Ancient origin and complex evolution of porcine endogenous retroviruses

Yicong Chen[1,2] ¶, Mingyue Chen[1¶], Xiaoyan Duan[1,2], and Jie Cui[1]*

[1] CAS Key Laboratory of Special Pathogens and Biosafety, Center for Emerging Infectious Diseases, Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan 430071, China

[2] University of Chinese Academy of Sciences, Beijing 100049, China

* Corresponding author

E-mail: jiecui@wh.iov.cn (JC)

¶ Y.C. and M.C. contributed equally to this work.

Abstract

Porcine endogenous retroviruses (PERVs) are potential infectious agents of xenotransplantation as they are able to infect human cells and can be endogenized. To trace the origin of PERVs, we performed large-scale genomic mining of 142 mammal and 14 pig genomes and investigated genomic dynamics and evolution of PERV-related viral "fossils". Large-scale genetic alterations were found in most PERVs with many indels discovered, indicating the ancient origin of these viruses. Remarkably, two none-porcine species, lesser Egyptian jerboa (*Jaculus jaculus*) and rock hyrax (*Procavia capensis*), harbor endogenous retroviruses (ERVs), named eJJRV and ePCRV, which are closely related to PERVs. Molecular dating and phylogenetic analyses suggest that ancestral PERV originated from recombination of JJRV and PCRV in ancient pigs, which likely occurred in the late Miocene in Africa. Furthermore, we have discovered evidence of genomic rearrangement via PERVs during porcine evolution. Taken together, we decipher a complex evolutionary history for the modern PERVs.

Introduction

Xenotransplantation, the transplantation of tissues and organs from one species to another, may alleviate shortages of human donor organs (1, 2). Porcine organs are suitable for xenotransplantation due to the similar size and function of porcine and human organs, and the fact that pigs can be bred in large numbers (3). However, the potential risk of cross-species transmission of porcine microorganism specific porcine endogenous retroviruses (PERVs) limits the xenotransplantation of porcine organs into humans (4). PERV, as a member of retroviruses, could potentially cause immunodeficiency and tumorigenesis (3, 5, 6).

PERVs are endogenous gammaretroviruses, and exist in the genomes of all pig strains (3, 7). The envelope (*env*) genes of three PERV classes (PERV-A, -B and -C) differ, especially with respect to the receptor-binding domain (RBD) (8). Although there is no evidence of PERV transmission in patients receiving encapsulated pig islets (9-11), PERV-A and -B have been observed to infect both human cells and pig cells while PERV-C infects only pig cells (12). PERVs may also integrate into the human genome in vitro (13, 14). In pig cells, PERV-C can recombine with the *env* of PERV-A to produce A/C recombinants, which can infect human cells more efficiently (12). This increases the inherent risk in xenotransplantation and xenogeneic cell therapies.

While several studies have examined the evolutionary relationships between PERVs and other viruses, the origin of PERVs remains unknown (15, 16). 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 (17). However, the common warthog (*Phacochoerus africanus*) carries PERVs, suggesting that an ancestral porcine species from the Miocene period (3.5 to 7.5 MYA) carried PERVs (17). 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 (18, 19). The 39-bp repeats carried by PERV-A and -B confer strong promoter activity and thus increase transcription.(18, 19). However, the 39-bp repeat structure is absent in some PERV-A and all PERV-C. These PERVs thus have low transcriptional activity (18, 19). 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) (20). Indeed, LTRs of PERV-A, -B and LTR-IS (a LTR family found solely in the mouse genome) have similar structure (20). The conserved LTR sequences across pigs and mice might had originated from a common exogenous viral element, but have evolved independently (20). Thus, little is known about the evolutionary history of PERVs, their history is increasingly traceable as the number of available mammalian genomes grows. Using genome mining, we find that PERVs are ancient (dating from the late Miocene period), and for the first time, we reveal that the PERV ancestor likely originated from co-infection and recombination of non-porcine endogenous retroviruses (ERVs).

