Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press



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

Pengju Zhao, Jiayi He, Chen Peng, et al.

Genome Res. published online March 23, 2026
Access the most recent version at doi:10.1101/gr.281289.125

P<P Published online March 23, 2026 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Creative Commons License This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see https://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**



To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

Published by Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

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.

Email addresses:
Pengju Zhao: zhaopengju2014@gmail.com; Jiayi He: 22017074@zju.edu.cn; Chen Peng: 22217072@zju.edu.cn; Yuelang Zhang: zhangyuelang@zju.edu.cn; Chong Wang: 12317021@zju.edu.cn; Dongyou Yu: dyyu@zju.edu.cn; Lingzhao Fang: Lingzhao.fang@qgg.au.dk; Zhengguang Wang: wzhguang68@zju.edu.cn

*Corresponding author:

Pengju Zhao, Ph.D.
Zhejiang University
Hainan Institute
Sanya, Hainan 572000, P. R. China
E-mail: zhaopengju2014@gmail.com

Zhengguang Wang, Ph.D.
Zhejiang University
College of Animal Sciences
Hangzhou, Zhejiang 310058, P. R. China
E-mail: wzhguang68@zju.edu.cn

Lingzhao Fang, Ph.D.
Aarhus University
Center for Quantitative Genetics and Genomics
Aarhus, 8000, Denmark
E-mail: lingzhao.fang@qgg.au.dk

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

Abstract

Transposable elements (TEs), which constitute approximately one-third of ruminant genomes, significantly influence genome evolution. However, the substantial sequence diversity across ruminant species complicates cross-species TE comparison and analysis of their interactions with Krüppel-associated box zinc finger proteins (KZFPs). Through a comprehensive analysis of 32 ruminant genomes, we identify taxon-specific expansions of KZFPs (approximately 271 per genome), which reflect established phylogenies. Among TEs, LINE/RTE-BovB dominates genomic content (55.7%), while LTR/ERVK subfamilies show significant taxon-specific variation. Our *k*-mer-based clustering approach effectively reconstructs ruminant phylogeny and reveals temporal patterns of TE expansion. Importantly, we uncover robust coevolutionary signatures between 90 KZFPs and specific TE *k*-mer clusters, notably involving 32 KZFPs that target Bovidae-specific LTR/ERVK clusters. Furthermore, Bovinae-specific TEs impact gene structure by exonizing 14 genes and generating tissue-enriched cis-regulatory elements (CREs), including 1,645 promoter- and enhancer-associated elements. This study establishes a robust computational framework for cross-species TE analysis and provides key insights into the coevolution of KZFPs and TEs, advancing our understanding of TE genomics and function in ruminants.

Keywords: Transposable elements; KZFPs; *k*-mer; Ruminants; Cross-species

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

Introduction

Transposable elements (TEs) are dynamic genomic sequences that jump within genomes, comprising 7-40% of genetic material in domestic animals (Pengju et al. 2023b). While the majority of TE insertions are either neutral or potentially harmful to host organisms, they have unpublished nonetheless been instrumental in shaping host genome evolution through their distinctive movement patterns (Irina 2018; Roy et al. 2018; Raquel et al. 2022). These TEs demonstrate remarkable versatility in their ability to create new transcripts or genes through exonization, while also modulating gene expression through strategic placement of enhancers or promoters near genes (Guillaume et al. 2018; Sophie and Gaël 2020). Current studies reveals important functional aspects of active TEs, as evidenced by *Alu* elements that insert themselves into introns of the human *TBXT* gene and encode in opposing genomic orientations to nearby ancestral *Alu* elements, enabling ape-specific alternative splicing that influences tail development (Bo et al. 2024). Studies indicate that the insertion of Endogenous retroviruses (ERVs) in the *APOB* gene causes cholesterol deficiency in Holstein-Friesian cattle (Lanhua et al. 2024). Furthermore, SINEs represent an important component of presence-absence variation in the pig genome, providing insights into population dynamics (Pengju et al. 2023a). Notably, during speciation and subspecies formation, TEs continuously evolve, leading to distinct functional characteristics among TE subfamilies through acquiring or losing specific transcription factor binding sites (Anna and Todd 2021). Consequently, understanding TE patterns is essential for comprehending the mechanisms of host genome evolution.

Krüppel-associated box zinc finger proteins (KZFPs, KRAB-ZFPs) function as essential transcriptional repressors that regulate TE transcriptional activity at the epigenetic level (Michaël et al. 2017). The human genome encompasses approximately 400 genes that encode KZFPs. From a structural perspective, KZFPs are composed of a KRAB domain and multiple ZNF zinc finger domains (Jonas de et al. 2023). Previous research has demonstrated that KZFPs has a remarkable ability to bind to TEs, exhibiting a continuous coevolutionary pattern (Gabriela et al. 2016). Through their intricate interactions, KZFPs and TEs establish

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

sophisticated epigenetic regulatory networks in humans (Michaël et al. 2017). Additionally, zinc finger proteins enhance cellular defense mechanisms against external retroviruses (including HIV-1) that contain PBS-Lys elements by effectively inhibiting their transcription and integration (Bo et al. 2022). The expression levels of cancer-associated KZFPs demonstrate significant correlations with patient survival outcomes, tumor histological characteristics, and molecular classifications (Marta et al. 2019b). Furthermore, studies have revealed that in mouse models, deletion of KZFPs gene clusters results in modest activation of ERV transposition activity (Gabriela et al. 2016). This evidence substantiates the essential regulatory function of KZFPs in TE regulation (Pierre-Yves et al. 2019). These discoveries emphasize the critical nature of understanding KZFP-TE interactions across species.

Ruminants are widely distributed herbivorous mammals with unique morphological and ecological characteristics (particularly head organs and multi-chambered stomachs). They also include several economically important livestock species, such as cattle, buffalo, sheep, and goats (Lei et al. 2019). Coupled with the genome evolution of ruminants, TEs formed distinct TE subfamilies, making ruminants ideal models for studying the dynamic evolutionary relationship between TEs and the KZFP family (Marta et al. 2019a). Furthermore, studying dynamic evolution of ruminant TE can help us reveal unique phenotypic traits in ruminant species (Siqian et al. 2022). Yet, the characterization of KZFPs and TEs as well as their coevolution patterns is limited in ruminants, because the conventional approachs based on conserved TE sequences, such as Repbase (Weidong et al. 2015) and Dfam (Jessica et al. 2021), failed in handling TEs of such high sequence diversity (Lei et al. 2019). Notably, *k*-mer-based methods offer a powerful, alignment-free alternative to reference-dependent techniques for TE phylogenetics. For instance, Hill *et al.* classified "Starship" elements using *k*-mer-based clustering, which grouped TEs by host taxonomy without requiring conserved sequence databases (R et al. 2025). This highlights the potential of such approaches for revealing TE expansions in understudied lineages like ruminants.

In this study, we aimed to comprehensively characterize the coevolutionary dynamics of KZFPs and TEs across 32 ruminant genomes. We sought to use a *k*-mer-based approach to

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

overcome limitations of conserved sequence methods and identify taxon-specific expansions of both elements. Our goals were to elucidate their coevolutionary signatures and assess the impact of TEs on gene regulation and structural variation, thereby advancing our understanding of genomic evolution in ruminants (**Fig. 1**).

Results

The KZFP catalog and TE composition in ruminants

From the Ruminant Genome Database (RGD) (Fu et al. 2022), we obtained 32 ruminant species with high-quality genomes (Lei et al. 2019). These comprised 1 species from Antilocapridae, 2 from Giraffidae, 5 from Cervidae, 1 from Moschidae, and 23 from Bovidae (Supplemental Table S1). The evolutionary relationships among these species had been thoroughly established in previous research, accurately showing the evolutionary radiation of each ruminant species (Lei et al. 2019) (**Fig. 2A**). We developed an atlas of KZFPs by searching these 32 ruminant genomes for sequences encoding KRAB domains adjacent to zinc-finger (ZNF) arrays (Supplemental Tables S2-S3). More specifically, we identified an average of 271 KZFPs (ranging from 127 to 376) through *in silico* scanning of the KZFP family in ruminant genomes using C2H2 zinc-finger, KRAB, SCAN, and DUF3669 domains (**Fig. 2B**). The number of KZFPs was similar across the ruminant family, but it was influenced by genome quality (R^2 to N50 value of 0.51) (Supplemental Fig S1). Our analysis also revealed that ruminant KZFPs were multi-exonic and typically contained a KRAB domain, with an average of 12.52 zf-C2H2 domains (Supplemental Fig S2). Notably, 101 of these proteins have developed SCAN accessory domains, while none possessed a DUF3669 domain (**Fig. 2C**). We constructed hierarchical clustering by leveraging the shared pattern of KZFPs among ruminants. This clustering showed a similar pattern to the established ruminant phylogeny (**Fig. 2D**), suggesting taxon-specific expansion of some KZFPs during ruminant evolution with differential KZFP ratios between different taxa (**Fig. 2E**). For example, eight KZFPs were found exclusively in Bovinae genomes but not in other ruminant genomes (**Fig. 2F**).

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

We conducted a *de novo* TE annotation on ruminant genomes using LongRepMarker (Liao et al. 2021) and RepeatMasker (Tarailo-Graovac and Chen 2009). LongRepMarker produced an average of 40 Mb consensus sequences per genome, with approximately 68.39% matching known TE families. Using these consensus sequences as a library in RepeatMasker, we found that an average of 20.94% of ruminant genomes consisted of high-precision TE sequences. The main TE families, by genomic size, included LINE/RTE-BovB (averaging 55.7% of total TE content), followed by LINE/L1 (18.13%), SINE/tRNA-Core-RTE (13.07%), and SINE/Core-RTE (8.16%) (**Fig. 2G**). When analyzing TE subfamily composition across ruminant genomes, we found that most subfamilies showed consistent proportions. However, some TE subfamilies varied significantly among different ruminants, particularly in the endogenous retroviruses LTR/ERVK family, such as ERV2-1-LTR_BT, BTLTR1, and ERV2-1-I_BT subfamilies (**Fig. 2H**). These subfamilies may have undergone unbalanced expansion across ruminants.

_k_-mer-based evolution analysis of TE families in ruminants

We performed a dynamic evolutionary analysis to examine how different TE families expanded unevenly in ruminants. Due to the challenges of comparing TE sequences directly across species, we used a reference-free *k*-mer clustering algorithm. This approach reduced sequence divergence bias among ruminant genomes while enabling comprehensive analysis of TE family distributions across ruminant species (**Fig. 3A**). Initially, we aggregated 218 Mb of TE consensus sequences from all ruminant species and generated *k*-mer spectra with k values ranging from 5 to 50 (Supplemental Fig S3). We identified the optimal k length of 42 based on the major peak in the spectrum. Using this optimal k value, we established the *k*-mer set for TE sequences. We then selected high-abundance *k*-mers that ranked in the top 1,000 and were detected in at least one species to represent the *k*-mer sets for each TE subfamily. In total, we retained 67,293 TE *k*-mers across 14 TE families, covering 205 subfamilies (Supplemental Table S4). Finally, we performed *k*-mer counting for this set on each ruminant genome to generate a TE *k*-mer count matrix across 32 ruminants.

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

We implemented hierarchical clustering analysis using the manhattan distance matrix calculated from the ruminant TE *k*-mer count matrix. As expected, the hierarchical relationships between these clusters corresponded with established ruminant species phylogeny (**Fig. 3B**). The analysis showed that ruminant species initially divided into two main clusters, aligning with the early Miocene period (~23–16 million years ago) (Lei et al. 2019). The first cluster contained most Bovidae, comprising a diverse group primarily from the Antilopinae and Capridae families. These taxonomic groups separated during the middle Miocene period, about 15–13 million years ago (Lei et al. 2019). The second cluster included Bovinae (a subdivision of Bovidae), Moschidae, Cervidae, Giraffidae, and Antilocapridae. The clustering method also effectively distinguished closely related ruminant species, such as black muntjac and Chinese muntjac within Muntiacinae. These findings showed that the *k*-mer-based clustering algorithm using TE successfully revealed evolutionary relationships among ruminant species.

To analyze the dynamic expansion of ruminant TEs, we performed hierarchical clustering using temporally specific TE families. These families were characterized by their divergence, which effectively differentiated the evolutionary ages of various TE families (**Fig. 3C**, Supplemental Fig S4). For example, MLT1F from LTR/ERVL-MaLR demonstrated predominant activity in earlier evolutionary periods, whereas ERV2-1-I_BT from LTR/ERVK exhibited more contemporary activity signatures (Supplemental Fig S5). Subsequently, we identified TE *k*-mer sets for each divergence bin to examine the temporal evolution of TE *k*-mer-based species divergence in ruminants. These sets were selected from the three most significant TE subfamilies (those with elevated ratios) within each divergence bin, representing the active TE *k*-mers during those respective temporal intervals. This facilitated the calculation of TE *k*-mer-based distances between each species and Pronghorn (a representative of the ancient ruminant lineage) (**Fig. 3D**). Our findings indicated that other ruminant species demonstrated reduced distance to Pronghorn when analyzing older TE subfamilies (divergence exceeding 20), and exhibited no substantial dispersal patterns among themselves. However, with decreasing TE family age, the TE *k*-mer-based distance

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

measurements increased, and the ordering became increasingly differentiated, with species from identical subfamilies exhibiting comparable divergence patterns. Notably, the genetic distance between Bovinae and Capridae had expanded considerably in recent evolutionary history, potentially related to the expansion of specific TE subfamilies.

