Articles

A Compendium of ERVs in the Genomes of Major Primates

DOI:https://dx.doi.org/10.71373/SBHJ1439
Submitted 5 November 2025
Accepted 5 December 2025
Pulished 10 December 2025

Yifeng Li[b#], Chunlan Jiang[d#], Lanxiang Li[b], Zhining Zhang[d], Yunlin He[d], Kangming He[d], Zhiwei Lu[d], Yongjian Huang[d], Boying Liang[ac*]

ERVs integrate into host DNA, forming vertically inherited proviral sequences. These sequences not only record the historical imprints of ancient retroviruses but also play significant biological roles in modern host genomes. The widespread distribution of ERVs in primate genomes suggests that constructing a unified system for ERV site distribution would facilitate the study of their impact on host genomic biology and evolution. In this study, we utilized LTR_harvest to predict candidate ERV sequences in primate genomes and further identified and annotated the characteristics of these candidate sequences using LTR_digest software. We have compiled a comprehensive and well-annotated database(*http://ervsdb.omics.henbio.com/home/index/index.html?idx=1*). This database enables the acquisition of ERV data and corresponding neighboring genes for 61 primate species, characterization of ERV proviral structures, assessment of integration sites, and prediction of potential regulatory gene functions associated with ERV-LTRs. Using Trachypithecus francoisi as an example, we have preliminarily explored the potential influence of ERVs on host resistance or susceptibility to diseases, providing a theoretical reference for further research on ERVs in other animals.

Introduction

ERVs are special genetic elements that represent the remnants of ancient interactions between retroviruses and vertebrates, which have been preserved in host genomes and are vertically transmitted through Mendelian inheritance mechanisms[1]. These viral remnants are widely present in the genomes of various species, including mammals, birds, reptiles, and amphibians, exhibiting distinct distribution and expression patterns across different species[2]. To date, approximately 43 viral families have been identified in eukaryotes that have formed endogenous relationships with their hosts. ERVs serve as important tools for studying the co-evolution of hosts and retroviruses, providing a genomic perspective on the roles of infection and immunity in the evolutionary history of retroviruses[3].

Throughout their evolutionary history, ERVs have undergone a series of events such as replication, recombination, mutation, and insertion, leading to a diversity in their sequence quantity, distribution, and expression patterns within host genomes[4]. For instance, a solo-LTR arises from homologous recombination between the 5' and 3' LTRs of ancestral viruses, with the intervening protein-coding sequences being deleted[5]. ERVs that contain the three major protein-coding domains of retroviruses—gag, pol, and env—are defined as complete ERVs[6]. In fact, at least 85% of ERVs exist in reference genomes as isolated LTRs[5]. ERVs are classified into three major categories based on differences in their reverse transcriptase polymerase, ranging from Class I (gamma-like retroviruses) to Class III (spumaretrovirinae), and extending to the more diverse Class 16, encompassing a wide range of ERV sequence diversity[7].

The regulatory sequences provided by ERVs play a crucial role in altering genetic networks, cellular functions, and gene regulation. Studies have shown that RNA, cDNA, and protein products encoded by ERVs are closely associated with the occurrence of various diseases, including cancer, aging, chronic inflammation, and neurological disorders[8-11]. Given the importance of primates in medical research, exploring their genomics and epigenomics is particularly critical, as primate genomics is a golden key to understanding the fundamental aspects of human evolution and disease[12]. The genetic, neuroanatomical, physiological, and immunological similarities between primates and humans make them ideal preclinical models[13,14]. Therefore, computational identification of transposable elements, especially ERVs, has become an urgent need. With the advancement of high-throughput sequencing technologies, we have the opportunity to deeply explore the distribution and activity of these ancient viral remnants in primate genomes. Our goal is to map the landscape of ERVs in primates. Through this information, we can more accurately assess the extent of ERV invasion and replication in the genome, as well as their potential impacts on the structure and function of the host genome.

Our research also focuses on the biological functions of ERV-encoded products, including investigating potential gene fusions between ERVs and host genes, and how these fusion events regulate the expression of host genes, affect cellular functions, and contribute to disease development. By analyzing ERV expression data and their interactions with host genes, we can better understand the new roles that these ancient viral remnants play in modern organisms. These findings not only help us comprehend the evolutionary history of ERVs but may also reveal their potential impacts on human health and disease, providing new insights for future medical research and therapeutic strategies.

a.Chizhou People's Hospital, Chizhou, 247100, Anhui, China *b.Wuzhou Conservation and Research Center for Francois' Langurs, Wuzhou543000, Guangxi, China* *c.Department of Immunology, School of Basic Medical Sciences, Guangxi Medical University, Nanning 530021, Guangxi, China* *d.Guangxi Henbio Biotechnology Co., Ltd., Nanning 530009, Guangxi, China*

#Li Yifeng, Jiang Chunlan contributed equally to this work.

**Correspondence: LBY1187169449@163.com*

Materials and Methods

Download and Processing of Whole Genomes from 61 Primate Species

In the NCBI database (https://www.ncbi.nlm.nih.gov/), we used the Latin names of primate species as search keywords to download the whole genome files and reference genome annotation files required for analysis. Information on the 61 primate species is provided in Table S1.

ERV-Related Protein Entries

In the Pfam database (http://pfam-legacy.xfam.org/), we utilized the "KEYWORD SEARCH" method with the query keyword "retro" to identify entries related to ERV proteins. Additionally, by integrating information from the literature, we ultimately established a library containing 55 ERV-related protein entries. This library was used to detect the presence of protein domains encoded by the gag, pol, and env genes in ERVs. The relevant information is presented in Table S2.

Run the code as follows:
(1) Download the relevant entry: wget http://pfam-legacy.xfam.org/family/PF13966/hmm.
(2) Convert to HMMER2 format: hmmconvert -2 hmm > newhmm.
(3) Integrate into the library: mv newhmm ./pfam/PF13966.hmm.

