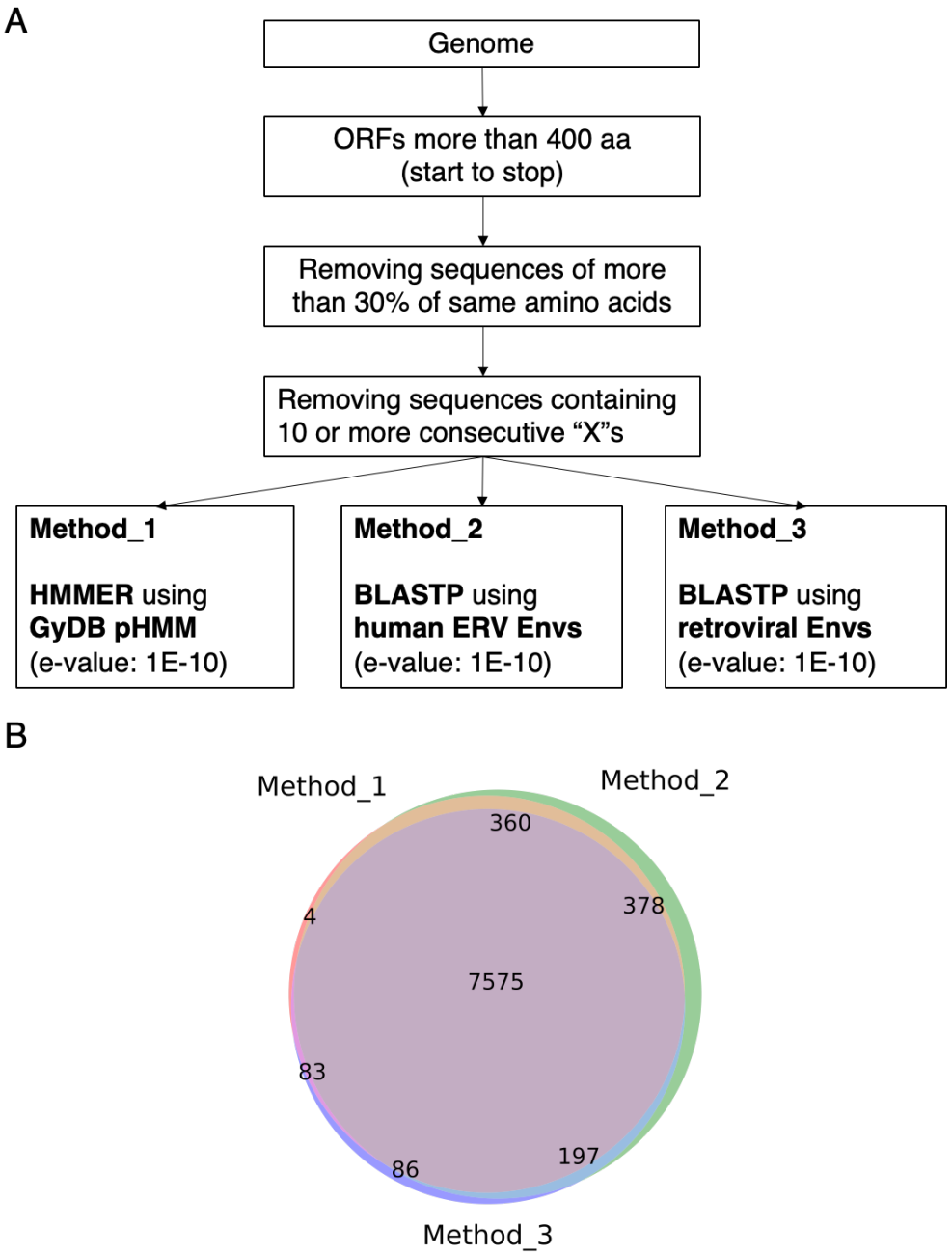

Fig. S1. Identification of *env*-ORFs in 247 primate genome assemblies. (A) Workflow for identifying *env*-ORFs in primate genome assemblies. We obtained *env*-ORFs by three methods of similarity search for retroviral reference databases. (B) Venn diagram illustrating the *env*-ORFs identified by three methods. In total, 8,683 *env*-ORFs were obtained.