We constructed dynamic trees utilizing TE *k*-mer sets for each divergence bin to illustrate taxon-specific expansion (**Fig. 3E**). Our analysis demonstrated that TE subfamilies from different active periods made distinct contributions to the evolutionary trajectories of various ruminants. At divergence 25—representing earlier evolutionary periods—the distance between these ruminants was minimal with no distinct group clustering evident. As divergence decreased from 20 to 5, however, the distance progressively increased with specific clustering patterns becoming apparent. The classification at divergence 5 aligned with the established evolutionary radiation of ruminant species, suggesting that TE differentiation during this period exhibited pronounced taxon-specificity. Notably, at divergence 1, we identified substantial differentiation in contemporary domesticated livestock compared to divergence 5, particularly regarding the increased genetic distance between species within the Bovinae family. These findings indicated a progressive expansion of TEs throughout ruminant genome evolution, pointing to a potential coevolutionary relationship between TEs and their host genomes.

Clustering of TE *k*-mers and their coevolution with the KZFP family

The interaction between TE-KZFP serves as a key regulatory mechanism for modulating TE activity in human genomes (Qian et al. 2024). To examine this interaction in ruminant species, we performed a comprehensive coevolution analysis of TE subfamilies and KZFPs. Initially, we applied *k*-means clustering to 38,451 *k*-mers based on their distribution across ruminant species to categorize *k*-mers with similar patterns. Our analysis identified 45 distinct TE *k*-mer clusters, with cluster sizes varying from 62 to 13,407 members (Supplemental Table S5). These TE *k*-mer clusters displayed distinctive distribution patterns across ruminants (**Fig. 4A**). The distribution included 17 pan-conserved clusters present in >90% of species, 14 Bovidae-shared clusters, 7 Bovidae-shared clusters (excluding Bovinae), and 7

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

subfamily-specific clusters (Supplemental Table S5). The largest cluster C1 (13,407 *k*-mers) demonstrated presence across all species examined. As Bovidae-specific clusters, the second and third largest—C2 (2,462 *k*-mers) and C3 (2,400 *k*-mers)—illustrated the evolutionary divergence between the Bovinae and other (non-Bovinae) subfamilies within Bovidae (**Fig. 4B**).

Our analysis mapping *k*-mer clusters to TE families identified significant enrichment of most TE families within the pan-conserved clusters (C1 cluster). LTR/ERVK and LTR/ERVL-MaLR, however, demonstrated distinct proliferation patterns with notable family/subfamily specificity (**Fig. 4C**). Of particular interest, LTR/ERVK subfamilies made substantial contributions to TE expansion within the Bovidae lineage (**Fig. 4D**), exhibiting three distinct patterns: (i) Bovinae-specific C2 clusters demonstrated predominant association with eight ERVK subfamilies; (ii) Bovidae-wide C3 clusters presented expansion bias in four ERVK subfamilies with reduced representation in Bovinae; and (iii) pan-Bovidae C5 clusters exhibited significant enrichment in four evolutionarily conserved ERVK subfamilies. These taxon-specific TE proliferation patterns indicated that *k*-mer clusters effectively represented the TE subfamilies that underwent specific expansion in ruminants.

To evaluate the coevolutionary relationship between KZFPs and TE families, we conducted a correlation analysis based on their distribution patterns across ruminants. Our analysis revealed that 90 KZFPs demonstrated a significant correlation ($R^2 > 0.6$, *p-value* < 0.01) with ruminant TE *k*-mer clusters (Supplemental Fig S6). Although TE *k*-mer clusters were targeted by multiple KZFPs, many KZFPs exhibited binding preferences for specific TE subfamilies. Notably, we identified three Bovidae-specific TE *k*-mer clusters, predominantly from the LTR/ERVK family, that were targeted by 32 KZFPs (**Fig. 4E**). These genes were primarily classified as LOC label genes (65.63%), exhibiting characteristic KZFP structure but lacking definitive annotations. Analysis of pairwise distance based on sequence indicated potential high orthology for these genes with taxon-specific TE *k*-mer clusters (**Fig. 4E**).

Exploration of TE-KZFP interactions

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

The distribution of KZFP-binding sites within human TEs suggests a significant prevalence of specific binding sites within these TE sequences (Julien et al. 2022). In the absence of ChIP-seq data for KZFPs in ruminants, we identified their DNA-binding sites through the analysis of potential DNA binding motifs in *k*-mer clusters. For comprehensive motif identification, we obtained potential DNA binding motifs from both JASPAR and predicted models (**Fig. 5A**). Specifically, we employed 25 unvalidated TFBS of KZFP from JASPAR 2024 (Jaime et al. 2021) and generated 75 DNA binding motifs of KZFPs by analyzing their tandem zinc finger domain sequences with the B1H-RC model (Najafabadi et al. 2015). Our analysis of 184 TE *k*-mer cluster-KZFP pairs (with $R^2 > 0.6$) demonstrated that 53.3% were identified (*p-value* of < 0.001), including 15 JASPAR-related and 53 model-predicted KZFP binding motifs (**Fig. 5B,** Supplemental Table S6). Within these pairs, numerous high-confidence KZFP binding sites exhibited robust correlations with species-specific clusters (**Fig. 5C**). The data revealed distinct evolutionary patterns: a ZNF789 isoform in C1 clusters ($R^2 = 0.67$) was present across ruminants; a ZNF10 isoform in C5 clusters ($R^2 = 0.70$) was exclusive to the Bovidae family; a ZNF565 isoform in C3 clusters ($R^2 = 0.69$) was found in the Bovidae family excluding the Bovinae subfamily; a ZNF182 isoform in C2 clusters ($R^2 = 0.91$) was characteristic of the Bovinae subfamily; and a ZNF484 isoform in C17 clusters ($R^2 = 0.99$) was consistently observed throughout Cervidae.

To investigate TE-KZFP interaction specificity, we examined the relationships between TE subfamilies and KZFPs through TE *k*-mer cluster analysis. Our results identified a notable divergence pattern within the ERVK family: despite high sequence similarity between certain ERVK subfamilies, they appeared in different TE *k*-mer clusters, indicating distinct KZFP interactions. Specifically, BTLTR1/BTLTRB1, which shared 88.06% sequence identity, separated into distinct *k*-mer clusters (C2 and C3, respectively) with different coevolutionary partners—ZNF182 isoform ($R^2 = 0.91$) for BTLTR1 and ZNF565 isoform ($R^2 = 0.91$) for BTLTRB1 (**Fig. 5D**). Further analysis of BTLTR1 and BTLTRB1's binding preferences demonstrated a significant correlation between KZFP-binding motif reliability (*p-value*) and coevolutionary patterns (R^2), substantiating these predicted interactions (**Fig. 5E**). These

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

findings indicated that sequence variations in key KZFP-binding regions facilitated evolutionary specialization in KZFP-TE partnerships, even between closely related TE subfamilies.

To validate these computational predictions, we compared our predicted TE-KZFP interactions with experimentally validated interactions in humans reported by Imbeault et al (Michaël et al. 2017). Despite substantial differences in TE sequences between humans and ruminants, we identified 12 known interactions supported by experimental evidence in humans. This successfully recapitulated 23.1% and 8.7% of predicted interactions for JASPAR and Predict KZFPs, respectively (Supplemental Table S7).

Functional effects of taxon-specific TEs on ruminant genomes

In our investigation of TE influence on functional evolution in ruminants, we conducted an analysis of sequence variations within protein-coding regions. For this analysis, we selected five well-annotated ruminant species that represent critical stages in Bovidae divergence: members of the Bovinae subfamily (Water Buffalo and Cattle), members of the Caprinae subfamily (Sheep and Goat), and Chinese Muntjac serving as an outgroup. Based on *k*-mer-based evolutionary analysis demonstrating that LTR/ERVK constituted the primary TE family exhibiting unbalanced expansion in Bovidae (**Fig. 4D**), we performed a comparative assessment of LTR/ERVK subfamily composition across these genomes. Our findings revealed that LTR/ERVK genome content was significantly elevated in Bovidae compared to Chinese Muntjac (**Fig. 6A**). Furthermore, distinct LTR/ERVK subfamilies demonstrated variable expansion patterns across taxonomic classifications (**Fig. 6B**). For example, BTLTR1B exhibited substantial expansion within Bovidae, while BTLTR1C demonstrated increased presence in the Caprinae subfamily relative to other species.

The evolutionary relationships among these species enabled us to classify the taxon-specific TE subfamilies into three distinct categories: Bovidae-specific, Bovinae-specific, and Caprinae-specific (**Fig. 6C**). Subsequently, we conducted gene annotation of these categorized TE subfamilies to examine TE-driven functional effects throughout various Bovidae evolutionary stages. Our analysis identified 14 genes exhibiting

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

exonization in their coding regions due to these taxon-specific TEs, comprising 3 Bovidae-specific, 6 Bovinae-specific, and 5 Caprinae-specific instances (**Fig. 6D,** Supplemental Table S8). The taxon-specific TE influenced on these genes predominantly manifested in the first-second exon (28.57%) and the last exon (71.43%). It was worth noting that when identical genes across different species were affected by taxon-specific TEs, the altered exon length and TE subfamily type remained consistent, suggesting that these TEs expanded prior to species differentiation and exerted uniform effects on these genes (**Fig. 6D**). For instance, regarding Bovidae-specific TEs affecting the *ACAD8* gene (**Fig. 6E**), BTLTR1D2 influenced the last exon, yielding a shorter exon compared to Chinese Muntjac, potentially resulting from TE insertion inducing premature termination. This gene functioned as a mitochondrial enzyme in the catabolism of the branched-chain amino acid valine and correlated with daily gain and beef tenderness in cattle (Hengde et al. 2007). Additionally, for Bovinae-specific TEs affecting the *IL18R1* gene (**Fig. 6F**), a significant immune regulation gene (Devasahayam Arokia Balaya et al. 2019), our observations revealed TE insertions in the first exon, with affected transcripts displaying markedly different lengths. Furthermore, Caprinae-specific TEs affecting the *IFI44* gene exhibited distinct exonization characteristics compared to other transcripts (**Fig. 6G**). Research indicates that *IFI44* influences the growth and development of Leiqiong cattle (Lei et al. 2020). Significantly, increased *IFI44* expression triggered an interferon response, consequently leading to the suppression of *MSTN* gene expression (Lei et al. 2020).

To explore the potential ongoing activity of taxon-specific TE in species, particularly as it related to environmental adaptation or domestication processes, we implemented a TE polymorphism analysis using cattle as our model system. Our investigation assessed the impact of eight Bovinae-specific TE subfamilies corresponding to C2 *k*-mer clusters on the cattle genome at the breed level (**Fig. 6C**). Analysis of 2,262 samples across 11 cattle breeds revealed 37,611 structural variants (PanCattle SV project, https://doi.org/10.6084/m9.figshare.28021577), with 59.34% attributable to TEs and 951 SVs specifically associated with Bovinae-specific TE subfamilies (Supplemental Table S9). The

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

data indicated that these SVs predominantly occur at high frequency throughout cattle populations (Supplemental Fig S7) while demonstrating minimal effects on exonic regions. TE-related SVs affected 204 genes within their gene bodies, with enrichment primarily observed in KEGG pathways related to neurodevelopment, including long-term depression and neurodegeneration (Supplemental Fig S8). Significantly, TE activity affected the exons of four specific genes: *LOC112446473*, *LOC112442367*, *LOC100848463*, and *LOC104975936*. Each of these affected genes exhibited first-exon alterations and were associated with endogenous retrovirus group K member Gag polyprotein-like proteins.

Regulatory roles of taxon-specific TEs in ruminants

To assess the regulatory influence of taxon-specific TEs in ruminants, we examined potential TE-derived cis-regulatory elements (CREs). Given its comprehensive epigenetic annotations across 16 tissues (S et al. 2022), the cattle genome was selected as the optimal representative model for ruminants. We performed an analysis of TE catalog enrichment from various age groups within different regulatory regions (**Fig. 7A**). Our findings indicated that the majority of TEs had lower levels of these epigenetic marks compared to other genomic regions. While this lower signal in younger TEs may be partially attributed to reduced sequence mappability within these highly repetitive regions, older TEs demonstrated a higher likelihood of enrichment relative to their younger counterparts. This pattern indicated that recently evolved TEs tend to minimize their genomic impact on the host, which corresponds with our previous research findings in porcine studies (Pengju et al. 2023a).

