PLOS ONE

RESEARCH ARTICLE

# Computational Evaluation of the Strict Master and Random Template Models of Endogenous Retrovirus Evolution

**Fabrícia F. Nascimento**[¤a]*, **Allen G. Rodrigo**[¤b]

National Evolutionary Synthesis Center, Durham, NC, United States of America

¤a Current address: Department of Zoology, University of Oxford, Oxford, United Kingdom
¤b Current address: Research School of Biology, The Australian National University, Acton, Australia
* fabricia.nascimento@zoo.ox.ac.uk

CrossMark
click for updates

OPEN ACCESS

**Citation:** Nascimento FF, Rodrigo AG (2016) Computational Evaluation of the Strict Master and Random Template Models of Endogenous Retrovirus Evolution. PLoS ONE 11(9): e0162454. doi:10.1371/journal.pone.0162454

**Editor:** Jean-Luc EPH Darlix, "INSERM", FRANCE

**Received:** December 7, 2015

**Accepted:** August 2, 2016

**Published:** September 20, 2016

**Copyright:** © 2016 Nascimento, Rodrigo. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** A Python code to simulate the true phylogenetic trees for the four ERV models described in this paper is available at https://github.com/thednainus/ERV_Simulations. A pipeline in R to calculate the same statistics for empirical phylogenetic trees is available at https://github.com/thednainus/R_Pipeline. The kNN classifiers trained with reconstructed phylogenetic trees and DNA sequences alignments of 1,000 bp and 10,000 bp can also be downloaded for future predictions of the proposed models described in this paper. Sequence alignment for porcine endogenous retrovirus used in this study can be downloaded at https://github.com/thednainus/R_Pipeline/tree/master/alignments.

## Abstract

Transposable elements (TEs) are DNA sequences that are able to replicate and move within and between host genomes. Their mechanism of replication is also shared with endogenous retroviruses (ERVs), which are also a type of TE that represent an ancient retroviral infection within animal genomes. Two models have been proposed to explain TE proliferation in host genomes: the strict master model (SMM), and the random template (or transposon) model (TM). In SMM only a single copy of a given TE lineage is able to replicate, and all other genomic copies of TEs are derived from that master copy. In TM, any element of a given family is able to replicate in the host genome. In this paper, we simulated ERV phylogenetic trees under variations of SMM and TM. To test whether current phylogenetic programs can recover the simulated ERV phylogenies, DNA sequence alignments were simulated and maximum likelihood trees were reconstructed and compared to the simulated phylogenies. Results indicate that visual inspection of phylogenetic trees alone can be misleading. However, if a set of statistical summaries is calculated, we are able to distinguish between models with high accuracy by using a data mining algorithm that we introduce here. We also demonstrate the use of our data mining algorithm with empirical data for the porcine endogenous retrovirus (PERV), an ERV that is able to replicate in human and pig cells *in vitro*.

## Introduction

Transposable elements (TEs) are DNA sequences able to move and replicate within, and occasionally between, host genomes [1]. These elements are present in almost all species including prokaryotes, and they are believed to constitute more than half of the human genome [1–5]. TEs are classified in two main groups. Class I elements, or retrotransposons, replicate through an RNA intermediate and insert a DNA copy into a new locus in the host genome, moving by a "copy and paste" mechanism [6]. In contrast, class II elements or DNA transposons replicate mainly by a "cut and paste" mechanism, excising themselves from one locus and reinserting in

**Funding:** This work was supported by the National Evolutionary Synthesis Center—National Science Foundation grant number EF-0905606 (http://www.nsf.gov/) to FFN and AR, and The Royal Society and British Academy—Newton International Fellowship NF 140338 to FFN. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

a different location in the host genome [6]. Given the very different mechanism of replication between the two TE classes, this study focuses only on the class I elements or retrotransposons.

The retrotransposon mechanism of replication is also shared with retroviruses [6]. Retroviruses have a dimer positive sense and single stranded RNA genome [7, 8], which is organized in four main coding domains: *gag* (encoding capsid, matrix and nucleocapsid proteins), *pro* (protease), *pol* (reverse transcriptase and integrase enzymes), and *env* (surface and transmembrane glycoproteins of the virus envelope) genes [8, 9] (Fig 1). While simple retroviruses are composed of these main genes, complex retroviruses have additional accessory genes [9]. Once a retrovirus integrates in the host genome it is referred to as a provirus. Proviruses have two long terminal repeat (LTR) sequences flanking their genome and containing regulatory elements [9, 10] (Fig 1). These LTR sequences are identical at the moment of integration, and often these two sequences recombine forming a solo LTR [7, 11].

Although retroviruses usually infect somatic cells, they can also infect and colonize germ cells [9]. Proviruses in germ cells are known as endogenous retroviruses (ERVs) [12–14]. ERVs are viewed as an ancient retroviral infection in animal genomes and are commonly referred to as viral "fossils" [8, 15, 16]. They are present in multiple copies; are passed to the offspring; and account for approximately 8% of the human genome [9, 17, 18]. Although the majority of ERVs observed in host genomes are classified as simple retroviruses, there are reports showing the integration of complex retroviruses in germ line cells [19, 20]. Because proviruses have the potential to disrupt host gene expression, they are negatively selected and typically lose their viral function and ability to reinfect [9]. However, some ERVs are still able to reinfect because they have intact viral genes; examples of ERVs that can reinfect are the koala retrovirus (KoRV) [21], the porcine endogenous retrovirus (PERV) [22], and the cervid endogenous retrovirus (CrERV) [23]. Even though ERVs are usually negatively selected, it is clear that ERV and other TEs play a role in shaping host genomes [14, 24, 25]. The most striking example is the role of an ERV gene in the formation of the placenta in mammals [26, 27]. Recently, Chuong *et al.* [28] have also shown the importance of human ERV (HERV) sequences for the innate immune response.

LTR sequences contain binding sites for cellular transcription factors that aim to promote the transcription of the provirus [7, 29]. An ERV is initially transcribed by the host polymerase, but it will increase in copy number following either reinfection or retrotransposition [29–31]. Reinfection involves the release of a virus that will reinfect another cell, a process that requires intact copies of all viral genes [30, 32]. Retrotransposition is the proliferation of a virus without the requirement of reinfecting another cell, and can occur either in *cis* or as complementation in *trans* [30]. Retrotransposition in *cis* requires functional *gag* and *pol* genes, while for complementation in *trans* no functional genes are required [31, 32]. In the latter case, the ERV needs to have an intact LTR for initial viral transcription to occur, with the other proteins necessary for viral replication provided by other viruses or TEs [30, 31]. In this case, a genome of a defective ERV will be integrated to the host genome if it successfully packages a reverse transcriptase and integrase enzymes [30]. An example of such process is the HERV-W that has used proteins from long interspaced elements (LINEs) to retrotranspose [33].

Two models of class I retrotransposition have been described: the strict master model (SMM) and the random template (or transposon) model (TM). In SMM, it is assumed that only one element of a given lineage in the genome–the "master"–is capable of producing a new copy, while in TM, it is assumed that all elements of a given lineage in the genome are equally able to produce new copies [34, 35]. Clough *et al.* [35] described the expected phylogenetic tree topology of retrotransposons under these two models, but did not investigate whether current phylogenetic methods would recover the expected tree topologies.

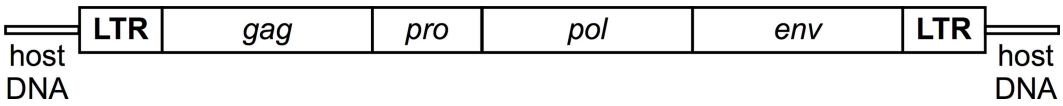


**Fig 1. A schematic illustration of a provirus genome.** The four main genes are depicted: *gag*, *pro*, *pol* and *env* genes. Proviruses are flaked by long terminal repeats (LTRs).

doi:10.1371/journal.pone.0162454.g001

Because of differences in genetic diversity, size, internal structure, and impact in host disease, it is important to understand the evolutionary dynamics of retrotransposons. Finding *in silico* models and evolutionary analyses that best explain their dynamics will advance our understanding of retrotransposons and their ability to retrotranspose in host genomes. In this paper, we focused on exploring these aspects by simulating SMMs and TMs on a type of class I elements or retrotransposons, the ERV. Our work is an extension of the work proposed by Clough *at al*. [35], but we include in our models ERV inactivation and ongoing activity related to their ability to retrotranspose or reinfect host cell genomes. Based on this information of ERV inactivation and ongoing activity, variations of SMM and TM were accessed in this study and named "SMM Mortal (SMM-m)", "TM Mortal (TM-m)", "SMM Immortal (SMM-i)" and "TM Immortal (TM-i)". These four extreme models were chosen to investigate whether a maximum likelihood approach would recover the expected tree topologies under the SMM and TM as described by Clough *et al*. [35].

