

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

Stylianos Bakoulis^{1,†}, Robert Krautz^{1,†}, Nicolas Alcaraz^{1,2}, Marco Salvatore¹ and Robin Andersson^{1,*}

¹The Bioinformatics Centre, Department of Biology, University of Copenhagen, 2200 Copenhagen, Denmark and

²Novo Nordisk Foundation Center for Protein Research (CPR), University of Copenhagen, 2200 Copenhagen, Denmark

Received June 21, 2021; Revised January 22, 2022; Editorial Decision January 25, 2022; Accepted January 27, 2022

ABSTRACT

Transposable elements are an abundant source of transcription factor binding sites, and favorable genomic integration may lead to their recruitment by the host genome for gene regulatory functions. However, it is unclear how frequent co-option of transposable elements as regulatory elements is, to which regulatory programs they contribute and how they compare to regulatory elements devoid of transposable elements. Here, we report a transcription initiation-centric, in-depth characterization of the transposon-derived regulatory landscape of mouse embryonic stem cells. We demonstrate that a substantial number of transposable element insertions, in particular endogenous retroviral elements, are associated with open chromatin regions that are divergently transcribed into unstable RNAs in a cell-type specific manner, and that these elements contribute to a sizable proportion of active enhancers and gene promoters. We further show that transposon subfamilies contribute differently and distinctly to the pluripotency regulatory program through their repertoires of transcription factor binding site sequences, 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 (1–4). The 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 (5–8). TF DNA-sequence preferences (9), gene expression (10,11) and the specific regulation of genes by TFs (12,13) are generally well conserved across eukaryotes. On the contrary, regulatory elements with enhancer activity are generally associated with high evolutionary turnover (14–16).

Transposable elements (TEs) are an abundant source of TF binding sites that contribute to the spread of sequences with regulatory potential (17). In mammals, the large majority of TEs are dormant, having lost their ability to replicate and are maintained repressed by H3K9me3, H3K27me3 and DNA methylation (18–22). However, favorable genomic integration may lead to the co-option of TEs as endogenous regulatory elements by utilizing their native or acquired TF binding sites (23–32). Consequently, the majority of species-specific open chromatin regions are associated with TEs (14) and TE-derived enhancers are generally not well conserved across evolution (25,33,34). 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 (29,31) and initiate transcription (35). In addition, species-specific TE-derived regulatory elements cause binding differences of pluripotency TFs between human and mouse ESCs (36). Distribution and fixation of mobile elements with readily available regulatory potential may therefore provide regulatory innovation but also stabilize the gene regulatory functions of TFs. Thus, characterizing the contribution of TEs to transcriptional regulation has the potential to provide insights into the formation of pluripotency regulatory programs, the origins of regulatory elements and thus the basis for their evolutionary turnover.

Most studies to date have inferred TE-associated regulatory elements from chromatin-accessible loci flanking nucleosomes with specific histone modifications indicative of enhancers (e.g., H3K4me1, H3K27ac). However, only

*To whom correspondence should be addressed. Tel: +45 35330245; Email: robin@binf.ku.dk

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

few of such predicted loci show enhancer activity (31,37). Rather, a growing body of literature points toward divergent transcription initiation as a key property of active regulatory elements with either enhancer or promoter function (4,38–45). While there is a general relationship between histone modifications and the transcriptional output of a regulatory element (4,44–46), the transcriptional status of a regulatory element better reveals its regulatory potential (40,44,47). Characterization of TE-derived enhancers from transcription initiation events therefore provide a more accurate picture of their regulatory contribution.

Divergent transcription of regulatory elements is established at closely spaced pairs of divergently oriented core promoters within open chromatin, 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 (39–42,44,48). Such transcription initiation events are accurately identified and quantified through 5' end sequencing of capped RNAs (CAGE, Cap Analysis of Gene Expression) (49,50). eRNAs and PROMPTs are generally non-polyadenylated and thus unprotected at their 3' ends, and as a consequence these RNA species are frequently targeted by the 3'-5' ribonucleolytic RNA exosome for degradation (39,51,52).

Although co-option of TEs as regulatory elements has been established as a mode of regulatory innovation, its extent and how TE-derived regulatory elements compare to non-TE associated regulatory elements (e.g., with regards to divergent transcription initiation, RNA metabolism and TF binding potential) remain unclear. Encouraged by the possibility to study TE-associated regulatory elements through transcription initiation mapping (35), we here investigate how the wide repertoire of mouse TEs contribute to the transcription initiation landscape and thus regulatory elements of mouse ESCs (mESCs). We demonstrate that many dormant TE insertions, in particular endogenous retroviral elements (ERVs), carry open chromatin regions that are divergently transcribed into unstable RNAs targeted by the exosome for degradation. Furthermore, open chromatin regions that are associated with transcribed ERVs show a high degree of species specificity and a large fraction of these either have enhancer function or contribute to gene promoters. This suggests that TE co-option as regulatory elements contributes to a sizable proportion of active species-specific regulatory elements in mESCs. Our transcription-centric approach allows for an unbiased systematic investigation of the regulatory potential across TE subfamilies, indicating that these contribute differently to the TF binding repertoire of the mouse genome, which can be linked to regulatory specificity in mESCs.

MATERIALS AND METHODS

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

Previously sequenced CAGE libraries for E14 mESCs, embryoid bodies sampled after 3 days of differentiation of mESCs (GEO ID GSE115710) (53) and human HeLa S2 cells (GEO ID GSE62047) (39) were collected. As

described in each report, CAGE libraries were prepared from exosome depleted samples as well as from control samples. E14 mESCs and differentiated embryoid bodies 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 two mismatches were discarded using rRNA dust (fantom.gsc.riken.jp/5/suppl/rRNA_dust/). 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 three additional alignment approaches (Supplementary Figure S1B,C). BWA (Burrows-Wheeler Aligner) version 0.7.15-r1140 (54) was used with the parameter *-n 2* (maximum distance for each alignment) and a subsequent mapping quality MAPQ > 20 threshold, using SAMtools (version 1.3.1) (55). BWA-PSSM, a modification of BWA using position specific scoring matrices (PSSM) (56), was used with parameters *-n2*, *-m 2000* (to allow for more suboptimal partial alignments to be tested) and a downstream MAPQ > 20 threshold. Finally, LAST (version 801) (57) was used with parameters *-u NEAR* and *-R 01* (lastdb) and *-Q 1* and *-D 100* (lastal), followed by a downstream MAPQ > 20 threshold.

Probabilistic multi-mapping rescue of CAGE tags

Following initial mapping of reads with Bowtie, we employed MUMRescueLite (58) 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 multi-mapping read. The probabilistic alignment of a multi-mapping read 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 were 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 (version 0.1.0) from the TEToolkit suite (59) to quantify transposable element expression at the locus level. The resulting quantification for each TE for both TElocal and multi-mapping rescuing output was normalized to tags per million mapped reads (TPM) and the results of the two approaches 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 transcription start sites (CTSSs). We then assigned CTSSs to transposable elements (TEs) as defined by RepeatMasker (version 4.0.7; <http://www.repeatmasker.org/>) and considered only instances with two or more CAGE tags. TEs denoted as Alu elements in RepeatMasker were considered B1 and proto-B1 (PB1) elements. Tag clusters (TCs) were generated from pooled CAGE libraries per condition, and wide TCs were narrowed or decomposed into sub-peak TCs if they contained multiple peaks (https://github.com/anderssonlab/CAGEfightR_extensions), as previously described (44). 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 <10% of the summit signal, it was discarded and decomposed TCs were merged if positioned within 20 bp from each other on the same strand. Expression quantification of each individual TC in each CAGE replicate was performed by aggregating the read counts of CTSSs falling into its genomic region. 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 TCs to TEs on the same strand through direct overlap using BEDtools (60).

Footprints of CAGE expression on TEs

To investigate the relationship between transposable element families and/or classes to CTSS locations, we plot-

ted the average binarized (presence or 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 (61). 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. The log₂ ratio of observed (CAGE libraries) versus expected (as estimated from the synthetic uniqueness track) average binarized signal was calculated in R (version 4.0.3; <http://www.R-project.org/>).

HeLa RNA-seq data processing

RNA-seq data from HeLa cells depleted of hRRP40 using siRNA-mediated knockdown as described elsewhere (62) (SRA accession: SRX365673) were considered. Briefly, after filtering of low-quality reads, removal of Illumina adaptors and reads <25 bp with Trimmomatic (version 0.36) (63), reads were mapped against the human reference genome (hg19) using HISAT2 (version 2.1.0) (64). Uniquely mapped and properly paired reads were selected with SAMtools (version 1.3.1). Gene-level expression quantification of mapped reads was performed with featureCounts (version 1.6.3) (65). 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 using the glm function (see Supplementary Note).

CAGE gene-level expression quantification

For statistical comparison of gene-level HeLa RNA-seq expression to gene-level HeLa expression as measured by CAGE, we quantified abundances of genes using CTSSs within ± 500 bp windows from the 5'end of GENCODE (version M10; GRCm38) transcripts (66), as CAGE signal saturates after ~ 500 bp from annotated gene TSSs (50). 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. Similarly, gene-level expression was quantified using CAGE data for mESC and embryoid bodies.

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

For 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 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 optimizing the cover-

age of proximal CAGE tags within flanking windows of size ± 300 bp around them, as previously described (44). The final set of 165 052 transcribed DHS was determined by filtering DHSs to not overlap any other DHS ± 300 bp 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 (39). 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 (44).

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 were 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.

Estimation of evolutionary conservation of TE-associated DHSs

Genomic regions spanning ± 150 bp around mESCs DHS signal peaks carrying CAGE tags and overlapping TEs were aligned to rat (rn7) and human (hg38) assemblies using the UCSC liftOver tool (67) with a *-minMatch* = 0.6 parameter. Similarly, 300 bp regions of nonTE-associated DHSs carrying CAGE tags, the mm10 genome assembly split in 300 bp fragments, the subset of those fragments not overlapping TEs, and all TE instances of RepeatMasker (full length) were aligned to rat and human assemblies. The genomic regions that had a $>60\%$ match (coverage) with those in the other species were considered orthologous.

