Mobile DNA

RESEARCH

Open Access

A genome-wide study of ruminants uncovers two endogenous retrovirus families recently active in goats

Marie Verneret[1,2], Caroline Leroux[1], Thomas Faraut[3], Vincent Navratil[4], Emmanuelle Lerat[2*†] and Jocelyn Turpin[1*†]

Abstract

Background Endogenous retroviruses (ERV) are traces of ancestral retroviral germline infections that constitute a significant portion of mammalian genomes and are classified as LTR-retrotransposons. The exploration of their dynamics and evolutionary history in ruminants remains limited, highlighting the need for a comprehensive and thorough investigation of the ERV landscape in the genomes of cattle, sheep and goat.

Results Through a de novo bioinformatic analysis, we characterized 24 Class I and II ERV families across four reference assemblies of domestic and wild sheep and goats, and one assembly of cattle. Among these families, 13 are represented by consensus sequences identified in the five analyzed species, while eight are exclusive to small ruminants and three to cattle. The similarity-based approach used to search for the presence of these families in other ruminant species revealed multiple endogenization events over the last 40 million years and distinct evolutionary dynamics among species. The ERV annotation resulted in a high-resolution dataset of 100,534 ERV insertions across the five genomes, representing between 0.5 and 1% of their genomes. Solo-LTRs account for 83.2% of the annotated insertions demonstrating that most of the ERVs are relics of past events. Two Class II families showed higher abundance and copy conservation in small ruminants. One of them is closely related to circulating exogenous retroviruses and is represented by 22 copies sharing identical LTRs and 12 with complete coding capacities in the domestic goat.

Conclusions Our results suggest the presence of two ERV families with recent transpositional activity in ruminant genomes, particularly in the domestic goat, illustrating distinct evolutionary dynamics among the analyzed species. This work highlights the ongoing influence of ERVs on genomic landscapes and call for further investigation of their evolutionary trajectories in these genomes.

Keywords ERV, Ruminant, Goat, Evolution, LTR retrotransposon, Genome annotation

†Emmanuelle Lerat and Jocelyn Turpin contributed equally to this work.

*Correspondence:
Emmanuelle Lerat
emmanuelle.lerat@univ-lyon1.fr
Jocelyn Turpin
jocelyn.turpin@inrae.fr
Full list of author information is available at the end of the article

Background

Endogenous retroviruses (ERVs) are remnants of ancient retroviral germline infections that have become permanently integrated into the host genome and are transmitted vertically to subsequent generations the same way as the host genes [1–3]. These ERVs make up a large proportion of mammalian genomes, representing 8% and 10% of the human [4, 5] and mouse genomes respectively [6]. The genomic structure of ERVs consists of four retroviral genes —*gag*, *pro*, *pol*, and *env*— which encode respectively the group-specific antigens, the protease, the polymerase and the envelope proteins, flanked in 5' and 3' by long terminal repeats (LTRs) [7]. They are considered as LTR-retrotransposons and can be classified into three groups according to the transposable element's classification [8, 9]. ERV-1, ERV-K or ERV-L, or Class I, II or III according to the retroviral taxonomy [7], reflecting their evolutionary relationship with the exogenous *Gammaretrovirus*, *Betaretrovirus* genera and *Spumaretrovirinae* subfamily respectively.

Once integrated into the host genome, most ERV insertions become non-functional due to the accumulation of mutations. In some cases, recombination between the two LTRs leads to the formation of solo-LTR insertions [10–12]. However, some ERV insertions have retained their coding capacity allowing them to potentially remain active, either through transposition or by reinfecting other cells [13, 14]. Some ERVs have been co-opted by their hosts and play crucial biological roles [2]. For example, the *syncytin* gene, derived from retroviral envelope proteins, has been independently captured in several mammalian species and is essential for placentation [15, 16]. In addition, ERV-derived LTRs can serve as regulatory elements, influencing host gene expression [17] while dysregulated ERV activity can lead to genome instability, contributing to diseases such as cancer [18–20], neurodegenerative disorders [21, 22], and autoimmune diseases [23–25].

ERVs have been extensively studied in humans [26], but their exploration in livestock species remains relatively understudied. Small ruminants like sheep and goats [27, 28], along with koalas [29, 30] and a few other vertebrate hosts [31, 32], provide a unique paradigm of coexistence between endogenous retroviruses and their exogenous counterparts. In small ruminants, the exogenous Jaagsiekte Sheep Retrovirus (JSRV) and Enzootic Nasal Tumor Virus (ENTV), responsible for respiratory cancers [33–35], are closely related to an ERV family previously named endogenous Jaagsiekte Sheep Retrovirus (enJSRV) [36]. Collectively, it has been shown that these ERVs have evolved over millions of years in the sheep, contributing to reproductive physiology [28, 37] and potentially protecting their host against exogenous retroviral infection [38–40], although the extent of this protection in vivo still remains unclear [41]. Other studies have shown that enJSRV insertions can have phenotypic effects [42] and are also highly polymorphic in sheep populations [43], but far less is known about their presence in goats. The global ERV landscape of small ruminants remains poorly documented [44–46], and no comprehensive comparison between cattle, sheep, and goats has been undertaken.

In this study, we aim to explore the unique model of small ruminants alongside cattle to uncover the full ERV repertoire in these genomes and provide new insights into their past and present evolutionary history.

Methods

Mining of ERV consensus sequences

1) De novo ERV identification

ERVs were characterized in five ruminant species: cattle (*Bos taurus*), domestic sheep (*Ovis aries*), wild sheep (*Ovis orientalis*), domestic goat (*Capra hircus*), and wild goat (*Capra aegagrus*). These five species were specifically selected because of the high quality of their assemblies available at the chromosome level and serving as reference genomes in their respective species. Each assembly was retrieved from the Genbank database (accession numbers in Supplementary Material 11 - Tab. S1) and ERV consensus sequences were identified using RepeatModeler (version 2.0.3) [47] with default parameters (see the complete workflow in the Supplementary Fig. S1A).

2) Manual curation of the consensus sequences (Supplementary Fig. S1B)

a) Filtering steps

The raw consensus sequences identified by RepeatModeler, were filtered out with an approach inspired from Goubert et al. [48]. Briefly, only sequences classified as ERV were conserved. The ones ranging from 1 to 15 kb were considered as potential internal parts (INT) with the retroviral genes and those smaller than 1 kb as long terminal repeats (LTRs). According to the 80-80-80 rule [8], sequences longer than 80 bp that share more than 80% identity over 80% of their sequences belong to the same family. To reduce redundancy, INT and LTR sequences were clustered separately using CD-HIT [49]. When sequences had more than 80% of sequence identity, only one representative sequence was kept for each cluster. The retroviral genes (*gag*, *pro*/*pol*, and *env*) were annotated within each INT consensus sequence using BLASTx [50] against a subset of the Dfam database (version 3.7) [51], that we named 'dfam_retro', obtained using

the keywords "gag", "pro", "pol", "env". To minimize false positives, only INT consensus sequences with at least one hit on a retroviral gene from "dfam_retro" were conserved. The different reading frames were also identified using ORFFinder (https://www.ncbi.nlm.nih.gov/orffinder/) [52].

b) INT and LTR consensus re-association

To re-associate the INT and LTR consensus sequences, the filtered INT and LTR consensus sequences were used as a custom library in RepeatMasker (version 4.1.5) [53] to identify hits corresponding to ERV copies in the five assemblies. Copies were reconstructed when INT and LTR hits were located within 500 bp from each other. Consensus sequences representing less than 10 putative copies were removed and the association between INT and LTR consensus sequences was established when they were connected in at least 10 copies.

c) Research of missing consensus sequences within species

Internal consensus parts not associated with any LTR consensus sequence were verified by examining the corresponding annotated copies. The position of these copies was extended by 5 kb on both the 5' and 3' sides and extracted from the genome using BEDTools (version 2.30.0) [54]. A multiple sequence alignment of the extended copies was performed using MAFFT (version 7.490) [55] and visualized using Geneious Prime 2023.0.2 (https://www.geneious.com). The start and the end of the copies were determined by identifying blocks of homology and the flanking sequences were manually cropped. The presence of flanking LTRs was determined using self dot-plot. When LTR sequences were identified, LTR consensus sequences were reconstructed using all the LTR sequences from the copies.

d) Research of missing consensus sequences between species

The consensus sequences of the five analyzed species were aligned using MAFFT [55] followed by phylogenetic analysis using Maximum Likelihood (ML) statistical method in IQ-TREE (version 1.5.3) [56, 57]. If a consensus was present in all the species but one, the corresponding sequence was searched in the raw transposable element library of the missing species. If no similar sequence was found, the absence of ERV insertion was confirmed using BLASTn [50] with default parameters using the consensus of the other species as query sequences. In cases where at least 10 copies were identified with more than 80% identity covering 80% of the consensus length, these were extracted from the assembly using BEDTools [54] and aligned using MAFFT [55] to reconstruct the missing consensus sequences.

3) Classification of the consensus sequences

The characterized consensus sequences were aligned with retroviral sequences from Genbank including gammaretroviruses such as Gibbon Ape Leukemia Virus (GaLV: NC_001885.3), Koala Retrovirus (KoRV: NC_039228.1), Murine Leukemia Virus (MuLV: NC_001501.1), Feline Leukemia Virus (FeLV: AF052723.1), as well as betaretroviruses with Mouse Mammary Tumor Virus (MMTV: NC_001503.1), Enzootic Nasal Tumor Virus Type 1 and 2 (ENTV1: NC_007015.1 and ENTV2: NC_004994.2), Jaagsiekte Sheep Retrovirus (JSRV: AF105220.1). Endogenous retrovirus sequences including human endogenous retroviruses from group E, K and L (HERV-E: M10976.1, HERV-K: M14123.1, HERV-L: X89211.1), feline endogenous retrovirus (enFeLV: AY364318.1), murine endogenous retrovirus (MuERV-L: Y12713.1), and cattle and sheep ERV reference consensus from Repbase (version 29.01) [58] were also included. The references available in Repbase for *Capra aegagrus* were not included as they contain only LTR sequences and no retroviral genes. Alignment was performed using MAFFT [55] followed by phylogenetic analysis using Maximum Likelihood (ML) statistical method in IQ-TREE (version 1.5.3) [56, 57] with 10,000 ultrafast bootstrap replicates [59]. Phylogenetic trees were visualized using the online version of the Interactive Tree Of Life (iTOL, https://itol.embl.de) [60]. Finally, each ERV family was categorized as Class I or Class II ERVs based on the closest exogenous retroviruses respectively gamma or betaretroviruses. The names of the ERV families were chosen according to their classification as Class I or II, together with an arbitrary attributed number. The final library of consensus sequences for each species is available as supplementary data (Supplementary Material 12 to 16).

