate later full activation, although we cannot rule out that this is an effect of mESC clonal heterogeneity.

Taken together, through several lines of evidence we demonstrate that dormant TEs, in particular endogenous retroviral elements, have frequently been repurposed into regulatory elements with enhancer and promoter activities in mESCs.

Endogenous retroviral subfamilies contribute to distinct regulatory programs

TF binding to regulatory elements is the key determinant of regulatory activity and the basis of cell-type specificity. To assess how TEs co-opted into regulatory elements contribute to transcriptional regulatory programs in mESCs, we performed a TF binding site enrichment analysis of DHS-associated transcribed TE insertions from 224 LTR subfamilies versus all CAGE-inferred TSS clusters. Several TFs, including pluripotency factors Nanog, Sox2 and Oct4, were enriched in transcribed LTRs compared to non-TE associated transcription initiation sites (Benjamini-Hochberg FDR-adjusted $p<1e-3$). These TFs were further found in a large number of LTR insertions (in 70.9%, 24.5% and 11,5% of all 19,436 transcribed DHS-associated LTRs versus in 40.3%, 9,6% and 5.2% of CAGE-inferred TSS clusters for Nanog, Sox2 and Oct4, respectively). These results confirm that LTRs are a major source of regulatory elements and binding sites for pluripotency factors.

Given their strong association with regulatory elements, we focused specifically on characterizing putative TF binding sites in DHSs of transcribed TEs across ERV subfamilies. The enrichment of TF binding site sequences indicates a diversity of regulatory potential across ERV subfamilies (Fig. 5A; Supplementary Fig. 12A; Supplementary Table 2). We observed a positive correlation between enrichments of

binding sites for pluripotency factors, e.g., Sox2 and Esrrb (Pearson's r = 0.54), Sox2 and Nanog (Pearson's r = 0.47; Fig. 5B; Supplementary Fig. 12B). The enrichments of Sox2, Nanog and Esrrb generally agreed with Isl1, Bcl6 and Oct4 (Fig 5B; Supplementary Fig. 12B), indicating that certain ERV subfamilies carry binding sites for a wide variety of pluripotency factors. Binding site sequences for all these TFs and the core pluripotency factor Oct4 were enriched in RLTR41, RLTR23, MMERVK9C_I_int, MLTR25A, and RLTR11A (mean $\log_2$ enrichment > 1).

We noted that SOX family related motifs, including that for Sox2, were enriched in specific ERV subfamilies (most notably in RLTR9D, RLTR9E, LTRIS2, RLTR13B1, RLTR41). Similarly, we identified a specific enrichment of KLF factors, including Klf4 (most highly enriched in RLTRETN_Mm, ORR1A1, ORR1A2, and RLTR10A). ERVL-MaLRs ORR1A1 and ORR1A2 were, in addition to Klf4, also enriched with Nanog, Isl1, Bcl6, and Lhx3, but not Esrrb and Oct4 binding sites, demonstrating that ERV subfamilies contribute differently and distinctly to the pluripotency regulatory program (Supplementary Fig. 12). Similarly, Lhx3 co-occurred with Esrrb in RLTR41 ERV1s, but was depleted from Esrrb-enriched ERVKs RLTR9D and RLTR13B1. Oct4 enriched ERVs also carried enrichments for Nanog and Sox2, but the top enriched ERVs for Oct4 (LTRIS2, RLTR17, RLTR16B_MM, RLTR41) were depleted from Klf4.

The specific enrichments or depletions of Klf4 or Oct4 in certain ERV subfamilies, compared to the overall agreement between Sox2, Esrrb and Nanog, indicates that the Klf4 and Oct4 ERV-derived regulatory networks have partially evolved independently from those of Sox2, Esrrb and Nanog. This is consistent with distinct binding preferences (Fig. 5C) and only context-dependent cooperativity between Sox2 and Oct4 (Li et al. 2019), which is limited during pluripotency maintenance (Malik et al. 2019, 2). We note that the motifs for Oct4 and Nanog could allow for the derivation of new binding sites for one TF from ancestral binding site sequences of the other through mutations (Fig. 5C). However, the dissimilarity of the Klf4 motif with those of Oct4 and Nanog indicates that evolutionary acquisition of Klf4 binding sites likely requires new transposition events.

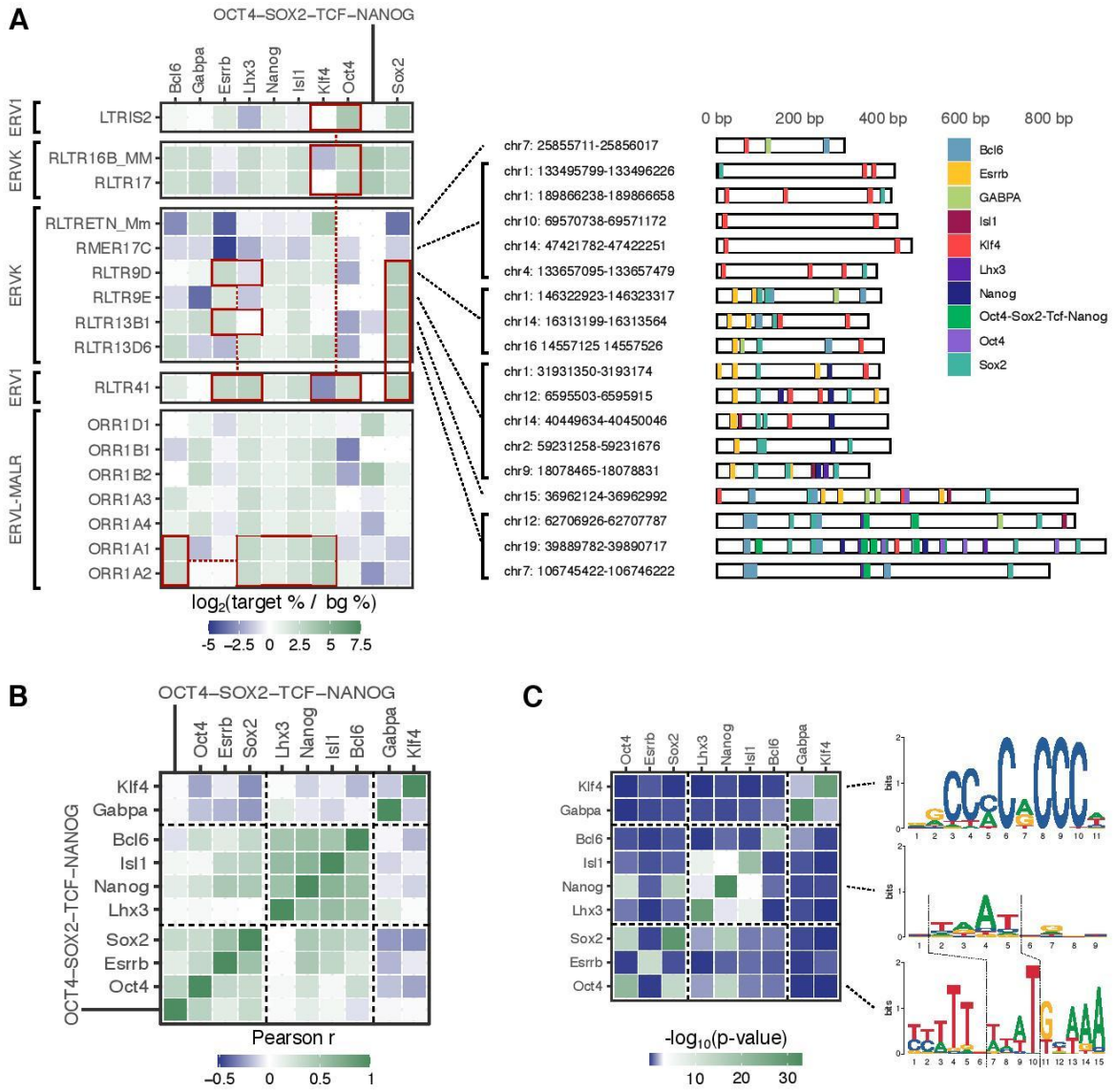


***Figure 5**: **ERV subfamilies contribute to distinct enrichments of binding sites for pluripotency factors. A:** Motif enrichments for selected TFs (columns) in transcribed TEs across selected ERV subfamilies (rows) versus a background of genomic regions +/- 200 bp around the summits of all CAGE-inferred TSS clusters. White cells indicate either no enrichment or non-reported values. The full TF enrichment heatmap is shown in Supplementary Figure 12A. Example genomic insertion sites for ERVK subfamilies are shown to illustrate their differences in carrying putative TF binding sites for Sox2, Esrrb and Oct4. **B:** Correlations (Pearson's r) between TF motif enrichments across all ERV subfamilies (as shown in Supplementary Fig. 12A). The full heatmap of correlations is given in Supplementary Figure 12B. **C:** Similarity (q-value) between binding motifs for selected TFs. Sequence logos (right) for Oct4, Nanog and Klf4 exemplify differences and similarities of TF binding sites.*

In addition to the enrichment of pluripotency TFs, we identified putative ERV-derived binding sites for a broad range of TFs (Supplementary Fig. 12A). For instance,

binding site sequences for ETS factors, involved in a large variety of gene regulatory programs and across cell types, were highly enriched among ERVL-MaLR subfamilies ORR1E and ORR1D2, which were further depleted of TF binding sites for Oct4 and Sox2. This demonstrates that transposition of TF binding sites native to LTR retrotransposons or the subsequent birth of new TF binding sites through mutations (Sundaram et al. 2017) is a general mechanism that could impact several regulatory programs, including those distinct from the naive pluripotency.

To further investigate the regulatory programs contributed to by ERVs co-opted into regulatory elements, we linked each gene-distal, DHS-associated transcribed ERV with its predicted target gene from activity-by-contact (ABC) modelling (Fulco et al. 2019). The resulting pathway and gene ontology enrichments of linked genes indicate a highly specific association of individual ERV subfamilies with gene regulatory programs (Fig. 6A; Supplementary Fig. 13), including regulation of genes involved in immunity by RLTR13B1 elements and regulation of genes involved in transcription, specifically basal TFs in the TFIID complex, by MLTR14 elements (Fig. 6B).

Taken together, our results demonstrate that the pluripotency regulatory network is shaped by ERV-derived regulatory elements, together providing binding sites to pluripotency TFs, including those for Oct4, Nanog, Sox2, Klf4 and Esrrb. However, although putative TF binding sites for multiple TFs were identified for each ERV subfamily, we note that the TF enrichments across ERV subfamilies are highly distinct (Fig. 5; Supplementary Fig. 12). Together with the diversity of TFs identified (Supplementary Fig. 12A) and the diversity of functions of target genes (Supplementary Fig. 13), this demonstrates that each ERV subfamily contributes to distinct regulatory programs and pathways.