We characterized TE-derived CREs as sequences where a minimum of 50% originated from a singular annotated TE. This methodology revealed that TEs constitute approximately 6.39% (469,286/7,340,427) of all cattle CREs. Analysis by annotation type demonstrates that TE contribution varied significantly, ranging from 2.49% in H3K4me3 to 8.11% in ATAC. Regarding epigenetic states, TEs comprise 7.04% of promoters and enhancers, with comparatively lower percentages—1.77% and 3.57%—associated with active transcription start sites (TssA) and active enhancers (EnhA), respectively (**Fig. 7B**). TE-derived CREs exhibited greater tissue specificity compared to all cattle CREs, with 91.04% manifesting in

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

fewer than 3 tissues, while only 1.59% were identified across more than 5 tissues (**Fig. 7C**). This finding aligned with previous reports indicating that TEs contribute to tissue-specific regulatory elements in humans (Ali and Liang 2024). Despite the general depletion of TEs in CREs—particularly the Bovinae-specific subfamily within the LTR/ERVK family, which accounted for only 1.42% of TE-derived CREs—our investigation identified 1,645 Bovinae-specific TE-derived CREs associated with promoter and enhancer regions, including 11 at TssA and 23 at EnhA (**Fig. 7D**). Significantly, the predominant portion of these TE-derived TssA and EnhA (79.41%) were associated with digestive tissues, primarily in the abomasum and cecum (Supplemental Table S10).

We conducted motif scanning for CREs derived from TEs to examine how TE-embedded sequences function as TFBSs. The analysis identified 105 TE-derived TFBSs (*q-value* < 0.01) associated with 109 subfamilies across 8 TE families. Consistent with TE enrichment patterns in epigenetic marks, TE-derived TFBSs demonstrated higher prevalence in older TE families compared to younger ones (Supplemental Fig S9). The proportion of TEs contributing to TFBSs varied substantially, ranging from 6.67% in LTR-ERVK to 51.43% in LINE-L1. These TE-sequence TFBSs were predominantly located in quiescent and repressed regions, exhibited limited association with promoters, and primarily displayed repressive functional states (**Fig. 7E**). Notably, certain TFBSs including THRA and NFE2 showed greater association with accessible islands and enhancers.

Discussion

This study presents the first comprehensive annotation of the KZFP family in ruminant TEs, identifying an average of 271 KZFP isoforms characterized by their zinc finger, KRAB, and SCAN domains. To validate the annotation pipeline, we applied it to KZFPs in the human genome, successfully annotating 298 of 350 previously documented human KZFPs (85.14%), thereby confirming the reliability of our approach for ruminant species (Supplemental Table S11). Notably, our annotation framework distinguishes between KZFP proteins and their specific isoforms, revealing that apparent lineage-specific losses, such as for ZNF182, often represent the expansion of particular isoforms in certain clades rather than broad gene loss

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

393 events. The distribution patterns of the KZFP family across ruminant genomes align well with
394 established evolutionary relationships. We observe that the taxon-specific expansion of
395 KZFPs not only validates our annotation methodology but also facilitates the examination of
396 KZFP-TE coevolutionary processes. To gain deeper insights into ruminant TE relationships,
397 we employed global *de novo* assembly utilizing LongRepMarker for TE consensus sequence
398 detection, as opposed to utilizing conventional conserved TE sequences from standard
399 databases (specifically Repbase and Dfam). This methodological approach yields more
400 precise species-specific TE sequences. Through our comparative examination of TE family
401 distributions across ruminant genomes, we have determined that while predominant TE
402 families maintain relatively stable levels across species, specific minor families—notably
403 ERVK subfamilies—exhibit marked inter-species variability. The resulting 21% genome
404 coverage, lower than the ~50% typically reported in RepBase-based studies, reflects our more
405 stringent filtering for well-annotated TE consensus sequences (>100 bp) to ensure k-mer
406 analysis precision.

407 When comparing TE expansion patterns among ruminant species, a major challenge
408 arose: the heterogeneity in TE family expansions across species, combined with sequence
409 variation within TE subfamilies, hindered direct comparative analysis. To address this, we
410 developed and applied an improved *k*-mer-based approach that effectively segments TE
411 sequences, distinguishing conserved from species-specific TEs. Using this *k*-mer strategy, we
412 conducted quantitative analyses of evolutionary relationships across multiple ruminant
413 genomes. The results confirmed the method's effectiveness in species differentiation,
414 corroborating established evolutionary relationships and enabling accurate classification at
415 both family and species levels based on TE *k*-mer sequences.

416 To address the potential ambiguity between lineage-specific TE gains and losses—a
417 limitation raised in our initial analysis with only Pronghorn as an outgroup—we applied the
418 *k*-mer framework using two non-ruminant outgroups: human and pig. This comparative
419 approach was designed to validate the accuracy of our *k*-mer strategy by clearly
420 distinguishing conserved TEs from those specifically acquired in the ruminant lineage. The

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

k-mer-based hierarchical clustering across all species (Supplemental Fig S10) visually confirms the lineage-specific expansion of these TEs in ruminants, thereby supporting the robustness of our original evolutionary interpretation. Therefore, our evidence indicates that TE *k*-mer sequences serve as reliable proxies for TE conserved sequences when examining the evolutionary relationship between TEs and the KZFP family.

Our analysis of TE *k*-mer clustering patterns in ruminant genomes revealed distinct enrichment profiles across various TE families and subfamilies. The analysis identified multiple taxon-specific TE *k*-mer clusters. Evidence indicated that Bovidae family-specific TE *k*-mer clusters correlated significantly with LTR2_BT, LTR2B_BT, and LTR2C_BT, whereas Cervidae family-specific TE *k*-mer clusters demonstrated a clear association with BTLTR1E1. Additionally, Bovinae family-specific TE *k*-mer clusters exhibited significant correlations with eight ERVK subfamilies. These results provided valuable insights into the complex evolutionary relationships between TE families and subfamilies within ruminant genomes, highlighting distinct patterns of genomic expansion across different TE families. Through systematic comparison of TE *k*-mer cluster expansion patterns and KZFP family distribution across ruminant species, we have identified specific KZFPs that appeared to have co-evolved with particular TE *k*-mer clusters. Notably, a greater number of KZFPs demonstrated association with taxon-specific TEs, potentially attributable to the enhanced activity observed in these taxon-specific elements.

Studies have demonstrated that KZFPs interact with TE sequences through binding mechanisms. In our study, we employed TFBS matrices from the JASPAR 2024 database and the KZFP motifs prediction model (B1H-RC model) to assess this relationship in ruminants. Given the current lack of tissue-specific H3K9me3 ChIP-seq data for most ruminants, our use of the B1H-RC model and k-mer enrichment serves as a robust computational proxy for identifying potential repressive regulatory interactions. Our examination of co-evolving KZFP transcription binding sites across various TE *k*-mer clusters indicated that most TEs contain corresponding TFBS. When analyzing TE subfamilies that share sequence similarities but show enrichment in different TE *k*-mer clusters, we observed a notable correlation

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

449 between KZFP binding motif reliability and their coevolutionary patterns with TE *k*-mer
450 clusters. This finding suggests increased sequence binding reliability between TEs and their
451 predicted co-evolving KZFPs. Our findings confirmed that the C2 cluster displays expected
452 binding characteristics in comparison to the C3 cluster. These outcomes provide support for
453 the binding mechanism between KZFPs and TEs, enhancing our understanding of TE
454 sequence specificity through KZFPs in ruminants.

455 We utilized TE *k*-mers to identify taxon-specific TE subfamilies that emerged during
456 various stages of ruminant family evolution. We systematically classified LTR/ERVK TE
457 families into Bovidae-specific, Bovinae-specific, or Caprinae-specific categories to
458 investigate how TE formation influences genomic functional regions. Analysis revealed that
459 these taxon-specific TEs play a significant role in functional evolution through exonization in
460 14 genes. This process primarily affects the last exon (71.43%) and less frequently the
461 first-second exon (28.57%). We speculate this 3' bias is driven by cryptic polyadenylation
462 signals within TEs and lower selective pressure on terminal exons, which avoids disrupting
463 upstream protein-coding frames. Notable examples include *ACAD8* (associated with growth
464 and tenderness in cattle), *IL18R1* (involved in immune function), and *IFI44* (regulates growth
465 through *MSTN* suppression). Using cattle as a model organism, we determined that TEs
466 contribute to numerous SVs occurring at high frequency in modern cattle populations, thus
467 enhancing genetic diversity. Several of these diversity-contributing SVs are located within
468 genes, with four gene formations specifically linked to young ERVK subfamilies. Regarding
469 regulatory functions, our findings indicate that certain taxon-specific TEs overlap with
470 promoter and enhancer regions, although TEs have a relatively limited role as cis-regulatory
471 elements (approximately 6.39%). Notably, these TEs exhibiting tissue-specific and functional
472 characteristics show preferential enrichment in digestive tissue. Additionally, TE-embedded
473 sequences serve as TFBSs, predominantly in repressive regulatory regions. Future
474 experimental validation, such as ChIP-exo in cattle, will be essential to further substantiate
475 these predicted interactions and their tissue-specific functional roles.

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

In summary, our study introduced the novel TE *k*-mer methodology, offering a thorough analysis of TE evolution patterns in ruminants. Our systematic annotation and comparative analysis of the KZFPs family identified key coevolutionary relationships, which were subsequently confirmed through TFBS analysis. Additionally, we performed a comprehensive assessment of how taxon-specific families affect genomic variation, gene structure, CREs, and TFBSs. These findings enhance our knowledge of TE formation, regulation, and functional evolution in ruminants while shedding light on the complex mechanisms of TE-genome coevolution.

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

Methods

Annotation of KZFPs in ruminant genomes

Thirty-two high-quality ruminant genomes were obtained from the Ruminant Genome Database (RGD; http://animal.omics.pro/code/index.php/RGD). To establish a robust foundation for KZFP identification, representative template protein sequences were curated from the NCBI annotations of three comprehensively annotated ruminant species – cattle (*Bos taurus*, assembly ARS-UCD1.2), sheep (*Ovis aries*, assembly ARS-UI_Ramb_v3.0), and goat (*Capra hircus*, assembly ARS1.2) – alongside the human reference genome (*Homo sapiens*, assembly GRCh38).

We established a KZFP database for each of the 32 ruminant species utilizing makeblastdb (Ján et al. 2019) (v2.5.0), drawing from template protein sequences of all potential KZFP proteins. Subsequently, we conducted preliminary protein localization analysis across each ruminant genome employing BLASTN (Ján et al. 2019) (v2.5.0). For identified localizations, we defined candidate KZFP regions by extending sequences 10 kb in both upstream and downstream directions. To ensure precision, we implemented Exonerate (Guy and Ewan 2005) (v2.2.0) to perform detailed protein localization analysis on these regions for KZFP structural annotation, applying the following parameters: "--model est2genome --querytype dna --targettype dna --showvulgar no --softmaskquery no --softmasktarget yes --showalignment no --showtargetgff yes --showcigar no --score 30 --bestn 1 --percent 50".

We performed a comprehensive domain architecture analysis focusing on key KZFP-associated domains (C2H2 zinc finger (ZNF), KRAB, SCAN, and DUF3669) using HMMER v3.1b2-icc-14.0.2 (Simon et al. 2018). The analysis utilized calibrated profile hidden Markov models (HMMs) from the Pfam database (Jaina et al. 2020) (models: PF00096 for ZNF_C2H2, PF01352 for KRAB, PF02023 for SCAN, PF12417 for DUF3669). We applied specific score thresholds to ensure accuracy: ZNF_C2H2: 0 (per domain), KRAB: 13, SCAN: 10, DUF3669: 21. To identify putative KZFP isoforms, we conducted proximity analysis by examining genomic loci for instances where KRAB domain-containing proteins

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

were located near C2H2 zinc finger domain arrays (typically within 5-10 kb on the same scaffold). Proteins exhibiting this KRAB-ZNF structural arrangement were designated as candidate KZFPs.

Anonation and classificaiton of TE family

To comprehensively characterize TE sequences within ruminant genomes, we implemented an integrated dual-methodology framework combining *de novo* discovery with homology-based classification. Initially, *de novo* prediction was executed using LongRepMarker (v2.0) (Liao et al. 2021), leveraging its optimized algorithms to identify novel repeats without a priori references. Subsequently, a homology-based classification phase was performed by cross-referencing all putative repeats against the Repbase (v26.12) database, applying stringent 80-80 criteria (≥80% sequence identity over ≥80% of the repeat length) to taxonomically classify conserved repeats and filter false positives. The *de novo*-derived repeat libraries, augmented with Repbase-validated annotations, were then utilized to mask repetitive regions across each ruminant genome assembly via RepeatMasker (v4.1.5) (Tarailo-Graovac and Chen 2009), employing default parameters with RMblastn (Ján et al. 2019) as the alignment engine.