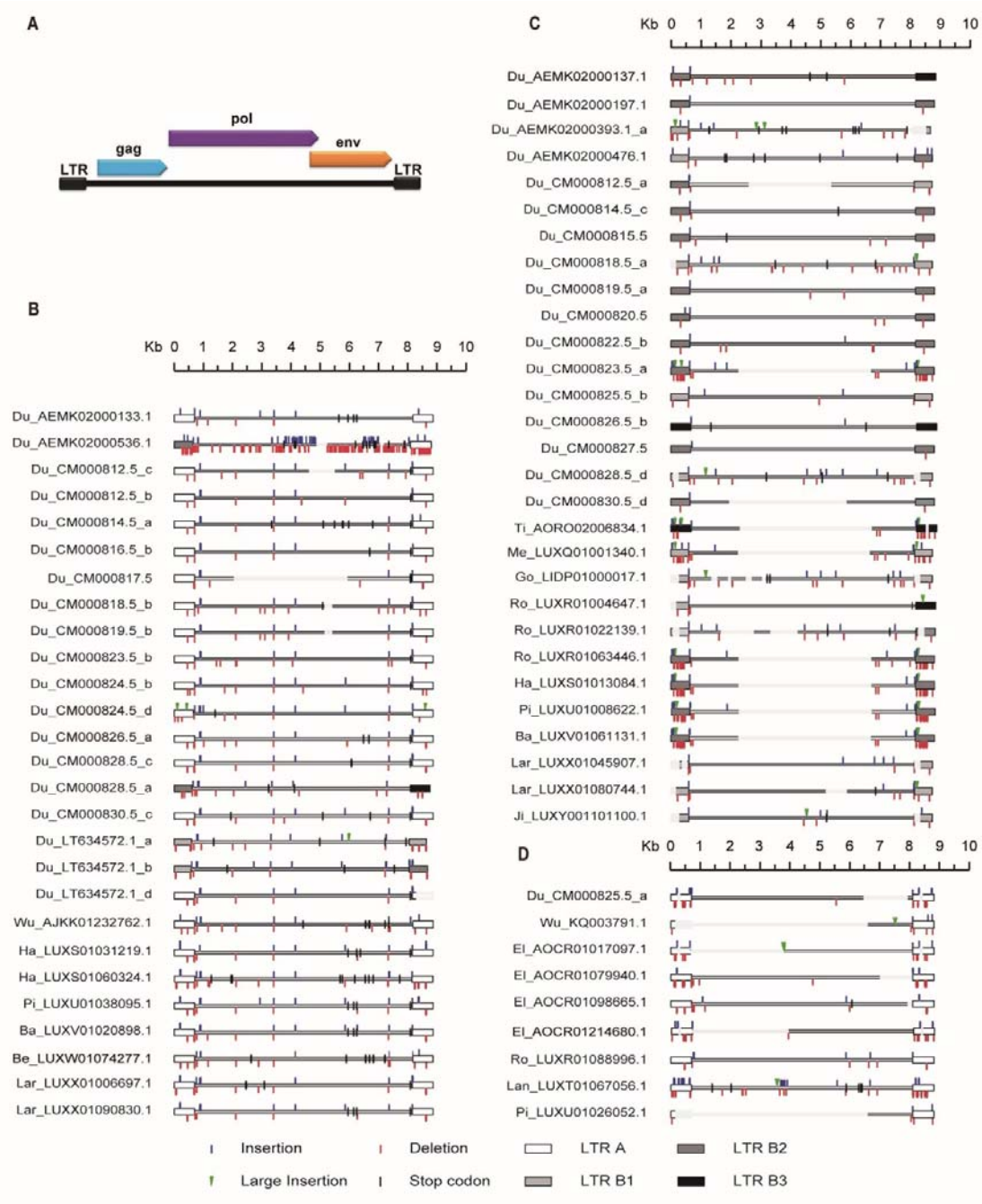
Results

Characterization of putative full-length PERVs.

Using previously reported PERV sequences as queries, we mined 14 pig genomes (Table S1) available in GenBank, and showed detailed genome-wide distribution of full-length PERVs with two flanking LTRs. We initially compiled a PERV dataset that included 84 previously classified (30 PERV-A, 39 PERV-B and 15 PERV-C) and 18 unclassified PERVs (i.e., lacking of *env* gene) (Table S2). The number of classified PERVs ranged from 38 copies in Duroc pig to 1 in the Tibetan pig. We identified 2–10 PERVs in each of 12 other pig breeds, including Meishan, Goettingen, and Large White. We removed 19 previously classified PERV sequences that were low quality fragments (> 200 “N” bases). The final dataset consisted of 65 high quality classified PERV (27 PERV-A, 29 PERV-B, and 9 PERV-C), and the genomic structures of the 65 PERVs are summarized in Fig. 1. PERVs had large-scale genetic alterations induced by indels and stop codons (Fig. 1), indicating a relatively long evolutionary history. PERV LTR were classified by the presence (LTR B) or absence (LTR A) of the 18 bp and 21 bp repeat structure reported previously (8, 18, 21).Three different type 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 65 high-quality PERVs we analyzed, we assigned 57, of which 32 (>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

97 contained different types of 5'- and 3'- LTR, which may reflect PERV recombination over

98 evolutionary time.



99

100 **Fig. 1. PERV proviruses in porcine genomes.** (**A**) Genomic structure of PERV, including

101 *gag*, *pol* and *env* genes and LTRs. Proviruses of PERV-A (**B**), PERV-B (**C**) and PERV-C

102 (**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). 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 3 subtypes (LTR B1, LTR B2 and LTR B3). 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. (Abbreviation: 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)

Recombination.

To identify recombined PERVs, we constructed a neighbor-joining tree representing the 5'- and 3'- LTR sequences of full-length PERVs across 14 genomes. The resulting phylogenetic tree was divided into three large clusters (Fig. S1), suggesting that the ages of individual PERVs varied and that three large integration events had occurred. Retrovirus integration creates a short duplication called target site duplication (TSD) flanking the LTR (22, 23). Here, 4 bp TSDs were flanking the provirus. Remarkably, 11 PERVs did not share the same TSD (Table 1, Table S3), likely due to chromosomal rearrangement through homologous recombination between distant PERVs, as mentioned in a previous study of primate ERV (24).

125 **Table 1. PERVs with different TSDs.**

Name	Accession number	Divergent LTRs based on structure [a]	Divergent LTR based on tree	Flanking TSD [b]	
				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

126 [a] LTRs of PERVs are divided into 4 types (LTR A, B1, B2, and B3). If two different types
127 of LTRs are flanking the PERV, the LTRs are divergent.

128 [b] Only TSDs flanking the intact 5' and 3' LTRs sequences were analyzed

129

130 **Detection of PERV-related sequences in mammalian genomes.**

131 After screening 142 mammalian genomes (Table S4) in Genbank, we identified a
132 sequence (accession number: NW_004504334.1) in the genome of lesser Egyptian jerboa
133 (*Jaculus jaculus*) that showed highly significant similarity (for *gag* and *pol*: >75%
134 nucleotide identity over 95% region; for *env*: >75% nucleotide identity over 55% region)
135 to PERVs using tBLASTn and choosing three major proteins (Gag, Pol and Env) of
136 PERVs as queries. Using this PERV-related sequence as query, three other possible PERV
137 related sequences were identified in *J. jaculus* (accession number: NW_004504375.1,

NW_004504378.1, and NW_004504445.1) with >85% nucleotide identity over 80% of the query sequence. The four PERV-related sequences identified in *J. jaculus* were designated eJJRVs. These sequences were located in large scaffolds > 5 Mb long, which indicated that the eJJRV sequences were relatively reliable. One full-length eJJRV (accession number: NW_004504334.1) is annotated in Fig. S2.

