Complex evolution of porcine endogenous retroviruses

Yicong Chen
Wuhan Institute of Virology Chinese Academy of Sciences
Mingyue Chen
Wuhan Institute of Virology Chinese Academy of Sciences
Xiaoyan Duan
Wuhan Institute of Virology Chinese Academy of Sciences
Jie Cui (jcui@ips.ac.cn)
Institut Pasteur of Shanghai Chinese Academy of Sciences https://orcid.org/0000-0001-8176-9951

Research

Keywords: porcine endogenous retrovirus, genomic rearrangement, cross-species transmission, evolution, origin

Posted Date: December 12th, 2019

DOI: https://doi.org/10.21203/rs.2.18651/v1

Abstract

Background: Porcine endogenous retroviruses (PERVs) are proviruses that can replicate in human cells. However, the evolutionary process leading to the generation of modern PERVs is not well understood.

Results: We mined 14 pig genomes and other available 304 mammalian genomes in silico, which led to the documentation of 185 full-length PERVs. Notably, we found two novel ERVs in the lesser Egyptian jerboa (Jaculus jaculus) and rock hyrax (Procavia capensis) named ERV-Gamma-Jja and ERV-Gamma-Pca, respectively, which were the source of the modern PERVs. Phylogenetic analyses provided evidence for the multiple origins of PERVs involving hosts of rodents, rock hyrax, and pigs.

Conclusion: These new findings help us to understand the complex evolution of the modern PERVs.

Background

Porcine endogenous retroviruses (PERVs) are endogenous gammaretroviruses, and exist in the genomes of all pig strains [1, 2]. Although there is no evidence of PERV transmission in patients receiving encapsulated pig islets [3–5], it has been found that PERV-A, -B, and recombinant -A/C are able to infect both human and pig cells, raising major concerns about the safety of xenotransplantation [1, 6].

The envelope (env) genes of three PERV classes (PERV-A, -B, and -C) differ with respect to the receptor-binding domain (RBD) [7]. PERVs have two different types of long terminal repeats (LTRs), one with a 39-bp repeat structure in the U3 region, and the other without this repeat structure [8, 9]. The 39-bp repeats carried by PERV-A and -B confer strong promoter activity and thus increase transcription [8, 9]. However, the 39-bp repeat structure is absent in some PERV-A and all PERV-C. Thus, these PERVs have low transcriptional activity [8, 9].

While several studies have examined the evolutionary relationships between PERVs and other viruses [7, 10–12], the origin and evolution of the modern PERVs remain uncertain. BLAST search analysis confirmed that the R and U5 regions of the PERV LTRs are highly conserved in the pig and mouse genomes (74–87% identity) [13], suggesting a murine origin of PERVs. At least two species that belong to the same order as pigs, Tayassu pecari (of Eocene origin) and Babyrousa babyrussa (of Miocene origin), lack PERVs [7], raising the possibility of a non-porcine origin of the PERVs. However, the common warthog (Phacochoerus africanus) carries PERVs, suggesting that an ancestral porcine species could have carried PERVs [7]. Here, we expand the genomic mining to all available mammalian genomes, aiming to discover non-porcine and non-murine ERVs that have high similarity to PERVs. Combined with our comprehensive phylogenetic analyses, we illustrate the evolutionary journey of the modern PERVs and reveal that the genesis of PERVs is much more complex than previously thought.

Results

In silico characterization of putative PERVs

Using previously reported PERV sequences as queries, we mined 14 pig genomes (Additional files 2: Table S1) available in GenBank and showed a detailed genome-wide distribution of full-length PERVs (i.e., containing two LTRs). We initially compiled a PERV data set that included 185 putative PERVs (containing at least one LTR) (Additional files 2: Table S2). A total of 84 classified (30 PERV-A, 39 PERV-B, and 15 PERV-C) and 18 unclassified PERVs (i.e., lacking the env gene) were retrieved from pig genomes. We identified 2–10 full-length PERVs in most pig breeds (12/14), including Meishan, Goettingen, and Large White (Additional files 2: Table S2). After removing 19 previously classified PERV sequences that were low-quality fragments (> 200 “N” bases), the final data set comprised 65 high-quality classified PERVs (27 PERV-A, 29 PERV-B, and 9 PERV-C). Their viral genomic structures are summarized in Figure 1.

Most of the PERVs exhibited large-scale genetic alterations induced by indels and stop codons (Fig. 1), suggestive of a relatively long evolutionary history. PERV LTRs were classified by the presence (LTR B) or absence (LTR A) of the 18 bp and 21 bp repeat structure reported previously [8, 14, 15], as shown in Fig. 1a. Three different types of B LTRs in the PERV were identified, distinguished by the number of 18 bp and 21 bp repeat sequences: LTR B1 (two 18-bp and one 21-bp repeats), LTR B2 (three 18-bp and two 21-bp repeats), and LTR B3 (four 18 bp and three 21 bp repeats). Of the high-quality PERVs we analyzed, 32 PERVs (>55%) carried LTR A, 10 carried LTR B1, 13 carried LTR B2, and 2 carried LTR B3. LTR A was identified in PERV-A and -C, and LTR B1 was identified in PERV-A and -B. LTR B2 and LTR B3 were only identified in PERV-B. The remaining eight PERVs contained different types of 5’- and 3’- LTR, which may reflect recombination between PERVs with different LTR types (Fig. 1). For example, we discovered one PERV (AEMK02000536.1) with LTR B2 at the 5’ end and LTR A at the 3’ end (Fig. 1).

Genomic rearrangement via PERVs

Retrovirus integration creates a short duplication called target site duplication (TSD) flanking the LTR [16, 17]. If chromosomal rearrangement through homologous recombination between distant proviruses occurred, the flanking TSDs should be different, as mentioned in a previous study of genomic rearrangement in primates via ERVs [18]. To identify pig genomic rearrangement via PERVs, we first constructed a maximum likelihood (ML) tree representing the 5’- and 3’- LTR sequences of full-length PERVs (Additional files 1: Fig. S1). The phylogenetic tree was divided into three large clusters (Additional files 1: Fig. S1), suggesting that three major integration events had occurred. We collected PERVs in which the 5’- and 3’- LTR sequences did not cluster together in the phylogenetic tree. Remarkably, 11 PERVs did not share the same TSD (4bp in length) (Table 1, Additional files 2: Table S3); thus, these PERVs could reflect porcine genomic rearrangement via PERV recombination during evolution.

Table 1. PERVs with different target site duplications (TSDs)

Name	Accession number	Divergent LTRs based on structure *	Divergent LTR based on tree	Flanking TSD **	
				5'	3'
AEMK02000536.1	AEMK02000536.1	yes	yes	AGCC	CTTT
CM000818.5_a	CM000818.5	no	yes	GTTC	CTTC
CM000826.5_c	CM000826.5	no	no	ACCA	AATC
CM000828.5_d	CM000828.5	no	yes	CCAC	CACC
KQ001967.1	KQ001967.1	no	yes	CCAC	CACC
LIDP01000017.1	LIDP01000017.1	no	yes	CCAC	CACC
LUXR01004647.1	LUXR01004647.1	yes	yes	GTTC	CTTC
LUXR01022139.1	LUXR01022139.1	yes	yes	CCAC	CACC
LUXX01045907.1	LUXX01045907.1	no	yes	CCAC	CACC
LUXX01080744.1	LUXX01080744.1	no	yes	GTTC	CTTC
LUXY01101100.1	LUXY01101100.1	no	yes	CCAC	CACC

*LTRs of PERVs comprise four types (LTR A, B1, B2, and B3). If two different types of LTRs are flanking the PERV, the LTRs are considered "divergent".