FANTOM enhancers and enhancer peak calling from STARR-seq data

Transcribed enhancers identified by remapped FANTOM5 CAGE libraries to mm10, deposited in Zenodo (68), were associated with TE-associated DHSs by overlap using BEDTools. STARR-seq data for 2iL grown mESCs E14Tg2a (E14) (69) (GEO ID GSE143546) were used to evaluate the enhancer potential of transcribed TEs. STARR-seq data were processed using Bowtie2 (70) and SAMtools (55). 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 (version 1.0) (71) 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 analysis 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 (version 2.5) (72), reads were mapped to the mouse genome assembly mm10 with bwa, duplicate reads were removed with Picard (version 2.20.7) (<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 ± 2000 bp, using deepTools (version 3.1.3). Hierarchical clustering, annotation and visualization were conducted in R with ChIPSeeker (73), ggplot2 (<https://ggplot2.tidyverse.org>) and profileplyr (version 1.6.0), using clustering parameters *rowMax* for summarizing the ranges across samples and *median* for defining proximity between clusters. The association of clusters with TE classes and families was done in R, using the Fisher's exact test with Bonferroni correction for multiple testing.

Chromatin state discovery and characterization using chromHMM

To annotate TE-associated DHSs and characterize which regulatory 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 (74). The multivariate hidden Markov model framework of chromHMM (75,76) 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 (version 3.1.1).

Transcription factor motif analysis

We scanned full ERV transposons carrying > 1 CAGE tag with known TF motifs in the HOMER motif database using findMotifsGenome.pl (77). 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 (± 200 bp regions) as the background set. The significantly enriched TF binding site motifs were selected with three additional conditions: (i) TF genes were expressed in mESCs using gene-level CAGE quantification, (ii) at least 10% of the target set contained the motif and (iii) known TF genes had 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. In order to account for cell type specific TE expression, as measured by CAGE in mESCs and embryoid bodies, we scanned TE-associated DHSs in mESCs, using scanMotifGenomeWide.pl (77), with the position weight matrices of nine pluripotency TFs that demonstrated enrichment across several transcribed ERV subfamilies (Figure 5A). ERV instances carrying a predicted binding site of at least one out of the nine TFs were considered to be associated with pluripotency TFs. 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 (version 11 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 enhancers (78) as target sets and all ABC-predicted target genes as background, were performed in R using clusterProfiler (79). 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 (80) and were associated with CAGE TCs by coordinate overlap (BedTools).

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 (4,40,44,45,81,82). To this end, we analyzed the genomic location of sequencing reads of capped RNA 5' ends in mESCs. 5' ends of capped RNAs reveal transcription start sites (TSSs) of RNA polymerase II transcripts and can be accurately assayed using CAGE (49). Here, we considered CAGE data from mESCs depleted for the RNA exosome core component Rrp40 (53). TSSs identified by these data thus also include those whose RNAs are targeted by the exosome for degradation.

To allow characterization of TSSs at base-pair resolution within TE insertions, we considered both uniquely mapping and multi-mapping rescued (58,83) CAGE reads (see Materials and Methods). Multi-mapping rescue increased the number of identified TE-associated TSSs and improved expression quantification of TE-associated loci in mESCs and, as a confirmation, also in HeLa cells (Supplementary note; Supplementary Figures S1–S4). Comparable expression levels were observed using an alternative strategy based on maximum likelihood alignments to annotated repeats (59), although the sets of detected, expressed TE insertions varied between methods (Supplementary note; Supplementary Figure S5).

The ability to infer the TSSs and expression levels of TE-derived RNAs from CAGE prompted us to characterize the transcription initiation patterns across TE families. A total of 82 383 mouse TE insertion events were associated with measurable RNA polymerase II transcripts in mESCs. Among these, endogenous retroviral TE family elements (ERV1, ERVK, ERVL, ERVL-MaLR) were most prevalent ($P < 2e-16$, Fisher's exact test; Supplementary Figures S2B and S3). These TE families and L1 LINE elements showed preferential TSS locations close to their repeat body boundaries (Figure 1A,B). Of note, TSS location preferences could in general not be explained by varying mappability over repeat bodies (Supplementary Figure S6). However, a lack of detected expression in DNA transposons could be due to low mappability, suggesting that expression quantification and TSS mapping of this class may require alternative approaches or longer sequencing reads. By taking mappability into account, a strong divergent pattern for L1 elements originating at their 5' and 3' ends was revealed (Supplementary Figure S6), suggesting that some of these may utilize their native TSSs. We note, however, that TSSs within TEs often deviate from their native TSSs, as seen for ERVs (Supplementary Figure S6).

The average profiles of TSS locations for ERVs and L1 LINEs (Figure 1A,B) indicated divergent transcription initiation, reminiscent of that of gene promoters and gene-distal enhancers (38–40,43). To investigate the functional relevance of individual genomic TE insertions, 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 Figure S7) and the majority of TE-associated DHSs displayed balanced bidirectional transcription initiation (Figure 1C), a hallmark of gene-distal regulatory elements with enhancer activity (40). This suggests that some of the investigated TEs may act as enhancers.

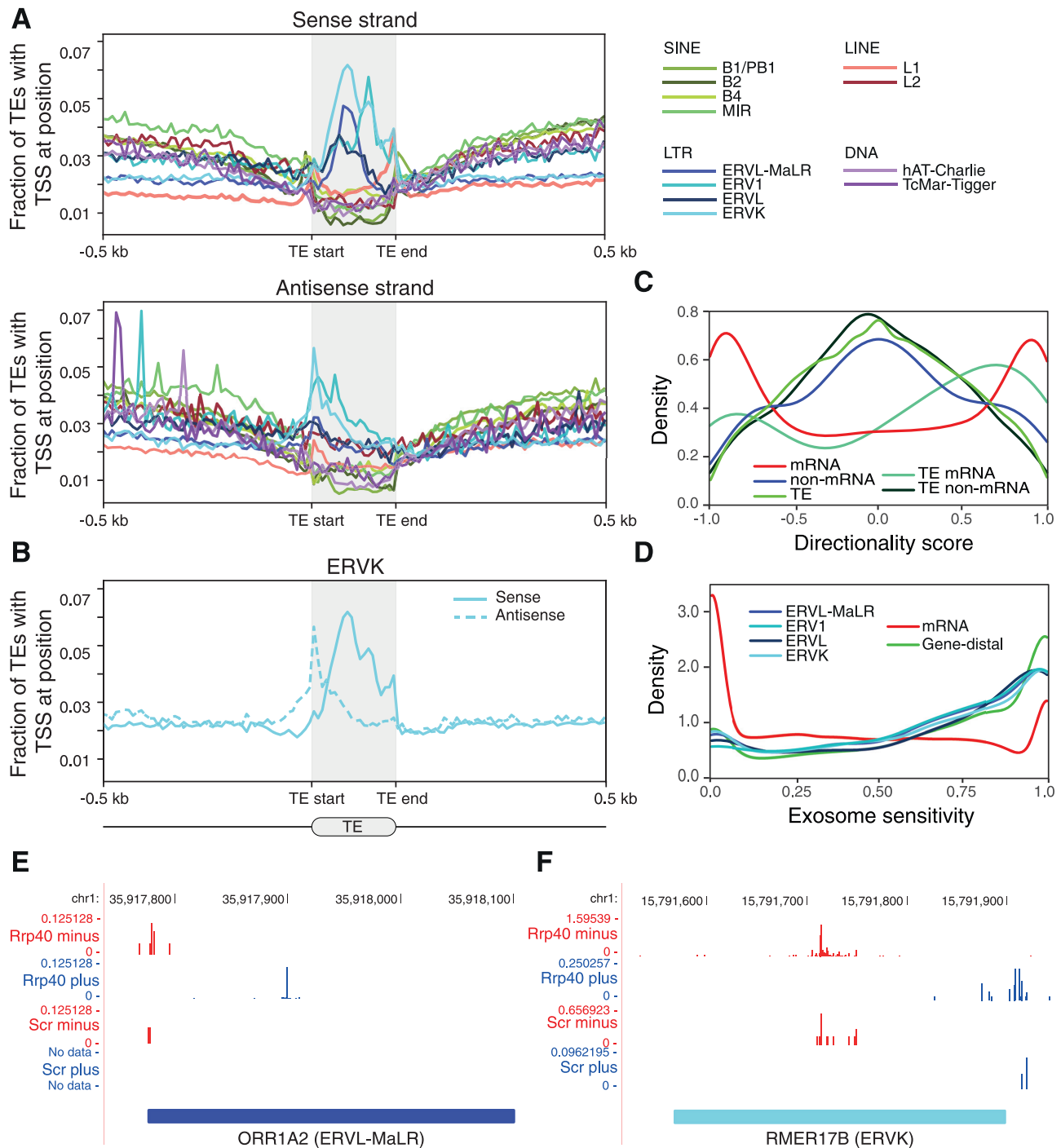


Figure 1. TEs are divergently transcribed into unstable RNAs. (A) Average distribution of CAGE-inferred TSS locations (vertical axis; expression agnostic) ± 500 bp upstream/downstream and across the body of major TE families (horizontal axis). TSS locations are visualized separately for the sense (upper 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 and non-mRNA (non-protein-coding GENCODE transcripts) as well as TE-associated and non-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. For comparison, exosome sensitivity is shown for mRNAs and gene-distal loci. (E and F) Genome browser tracks for two loci of unannotated transcripts with characteristic divergent expression patterns falling on TE insertions of ORR1A2 (ERV1-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 by the RNA exosome (39,51,52). Comparing CAGE data from exosome-depleted mESCs with wildtype mESCs (scrambled shRNA control) (53) confirmed that TE-derived RNAs are also degraded by the exosome (Figure 1D; Supplementary Figures S2–S4,8; Table S1), as exemplified by CAGE data at genomic loci containing insertion sites for ORR1A2 (ERVL-MaLR) and RMER17B (ERVK) (Figure 1E,F). The exosome sensitivity of TE-derived RNAs was similar to those of gene-distal loci, in contrast to mRNAs which are mostly protected against decay by the exosome (Figure 1D) (39,51,52). 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, more TEs with measurable transcription initiation (>1 CAGE tag, hereafter referred to as transcribed TEs) were detected in exosome-depleted compared to wildtype mESCs (26 079 in wild-type mESCs; 64 299 in exosome-depleted mESCs; 82 383 in pooled CAGE data of exosome-depleted and wild-type mESCs). Exosome-depleted cells further displayed an increased expression level of TE-derived RNAs (Supplementary Figure S2 and Table S1). Together, these results indicate that, although a prominent number of TEs are transcribed, the majority of derived RNAs are at least partially degraded.

