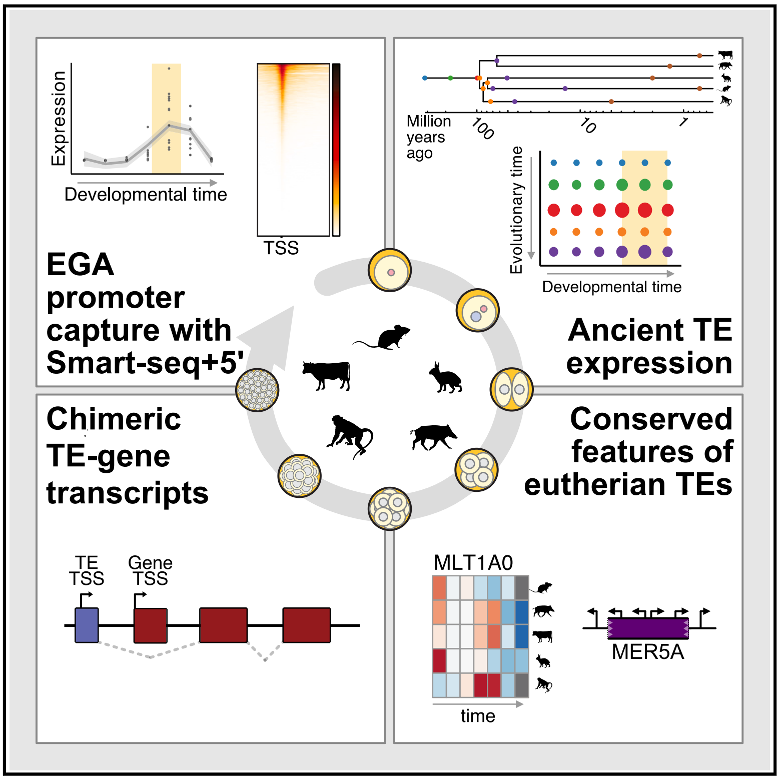
Cell

Resource

An atlas of transcription initiation reveals regulatory principles of gene and transposable element expression in early mammalian development

Graphical abstract



Highlights

- Mapping transcription start site reveals genome regulatory principles in early embryos
- Transposable elements exhibit convergent and divergent patterns across species
- LTRs, SINEs, LINEs, and DNA transposons drive chimeric transcripts in early embryos
- Ancient transposable elements are transcribed during mammalian development

Authors

Marlies E. Oomen,
Diego Rodriguez-Terrones,
Mayuko Kurome, ..., Eckhard Wolf,
Henrik Kaessmann,
Maria-Elena Torres-Padilla

Correspondence

torres-padilla@helmholtz-muenchen.de

In brief

Mapping transcription start sites across five mammalian species before, during, and after embryonic genome activation unveils widespread transposable element-driven transcription and co-option of evolutionary old elements for genome regulation.

Oomen et al., 2025, Cell *188*, 1156–1174
February 20, 2025 © 2024 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.cell.2024.12.013

CellPress

CellPress
OPEN ACCESS

Cell

Resource

An atlas of transcription initiation reveals regulatory principles of gene and transposable element expression in early mammalian development

Marlies E. Oomen,[1,9] Diego Rodriguez-Terrones,[1,9] Mayuko Kurome,[2] Valeri Zakhartchenko,[2] Lorenza Mottes,[1] Kilian Simmet,[2] Camille Noll,[1] Tsunetoshi Nakatani,[1] Carlos Michel Mourra-Diaz,[1] Irene Aksoy,[3] Pierre Savatier,[3,4] Jonathan Göke,[5,6] Eckhard Wolf,[2] Henrik Kaessmann,[7] and Maria-Elena Torres-Padilla[1,8,10,*]

[1]Institute of Epigenetics and Stem Cells, Helmholtz Munich, Munich, Germany
[2]Genzentrum, Ludwig-Maximilians-Universität, Munich, Germany
[3]Université Lyon 1, INSERM U1208, INRAE USC 1361, 69500 Bron, France
[4]Platform PrimaStem, INSERM U1208, INRAE USC 1361, 69500 Bron, France
[5]Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore, Singapore
[6]Department of Statistics and Data Science, National University of Singapore, Singapore, Singapore
[7]Center for Molecular Biology of Heidelberg University (ZMBH), DKFZ-ZMBH Alliance, Heidelberg, Germany
[8]Faculty of Biology, Ludwig-Maximilians Universität, Munich, Germany
[9]These authors contributed equally
[10]Lead contact
*Correspondence: torres-padilla@helmholtz-muenchen.de

https://doi.org/10.1016/j.cell.2024.12.013

SUMMARY

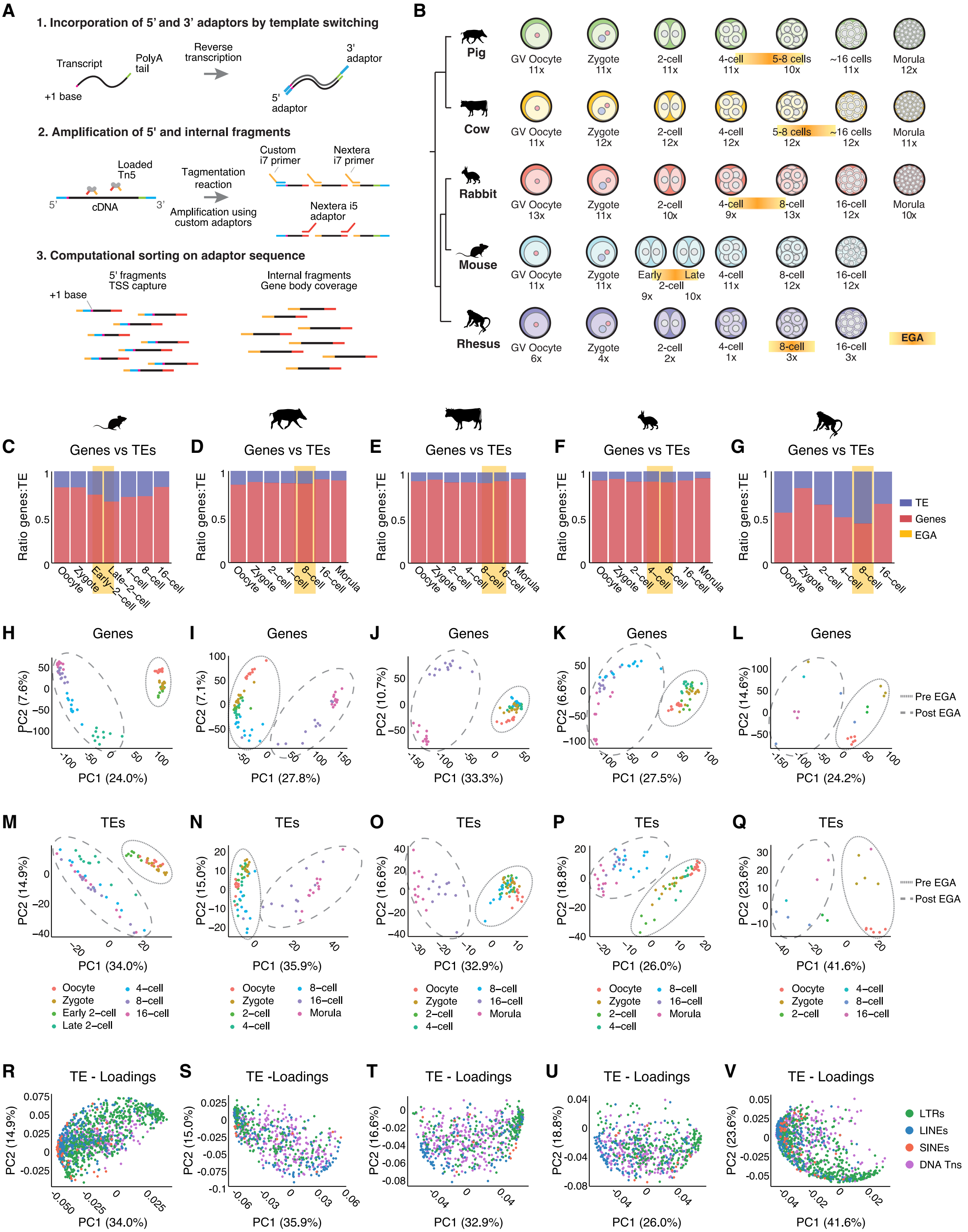
Transcriptional activation of the embryonic genome (EGA) is a major developmental landmark enabling the embryo to become independent from maternal control. The magnitude and control of transcriptional reprogramming during this event across mammals remains poorly understood. Here, we developed Smart-seq+5′ for high sensitivity, full-length transcript coverage and simultaneous capture of 5′ transcript information from single cells and single embryos. Using Smart-seq+5′, we profiled 34 developmental stages in 5 mammalian species and provide an extensive characterization of the transcriptional repertoire of early development before, during, and after EGA. We demonstrate widespread transposable element (TE)-driven transcription across species, including, remarkably, of DNA transposons. We identify 19,657 TE-driven genic transcripts, suggesting extensive TE co-option in early development over evolutionary timescales. TEs display similar expression dynamics across species and species-specific patterns, suggesting shared and divergent regulation. Our work provides a powerful resource for understanding transcriptional regulation of mammalian development.

INTRODUCTION

A hallmark of preimplantation development is embryonic genome activation (EGA), during which the embryo transitions from inherited maternal transcripts to genes transcribed from its own genome.[1,2] EGA coincides with extensive reprogramming of both parental chromatins,[3] as histone modifications are reestablished[4,5] and transcription factors (TFs) regain binding.[6] However, how the exquisite control of transcriptional regulation of thousands of genes at this precise time is achieved, is unknown.

Around half of the mouse and human genome is composed of transposable elements (TEs) and their remnants,[7] since many TEs have become fragmented over evolutionary timescales. TEs have been shown to be part of the embryonic transcriptome at EGA in some species.[7–11] Similar to genes, TE sequences contain *cis*-regulatory elements that enable recruitment of TFs and chromatin remodelers, which in turn regulate and initiate transcriptional activity.[12,13] Intriguingly, it has been shown for few individual TEs that transcriptional activation is not merely a side effect of heterochromatin remodeling after fertilization but that TEs play a role in early development.[14,15] These include the mouse ERVL (mERVL) LTR MT2_Mm,[9,16,17] which shows similar expression pattern and regulatory sequences as human ERVL (hERVL),[8,18] and LINE L1,[11] which regulate global chromatin accessibility in early mouse embryos.[10] However, a comprehensive study on the role and dynamics of TE expression across mammals enabling evolutionary investigations is lacking.

A significant obstacle in our understanding of TE transcription is the technical challenge to robustly differentiate transcription initiated within the TE as opposed to transcriptional initiation at neighboring genes. This has been particularly limiting in low-input



(legend on next page)

RNA sequencing (RNA-seq) techniques that do not provide information on transcription initiation.[19] To circumvent this, we developed a method based on Smart-seq2,[20,21] Smart-seq+5′, which allows robust profiling of 5′ ends of transcripts as well as of internal fragments. Smart-seq+5′ builds upon Smart-seq2 sensitivity and incorporates an alternative tagmentation strategy for the identification of transcription start sites (TSSs). Using this method, we undertook an evolutionary approach to profile and systematically characterize EGA in five mammalian species: mouse, rhesus, rabbit, cow, and pig. We report on their transcriptional genes and TE repertoire, which indicates evolutionary conserved patterns of expression and transcriptional regulation of TEs in mammalian preimplantation development. Our work constitutes an outstanding resource for the systematic investigation of the principles underlying transcriptional regulation at large during early mammalian development.

RESULTS

SMART-seq+5′ allows capture of 5′ end transcript information and reliable TE quantification

A challenge when studying TE expression and their regulation is to distinguish transcription events initiated from TE sequences themselves as opposed to read-through transcription driven by neighbouring host genes.[19,22] Mapping TSSs—and hence promoters—using approaches like cap analysis of gene expression (CAGE) based on capture of capped RNAs, indicative of transcript start,[23] can overcome this problem. Here, we devised a technique to quantify TE expression from single embryos, Smart-seq+5′, which allows for the detection and quantification of transcripts originating from TEs based on TSS information, as well as transcript body coverage. By revealing the 5′ end of transcripts, Smart-seq+5′ also provides information on TSS usage of single-copy genes. Briefly, we modified the Smart-seq2 protocol, which generates full-length transcript coverage based on polyA selection and template switching,[20,21] by incorporating molecular crowding[24] for improved reverse transcription efficiency and modified the tagmentation-based library preparation to enable identification of the 5′ transcript end (Figure 1A). To capture and differentiate 5′ fragments and internal fragments, we employed sequencing adaptors complementary to the Smart-seq2 adaptors to enable sequencing of the terminal ends in addition to the internal fragments generated during the standard tagmentation reaction (Figure 1A). We refer to this method as Smart-seq+5′ (see also STAR Methods). We validated the accuracy of Smart-seq+5′ for 5′ end mapping and therefore TSS identification.[23] First, we mapped internal and 5′ reads onto ERCCs. This confirmed that the first base pair (bp) of the 5′ reads corresponds to the synthetically defined 5′ coordinate of each ERCC molecule, contrary to the internal reads (Figures S1A and S1B). Second, we verified that Smart-seq+5′ recovers known TSSs by aligning 5′ reads over all annotated TSSs in refTSS ($n = 97{,}682$) (Figures S1C–S1E), compared to internal fragments, which span along the gene body (Figure S1C). Third, we generated Smart-seq+5′ of mouse embryonic stem cells (ESCs) and compared it with published CAGE data.[25] Mapping 5′ reads obtained using Smart-seq+5′ indicates a high concordance with TSSs captured using CAGE (Figure S1F). Fourth, we asked if Smart-seq+5′ captures well-studied mouse TEs and find indeed that 5′ reads map effectively to known TSSs of young LTRs (MT2_Mm), LINEs (L1 Mur2), and SINEs (B1 Mus1) (Figure S1G). Lastly, and importantly, Smart-seq+5′ produces results as robust as Smart-seq2, as validated against a published dataset[26] (Figure S1H). Thus, our protocol retains the high sensitivity, full-length transcript coverage and throughput of Smart-seq2 but also allows for simultaneous identification of TSSs, and hence promoters, which otherwise requires the use of additional techniques such as CAGE.[23,27,28]

Genes and TEs are expressed in a stage-specific manner throughout mammalian preimplantation development

We applied Smart-seq+5′ to 332 single preimplantation embryos from five eutherian species: mouse (*Mus musculus*), pig (*Sus scrofa*), cow (*Bos taurus*), rabbit (*Oryctolagus cuniculus*), and rhesus macaque (*Macaca mulatta*) (Figure 1B). These species were chosen because (1) they cover a substantial evolutionary time within the mammalian clade, (2) their genomes are relatively well annotated, and (3) they are common model systems for mammalian development. EGA occurs at different stages in these species,[2,29–32] and thus we collected embryos to cover the window before, during, and after EGA. Smart-seq+5′ libraries from 34 developmental stages were generated and filtered for quality control (Table S1). For TE analyses, we used the Dfam annotation since it is the best available annotation and is used in large-scale projects such as zoonomia,[33,34] thereby enabling direct comparisons between datasets. Notably, the majority of the reads mapped uniquely across the five species (Figure S1I; Table S1) and could be assigned to either a gene or a TE (Figure S1J). A global inspection of transcript abundance from genes vs. TEs revealed the presence of TE transcripts in all species and at all developmental

Figure 1. Developmental progression through mammalian EGA is demarcated by distinct TE expression

(A) Smart-seq+5′ overview.

(B) Single-embryo Smart-seq+5′ datasets. Number of embryos collected per stage passing QC is indicated.

(C–G) Fraction of 5′ fragments counted toward genes or TEs in mouse (C), pig (D), cow (E), rabbit (F), and rhesus (G) embryos.

(H–Q) PCA of single embryos by genes (H–L) and TEs (M–Q) in mouse (H and M), pig (I and N), cow (J and O), rabbit (K and P), and rhesus (L and Q) showing separation of pre-EGA (short-dashed line) and post-EGA stages (long-dashed lines). Each dot represents a single embryo at indicated developmental stage by color.

(R–V) PCA projections corresponding to (Q) in mouse (R), pig (S), cow (T), rabbit (U), and rhesus (V) showing the contribution of TE families to PC1 and PC2. Each dot is colored to represent TE (sub)class.

In (B) and (C)–(G), EGA timing for each species is marked in yellow.

See also Figures S1 and S2.

stages examined (Figures 1C–1G). Principal-component analysis (PCA) of Smart-seq+5′ genic reads resulted in a clear separation of pre- and post-EGA stages along the first principal component (PC1) (Figures 1H–1L), as before.[26] PCA using exclusively TE reads also revealed separation of pre- and post-EGA stages (Figures 1M–1Q). In addition, the PCA loading position of TEs suggests that LTRs contribute most dominantly to PC1, reflecting EGA timing (Figures 1R–1V). Thus, developmental progression through EGA in mammals is demarcated by expression of specific TEs.

Systematic investigation of EGA genes in five mammalian species using Smart-seq+5′

A pre-requisite to our molecular understanding of EGA is the identification of genes activated at EGA.[35] While this has been done for some mammalian species,[36–38] the protocols used are disparate, preventing robust cross-species comparisons. Our systematic single-embryo Smart-seq+5′ datasets provide an opportunity to deliver a resource to compare EGA genes across mammalian species. To identify EGA genes, we searched for genes differentially upregulated between zygotes and the respective stage of EGA in each species, using DE-seq.[39] Additionally, we only selected genes differentially downregulated between EGA and 16-cell/morula stages at which EGA has completed across all species (Table S2). By doing this, the summed expression levels of all identified EGA genes delineate the known temporal species-specific EGA profile (Figure S1K). Gene Ontology (GO) of EGA genes[37,39] revealed shared terms across species, listed in Table S2. For example, we find GO terms associated with gene expression. EGA genes corresponding to these terms display similar expression profiles across species (Figure S1L). Notably, the number of genes in the GO analysis is limited, in part due to different degrees of annotation, and thus the relevance of GO remains to be determined.

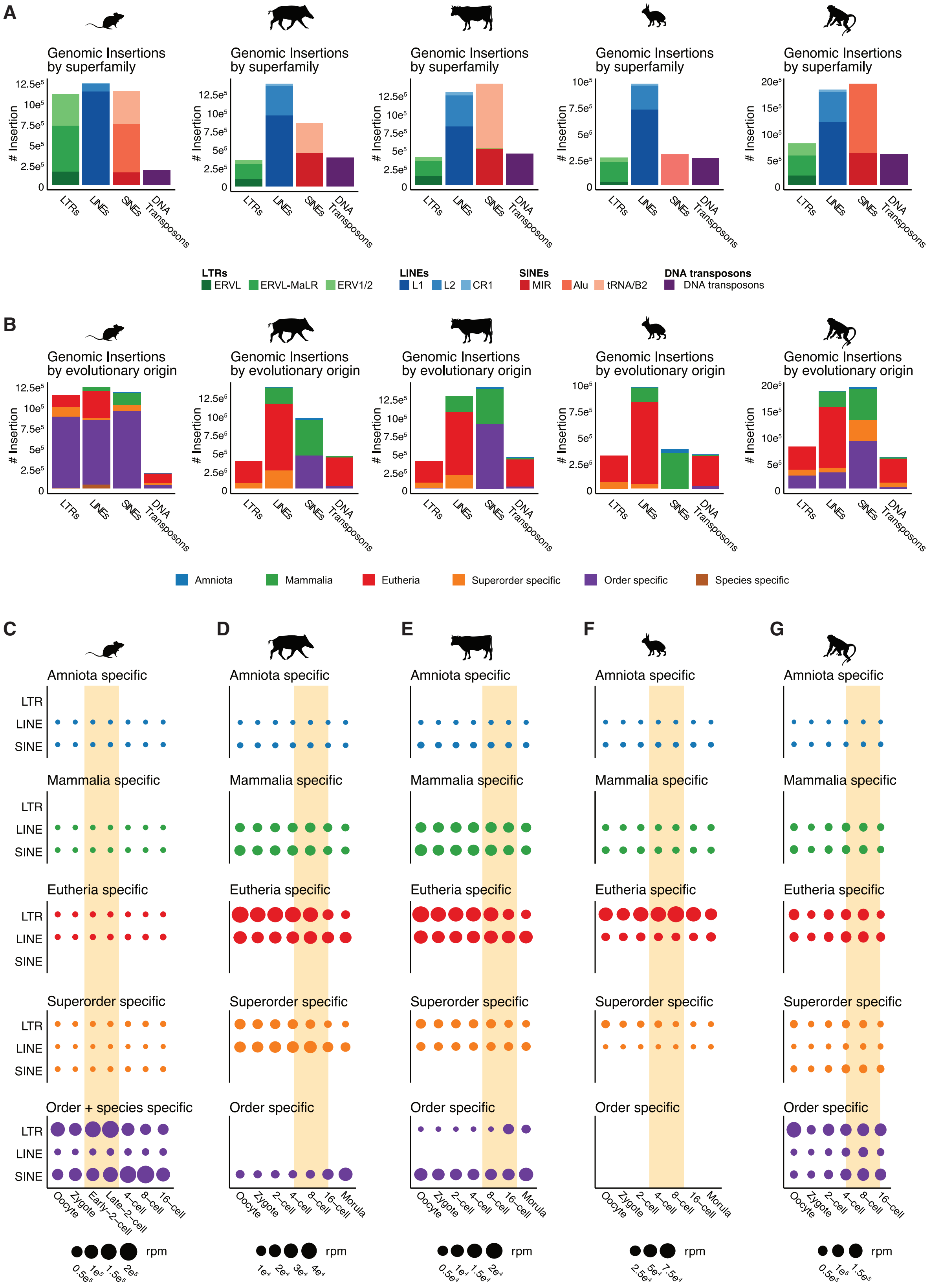
To explore potential conserved regulatory networks of EGA across species, we performed a TF motif search on EGA genes in all five species and analyzed their expression at EGA. We identified 26 TFs with motifs at EGA genes in two or more species, which included, for example, DUX and KLF proteins but also general TFs such as TBP and SP1 (Figure S2A). Individual TF motifs are rarely present at EGA genes in more than three species, except for the KLF family, which have motifs in EGA genes in all species excepting the rabbit, and the corresponding KLF TFs are expressed at EGA (Figure S2A). This suggests certain species specificity in the TF network regulating EGA. Indeed, expression analysis of TFs with known EGA functions in mouse revealed that while DUX orthologs show an EGA expression pattern in all species, NR5A2 is more variable and becomes activated post-EGA in all species excepting pig and rabbit (Figures S2B and S2C). This aligns with work showing that rodent-specific factors like OBOX contribute to mouse EGA regulation, while primate-specific factors like TPRX contribute to EGA in human embryos.[40,41]

While our subsequent analyses focus on TEs, our datasets provide a rich resource for the systematic investigation of genic expression and regulatory factors at large during early mammalian development.