ERV annotation in reference assemblies

The characterized consensus sequences were used as a custom library to annotate the ovine, caprine and bovine reference assemblies using RepeatMasker (version 4.1.5) [53]. Hits that shared at least 80% of sequence identity with the corresponding consensus sequence and were longer than 80 bp were conserved [8]. Features originating from the same family and located closer than 500 bp for solo-LTRs and 7 kb for other copies were merged to reconstruct the different ERV insertions. The complete annotation of the five selected assemblies and the number of insertions per family are available as supplementary

data (Supplementary Material 17 to 21 and Supplementary Material 11 - Tab. S3). The ERV genome fraction was computed as the proportion of bases in the genome covered by all the detected ERV insertions (ERV number of bases / total genome length × 100).

ERV insertion comparative analysis

1) Sequence divergence from consensus

Each copy was aligned to its consensus sequence using MAFFT [55] and a divergence score was calculated using the Kimura-2-parameter model (K80) with the ape package in R (version 4.2.3) [61, 62]. In cases where an insertion had multiple hits on different subfamilies or LTR consensus sequences, the one with the lowest divergence score was assigned. The metadata associated with the insertions are provided as supplementary data (Supplementary Material 22 to 26).

2) LTR comparison

The 5' and 3' LTR sequences were extracted from each ERV insertion and aligned pairwise using MAFFT [55] to estimate their sequence divergence.

3) Open reading frame annotation

The retroviral open reading frames (ORFs) in the ERV insertions were annotated using orfipy [63]. Only ORFs starting with an ATG codon and having a minimum length of 300 bp were conserved for *gag* and *env*. For *pro* and *pol*, as they are produced by frameshifts, ORFs with any sense codon longer than 300 bp were retained. Each ORF was translated into a protein sequence and aligned against the "dfam_retro" database (see above) using BLASTp [50]. Insertions with intact (ie. > 80% of the expected length) *gag*, *pro/pol* and *env* ORFs were designated as full-length copies potentially capable of reinfecting other cells. Those lacking only the *env* ORF were considered as copies capable of retrotransposition (Supplementary Material 11 - Tab. S3).

4) Detection of syntenic insertions sites

A comparative analysis of the families II-3 and II-5 insertion sites in the four small ruminant reference assemblies was performed. For each ERV insertion, excluding solo-LTRs, 5 kb flanking sequences on each side were extracted. ERV insertions at the edge of a chromosome or scaffold, resulting in at least one of the flanking sequences being shorter than 100 bp were excluded from the analysis. To identify the corresponding positions of each insertion in the other species, the flanking sequences were aligned to the other assemblies using Minimap2 (version 2.26) [64]. If the 5' and 3' flanking sequences matched in the same genomic region (within a 50 kb interval), and if the interval between them corresponded to an annotated ERV of the same family, the insertion site was considered to be syntenic between the species.

ERV detection in other ruminant assemblies

The presence of each ERV family was assessed in 20 additional ruminant species (Supplementary Material 11 - Tab. S1). In addition to the references, several assemblies from *C. hircus (n = 4)* and *O. aries (n = 23)* from different breeds were analyzed (Supplementary Material 11 - Tab. S1). The search was performed using BLASTn [50] with default parameters. The hits were filtered using the same criteria as for the reference assembly annotation to respect the 80–80-80 rule [8].

Statistical analyses

The statistical tests presented in the study were carried out with R (version 4.2.3) [62] using the chisq.test and ks.test functions from the "stats" package (version 4.2.3) and the Wilcoxon test with the compare_means function from the "ggpubr" package (version 0.6.0).

Results

Small ruminant species share the same ERV families

Currently, limited information is available on endogenous retrovirus (ERV) families in small ruminants, and databases lack consensus sequences for these species. In this study, we describe a total of 24 ERV families present in four reference assemblies of domestic and wild sheep and goats, as well as one domestic cattle assembly (Fig. 1, Supplementary Material 11 - Tab. S1). Among them, 14 were classified as Class I and 10 as Class II ERVs according to their relationship with exogenous retroviruses. We used a threshold of at least one conserved protein domain and a minimum of 10 copies to consider a consensus as a valid representative of an ERV family. This may explain why no Class III sequences have been detected; they may be present but as relics. Comparing the different species, consensus sequences from 13 families appeared to be shared by the four small ruminants and the cattle species. On the other hand, while three were exclusive to cattle (families II-1, II-4, II-8), consensus sequences of eight families were only reconstructed in small ruminants (families I-6, I-10, I-14, II-3, II-5, II-6, II-7, II-9). However, none of the families appeared to be specific to either one small ruminant species or genus.

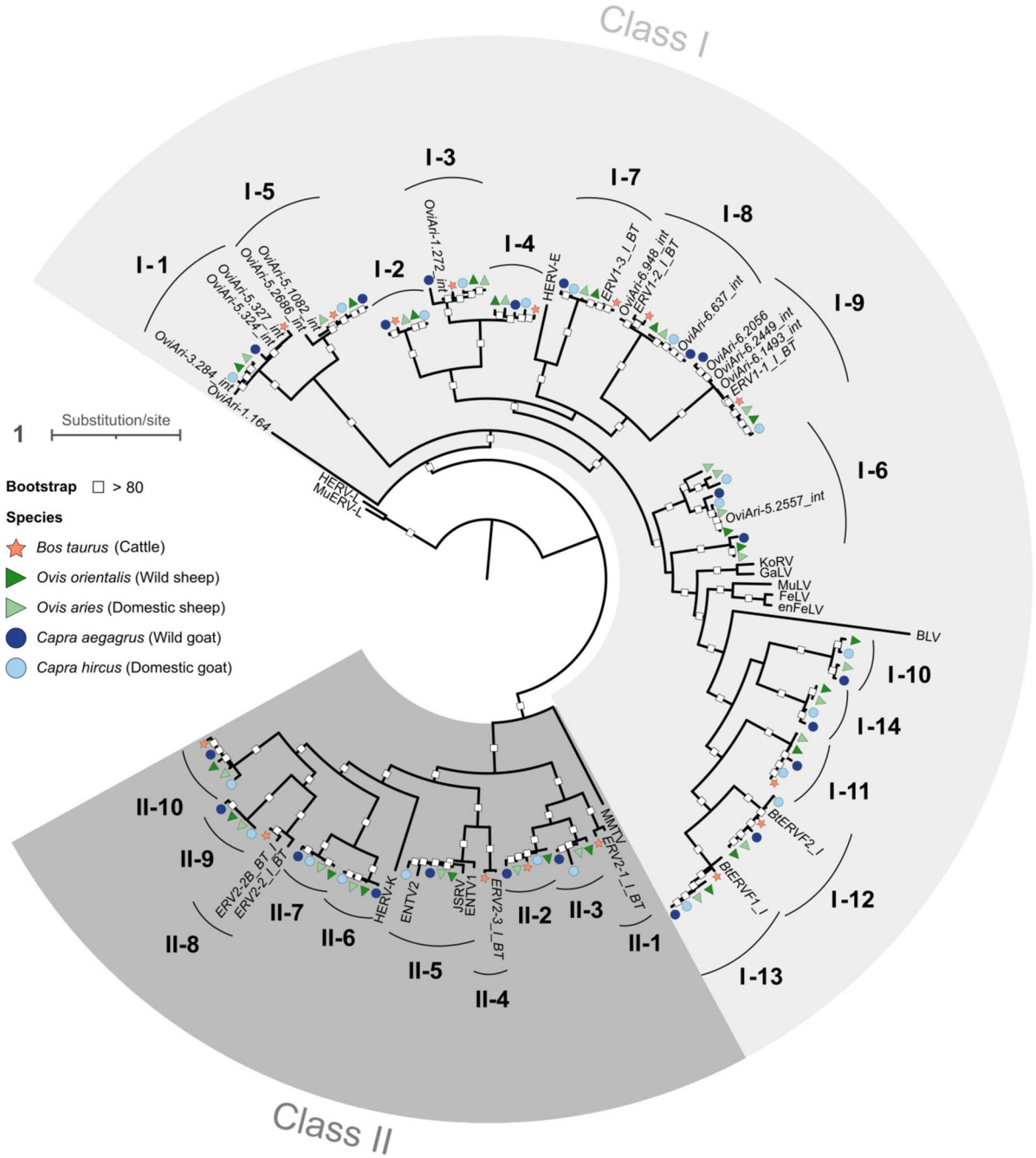

Fig. 1 ERV families in ruminant reference genomes. Maximum Likelihood phylogenetic tree reconstructed from the alignment of the consensus sequences (without LTRs) generated from the domestic and wild sheep and goat as well as cattle reference assemblies. Publicly available sequences of exogenous and endogenous retroviruses are indicated by their acronyms: GaLV, Gibbon Ape Leukemia Virus; KoRV, Koala Retrovirus; MuLV, Murine Leukemia Virus; FeLV, Feline Leukemia Virus; MMVT, Mouse Mammary Tumor Virus; ENTV1 and 2, Enzootic Nasal Tumor Virus Type 1 and 2; JSRV, Jaagsiekte Sheep Retrovirus; HERV, Human endogenous retrovirus; enFeLV, Feline endogenous retrovirus; MuERV-L, murine endogenous retrovirus. Sheep and cattle Repbase ERV references are represented in italic. The ERV family's names include their classification in Class I or II along with an arbitrary attributed number and are indicated next to each consensus sequence cluster

Although sheep and goats overall shared the same ERV families, the number of consensus sequences differed between these small ruminant species. This variation is explained by the identification of multiple consensus sequences for families I-6 (Fig. 1). On the other hand, although all the consensus sequences harbor retroviral