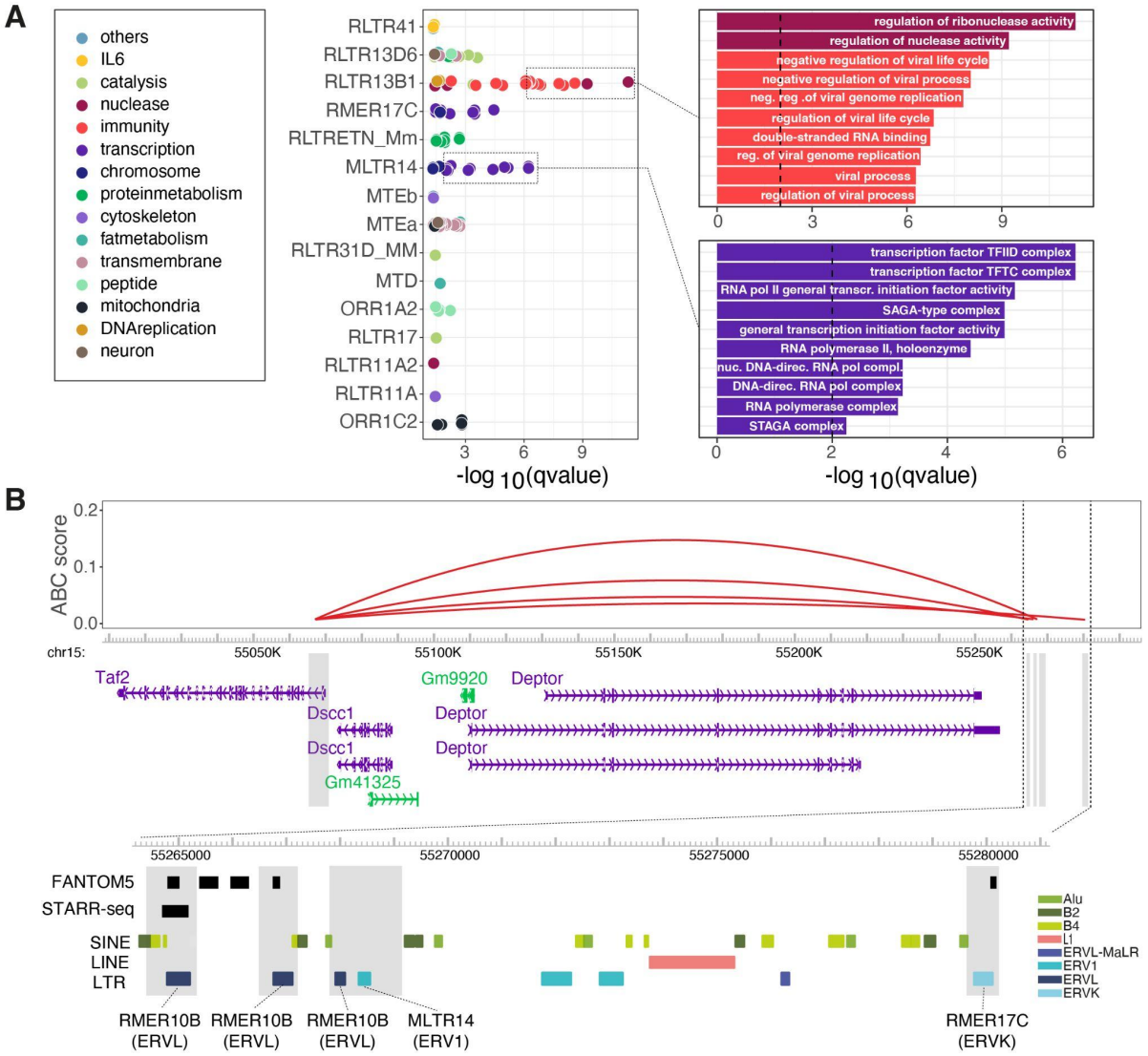


Figure 6: ERV subfamilies contribute to distinct gene regulatory programs. A: *GO term enrichment for putative target genes (ABC) of gene distal ERVs split by ERV subfamily (foreground) versus all ABC-predicted target genes (background). For ease of visualization, gene ontology terms are colored by manually curated process or function, and the underlying gene ontology term enrichments are shown for RLTR13B1 and MLTR14. Full results are provided in Supplementary Figure 13.* ***B:*** *Predicted enhancer interactions with the promoter of gene Taf2 (TATA-Box Binding Protein Associated Factor 2). The four enhancers are marked by grey boxes and a zoom-in is provided below, showing overlaps with ERV insertions of RMER10B, MLTR14 and RMER17C. Tracks for GENCODE (M19) transcripts, FANTOM5 enhancers, STARR-seq enhancers, and TEs provided by RepeatMasker are shown.*

Discussion

Transcriptional regulation of ESCs is multi-faceted. On one hand, regulatory elements act as binding platforms for transcription factors controlling the transcriptional activities of genes involved in activities not necessarily specific to ESCs, such as metabolism, transcription, stress response and replication. In parallel, ESCs must maintain plasticity to ensure potent differentiation capabilities. At the core of such activities are highly specialized and conserved transcription factors, including Oct4, Sox2 and Nanog, which, through targeted transcriptional regulation, maintain a naive pluripotent state. Their regulation of genes is highly conserved across Metazoa, although their binding sites are not, which is linked to a high evolutionary turnover of regulatory elements.

We here demonstrate that retrotransposons contribute to a sizable fraction of regulatory innovation. Through systematic characterization of the transcription initiation landscape of mESCs, we show that a fraction of TEs are transcribed and that these display balanced divergent transcription initiation patterns within sites of open chromatin. We provide, to the best of our knowledge, the first evidence that retrotransposon-derived RNAs are targeted by the nuclear exosome for degradation, and that the decay and initiation patterns are shared across cell types and species. As many as a third of divergently transcribed TEs are supported by *in vitro* enhancer potential derived from STARR-seq data (Peng et al. 2020) or by overlap with, rigorously validated, previously predicted enhancer sets from FANTOM (Andersson et al. 2014a; Arner et al. 2015; Dalby et al. 2018). In addition, we demonstrate a large overlap with annotated gene promoters (Frankish et al. 2019).

Endogenous retroviral elements are most frequently transcribed, and ERVKs stand out as the largest contributor of regulatory elements with enhancer potential in mESCs. This bias is likely explained by their general enrichment of binding site sequences for TFs regulating naive pluripotency, including Oct4, Sox2 and Nanog. However, we cannot rule out that differences in transcription frequencies across TE families are partly due to the cumbersome task of mapping short sequencing reads to repetitive regions. Approaches to meet this challenge include estimation of the relative abundances of transposable element expression on a transposable element

subfamily level from multi-mapping reads through expectation-maximization (Jin et al. 2015) and probabilistic rescue of multi-mapping reads guided by uniquely mapped reads (Faulkner et al. 2008; Hashimoto et al. 2009). We here show that the two approaches yield comparable expression estimates across TE subfamilies, which in our opinion favors the latter approach due to not focusing exclusively on annotated regions. However, we do acknowledge that uniquely mapping reads might skew focus to evolutionary older TEs (Simonti et al. 2017), which have accumulated more mutations than younger ones and therefore have a higher chance of mapping reads uniquely (Jin et al. 2015). Long sequencing reads, less affected by repeats in their mapping, will be instrumental for future investigations but pose other challenges, for instance lower library complexity. However, any mapping bias cannot explain differential TF enrichments across ERV subfamilies.

The diverse TF enrichments observed across ERV subfamilies indicate a major contribution of ERVs to the landscape of regulatory elements involved in a wide variety of gene regulatory programs. Interestingly, we observed varying degrees of over-representation of putative TF binding sites in ERV subfamilies, including those enriched with the wide repertoire of pluripotency TFs, those partially depleted of specific pluripotency TFs, and those enriched with non cell-type specific TFs, like the ETS superfamily. We noted a general strong co-occurrence of binding sites for Nanog, Sox2, and Esrrb, while Oct4 and Klf4 were depleted in some subfamilies. These enrichments argue for unique contributions by ERV subfamilies to the regulatory programs of mESCs, which is confirmed by the specific functions of putative target genes by gene-distal candidate enhancer co-opted from ERVs. Further, our data suggest that the regulatory landscape of naive pluripotency in mESCs has been shaped independently by multiple subfamilies of ERVs.

In addition to the core pluripotency TFs, we observed an enrichment of binding sites for multiple TFs in several ERV subfamilies. Since ERVs originate from ancestral genomic insertions of retroviral elements, the ancestral ERV sequences must therefore have either carried the full repertoire of TF binding sites that have been maintained through evolution or have served as substrates for the evolutionary acquisition of diverse TF binding sites through mutations (Sundaram et al. 2017; Sun et al. 2018). That way, ERVs may offer enough sequence diversity to allow encoding

for enhancer activity even in the absence of binding sites for specific master regulators of mESCs (Singh et al. 2021).

Taken together, we present a systematic characterization of the transcriptional activities, RNA decay patterns, chromatin signatures, and regulatory potential of TEs in mESCs. Our results demonstrate a sizable contribution of TEs to regulatory innovation and the regulatory landscape of naive pluripotency. Charting their cell-type specific activities and dynamics over development will be important next steps to further our understanding of the roles of TEs in transcriptional regulation.

Methods

E14 mESC and HeLa S2 CAGE libraries, data processing and mapping

Previously sequenced E14 mESCs (GEO ID GSE115710) (Lloret-Llinares et al. 2018) and human HeLa S2 cells (GEO ID GSE62047) (Andersson et al. 2014b) CAGE (Takahashi et al. 2012) libraries were collected. As described in each report, CAGE libraries were prepared from exosome depleted samples as well as from control samples. E14 mESCs were transduced with pLKO vectors encoding the shRNA: SHC002 (scrambled control - referred to as Scr control) and NM_025513.1-909s1c1 (referred to as Rrp40 exosome knockdown). HeLa cells were transfected with enhanced green fluorescent protein (eGFP - referred to as EGFP control) and the hRRP40 (*EXOSC3*) siRNA (referred to as RRP40 exosome knockdown). The CAGE libraries from mouse and human were processed as in the original publications, with some minor modifications. Reads were trimmed using the FASTX-Toolkit (Version 0.013 - http://hannonlab.cshl.edu/fastx_toolkit/) to remove linker sequences (Illumina adaptors) and then filtered for a minimum sequencing quality of 30 in 50% of the bases. Mapping to the mouse reference genome (mm10) was performed using Bowtie (Version 1.1.2), applying the following parameters to ensure several (up to 100) good alignments per read, which is essential for the rescue and analysis of TE-derived sequences: *-k 100* (report up to 100 good alignments per read), *-m 100* (eliminate reads that map > 100 times), *--best* and *--strata* (report alignments which have the highest quality). Reads that mapped to unplaced chromosome patches or chrM were discarded. Finally, all reads

corresponding to reference rRNA sequences (mouse: BK000964.3, human: U13369.1) with up to 2 mismatches were discarded using rRNAdust (fantom.gsc.riken.jp/5/suppl/rRNAdust/). Mapping to the human reference genome (hg19) was performed using the exact same parameters and version of Bowtie.
For exploratory/comparative purposes, the mapping of HeLa S2 CAGE libraries was performed with 3 additional alignment approaches (Supplementary Fig. 8 B,C). BWA (Burrows-Wheeler Aligner) version 0.7.15-r1140 (Li and Durbin 2009) was used with the parameter -n 2 (maximum distance for each alignment) and a subsequent mapping quality threshold (MAPQ >20), using SAMtools v1.3.1 (Li et al. 2009). BWA-PSSM, a modification of BWA using position specific scoring matrices (PSSM) (Kerpedjiev et al. 2014), was used with parameters -n 2 , -m 2000 (to allow for more suboptimal partial alignments to be tested) and a downstream MAPQ>20 threshold. Finally, LAST version 801 (Kiełbasa et al. 2011) was used with parameters -u NEAR -R 01 (*lastdb*) and -Q 1 and -D 100 (*lastal*), followed by a downstream MAPQ>20 threshold.

Probabilistic multi-mapping rescue approach for CAGE tags

Following initial mapping of reads with Bowtie, we employed *MUMRescueLite* (Hashimoto et al. 2009) to resolve short multi-mapping CAGE reads that aligned equally well to more than one genomic location. In short, this method examines the information about the local context of potential mapping positions given by uniquely mapping reads. By assuming that multi-mapping reads are more likely to come from regions which already have more uniquely mapping reads, MUMRescueLite probabilistically assigns the true source of a multimapping read. The probabilistic scheme behind the tool is weighted by the abundance and location of uniquely mapping reads. Thus, a nominal window parameter is required for which to identify unique mappers that occur around (upstream and downstream) each locus occupied by a multi mapper. In addition, a weight threshold is required over which one locus of a multi mapper is “rescued”, referring to the fraction out of the total number of unique mappers proximal to all loci associated with a specific multi mapper. The window parameter was cautiously selected to be 50 bp after a saturation investigation of how many reads are rescued. The weight threshold was set to 0.7, in order to select one locus as the true source of a multi-mapping read.

TElocal-inferred expression of mESC TE families

For statistical comparison purposes, CAGE unique and multi-mapping reads aligned with Bowtie, as described above, were supplied to TElocal v 0.1.0 from the TEToolkit suite (Jin et al. 2015) in order to quantify transposable element expression at the locus level. The resulting quantification for each transposon for both TElocal and multi mapping rescuing output was normalized to tags per million mapped reads (TPM) and the overlapping instances between the two techniques were compared using Spearman's rank-based correlations.