TE subclasses show both species-specific and shared patterns of developmental expression

The importance of TEs in shaping developmental programs prompted us to perform an in-depth TE transcriptomic analysis. We first examined the ability of Smart-seq+5′ to capture TEs, compared with Smart-seq2, and whether this is sensitive to poly-adenylation by performing Smart-seq2 with and without an *in vitro* poly-adenylation step using mouse 2-cell embryos. In general, Smart-seq+5′ detects TEs more efficiently than Smart-seq2 (Figure S2D), and a poly-adenylation step does not increase TE detection, with a few notable exceptions, including ERVL-MaLR and SINE B1/Alu but not LINE elements or other LTRs and SINEs (Figures S2E–S2H). This is in line with work indicating that some SINE B1s are not poly-adenylated.[42,43] We also asked whether the 5′ reads in Smart-seq+5′ detect known TSSs from retrotransposons. As expected, 5′ reads detect LTR sequences but not internal, coding ERVL sequences (Figure S2I). Interestingly, we find 5′ reads mapping to all LINE L1 structural elements, including 5′ end, ORF2, and 3′ end, suggesting that LINE L1 can initiate transcription from their internal elements (Figure S2J). Because many LINE L1 are ancient, fragmented elements, this suggests that LINE L1 fragments have retained and/or acquired the ability to be transcribed, despite having lost their structural integrity. We addressed this by comparing the young *Mus*-specific L1MdTf_I with the older, eutherian L1M5_orf2, which has become severely fragmented over evolutionary time in mouse (Figure S2K). We find that L1M5_orf2 fragments are expressed throughout mouse preimplantation development at levels comparable to those of L1MdTf_I, albeit with differences in their insertion frequency (Figure S4H). Finally, we examined the well-known SINE B1_Mus1, which was more efficiently detected by Smart-seq+5′ than Smart-seq2 (Figure S2L), likely due to increased efficiency of Smart-seq+5′ to capture shorter transcripts.

Having assessed the performance of Smart-seq+5′ to capture TEs, we analyzed their developmental and evolutionary expression patterns. We first focused on retrotransposons, which exhibit the highest regulatory potential described so far.[12,44,45] Specifically, we focused on the LINE, LTR, and SINE subclasses, present in high copy numbers in the genomes of all five species (Figure 2A). The absolute and relative copy numbers of TEs differ across species, with rhesus containing ∼2 million insertions of SINEs and pig only about 750,000 copies, for example (Figure 2A), while the genome length of all these species is comparable (2.5–2.85 Mbp). Notably, the mouse possesses the highest absolute number of LTRs (∼1.1 million), suggesting a more efficient LTR expansion strategy in this species (Figure 2A). Most TEs are located distally to genes (>5 kb), except for SINEs in the mouse and rhesus, where SINEs reside mostly inside genes or are equally enriched inside and distal to genes, respectively (Figure S3A). Despite their different copy numbers across species (Figure 2A), in general, LTRs show an expression increase at EGA, followed by a stark decrease after EGA in all species, except in bovine embryos in which maternal levels of LTR transcripts are globally similar to EGA levels (Figure S3B). In contrast, we observe more species-specific expression patterns for LINEs and SINEs, both in their temporal dynamics and abundance (Figure S3B). Importantly, insertion number alone does not explain



(legend on next page)

different expression dynamics between species (Figure 2A). To determine whether species-specific expression patterns and, in general, TE expression are linked to their evolutionary age, we categorized TEs as amniota (~319 mya), mammalia (~180 mya), eutheria (~99 mya), superorder (~87–94 mya), or order and species specific (less than ~70 mya) (Figure S3C). Analysis of the number of TE insertions according to their evolutionary origin revealed, for example, that mouse contains primarily younger, order- and species-specific elements (Figure 2B). SINEs are also mostly order and species specific in the bovine genome and, albeit to a lesser extent, in the rhesus genome (Figure 2B). In contrast, most LINEs tend to be mammalian and eutherian specific in the porcine, bovine, and rabbit genomes, suggestive of a less successful colonization in these species (Figure 2B). Examining the expression of these TE groups revealed species-specific patterns according to their age. Namely, rabbit embryos express primarily TEs conserved in eutheria rather than order-specific TEs (Figures 2C–2G). Additionally, expression patterns of order-specific TEs differ between species. LTRs and SINEs expressed in mouse embryos are largely species specific, whereas all SINEs transcripts in porcine embryos derive from species-specific SINEs (Figures 2C–2G). Thus, TE expression in early development is not exclusively restricted to young TEs. Additionally, while some species tend to express more young, species-specific TEs, others tend to express mostly TEs conserved throughout eutherians.

Because TE subclasses are highly sequence and evolutionary divergent,[7,34] we further studied the main superfamilies of each subclass. For LTR elements, ERVL-MaLR sequences were the most highly expressed in all species at almost all developmental stages, showing highly similar expression patterns relative to EGA across all species studied (Figure S3D). This contrasts to non-MaLR ERVL, ERV1, and ERV2 (also called ERVK) superfamilies, which show strong species-specific dynamics. For example, bovine embryos upregulate ERV1/2 after EGA, whereas in rabbit embryos ERV1/2 and ERVL transcripts are practically undetectable (Figure S3D). Interestingly, despite their proposed different evolutionary origin,[33] expression dynamics of LINE L1 are shared across all species, showing a consistent increase during and after EGA (Figure S3E). This suggests a degree of conservation in the transcriptional regulation of LINEs across species. Lastly, SINEs show the most striking species-specific patterns (Figure S3F). For example, the majority of SINEs detected in mouse embryos derive from tRNA sequences, whereas rhesus embryos express predominantly Alu elements (Figure S3F).

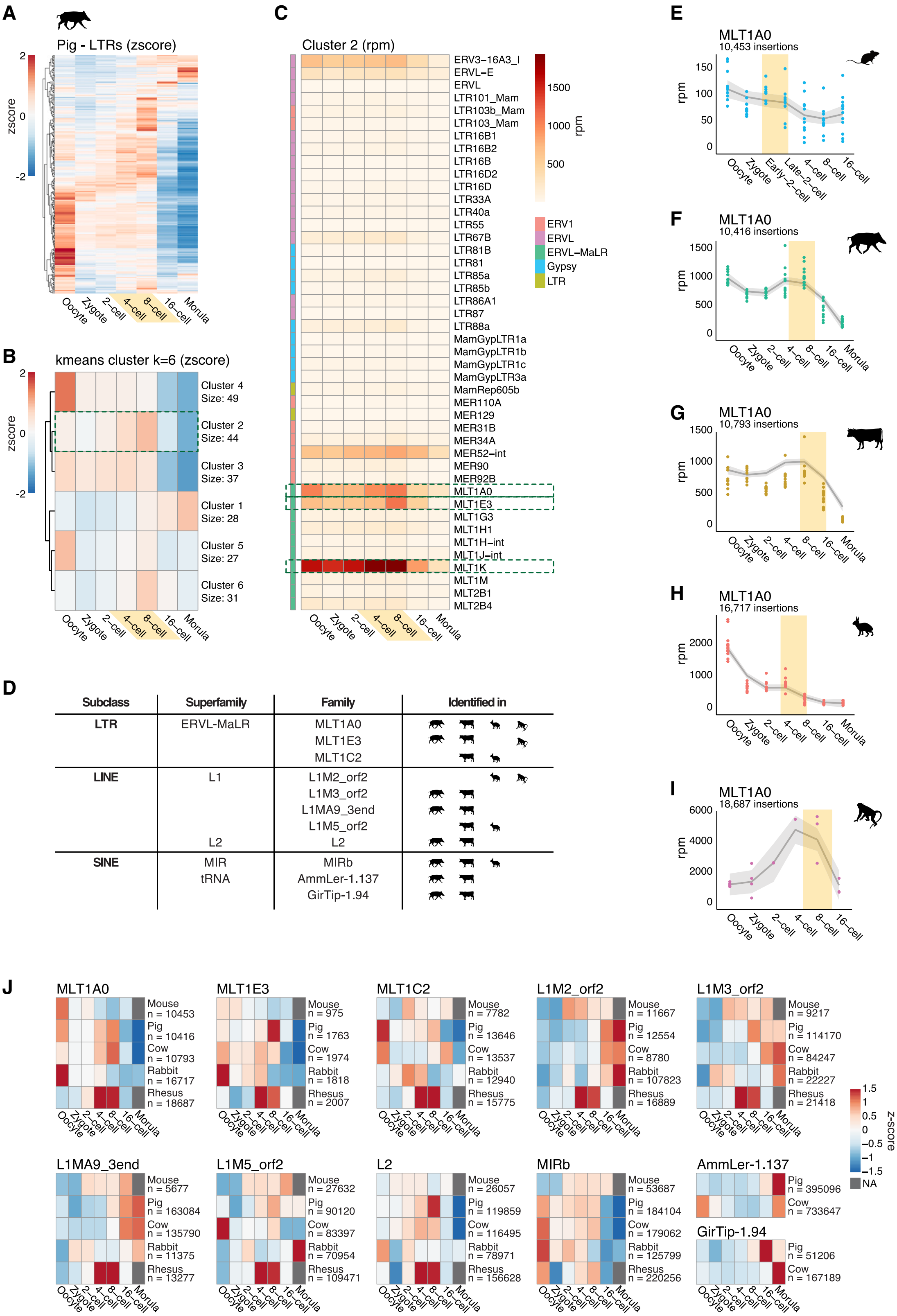
Because our observations rely on the analysis of transcription initiation at TE sequences and their expression abundance does not simply correspond to genomic abundance (Figure 2A), we conclude that there is both shared and species-specific TE expression at different stages of development. Thus, robust transcriptional activation of TEs during preimplantation development is a shared feature among mammals.

Retrotransposons with dynamic EGA expression share conserved transcriptional profiles

Next, we searched for TEs with potential regulatory roles throughout EGA. We used two criteria based on the assumption that such TEs should be highly expressed at EGA, compared with other stages. We extracted TEs that (1) undergo a stage-specific increase at EGA, similarly to the LTR of mERVL, MT2_Mm,[8,16,46] and/or (2) display a strong decrease in expression at or after EGA. We show examples of the analysis pipeline of LTRs in pig (Figures 3A–3C) and of LINEs and SINEs in rabbit and mouse, respectively (Figure S4). We first computed TE expression level as Z score values (Figures 3A, S4A, and S4E). Second, we used k-means clustering to categorize TEs by their expression profiles. This allowed us to identify specific patterns based on the criteria above (Figures 3B, S4B, and S4F). Lastly, we analyzed read counts of all individual TEs within these clusters to select for robust expression of the profile of interest (Figures 3C, S4C, S4D, S4G, and S4H). For example, within cluster 2 for pig LTRs, we identified several ERVL-MaLRs expressed during EGA undergoing a sharp downregulation after EGA (Figure 3C). Performing this analysis across species revealed 13 murine, 10 porcine, 15 bovine, 9 rabbit, and 11 rhesus TEs, which display a specific and robust EGA expression pattern representing all retrotransposon subclasses (Table S3). Many of these EGA TEs are ancient, highly conserved, and present throughout eutherians, except for the identified SINEs, which are either present throughout mammals (MIRs) or are conserved only at the order level (tRNAs) (Table S3). MT2_Mm was among those TEs in mouse embryos, as expected,[9] validating our analysis (Table S3). Additionally, while the developmental stages are not fully concordant with published human datasets, MLT2A1, which is expressed at EGA in rhesus (Table S3), is also expressed in day 3 human embryos.[47,48] Thus, a defined set of TEs shares expression dynamics at EGA during mammalian preimplantation development.

We next focused specifically on TEs with shared expression dynamics. Eleven TEs with an EGA profile are shared by two or more species (Figure 3D). Most notable was the LTR MTL1A0, an ERVL-MaLR, which we identified through our analysis in all species excepting the mouse. Indeed, MLT1A0 is robustly expressed at around 1,000 rpm (reads per million) value or more across all identified species and with similar temporal pattern, increasing its expression at EGA relative to the zygote, followed by a sharp downregulation after EGA (Figures 3E–3I). Additionally, oocytes from several species contain MLT1A0 transcripts, suggesting maternal inheritance (Figures 3E–3I). MLT1A0 did not appear as an EGA-related TE in mouse in our analysis because its expression is substantially lower compared with

Figure 2. Retrotransposon classes, their presence, and expression patterns relative to their evolutionary age
(A) Number of genomic insertions per retrotransposon (sub)class.
(B) Number of genomic insertions per evolutionary age group.
(C–G) Expression of retrotransposons according to the evolutionary age of TE families per Dfam annotation. Bubble size reflects mean expression levels at a given stage. Timing of EGA is highlighted in yellow.
See also Figure S3.



Subclass	Superfamily	Family	Identified in
LTR	ERVL-MaLR	MLT1A0	🐖 🐄 🐇 🐒
		MLT1E3	🐖 🐄 🐒
		MLT1C2	🐄 🐇
LINE	L1	L1M2_orf2	🐇 🐒
		L1M3_orf2	🐖 🐄
		L1MA9_3end	🐖 🐄
		L1M5_orf2	🐄 🐇
	L2	L2	🐖 🐄
SINE	MIR	MIRb	🐖 🐄 🐇
	tRNA	AmmLer-1.137	🐖 🐄
		GirTip-1.94	🐖 🐄

(legend on next page)

other species (Figure 3E). However, MLT1A0 displays a similar pattern during mouse preimplantation development, albeit with a less pronounced increase at EGA (Figures 3F–3I). Likewise, TEs that passed the above selection criteria in at least two of the five species show highly similar expression patterns relative to EGA timing in all species (Figure 3J). Additionally, among LTRs, our analysis identified primarily ERVL-MaLR LTRs displaying an EGA profile (Figure 3D; Table S3). Overall, our data reveal the full repertoire of TE transcription across preimplantation development and identify MLT1A0 as a TE with shared expression dynamics at EGA.

Retrotransposons employ characteristic types of TSS during early development

Several types of promoter architecture exist in metazoans,[27,28] which are linked to transcription initiation regulation and to groups of genes with specific functions.[49] Sharp promoters are characterized by a single well-defined TSS, whereas broad promoters, in which transcription initiates at multiple positions, use several TSSs.[49] Typically, sharp promoters occur at cell-type-specific genes of terminally differentiated cells, whereas broad promoters are more common in house-keeping genes across cell types and, in mammals, of developmental regulators.

We reasoned that an analysis of TSS usage could shed light on the regulation of transcriptional initiation of retrotransposons at EGA. Thus, we next asked whether TEs display specific promoter architecture in early mammalian embryos. The 5′ fragments in Smart-seq+5′ provide strand-specific, single-base-pair resolution to map TSSs, similarly to CAGE.[28] We investigated whether we could position the TSS within a TE. As a proof of principle, we first focused on MT2_Mm insertions in the mouse genome and aggregated the 5′ signal, which indicated a sharp positioning of one predominant TSS across all stages, with the strongest signal at the 2-cell stage (Figure 4A). Visualization of the 5′ signal as a heatmap, where each row is an individual MT2_Mm insertion, indicates that the vast majority of insertions are expressed and utilize the same TSS position (Figure 4B). We also asked whether MT2_Mm produces antisense transcripts but found predominantly sense transcripts (Figure 4C).

Next, we set out to compare TEs across species by categorizing EGA profiles, TSS profiles, and sense or antisense transcription specifically for retrotransposons that displayed EGA profile (Table S4). First, we distinguished four expression patterns, (1) upregulation specifically at EGA, (2) upregulation at EGA followed by a plateau in expression, (3) upregulation after EGA, and (4) downregulation at EGA (examples of each pattern are shown in Figures S4I–S4L). Overall, we found that several LTRs upregulated specifically at EGA. Downregulation of selected LTRs after EGA is a common feature of all the mammalian species studied, regardless of their expression pattern prior to EGA (Figure 4D), fitting with the overall trend observed for LTRs (Figure S3D). Conversely, LINEs display distinct expression behaviors, indicating dynamic regulation during early development in all species (Figure 4D). In general, SINEs with an EGA pattern tend to have more extended expression periods that typically plateau until later stages (Figure 4D), in line with work in the mouse blastocyst.[50] Second, we characterized six different TSS patterns: (1) a sharp sense-only TSS (Figure S4M), (2) one predominant sense TSS peak with a flanking antisense TSS (Figure S4N), (3) one predominant sense TSS with a weaker flanking antisense TSS (Figure S4O), (4) a sense TSS at the 5′ start of the TE sequence (Figure S4P), (5) a sense TSS at the 3′ end of the TE and antisense TSS at the 5′ start of the TE (Figure S4Q), and (6) a depletion of signal inside the TE (Figure S4R). The latter is primarily due to transcription initiation immediately downstream, within 500 nt, of the SINE element, for example, MIRb in the cow (Figure S4R). Remarkably, this analysis revealed very uniform TSS patterns for LTRs, LINEs, and SINEs (Figure 4D; Table S4). LTRs often rely on a single, sharp sense-only TSS (Figure 4E). Conversely, LINEs typically exhibit a sense TSS at the 3′ end and show antisense transcription from a TSS at the 5′ of ORF2 (Figure 4F), a feature previously described for a few individual full-length human L1 elements.[51–53] SINE elements display transcriptional initiation characterized by a sharp TSS in the sense orientation at the 5′ start of the TE signal across species (Figure 4G). Globally, we detect antisense transcription from within all LINEs but rarely at SINEs or LTRs. Also, some LTRs and SINEs with EGA profile are expressed at higher levels than LINEs (Figures 4E–4G), reflecting potentially higher promoter strength. These analyses of expression and TSS usage suggest a conserved mechanism of transcriptional regulation that is inherent to the TE and its subclass but not to the host or insertion frequency.

DNA transposons are transcriptionally active throughout mammalian preimplantation development

DNA transposons remain largely underinvestigated in mammals, mainly because they are considered predominantly extinct in terms of transposition potential in mammalian genomes.[54] Indeed, the last detected wave of DNA transposon amplification occurred over 40 mya, except in some bat species.[7,33] However, remnants of DNA transposons remain abundant,[7] with between 200 and 600 thousand copies in the species we analyzed

Figure 3. Analysis of TE families identifies shared TEs with EGA expression profile during early mammalian development

(A) Example of TE analysis pipeline using LTRs. Hierarchical clustering based on expression of individual LTR families per row in pig by *Z* score.

(B) k-means clustering (k = 6) of LTR family expression from (A) by *Z* score. Number of LTR families per cluster is indicated.

(C) Expression values from all embryos as mean rpm of individual TEs within cluster 2 from (B). Subclasses are indicated with color code.

(D) List of TEs with an EGA profile in two or more species analyzed.

(E–I) Expression levels of MLT1A0. Each dot represents sum rpm of all insertions per embryo. Shading indicates SD; number of MLT1A0 insertions per species is indicated.

(J) Heatmaps of expression profiles by *Z* score of 11 TEs from (D). SINE elements AmmLer-1.137 and GirTip-1.94 are only present in pig and cow. For mouse, late 2-cell stage is shown. Number of insertions (*n*) of TE in given species is indicated.

See also Figure S4.