LTR_harvest Prediction of Candidate ERVs

LTR_harvest is a highly efficient software tool for identifying LTRs in genomes. It employs a hidden Markov model algorithm to precisely recognize LTRs by aligning sequences and outputs their location, length, orientation, and completeness information. Due to the characteristic LTRs of LTR retrotransposons, they are ideal targets for computational identification[15]. Using LTR_harvest, candidate ERVs were filtered from whole-genome sequences with parameters set to similarity > 90%, length 1-15 kb, and TSD sequence length of 5-20 bp. These candidate ERVs possess LTR and TSD sequences at both ends, which is a distinct feature of retroviruses.

Run the code as follows:
(1) Decompress the genome file, taking Trachypithecus francoisi as an example: gunzip GCF_009764315.1_Tfra_2.0_genomic.fna.gz.
(2) Build the index: gt suffixerator -db GCF_009764315.1_Tfra_2.0_genomic.fna -indexname Tfra_genomic -tis -suf -lcp -des -ssp -sds –dna.
(3) Run LTR_harvest: gt ltrharvest -index Tfra_genomic -out Tfra.out -outinner Tfra.outinner -gff3 Tfra.gff -similar 80 -mindistltr 1000 -maxdistltr 15000 -mintsd 5 -maxtsd 20 > Tfra.result.

LTR_digest Identification and Annotation of Candidate ERVs

LTR_digest is a software tool designed for analyzing LTRs[16]. It employs local alignment and hidden Markov model algorithms to automatically detect retroviral protein domains, primer binding sites, and polypurine tracts within LTRs. In addition to identifying and annotating genes and reverse transcriptase (RT) coding sequences within LTR sequences, LTR_digest can determine various characteristics of LTRs, such as length, position, and orientation, and calculate relevant metrics. Moreover, LTR_digest can align LTR sequences with other sequences in the genome to reveal their origins, functions, and impacts, aiding in the classification and differentiation studies of LTR sequences. In this study, we combined the ERV-related protein domain library and used the LTR_digest program to conduct an in-depth analysis of the LTR regions of these candidate ERVs, with particular attention to sequence features such as PPT and PBS. We also integrated the HMMER algorithm to further analyze whether specific protein domains exist within the LTR sequences of these candidate ERVs.

Run the code as follows:
(1) Sort: gt gff3 -sort ./harvest_out/Tfra.gff > Tfra_sort_gff.
(2) Run LTR_digest with the constructed library: gt ltrdigest –hmms./pfam/*hmm -outfileprefix myspecies_ltrdigest Tfra_sort_gff./index/Tfra_genomic > myspecies_ltrdigest_output_gff.
(3) Filter out candidate ERVs that have no protein-domain hits at all (to help remove candidates that are probably not LTR-retrotransposon insertions; this step is implemented with a Lua script): gt select -rule_files filter_protein_match.lua -- < myspecies_ltrdigest_output_gff > myspecies_ltrdigest_output_gff2.

Fragmentation Analysis of ERVs in Primate Genomes

For the endogenous retrovirus sequences obtained, we conducted a fragmentation analysis based on their structure. The analysis included the following categories: complete open reading frame (ORF) endogenous retrovirus proviral sequences, sequences with deletions in the gag gene, sequences with deletions in the pol gene, sequences with deletions in the env gene, and sequences with deletions in either the 5'LTR or 3'LTR. Additionally, we quantified and determined the proportion of solo-LTR sequences of endogenous retroviruses in different primate genomes. The criterion for structural completeness was the presence of protein domains in the gag, pol, and env regions.

Acquisition and Functional Analysis of Genes Adjacent to ERVs

Based on the chromosomal distribution coordinates of complete ERVs in each primate genome, we utilized the bedtools software in conjunction with the corresponding annotation files downloaded from NCBI to perform a matching operation. The matching criteria were that gene sequences intersecting with the complete ERVs within a specified range upstream and downstream were identified. Subsequently, we conducted a functional analysis of these selected genes.

Classification of ERVs in Trachypithecus francoisi Genome

Generally, ERVs are classified based on their relationship with exogenous retroviruses, including seven genera[17,18]. For this classification, annotations primarily follow the guidelines of the International Committee on Taxonomy of Viruses (ICTV). Typically, ERVs are divided into three categories based on the sequence similarity of their pol region to the reverse transcriptase sequences of exogenous retroviruses: Class I is similar to gamma and epsilon retroviruses, Class II is similar to alpha, beta, and delta retroviruses, and Class III is similar to spumaviruses[19].

To facilitate classification analysis, we selected 21 well-annotated viral sequences with full-length sequences as reference sequences and constructed a phylogenetic tree together with ERV sequences from Trachypithecus francoisi genome that have complete structures (including gag, pol, and env protein domains). This allows us to categorize these complete ERVs into one of the three main ERV classes. The complete sequences of the 21 viruses were downloaded from the NCBI database based on their sequence numbers (Table1). Using the phylosuit software, we performed sequence alignment with MAFFT using default parameters and selected the optimal model as GTR+G. The phylogenetic tree was constructed using the maximum likelihood (ML) method with a bootstrap parameter of 1000. The tree was further refined using iTOL for visualization.