CAGE tag clustering, quantification and normalization

Following multi-mapping rescue, the number of overall CAGE tag 5′ends were counted for each genomic position to obtain base-pair (bp) resolution of CAGE TSSs (CTSSs). We then assigned CTSSs to transposable elements (TEs) as defined by RepeatMasker v4.0.7 (http://www.repeatmasker.org/) and considered only instances with two or more CAGE tags. Tag clusters (TCs) were generated from pooled CAGE libraries per condition, and wide TCs were narrowed or decomposed into sub-peak TCs if containing multiple peaks (https://github.com/anderssonlab/CAGEfightR_extensions), as previously described (Rennie et al. 2018). In short, CTSSs located within 20 bp from each other on the same strand were merged into initial TCs. For each TC, the bp with the most abundant count (summit position) or the median of multiple equal summits were identified. Next, the fraction of total CTSSs of each location within a TC to that of the summit position was calculated. If a position carried less than 10% of the summit signal, it was discarded and TCs split were merged if positioned within 20bp from each other on the same strand. Expression quantification of each individual TC in each CAGE replicate took place by adding up the CTSSs falling into them. Using the CAGE genomic background noise estimation (as described below), all TCs with expression values below the noise threshold were discarded from further analyses. Expression levels of TCs were normalized to tags per million mapped reads (TPM). Finally, we assigned normalized TCs to TEs on the same strand through direct overlap using BEDtools (Quinlan and Hall 2010).

Footprints of CAGE expression on TEs

To investigate the relationship between transposable element families and/or classes

to CAGE TSS locations, we plotted the average binarized (presence of absence of CTSS, regardless of expression value) pooled CTSS signal 500-bp upstream, across the body and 500-bp downstream of each TE instance using deepTools (Ramírez et al. 2016). Unique TE instance profiles were averaged for each TE family or class based on their Repeatmasker annotation. Furthermore, we constructed a synthetic CAGE uniqueness track by mapping the mm10 reference genome split in 25 bp long segments back to itself. Localization of the synthetic CAGE tags on a bp resolution was conducted as described above at the CTSS level, assigned to TE instances and the signal was binarized, representing the expected uniquely mapped background signal.

HeLa RNA-seq data processing

RNA-seq data from HeLa cells depleted of hRRP40 using siRNA-mediated knockdown as described elsewhere (Andersen et al. 2013) (SRA accession: SRX365673) were considered. Briefly, after filtering of low quality reads, removal of Illumina adaptors and reads shorter than 25 bp with Trimmomatic v0.36 (Bolger et al. 2014), reads were mapped against the human reference genome (hg19) using HISAT2 v2.1.0 (Kim et al. 2015). Uniquely mapped and properly paired reads were selected with SAMtools v1.3.1. Gene-level expression quantification of mapped reads was performed with featureCounts v1.6.3 (Liao et al. 2014). Further analyses and comparison to gene-level expression with CAGE using generalized linear Poisson regression models with backward elimination for variable selection was performed in R (http://www.R-project.org/) using the lm() function.

Processing of DNase-seq data and DHSs as focus points for transcription initiation

For accurate identification of TE-associated regulatory elements, sequencing reads from DNase-seq for the mouse ES-E14 cell line (GEO ID GSE37073 / GSM1014154) were processed using the ENCODE DNase-HS pipeline. Called DHS hotspot FDR 1% peaks in the mouse reference genome (mm10) were used as DNase I hypersensitive sites (DHSs). DHSs were used as focus points of minus and plus strand expression by defining DHS midpoints as positions optimising the coverage of proximal CAGE tags within flanking windows of size +/- 300 bps around them, as previously described (Rennie et al. 2018). The final set of 165,052

transcribed DHS was determined by filtering DHSs to not overlap any other DHS +/- the 300bp window on the same strand and to be supported by either control or exosome knockdown CAGE expression above the noise threshold (described below). Transcriptional directionality and exosome sensitivity scores were calculated considering this set of DHS regions, as defined previously (Andersson et al. 2014b). In short, the directionality score measures the expression level strand bias within transcribed DHSs, it ranges from -1 to 1 (100% minus or plus strand expression), while 0 indicates balanced bidirectional transcription. The exosome sensitivity score measures the relative amount of degraded RNAs by the exosome by quantifying the fraction of exosome-depleted CAGE expression seen only after exosome depletion: exosome sensitivity values closer to 1 are indicative of highly unstable RNAs.

Estimation of CAGE genomic background noise

The estimation of a CAGE genomic background noise threshold (https://github.com/anderssonlab/CAGEfightR_extensions) for robust assessment of lowly expressed regions was based on quantifying the CAGE 5'ends in randomly selected uniquely mappable regions of 200 bp distal to known TSSs, exons and DHSs, followed by extracting the 99th percentile of the empirical distribution of CAGE expression and using the max value across control libraries as a noise threshold for significant expression in further analyses, described in detail elsewhere (Rennie et al. 2018).

CAGE gene-level expression quantification

For statistical comparison of gene-level RNA-seq expression to gene-level expression as measured by CAGE, we quantified abundances of genes using CTSSs within +/- 500 bp windows from the 5'end of GENCODE primary mouse sequence and annotation (GRCm38) version M10 transcripts (Frankish et al. 2019), as CAGE signal from HeliScopeCAGE saturates after ~ 500bp from annotated gene TSSs (Kawaji et al. 2014). Gene-level abundances were quantified by first merging potentially overlapping TSS-centered windows per transcript belonging to the same gene and then summing the expression levels of all transcript windows for each gene.

Genomic annotation of transcribed TEs

We annotated TE-associated TCs based on different genomic regions as defined using GENCODE version M10 and BEDtools, ensuring there are no overlapping regions counted twice. Coordinates for all genic regions (exons, 5' UTRs, 3'UTRs) were extracted from the GENCODE annotation. Promoter regions were defined as regions at the starting positions of each transcript +/- 500 bp. To define intronic regions, we subtracted the exonic regions from the genic regions. Finally, distal/intergenic regions were defined as the remaining parts of the genome in-between annotated genes.

FANTOM enhancers and enhancer peak calling from STARR-seq data

Transcribed enhancers identified by remapped FANTOM5 CAGE libraries to mm10 (GRCh38), deposited in Zenodo (Dalby et al. 2018), were associated with TE-associated DHSs by overlap using BEDTools. STARR-seq data for 2iL grown mESCs E14Tg2a (E14) (Peng et al. 2020) (GEO ID GSE143546) were used to evaluate the enhancer potential of transcribed TEs. STARR-seq data were processed using Bowtie2 (Langmead and Salzberg 2012, 2) and SAMtools (Li et al. 2009). Reads were aligned to the mouse reference genome (mm10) using Bowtie2 (*--very sensitive*). The reads of the two replicates from each sample were sorted and merged and reads falling into regions from the ENCODE blacklist of the mouse reference genome were removed. STARRPeaker (Lee et al. 2020) was used to identify potential enhancers with default parameters and an adjusted p-value threshold of 0.05. Potential enhancers called from STARR-seq data were associated with expressed TE-associated DHSs by overlap using BEDTools.

Processing and analyses of histone modification ChIP-seq data

Mouse ENCODE E14Tg2a or E14 mESCs ChIP-seq data for six histone modifications: H3K4me3, H3K4me1, H3K27ac, H3K36me3, H3K9me3 (GEO ID GSE136479) and H3K27ac (GEO ID GSE31039) were processed using the ENCODE ChIP-seq processing pipeline (version 1.3.6). Adapters and low quality reads were filtered with cutadapt v2.5 (Martin 2011), reads were mapped to the mouse genome assembly mm10 with bwa, duplicate reads were removed with Picard v2.20.7 (http://broadinstitute.github.io/picard/), ENCODE mm10 blacklist regions were masked and only reads with mapping quality above 30 were considered for further downstream analyses. For the heatmap and footprint plots,

ChIP signal expressed as fold-over input control was averaged across sites in 10 bp bin intervals from the CAGE TC summit position up to a maximum of +/- 2000bp, using deepTools version 3.1.3. Hierarchical clustering, annotation and visualization were conducted in R with ChIPSeeker (Yu et al. 2015), *ggplot2* (https://ggplot2.tidyverse.org) and *profileplyr* (R package version 1.6.0), using clustering parameters *"rowMax"* for summarizing the ranges across samples and *"median"* for defining proximity between clusters.

Chromatin state discovery and characterization using chromHMM

To annotate TE-associated DHSs and characterize which cis-elements in mESCs (mm10) they are occupying, we constructed a 12-state model chromatin states map to identify genomic regions enriched in specific combinations of histone modifications and TF marks, as previously described (Pintacuda et al. 2017). The multivariate hidden Markov model framework of chromHMM (Ernst and Kellis 2012, 2017) was applied to the mouse reference genome (mm10) and ENCODE ChIP-seq data in E14 cells for the following ten marks: H3K4me1, H3K4me3, H3K27me3, H3K27ac, H3K9me3, H3K9ac, H3K36me3, CTCF, Nanog, Oct4. TE-associated DHSs and non-TE DHSs associated with TCs from control and exosome-depleted CAGE libraries were overlapped with the chromHMM states with BEDtools and the heatmaps were generated in R using heatmap.2 from the *gplots* package (R package version 3.1.1.).

Transcription factor motif analysis

We scanned full ERV transposons carrying at least one CAGE TC with known TF motifs using findMotifsGenome.pl from HOMER (Heinz et al. 2010). For each TF, findMotifsGenome.pl employs a hypergeometric test to compare the number of motifs found in the target set with that found in a specified background set. The tool was run using the set of expressed ERV transposons per subfamily as the target set and the full set of non-TE-associated CAGE TCs (+/- 200bp regions) as the background set. The significantly enriched TFBS motifs were selected with three additional conditions: (1) TF genes are expressed in mESCs using gene-level CAGE quantification; (2) at least 10% of the target set contained the motif; and (3) known TF genes have a match score > 0.9 to the *de novo* motifs found by HOMER. The motif enrichment score was calculated as $\log_2$ (% of target sequences with motif / %

of background sequences with motif). Heatmaps of TF motif enrichment were generated in R using the *ggplot2* package. Comparisons between binding sites of selected mouse TFs were performed using TomTom of the MEME suite including the three mouse-specific motif databases: HOCOMOCO Mouse (v11 FULL), UniPROBE Mouse (Sci09 Cell08), Embryonic Stem Cell TFs (Chen2008) both as query and target motifs.

Functional Enrichment Analysis

Gene ontology (GO) and pathway enrichment analyses, using TE-regulated ABC (activity-by-contact) predicted target genes (retrieved from https://osf.io/uhnb4/) for enhancer elements (Fulco et al. 2019) as target sets and all ABC predicted target genes as background, were performed in R using the package *clusterProfileR* (Yu et al. 2012). Only statistically significant results (p.adjust < 0.05) were considered after employing false discovery rate for multiple testing correction. ABC region coordinates were lifted from mm9 to mm10 using the UCSC liftOver tool (Rhead et al. 2010) and were associated with CAGE TCs by coordinate overlap (BedTools).

Acknowledgements

This project has received funding from the Danish Council for Independent Research (Grant 6108-00038) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant 638173). We would like to thank members of the Andersson lab for their rewarding comments and discussions.

Author contributions

R.A. and S.B conceived the project; S.B performed most analyses, with support from R.K., N.A. and M.S.; S.B., R.K. and R.A interpreted results and wrote the manuscript; all authors reviewed the final manuscript.

References

Andersen PR, Domanski M, Kristiansen MS, Storvall H, Ntini E, Verheggen C, Schein A, Bunkenborg J, Poser I, Hallais M, et al. 2013. The human cap-binding complex is functionally connected to the nuclear RNA exosome. *Nat Struct Mol Biol* **20**: 1367–1376.

Andersson R, Chen Y, Core L, Lis JT, Sandelin A, Jensen TH. 2015a. Human Gene Promoters Are Intrinsically Bidirectional. *Mol Cell* **60**: 346–347.

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014a. An atlas of active enhancers across human cell types and tissues. *Nature* **507**: 455–461.

Andersson R, Refsing Andersen P, Valen E, Core LJ, Bornholdt J, Boyd M, Heick Jensen T, Sandelin A. 2014b. Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nat Comms* **5**: 5336.

Andersson R, Sandelin A. 2020. Determinants of enhancer and promoter activities of regulatory elements. *Nature Reviews Genetics* **21**: 71–87.

Andersson R, Sandelin A, Danko CG. 2015b. A unified architecture of transcriptional regulatory elements. *Trends in Genetics* **31**: 426–433.

Arner E, Daub CO, Vitting-Seerup K, Andersson R, Lilje B, Drablos F, Lennartsson A, Rönnerblad M, Hrydziuszko O, Vitezic M, et al. 2015. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* **347**: 1010–1014.

Beagrie RA, Pombo A. 2016. Gene activation by metazoan enhancers: Diverse mechanisms stimulate distinct steps of transcription. *BioEssays* **38**: 881–893.

