Dynamic evolution of retroviral envelope-derived sequences in primates

Koichi Kitao[a,b,1], Kirill Kryukov[c,d,1], Lihua Jin[e,f], Yuta Shintaku[g,h], Takashi Hayakawa[i], So Nakagawa[e,f,j*]

[a]Laboratory of Genome and Epigenome Dynamics, Department of Animal Sciences, Graduate School of Bioagricultural Sciences, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan

[b]Laboratory for Retrotransposon Dynamics, RIKEN Center for Integrative Medical Sciences, Yokohama, 230-0045, Japan

[c]Center for Genome Informatics, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, Mishima, Shizuoka 411-8540, Japan

[d]Bioinformation and DDBJ Center, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan

[e]Department of Molecular Life Science, Tokai University School of Medicine, Isehara, Kanagawa 259-1193, Japan

[f]Division of Omics Sciences, Institute of Medical Sciences, Tokai University, Isehara, Kanagawa 259-1193, Japan.

[g]Japan Monkey Centre, Inuyama, Aichi 484-0081, Japan

[h]Wildlife Research Center, Kyoto University, Sakyo, Kyoto 606-8203, Japan

[i]Faculty of Environmental Earth Science, Hokkaido University, Sapporo, Hokkaido 060-0810, Japan

[j]Division of Interdisciplinary Merging of Health Research, Micro/Nano Technology Center, Tokai University, Hiratsuka, Kanagawa 259-1292, Japan.

*So Nakagawa

Email: so@tokai.ac.jp

Author Contributions: K. Kitao, K. Kryukov, and S. N. designed research; K. Kitao, K. Kryukov, Y. S., and T. H. performed research; K. Kitao, and K. Kryukov analyzed data; K. Kitao, Y. S., and T. H. performed experiment; and K. Kitao, K. Kryukov, L. J., and S. N. wrote the paper.

[1]Contributed equally to this work.

Competing Interest Statement: The authors declare that they have no conflict of interest.

Keywords: endogenous retrovirus, de novo genes, co-option, primate evolution

Abstract

It is known that several endogenous retroviruses (ERVs), remnants of ancient retroviral integrations, retain envelope (*env*) genes encoding fusogenic proteins in primates. While most *env* genes are degraded, a few have been co-opted by hosts, yet the entire evolutionary dynamics of *env* sequences remain poorly understood. To explore this, we screened and compared *env* open-reading frames (ORFs) from 247 primate genomes. In total, 8,683 nearly intact *env*-ORFs encoding over 400 amino acids were identified, and their copy numbers ranged from 3 to 429 across primate species. By conducting sequence similarity clustering, we found that the reported functional *env*-derived genes tend to have low copy numbers and to be retained across primate lineages. Notably, an evolutionary conserved *env* ortholog with low sequence similarity to other *env*-ORFs was identified in tarsiers, which exhibited cell fusion activity in vitro, suggesting a potential fusogenic function in tarsiers. Further, we found that certain co-opted *env*-derived genes may have lost their functions due to nonsense or indel mutations within specific primate lineages. Considering that many *env* genes tend to be maintained at low copy numbers and reported to be under natural selection, such dynamic evolutionary turnover of *env*-derived genes may be driven by host-virus arms races, as viruses and endogenous retroviruses often share cell-surface receptors.

Main Text

Introduction

It is known that approximately 8% of the human genome consists of LTR retrotransposons, primarily derived from endogenous retroviruses (ERVs) (1, 2). ERVs are remnants of ancient retroviral infections that became integrated into the host germline genome and are inherited across generations (3). Structurally, they typically consist of structural protein (*gag)*, polymerase (*pol*), and envelope (*env*) genes flanked by long terminal repeats (LTRs). Among them, 0.04% of ERVs retain open reading frames (ORFs) exceeding 80 amino acids in length (4). Although this fraction appears small, the human genome contains 1,731 *env*-derived ORFs encoding proteins longer than 80 amino acids (5). This number exceeds that of olfactory receptor genes, including pseudogenes, which constitute the largest multigene family in the human genome (6). Env proteins of retroviruses recognize cell receptors and fuse the viral and cellular membranes during infection. While most *env* genes are truncated, some near-intact ORFs have been co-opted for new functions by leveraging their inherent molecular activity (7).

Several Env proteins have acquired roles in placental development in primates. Syncytin-1 and Synytin-2, although derived from different ERV *env* genes; *ERVW-1* and *ERVFRD-1* respectively, promote syncytiotrophoblast cell fusion in Hominidea (apes) species (8–11). *ERVW-1* originated in the common ancestor of Catarrhini but lost function in many Cercopithecoidea (Old World monkey) species due to nonsense mutations (12). *ERVFRD-1* arose in the Simiiformes (simian) ancestor (10), but its protein appears non-fusogenic in some Platyrrhini (New World monkey) speceis (13). *ERVV-2* encodes Mac-syncytin-3, which shows placental fusogenicity in several Cercopithecoidea and Platyrrhini species, but not in Hominoidea species; its tandem duplicate, *ERVV-1*, is thought to be inactive due to disruptive mutations (14, 15).

The *env* genes are subject to dynamic turnover across different evolutionary lineages: some acquire functions, while others lose them. In previous studies, we termed this type of genetic shift as the “baton pass hypothesis”, proposing that a large abundance of ERV-derived sequences in the genome provides multiple candidates with similar functional potentials, occasionally resulting in stochastic functional replacement (**Fig. 1**) (16). Recently, various primate genomes have been sequenced and deposited in the public database, providing whole-genome assemblies for 247 subspecies from 238 primate species, representing 81% of genera and 94% of families in the primate order (17, 18). Leveraging this comprehensive genomic dataset enables investigating evolutionary changes in *env*-derived sequences with high resolution across species. In this study, utilizing 247 primate genome assemblies (**Table S1 and Fig. S1**), we aimed to examine the evolutionary dynamics of *env*-derived open reading frames, including functional envelope-derived genes in humans.

Results

Variation of *env*-ORFs in primates. To identify nearly intact *env* gene-encoding ORFs from 247 primate genome assemblies (**Dataset S1**), we first extracted all ORFs encoding more than 400 amino acids, defined from the initiation codon (i.e., “ATG”) to the stop codon. We then conducted protein domain searches using hidden Markov model (HMM) profiles obtained from the Gypsy database (19). In parallel, we performed blastp searches using representative human ERV Env proteins and retroviral Env proteins (**Dataset S2**). Then, the significant hits from both searches were combined, and long *env*-like ORFs, hereafter referred to as “*env*-ORFs”, were identified (**Fig. S1A**).