**Only TSDs flanking the intact 5’ and 3’ LTRs sequences were analyzed

Detection of PERV-related sequences in mammalian genomes

After screening 304 mammalian genomes (Additional files 2: Table S4) available on GenBank using tBLASTn and choosing three major proteins (Gag, Pol, and Env) of PERVs as queries, a significant sequence (accession number: NW_004504334.1) was found in the genome of lesser Egyptian jerboa (*Jaculus jaculus*) that exhibited strong sequence similarity (for *gag* and *pol*: >75% nucleotide identity over 95% region; for *env*: >75% nucleotide identity over 55% region) to PERVs. Using this PERV-like sequence as a query, three other possible PERV-like sequences were identified in *J. jaculus* with >85% nucleotide identity over 80% of the query sequence. The four PERV-like sequences identified in *J. jaculus* were designated as ERV-Gamma.n-Jja (Additional files 2: Table S5) (where n = 1–4). These four significant

hits are located in large scaffolds > 5 Mb in length and are flanked by several host genes, indicating that the ERVs-Gamma-Jja sequences were relatively reliable (Additional files 2: Table S6).

We were only able to identify one pair of ERV-Gamma-Jja LTRs. This full-length ERV-Gamma.1-Jja (containing 2 LTRs) is annotated in Additional files 1: Fig. S2. The length of the 3'- LTR of this ERV-Gamma-Jja is 674 bp, while the 5'- LTR is 932 bp with a 258 bp insertion. We aligned ERV-Gamma-Jja LTRs with PERV LTRs. The start of the U3 region and the end of the U5 region were distinct and were not included in the alignment (Additional files 1: Fig. S3). The ERV-Gamma-Jja LTRs include a similar repeat structure (three 18 bp and two 21 bp repeat sequences) in the U3 region, the sequence of which is identical to that of the PERV LTR B2 (identity ~73%). Alignment analysis revealed a closer relationship between the LTRs of the ERV-Gamma-Jja and LTR B2 of PERVs (Additional files 1: Fig. S3). Notably, the alignment of the conserved R region supported a close evolutionary relationship between the ERV-Gamma-Jja and PERVs (Fig. 2a). To highlight the similarity between PERVs and ERVs-Gamma-Jja, we generated pairwise alignments of ERV-Gamma.1-Jja and PERV nucleotides using the full-length ERVs and performed a sliding window analysis of these pairwise alignments (Fig. 2b) [19, 20]. For comparison, we determined the similarity of the HIV-1 provirus sequence to that of its closest relative (chimpanzee SIVcpz) [21, 22]. Interestingly, *gag* and *pol* were more similar between ERV-Gamma-Jja and PERV-A, -B, and -C than HIV-1 and SIVcpz (Fig. 2b). Therefore, our results indicate that ERVs-Gamma-Jja and PERVs are homologous. However, the RBD and the proline rich-region (PRR) of the surface subunit (SU) of *env* were dissimilar between ERV-Gamma-Jja and PERV-A, -B, and -C, as also seen in HIV and SIVcpz. As RBD determines the host range [23-26], this observation suggests that ERVs-Gamma-Jja and PERVs have distinct host ranges.

We used the RBD amino acid sequences from PERV-A, -B, and -C as queries to screen for homologous viral elements. The eight significant hits (>60% amino acid identity over 80% region) were obtained in rock hyrax (*Procavia capensis*) of *Procaviidae*, and all eight hits flanking with genes were located in large scaffolds >0.3 Mb in length (Additional files 2: Table S5). We examined the sequences flanking the eight hits (especially *pol*) and found that ERVs including these hits were endogenous gamma-retroviruses. These hits were therefore designated ERVs-Gamma.n-Pca (where n = 1–8). We aligned the RBDs of PERVs and ERVs-Gamma-Pca and found that ERVs-Gamma-Pca were highly similar to PERVs (Fig. 3). Pairwise comparisons revealed that ERV-Gamma.1-Pca and ERV-Gamma.2-Pca have high identity to PERV-B (63%) but a low identity with PERV-A, -C, and PERV-IM (40–43%). Therefore, the RBD of ERVs-Gamma-Pca and PERV-B are homologous.

We also found homologous LTRs of PERVs (~73% identity) in eight *Muroidea* species (*Mus caroli*, *M. pahari*, *M. musculus*, *M. spretus*, *Apodemus speciosus*, *A. sylvaticus*, *Rattus norvegicus*, and *Phodopus*

sungorus). The coding genes (*gag*, *pol*, and *env*) near these homologous LTRs were identified (Additional files 2: Table S5). Notably, *Muridae* ERVs also showed a similar repeat structure in the U3 region and share a 70% identity to LTR B2 (Fig. 2a; Additional files 1: Fig. S3), which indicates that PERV and these rodent ERVs may share the same origin. The large indels in *Muridae* ERVs LTR further suggest that they may have an ancient origin.

To characterize the relationships between ERV-Gamma-Jja, ERVs-Gamma-Pca, *Muridae* ERVs, and PERVs, we produced phylogenetic trees of Gag, Pol and Env, first by removing the variable RBD (Fig. 2c-e). Forty ERVs from two *Muroidea* species and one primate (*M. musculus*, *R. norvegicus*, and *Microcebus murinus)* found in a previous study showing the close relationship with PERVs were also included [27, 28]. Our maximum likelihood (ML) phylogenetic tree revealed that ERVs-Gamma-Jja and PERVs clustered together with high bootstrap supports in three phylogenies (Fig. 2c-e), suggesting that they share the most recent common ancestry. However, the Gag, Pol and Env (without RBD) of ERV-Gamma-Pca were distantly related to ERVs-Gamma-Jja and PERVs, while RBD was similar to PERV, which might indicate recombinant events among ancient retroviruses. As GALV and KoRV may also have a rodent origin [29, 30], *Muridae* ERVs, ERV-Gamma-Jja, and PERV might share the same ancestral rodent retrovirus via one or more intermediate hosts (Fig. 2c-e; Fig. 4). Failure to detect any other ERV-Gamma-Jja and PERV-like elements in the remaining rodent genomes indicated that these viruses were not vertically transmitted, suggesting instead that ancient horizontal transmission occurred during evolution.

Remarkably, a new lineage close to PERV-A and PERV-C was observed, named PERV-IM (designating a PERV-intermediate type), and presented in all 14 pig genomes (Fig. 2e, Additional files 2: Table S2). The Env proteins of PERV-IMs showed relatively low similarity to PERV-A, -B, and -C, and they were clearly distinct in the RBD region (Fig. 3).

Molecular dating analysis

To roughly estimate the integration time of PERVs, ERVs-Gamma-Jja and ERVs-Gamma-Pca, we used an LTR-divergence method based on the divergence between 5'- and 3'- LTR of ERVs with a known host nucleotide substitution rate [17, 31]. However, since the nucleotide substitution rates of *S. scrofa*, *J. jaculus*, and *P. canpensis* are unknown, we used an average mammal neutral substitution rate (2.2×10^{-9} per site per year) [32] for these three species. Our results indicated that PERV-A first invaded the *Suidae* ~6.6 million years ago (MYA), while PERV-B first invaded ~6.4 MYA. In contrast, the invasions of PERV-C and PERV-IM were relatively recent (~3.4 MYA and ~4.4 MYA, respectively) (Fig. 5, Additional files 2: Table S2). Thus, the oldest PERV-A and PERV-B invaded the host just after the *Suidae* split from the ancestral group (~7.3 MYA) [33]. PERV-A, -B, and -C have continued to integrate into pig genomes,

resulting in increasing numbers of insertions. However, the LTRs of another three ERVs-Gamma-Jja were incomplete, and the time estimation of these ERV-Gamma-Jja was based on only one provirus. ERVs-Gamma.1-Jja was estimated to have integrated ~17.2 MYA, which is well before *J. jaculus* speciated (~11.1 MYA), but later than the speciation of *Dipodidae* (~42.7 MYA) [34]. ERV-Gamma-Pca integration time was calculated based on two full-length ERV-Gamma-Pcas. ERVs-Gamma-Pca insertions were estimated to be much older than PERVs (~10.7 MYA and ~8.4 MYA).

