Screening Envelope Genes Across Primate Genomes Reveals Evolution and Diversity Patterns of Endogenous Retroviruses

Saili Chabukswar[1,2], Nicole Grandi[1], Elena Soddu[1,2], Liang-Tzung Lin[3,4], Enzo Tramontano[1]

[1]Laboratory of Molecular Virology, Department of Life and Environmental Sciences, University of Cagliari, 09042 Cagliari, Italy

[2]International Ph.D. Program in Medicine, College of Medicine, Taipei Medical University, Taipei 110, Taiwan

[3]Department of Microbiology and Immunology, School of Medicine, College of Medicine, Taipei Medical University, Taipei 110, Taiwan

[4]Graduate Institute of Medical Sciences, College of Medicine, Taipei Medical University, Taipei 110, Taiwan

*Corresponding author

Prof. Enzo Tramontano

E-mail address: tramon@unica.it

Abstract

Endogenous Retroviruses (ERVs) are integrated into the host DNA as result of ancient germ line infections, majorly by extinct exogenous retroviruses. In fact, vertebrates' genomes contain thousands of ERV copies, providing "fossil" records for the ancestral retroviral diversity and its evolution within the host. Like exogenous retroviruses, ERV proviral sequence consists of *gag, pro, pol,* and *env* genes flanked by long terminal repeats (LTRs). Among them, the characterization of *env* gene changes over time allows both to understand ERVs evolutionary trajectory and possible physiological and pathological domestication. To this aim, we reconstructed 32 Env sequences representing the prototypes of these ancestral proteins in Class I, Class II, and Class III HERVs. These reconstructed Envs were then employed in diverse methods comprising similarity search, phylogenetic analysis, and examination of recombination events occurred within primates' genomes that were applied to 43 primate species across the *Catarrhini* and *Platyrrhini* parvorders. Through a comprehensive pipeline we reconstitute a phylogenetic distribution of ERV based specifically on the *env* genes, showing that the ERVs have been prevalent and widely distributed across the primate lineage. We observed for the first time the presence of the HML groups in the *Platyrrhini* parvorder, possibly indicating initiation of spread of HML supergroup before the split between New World Monkeys (NWM) and Old World Monkeys (OWM) i.e. even before 40 mya. Importantly, we confirmed notable interclass and intra-class *env* recombination events showing the phenomenon of "*env* snatching" among primates' ERVs. As a result, we demonstrate that tracing the diversity patterns of ERVs' *env* provides relevant insights into the retroviral evolutionary history of ERVs in *Catarrhini* and *Platyrrhini* parvorders. Overall, our findings reveal that *env* recombination contributes to the diversification of ERVs, thereby broadening our comprehension of retroviral and primate evolution.

Keywords

Endogenous Retroviruses, Envelope Protein, Evolution, Recombination

Introduction

Retroviruses are among the first viruses discovered a century ago and comprises a diverse range of animal viruses, that are united by a common mechanism of replication (Mager & Stoye, 2015). Their replication cycle involves the reverse transcription of a ssRNA (single stranded RNA) into a dsDNA (double stranded DNA) molecule followed by the insertion of the viral DNA into the chromosomal DNA of host cells resulting into a provirus. Such proviruses contain promoters and enzymes essential for transcribing viral RNA as well as encoding all structural and regulatory proteins crucial for assembling new virions (Mager & Stoye, 2015). Usually, retroviruses infect specific types of somatic cells and are consequently integrated into the genomic DNA of these cells but are not passed on to the host progeny. However, some retroviruses can infect the germ line cells and thereby get integrated into the host genome and being vertically transmitted from one host to another over generations and are termed endogenous retroviruses (ERVs)(Sverdlov, 2000). Over the span of millions of years, thousands of ERV copies have been accumulated throughout the vertebrate lineage and can range from complete proviral structure to highly fragmented remnants of the provirus (Hayward, 2017; Mager & Stoye, 2015; Weiss, 2006).

Likewise, exogenous RVs, ERVs also consists of four major viral protein-coding genes i.e. *gag, pro, pol* and *env* flanked by identical long-terminal repeats (LTRs) at 5' and 3' ends of the provirus (Hayward, 2017; Johnson, 2019). The *gag* gene encodes for viral capsid core, *pro* encodes for protease enzyme, *pol* for the replicative enzymes i.e. reverse transcriptase (RT) and integrase (IN), and finally *env* encodes for envelope glycoproteins (Johnson, 2015; Weiss, 2016). The proviral LTRs derives from three regions of the viral ssRNA i.e. U3, R and U5, among which the R is repeated at both the ends whereas U5 and U3 are present as one copy each at 5' and 3' ends respectively, producing then identical LTRs with the structure U3-R-U5 at both the end of provirus. Among the four viral genes *gag, pro* and *pol* are well conserved across retroviral species while the *env* is said to be the most diverse, being responsible for the specific tropism of the infection (Henzy & Johnson, 2013). Given that the RT of *pol* gene is a highly conserved region, it is considered very useful in understanding phylogenetic relationships and classification of retroviral lineages (Henzy & Johnson, 2013). Based on RT sequences, ERVs are classified into three main classes i.e. Class I (Gammaretroviruses and Epsilonretroviruses), Class II (related to Deltaretroviruses, Lentiviruses, Betaretroviruses, and Alpharetroviruses) and Class III (Spumaretroviruses) (Blomberg et al., 2009; Gifford et al., 2018). Various phylogenetic relationships of ERVs are also based on the comparisons between

RT and the transmembrane (TM) region of the *env* gene, however discrepancies may arise due to the diversity of the *env* gene among different ERVs of the same class (Blomberg et al., 2009). While the RT sequences highlight the historical lineage that gives rise to a retrovirus genome, studying the *env* gene provides insights into retroviral evolution, signaling the emergence of new viruses leading to cross species transmission up to to the generation of new retroviral lineages.

In fact, the *env* gene is the first line determinant of the RVs' tissue/cell tropism: it encodes Env glycoprotein forming a trimer in the endoplasmic reticulum and is cleaved by the host protease in the late Golgi apparatus (Henzy & Johnson, 2013; Pancino et al., 1994). Each monomer of the Env glycoprotein trimer consists of surface (SU) and transmembrane (TM) domains. The SU is involved in the cellular receptor(s) recognition, initiating the infection process. Being exposed to the host immune system, the SU is under high selective pressure and is hence very diverse among RVs, while the TM subunit is involved in fusogenicity, i.e., fusion of virus and cell membrane during the entry process, and is moderately conserved among the RVs, that makes it more suitable for the phylogenetic comparisons (Johnson, 2019). The other Env domains used for its characterization and phylogenetic analysis are the immunosuppressive domain (ISD), the fusion peptide (FP), the heptad repeats (HR1 and HR2), and the CX(6)C motif(Chabukswar et al., 2023). The divergence in the *env* gene can be accounted by two main factors i.e. mutations and recombination. Mutations lead to modification in *env* nucleotide sequence, whereas recombination events involve the transfer of *env* portions across viral genomes. Recombination occurs between viral genomes present in the same host cell as they exchange one or more genetic portion(s), potentially leading to the emergence of new viral strains and viruses, expansion of viral lineages, increase in their virulence and pathogenicity, and resistance to antiviral drugs (Pérez-Losada et al., 2015). Mosaicism in the form of recombination or secondary integration in the *env* gene has been previously reported and tends to complicate the sequence-based HERV classification (Vargiu et al., 2016). Similarly, we also observed the presence of recombinant forms of HML2 in old world monkeys (OWMs) specifically in *Macaca mulatta* and *Macaca fascicularis* (Chabukswar et al., 2022).

Analyzing patterns of *env* evolution is an essential step to better understand the roles of ERVs and reconstructing the structure of the ancestral ERV viral genes can help in various functional analysis as the changes in the genes overtime has affected their functionality. Hence, in the present study we aimed to reconstructing the *env* gene from the different HERV groups by generating potentially full-length coding consensus and screening the *env* genes of ERVs in

the primate lineage searching for the patterns of diversity by implementing a similarity search strategy, followed by the study of their phylogeny and recombination pattern. With this approach, we generated a representative collection of 32 Envs for Class I, Class II and Class III HERVs that were used for the screening of representative species of *Catarrhini* (*Hominidae, Cercopithecinae* and *Colobinae*) and *Platyrrhini* parvorders. Overall, we provide a comprehensive description of the ERV *env* diversity in 43 primates species, showing the presence of several class II ERVs in the *Platyrrhini* parvorder that have not been previously reported, and detailing the presence of events of interclass and intraclass *env* recombination that significantly expand our understanding of retroviral evolution.

Methods

Generation of Representative HERV Envelope protein sequences

To begin with, we first collected the *env* sequences that have been previously reported in Table 4 and Table 6 of Vargiu et al.2016. Using the *env* consensus sequences (Vargiu et al., 2016) as a query, we further extracted additional nucleotide sequences for each HERV group from UCSC Genome Browser through the BLAT search function.

Secondly, group specific alignments were performed, using the Mafft algorithm (Katoh et al., 2018) of the Geneious software, for all the HERV *env* sequences collected from the above-mentioned resources. The *env* nt alignments were translated into forward frames in Geneious followed by extracting the portions from the translated regions from the alignments to avoid the stop codons. Lastly, the Env protein consensus sequences were generated for each group by aligning the extracted translated portions to the reference sequence to generate a complete representative Env sequence for each HERV group. These generated Env sequences were then aligned to the exogenous retroviruses of all classes. For this, we collected data from various databases for exogenous RVs and generated an exhaustive dataset of 58 envelope protein sequences, that included different strains as well as subtypes of exogenous retroviruses. This dataset included all the RVs from all the classes (i.e. alpha, beta, gamma, delta, epsilon, intermediate-Beta like, spuma and lentiviruses).

ERV Env Screening in Primate Genomes

Using the above mentioned reconstructed HERV Env sequences we screened the primate genomes using the BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi). We used reconstructed Envs as queries to perform group specific tBLASTn searches in *Hominidae*, *Cercopithecinae* and *Colobinae* families of the *Catarrhini* parvorder and *Platyrrhini* parvorder. All the genome assemblies were accesses from NCBI Assembly database (https://www.ncbi.nlm.nih.gov/assembly/). A 40% sequence identity and E-value of 1E-5 were used as filters in tBLASTn algorithm. The hits obtained from all the searches were aligned and refined by removing other repetitive elements present as secondary integrations. The refined significant hits obtained from the similarity search were used for further analysis.

Sequence Alignments and Phylogenetic Analysis

All the hits obtained from the tBLASTn search were filtered and were downloaded in fasta format. For each group, sequences whose length was less than 50% of the entire *env* gene were discarded. Followed by the refinements we performed group specific sequence alignments using MAFFT algorithm of Geneious software. These alignments were further used to carry out phylogenetic analysis with the IQTREE software (http://www.iqtree.org/) for Gamma-like, beta-like and spuma-like *envs*. The trees were inferred using maximum likelihood (ML) methods in IQ-TREE with a bootstrap of 1000 replicates. The trees generated were further visualized and annotated by the Figtree V1.4.3. software(http://tree.bio.ed.ac.uk/software/figtree/).

Recombination Events

After the phylogenetic analysis of gamma, beta and spuma-like ERVs, the recombination events were detected using RDP, Geneconv, Bootscan, Maxchi, Chimaera, Siscan and 3SEQ methods implemented in RDP4 software (http://web.cbio.uct.ac.za/~darren/rdp.html) (Martin & Rybicki, 2000). The default settings of the RDP4 software that were used are (i) only potential recombination events detected by four or more of the above-mentioned methods; (ii) sequences were treated as linear and (iii) window size of 30 variable nucleotide positions used for the RDP method. The breakpoint positions and the recombinant sequences indicated for the

potential recombination event were further confirmed through comparative genomics. The detected recombinant sequences were further used to perform BLAT search in the primate genomes available in UCSC Genome to confirm the position and location of recombinant sequence in the host genomes.