All primate genome assemblies analyzed in this study contained at least three *env*-ORFs, resulting in a total of 8,683 *env*-ORFs (**Fig. S1B; Dataset S3**). Among the primates analyzed, *Nycticebus coucang* (Sunda slow loris) exhibited the highest number of *env*-ORFs, with 429 identified. We then classified these *env*-ORFs based on their best hit to known retroviral Env proteins (**Fig. 2; Dataset S4**). The majority were assigned to the *Betaretrovirus* or *Gammaretrovirus* lineages. Notably, *env* genes belonging to the RD114-and-D-type-retrovirus (RDR) superfamily,

which are found in both *Betaretrovirus* and *Gammaretrovirus* (20), were counted separately. No *env*-ORFs related to *Deltaretrovirus* or *Epsilonretrovirus* were detected. Additionally, a number of unclassified *env*-ORFs showing no significant similarity to retroviral Env proteins by blastp were identified. *ERVMER34-1 (HEMO)*, derived from an ancient endogenous retroviral *env* gene (21), was included in this unclassified group. An *env*-ORF related to *Spumaretrivinae* (foamy viruses) was detected in *Cephalopachus bancanus* (Horsfeld's tarsier). Regarding this spumaretroviral *env*-ORF, *gag* and *pol* were also detected, and the identity between the 5'- and 3'-LTRs was almost identical (98.8%, **Fig. S2A**). The phylogenetic tree of its reverse transcriptase also supports that this ERV belongs to *Spumaretrovirinae* (**Fig. S2B**). ERVs of *Spumaretrovirinae* are rare in primate genomes, and an ERV with in-frame stop codons and frameshift mutations has been identified only in the aye-aye genome in primates (22).

Protein sequence clustering revealed different fates of *env*-ORFs. Next, we clustered the 8,683 *env*-ORFs based on amino acid sequence similarity (40% sequence identity). The clustering yielded 109 clusters, and a phylogenetic tree was constructed using representative amino acid sequences from each cluster (**Fig. 3**). For each cluster, we calculated the proportion of primate species retaining *env*-ORFs in the cluster and the copy number per genome (**Fig.3** and **Fig. S3**).

The phylogenetic tree shows that the known co-opted *env* genes (*ERV3-1*, *ERVMER34-1*, *ERVPb-1*, *ERVV-1*, *ERVV-2*, *ERVW-1*, and *ERVFRD-1*) were classified into separate clusters, excluding *ERVV-1* and *ERVV-2*. Each was included in a cluster present in many species with low copy numbers. For example, *ERVFRD-1*, which encodes Syncytin-2, was included in a cluster observed as a single copy in most species of Hominoidea, Cercopithecoidea, and Platyrrhini. Notably, *ERVV-1* and *ERVV-2* were observed in most of the species of Hominoidea and Cercopithecoidea but were poorly retained in Platyrrhini, although Mac-syncytin-3 encoded by *ERVV-2* exhibits fusogenic activity in several Platyrrhini species (14). These functionally co-opted *env* genes were acquired in ancient evolutionary periods and have subsequently been conserved, which could explain their presence in many species with low copy numbers. Besides these known co-opted genes, we did not find any clusters with high retention rates and low copy numbers in Hominoidea, Cercopithecoidea, and Platyrrhini.

In Tarsiiformes (tarsiers), however, one *env*-ORF cluster showed high retention and a low copy number. No clusters exhibited signatures of co-option in Lemuriformes and Lorisiformes. These findings suggest that while most *env*-ORF groups exhibit copy number variation, those encoding co-opted proteins tend to undergo copy loss and eventually stabilize at a low copy number during evolution.

Possible co-opted *env*-ORFs in Tarsiiformes. The clustering results suggested the presence of a conserved *env*-ORF in Tarsiiformes (**Fig. 3**). To our knowledge, evolutionarily conserved *env* genes remain to be elucidated in prosimians (i.e., Chiromyiformes, Lemuriformes, Lorisiformes and Tarsiiformes). To complement this result, we applied a more detailed phylogeny-based characterization for *env*-ORFs of prosimians. For Lemuriformes and Lorisiformes, one species from each genus was used for the phylogenetic analysis. In the phylogenetic tree of 217 *env*-ORFs from 13 species of Chiromyiformes (aye-aye) and Lemuriformes (lemurs), no conserved single-copy orthologs were observed (**Fig. S4**). Similarly, no single-copy ortholog was found from 604 *env*-ORFs in 7 species of Lorisiformes (galagos and lorises) (**Fig. S5**). These data suggest that there is no strong evidence of conserved *env* genes inherited from the last common ancestor of Chiromyiformes, Lemuriformes or Lorisiformes.

Four Tarsiiformes species diverged approximately 22 million years ago were analyzed in this study (**Fig. 4A**). A phylogenetic tree of 183 *env*-ORFs of these four Tarsiiformes species elucidated a single-copy ortholog (**Fig. 4B**). This gene was named *env-Tar1* (*env*-ORFs in Tarsiiformes 1). This *env-Tar1* was equivalent to the cluster shared in Tarsiiformes (**Fig. 3**). The *env-Tar1* ORFs encoded 446-448 amino acids, showing high pairwise identities (92.63-99.11%) and a purifying selection (**Fig. 4C**). Furthermore, either intact *gag-pol* ORFs or LTRs were not detected in the surrounding genomic regions, suggesting the selective selection to the *env-Tar1* gene (**Fig. S6**). The putative Env-Tar1 protein possesses characteristic features of retroviral Env

proteins, including an N-terminal signal peptide, a Furin cleavage site between the surface (SU) and transmembrane (TM) subunits, a hydrophobic transmembrane domain, and an immunosuppressive domain (**Fig. 4D**). These features suggest that the Env-Tar1 protein may retain cell fusion capability like the Syncytin proteins.

Env-Tar1 is a fusogenic Env protein. We experimentally evaluated whether the Env-Tar1 protein functions as a fusogen. We constructed the plasmid for expressing the Env-Tar1 protein of *Carlito syrichta* (Philippine tarsier). As a negative control, we generated a Furin cleavage-inactive mutant in which the arginine at position 272 was replaced with alanine (**Fig. 5A**). When the C-terminal Flag-tagged Env-Tar1 protein was expressed in human embryonic kidney 293T cells, the TM subunit was detected in the wild-type Env-Tar1, but not in the Furin mutant (**Fig. 5B**). This result confirmed that the Furin mutant serves as an appropriate negative control for functional assay. The fusogenic activity of Env-Tar1 was then assessed using a fusion-dependent LacZ assay (PMID: 37062963). In this assay, the Env-Tar1 construct did not contain a C-terminal Flag tag. Wild-type Env-Tar1 induced more syncytia compared to the Furin mutant (**Fig. 5C**). This demonstrates that Env-Tar1 functions as a fusogen.

Possible truncation of conserved *env*-ORFs in various species. As described above, clusters of co-opted *env* genes are often widely shared among species. However, upon closer inspection of the clustering results, we found that some species lacked *env*-ORFs that are expected to be conserved. This suggests that certain co-opted *env* genes may have lost their function within specific lineages. To investigate this possibility, we examined conservation of seven known human *env* genes: *ERVW-1* (*syncytin-1*), *ERVFRD-1* (*syncytin-2*), *ERVV-1*, *ERVV-2*, *ERV3-1*, and *ERVMER34-1* (*HEMO*) (**Fig. 6**). We performed the tblastn search using these seven genes as queries and investigated their ORF length (**Dataset S5**). The results suggest that several *env* gene ORFs are truncated in multiple species across various lineages.

In particular, we found various truncations in *ERVV* ORFs. Since *ERVV-1* and *ERVV-2* are tandemly duplicated genes and their sequences are highly similar to each other (23, 14), we obtained sequences that showed the highest similarity to either *ERVV-1* or *ERVV-2*. Nevertheless, several species were found to possess only short ORFs (**Fig. 6**). In *Macaca mulatta* (rhesus macaque), ERVV-2 is fusogenic and also referred to as Mac-syncytin-3 (14); however, it was found that the ORFs are shortened in multiple Cercopithecoidea species. Among them, the contigs of three *Semnopithecus* species (*S. hypoleucos*, *S. priam*, and *S. schistaceus*) were truncated at the same position, resulting in no intact *ERVV* ORFs being detected (**Fig. S7**), which may suggest that deletion events in the *ERVV* family have occurred in the ancestor of the lineage. Although we cannot rule out the possibility that the apparent truncations are due to simply low genome assembly accuracy (see Discussion), these results suggest that species-specific truncations may have occasionally occurred for *env*-derived genes during primate evolution.