To demonstrate the similarity between PERVs and eJJRVs, we generated pairwise alignments of eJJRV and PERV nucleotides using the most closely related full-length ERVs, and performed a sliding window analysis of these pairwise alignments (25, 26). For comparison, we determined the similarity of HIV-1 provirus sequence to that of its closest relative (chimpanzee SIVcpz) (27, 28). We found that between eJJRV and PERV-A, -B and -C, *gag* and *pol* were more similar than HIV-1 and SIVcpz (Fig. 2D). However, the RBD and the proline rich-region (PRR) of the surface subunit (SU) of *env* were dissimilar between eJJRV and PERV-A, -B and -C, and this pattern was also found between HIV and SIVcpz. In PERVs, the RBD and PRR determine the host range (29-32), suggesting that, although *gag* and *pol* were similar between eJJRVs and PERVs, they have a distinct host range. To characterize the relationship between eJJRVs and PERVs, we constructed phylogenetic trees of Gag, Pol and Env, first removing the divergent RBD. Data show that eJJRVs clustered with PERVs (Fig. 2A-C), which suggested that PERVs and eJJRVs might share common ancestor. In Fig. 2C, a sub-branch close to PERV-A and PERV-C was showed, and the PERVs in the branch were named as PERV-IMs, which were present in all 14 pig genomes. The Env proteins of PERV-IMs showed relatively low similarity to

that of PERV-A, -B, and -C. And RBD region alignment suggested that PERV-IM was distinct from PERV-A, -B and -C (Fig. 3). So PERV-IM could be a new class of PERVs.

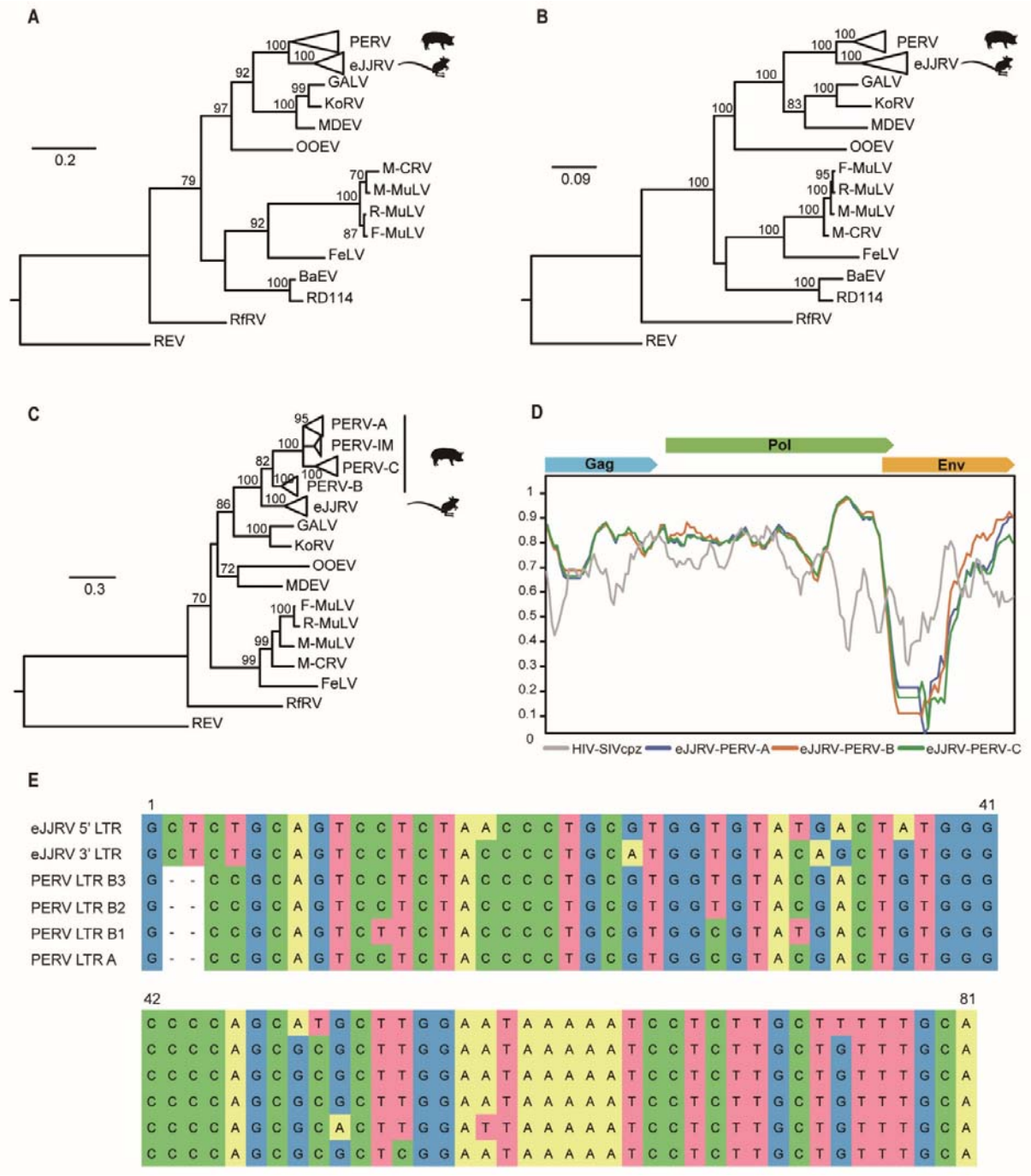


Fig. 2. The comparison of PERVs and eJJRVs. Phylogenetic trees of Gag (**A**), Pol (**B**) and Env (**C**) constructed using amino acid sequences of PERVs, eJJRVs and other

representative gammaretroviruses (Table S5). Bootstrap values <65% are not shown in phylogenetic trees. Trees were rooted using Reticuloendotheliosis virus (REV). The complete phylogenetic trees of Gag (**A**), Pol (**B**) and Env (**C**) are shown in Fig. S3-S5, respectively. (**D**) Sliding window analysis of percent sequence identity along pairwise alignments of proviruses without LTRs. (**E**) Alignment of R region of LTR in eJJRV and PERVs.