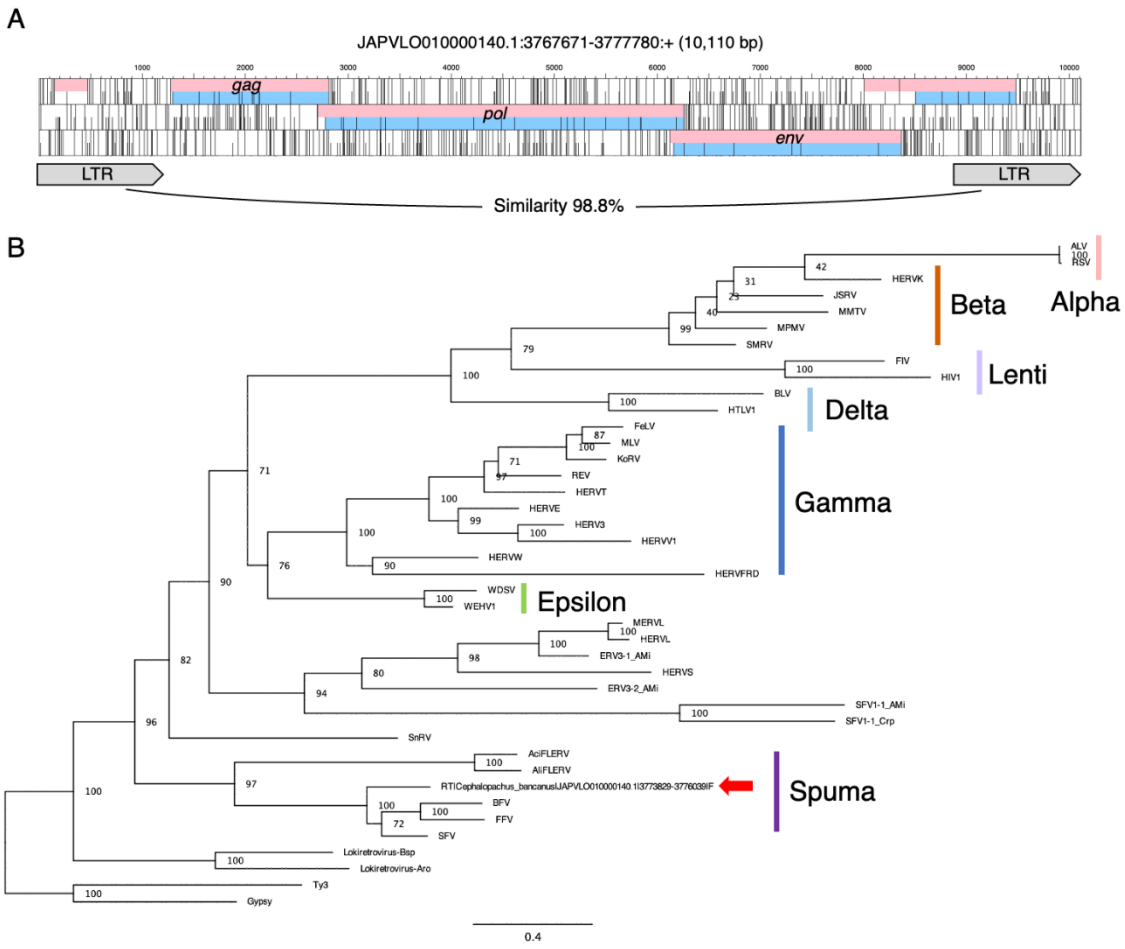

Fig. S2. The spumaretrovirus harboring the *env*-ORFs in *Cephalopachus bancanus*. (A) ORFs of the poroviral sequence. Short bars indicate start codons, and tall bars indicate stop codons. ORFs of more than 300 nt were filled. ORFs from a stop codon to a stop codon are shown at the top, and ORFs from an ATG triplet to a stop codon are shown at the bottom. ORFs of *gag*, *pol*, and *env* were indicated. 5′- and 3′-LTRs were filled with gray. (B) Phylogenetic tree of reverse transcriptase domains. The spumaretrovirus in *Cephalopachus bancanus* is indicated by an arrow.