Taken together, our results demonstrate that transcribed TEs in mESCs are associated with open chromatin regions accommodating divergent transcription initiation of RNAs that are subject to degradation. These properties imply the potential for TEs to function as enhancers.

Transcription of dormant ERVs reveals co-opted regulatory elements

The characteristics of transcribed TEs and the similarities between TE-derived RNAs, eRNAs and PROMPTs led us to investigate further similarities with transcribed 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 (Figure 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. Many CAGE-inferred TSSs of transcribed TEs overlapped with open chromatin regions as defined by DHSs (62 514, ~65%, of transcribed TEs overlapped via their TSSs with 7431 DHSs). This suggests that a prominent fraction of transcribed TEs may act, alone (34% of DHSs overlapped TSSs of one transcribed TE) or in combination with other transcribed TEs (66% of DHSs overlapped TSSs of multiple transcribed TEs), as gene regulatory elements, but may also reflect a selfish TE tropism for open chromatin regions (14,33,84,85).

In agreement with an association of TEs with species-specific DHSs (14), we observed that the majority of transcribed TE-associated DHSs in mESCs are rodent- or

mouse-specific (Figure 2B). Only 16.9% of DHS sequences had human orthologous sequences while 79.4% were orthologous to the rat genome, indicating that they derive from TEs that have accumulated after the primate-rodent split. In contrast, transcribed DHSs devoid of TEs were more conserved. Interestingly, transcribed DHSs associated with ERV subfamilies were even more mouse specific (Figure 2B and Supplementary Figure S10), in agreement with previous reports (25,32,33), suggesting regulatory innovation through TE exaptation and that these elements contribute to the high evolutionary turnover of enhancers.

To generally assess whether transcribed dormant TEs have been co-opted as transcriptional regulatory elements, we first evaluated their association with FANTOM enhancers (40,68,86), an extensively validated set of regulatory elements with predicted enhancer function inferred from bidirectional transcription initiation across a large number of cell types and tissues. Notably, 22.1% of FANTOM mouse enhancers with detectable expression in mESCs (1979 out of 8942) overlapped with transcribed TE-associated DHSs in mESCs (1504 out of 7431, 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 was composed of DHSs with LTR elements, in particular ERVKs and ERVL-MaLRs (Figure 2C and Supplementary Figure S9A), consistent with previously reported bidirectional transcription of ERVs (35). 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.

Since FANTOM enhancers were predicted based on the same property that we had observed for transcribed TEs, namely divergent transcription initiation, alternative experimental data are required to evaluate the enhancer potential of TEs (87). To this end, we considered the *in vitro* enhancer potential of transcribed TEs using genome-wide mESC STARR-seq data (69). About 18.9% of the transcribed TE-associated DHSs (1404 out of 7431) overlapped with open chromatin-associated STARR-seq enhancers (1949 out of 7078, 26.2%; Figure 2D; Supplementary Figure S9B). Again, LTRs, in particular ERVKs, constitute a sizable fraction of these enhancers, in agreement with the FANTOM enhancer overlap. In contrast, transcribed L1 LINE elements, despite frequently residing in gene-distal intergenic loci (Figure 2A), rarely overlapped with FANTOM or STARR-seq enhancers. L1 elements are therefore less likely to act as enhancers in ESCs, in line with human LINEs (29). Interestingly, in agreement with the strong association between transcription initiation and *in vitro* enhancer potential (40,44,47), we observed that non-transcribed TE-associated DHSs were to a lesser degree validated by STARR-seq enhancers than transcribed ones ($P < 2e-16$ for ERVKs and ERVL-MaLR, Fisher's exact test).

Combined, the STARR-seq and FANTOM sets indicate that 2378 (~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

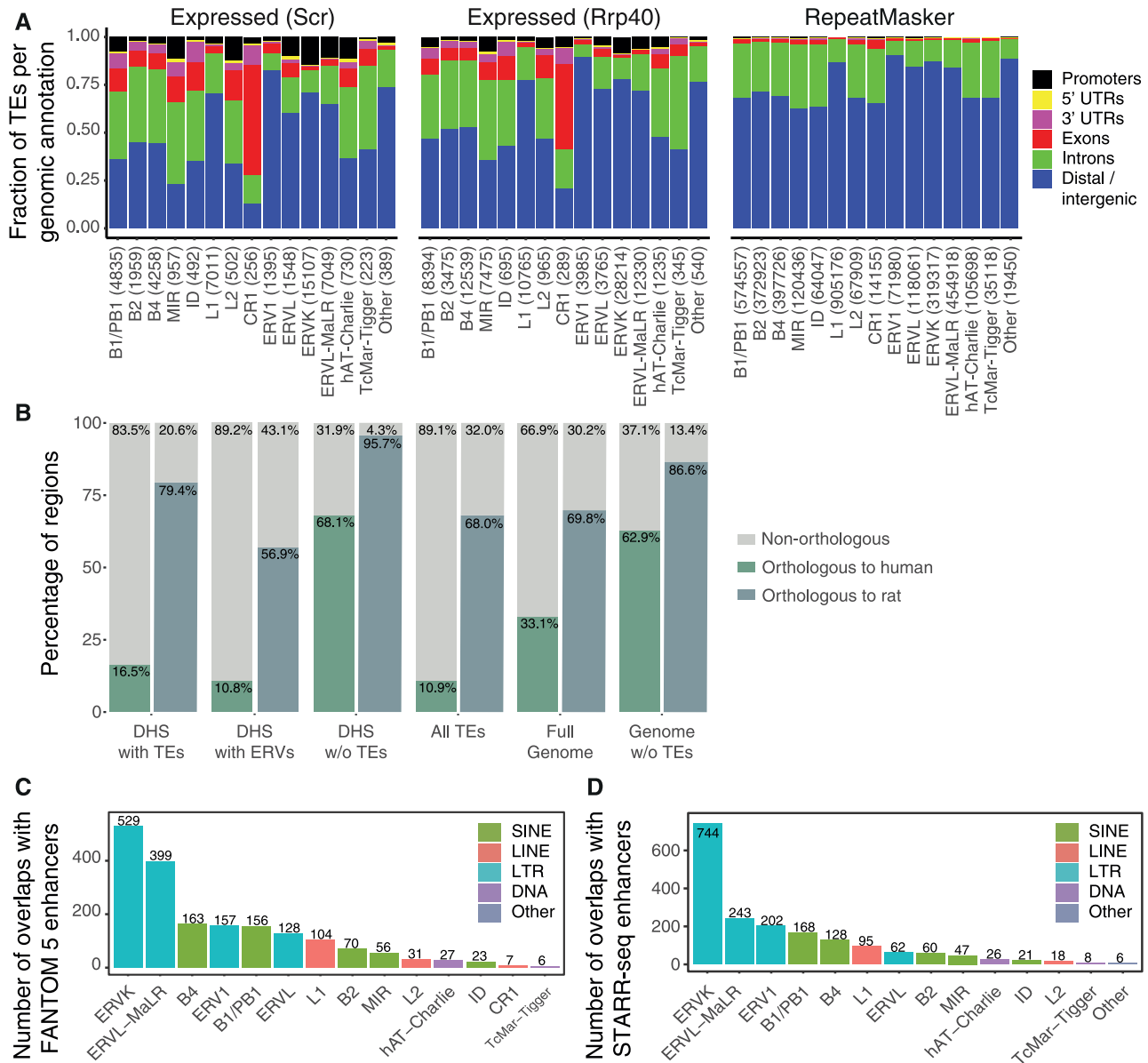


Figure 2. Transcription of transposable elements reveals co-opted regulatory elements. **(A)** Fraction of expressed TEs in control (Scr) and exosome KD (Rrp40) mESCs as well as all TEs annotated in RepeatMasker (expression agnostic) per genomic annotation group for each TE family. The number of instances of each TE family is shown in parenthesis. **(B)** Percentages of orthologous TE-associated, ERV-associated, and non-TE-associated expressed DHSs, as well as background regions between mouse and rat or human genomes. **(C)** The number of transcribed TEs overlapping FANTOM 5 mouse enhancers at the TE family level. TE subfamily counts are displayed in Supplementary Figure S9A. **(D)** The number of transcribed TEs overlapping STARR-seq mESC enhancers at the TE family level. TE subfamily counts are displayed in Supplementary Figure S9B.

individual TEs (Figure 3A; RLTR41 insertion) or, in some cases, pairs of TEs (Figure 3B; RLTR41 in pairs with MYSERV-int and RMER10B insertions).

Transcribed TEs carry chromatin features of regulatory elements

The strong overlap of ERVK elements and weak overlap of L1 elements with FANTOM and STARR-seq enhancer sets implies differences in the regulatory potential among TE families (Figure 2C,D). To investigate 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 TE-associated DHSs (88,89). Hierarchical clustering of these aggregated signals revealed six major chromatin states (Figure 4A and Supplementary Figure S11).