Results

Reconstruction of the Retroviral Envelope from Human ERVs

The initial step in understanding the evolutionary dynamics of ERVs – which encompasses their origin, proliferation within the host genome, and the mechanisms of inactivation and elimination – is to precisely reconstruct their viral genes. Reconstructing an ancestral prototype of ERVs is a very crucial step and needs to address two important factors i.e., the biases present in the ERV data used and the changes that had occured after the retrovirus was integrated into the host genome. Hence, with the purpose of providing prototype Env proteins for the defferent HERVs, we reconstructed the Env sequence of 32 HERV groups, including members of Class I (gamma- and epsilon-like, 20), II (beta-like, 10), and III (spuma-like, 2) (Table 1). To build these consensuses, we followed different approaches, we first extracted a total of 33 HERV elements that have been previously reported in Table 6 of Vargiu et al.2016, as the ones including the most intact *env* genes, belonging to the Class I groups HERV9, HERVFC, HERVH, HERVT, HERVW, HERVE, HERV3 and the Class II groups HML2 (18 out of 33 sequences), HML5, HML6, and HML8. For the other HERV groups not included in this list, we referred to the Table 4 of Vargiu et al., 2016, which provides a summary of 39 canonical HERV clades found in the human genome (GRCh37/hg19). More 190 sequences were extracted from the additional files of Vargiu et al., 2016 for which we generated a bed file and extracted them with the flanking of 200nt at both 3’ and 5’ ends using the table browser tool in the genome browser (Supplementary file1(S1)). These 190 sequences initially extracted were divided into canonical and non-canonical forms and further aligned group-specifically, to understand if any of the sequence codes for complete Env protein. Further, to have some sequences to be used as a reference, we extracted 19 consensus sequences from Dfam (nt and aa both), and 17 reported HERV-Env proteins from UniProt for which the accession numbers

are provided in Supplementary file 2 (S2). The sequences used to perform alignments for each HERV group are described in detail in Supplementary file 3 (S3), which were then aligned to references sequences taken from Dfam, NCBI and Uniprot sequences (Supplementary file2 (S2)).

For each HERV group alignment, we translated the nt sequences in 3 forward frames to check the presence of stop codons using the Geneious software. From these translated frames we selected the portions of sequences based on two conditions: the sequence should not have any stop codons and should be in line with the reference sequences that can be either Dfam sequence, Uniprot or the consensus sequences taken from the Supplementary files of Vargiu et al. 2016. The selected translated sequence portions were then extracted and realigned with reference sequences thus generating a consensus sequence for the HERV Envs. Therefore, we reconstructed a total of 32 Envs for HERVs, which includes 17 Env for Class I i.e. gammaretroviruses, 13 Env for Class II i.e betaretroviruses and 2 Env for Class III i.e. spumaretroviruses (Table1, Supplementary file1(S4)). A schematic representation of the Env reconstruction process if shown in Figure 1.

Since the Env reconstruction was performed based on the data available for human ERVs, to confirm the accuracy of reconstruction by its relatedness to the other RVs, we generated Env protein phylogeny of reconstructed Envs with 69 Env proteins of exogenous retroviruses (Figure 2). As expected, we observed a class specific clustering of the reconstructed Envs with that of exogenous RVs, except for HERVS which was clustered with Class I RVs and not with the Class III RVs, indicating that HERVS has a Class I Env that is related to PRIMA41, PABL and HERV1_Artodact, as also previously reported (Vargiu et al., 2016).

Distribution of ERVs in Primate Genomes

To track the spreading of the above HERV groups across primates' evolution and evaluate their conservation, we screened a total of 43 species of primates to identify the *env* genes of class I, class II and class III ERVs. Briefly, we performed a group specific tBLASTn analysis search for HERV Env consensus sequence in each primate genomes. From the tBLASTn similarity search, we obtained 1374 hits for class I i.e. gamma-like *env*, 873 hits for class II i.e. beta-like *env* and 96 for class III i.e. spuma-like *env*. For all the three classes, we generated nucleotide alignments and performed maximum likelihood (ML) phylogenies using IQTREE software with a bootstrap of 1000 replicates.

For the gamma-like *env* 1374 hits the best fit model generated for the phylogeny using the IQTREE was TVM+F+R4 (Figure 3a, Supplementary file 2). A group-specific clustering of the *env* genes was observed in the phylogeny with the major grouping of HERVHF supergroup, that comprises of HERVH, HERVFA, HERVFB and HERVFC members. The supergroup HERVHF is reported to be integrated into the *Catarrhini* parvorder around 30-45 million years ago (mya). A widespread distribution of this supergroup was observed in *Homo sapiens, Pan troglodytes, Gorilla gorilla, Pongo abelii, Nomascus leucogenys, Papio Anubis, Papio cynocephalus, Chlorocebus sabaeus, Macaca mulatta, Macaca fascicularis, Macaca thibetana, Colobus angolensis, Macaca nemestrina, Rhinopithecus bieti, Rhinopithecus roxellana Trachypithecus francoisi, Piliocolobis tephrosceles, Callithrix jacchus, Aotus nancymaae, Saimiri boliviensis, Cebus imitator* (Figure 3a, Supplementary file 2). Similarly, we also observed wide distribution of ERVT, ERVW, and ERVE and ERVFRD followed by some small clusters of ERVH48, ERV9 and ERVIP. Additionally, we noticed distinct occurrences and clusters of HERV15, along with HERVR, ERVV1, ERVV2 and PABL, despite their exclusion from the blast analysis. This indicates that the derived Env consensus sequences effectively represent other related members of the supergroup that were not initially included in the collection (Figure 3a).

Likewise, phylogenetic analysis was also performed for 873 beta-like *env* hits obtained from tBLASTn. The best fit model for the phylogeny obtained was TVM+F+I+R3, with the major grouping obtained for HML2, which was further divided into two types i.e. type I (Np9) and type II (Rec) (Cadeddu et al., 2013; Chabukswar et al., 2022; Subramanian et al., 2011). In fact, the HML2 members are divided into type I and type II due to the presence of alternative splicing variants of *env* thus producing Np9 (type I) and Rec (type II), differing in the presence and absence of a 292bp deletion in the *env* gene, respectively (Chabukswar et al., 2022; Heyne et al., 2015; Subramanian et al., 2011) (Figure 3b, Supplementary file 3). The HML2 group is present in abundance followed by HML8 and HML10. The complete HML supergroup was observed in *Homo sapiens, Pan troglodytes, Pan paniscus, Gorilla gorilla, Gorilla gorilla gorilla, Hylobates moloch, Nomascus leucogenys Macaca nemestrina, Macaca fascicularis, Macaca mulatta, Macaca thibetana, Cercocebus atys, Chlorocebus sabaeus, Papio anubis, Papio hamadryas, Theropithecus gelada, Madrillus sphinx, Rhinopithecus roxellana, Rhinopithecus bieti, Presbytis melalophos, Trachypithecus cristatus, Piliocolobus tephrosceles*. These species belong *Hominidae, Cercopithecinae* and *Colobinae* subfamilies within *Cercopithecidae* family (Figure 3b, Supplementary file 3). A very few sequences were obtained in the *Callithrix jacchus* that belong to *Platyrrhini* parvorder. Other than clusters for

HML2 Np9 and Rec, a cluster of HML2 Rec was observed which included sequences that are annotated as ERVK-7, ERVK-19 and ERVK-25 as the *env* of these sequences were observed to be different from HML2 Rec. The blast hits obtained in HML4 (HERVK-13) in *Callithrix jacchus* were clustered together with the ERVK-7, ERVK-19 and ERVK-25 group and hence to confirm the annotations, using ERVK-7, ERVK-19 and ERVK-25 sequences as query we performed the BLAT search in *Platyrrhini* genomes available on genome browser i.e. Marmoset (*Callithrix jacchus*) and Squirrel monkey (*Saimiri boliviensis*). Previous studies reported that the emergence of the HML2 group occurred approximately 30 mya i.e. in the *Catarrhini* parvorder and hence are found in abundance in the primates belonging to *Hominidae*, *Cercopithecinae* and *Colobinae* (Subramanian et al., 2011), but while performing the BLAT search, we obtained 6 copies of HML2 annotated as HERVK in the squirrel monkey genome while no HML2 sequences were detected in *Callithrix jacchus* (Table 2). Interestingly, previous studies that focuses on the characterization and identification of HERVs in the primate genome also reported the presence of HML supergroup in the *Catarrhini* parvorder (Grandi et al., 2017, 2021; Scognamiglio et al., 2023) and by far only HML5 has been known to be the most ancient member of the HERVK supergroup (Lavie et al., 2004), as its presence has been previously reported in marmoset and squirrel monkey. In this study, we identified the presence of members of HERVK supergroup i.e. HML1, HML2, HML4, HML8, HML9 and HML10 in the squirrel monkey genome that has never been previously reported (Table 2). The sequences identified are in few cases full-length ERV copies while in most of the cases are of very short length or fragmented. These results indicates two possible hypothesis for the presence of HML1, HML2, HML4, HML8, HML9 and HML10 in squirrel money is that either the onset of these HMLs began even before 30 mya i.e. before the split between New World Monkeys (NWM) and Old World Monkeys (OWM), that occurred around 43 mya or they got fixed in to the genome as a results of cross-species transmission of retroviruses after the split.

When spuma-like *envs* were investigated only 96 hits were obtained. An ML phylogenetic tree with the best fit model of K3Pu+F+R3 was generated with the bootstrap of 1000 replicates (Figure 3c, Supplementary file 4). While performing tBLASTn search for both HERVL and HERVS, hits were mostly obtained for gamma-like ERVs, specifically for ERV3-1, ERVFC-1, ERVFRD and ERV-W. These results were in line with our previous studies which states that HERVL is most similar to ERV3-1 in hyrax, tenrec, armadillo, alligator and turtle (Bénit et al., 1997; Cordonnier et al., 1995; Vargiu et al., 2016). The hits were further aligned and refined by excluding short elements and removing the sequence of other repetitive elements acquired by secondary integrations. The phylogenetic analysis performed for these groups showed that

a major distribution took place in *Homo sapiens* followed by *Macaca mulatta* and other primate genomes Apart from the group specific *env* clustering, we also observed some small clusters hinting towards the presence of recombinant ERVs, for example, the grouping of HML4 for *Callithrix jacchus* with HML2 in the beta-like *env* phylogeny.

Evidence of Recombination in the ERVs' Env

An important factor for the diversity of the *env* gene are recombination events, since recombination is a ubiquitous process and may give rise to new variants for the existing viruses, possibly accounting in changes of the original tropism. Various recombination mechanisms are known to occur in the ERVs such as template switching, *env* acquisition, *env* degradation, cross-species transmission, and ectopic recombination between homologous sequences (Chabukswar et al., 2023; Pérez-Losada et al., 2015). Such events can greatly increase the diversity among the ERVs and complicate their classification. Hence, we aimed to detect the potential recombination events in our gamma, beta and spuma-like *env* alignments. To this purpose, we used RDP4 software, which implements various inbuilt methods to detect recombination events, which are confirmed if four or more methods give a positive signal. The software detected multiple events of recombination in the gamma and beta alignment files generated in the present study (Figure 4a). Since the spuma-like *env* alignment includes only two groups i.e. HERVL and HERVS, and no recombination among them was detected, the spuma-like *env* were aligned with gamma-like *env*, to check the presence of recombination among the two classes. The analysis showed recombination events in HERVL (Figure 4a), but not HERVS. To confirm the presence of these recombinant forms in the primate genomes, we performed a BLAT search in the UCSC Genome Browser using the recombinant sequences detected by the RDP software as a query. While performing the BLAT search, we considered all the primate genomes available in Genome Browser (n=15) to also trace back the initiation of these recombination events along primates' evolution (Figure 4b).

