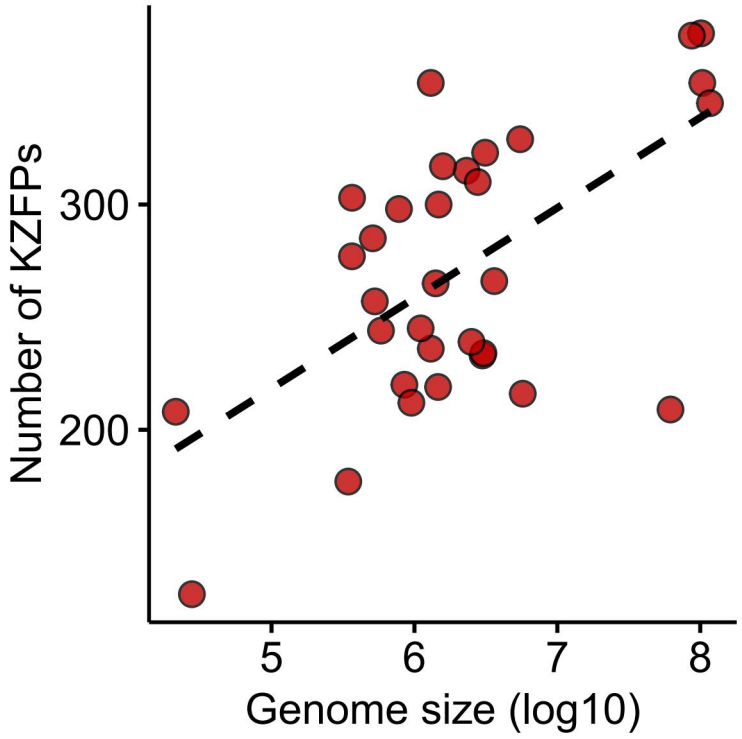
Supplemental Material for "*k*-mer cross-species profiling reveals taxon-specific TE expansions accompanied by KZFP co-option and functional impacts in ruminants"

Pengju Zhao[1,2], Jiayi He[1,2], Chen Peng[1,2], Yuelang Zhang[1,2], Chong Wang[1,2], Dongyou Yu[1,2], Lingzhao Fang[3], Zhengguang Wang[1,2]

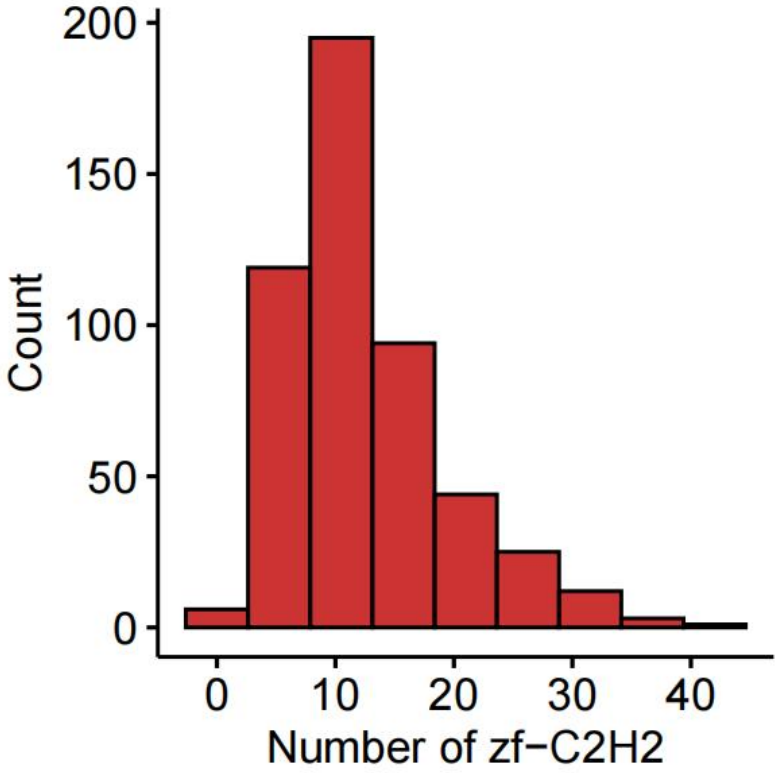
[1] Hainan Institute, Zhejiang University, Yazhou Bay Sci-Tech City, Sanya, Hainan 572000, China.
[2]College of Animal Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, China.
[3]Center for Quantitative Genetics and Genomics, Aarhus University, Aarhus, 8000, Denmark.

