Relative Age of Proviral Porcine Endogenous Retrovirus Sequences in *Sus scrofa* Based on the Molecular Clock Hypothesis

Ralf R. Tönjes1* and Marcus Niebert2

Paul-Ehrlich-Institut, D-63225 Langen,1 and Robert Koch-Institut, D-13353 Berlin,2 Germany

Received 17 June 2003/Accepted 10 August 2003

Porcine endogenous retroviruses (PERV) are discussed as putative infectious agents in xenotransplantation. PERV classes A, B, and C harbor different envelope proteins. Two different types of long terminal repeat (LTR) structures exist, of which both are present only in PERV-A. One type of LTR contains a distinct repeat structure in U3, while the other is repeatless, conferring a lower level of transcriptional activity. Since the different LTR structures are distributed unequally among the proviruses and, apparently, PERV is the only virus harboring two different LTR structures, we were interested in determining which LTR is the ancestor. Replication-competent viruses can still be found today, suggesting an evolutionary recent origin. Our studies revealed that the age of PERV is at most 7.6 × 106 years, whereas the repeatless LTR type evolved approximately 3.4 × 106 years ago, being the phylogenetically younger structure. The age determined for PERV correlates with the time of separation between pigs (Suidae, *Sus scrofa*) and their closest relatives, American-born peccaries (Tayassuidae, *Pecari tajacu*), 7.4 × 106 years ago.

Xenotransplantation involves the transfer of cells, tissues, or organs from one species to another for therapeutic reasons. While the technology is under development, it has the potential to overcome the significant shortage of human allotransplants (31). Using the pig (*Sus scrofa*; order, Suidae) as the favored donor animal presents a number of important microbiological safety issues (40) which may be solved regarding exogenous agents by proper handling of the animals, such as barrier breeding according to specific-pathogen-free conditions and application of vaccinations (38). However, porcine endogenous retroviruses (PERV) pose a more serious challenge, since they are present in almost every individual (26) and are transmitted vertically.

Endogenous retroviruses (ERV) are present in the genomes of all vertebrate species analyzed (8), and their general organization corresponds to exogenous retroviruses (8, 35, 41). Whereby multiple copies of ERV are present in their host genomes, most are truncated or mutated, rendering them replication incompetent. Nevertheless, some of the viral integration sites still display transcriptional activity and even produce viral particles by complementation in trans. Only a minority of the distinct proviruses are functional as reported for pigs (1, 20, 25, 28, 29). While PERV belong to the gamma-retroviruses, other ERV also share homologies with beta- and delta-retroviruses but not with lentiviruses (12, 16, 28, 41).

For PERV, three different classes, designated PERV-A, PERV-B, and PERV-C (1, 22), exist, whereby PERV-A and -B are polytropic and productively infect human cells in vitro, thus posing a serious risk in xenotransplantation and xenogeneic cell therapies. Ecotropic PERV-C (1) does not replicate on human cells and is therefore not included in this study. There are only minor genetic differences between the classes, which are most prominent in the receptor binding domain of the env protein (Fig. 1C and D). In addition, there are two different types of long terminal repeats (LTR) (Fig. 1B) that significantly affect the replication properties of single viruses (33) by instrumentalizing a special set of transcription factors (32). Both PERV-A and PERV-B proviruses demonstrate LTRs that harbor repeats in U3. On the other hand, PERV-A and PERV-C were found to display repeatless LTRs (33). LTRs displaying repeat structures and possible multimerization mechanisms were first described for murine retroviruses (23, 43). These retroviruses, along with other gamma-type retroviruses, share a common homology of approximately 60% among each other and with PERV. Furthermore, so as not to complicate the analysis, we excluded any of these viruses from this study and concentrated on determining the relative age of the two different PERV LTR types not found in any other virus. It was shown that PERV in cell culture actively adapt their LTR repeat structure, if present, to match the optimal but not the maximum replication performance in a given host cell (33). Conversely, the recombinant incorporation of an artificially created repeat structure, i.e., 10 times 39 bp, into U3 of the 5' LTR of a molecular clone caused rapid cell death after transfection into susceptible cells (Scheef and Tönjes, unpublished data). While different env isoforms are present in a variety of retroviruses (7), only PERV is known to harbor two related but profoundly different LTR types in addition to three env classes.