The recombination events detected in the gamma-like *env* alignment are ERVH-ERVE-MER11C, ERV15-LTR12, ERV15-MER11A, and PRIMA41-LTR5/LTR18. The first event was detected in ERVH, in which ERVE was present in the SU region while MER11C that is an LTR of HML8 group was present in the TM region of the *env* (Figure 4a). This event was only identified in 4 genomes i.e. humans (*Homo sapiens*, 8 copies), bonobo (*Pan paniscus*, 4 copies), chimpanzee (*Pan troglodytes*, 6 copies), and gorilla (*Gorilla gorilla*, 5 copies)

(Supplementary file 5), suggesting a very recent occurrence (i.e. less than 9 mya) (Figure 4b). It can be said that this recombinant form of ERVH emerged possibly due to the co-packaging of ERVH, ERVE and HML8 and thus the recombinant form ERVH-ERVE-MER11C got fixed in the genome of human, chimpanzee, gorilla, and bonobo with the very few copies. Other recombination events were observed in ERV15 and PRIMA41, in which a LTR was detected in the TM region of the *env* gene. The ERV15-LTR12/MER11A is detected in 12 primate genomes i.e. *Homo sapiens, Pan paniscus, Pan troglodytes, Gorilla gorilla, Pongo abelii, Nomascus leucogenys, Chlorocebus sabaeus, Macaca fascicularis, Macaca mulatta, Papio hamadryas, Nasalis larvatus* and *Rhinopithecus roxellana,* while the PRIMA41-LTR15/LTR18 that involves the LTR of ERV15 and ERV3 groups, respectively, was observed in only 6 genomes i.e. *Homo sapiens, Pan paniscus, Pan troglodytes, Gorilla gorilla* and *Pongo abelii* (Figure 4b). Of note, these two events were present as a single copy in each genome in just one chromosome, as indicated in supplementary file 5. These single events might have occurred due to the template switching mechanism during the reverse transcription process and eventually got fixed in the chromosome. In the template switching mechanism, the two copies of RVs are co-packaged into the virus particle, a template switch occurs as RT is estimated to dissociate from one template to another resulting in intermolecular i.e. homologous or nonhomologous recombination or intramolecular i.e. deletions or insertions (Delviks-Frankenberry et al., 2011). As in the case of the recombinants detected in the present study, a template switch could have occurred between the highly conserved TM region of the *env* of one ERV and the LTR of the other ERV due to some sequence similarity, resulting in the recombination, where an LTR is inserted into the TM region of the *env* gene. According to the mechanism of template switching, the LTR found within the *env* gene has the features of a viral RNA genome LTR (i.e. not with the complete U5-R-U3 structure that characterizes the proviral ones), as indeed it is observed in the sequences.

Similar events were detected in beta-like *env* alignment, where the LTRs are detected in the TM region of the HML *env* gene. The events detected were HML10-LTR13, HML2-LTR13 and HML2-MER11A/MER11B (Figure 4a). Single chromosomal recombination event was detected for HML10-LTR13 and HML2-LTR13, where the presence of LTR13 which is the LTR of HML4 group was detected in the HML10 and HML2 *env* gene in *Homo sapiens, Pan paniscus, Pan troglodytes, Gorilla gorilla, Pongo abelii, Nomascus leucogenys, Chlorocebus sabaeus, Macaca fascicularis* and *Rhinopithecus roxellana* (Figure 4b). Similarly to ERV15, these events are also present as the single copies in the primate genomes. While on the other hand, HML2-MER11A/MER11B is present in abundance in the genomes of *Cercopithecinae*

and *Colobinae* subfamilies and only one species of *Hominidae*. In the *Nomascus leucogenys* genome, 10 copies were detected followed by 17 copies in *Chlorocebus sabaeus*, 33 copies in *Macaca fascicularis*, 75 in *Macaca mulatta*, 18 in *Papio anubis*, 6 in *Nasalis larvatus*, and 23 in *Rhinopithecus roxellana*. The HML8 group contains three types of LTR i.e. MER11A, MER11B and MER11C, of which the presence of MER11A in HML2 *env* was detected in *Chlorosebus sabaeus, Macaca fascicularis, Macaca mulatta, Papio hamadryas, Nasalis larvatus* and *Rhinopithecus roxellana* while MER11B was detected in and *Nomascus leucogenys, Chlorocebus sabaeus, and Macaca fascicularis* (Figure 4b) suggesting that the template switching of HML2 happened with two different variants of HML8, as HML8 is divided into three variants based on the LTR type in the previous study(Scognamiglio et al., 2023). Since this event is absent in most of the species of *Hominidae* except *for Nomascus leucogenys*, it can be said that this event occurred after the split between *Colobinae* subfamily and *Platyrrhini*, simultaneously with the abundant spread of HML2. All the recombination detected in the HML group are hypothesized to be emerged due to the template switching mechanism as in the gamma-like HERVs and hence were fixed in the primate genomes.

Lastly for the spuma-like *env*, we detected two single chromosomal events for ERVL i.e. ERVL-MER11C and ERVL-LTR13 (Figure 4a): while on one hand presence of MER11C in the *env* was observed in *Chlorosebus sabaeus, Macaca fascicularis, Macaca mulatta, Papio hamadryas, Nasalis larvatus*, and *Rhinopithecus roxellana*, LTR13 was detected in *Homo sapiens, Pan paniscus, Pan troglodytes, Gorilla gorilla* and *Pongo abelii* (Figure 4b). As mentioned for all the above recombination events, ERVL seems to have also followed the same mechanism of template switching for both the LTRs.

Finally, to infer the functional effect of the presence of LTRs in the *env* region of ERVs, we performed the conserved domain (CD) search in NCBI. We first used the reconstructed Env to identify the domains present in the *env* and then performed the same search with the recombinant forms. In Table 3, all the CD are listed and their position in the *env* obtained from NCBI CD for the non-recombinant forms together with its presence or absence in the recombinant forms. Of note, a consistent pattern of loss of CD is observed for the ERVs which have the LTR in the *env* gene. The domains are either completely absent or truncated due to the LTR, leading to the loss of function in the corresponding protein. As shown in the Table 3, the only domain identified for ERVH is TLV coat, but other domains have been identified for the recombinant form of ERVH-ERVE-MER11A suggesting an *env* acquisition event for the ERVH group by the co-packaging with ERVE. From the overall results, we can say that the LTRs of HML8 (MER11A, MER11B and MER11C) are the most frequently found in the *env*

gene of different ERVs and are associated with recombination in primate genomes. Interestingly, the recombination events involving the presence of LTRs in the *env* gene are mostly found in the *Catarrhini* parvorder, suggesting that these events mostly occurred during the process of endogenization of the ERVs, leading to the intragenomic spreading of the recombinant forms. These events are usually associated with the Env loss of function, with phenomena termed "*env* snatching" or "*env* degradation". Such events can further lead to loss of the *env* gene followed by the intragenomic spreading of the *env*-less ERVs in the genomes. Interestingly, this is a widespread phenomenon among the retroviral integration in human lineage but have not been previously reported in the other primate genomes. The exception here is the ERVH, which is associated with the "*env* acquisition" i.e. the gain of *env* function from ERVE. Thus, such recombination events may lead not only to the generation of new retroviral elements but also to eventual spreading of ERVs throughout the genomes and therefore the co-evolution of ERVs in the primate lineage.

Discussion

ERVs are "fossils" of the ancient exogenous RVs that have become extinct, but they still carry out certain functional role under specific circumstances (Sverdlov, 2000). Indeed, the genomic integration of ERVs and its widespread dispersion has led to the domestication of some ERV proteins for useful functions. Some of their properties are linked to the conservation of an open reading frame (ORF), to their putative biochemical function and to their expression on appropriate cell or tissue. One such prime example is "syncytins" that expressed in human placenta that play a crucial role in human development. These syncytins have been domesticated in different mammalian species that co-opted different retroviral *env* genes for placentation. The two well characterized syncytins are syncytin-1 encoded by HERVW (ERVWE1) (Blond et al., 2000; Grandi & Tramontano, 2017) and syncytin-2 encoded by HERVFRD-1 (ERVFRD-1) (Rote et al., 2004). Another recurrent ERV co-option is acting as a restriction factor and interfering in steps of exogenous RV infection. In several vertebrates, the ERV-encoded Env protects the host cell from the viral entry by competing with the Env of exogenous RVs, a phenomenon similar to superinfection resistance(Frank & Feschotte, 2017). The ERV Env studies are mostly focused on finding the possible role in physiological, and sometime pathological, conditions, but limited studies have been performed to understand

ERVs' evolutionary dynamics and residual function. For this, a focus on how various factors may affect the Env divergence is required.

Reconstructing the ERVs' genes might help in understanding the evolutionary dynamics of ERVs, but can also have several functional significances as well such as host-virus interactions, ERVs contribution in disease development by understanding their functions can help in elucidating potential role of ERVs in various pathological conditions, as well as its utilization as antivirals, gene therapy vectors, etc. The reconstructed Env might aid in broadening the understanding and discovering various aspects of ERVs (Johnson, 2019) Hence, with the intention of studying the Env divergence, we firstly reconstructed 32 ancestral Env prototypes of ERVs that can help in various functional analysis such as gene regulation and expression, pathogenicity, evolutionary dynamics, host-ERV interaction, etc. (Johnson, 2019). The screening of 43 primate genomes in *Catarrhini* (*Hominidae, Cercopithecinae, Colobinae*) and *Platyrrhini* parvorders using the reconstructed Env sequences for each HERV class demonstrate that the majority of ERV *env* sequences hits are obtained in the *Catarrhini* parvorder specifically in Hominidae Superfamily. In the phylogenetic analyses performed for each Class we observe a very group specific clustering of all the ERVs. Over the period of ERVs genomic persistence numerous insertions, deletions, mutations as well as loss of the *env* gene are observed, leading to biases in the phylogenetic analysis. As a result, we report some mixed clustering among the ERV classes. Apart from the 17 gamma-like *env*, we also observe hits for ERV15, HERVR, ERVV1, ERVV2and PABL *env* genes which were initially not included in the dataset, meaning that the generated consensus sequences represent very well the members of the same supergroup. Our investigation demonstrates that the superfamilies HERVHF and HERVW9 of gamma-like *env*, as well as HML2 group of beta-like *env* integrated in the primate lineage over 30-45 mya, before than previously reported. In fact, several studies have reported the presence of HML groups in the genomes of *Catarrhini* parvorder indicating its integration began after the split between *Catarrhini* and *Platyrrhini* i.e. ~30 mya, as the HML sequences were not detected in *Callithrix jacchus* genome except for HML5, which is known to be the oldest HML member among the supergroup (Lavie et al., 2004; Subramanian et al., 2011). Even though other HML members are not present in the *Callithrix jacchus* genome, they are present in *Saimiri boliviensis*, which is also the member of *Platyrrhini* parvorder. Thus, we confirm the presence of integrations from HML1, HML2, HML4, HML8, HML9 and HML10 groups in the *Platyrrhini*, having possibly a lower rate of amplification, as they are less numerous in this parvorder, while they were spread more abundantly after the split of *Catarrhini* and *Platyrrhini*. In contrary to gamma-like and beta-

like, we detected only 96 hits for spuma-like *env* and hence one potential hypothesis explaining the lower prevalence of spuma-like ERVs in the genome could be that they either infrequently integrate or have a shorter persistence compared to gamma or beta ERVs (Hayward et al., 2015). Overall, the present phylogenomic approach suggests the widespread presence of the gamma ERVs, indicating they may have been integrated in the host genomes through multiple sources, and demonstrates the onset of class II ERV even before the split between *Catarrhini* and *Platyrrhini* parvorders. A complex evolutionary trajectory as compared to the one of spuma-like ERVs.

Numerous studies have reported multiple recombination events in ERVs in the vertebrate lineage leading to several modifications throughout the endogenization process (Chabukswar et al., 2023). A remarkable example of recombinant provirus with contributions from different ERVs is represented by Harlequin elements, which have a complicated structure of LTR2-HERVE-MER57I-LTR8-MER4I-HERVI-HERVE-LTR2 (Vargiu et al., 2016).