Discussion

Using systematic large-scale genome mining, we revealed hundreds of PERVs or PERV-like sequences, including some previously unidentified viruses from rodents and rock hyrax. Phylogenetic reconstruction, as well as sequence analysis, provided evidence that ERVs-Gamma-Jja from lesser Egyptian jerboa share the most recent ancestry with the modern PERVs. *Muridae* ERVs (with PERV-like LTRs) and ERVs-Gamma-Pca, from rock hyrax, also contributed to the origin of LTRs and the RBD region of modern PERVs, respectively (Fig. 2). Env phylogeny (Fig. 2c) revealed that the ancestral PERV was then speciated into different classes, in which the PERV-B emerged earlier than the other classes (Fig. 6). Previously, the insertion time of the most ancient PERV was estimated at 7.6 MYA [14], similar to our estimation that the oldest PERV was dated back to 6.6 MYA. In particular, the most ancient insertion times of ERVs-Gamma-Jja and ERVs-Gamma-Pca were roughly estimated as 17.2 and 10.7 MYA, respectively, which are much older events than the origination time of *Sus scrofa* (~7.3 MYA).

ERVs-Gamma-Jia were the closest relatives to PERVs (Fig. 2c-e). Their host, lesser Egyptian jerboa, was classed in the family *Dipodidae,* while the hosts of murine ERVs (with PERV-like LTRs) belong to family *Muridae.* Both of them were classified into the suborder *Myomorpha* and order *Rodentia*. However, the host of ERVs-Gamma-Pca, the source of PERV RBD region, was classified into the family *Procaviidae,* which is unrelated to *Rodentia*. All of this evidence supports a multiple origin history of the modern PERVs.

To reveal the evolutionary path of PERVs, we then compared the fossil records of these hosts. We noticed that Miocene (23–5.33 MY) *Suidae* fossils have been found in East Africa, Europe and Asia (http://fossilworks.org/?a=taxonInfo&taxon_no=42381); Miocene *Dipodidae* fossils have been found in North Africa, Europe, and Asia (http://fossilworks.org/?a=taxonInfo&taxon_no=41695); Pliocene (5.3–2.59 MY) *Dipodidae* fossils have been found in East Africa, thus suggesting that the *Dipodidae* may have spread to East Africa during the Miocene age. We also noticed that Miocene *Muridae* fossils have been found in Africa, Europe and Asia (http://fossilworks.org/?a=taxonInfo&taxon_no=42381), while Miocene *Procaviidae* fossils have been found in the South of Africa and East Africa (http://fossilworks.org/?a=taxonInfo&taxon_no=43293). Based on the above fossil records, the only shared geographical region for Miocene *Dipodidae*, *Procaviidae*, and *Suidae* fossils is in East Africa. We speculate that the ancestral

PERV was generated in this region via multiple recombination events involving *Rodentia* and *Dipodidae* species during Miocene (Fig. 6). This is also consistent with our dating results of the PERVs, and two PERV-like viruses—ERVs-Gamma-Jja and ERVs-Gamma-Pca.

We also checked the flanking regions of ERVs-Gamma-Jja, ERVs-Gamma-Pca, and PERVs, and they showed no similarity with each other, indicating that these ERVs were not vertically transmitted. However, we identified 29 full-length PERVs as ortholog sequences, while most PERVs, including some with same TSD (Additional files 2: Table S3), were not orthologs, revealing that PERVs were more invasive after the speciation of different pig breeds.

Conclusion

In sum, we decipher a complex evolutionary history for the modern PERVs. The ancestral PERV is likely to be derived from ancient retroviruses carried by non-porcine species via multiple recombination events. We also suggest that pig genomes have been shaped by PERVs, as specifically reflected by the PERV-associated genomic rearrangements that have occurred during porcine evolution. In other words, prior to their appearance in pigs, modern PERVs had an evolutionary history more complex than previously thought.

Materials And Methods

In silico identification of PERV and PERV-related proviruses.

To identify PERV-related elements in *Sus scrofa*, tBLASTn [35] was used and the amino acid sequences of Gag, Pol and Env of 20 representative PERV proviruses (accession numbers: HQ536016.1, HQ536015.1, HQ536013.1, KC116220.1, AY570980.1, HQ540592.1, HQ536007.1, AX546209.1, AF435967.1, AY953542.1, HQ540591.1, AY099323.1, AJ133817.1, EU523109.1, EF133960.1, AY056035.1, AY099324.1, A66553.1, HQ536011.1, and HQ536009.1) were chosen as queries to screen the 14 pig genomes available in GenBank. A 50% identity over 50% of the match region was used to filter significant hits. It has been shown that PERVs harbor two LTR structures, one with and one without a repeat structure in the U3 region[8, 14]. Using two typical LTRs as queries we extended flanking sequences of coding domains of PERVs to identify LTRs with BLASTn, and TSDs were used to define PERV boundaries. LTR lengths were defined as 100–1,000 bp. PERVs with at least one LTR and one coding gene were used in the evolutionary analysis.

To identify PERV-related proviruses in mammals, tBLASTn was used with the queries described above in 20 representative PERV proviruses to search the 304 mammal genomes available in GenBank. A 50% identity over 80% region was used to filter significant hits. LTRs were identified using LTR finder[36],

LTRharvest [37] and BLASTn. LTR lengths were also defined as 100–1,000 bp. These ERVs were named as ERV-Gamma.n-Jja, ERV-Gamma.n-Pca, ERV-Gamma.n-Asp, ERV-Gamma.n-Msp, ERV-Gamma.n-Mmu, ERV-Gamma.n-Rno, ERV-Gamma.n-Mca, ERV-Gamma.n-Mpa (in which n represents the number of the viral sequences extracted from host genome), consisted with previously study[38].

Detection of genomic rearrangement via PERVs.

To search for proviruses involved in recombination and genomic rearrangement, we constructed a maximum likelihood (ML) tree of the 5'- and 3'- LTRs of full-length PERVs using PhyML 3.1[39] with GTR+I+Γ nucleotide substitution model. LTRs less than 250 bp were not considered. Sequence alignment was performed with MAFFT 7.222[40].

Phylogenetic analyses.

To determine the evolutionary relationship among PERVs, ERVs-Gamma-Jja, ERVs-Gamma-Pca, *muridae* ERVs and representative gammaretroviruses (Additional files 2: Table S7), phylogenetic trees were inferred using the amino acid sequences of full-length PERVs and PERVs with one LTR and at least one coding gene. The length of protein sequences < 70% of the alignment are not considered for phylogenetic analyses. Significant bat viruses including *Eptesicus serotinus* bat retrovirus (EsRV) and the *Megaderma lyra* retrovirus (MlRV) are too short to include in analysis. However, their phylogenetic relationship with PERVs has been shown in previous study [41]. All Gag, Pol and Env protein sequences (additional files 3: dataset S1) were aligned in MAFFT 7.222 and confirmed manually in MEGA7/MEGA X [42, 43]. The phylogenetic history of these gammaretroviruses was then determined using the maximum likelihood (ML) method available in PhyML 3.1 [39], incorporating 100 bootstrap replicates to assess node robustness. The best-fit JTT+Γ amino acid substitution model was selected for Gag, Pol and JTT+Γ+I for Env using the ProtTest 3.4.2 [44].

