Comprehensive genomic analysis reveals dynamic evolution of mammalian transposable elements that code for viral-like protein domains

Mahoko Takahashi Ueda[1], Kirill Kryukov[2], Satomi Mitsuhashi[3], Hiroaki Mitsuhashi[4], Tadashi Imanishi[2,5], So Nakagawa[1,2,5*]

[1] Micro/Nano Technology Center, Tokai University, 4-1-1 Kitakaname, Hiratsuka, Kanagawa 259-1292, Japan.

[2] Department of Molecular Life Science, Tokai University School of Medicine, 143 Shimokasuya, Isehara, Kanagawa 259-1193, Japan

[3] Department of Human Genetics Yokohama City University Graduate School of Medicine, Yokohama, Kanagawa 236-0004, Japan

[4] Department of Applied Biochemistry, School of Engineering, Tokai University, Hiratsuka, Kanagawa, 259-1292, Japan.

[5] Institute of Medical Sciences, Tokai University, 143 Shimokasuya, Isehara, Kanagawa 259-1193, Japan.

*To whom correspondence should be addressed:

so@tokai.ac.jp

Tel: +81-463-93-1121 ext. 2661

Fax: +81-463-93-5418

Keywords:

endogenous viral element, mammalian evolution, transposon exaptation, co-option, ERV, LINE

Abstract

Transposable elements of endogenous retroviruses and long interspersed nuclear elements code for viral-like protein domains, called endogenous viral elements (EVEs). A large proportion of EVEs lose their open reading frames (ORFs), while some maintain them and have been exapted by the host species. However, it remains unclear how ORF-possessing EVEs (EVE-ORFs) have evolved while developing new functions in the process. To address these issues, we examined 19 mammalian genomes and investigated characteristics of approximately 600,000 EVE-ORFs containing high-quality annotations of viral-like protein domains. The observed divergence of EVEs from their consensus sequences suggested that a large number of EVE-ORFs either recently, or anciently, inserted themselves into mammalian genomes. Alternatively, very few EVEs lacking ORFs, were found to exhibit similar divergence patterns. In addition, the EVE-ORFs were determined to be less abundant downstream of transcription initiation sites, and DNase I hypersensitive sites, which was in contrast to EVEs lacking ORFs, indicating that the purifying selection process may serve to suppress unexpected expression of EVE-ORFs. In spite of such observations, we identified more than 1,000 EVE-ORFs in human and mouse RNA-sequencing data. This suggested a possibility to identify and express uncharacterized EVE-ORFs hidden within mammalian genomes, most of which have been confirmed to be lineage specific. We also demonstrated that fractions of each EVE-ORF-encoded viral-like protein domain varied among mammalian lineages. Together, our analyses suggest that EVE-ORFs, many of which have not uncovered yet, may be co-opted in a host-species specific manner and are likely to have contributed to mammalian evolution and diversification.

Significance Statement

Transposable elements (TEs), known as jumping genes, occupy up to 70% of mammalian genomes. Recently, TEs encoding viral-like protein domains have been reported to have acquired new functions during evolution. However, it is

unclear how these TEs possessing open reading frames (ORFs) have evolved while simultaneously obtaining new functions within their host species. We discovered that segments of each viral-like protein domain varied among 19 mammalian species. Furthermore, through expression analysis in humans and mice, more than 1000 TEs were detected as transcripts. These results suggest that many newly identified mammalian TEs possessing ORFs, may have been co-opted in a lineage-specific manner, and served to contribute to mammalian diversification.

Introduction

Transposable elements (TEs), also known as “jumping genes”, constitute large portions of mammalian genomes. It has been reported that up to 70%, of the human genome is originated from TEs (1-3). Generally, TEs are categorized as junk DNA (4); however, many studies have shown that TEs have contributed greatly to mammalian evolution. Furthermore, TEs promote rearrangements of chromosomal DNA (5), and can become sources of both coding and regulatory sequences during the evolution of host genomes (6-9). In particular, specific types of long terminal repeat (LTR) retrotransposons, including endogenous retroviruses (ERVs), have been shown to develop the ability to function as genes in several mammalian tissues (10-14). One of the most well-studied TE-co-opted genes is syncytin. Specifically, the human syncytin-1 (*ERVW-1*) gene, which is derived from an *env* (envelope) gene in human endogenous retrovirus type W (*HERV-W*), was exapted for cell fusion during placental development (15-18). Interestingly, many of the molecular functions of TE-co-opted genes remain the same as those in viruses (13-15).

TE groups that code viral-like protein domains are made up of LTR retrotransposons, including ERVs, and long interspersed nuclear elements (LINEs). Such groups of TEs, which are also called endogenous viral elements (EVEs), occupy approximately 30% of the human genome (1-3). Many EVEs lose their open reading frames (ORFs) by accumulating deletions or mutations after integration. Thus, it is unclear what percentages of EVEs in mammalian genomes possess viral-like protein ORFs and have maintained high transcriptional potential. Thus far, genome-wide comparative analyses using epigenomic and transcriptomic data have examined the regulatory regions of ERVs (19), however have failed to characterize their protein-coding. Recently, many whole transcriptome sequencing (RNA-seq) datasets have been accumulated in public databases. One of the most comprehensive RNA-seq

datasets was generated by the genotype-tissue expression (GTEx) study using 31 human tissues (20). The data was used to assemble a similar pipeline in the Comprehensive Human Expressed SequenceS (CHESS) project and served to generate 20,352 potential protein-coding genes, and 116,156 novel transcripts in the human genome (21). A comparative analysis study focusing on the protein-coding region of EVEs using the new transcript data, in combination with cap analysis of gene expression (CAGE) data (22), and epigenomic data will provide new insight into the transcriptional potential, and functionality of these regions.

To understand the characteristics of possible protein-coding EVEs that possess ORFs for viral-like protein domains (EVE-ORFs) in mammalian genomes, we comprehensively examined EVE-ORFs in 19 mammalian species. We previously developed a database called gEVE (http://geve.med.u-tokai.ac.jp) containing more than half a million EVE-ORFs containing at least one functional viral gene motif found in 19 mammalian genomes (23). In this study, we analyzed EVE-ORFs in the gEVE database using various transcriptome and epigenomic data in humans and mice including the abovementioned data. Several systematic searches were performed to obtain EVE-ORFs at the domain level; these studies were designed to detect ERVs in the human genome (24-27), and LINEs in both humans and mice (28). However, most of these studies were limited to protein-coding EVEs in humans or mice, and primarily examined EVE sequences that contained nearly full-length ORFs or specific domains. Importantly, EVE-derived genes have also been described that do not possess intact ORFs yet continue to play important roles in specific situations (29). Moreover, multi-exon genes have been identified that contain exons partially derived from EVE sequences (30). Note that EVE-ORFs stored in the gEVE database do not always possess full-length original ORFs with lengths ≥80 amino acid (aa); however, all must contain

viral-like protein domains that were predicted using Hidden Markov models (HMMs). The identified ORFs were primarily from four ERV genes [*gag* (viral core proteins), *pro* (proteases), *pol* (polymerase), and *env* (envelope)] and one LINE gene (*orf2*), all of which are commonly found in viral amino acid sequences. In this study, we performed genome-wide comparisons on the number, divergence, genomic distribution, and transcriptional potential for each protein domain in the EVE-ORFs of mammals. Our findings are expected to improve the understanding of the evolution and potential roles of unannotated EVE-derived genes.

Results

Characteristics of EVE-ORFs in 19 mammalian species

To characterize EVEs that are expressed as proteins in mammalian species, we first compared 614,488 possible protein-coding EVEs -ORFs from 19 mammalian species obtained from the gEVE database (23). EVE-ORFs include ORFs that contain ≥80 amino acid (aa) residues and are derived from protein domains in retrovirus-like genes, such as, *gag*, *pro*, *pol*, and *env*, or from a LINE *ORF2* gene that contains a virus-like motif of reverse transcriptase (RT). The number of EVE-ORFs vary among mammalian genomes (Figure 1A), as we have reported previously (23). In each genome, the percentage of possible protein-coding ERVs and LINEs (ERV-ORFs and LINE-ORFs, respectively) relative to the total ERVs and LINEs, were found to range from 0.05% to 0.15% (Figure 1A). Specifically, within mice and cows, the proportion of detectable EVE-ORFs was higher compared to other species. Moreover, the proportion of EVE-ORFs was significantly correlated with the total EVE genome length (ERV-ORF: $r = 0.71$, $p < 5.51 \times 10^{-4}$; LINE-ORF: $r = 0.46$, $p < 0.045$). Distributions of the ERV-ORF and LINE-ORF lengths among mammals are shown in Figures S1A-S1C.

We further compared mammalian ERV-ORFs at the level of gene domains. To reduce the influence of genome qualities on the number of domains detected, we calculated the relative fraction change (ΔF) from the mammalian average for each domain fraction (Figure 1B). The most species-specific characteristics observed in the ΔF values were determined to be in the *gag* gene, which exhibited a ΔF of more than ±1.5 in several species (Figure 1B). In addition, the *pro* gene, in mice expressed a dramatic increase ($\Delta F = 1.8$). Furthermore, platypuses showed quite unique compositions in their *gag* and *pro* genes with very small ΔF values (0.14 and 0.12, respectively). Alternatively, the

ΔF values for *pol* were similar among mammalian species, all of which had expression levels that were determined to be within ±1.5 of each other. Interestingly, although no species showed ΔF values greater than ±1.5 in the *env* gene, the values were relatively higher within non-human primates, specifically within rhesus macaque, which had a ΔF of nearly 1.5.

We next examined which TE groups from the Repbase database (31) were enriched in human and mice EVE-ORFs (herein we employ the term "group" rather than the Repbase classification of "clades" for LINEs, and "families/sub-families" for ERVs to avoid confusion with species classification). A large proportion of the detected ERVs in humans were found to be HERV-H and HERV-K, and IAPE (Intracisternal A-type Particles elements with an Envelope) in mice (Figure 1C). HERV-H represents the ancient HERV group, which was inserted into the common ancestral genome of simians and prosimians (32), while HERV-K is described as containing younger HERVs (33,34). In addition, the IAPE in mice is one of the ERV groups that continue to be active in mice (35). The most prominent LINE-ORFs were derived from L1, specifically members of the L1PA primate retrotransposon, which contain more recently identified families of L1PA2-6 in humans, and L1Md_T and L1Md_F in mice as well as older members, including L1Mus1 and L1Mus2 also in mice (36; Figure S2).

We identified divergent fragments of EVE-ORFs using the Repbase consensus sequences, and compared them with EVEs, excluding EVE-ORF regions (non-EVE-ORFs). We hypothesized that newer elements would be closer to the consensus, and older ones would be further. Thus, we were able to estimate the age of EVE-ORFs by comparing the divergence spectra to their consensus sequences. We first determined modes of EVE divergence based on kernel density estimates (KDE). For non-ERV-ORFs and non-LINE-ORFs, the modes of divergence were found to be approximately 20% and 30% in all mammalian genomes, respectively (Figures 2 and S3). However, the modes of

divergence for EVE-ORFs were quite different from those of non-EVE-ORFs. The ERV-ORFs showed clear bimodal distributions in many mammalian species, with modes of approximately 2~7% (younger) and 30~33% (older). It should be noted that in specific species the distribution patterns revealed weak or biased bimodality. Specifically, the bimodality was found to be weak and biased in dogs, sheep, and opossum, and small divergences were observed in humans. Although these distribution patterns were relatively similar among closely related species, such as non-human primates and rodents, overall, the distributions varied significantly between species (Figure 2). Furthermore, differences were noted in the divergence distributions between gene domains (Figure S3). Specifically, within the *pro* domain of primates (save for marmosets), rodents, and horses a high frequency of low divergence distribution was observed compared to that of other domains. Further, the *pol* domain exhibited a higher frequency of the older mode. Unlike the ERV, the modes of divergence in both LINE-ORFs and non-LINE-ORFs were determined to be unimodal (Figure S4). The mode values for LINE-ORFs were quite small compared to those for non-LINE-ORFs (<9%), which were relatively consistent across all examined mammals.