To detect the presence of such recombination events in primates we performed recombination analysis and identified 8 potential recombination events i.e. ERVH-ERVE-MER11C, ERV-15-LTR12, ERV-15-MER11A, HML2-LTR13, HML2-MER11A, HML2-MER11B, HML10-LTR13, ERVL-MER11C and ERVL-LTR13 in gamma-like and beta-like and spuma-like *env* sequences. We confirmed these events using the genome assemblies available in Genome Browser, hence assuming their presence also in the other primate genomes of the superfamily. The potential mechanism due to which such events occur in the genomes is the co-packaging of different ERVs in a virus particle, followed by the template switch from one ERV to another during the reverse transcription process. As in case of ERVH-ERVE-MER11C, ERVH seems to acquire the *env* gene of ERVE group, leading to the generation of a recombinant ERV that seems to have retained the potentially coding ERVE Env while for the rest of the recombinant sequences, an LTR exchange with the TM region of the *env gene* is observed which causes the disturbance in the *env* gene thus eventually leading to loss of the *env* function of ERVs further promoting the intragenomic spreading of the *env*-less ERVs (Magiorkinis et al., 2012). Interestingly, such swapping of the *env* region is a widespread phenomenon often leading to the loss of the *env* function but such events can also possibly lead to emergence of different variants of the existing retrovirus, as in case of HML2-MER11A/MER11B, where it is predicted that different HML8 variants swapped their LTR with the TM region of the HML2 *env*, and thus spreading in abundance in the *Cercopithecinae* subfamily members and creating diversity in the HML2 group. The recombination between HML2-MER11A/B in macaque genomes was previously reported (Chabukswar et al., 2022; Williams et al., 2024), suggesting

that HML8 and HML2 RNAs were co-packaged and the HML8 region appeared in HML2 *env* region due to their recombination during the reverse transcription. At this regards, Williams and colleagues showed that the HML-8-derived region provided a novel, Rec-independent constitutive transport element (CTE), which replaced the ancestral Rec–RcRE export system (Williams et al., 2024). In addition to the HML2-MER11A/B recombination, various similar events have been reported in this study for gamma, beta and spuma-like *envs*, having the LTRs in the TM region of the *env*. Hence, it is worth considering that such events were not limited to the HML2 group and occurred in several primate species, thus having implications on the biology of the different viruses.

Overall, the recombination events identified in the present study provide insights into the prevalence of similar recombinant forms within the host genome, underlying the significant role of recombination as a driving force behind the ERVs diversification.

Conclusion

ERVs are the remnants of ancestral exogenous RVs whose origin and diversity are still not completely understood and are a great tool to study the coevolution of host and RVs. The present study provides a phylogenomic perspective of the retroviral *env* gene diversity. With this study approach we provide the evidence for inter as well as intra group recombination events. Present results demonstrate the prevalence of ERVs across the primate lineage as well as the diversification of the retroviral Env through the process of recombination over the evolutionary timescale. Overall, this analysis gives further knowledge in studying the host-virus interactions on a larger scale.

References

Bénit, L., De Parseval, N., Casella, J. F., Callebaut, I., Cordonnier, A., & Heidmann, T. (1997). Cloning of a new murine endogenous retrovirus, MuERV-L, with strong similarity to the human HERV-L element and with a gag coding sequence closely related to the Fv1 restriction gene. *Journal of Virology*, *71*(7). https://doi.org/10.1128/jvi.71.7.5652-5657.1997

Blomberg, J., Benachenhou, F., Blikstad, V., Sperber, G., & Mayer, J. (2009). Classification and nomenclature of endogenous retroviral sequences (ERVs). Problems and recommendations. In *Gene* (Vol. 448, Issue 2). https://doi.org/10.1016/j.gene.2009.06.007

Blond, J.-L., Lavillette, D., Cheynet, V., Bouton, O., Oriol, G., Chapel-Fernandes, S., Mandrand, B., Mallet, F., & Cosset, F.-L. (2000). An Envelope Glycoprotein of the

Human Endogenous Retrovirus HERV-W Is Expressed in the Human Placenta and Fuses Cells Expressing the Type D Mammalian Retrovirus Receptor. *Journal of Virology*, *74*(7). https://doi.org/10.1128/jvi.74.7.3321-3329.2000

Cadeddu, M., Vargiu, L., Rodriguez-Tomé, P., Sperber, G. O., Blomberg, J., & Tramontano, E. (2013). Identification and analysis of HML2 sequences in human genome assembly GRCh37/hg19. *Retrovirology*, *10*(S1). https://doi.org/10.1186/1742-4690-10-s1-p9

Chabukswar, S., Grandi, N., Lin, L. T., & Tramontano, E. (2023). Envelope Recombination: A Major Driver in Shaping Retroviral Diversification and Evolution within the Host Genome. In *Viruses* (Vol. 15, Issue 9). https://doi.org/10.3390/v15091856

Chabukswar, S., Grandi, N., & Tramontano, E. (2022). Prolonged activity of HERV-K(HML2) in Old World Monkeys accounts for recent integrations and novel recombinant variants. *Frontiers in Microbiology*, *13*. https://doi.org/10.3389/fmicb.2022.1040792

Cordonnier, A., Casella, J. F., & Heidmann, T. (1995). Isolation of novel human endogenous retrovirus-like elements with foamy virus-related pol sequence. *Journal of Virology*, *69*(9). https://doi.org/10.1128/jvi.69.9.5890-5897.1995

Delviks-Frankenberry, K., Galli, A., Nikolaitchik, O., Mens, H., Pathak, V. K., & Hu, W. S. (2011). Mechanisms and factors that influence high frequency retroviral recombination. *Viruses*, *3*(9). https://doi.org/10.3390/v3091650

Frank, J. A., & Feschotte, C. (2017). Co-option of endogenous viral sequences for host cell function. In *Current Opinion in Virology* (Vol. 25). https://doi.org/10.1016/j.coviro.2017.07.021

Gifford, R. J., Blomberg, J., Coffin, J. M., Fan, H., Heidmann, T., Mayer, J., Stoye, J., Tristem, M., & Johnson, W. E. (2018). Nomenclature for endogenous retrovirus (ERV) loci. In *Retrovirology* (Vol. 15, Issue 1). https://doi.org/10.1186/s12977-018-0442-1

Grandi, N., Cadeddu, M., Pisano, M. P., Esposito, F., Blomberg, J., & Tramontano, E. (2017). Identification of a novel HERV-K(HML10): Comprehensive characterization and comparative analysis in non-human primates provide insights about HML10 proviruses structure and diffusion. *Mobile DNA*, *8*(1), 1–18. https://doi.org/10.1186/s13100-017-0099-7

Grandi, N., Pisano, M. P., Pessiu, E., Scognamiglio, S., & Tramontano, E. (2021). HERV-K(HML7) integrations in the human genome: Comprehensive characterization and comparative analysis in non-human primates. *Biology*, *10*(5). https://doi.org/10.3390/biology10050439

Grandi, N., & Tramontano, E. (2017). Type W human endogenous retrovirus (HERV-W) integrations and their mobilization by L1 machinery: Contribution to the human transcriptome and impact on the host physiopathology. *Viruses*, *9*(7), 1–37. https://doi.org/10.3390/v9070162

Hayward, A. (2017). Origin of the retroviruses: when, where, and how? In *Current Opinion in Virology* (Vol. 25). https://doi.org/10.1016/j.coviro.2017.06.006

Hayward, A., Cornwallis, C. K., & Jern, P. (2015). Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(2). https://doi.org/10.1073/pnas.1414980112

Henzy, J. E., & Johnson, W. E. (2013). Pushing the endogenous envelope. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1626). https://doi.org/10.1098/rstb.2012.0506

Heyne, K., Kölsch, K., Bruand, M., Kremmer, E., Grässer, F. A., Mayer, J., & Roemer, K. (2015). Np9, a cellular protein of retroviral ancestry restricted to human, chimpanzee and gorilla, binds and regulates ubiquitin ligase MDM2. *Cell Cycle*, *14*(16), 2619–2633. https://doi.org/10.1080/15384101.2015.1064565

Johnson, W. E. (2015). Endogenous Retroviruses in the Genomics Era. In *Annual Review of Virology* (Vol. 2). https://doi.org/10.1146/annurev-virology-100114-054945

Johnson, W. E. (2019). Origins and evolutionary consequences of ancient endogenous retroviruses. In *Nature Reviews Microbiology* (Vol. 17, Issue 6). https://doi.org/10.1038/s41579-019-0189-2

Katoh, K., Rozewicki, J., & Yamada, K. D. (2018). MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, *20*(4). https://doi.org/10.1093/bib/bbx108

Lavie, L., Medstrand, P., Schempp, W., Meese, E., & Mayer, J. (2004). Human Endogenous Retrovirus Family HERV-K(HML-5): Status, Evolution, and Reconstruction of an Ancient Betaretrovirus in the Human Genome. *Journal of Virology*, *78*(16). https://doi.org/10.1128/jvi.78.16.8788-8798.2004

Mager, D. L., & Stoye, J. P. (2015). Mammalian Endogenous Retroviruses. *Microbiology Spectrum*, *3*(1). https://doi.org/10.1128/microbiolspec.mdna3-0009-2014

Magiorkinis, G., Gifford, R. J., Katzourakis, A., De Ranter, J., & Belshaw, R. (2012). Env-less endogenous retroviruses are genomic superspreaders. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(19), 7385–7390. https://doi.org/10.1073/pnas.1200913109

Martin, D., & Rybicki, E. (2000). RDP: Detection of recombination amongst aligned sequences. *Bioinformatics*, *16*(6). https://doi.org/10.1093/bioinformatics/16.6.562

Pancino, G., Ellerbrok, H., Sitbon, M., & Sonigo, P. (1994). Conserved framework of envelope glycoproteins among lentiviruses. In *Current Topics in Microbiology and Immunology* (Vol. 188). https://doi.org/10.1007/978-3-642-78536-8_5

Pérez-Losada, M., Arenas, M., Galán, J. C., Palero, F., & González-Candelas, F. (2015). Recombination in viruses: Mechanisms, methods of study, and evolutionary consequences. In *Infection, Genetics and Evolution* (Vol. 30). https://doi.org/10.1016/j.meegid.2014.12.022

Rote, N. S., Chakrabarti, S., & Stetzer, B. P. (2004). The role of human endogenous retroviruses in trophoblast differentiation and placental development. *Placenta*, *25*(8–9). https://doi.org/10.1016/j.placenta.2004.02.008

Scognamiglio, S., Grandi, N., Pessiu, E., & Tramontano, E. (2023). Identification, comprehensive characterization, and comparative genomics of the HERV-K(HML8) integrations in the human genome. *Virus Research*, *323*. https://doi.org/10.1016/j.virusres.2022.198976

Subramanian, R. P., Wildschutte, J. H., Russo, C., & Coffin, J. M. (2011). Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology*, *8*, 90. https://doi.org/10.1186/1742-4690-8-90

Sverdlov, E. D. (2000). Retroviruses and primate evolution. In *BioEssays* (Vol. 22, Issue 2). https://doi.org/10.1002/(SICI)1521-1878(200002)22:2<161::AID-BIES7>3.0.CO;2-X

Vargiu, L., Rodriguez-Tomé, P., Sperber, G. O., Cadeddu, M., Grandi, N., Blikstad, V., Tramontano, E., & Blomberg, J. (2016). Classification and characterization of human endogenous retroviruses mosaic forms are common. *Retrovirology*, *13*(1), 1–29. https://doi.org/10.1186/s12977-015-0232-y

Weiss, R. A. (2006). The discovery of endogenous retroviruses. In *Retrovirology* (Vol. 3). https://doi.org/10.1186/1742-4690-3-67

Weiss, R. A. (2016). Human endogenous retroviruses: Friend or foe? In *APMIS* (Vol. 124, Issues 1–2). https://doi.org/10.1111/apm.12476

Williams, Z. H., Imedio, A. D., Gaucherand, L., Lee, D. C., Mostafa, S. M., Phelan, J. P., Coffin, J. M., & Johnson, W. E. (2024). Recombinant origin and interspecies

transmission of a HERV-K(HML-2)-related primate retrovirus with a novel RNA transport element. *ELife*, *13*. https://doi.org/10.7554/eLife.80216