Table 1 Virus sequence related information

Category	Abbreviation	Retrovirus Name	GenBank Number
ClassI	GALV	Gibbon ape leukemia virus	AAA46810.1
ClassI	KoRV	Koala retrovirus	AAF15098.1
ClassI	MDEV	Mus dunni endogenous virus	AAC31805.1
ClassI	PERV	Sus scrofa porcine endogenous retrovirus	AAC16767.1
ClassI	RMLV	Rauscher murine leukemia virus	AAB86912.1
ClassI	ZFERV	Danio rerio endogenous retrovirus	AAM34208.1
ClassI	WEHV1	Walleye epidermal hyperplasia virus 1	AAD30048.1
ClassI	WEHV2	Walleye epidermal hyperplasia virus 2	AAD30054.1
ClassII	ALV	Avian leukosis virus (ALV-A)	AMP18914.1
ClassII	RSV	Rous sarcoma virus	AAC08988.1
ClassII	RERV	Rabbit endogenous retrovirus	AAM81191.1
ClassII	SRV-1	Simian retrovirus 1	AAA47732.1
ClassII	BLV	Bovine leukemia virus	AAA42785.1
ClassII	HIV-1	Human immunodeficiency virus 1 (HIV-1)	AAA44325.1
ClassII	JSRV	Jaagsiekte sheep retrovirus	AAK38686.1
ClassII	SIV	Simian immunodeficiency virus	AAA47633.2
ClassII	FIV	Feline immunodeficiency virus	M25381.1
ClassIII	EFV	Equine foamy virus	AAF64414.1
ClassIII	FeFV	Feline foamy virus	CAA11581.1
ClassIII	HFV	Human foamy virus	CAA69003.1
ClassIII	SMRV	Squirrel monkey simian foamy virus	ADE05995.1

Results

Primate Classification and Overview of Whole-Genome ERVs and Solo-LTR Counts

This study encompasses 61 primate species, which are scientifically classified into 7 families and 20 genera(Figure 1A) . To gain a deeper understanding of the genomic characteristics of these primates, Figure 1B provides a visual representation of the distribution of ERVs and solo-LTRs across their whole genomes. ERV elements are widely present in primate genomes and play significant roles. Among the species studied, the representative of Old World monkeys, the Arabian baboon, has the highest total amount of ERVs in its genome. In contrast, the representative of New World monkeys, the white-eared titi monkey, exhibits a lower ERV content. Regarding the number of solo-LTRs, the genome of Trachypithecus phayrei shows the most significant presence. Conversely, the genome of Plecturocebus donacophilus has a relatively low number of solo-LTRs.

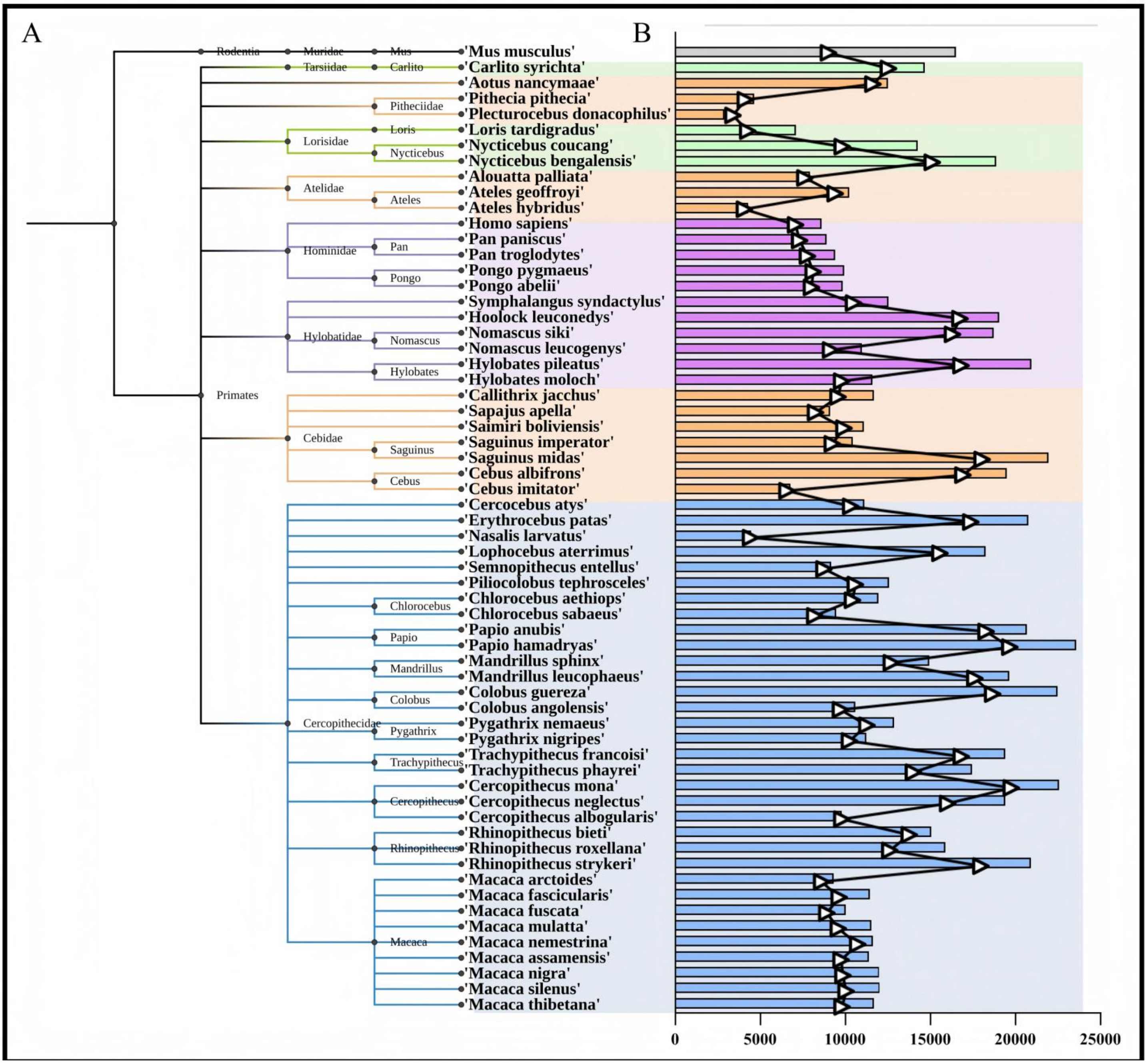

Figure 1. Phylogeny and endogenous retrovirus (ERV) load of 61 primate species. (A) Phylogenetic overview: outer colored bars indicate family-level clades, visually displaying the phylogenetic relationships and taxonomic affiliations of the 61 species.(B) ERV content versus genome size: bars represent the genome size of each species; the line plot shows the total number of ERV copies in each genome. The two panels together allow direct comparison of ERV burden across primates against their phylogenetic background.