Our results show that one is more likely to recover trees similar to the expected phylogenetic trees when phylogenies were reconstructed using alignments of 10,000 base pairs (bp) rather than 1,000 bp. In general, it was also more likely to recover the expected topologies when the rate of ERV replication per host generation was low. By increasing the rate of ERV replication per host generation, it also became more difficult to distinguish tree topologies under SMM and TM. Nonetheless, when appropriate statistics were calculated for phylogenetic trees, we were able to correctly identify 84% and 93% of the different models when trees were reconstructed with alignments of 1,000 bp and 10,000 bp, respectively. Our statistical approach was also able to recover the expected replication patterns for porcine endogenous retroviruses (PERVs).

Our study showed the importance of thoroughly analyzing extreme models of ERV dynamics and evolution before more complex models could be proposed. For example, a more complex model could involve only a proportion of elements able to replicate in a host genome. If we were unable to correctly identify the models herein proposed, it would be unlikely to do so by using more complex models of ERV dynamics.

## Materials and Methods

### The models

We assume that a single exogenous retrovirus colonizes a host germ cell genome and simulations start from this single copy. This was considered the initial time in all simulations, and it was also the time this retrovirus is endogenized. One simulation run represented the evolution of a single ERV lineage.

We have applied to ERVs both SMMs and TMs [34, 35] generally described for TE replication in host genomes. SMMs assumes that only one element of a given lineage–the "master"–is able of producing a new copy, while TMs assumes that all elements of a given lineage are equally able to produce new copies in the host genome. Based on SMMs and TMs four models were assessed in this study and named "SMM Mortal (SMM-m)", "TM Mortal (TM-m)", "SMM Immortal (SMM-i)" and "TM Immortal (TM-i)".

For the Mortal models, it was assumed that replication of an ERV lineage stopped after a fixed number of copies in the genome were attained. This follows the biological assumption that full-length ERVs are unable to reinfect when, for example, mutations cause gene inactivation of all ERV genes in all copies [9, 11]. A fixed number of elements was used based on information of copy number of ERV lineages in different host genomes [36]. In contrast, for the Immortal models, it was assumed that (*i*) an ERV lineage was able to replicate indefinitely and, (*ii*) the newly generated copy could occupy the locus of a previous copy by replacement. This follows the biological observation of full-length ERVs, such as the porcine endogenous retrovirus (PERV) [22] and the koala retrovirus (KoRV) [21], which still have the ability to reinfect or retrotranspose.

In our models, we do not distinguish between retrotransposition and reinfection. We assume that a newly generated ERV will be successfully reinserted in the host genome and become fixed, unless it is replaced as for the Immortal models.

## Computer Simulations

**1. Simulation of true phylogenetic trees.** Computer simulations were carried out using two variables. The first variable was the ERV mutation rate ($\mu_{erv}$) set to $1.0\times10^{-4}$, $1.0\times10^{-5}$, $3.0\times10^{-5}$, $1.0\times10^{-6}$, $1.0\times10^{-7}$, and $1.2\times10^{-8}$ substitutions per nucleotide per infection (*s/n/i*). A mutation rate of $3.0\times10^{-5}$ *s/n/i* has been estimated for Murine Leukemia Virus (MLV) [37], a virus that can be found in both exogenous and endogenous form [38]. We used other four ERV mutation rates to test the influence of this parameter on phylogenetic tree branch lengths. The second variable was the rate of ERV retrotransposition or reinfection per host generation that in this paper will be solely referred to as ERV replication (λ). No accurate information is available for ERV replication in host genomes. For this reason, arbitrary values of ERV replication were used in our simulations and were set to $1.0\times10^{-4}$, $2.0\times10^{-4}$, $3.0\times10^{-4}$, $4.0\times10^{-4}$, $5.0\times10^{-4}$, $6.0\times10^{-4}$, $7.0\times10^{-4}$, $8.0\times10^{-4}$, and $9.0\times10^{-4}$ retrotranspositions or reinfections per host generation (*r/g*). Host substitution rate ($\mu_h$) was fixed at $1.2\times10^{-8}$ substitutions per nucleotide per host generation (*s/n/g*), which is the described substitution rate for humans [39].

Because two mutation rates were used to simulate phylogenetic trees, branch lengths in substitutions per site represented a composite rate between host and ERV mutation rates. At least two different mutation rates are associated with the evolution of an ERV lineage [15, 40]. First, a new ERV copy will be the consequence of retrotransposition or reinfection because both mechanisms involve the reverse transcription of a viral RNA intermediate by the ERV encoded reverse transcriptase enzyme. Second, an ERV lineage can replicate because of host DNA replications by its DNA polymerase. Retroviral reverse transcriptase has a higher substitution rate than the host DNA polymerase [37, 40, 41].

The waiting time for an ERV to release a new copy in the host genome was simulated as an exponential random variable, with rate λ for SMMs and Nλ for TMs, where λ is the rate of ERV replication per host generation; and *N* is the number of ERV copies already generated. The ERV chosen to release a new copy in the host genome will release this new copy under the viral mutation rate ($\mu_{erv}$), while all other copies that remained in the genome will accumulate mutations according to the host substitution rate ($\mu_h$).

Different ERV lineages show different copy number, e.g. [36]. Because we are simulating simpler models to understand the dynamics of ERVs, the maximum number of ERV copies in the phylogenetic tree ($n_{max}$) was set to 50. This number was chosen based on the copy number described for porcine endogenous retroviruses (PERVs) [42] and some human endogenous retroviruses (HERVs) [36]. Finally, 100 simulations were carried out for each combination of ERV mutation rate and ERV replication for a total of 1,000,000 host generations. We refer to simulated phylogenetic trees as true trees.

**2. Algorithm to simulate true phylogenetic trees for Mortal models.** (1) A waiting time ($t$) is generated, which represents the time in host generations before an ERV releases a new copy in the host genome.

$$t = \frac{-ln(U)}{N\lambda}$$

Where $U$ is uniform random number, $N$ is the number of elements that is able to generate a new copy in the host genome (for SMM, $N$ is always 1), and λ is the ERV replication rate per host generation.

(2) If the sum of waiting times is less than the maximum number of host generations (1,000,000), a new element will be added to the phylogenetic tree.

(3) If the total number of elements is less than the maximum number of elements ($n_{\max}$), branch lengths ($l$) will be calculated as follows:

$$l = \mu_h \times t$$

$$l_{nc} = (\mu_h \times t) + \mu_{erv}$$

where $l$ and $l_{nc}$ are the branch lengths of genomic ERVs and the new ERV, respectively.

(4) If the total number of elements is equal to the maximum number of elements, all copies will accumulate mutations according to the host mutation rate, and the final branch length ($l_f$) will be calculated as follow:

$$T_f = T - \sum t$$

$$l_f = \mu_h \times T_f$$

where $T$ is the maximum number of host generations, and $T_f$ is the time at which the phylogenetic tree was composed by 50 elements (the maximum number of elements).

(5) If number of elements in the tree is less than the maximum number of elements, and the sum of $t$ is less than the maximum number of host generations, then return to Step 1, otherwise STOP.

**3. Algorithm to simulate true phylogenetic trees for Immortal models.** Steps (1), (2), and (3) are the same as described for the Mortal models.

(4) For Immortal models, replacement of elements is allowed. Replacement represents a homologous recombination between two proviruses [11]. The probability a replacement ($R$) will occur was calculated as:

$$P(R) = \frac{n-1}{n_{max}-1}$$

Where $n$ represents the current number of elements (or number of tips) in the phylogenetic tree. The randomly chosen ERV that will give birth to a new element was not allowed to be replaced. For this reason, we subtract 1 element from the equation above. This probability was chosen following the biological assumption that as the number of elements in the tree increases, the probability of replacement also increases.

(5) Repeat Steps 3 and 4 until the sum of $t$ reaches the maximum number of host generations.

**4. Simulation of DNA sequence alignments and phylogenetic reconstructions.** Seq-Gen 1.3.3 [43] was used to simulate DNA sequence alignments of 1,000, 10,000, and 100,000 bp under the Jukes and Cantor (JC) substitution model [44] for each true tree generated under the

four ERV models. The approximate size of an ERV genome is 10,000 bp [9]. However, it is common to reconstruct ERV phylogenies using partial genes of approximately 1,000 bp [45–48]. Simulations using 100,000 bp alignments were carried out to understand the effect of sampling errors in reconstructing ERV phylogenies.