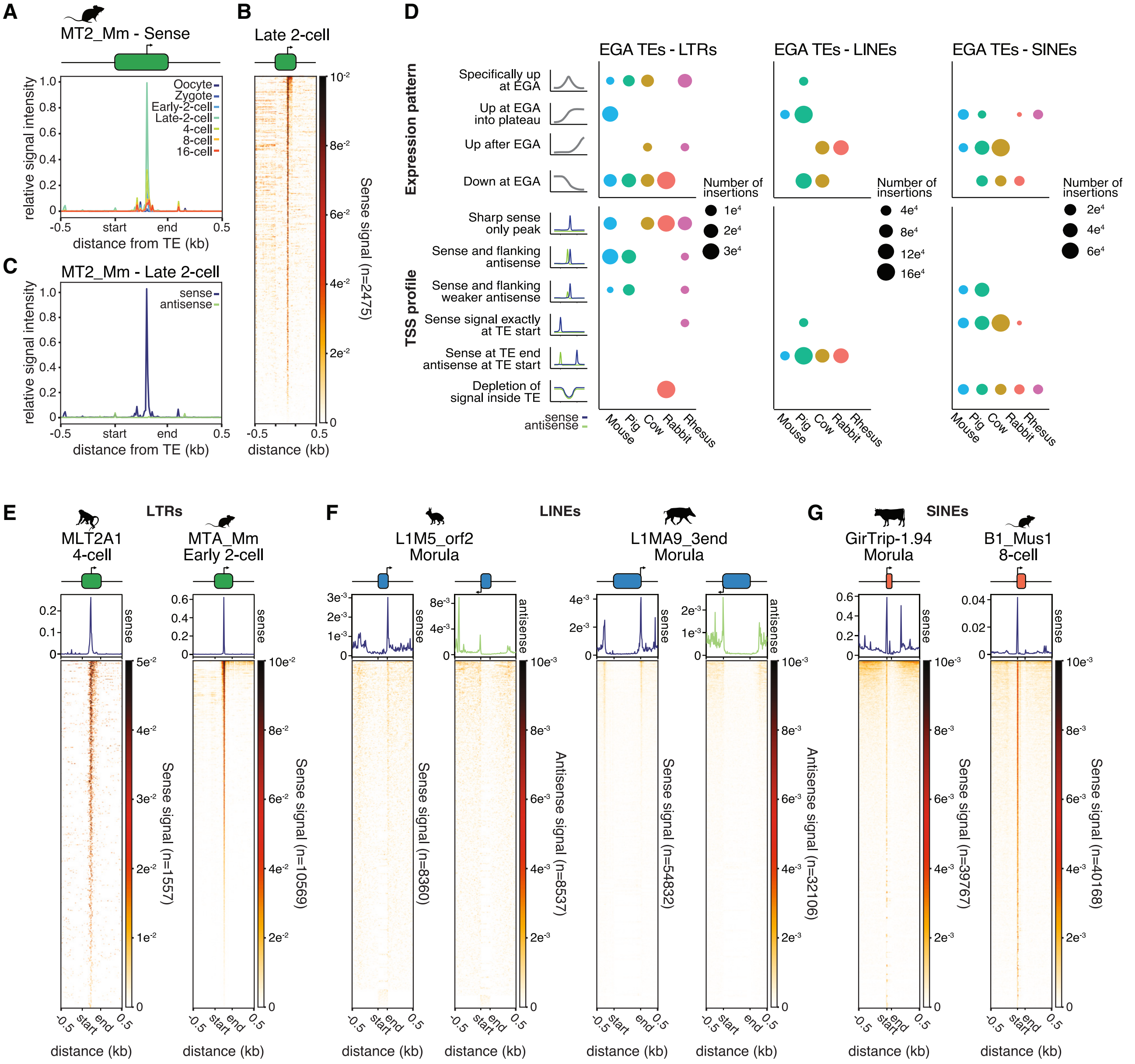


Figure 4. Analysis of TSS and expression patterns suggests a conserved mechanism of transcriptional regulation inherent to TE classes

(A) 5′ signal of mouse LTR MT2_Mm sequence across developmental stages in sense orientation indicating a sharp TSS within the LTR. Start and end refer to the position of MT2_Mm.

(B) Heatmap of 5′ sense signal illustrating the TSS across individual MT2_Mm insertions in late 2-cell embryos. Number (*n*) of insertions shown is indicated.

(C) 5′ signal from sense (blue) and antisense (green) transcripts over MT2_Mm sequences at late 2-cell stage. Start and end refer to the position of MT2_Mm.

(D) Summary of EGA TEs showing expression pattern and TSS profile. Bubble size represents the number of insertions showing a given expression pattern or TSS profile.

(E–G) Examples of conserved TSS patterns for LTRs, LINEs, and SINEs depicted as a heatmap of sense or antisense 5′ signal over the genomic insertions of each TE. Only uniquely mapped reads were used. Lower mappability within the TE cannot be excluded. Number (*n*) of insertions shown are indicated.

See also Figure S4.

(Figure S5A). Remarkably, we observed transcription of DNA transposons at all stages of development in all investigated species (Figure S5B). Their combined transcriptional dynamics varied across species, with a strong maternal contribution of DNA transposon transcripts in pig, cow, rabbit, and rhesus but less in mouse (Figure S5B). We asked whether specific DNA transposons are transcribed at EGA. Hierarchical clustering of DNA transposons indicated several types of expression patterns, for example, as seen in cow embryos (Figure 5A). These include not only maternally inherited DNA transposon transcripts but

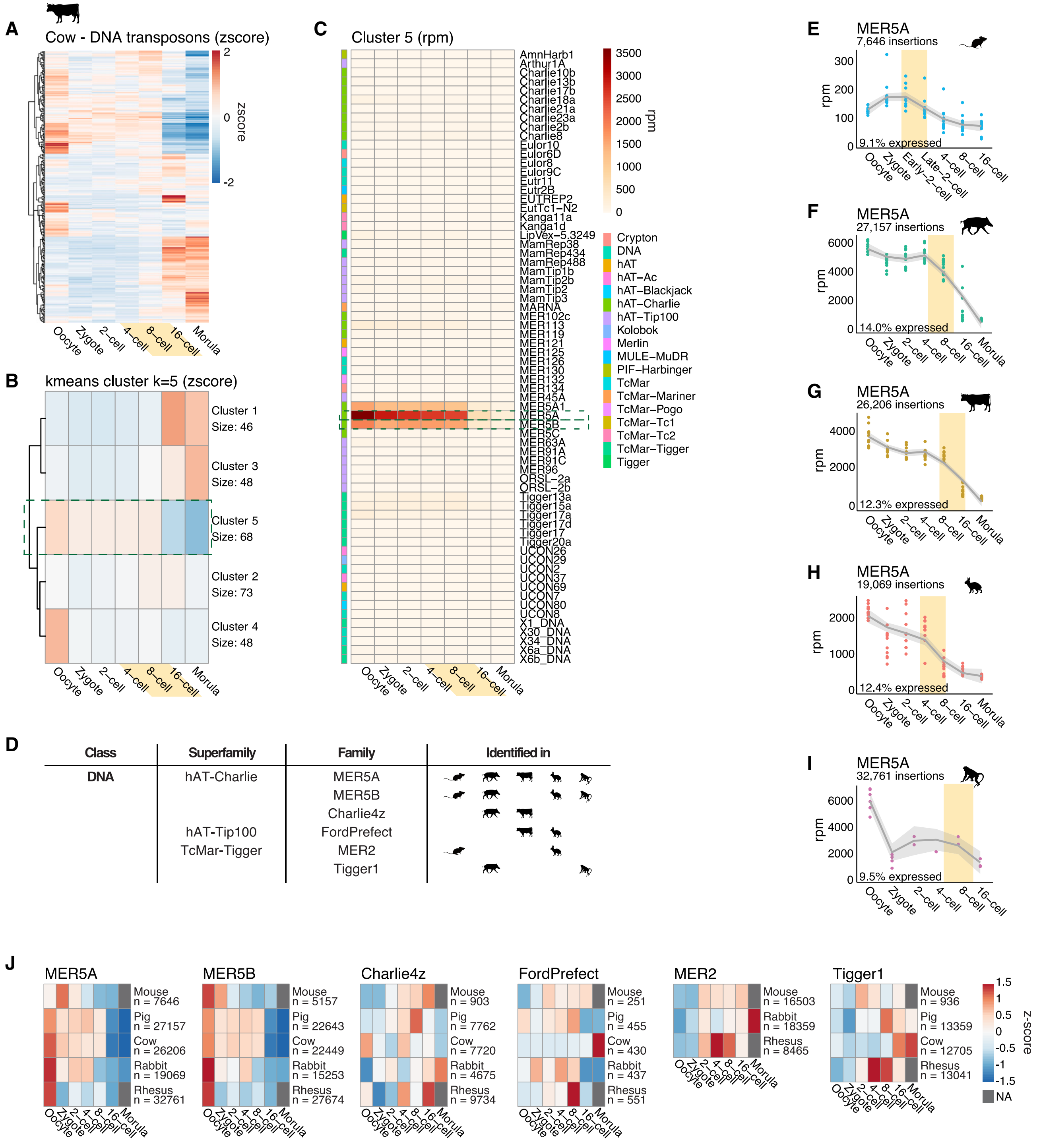


Figure 5. DNA transposons are actively transcribed during early mammalian development and some show conserved EGA-specific expression

(A) Example of TE analysis pipeline for DNA transposons. Hierarchical clustering based on family expression in cow by *Z* score. Each row corresponds to an individual transposon.

(B) k-means clustering (k = 5) of DNA transposon family expression from (A) by *Z* score. Number per cluster is indicated.

(C) Expression values from all embryos as mean rpm values of individual TEs within cluster 5 in (B). Subclasses are indicated with color code.

(D) List of TEs with an EGA profile in two or more species analyzed.

(legend continued on next page)

also transcripts displaying increasing levels at different stages of preimplantation development (Figure 5A). k-means clustering (Figure 5B) followed by the identification of clusters displaying dynamic expression patterns throughout EGA and a robust signal allowed for selecting specific DNA transposons within such clusters (Figure 5C; see cluster 5). This led to the identification of a handful of DNA transposon families with an EGA profile for each species, listed in Table S3, several of which are shared across species (Figure 5D). Most notably, MER5A is strongly expressed in all five species with an EGA-specific pattern (Figures 5E–5I). MER5A is a DNA transposon from the Charlie-hAT superfamily found across the eutherian clade and therefore >99 mya.[34,55] Only 9%–14% MER5A insertions are transcribed (694 in mouse, 2,366 in rabbit, 3,797 in pig, 3,215 in cow, 3,111 in rhesus; Figures 5E–5I), perhaps due to their old evolutionary age and loss of sequence integrity. While there is no strong correlation between expression levels and sequence fragmentation, the MER5A insertions that are expressed tend to be less fragmented, most visibly in rhesus, bovine, and pig (Figure S5C). Over 30% of transcriptionally active MER5A insertions in mouse (33%), pig (35%), rabbit (36%), and rhesus (47%) have an orthologous insertion in human, which is highly similar to the total number of MER5A insertions that is syntenic with human (33%, 33%, 34%, and 43%, respectively). TSS analysis confirmed that MER5A initiates transcription from its own sequence (Figures S5D–S5H). Additionally, and in contrast to LTRs, MER5A uses a broad promoter type, with several initiation sites within the TE as well as directly downstream of the TE (Figure S5I). We observed sense and antisense transcription, likely because MER5A is a palindrome.[34,55] We next analyzed enrichment of TF motifs across all MER5A insertions in the five species to address whether a similar TF network is associated with MER5A across species. We identified motifs for 21 TFs in total, among which ZNF692 contains a motif across at least ∼31% MER5A insertions in all species (Figure S5J). Additionally, while the TEAD2 motif is enriched in rabbit, pig, cow, and mouse MER5A, it is not present in rhesus, which instead has >59% of MER5A insertions containing a TEAD4 TF motif. Thus, we find conserved motifs as well as diverse TF motifs across MER5A insertions.

We extended our expression analysis toward other DNA transposons, which revealed several additional DNA transposons with an EGA profile in at least two species (Figure 5J). Thus, DNA transposons share the regulatory burst of transcriptional activation of retrotransposons during mammalian EGA.

These findings are particularly important because DNA transposons are much more ancestral sequences.[33] Despite their assumed loss of transposition potential, our data indicate that DNA transposons are transcriptionally active and show that specific DNA transposons become activated in early development with expression patterns shared across species.

MLT1A0 shows evolutionary conservation and can drive gene expression across species

To further understand the potential roles of TEs at mammalian EGA, we focused on the ERVL-MaLR LTR MLT1A0 as it displays conserved EGA expression dynamics. To study the conservation of MLT1A0, we first generated species-specific consensus MLT1A0 sequences[56–58] including seven additional mammalian species, using all predominant full-length MLT1A0 insertions in a given genome (Figure S6A). It is notably full-length MLT1A0 that contributes most to its EGA profile (Figure S6B). To reveal the evolutionary relationship of MLT1A0 sequences, we performed a phylogenetic analysis of consensus sequences of 12 mammalian species and the MLT1A0 consensus curated by Dfam.[34] As expected, the Dfam consensus clusters most distantly (Figure 6A). The phylogenetic MLT1A0 tree follows a highly similar structure to the general genome diversification of species and orders.[59] Specifically, MLT1A0 of primates (gorilla, rhesus, chimpanzee, and human) clusters together, the glires (rabbit, mouse, and rat) form a second cluster, and a third cluster forms with artiodactyls (goat, alpaca, pig, cow, and sheep) (Figure 6A). Additionally, MLT1A0 consensus displays very high sequence identity with the Dfam consensus (between 92% and 96% identity) (Figure 6B).

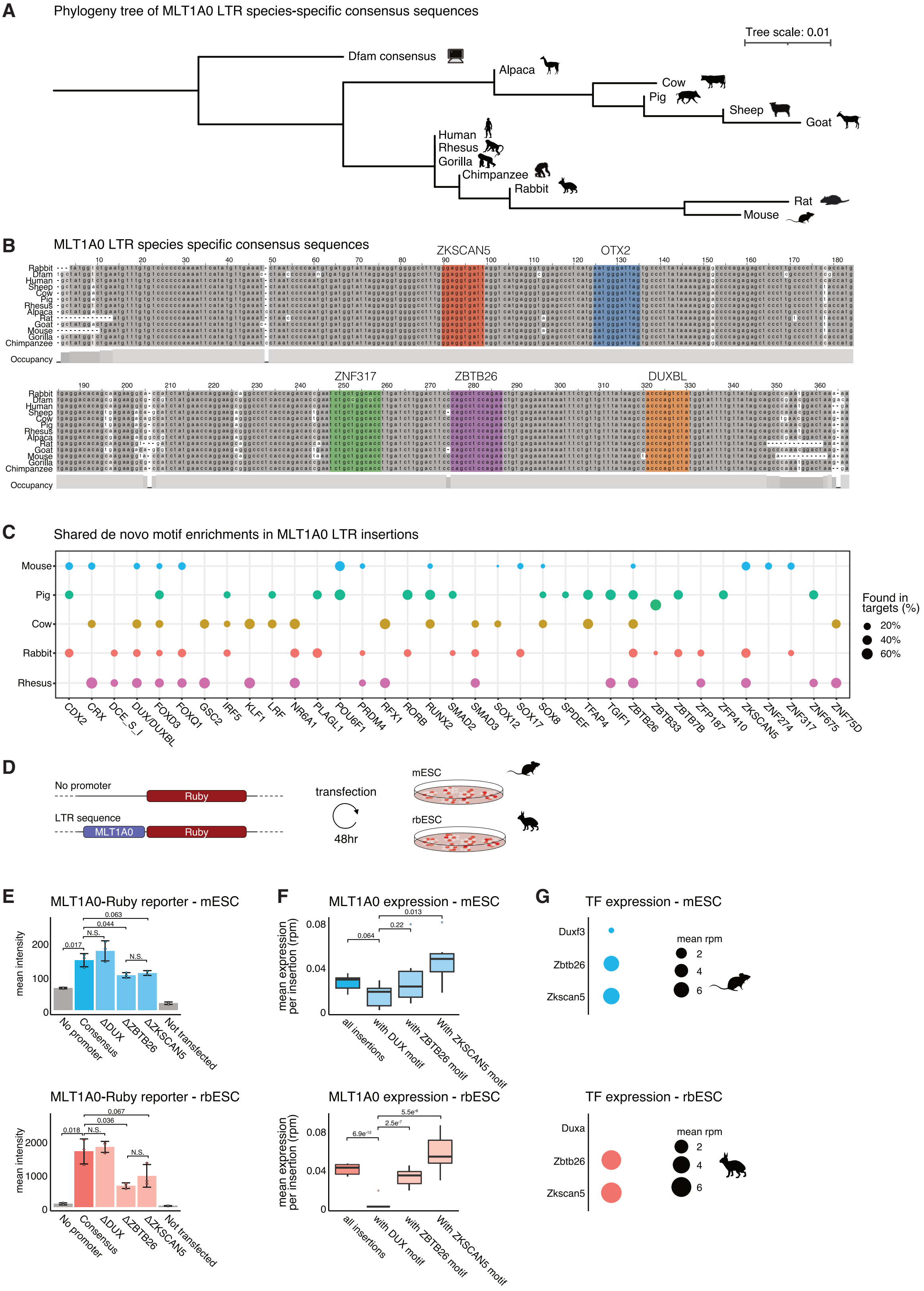
MLT1A0 LTR possesses a strong sense-only TSS in pig, cow, rabbit, and rhesus embryos, but not in mouse embryos, at the time of EGA when MLT1A0 expression in mice is low (Figure S6C). To investigate whether MLT1A0 regulation is conserved, we first sought to identify its transcriptional regulators. We performed a TF motif search and identified known motifs within the MLT1A0 consensus (Figure 6B). These TF motifs were also found when performing a *de novo* motif search on individual full-length MLT1A0 insertions (Figure 6C). Several TFs with binding motifs in MLT1A0 are conserved and include OTX2 and ZKSCAN5, which are expressed throughout preimplantation development (Figures S6D and S6E), further supporting that these TFs could be involved in the transcriptional regulation of MLT1A0. Interestingly, OTX2 has been suggested as regulator of human EGA,[41] and we find MLT1A0 is also expressed in early human embryos from publicly available data.[47]

Second, we assessed whether MLT1A0 possesses intrinsic transcriptional potential across species. We tested whether MLT1A0 can drive expression of a flanking gene in a heterologous reporter, for which we cloned MLT1A0 LTR from eight different species of rodents, lagomorphs, primates, and artiodactyls upstream of the coding region for red fluorescent protein Ruby (Figures 6D and S6F). Remarkably, we find all can drive reporter expression when transfected into mouse or rabbit ESCs, compared with the control plasmid lacking a promoter (Figure S6F). Interestingly, the transactivation capacity of MLT1A0, as measured based on expression levels of the reporter, follows the tree structure of the phylogeny analysis. Thus, we conclude that MLT1A0 possesses transcriptional activity across species.

(E–I) Expression levels of MER5A. Each dot represents sum rpm of all insertions in each species, per embryo. Shading represents SD; number of MER5A insertions for each species and percentage expressed (>1 rpm in at least two replicates) are indicated.

(J) Heatmaps of expression profiles by *Z* score of the six DNA transposons with EGA profile in (D). For MER2 and Tigger1, only data for the species in which they are annotated are shown. For mouse, late 2-cell stage is shown. The number of insertions (*n*) per species is indicated.

See also Figure S5.



(legend on next page)

Lastly, to directly probe regulation of MLT1A0 by specific TFs, we performed deletions of the TF motifs we identified above. We used the mouse and the rabbit MLT1A0 sequences on which we performed individual deletions of predicted DUX, ZBTB26, and ZKSCAN5 motifs, as these were present in most species. We performed reporter assays with either wild type or MLT1A0 containing deletions in mouse and rabbit ESCs. The MLT1A0 reporter activity was reduced upon deletion of ZBTB26 or ZKSCAN5 motifs but not that of DUX (Figure 6E). Examining the expression of MLT1A0 insertions based on whether they possess DUX, ZBTB26, or ZKSCAN5 motifs indicated that those insertions containing the DUX motif are expressed at lower levels than those containing the ZBTB26 and ZKSCAN5 motifs (Figure 6F), in line with the expression abundance of these TFs in ESCs (Figure 6G). Thus, we conclude that ZBTB26 and ZKSCAN5 contribute mostly to MLT1A0 transcriptional activity in stem cells.

Systematic identification of chimeric transcripts reveals widespread influence of TEs initiating genic transcription in mammalian embryos

Our observations above indicate that TEs occupy a vast place in the embryonic transcriptome. To investigate whether TEs have potential regulatory functions, we asked whether TEs can initiate genic transcription. To address this and taking advantage of our ability to capture TSSs, we set out to identify chimeric transcripts initiating at a TE (Figure 7A). We employed ChimeraTE,[60] which uses directionality information to distinguish chimeric reads that use TE sequences to initiate or terminate a chimeric transcript as well as exonized TE transcripts. Applying ChimeraTE to all developmental stages and species, we identified a total of 19,657 unique chimeric transcripts initiated at a TE (Figure 7B; Table S5). All retrotransposon subclasses and DNA transposons can initiate chimeric transcripts in embryos (Figure 7B), albeit at different proportions between stages and species (Figures 7B and 7C). Some species have a preference for using certain retrotransposons as transcript-initiating TEs, reflecting in part their higher genomic content in the respective species (Figure 2A). For example, LTRs are the dominant TE-initiating transcripts in mouse embryos (59%, 3,614 out of 6,112), but LINEs are most common in rabbit (40%, 646 out of 1,612), and SINEs in rhesus (42%, 3,910 out of 9,121). Despite their typically evolutionary younger origin, SINEs form chimeric transcripts in all species: 1,863 transcripts in mouse, 356 in rabbit, 112 in bovine, 3,910 in rhesus, and 495 in pig embryos (Figure 7C). We find that there are specific stages at which chimeric reads are more abundant: TE-initiated chimeric transcripts are most prevalent in oocytes followed by the EGA stage (Figure 7C). This pattern is conserved in all analyzed species.

We also investigated whether TEs have a different ability to regulate neighboring genes in different species by performing a distance analysis between the TE-initiating transcription and the start position of the host gene in the chimeric transcript. The vast majority of TE-gene interactions occur at <50 kb (Figure S7A). SINEs initiate host transcripts at the shortest distances (mostly within 5 kb) in all species, except for cow (Figure S7B). Inversely, LTRs typically initiate transcription of host genes over larger genomic distances, which is a conserved feature (Figure S7B). Importantly, the relative representation of LTRs (Figure S7C), LINEs (Figure S7D), and SINEs (Figure S7E), which initiate chimeric transcripts as alternative promoters, does not necessarily reflect their expression levels (compare Figures S7C–S7E with Figures S3D–S3F).