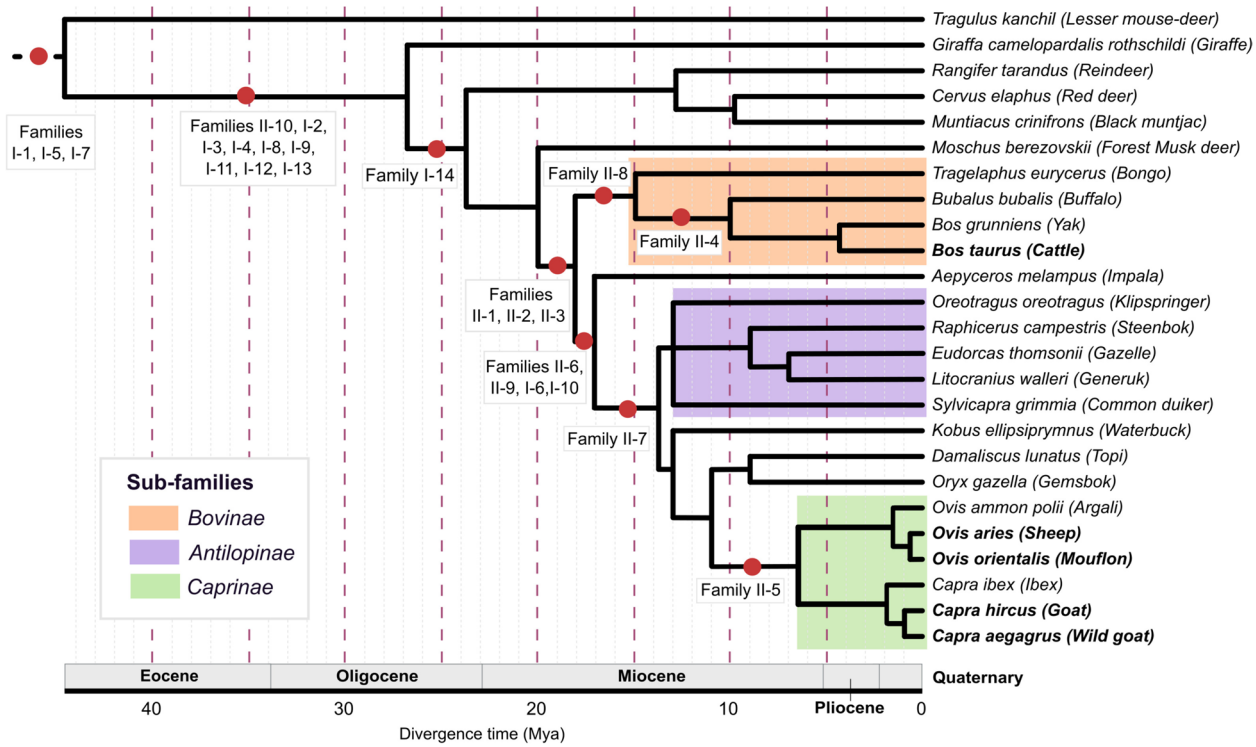

Fig. 2 ERV family integration events across ruminant evolution. Each ERV family was detected in both small ruminant and cattle reference assemblies (highlighted in bold), along with 20 additional ruminant species represented on the tree (see Supplementary Material 11 - Tab. S1 for accession numbers). Red dots, which represent the integration events, were placed before the oldest node including all the species in which each family was found present. The phylogeny was produced using TimeTree [65] coupled with divergence time from [66, 67]

genes, their completeness is different between species and ERV families (Supplementary Fig. S8 and S9). For instance, seven families (I-6, I-9, I-12, I-13, II-5, II-6, II-7) are represented by consensus sequences with intact coding sequences, whereas the families I-1 and I-5 are represented by incomplete consensus containing only parts of retroviral genes. Moreover, some internal consensus sequences are associated with multiple LTR consensus sequences, going up to five LTR consensus for family I-1 in the wild sheep, highlighting the complexity of the evolutionary history of certain families (Supplementary Material 11 - Tab. S6).

The consensus sequences generated in this study were compared with those present in Repbase (Supplementary Fig. S2 and S3). We successfully recovered all the cattle reference families from Repbase, but we also discovered eight new bovine families shared with the small ruminants (families I-1, I-2, I-3, I-4, I-5, I-11, II-2, II-10). Notably, all the Class II cattle Repbase consensus sequences corresponded to families found exclusively in cattle. For small ruminants, we also identified consensus corresponding to the Repbase references, except for OviAri_1.164 and OviAri_5.2686 and we described 15 additional small ruminant families. The consensus sequences identified in this study are characterized by longer retroviral genes compared to those in Repbase which includes an incomplete set of Class I consensus sequences for domestic sheep and only LTR sequences for wild goats. Furthermore, we observed redundancy among the Repbase consensus sequences, with multiple sequences related to the same family.

Multiple integration events of ERV families across ruminant evolution

Considering the diversity of the observed ERV families, we inferred their approximate integration time during ruminant evolution. With this objective, we looked for the presence of these ERV families in 20 other ruminant species (Fig. 2, Supplementary Material 11 - Tab. S2). Three families (I-1, I-5, I-7) were identified as the oldest ones. They were found in all the analyzed species, suggesting that their first integration occurred more than 40 million years (Myr) ago during the Eocene period. The second oldest families (I-2, I-3, I-4, I-8, I-9, I-11, I-12, I-13, II-10) likely integrated during the Oligocene period, between 27 and 44 Myr ago. The eight ERV families specific to the small ruminants emerged from multiple integration events since the divergence from the *Bovidae*

species. Six families (I-6, I-10, I-14, II-3, II-6, II-7, II-9) were found in both *Antilopinae* and *Caprinae* species with four of them (I-6, I-10, II-3, II-9) also identified in the impala, suggesting an initial integration between 17 and 18 Myr ago. In contrast, families I-14 and II-7 were not present in the impala, but family I-14 was identified in *Cervidae*, suggesting an integration time of 14–17 Myr ago for family II-7 and between 24 and 27 Myr ago for family I-14. Only family II-5 was exclusively found in *Caprinae* species, suggesting an integration between 6 and 11 Myr ago. Noteworthy, this family exhibits a close relationship with the exogenous retroviruses ENTV and JSRV (Fig. 1).

Among the three families with consensus sequences identified exclusively in cattle (II-1, II-4, II-8), two were present only in *Bovinae* species suggesting integration events from 15 to 18 Myr ago (family II-8) and 10 to 15 Myr ago (family II-4), respectively before the *Bovinae* and the *Bovini* speciation events. For the third family, family II-1, no consensus sequence was reconstructed in the small ruminant species. However, traces of highly degraded copies from this family were detected in *Antilopinae* and *Caprinae* species using presence/absence search analysis (Supplementary Material 11 - Tab. S2). Similarly, for family II-3 identified in small ruminants, no consensus sequence was reconstructed from the cattle assembly although traces of insertions were identified in *Bovinae* and *Antilopinae* species. This suggests that these two families, along with family II-2, were present in their common ancestor prior to *Bovidae* speciation between 18 and 20 Myr ago (Fig. 2). In summary, ERV families emerged from multiple integration events at different times through the ruminant evolutionary history from over 40 Myr, but no family emerged more recently than 6 Myr ago, spanning from the end of the Pliocene through the Quaternary periods. Class I families appeared to be older than Class II with only family II-10 older than 25 Myr and families I-6 and I-10 that appeared after *Bovinae* and *Caprinae* speciation.

Differential insertion dynamics of ERV between species

The annotation of ERVs allowed the estimation of their proportion in different genomes, ranging from approximately 18,000 to 23,000 insertions (Tab. 1). These insertions represent between 0.65% and 1.07% of the different analyzed genomes. Global proportion analysis revealed a predominance of Class I over Class II ERV insertions. Comparison among small ruminants showed a similar ERV profile between wild and domestic sheep with a significant over-representation of Class I insertions and an under-representation of Class II insertions compared to the wild goat. On the other hand, even though the proportion of Class II insertions in the domestic goat is lower compared to the sheep or the wild goat, they constitute a genome fraction nearly twice as high as in the other small ruminants (Table 1).

Table 1 Global ERV proportion in ruminant reference assemblies

Species	Common name	Number of insertions		Genome fraction (%)	
		Class I	Class II	Class I	Class II
Bos taurus	Cattle	14,489[b]	6,035[a]	0.52	0.52
Ovis orientalis	Wild sheep	15,165[a]	4,648[b]	0.62	0.21
Ovis aries	Domestic sheep	15,161[a]	4,405[b]	0.66	0.19
Capra aegagrus	Wild goat	13,200[b]	4,734[a]	0.48	0.17
Capra hircus	Domestic goat	18,683[a]	4,014[b]	0.73	0.34

The dependency between species and the number of ERV insertions was tested using Pearson's χ^2 test ($\chi^2 = 911.1824$, df = 4, $p < 2.2^e{-}16$)

[a] Over-represented classes of ERV

[b] Under-represented classes of ERV

To better understand the differences observed among small ruminant species, the proportion of each ERV family was compared (Fig. 3A). The analysis revealed significant variations in abundance among the different ERV families and across the species for Class I (Pearson's $\chi^2 = 6180328$, $p < 2.2^e{-}16$) and Class II insertions (Pearson's $\chi^2 = 35864259$, $p < 2.2^e{-}16$). The proportion of the different ERV families greatly varied, some being very abundant, with more than 1,000 insertions in the five analyzed assemblies (I-1, I-6, I-7, I-8, I-9, and II-3), while others contain less than 200 insertions (I-11, I-12, I-14 and II-9). Among the small ruminant species, distinct quantitative patterns of the different families were observed (Supplementary Fig. S4). For Class I, the family I-1 showed a significant over-representation in the domestic goat whereas it is under-represented in the wild goat, together with family I-6. For Class II, inter-genus differences between *Ovis* and *Capra* were observed for some families such as II-7 being over-represented in sheep genomes. Species-specific dynamics were observed with families II-3 and II-5 being over-represented in domestic goat (Supplementary Fig. S4). Furthermore, in cattle, a profile different from the small ruminants was observed, particularly in the context of Class II ERVs (Fig. 3).

Different numbers of insertions were identified between ERV families and species, highlighting the importance of deciphering whether these differences were caused by recent or ancient peaks of transposition activity. Further analyses revealed distinct divergence landscapes between Class I and Class II ERVs (Fig. 3B and Supplementary Material 11 - Tab. S4). For Class I, most of the insertions formed a peak with a sequence