Discussion

In this study, we found that primate *env*-ORFs are mainly derived from *Betaretrovirus* and *Gammaretrovirus*, and that the number of *env*-ORFs in primates varies greatly, even within the same genus (**Fig. 2**). These results suggested that the turnover of *env*-ORFs in primate genomes is highly active, illustrating the dynamic evolution of *env* genes in primates. On the other hand, we found that *env*-derived genes that have acquired functions are conserved across many lineages, and similar sequences are present in only a few copies for each species (**Fig. 3**). Indeed, such an *env* gene found in Tarsiiformes retained membrane fusion activity, suggesting that it had acquired a novel function in Tarsiiformes and been under negative selection (**Figs. 4 and 5**). However, we also found that some *env* genes that had once acquired functions may have subsequently lost them (**Fig. 6**), suggesting that these *env* genes may have been functionally replaced (**Fig. 1**).

Indeed, in ruminants, *syncytin-Rum1* was reported to encode a fusogenic protein and be highly expressed in the placenta (24). However, in the Bovidae family, *fematrin-1*, another *env*-derived gene found only in Bovidae, shows higher expression levels and greater fusion capacity

than *syncytin-Rum1* (25). This likely indicates that *fematrin-1* has replaced the molecular function related to the cell-cell fusion process of *syncytin-Rum1* in Bovidae (16, 26, 27). Our findings also support the idea that functional *env*-derived gene replacements may frequently occur during mammalian evolution.

Why do *env*-derived genes change so dynamically during evolution? One possible explanation is the shared usage of receptors of host cells by viruses and ERVs. For example, it is known that envelope proteins of RDR interference group viruses, such as Syncytin-1, Env-Tac1, suppressyn, and RD-114, use host ASCT-2 protein on the surface of cells as their receptor (28–31). For *ASCT-2* gene, it is reported that independent insertion mutations occurred in rodents and birds at the sites where these proteins interact with viral envelope proteins (32). Indeed, many positively selected sites have been detected in receptor genes of primates, particularly at the virus-receptor interfaces (33). This suggests that even if *env*-derived genes can functionally utilize membrane fusion activity, mutations in the host receptors may be advantageous to the host if infectious viruses use the same receptor. This may contribute to the continuous turnover of *env*-derived genes during evolution. In fact, various *env*-derived genes are known to be under purifying selection (24, 13), which our study also shows that they tend to be maintained at low copy numbers (**Fig. 3**). Nevertheless, the fact that they still undergo frequent changes may be due to positive selection acting on their receptors, which are involved in viral escape. For example, retroviral proteins derived from genes other than *env,* such as *PEG10* and *PEG11*/*RTL1,* have been reported to be functionally conserved across many mammalian species (34, 35). Therefore, host-virus arms races may also drive dynamic changes in *env*-derived genes during evolution.

Note that potential genome misassemblies may have led to misannotations of truncations in some *env*-derived genes. Indeed, we found that genomes with shorter contigs tend to contain fewer *env*-derived ORFs (**Fig. S8**). Further, *ERVW-1*, which encodes Syncytin-1 and is expected to be conserved in Hominoidea, appears truncated in *Nomascus gabriellae* (Yellow-cheeked gibbon) (**Fig. 6**). Two syncytin-1-like ORFs found in this species contain nonsense mutations (**Fig. S9A**), suggesting that they may lose *ERVW-1* or the corresponding region was not detected due to the low accuracy of the genome assembly. For *ERVFRD-1* encoding Syncytin-2, most simians retain intact ORFs, but six species (two Cercopithecoidea and four Platyrrhini species) carry truncated forms that seem to have occurred independently (**Fig. 6**). In *Papio hamadryas* (Hamadryas baboon), the reference genome shows a 1-bp deletion causing ORF truncation, yet PCR analyses from two individuals revealed no such deletion (**Fig. S9B**). Comparative analyses further indicate that the reference sequence likely represents a chimeric artifact involving another ERV copy on chromosome 3 (**Fig. S9C**). These findings suggest that several apparent truncations of *syncytin* genes may result from genome misassemblies.

On the other hand, a shortened *env*-derived gene does not necessarily mean that the gene has lost its function. In fact, several truncated *env*-derived genes have been shown to retain biological activity, such as *suppressyn* and others in primates (30, 36, 37) and *refrex-1* in cats (38). However, their functions differ from the original membrane fusion ability of envelope proteins, as they act by interfering with host receptors to block viral infection. Since potential functions such as interference-mediated inhibition of viral infection were not examined in this study, the precise roles of these truncated *env*-derived genes remain largely unclear.

Although there are several limitations, our study indicates that *env*-derived genes exhibit considerable diversity, and functional genes may also undergo frequent changes in primates. As the accuracy of genome sequences improves and more genomes of various species and individuals become accessible, our understanding of this area will advance even further.

Materials and Methods

DNA experiment ethics. The primate DNA samples used in this study were collected at the Japan Monkey Centre. This experiment is authorized by the ethics community of Japan Monkey Centre under 2024009.

PCR. To determine the nucleotide sequence of *ERVFRD-1* of *Papio hamadryas*, genomic DNA was first extracted from muscle tissue (Sample ID: Pr6767) and blood (Sample ID: 125) using the Purelink Genomic DNA Mini kit (K182001, Invitrogen). A PCR reaction was set up with the purified genomic DNA, employing the forward (5'-CCTCAAAGTAAGGTGATCCCAATA-3') and reverse (5'-TATCCAGCCAGATCCACAGGAGA-3') primers. Using KOD One polymerase (KMM-101, TOYOBO), the DNA was amplified 30 cycles under the following conditions: denaturation at 98°C for 10 s, annealing at 60°C for 5 s, and extension at 68°C for 20 s. The sequence of the amplicon was subsequently verified through Sanger sequencing (Eurofins genomics).

Genome assemblies. 247 primate genome assemblies were used: 20 Hominoidea (apes), 87 Cercopithecoidea (catarrhine monkeys), 81 Platyrrhini (platyrrhine monkeys), 4 Tarsiiformes, and 55 Strepsirrhini. A complete list of the species and the accession numbers of their genome assemblies were summarized in **Dataset S4**. BEDtools v2.30.0 (39) was used to extract genomic sequences.