Finally, we asked whether MER5A and MLT1A0 can initiate expression of host genes. We find that MER5A can act as alternative promoter for host genes in all species, amounting to 210 genes in all 5 species and stages, including protein-coding and noncoding transcripts (Table S5). Likewise, MLT1A0 forms chimeric transcripts in all 5 species (93 in total), indicating that MLT1A0 can regulate transcription of host genes in embryos. Among those, C1D, a nuclear co-repressor is among the highest expressed TE-gene chimera in porcine, bovine, and mouse embryos (Table S5). While MLT1A0 initiates C1D transcription in porcine and bovine embryos, it is MT2C, a ∼30-mya-old LTR ancestor to MT2_Mm, that initiates C1D transcription at EGA in mouse (Figures 7D–7F). Visualization of TSS usage[61,62] indicates that the TE TSS is most used at EGA, regardless of whether MLT1A0 or MT2C_Mm are used as alternative promoter (Figures 7D–7F). However, across all chimeric transcripts, we did not find shared ortholog genes that use the same TE as alternative TSS in all five mammalian species (Table S5). We further probed directly whether MLT1A0 can drive transcription in embryos during EGA by performing microinjections of the MLT1A0 reporters into mouse and bovine embryos (Figure 7G). MLT1A0 drives reporter transcription in bovine embryos shortly before major EGA, at the eight-cell stage (Figure 7H). In addition, while the *Mus*-specific MERVL MT2_Mm drives transcription of the Ruby reporter in mouse embryos at the 2-cell stage, MLT1A0 does not

Figure 6. MLT1A0 shows conserved sequence, expression features, and intrinsic transcriptional activity

(A) Maximum likelihood phylogenetic tree of MLT1A0 LTR consensus sequences in 12 species and Dfam consensus. Numbers indicate relative genetic distance between nodes.

(B) Alignment of species-specific full-length MLT1A0 consensus and Dfam consensus with indicated TF motifs.

(C) *De novo* TF motif enrichment across all full-length insertions in each species. Bubble size represents the percentage of insertions in each genome enriched for a given motif over genomic background. Only TFs shared across at least two species are shown. Motifs belonging to the same TF were pooled.

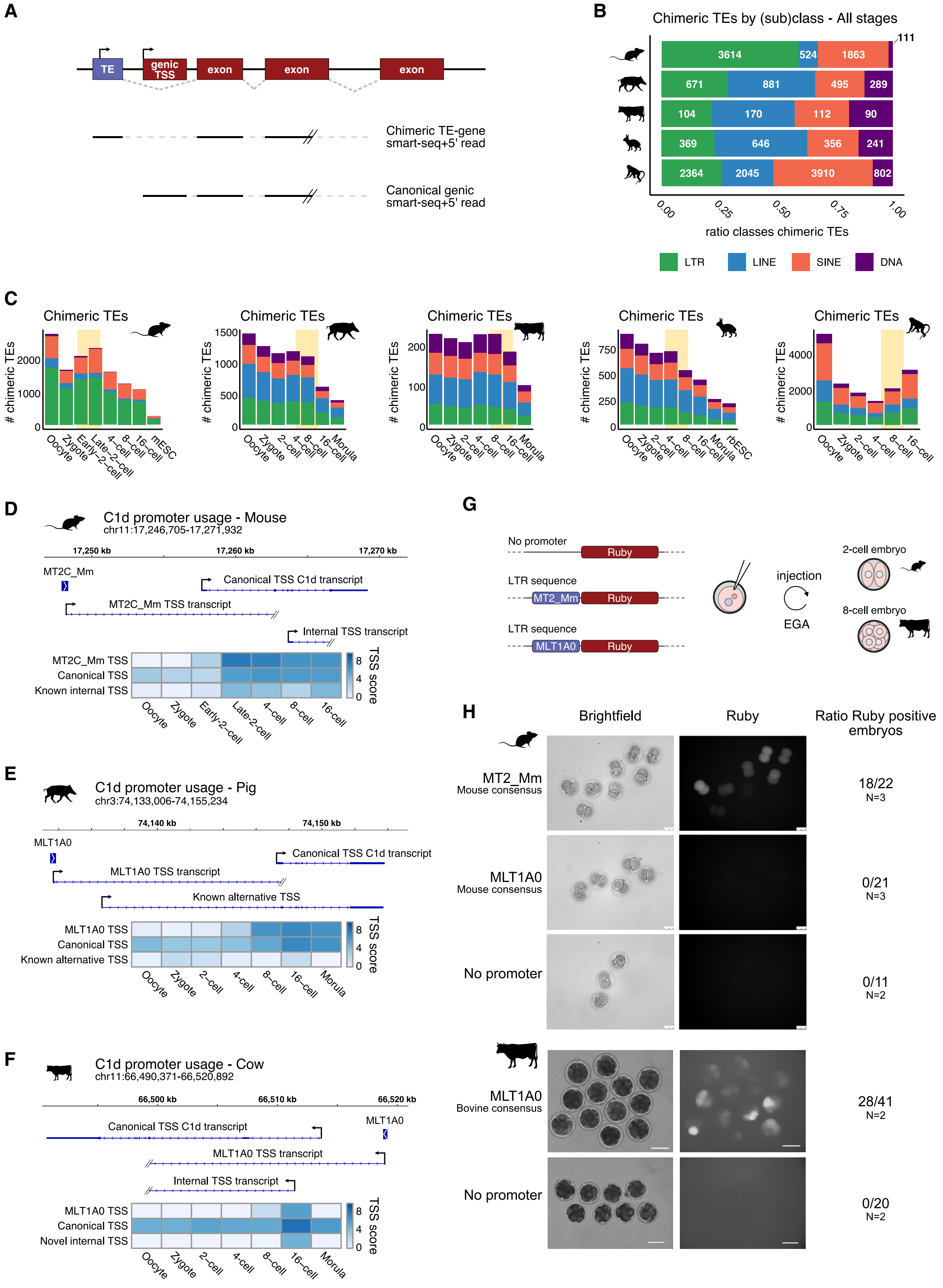
(D) Experimental design to test transcriptional activity of mouse and rabbit MLT1A0 consensus in mouse (mESCs) and rabbit (rbESCs) stem cells.

(E) Mean intensity of Ruby fluorescence by flow cytometry in mESC (top) and rbESC (bottom) upon transfection of relevant species-specific MTL1A0 consensus and TF-motif deletions. Barplots show median ± SD; individual biological replicates are shown as dots. *p* values below 0.1 are shown, *p* values above 0.1 indicated as non-significant.

(F) Mean expression of individual MLT1A0 insertions with indicated TF motifs in mESC (top) and rbESC (bottom). Boxplots indicate median with upper/lower quartiles as box limits and quartile range as whiskers. *p* values are shown above boxplots.

(G) Expression of indicated TFs in mESC (top) and rbESC (bottom).

See also Figure S6.



(legend on next page)

(Figure 7H). These observations are in line with our findings that MLT1A0 has an EGA pattern in all species but with substantially lower expression in mice, especially compared with the *Mus*-specific LTR MT2_Mm, and they demonstrate that MLT1A0 can drive transcription *in vivo* at EGA. Overall, these data suggest that mouse-specific LTRs such as MT2_Mm and MT2C_Mm have overtaken the role of older, eutherian LTRs like MLT1A0 during evolution in driving chimeric gene expression.

DISCUSSION

Here, we developed a protocol and generated a resource enabling the systematic interrogation of the transcriptional regulatory landscape of mammalian preimplantation development. In addition to providing reference EGA datasets, we provide a comprehensive dissection of the TE transcriptional repertoire.

A long-standing hypothesis is that transposition of TEs in their host genomes allowed for dispersed integration of TF motifs, promoter, enhancer, and/or repressive sequences.[12,14,48,63,64] Although the majority of TEs in vertebrate genomes are no longer mobile,[65] their transcriptional regulation would allow simultaneous regulation of many genes flanking TEs[14,16,63] and chromatin regulation genome-wide,[10] orchestrating entire networks of genes.[15,66] Using our low-input Smart-seq+5′, we identified TEs with conserved dynamics at and around mammalian EGA, when large gene networks are turned on simultaneously,[67] hinting toward a co-option for regulatory roles. Alternatively, conserved TE regulation may be a testament of TE colonization of cellular niches promoting their evolutionary persistence. The preimplantation embryo may be a desirable niche for TE expression, providing an opportunity to produce inheritable insertions.[67] The findings presented here are consistent with this idea and suggest that TEs' strategy to hijack EGA is evolutionary stable.

Expression profiles and TSS features are shared across subclasses of TEs and individual TE families, suggesting evolutionary conservation of certain TE characteristics and regulation. Among them, the eutherian LTR MLT1A0 shows strong evolutionary conservation of its expression patterns, sequence, TF motifs, and ability to drive gene expression. While the EGA profile of MLT1A0 had been observed in bovine embryos,[68] our work extends these observations to other species and provides functional and evolutionary characterization.

Smart-seq+5′ will allow for accurately quantifying TE expression in low-input samples and single cells. We anticipate that our data will be key to explore the use of alternative isoforms and promoter usage during mammalian development. This will be particularly relevant to further uncover promoter architectures across TE families and mammalian species, of which only few have been characterized. For example, human LINE L1 promoters have been functionally characterized but remain largely unstudied in other species.

We expect our Smart-seq+5′ datasets will allow to address many other outstanding questions beyond TE biology, as shown by the characterization of EGA genes across species, including deciphering genome-wide rules of core promoter usage of genes and characterization of TFs that regulate EGA in different mammalian species.

In sum, our work highlights evolutionary conservation of TEs and their transcriptional activity while providing a powerful resource for understanding transcriptional regulation of host genes and TEs and their potential co-option during mammalian development.

Limitations of the study

Like many, our RNA-seq methodology cannot distinguish between steady-state and nascent transcripts. Additionally, quality differences between genome assemblies and annotations remain challenging for comparative genomics approaches. Also, some aspects for mapping repetitive elements, including TEs, remain technically complex. While our study presents both experimental and computational efforts in this direction, we cannot formally exclude loss of signal due to loss of multimapping reads, for example, in the visualization in heatmaps. While we cannot exclude additional roles of MLT1A0, MER5A, and other ancient TEs, for example, as enhancer, in splicing donor sequencing, or through non-coding functions, our data indicate that one of their functions may be to generate chimeric transcripts. Whether these alternative transcripts are translated and result in functional proteins remains to be determined. Indeed, Smart-seq+5′ cannot be used to assemble full-length transcripts since internal fragments lack strand information. Furthermore, investigating the expression and regulation of and by MTL1A0 would be of interest in the context of early embryonic-like cells[69] such as mouse 2-cell-like cells (2CLCs)[70,71] and human 8-cell-like cells (8CLCs).[72] The deployment of targeted approaches such as CRISPRi/a will allow for functional studies of TEs expressed during early embryogenesis in the future. Currently, performing such analyses in non-mouse

Figure 7. TE insertions are used as alternative TSS for genic transcription during mammalian preimplantation development

(A) Schematic of TE usage as promoters. Owing to sequence read length, full-length transcripts are not always captured, indicated by slashed lines.

(B) Relative abundance of chimeric TE-gene transcripts per type and per species across stages. Absolute numbers of chimeric TEs by subclass are indicated.

(C) Number of chimeric TE-gene transcripts per species and subclass. EGA timing is highlighted in yellow. Only chimeric transcripts expressed in at least two embryos in at least one stage are shown.

(D) Promoter usage (TSS score) of C1d in mouse embryos. The position of MT2C_Mm or canonical (genic) TSS is indicated.

(E and F) As in (D), for pig (E) and bovine (F). Shown are the positions of the MLT1A0 TSS and the canonical (genic) annonated TSSs. Heatmap indicates the TSS usage as TSS score at indicated stages. (D–F) End of transcript information is indicated by slashed lines.

(G) Schematic of TE-Ruby reporter plasmid microinjected in mouse and bovine zygotes.

(H) Representative bright-field and fluorescence images at timing of EGA in mouse 2-cell and bovine 8-cell embryos after microinjection with indicated plasmids. *n*, number of embryos with fluorescence signal over total number of embryos microinjected in two independent experiments ($N = 2$). Fluorescence intensity is heterogeneous between blastomeres in bovine embryos, reflecting known asynchrony of cells entering EGA in cow.

See also Figure S7.

mammalian models remains challenging. An interesting question is why certain insertions of a TE family become expressed during development while others do not. Techniques such as assay for transposase-accessible chromatin using sequencing (ATAC-seq) could address whether accessibility of certain genomic regions affects the transcription of a TE insertion. Indeed, recent work on the TF NR5A2 in mouse embryos suggests that chromatin accessibility, TF binding, and TE expression are not necessarily always correlated.[73] Similarly, analysis of chromatin state across species will shed molecular information on why some TEs are expressed more robustly in different species.

RESOURCE AVAILABILITY

Lead contact

Further information should be directed to the lead contact, Maria-Elena Torres-Padilla (torres-padilla@helmholtz-muenchen.de).

Materials availability

No new materials were generated in this study.

Data and code availability

- Generated genomics data is available through GEO: GSE225056.
- Data can be explored through embryo.helmholtz-munich.de.
- All original code is available through GitHub (https://github.com/meoomen/Smartseq5).
- Any additional information required to reanalyze the data reported in this work is available from the lead contact upon request.

ACKNOWLEDGMENTS

We thank M. Genet for advice, M. Afanassieff (Stem Cell and Brain Research Institute, Lyon, France) for the rabbit stem cell line, I. de la Rosa Velazquez (Helmholtz Munich Genomics Facility) for sequencing, C. Hermant and A. Burton for critical reading of the manuscript, and T. Yamada and Torres-Padilla lab members for helpful suggestions. M.-E.T.-P. acknowledges funding from the Helmholtz Association and the German Research Foundation (DFG) with project ID 213249687 (SFB 1064). P.S. acknowledges support from the Infrastructure Nationale en Biologie et Sante INGESTEM (ANR-11-INBS-0009). M.E.O. is an EMBO long-term fellow (ALTF 91-2021). M.K. is a fellow of the Leducq Network 23CVD01 "XenExoCor."

AUTHOR CONTRIBUTIONS

M.E.O. performed experimental work and computational analyses. D.R.-T. developed Smart-seq+5′ and prepared embryo libraries. M.K. and T.N. performed microinjections and embryo collection. V.Z., K.S., C.N., and I.A. collected embryos. L.M. performed TF deletion experiments. C.M.M.-D. performed phylogenetic analyses. P.S., J.G., E.W., and H.K. provided essential study support. M.E.O. and M.-E.T.-P. wrote the manuscript with input from all authors. M.-E.T.-P. conceived and supervised the study.

DECLARATION OF INTERESTS

M.-E.T.-P. is member of the Advisory Ethics Panel of Merck.

STAR★METHODS

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

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - ○ Mouse embryo collection
  - ○ Pig and cow embryo collection
  - ○ Rabbit GV oocytes and in vivo preimplantation stage embryos
  - ○ Rhesus macaque embryo collection
  - ○ mESC cell culture
  - ○ rbESC cell culture
- METHOD DETAILS
  - ○ Embryo collection for Smart-seq+5'
  - ○ Stem cell collection for Smart-seq+5'
  - ○ Preparation of Smart-seq+5' samples
  - ○ Smart-seq2 with poly-adenylation step
  - ○ MLT1A0-reporter assay in stem cells
  - ○ LTR reporter assay in mouse embryos
  - ○ LTR reporter assay in bovine embryos
  - ○ Mapping and downstream processing of Smart-seq+5' libraries
  - ○ Mapping and analysis of Smart-seq2 libraries
  - ○ Analysis of CAGE data
  - ○ EGA gene candidate selection
  - ○ TSS profiling of TEs
  - ○ TE age analysis
  - ○ TElocal MTL1A0 analysis
  - ○ Phylogeny maximum likelihood analysis
  - ○ De novo motif search on TEs
  - ○ Chimeric TE-gene interaction analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2024.12.013.

Received: April 11, 2023
Revised: October 26, 2024
Accepted: December 10, 2024
Published: January 20, 2025

REFERENCES

1. Aoki, F., Worrad, D.M., and Schultz, R.M. (1997). Regulation of transcriptional activity during the first and second cell cycles in the preimplantation mouse embryo. Dev. Biol. *181*, 296–307. https://doi.org/10.1006/dbio.1996.8466.

2. Jukam, D., Shariati, S.A.M., and Skotheim, J.M. (2017). Zygotic genome activation in vertebrates. Dev. Cell *42*, 316–332. https://doi.org/10.1016/j.devcel.2017.07.026.

3. Reik, W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. Nature *447*, 425–432. https://doi.org/10.1038/nature05918.

4. Burton, A., and Torres-Padilla, M.-E. (2014). Chromatin dynamics in the regulation of cell fate allocation during early embryogenesis. Nat. Rev. Mol. Cell Biol. *15*, 723–734. https://doi.org/10.1038/nrm3885.

5. Xia, W., and Xie, W. (2020). Rebooting the Epigenomes during Mammalian Early Embryogenesis. Stem Cell Rep. *15*, 1158–1175. https://doi.org/10.1016/j.stemcr.2020.09.005.

6. De Iaco, A., Planet, E., Coluccio, A., Verp, S., Duc, J., and Trono, D. (2017). DUX-family transcription factors regulate zygotic genome activation in placental mammals. Nat. Genet. *49*, 941–945. https://doi.org/10.1038/ng.3858.

7. Rodriguez-Terrones, D., and Torres-Padilla, M.-E. (2018). Nimble and ready to mingle: transposon outbursts of early development. Trends Genet. *34*, 806–820. https://doi.org/10.1016/j.tig.2018.06.006.

8. Hendrickson, P.G., Doráis, J.A., Grow, E.J., Whiddon, J.L., Lim, J.-W., Wike, C.L., Weaver, B.D., Pflueger, C., Emery, B.R., Wilcox, A.L., et al. (2017). Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. Nat. Genet. *49*, 925–934. https://doi.org/10.1038/ng.3844.

9. Sakashita, A., Kitano, T., Ishizu, H., Guo, Y., Masuda, H., Ariura, M., Murano, K., and Siomi, H. (2023). Transcription of MERVL retrotransposons is required for preimplantation embryo development. Nat. Genet. *55*, 484–495. https://doi.org/10.1038/s41588-023-01324-y.

10. Jachowicz, J.W., Bing, X., Pontabry, J., Bošković, A., Rando, O.J., and Torres-Padilla, M.E. (2017). LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. Nat. Genet. *49*, 1502–1510. https://doi.org/10.1038/ng.3945.

11. Fadloun, A., Le Gras, S., Jost, B., Ziegler-Birling, C., Takahashi, H., Gorab, E., Carninci, P., and Torres-Padilla, M.-E. (2013). Chromatin signatures and retrotransposon profiling in mouse embryos reveal regulation of LINE-1 by RNA. Nat. Struct. Mol. Biol. *20*, 332–338. https://doi.org/10.1038/nsmb.2495.

12. Hermant, C., and Torres-Padilla, M.-E. (2021). TFs for TEs: the transcription factor repertoire of mammalian transposable elements. Genes Dev. *35*, 22–39. https://doi.org/10.1101/gad.344473.120.

13. Wells, J.N., and Feschotte, C. (2020). A field guide to eukaryotic transposable elements. Annu. Rev. Genet. *54*, 539–561. https://doi.org/10.1146/annurev-genet-040620-022145.

14. Senft, A.D., and Macfarlan, T.S. (2021). Transposable elements shape the evolution of mammalian development. Nat. Rev. Genet. *22*, 691–711. https://doi.org/10.1038/s41576-021-00385-1.

15. Garcia-Perez, J.L., Widmann, T.J., and Adams, I.R. (2016). The impact of transposable elements on mammalian development. Development *143*, 4101–4114. https://doi.org/10.1242/dev.132639.

16. Peaston, A.E., Evsikov, A.V., Graber, J.H., de Vries, W.N., Holbrook, A.E., Solter, D., and Knowles, B.B. (2004). Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. Dev. Cell *7*, 597–606. https://doi.org/10.1016/j.devcel.2004.09.004.

17. Franke, V., Ganesh, S., Karlic, R., Malik, R., Pasulka, J., Horvat, F., Kuzman, M., Fulka, H., Cernohorska, M., Urbanova, J., et al. (2017). Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes. Genome Res. *27*, 1384–1394. https://doi.org/10.1101/gr.216150.116.

18. Göke, J., Lu, X., Chan, Y.-S., Ng, H.-H., Ly, L.-H., Sachs, F., and Szczerbinska, I. (2015). Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. Cell Stem Cell *16*, 135–141. https://doi.org/10.1016/j.stem.2015.01.005.

19. Lanciano, S., and Cristofari, G. (2020). Measuring and interpreting transposable element expression. Nat. Rev. Genet. *21*, 721–736. https://doi.org/10.1038/s41576-020-0251-y.

20. Picelli, S., Björklund, Å.K., Faridani, O.R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat. Methods *10*, 1096–1098. https://doi.org/10.1038/nmeth.2639.

21. Picelli, S., Faridani, O.R., Björklund, A.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. *9*, 171–181. https://doi.org/10.1038/nprot.2014.006.

22. Rebollo, R., Farivar, S., and Mager, D.L. (2012). C-GATE - Catalogue of genes affected by transposable elements. Mobile DNA *3*, 9. https://doi.org/10.1186/1759-8753-3-9.

23. Takahashi, H., Kato, S., Murata, M., and Carninci, P. (2012). CAGE (Cap Analysis of Gene Expression): A Protocol for the Detection of Promoter and Transcriptional Networks. Methods Mol. Biol. *786*, 181–200. https://doi.org/10.1007/978-1-61779-292-2_11.

24. Bagnoli, J.W., Ziegenhain, C., Janjic, A., Wange, L.E., Vieth, B., Parekh, S., Geuder, J., Hellmann, I., and Enard, W. (2018). Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. Nat. Commun. *9*, 2937. https://doi.org/10.1038/s41467-018-05347-6.

25. 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. *46*, 11502–11513. https://doi.org/10.1093/nar/gky817.

26. Deng, Q., Ramsköld, D., Reinius, B., and Sandberg, R. (2014). Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells. Science *343*, 193–196. https://doi.org/10.1126/science.1245316.

27. Nepal, C., Hadzhiev, Y., Previti, C., Haberle, V., Li, N., Takahashi, H., Suzuki, A.M.M., Sheng, Y., Abdelhamid, R.F., Anand, S., et al. (2013). Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis. Genome Res. *23*, 1938–1950. https://doi.org/10.1101/gr.153692.112.

28. Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A.M., Taylor, M.S., Engström, P.G., Frith, M.C., et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet. *38*, 626–635. https://doi.org/10.1038/ng1789.

29. Zhai, Y., Yu, H., An, X., Zhang, Z., Zhang, M., Zhang, S., Li, Q., and Li, Z. (2022). Profiling the transcriptomic signatures and identifying the patterns of zygotic genome activation – a comparative analysis between early porcine embryos and their counterparts in other three mammalian species. BMC Genomics *23*, 772. https://doi.org/10.1186/s12864-022-09015-4.

30. Brunet-Simon, A., Henrion, G., Renard, J.P., and Duranthon, V. (2001). Onset of Zygotic Transcription and Maternal Transcript Legacy in the Rabbit embryo. Mol. Reprod. Dev. *58*, 127–136. https://doi.org/10.1002/1098-2795(200102)58:2<127::AID-MRD1>3.0.CO;2-A.

31. Wolf, D.P. (2004). Assisted reproductive technologies in rhesus macaques. Reprod. Biol. Endocrinol. *2*, 37. https://doi.org/10.1186/1477-7827-2-37.

32. Graf, A., Krebs, S., Zakhartchenko, V., Schwalb, B., Blum, H., and Wolf, E. (2014). Fine mapping of genome activation in bovine embryos by RNA sequencing. Proc. Natl. Acad. Sci. USA *111*, 4139–4144. https://doi.org/10.1073/pnas.1321569111.

33. Osmanski, A.B., Paulat, N.S., Korstian, J., Grimshaw, J.R., Halsey, M., Sullivan, K.A.M., Moreno-Santillán, D.D., Crookshanks, C., Roberts, J., Garcia, C., et al. (2023). Insights into mammalian TE diversity through the curation of 248 genome assemblies. Science *380*, eabn1430. https://doi.org/10.1126/science.abn1430.

34. Storer, J., Hubley, R., Rosen, J., Wheeler, T.J., and Smit, A.F. (2021). The Dfam community resource of transposable element families, sequence models, and genome annotations. Mobile DNA *12*, 2. https://doi.org/10.1186/s13100-020-00230-y.

35. Lee, M.T., Bonneau, A.R., and Giraldez, A.J. (2014). Zygotic genome activation during the maternal-to-zygotic transition. Annu. Rev. Cell Dev. Biol. *30*, 581–613. https://doi.org/10.1146/annurev-cellbio-100913-013027.

36. Henderson, G.R.W., Brahmasani, S.R., Yelisetti, U.M., Konijeti, S., Katari, V.C., and Sisinthy, S. (2014). Candidate gene expression patterns in rabbit preimplantation embryos developed in vivo and in vitro. J. Assist. Reprod. Genet. *31*, 899–911. https://doi.org/10.1007/s10815-014-0233-0.

37. Park, S.-J., Komata, M., Inoue, F., Yamada, K., Nakai, K., Ohsugi, M., and Shirahige, K. (2013). Inferring the choreography of parental genomes during fertilization from ultralarge-scale whole-transcriptome analysis. Genes Dev. *27*, 2736–2748. https://doi.org/10.1101/gad.227926.113.

38. Khan, D.R., Dubé, D., Gall, L., Peynot, N., Ruffini, S., Laffont, L., Le Bourhis, D., Degrelle, S., Jouneau, A., and Duranthon, V. (2012). Expression of Pluripotency Master Regulators during Two Key Developmental Transitions: EGA and Early Lineage Specification in the Bovine Embryo. PLoS One *7*, e34110. https://doi.org/10.1371/journal.pone.0034110.

39. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.

40. Ji, S., Chen, F., Stein, P., Wang, J., Zhou, Z., Wang, L., Zhao, Q., Lin, Z., Liu, B., Xu, K., et al. (2023). OBOX regulates mouse zygotic genome

activation and early development. Nature *620*, 1047–1053. https://doi.org/10.1038/s41586-023-06428-3.

41. Zou, Z., Zhang, C., Wang, Q., Hou, Z., Xiong, Z., Kong, F., Wang, Q., Song, J., Liu, B., Liu, B., et al. (2022). Translatome and transcriptome co-profiling reveals a role of TPRXs in human zygotic genome activation. Science *378*, abo7923. https://doi.org/10.1126/science.abo7923.

42. Borodulina, O.R., Golubchikova, J.S., Ustyantsev, I.G., and Kramerov, D.A. (2016). Polyadenylation of RNA transcribed from mammalian SINEs by RNA polymerase III: complex requirements for nucleotide sequences. Biochim. Biophys. Acta *1859*, 355–365. https://doi.org/10.1016/j.bbagrm.2015.12.003.

43. Ustyantsev, I.G., Borodulina, O.R., and Kramerov, D.A. (2021). Identification of nucleotide sequences and some proteins involved in polyadenylation of RNA transcribed by Pol III from SINEs. RNA Biol. *18*, 1475–1488. https://doi.org/10.1080/15476286.2020.1857942.

44. Platt, R.N., Vandewege, M.W., and Ray, D.A. (2018). Mammalian transposable elements and their impacts on genome evolution. Chromosome Res. *26*, 25–43. https://doi.org/10.1007/s10577-017-9570-z.

45. 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. https://doi.org/10.1038/nrg.2016.139.

46. Kigami, D., Minami, N., Takayama, H., and Imai, H. (2003). MuERV-L is one of the earliest transcribed genes in mouse one-cell embryos. Biol. Reprod. *68*, 651–654. https://doi.org/10.1095/biolreprod.102.007906.

47. Petropoulos, S., Edsgärd, D., Reinius, B., Deng, Q., Panula, S.P., Codeluppi, S., Plaza Reyes, A., Linnarsson, S., Sandberg, R., and Lanner, F. (2016). Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. Cell *165*, 1012–1026. https://doi.org/10.1016/j.cell.2016.03.023.

48. Töhönen, V., Katayama, S., Vesterlund, L., Jouhilahti, E.-M., Sheikhi, M., Madissoon, E., Filippini-Cattaneo, G., Jaconi, M., Johnsson, A., Bürglin, T.R., et al. (2015). Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development. Nat. Commun. *6*, 8207. https://doi.org/10.1038/ncomms9207.

49. Haberle, V., and Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. Nat. Rev. Mol. Cell Biol. *19*, 621–637. https://doi.org/10.1038/s41580-018-0028-8.

50. Ohnishi, Y., Totoki, Y., Toyoda, A., Watanabe, T., Yamamoto, Y., Tokunaga, K., Sakaki, Y., Sasaki, H., and Hohjoh, H. (2012). Active role of small non-coding RNAs derived from SINE/B1 retrotransposon during early mouse development. Mol. Biol. Rep. *39*, 903–909. https://doi.org/10.1007/s11033-011-0815-1.

51. Speek, M. (2001). Antisense Promoter of Human L1 Retrotransposon Drives Transcription of Adjacent Cellular Genes. Mol. Cell. Biol. *21*, 1973–1985. https://doi.org/10.1128/MCB.21.6.1973-1985.2001.

52. Cruickshanks, H.A., and Tufarelli, C. (2009). Isolation of cancer-specific chimeric transcripts induced by hypomethylation of the LINE-1 antisense promoter. Genomics *94*, 397–406. https://doi.org/10.1016/j.ygeno.2009.08.013.

53. Li, J., Kannan, M., Trivett, A.L., Liao, H., Wu, X., Akagi, K., and Symer, D.E. (2014). An antisense promoter in mouse L1 retrotransposon open reading frame-1 initiates expression of diverse fusion transcripts and limits retrotransposition. Nucleic Acids Res. *42*, 4546–4562. https://doi.org/10.1093/nar/gku091.

54. Pace, J.K., and Feschotte, C. (2007). The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. Genome Res. *17*, 422–432. https://doi.org/10.1101/gr.5826307.

55. Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. (2017). TimeTree: A resource for timelines, timetrees, and divergence times. Mol. Biol. Evol. *34*, 1812–1819. https://doi.org/10.1093/molbev/msx116.

56. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348.

57. Edgar, R.C. (2004). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113. https://doi.org/10.1186/1471-2105-5-113.

58. Edgar, R.C. (2022). Muscle5: high-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. Nat. Commun. *13*, 6968. https://doi.org/10.1038/s41467-022-34630-w.

59. Bininda-Emonds, O.R.P., Cardillo, M., Jones, K.E., MacPhee, R.D.E., Beck, R.M.D., Grenyer, R., Price, S.A., Vos, R.A., Gittleman, J.L., and Purvis, A. (2007). The delayed rise of present-day mammals. Nature *446*, 507–512. https://doi.org/10.1038/nature05634.

60. Oliveira, D.S., Fablet, M., Larue, A., Vallier, A., Carareto, C.M.A., Rebollo, R., and Vieira, C. (2023). ChimeraTE: a pipeline to detect chimeric transcripts derived from genes and transposable elements. Nucleic Acids Res. *51*, 9764–9784. https://doi.org/10.1093/nar/gkad671.

61. Chen, Y., Sim, A., Wan, Y.K., Yeo, K., Lee, J.J.X., Ling, M.H., Love, M.I., and Göke, J. (2023). Context-aware transcript quantification from long-read RNA-seq data with Bambu. Nat. Methods *20*, 1187–1195. https://doi.org/10.1038/s41592-023-01908-w.

62. Demircioğlu, D., Cukuroglu, E., Kindermans, M., Nandi, T., Calabrese, C., Fonseca, N.A., Kahles, A., Van Lehmann, K.V., Stegle, O., Brazma, A., et al. (2019). A pan-cancer transcriptome analysis reveals pervasive regulation through alternative promoters. Cell *178*, 1465–1477.e17. https://doi.org/10.1016/j.cell.2019.08.018.

63. Fueyo, R., Judd, J., Feschotte, C., and Wysocka, J. (2022). Roles of transposable elements in the regulation of mammalian transcription. Nat. Rev. Mol. Cell Biol. *23*, 481–497. https://doi.org/10.1038/s41580-022-00457-y.

64. Oomen, M.E., and Torres-Padilla, M.E. (2024). Jump-starting life: balancing transposable element co-option and genome integrity in the developing mammalian embryo. EMBO Rep. *25*, 1721–1733. https://doi.org/10.1038/s44319-024-00118-5.

65. Almeida, M.V., Vernaz, G., Putman, A.L.K., and Miska, E.A. (2022). Taming transposable elements in vertebrates: from epigenetic silencing to domestication. Trends Genet. *38*, 529–553. https://doi.org/10.1016/j.tig.2022.02.009.

66. Friedli, M., and Trono, D. (2015). The developmental control of transposable elements and the evolution of higher species. Annu. Rev. Cell Dev. Biol. *31*, 429–451. https://doi.org/10.1146/annurev-cellbio-100814-125514.

67. Haig, D. (2016). Transposable elements: self-seekers of the germline, team-players of the soma. BioEssays *38*, 1158–1166. https://doi.org/10.1002/bies.201600125.

68. Halstead, M.M., Ma, X., Zhou, C., Schultz, R.M., and Ross, P.J. (2020). Chromatin remodeling in bovine embryos indicates species-specific regulation of genome activation. Nat. Commun. *11*, 4654. https://doi.org/10.1038/s41467-020-18508-3.

69. Genet, M., and Torres-Padilla, M.-E. (2020). The molecular and cellular features of 2-cell-like cells: a reference guide. Development *147*, dev189688. https://doi.org/10.1242/dev.189688.

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

71. Rodriguez-Terrones, D., Gaume, X., Ishiuchi, T., Weiss, A., Kopp, A., Kruse, K., Penning, A., Vaquerizas, J.M., Brino, L., and Torres-Padilla, M.-E. (2018). A molecular roadmap for the emergence of early-embryonic-like cells in culture. Nat. Genet. *50*, 106–119. https://doi.org/10.1038/s41588-017-0016-5.

72. Taubenschmid-Stowers, J., Rostovskaya, M., Santos, F., Ljung, S., Argelaguet, R., Krueger, F., Nichols, J., and Reik, W. (2022). 8C-like cells

capture the human zygotic genome activation program in vitro. Cell Stem Cell *29*, 449–459.e6. https://doi.org/10.1016/j.stem.2022.01.014.

73. Festuccia, N., Vandormael-Pournin, S., Chervova, A., Geiselmann, A., Langa-Vives, F., Coux, R.X., Gonzalez, I., Collet, G.G., Cohen-Tannoudji, M., and Navarro, P. (2024). Nr5a2 is dispensable for zygotic genome activation but essential for morula development. Science *386*, eadg7325. https://doi.org/10.1126/science.adg7325.

74. Miyanari, Y., and Torres-Padilla, M.E. (2012). Control of ground-state pluripotency by allelic regulation of Nanog. Nature *483*, 470–473. https://doi.org/10.1038/nature10807.

75. Osteil, P., Moulin, A., Santamaria, C., Joly, T., Jouneau, L., Aubry, M., Tapponnier, Y., Archilla, C., Schmaltz-Panneau, B., Lecardonnel, J., et al. (2016). A panel of embryonic stem cell lines reveals the variety and dynamic of pluripotent states in rabbits. Stem Cell Rep. *7*, 383–398. https://doi.org/10.1016/j.stemcr.2016.07.022.

76. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/BIOINFORMATICS/BTU170.

77. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

78. 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. https://doi.org/10.1093/bioinformatics/btv422.

79. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Camb) *2*, 100141. https://doi.org/10.1016/J.XINN.2021.100141.

80. 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. https://doi.org/10.1093/nar/gkw257.

81. Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274. https://doi.org/10.1093/molbev/msu300.

82. 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. https://doi.org/10.1016/j.molcel.2010.05.004.

83. Simmet, K., Zakhartchenko, V., Philippou-Massier, J., Blum, H., Klymiuk, N., and Wolf, E. (2018). OCT4/POU5F1 is required for NANOG expression in bovine blastocysts. Proc. Natl. Acad. Sci. USA *115*, 2770–2775. https://doi.org/10.1073/pnas.1718833115.

84. Kurome, M., Kessler, B., Wuensch, A., Nagashima, H., and Wolf, E. (2015). Nuclear transfer and transgenesis in the pig. Methods Mol. Biol. *1222*, 37–59. https://doi.org/10.1007/978-1-4939-1594-1_4.

85. Yuan, Y., Spate, L.D., Redel, B.K., Tian, Y., Zhou, J., Prather, R.S., and Roberts, R.M. (2017). Quadrupling efficiency in production of genetically modified pigs through improved oocyte maturation. Proc. Natl. Acad. Sci. USA *114*, E5796–E5804. https://doi.org/10.1073/pnas.1703998114.

86. Tachibana, M., Sparman, M., Ramsey, C., Ma, H., Lee, H.S., Penedo, M.C.T., and Mitalipov, S. (2012). Generation of chimeric rhesus monkeys. Cell *148*, 285–295. https://doi.org/10.1016/J.CELL.2011.12.007.

87. Aksoy, I., Rognard, C., Moulin, A., Marcy, G., Masfaraud, E., Wianny, F., Cortay, V., Bellemin-Ménard, A., Doerflinger, N., Dirheimer, M., et al. (2021). Apoptosis, G1 Phase Stall, and Premature Differentiation Account for Low Chimeric Competence of Human and Rhesus Monkey Naive Pluripotent Stem Cells. Stem Cell Rep. *16*, 56–74. https://doi.org/10.1016/j.stemcr.2020.12.004.

88. Lam, A.J., St-Pierre, F., Gong, Y., Marshall, J.D., Cranfill, P.J., Baird, M.A., McKeown, M.R., Wiedenmann, J., Davidson, M.W., Schnitzer, M.J., et al. (2012). Improving FRET dynamic range with bright green and red fluorescent proteins. Nat. Methods *9*, 1005–1012. https://doi.org/10.1038/nmeth.2171.

89. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/GIGASCIENCE/GIAB008.

90. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

91. Nassar, L.R., Barber, G.P., Benet-Pagès, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J.N., Hinrichs, A.S., Lee, B.T., et al. (2023). The UCSC Genome Browser database: 2023 update. Nucleic Acids Res. *51*, D1188–D1195. https://doi.org/10.1093/nar/gkac1072.

92. Amezquita, R.A., Lun, A.T.L., Becht, E., Carey, V.J., Carpp, L.N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., et al. (2020). Orchestrating single-cell analysis with Bioconductor. Nat. Methods *17*, 137–145. https://doi.org/10.1038/s41592-019-0654-x.

93. Stacklies, W., Redestig, H., Scholz, M., Walther, D., and Selbig, J. (2007). pcaMethods—a bioconductor package providing PCA methods for incomplete data. Bioinformatics *23*, 1164–1167. https://doi.org/10.1093/BIOINFORMATICS/BTM069.

94. Hahne, F., and Ivanek, R. (2016). Visualizing genomic data using Gviz and bioconductor. Methods Mol. Biol. *1418*, 335–351. https://doi.org/10.1007/978-1-4939-3578-9_16.

95. Wickham, H. (2016). ggplot2 (Springer International Publishing) https://doi.org/10.1007/978-3-319-24277-4.

96. Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191. https://doi.org/10.1038/nprot.2009.97.

97. Lin, Y., Ghazanfar, S., Wang, K.Y.X., Gagnon-Bartsch, J.A., Lo, K.K., Su, X., Han, Z.G., Ormerod, J.T., Speed, T.P., Yang, P., et al. (2019). ScMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. Proc. Natl. Acad. Sci. USA *116*, 9775–9784. https://doi.org/10.1073/pnas.1820006116.

98. Abugessaisa, I., Noguchi, S., Hasegawa, A., Kondo, A., Kawaji, H., Carninci, P., and Kasukawa, T. (2019). refTSS: A reference data set for human and mouse transcription start sites. J. Mol. Biol. *431*, 2407–2422. https://doi.org/10.1016/j.jmb.2019.04.045.

99. Storer, J.M., Hubley, R., Rosen, J., and Smit, A.F.A. (2021). Curation Guidelines for de novo Generated Transposable Element Families. Curr. Protoc. *1*, e154. https://doi.org/10.1002/cpz1.154.

100. RepeatMasker Home Page. https://www.repeatmasker.org/.

101. Letunic, I., and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics *23*, 127–128. https://doi.org/10.1093/bioinformatics/btl529.

102. Park, S.J., Shirahige, K., Ohsugi, M., and Nakai, K. (2015). DBTMEE: A database of transcriptome in mouse early embryos. Nucleic Acids Res. *43*, D771–D776. https://doi.org/10.1093/nar/gku1001.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
E. coli competent cells	Torres-Padilla lab	N/A
Biological samples		
Mouse preimplantation embryos (oocyte to 16-cell)	Torres-Padilla lab	N/A
Pig preimplantation embryos (oocyte to Morula)	Wolf lab	N/A
Cow preimplantation embryos (oocyte to Morula)	Wolf lab	N/A
Rabbit preimplantation embryos (oocyte to Morula)	Wolf lab	N/A
Rhesus Macaque preimplantation embryos (oocyte to 16-cell)	Savatier lab	N/A
Chemicals, peptides, and recombinant proteins		
10x lysis buffer	Clontech	635013
Hyaluronidase	Sigma	H3506
0.1% polyvinylalcohol	Sigma	P8136
DMEM-Glutamax	Gibco	31966047
Fetal Bovine Serum for mESC culture	Panbiotech	P303302
Non-essential Amino Acids	ThermoScientific	11140-035
recombinant hLIF	Produced in-house	N/A
penicillin–streptomycin	Gibco	15070063
2-mercaptoethanol	ThermoScientific	31350010
CHIR99021	Cayman	13122-25
PD0325901	Miltenyi	130-106-541
Gelatin	Panbiotech	P06-20410
DMEM/F12 media	ThermoScientific	21331020
KOSR	LifeTechnologies	10828-028
Fetal Bovine Serum for rbESC culture	Gibco	10270-106
Sodium pyruvate	LifeTechnologies	11360-039
Puromycin	Sigma	P8833
Accutase	Sigma	A6964
Mitomycin-C	Sigma	M4287
dNTP mix	ThermoFisher	R0192
RNAse inhibitor	TAKARA	2313A
Superscript II RT	ThermoFisher	18064014
Betaine	Sigma	B0300
PEG-8000	Sigma	P1458
HiFi ReadyMix	KAPA	KM2605
EVAGreen	Biotium	31000
Poly(A) polymerase	NEB	M0276
Dextran 488	Thermo	D34682
Critical commercial assays		
Q5 mutagenesis kit	NEB	E0554S
NucleoSpin Plasmid EasyPure, Mini Kit	Macherey-Nagel	740727.50
AMPure RNA magnetic beads	Beckman Coulter	A63987
AMPure XP DNA magnetic beads	Beckman Coulter	A63881
Nextera XT	Illumina	15032354

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
ReliaPrep RNA extraction kit	Promega	Z6011
Q5 mutagenesis kit	NEB	E0554S
NucleoSpin Plasmid EasyPure, Mini Kit	Macherey-Nagel	740727.50
NucleoBond Xtra Midi Kit	Macherey-Nagel	740410.50
JetPrime transfection reagent	Polyplus	101000015
Deposited data		
mESC CAGE data	Lloret-Llinares et al.[25]	GSE115727
Mouse Embryo Smart-seq2 data	Deng et al.[26]	GSE45719
All Smart-seq+5, Smart-seq2 and Smart-seq2-polyA data generated in this study	This study	GSE225056
Experimental models: Cell lines		
Mouse: E14 embryonic stem cells	Miyanari and Torres-Padilla[74]	RRID:CVCL_C320
Rabbit ESC cell line AKSL20-EOS	Osteil et al.[75]	N/A
Mouse embryonic fibroblast cells	Derived from mouse embryos in Torres-Padilla lab	N/A
Oligonucleotides		
ERCC RNA spike ins	Ambion	4456653
Oligo-dT30	Sigma	https://github.com/meoomen/Smartseq5
rGrG+G TSO	TIB MolBiol	https://github.com/meoomen/Smartseq5
Sequencing adaptor mix	IDT	https://github.com/meoomen/Smartseq5
Primers for Q5 mutagenesis	Eurofins	N/A
Species-specific MLT1A0 sequences (see Figure 6B)	Synthesized by Eurofins	N/A
Recombinant DNA		
Plasmid pcDNA3-mRuby2 (#40260)	Addgene	40260
Software and algorithms		
Trimmomatic	Bolger et al.[76]	https://anaconda.org/bioconda/trimmomatic
STAR	Dobin et al.[77]	https://anaconda.org/bioconda/star
TEtranscripts & TElocal	Jin et al.[78]	https://hammelllab.labsites.cshl.edu/software/
DEseq2	Love et al.[39]	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
clusterProfiler	Wu et al.[79]	https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html
Deeptools	Ramírez et al.[80]	https://anaconda.org/bioconda/deeptools
Muscle	Edgar et al.[57,58]	https://www.drive5.com/muscle/
Trimal	Capalla-Gutierrez et al.[56]	https://anaconda.org/bioconda/trimal
IQ-TREE	Nguyen et al.[81]	http://www.iqtree.org
HOMER	Heinz et al.[82]	http://homer.ucsd.edu/homer/
ChimeraTE	Oliveira et al.[60]	https://github.com/OliveiraDS-hub/ChimeraTE
Bambu	Chen et al.[61]	https://github.com/GoekeLab/bambu
proActive	Demircioğlu et al.[62]	https://github.com/GoekeLab/proActiv
Other		
Resource for computational code for this paper	Torres-Padilla lab	https://github.com/meoomen/Smartseq5
TE annotations (as listed in Table S5)	Dfam consortium	https://github.com/Dfam-consortium/FamDB
Genomes (as listed in Table S5)	NCBI	https://ncbi.nlm.nih.gov/datasets/genome/
Gene annotations (as listed in Table S5)	Ensembl	https://jul2019.archive.ensembl.org/

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Mouse embryo collection

C57BL6J mouse embryos were collected from natural mating and washed in PBS, prior to transfer into single-embryo aliquots in Smart-seq+5' lysis buffer mix. Embryos were snap frozen and stored at -80°C until further processing.