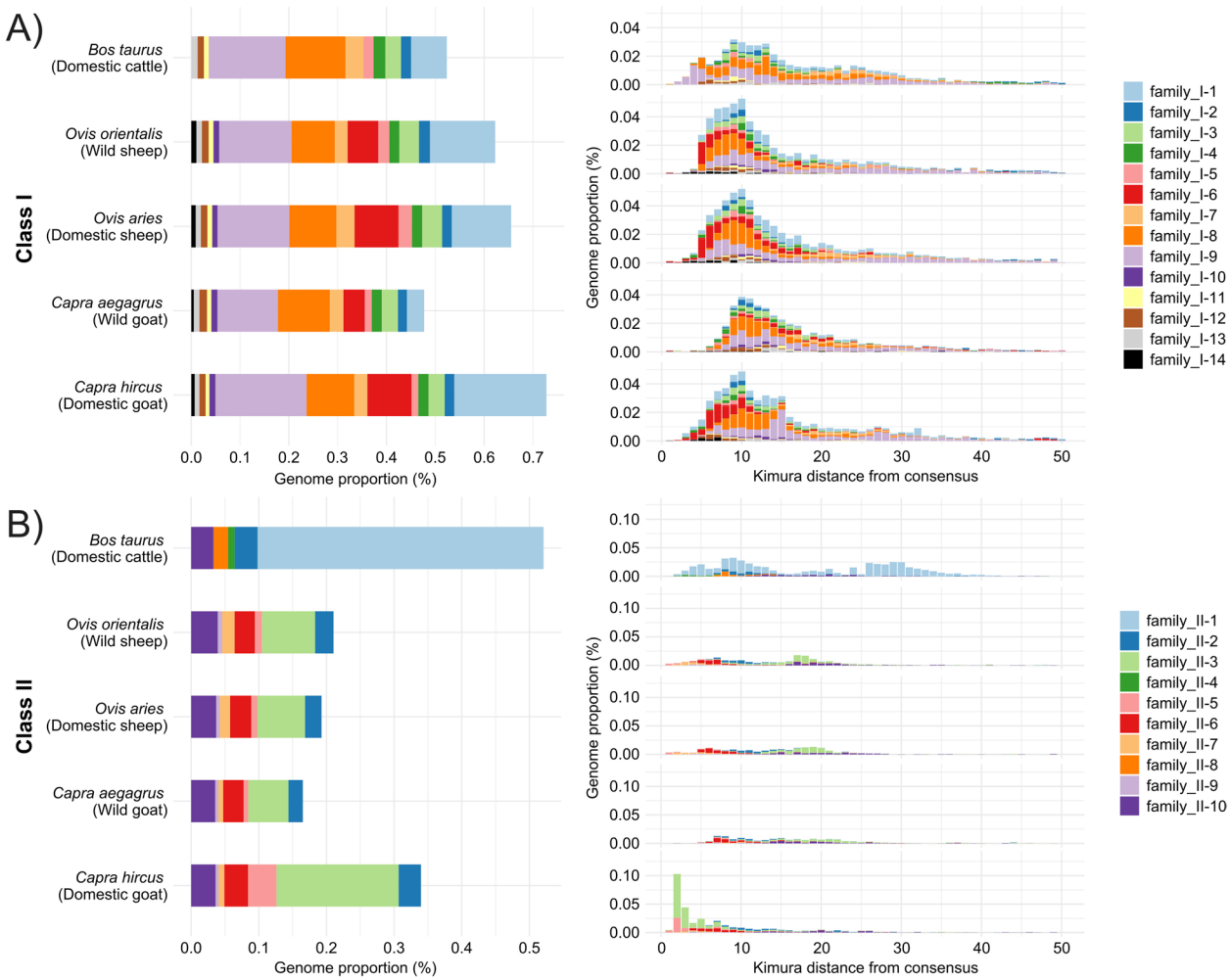

Fig. 3 Family proportion and divergence landscapes of ERV in ruminant reference assemblies. Class I and Class II are represented separately in **A**) and **B**) panels respectively. The left panels represent ERV family proportions in cattle and small ruminant assemblies. Over- and under-represented families were identified comparing *Caprinae* species (see Pearson's χ^2 residuals in Supplementary Fig S4). In the right panels, for each family and Kimura-2 sequence divergence interval, the genome coverage was computed as the percentage of the total ERV insertion length on the total genome length. The divergence distribution was compared between species using the discrete Kolmogorov–Smirnov test and the Benjamini & Hochberg correction (Statistics in Supplementary Material 11 - Tab. S4)

divergence of around 10% indicating that they are overall no longer active although relatively recent. Only a few copies had a sequence divergence from the consensus close to 0% (Supplementary Fig. S5) suggesting that they transposed very recently. Interestingly, these copies are from the families I-6 and I-10 that appeared the most recently in small ruminants and include copies with intact ORFs for family I-6 suggesting that they could be able to retrotranspose (Supplementary Material 11 - Tab. S3).

Class II ERVs exhibited diverse sequence divergence profiles, characterized by multiple peaks suggesting successive waves of ERV activity (Fig. 3B, Supplementary Fig. S6). Families II-3, and II-10 displayed peaks of sequence divergence around 20%, consistent with their potentially more ancient emergence among Class II families in small ruminants (Fig. 2). On the other hand, in cattle, the family II-1 distinctly exhibited two peaks around 30% and 10% of sequence divergence and harbored eight insertions with less than 5% of sequence divergence with intact ORFs (Supplementary Material 11 - Tab. S3), suggesting that these copies could be at the origin of a new activity wave since they are potentially still able to

retrotranspose. Consistent with their time of appearance after *Bovidae* speciation, families II-5 and II-7 in small ruminants and families II-4 and II-8 in cattle showed peaks below 10% of sequence divergence, indicating more recent activity. Family II-5 exhibited recent copies across all small ruminant species but representing a small fraction of their genomes. The domestic goat displayed a unique pattern with a significantly higher number of recent insertions (sequence divergence < 5%) for family II-5 but also for the more ancient family II-3. These results suggest a recent burst of transposition of these families exclusively in the domestic goat. Although most of the divergence distributions were significantly different across species (Supplementary Material 11 - Tab. S4) similar distributions were observed for several families including family II-9 among all small ruminants, family II-10 in domestic species and family II-5 among domestic and wild sheep.

Reactivation of an old family in domestic goat: focus on the family II-3

Copies from the family II-3 were searched in 27 other small ruminant assemblies from different breeds (Fig. 4A, Supplementary Material 11 - Tab. S1). In the other goat assemblies, the number of family II-3 ERV copies was also higher compared to domestic sheep, with a median of 185 copies in sheep and 380 in goats (Wilcoxon test, *p*-value = 0.0066). However, the results remained heterogeneous across goat assemblies and may be linked to a specific goat breed. To investigate why this family contains more copies in the domestic goat, we examined the size distribution of the insertions (Fig. 4B). The mean length of the copies was higher in the domestic goat compared to other small ruminant reference assemblies (Wilcoxon test, *p*-value < 2.2e-16) whereas it was lower in the wild goat (Wilcoxon test, *p*-value < 2.2e-16). The length difference is partly explained by the number of copies identified in the domestic goat genome. However, a significant proportion of them fall within the 2.5 to 6 kb range suggesting that the family is globally degraded in small ruminants, as reflected by the structure of the consensus sequences (Fig. 4E). Indeed, these consensus sequences are composed of partial retroviral genes and are approximately 4.5 kb in length, in contrast to the Class II ERVs with complete ORFs that range from 7 to 8 kb. We also compared the sequence divergence between the two LTRs of each insertion as a proxy to estimate its insertion date (Supplementary Fig. S7A). The 5' and 3' LTRs were better conserved in the domestic goat insertions with a median of 97.25% sequence identity compared to the wild goat (85.02%, Wilcoxon test, p-adj = 0.00027) and the wild sheep (90.02%, Wilcoxon test, p-adj = 0.014) but not to the domestic sheep (96.37%, Wilcoxon test, p-adj = 0.52).

We then evaluated if the copies were inserted at the same position in the four small ruminant genomes. Thus, insertion sites of the copies without considering solo-LTRs, were compared between species using a liftover method to convert the ERV positions in other genomes (Fig. 4C). A total of 81 insertion sites were shared between domestic and wild sheep. On the other hand, among the 79 insertions shared between the domestic and wild goat, six were also shared with one of the sheep species and 54 were present in the four species suggesting that they integrated before *Ovis* and *Capra* speciation. A low number of insertions were species-specific except for the domestic goat in which 85% of the analyzed insertions were not found in the other species. Phylogenetic analysis of the domestic goat copies showed that the goat-specific insertions were closely related with up to 100% sequence identity (Fig. 4D). Almost all of the 54 syntenic, and therefore older, insertions clustered within the same clade. In addition, the presence of multiple clades containing a large number of nearly identical sequences, going up to 378 insertions, suggests recent reactivations of family II-3 members in the domestic goat through successive waves of transposition.

The most recent ERV family could be still active in small ruminants: focus on the family II-5

Domestic goat reference genome also appeared to be representative of the *C. hircus* species for the family II-5, as evidenced by a significantly higher number of family II-5 ERV copies in several goat breeds compared to sheep (Fig. 5A, Wilcoxon test, *p*-value = 0.00058). The median number of ERV copies, excluding solo-LTRs, in goats was 81, whereas it was 52 in sheep assemblies, although the number of insertions varied more in goats than in sheep, ranging from 79 to 189 in goats to 24 to 63 in sheep. To explore why this family contains more copies in goats than in sheep, we used the same methodology as for family II-3 and examined the size distribution of the insertions (Fig. 5B). While the mean length of the copies was not significantly higher in domestic goats than in other small ruminants, wild goats had significantly shorter copies than the domestic goat (Wilcoxon test, p-adj < 2.2e-16), and both domestic and wild sheep (Wilcoxon tests, p-adj = 1.2e-14 and p-adj = 1.5e-12 respectively) mainly caused by the absence of full-length copies. In contrast, wild and domestic sheep and the domestic goat have 20, 30 and 78 copies respectively with lengths greater than 7,5 kp. The median distance between the two LTRs of each insertion was similar among the domestic goat