Finally, we determined how many EVE-ORFs act as raw materials for the assembly of multi-exon genes. Comparison of our observed EVE-ORFs with those in Ensembl gene and protein annotations revealed that 13, and 24 EVE-ORFs were constituents of ordinary multi-exon genes in humans and mice, respectively (Figure S5 and Table S1). It is also intriguing that large numbers of these multi-exon genes were found in horses (180 genes), sheep (337 genes), and opossum (737 genes). Specifically, within horses and sheep, the number of multi-exon genes containing EVE-ORFs, as predicted by HMM not RepeatMasker (3), were quite large (25: horses, 18: sheep, Figure S5).

Overlap between various genomic features

To predict whether EVE-ORFs are expressed as proteins, we performed comparative analyses using a wide-range of transcriptome data generated by next-generation sequencing. We compared observed EVE-ORFs from our study with quantified transcriptome data derived from 31 human tissues generated in the GTEx study (20) from the CHESS database (21). We found that a total of 343 EVE-ORFs (282 in ERVs and 61 in LINE-ORFs) overlapped with transcripts in the CHESS database. Of these, 7 EVE-ORFs were found to correspond to genes containing exons derived from EVE sequences predicted only by HMM, not by RepeatMasker (Table S1). The proportion of ERV-ORFs overlapping with CHESS transcripts in each domain were 3.0, 2.9, 3.5, and 4.7 in *gag*, *pro*, *pol*, and *env*, respectively. This suggests that the *env* segment was relatively large compared to the other domains, however, this result was not statistically significant (chi-square test $p < 0.16$).

Since EVEs are exclusively expressed during embryogenesis and cell differentiation in all mammals (37, 38), the RNA-seq data acquired by the GTEx project is insufficient to obtain accurate information on EVE-ORFs because it only analyzes samples taken from fully differentiated human tissues. We, therefore, also analyzed RNA-seq data from human and mouse differentiating myoblasts. The relevant expression data from human primary myoblasts (12 runs) and mouse C2C12 cells (16 runs) were collected from the sequence read archive (SRA) database (see Supplementary Table 2 for details). Following gene mapping and quantification, 170 and 694 EVE-ORFs were identified as having at least 10 normalized read counts in the human and mouse myoblasts, respectively. Furthermore, the RNA-seq data were analyzed from 4 (Days 0-3) and 3 (Days 0, 3, 6) time points after the differentiation had begun in humans and mice, respectively. Principal component analysis (PCA) plots successfully captured the difference in

EVE-ORF expression at each time point (Figures 3A and S6). It is noteworthy to mention that in EVE-ORFs expressed during human muscle cell differentiation, only 37 of the 170 (21.8%) were also identified in GTEx transcripts. Furthermore, in humans, the number of expressed ERV-ORFs and LINE-ORFs were relatively similar (Figure 3B). Further, some EVE-ORFs exhibited high expression levels throughout the entire differentiation process, many of which were identified as *env*-derived transcripts (Figure 3B). Alternatively, in mice, there were significantly higher levels of ERV-ORFs expressed compared to LINE-ORFs. For example, while only 86 ERV-ORFs (50.1% of the total EVE-ORFs) were detected in humans, 602 ERV-ORFs (86.6%) were detected in mice (Figure S7).

To further investigate whether EVE-ORFs are expressed as mRNAs, we examined the relationship between the presence of transcription start sites (TSS) and EVE-ORFs using the cap analysis of gene expression (CAGE) data in the FANTOM5 database (22). CAGE data are derived from 1,816 human samples from various tissues, primary cells, and cell lines. Using the data, we assessed the relationship between ERV-ORFs located within 10 known EVE-derived single-exon genes and the closest TSSs in humans and mice; these ERV-ORFs were located approximately 7,000 bp from the TSS (Table S1). Four ERV-ORFs were located within 1,000 bp of the TSS; however the rest, specifically those within the *pol* and *env* domains, were found 1,821 bp to 7,086 bp downstream of the TSS (Table S1). Within mice, the ERV-ORFs demonstrated a similar pattern as that observed in humans (Table S1). This suggests that the structure of EVE affects the location at which it integrates into the host genome, and vice versa. We applied this analysis to all our EVE data sets, and further identified EVEs with or without ORFs that were located downstream of the TSS. We first calculated the number of EVE-ORFs and non-EVE-ORFs found within each bin of 1,000 bp from the TSS, and found that

the percentage of EVE-ORFs was smaller than that of non-EVE-ORFs (Figures 3C and S6). Interestingly, the number of ERV-ORFs and LINE-ORFs appearing close to TSSs varied significantly in humans; specifically, there were significantly fewer ERV-ORFs and LINE-ORFs located within 19,000 bp, and 2,000 bp of the TSS compared to non-ERV-ORFs and non-LINE-ORFs, respectively. Comparatively, in mice, the distance from TSSs that exhibited reduced expression of EVE-ORFs was relatively small compared to those in humans; ERV-ORFs and LINE-ORFs were markedly reduced within 9,000 bp and 3,000 bp of TSS, respectively (Figure S8). Similar results were achieved following analysis using DNase I hypersensitive sites (DHSs) from 1,511 human and mice cells, obtained from the ChIP-Atlas database (39). DHSs are genomic indicators for accessible chromatin regions, which is directly related to transcriptional activity (40-42). Further, TSSs and potentially all expressed genes are marked by DHS (42). In both humans and mice, the proportion of EVE-ORFs located within a DHS was found to be significantly less than for non-EVE-ORFs [chi-squared < 0.01, false discovery rate (FDR) corrected, Figure S9].

We next investigated which cell types contain active TSSs located upstream of EVE-ORFs in human expression profiles within CAGE data in the FANTOM5 database. We performed PCA analyses using profiles from 38 different cell lines or tissues types, all of which had a minimum of 10 available samples for analysis, and found that these active TSSs located near EVE-ORFs were segregated into clearly defined groups for different cell types. Principally, in the expression profile of ERVs, embryonic stem (ES) cells, embryoid body (EB) cells, and induced pluripotent stem (iPS) cells were separated into 3 groups (Figure 3D). Conversely, in the LINE-ORF profile, these cell lines were separated but were not clearly demarcated (Figure 3D). Distinct groups were also identified in the profiles of EVE-ORF in osteoblasts, blood, and brain samples. Specifically, within the ERV profile, osteoblasts. Alternatively, in the

LINE-ORF profile, blood cells formed a distinct group spanning a large range, and was partly overlapped by osteoblasts. Moreover, human body tissue samples such as, ovary, muscle, and retinal pigment epithelium (RPE) formed individual groups, many of which overlapped with each other, and merged to form one large group in the profile of ERVs (Figure 3D). This tendency became much more pronounced in the LINE-ORF profile with body tissues, excluding brain samples, not being clearly separated into distinct groupings (Figure 3D).

The Repbase annotations for EVE-ORFs found near FANTOM TSSs or within CHESS transcripts were identified as different from those of all other EVE-ORFs in the human genome (Figure 3E). In the ERV-ORF near the TSS, there was an enrichment of HERV-E and HERV-H compared to all other ERV-ORFs. Conversely, HERV-H was significantly under-expressed in the ERV-ORFs detected in the CHESS database compared to all other ERV-ORFs. Similarly, differences were observed in the Repbase annotations for the ERV-ORFs located near TSSs, and within CHESS transcripts. For example, high levels of HERV-E were observed in ERV-ORFs from the CHESS database, however, were not seen in ERV-ORFs found near TSSs. Conversely, HERV-H was found to be highly expressed in FANTOM TSSs compared to those detected in CHESS datasets. FANTOM datasets contains a larger variety of samples, including cell lines such as iPS, ES, and cancer cells, while CHESS datasets are obtained from human-tissues alone. The observed differences in the annotations of EVE-ORFs may result from differences in sample types used in each project. We thus limited the FANTOM TSSs located within 2,000 bp of EVE-ORFs, to those that were determined to be active in human tissues alone. As a result, the fraction of HERV-E increased from 9.9% to 14.6%, and the fraction of HERV-H decreased from 27.0 to 16.3% (Figure S10) in the Repbase annotation of EVE-ORFs located near active tissue TSSs, which were found to be similar to those of EVE-ORFs detected using the CHESS database.

Our analyses, thus far, using expression data such as RNA-seq and CAGE data, have shown that only a small fraction of EVE-ORFs become transcribed, which suggests that the remaining EVE-ORFs may not be transcribed. To determine if this is true, we examined genomic regions containing EVE-ORFs and compared them with the same regions in “ordinary” genes. In a cell nucleus, chromosomes occupy their own territories (43, 44) and recent studies on 3D genome structure have revealed that chromosomes in the mammalian genome are compartmentalized into active (open: A-type) and inactive (closed: B-type) chromatin domains (45), which are called topologically associated domains (TAD) (46, 47). Recent advances in high-resolution TAD maps allowed for further organization of human chromatin into six smaller contact domains, which are associated with distinct patterns of the histone markers; A1, A2, and B1 to B4 (48). Both A1 and A2 are gene-dense, and actively transcribed regions, however, A1 replication is complete at the beginning of S phase, whereas A2 continues replicating into mid S phase. Furthermore, B1 is composed of facultative heterochromatin and replicates during the middle of S phase. Histones B2 and B3, however, do not being replication until the end of S phase. B2 contains pericentromeric heterochromatin and is enriched in the nuclear lamina and nucleolus-associated domains. B3 is enriched only in the nuclear lamina. In addition, subcompartment B4 contains many Krüppel associated box-containing zinc-finger repressor proteins (*KRAB-ZNF*) superfamily genes, which exhibit a highly distinct chromatin pattern containing both active and heterochromatin-associated markers. All six of these subcompartments were detected using human lymphoblastic cells, and the contact map was also found to be often conserved among human cell lines and in mice (48).