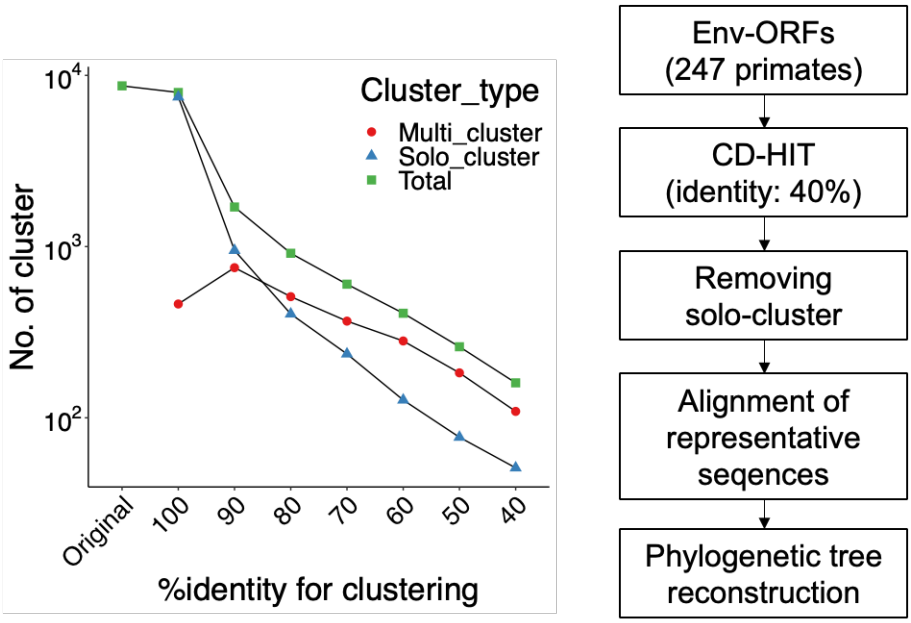

Fig. S3. Clustering of *env*-ORFs. The x-axis represents the sequence identity established via clustering. The y-axis represents the number of clusters and also shows the original number of sequences before clustering. The "Multi_cluster" is comprised of more than two sequences and the "Solo_cluster" consists of one sequence. A logarithmic scale was used on the y-axis. The phylogenetic tree of Figure 3 was constructed by representative sequences of clusters with 40% identity.