In anthropological studies, human endogenous retroviruses (HERV) have been used successfully to designate the age of the proviruses in hominoids as well as the time points of species separation events and phylogenetic relationships of different hominoids (17, 18, 34). These studies had the benefit of comparing many different HERV from different hominoids, ranging from humans to apes to monkeys. To repeat such a study for PERV is complicated due to the present diversity of the *Artiodactyla* (27, 37). Therefore, the closest relatives to...
pigs, the American-born peccaries (Tayassuidae, Pecari tajacu; see below), were included in this study to analyze an archaeological fixed relationship between two species. Besides determining the age of PERV, we were particularly interested in investigating which of the two LTR types is the phylogenetic predecessor. We have analyzed the prevalence of six well-characterized full-length PERV, five of them being replication competent and four of them being chromosomally assigned (20, 25). These analyses revealed a heterogeneous distribution of PERV among individuals (26), and since no PERV is present in every pig, it seems feasible to generate pigs free of functional PERV by conventional breeding. In addition, specific proviruses show internal point mutations which significantly affect their replication capacities. Since there are two different types of PERV LTR structures (Fig. 1B) showing various levels of transcriptional capacity (33), an analysis of 21 distinct chromosomal locations revealed that PERV which harbor highly active LTR with repeat elements in U3 are dominant (26). In addition, the two polytropic envelope genes were assayed for sequence variations, displaying class-specific hot spots of variation, as well as variations in the R peptide region (Fig. 1C and D).

Sample acquisition. In order to study the phylogenetics of PERV, we have analyzed the LTRs and env genes of 17 mostly

FIG. 1. Genomic organization of proviral PERV. (A) PERV displays genes for group specific antigen (gag), protease/polymerase polypolyprotein (pro/pol), and envelope protein (env), flanked by LTR. Both LTR and env vary significantly between individual proviruses, with the LTR determining the transcriptional activity and the envelope protein determining the host range with two polytropic virus classes, designated A and B. The gag and pol genes show only minor variations between proviruses. (B) Two different PERV LTR structures exist, one type harboring different numbers of a distinct 39-bp repeat in U3 (I to III) which is composed of subrepeat I (18 bp) and subrepeat II (21 bp) (26, 33). The second LTR type has no repeat structure, although sequences homologous to the subrepeats (designated Ia and IIa) can be found scattered across U3. The numbers given for each LTR designate the presence of proviral PERV LTR found in 21 BAC clones of a genomic large white library (26, 30). (C and D) Entropy plots for PERV envelope genes (panel C, PERV-A; panel D, PERV-B). For calculating entropy, all env sequences (see Appendix) were aligned, and the number of nucleotide differences was determined for every position. This difference is plotted against the position, thus indicating regions of high sequence variation. The sequence variation at the 3′ end of both env genes is caused by two factors. Since sequences used for entropy analysis were taken unmodified from GenBank, some truncated sequences led to an overall increase in diversity. The more important fact, however, is a structural divergence of the R-peptide structure of various PERV. This divergence corresponds in large part, but not exclusively, with the LTR structure (see “Phylogenetic analysis of env” in the text for details). Broadly striped arrowhead, hot spot with very high sequence variation; finely striped arrow head, variations at the C terminus of Env; cap, cap site; SD, splice donor; PBS, primer binding site; SA, splice acceptor; ppt, polypurine tract; p(A), poly adenylation site.
intact and chromosomally localized proviruses (25, 30) by PCR and subsequent sequencing. These sequences were submitted to GenBank, and the accession numbers of those as well as of instrumental full-length proviral sequences are given in the Appendix. In addition, sequences of 27 envelope genes published in GenBank were analyzed. Their accession numbers are given in the Appendix as well.

Molecular clock analysis. In order to evaluate the clocklike behavior of PERV proviruses, distinct trees were constructed using the datasets of LTR and env sequences, proving that the molecular clock (5, 24, 45, 46) is the best approximation and could not be rejected assuming a constant rate of evolution (Table 1). When both data sets were tested for substitution saturation using the DAMBE software package (44), no saturation could be observed when transitions (s) and transversions (v) were plotted versus the evolutionary distance (Fig. 2). The plot shows a linear increase of s and v values based on increasing divergence between different PERV. Since s occurs much more often than v, s should increase faster than v. In case of substitution saturation, where multiple substitutions have occurred at each site, the phylogenetic signal is essentially lost, and its effect is detectable because v gradually outnumbers s. Thus, the established graphs (Fig. 2) indicate that no substitution saturation has occurred in the investigated data sets of PERV LTR (Fig. 2A) and env (Fig. 2B) sequences, suggesting the reliability of the chronological dating based on the molecular clock.