***k*-mer-based TE evolution analysis**

k-mer-based TE analysis began with systematic organization and classification of TE sequences across ruminant species. We conducted this process through sequence alignments against established TE reference databases using BLASTN (v2.5.0) (Ján et al. 2019). To identify the optimal *k*-mer length, we performed a comprehensive spectrum distribution analysis across a range from k = 5 to k = 200 using ntcard (v1.2.2) (Hamid et al. 2017). This evaluation determined k = 42 as the optimal length, representing maximum *k*-mer abundance.

We used Jellyfish (v2.2.10) (Guillaume and Carl 2011) to count all possible 42 *k*-mers within the consolidated ruminant TE sequences, creating a comprehensive TE-derived *k*-mer dataset. To focus on biologically significant signals, we extracted the 1,000 most abundant *k*-mers for each TE subfamily, requiring each selected *k*-mer to appear in at least one representative species within the subfamily to ensure broad taxonomic representation.

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

We quantified these representative TE *k*-mers across all ruminant genomes using BWA (v0.7.17-r1188) (Heng and Richard 2009) in "aln" alignment mode. This produced a structured *k*-mer count matrix—with rows representing species and columns representing TE subfamily-specific *k*-mers—that served as our foundation for subsequent analyses. For phylogenetic relationship analysis based on TE *k*-mer profiles, we calculated a Manhattan distance matrix from the normalized *k*-mer counts and applied hierarchical clustering analysis (hclust function in R). Additionally, to examine TE expansion dynamics over time, we conducted hierarchical clustering specifically on TE families grouped by their estimated evolutionary age, defined by sequence divergence levels from RepeatMasker outputs.

Finally, to identify patterns within the *k*-mer space, we applied *k*-means clustering (*kmeans* function in R) to the complete set of 38,451 identified TE *k*-mers based on their distribution profiles across all ruminant species. This grouped *k*-mers with similar cross-species distribution patterns regardless of their original TE subfamily classification. We calculate the ratio of specific *k*-mer clustering in each species by dividing the number of *k*-mers from a particular *k*-mer clustering present in that species by the total number of *k*-mers in that clustering.

TE-KZFP interactions analysis

To comprehensively identify DNA-binding sites, we employed an integrated approach combining database mining and de novo motif prediction. For established motif references, we retrieved 25 experimentally unvalidated TFBS profiles for KZFP from the JASPAR 2024 database (Supplemental Table S12). We then generated 78 additional DNA-binding motifs through computational prediction using the B1H-RC model (http://zifrc.ccbr.utoronto.ca/), which leverages protein sequence information from tandem zinc finger domains to infer binding specificities. All 130 combined motifs were scanned against TE-derived *k*-mer sequences using the FIMO tool (Charles et al. 2011) (version 4.11.2) from the MEME Suite (Timothy et al. 2015). This scanning was performed with default parameters and a significance threshold of *p-value* < 0.01 (--thresh 0.01), enabling genome-wide detection of putative binding sites.

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

Gene annotation

To comprehensively annotate all identified TE locations and elucidate their potential functional impacts, we utilized ANNOVAR (Wang et al. 2010) (version 20221005) for gene-centric annotation. This analysis integrated multi-species genome annotations from five ruminant lineages: Chinese muntjac (*Muntiacus reevesi*, GCA_028583565.1), water buffalo (*Bubalus bubalis*, GCF_019923935.1), cattle (*Bos taurus*, GCF_002263795.1), sheep (*Ovis aries*, GCF_016772045.2), and goat (*Capra hircus*, GCF_001704415.2). Each species'annotation file provided a hierarchical genomic context, enabling precise categorization of TE positions relative to functional regions. Specifically, we extracted: (1) exonic regions (further stratified by synonymous/non-synonymous effects), (2) splicing sites (±2 bp boundaries), (3) intronic regions, (4) 5′ and 3′ untranslated regions (UTRs), (5) regulatory regions (1-kb upstream/downstream of transcription start/end sites), and (6) intergenic regions.

KEGG enrichment analysis

We performed KEGG enrichment analysis on genes associated with SVs derived from TE polymorphism utilizing Metascape (Karnovsky et al. 2012). For the KEGG pathway enrichment analysis, we employed cattle gene IDs mapped against a comprehensive cattle gene background list after completing the necessary ID conversion process.

Definition of cis-regulatory elements

To comprehensively annotate TEs in the cattle genome, we leveraged five key epigenetic markers—ATAC-seq, H3K27ac, H3K27me3, H3K4me1 and H3K4me3—across 16 biologically diverse tissue samples derived from the Cattle functional annotation (Lingzhao et al. 2019). These markers were systematically integrated using a unified computational pipeline to characterize chromatin accessibility, active transcription, repressive domains, and enhancer signatures. Building on established chromatin state segmentation frameworks by ChromHMM (Ernst and Kellis 2017), 15 distinct functional states were decoded through combinatorial patterns of these epigenetic signals. Promoter-associated chromatin states into three subtypes: TssA (active promoters with strong transcription), TssAHet (flanking active

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

transcription start sites lacking ATAC-seq signal), and TssBiv (bivalent promoters). Enhancers were subdivided into five states: EnhA (strong active enhancers), EnhAMe (medium enhancers with ATAC-seq signal), EnhAWk (weak active enhancers), EnhAHet (active enhancers without ATAC-seq signal), and EnhPois (poised enhancers). Transcription start site (TSS)-proximal regions included TxFlnk (transcribed regions at gene bodies), TxFlnkWk (weakly transcribed regions), and TxFlnkHet (transcribed regions lacking ATAC-seq signal). The remaining states comprised Repr (repressed heterochromatin), ReprWk (weakly repressed regions), and Qui (quiescent/inactive regions).

Permetation test of TEs on functional regions

To comprehensively assess the enrichment patterns of specific TE families within distinct functional genomic regions, we employed a rigorous permutation-based statistical framework. This approach systematically evaluates whether the observed spatial association between TEs and functional annotations (e.g., promoters, enhancers, or repressive domains) exceeds random expectation. For each TE family, we conducted this analysis independently using the regionR (Gel et al. 2016) package (v1.34.0) in R (Null et al. 2011), leveraging its built-in functions for permutation, null modeling, and significance estimation. The permutation count (n = 100) was validated through saturation analysis to ensure stable Z-score estimation.

TFBS scaning on TE sequence

TE sequences from the cattle reference genome were extracted to identify TFBS within the TE sequences. Then, the FIMO software (Charles et al. 2011) (version 4.11.2), part of the MEME Suite (Timothy et al. 2015), was used to scan for TFBS occurrences across these sequences. The FIMO results with motif information, using the 841 validated transcription factors (TFs), were annotated from the Vertebrate JASPAR 2024 database (Jaime et al. 2021) (https://jaspar.elixir.no/). Eighty-eight unvalidated TFBS for KZFPs were used to predict the protein-DNA sequence binding. The FIMO software parameters were set to: *p-value* = 1e-4, --max-stored-scores 100000000, and --alpha = 1.

Ethics statement

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

The experimental data acquisition in this study strictly complied with the relevant provisions of the "Regulations for the Administration of Laboratory Animals" (March 2017 revision) of the Ministry of Science and Technology of China.

Sources of published datasets

The genome accession records from previous publications are accessible through the RGD database (http://animal.omics.pro/code/index.php/RGD). The genomic annotation files were obtained from the NCBI database for the following species: Chinese muntjac (GCA_028583565.1), water buffalo (GCF_019923935.1), cattle (GCF_002263795.1), sheep (GCF_016772045.2), and goat (GCF_001704415.2). Cattle genetic variation data, comprising 37,611 SVs, were obtained from 2,262 samples across 11 populations in our previous PanCattle SV project (figshare: https://doi.org/10.6084/m9.figshare.28021577) and GitHub repository (https://github.com/PengjuZ/CattleSV).

Data access

The TE *k*-mer matrix is available at https://github.com/PengjuZ/RuminantTE and in the Supplemental Data.

Software availability

All software and their respective versions used in this study are publicly available, as described in the methods section. The pipeline for KZFP annotation, *de novo* TE annotation, and *k*-mer-based TE analysis is available in Supplemental Code and at https://github.com/PengjuZ/RuminantTE.

Acknowledgements

This work is financially supported by National Natural Science Foundation of China General Program (32370678), Research Startup Funding from Hainan Institute of Zhejiang University (HZY20210401), High-performance Computing Platform of YZBSTCACC.

Author's contribution

P.Z., Z.W., and L.F. conceived and designed the experiments. P.Z. designed an analytical strategy and performed all analyses. J.H. and C.P. assisted in the bioinformatics

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

652 analysis. Y.Z. and C.W. collected for sequencing data. P.Z., Z.W., L.F., and D.Y. wrote and
653 revised the paper. All authors read and approved the final manuscript.

654 **Competing interests**

655 The authors declare no competing interests.

656

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

References

Ali A, Liang P. 2024. Transposable elements contribute to tissue-specific gene regulation in humans. *Genes Genomics* **46**: 1327-1343.

Anna DS, Todd SM. 2021. Transposable elements shape the evolution of mammalian development. *Nature Reviews Genetics* doi:10.1038/s41576-021-00385-1.

Bo X, Weimin Z, Guanglei Z, Xinru Z, Jiangshan B, Ran B, Aleksandra W, Emily H, Hannah A, Gwen E et al. 2024. On the genetic basis of tail-loss evolution in humans and apes. *Nature* doi:10.1038/s41586-024-07095-8.

Bo Y, Lu F, Qianqian G, Ce X, Junqin X, Zhenxia C, Yixuan W, Peng Y. 2022. Species-specific KRAB-ZFPs function as repressors of retroviruses by targeting PBS regions. *Proceedings of the National Academy of Sciences of the United States of America* doi:10.1073/pnas.2119415119.

Charles EG, Timothy LB, William Stafford N. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* doi:10.1093/bioinformatics/btr064.

Devasahayam Arokia Balaya R, Nupur A, Thottethodi Subrahmanya Keshava P, Richard KK, Yashwanth S, Sneha MP. 2019. A comprehensive pathway map of IL-18-mediated signalling. *Journal of Cell Communication and Signaling* doi:10.1007/s12079-019-00544-4.

Ernst J, Kellis M. 2017. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc* **12**: 2478-2492.

Fu W, Wang R, Nanaei HA, Wang J, Hu D, Jiang Y. 2022. RGD v2.0: a major update of the ruminant functional and evolutionary genomics database. *Nucleic Acids Res* **50**: D1091-D1099.

Gabriela E, Marco C, Annamaria K, Julien D, Andrea C, Sandra O, Michaël I, Helen MR, Priscilla T, Didier T. 2016. Transposable Elements and Their KRAB-ZFP Controllers Regulate Gene Expression in Adult Tissues. *Developmental Cell* doi:10.1016/j.devcel.2016.02.024.

Gel B, Diez-Villanueva A, Serra E, Buschbeck M, Peinado MA, Malinverni R. 2016. regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics* **32**: 289-291.

Guillaume B, Kathleen HB, Mary G, Vera G, Andrei S, Molly Gale H, Michaël I, Zsuzsanna I, Henry LL, Todd SM et al. 2018. Ten things you should know about transposable elements. *Genome Biology* doi:10.1186/s13059-018-1577-z.

Guillaume M, Carl K. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* doi:10.1093/bioinformatics/btr011.

Guy S, Ewan B. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* doi:10.1186/1471-2105-6-31.

Hamid M, Hira K, İnanç B. 2017. ntCard: a streaming algorithm for cardinality estimation in genomics data. *Bioinformatics* doi:10.1093/bioinformatics/btw832.

Heng L, Richard D. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* doi:10.1093/bioinformatics/btp324.

Hengde L, Shaohang X, Xue G, Hongyan R. 2007. Structure of the Bovine ACAD8 Gene and the Association of Its Polymorphism with the Production Traits. *Journal of Genetics and Genomics* doi:10.1016/s1673-8527(07)60033-2.

Irina RA. 2018. Neutral Theory, Transposable Elements, and Eukaryotic Genome Evolution. *Molecular Biology and Evolution* doi:10.1093/molbev/msy083.

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

Jaime AC-M, Rafael R-P, Ieva R, Roza Berhanu L, Laura T, Romain B-M, Jérémy L, Paul B, Aziz K, Nicolás Manosalva P et al. 2021. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* doi:10.1093/nar/gkab1113.

Jaina M, Sara C, Lowri W, Matloob Q, Gustavo S, Erik LLS, Silvio CET, Lisanna P, Shriya R, Lorna R et al. 2020. Pfam: The protein families database in 2021. *Nucleic Acids Research* doi:10.1093/nar/gkaa913.