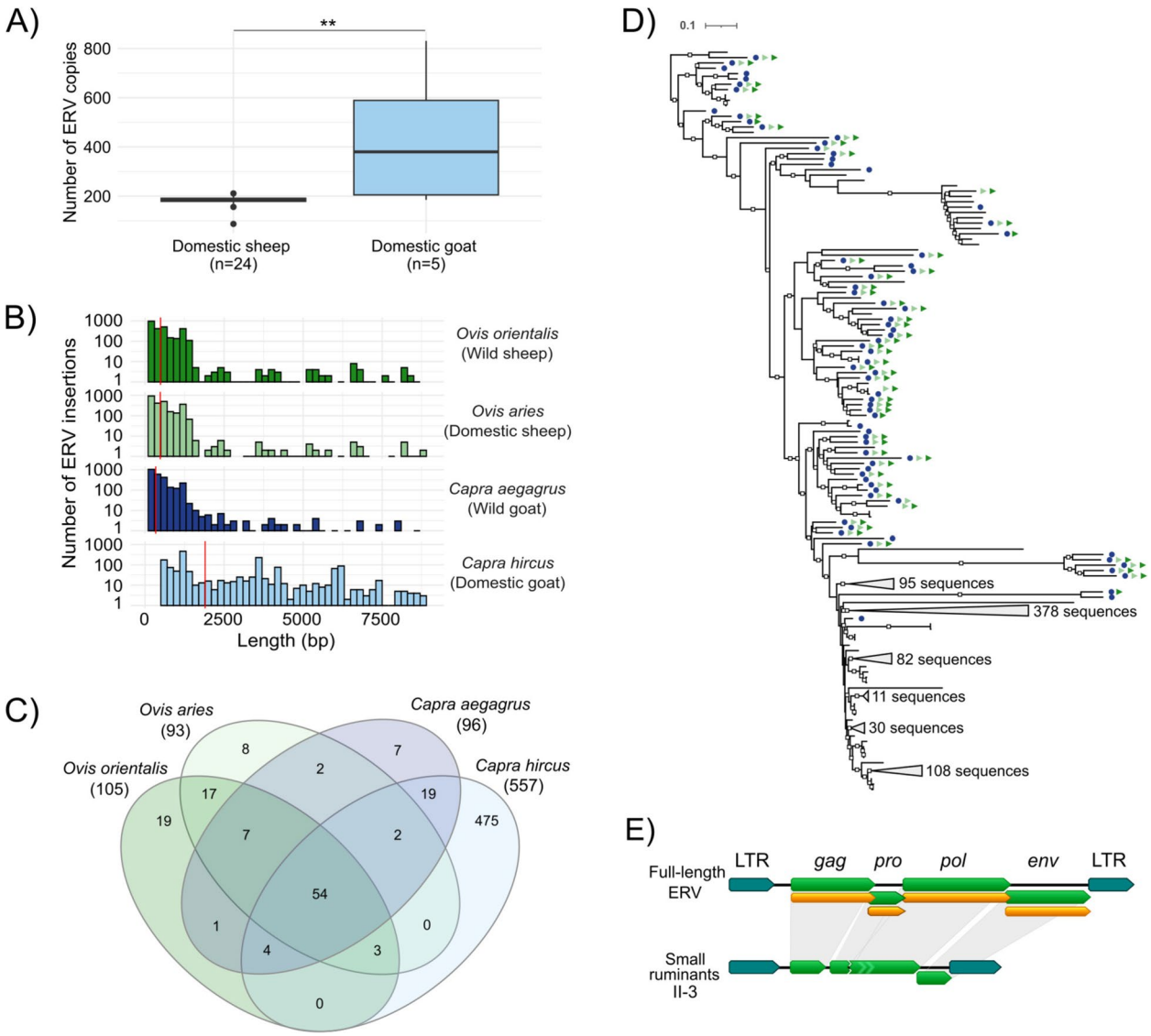

Fig. 4 Characteristics of the family II-3 copies in small ruminant genomes. **A** Number of family II-3 ERV copies excluding solo-LTRs in 29 assemblies from domestic sheep and goat of different breeds (accession numbers in Supplementary Material 11 - Tab. S1). **B** Length distribution of the ERV insertions for each of the small ruminant reference assembly. The red line indicates the mean length. **C** Number of common ERV *loci* excluding solo-LTRs between the species. Only insertion sites flanked by at least 100 bp of sequence on both sides were retained. **D** Phylogenetic tree of the family II-3 copies in *C. hircus* reference genome excluding solo-LTRs. Copies sharing insertion sites with other species are reported with symbols: dark blue circle for wild goat, dark green triangle for wild sheep and soft green circle for domestic sheep. To clarify the tree, nodes including only domestic goat specific insertions without any synteny and nearly identical sequences were collapsed. The white squares indicate branches supported by a bootstrap higher than 80%. Branch lengths are expressed as the number of substitutions per site. **E** Comparison of the expected complete consensus sequence with the one from family II-3 in small ruminants. The green boxes represent the retroviral genes and the orange ones the coding sequence (ORF)

and the domestic and wild sheep with a sequence divergence lower than 1% (Supplementary Fig. S7B). However, in the wild goat, the LTRs were significantly more divergent than in the domestic goat (99.30%, Wilcoxon test, p-adj = 0.000088) and both the domestic and wild sheep (99.29% and 99.10% respectively, Wilcoxon test, p-adj = 0.00075 and p-adj = 0.00059). This suggests that this family is well conserved in the small ruminants as confirmed by the structure of the consensus sequences that contain intact ORFs (Fig. 5E). Surprisingly, a very well conserved consensus was obtained for the wild goat whereas only incomplete insertions were annotated. In comparison, the domestic goat and wild sheep harbored

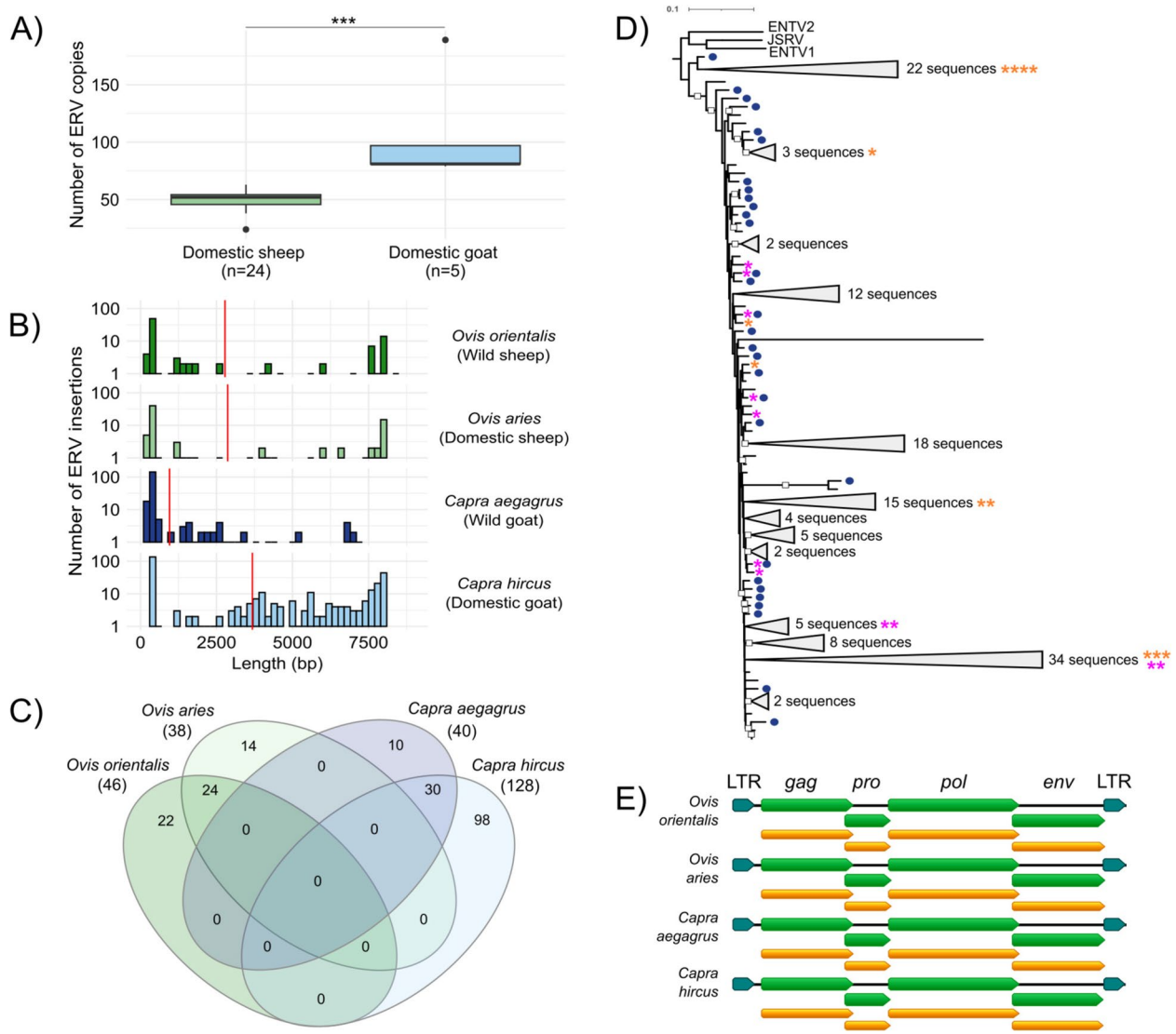

Fig. 5 Characteristics of the family II-5 copies in small ruminant genomes. **A** Number of family II-5 ERV copies excluding solo-LTRs in 29 assemblies from domestic sheep and goat of different breeds (accession numbers in Supplementary Material 11 - Tab. S1). **B** Length distribution of the ERV insertions for each of the small ruminant reference assembly. The red line indicates the mean length. **C** Number of common ERV *loci* excluding solo-LTRs between the species. Only insertions sites flanked by at least 100 bp of sequence on both sides were retained. **D** Phylogenetic tree of the family II-5 copies in *C. hircus* reference genome excluding solo-LTRs. Copies sharing insertion sites with wild goat are represented by dark blue circle. The orange stars represent the insertions with complete *gag*, *pro*, *pol* and *env* ORFs and the pink stars, the copies missing only the complete *env* one. To clarify the tree, nodes including only domestic goat specific insertions without any synteny and nearly identical sequences were collapsed. The white squares indicate the branch supported by a bootstrap of 100%. Branch lengths are expressed as the number of substitutions per site. **E** Comparison of the family II-5 consensus sequences between small ruminants. The green boxes represent the retroviral genes and the orange ones the coding sequence (ORF)

respectively 12 and seven copies with complete coding sequences while only partially conserved copies were annotated in the domestic sheep assembly.

We described insertions with low sequence divergence compared to their corresponding consensus sequences, with conserved ORFs and with almost identical LTRs indicating that they are very recent copies or that they have been selectively conserved in the small ruminant genomes. To better estimate their integration dates, the different insertion sites of the copies, excluding solo-LTRs, were compared between the species (Fig. 5C). Respectively 24 and 30 insertions have been found in the same genomic regions between domestic and wild sheep,

or between domestic and wild goats. Regarding the species-specific insertions, 98 insertions have been found only in domestic goat representing 76% of the analyzed insertion sites in goat (excluding solo-LTRs). The analysis did not reveal any common insertions between sheep and goats, suggesting that the present copies integrated after the speciation of *Ovis* and *Capra* between 1 and 6 Myr ago (Fig. 2).

Phylogenetic analysis of the domestic goat copies revealed no discernible clade and indicated close sequence identity between the copies (Fig. 5D). Of the identified sequences, 12 contained complete coding sequences, while 11 lacked the *env* ORF. None of the complete coding sequences corresponded to syntenic copies in the wild goats, while three sequences missing only the *env* ORF are in the same genomic region as wild goat copies. Remarkably, the syntenic copies in wild goats showed poor sequence conservation and are all truncated copies highlighting different evolutionary mechanisms involved between wild and domestic goats.

Discussion

Using a combination of bioinformatic tools, we characterized 24 ERV families across five ruminant reference genomes, including cattle and both domestic and wild sheep and goat species. For each ERV family, consensus sequences were generated after applying stringent filtering steps, allowing us to establish reliable representatives of these ERV families. Our approach follows the manual curation methodology outlined by Goubert et al. [48]. Other methods also have been published [68], and several automated pipelines have been recently developed [69, 70], highlighting the importance of this step and the need to adapt it to specific research questions.