The gap in the s/v plot of the LTR data set (Fig. 2A, II) is due to the repeat structure present in U3 of the PERV LTR (Fig. 1B). The comparison of LTRs with a different repeat count showed a stepwise increase in sequence variations. When comparing viruses with different LTR types, each repeat-harboring LTR type increases genetic distance by 39 bp, while the repeatless LTRs deviate considerably from the repeat-harboring LTRs on the basis of pairwise comparison, thus producing the gap in genetic distance that is observed in the LTR s/v plot. The significant dominance of transitions and no crossing of the different symbols, representing transitions and transversions, suggest no substitution saturation and proof of a constant rate of evolution, therefore allowing a determination of age by molecular clock. Each × or △ in the plot represent a transition or transversion event at a given genetic distance, respectively. n, number of sequences analyzed for each plot (see Appendix).

![Fig. 2](http://jvi.asm.org/)

**TABLE 1. Maximum-likelihood test for the molecular clock hypothesis**

<table>
<thead>
<tr>
<th>Data set</th>
<th>Log likelihood</th>
<th>2(\text{In ML} - \text{In MLK})</th>
<th>χ² (0.95)</th>
<th>Statistical significance level (P)</th>
<th>Rejection of the molecular clock hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTR</td>
<td>−1,968.13</td>
<td>−2,727.04</td>
<td>1,517.81</td>
<td>23</td>
<td>0.001</td>
</tr>
<tr>
<td>env</td>
<td>−7,334.43</td>
<td>−10,017.67</td>
<td>5,366.47</td>
<td>35</td>
<td>0.001</td>
</tr>
</tbody>
</table>

*a Calculation of the statistical significance of rejecting the molecular clock hypothesis using the χ² test comparing maximum likelihood and maximum likelihood with enforced molecular clock. The hypothesis that PERV does evolve according to a molecular clock is rejected, if the significance p is larger than 0.05. ML, maximum likelihood method; MLK, likelihood method enforcing a molecular clock using the ML topology. The calculation was performed using the software package DAMBE (44) according to the instructions given in reference 39.*
The 5’ and 3’ LTRs of 21 chromosomally assigned proviral clones were amplified and sequenced, thus revealing nucleotide differences in the corresponding LTRs. With estimated mutation rates of approximately $2.3 \times 10^{-9}$ to $5.0 \times 10^{-9}$ substitutions per nucleotide per year, any substitution translates to 250,000 and 650,000 years of evolution, respectively. The mean of 450,000 years was used for calculating the age of PERV, multiplying this value with the average number of mutations per LTR type. The appearance of very low or very high mutation numbers in a single copy of an LTR giving rise to high standard deviation was corrected for the final estimation of age using the median. No occurrences of the respective LTR type in the 21 BAC clones assayed; $\delta$ no. of mutations, mean number of mutations found in all LTRs of a given type, given as mean number with standard deviation; approx. age, the estimation of age is given in years with the minimum (Min) and maximum (Max) times since the proviral integrations into the host genome have occurred. w/o, without.

number results in this steplike increase of genetic distance as insertions or deletions of 39 nucleotides happen at once. The separation of env sequences into two subpopulations according to their class (A or B) results in similar s/s plots, as observed for the combined sequence set (data not shown), suggesting similar rates of evolution for both env classes.

**Phylogenetic analysis of LTR sequences.** The clocklike evolution of PERV allowed an attempt of dating by sequence comparison of the 5’ and 3’ LTR of distinct proviruses. Since the number of fully sequenced PERV is not large enough to allow for phylogenetic analyses based on intact and defect proviral sequences, we amplified the LTRs of distinct chromosomally assigned PERV present in a genomic pig library constructed in bacterial artificial chromosomes (BAC) (30). Because every BAC statistically harbors only one provirus (checked by PCR; data not shown), we could amplify 5’ and 3’ LTRs of different PERV from this quasigenomic template without interference with LTRs of unrelated proviruses.

Analyses of primate ERV showed that these ERV accumulate mutations with a rate of approximately $2.3 \times 10^{-9}$ to $5.0 \times 10^{-9}$ substitutions per nucleotide per year (17). Assuming the same mutation rate and with a mean LTR length of 700 bp, this translates to 250,000 to 650,000 years per mutation for PERV, with a mean of 450,000 years used for the calculations in Table 2. The fact that transcriptionally active viruses exist (11, 19, 25) favors a relatively recent phylogenetic origin of PERV. While the results of this attempt of chronological dating are given in Table 2 and the actual estimation of time may be improved in the future, it is clear that the repeatless LTR (Fig. 1B) is phylogenetically younger than the repeat-harboring LTR (Fig. 1B).