Berthelot C, Villar D, Horvath JE, Odom DT, Flicek P. 2018. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nature Ecology & Evolution 2017 2:1* **2**: 152–163.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew J-L, Ruan Y, Wei C-L, Ng HH, et al. 2008. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res* **18**: 1752–1762.

Cao Y, Chen G, Wu G, Zhang X, McDermott J, Chen X, Xu C, Jiang Q, Chen Z, Zeng Y, et al. 2019. Widespread roles of enhancer-like transposable elements in cell identity and long-range genomic interactions. *Genome Res* **29**: 40–52.

Chan ET, Quon GT, Chua G, Babak T, Trochesset M, Zirngibl RA, Aubin J, Ratcliffe MJ, Wilde A, Brudno M, et al. 2009. Conservation of core gene expression in vertebrate tissues. *Journal of Biology* **8**: 33.

Chen Y, Pai AA, Herudek J, Lubas M, Meola N, Järvelin AI, Andersson R, Pelechano V, Steinmetz LM, Jensen TH, et al. 2016. Principles for RNA metabolism and alternative transcription initiation within closely spaced promoters. *Nat Genet* **48**: 984–994.

Chuong EB, Rumi MAK, Soares MJ, Baker JC. 2013. Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nature Genetics* **45**: 325–329.

Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. 2014. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet* **46**: 1311–1320.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci USA* **107**: 21931–21936.

Dalby M, Rennie S, Andersson R. 2018. FANTOM5 transcribed enhancers in mm10. https://zenodo.org/record/1411211#.YKInG2YzbDE (Accessed May 17, 2021).

Ernst J, Kellis M. 2017. Chromatin-state discovery and genome annotation with ChromHMM.

Nature Protocols **12**: 2478–2492.
Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**: 215–216.
Faulkner GJ, Forrest ARR, Chalk AM, Schroder K, Hayashizaki Y, Carninci P, Hume DA, Grimmond SM. 2008. A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics* **91**: 281–288.
Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, et al. 2009. The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* **41**: 563–571.
Fort A, Hashimoto K, Yamada D, Salimullah M, Keya CA, Saxena A, Bonetti A, Voineagu I, Bertin N, Kratz A, et al. 2014. Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nat Genet* **46**: 558–566.
Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al. 2019. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**: D766–D773.
Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, Grossman SR, Anyoha R, Doughty BR, Patwardhan TA, et al. 2019. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature Genetics* **51**: 1664–1669.
Grossman SR, Zhang X, Wang L, Engreitz J, Melnikov A, Rogov P, Tewhey R, Isakova A, Deplancke B, Bernstein BE, et al. 2017. Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proc Natl Acad Sci USA* 201621150.
Haberle V, Stark A. 2018. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol* **19**: 621–637.
Halfon MS. 2019. Studying Transcriptional Enhancers: The Founder Fallacy, Validation Creep, and Other Biases. *Trends in Genetics* **35**: 93–103.
Hashimoto T, de Hoon MJL, Grimmond SM, Daub CO, Hayashizaki Y, Faulkner GJ. 2009. Probabilistic resolution of multi-mapping reads in massively parallel sequencing data using MuMRescueLite. *Bioinformatics* **25**: 2613–2614.
He J, Fu X, Zhang M, He F, Li W, Abdul MM, Zhou J, Sun L, Chang C, Li Y, et al. 2019. Transposable elements are regulated by context-specific patterns of chromatin marks in mouse embryonic stem cells. *Nature Communications* **10**: 34.
Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**: 311–318.
Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell* **38**: 576–589.
Henriques T, Scruggs BS, Inouye MO, Muse GW, Williams LH, Burkholder AB, Lavender CA, Fargo DC, Adelman K. 2018. Widespread transcriptional pausing and elongation control at enhancers. *Genes Dev*. http://genesdev.cshlp.org/lookup/doi/10.1101/gad.309351.117.
Jin Y, Tam OH, Paniagua E, Hammell M. 2015. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31**: 3593–3599.
Karimi MM, Goyal P, Maksakova IA, Bilenky M, Leung D, Tang JX, Shinkai Y, Mager DL, Jones S, Hirst M, et al. 2011. DNA Methylation and SETDB1/H3K9me3 Regulate Predominantly Distinct Sets of Genes, Retroelements, and Chimeric Transcripts in mESCs. *Cell Stem Cell* **8**: 676–687.
Kawaji H, Lizio M, Itoh M, Kanamori-Katayama M, Kaiho A, Nishiyori-Sueki H, Shin JW, Kojima-Ishiyama M, Kawano M, Murata M, et al. 2014. Comparison of CAGE and

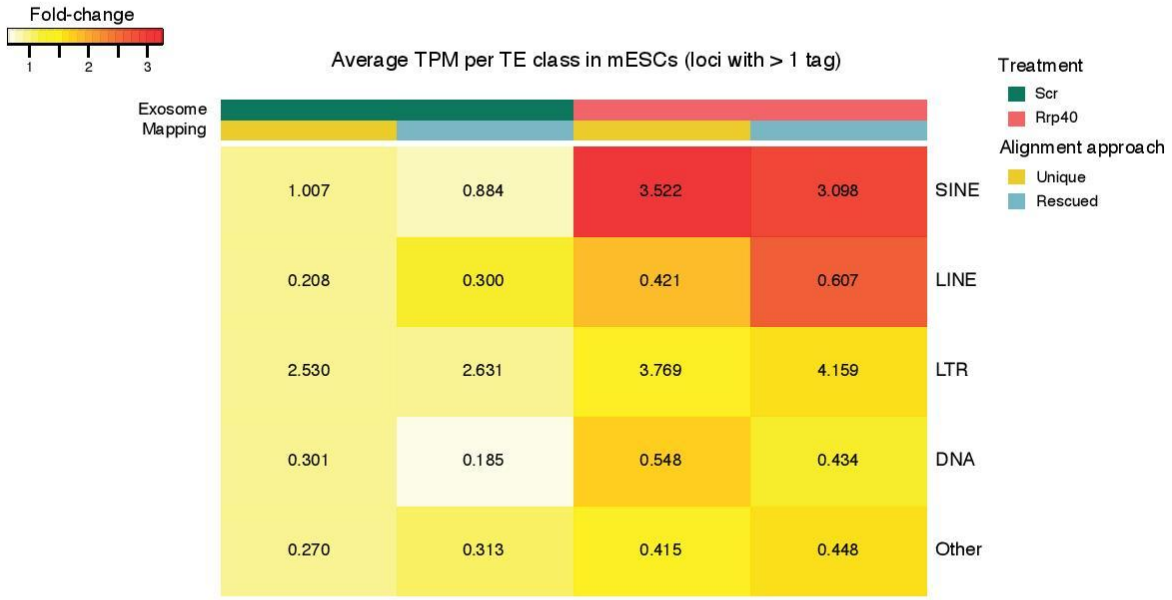
RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. *Genome Res* **24**: 708–717.
Kerpedjiev P, Frellsen J, Lindgreen S, Krogh A. 2014. Adaptable probabilistic mapping of short reads using position specific scoring matrices. *BMC Bioinformatics* **15**: 100.
Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. 2011. Adaptive seeds tame genomic sequence comparison. *Genome Res* **21**: 487–493.
Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**: 357–360.
Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**: 182–187.
Koch F, Fenouil R, Gut M, Cauchy P, Albert TK, Zacarias-Cabeza J, Spicuglia S, de la Chapelle AL, Heidemann M, Hintermair C, et al. 2011. Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nat Struct Mol Biol* **18**: 956–963.
Kunarso G, Chia N-Y, Jeyakani J, Hwang C, Lu X, Chan Y-S, Ng HH, Bourque G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* **42**: 631–634.
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357–359.
Lee D, Shi M, Moran J, Wall M, Zhang J, Liu J, Fitzgerald D, Kyono Y, Ma L, White KP, et al. 2020. STARRPeaker: uniform processing and accurate identification of STARR-seq active regions. *Genome Biology* **21**: 298.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**: 1754–1760.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.
Li S, Zheng EB, Zhao L, Liu S. 2019. Nonreciprocal and Conditional Cooperativity Directs the Pioneer Activity of Pluripotency Transcription Factors. *Cell Reports* **28**: 2689-2703.e4.
Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**: 923–930.
Lloret-Llinares M, Karadoulama E, Chen Y, Wojenski LA, Villafano GJ, Bornholdt J, Andersson R, Core L, Sandelin A, Jensen TH. 2018. The RNA exosome contributes to gene expression regulation during stem cell differentiation. *Nucleic Acids Res* **4**: a008128.
Maksakova IA, Thompson PJ, Goyal P, Jones SJ, Singh PB, Karimi MM, Lorincz MC. 2013. Distinct roles of KAP1, HP1 and G9a/GLP in silencing of the two-cell-specific retrotransposon MERVL in mouse ES cells. *Epigenetics & Chromatin* **6**: 15.
Malik V, Glaser LV, Zimmer D, Velychko S, Weng M, Holzner M, Arend M, Chen Y, Srivastava Y, Veerapandian V, et al. 2019. Pluripotency reprogramming by competent and incompetent POU factors uncovers temporal dependency for Oct4 and Sox2. *Nature Communications* **10**: 3477.
Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet j* **17**: 10.
Matsui T, Leung D, Miyashita H, Maksakova IA, Miyachi H, Kimura H, Tachibana M, Lorincz MC, Shinkai Y. 2010. Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. *Nature* **464**: 927–931.
Miao B, Fu S, Lyu C, Gontarz P, Wang T, Zhang B. 2020. Tissue-specific usage of transposable element-derived promoters in mouse development. *Genome Biol* **21**: 255.
Mikhaylichenko O, Bondarenko V, Harnett D, Schor IE, Males M, Viales RR, Furlong EEM. 2018. The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. *Genes Dev*.

http://genesdev.cshlp.org/lookup/doi/10.1101/gad.308619.117.
Nguyen TA, Jones RD, Snavely AR, Pfenning AR, Kirchner R, Hemberg M, Gray JM. 2016. High-throughput functional comparison of promoter and enhancer activities. *Genome Res* **26**: 1023–1033.
Nitta KR, Jolma A, Yin Y, Morgunova E, Kivioja T, Akhtar J, Hens K, Toivonen J, Deplancke B, Furlong EEM, et al. 2015. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution ed. B. Ren. *eLife* **4**: e04837.
Ntini E, Järvelin AI, Bornholdt J, Chen Y, Boyd M, Jørgensen M, Andersson R, Hoof I, Schein A, Andersen PR, et al. 2013. Polyadenylation site–induced decay of upstream transcripts enforces promoter directionality. *Nat Struct Mol Biol* **20**: 923–928.
Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E. 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* **39**: 730–732.
Peng T, Zhai Y, Atlasi Y, ter Huurne M, Marks H, Stunnenberg HG, Megchelenbrink W. 2020. STARR-seq identifies active, chromatin-masked, and dormant enhancers in pluripotent mouse embryonic stem cells. *Genome Biology* **21**: 243.
Pintacuda G, Wei G, Roustan C, Kirmizitas BA, Solcan N, Cerase A, Castello A, Mohammed S, Moindrot B, Nesterova TB, et al. 2017. hnRNPK Recruits PCGF3/5-PRC1 to the Xist RNA B-Repeat to Establish Polycomb-Mediated Chromosomal Silencing. *Molecular Cell* **68**: 955-969.e10.
Preker P, Nielsen J, Kammler S, Lykke-Andersen S, Christensen MS, Mapendano CK, Schierup MH, Jensen TH. 2008. RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters. *Science* **322**: 1851–1854.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.
Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research* **44**: W160–W165.
Rennie S, Dalby M, Lloret-Llinares M, Bakoulis S, Dalager Vaagensø C, Heick Jensen T, Andersson R. 2018. Transcription start site analysis reveals widespread divergent transcription in D. melanogaster and core promoter-encoded enhancer activities. *Nucleic Acids Res* **39**: 311.
Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, et al. 2010. The UCSC Genome Browser database: update 2010. *Nucleic Acids Research* **38**: D613–D619.
Robertson AG, Bilenky M, Tam A, Zhao Y, Zeng T, Thiessen N, Cezard T, Fejes AP, Wederell ED, Cullum R, et al. 2008. Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. *Genome Res* **18**: 1906–1917.
Rowe HM, Kapopoulou A, Corsinotti A, Fasching L, Macfarlan TS, Tarabay Y, Viville S, Jakobsson J, Pfaff SL, Trono D. 2013. TRIM28 repression of retrotransposon-based enhancers is necessary to preserve transcriptional dynamics in embryonic stem cells. *Genome Res* **23**: 452–461.
Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**: 1036–1040.
Scruggs BS, Gilchrist DA, Nechaev S, Muse GW, Burkholder A, Fargo DC, Adelman K. 2015. Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. *Mol Cell* **58**: 1101–1112.
Shlyueva D, Stampfel G, Stark A. 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* **15**: 272–286.
Simonti CN, Pavličev M, Capra JA. 2017. Transposable Element Exaptation into Regulatory Regions Is Rare, Influenced by Evolutionary Age, and Subject to Pleiotropic Constraints. *Molecular Biology and Evolution* **34**: 2856–2869.
Singh G, Mullany S, Moorthy SD, Zhang R, Mehdi T, Tian R, Duncan AG, Moses AM,