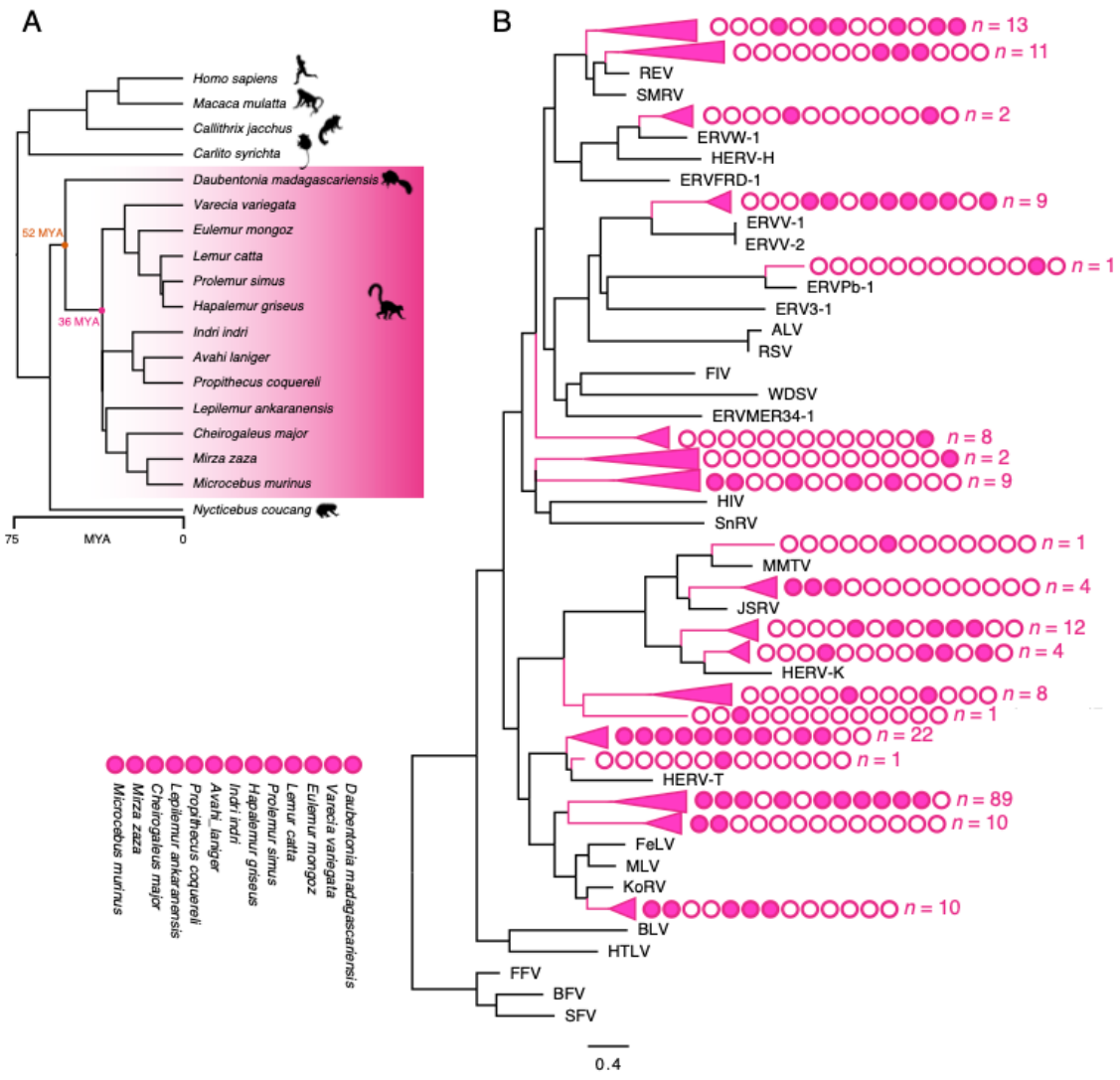

Fig. S4. Maximum-likelihood-based phylogenetic tree reconstructed from representative retroviral Env proteins and Env proteins encoded in env-ORFs of Chiromyiformes (1 species) and Lemuriformes (12 species). Clades with env-ORFs were collapsed with the number of sequences. Filled and open circles indicate species in which the *env* clades were found or not found, respectively.Table S1. Type or paste table title here. Paste table below the title.