Finally, using the simulated DNA sequence alignments, phylogenetic trees were reconstructed by maximum-likelihood (ML) with RAxML 8.0.19 [49] and setting the nucleotide substitution model to general time reversible (GTR) [50–52] (estimated values). Trees were rooted with a midpoint root using a script in R 3.0.3 [53] and the package *phangorn* [54]. Reconstructed phylogenetic trees will be referred to as ML trees.

## Statistical analysis

To compare true with ML trees reconstructed with alignments of 1,000, 10,000 and 100,000 bp, we used the Robinson-Foulds (RF) metric [55, 56] in the R package *phangorn*. Because RF metric is a partition metric, its range is 0 for identical trees with a maximum value of $2n - 6$, where $n$ is the total number of tips (or number of elements) in the tree [55]. The RF metric was calculated for rooted and unrooted trees to study the effect of midpoint rooting in ML trees. Comparison using ML trees reconstructed with alignments of 100,000 bp were carried out to understand the effects of sampling error in reconstructing the evolution of ERVs following the four proposed models.

Because the ERV genome size is approximately 10,000 bp, and because we would like to understand whether it is possible to distinguish phylogenies under different ERV models, tree statistics were calculated only for ML trees reconstructed using alignments of 1,000 and 10,000 bp. ML trees reconstructed using alignments of 100,000 bp were used solely to understand the effect of sampling error in reconstructing ERV phylogenetic trees.

The following 10 statistics were calculated as candidate variables for model classification, allowing us to test the best combination of statistics that is able to predict the correct ERV model proposed in this study:

(*i*) The tree shape statistic beta-splitting model (Beta) [57] was calculated using the R package *apTreeshape* [58]. Beta values equal to −2 represent completely unbalanced trees, which is expected for phylogenetic trees simulated under SMMs [35]. Increasing values of Beta correspond to greater tree balance [57], which is expected for phylogenetic trees simulated under TMs [35].

The other following seven tree shape statistics (*ii* to *viii*) were calculated using the R package *phyloTop* [59] following Colijn and Gardy [60], in which definitions are summarized below. For further information on statistics *ii* to *vii*, please see Colijn and Gardy [60].

(*ii*) Ladder length is defined by the maximum number of connected internal branches with a single terminal descendant branch (Max. ladder);

(*iii*) "IL" branches are defined as the portion of internal branches with a single terminal branch as descendant ("IL" portion);

(*iv*) Maximum depth and (*v*) maximum width: The depth of a branch is defined as the number of branches between that branch and the tree's root, while the tree width at depth $d$ is defined as the number of branches with depth $d$;

(*vi*) Maximum width over maximum depth: The ratio between maximum width and maximum depth;

(*vii*) Maximum difference in widths is defined as the maximum absolute difference in widths from one depth to the next, over all depths in the tree;

(*viii*) Number of cherries was also calculated; a cherry is defined as a pair of terminal branches that are adjacent to a common ancestor node [61].

To account for tree size, values obtained for summaries *ii*, *iii*, *iv*, *v*, *vii* and *viii* were divided by the number of terminal branches (or number of elements) in the tree.

In addition to tree shape statistics, (*ix*) the proportion of terminal branch lengths that contributed to the total tree branch length ("prop") was calculated using an R script and the *ape* package [62]. A higher proportion is expected for the Mortal models. Finally, (*x*) nucleotide diversity for simulated DNA sequences was calculated for alignments of 1,000 bp and 10,000 bp also using an R script and the *pegas* package [63].

## Comparison of true and ML trees and classification of ERV models

To compare the distribution of each of the 10 statistics calculated in the previous section for SMM and TM, we used the Jensen-Shannon divergence (JSD). The JSD is a symmetric divergence statistic that can be used to measure similarities between two distributions [64]. Comparisons were performed in pairs for SMM-m and TM-m as well as SMM-i and TM-i: JSD was calculated for true trees as well as for ML trees for each statistics described in the previous section. If two distributions are identical JSD = 0, and larger values of JSD represents dissimilar distributions. In the context of this study, if JSD = 0 there is no difference between trees under SMM and TM.

Because of the different tree topologies expected for trees generated under the SMM and TM [35], we would expect larger values for JSD calculated for the true trees under SMM and TMs for each of the tree shape statistics. With finite sequence data, we expect that errors in phylogenetic reconstruction will introduce variation in the differences of JSD statistics.

Because we would like to distinguish between trees reconstructed under the SMM and TM, we calculated JSD for each statistics for all combination of models in pairs (for example, TM-m *vs* TM-i, SMM-m *vs* TM-i, *etc*) using ML trees. We chose the statistics in which the JSD was larger and different from zero. This was used as a pre-screening of which variables should be included in a *k*-nearest neighbor (*k*NN) classifier. We also trained a *k*NN classifier using only the Beta statistics, which is a metric to detect phylogenetic tree imbalance.

A *k*NN was trained using R and the function IBk of package RWeka [65]. We let the function automatically find the best number of nearest neighbor value *k* between 1 and 30. We also tested *k* varying between 1 and 100. This training was performed on values of tree statistics and nucleotide diversity calculated for 21,600 ML trees / DNA sequence alignments of 1,000 and 21,600 ML trees / DNA sequence alignments of 10,000 bp. Results are reported using a 10-fold cross-validation and the *k*NN classifier trained with alignments of 1,000 bp were cross-validated using only 1,000 bp alignments. Similarly, a *k*NN classifier trained with alignments of 10,000 bp was cross-validated using only 10,000 bp alignments.

There is a lack of information regarding ERV model of replication (SMM or TM), ERV replication rate per host generation and ERV mutation rate. For this reason, and because few ERVs are know to be able to replicate in host genomes, we decided to train the *k*NN to classify between the four proposed models rather than trying to improve performance by classifying between SMM-m/TM-m and between SMM-i/TM-i, for example.

## Empirical data

We used datasets of PERV DNA sequences from two different lineages of PERVs, the gamma1 and gamma2 PERVs. Because of polymorphism in their *env* gene, gamma 1 PERVs is further divided into A, B and C classes [22, 66, 67], while gamma2 PERVs comprises only PERV class E [68, 69]. While gamma1 has the ability to replicate [22, 70], this does not seem to be the case for gamma2 PERVs [47, 68, 69]. To demonstrate how the framework developed in this paper could be used for empirical data, we analyzed 46 sequences comprising genomic data for *Sus*

*scrofa* gamma1 PERVs (classes A, B and C), with an alignment of 9,017 bp (including gaps). Because some alignment differences were observed between PERV-A, -B and -C, we also analyzed a subset of 24 sequences comprising only PERV class B. PERV-B alignment comprised 8,762 bp (including gaps). From those 46 genomic data, we also analyzed 1,000 bp of the *pol* gene. Finally, we analyzed 999 bp of 50 sequences of *env* type E for gamma 2 PERVs in *Sus* species.

Sequences were obtained from GenBank (for accession numbers see alignments at GitHub. Information is available in Code and Data availability section), and to increase sample size, PERV genomic sequences were also mined from the *Sus scrofa* genome (version Sscrofa 10.2) using blastn. All sequences were aligned using Muscle [71] with default options implemented in the program seaview [72]. Alignments were manually curated according to Yang [73].

Phylogenetic trees were reconstructed using the same methodology as described for simulated DNA sequences and using the GTR + Γ [50–52] as the DNA substitution model. The same statistics for tree shape, nucleotide diversity and proportion of terminal branch lengths that contributed to the total tree branch length were calculated using the same approach described for simulated data. We used the *k*NN algorithm trained with ML trees reconstructed with 1,000 bp and 10,000 bp to make predictions using partial gene and genomic sequence data, respectively.

## Code and Data availability

Algorithms to simulate the four ERV models described in this paper were written in Python and used the Python package ETE2 [74] to simulate rooted phylogenetic trees with branch lengths. This Python code is available at https://github.com/thednainus/ERV_Simulations

A pipeline in R to calculate the same statistics for empirical phylogenetic trees is available at https://github.com/thednainus/R_Pipeline. The *k*NN classifiers trained with reconstructed phylogenetic trees and DNA sequences alignments of 1,000 bp and 10,000 bp can also be downloaded for future predictions of the proposed models described in this paper.

Sequence alignments for porcine endogenous retrovirus used in this study can be downloaded at https://github.com/thednainus/R_Pipeline/tree/master/alignments. Information on GenBank accession numbers can also be found in these alignments.