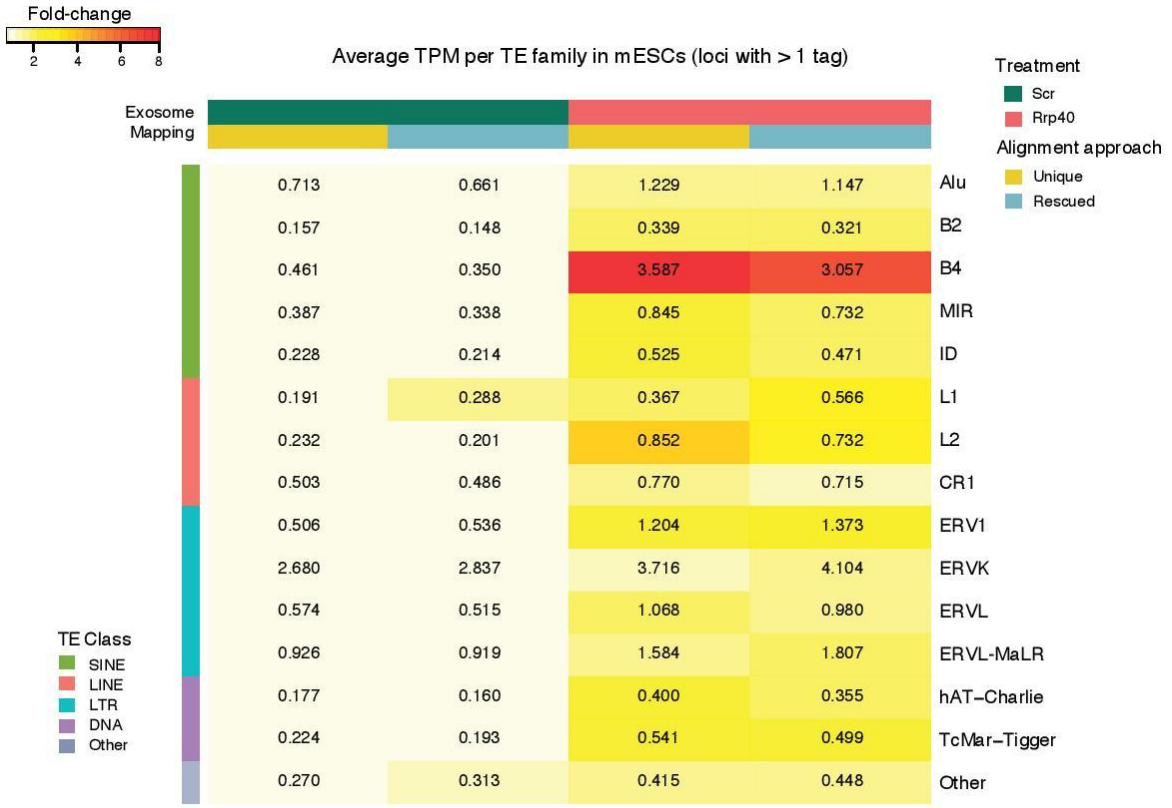
Mitchell JA. 2021. A flexible repertoire of transcription factor binding sites and a diversity threshold determines enhancer activity in embryonic stem cells. *Genome Res* **31**: 564–575.
Smith RP, Taher L, Patwardhan RP, Kim MJ, Inoue F, Shendure J, Ovcharenko I, Ahituv N. 2013. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat Genet* **45**: 1021–1028.
Sun X, Wang X, Tang Z, Grivainis M, Kahler D, Yun C, Mita P, Fenyö D, Boeke JD. 2018. Transcription factor profiling reveals molecular choreography and key regulators of human retrotransposon expression. *PNAS* **115**: E5526–E5535.
Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T. 2014. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res* **24**: 1963–1976.
Sundaram V, Choudhary MNK, Pehrsson E, Xing X, Fiore C, Pandey M, Maricque B, Udawatta M, Ngo D, Chen Y, et al. 2017. Functional cis -regulatory modules encoded by mouse-specific endogenous retrovirus. *Nature Communications* **8**: 14550.
Takahashi H, Lassmann T, Murata M, Carninci P. 2012. 5′ end–centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat Protoc* **7**: 542–561.
The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.
Todd CD, Deniz Ö, Taylor D, Branco MR. 2019. Functional evaluation of transposable elements as enhancers in mouse embryonic and trophoblast stem cells. *Elife* **8**: 1473.
Trizzino M, Park Y, Holsbach-Beltrame M, Aracena K, Mika K, Caliskan M, Perry GH, Lynch VJ, Brown CD. 2017. Transposable elements are the primary source of novelty in primate gene regulation. *Genome Res* gr.218149.116.
Vierstra J, Rynes E, Sandstrom R, Zhang M, Canfield T, Hansen RS, Stehling-Sun S, Sabo PJ, Byron R, Humbert R, et al. 2014. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* **346**: 1007–1012.
Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT, Jasinska AJ, et al. 2015. Enhancer Evolution across 20 Mammalian Species. *Cell* **160**: 554–566.
Walter M, Teissandier A, Pérez-Palacios R, Bourc'his D. 2016. An epigenetic switch ensures transposon repression upon dynamic loss of DNA methylation in embryonic stem cells ed. A.C. Ferguson-Smith. *eLife* **5**: e11418.
Weingarten-Gabbay S, Nir R, Lubliner S, Sharon E, Kalma Y, Weinberger A, Segal E. 2019. Systematic interrogation of human promoters. *Genome Res* **29**: 171–183.
Wu H, Nord AS, Akiyama JA, Shoukry M, Afzal V, Rubin EM, Pennacchio LA, Visel A. 2014. Tissue-Specific RNA Expression Marks Distant-Acting Developmental Enhancers ed. G.S. Barsh. *PLoS Genet* **10**: e1004610.
Young RS, Hayashizaki Y, Andersson R, Sandelin A, Kawaji H, Itoh M, Lassmann T, Carninci P, The FANTOM Consortium, Bickmore WA, et al. 2015. The frequent evolutionary birth and death of functional promoters in mouse and human. *Genome Res* **25**: 1546–1557.
Yu G, Wang L-G, Han Y, He Q-Y. 2012. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology* **16**: 284–287.
Yu G, Wang L-G, He Q-Y. 2015. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**: 2382–2383.
Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD, et al. 2014. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**: 355–364.