Dating estimation of PERV, ERV-Gamma-Jja and ERV-Gamma-Pca.

The 5' and 3' LTRs of ERVs are identical at the point of integration, and then diverge and evolve independently [31]. So the ERV integration time can be calculated using the following relation: T = (D/R)/2, in which T is the invasion time (million years, MY), D is the number of nucleotide differences per site between the two LTRs, and R is the genomic substitution rate (nucleotide substitutions per site, per year). We used the previously estimated average mammal substitution rate (2.2×10^{-9} per site per year) [32], as no substitution rate (r) has yet been estimated for the *S. Scrofa*, *J. jaculus* and *P. canpensis*.

Abbreviations

ERV: endogenous retroviruses PERV: porcine endogenous retroviruses RBD: receptor binding domain LTR: long terminal repeat TSD: target site duplication PRR: proline rich-region SU: surface subunit EsRV; Eptesicus serotinus bat retrovirus MlRV: Megaderma lyra retrovirus

Declarations

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Availability of data and materials. All datasets used and analyzed during the current this study are included in this published article (and its additional files).

Competing interests. The authors declare that they have no competing interests.

Funding. This work was supported by the Special Key Project of Biosafety Technologies (2017YFC1200800) for the National Major Research & Development Program of China. J.C. is supported by National Natural Science Foundation of China under grant no. 31671324 and CAS Pioneer Hundred Talents Program.

Author contributions. J.C. conceived and designed the research. Y.C. and M.C. conducted the analyses. X.D. participated in the analysis of full-length PERVs. J.C. supervised the whole project. All authors participated in the project discussion and manuscript preparation.

Acknowledgement. Not applicable.

Additional Files

Additional files 1. Fig. S1. Maximum likelihood (ML) tree of the 5’ and 3’ LTRs of all full-length PERVs. **Fig. S2.** Detailed descriptions of ERV-Gamma.1-Jja genome. **Fig. S3.** The alignment of LTRs of PERVs, ERV-Gamma-Jja, and one of Muridae ERVs. **Fig. S4.** The complete phylogenetic tree of Gag. Fig. **S5.** The complete phylogenetic tree of Pol. **Fig. S6.** The complete phylogenetic tree of Env.

Additional files 2. Table S1. The information of pig, rodent, and rock hyrax genomes used for data mining. **Table S2.** The information of full-length and near full-length PERVs. **Table S3.** The recombination-related information of full-length PERVs shown in Fig. S1. **Table S4.** The information of 304 mammals used for PERV-like sequences mining. **Table S5.** The matching contigs identified in mammal genomes. **Table S6.** The information of genes flanking the ERVs-Gamma-Jja and ERVs-Gamma-Pca. **Table S7**. The information of representative retroviruses used for phylogenetic analysis.

Additional files 3. Data set S1. The alignments used to build the phylogenetic trees of Gag, Pol and Env represented in Fig. S4, S5, and S6, respectively.

References

1. Niu D, Wei HJ, Lin L, George H, Wang T, Lee IH, Zhao HY, Wang Y, Kan Y, Shrock E, et al: **Inactivation of porcine endogenous retrovirus in pigs using CRISPR-Cas9.** *Science* 2017, **357:**1303-1307.
2. Denner J, Tonjes RR: **Infection barriers to successful xenotransplantation focusing on porcine endogenous retroviruses.** *Clin Microbiol Rev* 2012, **25:**318-343.
3. Morozov VA, Wynyard S, Matsumoto S, Abalovich A, Denner J, Elliott R: **No PERV transmission during a clinical trial of pig islet cell transplantation.** *Virus Res* 2017, **227:**34-40.
4. Crossan C, Mourad NI, Smith K, Gianello P, Scobie L: **Assessment of porcine endogenous retrovirus transmission across an alginate barrier used for the encapsulation of porcine islets.** *Xenotransplantation* 2018**:**e12409.
5. Wynyard S, Nathu D, Garkavenko O, Denner J, Elliott R: **Microbiological safety of the first clinical pig islet xenotransplantation trial in New Zealand.** *Xenotransplantation* 2014, **21:**309-323.
6. Denner J: **Paving the Path toward Porcine Organs for Transplantation.** *N Engl J Med* 2017, **377:**1891-1893.
7. Niebert M, Tonjes RR: **Evolutionary spread and recombination of porcine endogenous retroviruses in the suiformes.** *J Virol* 2005, **79:**649-654.
8. Scheef G, Fischer N, Krach U, Tonjes RR: **The number of a U3 repeat box acting as an enhancer in long terminal repeats of polytropic replication-competent porcine endogenous retroviruses dynamically fluctuates during serial virus passages in human cells.** *J Virol* 2001, **75:**6933-6940.
9. Wilson CA, Laeeq S, Ritzhaupt A, Colon-Moran W, Yoshimura FK: **Sequence analysis of porcine endogenous retrovirus long terminal repeats and identification of transcriptional regulatory regions.** *J Virol* 2003, **77:**142-149.
10. Li Z, Ping Y, Shengfu L, Hong B, Youping L, Yangzhi Z, Jingqiu C: **Phylogenetic relationship of porcine endogenous retrovirus (PERV) in Chinese pigs with some type C retroviruses.** *Virus Res* 2004, **105:**167-173.
11. Cui J, Tachedjian G, Tachedjian M, Holmes EC, Zhang S, Wang LF: **Identification of diverse groups of endogenous gammaretroviruses in mega- and microbats.** *J Gen Virol* 2012, **93:**2037-2045.
12. Benveniste RE, Todaro GJ: **Evolution of type C viral genes: preservation of ancestral murine type C viral sequences in pig cellular DNA.** *Proc Natl Acad Sci U S A* 1975, **72:**4090-4094.
13. Huh JW, Cho BW, Kim DS, Ha HS, Noh YN, Yi JM, Lee WH, Kim HS: **Long terminal repeats of porcine endogenous retroviruses in Sus scrofa.** *Arch Virol* 2007, **152:**2271-2276.
14. Tonjes RR, Niebert M: **Relative age of proviral porcine endogenous retrovirus sequences in Sus scrofa based on the molecular clock hypothesis.** *J Virol* 2003, **77:**12363-12368.