Detection of *env*-ORFs. ORFs that were greater than 400 amino acids in length, beginning with a start codon and ending with a stop codon, were retrieved using the getorf program from the European Molecular Biology Open Software Suite v6.5.7.0 (40). To remove ORFs that were generated by simple repeat sequences, ORFs with more than 30% of the sequence consisting of a single amino acid were excluded. ORFs with 10 or more consecutive "X" amino acids were also excluded, to remove ORFs derived from "N" sequences used to fill unknown bases or gaps in the genome assemblies. To identify *env*-ORFs from the retrieved ORFs, we employed the following three methods. First, we used the hmmscan program from HMMER3 v3.2.2 (41) to search for Env domains with an expected threshold of 1E-10, using Env protein HMM profiles from the GyDB collection (19). Second, a blastp search (E-value ≤ 1E-10) was performed using a set of representative human endogenous Env proteins, including ERVW-1/Syncytin-1 (GenBank ID: AAD14546.2), ERVFRD-1/Syncytin-2 (GenBank ID: NP_997465.1), ERVV-1 (GenBank ID: NP_689686.2), ERVV-2 (GenBank ID: NP_001177984.1), ERV3-1 (GenBank ID: NP_001007254.2), ERV-Pb1 (GenBank ID: ABB52637.1), HEMO (GenBank ID: NP_001229619.1), HERV-T (GenBank ID: BAG06168.1), HERV-H (GenBank ID: AAD34324.1), and HERV-K115 (GenBank ID: AAL18259.1). Third, another blastp search (E-value ≤ 1E-10) was conducted using representative Env proteins of exogenous retroviruses (see **Dataset S2**). The results from all three methods were then merged to create a non-redundant list of *env*-ORFs for each species.

***env*-ORF clustering and phylogenetic analysis.** The total set of *env*-ORFs from the primate genomes was clustered at a 40% amino acid sequence identity threshold using CD-HIT v4.8.1 (42). Note that any singletons—ORFs that did not cluster with any other sequences—were excluded from this analysis. The amino acid sequences of representative *env*-ORFs in each cluster suggested by CD-HIT were aligned using MAFFT v7.487 with the "L-INS-i" method (43). Following this, the resulting alignment was refined by removing all gaps using the trimAl v1.4.rev15 program with the "-gappyout" option (44). The trimmed alignment was used for the phylogenetic tree construction using IQ-TREE2 v2.0.3 (45). To assess branch reliability, the ultrafast bootstrap values with 1000 replicates were computed for each node (46).

Plasmid construction. The ERVW-1/Syncytin-1 expression plasmid (phCMV3-Syn1+100) was previously constructed (47). To construct the Env-Tar1 expression plasmid (phCMV3-Env-Tar1), dsDNA encoding the *env-Tar1* ORFs of *Carlito syrichta* was synthesized by Eurofins Genomics (Tokyo, Japan). The synthesized dsDNA was inserted into EcoRI and BamHI sites of phCMV3 by NEBuilder. For the Furin-cleavage inactive mutant, phCMV3-Env-Tar1 was linearized by PCR with the forward (5'-CGGTGGACAGCCTCAGTCTTTCACTGGTATG-3') and reverse (5'-TGAGGCTGTCCACCGATCGACCAAATGAGGC-3') primers and circularized by NEBuilder. For the Flag-tagged mutant, the ORF with Flag sequence was amplified from phCMV3-Env-Tar1 and phCMV3-Env-Tar1-R272A using the forward (5'-

CGGTGGACAGCCTCAGTCTTTCACTGGTATG-3') and reverse (5'-AACATCGTATGGGTAGGGTTTCGGGGACAACAG-3') primers and inserted into phCMV3 using NEBuilder. All PCRs described above were carried out using KOD One Master Mix with a C1000 Touch thermal cycler (Bio-Rad).

Western blotting. 293T human embryonic kidney cells (RCB2202) was provided by the RIKEN BRC through the National BioResource Project of the MEXT, Japan. 293T cells were seeded on 24-well plates at 2.5×10^5 cells per well. The next day, cells were transfected with phCMV3, phCMV3-Env-Tar1-Flag or phCMV3-Env-Tar1-R272A-Flag, using Avalanche everyday transfection reagent (EZTEVDY-1, EZ Biosystems). The cells were lysed in sample buffer for SDS-PAGE (09499-14, Nacalai Tesque) 24 h after transfection. SDS/PAGE was performed, and peptides were transferred from the gel to polyvinylidene difluoride membranes. The membranes were reacted with a mouse anti-Flag antibody (M2 clone, Sigma-Aldrich) as a primary antibody. A goat anti-mouse IgG antibody (32430, Invitrogen) was used as a second antibody. Signals were detected using SuperSignal west femto (34095, Thermo Fisher Scientific).

Fusion-dependent LacZ assay. 293T cells as donors were seeded at 2.5×10^5 cells per well on 24-well plates and were transfected with 0.5 μg of phCMV3, phCMV3-Env-Tar1 or phCMV3-Env-Tar1-R272A and 0.5 μg of the LacZ reporter plasmid (pT7EMCV-LacZ), which expresses LacZ with internal ribosome entry site from encephalomyocarditis virus in the presence of T7 polymerase. The 293T cells as recipients were transfected with 0.5 μg of an expression plasmid of T7 polymerase (pCAG-T7Pol). Transfection was conducted with 0.8 μL of Avalanche Everyday Transfection Reagent. Donor and recipient cells were cocultured 6 h after transfection, and the LacZ-positive fused cells were stained by 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal) 42 h after coculture.

Similarity search for conserved *env*-ORFs. We conducted a tblastn (48) search to find homologs of human ERV genes in primate genomes. We used tblastn 2.15.0 with parameters "-evalue 1e-5 -dbsize 3200000000 -outfmt 6 -num_alignments 500000 -show_gis -mt_mode 2 -num_threads 8 -seg no". For a query we used 35 amino acid sequences of human ERV-related genes (**Dataset S5**). We used 247 primate genome assemblies as a database (**Table S1**).

Acknowledgments

We acknowledge the use of ChatGPT5 for improving the English language and flow of the manuscript. This work was supported by JSPS KAKENHI Grant Numbers JP25K02265 (to K. Kitao and S.N) and JP21H04919 and JP21KK0106 (to T.H.), and by Core 1 Project of the Institute of Medical Sciences, Tokai University (to L.J. and S.N). The supercomputing resource was partially supported by the NIG supercomputer at ROIS National Institute of Genetics.

References

1. E.S. Lander *et al.*, Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. S.J. Hoyt *et al.*, From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science* **376**, eabk3112 (2022).
3. W.E. Johnson, Origins and evolutionary consequences of ancient endogenous retroviruses. *Nat. Rev. Microbiol*. **17**, 355–370 (2019).
4. M.T. Ueda *et al.*, Comprehensive genomic analysis reveals dynamic evolution of endogenous retroviruses that code for retroviral-like protein domains. *Mob. DNA* **11**, 29 (2020).
5. S. Nakagawa, M.U. Takahashi, gEVE: a genome-based endogenous viral element database provides comprehensive viral protein-coding sequences in mammalian genomes. *Database* **2016**, baw087 (2016).
6. Y. Niimura, M. Nei, Evolution of olfactory receptor genes in the human genome. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 12235–12240 (2003).