Figure legends

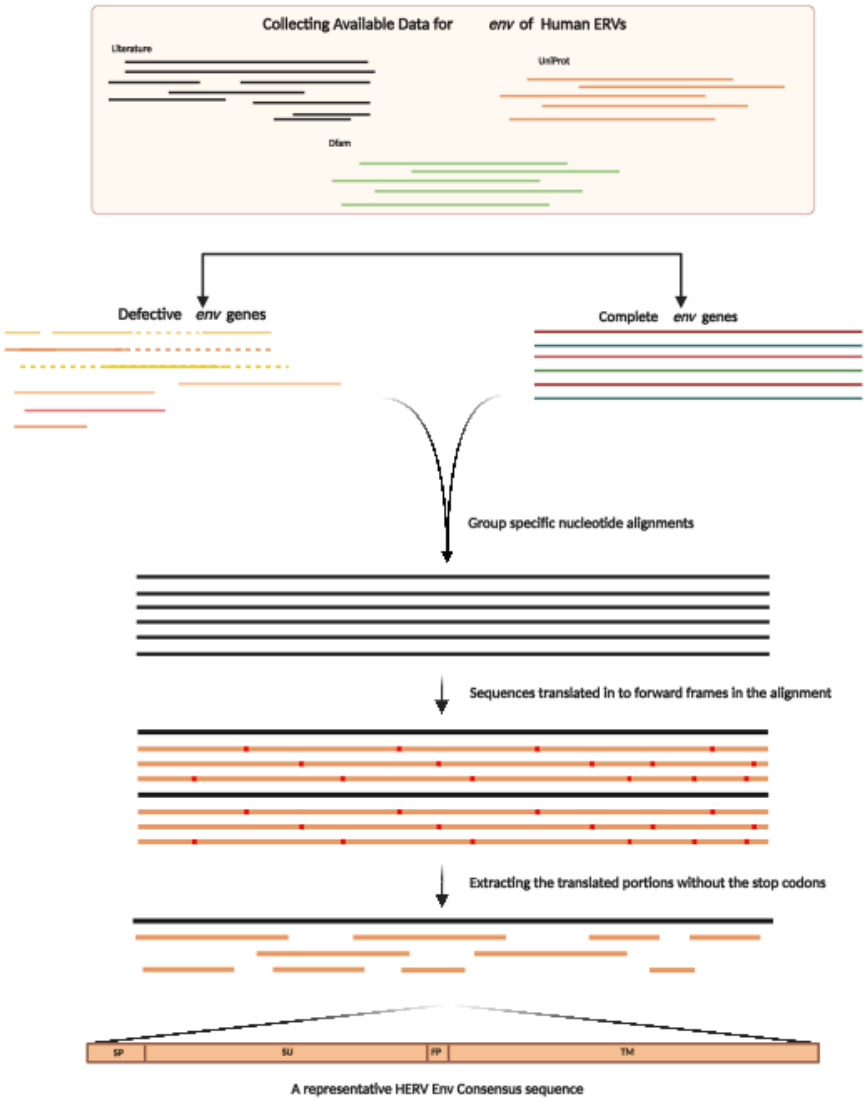


Figure 1: Illustration of the reconstruction process of HERV Env proteins. The first step was to extract *env* genes for both endogenous and exogenous retroviruses from Literature, Dfam, NCBI, Uniprot, and PDB (complete env sequence as well as defective *env* genes). Second, we performed group specific alignments for all the ERV sequences with Mafft and

further divided the sequences into subtypes based on the alignments, as in case of HERV3 and HML3. Then, we translated the sequences into three forward frames in the alignment and extracted only the translated portions that did not have any stop codons. Lastly, the extracted aa portions were aligned to the reference sequence of each group and hence, a reconstructed env sequence was generated for HERV groups as mentioned in Table 1.

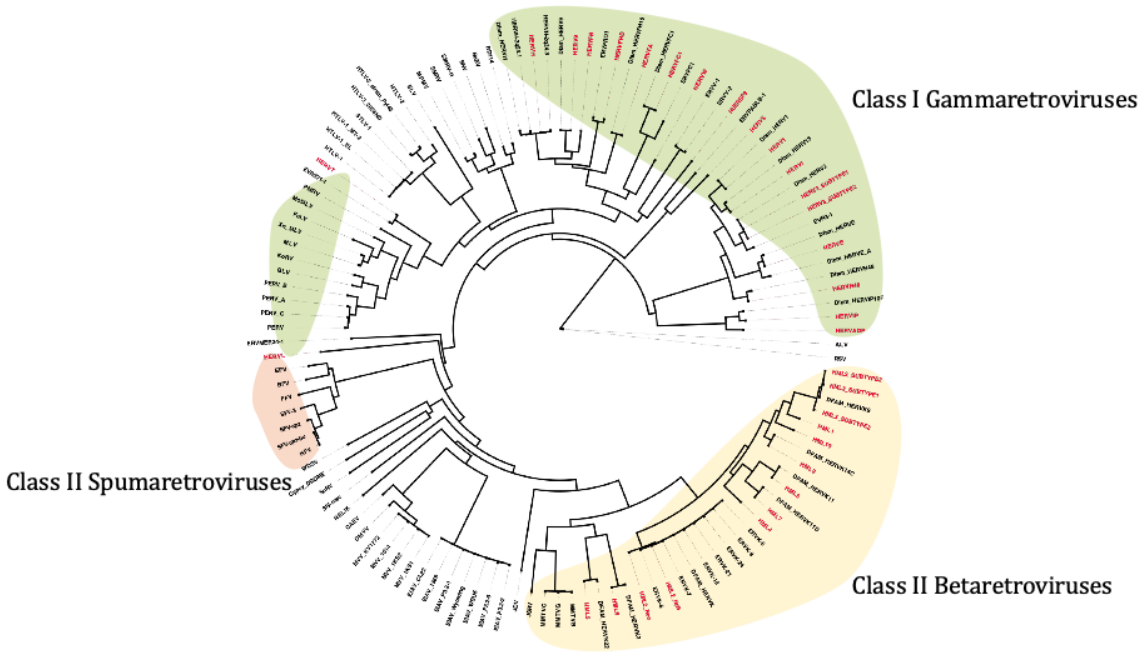


Figure 2: Phylogeny of the reconstructed Envs with the envelope protein of exogenous retroviruses. The phylogenetic tree was generated for the reconstructed HERV Env proteins with the 69 exogenous envelope proteins for Class I, Class II and Class III retroviruses. The class I exogenous and endogenous retroviruses are highlighted in green, class II in yellow and class III in orange color. The reconstructed Envs are presented in red color in the tree. The tree also contains RVs of other class with which no ERVs were clustered and hence are not highlighted.

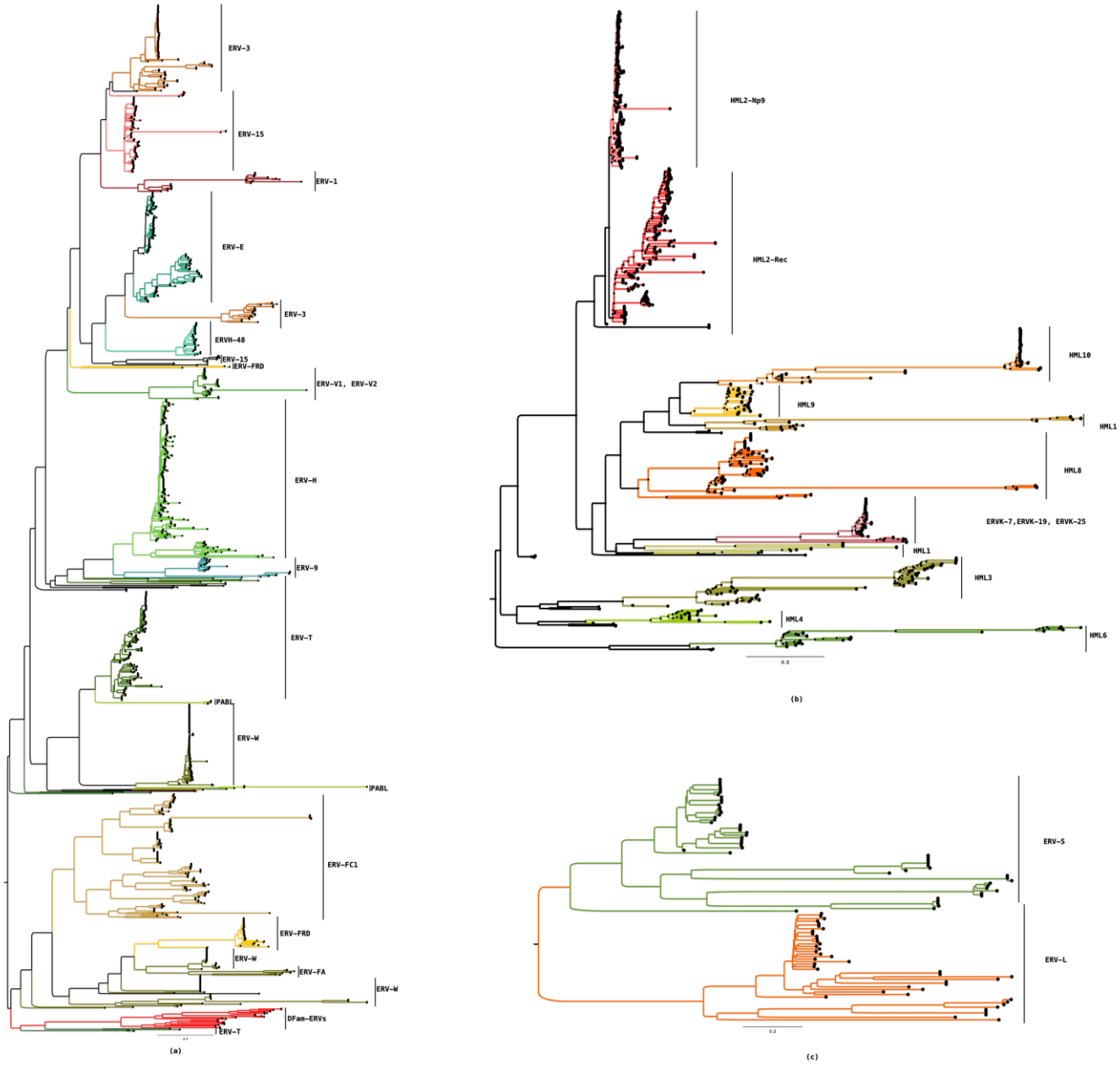


Figure 3: Phylogenetic analysis of the envelope hits obtained from the tBLASTn analysis for the three classes i.e. Class I (gamma-like *env*), Class II (beta-like *env*) and Class III (spuma-like *env*). The trees were generated using IQ-TREE software with the bootstrap of 1000 replicates. (a)Phylogeny for all the hits obtained for Class I ERVs with the best fit model of TVM+F+R4. (b) Phylogeny for all the hits obtained for Class II ERVs with the best fit model of TVM+F+I+R3 and (c) Phylogeny for all the hits obtained for Class III ERVs with the best fit model of K3Pu+F+R3. All the groups obtained from the tBLASTn hits and further clustered together with the phylogenetic analysis are labelled in the tree. Similar trees are provided as

supplementary files with the tip labels having all the annotations and the accession numbers for all the three classes (Supplementary file 2,3 and 4).