# Results

A total of 21,600 trees were simulated following the different ERV models proposed in this study (see Materials and Methods). The maximum number of elements or tips ($n_{max}$ = 50) per phylogenetic tree was achieved in all simulations with the exception of simulations following the SMM-i. In this case, $n_{max}$ = 50 was achieved when ERV replication was set to $6.0\times10^{-4}$, $7.0\times10^{-4}$, $8.0\times10^{-4}$, and $9.0\times10^{-4}$ retrotranspositions or reinfections per host generation ($r/g$). For other ERV replication variables, the total number of elements or tips ($n$) in the phylogenetic tree ranged from 36 to 50 elements, with the majority of observations between 48 to 50 elements (Table 1).

Unsurprisingly, the agreement between true and estimated phylogenies improved as sequence length increased: Maximum likelihood (ML) trees reconstructed using longer alignments of 100,000 bp had the lowest Robinson-Foulds (RF) distance [55, 56] to the true trees, followed by ML trees reconstructed using alignments of 10,000 bp and lastly, by ML trees from 1,000 bp alignments (S1–S3 Figs). In general, using the midpoint root on reconstructed trees was sufficiently robust; no strong difference was observed between true and ML trees when different ERV mutation rates were considered (S1–S3 Figs). However, as the rate of ERV

**Table 1. Frequency table showing the number of elements or tips (*n*) observed for each phylogenetic tree following the SMM-i when ERV replication was set to $1.0\times10^{-4}$ to $5.0\times10^{-4}$.**

| | *n* | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ERV replication** | **36** | **37** | **38** | **39** | **40** | **41** | **42** | **43** | **44** | **45** | **46** | **47** | **48** | **49** | **50** |
| $1.0\times10^{-4}$ | 1 | 3 | 10 | 13 | 39 | 53 | 78 | 89 | 100 | 95 | 58 | 40 | 16 | 5 | 0 |
| $2.0\times10^{-4}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 3 | 23 | 97 | 214 | 261 |
| $3.0\times10^{-4}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 60 | 535 |
| $4.0\times10^{-4}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 595 |
| $5.0\times10^{-4}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 597 |

doi:10.1371/journal.pone.0162454.t001

replication increased, it became more difficult to reconstruct trees similar to the true trees (S1–S3 Figs).

The Jensen-Shannon divergence (JSD) was calculated for the distribution of each statistics (see Material and Methods) between SMM-m and TM-m as well as SMM-i and TM-i (Figs 2–5).

For almost all true trees, the values of JSD were higher than zero (Figs 2 and 3) suggesting very different distributions for SMM-m and TM-m as well as SMM-i and TM-i. Visual inspection of boxplots (data not shown) for the distribution of each statistics confirmed that it is possible to distinguish between SMM-m and TM-m and between SMM-i and TM-i.

An exception was JSD calculated between SMM-m and TM-m for the proportion of terminal branch lengths that contributed to the total tree branch length ("prop") (Fig 2). In that

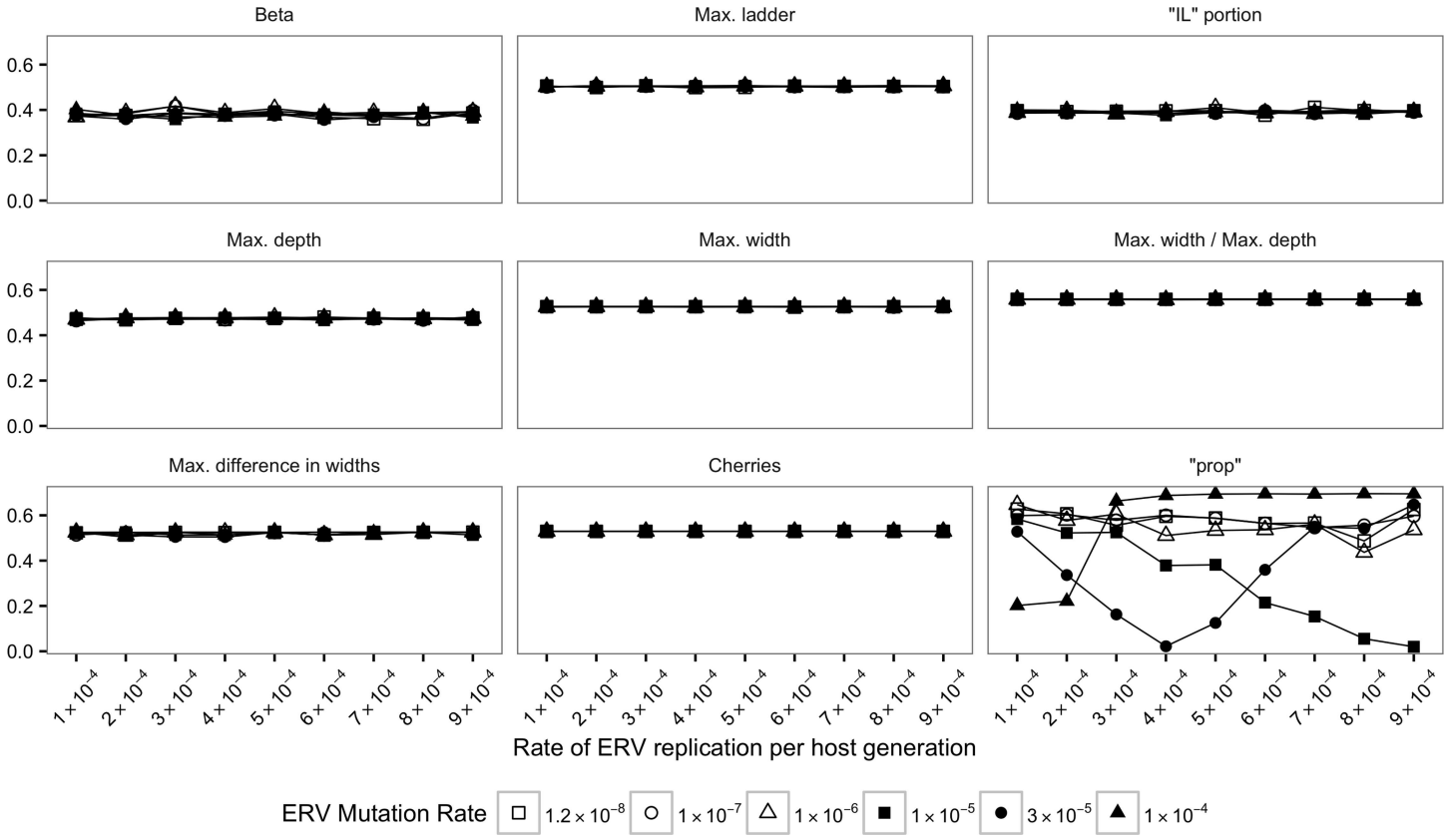


**Fig 2. Jensen-Shannon divergences (JSDs) between SMM-m and TM-m for true trees.** Each plot and its y-axis represent a statistic summary. ERV replication rate per host generation is depicted in the x-axis.