7. N. Grandi, E. Tramontano, HERV Envelope Proteins: Physiological Role and Pathogenic Potential in Cancer and Autoimmunity. *Front. Microbiol.* **9**, 462 (2018).
8. P. Priščáková *et al.*, Syncytin-1, syncytin-2 and suppressyn in human health and disease. *J. Mol. Med.* **101**, 1527–1542 (2023).
9. S. Mi *et al.*, Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* **403**, 785–789 (2000).
10. S. Blaise, N. de Parseval, L. Bénit, T. Heidmann, Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 13013–13018 (2003).
11. M. Gholami Barzoki, S. Shatizadeh Malekshahi, Z. Heydarifard, M.J. Mahmodi, H. Soltanghoraee, The important biological roles of Syncytin-1 of human endogenous retrovirus W (HERV-W) and Syncytin-2 of HERV-FRD in the human placenta development. *Mol. Biol. Rep.* **50**, 7901–7907 (2023).
12. B. Bonnaud *et al.*, Natural history of the ERVWE1 endogenous retroviral locus. *Retrovirology* **2**, 57 (2005).
13. H. Shoji, K. Kitao, T. Miyazawa, S. Nakagawa, Potentially reduced fusogenicity of syncytin-2 in New World monkeys. *FEBS Open Bio.* **13**, 459–467 (2023).
14. C. Esnault, G. Cornelis, O. Heidmann, T. Heidmann, Differential evolutionary fate of an ancestral primate endogenous retrovirus envelope gene, the EnvV syncytin, captured for a function in placentation. *PLoS Genet.* **9**, e1003400 (2013).
15. S. Blaise, N. de Parseval, T. Heidmann, Functional characterization of two newly identified Human Endogenous Retrovirus coding envelope genes. *Retrovirology* **2**, 19 (2005).
16. K. Imakawa, S. Nakagawa, T. Miyazawa, Baton pass hypothesis: successive incorporation of unconserved endogenous retroviral genes for placentation during mammalian evolution. *Genes Cells* **20**, 771–788 (2015).
17. L.F.K. Kuderna *et al.*, A global catalog of whole-genome diversity from 233 primate species. *Science* **380**, 906–913 (2023).
18. L.F.K. Kuderna *et al.*, Identification of constrained sequence elements across 239 primate genomes. *Nature* **625**, 735–742 (2024).
19. C. Llorens, The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res.* **39**, D70–D74 (2011).
20. A. Sinha, W.E. Johnson, Retroviruses of the RDR superinfection interference group: ancient origins and broad host distribution of a promiscuous Env gene. *Curr. Opin. Virol.* **25**, 105–112 (2017).
21. O. Heidmann, HEMO, an ancestral endogenous retroviral envelope protein shed in the blood of pregnant women and expressed in pluripotent stem cells and tumors. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E6642–E6651 (2017).
22. G.Z. Han, M. Worobey, An endogenous foamy virus in the aye-aye (Daubentonia madagascariensis). *J. Virol.* **86**, 7696–7698 (2012).
23. A.L. Kjeldbjerg, P. Villesen, L. Aagaard, F.S. Pedersen, Gene conversion and purifying selection of a placenta-specific ERV-V envelope gene during simian evolution. *BMC Evol. Biol.* **8**, 266 (2008).
24. G. Cornelis *et al.*, Captured retroviral envelope syncytin gene associated with the unique placental structure of higher ruminants. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E828–E837 (2013).
25. Y. Nakaya, K. Koshi, S. Nakagawa, K. Hashizume, T. Miyazawa, Fematrin-1 is involved in fetomaternal cell-to-cell fusion in Bovinae placenta and has contributed to diversity of ruminant placentation. *J. Virol.* **87**, 10563–10572 (2013).
26. Y. Nakaya, T. Miyazawa, The Roles of Syncytin-Like Proteins in Ruminant Placentation. *Viruses* **7**, 2928–2942 (2015).
27. K. Imakawa, Endogenous Retroviruses and Placental Evolution, Development, and Diversity. *Cells* **11**, 2458 (2022).
28. K. Štafl *et al.*, Receptor usage of Syncytin-1: ASCT2, but not ASCT1, is a functional receptor and effector of cell fusion in the human placenta. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2407519121 (2024).

29. K. Kitao, H. Shoji, T. Miyazawa, S. Nakagawa, Dynamic Evolution of Retroviral Envelope Genes in Egg-Laying Mammalian Genomes. *Mol. Biol. Evol.* **40**, msad090 (2023).
30. J. Sugimoto, M. Sugimoto, H. Bernstein, Y. Jinno, D. Schust, A novel human endogenous retroviral protein inhibits cell-cell fusion. *Sci. Rep.* **3**, 1462 (2013).
31. M. Marin, C.S. Tailor, A. Nouri, D. Kabat, Sodium-dependent neutral amino acid transporter type 1 is an auxiliary receptor for baboon endogenous retrovirus. *J. Virol.* **74**, 8085–8093 (2000).
32. R.N. Miyaho *et al.*, Susceptibility of domestic animals to a pseudotype virus bearing RD-114 virus envelope protein. *Gene* **567**, 189–195 (2015).
33. W. Wang, H. Zhao, G.Z. Han, Host-Virus Arms Races Drive Elevated Adaptive Evolution in Viral Receptors. *J. Virol.* **94**, e00684-20 (2020).
34. S. Suzuki *et al.*, Retrotransposon silencing by DNA methylation can drive mammalian genomic imprinting. *PLoS Genet.* **3**, e55 (2007).
35. H. Shiura, M. Kitazawa, F. Ishino, T. Kaneko-Ishino, Roles of retrovirus-derived *PEG10* and *PEG11/RTL1* in mammalian development and evolution and their involvement in human disease. *Front. Cell. Dev. Biol.* **11**, 1273638 (2023).
36. J.A. Frank *et al.*, Evolution and antiviral activity of a human protein of retroviral origin. *Science* **378**, 422–428 (2022).
37. A. Miyake *et al.*, Convergent evolution of antiviral machinery derived from endogenous retrovirus truncated envelope genes in multiple species. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2114441119 (2022).
38. J. Ito *et al.*, Refrex-1, a soluble restriction factor against feline endogenous and exogenous retroviruses. *J. Virol.* **87**, 12029–12040 (2013).
39. A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
40. P. Rice, I. Longden, A. Bleasby, EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
41. S.R. Eddy, Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
42. W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
43. K. Katoh, D.M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
44. S. Capella-Gutiérrez, J.M. Silla-Martínez, T. Gabaldón, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
45. B.Q. Minh *et al.*, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
46. D.T. Hoang, O. Chernomor, A. von Haeseler, B.Q. Minh, L.S. Vinh, UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
47. K. Kitao, T. Tanikaga, T. Miyazawa, Identification of a post-transcriptional regulatory element in the human endogenous retroviral syncytin-1. *J. Gen. Virol.* **100**, 662–668 (2019).
48. C. Camacho *et al.*, BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