15. Niebert M, Kurth R, Tonjes RR: **Retroviral safety: analyses of phylogeny, prevalence and polymorphisms of porcine endogenous retroviruses.** *Ann Transplant* 2003, **8:**56-64.
16. Mayer J, Blomberg J, Seal RL: **A revised nomenclature for transcribed human endogenous retroviral loci.** *Mob DNA* 2011, **2:**7.
17. Johnson WE, Coffin JM: **Constructing primate phylogenies from ancient retrovirus sequences.** *Proc Natl Acad Sci U S A* 1999, **96:**10254-10260.
18. Hughes JF, Coffin JM: **Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution.** *Nat Genet* 2001, **29:**487-489.
19. Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, Ingersoll R, Sheppard HW, Ray SC: **Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination.** *J Virol* 1999, **73:**152-160.
20. Zhuo X, Feschotte C: **Cross-Species Transmission and Differential Fate of an Endogenous Retrovirus in Three Mammal Lineages.** *PLoS Pathog* 2015, **11:**e1005279.
21. Martoglio B, Graf R, Dobberstein B: **Signal peptide fragments of preprolactin and HIV-1 p-gp160 interact with calmodulin.** *Embo j* 1997, **16:**6636-6645.
22. Corbet S, Muller-Trutwin MC, Versmisse P, Delarue S, Ayouba A, Lewis J, Brunak S, Martin P, Brun-Vezinet F, Simon F, et al: **env sequences of simian immunodeficiency viruses from chimpanzees in Cameroon are strongly related to those of human immunodeficiency virus group N from the same geographic area.** *J Virol* 2000, **74:**529-534.
23. Argaw T, Wilson CA: **Detailed mapping of determinants within the porcine endogenous retrovirus envelope surface unit identifies critical residues for human cell infection within the proline-rich region.** *J Virol* 2012, **86:**9096-9104.
24. Watanabe R, Miyazawa T, Matsuura Y: **Cell-binding properties of the envelope proteins of porcine endogenous retroviruses.** *Microbes Infect* 2005, **7:**658-665.
25. Denner J: **Recombinant porcine endogenous retroviruses (PERV-A/C): a new risk for xenotransplantation?** *Arch Virol* 2008, **153:**1421-1426.
26. Ericsson TA, Takeuchi Y, Templin C, Quinn G, Farhadian SF, Wood JC, Oldmixon BA, Suling KM, Ishii JK, Kitagawa Y, et al: **Identification of receptors for pig endogenous retrovirus.** *Proc Natl Acad Sci U S A* 2003, **100:**6759-6764.
27. Hayward A, Cornwallis CK, Jern P: **Pan-vertebrate comparative genomics unmasks retrovirus macroevolution.** *Proc Natl Acad Sci U S A* 2015, **112:**464-469.
28. Hayward A, Grabherr M, Jern P: **Broad-scale phylogenomics provides insights into retrovirus-host evolution.** *Proc Natl Acad Sci U S A* 2013, **110:**20146-20151.
29. Wolgamot G, Miller AD: **Replication of Mus dunni endogenous retrovirus depends on promoter activation followed by enhancer multimerization.** *J Virol* 1999, **73:**9803-9809.
30. Lieber MM, Sherr CJ, Todaro GJ, Benveniste RE, Callahan R, Coon HG: **Isolation from the asian mouse Mus caroli of an endogenous type C virus related to infectious primate type C viruses.** *Proc*

Natl Acad Sci U S A 1975, **72:**2315-2319.

31. Dangel AW, Baker BJ, Mendoza AR, Yu CY: **Complement component C4 gene intron 9 as a phylogenetic marker for primates: long terminal repeats of the endogenous retrovirus ERV-K(C4) are a molecular clock of evolution.** *Immunogenetics* 1995, **42:**41-52.
32. Kumar S, Subramanian S: **Mutation rates in mammalian genomes.** *Proc Natl Acad Sci U S A* 2002, **99:**803-808.
33. Frantz LAF: **Speciation and domestication in Suiformes: a genomic perspective.** *Wageningen University* 2015.
34. Zhang Q, Xia L, Kimura Y, Shenbrot G, Zhang Z, Ge D, Yang Q: **Tracing the Origin and Diversification of Dipodoidea (Order: Rodentia): Evidence from Fossil Record and Molecular Phylogeny.** *Evolutionary Biology* 2013, **40:**32-44.
35. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
36. Xu Z, Wang H: **LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons.** *Nucleic Acids Res* 2007, **35:**W265-268.
37. Ellinghaus D, Kurtz S, Willhoeft U: **LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons.** *BMC Bioinformatics* 2008.
38. Gifford RJ, Blomberg J, Coffin JM, Fan H, Heidmann T, Mayer J, Stoye J, Tristem M, Johnson WE: **Nomenclature for endogenous retrovirus (ERV) loci.** *Retrovirology* 2018, **15:**59.
39. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O: **New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.** *Syst Biol* 2010, **59:**307-321.
40. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability.** *Mol Biol Evol* 2013, **30:**772-780.
41. Denner J: **Transspecies Transmission of Gammaretroviruses and the Origin of the Gibbon Ape Leukaemia Virus (GaLV) and the Koala Retrovirus (KoRV).** *Viruses* 2016, **8**.
42. Kumar S, Stecher G, Tamura K: **MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets.** *Mol Biol Evol* 2016, **33:**1870-1874.
43. Kumar S, Stecher G, Li M, Knyaz C, Tamura K: **MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms.** *Mol Biol Evol* 2018, **35:**1547-1549.
44. Abascal F, Zardoya R, Posada D: **ProtTest: selection of best-fit models of protein evolution.** *Bioinformatics* 2005, **21:**2104-2105.