Contents

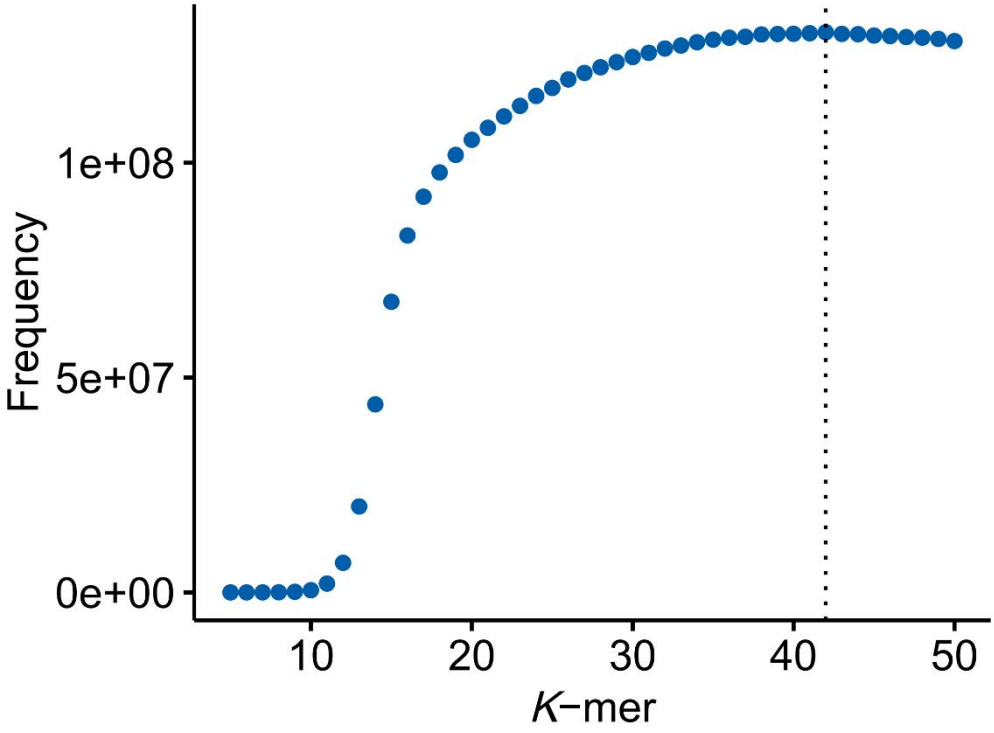
A1 Supplemental Figures



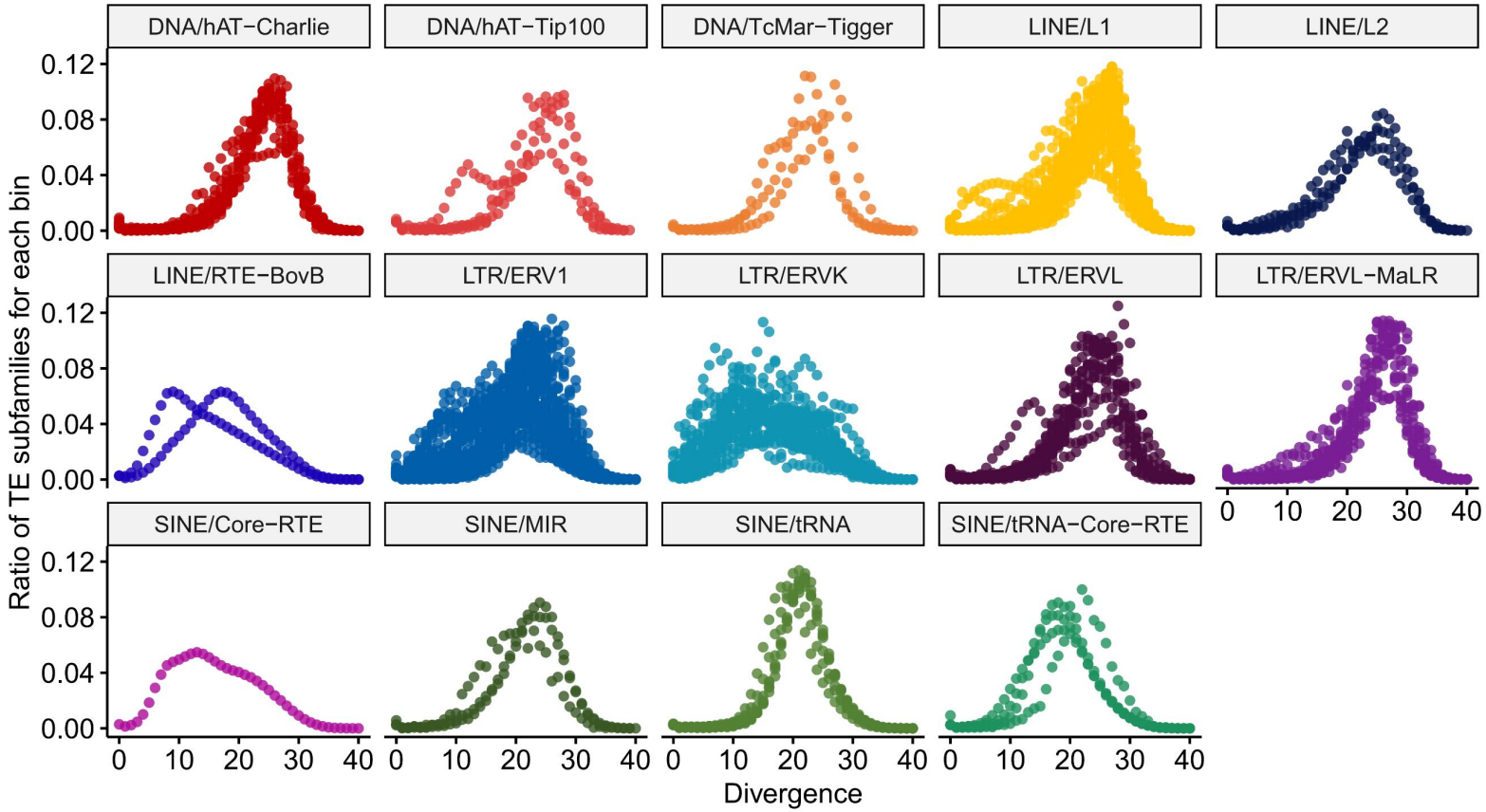
Supplemental Fig S1 | Correlation between the number of KZFPs and genome assembly quality as measured by N50 values. The x-axis represents the genome size, while the y-axis represents the total count of KZFPs identified in each genome.