Ján L, Richard PB, Miroslav F, Günter G, Zdeněk G, Ivan H, Jan H, Ferenc K, Anikó K, Lenka M et al. 2019. Basic Local Alignment Search Tool. *IUPAC Standards Online* doi:10.1515/iupac.90.0375.

Jessica MS, Robert H, Jeb RR, Travis JW, Arian FAS. 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA* doi:10.1186/s13100-020-00230-y.

Jonas de T-H, Christian WT, Romain F, Evarist P, Julien D, Alexandre C, Bara K, Sandra O, Cyril P, Jacques F et al. 2023. Genetic features and genomic targets of human KRAB-zinc finger proteins. *Genome Research* doi:10.1101/gr.277722.123.

Julien P, Cyril P, Christopher JP, Evarist P, Delphine G, Sandra O, Julien D, Andrea M, Matthias PL, Didier T. 2022. Primate-specific transposable elements shape transcriptional networks during human development. *Nature Communications* doi:10.1038/s41467-022-34800-w.

Karnovsky A, Weymouth T, Hull T, Tarcea VG, Scardoni G, Laudanna C, Sartor MA, Stringer KA, Jagadish HV, Burant C et al. 2012. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. *Bioinformatics* **28**: 373-380.

Lanhua T, Benjamin S, Sébastien D, Chad H, Gabriel Costa Monteiro M, Keith D, Maria A, Erik M, Arnaud S, Latifa K et al. 2024. GWAS reveals determinants of mobilization rate and dynamics of an active endogenous retrovirus of cattle. *Nature Communications* doi:10.1038/s41467-024-46434-1.

Lei C, Qiang Q, Yu J, Kun W, Qisheng W, Zhipeng L, Faysal B, Yongzhi Y, Jinhuan W, Wei N et al. 2019. Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. *Science* doi:10.1126/science.aav6202.

Lei K, Guangbin L, Ming D, Zhiquan L, Yinru H, Baoli S, Yongqing G, Dewu L, Yaokun L. 2020. Growth retardation-responsive analysis of mRNAs and long noncoding RNAs in the liver tissue of Leiqiong cattle. *Scientific Reports* doi:10.1038/s41598-020-71206-4.

Liao X, Li M, Hu K, Wu F-X, Gao X, Wang J. 2021. A sensitive repeat identification framework based on short and long reads. *Nucleic Acids Res* **49**: e100.

Lingzhao F, Shuli L, Mei L, Xiaolong K, Shudai L, Bingjie L, Erin EC, Baldwin VIRL, Albert T, Li M et al. 2019. Functional annotation of the cattle genome through systematic discovery and characterization of chromatin states and butyrate-induced variations. *BMC Biology* doi:10.1186/s12915-019-0687-8.

Marta F, Jaebum K, Anastasia AP, Yang Z, Anastasia IK, Qiye L, Yang Z, Yingqi X, Jennifer LJ, Polina LP et al. 2019a. Evolution of gene regulation in ruminants differs between evolutionary breakpoint regions and homologous synteny blocks. *Genome Research* doi:10.1101/gr.239863.118.

Marta M, Rafał C, Kornel K, Przemysław B, Triantafillos L, Andrzej M, Urszula O. 2019b. The expression signature of cancer□associated KRAB□ZNF factors identified in TCGA pan□cancer transcriptomic data. *Molecular Oncology* doi:10.1002/1878-0261.12407.

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

Michaël I, Pierre-Yves H, Didier T. 2017. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* doi:10.1038/nature21683.

Najafabadi HS, Mnaimneh S, Schmitges FW, Garton M, Lam KN, Yang A, Albu M, Weirauch MT, Radovani E, Kim PM et al. 2015. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat Biotechnol* **33**: 555-562.

Null RCTR, Team R, Null RCT, Core Writing T, Null R, ., Team R, Null RDCT, Core R, Team R, Team RDC. 2011. R: A language and environment for statistical computing. *Computing* **1**: 12-21.

Pengju Z, Lihong G, Yahui G, Zhangyuan P, Lei L, Xingzheng L, Huaijun Z, Dongmin Y, Xinyan H, Lichun Q et al. 2023a. Young SINEs in pig genomes impact gene regulation, genetic diversity, and complex traits. *Communications biology* doi:10.1038/s42003-023-05234-x.

Pengju Z, Peng C, Lingzhao F, Zhengguang W, George EL. 2023b. Taming transposable elements in livestock and poultry: a review of their roles and applications. *Genetics Selection Evolution* doi:10.1186/s12711-023-00821-2.

Pierre-Yves H, Moritz H, Julien D, Cécile P, Christian WT, Andrea C, Julien P, Michaël I, Priscilla T, Ruedi A et al. 2019. The interactome of KRAB zinc finger proteins reveals the evolutionary history of their functional diversification. *The EMBO Journal* doi:10.15252/embj.2018101220.

Qian W, Junqin X, Jiaxin Z, Yixuan W, Peng Y. 2024. Exploration of Evolutionary Mechanisms Shaping KZFP and Retrotransposon Interactions. *Science Bulletin* doi:10.1016/j.scib.2024.07.023.

R H, D S, G C, M G, KE H-K, M M. 2025. Starship giant transposable elements cluster by host taxonomy using k-mer-based phylogenetics. *G3 (Bethesda, Md)* **15**.

Raquel F, Julius J, Cédric F, Joanna W. 2022. Roles of transposable elements in the regulation of mammalian transcription. *Nature Reviews Molecular Cell Biology* doi:10.1038/s41580-022-00457-y.

Roy NP, Michael WV, David AR. 2018. Mammalian transposable elements and their impacts on genome evolution. *Chromosome Research* doi:10.1007/s10577-017-9570-z.

S L, Y G, O C-X, S W, Y Y, W C, B L, R X, AJ C, E P-C et al. 2022. A multi-tissue atlas of regulatory variants in cattle. *Nature genetics* **54**: 1438-1447.

Simon P, Aurélien L, Sean RE, Young Mi P, Rodrigo L, Alex B. 2018. HMMER web server: 2018 update. *Nucleic Acids Research* doi:10.1093/nar/gky448.

Siqian C, Shuli L, Shaolei S, Yifan J, Mengli C, Yongjie T, Wenlong L, Jianfeng L, Lingzhao F, Ying Y et al. 2022. Comparative epigenomics reveals the impact of ruminant-specific regulatory elements on complex traits. *BMC Biology* doi:10.1186/s12915-022-01459-0.

Sophie L, Gaël C. 2020. Measuring and interpreting transposable element expression. *Nature Reviews Genetics* doi:10.1038/s41576-020-0251-y.

Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **Chapter 4**: 4 10 11-14 10 14.

Timothy LB, James EJ, Charles EG, William Stafford N. 2015. The MEME Suite. *Nucleic Acids Research* doi:10.1093/nar/gkv416.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**: e164.

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

787 Weidong B, Kenji K, Oleksiy K. 2015. Repbase Update, a database of repetitive elements in eukaryotic
788 genomes. *Mobile DNA* doi:10.1186/s13100-015-0041-9.
789

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure legends

Figure 1. Ruminant KZFP-TE regulatory dynamics. The top-left panel illustrates the study's focus on diverse ruminant species with complex phylogenetic relationships. The bottom-left panel highlights the central scientific question regarding the potential arms race between TE and KZFP families and its implications for genome evolution and biology. The right panel outlines our integrated analysis pipeline: top, KZFP annotation based on domain structures (KRAB, SCAN, and zf-C2H2) and TE annotation utilizing a novel *k*-mer-based strategy; bottom, the key analytical outcomes, including taxon-specific TE expansion, KZFP co-evolutionary patterns, and the preferential enrichment of TEs in cis-regulatory elements (CREs).

Figure 2. Annotation of KZFPs and TE Families. A. The phylogenetic tree illustrates the evolutionary diversification across 32 ruminant species, showing speciation events and divergence timescales. **B.** The bar plot displays the number of annotated KZFPs for each ruminant species. **C.** The multi-segmented pie chart shows the proportional distribution of KZFP domain architectures, categorizing structural configurations for KRAB, SCAN, and zf-C2H2. **D.** The hierarchical clustering based on KZFP distribution patterns demonstrates evolutionary relationships among ruminant species. **E.** The scatter plot characterizes rank-ordered family-differential KZFP proportions, visualizing taxon-specific expansions or contractions of particular KZFP families. The x-axis represents the rank distribution ordered from highest to lowest differential KZFP ratio. The y-axis represents the differential KZFP ratio between the specific taxon and other ruminants. **F.** The heatmap tracks the presence/absence status of eight KZFPs across all 32 species. **G.** The box plot shows genomic landscape occupation by TE families, revealing substantial variation in TE content proportions. **H.** The box plot illustrates the variation coefficient (CV) in TE subfamily frequency across the analyzed ruminant species.

Figure 3. Evolutionary dynamics of TE families across ruminants based on *k*-mer-based analyses. A. Reference-free TE subfamily clustering pipeline. Schematic workflow for identification and comparison of TE subfamilies using *k*-mer frequency spectra. This

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

approach enables cross-species comparison of TE landscapes without genome assembly, highlighting shared and taxon-specific TE distributions. **B.** Hierarchical clustering of ruminant species based on TE-derived *k*-mers. Unrooted tree illustrating phylogenetic relationships inferred from genome-wide TE *k*-mer similarity. **C.** Correlation between TE subfamily expansion and evolutionary age. Dot plot depicting the proportional abundance of TE subfamilies (y-axis) relative to their estimated evolutionary ages (x-axis, inferred from Kimura divergence). Dots are binned by divergence intervals, revealing episodic bursts of TE activity. **D.** Pairwise *k*-mer distance comparison of Pronghorn (Antilocapridae) TEs. Dot plot showing genetic distances (y-axis) between Pronghorn and other ruminant families based on TE *k*-mer profiles. **E.** Phylogenetic trees of 32 ruminant species reconstructed from TE *k*-mers at varying divergence thresholds. Phylogenetic topologies generated using TE *k*-mers filtered at divergence levels of 1%, 5%, 15%, 20%, and 25%. For each phylogenetic tree, different colors distinguish major taxonomic groups at various divergence levels.

Figure 4. Coevolution analysis between the KZFP and TE families. A. TE *k*-mer cluster enrichment across ruminant genomes. Heatmap displaying the distribution and enrichment patterns of TE *k*-mer clusters in 32 ruminant genomes. Rows represent specific TE *k*-mer clusters, columns denote individual genomes. Color intensity indicates the ratio of specific kmer clustering present in each species. **B.** Compositional dynamics of major TE *k*-mer clusters. Dot plot illustrating the proportional representation of four conserved TE *k*-mer clusters (C1/C2/C3/C5) across ruminant genomes. **C-D.** TE *k*-mer cluster enrichment in family and subfamily. Heatmap correlating eight major TE families and 28 LTR/ERVK subfamilies with *k*-mer clusters. Color intensity reflects enrichment levels of family or subfamily within each cluster. **E.** KZFP-TE coevolution and sequence divergence. Left: correlation heatmap between KZFPs and TE *k*-mer clusters. Rows: KZFP genes; columns: TE clusters (C2/C3/C5). Red indicates a positive correlation, suggesting KZFP putative binding preferences for specific TE *k*-mer clusters. Right: heatmap showing pairwise distance matrix of KZFP protein sequences.

Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

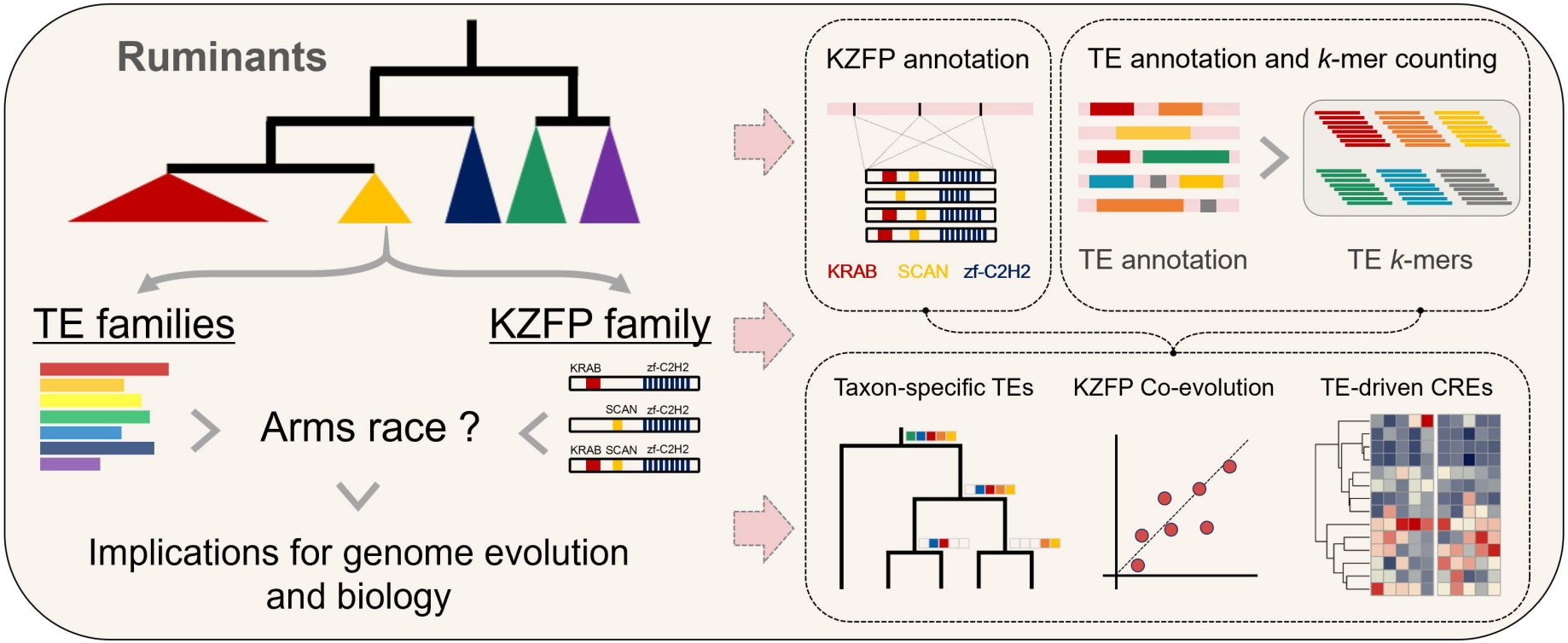
Figure 5. Analysis and verification of ruminant TE-KZFP interactions. A. Schematic of TE-KZFP interaction and binding site identification. KZFP-binding sites were identified by integrating *de novo* motifs from JASPAR database and B1H-RC model predictions. **B.** Proportion of TE-derived *k*-mers containing KZFP-Binding motifs. Pie charts show the proportion of KZFP-binding motifs among all motif-containing k-mers derived from TEs. **C.** Position weight matrices of co-evolved KZFP motifs. Sequence logos and position weight matrices (PWMs) representing DNA-binding motifs of KZFPs showing coevolutionary with TE *k*-mer clusters. Motifs are stratified by TE family specificity, with matrix height reflecting nucleotide conservation at each position. **D.** ERVK family phylogeny and TE *k*-mer cluster enrichment. Left: Maximum-likelihood phylogenetic tree of ERVK subfamilies, reconstructed from consensus TE sequences. Right: Heatmap showing enrichment scores of TE *k*-mer clusters across ERVK subfamilies. Color intensity denotes enrichment levels of ERVK family within each cluster. **E.** Correlation between KZFP motif reliability and coevolution with TE *k*-mers. Scatter plot evaluating the relationship between KZFP motif prediction confidence and coevolutionary patterns with TE *k*-mer clusters. X-axis: Difference in correlation R^2 between C2 and C3 clusters. Y-axis: Negative log-transformed ratio of motif detection *p-values* (C2/C3).

Figure 6. Impact of taxon-specific TE genes on ruminant genomes. A. Bar plots quantify the total genomic content of the ERVK family across five ruminant species. **B.** A dot-based heatmap visualizes the abundance of individual ERVK subfamilies (rows) in each ruminant genome (columns). Dot size and red color intensity scale with genomic occupancy. This reveals subfamily-specific proliferation bursts, with larger/darker dots indicating subfamilies that underwent significant expansion in particular species. **C.** Phylogenetic distribution of taxon-specific TE. A maximum-likelihood tree depicts evolutionary relationships among the five ruminant species. Concentric arcs around the tree illustrate the distribution of genes disrupted by family-specific and subfamily-specific TEs. **D.** Exonization events driven by taxon-specific TEs. A table catalogs 14 genes exhibiting TE-induced exonization. **E.**

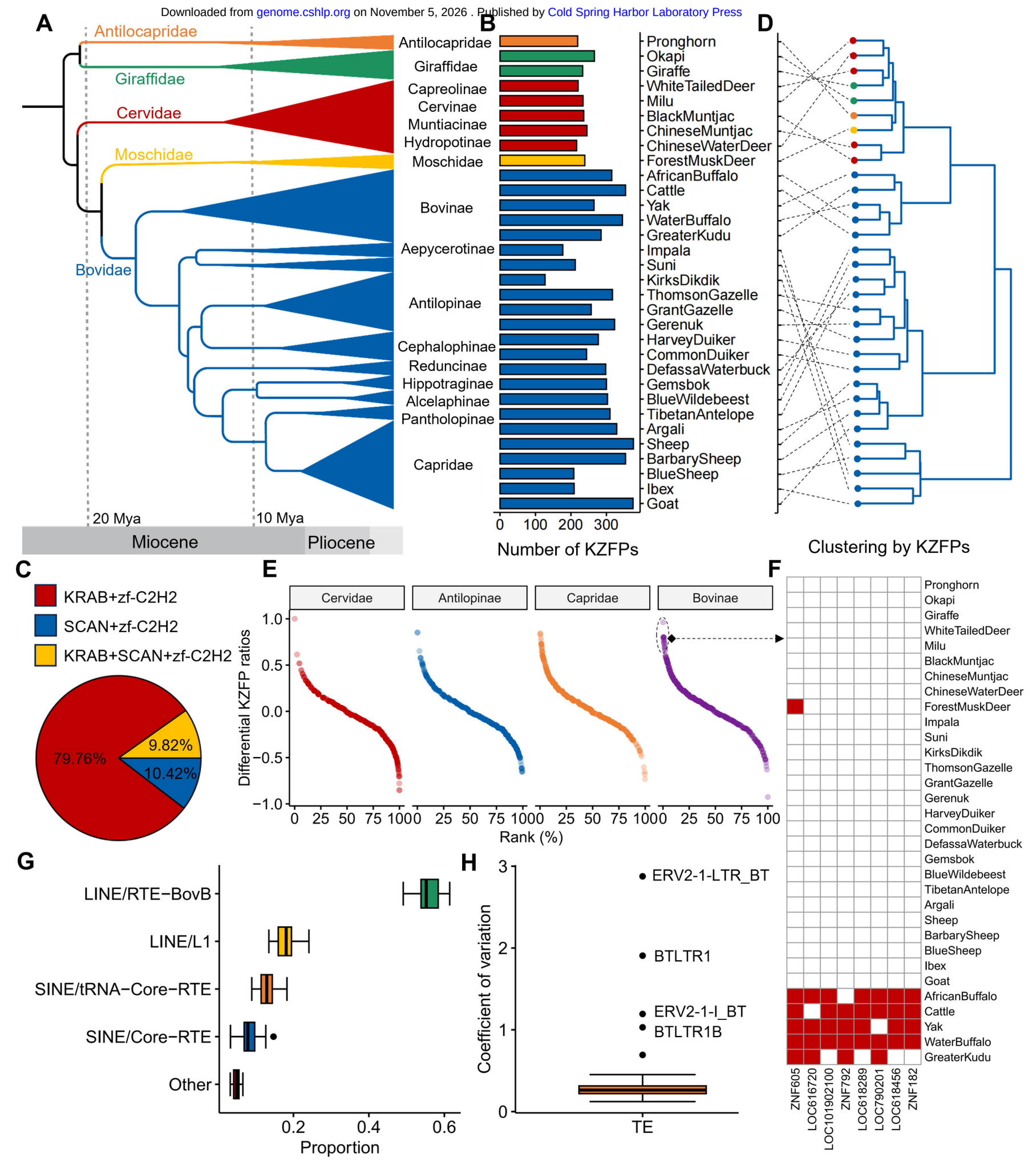
Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press

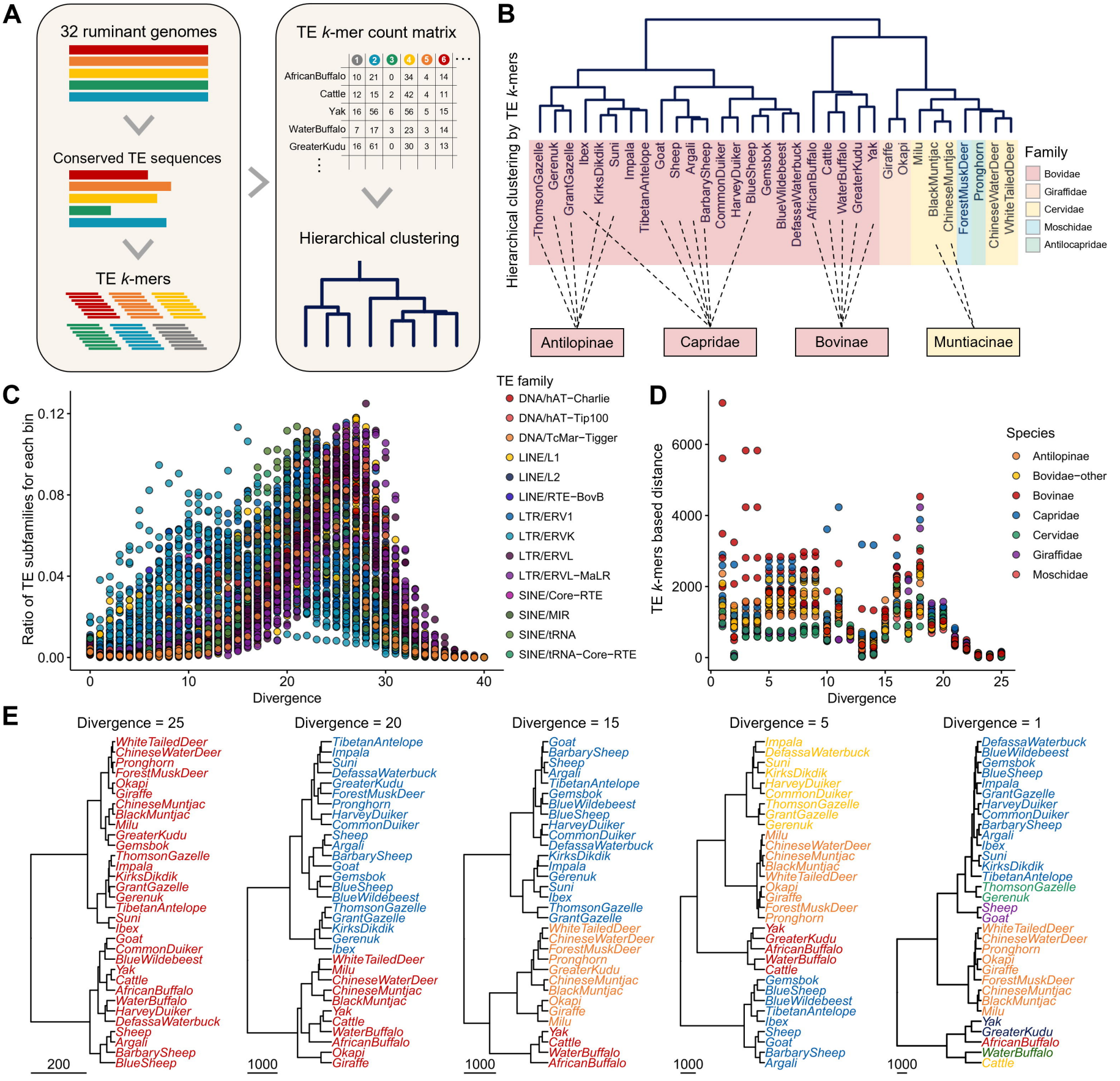
Bovidae-specific exonization in *ACAD8*. **F.** Bovinae-specific exonization in *IL18R1*. **G.** Caprinae-specific exonization in *IFI44*.