Figures

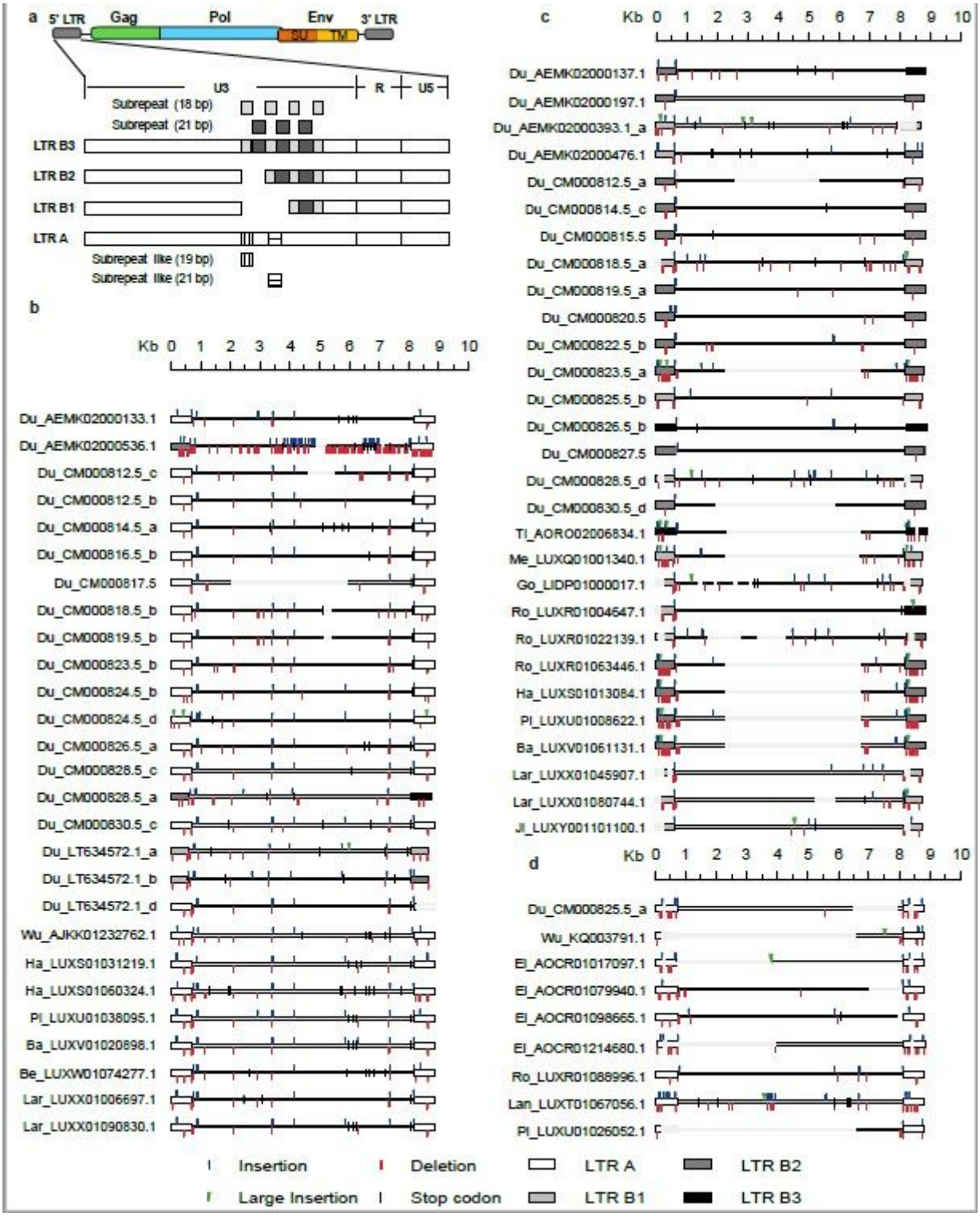

Figure 1

PERV proviruses in porcine genomes. (a) Organization and genomic structure of PERVs, including gag, pol and env genes and different types of LTRs. LTRs of PERVs were classified by the presence (type B) or absence (type A) of the 18 bp and 21 bp repeat structure. Type B LTRs were divided into three subtypes (LTR B1, LTR B2 and LTR B3). Proviruses of PERV-A (b), PERV-B and (c) and PERV-C (d) groups depicted based on reference PERV-A (accession number: AF435967.1), PERV-B (accession number: EU523109.1) and PERV-C (accession number: HQ536015.1). LTR A, B1, B2 and LTR B3 are presented in white, light

gray, dark gray and black, respectively. Insertions and deletions (< 50 bp) are depicted with blue and red flags, respectively. Larger insertions (>50 bp) are labeled with green arrow; large deletions (>50 bp) are shown without lines. Stop codons are showed with a black flag. (Du, Duroc pig; Wu, Wuzhishan pig; El, Ellegaard pig; Ti, Tibetan pig; Go, Goettingen pig; Me, Meishan pig; Ro, Rongchang pig; Ha, Hampshire pig; Lan, Landrace pig; Pi, Pietrain pig; Ba, Bamei pig; Be, Bekshire pig; Lar, LargeWhite pig; Ji, Jinhua pig).

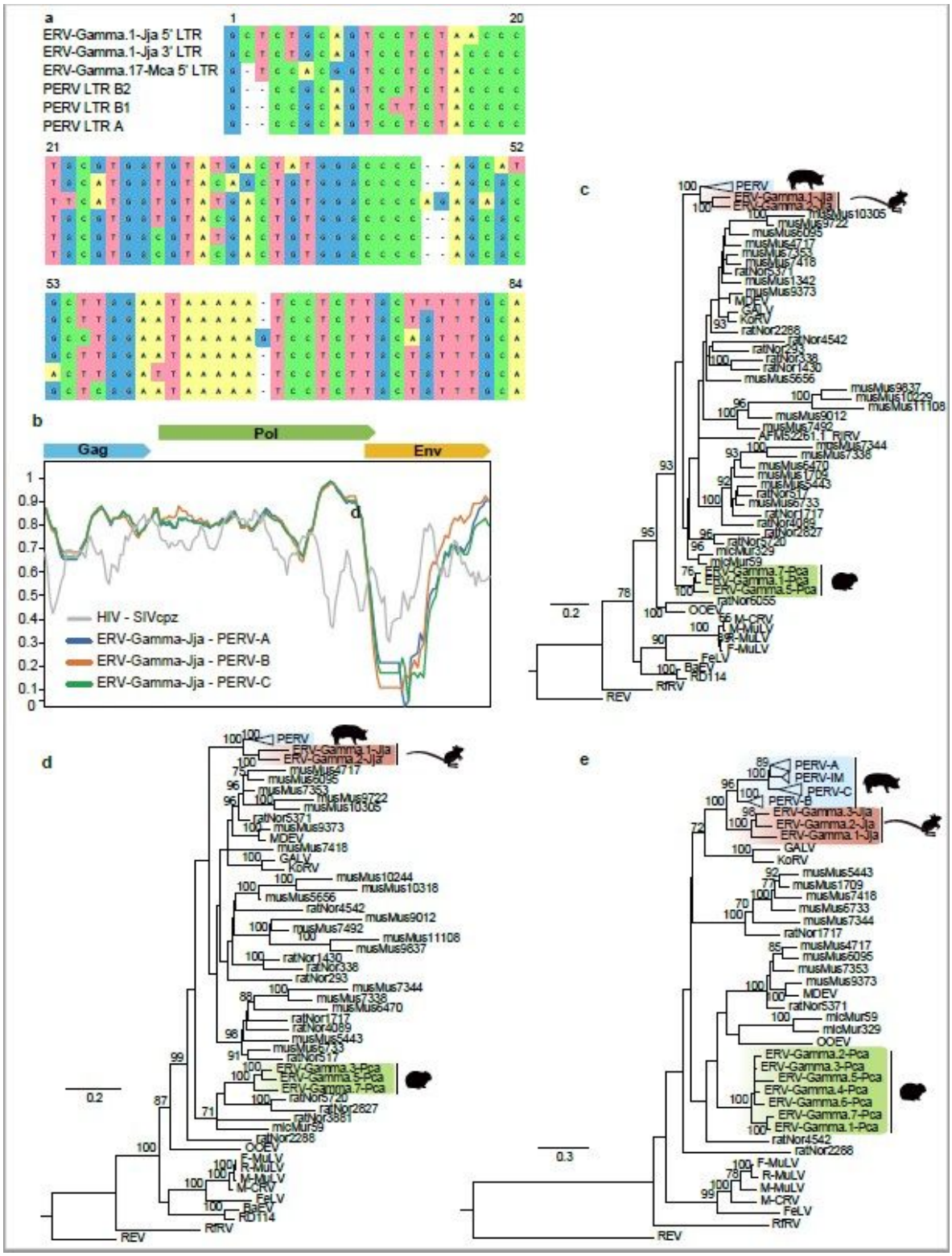

Figure 2

Comparison of PERVs, ERVs-Gamma-Jja, Muridae ERVs and ERVs-Gamma-Pca. (a) Alignment of R region of LTR in ERV-Gamma-Jja, Muridae ERVs and PERVs. (b) Sliding window analysis of percent sequence identity along pairwise alignments of proviruses without LTRs. Phylogenetic trees of Gag (c), Pol (d) and Env (e) inferred using the amino acid sequences of PERVs, ERVs-Gamma-Jja, Muridae ERVs, ERVs-Gamma-Pca and other representative gammaretroviruses (Table S7). Bootstrap values < 70% are not shown in phylogenetic trees. Trees were rooted to Reticuloendotheliosis virus (REV). The complete phylogenetic trees of Gag (c), Pol (d) and Env (e) are presented in additional files 1: Fig. S4-S6, respectively. All abbreviations can be found in Table S7.