Through a comparison with Repbase reference families [58], our results have allowed the refinement of the existing sequences for these species but also the introduction of additional reference sequences, especially for species like wild and domestic goats, and wild sheep, in which only LTRs or no consensus sequences were previously available. Our analysis revealed the presence of both Class I and Class II ERV families in ruminant genomes, in agreement with previous reports in cattle [44, 71, 72]. Although previous studies in sheep mainly focused on Class II ERV, in particular family II-5 named as enJSRV [28, 37, 40, 41, 73–76], nine partial sequences of the *pol* gene from Class I copies have been described [77] corresponding to the families I-4, I-8, I-6 and I-10 in our study. A recent study on *Caprinae* species described 28 ERV families [46]. Three families from their analysis (CapERV-1, CapERV-11, CapERV-26) are not present in our study likely due to the different methodology used for de novo ERV identification. However, our approach led to the characterization of seven other families including the oldest families (I-1, I-5, I-7) and four additional ones (I-2, I-12, I-13, II-3), expanding the known ERV repertoire. Interestingly, two families that we identified with multiple LTR consensus sequences (I-6 and I-10) corresponded to multiple sequences in their analysis confirming the complex evolutionary history of these families. Using multiple sequence alignment of the I-6 copies, we identified distinct groups of copies with different LTR sequences in the domestic sheep but it appears that a single internal region coupled with multiple LTR consensus sequences is sufficient to correctly annotate these copies (Supplementary Fig. S10).

An important contribution of our work is the ERV comparative analysis between cattle, sheep and goat species. While small ruminants share the same ERV families, notable differences are observed when compared to cattle. We showed that ruminant ERV families emerged from multiple integration events across evolution, resulting in ERV families common to both *Bovinae* and *Caprinae*, as well as families specific to each of these *Bovidae* sub-families. Family II-5 exclusively detected in small ruminants, in concordance with some studies but contrasting with other previous works reporting its presence in some cattle breeds and in the river buffalo [72, 78, 79]. Our findings suggest divergent evolutionary trajectories of ERVs driven by both family and species-specific factors.

Dating analysis revealed that ERV Class I families tend to be older and more degraded than Class II families. By analyzing the number of ERV copies, we estimated that these elements represents approximately 1% of each ruminant genome, consistent with previous global studies that estimated the total ERV genome fraction, including Class III insertions, at approximately 3% [80–82]. Our stringent filtering criteria may have led to an underestimation of their genome fraction by missing the most ancient and degraded copies, explaining why we did not encounter any Class III families in our analysis. The high proportion of solo-LTRs (82.3%) suggests that most ERV transpositional activity occurred in the distant past, with a few recent insertions observed, except for families II-3 and II-5 which have been recently and might still potentially be active, especially in the domestic goat.

Family II-3 showed a higher number of copies in the domestic goat compared to other small ruminants. While many insertion sites are shared across the four small ruminant species, suggesting integration prior speciation, numerous goat-specific copies were identified. Some of these copies, located on scaffolds, may originate from the X chromosome which remain unassembled in the goat reference genome [83]. Comparison of the

flanking sequences of each goat scaffold-located insertion revealed different insertion sites for some, confirming their existence, but also identical ones for others, which could be the result of satellites from ectopic recombination of LTR-retrotransposon [84–86]. Phylogenetic analysis revealed different groups among the family II-3 ERV copies in the domestic goat. One group contained almost all syntenic copies together while the others mainly contained the goat-specific insertions, suggesting recent bursts of transposition, possibly mediated by trans-regulatory mechanisms [87–91]. Nonetheless, we did not detect any copy with intact ORFs, raising questions about its transpositional mechanism, which might rely on trans-complementation by other elements, as already described for the family II-1 in cattle known as ERVK[2–1-LTR] [92], for which most of the de novo insertions originated from non-autonomous elements.

Considering the family II-5, no syntenic insertions were found between *Ovis* and *Capra*, suggesting that all insertions probably occurred after speciation, contrasting with other studies that showed at least two insertions (enJSRV-6 and enJSRV-10) shared between sheep and goats suggesting an integration before speciation [27, 28, 43]. The identification of the insertion sites of these copies in the available small ruminant genomes using their flanking sequences allowed us to confirm that these two copies were absent from both domestic and wild goat genomes but present in most of sheep genomes (Supplementary Material 11 - Tab. S5). These results suggest multiple events of transposition activity and the possible loss of these older insertions in the current goat populations.

Family II-5 also stands out as the most recent families, with highly conserved copies across all four small ruminant species. Some of these copies contain intact ORFs, raising the possibility of active autonomous retrotransposition and virus-like particles production initiating new insertions [27, 93–96]. Family II-5 is less represented in wild sheep, but contrary to the domestic sheep intact ORFs are more prevalent in this species, suggesting differences in regulatory systems between domestic and wild ruminants. Certain copies from this family have previously been identified as potential invaders of small ruminant genomes [74, 97], suggesting that this family has not been completely silenced by transposable element regulatory processes. Other studies have highlighted the importance of family II-5 in placental morphogenesis in sheep [98, 99] and its expression in similar tissues in goats [100]. However, the lack of syntenic insertions between the two genera raises the question of whether these insertions actually play the same role in the two species, and if the family has been co-opted and is the result of evolutionary convergence. Further investigation is needed to determine the extent of functional co-option of this family across species and its evolutionary implications including the interplay of the ERVs with their exogenous counterparts JSRV and ENTV [28, 43, 74].

Conclusions

In this study, we have generated a robust library of ERV consensus sequences and high-resolution ERV annotations in ruminant species. Our results provide new insights into the evolutionary history of ERVs, tracing their activity over the last 40 million years and their ongoing role in shaping ruminant genomic landscapes. Two of the identified ERV families appear to have been recently active, particularly in the domestic goat, and may still have transpositional potential, although further investigation is needed to confirm this activity. These results open up new research opportunities to explore the complex interplay between ERVs and their host genome focusing on ERV transcriptional activity, regulatory mechanisms, and functional implications. Additionally, our work provides important resources for population genomics studies to investigate ERV insertion polymorphisms, which could help to further unravel the selective forces acting on these genomic elements across ruminant species.

Abbreviations

Bp	Base pair
enJSRV	Endogenous Jaagsiekte Sheep Retrovirus
ENTV	Enzootic Nasal Tumor Virus
ERV	Endogenous Retrovirus
INT	ERV internal part including the retroviral genes
JSRV	Jaagsiekte Sheep Retrovirus
LTR	Long Terminal Repeat
Myr	Million years
ORF	Open Reading Frame
p-adj	Adjusted p-value

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13100-024-00337-6.

Supplementary Material 1: Fig. S1: Pipeline used to characterize ERVs in ruminant reference genomes. (A) Simplified representation of the pipeline composed of two mains steps including the ERV mining and then the annotation of the insertions in the reference genomes. (B) Details of the manual curation steps to obtain the final set of consensus sequences. The number of consensus sequences along the pipeline is given in the upper table

Supplementary Material 2: Fig. S2: Comparison between *Ovis aries* ERV Repbase consensus sequences with the ones from this study. The sequences excluding the LTR parts were aligned using MAFFT and visualized with Geneious Prime® (version 2023.0.2). Each panel corresponds to a different ERV family. The consensus sequences characterized in this study are first in raw in bold. The gray boxes indicate similar sequences, the black boxes show sequence differences and the horizontal black lines long deletions. The green arrows represent the retroviral genes

Supplementary Material 3: Fig. S3: Comparison between *Bos taurus* ERV Repbase consensus sequences with the ones from this study. The

sequences excluding the LTR parts were aligned using MAFFT and visualized with Geneious Prime® (version 2023.0.2). Each panel corresponds to a different ERV family. The consensus sequences characterized in this study are first in raw in bold. The gray boxes indicate similar sequences, the black boxes show sequence differences and the horizontal black lines long deletions. The green arrows represent the retroviral genes

Supplementary Material 4: Fig. S4: ERV representation between families and small ruminant species. The proportion of each ERV family was compared with a Pearson's χ^2 test (Class I: $\chi^2 = 6180328$, df=36, $p < 2.2^e$-16 and Class II: $\chi^2 = 35864259$, df=18, $p < 2.2^e$-16). The χ^2 residuals are shown here. Families over-represented are in red and the one down-represented in blue. No values were computed for families II-1, II-4 and II-8 only present in cattle but not in small ruminants.

Supplementary Material 5: Fig. S5: ERV family Class I divergence landscapes in reference ruminant assemblies

Supplementary Material 6: Fig. S6: ERV family Class II divergence landscapes in reference ruminant assemblies.

Supplementary Material 7: Fig. S7: Sequence divergence between ERV family II-3 and II-5 LTRs in small ruminant reference assemblies. The 5' and 3' LTRs of each annotated ERV insertion were aligned separately using MAAFT to obtain a pairwise identity score for each LTR pair. Family II-3 is represented in A) and family II-5 in B). For each species, the red line indicates the mean % of identity

Supplementary Material 8: Fig. S8: Class I consensus sequences' genic structure comparison between species. The sequences excluding the LTR parts were aligned using MAFFT and visualized with Geneious Prime® (version 2023.0.2). The gray boxes indicate similar sequences, the black boxes show sequence differences and the horizontal black lines long deletions. The mean pairwise identity over all the sequences for each position is shown above the alignment. Green bars indicate 100% identity, green-brown between 30 and 100% identity and red below 30% identity. The green arrows represent the retroviral genes. Only the internal parts are represented (without the LTR). BT: B. taurus; OO: O. orientalis; OA: O. aries; CA: C. aegagrus; CH: C. hircus

Supplementary Material 9: Fig. S9: Class II consensus sequences' genic structure comparison between species. The sequences excluding the LTR parts were aligned using MAFFT and visualized with Geneious Prime® (version 2023.0.2). The gray boxes indicate similar sequences, the black boxes show sequence differences and the horizontal black lines long deletions. The mean pairwise identity over all the sequences for each position is shown above the alignment. Green bars indicate 100% identity, green-brown between 30 and 100% identity and red below 30% identity. The green arrows represent the retroviral genes. Only the internal parts are represented (without the LTR). BT: B. taurus; OO: O. orientalis; OA: O. aries; CA: C. aegagrus; CH: C. hircus

Supplementary Material 10: Fig. S10: The sub-families of family I-6 in small ruminant genomes. (A) ERV family I-6 consensus sequences phylogeny. Maximum Likelihood phylogenetic tree reconstructed for the family I-6, I-3 and I-4 consensus sequences (without LTRs) generated for the domestic and wild sheep and the goat as well as the cattle reference assemblies. The tree was rooted with Bovine Leukemia virus (BLV). Branch lengths are expressed as the number of substitutions per site. (B) Multiple LTR sequences associated to the family I-6b in the domestic sheep genome. A single internal consensus sequence was identified for the family I-6b in the domestic sheep genome, but it was associated to three distinct LTR consensus sequences. All the ERV insertions from family I-6b were aligned to the internal consensus sequence. The 5' LTR sequences of these ERV insertions were aligned with the different LTR consensus sequences. The first panel displays the average percentage of identity per sub-families at each position of the internal consensus sequence. The second panel presents the alignment of the LTRs, with color-coded boxes representing the sequence differences

Supplementary Material 11: Table S1: Ruminant genomes used for ERV detection. The five reference assemblies used in this study are indicated in bold. NA: Non-Applicable. Table S2: Presence/absence of each ERV families in ruminant species. Table S3: Number of ERV identified in five reference ruminant genomes. NA: Non-Applicable. Table S4: Kolmogrorov-Smirnov test results on the ERV divergence distribution. ERV families with non-significantly different distributions between species (*p*-adjusted > 0.05) are shown in bold. NA: Non-Applicable. Table S5: enJSRV insertion sites in small ruminant assemblies. Table S6: Retroviral gene positions and number of ERV LTR consensus sequences associated to the INT consensus sequences. NA: Non-Applicable.