As the quality of another three of the eJJRVs were poor, we were only able to identify one pairwise eJJRV LTRs. The length of 3'- LTR of the eJJRV is 674 bp while 5'- LTR is 932bp which has a 258 bp insertion. We aligned eJJRV LTRs with PERV LTRs. The start of the U3 region and the end of the U5 region are distinct and not included in the alignment (Fig. S6). The alignment of the R region supported a close relationship between the eJJRV and the PERVs (Fig. 2E). The eJJRV LTRs included a repeat structure (three 18 bp and two 21 bp repeat sequences) in the U3 region, identical to that of the PERV LTR B2. Alignment analysis of the repeat structure revealed a closer relationship between LTRs of the eJJRV and LTR B2 of PERVs (Fig. S6). Furthermore, 3' LTR of eJJRV had high identity with LTR B2 of PERV (~73%). Therefore, our results indicated that eJJRVs and PERVs were homologous.

The RBDs of eJJRVs and PERVs were distinct, so we used the RBD amino acid sequences from PERV-A, -B and -C as queries to screen homologous viral elements. The eight

significant hits (>60% amino acid identity over 80% region) were obtained in rock hyrax (*Procavia canpensis*) of *Procaviidae*, and all 8 hits were located in large scaffolds >0.3 Mb long (accession number: KN678690.1, KN676491.1, KN678005.1, KN677924.1, KN676905.1, KN676182.1, KN680906.1, and KN676638.1). We examined the gene flanking the eight hits (especially *pol*), and found that ERVs including these hits were endogenous gammaretroviruses. These hits were therefore designated ePCRVs. We aligned the RBDs of PERVs and ePCRVs, and found that ePCRVs were highly similar to PERVs (Fig. 3). To quantify the homology between ePCRVs and PERVs, we made pairwise comparisons. Our comparisons suggested that ePCRV_1 and ePCRV_2 had a high identity with PERV-B (63%) but a low identity with PERV-A, -C and PERV-IM (40–43%). Therefore, PERV RBD might be derived from ePCRVs, and the divergence of RBD of PERVs might have occurred after the recombination of PCRVs and JJRVs.

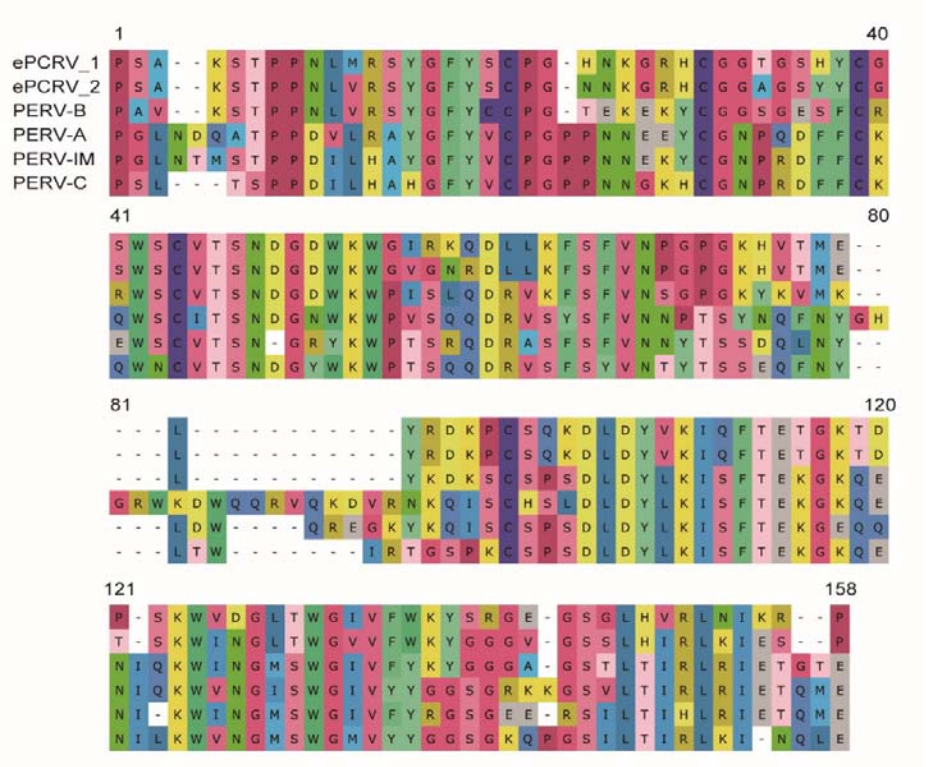


Fig. 3. Amino acid sequence comparison of RBD of PERVs and ePCRVs. Two ePCRVs (ePCRV_1 and ePCRV_2) predicted to harbor both the 5' and 3' LTRs were selected to align with PERVs.

Molecular dating analysis.

To better understand the integration time of PERVs, we used an LTR-divergence method to estimate when PERVs and eJJRVs invaded the host genome. This estimation method was based on divergence between 5'- and 3'- LTR of ERVs. Because nucleotide substitution rates of *S. scrofa*, *J. jaculus* and *P. canpensis* were unknown, we used an average mammal neutral substitution rate (2.2×10^{-9} per site per year) (33) for these three species. Our results indicated that PERV-A first invaded the *Suidae* ~6.6 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. 4, 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) (34). PERV-A, -B and -C has continued to integrate into pig genomes, resulting in increasing numbers of insertions. Because the LTRs of three eJJRVs were incomplete, our eJJRV results were based on only one provirus. eJJRVs 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) (35). ePCRV integration time was calculated based on two full-length ePCRVs. ePCRVs insertions were estimated to be much older than PERVs (~10.7 MYA and ~8.4 MYA).