doi:10.1371/journal.pone.0162454.g002

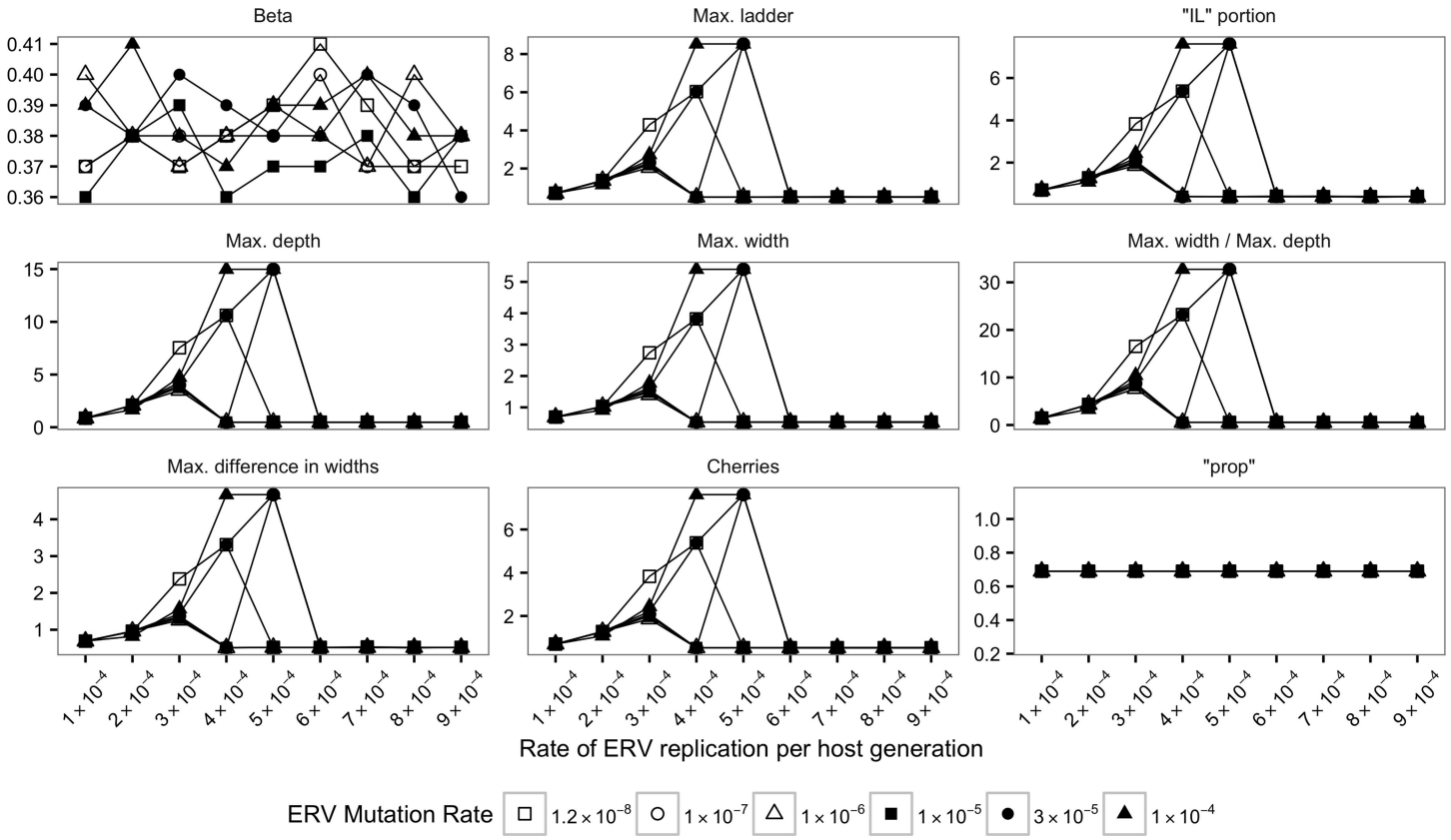


**Fig 3. Jensen-Shannon divergences (JSDs) between SMM-i and TM-i for true trees.** Each plot and its y-axis represent a statistic summary. ERV replication rate per host generation is depicted in the x-axis.

doi:10.1371/journal.pone.0162454.g003

case, when ERV mutation rate ($\mu_{erv}$) was $1.0\times10^{-5}$ *s/n/i*, JSD tended to zero as the rate of ERV replication per host generation (λ) increased. When $\mu_{erv} = 3.0\times10^{-5}$ *s/n/i*, JSD tended to zero until $\lambda = 4.0\times10^{-4}$ *r/g*. After this point JSD increased (Fig 2). Visual inspection of boxplots (data not shown) showed that for $\lambda < 4.0\times10^{-4}$ *r/g* the distribution of "prop" values obtained for TM-m are higher than those obtained for SMM-m. For $\lambda > 4.0\times10^{-4}$ *r/g*, the opposite was observed: the distribution of "prop" values obtained for SMM-m are higher than those obtained for TM-m. When $\lambda = 4.0\times10^{-4}$ *r/g*, the distribution of "prop" values obtained for TM-m and SMM-m were very similar. Finally, when $\mu_{erv} = 1.0\times10^{-4}$ *s/n/i*, and $\lambda \leq 2.0\times10^{-4}$ *r/g*, JSD was approximately 0.2. Visual inspection of boxplots (data not shown) showed some overlap of "prop" values. As ERV replication increased ($\lambda > 2.0\times10^{-4}$ *r/g*), JSD was higher than 0.6 showing that is possible to distinguish between SMM-m and TM-m (also confirmed by visual inspection of boxplots).

For ML trees, JSD calculated for the distribution of the 10 statistics (Figs 4 and 5) indicated that with long enough sequences it is possible to use shape statistics, "prop" and nucleotide diversity from ML trees to identify the underlying model of ERV evolution.

However, JSD values for tree shape statistics for ML trees reconstructed with alignments of 1,000 bp showed that in most cases it was difficult to distinguish between SMM-m and TM-m models (Fig 4) and between SMM-i and TM-i models (Fig 5). JSD for the distribution of tree shape statistics using alignments of 1,000 bp showed, in general, similar distributions for SMM-m and TM-m, suggested by values of JSD closer to zero. Similar pattern was observed for JSD calculated for SMM-i and TM-i and alignment of 1,000 bp. However, JSD for "IL"

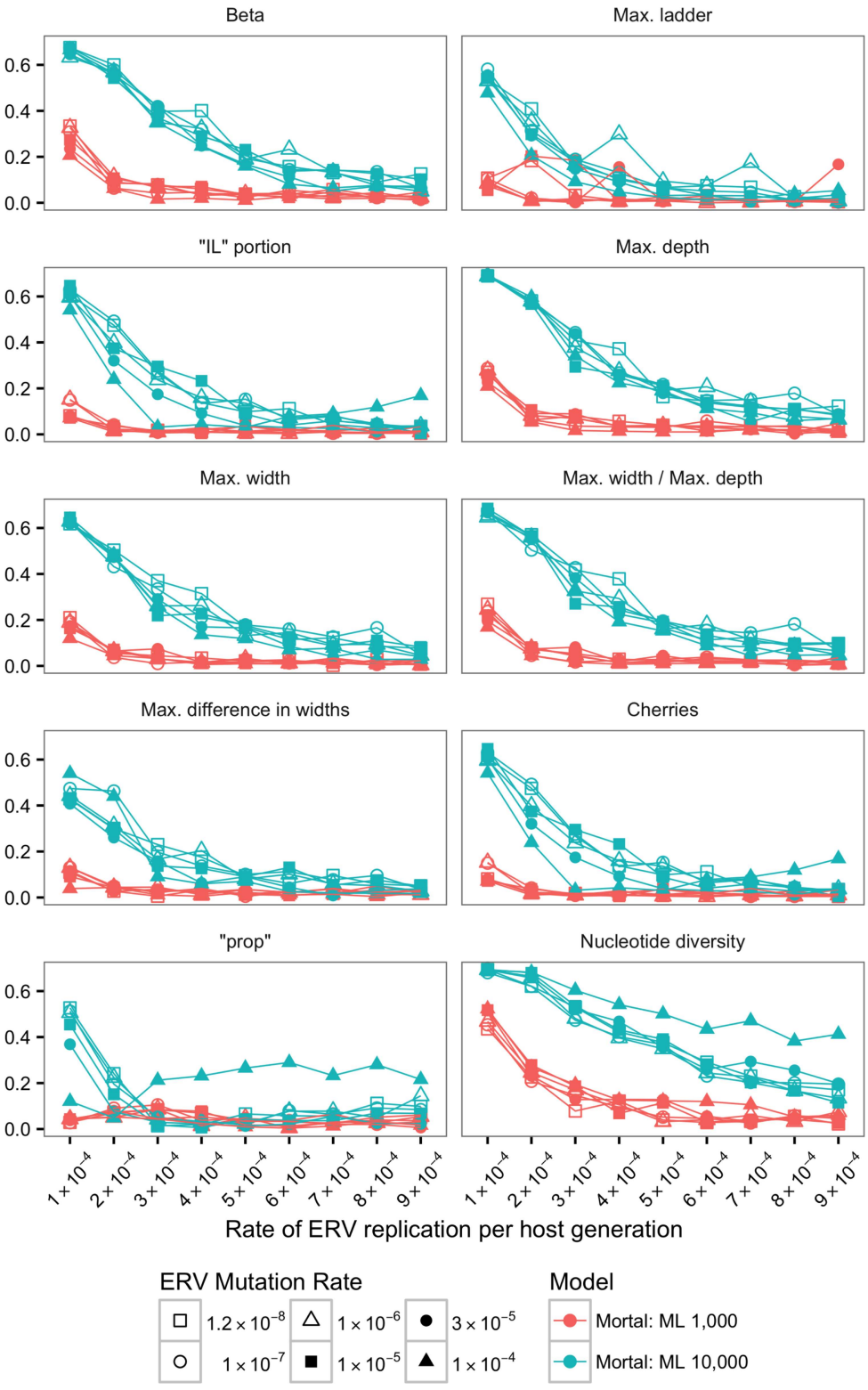


**Fig 4. Jensen-Shannon divergences (JSDs) between SMM-m and TM-m for ML trees using alignments of 1,000 bp and 10,000 bp.** Each plot and its y-axis represent a statistic summary. For all 10 plots, ERV replication rate per host generation is depicted in the x-axis.