Supplementary figures

A

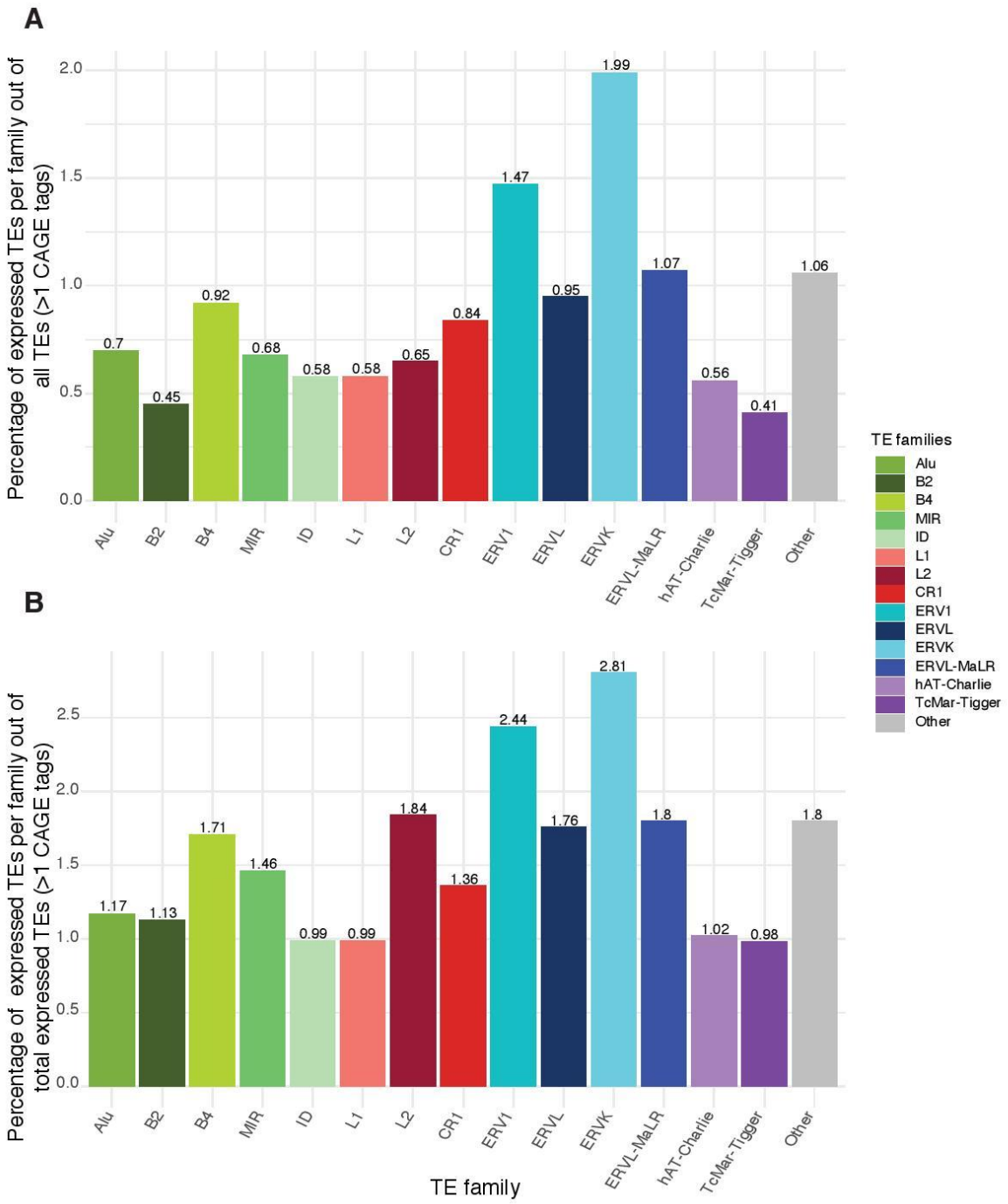


B

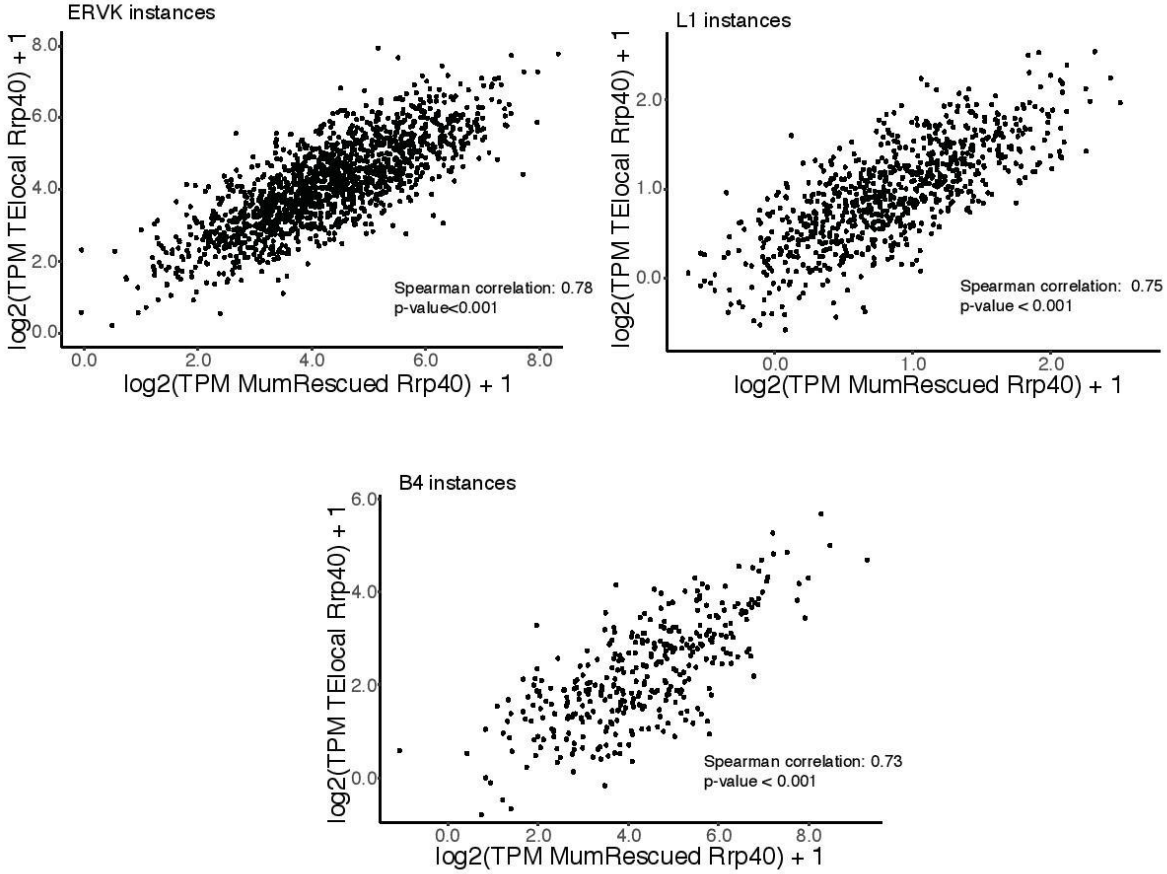


***Supplementary Figure 1**. **A-B**: Heatmaps of average TPM-normalized expression values quantified at the CTSS level for major TE classes (**A**) and major TE families (**B**) in mESCs. Values for control (Scr) or exosome-depleted (Rrp40) CAGE libraries and using only uniquely mappable reads (Unique) or by employing the MumRescueLite algorithm (Rescued) are shown per column. The values in the*

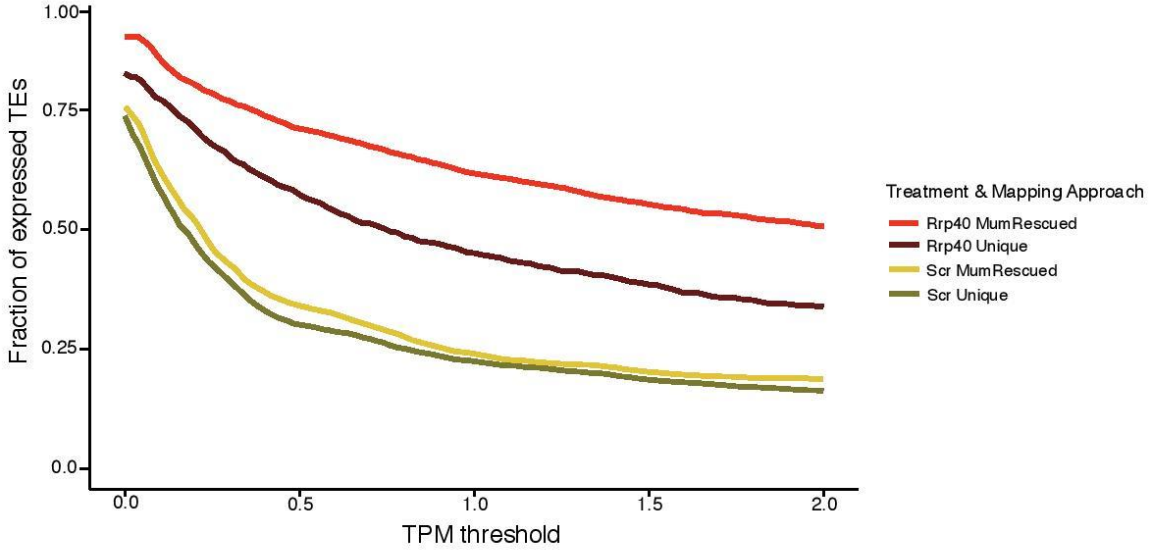
cells are calculated as TPM values divided by the union of TEs expressed in at least one of the columns/conditions, thus representing a comparable average TPM value. The colour key represents the fold-change versus the value for uniquely mappable reads in control samples.



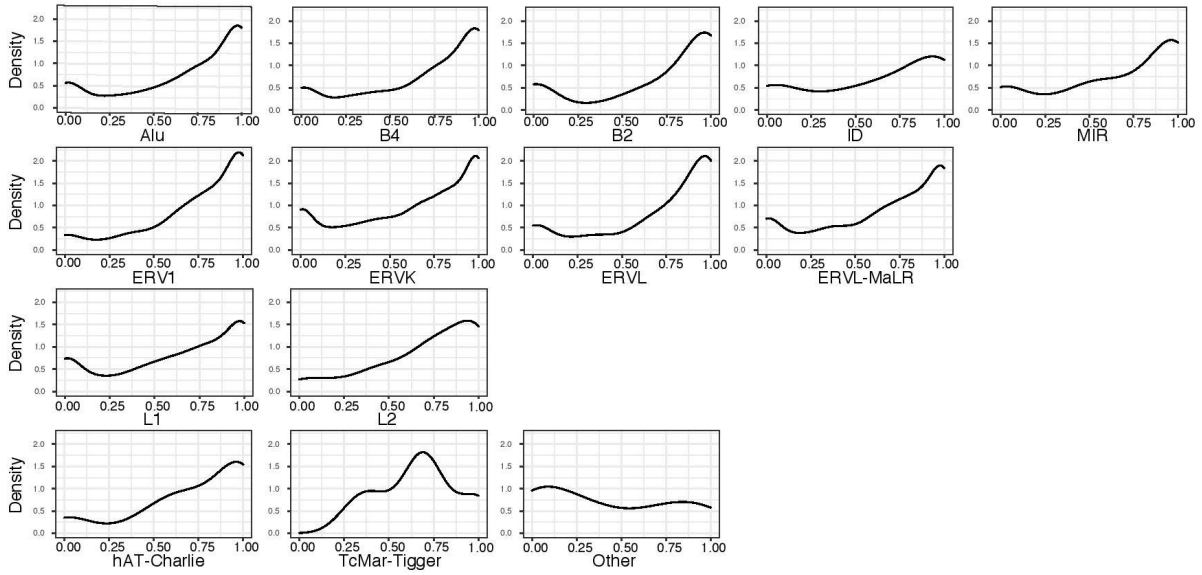
***Supplementary Figure 2**. **A-B:** Percentage of transcribed TEs at the TE family level out of all TE insertions, as annotated in RepeatMasker, (**A**) and out of all identified TEs with two or more CAGE tags (**B**).*



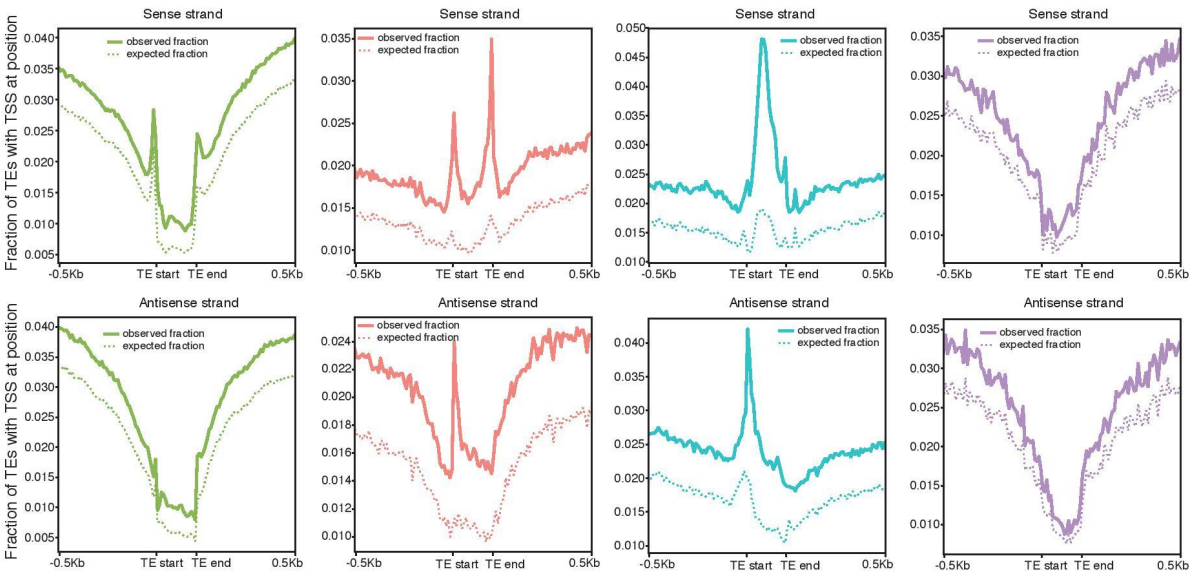
Supplementary Figure 3*: Quantified expression levels (log_2-transformed TPM-normalized counts) by TElocal (vertical axis) and MumRescueLite (horizontal axis) from pooled CAGE libraries. Expressed TE insertions detected by both approaches for three major TE families (ERVK, L1 and B4) are shown.*



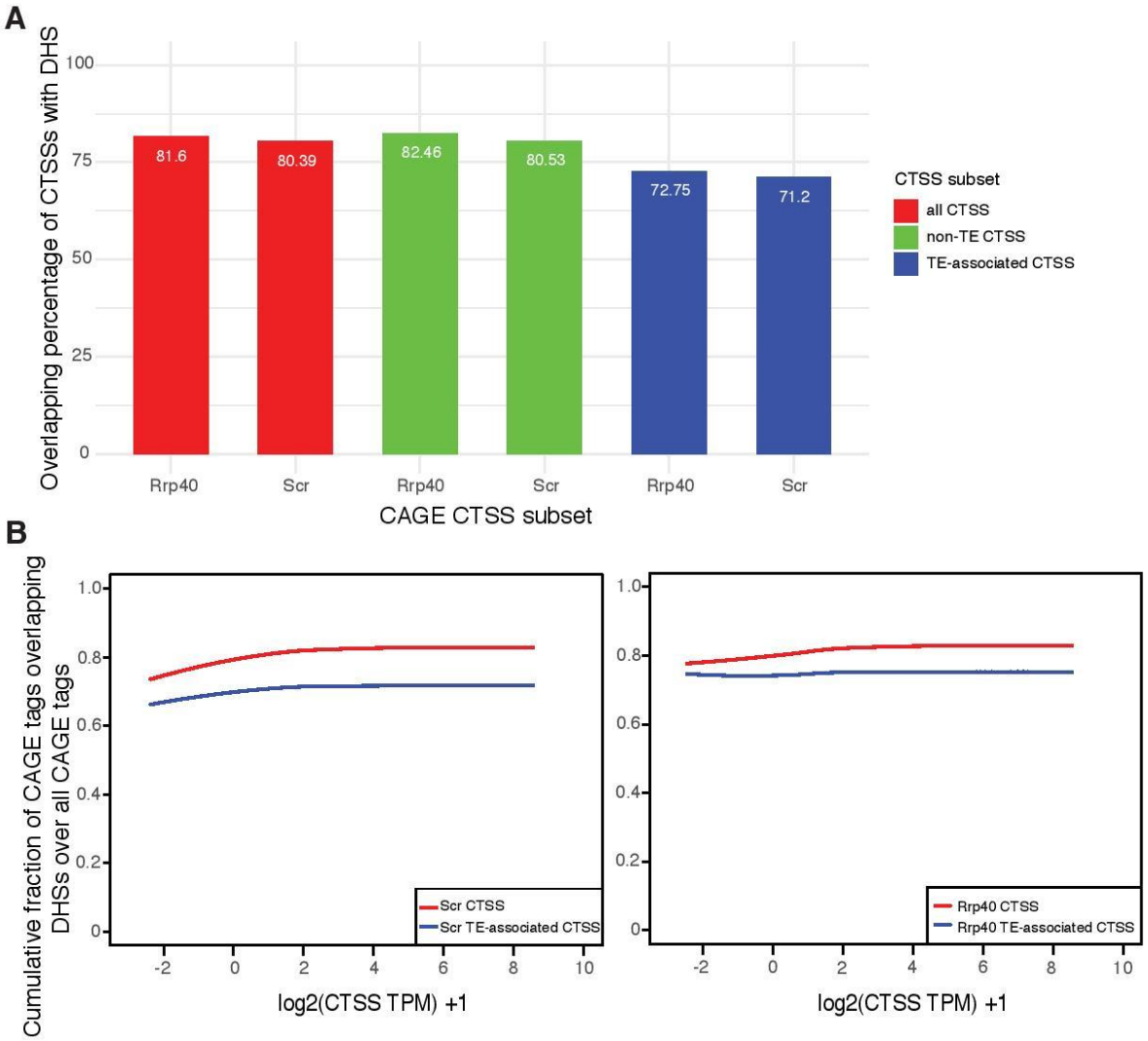
***Supplementary Figure 4**: Fraction of expressed TE insertions versus TPM expression threshold (horizontal axis) over all expressed TEs in control (Scr) and exosome-depleted (Rrp40) libraries, counting only uniquely mappable (Unique) or, in addition, also MumRescueLite-processed (MumRescued) reads.*