Research on the Proportion and Fragmentation Distribution Characteristics of ERVs in Primate Genomes

We conducted an in-depth comparison of the proportion of ERVs in the genomes of 61 primate species, revealing significant differences in the quantity and distribution of ERVs among different primates. The genome of Trachypithecus francoisi exhibited the highest proportion of ERV sequences, reaching 14.2%, while the genome of Symphalangus syndactylus had the lowest proportion at only 2.78%(Figure 2A). ERVs in primates predominantly exist as solo-LTRs (long terminal repeats)((Figure 2B)). To gain a deeper understanding of the fragmentation of ERVs in primate genomes, we performed a comprehensive analysis of ERVs (domains ≥ 1) across the 61 primate species. The results showed that ERVs exhibit similar fragmentation distribution trends within the primate population. Among these primates, the pol gene dominates, followed by the gag-pol-env structure(Figure 2C).

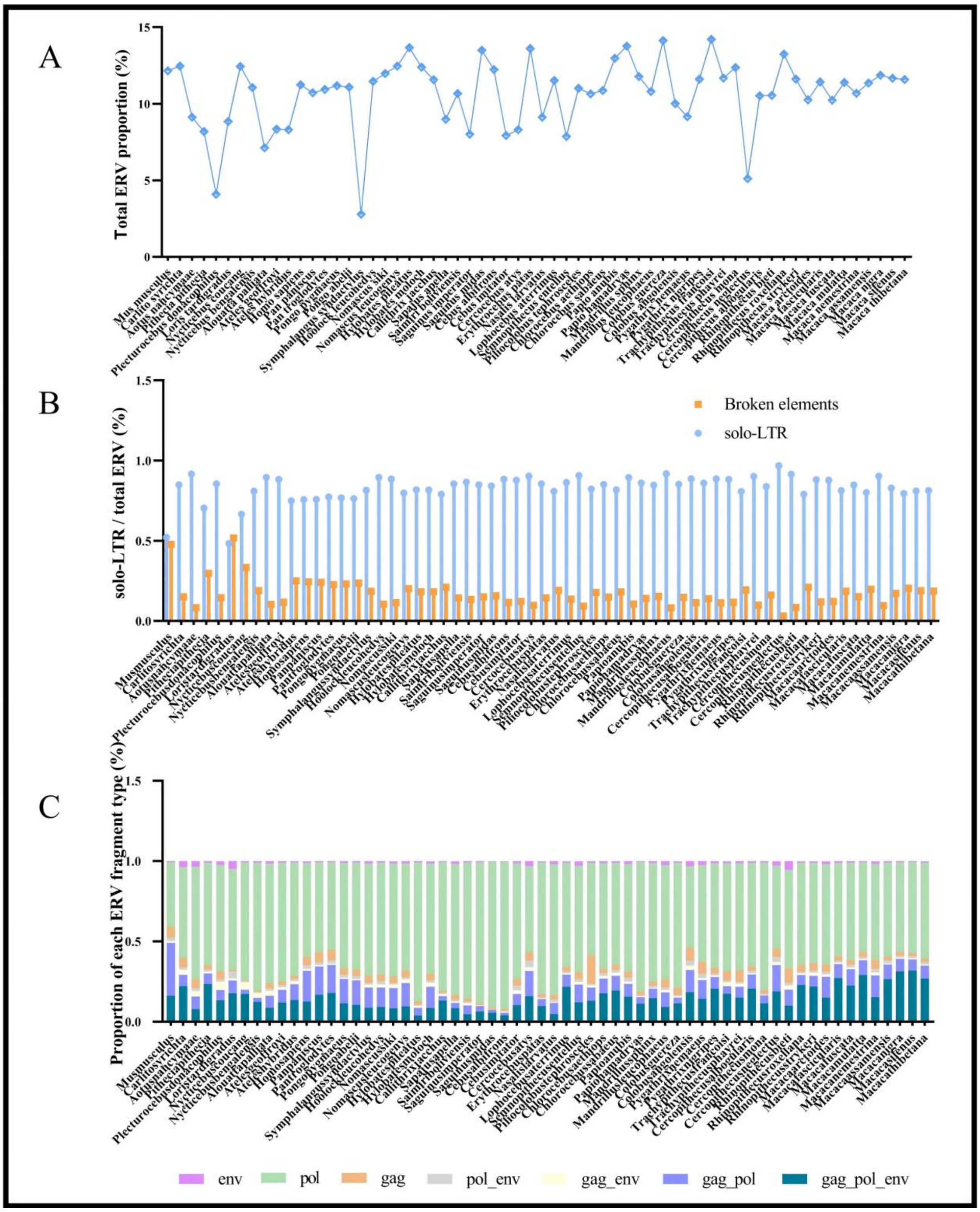

Figure 2 Fragmentation signatures and relative abundance of ERVs in primate genomes.(A) ERV proportion: line plot shows the percentage of ERV sequences relative to the whole genome across 61 primate species.(B) solo-LTR proportion: blue bars indicate the percentage of solo-LTRs within the total ERV content of each species.(C) Fragmentation landscape: stacked bars display the relative proportion (%) of each ERV fragment type (gag, pol, env, gag-env, etc.) within the species-specific ERV repertoire; colors denote fragment categories, enabling direct comparison of ERV fragmentation patterns among primates.