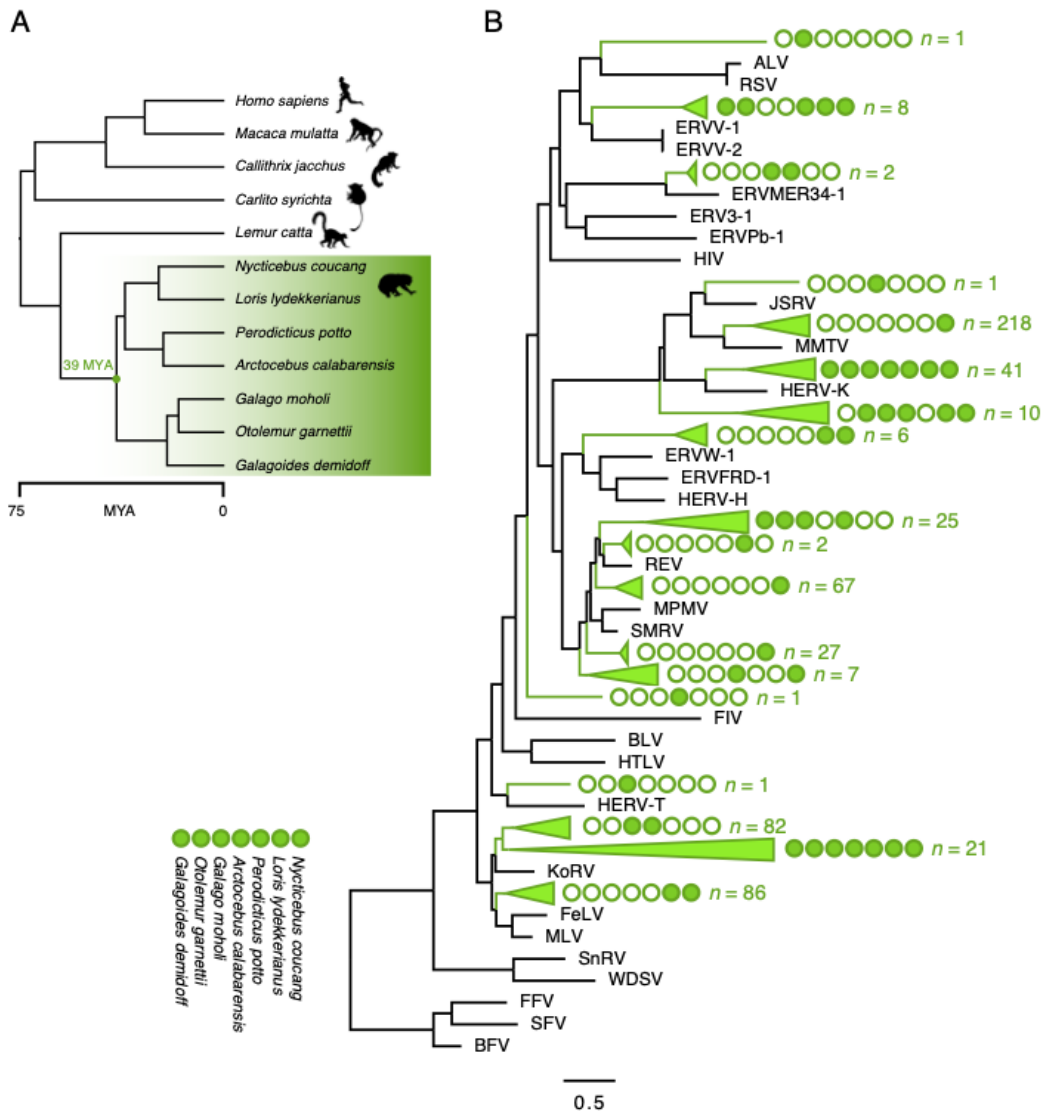

Fig. S5. Maximum-likelihood-based phylogenetic tree reconstructed from representative retroviral Env proteins and Env proteins encoded in env-ORFs of Lorisiformes (7 species). Clades with env-ORFs were collapsed and the number of sequences were indicated on the right. Filled and open circles indicate species in which the *env* clades were found or not found, respectively.

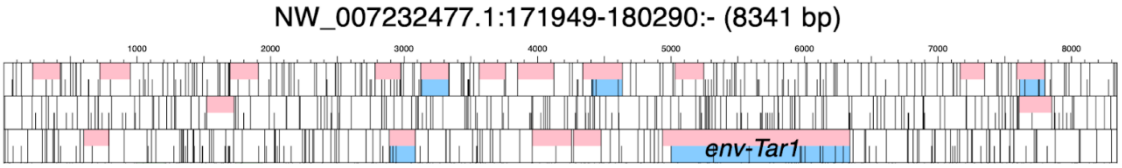

Fig. S6. Distribution of ORFs within 5 kbp upstream and 2 kbp downstream of the *env-Tar1* ORF in *Carlito syrichta*. Vertical lines indicate stop codons in each reading frame, while half-length vertical lines indicate start codons. ORFs longer than 180 codons are highlighted. The upper section (pink) shows ORFs spanning from stop codon to stop codon, while the lower section (blue) shows ORFs spanning from start codon to stop codon.