Pig and cow embryo collection

Pig and cow embryos were produced in vitro as described previously.[83–85] Briefly, cumulus-ooycte-complexes (COCs) were aspirated from abattoir derived ovaries and for sampling of germinal vesicle (GV) stage oocytes, cumulus cells were immediately removed by repeated pipetting in medium containing 0.1 mg/mL hyaluronidase. For in vitro fertilization, COCs were matured and subsequently fertilized with frozen-thawed semen from a commercially available Duroc boar and a Fleckvieh bull, respectively. In pig, zygotes, 2-cell and 4-cell stage embryos and morulae were collected 10h, 25h, 34h and 97h after fertilization, respectively. Due to a high lipid content, it is not possible to doubtlessly identify 8-cell and 16-cell embryos. Therefore, we preselected pig 2-cell stage embryos at 25h and subsequently 4-cell stage embryos at 34h after fertilization to avoid sampling of abnormal embryos. From this population, we collected 5-8-cell embryos 48h and embryos consisting of approximately 16 cells 72h after fertilization. In cow, zygotes, 2-cell and 4-cell stage embryos were collected at 18h, 22-27h and 40-42h after fertilization, respectively. Similar to the sampling of pig embryos, 2-cell stage and 4-cell stage embryos were preselected at 27h and 42h. From this population, 5-8-cell and approximately 16-cell embryos and morulae were collected at 48-50h, 72h and 120h after fertilization, respectively. Finally, pig and cow samples were washed twice in PBS supplemented with 0.1% polyvinylalcohol (PVA) and once in PBS without Mg^{2+}/Ca^{2+} before transfer to Smart-seq+5' lysis buffer. After shock-freeze in liquid nitrogen, samples were stored at -80° C until further analysis.

Rabbit GV oocytes and in vivo preimplantation stage embryos

Non-superovulated Zika female rabbits served as donors of GV oocytes and in vivo embryos. GV oocytes were obtained by follicular aspiration of the ovaries. In vivo embryos were obtained by natural mating of donor females with Zika males and collection at appropriate hours post-mating (hpm); zygote (18-20 hpm), 2-cell (24-26 hpm), 4-cell (30-32 hpm), 8-cell (38-40 hpm), 16-cell late (46-48 hpm) and morula (54-56 hpm). Embryos were flushed from the oviducts of donor females using PBS supplemented with 4 mg bovine serum albumin (BSA). Finally, isolated rabbit embryos were washed twice in PBS supplemented with 0.1% PVA and once in PBS without Mg^{2+}/Ca^{2+} before transfer to Smart-seq+5' lysis buffer mix. After shock-freeze in liquid nitrogen, samples were stored at -80° C until further analysis.

Rhesus macaque embryo collection

Procedures for superovulation, oocytes collection, fertilization and embryo culture were performed as previously described.[31,86,87] In brief, healthy female rhesus monkeys with regular menstrual cycles were selected for superovulation. Starting at days 1–4 of the menstrual cycle, females received twice-daily injections of recombinant human FSH for 8 days and recombinant human LH on days 7–8 of the stimulation protocol. In addition, animals received a GnRH antagonist and human chorionic gonadotropin on day 7 of the protocol, approximately 36hr prior to laparoscopic follicle aspiration and oocyte retrieval. Cumulus-oocyte complexes were collected from anesthetized females by laparoscopic follicular aspiration and placed in HEPES-buffered TALP (modified Tyrode solution with albumin, lactate, and pyruvate) containing 0.3% bovine serum albumin (TH3) at 37°C and treated with hyaluronidase to remove cumulus and granula cells. Oocytes were then placed in hamster embryo culture medium (HECM-9) covered with LP at 37°C in 5% CO_2, 5% O_2, and 90% N_2 and covered with liquid paraffin (LP) until use. Only oocytes at the metaphase II (MII) stage were used to perform intracytoplasmic sperm injection (ICSI). Injected oocytes were transferred in HECM-9 medium covered with LP and cultured at 37°C in 5% CO_2, 5% O_2, and 90% N_2. Embryos at the eight-cell stage were transferred to fresh plates of HECM-9 medium supplemented with 5% fetal bovine serum with medium change every other day.

mESC cell culture

For the purpose of the MLT1A0-Ruby reporter experiment, mouse E14 ESC line[74] was cultured in DMEM-Glutamax (Gibco # 31966047) containing 15% FBS (Panbiotech #P303302), NEAA (ThermoScientific #11140-035), recombinant hLIF (produced in-house), penicillin–streptomycin (Gibco #15070063), 0.1 mM 2-mercaptoethanol (ThermoScientific #31350010), 3 μM CHIR99021 (Cayman #13122-25) and 1 μM PD0325901 (Miltenyi #130-106-541) on plates coated with 0.1% gelatin in PBS (Panbiotech #P06-20410).

rbESC cell culture

RbESC cell line AKSL20-EOS[75] was used for the MLT1A0-Ruby reporter assay. Cells were grown in DMEM/F12 media (ThermoScientific #21331020) containing 10% KOSR (LifeTechnologies #10828-028), 10% FBS (Gibco #10270-106), 1% NEAA (ThermoScientific #11140-035), 1mM sodium pyruvate (LifeTechnologies #11360-039), 1x PSG (ThermoScientific #10378-016), 50mM 2-mercaptoethanol (ThermoScientific #31350010), 0.1% hLIF (gift from Savatier lab) and 0.5ng/mL puromycin (Sigma #P8833). Cells were cultured at 38°C in a gas-filled incubator chamber (5% CO_2, 5% O_2 and 90% N_2). Media was changed daily. Cells were passaged every 2-3 days using Accutase (Sigma #A6964). Cells were grown on a layer of mitomycin-C

(Sigma #M4287) treated mouse embryonic fibroblast (OF-1) feeder cells. OF-1 cells were grown in DMEM-Glutamax (Gibco # 31966047) supplemented with 10% FBS, 1% PSG and 1% NEAA prior to mitomycin treatment. Treated feeders were plated on 6-well plates coated with 0.1% gelatin in PBS (Panbiotech #P06-20410).

METHOD DETAILS

Embryo collection for Smart-seq+5’

Clontech 10X lysis buffer (Cat. #635013) was used for all embryo collections. ERCC RNA spike-ins were added at the moment of lysis buffer preparation at a final concentration of 1:581,000. Aliquots of this lysis buffer mix preparation were used throughout embryo collection of all species. Timing and mating strategies for all embryo collection can be found in Table S6.

Stem cell collection for Smart-seq+5’

Mouse and rabbit stem cells were harvested after 2-3 days of culture since passage (described above). rbESC were separated from feeder cells by flow cytometry using the *Oct4*-GFP label.[75] RNA was extracted from cells using the reliaprep kit (Promega #Z6011) and 250pg diluted in ultrapure MilliQ was used as input for the Smart-seq+5’ protocol.

Preparation of Smart-seq+5’ samples

Samples were prepared largely following the Smart-seq2 protocol[20,21] with several modifications. We specifically chose to adapt Smart-seq2 because of its high sensitivity, full-length transcript coverage and adequate throughput and because Smart-seq2 has been robustly used for early embryos to efficiently study the maternal-to-zygotic transition.[26,47] Briefly, following unfreezing of the embryo lysates at room temperature, RNA was purified using AMPure RNA magnetic beads (Beckman Coulter #A63987), washed once in 100 μL of 80% ethanol and resuspended in 3 μL of annealing mix containing 1 μL dNTP mix (ThermoFisher, R0192), 1 μL oligo-dT30 (Sigma, 5’-AAGCAGTGGTATCAACGCAGAGTACT30V-3’) at 10 μM and 1 μL nuclease free water with 5% RNAse inhibitor (Clontech 2313A). Following a 3 min incubation at 72° C, samples were held at 4° C or in ice. Afterwards, 7 μL of reverse transcription mix was added (2 μL Superscript II RT buffer (ThermoFisher, 18064014), 2 μL 5M Betaine solution (Sigma, B0300-1VL), 0.5 μL DTT, 0.5 μL Superscript II RT, 0.25 μL RNAse inhibitor (TAKARA 2313A), 0.1 μL rGrG+G TSO (TIB MolBiol, 5’-AAGCAGTGGT ATCAACGCAGAGTACATrGrG+G-3’) at 100 μM, 0.06 μL 1M $MgCl_2$ (Sigma, M1028) and 1.6 μL of 40% PEG-8000 solution (Sigma, P1458)). Samples were then incubated for 90 min at 42° C and the enzyme was afterwards inactivated at 70° C for 15 min. PCR was performed using KAPA HiFi ReadyMix (KM2605) for 14 cycles as described in the Smart-seq2 protocol and purification was carried out using 12.5 μL of AMPure beads. The optimal number of PCR cycles was determined beforehand by performing a qPCR on a representative sample using identical PCR conditions but supplemented with 1 μL EVAGreen (Biotium, 31000). Finally, tagmentation was carried out with the Nextera XT kit (Illumina, 15032354) as described in the Smart-seq2 protocol, with a few modifications; 2.5 μL of 120 pg/uL cDNA were used per reaction (for a total of 300 pg) and mixed with 2.5 μL of Amplicon Tagment Mix and 5 μL of Tagment DNA buffer. After incubation for 5 min at 55° C, 2.5 μL of NT buffer were pipetted into each sample and another incubation for 5 minutes at room temperature was performed. Finally, 5 μL of a custom sequencing adaptor mix (IDT) was added, containing two standard i5 and i7 Nextera Unique Double Indexes and an additional tailed i7 index with an overhang complementary to the Smart-seq2 adaptors (Index and overhang shown in bold: 5’-CAAGCAGAAGACGGCATACGAGAT**NNNNNNNN**GTCTCGTGGGCTCGGAGATG TGTATAAGAGACAG**AAGCAGTGGTATCAACGC***__A__*G-3’), all at 2.5 μM concentration. 7.5 μL of NPM mix was added and PCR was carried out for 12 cycles. Libraries were sequenced in a Hiseq4000 using 2 x 150 bp reads. All oligo sequences will be available for download through github code repository (https://github.com/meoomen/Smartseq5).

Smart-seq2 with poly-adenylation step

2-cell stage mouse embryos were collected in 10 μL Qiagen TCL lysis buffer (Cat. #1031576) and flash frozen. Upon thawing, RNA was purified using AMPure RNA magnetic beads (Beckman Coulter, A63987), washed once in 100 μL of 80% ethanol and resuspended in 10 μL of Polyadenylation mix containing 7 μL water, 0.5 μL E. coli Poly(A) polymerase (NEB, #M0276), 1 μL 10 Polymerase reaction buffer, 1 μL ATP and 0.5 μL RNAse inhibitor (Clontech 2313A). Following a 10 min incubation at 37° C, RNA was purified using magnetic beads once again and resuspended in annealing mix as per the standard Smart-seq2 protocol.[20,21]

MLT1A0-reporter assay in stem cells

Addgene Plasmid pcDNA3-mRuby2 (#40260)[88] containing mRuby was mutated using the Q5 mutagenesis kit (NEB # E0554S) to remove the full CMV promoter sequence. MLT1A0 species-specific consensus sequences were ordered from Eurofins and cloned into the CMV-less plasmid upstream of the mRuby2 sequence. TF motifs were removed from consensus sequence using Q5 site-directed mutagenesis kit (NEB #E0554S). Mini- and midipreps kits were used to isolate the plasmid (MN# 740727.50 and #740410.50). Plasmids were sequenced to ensure correct directionality and integrity of the inserted MLT1A0 sequences. Cells were transfected with 2μg of plasmid during passaging as described above using JetPrime (PolyPlus #101000015) and plated in individual wells of a 6-well plates. Media was exchanged for fresh media after 24hr post-transfection. 48hr after transfection, Ruby signals were measured by flow cytometry. For Ruby quantification, *Oct4*-GFP labelled rbESC cells[75] were separated from uncolored MEF feeder cells during analysis.

LTR reporter assay in mouse embryos

F1 female mice (C57BL/6J × CBA) were super-ovulated by intraperitoneal injection of 10 U of pregnant mare serum gonadotropin (PMSG) followed by 10 U of human chorionic gonadotropin (hCG) 48 hours later, and then mated with same genetic background of F1 male mice. Zygotes were collected at 18-19 hours post-hCG injection. 20 ng/μL MT2_Mm, MLT1A0, or negative control (no promoter) reporter plasmids were injected into zygotes at 18-19 hours post-hCG injection and embryos were cultured until the late 2-cell stage. At 46-48 hours post-hCG, the ruby signal was observed under a fluorescence microscope and the number of positive and negative 2-cell embryos were counted.

LTR reporter assay in bovine embryos

In vitro fertilized (IVF) bovine embryos were produced as described above for Smart-seq+5′ data generation. MLT1A0 or negative control reporter (no promoter) plasmids were diluted to a concentration of 10ng/μL with Tris-EDTA buffer (pH 7.4) containing dextran 488 (0.1mg/mL; Thermo # D34682) to detect successfully DNA-injected embryos. Cytoplasmic injection of the plasmids was performed on presumptive zygotes at 15 hours after IVF. After culturing for 3 days, FITC signal-positive 5-8-cell embryos were collected, and the ratio of Ruby-positive embryos over all injected embryos was quantified with N=2.

Mapping and downstream processing of Smart-seq+5' libraries

Sequence quality of all datasets was checked using FastQC and trimmed for adaptor sequences and low-quality sequence ends using trimmomatic.[76] Reads originating from 5' transcript ends or internal transcript fragments were sorted based on adaptor sequence using a custom python script available on github upon publication (https://github.com/meoomen/Smartseq5). Further downstream processing was done separately and in parallel for 5' reads and internal reads. Reads were mapped to the appropriate reference genome using STAR[77] and filtered for quality, secondary alignments, singletons and unmapped reads using samtools.[89] Two files were created for unique-mapping only reads and multimapping + unique-mapping reads to be used for further analyses as indicated below. Only Illumina-format read2 was used for downstream analyses of 5′ reads, as this contains the true 5' transcript information. rRNA reads were removed using repeatmasker annotated positions and tools provided by bedtools and picard. In order to assign multimapping reads, we used TEcount, TElocal and chimeraTE as described below. For downstream analysis using deeptools and bambu, we solely used unique mapping reads and excluded all multimapping reads. For quantification of gene and TE counts, we used TEcount from the TEtranscript toolkit.[78] To generate genome browser tracks, we used unique mapping reads only, using tools from ucsc bedgraphToBigWig and bedtools.[90,91] All reference genomes and genome annotations used in this study are listed in Table S6.[34] Several R/Bioconductor packages were used for the downstream processing and visualization of the data in R; SingleCellExperiment,[92] pcaMethods,[93] Gviz,[94] ggplot2,[95] biomaRt,[96] pheatmap and scMerge.[97]

Mapping and analysis of Smart-seq2 libraries

Smart-seq2 libraries from the study by Deng et al. were obtained from GEO access number GSE45719.[26] Both publicly available libraries as well as the in-house generated Smart-seq2 and Smart-seq2+polyA libraries were processed similar to Smart-seq+5' samples with few adaptations. As Smart-seq2 libraries only captures internal transcripts fragments, reads were not separated as 5' or internal fragments. Additionally, the Deng et al. dataset was sequenced as a single-end library and therefore mapped accordingly using STAR to the mouse genome mm10/GRCm38.

Analysis of CAGE data

Processed CAGE data of mESCs were downloaded from GEO (accession code GSE115727)[25] and signal was aggregated on the annotated TSS from refTSS[98] using deeptools.[80]

EGA gene candidate selection

EGA genes were selected using DEseq2[39] using the raw read count table containing only 5' fragments mapping to genes. Only genes containing more than 10 reads in at least 10 or more datapoints were used. An exemption was made for rhesus as fewer embryos were available for collection and genes containing more than 10 reads in at least 3 or more datapoints were used instead. To identify genes as EGA genes, we compared their expression between zygote to timing of EGA and from timing of EGA to the 16-cell/Morula stage. Specifically, EGA genes were selected as differential genes with a LFC of >1.5 from zygote to timing of EGA (mouse Late-2-cell; pig, cow and rhesus 8-cell; rabbit 4-cell) and padj<0.05, followed by a LFC<-0.25 and padj<0.1 from timing of EGA to the latest stage available in the dataset (Morula for pig, cow and rabbit, 16-cell for mouse and rhesus). As mouse has a very strong EGA profile, a cutoff of LFC<-0.5 was used for the selection of differential downregulated genes post-EGA for mouse EGA genes. These criteria were chosen because not all species have a clearly sharp and distinct EGA time. Instead, species like cow and rabbit undergo a slower ramp up of transcription activation. Thus, the criteria that we used to extract EGA genes, namely using the zygotes as pre-EGA reference point, is expected to include all EGA stages (and genes). Enriched GOterms were found using Enricher in the clusterProfiler toolbox.[79]

TSS profiling of TEs

Unique mapping 5' reads were used to visualize the TSS profile of TE inserts with deeptools computeMatrix, followed by plotting of signal using plotHeatmap and plotProfile.[80] Signal was plotted using scale-regions feature, setting the average full length TE size as regionBodylength, and a window of 500bp around the TE sequence. TE insertions with no signal were omitted from the analysis (using –skipZeros parameter) unless stated otherwise.

TE age analysis

Evolutionary age of TE families was extracted from the Dfam annotation (v3.8),[99] using the famdb tools available on github (https://github.com/Dfam-consortium/FamDB). The TimeTree database was used to estimate the evolutionary age for each category.[55]

TElocal MTL1A0 analysis

For the analysis of MLT1A0 expression of insertions with specific TF motifs (Figure 6F), we quantified insertions level specific expression levels using TElocal,[78] following the same mapping and normalization strategies as described above. ZKSCAN5, ZBTB26 and/or DUX motifs were found in the respective genome using HOMER (scanMotifGenomeWide.pl) and MLT1A0 insertions were categorized as having one or more TF motifs based on overlap with these found motifs.[82]

Phylogeny maximum likelihood analysis

We used the Dfam.h5 file curated by the Dfam consortium[34] and repeatMasker[100] to annotate the additional 7 mammalian genomes used in the phylogeny analysis. After selecting for insertions of appropriate insert length, we retrieved all insert sequences using bedtools getfasta.[90] These sequences were aligned using muscle,[57,58] followed by trimming by trimAl using settings -gt 0.01 -clustal[56] to compute the specific-specific consensus sequence. The species-specific consensus sequences were then in turn aligned using muscle with default parameters, as recommended for small data sets[57] and non-informative regions were trimmed from the consensus alignment using Trimal with a gap threshold of 0.01.[56] The resulting alignment was analysed with IQ-TREE (v2.2.5) using the '-m MFP' option for automatic model selection, a minimum split support value of 0.95 for internal node reconstruction, and a seed value of 42 for replicability.[81] The resulting phylogenetic trees were uploaded to iTOL (v6.9) and re-rooted to the Dfam consensus, which was used as a reference point.[101]

De novo motif search on TEs

Enriched motifs were identified at TE insertions and TE consensus sequences using HOMER scripts findmotifs.pl and findMotifsGenome.pl with parameters -mset vertebrates and -size given.[82] De novo motifs were matched with known motifs in the HOMER vertebrate database for motif searches at TE sequences. Known motifs were shown for enrichment at EGA genes together with their expression level at timing of genome activation when annotated in the listed gene annotation.