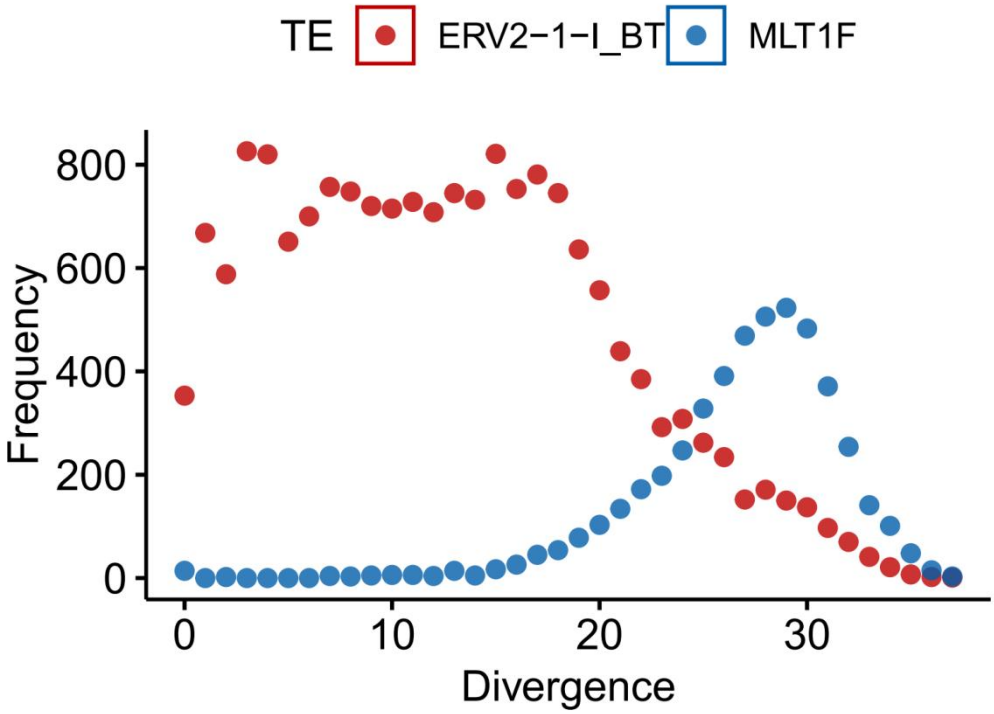
Supplemental Fig S2 | Distribution of zf-C2H2 Domains in Ruminant KZFPs. Bar plot illustrates the distribution of KZFPs in ruminants based on the number of zf-C2H2 domains per protein. The x-axis represents the count of zf-C2H2 domains per KZFP, while the y-axis shows the frequency of KZFPs.