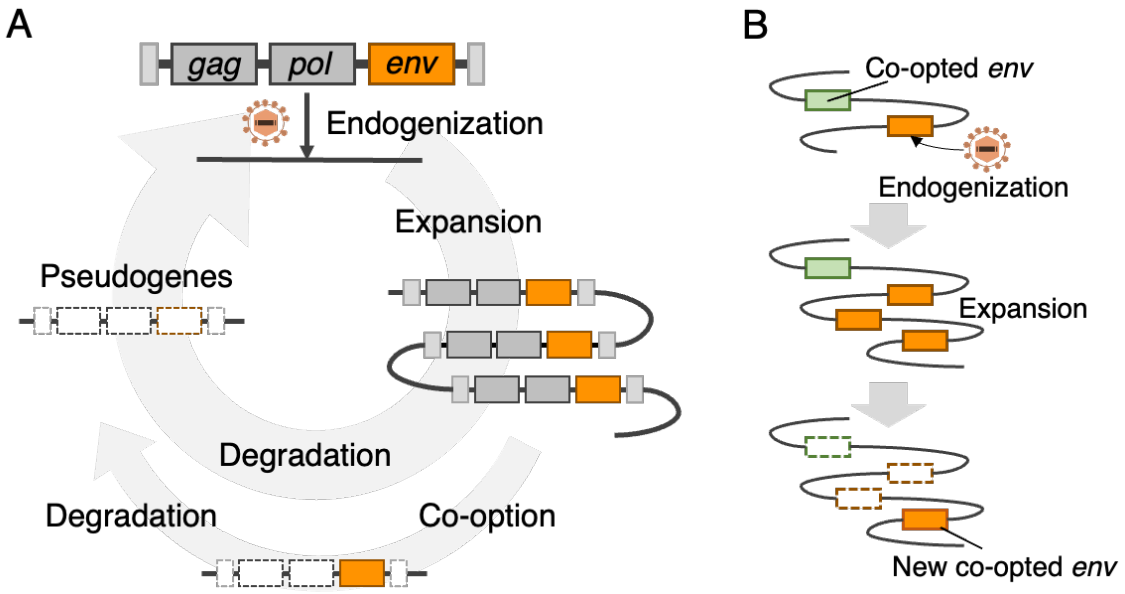


Fig. 1. Dynamic evolution of *env* genes in host genomes. (A) The cycle of copy number change of retroviral *env* genes. Once ERV is integrated into the host genome by infection to germ cells, its copy number will increase by retrotransposition. However, mutation causes loss of retrotransposition ability, and the copy number increase is halted. Most *env* genes and other retroviral genes evolve neutrally and lose coding capacity through the accumulation of mutations, becoming pseudogenes. A small number of *env* genes are co-opted and conserved. However, co-opted *env* genes are sometimes lost. (B) Scheme of the baton pass hypothesis. For simplicity, only *env* genes are shown. When a new *env* gene is introduced and expanded in the genome, the selection pressure on the existing co-opted *env* gene is relaxed. Then, one of the new *env* genes takes over the function and is evolutionarily conserved.

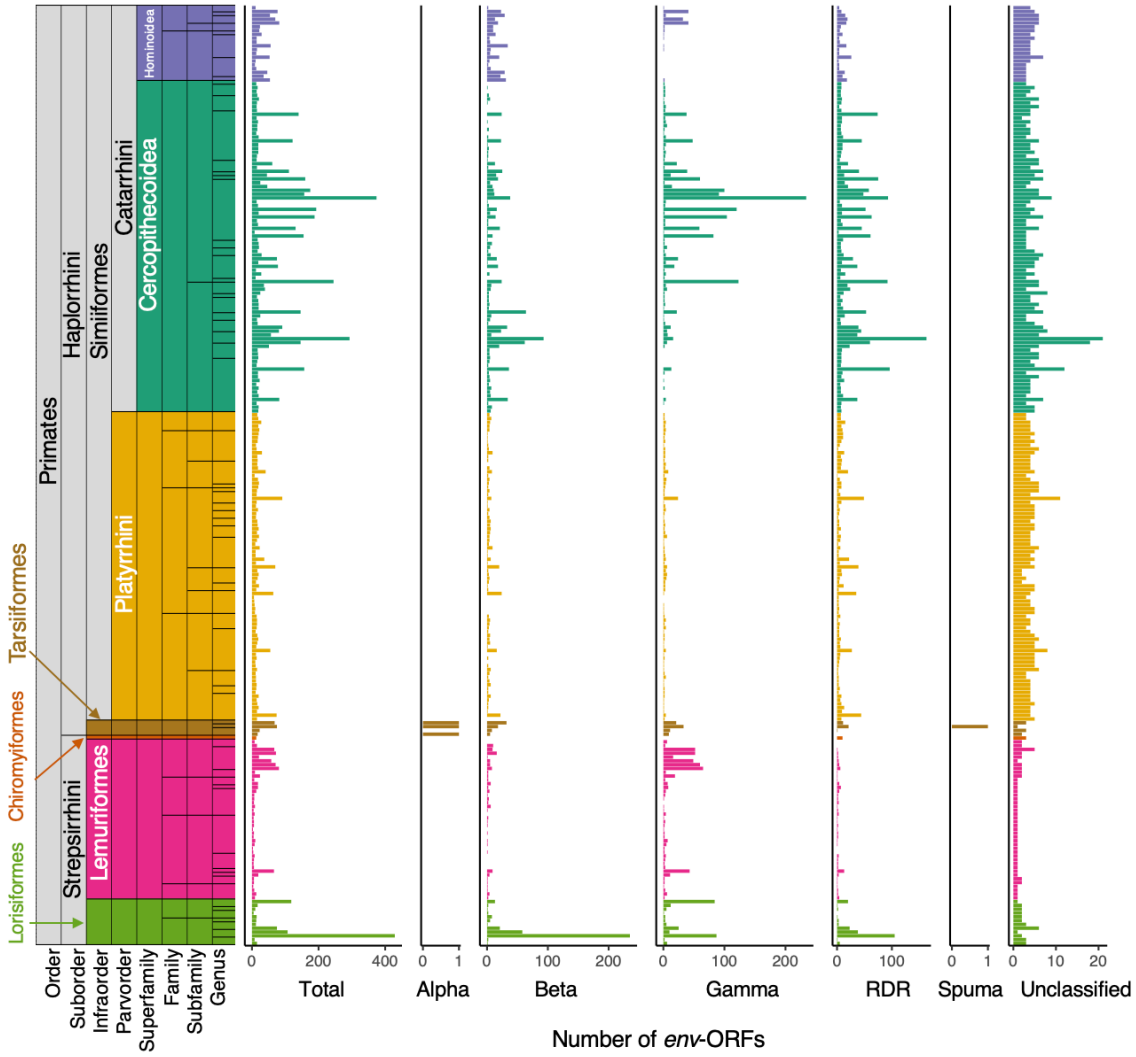
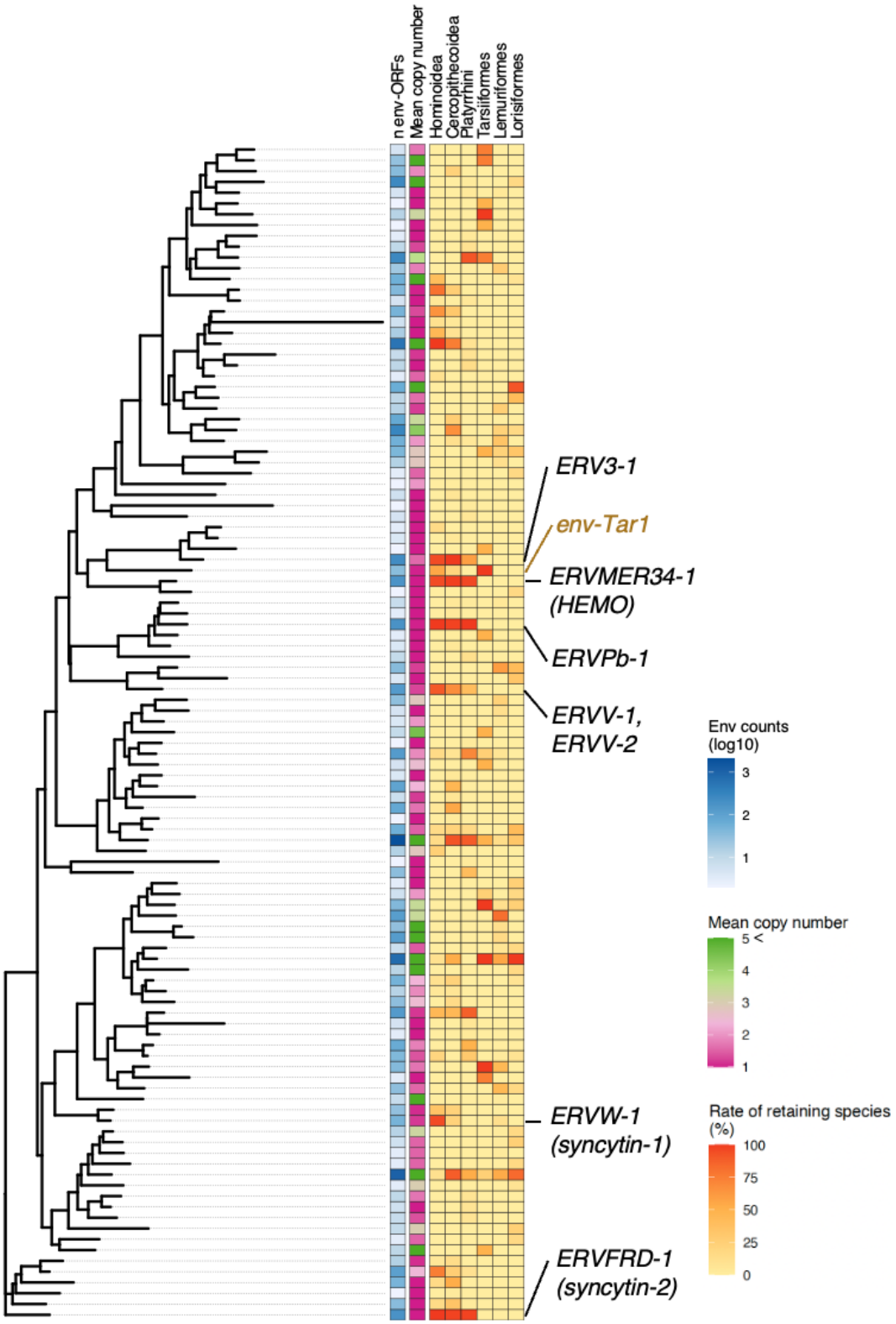