Chimeric TE-gene interaction analysis

Chimeric TE-gene interactions were identified using mode1 of the chimeraTE analysis tool.[60] We used only read 2 (R2) of our paired-end Smart-seq+5' data containing the 5' fragment barcode, as this contains the 5' fragment end of the transcripts. ChimeraTE was ran using –strand rf-stranded and –window 150000 as parameters, and using TE and gene annotations as listed in Table S6, except for the bovine chimeric analysis. Due to a lower quality in gene annotation of the bosTau5 genome assemble, we instead used the de novo repeatmasker annotation of latest genome assembly bosTau9/ARS-UCD1.2 using the Dfam v3.5 consensus sequences and the corresponding refSeq gene annotation. Only TE-initiated chimeric transcripts present in 2 or more replicates per stage were used in downstream analysis and visualization, with the exception of the rhesus data as the low sample count did not allow us to use this filter. De novo transcript isoforms were annotated using bambu[61] providing bam files containing only R2 originating from 5' fragment in our Smart-seq+5' data as input and genome information as listed in Table S6 (except for bovine, for which we used bosTau9 as mentioned above). Lastly, relative promoter usage of canonical or chimeric promoters were visualized using proActiv.[62]

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analysis was performed using R (version 4.2.2). Smoothing of expression profiles in Figures 3E–3I, 5E–5I, S1K, S1L, S2B, S2C, S2K, S3B, S3D–S3F, S4I–S4L, S5B, S6B, S6D, and S6E was done using the stat_smooth function (level=0.95) and indicating the standard error in grey.

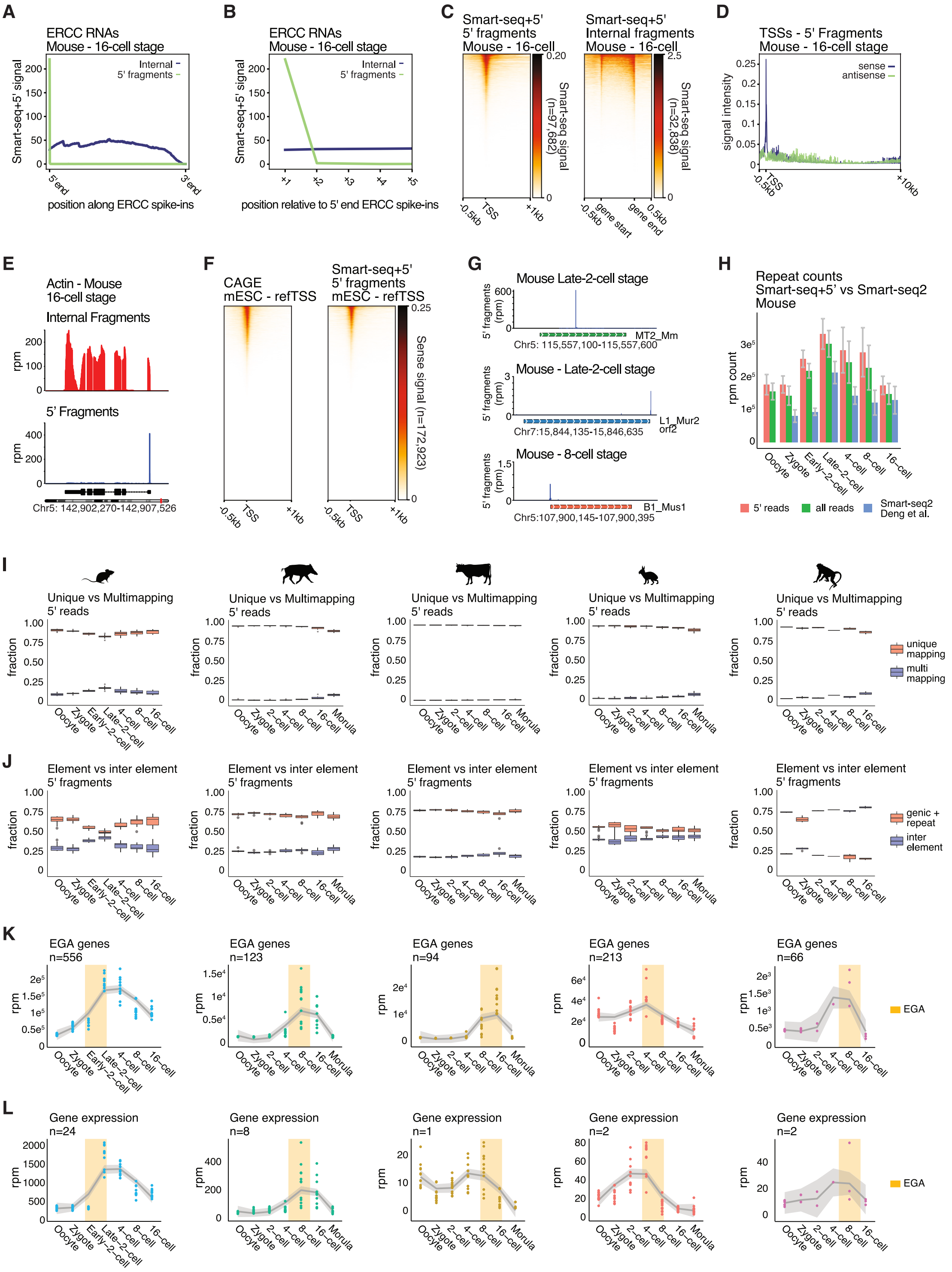
Barplots in Figures 6E and S6F indicate the median ± s.d. with individual biological replicates shown as dots. Boxplots in Figure 6F indicate the median with upper and lower quartiles as box limits and quartile range as whiskers. In Figures S1I and S1J, the box plots indicate mean and upper and lower quartiles for all single embryos analyzed per stage and for the species indicated. For statistical tests in Figures 6E and 6F, a Welsch t-test was used. In Figures 3J, 4B, 4E–4G, 5J, S1C, S1F, S1K, S1L, S5D, S6B, S7A, and S7B, n indicates number of genomic elements represented in the analysis. Number of embryos included in the Smart-seq+5′ dataset (and passed QC) is listed per stage and per species in Figure 1B. For mouse and bovine embryo experiments in Figure 7, N represents the experimental replicates performed on separate days. No specific method was used to determine statistical assumption. All statistical details for experiments can be found the corresponding figure legends.

ADDITIONAL RESOURCES

Detailed description, example scripts and software versions related to bioinformatic analyses used in this study can be found on github (https://github.com/meoomen/Smartseq5/). Data can be explored on embryo.helmholtz-munich.de.

CellPress
OPEN ACCESS

Supplemental figures



(legend on next page)

Figure S1. Smart-seq+5′ robustly detects 5′ transcript ends and improves quantification of TE expression and curation of EGA genes and their GO terms, related to Figure 1

(A and B) Smart-seq+5′ signal from all ERCC RNA spike-ins split by internal vs. 5′ fragments averaged (B) and over the first 5 base pairs (C) originated from single mouse 16-cell stage embryo libraries. Note that Smart-seq+5′ leads to a robust coverage of the full-length ERCC from the internal fragments as well as of the 5′ end of the transcript.

(C) Aggregate sense signal of 5′ fragments (left) on all annotated TSSs with signal and internal fragments (right) on all annotated genes scaled from gene start to gene end, using Smart-seq+5′ 16-cell mouse embryo data, represented as an heatmap. Rows are sorted by signal intensity, shown over a 1-kb window from the annotated TSS or scaled relative from gene start to gene end. The number of genomic elements is listed along the y axis.

(D) Aggregate signal of 5′ fragments of Smart-seq+5′ 16-cell mouse embryo data on all annotated TSS, separated by sense (blue) and antisense (green) signal and shown over a 10-kb window from the annotated TSS.

(E) Transcriptional profiling with Smart-seq+5′ captures internal fragments (top) and 5′ fragments (bottom) of transcripts exemplified by β-actin in 16-cell-stage mouse embryos.

(F) Aggregate sense signal of public CAGE data[25] and 5′ fragments from Smart-seq+5′ (right) in mESCs on all annotated TSSs from the refTSS database.[98] The number of annotated TSSs is indicated on the y axis.

(G) 5′ fragment signal Smart-seq+5′ of representative TE insertions of LTR MT2_Mm in the late 2-cell mouse embryo (G), L1_Mur2_orf2 in the late 2-cell mouse embryo (H), and B1_Mus1 in the eight-cell mouse embryo (I). Exact chromosome position of the TE insertion are listed along the x axis.

(H) Quantification (rpm) of reads mapped to repeat elements (TE and non-TE repeats) using only 5′ fragments or all reads from single-embryo Smart-seq+5′ libraries in the indicated mouse preimplantation stages, compared with rpm quantification of reads aligned to repeat elements from a published Smart-seq2 dataset of the same stages.[26]

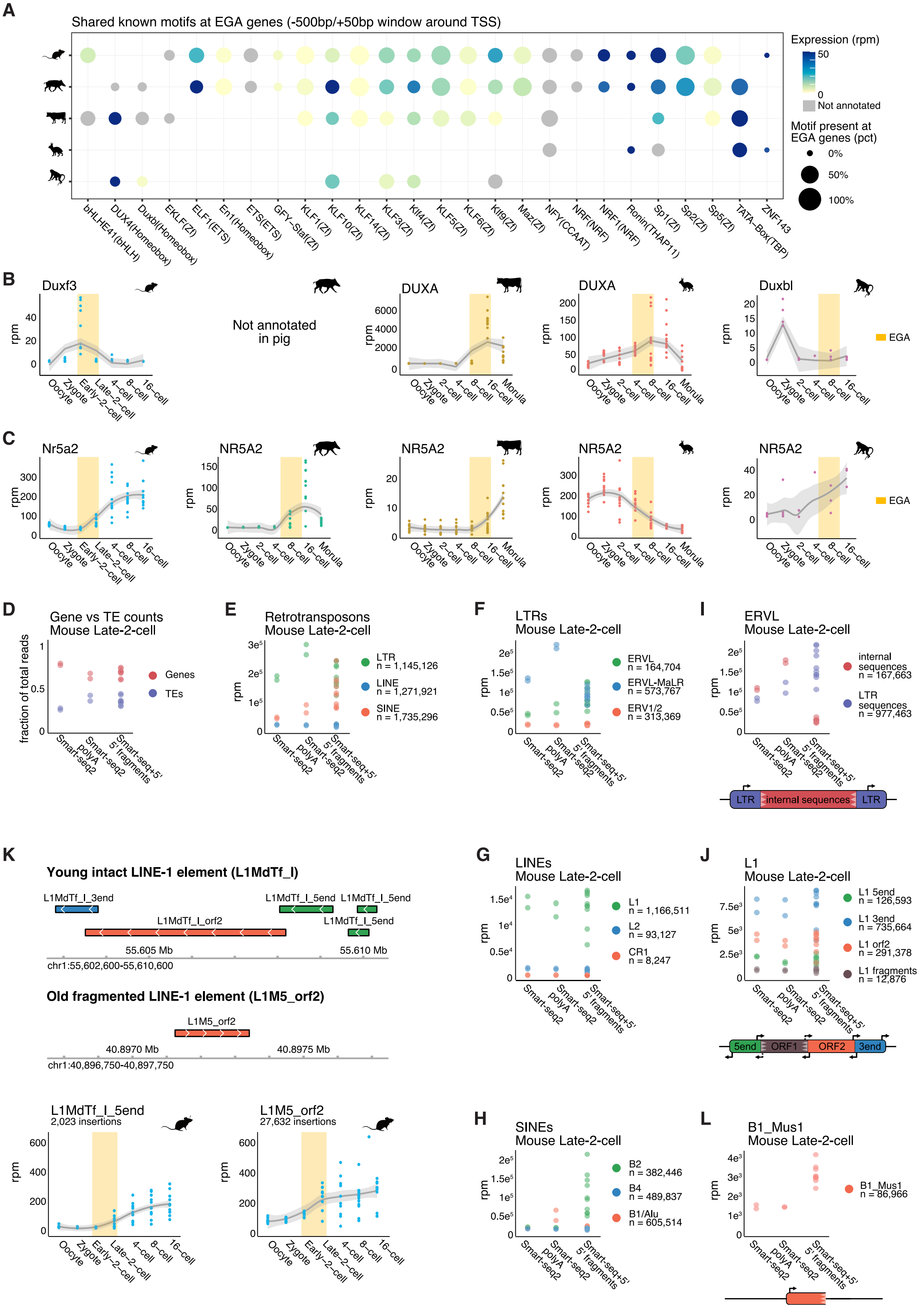
(I) Unique (red) or multimapping (blue) reads using the 5′ fragments of Smart-seq+5′. The boxplots indicate mean and upper and lower quartiles for all single embryos analyzed per stage and for the species indicated. The number of embryos per stage and per species is indicated in Figure 1C.

(J) Fraction of 5′ fragments mapped to a gene or repeat (all annotated TE and non-TE repeats), labeled "element" or not assigned to either, labeled "interelement." The boxplots indicate mean with limits at upper and lower quartile for all single embryos analyzed per stage and for the species indicated. The number of embryos per stage and per species is indicated in Figure 1C.

(K) EGA genes were selected based on differential increased expression from zygote to timing of EGA and differential decreased expression from timing of EGA to 16-cell/morula in all species. Summed expression levels as rpm values of all EGA genes are shown in mouse, pig, cow, rabbit, and rhesus.

(L) Summed expression values as rpm of EGA genes that have GO terms associated with gene expression (GO:0010468 and GO:0010628) in mouse , pig, cow, rabbit and rhesus.

In all panels, each dot represents the summed rpm value per embryo for each stage and species. The shaded line indicates SD, and the yellow rectangle depicts the time of EGA.



(legend on next page)

Figure S2. TF motif analysis of EGA genes and TF expression levels across species, capture of genes and TEs by Smart-seq+5′ vs. Smart-seq2 and polyadenylated Smart-seq2, related to Figure 1

(A) Known TF motifs found in a ±500-bp window around the TSS of all EGA genes in a given species. Size of the dots indicate the pct of EGA genes that were found to contain the TF, and color of the dot shows the expression level of this TF in a given species at timing of EGA. Only TF motifs that are shared across two or more species are shown. Note that DUX is known to bind neighboring MT2_Mm to EGA genes, but it is not enriched at EGA genes.

(B) Expression levels for Duxf3, DUXA, or Duxbl orthologs across early developmental stages in mouse, pig, cow, rabbit, and rhesus. Note that there is no annotated DUXA or Duxbl ortholog in pig and that Duxbl and DUXA share the same annotated binding motif. Expression levels are plotted as rpm values for the indicated genes in individual embryos, represented by dots. The shading indicates SD and the yellow rectangle the EGA time for the corresponding species.

(C) Expression levels as in (L) for Nr5A2 orthologs across early developmental stages in each species.

(D) Relative fraction of genic and TE counts in late 2-cell mouse embryos, detected by variations of Smart-seq. Dots represent values of individual embryos.

(E) rpm-normalized expression levels of retrotransposons in late 2-cell mouse embryos, detected by variations of Smart-seq as indicated. Dots represent values of individual embryos. The number of insertions (*n*) of the reflected TE subclass is listed in the color legend.

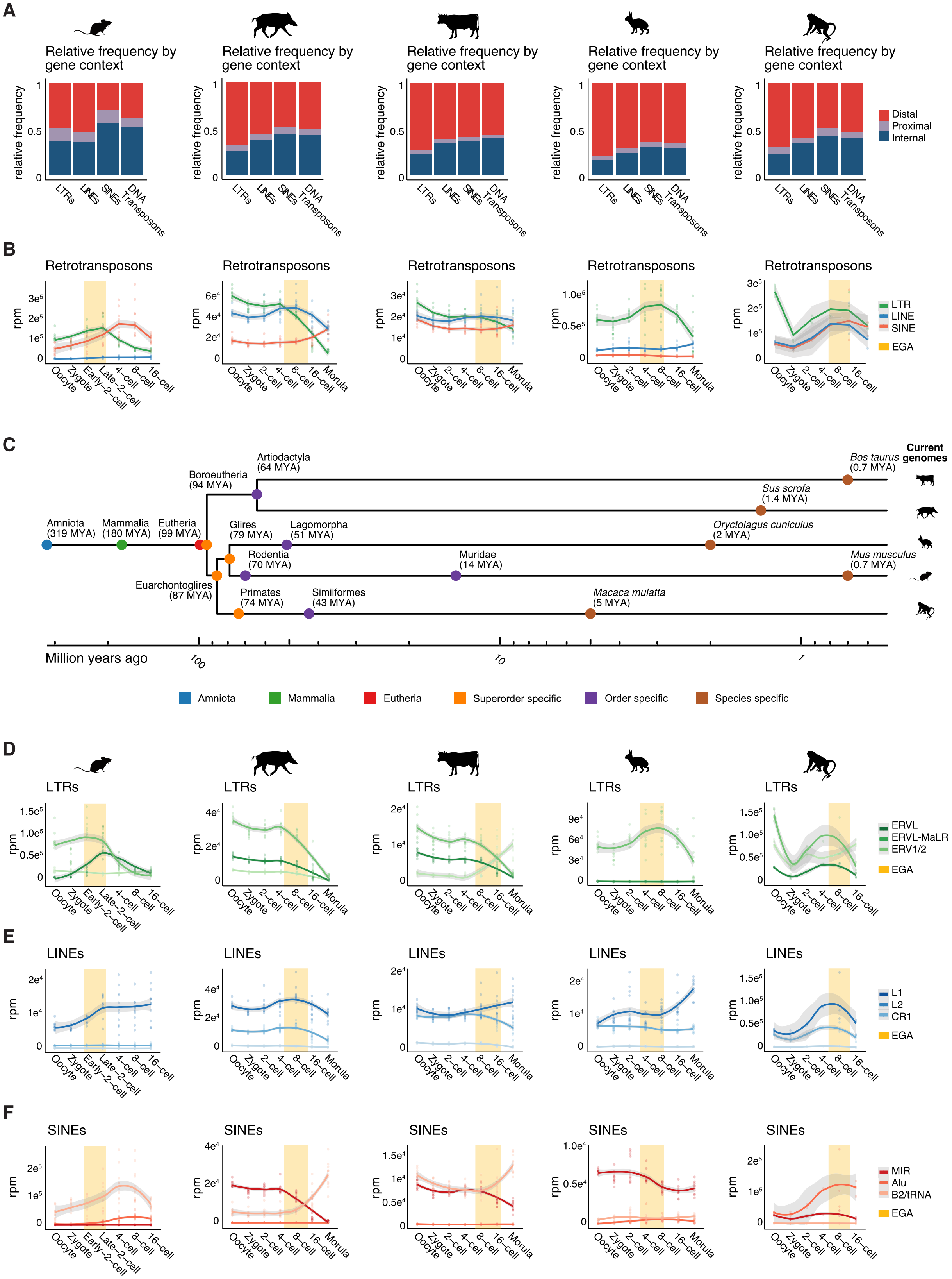
(F–H) rpm-normalized expression of LTR (F), LINEs (G), and SINEs (H) elements by superfamilies, detected by variations of Smart-seq as indicated. Dots represent values of individual embryos. The number of insertions (*n*) of the reflected TE super family is listed in the color legend.

(I) Expression levels of different sequences features from ERV/LTR TEs in the mouse late 2-cell embryo, as detected by variations of Smart-seq2 as indicated. Dots represent values of individual embryos. A schematic figure below represents the typical ERV/LTR element structure originally inserted at time of transposition. Number of insertions (*n*) of ERVL internal sequences and number of ERVL LTRs are listed in the color legend.

(J) Expression levels of different sequence features from L1 LINE elements in the mouse late 2-cell embryo, as detected by variations of Smart-seq2 as indicated. Dots represent values of individual embryos. A schematic figure below represents the typical L1 element structure originally inserted at time of transposition. The number of insertions (*n*) of L1 structural elements is listed in the color legend. The structural elements (L1_5end, L1_3end, L1 orf, and L1 fragments) are from Dfam annotation, but Repbase does not distinguish such structural elements.

(K) Representative view of a young intact LINE L1 element L1MdTf_I on chromosome 3 (top) and ancient fragmented LINE L1M5_orf2 on chromosome 1 (bottom) in the mouse genome. For visualization, we only show structural elements belonging to these LINE L1 elements at this genomic location. Left panels show the expression levels of all L1MdTf_I_5end (top, 2,023 insertions) and L1M5_orf2 (bottom, 27,632 insertions) across mouse developmental stages. In all panels, each dot represents the rpm value per embryo for each stage and species. The shaded line indicates SD, and the yellow rectangle depicts the time of EGA.

(L) Expression levels B1_Mus1 in the mouse late 2-cell embryo as detected by variations of Smart-seq2 as indicated. Dots represent values of individual embryos. A schematic figure below represents the typical SINE structure found in mammalian genomes. Number of B1_Mus1 insertions (*n*) is listed in the color legend.