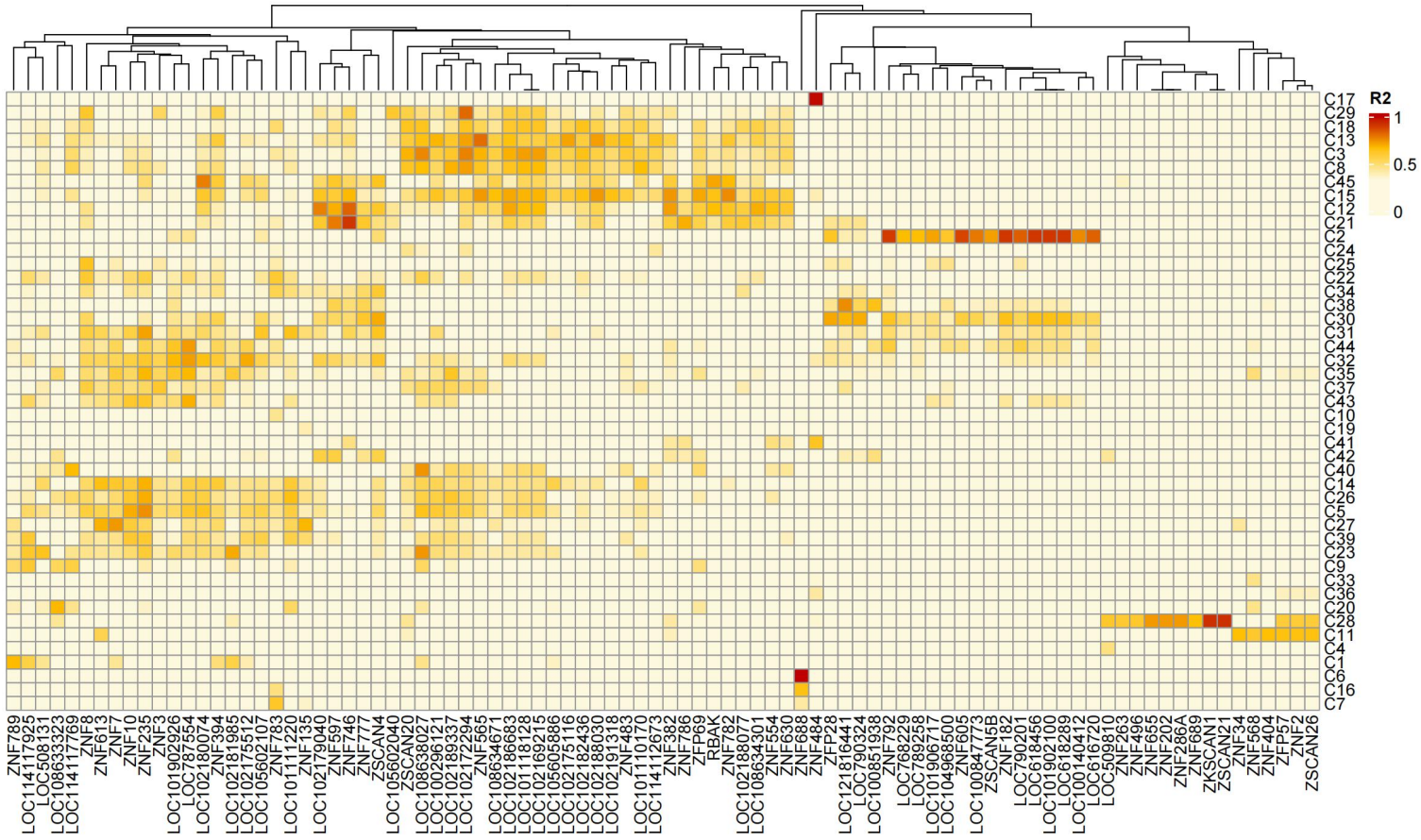
Supplemental Fig S3 | Frequency distribution of *k*-mers within 218 megabases of ruminant TE consensus sequences. The x-axis represents the *k*-mer sequences, while the y-axis represents their corresponding frequency of occurrence in the TE consensus library