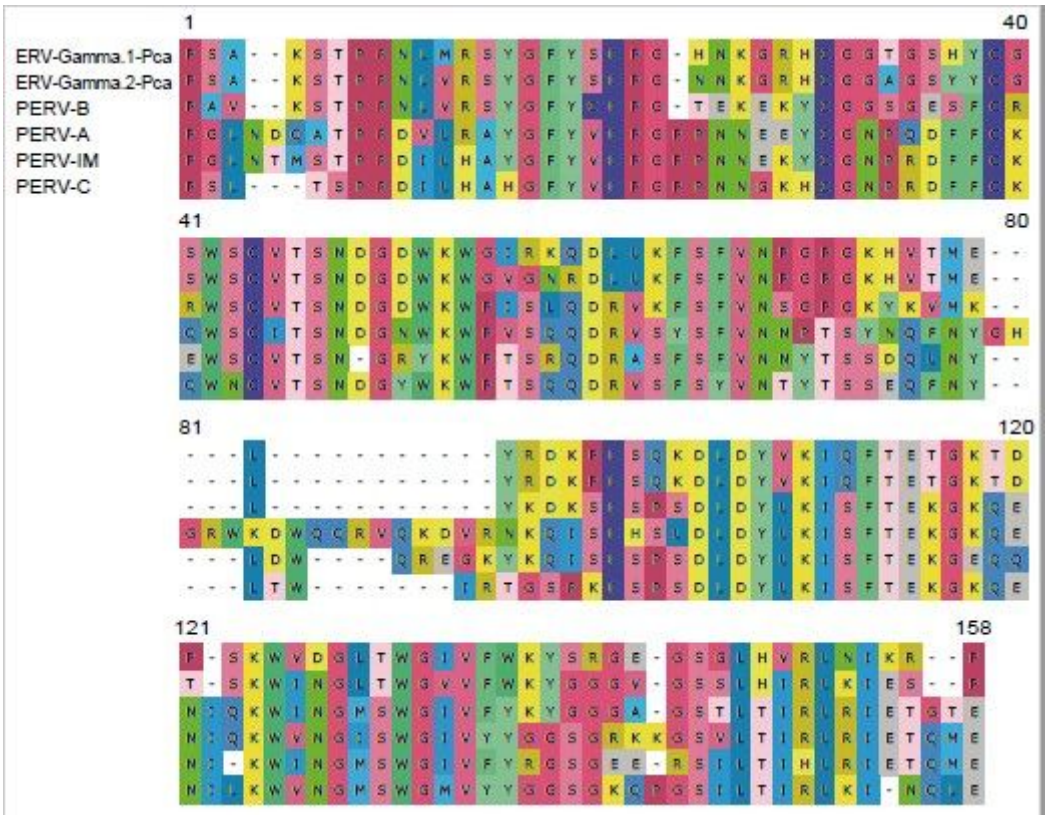

Figure 3

Amino acid sequence comparison of the RBD of PERVs and ERVs-Gamma-Pca. Two ERVs-Gamma-Pca were aligned to PERV-A, -B, -C, and the newly discovered PERV-IM.

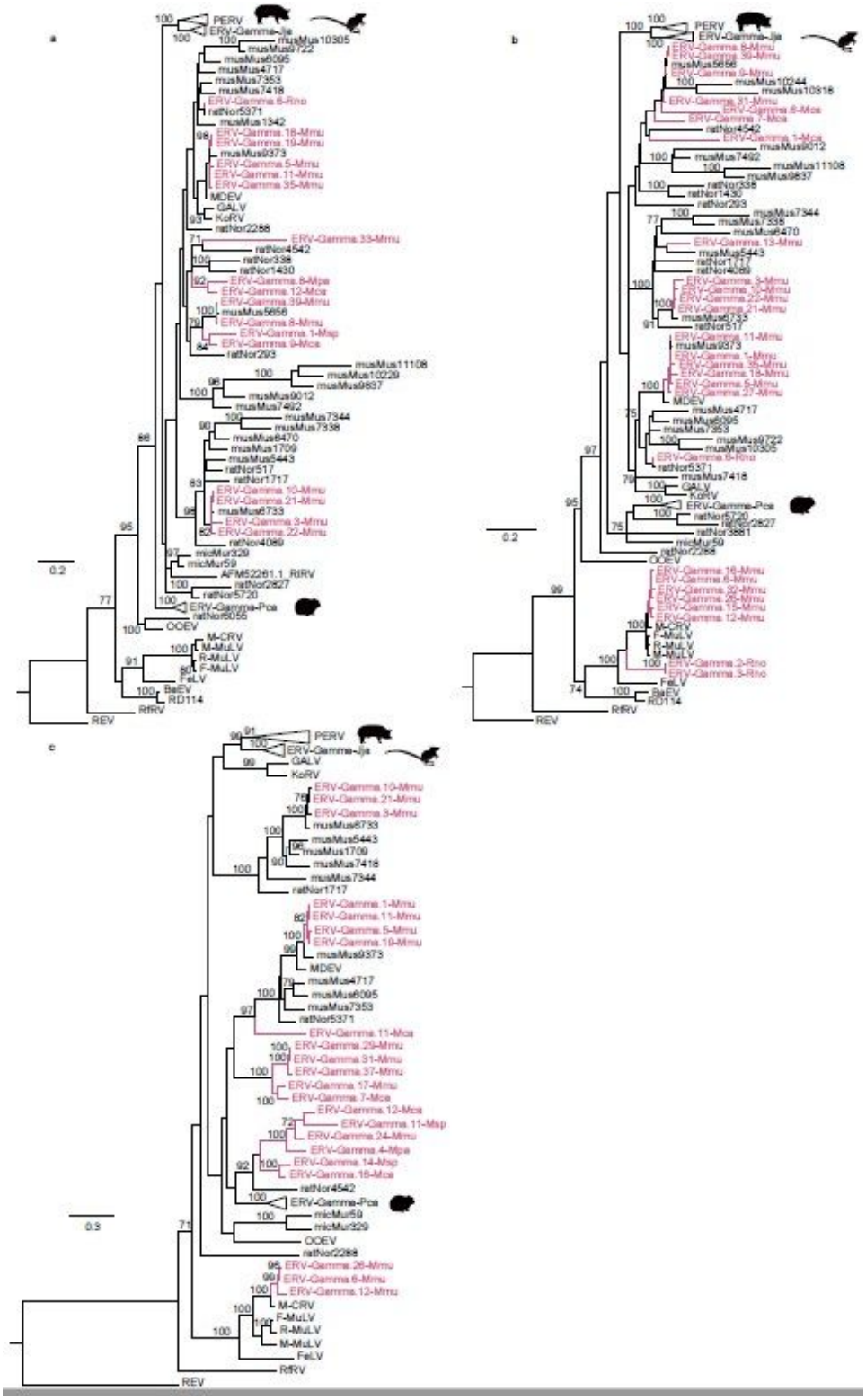

Figure 4

The complete phylogenetic tree of Gag, Pol and Env of Muroidea ERVs, ERVs-Gamma-Jja, ERVs-Gamma-Pca, PERVs and representative retroviruses. Phylogenetic trees of Gag (a), Pol (b) and Env (c) constructed using amino acid (aa) sequences of PERVs, ERVs-Gamma-Jja, ERVs-Gamma-Pca, Muroidea ERVs (Additional file 2: Table S5) and other representative gammaretroviruses. he best-fit JTT+Γ amino acid substitution model was selected for Gag, Pol and JTT+Γ+I for Env. Bootstrap values <70% are not

presented in the phylogenetic trees. Trees were rooted using Reticuloendotheliosis virus (REV). The information of Muroidea genomes presents in Table S1. Previous found ERVs are labeled in blue. Muroidea ERVs are labeled in red. The alignment used to build the phylogenetic tree is represented in Dataset S2.