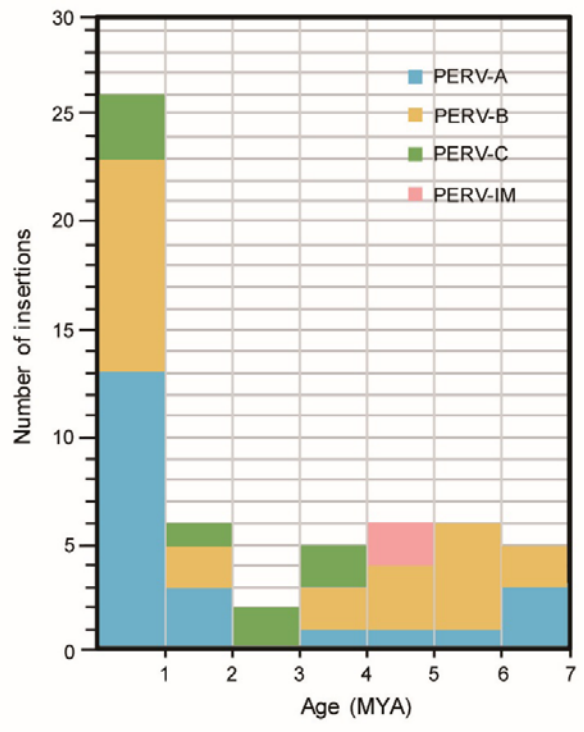
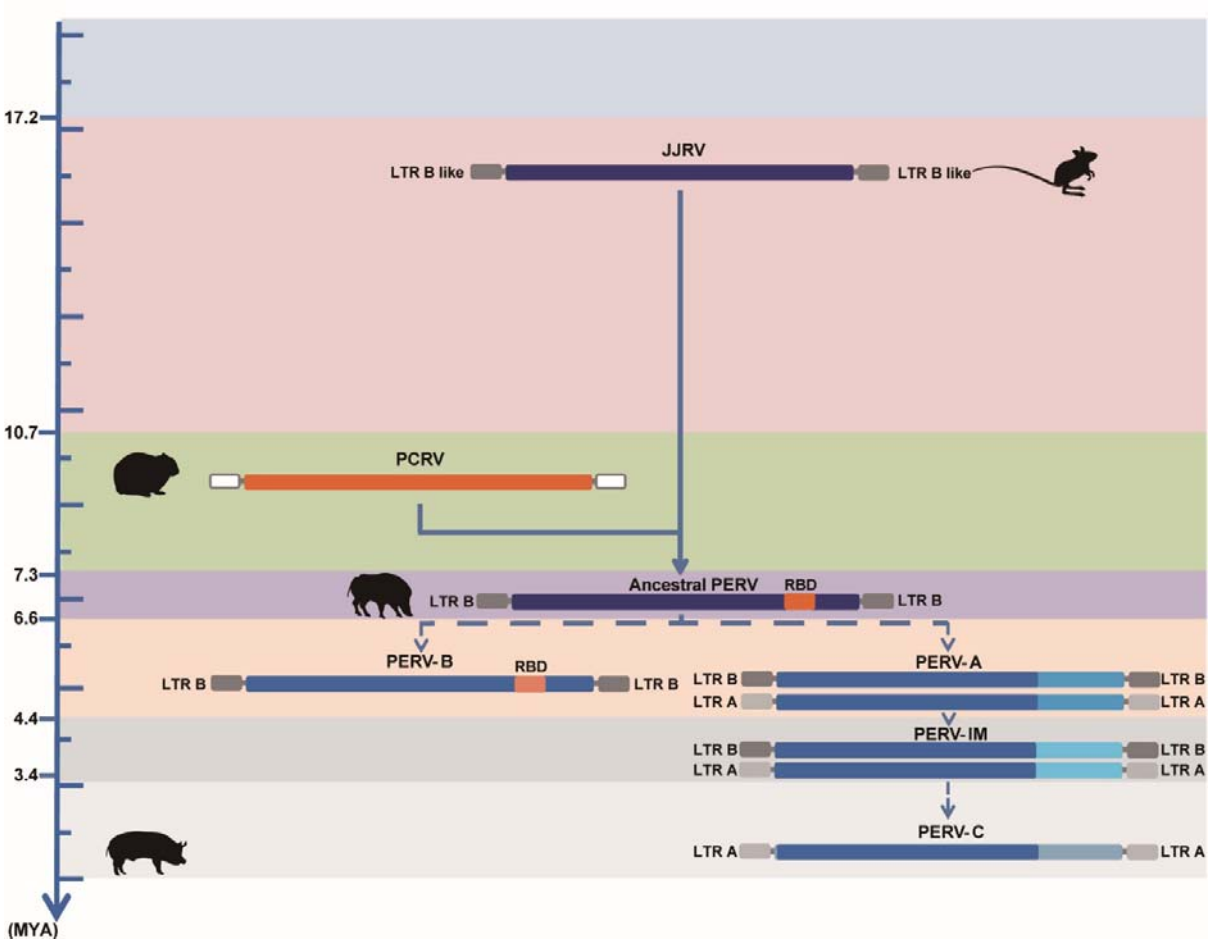


Fig. 4. Dating of PERVs insertion based on LTR-LTR divergence. The Y axis shows the number of insertions for different classes and X axis indicates the putative insertion time using MY as a unit. PERVs with LTR > 300 bp are used for estimation.

Evolutionary history of PERVs.