The distribution and quantity of different ERV fragmentation types vary across the genomes of various primate species. The proportions of complete ERVs(gag-pol-env), gag-env, pol and env fragmentation patterns are significantly higher in the genomes of Prosimians compared to other primate groups, while the proportion of solo-LTRs is significantly lower. In contrast, there are no significant differences in the proportions of all fragmentation types between New World monkeys and Old World monkeys. The proportion of solo-LTR structures among candidate ERVs containing domains varies significantly across different primate species, with the highest proportion observed in the genomes of Old World monkeys. The proportion of complete ERVs is highest in the genomes of Prosimians, followed by Old World monkeys, while New World monkeys exhibit the lowest proportion ($p < 0.05$) (Figure 3).

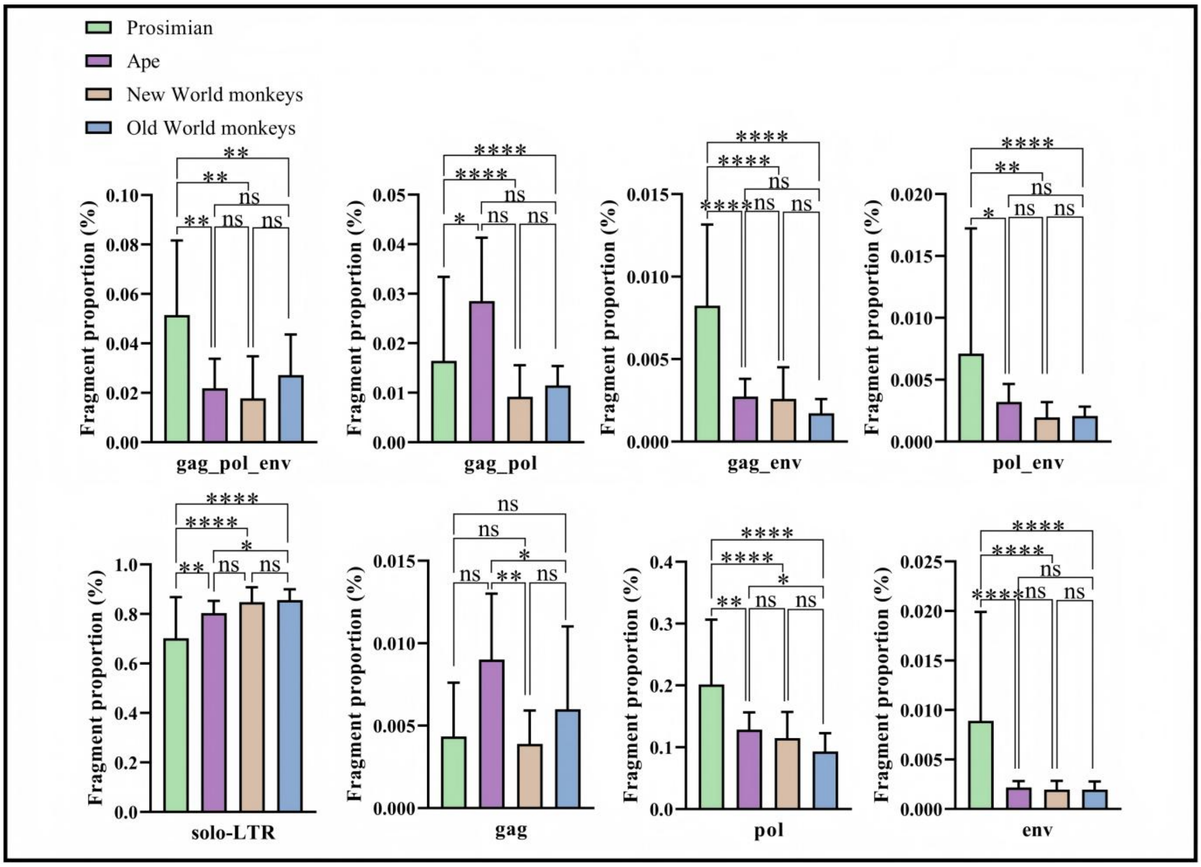

Figure 3 Composition of ERV fragment types across primate genomes. Bars represent mean copy numbers (mean ± SEM; n = number of species per family) of the main ERV fragments (gag, pol, env, gag-env, etc.) in four primate families. Inter-family differences for each fragment category were tested with Kruskal−Wallis tests; significance levels are indicated above bars: *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001; NS, not significant.Colors denote fragment classes, allowing direct comparison of ERV fragmentation patterns among major primate lineages.

Distribution of Genes Adjacent to ERVs in Primate Genomes

Among the 61 primate species, only 24 have both complete whole-genome sequencing files and genome annotation files available. For these 24 primate genomes, genes located within 10 kb upstream and downstream of the identified ERV sequences were identified. Any genes with overlapping sequences within the specified range were defined as adjacent to ERVs. Additionally, some genes overlapped with ERV sequences, suggesting potential gene fusion events with ERVs. Details are provided in Table 2.