A possible mechanism could be an insertional mutation in an LTR with 1.5 repeats, disrupting this motif, since highly homologous but spatially not organized subrepeat fragments can be found in the LTRs that are devoid of 39-bp repeats (Fig. 1B, subrepeats Ia and IIa). An insertional mutation only needs to happen in the 3’ LTR, since the replication mechanism of retroviruses will copy that mutation to the 5’ LTR as well. Unfortunately, the stretch of DNA inserted into the LTR is too small to allow tracing of its origin confidently in the databases, producing matches with many genes of different species. The repeatless LTRs show a significantly lower level of transcriptional activity because the binding site of the major transcription factor NF-Y is located directly at the junction of the 18-bp and 21-bp-repeat and, therefore, is missing in the repeatless LTRs (33). We assume that repeatless LTRs are an adaptation to the acquired endogenous replication cycle of PERV, since this LTR structure largely reduces the transcriptional activity (33), which is in line with observations that artificially elongated LTRs kill their host cells rather rapidly (see above) and that replication-competent viruses dynamically adjust their repeat numbers to host cells, favoring lower copy numbers of these repeats in U3 (33). Furthermore, adaptations reducing the replicative performance have been described for a number of virus families, such as endogenously replicating members like MLV or HERV (10). While PERV are discussed as major obstacles to xenotransplantation applications, since they productively infect human cells in vitro, some of our results indicate that viruses carrying repeatless LTRs, though capable of minor replication in vitro, are not able to cross the species barrier on their own. Proviruses isolated from PERV-infected human cell line 293-PERV-PK (29) all showed a repeat-harboring LTR, and no integrated repeatless LTR could be identified in genomic DNA of that cell line by PCR (11, 25).

**Phylogenetic analysis of env sequences.** Since the reliability of the molecular clock was proven for PERV LTR and env sequences (Table 1 and Fig. 2), an alternative method for phylogenetic dating is the comparison of coding sequences generating phylogenetic trees based on Env (Fig. 3). In general, the homologies between PERV-A and PERV-C are approximately 85%, while the similarities between PERV-B and both PERV-A and PERV-C barely exceed 70%. This and the occurrence of repeatless LTRs in both PERV-A and PERV-C but not in PERV-B (see above) lead to the assumption of a common evolutionary origin of these two classes, both seeming younger than PERV-B. Unfortunately, there are not enough PERV-C sequences available currently to conduct a thorough analysis. The exclusion of PERV-C sequences may present a weak point. There is the possibility that PERV-C recombined with PERV-A, which is evident by A/C env recombinants (13, 42). Therefore, a PERV-A population with mixed LTR sequences is conceivable. PERV-C obviously was not able to spread in the porcine genome (initial results of screening several porcine genomic libraries suggest a very limited number of PERV-C sequences in the pig genome; Preuβ and Tönjes, unpublished data); however, PERV-A with both types of LTR did. We admit that if more PERV-C sequences become available the study must be extended. While the observed hot spots of variation specific for each class most likely are due to different selective pressures (Fig. 1C and D), there is a principle sequence difference between both classes in the receptor binding region, while other regions display high levels of homology. When using such sequences for computational analysis, this may lead to intrinsic bias. Therefore, both subpopulations were assayed separately. However, the comparison of coding sequences, in contrast to the comparison of LTR sequences, lacks an internal reference and must be calibrated with an externally generated fixed time point, most preferably related to archaeological data. Since those data are limited for pigs, an
samples failed to identify any PERV-specific signals in any vertebrate species (28). The absence of PERV sequences in peccaries (28).

FIG. 3. Phylogenetic tree of PERV-A envelope protein sequences aligned for evolutionary time, calculated as in Table 2. Env sequences were shortened to amino acid position 654 to avoid any bias due to R-peptide variations or truncation (see the text and Fig. 1C and D for details). The analyzed PERV-A Env sequences cluster into three groups, with their branching points indicated by letters. The most divergent group (A) is the oldest one, the second group (B) being only slightly younger. This behavior correlates with LTR types with 1 1/2 and 2 1/2 repeats, respectively. The less-divergent group (C) is the youngest one, with all members belonging to proviruses with repeatless LTRs. The distance between cluster A and B is approximately one-fourth of the distance between cluster C and cluster A. When assuming a difference of approximately 0.9 million years between LTRs with 1 1/2 and 2 1/2 repeats (white box) (see Table 2), these two groups are approximately 7.4 million years old. This age correlates with the separation of pigs and their closest relatives, the peccaries, approximately 7.6 × 10^6 years ago, and the event that led to the generation of the repeatless type of LTR occurred between 400,000 and 3 million years ago (Table 2). Upon integration into the germ line, both LTR variants will have coexisted exo- and endogenously for some time. It is evident from observations of other endogenous retroviruses (especially HERV) that these ERV acquire mutations which reduce and finally abolish their replicative performance. The acquisition of a weakly performing promoter may help the virus to do no damage to its host cell while retaining some levels of replication competence. However, the generation of repeatless LTRs may reflect an adaptation process of the virus, switching from an exogenous to an endogenous life cycle. The repeatless LTR is present in the relatively closely related PERV-A and PERV-C proviruses, while the more distant PERV-B proviruses exclusively display repeat-harboring LTRs.