Using the genome structural information from human lymphoblastic cells as described above, we examined the genomic regions where EVE-ORFs

are located in the human genome. We calculated the fraction of EVE-ORFs found in the six histone-based subcompartments and compared them with those in Ensembl genes (Figure 3F, top). For comparison, we examined an additional 30 known genes that contain EVE-ORFs (EVE genes), which were also detected in our EVE-ORF data (see list in Table S1). We found that the distribution of EVE-ORFs in each subcompartment differed from that of Ensembl genes as well as EVE genes, and the distribution of Ensembl and EVE genes also differed from each other (Figure 3F). Moreover, large fractions of Ensembl genes were found to be located within sub section A1, while most of the EVE genes were located in A2 and B2. Unlike these known genes, large segments of EVE-ORFs were found in B3. The distribution of ERV-ORFs and LINE-ORFs was similar compared to those of known genes. However, when compared with non-EVE-ORFs, distribution patterns for EVE-ORFs occurred in opposite directions in all subcompartments save for in B1 (Figure 3F, middle and bottom). Further, all observed ERV-ORFs and LINE-ORFs were significantly enriched in A-type and B-type states when compared to non-ERV-ORFs and non-LINE-ORFs, respectively. Also, at the gene domain level in the ERV-ORF, the *gag* domain was found to be significantly enriched in A1, and depleted in B3 when compared to the expected values that were based on non-EVE-ORF fractions ($p < 0.001$, FDR corrected) (Figure S11). Conversely, *pol* domains were significantly depleted in A1 and enriched in B3 ($p < 0.01$, FDR corrected), while *env* domains were found to be depleted in B3 ($p < 0.05$, FDR corrected). Alternatively, the *pro* domains showed no significant change in any subcompartments. However, similar to the *pol* genes in ERV-ORFs, LINE-ORFs were enriched in A1 and depleted in B3 ($p < 10^{-5}$, FDR corrected).

We further compared the genomic loci of EVE-ORFs and ordinary genes within the human genome. We plotted the number of EVE-ORFs in a bin size of 500 kb on the human genome (Figure S12). Results show that the

EVE-ORF distributions in the human genome were significantly different from those of ordinary genes obtained from the Ensembl database (Figure S12C). Genomic regions containing a high density of ERV-ORFs were also found to be different from those of LINE-ORFs. Ten high-density bins were identified containing ERV-ORFs in chromosomes 1, 4, 7, 8, 19, and Y (Figure S12A). However, LINE-ORFs were more uniformly distributed in all chromosomes, and showed only 4 high-density bins. Most significant, LINE-ORF enriched regions were found in chromosome X (Figure S12B).

Cluster analysis of EVE-ORFs

To understand long-range relationships between mammalian EVEs, we conducted a genome-wide cluster analysis on EVE-ORFs using CD-HIT (49). Since the current annotations by Repbase are based on nucleotide identity, analysis of the amino acid sequences enabled us to better understand the relationship between EVE sequences. Cross-species EVE clustering also provided important predictions for the number of sequence types (different amino acid sequences) in the 19 examined mammalian species, as well as deeper evolutionally relationships. The CD-HIT program clusters sequences by finding the longest reference sequence in each cluster group, and performs subsequent calculations to identify all sequences that contain the reference sequence (49). This clustering step was repeated in a round-robin fashion for each EVE domain category. After clustering EVE-ORF sequences, we obtained 123,035 to 4,092 clusters at the levels of ≥90% to ≥30% sequence identity, respectively (Figure 4A). The fewest unique sequences, and thus the highest level of clustering, were identified in the LINE-ORF analysis (Figure 4A), which suggests that the sequence similarity between LINE-ORFs was quite high. This is congruent with the results obtained through divergent analysis (Figures 2 and S3). Furthermore, among ERV genes, the number of clusters identified for *gag*

genes (Figure 4A) were more variable, while that for *pro* genes was less variable. We assigned cluster IDs for each cluster by corresponding the number of sequences in the cluster with the length of the reference sequences (Figures S13 and S14). In our analyses, cluster IDs below 200 generally contained long EVE-ORF sequences (≥ 500 aa in length).

The cluster analysis identified many clusters shared between multiple species, which had not been detected by clustering of nucleotide EVE sequences. Many clusters were shown to be lineage specific; at ≥60% identity, many clusters were found to be shared between primates, rodents, or bovids, whereas less, or almost no, shared clustering occurred in opossum and platypuses (Figure 4B). Using the clusters and human RNA-seq data, we determined the degree of shared EVE-ORFs detected in human RNA-seq by examining whether the EVE-ORFs were derived from human, primate, or multi-species clusters. Many of the human EVE-ORFs found in transcripts in the CHESS database were also found in clusters of primates (apes and old-world monkeys) at the level of ≥60% identity (Figure 4C). Furthermore, more than half of the ERV clusters contained ≥100 sequences from various mammalian species that contained ERVs found in the CHESS database. Overall, the clustering of *gag* and *env* domains were shared between similar species; many of which were shared among closely related species such as primates, while other *gag* clusters were shared among distantly related species. Additionally, the *pro* domain cluster, containing transcripts found in the CHESS database, was relatively small, and appeared to be more specific to apes compared to other gene categories. By contrast, LINE-ORFs found within clusters containing transcripts from the CHESS database, were mostly shared among a broad range of mammals including even-toed ungulates, carnivores, and primates. The clusters that contained human EVE-ORFs and overlapped with CHESS transcripts with known annotations as EVE genes, are highlighted in Figure 4C. (red triangles in

the figure). Many of these genes, save for syncytin-1, were found in clusters made up of a smaller number of sequences (≤100). Of the total 33,966 human EVE-ORFs examined in this study, only 223 sequences were human-specific (un-clustered unique sequences or human-specific clusters). Moreover, in the EVE-ORFs overlapping with CHESS transcripts, only 4 sequences were found to be human-specific ones.

To understand the relationship between sequence similarity and functionality in each cluster, we analyzed a cluster containing *ERVW-1* that encodes syncytin-1 as a representative of a functional EVE gene. We extracted 44 unique EVE-ORF sequences with ≥30% similarity and that were ≥250 aa in length from the cluster containing *ERVW-1* gene (Figure 4C, *env* panel), and generated a maximum likelihood phylogeny by using the amino acid sequences using RAxML (50: Figure 4D). Syncytin-1 has been found only in apes, and thus *ERVW-1* formed a small clade with apes, not with monkeys (Figure 4D). Additionally, the branch length for apes within the syncytin-1 clade were shorter compared to other clades in the tree, which may be a sign of evolutional conservation of sequences. We also found another small clade containing short branch lengths among apes. The human EVE-ORF in this clade was also identified in the CHESS database as well as via RNA-seq myoblast differentiation (Figure 4D, red circle). Note that other EVE-ORFs overlapping with CHESS transcripts had no clade at this specific cut-off, but it did have a similar EVE-ORF in chimpanzees when the length cut-off was relaxed (the length of chimpanzee EVE-ORF was 229 aa), and exhibited a short branch length. In addition, we examined the relationship between the human EVE-ORFs in the phylogeny and TSSs using FANTOM5 datasets, and found that 4 human EVE-ORFs were located 2,000 bp downstream of the TSSs (Figure 4D, purple circle).

Many human EVE-ORFs identified via RNA-seq analysis were found to

be shared among primates, while some EVE-ORFs were found in multiple shared clusters. Indeed, of the clusters at the level of 60% identity, we identified 156 clusters that were shared among ≥8 mammalian species (Figure 5). Approximately half of these shared clusters contained LINE-ORFs (79 clusters). The second largest shared EVE-ORF domain category was for the *pol* gene, which had 44 clusters shared between species. The remaining ERV gene domains contained a similar number of cross-species clusters (12, 11, and 10 for *gag*, *pro*, and *env*, respectively). Although 12 clusters contained known annotated genes or reported transcripts (Figure 5 and Table S1), many of the EVE-ORFs identified in the clusters were previously unreported. Interestingly, 21 of these multi-species clusters did not contain human EVE-ORFs. It is also worthy to note that 14 of these clusters remained following the extraction of ≥8 shared species clusters at the level of ≥80% identity, and 58 clusters remained when extracting clusters shared among ≥10 mammalian species (Figures S15A and S15B). Further, there were fewer multi-species EVE-ORF clusters identified in rodents and opossum. Instead, these species exhibited larger numbers of species-specific clusters and unique EVE sequences (Figure 5, right panels). The species-specific clusters in these species were abundant in *pol* genes, while there were no multi-species shared clusters observed in analysis of platypuses.

Discussions

We comprehensively identified EVE-ORFs in 19 mammalian species, with differences in their expression patterns noted between species (Figures 1A and 1B). This is the first report, to our knowledge, that presents data on the genome-wide prediction of EVEs that may code for viral proteins in a wide range of mammalian species. Interesting observations were made in mice, dogs, opossum and platypuses regarding the presence of EVE-ORFs. In particular, mice and platypuses exhibited extreme characteristics in the location and

number of EVE-ORFs. The high percentage of ERV-ORFs identified in the mouse genome is likely due to the presence of many active EVEs (35). This is supported by the observation that high densities of ≥500 aa long EVE-ORFs were observed in all examined domain, which was particularly significant in the mouse *gag* and *pro* domains (Figure S1C). Furthermore, the dominant groups of Repbase annotations in the mouse genome were found to be IAPE (Figure 1C). Alternatively, in the platypus genome, we identified very few ERV-ORFs (Figure 1A), which is consistent with a previous report stating that ERVs account for only 0.1% of the platypus genome (51). Moreover, in all species examined, we observed higher numbers of LINE-ORFs compared to ERV-ORFs, which may be due to the presence of active LINE-ORFs in mammals, and thus the presence of young sequences (Figures 1A, 2 and S3). Specifically, many mammals contain active L1 retrotransposons (52). In addition to expressing active L1, cows also contain the LINE, RTE-BovB, which has a high integration rate (53).

All Identified EVE-ORFs may not be candidates for novel genes. It is possible that the observed ORFs were present by chance after being recently inserted into the host genome, as was observed with an endogenous bornavirus-like nucleoprotein element (EBLNs) expressed in Old World and New World monkeys (54). The observed positive correlation between the fraction of EVE-ORFs and the total EVE length may reflect this. To exclude EVE-ORFs that were simply present by chance and did not have an active function, we compared the divergence distribution of our observed EVE-ORFs to those presented by Repbase consensus sequences. We limited these EVE-ORFs to those located ≤2,000 bp downstream of the TSSs. These identified EVE-ORFs were, therefore, considered to exhibit high transcription potential. If the frequency of EVE-ORFs located near TSSs is determined to be either smaller or larger for a specific divergence than that of all EVE-ORFs in the genome, the EVE-ORFs at that divergence are presumed to contain more EVEs that occur

either by chance or targeted selection, respectively. When comparing divergence patterns, we observed an interesting change in the frequency of EVE-ORFs near TSSs. At a divergence of approximately 10% (7-10% for ERV-ORF, and 7-13% for LINE-ORF), the frequency of EVE-ORFs located near TSSs in humans and mice were determined to be higher than the total number of EVE-ORFs within these genomes (Figures S16A and S16B). Conversely, the reduced divergence frequencies observed in the EVE-ORFs near TSSs as well as for all EVE-ORFs in the genomes, were determined to species-specific. For example, we found that only the ERV-ORFs located near TSSs in mice exhibited a large reduction in the frequency of divergence at approximately 31%. Overall, in EVE-ORFs, the frequency of divergence at ≤5% tended to be smaller than those of all EVE-ORFs in the genomes. This confirmed that EVE-ORFs with relatively small divergence (approximately ≤5%), were likely recently inserted into the genomes, and may include EVE-ORFs that remained by chance alone.

The EVE-ORF expression results may also suggest that the promoter regions of EVEs were co-opted rather than the ORFs. Previous studies have reported that promoter regions, especially for ERV LTRs, have been co-opted as tissue-specific, or alternative promoters in mammalian cells (55). In this analysis, we are not able to exclude this as a possibility. The LTR promoter usage in a given cell may also be inferred from RNA-seq data. When analyzing promoter usage, the expression levels were often accumulated and applied to each EVE group, such as HERV-K and MERV-L. Thus, if only a certain EVE-ORF is functional, this accumulated data may cause underestimation of the actual ORF expression levels. In our analysis of human myoblast RNA-seq data, we presented the expression of each EVE-ORF loci separately, and confirmed that the expression of each EVE-ORF loci was unique during muscle cell differentiation (Figure 3B). We believe this level of detailed expression analysis will assist in revealing the function of EVE-ORFs by

identifying individual EVE-ORF loci.