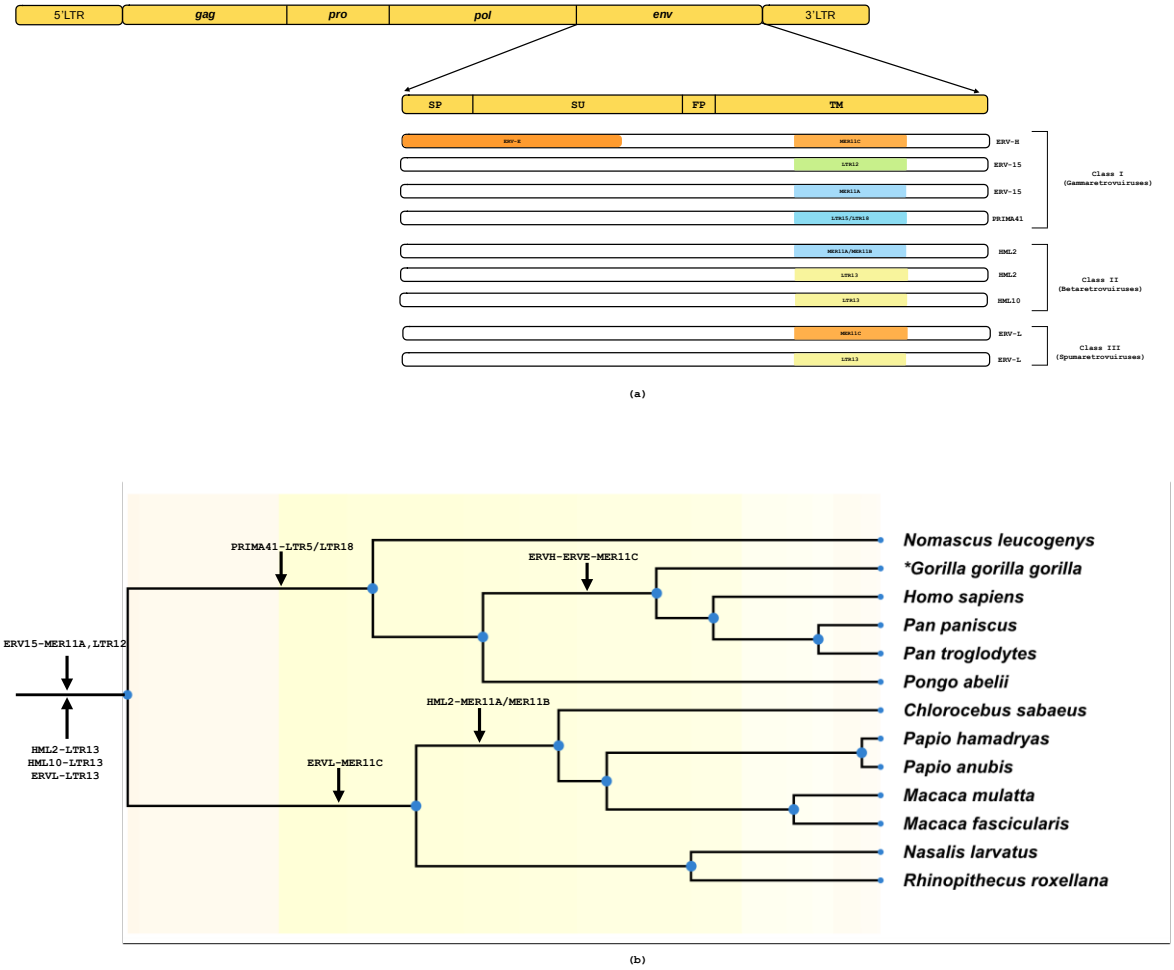


Figure 4: Detecting the recombination events in the ERV's *env* gene. (a) Schematic representation *env* recombination events. The representation indicates the presence of recombination in the ERVs' *env* sequence as well as its position in the *env*. The coordinates for these recombination events are provided in the supplementary file 5. (b) Phylogenetic representation of the initiation of the recombination event in the primate lineage. The phylogenetic tree was build using TimeTree of Mega software with only the *Catarrhini* genomes available in the Genome Browser. The recombinations events were tested doing the blat search in genome browser to detect its presence and to traceback the time when the recombination initiated. The genome browser contains the genome of *Gorilla gorilla* but the Timetree used the *Gorilla gorilla gorilla* genome to build the tree and hence if marked with *.

Tables

Table1: Dataset of the reconstructed HERV Envelope proteins

	Supergroup	Group	Subtypes	aa Length
Class I	MLLV	HERVT		623
	HERVERI	HERVE		671
		HERV3	Subtype1	666
			Subtype2	661
		HERV1		777
		HERVI		728
	HERVW9	HERVW		544
		HERV9		554
	HERVIADP	HERVIP		780
		HERVADP		672
	HERVHF	HERVH		581
		HERVH48		244
		HERVFA		561
		HERVFB		542
		HERVFC		570
	HERVFRDlike	HERVFRD		599
	HUERSP	HUERSP3		536
Class II	HERVK	HML1		667
		HML2	Type1(Np9)	588
			Type2(Rec)	684
		HML3	Subtype1	557
			Subtype2	465
			Subtype3	390
		HML4		657
		HML5		644
		HML6		639
		HML7		664
		HML8		776
		HML9		682
		HML10		625
Class III	HSERVIII	HERVL		460
		HERVS		420

Table2: Coordinates of HML supergroup in *Platyrrhini* (Squirrel Monkey Genome)

SUPERGROUP	HML	COORDINATES	STRAND	LENGTH
HERVK	HML1	JH378338:275937-278501	+	2564
		JH378138:15705974-15707323	-	1349
		JH378108:32563432-32571361	+	7929
		JH378224:2469781-2470931	-	1150
		JH378134:11672169-11673824	-	1655
	HML2	JH378134:10458979-10463351	-	4372
		JH378105:56691102-56691325	-	223
		JH378105:56692918-56693866	-	948
		JH378105:56695565-56696715	-	1150
		JH378105:56697213-56697652	-	439
		JH378138:15703105-15705969	-	2864
	HML4	JH378161:520931-521612	-	681
		JH378134:16151196-16152689	+	1493
		JH378134:7690429-7691554	-	1125
		JH378146:13346796-13353682	-	6886
		JH378137:7471310-7471738	-	428
		JH378106:35578339-35578392	+	53
	HML8	JH378141:3704815-3715095	+	10280*
		JH378112:24913201-24913256	-	55
	HML9	JH378336:397006-400978	-	3972
		JH378132:8543270-8546474	+	3204
		JH378113:32170092-32175530	-	5438
		JH378152:4554178-4561789	-	7611
		JH378362:166905-169335	-	2430
		JH378344:454424-459278	-	4854
		JH378332:381766-384506	-	2790
	HML10	JH378114:8677151-8678160	-	1009
		JH378114:15213478-15214880	+	1402
		JH378130:3582370-3583730	+	1360
		JH378233:1994071-1995016	-	945
		JH378110:37554414-37555367	-	953

Note: * indicates the truncated sequence

761

Table3: Detection of the conserved domains in the reference ERV sequences as well as recombinant sequences.

Recombinant ERV	Conserved Domains Features								
	HERV-K_REC Super family	HERVK_env_2	TLV Coat	Ebola-like_HR1-HR2 Superfamily					
				HR1-GP1 interface	HR2	Homotrimer interface	C1 binding site	CX(6,7)C Motif	Immuno-suppressive region
***ERV15**	-	-	1064-1258	1439-1552	1631-1669	1439-1669	1517-1528	1571-1598	1521-1569
ERV15-LTR12	-	-	-	√	-	√*	-	-	√
ERV15-MER11A	-	-	-	-	-	-	-	-	-
***ERVH**	-	-	825-1487	-	-	-	-	-	-
ERVH-ERVE-MER11C	-	-	-	√	-	√	√	√	√
***PRIMA41**	-	-	-	-	-	-	-	-	929-955
PRIMA41-LTR5/LTR18	-	-	-	-	-	-	-	-	-
***HML2**	55-315	346-849	-	1501-1659	1753-1878	1507-1878	-	-	-
HML2-LTR13		-	-	-	-	-	-	-	-
HML2-MER11A/MER11B	√	√		√	-	√*	-	-	-
***HML10**		157-636	-	1339-1437	-	1339-1497	-	-	1421-1447
HML10-LTR13		-	-	-	-	-	-	-	-
***ERVL**	-	-	455-1051	716-814	908-934	716-934	-	-	797-823
ERVL-MER11C	-	-	-	-	-	-	-	-	-
ERVL-LTR13	-	-	-	-	-	-	-	-	-

Note: √ indicates presence of the Domain; √* indicates the domain is truncated; - indicates absence of the domain; * indicates the non-recombinant ERV

Supplementary Files

Supplementary File1: Excel sheet for all the HERVs used for the Reconstruction of HERVs' Env (S1, S2 and S3). The S4 sheet contains Fasta File of gamma-like, beta-like and spuma-like reconstructed HERV Envelope sequences.

Supplementary File2: Complete phylogenetic tree with the tip labels for Gamma-like ERVs' env.

Supplementary File3: Complete phylogenetic tree with the tip labels for Beta-like ERVs' env.

Supplementary File4: Complete phylogenetic tree with the tip labels for Spuma-like ERVs' env.

Supplementary File5: Excel file for the Coordinates of the recombinant sequences in the 15 genome assemblies extracted from the BLAT search.

Data availability statement

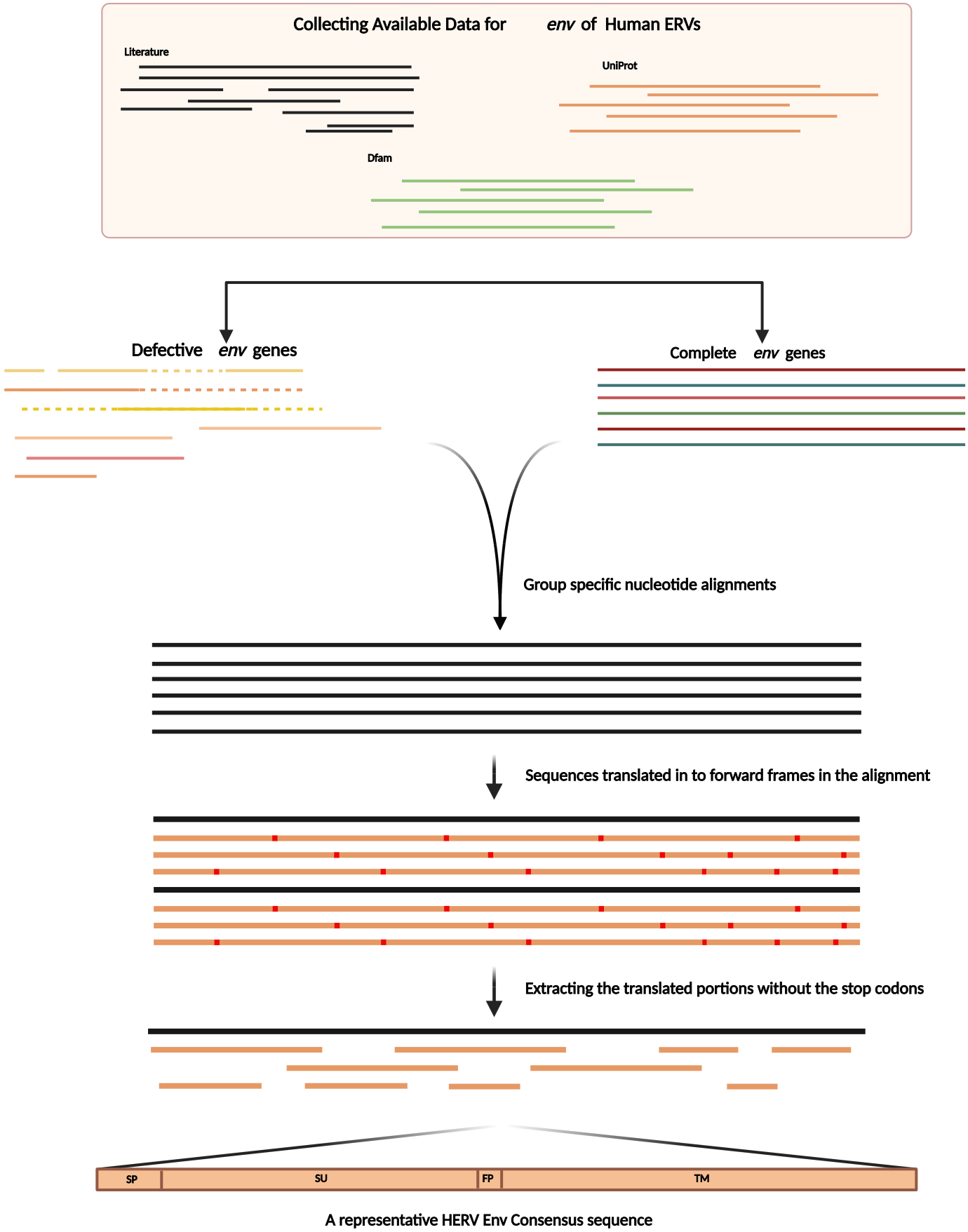
The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