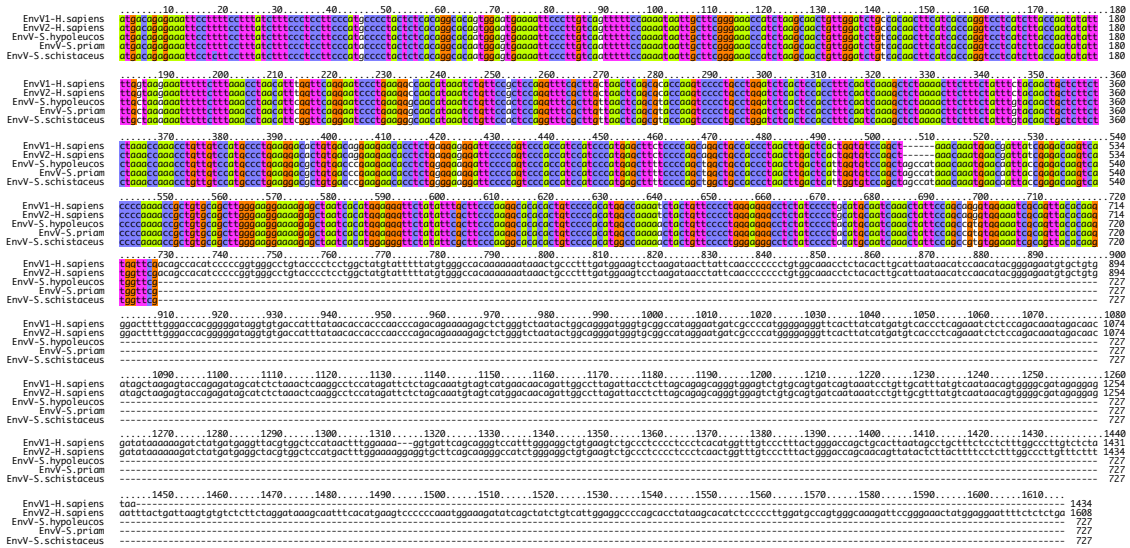

Fig. S7. Multiple alignment of nucleotide sequences of ERVV from humans and three *Semnopithecus* species. For the Semnopithecus species, each contig was obtained as follows: S. hypoleucos, CAUYNQ010114950.1; S. priam, CAUYPN010033499.1; S. schistaceus, CAUYPV010179808.1.

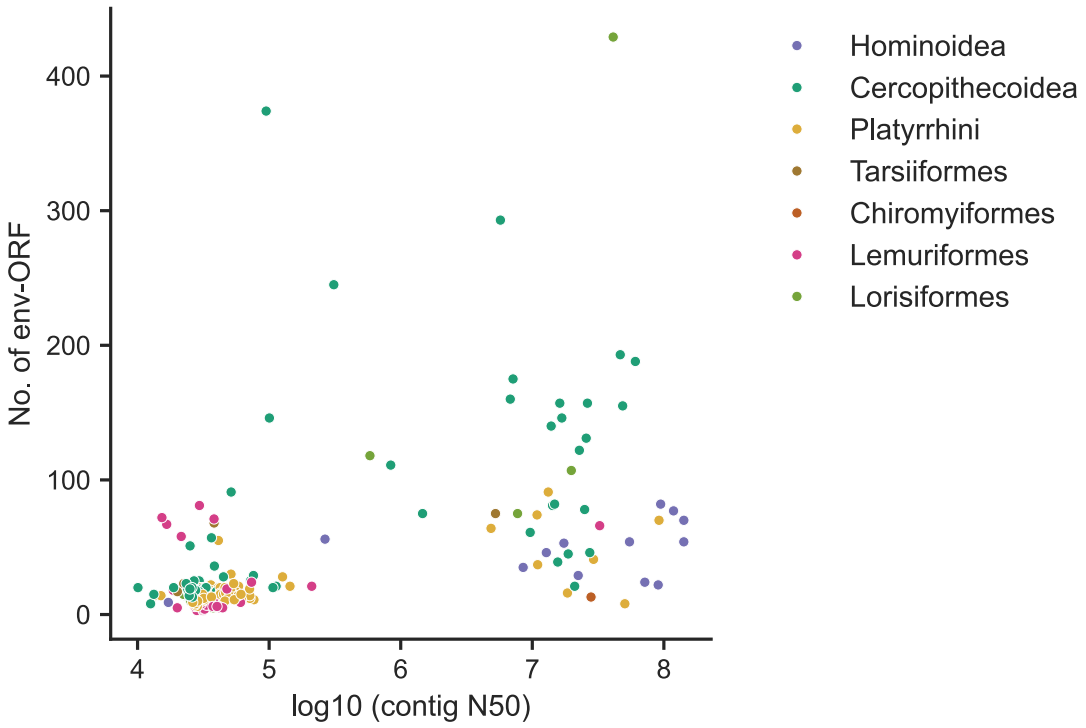

Fig. S8. A scatter plot illustrating the relationship between N50 length of contigs and the count of *env* ORFs (≥400 amino acids) across primate species.