doi:10.1371/journal.pone.0162454.g004

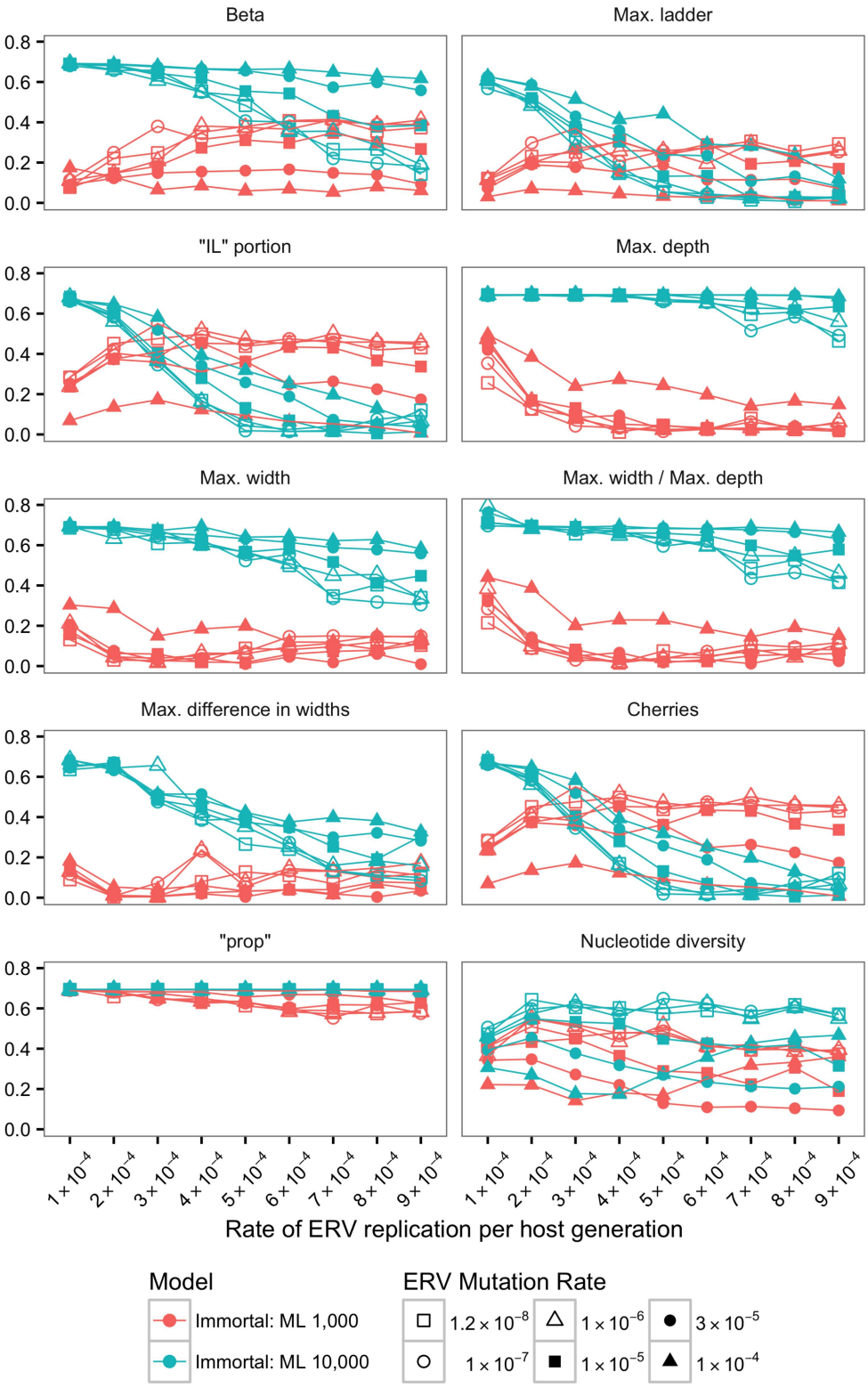


**Fig 5. Jensen-Shannon divergences (JSDs) between SMM-i and TM-i for ML trees using alignments of 1,000 bp and 10,000 bp.** Each plot and its y-axis represent a statistic summary. For all 10 plots, ERV replication rate per host generation is depicted in the x-axis.

doi:10.1371/journal.pone.0162454.g005

portion and number of cherries for immortal models and alignments of 1,000 bp suggested that it is possible to distinguish between SMM-i and TM-i (Fig 5).

In contrast, JSD values calculated for the distribution of tree shape statistics for ML trees reconstructed with alignments of 10,000 bp showed that it was possible to distinguish between SMM-m and TM-m models and between SMM-i and TM-i models when the rate of ERV replication was low (Figs 4 and 5). However, as this rate increased, it became, in general, more difficult to distinguish between SMM-m and TM-m models (Fig 4), and between SMM-i and TM-i models (Fig 5) as suggested by values of JSD approaching zero. However, JSD for max. depth, max. width, and max. width / max. depth for immortal models and alignments of 10,000 bp suggested that it was possible to distinguish between SMM-i and TM-i even when the rate of ERV replication was high.

JSD calculated for SMM-m and TM-m for ML trees reconstructed with alignments of 1,000 bp for "prop" (Fig 4) was close to zero for all ERV mutation and ERV replication rates analyzed, suggesting that it was not possible to distinguish between SMM-m and TM-m when using only "prop". In contrast, JSD calculated for SMM-m and TM-m for ML trees reconstructed with alignments of 10,000 bp (Fig 4) suggests that it was possible to distinguish between the two mortal models when ERV replication rate was the lowest.

JSD calculated for SMM-i and TM-i for ML trees reconstructed with alignments of 1,000 bp (Fig 5) for "prop" showed higher values than those obtained for the mortal models (Fig 4), suggesting that it was possible to distinguish between SMM-i and TM-i (Fig 5).

Our simulations show that it is also possible to distinguish between Mortal and Immortal models by calculating the proportion of terminal branch lengths to total tree branch length: True trees for Mortal models had longer terminal branches than true trees for Immortal models. This was true even when ML trees were reconstructed with alignments of 1,000 bp. In this case, only a small overlap of values calculated for SMM-m and SMM-i models were observed (data not shown). Nucleotide diversity was also a good measure to distinguish between Mortal and Immortal models as a higher diversity was always observed for Mortal models.

## Classification of ERV models

Although the four ERV models used for simulations are very different and indeed statistical analysis of true trees showed distinct values, it was still possible to distinguish between phylogenetic trees simulated under a Master or a Transposon model. The same was not observed when analyzing ML trees. For ML trees, JSD was closer to zero (indicating similar distributions) for several combinations of ERV mutation rate and replication, and the use of a classification method to distinguish between the different models, in this case, is very useful. In this context, we developed a *k*-nearest neighbor (*k*NN) classifier using ML trees reconstructed with alignments of 1,000 bp and 10,000 bp to identify models of ERV evolution. The *k*NN classifier was trained with ML trees and not true trees, because our results show that was not possible to fully recover true trees using alignments of 1,000 bp or 10,000 bp.

Pre-screening of JSD of the 10 statistics (data not shown) between all possible combinations of models (for example, TM-m *vs* TM-i, SMM-m *vs* TM-i, etc.) was used to detect the best combination of variables to be used in a *k*NN classifier. We chose the statistics with JSD values higher than 0.6 for the majority of observations. This suggested that statistics Beta, "IL" portion, number of cherries, "prop" and nucleotide diversity should be used to construct a classifier using ML trees reconstructed with alignments of 1,000 bp. Similarly, statistics Beta, max. depth, "prop" and nucleotide diversity should be used to construct a classifier using ML trees reconstructed with alignments of 10,000 bp.