Table 2 Neighboring Genes with Sequence Overlap with ERVs

NCBI Number	Species Latin Name	Complete ERVs Number	Crossing-over Genes Number
GCF_000952055	Aotus nancymaae	82	37
GCF_000955945	Cercocebus atys	173	79
GCF_000956065	Macaca nemestrina	171	93
GCF_001604975	Cebus imitator	86	40
GCF_001698545	Rhinopithecus bieti	127	70
GCF_002776525	Piliocolobus tephrosceles	292	143
GCF_003339765	Macaca mulatta	666	250
GCF_006542625	Nomascus leucogenys	211	122
GCF_007565055	Rhinopithecus roxellana	765	324
GCF_009761245	Sapajus apella	111	48
GCF_027406575	Nycticebus coucang	823	300
GCF_012559485	Macaca fascicularis	580	242
GCF_015252025	Chlorocebus sabaeus	332	113
GCF_016699345	Saimiri boliviensis	69	47
GCF_009764315	Trachypithecus francoisi	589	358

Among the 358 adjacent host genes identified in Trachypithecus francoisi, 30 are related to immunity, including *CCL15, SGCZ, VPS37A, SH2D4A, PIWIL2, NRG1, ADAM32, HOOK3, BRCA2, C1QTNF9, SACS, RAB27B, TLR5, ZNF678, ATF6, PGLYRP4, THEM4, THEM5, MAGI3, ST7L, SLC30A7, CCDC18, EVI5, GLMN, MAGEA8, ZNF75D, SPRY3, F8, CTAG2, COL4A6*, and *ATG4A*. KEGG enrichment analysis, revealed that the genes overlapping with ERVs upstream and downstream are primarily enriched in pathways related to substance dependence, nucleotide metabolism, replication and repair, cardiovascular diseases, cellular motility, cofactor and vitamin metabolism, amino acid metabolism, infectious diseases: bacterial, nervous system, amino acid metabolism, and signaling molecules and interactions (Figure 4).

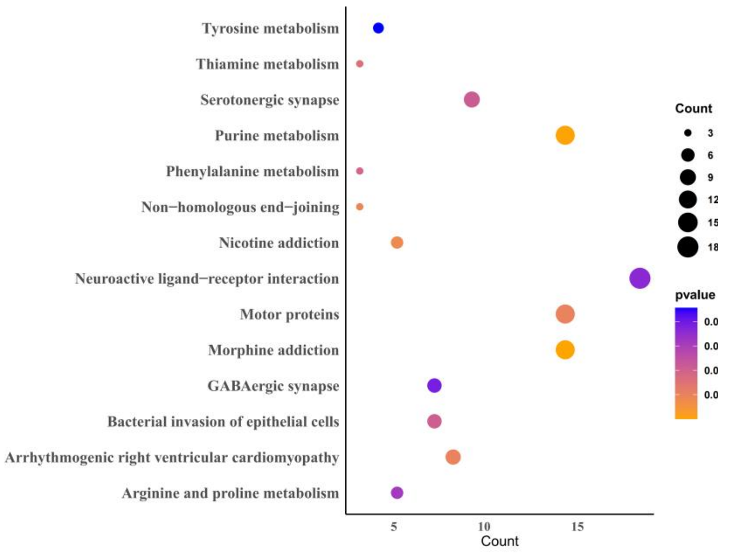

Figure 4 KEGG pathway enrichment bubble plot. X-axis: number of significantly enriched genes (Count); Y-axis: pathway names; dot size: enrichment factor; color scale: adjusted p-value (red–blue gradient).

The Classification of ERVs in the Genome of Trachypithecus francoisi

In the genome of Trachypithecus francoisi, a total of 17,436 candidate ERVs were identified, with related sequences accounting for 14.2% of the total genome. Among the protein domains of the identified ERVs, the pol gene represented the highest proportion. Out of the 17,436 candidate ERVs, 589 sequences were structurally intact, with an average length of 27,802 bp, and the average length of their LTR sequences was 376 bp. The distribution density of these structurally intact ERVs across the sequences is shown in Table S3. We constructed a reference phylogenetic tree using MEGA11 with the RT sequences of 21 known structurally intact retroviruses(Figure 5A) . In the figure, yellow represents Class I, which includes the genera Gammaretrovirus and Epsilonretrovirus; green represents Class II, which includes the genera Alpharetrovirus, Betaretrovirus, Deltaretrovirus, and Lentivirus; and purple represents Class III, which belongs to the genus Spumavirus.

A phylogenetic tree was constructed using the 589 structurally intact ERVs from the black snub-nosed monkey with sufficiently conserved RT sequences, along with the RT sequences of the aforementioned 21 known structurally intact retroviruses(Figure 5B) . We identified 295 Class I ERVs, accounting for 48.7% of the total; 311 Class II ERVs, accounting for 51.3%.

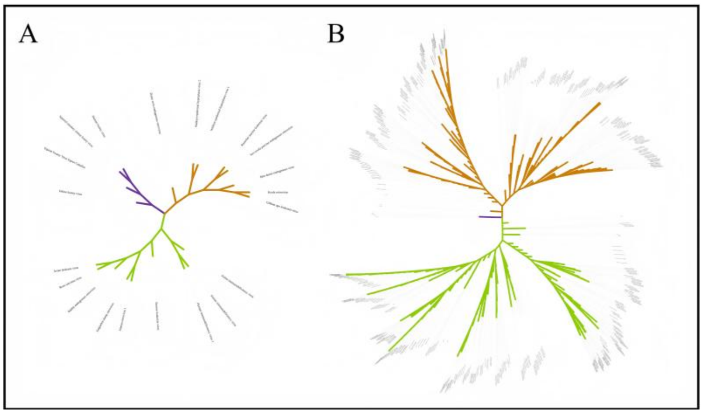

Figure 5 Classification and phylogenetic placement of Trachypithecus francoisi endogenous retroviruses (ERVs). (A) Reference phylogeny; (B) Trachypithecus francoisi ERV tree; node support shown as SH-aLRT values >80 %.branches are colored according to the three ERV clades identified in this species.

Discussion

In the vast field of life sciences, ERVs, as ancient and enigmatic components of the genome, have consistently been a focal point of research. They are not only key players in genome evolution but may also profoundly influence the physiological and pathological processes of their hosts. This study employs bioinformatics methods to thoroughly investigate the distribution of ERVs across the entire genome in major primate species.