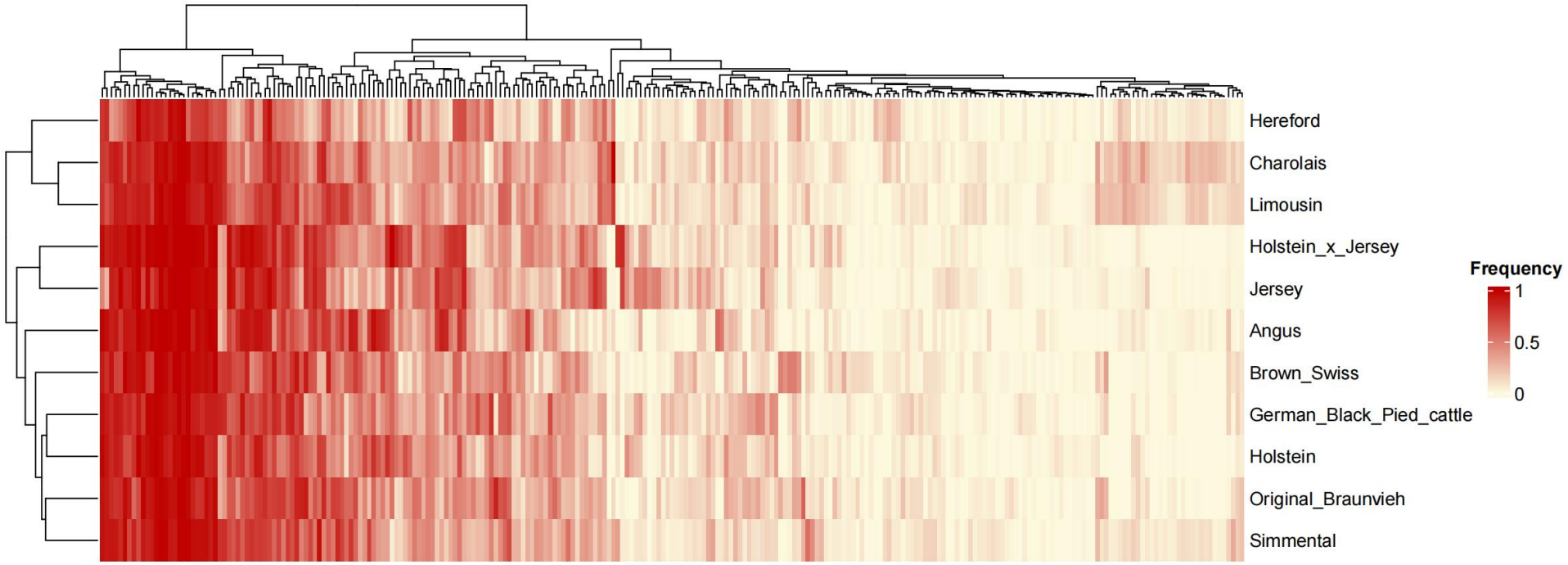
Supplemental Fig S4 | Detailed relationship between TE family expansion and evolutionary age, analyzed by individual families. Line plots depicting the proportional abundance (y-axis) of major TE subfamilies over their evolutionary age (x-axis). Each panel represents a distinct TE family, illustrating how their respective activity bursts are distributed across evolutionary time. This breakdown validates and supplements the integrated pattern shown in Fig. 2C by revealing subfamily-specific temporal dynamics of transpositional activity.