Fig. 2. *env*-ORFs in primate genomes. The number of *env*-ORFs in the primate genome assemblies. The *env*-ORFs were classified into 5 groups based on the best hit to retroviral Env proteins. The source data for taxonomy and *env*-ORF counts are available in **Dataset S3**.


n env-ORFs
Mean copy number
Hominoidea
Cercopithecoidea
Platyrrhini
Tarsiiformes
Lemuriformes
Lorisiformes
ERV3-1
env-Tar1
ERVMER34-1
(HEMO)
ERVPb-1
ERVV-1,
ERVV-2
ERVW-1
(syncytin-1)
ERVFRD-1
(syncytin-2)
Env counts
(log10)
3
2
1
Mean copy number
5 <
4
3
2
1
Rate of retaining species
(%)
100
75
50
25
0

460
461

Fig. 3. Clustering of *env*-ORFs. The 8,683 *env*-ORFs were grouped into 109 clusters based on the amino acid sequence similarity. The left maximum-likelihood-based phylogenetic tree is constructed from the amino acid sequences of representative *env*-ORFs of clusters. “Mean copy number” indicates the mean copy number of *env*-ORFs per genome assembly. “Rate of retaining species” indicates the proportion of species with at least one *env*-ORF in each taxonomic group. Clusters consisting of conserved single-copy *env* genes are expected to show a low “Mean copy number” and a high “Rate of retaining species”. The representative conserved *env* genes are labeled on the clusters, including them. The tarsier “*env-Tar1*” is described in **Figs. 4** and **5**.

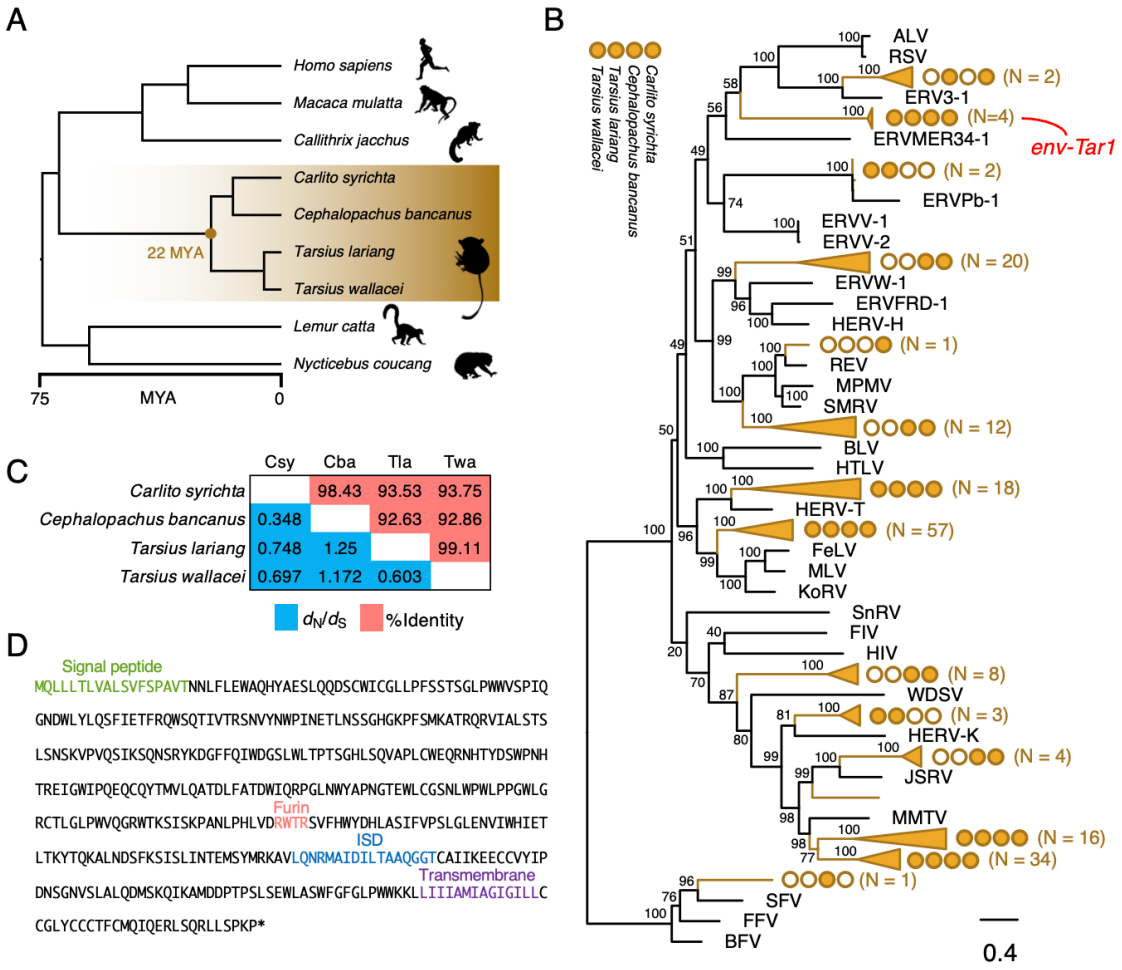


Fig. 4. Identification of a single orthologous *env*-ORFs in Tarsiiformes. (A) Species tree with estimated divergence time obtained from TimeTree (https://timetree.org/). (B) Maximum-likelihood-based phylogenetic tree reconstructed from representative retroviral Env proteins and tarsier *env*-ORF-encoded proteins. Clades containing tarsier *env*-ORFs were collapsed, with the number of sequences shown in parentheses. Filled and open circles indicate species in which the clades were found and not found, respectively. The *env-Tar1* was present in all four species, one copy each. (C) Pairwise amino acid sequence identities and corresponding d_N/d_S ratios among the *env-Tar1* ORFs. (D) Amino acid sequence of Env-Tar1 from *Carlito syrichta*. Characteristic retroviral Env features are highlighted. 60 aa was shown for each line.