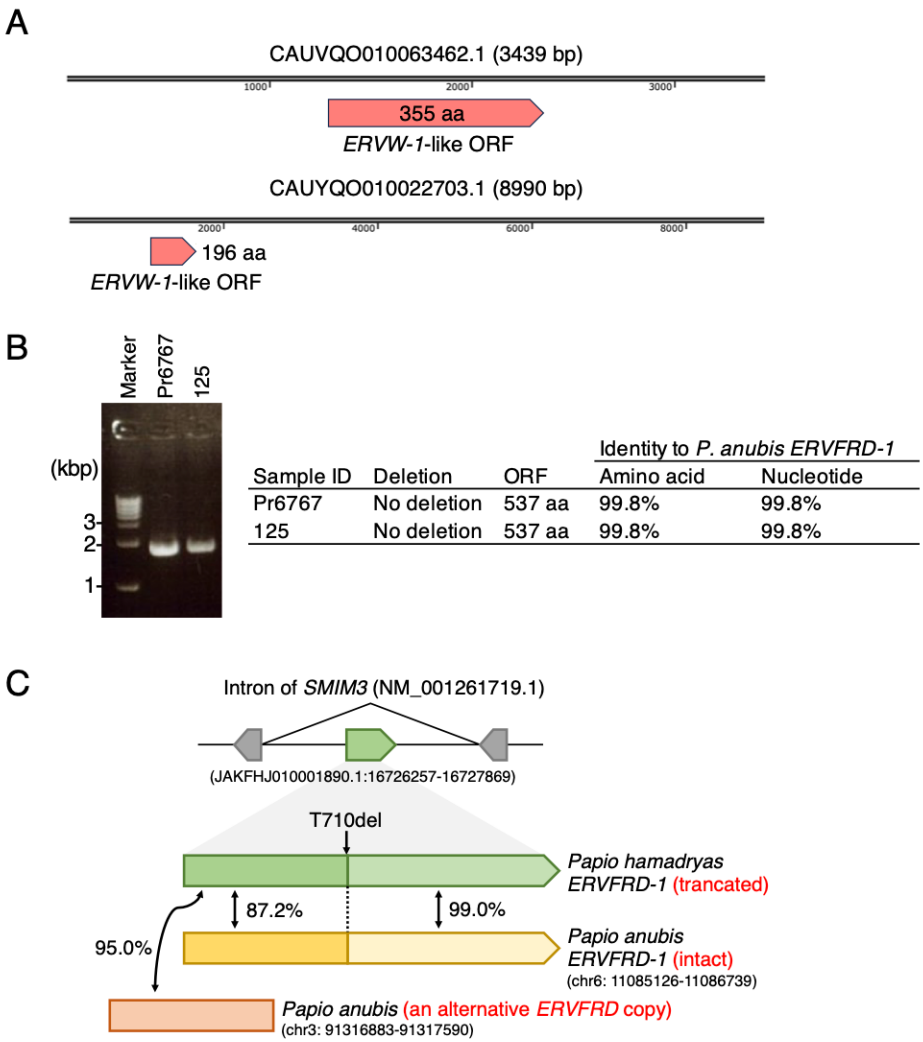

Fig. S9. Variation of *ERVW-1* (*syncytin-1*) and *ERVFRD-1* (*syncytin-2*) with truncated mutations in genome assemblies. (A) The intact ORF encoding Syncytin-1 was not identified in the *Nomascus gabriellae* genome assembly (GCA_963574555.1). *ERVW-1*-like sequences were still identified; however, these two ORFs were shorter than the intact *syncytin-1* ORF and were located in short contigs. (B) The *ERVFRD-1* encoding Syncytin-2 was truncated by mutations in the *Papio hamadryas* genome assembly (GCA_023781915.1). We performed PCR for *ERVFRD-1* from the *P. hamadryas* genomic DNA. The PCR products were electrophoresed on a 1% agarose and were subjected to Sanger sequencing to determine its nucleotide sequence. The amplified sequences encoded a complete ORFs for Syncytin-2. (C) Analysis of the genomic region corresponding to *ERVFRD-1* in the *P. hamadryas* genome assembly. A single-base deletion was identified within the *ERVFRD-1* ORF, resulting in the loss of the open reading frame. The sequences flanking this deletion were compared with those of the closely related species *P. anubis*.

Dataset legends

Dataset S1. List of 247 primate genome assemblies used in this study.

Dataset S2. Queries used for the env-ORF search.

Dataset S3. Annotations for env-ORFs. Species name, genome assemblies, location on the assemblies, and the hit queries.

Dataset S4. The number of env-ORFs for each genome assembly.

Dataset S5. Coverage of the best hits for human env-derived genes in the genome assemblies.

Dataset S6. List of 35 human amino acid queries used for the tblastn search.