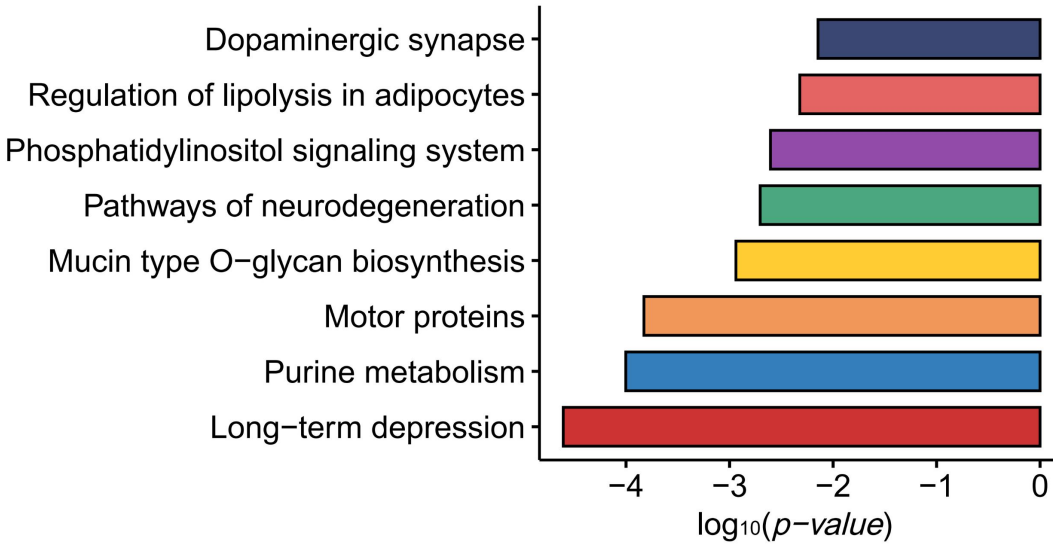
Supplemental Fig S5 | Activity patterns of ERV2-1-I_BT and MLT1F TEs. ERV2-1-I_BT (red) and MLT1F (blue) display distinct patterns across evolutionary divergence levels. The x-axis shows sequence divergence from the consensus, while the y-axis indicates relative activity. ERV2-1-I_BT exhibits higher recent activity (low divergence), whereas MLT1F shows sustained activity across intermediate divergence levels, suggesting distinct expansion timelines in ruminant genomes.