The significant reduction in EVE-ORF located near TSSs in the FANTOM5 database showed that a selective pressure may have been present to induce mutation and destroy any EVE-ORFs within the transcriptionally active regions. However, our analysis of EVE-ORF using transcriptomic data from the CHESS database and myoblast RNA-seq data identified previously unreported EVE-ORFs which may function to express proteins in healthy human and mouse tissues. Importantly, we also determined that a higher fraction of ERV-ORFs exhibited the ability to be transcribed compared to LINE-ORFs. LINE-ORFs were found to be significantly enriched in heterochromatin (subcompartments B2 and B3). This is consistent with the result in a previous study that stated that long L1s preferentially locate in subcompartments B2 and B3 (56; Figure 3E). In addition, the average lengths of LINE-ORFs were found to be longer than for non-LINE-ORF so that the long L1 described by Babenko et al (2017) is considered to correspond with the LINE-ORFs in this study. This is an interesting observation because in somatic tissues, the relaxation of heterochromatin is associated with increased transcription of retrotransposons (57).

The most important characteristic of EVE-ORFs is the consistent harboring of viral-like protein domains. The percentage of EVE-ORFs that may function as potential candidates for functional genes was determined to be very small; in humans, we identified only 480 EVE-ORFs, using public transcriptomic datasets, that were transcriptionally active. An additional 2,525 EVE-ORFs were found to be located 2,000 bp downstream of TSSs, and thus a total of 2,881 EVE-ORFs were determined to exhibit transcriptional potential, which accounts for approximately 8.8% of the total EVE-ORFs found in humans. Significant depletions of EVE-ORFs near the TSS may also supports the low transcription potential of EVE-ORFs. This suggests that the EVE-ORFs near TSSs may be subjected to purifying selection to avoid their being expressed. However, if

specific epigenetic changes occur in close proximity to EVE-ORFs with low transcription potential and they subsequently begin to be expressed, they may become functional, however, their expression may lead to development of diseases as described in previous studies (58, 59). This may be why EVE-ORFs are seen to be depleted close to TSS in transcriptionally active regions, to avoid potential detrimental secondary effects of the expression. Moreover, EVE-ORFs may function as the raw materials required for constructing new exons for multi-exon genes composed of non-EVE-derived exons. Indeed, we identified EVE-ORFs that were exonized in known Ensembl multi-exon genes. Further, the numbers of EVE-containing genes varied among mammalian species (Figure S5); horse, sheep, and opossum contained a larger number of these genes compared to other mammalian species. However, it is not clear whether this observation reflected the difference in EVE-ORF co-opted rates or just in gene annotation quality. Nevertheless, differences, although small, were observed in the number of EVE-ORFs identified in well-annotated human and mouse genomes (Figure S5). We also found that some EVE-ORFs partially overlapped with known gene exons (e.g. *AKR1B15, UBXN8,* and *PPHLN1* in humans) that contain splice variants. This suggests that, as reported in a previous study (60), EVE-ORFs may be transcribed as alternative transcript variants. It is interesting to note that known multi-exon genes have been identified as containing EVE-ORFs as predicted by HMM alone, suggesting that these ORFs were newly identified. Moreover, similar numbers of these genes were identified in all examined mammals, save for horses, sheep, and platypuses. Considering that orthologs of these genes were found in different species (e.g. *CNBP, GIN1, SUGP2*, in chickens and *CTSE* in humans and mice), these EVE-ORFs may have been co-opted before the evolutionary divergence of birds and mammals.

Combinatorial analysis of functional datasets and cluster analysis for

EVE-ORFs in humans revealed that many expressed EVE-ORFs were primate specific. This is consistent with previous studies which showed that specific ERV genes, such as *ERVW-1,* are only shared among apes. In addition, we demonstrated that fractions of each viral-like protein domain in EVE-ORFs varied among mammalian lineages.

Together, our results suggest that mammalian EVE-ORFs, many of which were not previously described, may be co-opted in a cell lineage specific manner, and are likely to have contributed to mammalian evolution and diversification. However, functional data is currently very limited. Thus, further EVE-ORF studies employing the functional data from multiple species may provide a deeper understanding of mammalian evolution.

Materials and methods

Extraction of EVE-ORFs

All sequences and annotations for potential protein-coding EVE-ORFs from 19 mammalian genomes that were used in this study were obtained from the gEVE database (23; http://geve.med.u-tokai.ac.jp). Species names and corresponding genome versions used in this study are summarized in Table S2. We classified EVE-ORF sequences by domains predicted by HMMER version 3.1b1 (hmmer.org) with HMM profiles and/or RetroTector (61) as provided in the gEVE database. We further separated ERV *pol* and LINE-ORF2 domains as previously described (23). We first calculated the fraction of EVE-ORF domains within a species, and obtained the average for each domain across all 19 mammals. We then further divided the original domain fractions for each species by the mammalian average and obtained fold changes for all domains.

Divergence of EVE-ORFs

In our previous study (Nakagawa and Takahashi, 2016), we used RepeatMasker (3) to identify EVE-ORFs in the mammalian genomes. From those RepeatMasker outputs, we obtained divergence information (% substitutions in matching region compared to the consensus) for EVE-ORFs. The R package version 3.4.4 (62) was used to calculate the means for each specie's genome. We determined the modes of divergence using the R package LaplacesDemon version 16.1.1 (63).

RNA-seq analysis

RNA-seq data for human and mouse myoblast differentiation were downloaded from the NCBI Sequence Read Archive (SRA; https:// www.ncbi.nlm.nih.gov/sra). For humans, we used 12 RNA-seq datasets from skeletal muscle cells under the

SRA study number SRP033135 (64). We obtained 14 RNA-seq datasets from C2C12 cell differentiation under the SRA number SRP036149 to analyze mouse myoblast differentiation. The accession number of runs used in this analysis are summarized in Table S3. After trimming the sequences using fastp version 0.12.5 (65) with the following set parameters: -q 20 -l 30, reads were mapped onto the human (GRCh38), and mouse (mm10) genomes using HISAT2 version 2.1.0 (66) with the default option. StringTie version 1.3.4b (67) and DESeq2 version 1.18.1 (68) were used to quantify and evaluate EVE expression. In this analysis, we performed mapping and quantification for EVE-ORF transcripts by using in-house EVE gene transfer format (GTF) files, which were available in the gEVE database. After normalizing for size and filtering the data (removing small read counts < 10 aa), we log transformed the count data using the rlog transformation function of the DESeq2, which minimizes the detection of sample differences for transcripts with small counts, and normalizes the counts with respect to the library size. We then performed PCA analysis using the standardized read counts. PCA plots were generated using the built-in R function prcomp. For human RNA-seq data, expression data was sorted based on the total sum of each EVE-ORF expression profile, and the extracted top 100 EVE-ORF transcripts. The standardized counts were then represented in a heatmap, which was generated using the R package pheatmap version 1.0.10 (69).

Analysis using CHESS

To investigate the functionality of EVE-ORF, we obtained RNA-seq datasets from the GTEx project (20) in GFF format from the CHESS v.2.1 database (21). We then identified overlapping EVE-ORFs with exonic regions of transcripts in the GFF file using the intersect function of BEDTools version 2.26.0 (70).

Analysis using FANTOM data

To predict which EVE-ORFs can be transcribed, we analyzed CAGE data from FANTOM5 (22). Human and mouse CAGE data (bed files, read count data, and sample information table) were retrieved from the FANTOM5 data repository. We converted EVE-ORF coordinates into hg19 assembly using the UCSC LiftOver tool (71), and identified the EVE-ORF and non-EVE-ORF located downstream of TSSs using the closest function in BEDTools version 2.26.0. To determine of the differences in the numbers of EVE-ORF and non-EVE-ORF near TSS was significant, chi-square independence analysis was performed using built-in R function chisq.test. The expected number of EVE-ORF located ≤5000 bp downstream of TSSs with a bin size of 1000 bp was estimated using the fraction of non-EVE-ORF with the same bin size. To control for the false discovery rate (FDR), we performed multiple-test correction with FDR (72) using the p.adjust function in the R package. PCA analysis for TSS data was performed using the same procedure as described for the RNA-seq data.

Subcompartmental analysis

We obtained six subcompartment coordinates for human lymphoblastoid cells from the Gene Expression Omnibus (GEO) database with accession number GSE63525 (48). We converted these coordinates from hg19 to hg38 (GRCh38) using the UCSC LiftOver tool. We then identified overlapping EVE-ORF coordinates within six subcompartments using the intersect function of BEDTools version 2.26.0. Statistical analysis was performed using the built-in R function chisq.test. The expected number of EVE-ORFs found in each subcompartment was calculated using the fraction of non-EVE-ORF identified in the corresponding subcompartment.

Cluster analysis

We obtained EVE-ORF alignments, and clustered each domain with the CD-HIT program (49) for 40-90% shared sequence identity, and we used the PSI-CD-HIT program (73) to detect ≤30% level identity. We then extracted the sequence that contained the highest level of homology with the reference sequence for each species within each cluster and generated matrices for heatmaps. Manipulation of this data was accomplished using in-house Perl scripts. Heatmaps was visualized using the geom_tile function of the R package ggplot2 (74).

Phylogenetic analysis

We first identified clusters that contained *ERVW-1* in the *env* domain with ≥30% identity, and extracted amino acid sequences that were ≥200 aa in length from that cluster. After multiple alignments of the sequences with MAFFT version 7.394 using the auto option (75), we generated maximum likelihood trees with RAxML version 8.2.11 and the following options: -f a -m PROTGAMMAJTTF -#500; 50. The phylogenetic tree was visualized using the R package ggtree version 1.10.5 (76).

Data manipulations and visualizations

For all analyses employed in this study, we used in-house programs developed with Perl, Python, AWK, and Shell script as well as R package dplyr (77) and rehape (78) for processing and manipulation of the data. For visualization, we used ggplot2 version 2.2.1 with RColorBrewer (79) in the R package.

Acknowledgments

We would like to acknowledge the financial support that this study received from JSPS KAKENHI Grants-in-Aid for Challenging Exploratory Research (17K19359 to MTU and SM), Young Scientists (16K21386 to SN), Scientific Research on Innovative Areas (16H06429, 16K21723, 17H05823, 19H04843 to SN) and by MEXT-Supported program for the Strategic Research Foundation at Private Universities (S1411010, to MTU and SN).

Author contributions

MTU and SN designed research; MTU and KK conducted data analysis; MTU, SM and HM analyzed muscle sequencing data; MTU, TI and SN interpreted the results; MTU and SN wrote the paper; All authors read and approved the final manuscript.