The majority of loci (clusters 1, 3, 4) carried histone modification signal associated with active regulatory elements (90–92): H3K4me1/3 and/or H3K27ac (Figure 4A and Supplementary Figure S12). The elements in cluster 3 and 4 were mainly intronic or gene-distal (>80%; Figure 4B) and displayed H3K4me1 signal, associated with weak transcription (4,46), in combination with H3K27ac.

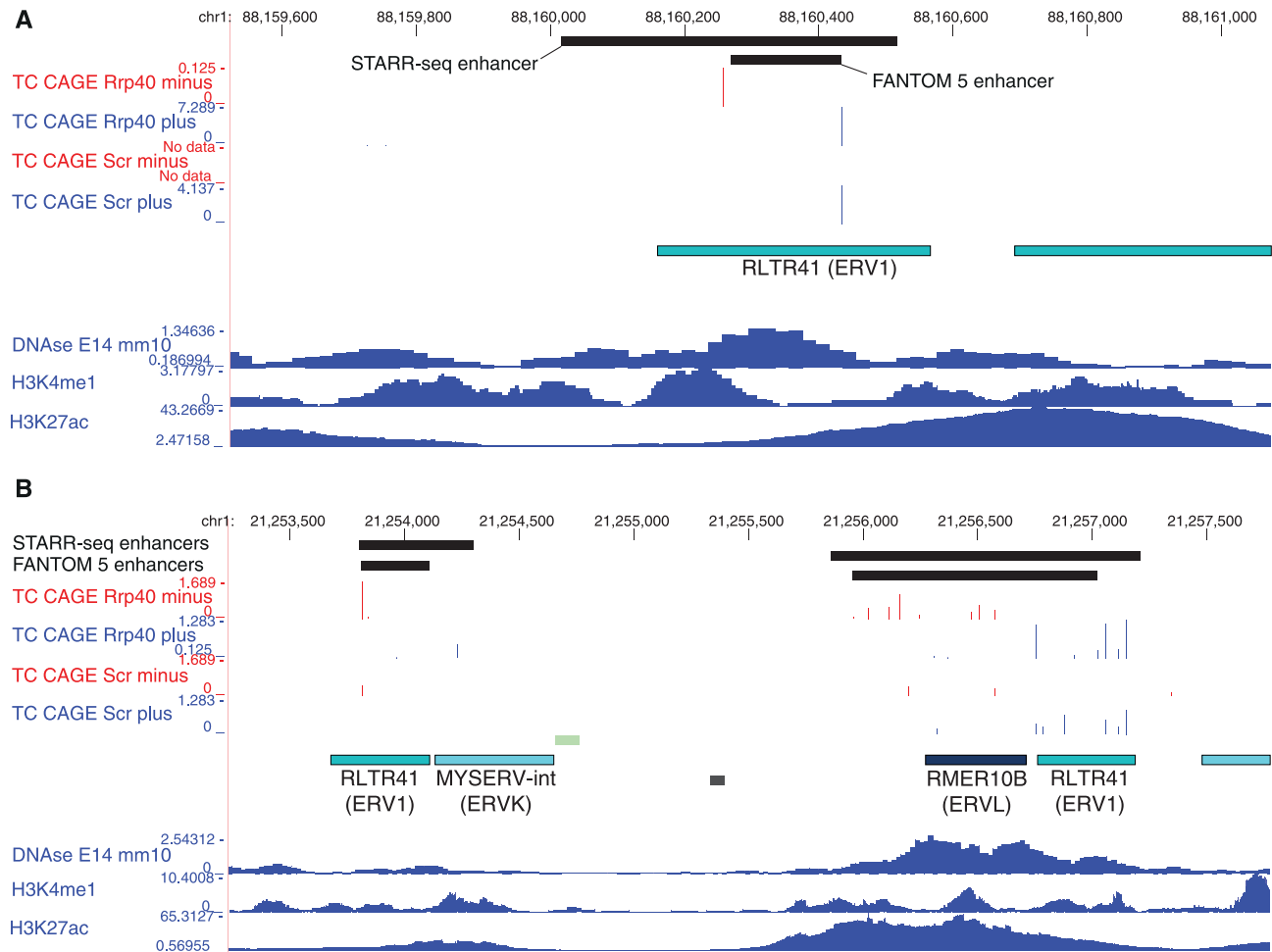


Figure 3. TE insertions co-opted as divergently transcribed enhancers. (A and B) Genome browser tracks for two intergenic loci showing TPM-normalized CAGE data pooled across replicates and split by plus (blue) and minus (red) strands. Shown are also the locations of FANTOM5 mouse enhancers and STARR-seq mESC enhancers and signal tracks for ENCODE DNase-seq data and H3K4me1 and H3K27ac ChIP-seq data for E14 mESC. 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.

We observed a strong association with LTRs (Figure 4B,C), in particular ERVs (Figure 4B,D), in these two clusters, reinforcing the association statistics with STARR-seq and FANTOM enhancers. Although cluster 3 loci were associated with many TE families, a considerable fraction of TEs belong to the ERVL, ERVL-MaLR and B4 TE families (ERVL: 871, 33.9% of clustered ERVLs; ERVL-MaLR: 4106, 37.4% of clustered ERVL-MaLRs; B4: 1597, 35.5% of clustered B4s). ERV1s and ERVLs were enriched among cluster 4 TEs (Figure 4D). In addition, transcribed STARR-seq associated DHSs displayed comparable expression levels and similar histone modifications at TE and non-TE loci (Supplementary Figure S13), suggesting that TE-derived enhancers operate at similar activity levels as non-TE enhancers.

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 (4,46). The majority of these fell close to GENCODE annotated TSSs and were enriched with SINE (B1 and proto-B1) and LTR (ERVK) elements (Figure 4B,D).

Hence, TEs are not only evolutionary co-opted into distal regulatory elements, but also gene promoters. Indeed, out of 7431 transcribed TE-associated DHSs, 1012 (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 Figure S14).

Interestingly, we found some ERV1s and ERVKs, in addition to L1 LINEs, to be associated with both H3K27ac and H3K9me3 (cluster 6). The latter is present at repressed TEs in heterochromatin (19–22) and the 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 (93). We also identified a group of transcribed TE loci (cluster 5) that was highly enriched with L1s but overall, only carried low levels of histone modifications. Some of these loci were associated with H3K27me3 signal, indicative of polycomb-mediated repression. TE loci of cluster 5 and 6

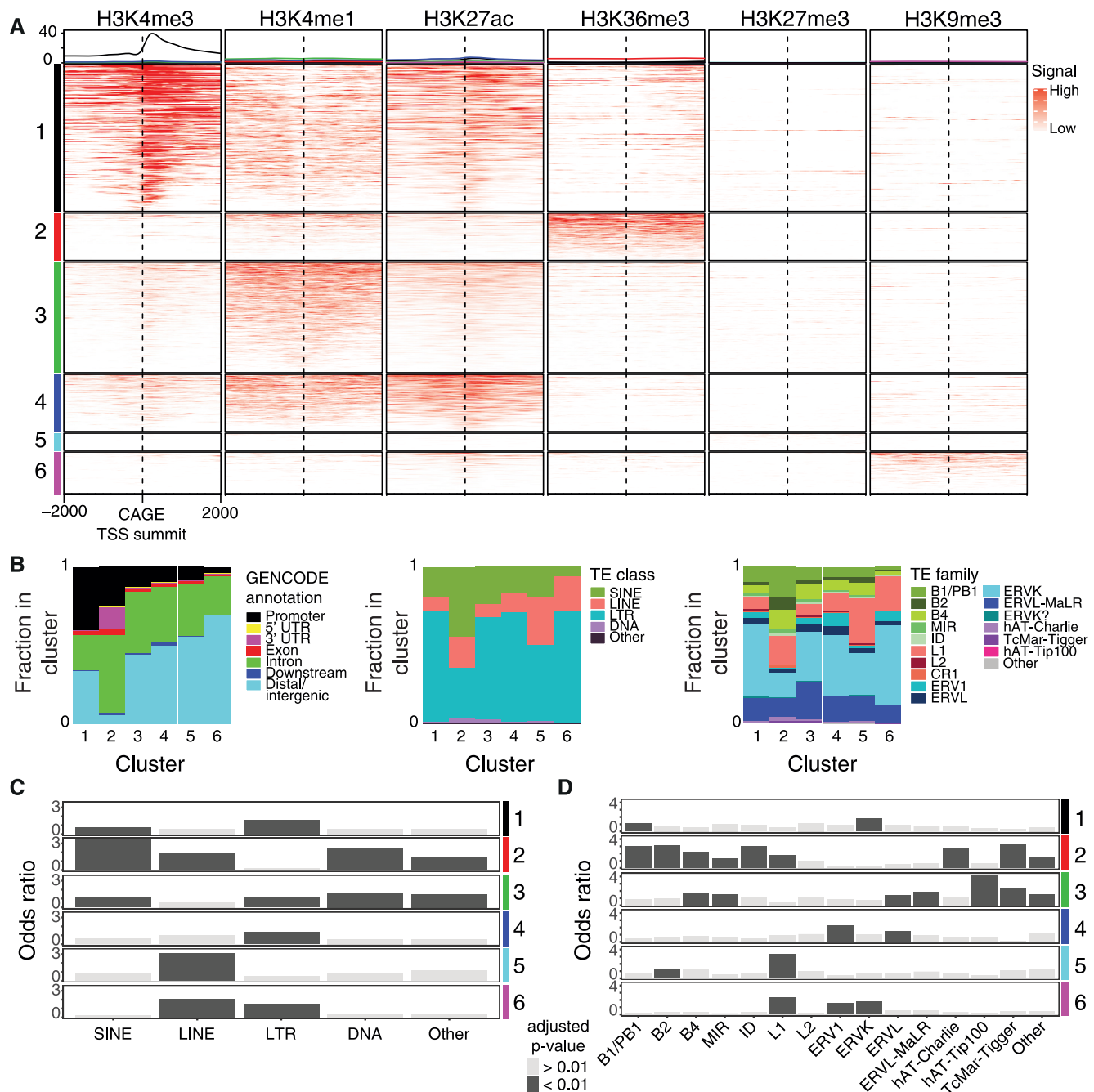


Figure 4. Transcribed TEs exhibit chromatin features of regulatory elements. (A) Hierarchical clustering of histone modification (ChIP-seq) signals ± 2000 bp around the summits of TE-associated clusters of CAGE-inferred TSSs (CAGE tag clusters, Materials and Methods). The ChIP-seq signal is shown as 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 and D) Bar plots of odds ratios of enrichments (Fisher's exact test) of TE classes (C) and TE families (D) in each cluster.

thus likely represent 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.