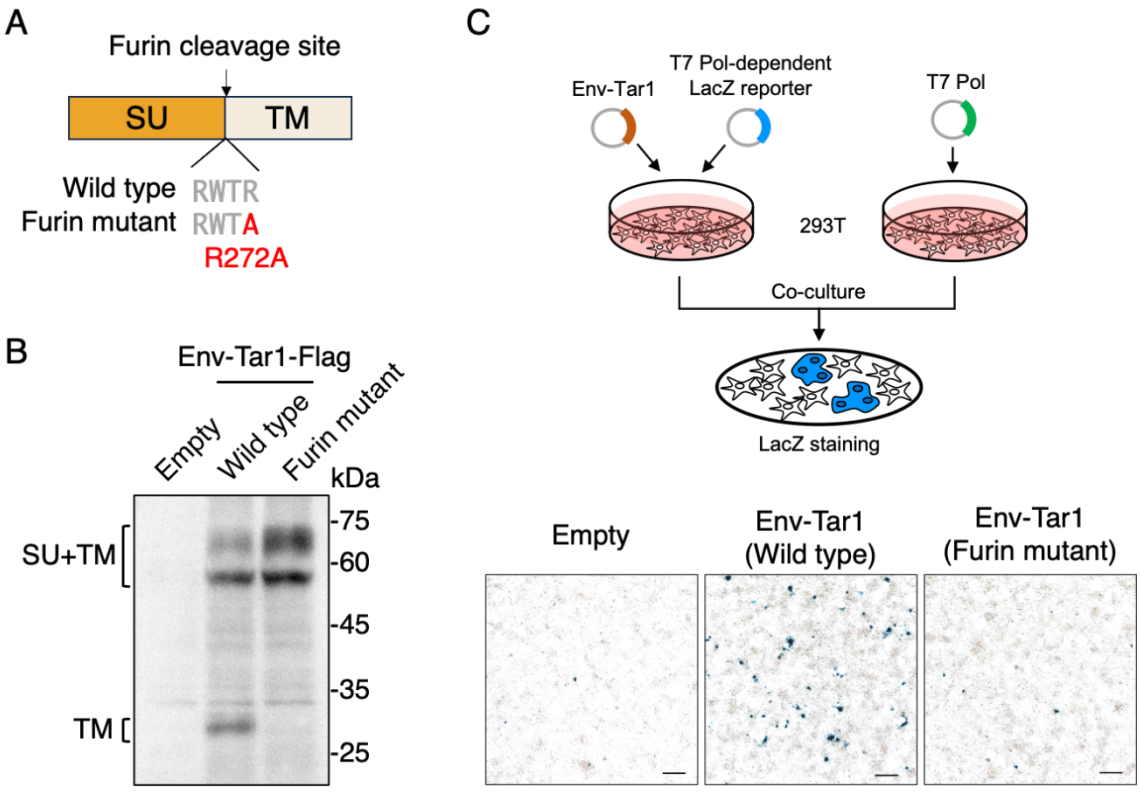


Fig. 5. The Env-Tar1 protein is a fusogen. (A) Env-Tar1 possesses a Furin cleavage site that separates the surface (SU) and transmembrane (TM) subunits. A point mutation was introduced into the Furin cleavage site to generate a negative control (referred to as the Furin mutant). (B) C-terminal Flag-tagged Env-Tar1 was expressed in 293T cells. The TM subunit was absent in the Furin mutant, indicating impaired cleavage. (C) Cell-cell fusion-dependent LacZ assay. In donor 293T cells, the Env-Tar1 expression plasmid was co-transfected with a reporter plasmid that expresses lacZ in the presence of T7 polymerase. Recipient 293T cells were transfected with a plasmid expressing T7 polymerase. After co-culture of donor and recipient cells, fused cells expressing LacZ were visualized by X-gal staining. The wild-type Env-Tar1 induced more LacZ-positive fused cells than the empty vector or the Furin mutant. The scale bars represent 100 μm.

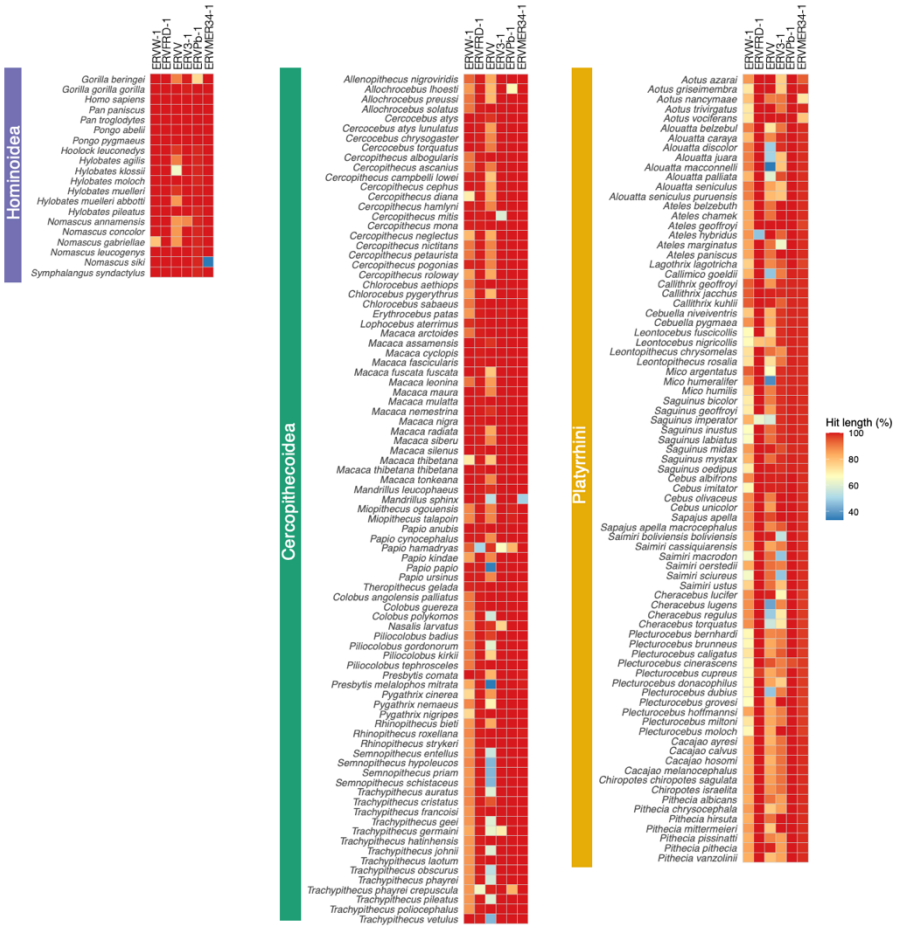


Fig. 6. Conservation of the human *env* gene in primate genomes. We performed tblastn against primate genome assemblies using the human *env* genes, which are known to be evolutionarily conserved, as queries. We then examined the coverage of each query *env* gene by the best hits. Query coverage in each genome assembly is shown as a heat map.