This study was supported by a grant from the German Ministry of Health (Bundesministerium für Gesundheit), Bonn, Germany.

**APPENDIX**

Specific PERV loci isolated from the BAC library (30) are grouped according to their respective PERV subgroup, while the accession numbers are given for 5′ LTR, 3′ LTR, and env for the respective clone. PERV-A: 242D4 (AY312539, AY312555, AY312523); 141G12 (AY312537, AY312553, AY312521); 305F5 (AY312542, AY312558, AY312526); 258A11 (AY312541, AY312557, AY312525); 135E5 (AY312536, AY312552, AY312520); 253B6 (AY312540, AY312556, AY312524); 383E10 (AY312543, AY312559, AY312527); and 1079D9 (AY312535, AY312551, AY312519). PERV-B: 161B7 (AY312538, AY312554, AY312522); 484G4 (AY312544, AY312560, AY312528); 498D8 (AY312545, AY312561, AY312529); 534G4 (AY312546, AY312562, AY312530); 647G4 (AY312547, AY312563, AY312531); 667G4 (AY312548, AY312564, AY312532); 783D7 (AY312549, AY312562, AY312518); PERV-C: none.

In addition, we have used full-length sequences of previously described proviruses isolated in our laboratory as well as sequences also did not detect any PERV sequences in peccaries (28). The absence of PERV sequences in peccaries indicates that the age of PERV cannot exceed the calculated 7.4 × 10^6 to 7.6 × 10^6 years. According to calculations based on the LTR data, the repeatless LTR evolved between 400,000 and 3.1 × 10^6 years ago (Table 2). We assume that the appearance of that LTR which shows a low transcriptional activity enabled the virus to pursue an endogenous replication cycle more easily, since the low replication level does not damage the host cell while the virus still retains its ability to replicate outside its host.

Due to missing external references, protein sequences could be employed only for dating the relative age, thereby reinforcing the LTR analysis and indicating that the repeatless LTRs are in fact phylogenetically younger (data not shown).

**Conclusion.** A number of authors have pointed out that molecular clock calibrations are subject to a wide margin of error (3, 4, 15). Therefore, the calculations given in Table 2 are only rough estimates of absolute time, but they are nonetheless useful for comparing the relative ages of different PERV. In summary, we have shown that the repeatless PERV LTR evolved from the repeat-harboring LTR, most likely by insertional mutation. This can be proven by employing two different sets of data (LTR and env/Env). PERV is calculated to be approximately 7.4 million years old. This age correlates with the separation of pigs and their closest relatives, the peccaries, approximately 7.6 × 10^6 years ago, and the event that led to the generation of the repeatless type of LTR occurred between 400,000 and 3 million years ago (Table 2). Upon integration into the germ line, both LTR variants will have coexisted exo- and endogenously for some time. It is evident from observations of other endogenous retroviruses (especially HERV) that these ERV acquire mutations which reduce and finally abolish their replicative performance. The acquisition of a weakly performing promoter may help the virus to do no damage to its host cell while retaining some levels of replication competence. However, the generation of repeatless LTRs may reflect an adaptation process of the virus, switching from an exogenous to an endogenous life cycle. The repeatless LTR is present in the relatively closely related PERV-A and PERV-C proviruses, while the more distant PERV-B proviruses exclusively display repeat-harboring LTRs.
available in the international databases. PERV-A: YA099323 (6); AJ285885, AJ285884 (9); AJ133817 (11); AJ293656 (20); AF24067, AF240629, AF240624, AF240625, AF240627, AF240628, AF240642, (21); AJ279056, AJ279056, AF358696 (25); and AF296168 (Chang et al., unpublished data). PERV-B: YA099324 (6); AJ285859, AJ285892, AJ285859, AJ285859, AJ285859, AJ133816, AJ133818 (11); AJ293657 (20); AF296916, AF296940, AF296937, AF296936, AF296935, AF296933, AF296946, (21); Y122293 (22); and AJ293656 (25). PERV-C: AF038600 (1).

REFERENCES