The TF binding repertoires and regulatory topologies of ERV subfamilies indicate involvement in 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 as regulatory elements may contribute to transcriptional regulatory programs in mESCs, we performed a TF binding site enrichment analysis of transcribed DHSs associated with TE insertions from 224 LTR

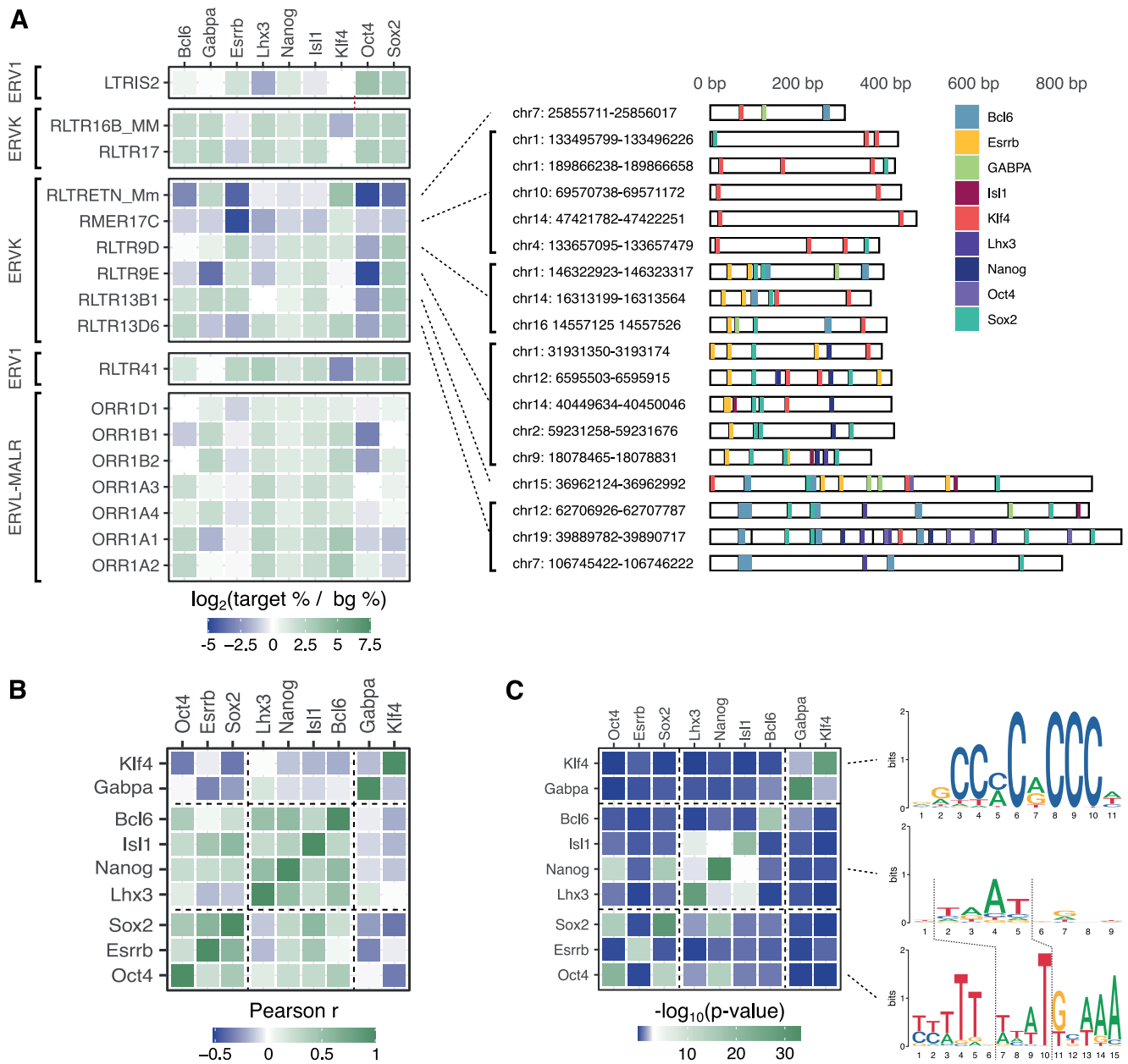


Figure 5. ERV subfamilies contribute to distinct enrichments of binding sites for pluripotency factors. (A) Motif enrichments for selected TFs (columns) in transcribed TEs across selected ERV subfamilies (rows) versus a background of non-TE genomic regions ± 200 bp around the summits of all CAGE-inferred TSS clusters. White cells indicate no enrichment and cases of complete depletion were assigned the lowest detected score, represented in dark blue according to the scale. The full TF enrichment heatmap is shown in Supplementary Figure S15A. 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 Figure S15A). The full heatmap of correlations is given in Supplementary Figure S15B. (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.

subfamilies versus all genomic loci of CAGE-inferred TSSs in mESCs (Materials and Methods section). This analysis thus reveals TFs that are enriched or depleted in TEs compared to what we would expect from regulatory elements in general. Interestingly, predicted binding sites for several TFs, including pluripotency factors Nanog, Sox2 and Oct4, were enriched in transcribed LTRs compared to non-TE associated TSSs (Benjamini–Hochberg FDR-adjusted $P < 1e-3$). Binding site sequences for 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 TSSs for Nanog, Sox2 and Oct4, respectively). These results indicate that LTRs are a major source of regulatory elements controlled by pluripotency factors in mESCs. These results are in line with those observed for human ESCs (29,36), despite the low conservation of LTRs between species (Figure 2B) (34).

Given their strong association with regulatory elements, we focused on putative TF binding sites in DHSs of transcribed TEs across ERV subfamilies. In fact, we observe a variability in enrichment of various TF binding site sequences in specific ERV subfamilies indicating diverse regulatory potentials of TEs and specific TF-TE-associations (Figure 5A, Supplementary Figure S15A and Table S2). We observed a positive correlation between enrichments of binding site sequences for pluripotency factors, e.g., Sox2 and Esrrb (Pearson's $r = 0.54$), as well as Sox2 and Nanog (Pearson's $r = 0.47$; Figure 5B; Supplementary Figure S15B). Differences in enrichments for Sox2 and Nanog were, however, found for some ERV subfamilies, as seen for instance for ORR1A1 and ORR1A2 (Figure 5A), indicating that calculated motif enrichment similarities between these two TFs are not necessarily driven by motif similarities. The overall enrichments of Sox2, Nanog and Esrrb generally agreed with Isl1, Bcl6 and Oct4 (Figure 5B and Supplementary Figure S15B), 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 RLTR13B1, ORR1A1, ORR1A2 and RLTR10A). ERV-MaLRs ORR1A1 and ORR1A2 were, in addition to Klf4, also enriched with Nanog, Isl1, Bcl6, and Lhx3, but interestingly not Esrrb and Oct4 binding site sequences. This suggests that, by carrying different repertoires of TF binding sites, ERV subfamilies may contribute differently and distinctly to the pluripotency regulatory program (Supplementary Figure S15). Similarly, Lhx3 binding site sequences co-occurred with those of Esrrb in RLTR41 ERVs but were depleted from Esrrb-enriched ERVs RLTR9D and RLTR13B1. In addition, the top enriched ERVs for putative Oct4 binding sites (LTRIS2, RLTR17, RLTR16B.MM, RLTR41) were depleted from those of Klf4.

The specific enrichments or depletions of Klf4 or Oct4 binding site sequences in certain ERV subfamilies indicate that the Klf4 and Oct4 ERV-derived regulatory networks have partially evolved independently from those of Sox2 and Nanog. This is consistent with distinct binding preferences (Figure 5C) and context-dependent cooperativity between Sox2 and Oct4 (94), which is limited during pluripotency maintenance (95). We note that the binding preferences 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 (Figure 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.

In addition to the enrichment of pluripotency TFs, we identified putative ERV-derived binding sites for a broad range of TFs (Supplementary Figure S15A). 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 ERV-MaLR subfamilies ORR1E and ORR1D2, which were further depleted of putative TF binding sites for Oct4 and Sox2. This suggests that transposition of TF binding sites native to LTR retrotransposons or the subsequent birth of new TF binding sites through mutations (27) could impact several regulatory programs, including those distinct from the naive pluripotency.

To further investigate the regulatory programs contributed to by ERVs co-opted as regulatory elements, we linked each gene-distal, transcribed ERV-associated DHS with its predicted target gene from activity-by-contact (ABC) modeling (78). The resulting gene ontology enrichments of linked genes indicate a highly specific association of individual ERV subfamilies with gene regulatory programs (Figure 6A and Supplementary Figure S16), 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 (Figure 6B).

The enrichment of pluripotency factors in transcribed ERV-associated DHSs suggests that the regulatory activities of these elements have a specificity toward cell types in which these TFs are active. To investigate their putative cell-type restricted activity, we quantified their expression using CAGE data for exosome-depleted samples derived after 3 days differentiation of mESCs into embryoid bodies (EBs) (53). In agreement with their predicted role in mESCs, the expression of ERV-associated regulatory elements with inferred binding sites for pluripotency TFs (as given in Figure 5A) were generally reduced in EBs, while those not containing such sites displayed a similar expression level in mESCs and EBs (Figure 6C). Accordingly, genes whose ABC-linked ERV-associated enhancers contained binding site sequences for pluripotency factors displayed lower expression in EBs (Figure 6D). These observations reflect a reduced activity of pluripotency TFs in EBs and demonstrate that expression of ERV-associated DHSs can be used as a marker for their cell-type specificity in regulatory activity.