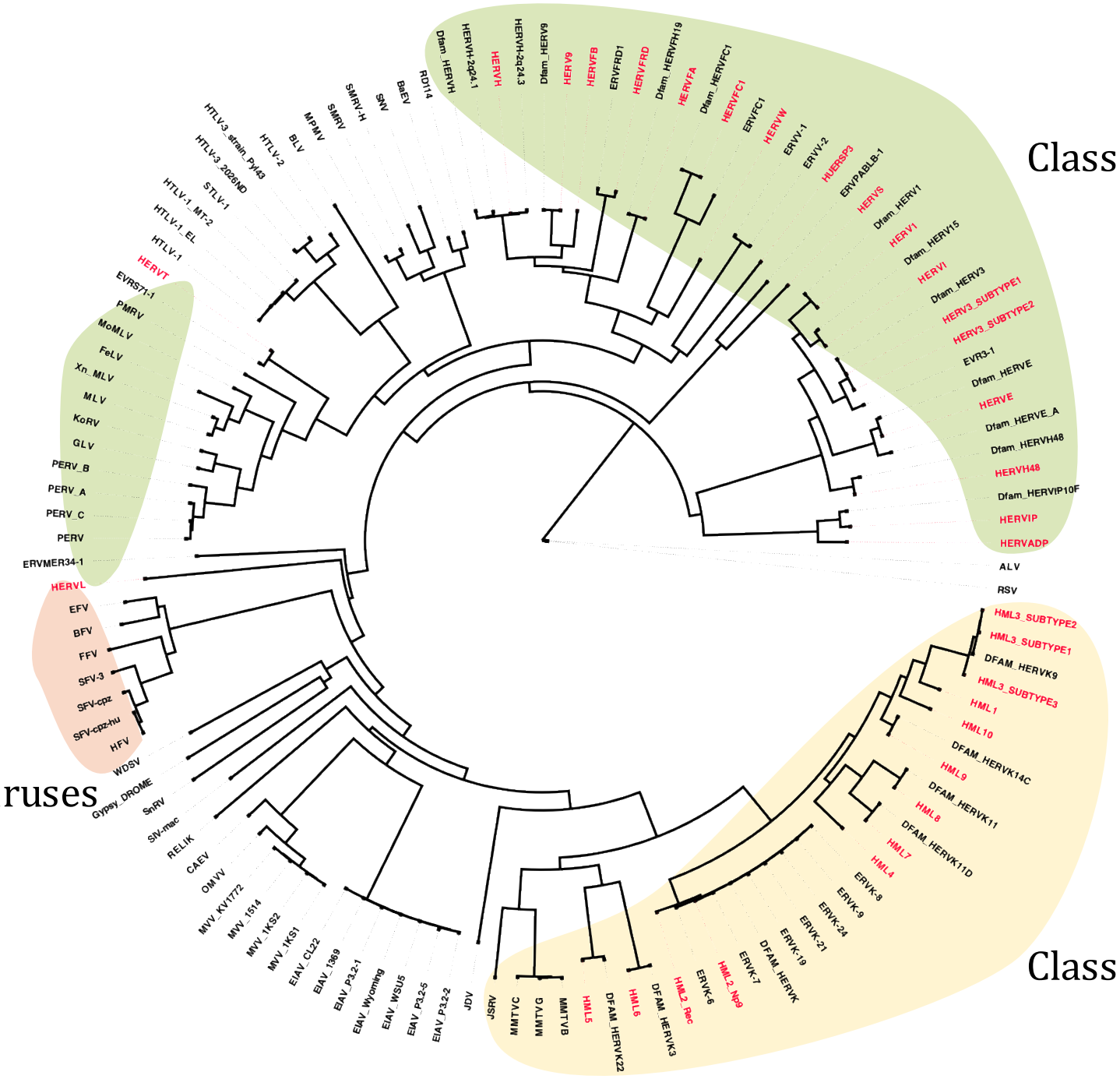
Author contributions

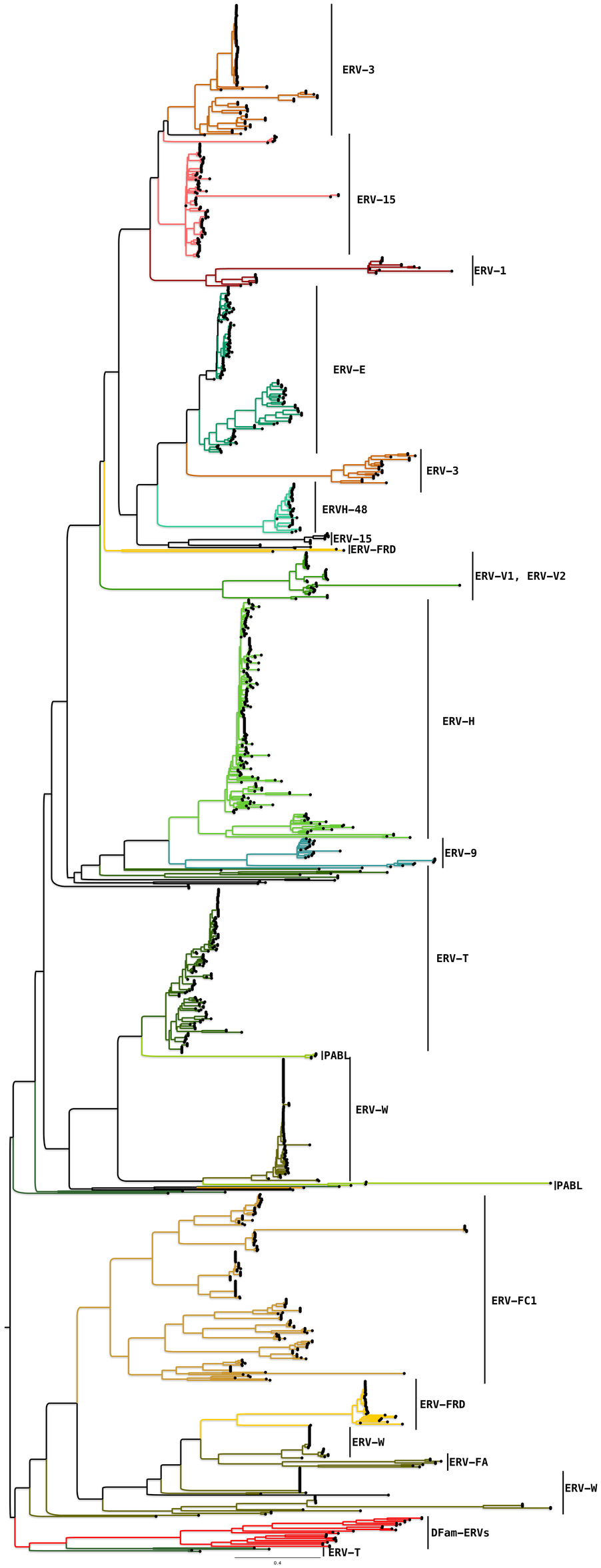
SC performed the analysis and wrote the manuscript. ES participated in performing the analysis. NG supervised the analyses, checked the data, and participated in editing the manuscript. ET and LTL conceived and coordinated the study. All authors contributed to the article and approved the submitted version.

Conflict of interest

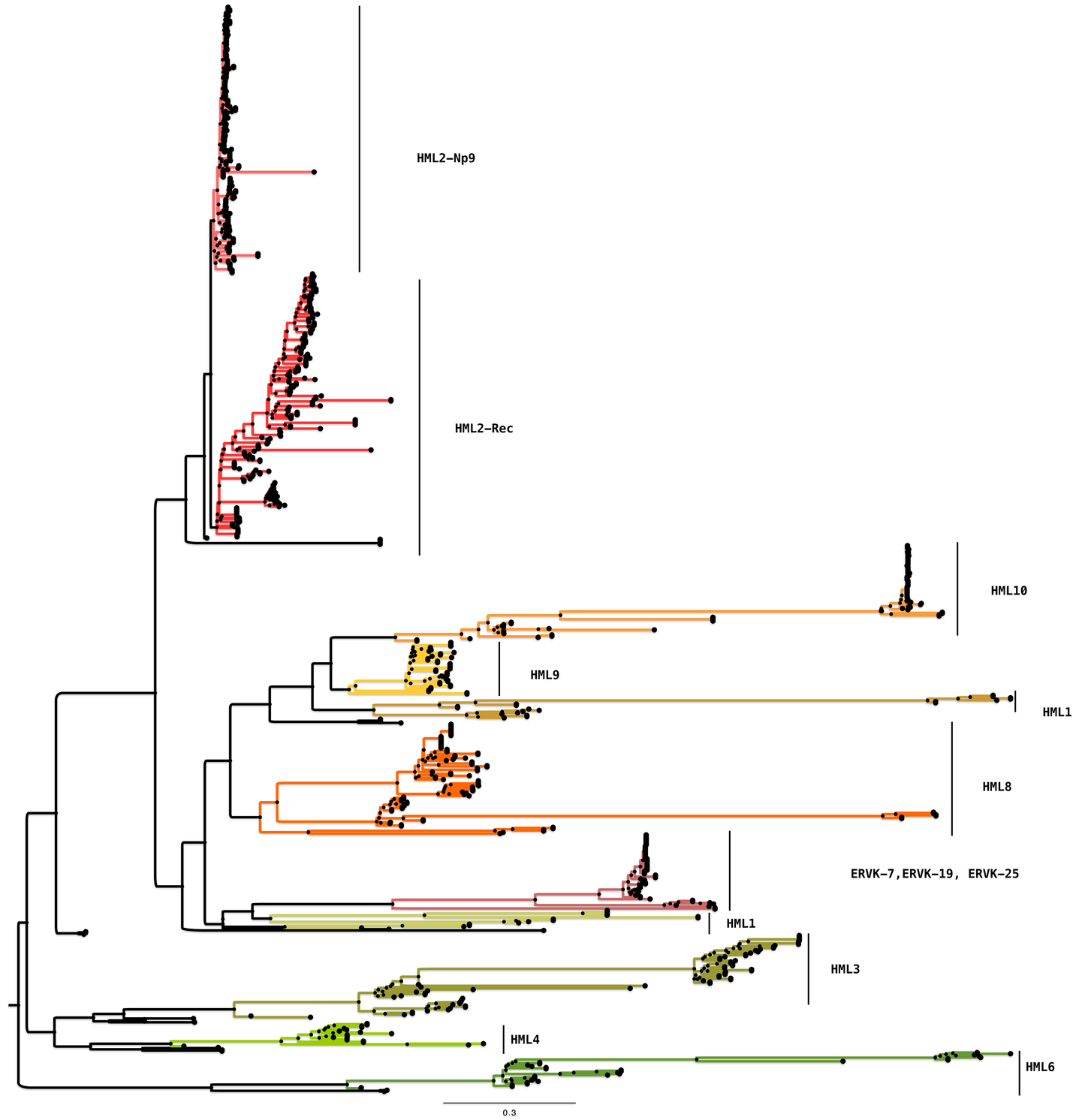
The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.