Taken together, evolutionary history of PERVs could be divided into four stages. First, JJRVs, the most closely related ERVs to PERVs, integrated into *Dipodidae* ~17.2 MYA. But the SU subunits of *env* were dissimilar between eJJRVs and PERVs, indicating that the subunit may be derived from other ERVs (Fig. 5). Then ePCRVs emerged ~10.7 MYA, and the SU subunit of *env*, especially RBD, was highly similar to that of PERVs, which suggested that PCRVs may also be a donor of the ancestral PERV. Third, the ancestral PERV emerged. The oldest modern PERV, PERV-A integrated into *Suidae* ~6.6 MYA just after the emergence of *Suidae* (~7.3 MYA). It is possible that the ancestral PERV

originated from the co-infection and recombination of JJRVs and PCRVs, and originally appeared around the late Miocene (~6.6 - 7.3 MYA) after the emergence of *Suidae*. Finally, after rapid adaptation in *Suidae*, PERV-A and -B diverged from the ancestral PERV ~6.6 MYA and ~6.4 MYA, respectively. The integration of PERV-IM occurred around ~4.4 MYA, between the integration of earliest PERV-A (~6.6 MYA) and PERV-C (~3.4 MYA). The homologies between PERV-A and -C are reported to be ~85%, while those between PERV-B and both PERV-A and PERV-C barely exceed 70% (17). Moreover, PERV-C harbors only one type of LTR, LTR A, which is also present in PERV-A, but not PERV-B. PERV-A and -B can infect cells from several species including humans while PERV-C infects only pig cells (7, 12, 36). It is possible that PERV-C descended from PERV-A, but lost the ability to infect other species in order to increase its adapt ability.



247

248 **Fig. 5. Evolution history of PERVs.** Time-calibrated arrow indicates putative insertion
249 time of each type of ERVs. Different background colors illustrate evolution periods of
250 PERVs. Solid arrows show recombination events and dotted arrows show speciation.
251 Evolutionary history of PERVs could be divided into four stages. Pink background (the
252 first stage) represents eJJRVs appearance (~17.2 MYA), before the emergence of ePCRVs
253 (~10.7 MYA). Green (the second stage) represents the emergence of ePCRVs (~10.7
254 MYA), just before the emergence of *Suidae* (~7.3 MYA). Purple (the third stage)
255 represents the period when eJJRVs and ePCRVs co-infected and recombined, which
256 resulted in the emergence of ancestral PERV, just after the emergence of *Suidae* (~7.3
257 MYA). Until ~6.6 MYA the modern PERV-A emerged. Orange, blue and light grey (the
258 fourth stage) represents the period when modern PERVs (PERV-A, -B, -C, and -IM)
259 emerged and evolved. Orange represents the emergence of PERV-A (~6.6 MYA) and -B
260 (~6.4 MYA), just before the emergence of PERV-IM (~4.4 MYA). Blue represents the
261 emergence of PERV-IM (~4.4 MYA), just before the emergence of PERV-C (~3.4 MYA).
262 Light grey represents the period when PERV-C emerged, which can be dated back to ~3.4
263 MYA.

264

265

266 **Discussion**

267 Using systematic large-scale genome mining, we analyzed the origin and evolution of
268 PERVs. eJJRV, the most closely related ERV to PERVs, can be traced back to ~17.2 MYA,

which is well before *J. jaculus* speciated (~11.1 MYA), but later than the speciation of *Dipodidae* (~42.7 MYA). Unexpectedly, homologous LTRs of PERVs (~73% identity) were also found in 8 *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. Also, previously study found ERV in 2 *Muroidea* species (*M. musculus*, *R. norvegicus*)(37). But phylogenetic analysis suggested that coding genes were distantly related to PERVs and eJJRVs (Fig. S7). The homologous LTRs in *Muroidea* and *Dipodoidea* (especially eJJRVs) indicated that PERV-related LTRs have integrated into rodents before the divergence of *Muroidea* and *Dipodoidea* from a common ancestor (~53.0 MYA). Then the LTR-related ERVs in *Muroidea* and *Dipodoidea* evolved separately. *Dipodoidea* became the most closely related ancestor of PERVs until eJJRVs emerged (~17.2 MYA).

PERVs (~6.6 MYA), JJRVs (~17.2 MYA), and PCRVs (~10.7 MYA) integrated into *Suidae*, *Dipodidae*, and *Procaviidae*, respectively. The fossil records of *Suidae Dipodidae*, and *Procaviidae* also support this speculation of evolution of PERVs. 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 identified in East Africa, thus suggesting that the *Dipodidae* may have spread to East Africa during the Miocene. Miocene Procaviidae fossils have

291 been found in South of Africa and East Africa
292 (http://fossilworks.org/?a=taxonInfo&taxon_no=43293). According to the current fossil
293 records, the only shared region for Miocene *Dipodidae*, *Procaviidae* and *Suidae* fossils is
294 East Africa. So co-infection and recombination may occurred between retroviruses carried
295 by *Dipodidae* and *Procaviidae* in East Africa during the Miocene, and then recombinants
296 may invaded the *Suidae*, producing ancestral PERV.
297
298 In summary, for the first time, we decipher a complex evolutionary history for the PERVs.
299 The ancestral PERV might derive from recombination and co-infection of JJRVs and
300 PCRVs from *Dipodidae* and *Procaviidae*. Then the ancestral PERV split into two classes
301 (PERV-A and PERV-B). Finally, PERV-C diverged from PERV-A. We also suggest that pig
302 genomes have been shaped by PERV invasions, as specifically reflected by PERV-
303 associated genomic rearrangement that have occurred during porcine evolution. In a word,
304 modern PERVs have a complex evolutionary history prior to their appearance in pigs.
305

Materials and Methods

In silico identification of PERV and PERV-related proviruses.

To identify PERV proviruses in *Sus scrofa*, tBLASTn (38) was used and amino acid sequences of Gag, Pol and Env of 20 representative PERV proviruses (accession number: 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 search the 14 pig genomes available in Genbank that were released before November 2017. A 50% identity over 50% 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, 18). 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 boundaries of PERV. LTR lengths were defined as 100–1,000 bp. PERVs with at least one LTR and one coding gene were screened for the next analysis.

To identify PERV-related proviruses in mammals, tBLASTn was used with the queries mentioned above in 20 representative PERV proviruses to search 142 mammal genomes available in Genbank that were released before November 2017. A 50% identity over 80% region was used to filter significant hits. LTRs were identified using LTR finder (39), LTRharvest (40) and BLASTn. LTR lengths were also defined as 100–1,000 bp.