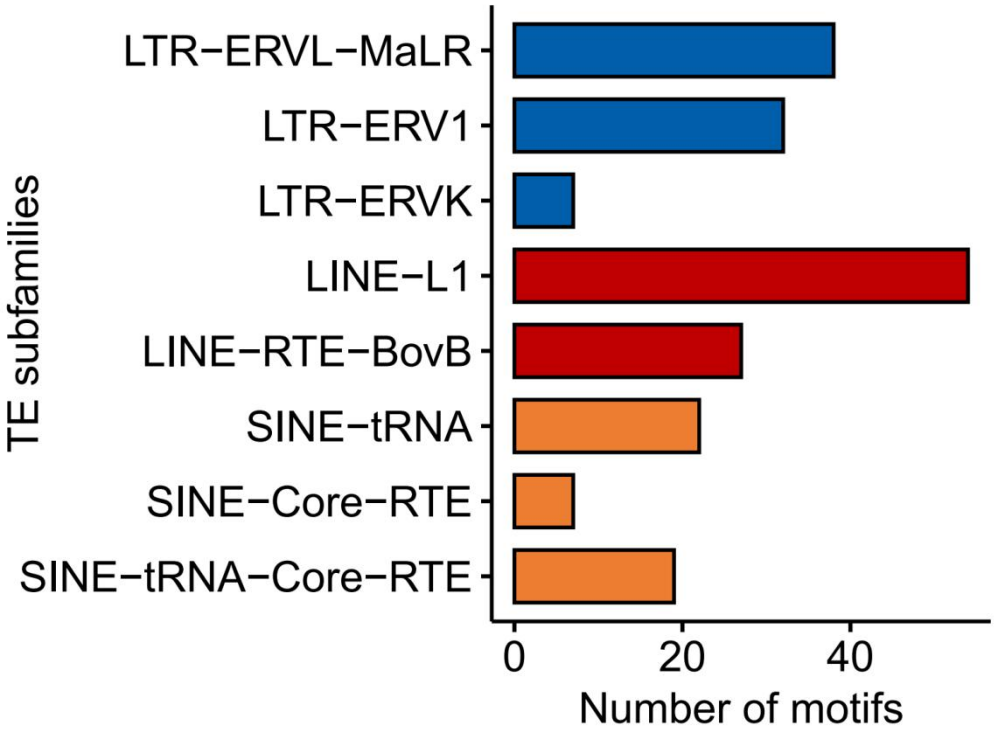
Supplemental Fig S6 | Heatmap for KZFP-TE *k*-mer cluster correlations. Heatmap of significant correlations ($R^2 \geqslant 0.6$) between KZFPs (columns) and TE-derived *k*-mer clusters (rows). Color intensity reflects the strength of positive (red) correlations. Clusters represent TE sequences bound by specific KZFP subsets, indicating co-evolutionary relationships where KZFPs target TEs.



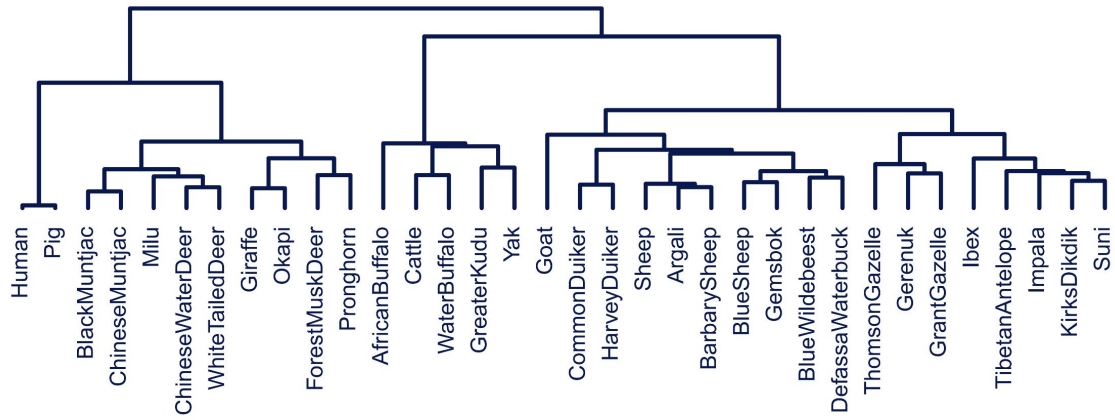
Supplemental Fig S7 | Frequency spectrum of TE-driven structural variants. The population frequency distribution of TE-induced SVs in cattle is shown. Columns represent individual SVs, rows denote cattle populations, and color intensity indicates abundance.