Collecting Available Data for *env* of Human ERVs
Literature
UniProt
Dfam
Defective *env* genes
Complete *env* genes
Group specific nucleotide alignments
Sequences translated in to forward frames in the alignment
Extracting the translated portions without the stop codons
SP
SU
FP
TM
A representative HERV Env Consensus sequence

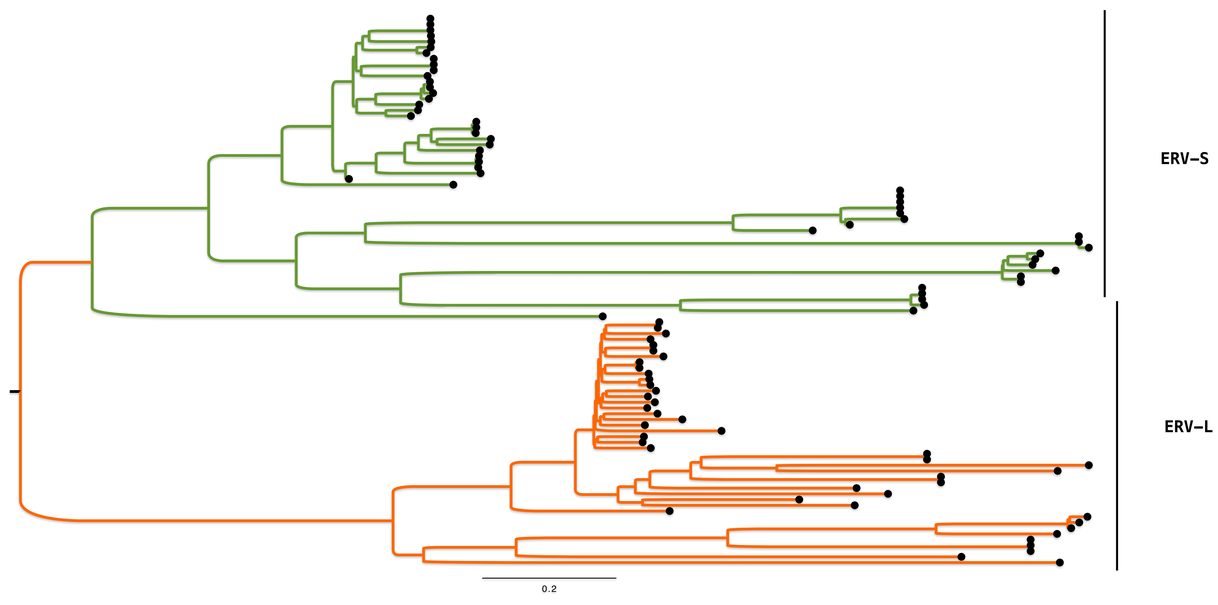

Dfam_HERVH
HERVH-2q24-1
HERVH
HERVH-2q24-3
Dfam_HERV6
HERV9
HERVFB
ERVFRD1
HERVFRD
Dfam_HERVFH19
HERVFA
Dfam_HERVFC1
HERVFC1
ERVFC1
HERVW
ERVV-1
ERVV-2
HUERSP3
ERVPABLB-1
HERVS
Dfam_HERV1
HERV1
Dfam_HERV15
HERV1
Dfam_HERV3
HERV3_SUBTYPE1
HERV3_SUBTYPE2
EVR3-1
Dfam_HERVE
HERVE
Dfam_HERVE_A
Dfam_HERVH48
HERVH48
Dfam_HERVIP10F
HERVIP
HERVADP
ALV
RSV
HML3_SUBTYPE2
HML3_SUBTYPE1
DFAM_HERVK9
HML3_SUBTYPE3
HML1
HML10
DFAM_HERVK14C
HML9
DFAM_HERVK11
HML8
DFAM_HERVK11D
HML7
HML4
ERVK-8
ERVK-9
ERVK-24
ERVK-21
ERVK-19
DFAM_HERVK
ERVK-7
HML2_Np9
ERVK-6
HML2_Rec
DFAM_HERVK3
HML6
DFAM_HERVK22
HML5
MMTVB
MMTVG
MMTVC
JSRV
JDV
EIAV_P3.2-2
EIAV_P3.2-5
EIAV_P3.2-1
EIAV_Wyoming
EIAV_WSU5
EIAV_1369
EIAV_CL22
MVV_1KS1
MVV_1KS2
MVV_1514
MVV_KV1772
OMVV
CAEV
RELIK
SIV-mac
SnRV
Gypsy_DROME
WDSV
HFV
SFV-cpz-hu
SFV-cpz
SFV-3
FFV
BFV
EFV
HERVL
ERVMER34-1
PERV
PERV_C
PERV_A
PERV_B
GLV
KoRV
MLV
Xn_MLV
FeLV
McMLV
PMRV
EVRS71-1
HERVT
HTLV-1
HTLV-1_EL
HTLV-1_MT-2
STLV-1
HTLV-3_2026ND
HTLV-3_strain_Pyl43
HTLV-2
BLV
MPMV
SMRV
SMRV-H
SIV
BaEV
RD114
Class I Gammaretroviruses
Class II Spumaretroviruses
Class II Betaretroviruses


ERV-3
ERV-15
ERV-1
ERV-E
ERV-3
ERVH-48
ERV-15
ERV-FRD
ERV-V1, ERV-V2
ERV-H
ERV-9
ERV-T
IPABL
ERV-W
IPABL
ERV-FC1
ERV-FRD
ERV-W
ERV-FA
ERV-W
DFam-ERVs
ERV-T
0.4

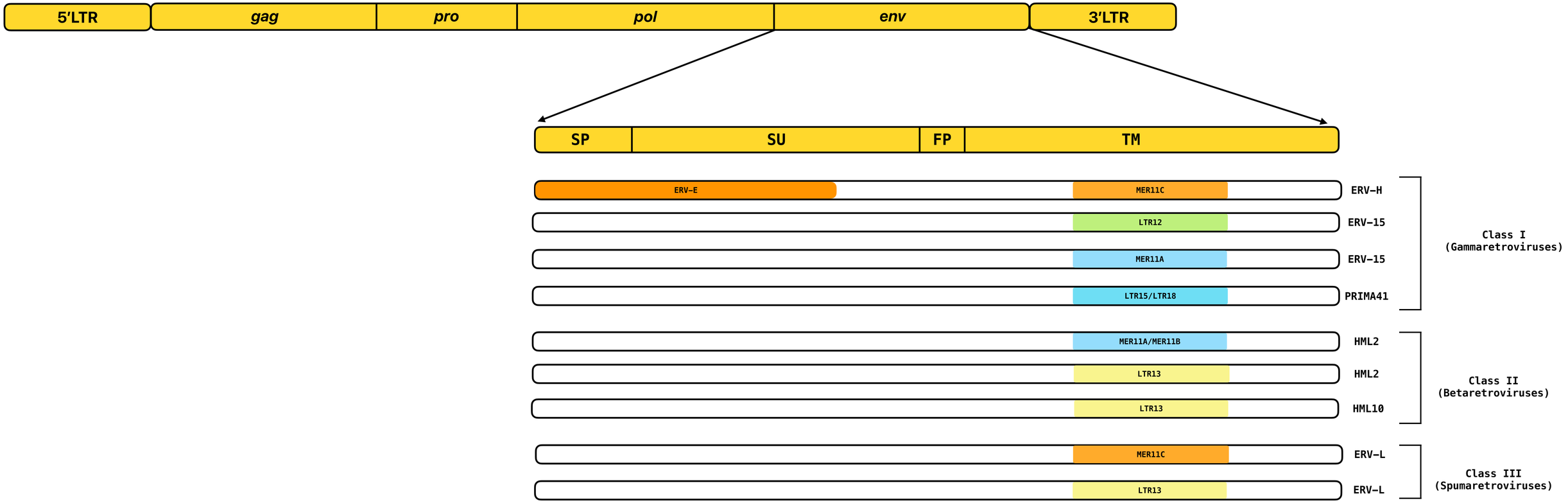
(a)


HML2-Np9
HML2-Rec
HML10
HML9
HML1
HML8
ERVK-7,ERVK-19, ERVK-25
HML1
HML3
HML4
HML6
0.3

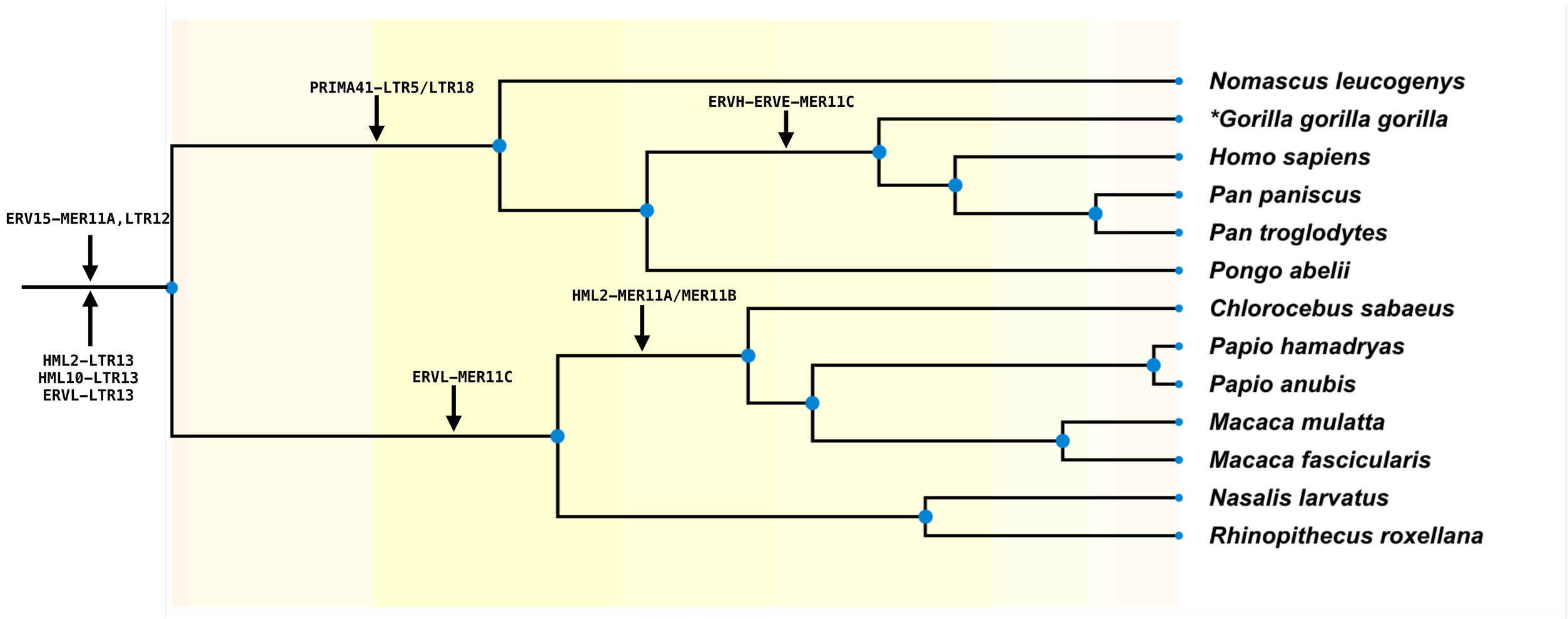
(b)


ERV-S
ERV-L
0.2

(c)


5'LTR
gag
pro
pol
env
3'LTR
SP
SU
FP
TM
ERV-E
MER11C
ERV-H
LTR12
ERV-15
MER11A
ERV-15
LTR15/LTR18
PRIMA41
Class I
(Gammaretroviruses)
MER11A/MER11B
HML2
LTR13
HML2
LTR13
HML10
Class II
(Betaretroviruses)
MER11C
ERV-L
LTR13
ERV-L
Class III
(Spumaretroviruses)

(a)


PRIMA41-LTR5/LTR18
ERVH-ERVE-MER11C
ERV15-MER11A,LTR12
HML2-MER11A/MER11B
HML2-LTR13
HML10-LTR13
ERVL-LTR13
ERVL-MER11C
Nomascus leucogenys
*Gorilla gorilla gorilla
Homo sapiens
Pan paniscus
Pan troglodytes
Pongo abelii
Chlorocebus sabaeus
Papio hamadryas
Papio anubis
Macaca mulatta
Macaca fascicularis
Nasalis larvatus
Rhinopithecus roxellana

(b)