Detection of recombination mediated by PERVs.

328 To search for proviruses involved in recombination and chromosomal rearrangements, we
329 constructed a neighbor-joining tree of 5'- and 3'- LTR of full-length PERVs using
330 MEGA7 (41) with Kimura 2-parameter distance estimates. LTRs less than 250 bp were
331 not considered. Alignment was carried out with MAFFT 7.222 (42).
332
333 **Phylogenetic analyses.**
334 To determine the evolutionary relationship among PERVs, eJJRVs and representative
335 gammaretroviruses (S5 Table), phylogenetic trees were inferred with amino acid
336 sequences. Full-length PERVs and PERVs with one LTR and at least one coding gene
337 were used to construct phylogenetic trees. All Gag, Pol and Env protein sequences were
338 aligned in MAFFT 7.222 and confirmed manually in MEGA7. The evolutionary history of
339 these gammaretroviruses was then determined using the maximum-likelihood (ML)
340 phylogenetic method available in PhyML 3.1 (43), incorporating 100 bootstrap replicates
341 to determine the robustness. The best-fit JTT+Γ amino acid substitution model was
342 selected for Gag, Pol and JTT+Γ+I for Env using the ProtTest 3.4.2 (44). All alignments
343 can be found in Dataset S1
344
345 **Molecular dating of PERV, eJJRV and ePCRV.**
346 The 5' and 3' LTRs of ERVs are identical at the point of integration, and then diverge and
347 evolve independently (45). So the ERV integration time can be calculated using the
348 following relation: T = (D/R)/2, in which T is the invasion time (million years, MY), D is
349 the number of nucleotide differences per site between the two LTRs, and R is the genomic

350 substitution rate (nucleotide substitutions per site, per year). We used the previously
351 estimated average mammal substitution rate (2.2×10^{-9} per site per year) (33), as no
352 substitution rate (r) has yet been estimated for the *S. Scrofa*, *J. jaculus* and *P. canpensis*.
353

354 ## References

355 1. **Ekser B, Cooper DKC, Tector AJ.** 2015. The need for xenotransplantation as a
356 source of organs and cells for clinical transplantation. Int J Surg **23:**199-204.
357 2. **Ekser B, Ezzelarab M, Hara H, van der Windt DJ, Wijkstrom M, Bottino R,**
358 **Trucco M, Cooper DK.** 2012. Clinical xenotransplantation: the next medical
359 revolution? Lancet **379:**672-683.
360 3. **Niu D, Wei HJ, Lin L, George H, Wang T, Lee IH, Zhao HY, Wang Y, Kan Y,**
361 **Shrock E, Lesha E, Wang G, Luo Y, Qing Y, Jiao D, Zhao H, Zhou X, Wang S,**
362 **Wei H, Guell M, Church GM, Yang L.** 2017. Inactivation of porcine endogenous
363 retrovirus in pigs using CRISPR-Cas9. Science **357:**1303-1307.
364 4. **Denner J.** 2017. Paving the Path toward Porcine Organs for Transplantation. N
365 Engl J Med **377:**1891-1893.
366 5. **Wegman-Points LJ, Teoh-Fitzgerald ML, Mao G, Zhu Y, Fath MA, Spitz DR,**
367 **Domann FE.** 2014. Retroviral-infection increases tumorigenic potential of MDA-
368 MB-231 breast carcinoma cells by expanding an aldehyde dehydrogenase
369 (ALDH1) positive stem-cell like population. Redox Biol **2:**847-854.
370 6. **Denner J, Young PR.** 2013. Koala retroviruses: characterization and impact on the
371 life of koalas. Retrovirology **10:**108.

7. **Denner J, Tonjes RR.** 2012. Infection barriers to successful xenotransplantation focusing on porcine endogenous retroviruses. Clin Microbiol Rev **25:**318-343.

8. **Tonjes RR, Niebert M.** 2003. Relative age of proviral porcine endogenous retrovirus sequences in Sus scrofa based on the molecular clock hypothesis. J Virol **77:**12363-12368.

9. **Morozov VA, Wynyard S, Matsumoto S, Abalovich A, Denner J, Elliott R.** 2017. No PERV transmission during a clinical trial of pig islet cell transplantation. Virus Res **227:**34-40.

10. **Crossan C, Mourad NI, Smith K, Gianello P, Scobie L.** 2018. Assessment of porcine endogenous retrovirus transmission across an alginate barrier used for the encapsulation of porcine islets. Xenotransplantation doi:10.1111/xen.12409**:**e12409.

11. **Wynyard S, Nathu D, Garkavenko O, Denner J, Elliott R.** 2014. Microbiological safety of the first clinical pig islet xenotransplantation trial in New Zealand. Xenotransplantation **21:**309-323.

12. **Moalic Y, Blanchard Y, Félix H, Jestin A.** 2006. Porcine Endogenous Retrovirus Integration Sites in the Human Genome: Features in Common with Those of Murine Leukemia Virus. J Virol **80:**10980-10988.

13. **Czauderna F, Fischer N, Boller K, Kurth R, Tonjes RR.** 2000. Establishment and characterization of molecular clones of porcine endogenous retroviruses replicating on human cells. J Virol **74:**4028-4038.

14. **Patience C, Takeuchi Y, Weiss RA.** 1997. Infection of human cells by an

endogenous retrovirus of pigs. Nat Med **3:**282-286.

15. **Li Z, Ping Y, Shengfu L, Hong B, Youping L, Yangzhi Z, Jingqiu C.** 2004. Phylogenetic relationship of porcine endogenous retrovirus (PERV) in Chinese pigs with some type C retroviruses. Virus Res **105:**167-173.

16. **Cui J, Tachedjian G, Tachedjian M, Holmes EC, Zhang S, Wang LF.** 2012. Identification of diverse groups of endogenous gammaretroviruses in mega- and microbats. J Gen Virol **93:**2037-2045.