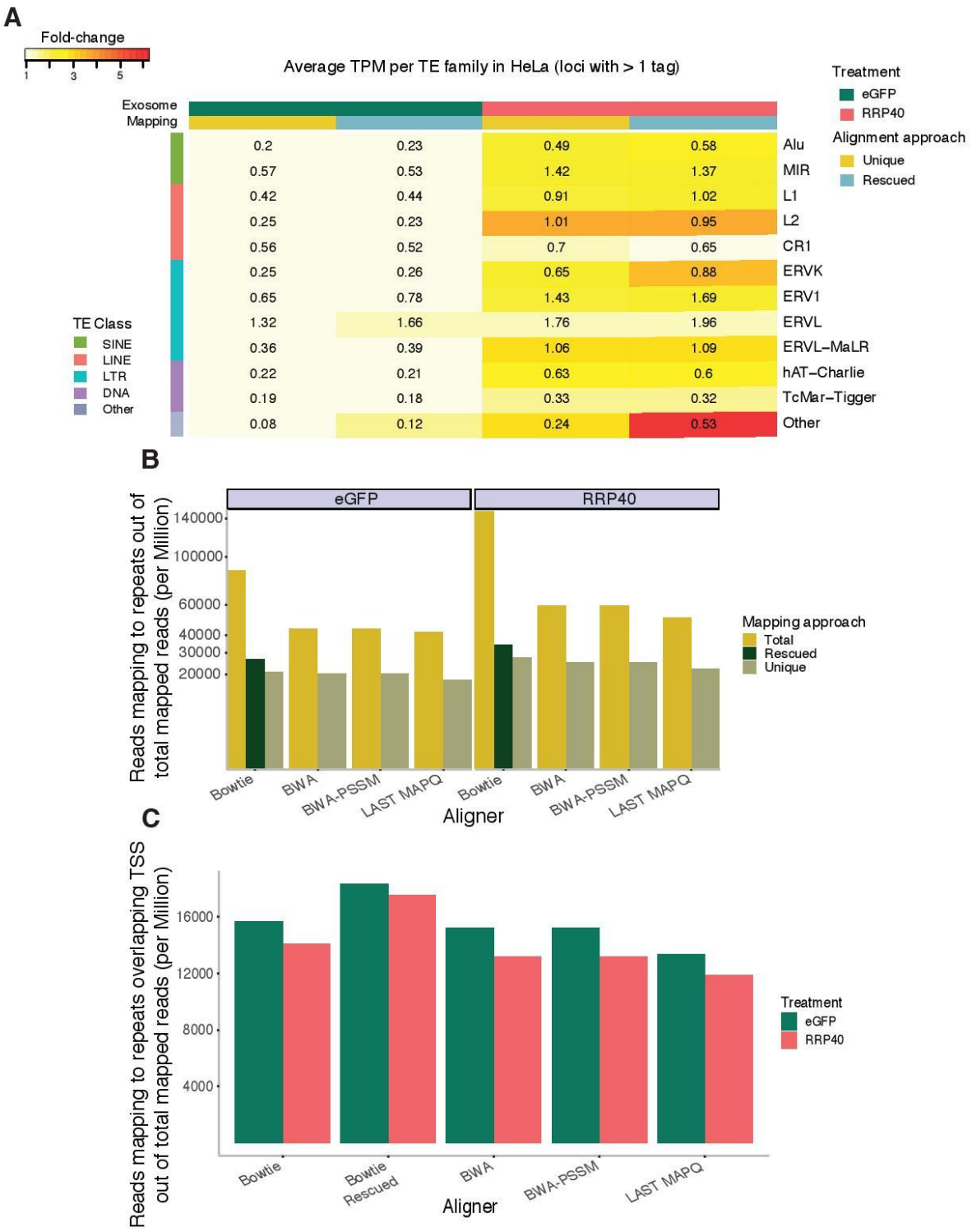
***Supplementary Figure 5**: Densities of exosome sensitivity score, measuring the relative amount of exosome degraded RNAs, for transcripts associated with major TE families based on DHS-associated strand-specific expression levels in control and exosome-depleted CAGE libraries. The exosome sensitivity score ranges from 0 (fully captured by control CAGE) to 1 (only observed upon exosome depletion).*



Supplementary Figure 6*: Average distribution of CAGE-inferred TSS locations (vertical axis; solid lines) +/- 500 bp upstream/downstream and across the body of major TE classes (horizontal axis). TSS locations are visualized separately for the sense (upper panel) and antisense (middle panel) strands. The average expected distributions of TSS locations as determined by a synthetic CAGE uniqueness track (dotted lines) are also shown.*

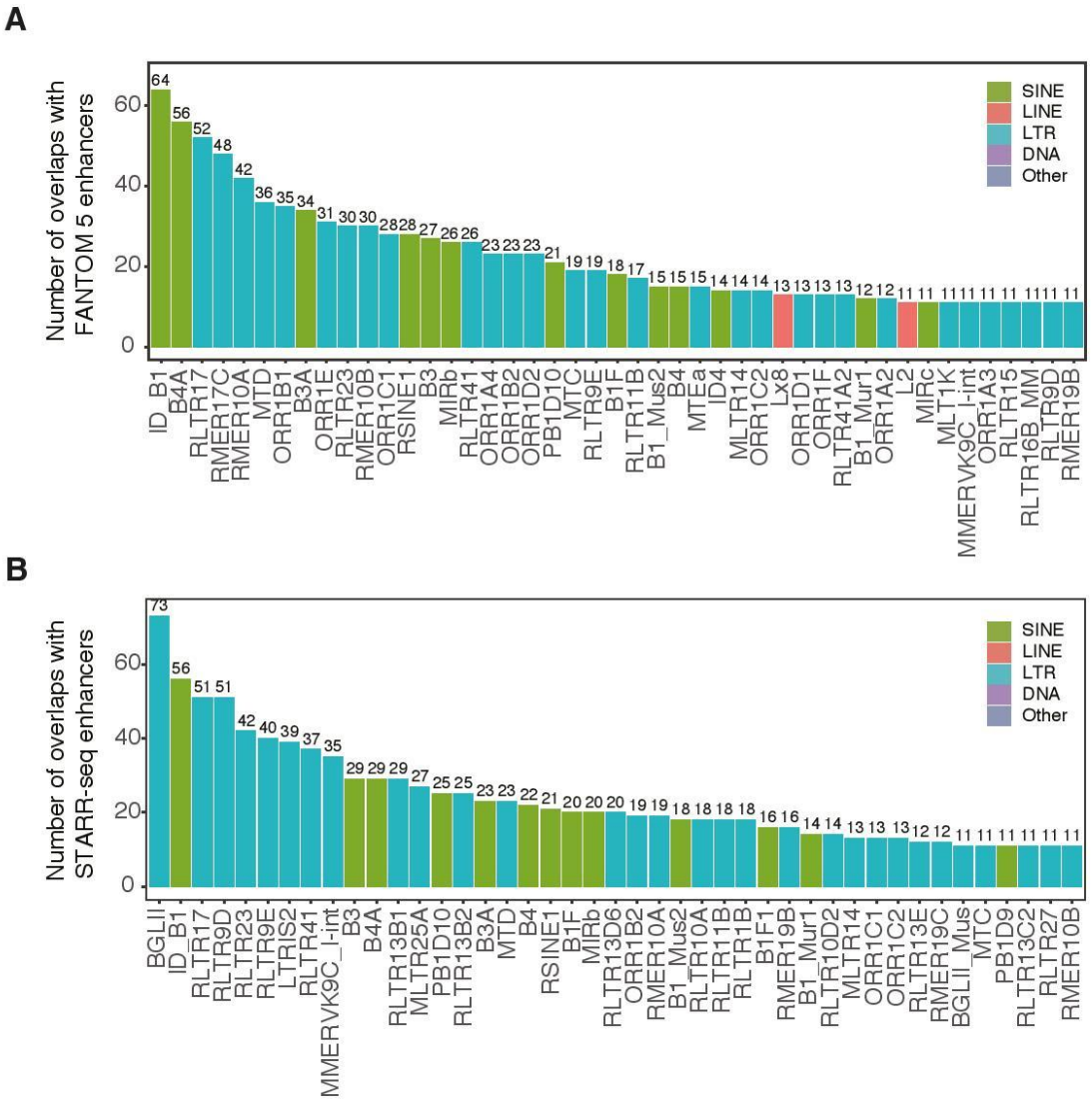


Supplementary Figure 7. ***A:*** *Percentage of CAGE CTSSs within +/- 250 bp of DHS midpoints out of total in each of the subsets: all called (red), all non-TE associated (green) and all TE-associated (blue) CAGE CTSS in control and exosome-depleted CAGE libraries, respectively.* ***B:*** *Cumulative fraction of CAGE tags within +/- 250 bp of DHS midpoints out of total CAGE ctss in control (Scr, left panel) and exosome-depleted (Rrp40, right panel) CAGE libraries versus CTSS expression level (log_2-transformed TPM-normalized counts). The cumulative fraction for all CAGE CTSS (red) and TE-associated CTSSs (blue) is shown.*

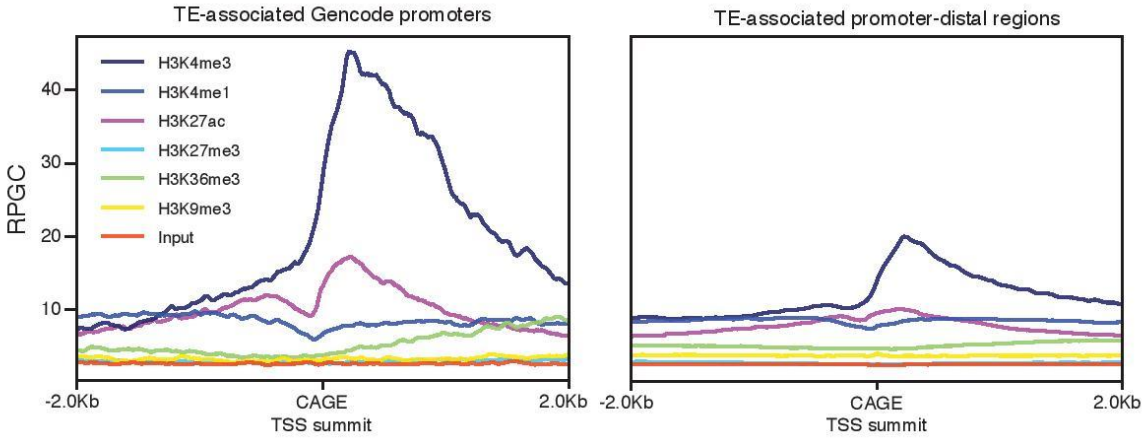


Supplementary Figure 8. A: *Heatmap of average TPM-normalized expression values quantified at the CTSS level for major TE families in HeLa S2 cells. Values for control (eGFP) or exosome-depleted (RRP40) CAGE libraries and using only uniquely mappable reads (Unique) or by employing the MumRescueLite algorithm (Rescued) are shown per column. The values in the cells are calculated as TPM values divided by the union of TEs expressed in at least one of the columns/conditions, thus representing a comparable average TPM value. The colour key represents the fold-change versus the value for uniquely mappable reads in control samples.* ***B:*** *The number of*