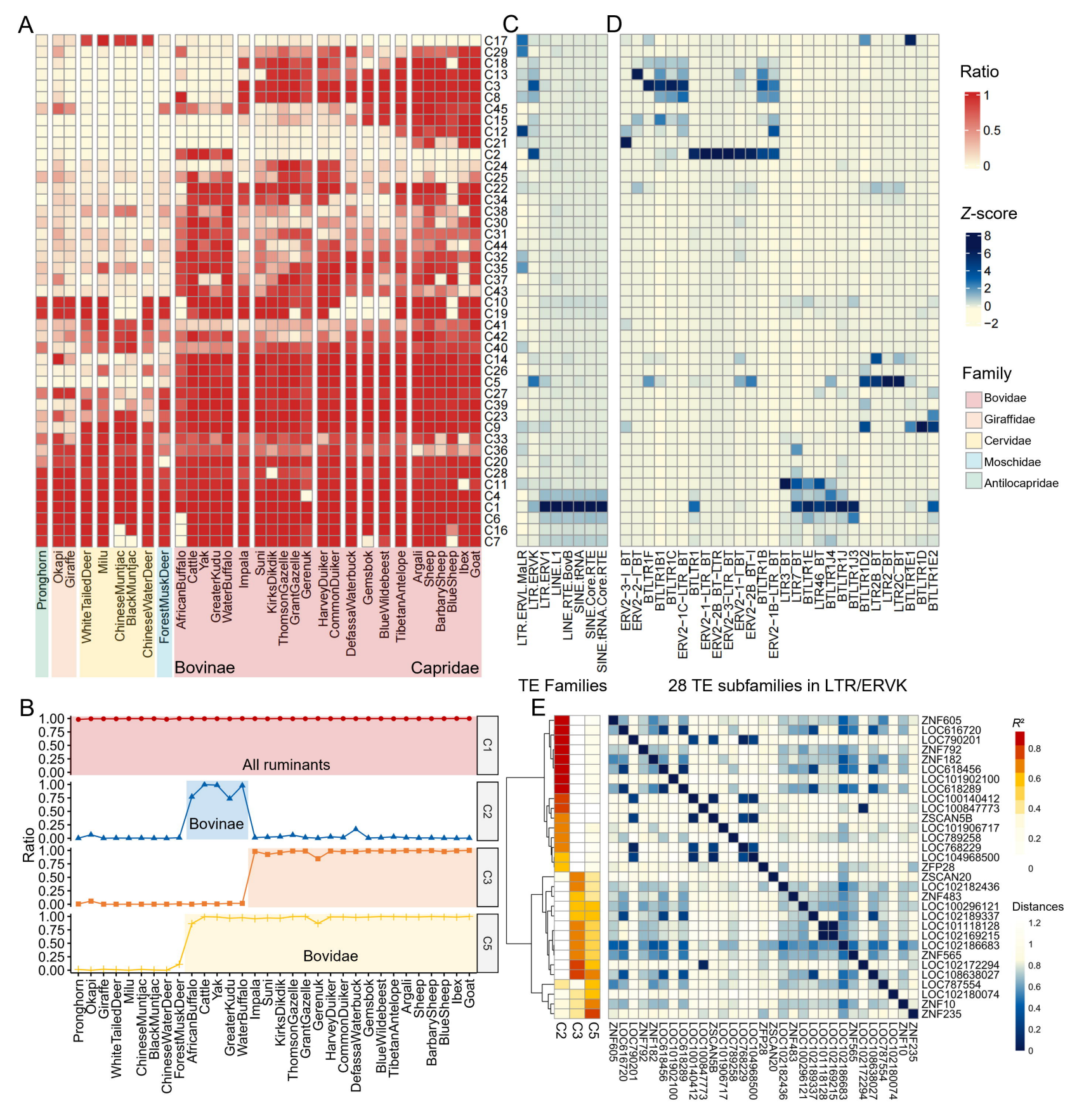
Figure 7. Epigenetic regulation of recent TEs in ruminants. A. Distribution of TE enrichment across epigenetic markers. Boxplot illustrates systematic enrichment patterns of 8 TE families relative to five functionally distinct epigenetic markers (ATAC, H3K27ac, H3K4me1, H3K4me3, and H3K27me3). Peaks signify genomic regions where TEs co-localize with active/repressive marks, suggesting TE roles in regulatory elements. Significance levels: * *p-value* < 0.05, ** *p-value* < 0.01, *** *p-value* < 0.001. **B.** Contribution of TE-derived CREs by functional class. Barplot quantifies the percentage of TE-derived CREs within distinct functional annotations (y-axis: e.g., enhancers, promoters, insulators, and repressed regions). X-axis spans 0–100%, revealing TEs' preferential contribution to specific regulatory classes. **C.** Tissue specificity of TE-derived vs. all CREs. Grouped bar plot compares the distribution of all CREs (red fill) and TE-derived CREs (blue fill) across tissues. **D.** Tissue-gene association of TE-derived CREs. Heatmap shows rows: genes with TE-derived CREs; columns: tissues. Red = CRE presence; white = CRE absence. **E.** ERVK-derived TFBSs. Left: Bar plot showing TFBS counts (y-axis) for key transcription factors within ERVK-derived CREs. Right: Proportional bar plot depicting the genomic context of ERVK TFBSs.

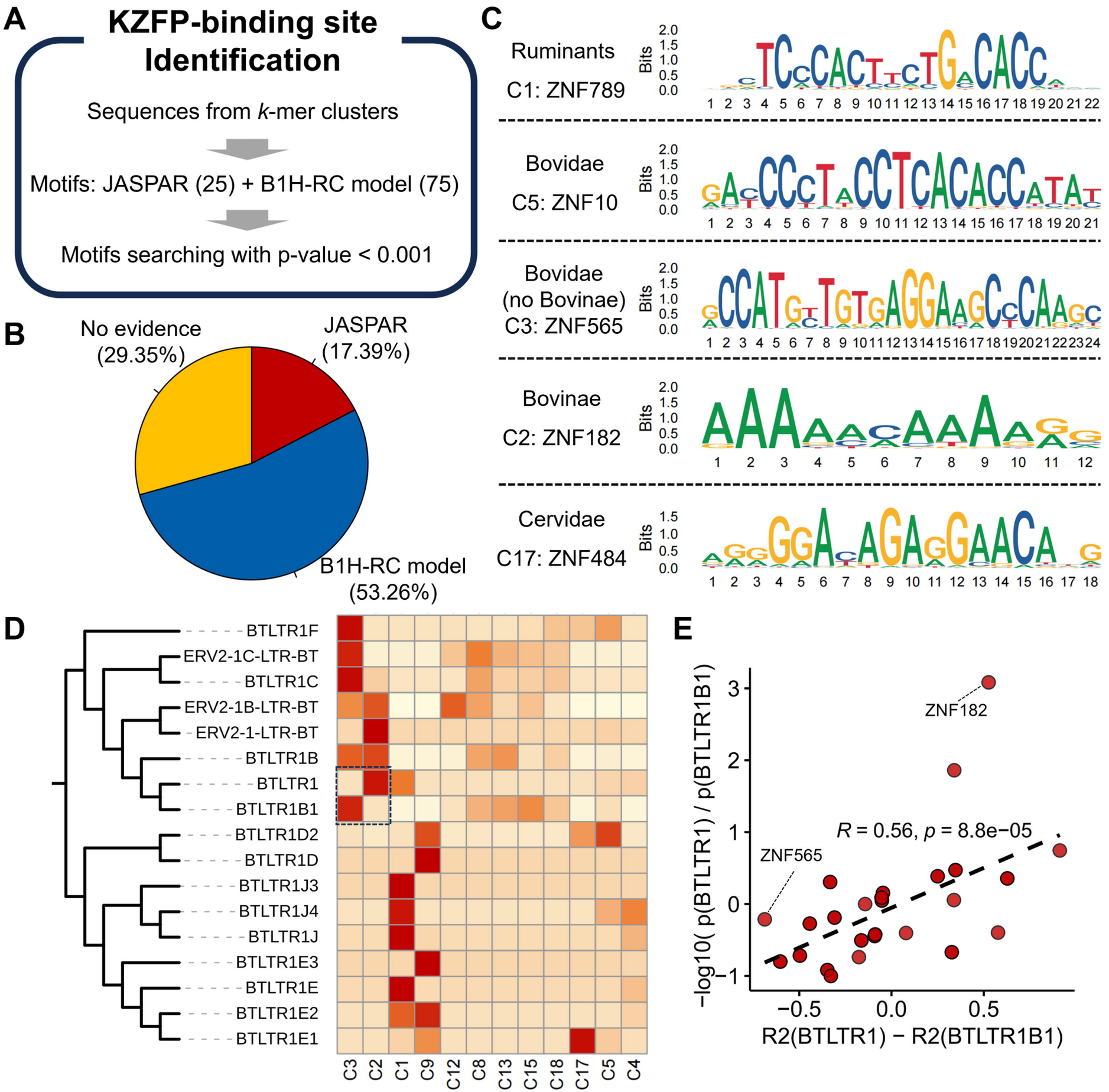

Ruminants
TE families
KZFP family
KRAB
zf-C2H2
SCAN
zf-C2H2
KRAB SCAN
zf-C2H2
Arms race ?
Implications for genome evolution and biology
KZFP annotation
KRAB
SCAN
zf-C2H2
TE annotation and k-mer counting
TE annotation
TE k-mers
Taxon-specific TEs
KZFP Co-evolution
TE-driven CREs

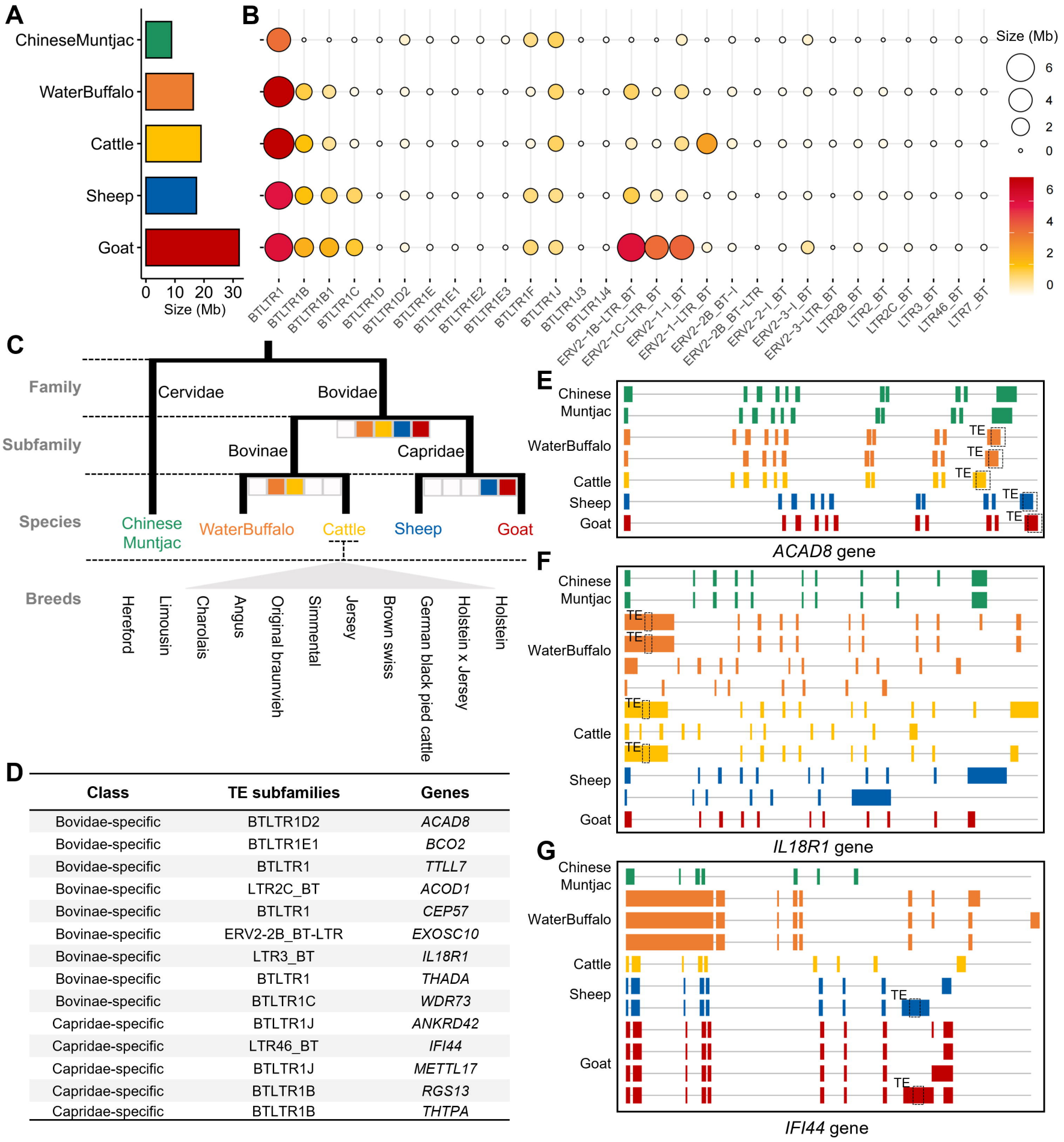
Downloaded from genome.cshlp.org on November 5, 2026 . Published by Cold Spring Harbor Laboratory Press