Taken together, our results indicate that ERV-derived regulatory elements transcribed in mESCs contribute in a specific manner to the pluripotency regulatory network through their binding sites for pluripotency TFs. Although we identified putative TF binding sites for multiple TFs for each ERV subfamily, the TF enrichments across ERV subfamilies were highly distinct (Figure 5 and Supplementary Figure S15A). Together with the diversity of functions of predicted target genes (Supplementary Figure S16), this suggests that each ERV subfamily contributes to distinct regulatory programs and pathways.

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 ac-

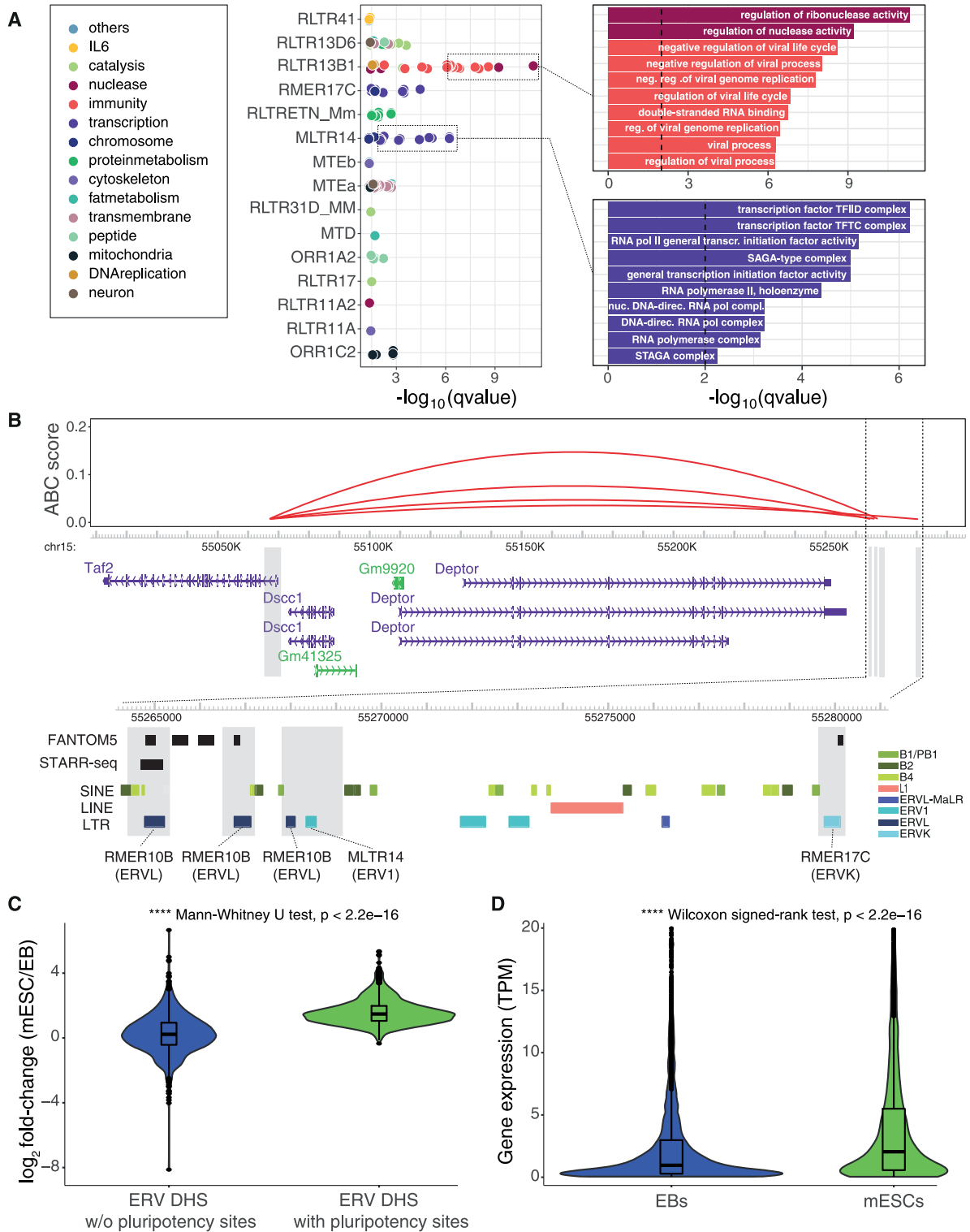


Figure 6. ERV subfamilies contribute to distinct gene regulatory programs. (A) GO term enrichment for putative target genes (ABC) of gene distal ERVs split by ERV subfamily (foreground) versus all ABC-predicted target genes (background). For ease of visualization, gene ontology terms are colored by manually curated process or function, and the underlying gene ontology term enrichments are shown for RLTR13B1 and MLTR14. Full results are provided in Supplementary Figure S16. (B) Predicted enhancer interactions with the promoter of gene *Taf2* (TATA-Box Binding Protein Associated Factor 2). The four enhancers are marked by gray 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. (C) Expression fold-change (\log_2), as measured by CAGE, in mESCs versus EBs at ERV-associated DHSs that carry (right) or not (left) predicted binding sites for pluripotency factors. (D) Gene-level expression (TPM normalized), as measured by CAGE, quantified for ABC-linked genes of transcribed TE-associated DHSs that carry predicted binding sites for pluripotency factors.

tivities are highly specialized and conserved TFs, including Oct4, Sox2 and Nanog, which, through targeted transcriptional regulation, maintain a naive pluripotent state. While gene regulation in ESCs by these TFs is highly conserved across Metazoa, their respective binding sites are not. TEs have been suggested to contribute to stabilizing the gene regulatory functions of TFs by providing regulatory sequences with the required binding sites.

We here demonstrate that retrotransposons contribute to a sizable fraction of such regulatory innovation. Using an accurate and unbiased approach based on genome-wide profiling of TSSs, we systematically investigate the regulatory potential and transcriptional activities of the wide repertoire of mouse TEs in 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 exosome for degradation. As many as a third of divergently transcribed TE-associated DHSs are supported by *in vitro* enhancer potential derived from STARR-seq data or by overlap with, rigorously validated, previously predicted enhancer sets from FANTOM. In addition, we find a large overlap with annotated gene promoters, demonstrating that at least a half of transcribed TE-associated DHSs in mESCs are regulatory elements (enhancers or promoters). Transcription start site profiling by CAGE thus lends itself as an accurate approach for genome-wide surveys of the regulatory activity of TEs, for which predictions based on open chromatin and histone modifications yield considerably lower validation rates (29,31). We observed that a substantial fraction of transcribed TE-derived regulatory elements were non-orthologous to rat or human genomes, demonstrating a high degree of regulatory innovation by TEs. Our TSS-centric approach thus allows for an unbiased, systematic investigation of the regulatory potential across TE subfamilies, beyond a selected few subfamilies.

We note that disagreements in validation rates to a recent study (31) could further be due to difference in validation assays used. While CRISPR interference (CRISPRi), used by Todd *et al.* (31), has the potential to better reflect *in vivo* activities, the false-negative rate of CRISPRi remains a concern (96). In addition, redundant enhancers may buffer regulatory effects (97,98). Therefore, an enhancer could still have a causal role in gene regulation even though no observable effect can be measured by CRISPRi. As such, *in vitro* measurements, e.g., derived from STARR-seq, are therefore better suited to reveal the enhancer potential of a regulatory element, even though such an approach may suggest enhancers that are not active *in vivo*. Further studies are necessary to properly compare the quantified activities of regulatory elements between STARR-seq and CRISPRi.

Endogenous retroviral elements were 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 sites (23,27) and binding site sequences (Figure 5A) for TFs regulating naive pluripotency, including Oct4, Sox2 and Nanog. However, we do acknowledge

that mapping of sequencing reads might skew our focus to evolutionary older ERV subfamilies (99), which have accumulated more mutations than younger ones and therefore have a higher chance of mapping reads uniquely (59). However, any mapping bias cannot explain differential TF enrichments across ERV subfamilies or their cell-type specific expression patterns.

The diverse TF enrichments observed across ERV subfamilies indicate a major contribution of ERVs to the landscape of regulatory elements and 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 site sequences for Nanog, Sox2, and Esrrb, while Oct4 and Klf4 were depleted in some subfamilies. While some of these TFs have similar motifs, e.g., Nanog, Sox2, and may therefore overestimate co-occurrences, the specific enrichments and depletions 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 ERVs predicted to be co-opted as gene-distal enhancers. Further, our data suggest that the regulatory landscape of naive pluripotency in mESCs has been shaped independently by multiple subfamilies of ERVs.

In agreement with previous work (23,29,36), we observed a substantial enrichment of binding site sequences for pluripotency factors in ERV-associated regulatory elements. However, our transcription-centric approach identified many ERV subfamilies with enrichment for Nanog or Oct4 binding site sequences not revealed when focusing on their binding sites as inferred from ChIP-seq (36), including ORR1A1, ORR1A2, RMER17B, LTRIS2, RLTR9E, RMER19B for Nanog and LTRIS2, RLTR16B_MM, MLTR25A, RLTR23, BGLII_B, RMER1B for Oct4. While we acknowledge that such differences can be due to false positives from motif scanning (100), by focusing on sites of transcription initiation we may better capture the regulatory active subset of TF binding events (101). We speculate that utilizing transcription initiation profiling to infer the regulatory activity of TEs specifically identifies TE-derived regulatory elements for which relevant DNA sequences, e.g., binding sites for pluripotency factors, are linked with cell-type specific transcriptional and regulatory activity. This is confirmed by seemingly cell-type restricted expression of pluripotency-associated ERVs and that of their putative target genes, when expression is compared between mESCs and EBs (Figure 6C,D).

In addition to the core pluripotency TFs, we observed an enrichment of binding site sequences 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 (27,28). 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 (102).