Supplementary Material 12. ERV consensus sequences generated from *Bos taurus* ARS-UCD1.3 assembly

Supplementary Material 13. ERV consensus sequences generated from *Ovis orientalis* CAU_Oori_1.0 assembly

Supplementary Material 14. ERV consensus sequences generated from *Ovis aries* ARS-UI_Ramb_v2.0 assembly

Supplementary Material 15. ERV consensus sequences generated from *Capra aegagrus* CapAeg_1.0 assembly

Supplementary Material 16. ERV consensus sequences generated from *Capra hircus* ARS1.2 assembly

Supplementary Material 17. ERV insertion localization in *Bos taurus* ARS-UCD1.3 assembly

Supplementary Material 18. ERV insertion localization in *Ovis orientalis* CAU_Oori_1.0 assembly

Supplementary Material 19. ERV insertion localization in *Ovis aries* ARS-UI_Ramb_v2.0 assembly

Supplementary Material 20. ERV insertion localization in *Capra aegagrus* CapAeg_1.0 assembly

Supplementary Material 21. ERV insertion localization in *Capra hircus* ARS1.2 assembly

Supplementary Material 22. ERV insertion characteristics in *Bos taurus* ARS-UCD1.3 assembly

Supplementary Material 23. ERV insertion characteristics in *Ovis orientalis* CAU_Oori_1.0 assembly

Supplementary Material 24. ERV insertion characteristics in *Ovis aries* ARS-UI_Ramb_v2.0 assembly

Supplementary Material 25. ERV insertion characteristics in *Capra aegagrus* CapAeg_1.0 assembly

Supplementary Material 26. ERV insertion characteristics in *Capra hircus* ARS1.2 assembly

Acknowledgements
This work was performed using the computing facilities of the CC LBBE/PRABI. We thank F. Arnaud for sharing the unpublished flanking genome sequences of the enJSRV copies reported in Arnaud et al, PLoS PAthog 2007.

Authors' contributions
M.V. performed the analyses and wrote the manuscript. J.T., E.L. and C.L. conceived the project. J.T. and E.L. supervised the work. M.V., J.T., E.L., C.L., T.F., V.N. contributed to the design and implementation of the research, participated to the interpretation of the data and revised the manuscript. Correspondence to JT [jocelyn.turpin@univ-lyon1.fr] and EL [emmanuelle.lerat@univ-lyon1.fr]. JT and EL contributed equally to this work.

Funding
This research was funded by the Agence Nationale de la Recherche (ANR), grant ANR-22-CE35-0002–01 and by INRAE GA-SA joint program GoatRetrovirome.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]IVPC UMR754, INRAE, Universite Claude Bernard Lyon 1, EPHE, PSL Research University, 69007 Lyon, France. [2]Universite Claude Bernard Lyon 1, LBBE, UMR 5558, CNRS, VAS, 69622 Villeurbanne, France. [3]GenPhySE, Universite de Toulouse, INRAE, INPT, ENVT, 31326 Castanet Tolosan, France. [4]PRABI, Pôle Rhône-Alpes Bioinformatics Center, Universite Claude Bernard Lyon 1, 69622 Villeurbanne, France.

Received: 27 June 2024 Accepted: 21 November 2024
Published online: 17 February 2025

References

1. Weiss RA. The discovery of endogenous retroviruses. Retrovirology. 2006;3(1):67.
2. Stoye JP. Studies of endogenous retroviruses reveal a continuing evolutionary saga. Nat Rev Microbiol. 2012;10(6):395–406.
3. Feschotte C, Gilbert C. Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet. 2012;13(4):283–96.
4. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860–921.
5. Griffiths DJ. Endogenous retroviruses in the human genome sequence. Genome Biol. 2001;2(6):reviews1017.1-reviews1017.5.
6. Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002;420(6915):520–62.
7. Gifford RJ, Blomberg J, Coffin JM, Fan H, Heidmann T, Mayer J, et al. Nomenclature for endogenous retrovirus (ERV) loci. Retrovirology. 2018;28(15):59.
8. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8(12):973–82.
9. Kapitonov VV, Jurka J. A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet. 2008;9(5):411–2.
10. Mager DL, Stoye JP. Mammalian Endogenous Retroviruses. Microbiol Spectr. 2015 Feb 5;3(1). https://doi.org/10.1128/microbiolspec.MDNA3-0009-2014.
11. Zheng J, Wei Y, Han GZ. The diversity and evolution of retroviruses: perspectives from viral "fossils." Virol Sin. 2022;37(1):11–8.
12. Sverdlov ED. Perpetually mobile footprints of ancient infections in human genome. FEBS Lett. 1998;428(1–2):1–6.
13. Miyazawa T, Yoshikawa R, Golder M, Okada M, Stewart H, Palmarini M. Isolation of an infectious endogenous retrovirus in a proportion of live attenuated vaccines for pets. J Virol. 2010;84(7):3690–4.
14. Patience C, Takeuchi Y, Weiss RA. Infection of human cells by an endogenous retrovirus of pigs. Nat Med. 1997;3(3):282–6.
15. Dupressoir A, Lavialle C, Heidmann T. From ancestral infectious retroviruses to bona fide cellular genes: Role of the captured *syncytins* in placentation. Placenta. 2012;33(9):663–71.
16. Lavialle C, Cornelis G, Dupressoir A, Esnault C, Heidmann O, Vernochet C, et al. Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation. Philos Trans R Soc B Biol Sci. 2013;368(1626):20120507.
17. Feschotte C. The contribution of transposable elements to the evolution of regulatory networks. Nat Rev Genet. 2008;9(5):397–405.
18. Záveský L, Jandáková E, Weinberger V, Minář L, Kohoutová M, Slanař O. Human endogenous retroviruses in breast cancer: altered expression pattern implicates divergent roles in carcinogenesis. Oncology. 2024;26:858–67.
19. Kumar V, McClelland M, Nguyen J, De Robles G, Ittmann M, Castro P, et al. Expression of endogenous retroviral RNA in prostate tumors has prognostic value and shows differences among Americans of African Versus European/Middle Eastern Ancestry. Cancers. 2021;13(24):6347.
20. Alldredge J, Kumar V, Nguyen J, Sanders BE, Gomez K, Jayachandran K, et al. Endogenous Retrovirus RNA expression differences between race, stage and hpv status offer improved prognostication among women with cervical cancer. Int J Mol Sci. 2023;24(2):1492.
21. Licastro F, Porcellini E. Activation of endogenous retrovirus, brain infections and environmental insults in neurodegeneration and Alzheimer's Disease. Int J Mol Sci. 2021;22(14):7263.
22. Römer C. Viruses and endogenous retroviruses as roots for neuroinflammation and neurodegenerative diseases. Front Neurosci. 2021;12(15):648629.
23. Dhillon P, Mulholland KA, Hu H, Park J, Sheng X, Abedini A, et al. Increased levels of endogenous retroviruses trigger fibroinflammation and play a role in kidney disease development. Nat Commun. 2023;14(1):559.
24. Rangel SC, da Silva MD, da Silva AL, dos Santos J de MB, Neves LM, Pedrosa A, et al. Human endogenous retroviruses and the inflammatory response: A vicious circle associated with health and illness. Front Immunol. 2022;13:1057791.
25. Greenig M. HERVs, immunity, and autoimmunity: understanding the connection. PeerJ. 2019;5(7):e6711.
26. Vargiu L, Rodriguez-Tomé P, Sperber GO, Cadeddu M, Grandi N, Blikstad V, et al. Classification and characterization of human endogenous retroviruses; mosaic forms are common. Retrovirology. 2016;22(13):7.
27. Arnaud F, Caporale M, Varela M, Biek R, Chessa B, Alberti A, et al. A paradigm for virus-host coevolution: sequential counter-adaptations between endogenous and exogenous retroviruses. PLoS Pathog. 2007;3(11):e170.
28. Armezzani A, Varela M, Spencer TE, Palmarini M, Arnaud F. "Ménage à Trois": The Evolutionary Interplay between JSRV, enJSRVs and Domestic Sheep. Viruses. 2014;6(12):4926–45.
29. Tarlinton RE, Meers J, Young PR. Retroviral invasion of the koala genome. Nature. 2006;442(7098):79–81.
30. Lillie M, Hill J, Pettersson ME, Jern P. Expansion of a retrovirus lineage in the koala genome. Proc Natl Acad Sci. 2022;119(25):e2201844119.
31. Chiu ES, VandeWoude S. Endogenous retroviruses drive resistance and promotion of exogenous retroviral homologs. Annu Rev Anim Biosci. 2021;9:225–48.
32. Mottaghinia S, Stenzel S, Tsangaras K, Nikolaidis N, Laue M, Müller K, et al. A recent gibbon ape leukemia virus germline integration in a rodent from New Guinea. Proc Natl Acad Sci. 2024;121(6):e2220392121.
33. Leroux C, Girard N, Cottin V, Greenland T, Mornex JF, Archer F. Jaagsiekte Sheep Retrovirus (JSRV): from virus to lung cancer in sheep. Vet Res. 2007;38(2):211–28.
34. Leroux C, Mornex JF. Retroviral infections in sheep and the associated diseases. Small Rumin Res. 2008;76(1):68–76.
35. Monot M, Archer F, Gomes M, Mornex JF, Leroux C. Advances in the study of transmissible respiratory tumours in small ruminants. Vet Microbiol. 2015;181(1):170–7.
36. DeMartini JC, Carlson JO, Leroux C, Spencer T, Palmarini M. Endogenous retroviruses related to jaagsiekte sheep retrovirus. Curr Top Microbiol Immunol. 2003;275:117–37.
37. Wang X, Liu S. Endogenous Jaagsiekte sheep retrovirus envelope protein promotes sheep trophoblast cell fusion by activating PKA/MEK/ERK1/2 signaling. Theriogenology. 2022;1(193):58–67.
38. Murcia PR, Arnaud F, Palmarini M. The transdominant endogenous retrovirus enJS56A1 associates with and blocks intracellular trafficking of jaagsiekte sheep retrovirus gag. J Virol. 2007;81(4):1762–72.
39. Arnaud F, Murcia PR, Palmarini M. Mechanisms of late restriction induced by an endogenous retrovirus. J Virol. 2007;81(20):11441–51.
40. Armezzani A, Arnaud F, Caporale M, di Meo G, Iannuzzi L, Murgia C, et al. The signal peptide of a recently integrated endogenous sheep betaretrovirus envelope plays a major role in eluding gag-mediated late restriction. J Virol. 2011;85(14):7118–28.