Previous studies on ERVs in primate genomes have revealed the following findings: 425 full-length ERVs containing reverse transcriptase (RT) sequences were identified in the gorilla genome[20], 284 ERVs were detected in the golden snub-nosed monkey (Rhinopithecus roxellana), with 85 structurally intact; the black snub-nosed monkey (Rhinopithecus bieti) harbored 56 intact ERVs; while the François' langur (Trachypithecus francoisi) exhibited 160 structurally complete ERVs, accounting for 40% of its total ERV content[21]. Concurrently, ERVs constitute approximately 8% of the human genome[22] and about 10% in mice. However, discrepancies exist between these findings and the results of the present study, which may be attributed to methodological variations and parameter settings across different studies. In previous research, gorilla ERVs were identified using LTR_struc combined with BLAST alignment of conserved RT sequences, whereas ERVs in golden snub-nosed monkeys, black snub-nosed monkeys, and François' langurs were screened through BLAST analysis using gibbon leukemia virus gag and env sequences as probes, followed by RetroTecter software validation. In contrast, our study employed LTR_harvest for candidate ERV prediction and performed comprehensive sequence characterization and annotation using LTR_digest, enabling a more systematic and exhaustive screening and analysis of ERVs across genomes.

The study revealed that ERVs are classified into intact and various fragmented types based on their canonical structural features (5' LTR-gag-pol-env-3' LTR), exhibiting significant distribution disparities across primate genomes. The François' langur (Trachypithecus francoisi) displayed the highest ERV content at 14.2%, whereas the siamang (Symphalangus syndactylus) genome contained the lowest proportion at merely 2.78%. This divergence likely stems from the combined effects of evolutionary history, genome size and complexity, ecological pressures, and genetic variation, collectively driving heterogeneous patterns of ERV insertion, deletion, rearrangement, and population distribution[23]. Notably, ERVs predominantly exist in fragmented forms among primates, with solo-LTRs constituting the primary remnants, while pol genes remain most prevalent. The pol gene, encoding reverse transcriptase and integrase—core components for ERV replication and genomic integration—suggests these fragmented sequences may retain residual functionality or regulatory potential within host genomes[24]. Conversely, the widespread loss of env genes correlates with their distinct evolutionary fate and potential detrimental impacts on both ERVs and hosts. Further analysis demonstrated that most ERVs in François' langur genomes are incomplete, with env gene absence correlating with a 30-fold increase in ERV proliferation rates, rendering them genomic super-spreaders. Additionally, the potential immunosuppressive properties of Env proteins may disadvantage both ERVs and hosts, leading to gradual evolutionary devaluation and subsequent loss of env gene functionality through natural selection.

From a geographic distribution perspective, Old World monkeys and New World monkeys—separated by continental drift—exhibit distinct ERV sequence divergence. However, the similarity between human ERV sequences and those of Old World monkeys, particularly in terms of quantity and structural integrity, supports their shared ancestral origin. In this study, we conducted detailed classification of ERVs within the François' langur genome, revealing their predominant classification into Class I and Class II ERVs, with balanced numerical distribution between these two categories. Notably, while Cui et al. identified 412 foamy viruses (Class III) representing six distinct lineages in amphibians—demonstrating their broad distribution in certain species—our analysis failed to detect Class III sequences in François' langurs, suggesting either extreme rarity or complete absence of foamy viruses in this species[25,26]. Different ERV classes exhibit distinct amplification and retention mechanisms during evolution, potentially explaining their variable genomic abundance, distribution patterns, and evolutionary trajectories.

Research indicates that ERV LTRs frequently demonstrate promoter or enhancer activity, functioning as cis-regulatory elements to modulate host gene expression near integration sites. Wang et al. demonstrated that p53 binding sites within HERV LTRs regulate host gene expression, with over one-third of p53 binding sites originating from ERV LTRs[27]. Furthermore, ERV-host gene fusion events can alter transcriptional regulation, exemplified by ERV9 insertions upstream of human TP63 that enhance specific p63 isoform expression—a critical mechanism for germline stability[28]. Among the 358 protein-coding genes potentially fused with intact ERVs in François' langurs, enriched signaling pathways suggest that ERVs may significantly influence host gene expression through either fusion events or cis-regulatory mechanisms. These interactions likely participate in regulating fundamental cellular processes, environmental adaptation, disease pathogenesis, and complex behavioral modulation in the host species.Taking immune-related genes as an example, the NF-κB binding sites within ERV-LTRs can bidirectionally regulate the expression of TLR5 (the bacterial flagellin receptor), PGLYRP4 (a peptidoglycan recognition protein), and BRCA2 (a core factor in homologous recombination repair), providing a testable molecular model for "ERV-immune co-evolution."

ERVsDb compiles the first comprehensive atlas of endogenous retroviruses (ERVs) across 61 primate species, couples deep annotation of adjacent genes with data-driven prediction of ERV activation potential, and delivers all analyses through a zero-code interface—thereby closing a long-standing gap for a dedicated, systematic yet user-friendly resource. By systematically mining the regulatory power of long terminal repeats (LTRs), the database converts ERVs from "fossil sequences" into "functional elements". Each ERV is assigned a quantifiable and experimentally addressable identity, propelling the field from descriptive correlation to mechanistic validation and providing a sustainable, expandable data foundation for deciphering how ERVs continuously reshape primate genome evolution, immune networks and disease susceptibility.