References

1. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921.
2. de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD (2011) Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet* 7:e1002384.
3. Smit AFA, Hubley R, Green P (2013) RepeatMasker Open-4.0., 2013-2015. http://www.repeatmasker.org.
4. Smit AFA (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Develop* 9: 657–663.
5. McVean G (2010) What drives recombination hotspots to repeat DNA in humans? *Phillos Trans R Soc B Biol Sci* 365, 1213–1218.
6. Thornburg BG, Gotea V, Makalowski W (2006) Transposable elements as a significant source of transcription regulating signals. *Gene* 365, 104–110.
7. Nishihara H, Kobayashi N, Kimura-Yoshida C, Yan K, Bormuth O, Ding Q, Nakanishi A, Sasaki T, Hirakawa M, Sumiyama K, Furuta Y, Tarabykin V, Matsuo I, Okada N (2016) Coordinately co-opted multiple transposable elements constitute an enhancer for wnt5a expression in the mammalian secondary palate. *PLoS Genet.* 12:e1006380.
8. Chuong EB, Elde NC, Feschotte C (2016) Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* 351:1083–1087.
9. Diwash J, Feschotte C, Betran E (2017) Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. *Trends in Genetics* 33: 817–831.
10. Ono R, Nakamura K, Inoue K, Naruse M, Usami T, Wakisaka-Saito N, dHino T, Suzuki-Migishima R, Ogonuki N, Miki H, Kohda T, Ogura A, Yokoyama M, Kaneko-Ishino T, Ishino F (2006) Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality, Nat Genet.38: 101–106.
11. Matsui T, Miyamoto K, Kubo A, Kawasaki H, Ebihara T, Hata K, Tanahashi S, Ichinose S, Imoto I, Inazawa J, Kudoh J, Amagai M (2011) SASPase regulates stratum corneum hydration through profilaggrin-to-filaggrin processing. *EMBO Mol Med* 3:320–333.
12. Nakaya Y, Koshi K, Nakagawa S, Hashizume K, Miyazawa T (2013) 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.

13. Pastuzyn ED, Day CE, Kearns RB, Kyrke-Smith M, Taibi AV, McCormick J, Yoder N, Belnap DM, Erlendsson S, Morado DR, Briggs JAG, Feschotte C, Shepherd JD (2018) The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. *Cell* 172:275-288.e18.
14. Ashley J, Cordy B, Lucia D, Fradkin LG, Budnik V, Thomson T (2018) Retrovirus-like Gag Protein Arc1 Binds RNA and Traffics across Synaptic Boutons. *Cell* 172:262–274.
15. Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, LaVallie E, Tang XY, Edouard P, Howes S, Keith JC Jr, McCoy JM. (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis, Nature 403: 785–789.
16. Dupressoir A, Lavialle C, Heidmann T. (2012) From ancestral infectious retroviruses to bona fide cellular genes: role of the captured syncytins in placentation. Placenta 33, 663–667.
17. Lavialle C, Cornelis G, Dupressoir A, Esnault C, Heidmann O, Vernochet C, Heidmann T (2013) Paleovirology of 'syncytins', retroviral *env* genes exapted for a role in placentation. *Philos Trans R Soc Lond B* 368:20120507.
18. Bolze PA, Mommert M, Mallet F (2017) Contribution of syncytins and other endogenous retroviral envelopes to human placenta pathologies. *Prog Mol Biol Transl Sci* 145: 111–162.
19. Ito J. Sugimoto R. Nakaoka H, Yamada S, Kimura T, Hayano T, Inoue (2017) Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. *PLoS Genet* 13: e1006883.
20. Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, Compton CC, DeLuca DS, Peter-Demchok J, Gelfand ET, Guan P, Korzeniewski GE, Lockhart NC, Rabiner CA, Rao AK, Robinson KL, Roche NV, Sawyer SJ, Segrè AV, Shive CE, Smith AM, Sobin LH, Undale AH, Valentino KM, Vaught J, Young TR, Moore HM, on behalf of the GTEx consortium (2015) A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreserv Biobank* 13:311–319.
21. Pertea M, Shumate A, Pertea G, arabyou A, Breitwieser FP, Chang Y, Madugundu AK, Pandey A, Salzberg SL (2018) CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome biol* 19:208.

22. Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, Abugessaisa I, Fukuda S, Hori F, Ishikawa-Kato S, Mungall CJ, Arner E, Baillie JK, Bertin N, Bono H, de Hoon M, Diehl AD, Dimont E, Freeman TC, Fujieda K, Hide W, Kaliyaperumal R, Katayama T, Lassmann T, Meehan TF, Nishikata K, Ono H, Rehli M, Sandelin AG, Schultes EA, 't Hoen PAC, Tatum Z, Thompson M, Toyoda T, Wright DW, Daub CO, Itoh M, Carninci P, Hayashizaki Y, Forrest ARR, Kawaji H (2015) Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol* 16:22.
23. Nakagawa S, Takahashi MU. (2016) gEVE: a genome-based endogenous viral element database provides comprehensive viral protein-coding sequences in mammalian genomes. *Database* baw087.
24. Paces J, Pavlícek A., Paces V (2002) HERVd: database of human endogenous retroviruses. *Nucleic Acids Res* 30:205–206.
25. de Parseval N, Lazar V, Casella JF, Benit L, Heidmann T (2003) Survey of human genes of retroviral origin: Identification and transcriptome of the genes with coding capacity for complete envelope proteins. *J Virol* 77:10414–10422.
26. Villesen P, Aagaard L, Wiuf C, Pedersen FS (2004) Identification of endogenous retroviral reading frames in the human genome. *Retrovirology* 1:32.
27. Tokuyama M, Kong Y, Song E, Jayewickreme T, Kang I, Iwasaki A (2018) ERVmap analysis reveals genome-wide transcription of human endogenous retroviruses. *Proc Natl Acad Sci USA* 115:12565-12572.
28. Penzkofer T, Dandekar T, Zemojtel T (2005) L1Base: from functional annotation to prediction of active LINE-1 elements. *Nucleic Acids Res* 33(Database issue): D498–500.
29. Sugimoto J. Sugimoto M. Bernstein H, Jinno Y, Schust D (2013) A novel human endogenous retroviral protein inhibits cell-cell fusion. *Sci Rep* 3:1462.
30. Sela N, Mersch B, Gal-Mark N, Lev-Maor G, Hotz-Wagenblatt A, Ast G (2007) Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. *Genome Biol* 8:R127.
31. Bao W, Kojima KK, Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* 6:11.

32. Bénit L, Calteau A, Heidmann T (2003) Characterization of the low-copy HERV-Fc family: evidence for recent integrations in primates of elements with coding envelope genes. *Virology* 312:159–168.
33. Bannert N, Kurth R (2006) The evolutionary dynamics of human endogenous retroviral families. *Annu Rev Genomics Hum Genet* 7:149–173.
34. Subramanian R, Wildschutte J, Russo C, Coffin J (2011) Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology* 8:90.
35. Mouse Genome Sequencing Consrotium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–562.
36. Sookdeo A, Hepp CM, McClure MA, Boissinot S (2013) Revisiting the evolution of mouse LINE-1 in the genomic era. *Mob DNA* 4:3.
37. Lin L, Xu B, Rote NS (1999) Expression of endogenous retrovirus ERV-3 induces differentiation in BeWo, a choriocarcinoma model of human placental trophoblast. Placenta 20:109–118.
38. Grow EJ, Flynn RA, Chavez SL, Bayless NL, Wossidlo M, Wesche DJ, Martin L, Ware CB, Blish CA, Chang HY, Pera RA, Wysocka J (2015) Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* 522:221–225.
39. Oki S, Ohta T, Shioi G, Hatanaka H, Ogasawara O5, Okuda Y5, Kawaji H, Nakaki R, Sese J, Meno C (2018) ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep* 19: e46255.
40. Wu C (1980) The 5′ ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. *Nature* 286: 854–860.
41. Gross DS, Garrard WT (1988) Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* 57:159–197.
42. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell* 132: 311–322.
43. Dietzel, Jauch A, Kienle D, Qu G, Holtgreve-Grez H, Eils R, Munkel C, Bittner M, Meltzer PS, Trent JM (1998) Separate and variably shaped chromosome arm domains are disclosed by chromosome arm painting in human cell nuclei. *Chromosome Res* 6:25–33.

44. Bolzer A, Kreth G, Solovei I, Koehler D, Saracoglu K, Fauth C, Müller S, Eils R, Cremer C, Speicher MR, Cremer T (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol* 3:e157.
45. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326:289–293.
46. Dixon, JR, Selvaraj S., Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376–380.
47. Nora EP, Lajoie B, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J, Gribnau J, Barillot E, Blüthgen, N, Dekker J, Heard E (2012) Spatial partitioning of the regulatory landscape of the X- inactivation centre. *Nature* 485:381–385.
48. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159:1665–1680.
49. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–1659.
50. Stamatakis . (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688–2690.
51. Warren WC, Hillier LW, Marshall Graves JA, Birney E, Ponting CP, Grützner F, Belov K, Miller W, Clarke L, Chinwalla AT, Yang SP, Heger A, Locke DP, Miethke P, Waters PD, Veyrunes F, Fulton L, Fulton B, Graves T, Wallis J, Puente XS, López-Otín C, Ordóñez GR, Eichler EE, Chen L, Cheng Z, Deakin JE, Alsop A, Thompson K, Kirby P, Papenfuss AT, Wakefield MJ, Olender T, Lancet D, Huttley GA, Smit AF, Pask A, Temple-Smith P, Batzer MA, Walker JA, Konkel MK, Harris RS, Whittington CM, Wong ES, Gemmell NJ, Buschiazzo E, Vargas Jentzsch IM, Merkel A, Schmitz J, Zemann A, Churakov G, Kriegs JO, Brosius J, Murchison EP, Sachidanandam R, Smith C, Hannon GJ, Tsend-Ayush E, McMillan D, Attenborough R, Rens W, Ferguson-Smith M, Lefèvre CM, Sharp JA, Nicholas KR, Ray DA, Kube M, Reinhardt

R, Pringle TH, Taylor J, Jones RC, Nixon B, Dacheux JL, Niwa H, Sekita Y, Huang X, Stark A, Kheradpour P, Kellis M, Flicek P, Chen Y, Webber C, Hardison R, Nelson J, Hallsworth-Pepin K, Delehaunty K, Markovic C, Minx P, Feng Y, Kremitzki C, Mitreva M, Glasscock J, Wylie T, Wohldmann P, Thiru P, Nhan MN, Pohl CS, Smith SM, Hou S, Nefedov M, de Jong PJ, Renfree MB, Mardis ER, Wilson RK (2008) Genome analysis of the platypus reveals unique signatures of evolution. *Nature* 453:175–183.

52. Rodriguez-Terrones D, Torres-Padilla ME (2018) Nimble and ready to mingle: Transposon outbursts of early development. *Trends in Genetics* 34:806–820.
53. Adelson DL, Raison JM, Edgar RC (2009) Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. *Proc Natl Acad Sci USA* 106:12855–12860.
54. Kobayashi Y, Horie M, Tomonaga K, Suzuki Y (2011) No evidence for natural selection on endogenous borna-like nucleoprotein elements after the divergence of Old World and New World monkeys. *PLoS One* 6:e24403.
55. Thompson PJ, Macfarlan TS, Lorincz MC (2016) Long Terminal Repeats: From Parasitic Elements to Building Blocks of the Transcriptional Regulatory Repertoire Mol Cell 62:766-76.
56. Babenko VN, Chadaeva IV, Orlov_YL (2017) Genomic landscape of CpG rich elements in human. *BMC Evol Biol* 17(Suppl 1):19.
57. Cecco MD, Criscione SW, Peterson AL, Neretti N, Sedivy JM, Kreiling JA (2013) Transposable elements become active and mobile in the genomes of aging mammalian somatic tissues. *Aging (Albany NY)* 5:867–83.
58. Perron H, Germi R, Bernard C, Garcia-Montojo M, Deluen C, Farinelli L, Faucard R, Veas F, Stefas I, Fabriek BO, Van-Horssen J, Van-der-Valk P, Gerdil C, Mancuso R, Saresella M, Clerici M, Marcel S, Creange A, Cavaretta R, Caputo D, Arru G, Morand P, Lang AB, Sotgiu S, Ruprecht K, Rieckmann P, Villoslada P, Chofflon M, Boucraut J, Pelletier J, Hartung HP (2012) Human endogenous retrovirus type W envelope expression in blood and brain cells provides new insights into multiple sclerosis disease. *Mult. Scler* 18:1721–1736.
59. Kassiotis G (2014) Endogenous retroviruses and the development of cancer. *J Immunol* 192:1343–1349.