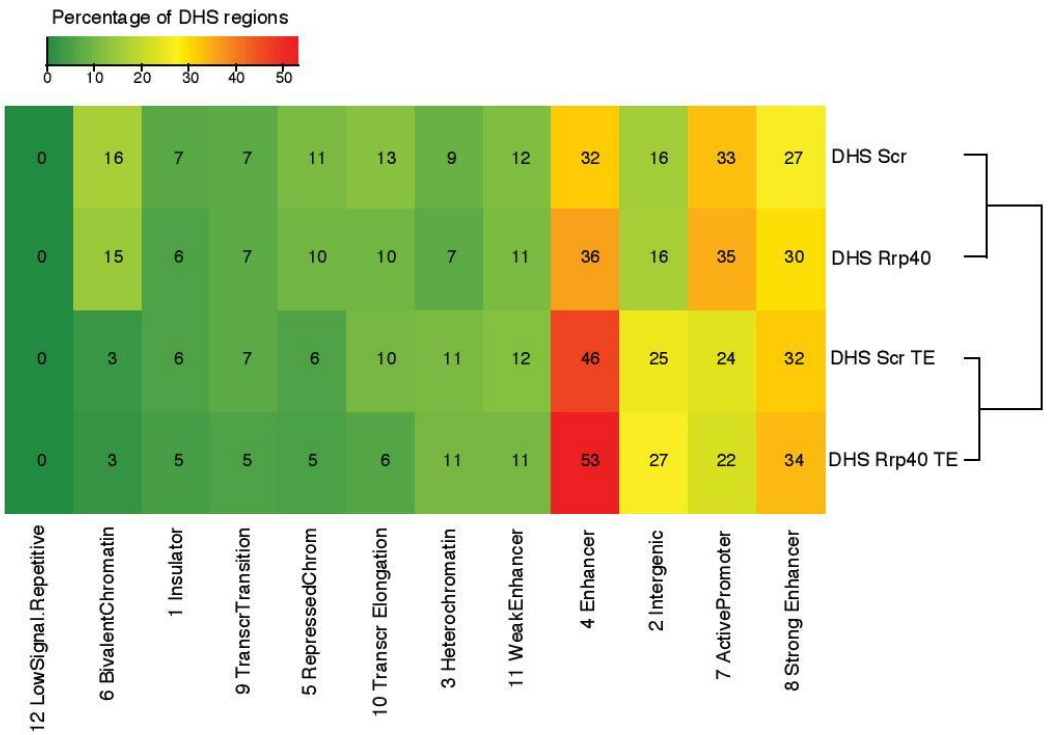
CAGE tags mapping to TEs out of the total mapped reads (per million reads) using different aligners (horizontal axis) and different downstream approaches (keeping all multi-mapping reads, all uniquely mappable reads and Multimapping rescued reads with Bowtie). ***C:*** *The number of CAGE tags mapping to TEs overlapping GENCODE TSSs out of total mapped reads (per million reads) using different aligners and downstream approaches (horizontal axis) as quantified in control (eGFP) and exosome-depleted (RRP40) CAGE libraries.*



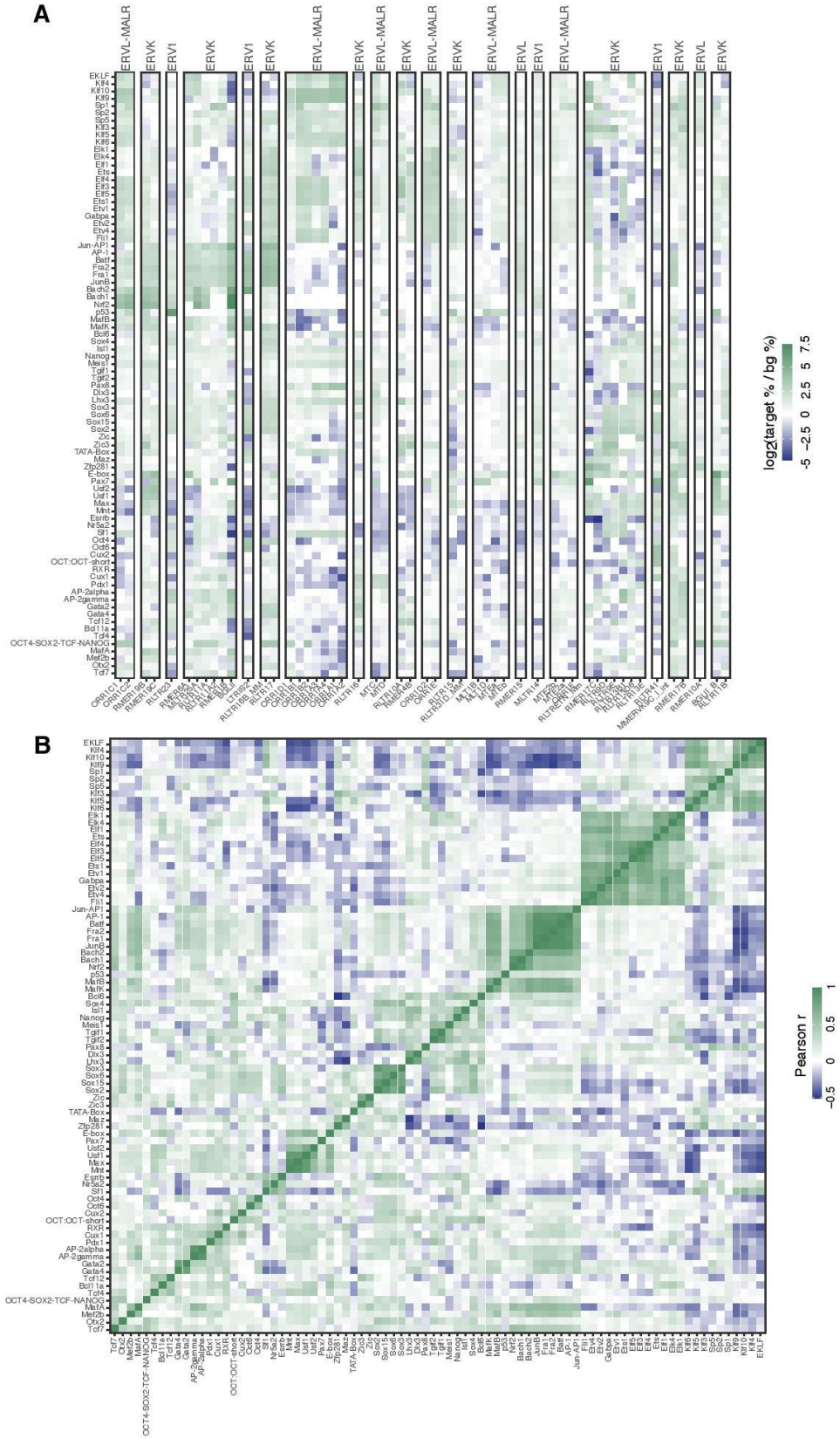
Supplementary Figure 9. A,B: *The number of transcribed TEs overlapping FANTOM5 enhancers (**A**) and STARR-seq mESC enhancers (**B**) at the TE subfamily level.*



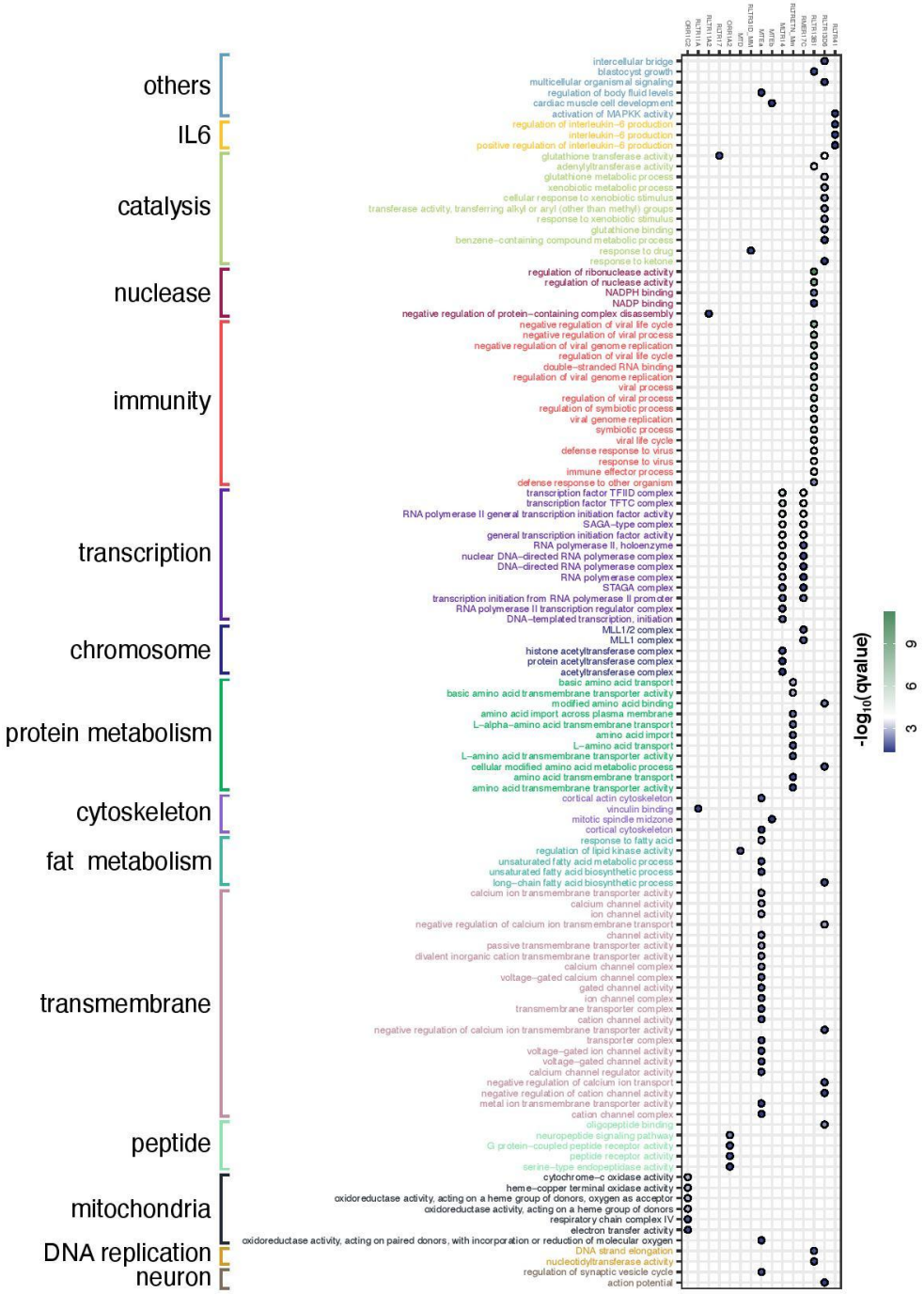
***Supplementary Figure 10**: Average ChIP-seq signal for histone modifications (H3K4me3, H3K4me1, H3K27ac, H3K27me3, H3K36me3, H3K9me3) and input +/- 2 kb around the summits of TE-associated CAGE-inferred TSS clusters at GENCODE promoters (left) and at gene-distal locations (right). Signals shown are reads per genomic context (RPGC), normalized for sequencing depth to 1x genome coverage.*



Supplementary Figure 11: 12-state ChromHMM model revealing the enrichment of specific combinations of histone modifications and TF marks at expressed DHS-associated genomic regions. The colour key (from green (low) to red (high)) and the values in the heatmap cells show the percentage of DHSs that overlap each different chromHMM state (horizontal axis) that are either expressed in control (Scr) and exosome-depleted (Rrp40) CAGE libraries, split by considering either all expressed DHSs or those that also overlap transcribed TEs.



Supplementary Figure 12. A: *Motif enrichments for TFs (columns) in transcribed TEs for all ERV subfamilies (rows) versus a background of genomic regions +/- 200 bp around the summits of all CAGE-inferred TSS clusters. White cells indicate either no enrichment or non-reported values.* ***B:*** *Pearson's correlation coefficients for pairwise TF motif co-occurence across all ERV subfamilies.*



Supplementary Figure 13: *Enriched GO terms based on putative target genes (ABC) of gene distal ERVs for individual ERV subfamilies (foreground, rows) versus all ABC-predicted target genes (background). All enriched GO terms are manually curated into processes or functions and coloured accordingly.*