Reference

1. JOHNSON W E. Origins and evolutionary consequences of ancient endogenous retroviruses [J]. Nat Rev Microbiol, 2019, 17(6): 355-70.
2. WANG J, HAN G Z. Genome mining shows that retroviruses are pervasively invading vertebrate genomes [J]. Nat Commun, 2023, 14(1): 4968.
3. YU J, QIU P, AI J, et al. Endogenous retrovirus activation: potential for immunology and clinical applications [J]. Natl Sci Rev, 2024, 11(4): nwae034.
4. MALIK H S, HENIKOFF S, EICKBUSH T H. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses [J]. Genome Res, 2000, 10(9): 1307-18.
5. THOMAS J, PERRON H, FESCHOTTE C. Variation in proviral content among human genomes mediated by LTR recombination [J]. Mob DNA, 2018, 9: 36.
6. DENG R, HAN C, ZHAO L, et al. Identification and characterization of ERV transcripts in goat embryos [J]. Reproduction, 2019, 157(1): 115-26.
7. LAMBOWITZ A M, BELFORT M. Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution [J]. Microbiol Spectr, 2015, 3(1): MDNA3-0050-2014.
8. MODENINI G, ABONDIO P, BOATTINI A. The coevolution between APOBEC3 and retrotransposons in primates [J]. Mob DNA, 2022, 13(1): 27.
9. [9] HONG Y, HU C B, BAI J, et al. Essential role of an ERV-derived Env38 protein in adaptive humoral immunity against an exogenous SVCV infection in a zebrafish model [J]. PLoS Pathog, 2023, 19(4): e1011222.
10. TAM O H, OSTROW L W, GALE HAMMELL M. Diseases of the nERVous system: retrotransposon activity in neurodegenerative disease [J]. Mob DNA, 2019, 10: 32.
11. CANADAS I, THUMMALAPALLI R, KIM J W, et al. Tumor innate immunity primed by specific interferon-stimulated endogenous retroviruses [J]. Nat Med, 2018, 24(8): 1143-50.
12. SALAVATIHA Z, SOLEIMANI-JELODAR R, JALILVAND S. The role of endogenous retroviruses-K in human cancer [J]. Rev Med Virol, 2020, 30(6): 1-13.
13. PAN M T, ZHANG H, LI X J, et al. Genetically modified non-human primate models for research on neurodegenerative diseases [J]. Zool Res, 2024, 45(2): 263-74.
14. LIN X, WANG H, CHEN J, et al. Nonhuman primate models of ischemic stroke and neurological evaluation after stroke [J]. J Neurosci Methods, 2022, 376: 109611.
15. ELLINGHAUS D, KURTZ S, WILLHOEFT U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons [J]. BMC Bioinformatics, 2008, 9: 18.
16. STEINBISS S, WILLHOEFT U, GREMME G, et al. Fine-grained annotation and classification of de novo predicted LTR retrotransposons [J]. Nucleic Acids Res, 2009, 37(21): 7002-13.
17. GIFFORD R J, BLOMBERG J, COFFIN J M, et al. Nomenclature for endogenous retrovirus (ERV) loci [J]. Retrovirology, 2018, 15(1): 59.
18. KRUPOVIC M, BLOMBERG J, COFFIN J M, et al. Ortervirales: New Virus Order Unifying Five Families of Reverse-Transcribing Viruses [J]. J Virol, 2018, 92(12).
19. VARGIU L, RODRIGUEZ-TOME P, SPERBER G O, et al. Classification and characterization of human endogenous retroviruses; mosaic forms are common [J]. Retrovirology, 2016, 13: 7.
20. POLAVARAPU N, BOWEN N J, MCDONALD J F. Identification, characterization and comparative genomics of chim-

panzee endogenous retroviruses [J]. Genome Biol, 2006, 7(6): R51.

21. WANG X, WANG B, LIU Z, et al. Genome-wide characterization of endogenous retroviruses in snub-nosed monkeys [J]. PeerJ, 2019, 7: e6602.
22. GEIS F K, GOFF S P. Silencing and Transcriptional Regulation of Endogenous Retroviruses: An Overview [J]. Viruses, 2020, 12(8).
23. ZHANG R, WU M, XIANG D, et al. A primate-specific endogenous retroviral envelope protein sequesters SFRP2 to regulate human cardiomyocyte development [J]. Cell Stem Cell, 2024, 31(9): 1298-314 e8.
24. WANG J, LU X, ZHANG W, et al. Endogenous retroviruses in development and health [J]. Trends Microbiol, 2024, 32(4): 342-54.
25. YEDAVALLI V R K, PATIL A, PARRISH J, et al. A novel class III endogenous retrovirus with a class I envelope gene in African frogs with an intact genome and developmentally regulated transcripts in Xenopus tropicalis [J]. Retrovirology, 2021, 18(1): 20.
26. CHEN Y, ZHANG Y Y, WEI X, et al. Multiple Infiltration and Cross-Species Transmission of Foamy Viruses across the Paleozoic to the Cenozoic Era [J]. J Virol, 2021, 95(14): e0048421.
27. WANG T, ZENG J, LOWE C B, et al. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53 [J]. Proc Natl Acad Sci U S A, 2007, 104(47): 18613-8.
28. BEYER U, KRONUNG S K, LEHA A, et al. Comprehensive identification of genes driven by ERV9-LTRs reveals TN FRSF10B as a re-activatable mediator of testicular cancer cell death [J]. Cell Death Differ, 2016, 23(1): 64-75.

Acknowledgments

The authors express their gratitude to the data providers and to all participants in this study.

Consent for publication

It is not applicable.

Ethics Statement.

It is not applicable.

Data and Code Availability.

The primate genomic data used in this project were obtained from the public database NCBI, the analysis code is provided within the main text, and the database information is publicly accessible free of charge.

Competing interests

All authors declared no competing interests.

Funding

This work is supported by the Natural Science Foundation (2025KY0127).

Authors’contributions

YFL, CLJ and LXL designed the research and provided relevant knowledge; ZNZ, YLH download the required data; ZWL, YJH generated the Figures and tables; BYL wrote and edited the manuscript. All of the authors read and approved the final manuscript.

SUPPORTING INFORMATION

Additional supplementary information is available for download and review in the supplementary information section located on the right-hand side of this article's HTML page.

supplementary tables.xlsx:

- **Table S1.**
- **Table S2.**
- **Table S3.**