60. Bae MI, Kim YJ, Lee JR, Jung YD, Kim HS (2013) A new exon derived from a mammalian apparent LTR retrotransposon of the SUPT16H gene. *Int J Genomics* 2013:387594.
61. Sperber GO, Airola T, Jern P, Blomberg J (2007) Automated recognition of retroviral sequences in genomic data–RetroTector. *Nucleic Acids Res* 35:4964–4976.
62. R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
63. Statisticat LLC (2018) LaplacesDemon: Complete Environment for Bayesian Inference. Bayesian-Inference.com. R package version 16.1.1. https://web.archive.org/web/20150206004624/http://www.bayesian-inference.com/software
64. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* 32:381– 386.
65. Chen S, Zhou Y, Chen Y, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34: i884–i890.
66. Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 12:357–360.
67. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology 2015;33:290–295.
68. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:1–21.
69. Kolde R (2015) pheatmap: Pretty Heatmaps. R package version 1.0.8. https://CRAN.R-project.org/package=pheatmap
70. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.
71. Hinrichs, AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, Hillman-Jackson J, Kuhn RM, Pedersen JS, Pohl A, Raney BJ, Rosenbloom KR, Siepel A, Smith KE, Sugnet CW, Sultan-Qurraie A, Thomas DJ, Trumbower H, Weber RJ, Weirauch M, Zweig AS, Haussler D, Kent WJ

(2006) The ucsc genome browser database: update 2006. *Nucleic Acids Res* 34(suppl 1):D590–D598.

72. Benjamini, Y. Hochberg (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser A Stat Soc* 57: 289–300.
73. Altschul SF, Madden TL, Schaff er AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
74. Wickham H (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
75. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30: 772–780.
76. Yu G, David Smith, Zhu H, Guan U, Lam TT (2017) ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* 8:28–36.
77. Wickham H, François R, Henry L, Müller K (2018) dplyr: A Grammar of Data Manipulation. R package version 0.7.5. https://CRAN.R-project.org/package=dplyr
78. Wickham H (2007) Reshaping Data with the reshape Package. *J Stat Softw* 21:1–20. http://www.jstatsoft.org/v21/i12/
79. Neuwirth E (2014) RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer

Figure legends

Figure 1. Number of identified ERVs and Repbase annotations in 19 mammalian species

A) The mammalian phylogeny showing the number of identified EVE-ORFs. Background colors for each specie's names indicate that they have shared species classification equal or below the level of order. Colors in a bar represent each domain. The proportion of EVE-ORFs identified compared to all EVE sequences in each species, are shown on the right. Chimp: chimpanzee, Rhesus: rhesus macaque monkey. B) Relative change in the number of each protein domain identified compared to the mammalian average (ΔF) is shown on the y-axis. Colors of protein domains and mammalian classification are the same as those used in panel A. Any ΔF that was determined to be more than ±1.5 are highlighted by red triangles. The ΔF for the *env* gene of rhesus macaques, which was approximately 1.5 (1.49), is also highlighted by an open triangle. C) Top 20 Repbase annotations for ERVs in the human and mouse genomes. The x-axis represents the number of EVE-ORFs in each annotated group. In the Repbase annotation, ERV internal regions (non-LTR regions) are expressed as "-int"; however, only the group names without "-int" are shown due to the limited space.

Figure 2. Divergence from Repbase consensus sequences

The divergence frequencies for the consensus sequences in ERV-ORF (blue) and non-ERV-ORF (gray) are shown. Background colors for each panel represent the corresponding mammalian classification as indicated in Figure 1A. For each species, the mode and average values for ERV divergence are shown in blue and red, respectively.

Figure 3. Overlapping human EVE-ORFs with functional data in various tissues and cell lines

(A) PCA plot illustrating the EVE-ORFs expressed during the 4 stages of human primary myoblast differentiation (3 replicates in each stage) on the first two principal components (PC1 and PC2). Differentiation status was indicated by color. D0: undifferentiated myoblasts, D1-3: 1-3 days after myoblast differentiation began. The percentage of contributing variables for a given PC is shown on the axes. B) Expression levels for EVE-ORFs with ≥10 normalized read counts during human primary myoblast differentiation. EVE type (ERV or LINE) are displayed on the left. The regularized logarithm (rlog) transformed read counts for EVE-ORFs are color-coded from blue (low expression) to red (high expression). Differentiation status is presented in the same color as in panel A. C) Percentages of EVE-ORFs downstream of TSS as obtained from FANTOM datasets in each 1,000 bp bin (light blue). EVEs without ORFs (non-EVE-ORFs) are shown for comparison (grey). The x-axis represents the distance between EVE-ORFs and the closest TSS. An asterisk (*) indicates statistically significant differences when comparing numbers of the observed EVE-ORF to those expected using fractions of the non-EVE-ORF for each bin ($p<0.001$, chi-squared test, FDR corrected). D) PCA plots for TSSs from CAGE datasets, located within 2,000 bp upstream of EVE-ORFs. Colors represent different tissues/cell lines. COBLa_rind: COBL-a (a cell line established from human umbilical cord blood) infected by rinderpest; H9EB/ES: H9 embryoid bodies/embryonic stem cells; MSC: mesenchymal stem cells; RPE: retinal pigment epithelium. E) Percentage of the Repbase annotations exhibiting ERV-ORFs. The different ERV sets for all ERV-ORFs in this study (gray bar), detected from the CHESS datasets (pink circle), and FANTOM datasets (blue circle) are indicated. F) Fractions of EVE-ORFs located within different

subcompartments. Asterisks (*) and (**) indicate statistically significant differences $p < 0.01$, and $p < 10^{-3}$, respectively (chi-square analysis, FDR corrected) when compared to EVE-ORF numbers expected as determined by non-EVE-ORF fractions in each subcompartment.

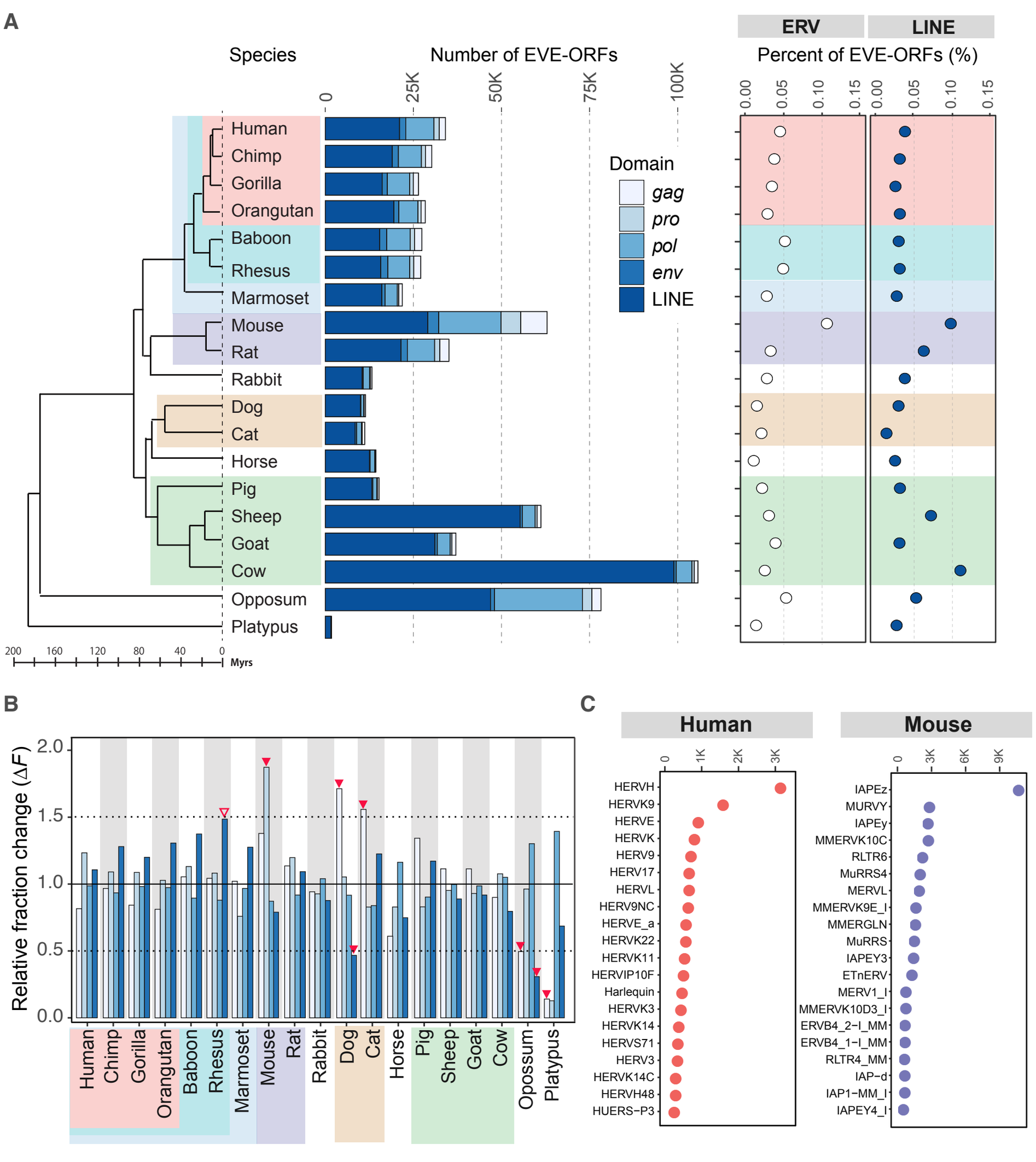
Figure 4. Summary of ERV sequence clustering