17. **Niebert M, Tonjes RR.** 2005. Evolutionary spread and recombination of porcine endogenous retroviruses in the suiformes. J Virol **79:**649-654.

18. **Scheef G, Fischer N, Krach U, Tonjes RR.** 2001. 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 **75:**6933-6940.

19. **Wilson CA, Laeeq S, Ritzhaupt A, Colon-Moran W, Yoshimura FK.** 2003. Sequence analysis of porcine endogenous retrovirus long terminal repeats and identification of transcriptional regulatory regions. J Virol **77:**142-149.

20. **Huh JW, Cho BW, Kim DS, Ha HS, Noh YN, Yi JM, Lee WH, Kim HS.** 2007. Long terminal repeats of porcine endogenous retroviruses in Sus scrofa. Arch Virol **152:**2271-2276.

21. **Niebert M, Kurth R, Tonjes RR.** 2003. Retroviral safety: analyses of phylogeny, prevalence and polymorphisms of porcine endogenous retroviruses. Ann Transplant **8:**56-64.

22. **Mayer J, Blomberg J, Seal RL.** 2011. A revised nomenclature for transcribed human endogenous retroviral loci. Mob DNA **2:**7.

23. **Johnson WE, Coffin JM.** 1999. Constructing primate phylogenies from ancient retrovirus sequences. Proc Natl Acad Sci U S A **96:**10254-10260.

24. **Hughes JF, Coffin JM.** 2001. Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. Nat Genet **29:**487-489.

25. **Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, Ingersoll R, Sheppard HW, Ray SC.** 1999. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J Virol **73:**152-160.

26. **Zhuo X, Feschotte C.** 2015. Cross-Species Transmission and Differential Fate of an Endogenous Retrovirus in Three Mammal Lineages. PLoS Pathog **11:**e1005279.

27. **Martoglio B, Graf R, Dobberstein B.** 1997. Signal peptide fragments of preprolactin and HIV-1 p-gp160 interact with calmodulin. Embo j **16:**6636-6645.

28. **Corbet S, Muller-Trutwin MC, Versmisse P, Delarue S, Ayouba A, Lewis J, Brunak S, Martin P, Brun-Vezinet F, Simon F, Barre-Sinoussi F, Mauclere P.** 2000. 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 **74:**529-534.

29. **Argaw T, Wilson CA.** 2012. Detailed mapping of determinants within the porcine endogenous retrovirus envelope surface unit identifies critical residues for human

438 cell infection within the proline-rich region. J Virol **86:**9096-9104.

439 30. **Watanabe R, Miyazawa T, Matsuura Y.** 2005. Cell-binding properties of the
440 envelope proteins of porcine endogenous retroviruses. Microbes Infect **7:**658-665.

441 31. **Denner J.** 2008. Recombinant porcine endogenous retroviruses (PERV-A/C): a
442 new risk for xenotransplantation? Arch Virol **153:**1421-1426.

443 32. **Ericsson TA, Takeuchi Y, Templin C, Quinn G, Farhadian SF, Wood JC,**
444 **Oldmixon BA, Suling KM, Ishii JK, Kitagawa Y, Miyazawa T, Salomon DR,**
445 **Weiss RA, Patience C.** 2003. Identification of receptors for pig endogenous
446 retrovirus. Proc Natl Acad Sci U S A **100:**6759-6764.

447 33. **Kumar S, Subramanian S.** 2002. Mutation rates in mammalian genomes. Proc
448 Natl Acad Sci U S A **99:**803-808.

449 34. **Frantz LAF.** 2015. Speciation and domestication in Suiformes: a genomic
450 perspective. Wageningen University.

451 35. **Zhang Q, Xia L, Kimura Y, Shenbrot G, Zhang Z, Ge D, Yang Q.** 2013.
452 Tracing the Origin and Diversification of Dipodoidea (Order: Rodentia): Evidence
453 from Fossil Record and Molecular Phylogeny. Evolutionary Biology **40:**32-44.

454 36. **Denner J.** 2016. How Active Are Porcine Endogenous Retroviruses (PERVs)?
455 Viruses **8**.

456 37. **Hayward A, Cornwallis CK, Jern P.** 2015. Pan-vertebrate comparative genomics
457 unmasks retrovirus macroevolution. Proc Natl Acad Sci U S A **112:**464-469.

458 38. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local
459 alignment search tool. J Mol Biol **215:**403-410.

460 39. **Xu Z, Wang H.** 2007. LTR_FINDER: an efficient tool for the prediction of full-
461 length LTR retrotransposons. Nucleic Acids Res **35:**W265-268.
462 40. **Ellinghaus D, Kurtz S, Willhoeft U.** 2008. LTRharvest, an efficient and flexible
463 software for de novo detection of LTR retrotransposons. BMC Bioinformatics
464 doi:10.1186/1471-2105-9-18.
465 41. **Kumar S, Stecher G, Tamura K.** 2016. MEGA7: Molecular Evolutionary
466 Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol **33:**1870-1874.
467 42. **Katoh K, Standley DM.** 2013. MAFFT multiple sequence alignment software
468 version 7: improvements in performance and usability. Mol Biol Evol **30:**772-780.
469 43. **Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O.**
470 2010. New algorithms and methods to estimate maximum-likelihood phylogenies:
471 assessing the performance of PhyML 3.0. Syst Biol **59:**307-321.
472 44. **Abascal F, Zardoya R, Posada D.** 2005. ProtTest: selection of best-fit models of
473 protein evolution. Bioinformatics **21:**2104-2105.
474 45. **Dangel AW, Baker BJ, Mendoza AR, Yu CY.** 1995. Complement component C4
475 gene intron 9 as a phylogenetic marker for primates: long terminal repeats of the
476 endogenous retrovirus ERV-K(C4) are a molecular clock of evolution.
477 Immunogenetics **42:**41-52.

478
479