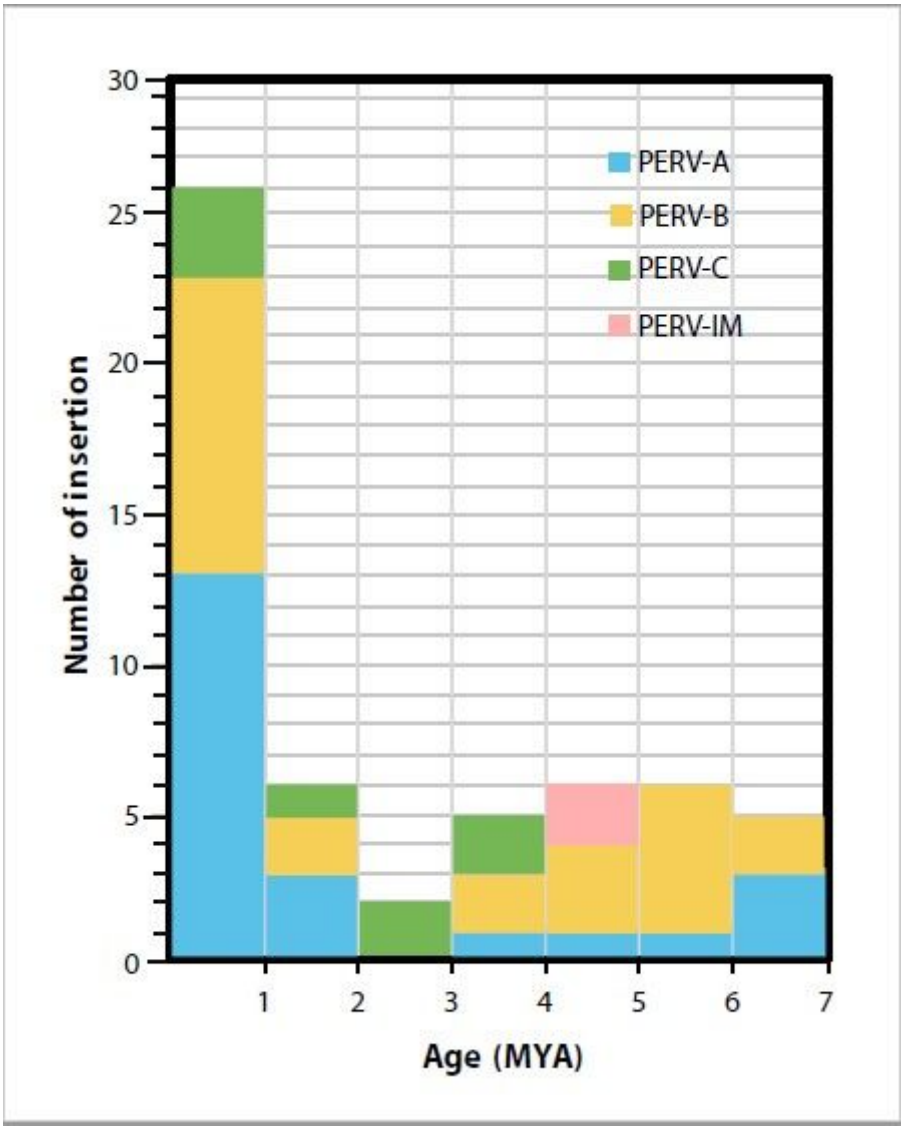

Figure 5

Dating of PERVs insertion based on LTR-LTR divergence. The Y axis shows the number of insertions for different classes and X axis indicates the putative insertion time using MY as a unit. PERVs with LTR > 300 bp are used for estimation.

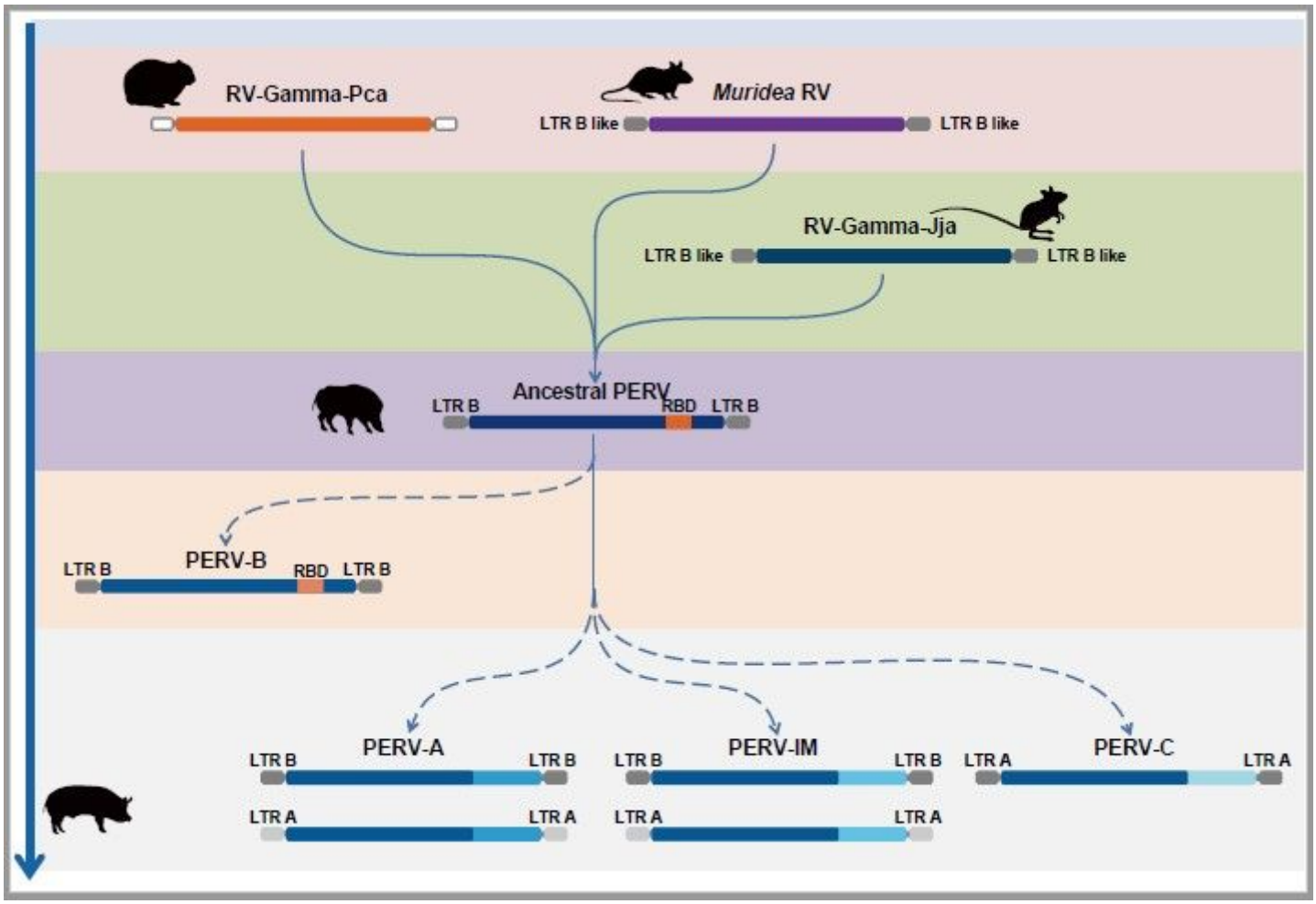

Figure 6

Scenario of genesis of PERVs. A schematic representation of PERV evolutionary history is shown, summarizing our hypothesis regarding the origin and evolution of PERVs. Arrow on the left indicates putative timeline of each type of ERVs. Different background colors illustrate putative evolution periods of modern PERVs. RV-Gamma-Jja, Muridae RV, RV-Gamma-Pca and ancestral PERV represent the exogenous forms of the retroviruses from their hosts. Solid blue arrows show origination and dotted blue arrows show speciation. Abbreviation: RBD, receptor binding domain; LTR, long terminal repeat.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfiles312.5.txt
- Additionalfiles112.5.docx
- Additionalfiles212.5.docx