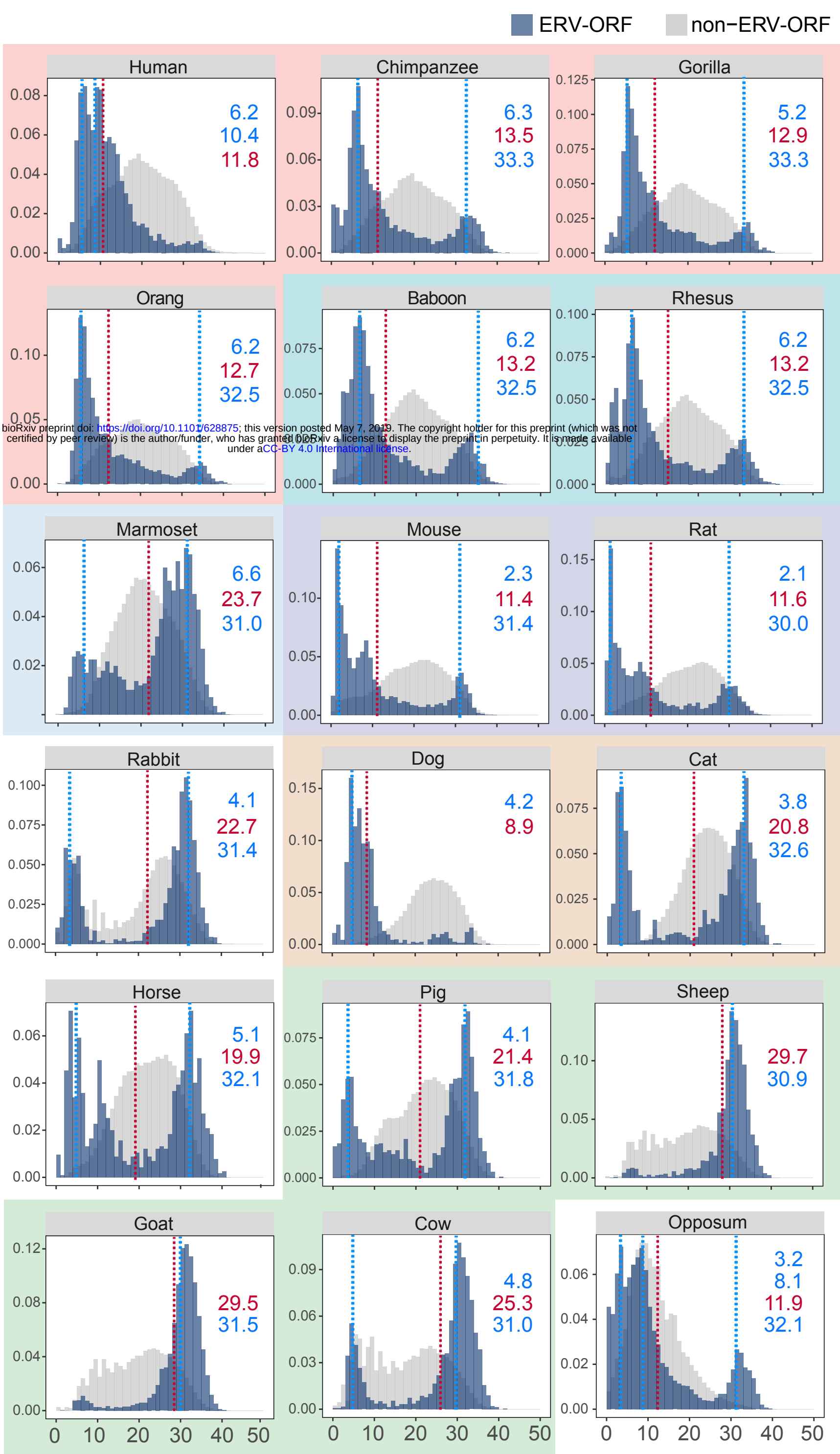
A) Number of EVE-ORF clusters associated with each EVE-ORF domain. The x-axis represents the sequence identity established via clustering. The y-axis represents the number of clusters and also shows the original number of sequences before clustering occurred (gray shaded area) for comparison. A logarithmic scale was used on the y-axis. B) Pairwise comparison of shared sequences clustered at ≥60% identity level in 19 mammalian species. The color bar represents the number of shared sequences. Gray indicates that no shared sequences were identified. C) The number of EVE-ORF sequences in each cluster containing at least one human EVE-ORF identified in the CHESS database. The x-axis represents clusters with ≥60% identity; however, the cluster name is not shown due to limited space. Individual bar colors indicate which EVE-ORFs in each cluster are derived from which species shown in Figure 1A [*e.g*. Apes (blue) contain EVE-ORFs from human, chimpanzee, gorilla, and/or orangutan]. Clusters containing EVE-derived genes are indicated by red triangles and the specific gene name. D) Phylogenetic tree of an EVE-ORF cluster containing *ERVW-1*. The color of each node represents different EVE-ORF species. Red and pink circles indicate human EVE-ORFs detected in the expression data (CHESS data or myoblast RNA-seq data) and FANTOM datasets, respectively. The *ERVW-1* clade is indicated by red font.

Figure 5 EVE-ORF clusters shared among at least 8 species

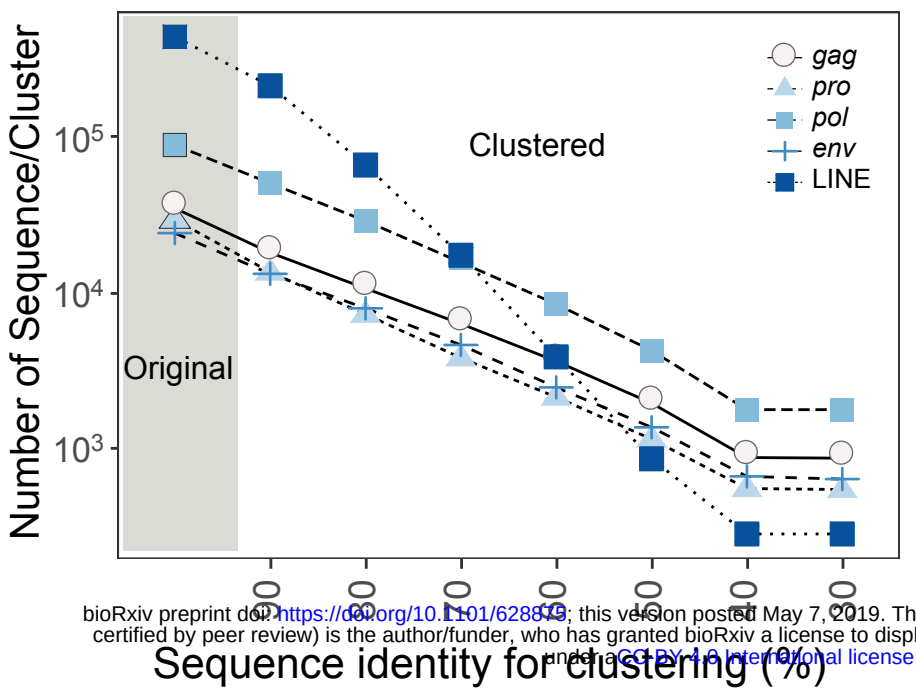
Clusters of EVE-ORF sequences shared among mammals at ≥60% identity

levels are shown. Domain names are shown on the top. The red-scale color bar represents the percent identities for sequences against a reference sequence within the cluster. The highest identity in each species is shown. Blue represents the absence of a specific sequence in the given cluster. The amount of sequences forming single-species clusters (middle) and those that failed to form any cluster (right) are also shown for each species.

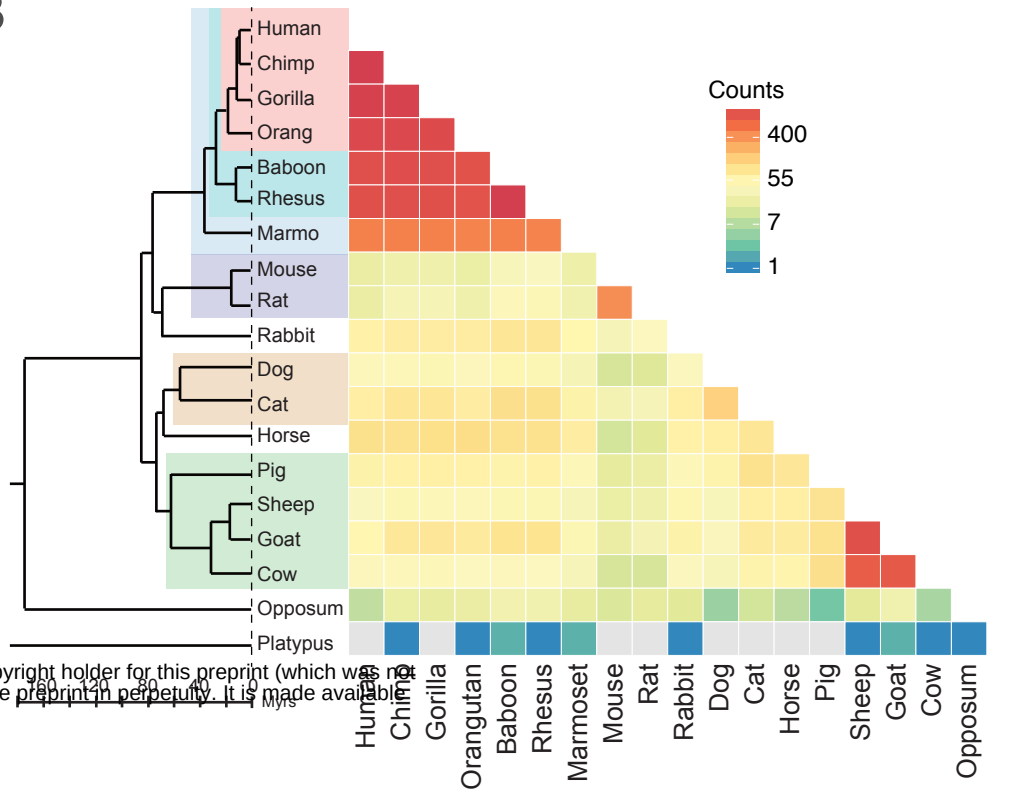

A
Species
Number of EVE-ORFs
0
25K
50K
75K
100K
Human
Chimp
Gorilla
Orangutan
Baboon
Rhesus
Marmoset
Mouse
Rat
Rabbit
Dog
Cat
Horse
Pig
Sheep
Goat
Cow
Opposum
Platypus
Domain
gag
pro
pol
env
LINE
ERV
LINE
Percent of EVE-ORFs (%)
0.00
0.05
0.10
0.15
200
160
120
80
40
0
Myrs
B
Relative fraction change (ΔF)
2.0
1.5
1.0
0.5
0.0
C
Human
Mouse
0
1K
2K
3K
HERVH
HERVK9
HERVE
HERVK
HERV9
HERV17
HERVL
HERV9NC
HERVE_a
HERVK22
HERVK11
HERVIP10F
Harlequin
HERVK3
HERVK14
HERVS71
HERV3
HERVK14C
HERVH48
HUERS-P3
0
3K
6K
9K
IAPEz
MURVY
IAPEy
MMERVK10C
RLTR6
MuRRS4
MERVL
MMERVK9E_I
MMERGLN
MuRRS
IAPEY3
ETnERV
MERV1_I
MMERVK10D3_I
ERVB4_2-I_MM
ERVB4_1-I_MM
RLTR4_MM
IAP-d
IAP1-MM_I
IAPEY4_I