Supplemental Fig S8 | KEGG pathways of genes impacted by TE insertions. The figure shows enriched KEGG pathways for 204 cattle genes containing TE insertions in gene bodies (exons/introns). The x-axis displays the pathways, while the y-axis represents enrichment significance ($\log_{10}$(*p-value*)).



Supplemental Fig S9 | Number of predicted regulatory motifs per TE subfamily. Number of predicted transcription factor binding motifs within sequences of major TE subfamilies.



Supplemental Fig S10 | Hierarchical clustering of ruminant species based on TE-derived *k*-mer profiles. Unrooted phylogenetic tree illustrating relationships inferred from genome-wide TE *k*-mer similarity.

A2 Supplemental Tables

Supplemental Table S1 | Genome assembly and sample information for the 32 ruminant species. Shown are the family classification, subfamily name, species name, and genome N50 for each ruminant species.

Supplemental Table S2 | Catalog and domain structure of all KZFPs. Shown are the gene name, representative isoform identifier, and counts of key protein domains (KRAB, C2H2-type zinc finger, and SCAN) for each KZFP identified from the 32 ruminant genomes.

Supplemental Table S3 | Presence and absence of KZFPs across 32 ruminant species. Shown are the KZFP name, a representative isoform identifier, and a matrix indicating the presence ("1") or absence ("0") of an KZFP isoforms for each predicted KZFP gene across all 32 ruminant genomes analyzed.

Supplemental Table S4 | Reference library of TE *k*-mers across 14 TE families in ruminant genomes. Shown are the TE family classification, specific TE subfamily designation, and the exact nucleotide sequence for each of the 67,293 unique *k*-mers retained as markers for profiling TE sequence abundance across 32 ruminant genomes.

Supplemental Table S5 | Summary of TE *k*-mer clusters and their taxonomic distribution patterns. Shown are the cluster identifier, the number of TE *k*-mer sequences assigned to each cluster, and the evolutionary distribution classification based on their presence.

Supplemental Table S6 | Association details between TE *k*-mer clusters and predicted KZFP binding motifs. Shown are the data source, KZFP name, associated TE classification, correlation strength (R^2), predicted binding motif sequence, and statistical significance (p-value) for each of the 68 significant TE *k*-mer cluster-KZFP pairs ($R^2 > 0.6$, p-value < 0.001).

Supplemental Table S7 | Experimentally validated KZFP-TE interactions conserved between humans and ruminants. Shown are the data source, KZFP name, targeted TE family class, the predicted binding motif sequence, and the key experimental study providing supporting evidence for each of the 12 conserved interactions identified.

Supplemental Table S8 | Genomic details of 14 genes exhibiting taxon-specific TE exonization events. Shown are the genomic coordinates, identities, and orientation of

both the integrated TE sequences and the host genes, including specific exon and transcript identifiers, for each of the 14 confirmed TE exonization events.

Supplemental Table S9 | Genomic details of 951 TE-associated structural variants identified across 11 cattle breeds. Shown are the genomic coordinates, reference and alternative sequences, genomic context, and gene annotations for each of the 951 structural variants attributable to Bovinae-specific TE subfamilies.

Supplemental Table S10 | Genomic annotation and tissue specificity of TE-derived cis-regulatory elements. Shown are the genomic coordinates, tissue source, regulatory state, functional annotation, TE origin, genomic context, and nearest gene for each TE-derived promoter (TssA) and enhancer (EnhA) region identified in ruminant genomes.

Supplemental Table S11 | Annotation results of human KZFP genes validated by the pipeline. Shown are gene names, transcript IDs, and counts of key domains (KRAB, C2H2 zinc fingers, SCAN, and DUF3669) for each isoform.

Supplemental Table S12 | Profile information for 25 experimentally unvalidated KZFP TFBS retrieved from the JASPAR 2024 database. Shown are the matrix identifier, transcription factor name, taxonomic group, source species, experimental data type, and source publication ID for each TFBS motif profile used as reference for downstream analyses.