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. We further show that expression of TE-associated open chromatin regions is indicative of cell-type restricted regulatory activity. Charting their dynamic activities over development will thus be an important next step to further our understanding of the cell-type specific roles of TEs in transcriptional regulation.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank members of the Andersson lab for rewarding comments and discussions.

Author Contributions: R.A. and S.B conceived the project. S.B., R.K., N.A. and M.S. performed analyses. S.B., R.K. and R.A interpreted results and wrote the manuscript. All authors reviewed the final manuscript.

FUNDING

Danish Council for Independent Research [6108–00038]; European Research Council (ERC) [638173]; Novo Nordisk Foundation [NNF18OC0052570, NNF20OC0059796]. Funding for open access charge: Novo Nordisk Foundation [NNF20OC0059796].

Conflict of interest statement. None declared.

REFERENCES

- Haberle, V. and Stark, A. (2018) Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell Biol.*, **19**, 621–637.
- Beagrie, R.A. and Pombo, A. (2016) Gene activation by metazoan enhancers: diverse mechanisms stimulate distinct steps of transcription. *Bioessays*, **38**, 881–893.
- Shlyueva, D., Stampfel, G. and Stark, A. (2014) Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.*, **15**, 272–286.
- Andersson, R. and Sandelin, A. (2020) Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.*, **21**, 71–87.
- Nguyen, T.A., Jones, R.D., Snavely, A.R., Pfenning, A.R., Kirchner, R., Hemberg, M. and Gray, J.M. (2016) High-throughput functional comparison of promoter and enhancer activities. *Genome Res.*, **26**, 1023–1033.
- Weingarten-Gabbay, S., Nir, R., Lubliner, S., Sharon, E., Kalma, Y., Weinberger, A. and Segal, E. (2019) Systematic interrogation of human promoters. *Genome Res.*, **29**, 171–183.
- Smith, R.P., Taher, L., Patwardhan, R.P., Kim, M.J., Inoue, F., Shendure, J., Ovcharenko, I. and Ahituv, N. (2013) Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat. Genet.*, **45**, 1021–1028.
- Grossman, S.R., Zhang, X., Wang, L., Engreitz, J., Melnikov, A., Rogov, P., Tewhey, R., Isakova, A., Deplancke, B., Bernstein, B.E. *et al.* (2017) Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, E1291–E1300.
- Nitta, K.R., Jolma, A., Yin, Y., Morgunova, E., Kivioja, T., Akhtar, J., Hens, K., Toivonen, J., Deplancke, B., Furlong, E.E.M. *et al.* (2015) Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife*, **4**, e04837.
- Chan, E.T., Quon, G.T., Chua, G., Babak, T., Trochesset, M., Zirngibl, R.A., Aubin, J., Clarcliffe, M.J., Wilde, A., Brudno, M. *et al.* (2009) Conservation of core gene expression in vertebrate tissues. *Journal of Biology*, **8**, 33.
- Berthelot, C., Villar, D., Horvath, J.E., Odom, D.T. and 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.
- Odom, D.T., Dowell, R.D., Jacobsen, E.S., Gordon, W., Danford, T.W., MacIsaac, K.D., Rolfe, P.A., Conboy, C.M., Gifford, D.K. and Fraenkel, E. (2007) Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat. Genet.*, **39**, 730–732.
- Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C.P., Mackay, S. *et al.* (2010) Five-vertebrate chip-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, **328**, 1036–1040.
- Vierstra, J., Rynes, E., Sandstrom, R., Zhang, M., Canfield, T., Hansen, R.S., Stehling-Sun, S., Sabo, P.J., Byron, R., Humbert, R. *et al.* (2014) Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science*, **346**, 1007–1012.
- Young, R.S., Hayashizaki, Y., Andersson, R., Sandelin, A., Kawaji, H., Itoh, M., Lassmann, T., Carninci, P., The FANTOM Consortium, Bickmore, W.A. *et al.* (2015) The frequent evolutionary birth and death of functional promoters in mouse and human. *Genome Res.*, **25**, 1546–1557.
- Villar, D., Berthelot, C., Aldridge, S., Rayner, T.F., Lukk, M., Pignatelli, M., Park, T.J., Deaville, R., Erichsen, J.T., Jasinska, A.J. *et al.* (2015) Enhancer evolution across 20 mammalian species. *Cell*, **160**, 554–566.
- Chuong, E.B., Elde, N.C. and Feschotte, C. (2017) Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.*, **18**, 71–86.
- 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.
- Maksakova, I.A., Thompson, P.J., Goyal, P., Jones, S.J., Singh, P.B., Karimi, M.M. and Lorincz, M.C. (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.
- Matsui, T., Leung, D., Miyashita, H., Maksakova, I.A., Miyachi, H., Kimura, H., Tachibana, M., Lorincz, M.C. and Shinkai, Y. (2010) Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. *Nature*, **464**, 927–931.
- Karimi, M.M., Goyal, P., Maksakova, I.A., Bilenky, M., Leung, D., Tang, J.X., Shinkai, Y., Mager, D.L., 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.
- Rowe, H.M., Kapopoulou, A., Corsinotti, A., Fasching, L., Macfarlan, T.S., Tarabay, Y., Vville, 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.
- Bourque, G., Leong, B., Vega, V.B., Chen, X., Lee, Y.L., Srinivasan, K.G., Chew, J.-L., Ruan, Y., Wei, C.-L., Ng, H.H. *et al.* (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.*, **18**, 1752–1762.
- Chuong, E.B., Rumi, M.A.K., Soares, M.J. and Baker, J.C. (2013) Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nat. Genet.*, **45**, 325–329.
- Sundaram, V., Cheng, Y., Ma, Z., Li, D., Xing, X., Edge, P., Snyder, M.P. and Wang, T. (2014) Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res.*, **24**, 1963–1976.
- Trizzino, M., Park, Y., Holsbach-Beltrame, M., Aracena, K., Mika, K., Caliskan, M., Perry, G.H., Lynch, V.J. and Brown, C.D. (2017) Transposable elements are the primary source of novelty in primate gene regulation. *Genome Res.*, **27**, 1623–1633.

27. Sundaram,V., Choudhary,M.N.K., 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. *Nat. Commun.*, **8**, 14550.
28. Sun,X., Wang,X., Tang,Z., Grivainis,M., Kahler,D., Yun,C., Mita,P., Fenyö,D. and Boeke,J.D. (2018) Transcription factor profiling reveals molecular choreography and key regulators of human retrotransposon expression. *Proc. Natl Acad. Sci.*, **115**, E5526–E5535.
29. Barakat,T.S., Halbritter,F., Zhang,M., Rendeiro,A.F., Perenthaler,E., Bock,C. and Chambers,I. (2018) Functional dissection of the enhancer repertoire in human embryonic stem cells. *Cell Stem Cell*, **23**, 276–288.
30. 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.
31. Todd,C.D., Deniz,Ö., Taylor,D. and Branco,M.R. (2019) Functional evaluation of transposable elements as enhancers in mouse embryonic and trophoblast stem cells. *Elife*, **8**, 1473.
32. Miao,B., Fu,S., Lyu,C., Gontarz,P., Wang,T. and Zhang,B. (2020) Tissue-specific usage of transposable element-derived promoters in mouse development. *Genome Biol.*, **21**, 255.
33. Jacques,P.-É., Jeyakani,J. and Bourque,G. (2013) The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genet.*, **9**, e1003504.
34. Glinsky,G. and Barakat,T.S. (2019) The evolution of great apes has shaped the functional enhancers' landscape in human embryonic stem cells. *Stem Cell Res.*, **37**, 101456.
35. Fort,A., Hashimoto,K., Yamada,D., Salimullah,M., Keya,C.A., 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.
36. Kunarso,G., Chia,N.-Y., Jeyakani,J., Hwang,C., Lu,X., Chan,Y.-S., Ng,H.H. and Bourque,G. (2010) Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.*, **42**, 631–634.
37. Kheradpour,P., Ernst,J., Melnikov,A., Rogov,P., Wang,L., Zhang,X., Alston,J., Mikkelsen,T.S. and Kellis,M. (2013) Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.*, **23**, 800–811.
38. Kim,T.-K., Hemberg,M., Gray,J.M., Costa,A.M., Bear,D.M., Wu,J., Harmin,D.A., Laptewicz,M., Barbara-Haley,K., Kuersten,S. *et al.* (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature*, **465**, 182–187.
39. Andersson,R., Refsing Andersen,P., Valen,E., Core,L.J., Bornholdt,J., Boyd,M., Heick Jensen,T. and Sandelin,A. (2014) Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nat Commun.*, **5**, 5336.
40. Andersson,R., Gebhard,C., Miguel-Escalada,I., Hoof,I., Bornholdt,J., Boyd,M., Chen,Y., Zhao,X., Schmidl,C., Suzuki,T. *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature*, **507**, 455–461.
41. Andersson,R., Chen,Y., Core,L., Lis,J.T., Sandelin,A. and Jensen,T.H. (2015) Human gene promoters are intrinsically bidirectional. *Mol. Cell*, **60**, 346–347.
42. Scruggs,B.S., Gilchrist,D.A., Nechaev,S., Muse,G.W., Burkholder,A., Fargo,D.C. and Adelman,K. (2015) Bidirectional transcription arises from two distinct hubs of transcription factor binding and active chromatin. *Mol. Cell*, **58**, 1101–1112.
43. Chen,Y., Pai,A.A., Herudek,J., Lubas,M., Meola,N., Järvelin,A.I., Andersson,R., Pelechano,V., Steinmetz,L.M., Jensen,T.H. *et al.* (2016) Principles for RNA metabolism and alternative transcription initiation within closely spaced promoters. *Nat. Genet.*, **48**, 984–994.
44. Rennie,S., Dalby,M., Lloret-Llinares,M., Bakoulis,S., Dalager Vaagenso,C., Heick Jensen,T. and 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.
45. Henriques,T., Scruggs,B.S., Inouye,M.O., Muse,G.W., Williams,L.H., Burkholder,A.B., Lavender,C.A., Fargo,D.C. and Adelman,K. (2018) Widespread transcriptional pausing and elongation control at enhancers. *Genes Dev.*, **32**, 26–41.
46. Core,L.J., Martins,A.L., Danko,C.G., Waters,C.T., Siepel,A. and Lis,J.T. (2014) Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.*, **46**, 1311–1320.
47. Wu,H., Nord,A.S., Akiyama,J.A., Shoukry,M., Afzal,V., Rubin,E.M., Pennacchio,L.A. and Visel,A. (2014) Tissue-Specific RNA expression marks distant-acting developmental enhancers. *PLoS Genet.*, **10**, e1004610.
48. Koch,F., Fenouil,R., Gut,M., Cauchy,P., Albert,T.K., Zacarias-Cabeza,J., Spicuglia,S., de la Chapelle,A.L., 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.
49. Takahashi,H., Lassmann,T., Murata,M. and Carninci,P. (2012) 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat. Protoc.*, **7**, 542–561.
50. Kawaji,H., Lizio,M., Itoh,M., Kanamori-Katayama,M., Kaiho,A., Nishiyori-Sueki,H., Shin,J.W., 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.
51. Preker,P., Nielsen,J., Kammler,S., Lykke-Andersen,S., Christensen,M.S., Mapendano,C.K., Schierup,M.H. and Jensen,T.H. (2008) RNA exosome depletion reveals transcription upstream of active human promoters. *Science*, **322**, 1851–1854.
52. Ntini,E., Järvelin,A.I., Bornholdt,J., Chen,Y., Boyd,M., Jørgensen,M., Andersson,R., Hoof,I., Schein,A., Andersen,P.R. *et al.* (2013) Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. *Nat. Struct. Mol. Biol.*, **20**, 923–928.
53. Lloret-Llinares,M., Karadoulama,E., Chen,Y., Wojenski,L.A., Villafano,G.J., Bornholdt,J., Andersson,R., Core,L., Sandelin,A. and Jensen,T.H. (2018) The RNA exosome contributes to gene expression regulation during stem cell differentiation. *Nucleic Acids Res.*, **4**, a008128.
54. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, **25**, 1754–1760.
55. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
56. Kerpedjiev,P., Frellsen,J., Lindgreen,S. and Krogh,A. (2014) Adaptable probabilistic mapping of short reads using position specific scoring matrices. *BMC Bioinf.*, **15**, 100.
57. Kielbasa,S.M., Wan,R., Sato,K., Horton,P. and Frith,M.C. (2011) Adaptive seeds tame genomic sequence comparison. *Genome Res.*, **21**, 487–493.
58. Hashimoto,T., de Hoon,M.J.L., Grimmond,S.M., Daub,C.O., Hayashizaki,Y. and Faulkner,G.J. (2009) Probabilistic resolution of multi-mapping reads in massively parallel sequencing data using mumrescue-lite. *Bioinformatics*, **25**, 2613–2614.
59. Jin,Y., Tam,O.H., Paniagua,E. and Hammell,M. (2015) TETranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics*, **31**, 3593–3599.
60. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
61. Ramírez,F., Ryan,D.P., Grüning,B., Bhardwaj,V., Kilpert,F., Richter,A.S., Heyne,S., Dündar,F. and Manke,T. (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.*, **44**, W160–W165.
62. Andersen,P.R., Domanski,M., Kristiansen,M.S., 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.
63. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
64. Kim,D., Langmead,B. and Salzberg,S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.