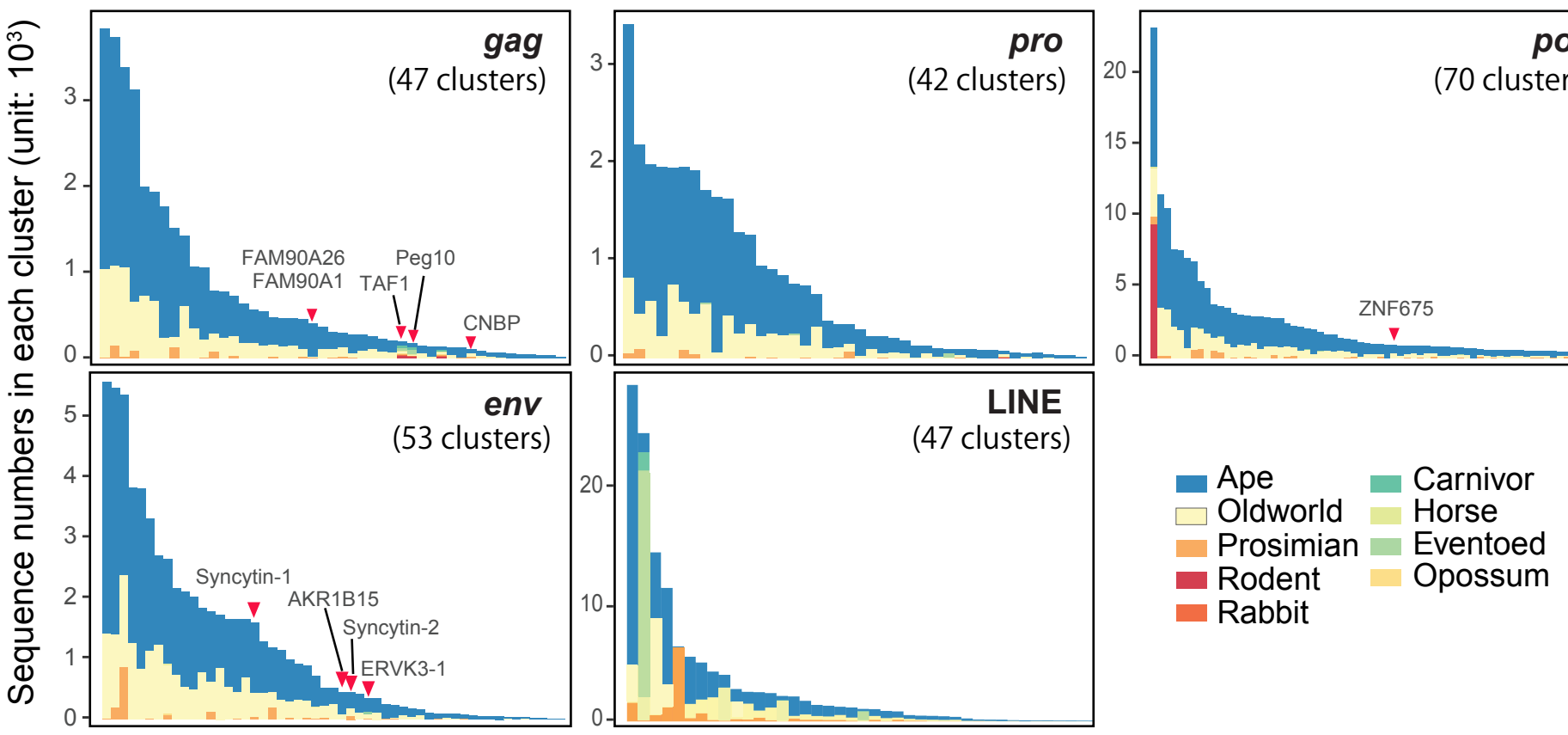
A



B

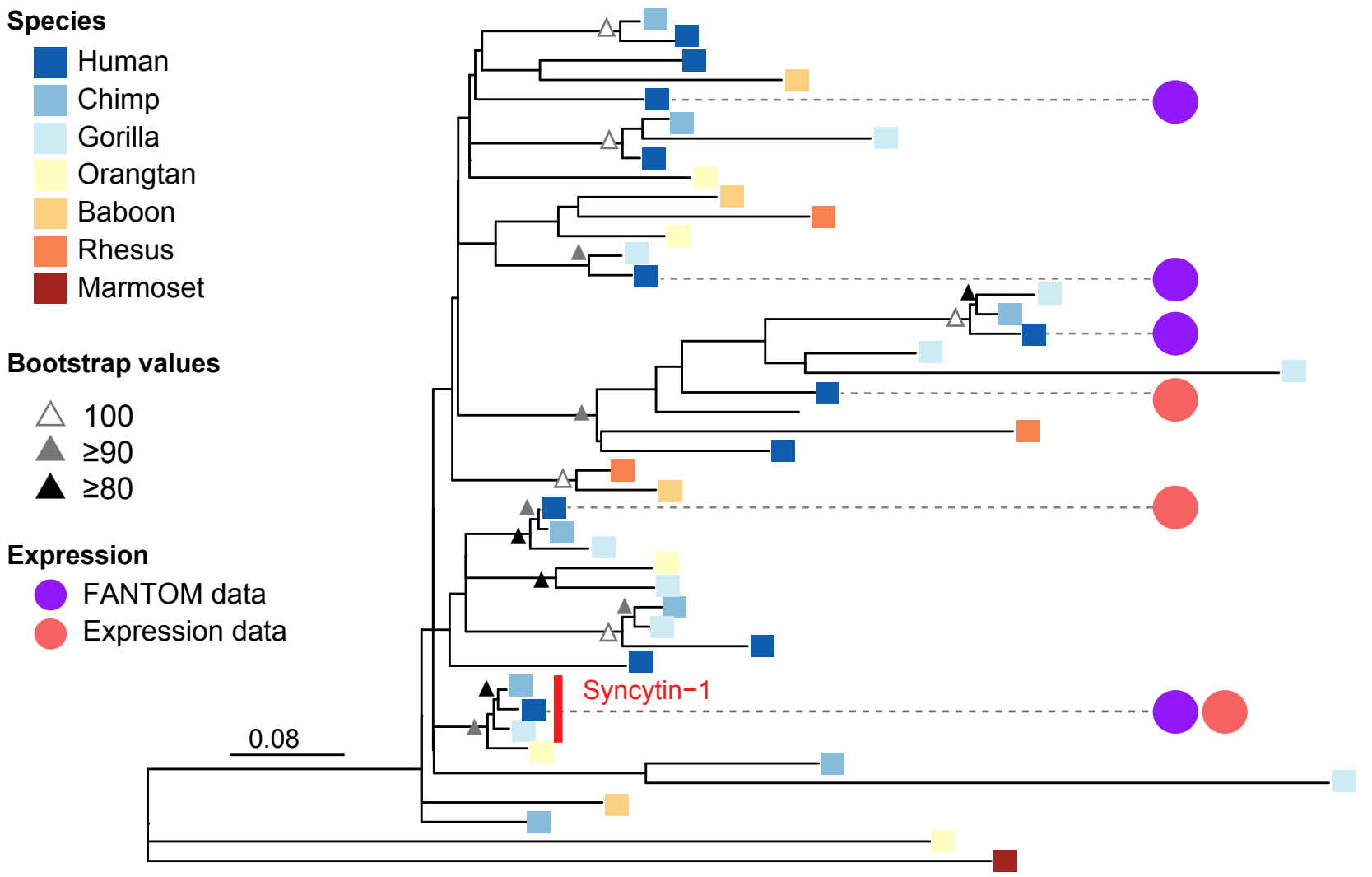


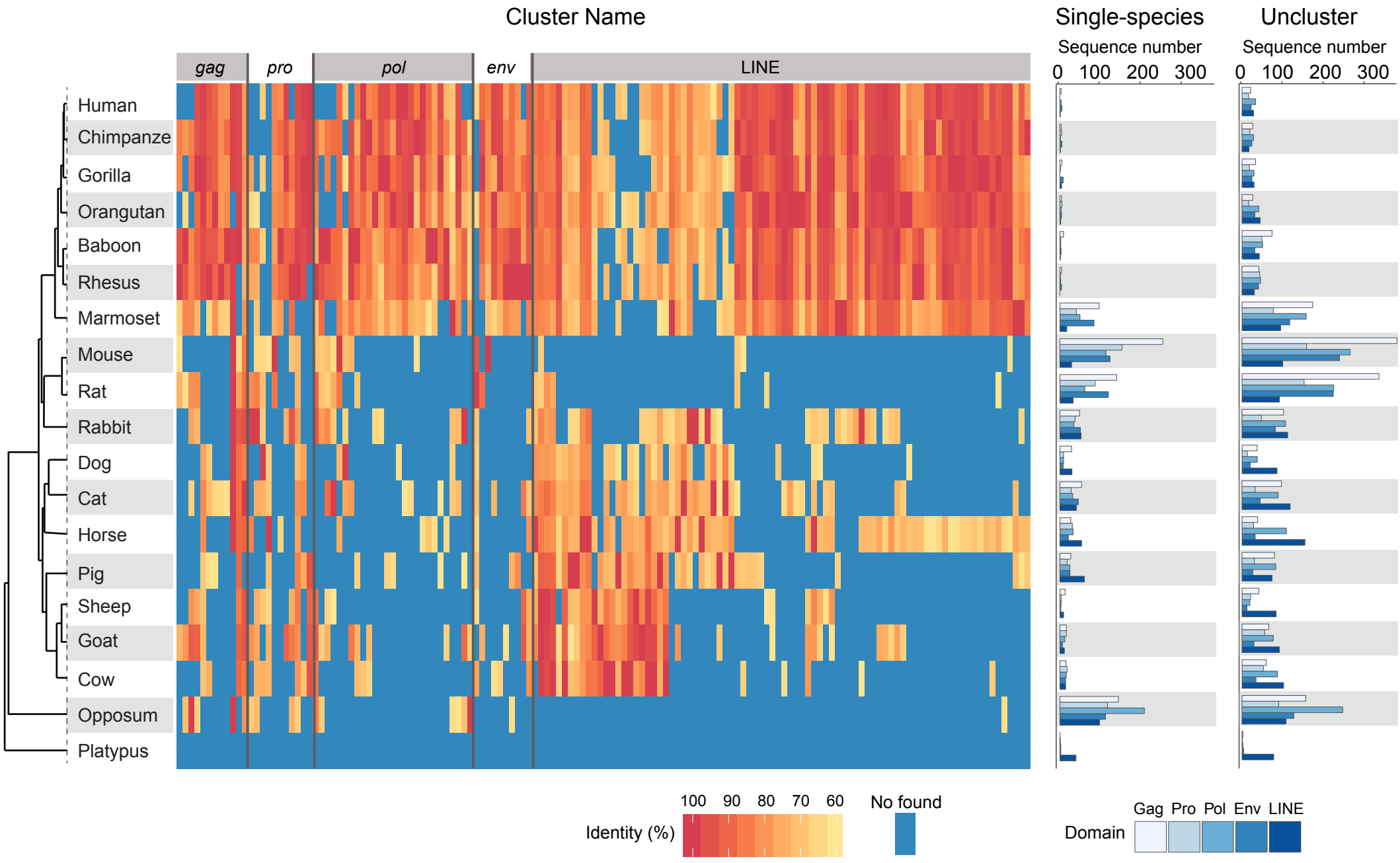
C



Clusters containing human EVE-ORFs detected in the CHESS data

D




Cluster Name
gag
pro
pol
env
LINE
Human
Chimpanze
Gorilla
Orangutan
Baboon
Rhesus
Marmoset
Mouse
Rat
Rabbit
Dog
Cat
Horse
Pig
Sheep
Goat
Cow
Opposum
Platypus
Single-species
Sequence number
0 100 200 300
Uncluster
Sequence number
0 100 200 300
Identity (%)
100 90 80 70 60
No found
Domain
Gag Pro Pol Env LINE