(legend on next page)

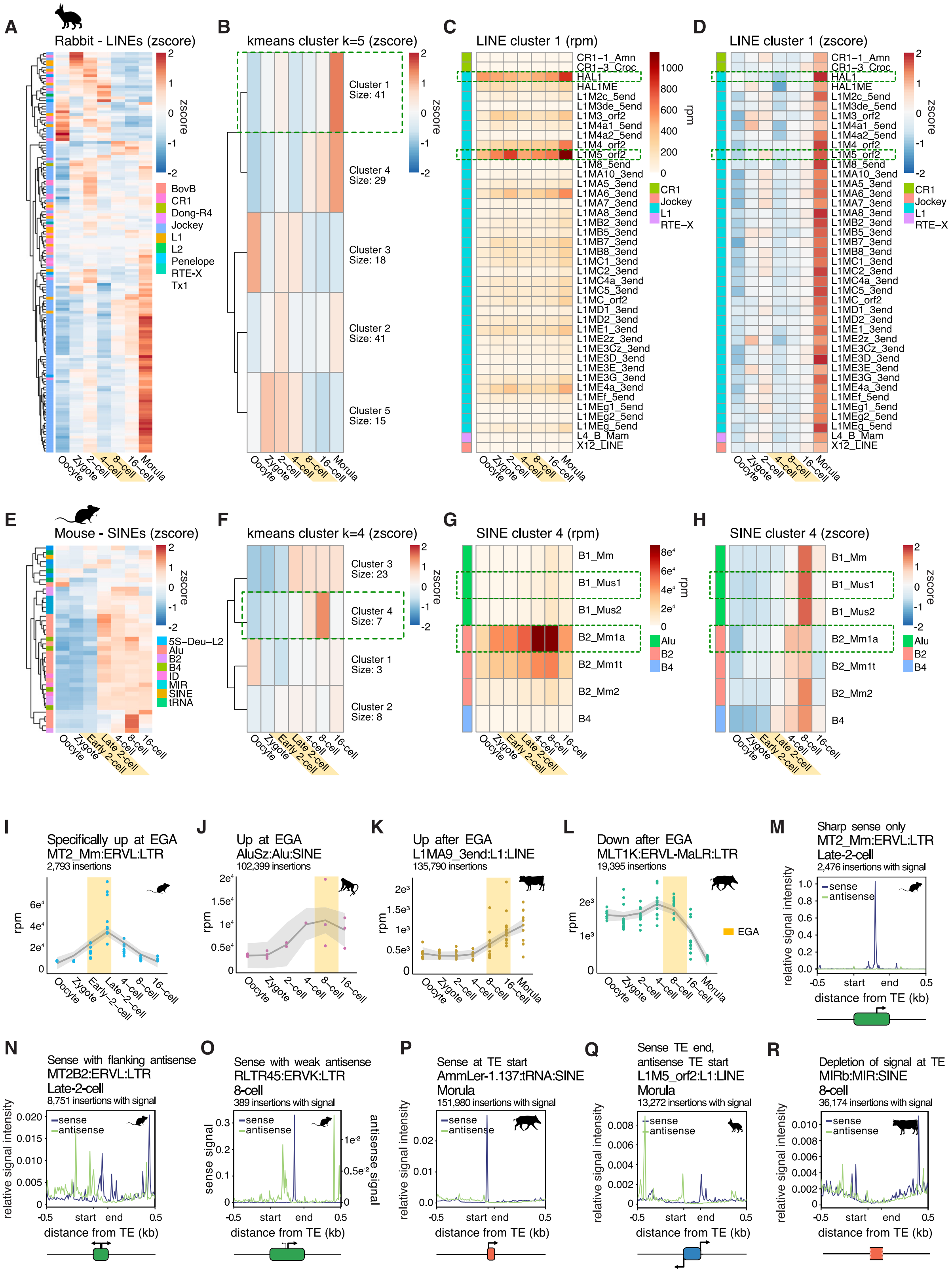
Figure S3. TE localization across genomes relative to genes and expression of the retrotransposon subclasses of LTRs, LINEs, and SINEs across stages and species, related to Figure 2

(A) Relative frequency of LTRs, LINEs, SINEs, and DNA transposons by genomic context (internal to genic sequence, proximal [<5 kb to gene], and distal [>5 kb to gene]) in mouse, pig, cow, rabbit, and rhesus.

(B) Expression of retrotransposon classes LTR, LINE, and SINE at the developmental stages indicated in the mouse, pig, cow, rabbit, and rhesus. Individual dots represent the summed rpm value per class per single embryo. The shade indicates SD.

(C) Evolutionary age throughout speciation. Approximation of age at each node and/or species emergence derives from the TimeTree database.[55] Dots are color coded to reflect evolutionary age of TE families shown in Figures 2B–2G.

(D–F) Expression profiles of LTR superfamilies ERV1 and ERV2 (ERVK) combined; ERVL and ERVL-MaLR (D); LINE superfamilies CR1, L1, and L2 (E); and SINE superfamlies Alu, MIR, and tRNA (F) in all species at the indicated stages, where each dot represents the summed rpm value of all elements per subclass, per single embryo. The shade indicates SD.



(legend on next page)

Figure S4. Analysis pipeline and examples of selected LINEs and SINEs with an EGA profile and examples of conserved expression patterns and TSS profiles for TEs with an EGA profile , related to Figures 3 and 4

(A) Hierarchical clustering based on expression of LINE families in rabbit embryos by *Z* score, whereby each row corresponds to an individual LINE from the different subclasses as indicated by the color code.

(B) k-means clustering (*k* = 5) based on the expression of LINE families by *Z* score from (A). Number of elements per cluster is indicated.

(C and D) Expression values as mean rpm (C) and *Z* score (D) of the individual LINEs from cluster 1 in (B), in all embryos from indicated stages. Corresponding subclasses are indicated by the color code.

(E) Hierarchical clustering based on expression of SINE families in mouse embryos by *Z* score. Each row corresponds to an individual SINE from the different subclasses as indicated by the color code.

(F) k-means (k = 4) clustering based on the expression of SINE families by *Z* score from (E), with number of elements per cluster as indicated.

(G and H) Expression values as mean rpm (G) and *Z* score (H) of the individual SINEs from cluster 4 in (B), in all embryos from indicated stages. Corresponding subclasses are indicated by the color code.

(I) Mouse LTR MT2_Mm increases expression specifically at the time of EGA.

(J) Rhesus SINE AluSz becomes expressed at the time of EGA in eight-cell-stage rhesus embryos, and its expression persists after EGA.

(K) Bovine LINE L1MA9 continues to increase expression after EGA.

(L) Pig LTR MLT1K is an example of a TE that decreases in expression after EGA.

From (I) to (L), individual dots represent single embryos and indicate the rpm value of the sum of the 5′ reads mapping to all insertions for each indicated TE. The shaded line indicates SD, and the yellow rectangle indicates the timing of EGA for each species.

(M) Mouse LTR MT2_Mm contains a sharply positioned TSS, transcribed in the sense direction only in late 2-cell-stage mouse embryos.

(N) Mouse LTR MT2B2 shows transcription in both sense and flanking antisense direction at its TSS in the late 2-cell-stage embryo.

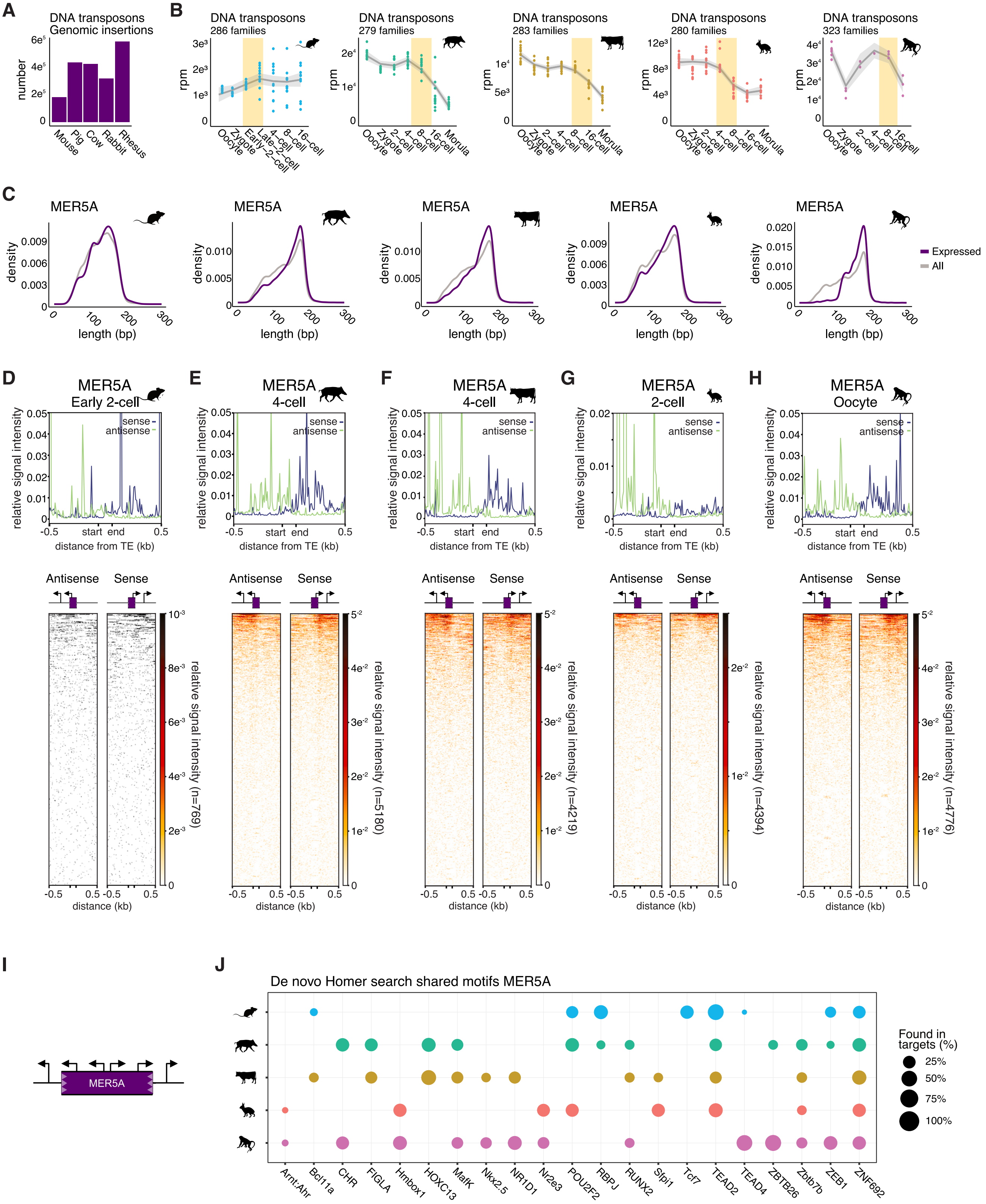
(O) Mouse LTR RLTR45 is transcribed in eight-cell embryos from its TSS in sense direction with weak flanking antisense transcription (note the adjusted y axis scale for antisense signal).

(P) Pig SINE AmmLer-1.137 contains a sharply positioned TSS at the start of the TE sequence.

(Q) Rabbit LINE L1M5_orf2 contains two TSSs, initiating transcription in sense direction at the end of the TE sequence and in antisense direction at the start of the TE sequence.

(R) Cow SINE MIRb is an example of a TE that shows a depletion of signal around the TE sequence.

From (M) to (R), shown are aggregate signals of all genomic insertions with unique mapping signal. Reads in the sense direction are indicated in blue and in the antisense direction in cyan. Number of insertions with captured unique mapping signal is indicated.



(legend on next page)

Figure S5. MER5A shows similar TSS profiles across species and shared *de novo* TF motifs, related to Figure 5

(A) Number of total DNA transposon genomic insertions per species.

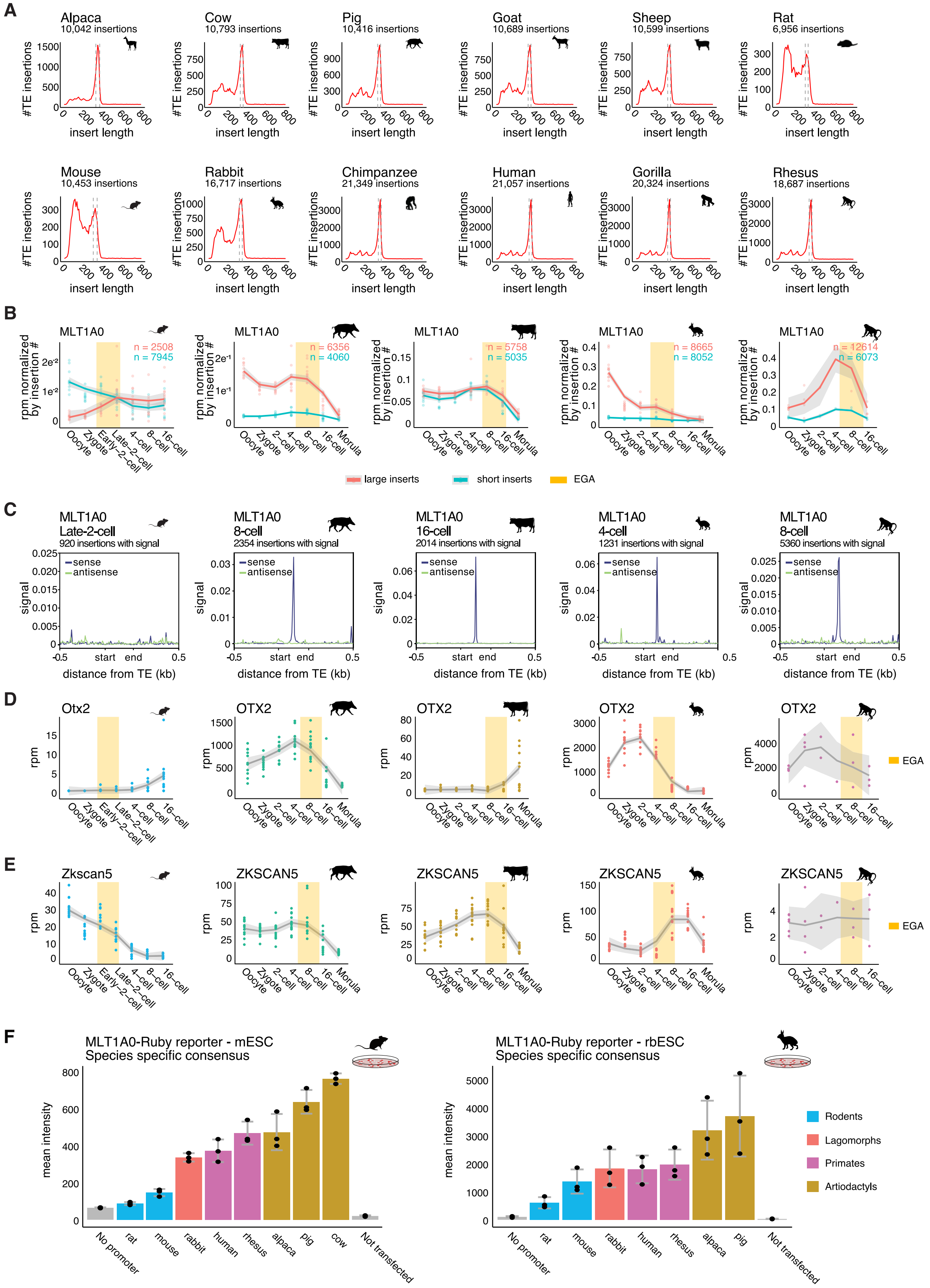
(B) Expression levels of all DNA transposons calculated as rpm of the sum of 5′ fragments aligned to DNA transposons in mouse, pig, cow, rabbit and rhesus. Each dot represents the rpm value per embryo for each stage and species. The dashed line depicts SD. The number of embryos per stage and per species is indicated in Figure 1B. The number of DNA transposon families represented in expression profiles is listed.

(C) Density plots of MER5A insert size distrutions for all insertions (gray) and all transcriptionally active insertions (purple).

(D–H) Aggregate signal analysis of MER5A in mouse early 2-cell (D), pig 4-cell (E), cow 4-cell (F), and rabbit 2-cell (G) embryos and in rhesus oocyte (H). Pileup of sense (blue) and antisense (green) signal on MER5A insertions at a representative development stage (top). Heatmap of the 5′ fragment sense and antisense signal illustrating the TSS across individual MER5A insertions in embryonic stage as listed. Number (*n*) of insertions shown in the heatmap is indicated along the y axis. Insertions with no Smart-seq+5′ signal were excluded from this analysis.

(I) Schematic drawing of MER5A element features.

(J) *De novo* TF motif enrichment across all MER5A insertions in the mouse, pig, cow, rabbit, and rhesus genomes. Bubble size represents the percentage of TE insertions in each genome, which were enriched for a given TF motif over the background genomic sequence. Only TFs that were shared across at least two species are shown. Motifs that belong to the same TF were pooled.



(legend on next page)

Figure S6. Analysis of MTL1A0 LTR sequence conservation, related to Figure 6

(A) Histograms of MLT1A0 LTR insert length (in bp) distribution of all the corresponding insertions in 12 different mammalian genomes. Dashed lines depict the size cutoffs used for phylogeny analysis.

(B) Expression of MLT1A0 LTR insertions split by the size of the LTR in either larger than 300 bp (red) or shorter than 300 bp (blue). Number of insertions (*n*) in each category is listed. Each dot represents the rpm value for all insertions in the corresponding size class per embryo, normalized by the number of insertions of that class across the genome of each species. The shaded line indicates SD, and the yellow rectangle depicts the time of EGA.

(C) TSS profile of size-selected MLT1A0 LTR sequences (as shown in Figures S7A–S7L) in mouse, pig, cow, rabbit, and rhesus embryos at the indicated developmental stages. The plots show the aggregate of 5′ transcript signal over the MLT1A0 LTR insertions, relative to its "start" and "end," in the sense (blue) and the antisense (cyan) direction and thus an approximation of the position of the TSS relative to the start and end of MTL1A0. Number of insertions reflected in this analysis is listed.

(D) Otx2 or Otx2 ortholog expression across early developmental stages in mouse, pig, cow, rabbit, and rhesus. Expression levels are plotted as rpm values for the indicated genes in individual embryos, represented by dots. The shading indicates SD and the yellow rectangle the EGA time for the corresponding species.

(E) Expression levels as in (D), for Zkscan5 or Zkscan5 orthologs across early developmental stages in each species as indicated.

(F) MLT1A0-Ruby reporter assay as measured by flow cytometry in mouse ESCs (left) and rabbit ESCs (right), following experimental setup as shown in Figure 6D. Species-specific consensus sequences were used as shown in Figure 6B.

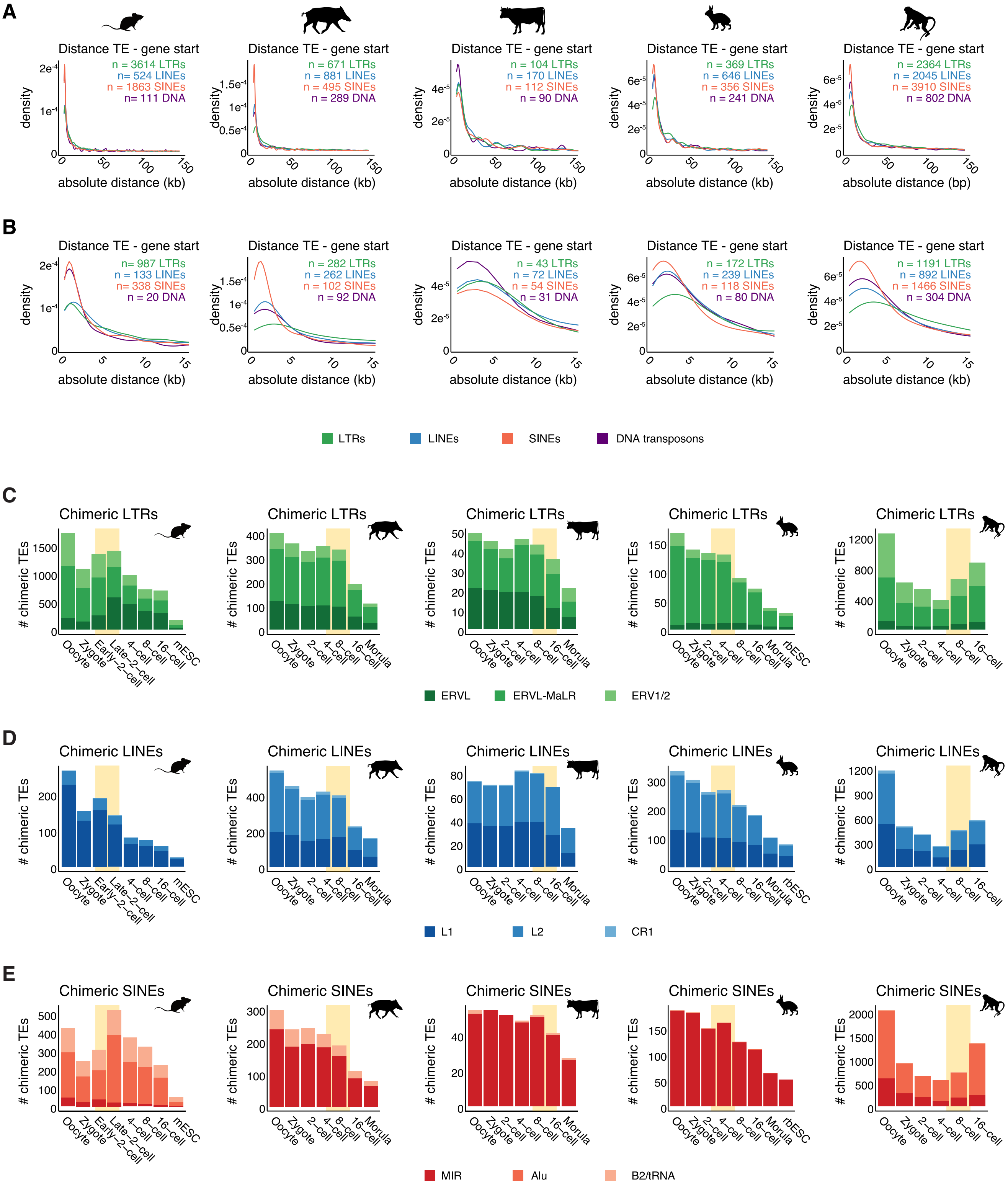


Figure S7. Genomic distance between TE and gene start across all chimeric TE-gene 5′ fragments from Figure 7

(A and B) Density plot of absolute distance between gene start and TE found in chimeric 5′ fragments separated by subclass in mouse, pig, cow, rabbit, and rhesus on a unrestricted x axis (top) and zoomed in to 0–15 kb distance (bottom). Number of chimericTEs reflected in the density plots (*n*) is listed, following the color coding of the plots.

(C–E) Number of chimericTEs indicated by superfamily in LTRs (C), LINEs (D), and SINEs (E) across species and across developmental stages.