65. Liao, Y., Smyth, G.K. and Shi, W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
66. Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J. *et al.* (2019) GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.*, **47**, D766–D773.
67. Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F. *et al.* (2006) The UCSC genome browser database: update 2006. *Nucleic Acids Res.*, **34**, D590–D598.
68. Dalby, M., Rennie, S. and Andersson, R. (2018) FANTOM5 transcribed enhancers in mm10. *Zenodo*, <https://doi.org/10.5281/zenodo.1411211>.
69. Peng, T., Zhai, Y., Atlasi, Y., ter Huurne, M., Marks, H., Stunnenberg, H.G. and Megchelenbrink, W. (2020) STARR-seq identifies active, chromatin-masked, and dormant enhancers in pluripotent mouse embryonic stem cells. *Genome Biol.*, **21**, 243.
70. Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with bowtie 2. *Nat. Methods*, **9**, 357–359.
71. Lee, D., Shi, M., Moran, J., Wall, M., Zhang, J., Liu, J., Fitzgerald, D., Kyono, Y., Ma, L., White, K.P. *et al.* (2020) STARRPeaker: uniform processing and accurate identification of STARR-seq active regions. *Genome Biol.*, **21**, 298.
72. Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.*, **17**, 10.
73. Yu, G., Wang, L.-G. and He, Q.-Y. (2015) ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, **31**, 2382–2383.
74. Pintacuda, G., Wei, G., Roustan, C., Kirmizitas, B.A., Solcan, N., Cerase, A., Castello, A., Mohammed, S., Moindrot, B., Nesterova, T.B. *et al.* (2017) hnRNP K recruits PCGF3/5-PRC1 to the xist RNA B-Repeat to establish polycomb-mediated chromosomal silencing. *Mol. Cell*, **68**, 955–969.
75. Ernst, J. and Kellis, M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.
76. Ernst, J. and Kellis, M. (2017) Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.*, **12**, 2478–2492.
77. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H. and Glass, C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Mol. Cell*, **38**, 576–589.
78. Fulco, C.P., Nasser, J., Jones, T.R., Munson, G., Bergman, D.T., Subramanian, V., Grossman, S.R., Anyoha, R., Doughty, B.R., Patwardhan, T.A. *et al.* (2019) Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.*, **51**, 1664–1669.
79. Yu, G., Wang, L.-G., Han, Y. and He, Q.-Y. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*, **16**, 284–287.
80. Rhead, B., Karolchik, D., Kuhn, R.M., Hinrichs, A.S., Zweig, A.S., Fujita, P.A., Diekhans, M., Smith, K.E., Rosenbloom, K.R., Raney, B.J. *et al.* (2010) The UCSC genome browser database: update 2010. *Nucleic Acids Res.*, **38**, D613–D619.
81. Andersson, R., Sandelin, A. and Danko, C.G. (2015) A unified architecture of transcriptional regulatory elements. *Trends Genet.*, **31**, 426–433.
82. Mikhaylichenko, O., Bondarenko, V., Harnett, D., Schor, I.E., Males, M., Viales, R.R. and Furlong, E.E.M. (2018) The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. *Genes Dev.*, **32**, 42–57.
83. Faulkner, G.J., Forrest, A.R.R., Chalk, A.M., Schroder, K., Hayashizaki, Y., Carninci, P., Hume, D.A. and Grimmond, S.M. (2008) A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics*, **91**, 281–288.
84. Dewannieux, M., Dupressoir, A., Harper, F., Pierron, G. and Heidmann, T. (2004) Identification of autonomous IAP LTR retrotransposons mobile in mammalian cells. *Nat. Genet.*, **36**, 534–539.
85. Brady, T., Lee, Y.N., Ronen, K., Malani, N., Berry, C.C., Bieniasz, P.D. and Bushman, F.D. (2009) Integration target site selection by a resurrected human endogenous retrovirus. *Genes Dev.*, **23**, 633–642.
86. Arner, E., Daub, C.O., Vitting-Seerup, K., Andersson, R., Lilje, B., Drablos, F., Lennartsson, A., Rönnerblad, M., Hrydziszko, O., Vitezic, M. *et al.* (2015) Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science*, **347**, 1010–1014.
87. Halfon, M.S. (2019) Studying transcriptional enhancers: the founder fallacy, validation creep, and other biases. *Trends Genet.*, **35**, 93–103.
88. The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
89. Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B.D. *et al.* (2014) A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, **515**, 355–364.
90. Creighton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A. *et al.* (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. USA.*, **107**, 21931–21936.
91. Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A. *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311–318.
92. Robertson, A.G., Bilenky, M., Tam, A., Zhao, Y., Zeng, T., Thiessen, N., Cezard, T., Fejes, A.P., Wederell, E.D., 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.
93. He, J., Fu, X., Zhang, M., He, F., Li, W., Abdul, M.M., 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. *Nat. Commun.*, **10**, 34.
94. Li, S., Zheng, E.B., Zhao, L. and Liu, S. (2019) Nonreciprocal and conditional cooperativity directs the pioneer activity of pluripotency transcription factors. *Cell Rep.*, **28**, 2689–2703.
95. Malik, V., Glaser, L.V., 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. *Nat. Commun.*, **10**, 3477.
96. Gasperini, M., Tome, J.M. and Shendure, J. (2020) Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat. Rev. Genet.*, **21**, 292–310.
97. Hong, J.W., Hendrix, D.A. and Levine, M.S. (2008) Shadow enhancers as a source of evolutionary novelty. *Science*, **321**, 1314–1314.
98. Cannavò, E., Khoeiry, P., Garfield, D.A., Geleher, P., Zichner, T., Gustafson, E.H., Ciglar, L., Korbel, J.O. and Furlong, E.E.M. (2016) Shadow enhancers are pervasive features of developmental regulatory networks. *Curr. Biol.*, **26**, 38–51.
99. Simonti, C.N., Pavličev, M. and Capra, J.A. (2017) Transposable element exaptation into regulatory regions is rare, influenced by evolutionary age, and subject to pleiotropic constraints. *Mol. Biol. Evol.*, **34**, 2856–2869.
100. Wasserman, W.W. and Sandelin, A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.
101. Azofeifa, J.G., Allen, M.A., Hendrix, J.R., Read, T., Rubin, J.D. and Dowell, R.D. (2018) Enhancer RNA profiling predicts transcription factor activity. *Genome Res.*, **28**, 334–344.
102. Singh, G., Mullany, S., Moorthy, S.D., Zhang, R., Mehdi, T., Tian, R., Duncan, A.G., Moses, A.M. and Mitchell, J.A. (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.