A
32 ruminant genomes
Conserved TE sequences
TE k-mers
TE k-mer count matrix
AfricanBuffalo 10 21 0 34 4 14
Cattle 12 15 2 42 4 11
Yak 16 56 6 56 5 15
WaterBuffalo 7 17 3 23 3 14
GreaterKudu 16 61 0 30 3 13
Hierarchical clustering
B
Hierarchical clustering by TE k-mers
ThomsonGazelle Gerenuk GrantGazelle Ibex KirksDikdik Suni Impala TibetanAntelope Goat Sheep Argali BarbarySheep CommonDuiker HarveyDuiker BlueSheep Gemsbok BlueWildebeest DefassaWaterbuck AfricanBuffalo Cattle WaterBuffalo GreaterKudu Yak Giraffe Okapi Milu BlackMuntjac ChineseMuntjac ForestMuskDeer Pronghorn ChineseWaterDeer WhiteTailedDeer
Family
Bovidae
Giraffidae
Cervidae
Moschidae
Antilocapridae
Antilopinae
Capridae
Bovinae
Muntiacinae
C
Ratio of TE subfamilies for each bin
0.12
0.08
0.04
0.00
0
10
20
30
40
Divergence
TE family
DNA/hAT−Charlie
DNA/hAT−Tip100
DNA/TcMar−Tigger
LINE/L1
LINE/L2
LINE/RTE−BovB
LTR/ERV1
LTR/ERVK
LTR/ERVL
LTR/ERVL−MaLR
SINE/Core−RTE
SINE/MIR
SINE/tRNA
SINE/tRNA−Core−RTE
D
TE k-mers based distance
6000
4000
2000
0
0
5
10
15
20
25
Divergence
Species
Antilopinae
Bovidae−other
Bovinae
Capridae
Cervidae
Giraffidae
Moschidae
E
Divergence = 25
WhiteTailedDeer ChineseWaterDeer Pronghorn ForestMuskDeer Okapi Giraffe ChineseMuntjac BlackMuntjac Milu GreaterKudu Gemsbok ThomsonGazelle Impala KirksDikdik GrantGazelle Gerenuk TibetanAntelope Suni Ibex Goat CommonDuiker BlueWildebeest Yak Cattle AfricanBuffalo WaterBuffalo HarveyDuiker DefassaWaterbuck Sheep Argali BarbarySheep BlueSheep
200
Divergence = 20
TibetanAntelope Impala Suni DefassaWaterbuck GreaterKudu ForestMuskDeer Pronghorn HarveyDuiker CommonDuiker Sheep Argali BarbarySheep Goat Gemsbok BlueSheep BlueWildebeest ThomsonGazelle GrantGazelle KirksDikdik Gerenuk Ibex WhiteTailedDeer Milu ChineseWaterDeer ChineseMuntjac BlackMuntjac Yak Cattle WaterBuffalo AfricanBuffalo Okapi Giraffe
1000
Divergence = 15
Goat BarbarySheep Sheep Argali TibetanAntelope Gemsbok BlueWildebeest BlueSheep HarveyDuiker CommonDuiker DefassaWaterbuck KirksDikdik Impala Gerenuk Suni Ibex ThomsonGazelle GrantGazelle WhiteTailedDeer ChineseWaterDeer ForestMuskDeer Pronghorn GreaterKudu ChineseMuntjac BlackMuntjac Okapi Giraffe Milu Yak Cattle WaterBuffalo AfricanBuffalo
1000
Divergence = 5
Impala DefassaWaterbuck Suni KirksDikdik HarveyDuiker CommonDuiker ThomsonGazelle GrantGazelle Gerenuk Milu ChineseWaterDeer ChineseMuntjac BlackMuntjac WhiteTailedDeer Okapi Giraffe ForestMuskDeer Pronghorn Yak GreaterKudu AfricanBuffalo WaterBuffalo Cattle Gemsbok BlueSheep BlueWildebeest TibetanAntelope Ibex Sheep Goat BarbarySheep Argali
1000
Divergence = 1
DefassaWaterbuck BlueWildebeest Gemsbok BlueSheep Impala GrantGazelle HarveyDuiker CommonDuiker BarbarySheep Argali Ibex Suni KirksDikdik TibetanAntelope ThomsonGazelle Gerenuk Sheep Goat WhiteTailedDeer ChineseWaterDeer Pronghorn Okapi Giraffe ForestMuskDeer ChineseMuntjac BlackMuntjac Milu Yak GreaterKudu AfricanBuffalo WaterBuffalo Cattle
1000

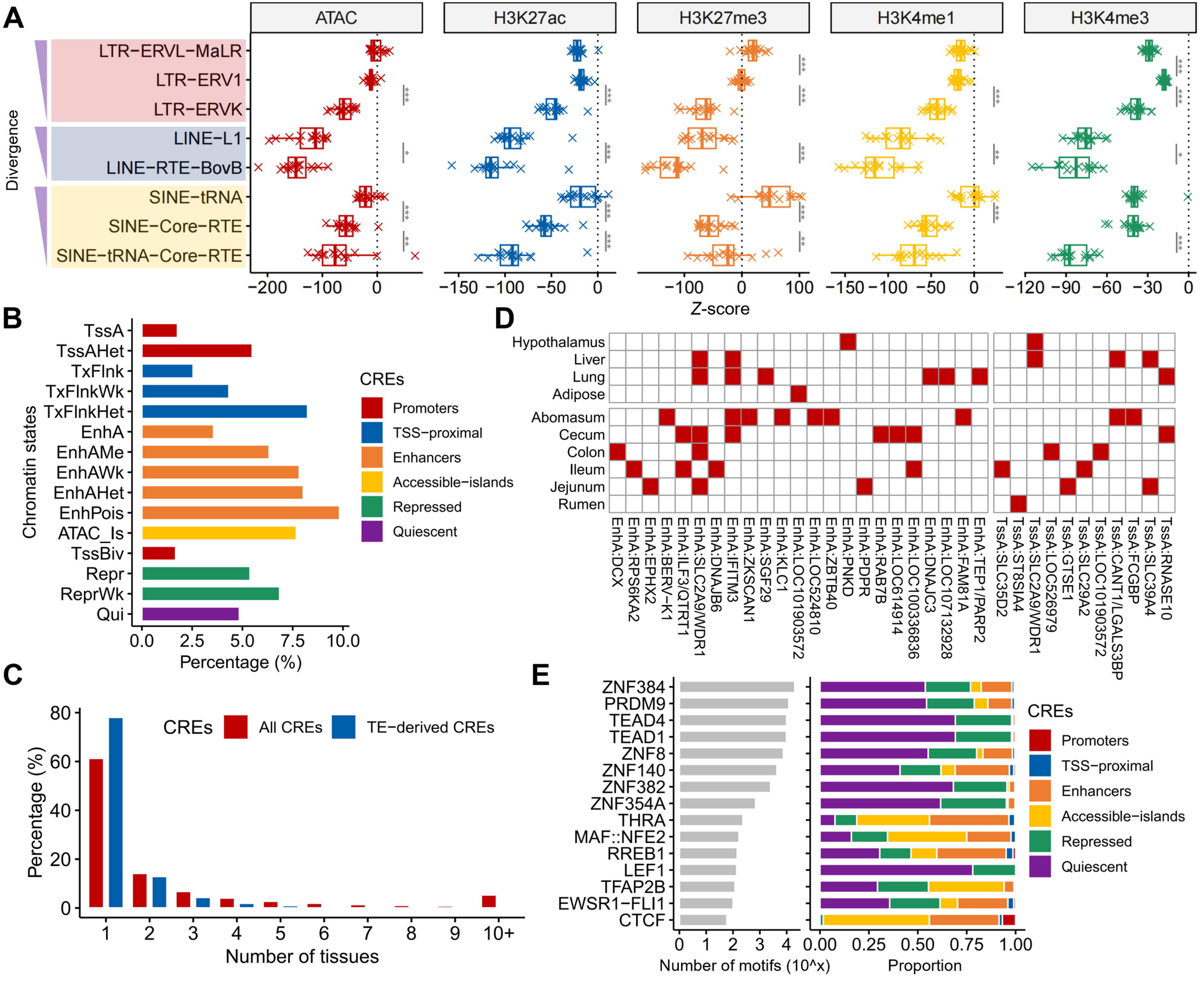

A
C
D
B
E
Ratio
Z-score
Family
Bovidae
Giraffidae
Cervidae
Moschidae
Antilocapridae
Pronghorn
Okapi
Giraffe
WhiteTailedDeer
Milu
ChineseMuntjac
BlackMuntjac
ChineseWaterDeer
ForestMuskDeer
AfricanBuffalo
Cattle
Yak
GreaterKudu
WaterBuffalo
Impala
Suni
KirksDikdik
ThomsonGazelle
GrantGazelle
Gerenuk
HarveyDuiker
CommonDuiker
DefassaWaterbuck
Gemsbok
BlueWildebeest
TibetanAntelope
Argali
Sheep
BarbarySheep
BlueSheep
Ibex
Goat
Bovinae
Capridae
TE Families
28 TE subfamilies in LTR/ERVK
All ruminants
Bovinae
Bovidae
R²
Distances


A
KZFP-binding site Identification
Sequences from k-mer clusters
Motifs: JASPAR (25) + B1H-RC model (75)
Motifs searching with p-value < 0.001
B
No evidence (29.35%)
JASPAR (17.39%)
B1H-RC model (53.26%)
C
Ruminants
C1: ZNF789
Bovidae
C5: ZNF10
Bovidae (no Bovinae)
C3: ZNF565
Bovinae
C2: ZNF182
Cervidae
C17: ZNF484
Bits
D
BTLTR1F
ERV2-1C-LTR-BT
BTLTR1C
ERV2-1B-LTR-BT
ERV2-1-LTR-BT
BTLTR1B
BTLTR1
BTLTR1B1
BTLTR1D2
BTLTR1D
BTLTR1J3
BTLTR1J4
BTLTR1J
BTLTR1E3
BTLTR1E
BTLTR1E2
BTLTR1E1
C3 C2 C1 C9 C12 C8 C13 C15 C18 C17 C5 C4
E
ZNF182
ZNF565
R = 0.56, p = 8.8e−05
−log10(p(BTLTR1) / p(BTLTR1B1)
R2(BTLTR1) − R2(BTLTR1B1)



D

Class	TE subfamilies	Genes
Bovidae-specific	BTLTR1D2	*ACAD8*
Bovidae-specific	BTLTR1E1	*BCO2*
Bovidae-specific	BTLTR1	*TTLL7*
Bovinae-specific	LTR2C_BT	*ACOD1*
Bovinae-specific	BTLTR1	*CEP57*
Bovinae-specific	ERV2-2B_BT-LTR	*EXOSC10*
Bovinae-specific	LTR3_BT	*IL18R1*
Bovinae-specific	BTLTR1	*THADA*
Bovinae-specific	BTLTR1C	*WDR73*
Capridae-specific	BTLTR1J	*ANKRD42*
Capridae-specific	LTR46_BT	*IFI44*
Capridae-specific	BTLTR1J	*METTL17*
Capridae-specific	BTLTR1B	*RGS13*
Capridae-specific	BTLTR1B	*THTPA*


A
ATAC
H3K27ac
H3K27me3
H3K4me1
H3K4me3
Divergence
LTR−ERVL−MaLR
LTR−ERV1
LTR−ERVK
LINE−L1
LINE−RTE−BovB
SINE−tRNA
SINE−Core−RTE
SINE−tRNA−Core−RTE
Z-score
B
Chromatin states
TssA
TssAHet
TxFlnk
TxFlnkWk
TxFlnkHet
EnhA
EnhAMe
EnhAWk
EnhAHet
EnhPois
ATAC_Is
TssBiv
Repr
ReprWk
Qui
Percentage (%)
CREs
Promoters
TSS−proximal
Enhancers
Accessible−islands
Repressed
Quiescent
C
CREs
All CREs
TE−derived CREs
Percentage (%)
Number of tissues
D
Hypothalamus
Liver
Lung
Adipose
Abomasum
Cecum
Colon
Ileum
Jejunum
Rumen
EnhA:DCX
EnhA:RPS6KA2
EnhA:EPHX2
EnhA:BERV−K1
EnhA:ILF3/QTRT1
EnhA:SLC2A9/WDR1
EnhA:DNAJB6
EnhA:IFITM3
EnhA:ZKSCAN1
EnhA:SGF29
EnhA:KLC1
EnhA:LOC101903572
EnhA:LOC524810
EnhA:ZBTB40
EnhA:PNKD
EnhA:PDPR
EnhA:RAB7B
EnhA:LOC614914
EnhA:LOC100336836
EnhA:DNAJC3
EnhA:LOC107132928
EnhA:FAM81A
EnhA:TEP1/PARP2
TssA:SLC35D2
TssA:ST8SIA4
TssA:SLC2A9/WDR1
TssA:LOC526979
TssA:GTSE1
TssA:SLC29A2
TssA:LOC101903572
TssA:CANT1/LGALS3BP
TssA:FCGBP
TssA:SLC39A4
TssA:RNASE10
E
ZNF384
PRDM9
TEAD4
TEAD1
ZNF8
ZNF140
ZNF382
ZNF354A
THRA
MAF::NFE2
RREB1
LEF1
TFAP2B
EWSR1−FLI1
CTCF
Number of motifs (10^x)
Proportion
CREs
Promoters
TSS−proximal
Enhancers
Accessible−islands
Repressed
Quiescent