41. Viginier B, Dolmazon C, Lantier I, Lantier F, Archer F, Leroux C, et al. Copy number variation and differential expression of a protective endogenous retrovirus in sheep. PLoS ONE. 2012;7(7):e41965.
42. Kent M, Moser M, Boman IA, Lindtveit K, Árnyasi M, Sundsaasen KK, et al. Insertion of an endogenous Jaagsiekte sheep retrovirus element into the BCO2 - gene abolishes its function and leads to yellow discoloration of adipose tissue in Norwegian Spælsau (Ovis aries). BMC Genomics. 2021;22(1):492.
43. Cumer T, Pompanon F, Boyer F. Old origin of a protective endogenous retrovirus (enJSRV) in the Ovis genus. Heredity. 2019;122(2):187–94.
44. Garcia-Etxebarria K, Jugo BM. Evolutionary history of bovine endogenous retroviruses in the Bovidae family. BMC Evol Biol. 2013;20(13):256.
45. Deng R, Han C, Zhao L, Zhang Q, Yan B, Cheng R, et al. Identification and characterization of ERV transcripts in goat embryos. 2019 Jan 1; Available from: https://rep.bioscientifica.com/view/journals/rep/157/1/REP-18-0336.xml. Cited 2024 Oct 2.
46. Moawad AS, Wang F, Zheng Y, Chen C, Saleh AA, Hou J, et al. Evolution of endogenous retroviruses in the subfamily of caprinae. Viruses. 2024;16(3):398.
47. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci. 2020;117(17):9451–7.
48. Goubert C, Craig RJ, Bilat AF, Peona V, Vogan AA, Protasio AV. A beginner's guide to manual curation of transposable elements. Mob DNA. 2022;13(1):7.
49. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658–9.
50. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
51. Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mob DNA. 2021;12(1):2.
52. Rombel IT, Sykes KF, Rayner S, Johnston SA. ORF-FINDER: a vector for high-throughput gene identification. Gene. 2002;282(1):33–41.
53. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013–2015. Available from: https://www.repeatmasker.org.
54. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
55. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform. 2019;20(4):1160–6.
56. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59(3):307–21.
57. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–74.
58. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6(1):11.
59. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol. 2018;35(2):518–22.
60. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021;49(W1):W293–6.
61. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinforma Oxf Engl. 2019;35(3):526–8.
62. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2023. Available from: https://www.R-project.org/.
63. Singh U, Wurtele ES. orfipy: a fast and flexible tool for extracting ORFs. Birol I, editor. Bioinformatics. 2021;37(18):3019–20.
64. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
65. Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, et al. TimeTree 5: an expanded resource for species divergence times. Mol Biol Evol. 2022;39(8):msac174.
66. Chen L, Qiu Q, Jiang Y, Wang K, Lin Z, Li Z, et al. Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. Science. 2019;364(6446):eaav6202.
67. Chen ZH, Xu YX, Xie XL, Wang DF, Aguilar-Gómez D, Liu GJ, et al. Whole-genome sequence analysis unveils different origins of European and Asiatic mouflon and domestication-related genes in sheep. Commun Biol. 2021;4(1):1–15.
68. Storer J, Hubley R, Rosen J, Smit AFA. Curation guidelines for de novo generated transposable element families. Curr Protoc. 2021;1(6):e154.
69. Jiangzhao_Qian.qjiangzhao/TEtrimmer.2024.Available from: https://github.com/qjiangzhao/TEtrimmer.
70. Orozco-Arias S, Sierra P, Durbin R, González J. MCHelper automatically curates transposable element libraries across species [Internet]. bioRxiv; 2023. p. 2023.10.17.562682. Available from: https://www.biorxiv.org/content/https://doi.org/10.1101/2023.10.17.562682v1. Cited 2024 May 21.
71. Xiao R, Park K, Lee H, Kim J, Park C. Identification and classification of endogenous retroviruses in cattle. J Virol. 2008;82(1):582–7.
72. Garcia-Etxebarria K, Jugo BM. Genome-wide detection and characterization of endogenous retroviruses in bos taurus. J Virol. 2010;84(20):10852–62.
73. Sistiaga-Poveda M, Jugo BM. Evolutionary dynamics of endogenous Jaagsiekte sheep retroviruses proliferation in the domestic sheep, mouflon and Pyrenean chamois. Heredity. 2014;112(6):571–8.
74. Arnaud F, Varela M, Spencer TE, Palmarini M. Coevolution of endogenous betaretroviruses of sheep and their host. Cell Mol Life Sci CMLS. 2008;65(21):3422–32.
75. Spencer TE, Palmarini M. Endogenous retroviruses of sheep: a model system for understanding physiological adaptation to an evolving ruminant genome. J Reprod Dev. 2012;58(1):33–7.
76. Qi J wei, Xu M jie, Liu S ying, Zhang Y fei, Liu Y, Zhang Y kun, et al. Identification of Sheep Endogenous Beta-Retroviruses with Uterus-Specific Expression in the Pregnant Mongolian Ewe. J Integr Agric. 2013;12(5):884–91.
77. Klymiuk N, Müller M, Brem G, Aigner B. Characterization of endogenous retroviruses in sheep. J Virol. 2003;77(20):11268–73.
78. Morozov VA, Morozov AV, Lagaye S. Endogenous JSRV-like proviruses in domestic cattle: analysis of sequences and transcripts. Virology. 2007;367(1):59–70.
79. Perucatti A, Iannuzzi A, Armezzani A, Palmarini M, Iannuzzi L. Comparative Fluorescence In Situ Hybridization (FISH) Mapping of Twenty-Three Endogenous Jaagsiekte Sheep Retrovirus (enJSRVs) in Sheep (Ovis aries) and River Buffalo (Bubalus bubalis) Chromosomes. Animals. 2022;12(20):2834.
80. Ito J, Gifford RJ, Sato K. Retroviruses drive the rapid evolution of mammalian APOBEC3 genes. Proc Natl Acad Sci. 2020;117(1):610–8.
81. Garcia-Etxebarria K, Sistiaga-Poveda M, Jugo BM. Endogenous retroviruses in domestic animals. Curr Genomics. 2014;15(4):256–65.
82. Hayward A, Cornwallis CK, Jern P. Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. Proc Natl Acad Sci. 2015;112(2):464–9.
83. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat Genet. 2017;49(4):643–50.
84. Dias GB, Svartman M, Delprat A, Ruiz A, Kuhn GCS. Tetris is a foldback transposon that provided the building blocks for an emerging satellite DNA of Drosophila virilis. Genome Biol Evol. 2014;6(6):1302–13.
85. Sharma A, Wolfgruber TK, Presting GG. Tandem repeats derived from centromeric retrotransposons. BMC Genomics. 2013;4(14):142.
86. Macas J, Koblížková A, Navrátilová A, Neumann P. Hypervariable 3′ UTR region of plant LTR-retrotransposons as a source of novel satellite repeats. Gene. 2009;448(2):198–206.
87. Fultz D, Choudury SG, Slotkin RK. Silencing of active transposable elements in plants. Curr Opin Plant Biol. 2015;1(27):67–76.
88. Almeida MV, Vernaz G, Putman ALK, Miska EA. Taming transposable elements in vertebrates: from epigenetic silencing to domestication. Trends Genet TIG. 2022;38(6):529–53.
89. Czech B, Hannon GJ. one loop to rule them all: the ping-pong cycle and piRNA-guided silencing. Trends Biochem Sci. 2016;41(4):324–37.
90. Wang J, Yuan L, Tang J, Liu J, Sun C, Itgen MW, et al. Transposable element and host silencing activity in gigantic genomes. Front Cell Dev Biol. 2023;24(11):1124374.
91. Deniz Ö, Frost JM, Branco MR. Regulation of transposable elements by DNA modifications. Nat Rev Genet. 2019;20(7):417–31.

92. Tang L, Swedlund B, Dupont S, Harland C, Costa Monteiro Moreira G, Durkin K, et al. GWAS reveals determinants of mobilization rate and dynamics of an active endogenous retrovirus of cattle. Nat Commun. 2024;15(1):2154.
93. Arnaud F, Black SG, Murphy L, Griffiths DJ, Neil SJ, Spencer TE, et al. Interplay between ovine bone marrow stromal cell antigen 2/Tetherin and endogenous retroviruses. J Virol. 2010;84(9):4415–25.
94. Black SG, Arnaud F, Burghardt RC, Satterfield MC, Fleming JAGW, Long CR, et al. Viral particles of endogenous betaretroviruses are released in the sheep uterus and infect the conceptus trophectoderm in a transspecies embryo transfer model. J Virol. 2010;84(18):9078–85.
95. Fábryová H, Hron T, Kabíčková H, Poss M, Elleder D. Induction and characterization of a replication competent cervid endogenous gammaretrovirus (CrERV) from mule deer cells. Virology. 2015;1(485):96–103.
96. Preuss T, Fischer N, Boller K, Tönjes RR. Isolation and characterization of an infectious replication-competent molecular clone of ecotropic porcine endogenous retrovirus class C. J Virol. 2006;80(20):10258–61.
97. Wang J, Han GZ. Genome mining shows that retroviruses are pervasively invading vertebrate genomes. Nat Commun. 2023;14(1):4968.
98. Dunlap KA, Palmarini M, Varela M, Burghardt RC, Hayashi K, Farmer JL, et al. Endogenous retroviruses regulate periimplantation placental growth and differentiation. Proc Natl Acad Sci U S A. 2006;103(39):14390–5.
99. Cornelis G, Heidmann O, Degrelle SA, Vernochet C, Lavialle C, Letzelter C, et al. Captured retroviral envelope syncytin gene associated with the unique placental structure of higher ruminants. Proc Natl Acad Sci. 2013;110(9):E828–37.
100. Caporale M, Martineau H, De las Heras M, Murgia C, Huang R, Centorame P, et al. Host species barriers to Jaagsiekte sheep retrovirus replication and carcinogenesis. J Virol. 2013;87(19):10752–62.

Publisher's Note