To let the function automatically find the best number of nearest neighbor (see Materials and Methods), two sets of values of *k* were tested: *k* ranging from 1 to 30, and *k* ranging from

**Table 2. Table showing the results of a *k*-nearest neighbor (*k*NN) classifier using maximum likelihood (ML) trees reconstructed with alignments of 1,000 bp and 10,000 bp.** Precision is the proportion of the examples which truly have class x among all those which were classified as class x [75].

| | ML 1,000 bp | | | ML 10,000 bp | | |
|---|---|---|---|---|---|---|
| Correctly Classified Instances | 18,226 (84.37%) | | | 20,117 (93.13%) | | |
| Incorrectly Classified Instances | 3,374 (15.63%) | | | 1,483 (6.87%) | | |
| **ERV Model** | **TP Rate[1]** | **FP Rate[2]** | **Precision** | **TP Rate[1]** | **FP Rate[2]** | **Precision** |
| Master Immortal (SMM-i) | 0.999 | 0.001 | 0.996 | 1.000 | 0.000 | 1.000 |
| Master Mortal (SMM-m) | 0.623 | 0.081 | 0.719 | 0.839 | 0.038 | 0.881 |
| Transposon Immortal (TM-i) | 0.996 | 0.000 | 1.000 | 1.000 | 0.000 | 1.000 |
| Transposon Mortal (TM-m) | 0.756 | 0.126 | 0.667 | 0.887 | 0.054 | 0.846 |

[1] True Positive Rate

[2] False Positive Rate

doi:10.1371/journal.pone.0162454.t002

1 to 100. Very similar results were obtained. Below we report results obtained when *k* varied from 1 to 30.

With 1,000 bp alignments, 28 nearest neighbors were used in the *k*NN classifier. Using this classifier to predict the model underlining our own ML trees reconstructed with 1,000 bp alignments, we were able to correctly classify 84.37% of the four ERV models assessed in this study (Table 2). No misclassification was observed between Mortal and Immortal models, and all SMM-m was correctly classified. A higher misclassification rate was observed between SMM-m and TM-m models than between SMM-i and TM-i models.

With 10,000 bp alignments, 12 nearest neighbors were used in the *k*NN classifier. Using this classifier to predict the model underlining our own ML trees reconstructed with 10,000 bp alignments, we were able to correctly classify 93.13% of all models (Table 2). No misclassification was observed between Mortal and Immortal models, and all SMM-i and all TM-i were correctly classified.

Using both classifiers mentioned above, it was difficult to classify SMM-m and TM-m models, but a better classification was achieved when training a *k*NN using data calculated for ML trees reconstructed with alignments of 10,000 bp (Table 2).

## Empirical data

All data analyzed for PERV gamma1 was classified as TM-i with a very high probability (99.9%). The *env* E gene of the gamma2 PERV was classified as TM-m also with a very high probability (92.8%).

# Discussion

Transposable elements make up a part of the large fraction of what is considered non-coding DNA in eukaryotic genomes [76, 77], and recent studies are showing the significant role of TEs in shaping host genomes by restructuring genes and providing new regulatory sequences [2, 78, 79]. Although TEs are considered as non-coding DNA in their host genomes, TEs can encode their own proteins responsible for their replication. TE-encoded proteins can be co-opted into functional proteins within the host in an evolutionary process referred to as "molecular domestication" [79]. The most interesting example of co-option between ERV and hosts is the use of the *env* ERV gene in the formation of the placenta in mammals, including primates, rodents, lagomorphs and marsupials [26, 27, 80]. TE mobility can also negatively affect the host, as they are associated with disease by insertional mutagenesis and homologous recombination [81, 82].

Several studies suggest that retrotransposons, including ERVs, replicate following a SMM or TM [35, 83, 84]. In these cases, phylogenetic tree topologies were used as a reliable indication to determine whether an ERV lineage replicates following one of these models. ERV lineages following a SMM always generate completely unbalanced phylogenetic trees, while those following a TM tend to generate more balanced trees [35]. Our results from analyzing true trees also confirm this previous study [35]. Even though our study confirms the phylogenetic trees expected for SMM and TM as described in Clough *et al.* [35], we demonstrated that to accurately reconstruct these trees is not possible either using alignments of 1,000 bp or 10,000 bp. Unsurprisingly, it was more difficult to reconstruct ML trees that were similar to the true trees with shorter sequences. Our analyses, using alignments of 1,000, 10,000 and 100,000 bp, show that a likely reason for not recovering the true tree when ML trees are reconstructed using alignments of 1,000 and 10,000 bp is sampling error induced by limited numbers of sites.

Because an ERV genome size is approximately 10,000 bp and because ERV genomes can be mined from publicly available host genomes, we focused on understanding whether ML trees reconstructed with alignments of 1,000 or 10,000 bp would be consistent with a SMM or TM. Even though it is difficult to distinguish between different models by visual inspection of ML trees, we were still able to classify with high accuracy the four ERV models proposed in this study using the *k*NN classifier and the summary statistics for ML trees reconstructed with alignments of 10,000 bp. Interestingly, we were also able to correctly classify ML trees reconstructed with alignments of 1,000 bp, although the false positive rate for the Mortal models was higher than that obtained using 10,000 bp alignments.

There is a great interest in knowing whether an ERV lineage is still able to proliferate in a host genome or whether this ability has been lost. The ability to proliferate involves several mechanisms from reinfection to retrotransposition in *cis* and complementation in *trans*. Reinfection can lead to cross-species transmission of ERVs, which have been documented and occurs more than previously thought [85]. Even though our models do not distinguish between these mechanisms or whether horizontal transmissions have occurred, our models can detect whether a retrotransposon is still able to retrotranspose to different loci in a group of closely related ERVs. New ERV integrations may have several consequences for the host, from beneficial to detrimental and these new integrations may disrupt a host gene or cause diseases [7, 13, 86]. Our results indicated that although it was difficult to distinguish between SMM-m and TM-m models, a total separation was achieved between Mortal and Immortal models when using ML trees reconstructed with alignments of either 1,000 bp or 10,000 bp. These results suggest that it is possible to understand whether an ERV lineage is still able to show ongoing activity by retrotransposing or reinfecting host cells, or whether it lost this ability a very long time ago. This would be a cheaper alternative to pre-screening for ERVs that can or not retrotranspose or reinfect.

Other explanations have also been proposed to describe how retrotransposons replicate in host genomes [34]. For example, there are suggestions that more than one master template exists for an ERV lineage; or during the course of an ERV lineage evolution, a master template may become inactive with another ERV copy occupying its position [45, 87]. In this paper, we focused on understanding whether it was possible to distinguish evolutionary patterns with simpler models before simulating more complex models. According to our results, we would expect that in cases where more than one master templates are present in an ERV lineage, ML trees would resemble those simulated under the simpler SMM models proposed in this study.

Our results were consistent in showing that visual inspection of phylogenetic trees, e.g. [45, 83], is not the appropriate method to decide whether an ERV lineage is replicating following a SMM or TM model, and it is likely that this result can be extended to any TE lineage with limited genome size. In addition, using only one metric to check tree imbalance–in this paper, the

Beta statistic–is also not sufficient in distinguishing between the models proposed here. This was evaluated by training a classifier using only the Beta statistics. In this case, only 63.24% and 58.98% of the models were correctly classified when a *k*NN was trained with ML trees reconstructed with alignments of 1,000 bp and 10,000 bp, respectively. We suggest that for a better classification of models, the Beta, "IL" portion, number of cherries, "prop" and nucleotide diversity should be used in a classifier when using alignments of approximately 1,000 bp. Similarly, the Beta, max. depth, "prop" and nucleotide diversity should be used in a classifier when using alignments of approximately 10,000 bp. When a classifier trained using these statistics was used, we were able to correctly classify ≈84% and ≈93% of all models when using summary statistics calculated for ML trees reconstructed with alignments of 1,000 bp and 10,000 bp, respectively.

Analysis of empirical data using our approach suggested that gamma1 PERVs are still able to replicate. In fact, the ability of gamma1 PERVs to replicate in host genomes is supported by *in vitro* studies [22, 70]. On the other hand, analysis of gamma2 PERVs using our approach suggested that these ERVs may have lost this ability; again, genetic studies showing that these ERVs have several stop codons and frame-shift mutations in all their genes [68] are consistent with this conclusion. Expression analysis of gamma2 PERVs also showed an inconsistent pattern when different samples were analyzed corroborating to the hypothesis that gamma2 may not be replicating [47, 69]. Our analysis also indicated that gamma1 PERVs in *Sus scrofa* is possibly replicating in accordance with a TM-i, while gamma2 PERVs in *Sus* species are replicating following a TM-m.

## Conclusion

We confirmed a previous study [35] that SMM and TM show very distinct phylogenetic tree shape. However, we demonstrated for the first time that it is not possible to accurately reconstruct these true trees using either alignments of 1,000 bp or 10,000 bp. A likely reason for this was sampling errors induced by limited number of sites, as reconstruction of true trees using alignments of 100,000 bp showed the lowest Robinson-Foulds distance (S1–S3 Figs). Given that the size of an ERV genome is limited and approximately 10,000 bp, and based on information obtained in this study we developed a *k*NN classifier to predict the likely model of TE replication and evolution in host genomes.

We suggest that instead of visual inspection of phylogenetic tree as used in some studies, e.g. [45, 83], one should calculate the statistics proposed in this study and use the respective classifier to gain a better understanding of the underlying model of TE replication and evolution. This developed classifier could also be used to predict whether a retrotransposon lineage is still able to proliferate or lost this ability a long time ago.

Although the proposed models described in this study represent simplistic models of ERV replication and evolution. This study represents an important step to understand whether it is possible to reconstruct trees similar to the expected trees under the SMM and TM. With the development of a *k*NN classifier we were able to distinguish between models with high accuracy. If we were unable to predict whether phylogenetic trees were from a SMM or TM, it would be unlikely to do so using more complex models. This is because more complex models of ERV evolution would involve variations of the simplistic models we are analyzing in this study. Our results are promising for the future development of more complex models of ERV replication and evolution in host genomes.

## Supporting Information

**S1 Fig. Robinson-Foulds (RF) metric for true and ML trees reconstructed with alignments of 1,000 bp.** Plots for each ERV mutation rate showing RF metric (y-axis) for the Strict Master

(SMM) and Transposon (TM) mortal and immortal models for ERV replication per host generation (x-axis). RF metrics were calculated for rooted and unrooted trees comparing true phylogenetic trees with ML trees reconstructed using alignments of 1,000 bp.
(PDF)

**S2 Fig. Robinson-Foulds (RF) metric for true and ML trees reconstructed with alignments of 10,000 bp.** Plots for each ERV mutation rate showing RF metric (y-axis) for the Strict Master (SMM) and Transposon (TM) mortal and immortal models for ERV replication per host generation (x-axis). RF metrics were calculated for rooted and unrooted trees comparing true phylogenetic trees with ML trees reconstructed using alignments of 10,000 bp.
(PDF)

**S3 Fig. Robinson-Foulds (RF) metric for true and ML trees reconstructed with alignments of 100,000 bp.** Plots for each ERV mutation rate showing RF metric (y-axis) for the Strict Master (SMM) and Transposon (TM) mortal and immortal models for ERV replication per host generation (x-axis). RF metrics were calculated for rooted and unrooted trees comparing true phylogenetic trees with ML trees reconstructed using alignments of 100,000 bp.
(PDF)

## Acknowledgments

We would like to thank Nicole Duncan for proofreading previous versions of this manuscript. We would like to thank Drs. Aris Katzourakis, Jeet Sukumaran, David Swofford, and Steven Wu for helpful discussion during the development of this manuscript.

## Author Contributions

**Conceptualization:** FFN.

**Formal analysis:** FFN AGR.

**Funding acquisition:** FFN.

**Methodology:** FFN AGR.

**Resources:** FFN AGR.

**Software:** FFN.

**Supervision:** FFN.

**Visualization:** FFN.

**Writing – original draft:** FFN.

**Writing – review & editing:** FFN AGR.

## References

1. Flutre T, Permal E, Quesneville H. In search of lost trajectories: Recovering the diversification of transposable elements. Mob Genet Elements. 2011; 1(2):151–4. PMID: 22016865
2. Fedoroff NV. Presidential address. Transposable elements, epigenetics, and genome evolution. Science. 2012; 338(6108):758–67. PMID: 23145453
3. de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 2011; 7(12):e1002384. doi: 10.1371/journal.pgen.1002384 PMID: 22144907

4. Siguier P, Filee J, Chandler M. Insertion sequences in prokaryotic genomes. Curr Opin Microbiol. 2006; 9(5):526–31. PMID: 16935554
5. Richardson SR, Doucet AJ, Kopera HC, Moldovan JB, Garcia-Perez JL, Moran JV. The influence of LINE-1 and SINE retrotransposons on mammalian genomes. Microbiol Spectr. 2015; 3(2):MDNA3-0061-2014.
6. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007; 8(12):973–82. PMID: 17984973
7. Jern P, Coffin JM. Effects of retroviruses on host genome function. Annu Rev Genet. 2008; 42:709–32. doi: 10.1146/annurev.genet.42.110807.091501 PMID: 18694346
8. Johnson WE. Endogenous retroviruses in the genomics era. Annu Rev Virol. 2015; 2(1):135–59. doi: 10.1146/annurev-virology-100114-054945 PMID: 26958910
9. Gifford R, Tristem M. The evolution, distribution and diversity of endogenous retroviruses. Virus Genes. 2003; 26(3):291–315. PMID: 12876457
10. Coffin JM. Retroviridae: The viruses and their replication. In: Fields BN, Knipe PM, Howler PM, Chanock RM, Monath TP, Melnick JL, et al., editors. Fields Virology. Philadelphia: Lippincott—Raven Publishers; 1996. p. 1767–847.
11. Stoye JP. Endogenous retroviruses: still active after all these years? Curr Biol. 2001; 11(22):R914–6. PMID: 11719237
12. Feschotte C, Gilbert C. Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet. 2012; 13(4):283–96. doi: 10.1038/nrg3199 PMID: 22421730
13. Weiss RA. On the concept and elucidation of endogenous retroviruses. Phil Trans R Soc Lond B Biol Sci. 2013; 368(1626):20120494.
14. Friedli M, Trono D. The developmental control of transposable elements and the evolution of higher species. Annu Rev Cell Dev Biol. 2015; 31:429–51. doi: 10.1146/annurev-cellbio-100814-125514 PMID: 26393776
15. Stoye JP. Studies of endogenous retroviruses reveal a continuing evolutionary saga. Nat Rev Microbiol. 2012; 10(6):395–406. doi: 10.1038/nrmicro2783 PMID: 22565131
16. Gifford RJ. Viral evolution in deep time: lentiviruses and mammals. Trends Genet. 2012; 28(2):89–100. doi: 10.1016/j.tig.2011.11.003 PMID: 22197521
17. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001; 409(6822):860–921. PMID: 11237011
18. Magiorkinis G, Gifford RJ, Katzourakis A, De Ranter J, Belshaw R. Env-less endogenous retroviruses are genomic superspreaders. Proc Natl Acad Sci USA. 2012; 109(19):7385–90. doi: 10.1073/pnas.1200913109 PMID: 22529376
19. Gifford RJ, Katzourakis A, Tristem M, Pybus OG, Winters M, Shafer RW. A transitional endogenous lentivirus from the genome of a basal primate and implications for lentivirus evolution. Proc Natl Acad Sci USA. 2008; 105(51):20362–7. doi: 10.1073/pnas.0807873105 PMID: 19075221
20. Katzourakis A, Tristem M, Pybus OG, Gifford RJ. Discovery and analysis of the first endogenous lentivirus. Proc Natl Acad Sci USA. 2007; 104(15):6261–5. PMID: 17384150
21. Hanger JJ, Bromham LD, McKee JJ, O'Brien TM, Robinson WF. The nucleotide sequence of koala (*Phascolarctos cinereus*) retrovirus: a novel type C endogenous virus related to Gibbon ape leukemia virus. J Virol. 2000; 74(9):4264–72. PMID: 10756041
22. Le Tissier P, Stoye JP, Takeuchi Y, Patience C, Weiss RA. Two sets of human-tropic pig retrovirus. Nature. 1997; 389(6652):681–2. PMID: 9338777
23. Fabryova H, Hron T, Kabickova H, Poss M, Elleder D. Induction and characterization of a replication competent cervid endogenous gammaretrovirus (CrERV) from mule deer cells. Virology. 2015; 485:96–103. doi: 10.1016/j.virol.2015.07.003 PMID: 26218214
24. Rowe HM, Trono D. Dynamic control of endogenous retroviruses during development. Virology. 2011; 411(2):273–87. doi: 10.1016/j.virol.2010.12.007 PMID: 21251689
25. Mita P, Boeke JD. How retrotransposons shape genome regulation. Curr Opin Genet Dev. 2016; 37:90–100. doi: 10.1016/j.gde.2016.01.001 PMID: 26855260
26. Dupressoir A, Lavialle C, Heidmann T. From ancestral infectious retroviruses to bona fide cellular genes: role of the captured syncytins in placentation. Placenta. 2012; 33(9):663–71. doi: 10.1016/j.placenta.2012.05.005 PMID: 22695103
27. Cornelis G, Vernochet C, Carradec Q, Souquere S, Mulot B, Catzeflis F, et al. Retroviral envelope gene captures and syncytin exaptation for placentation in marsupials. Proc Natl Acad Sci USA. 2015; 112(5):E487–96. doi: 10.1073/pnas.1417000112 PMID: 25605903

28. Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science. 2016; 351(6277):1083–7. doi: 10.1126/science.aad5497 PMID: 26941318

29. Levin HL, Moran JV. Dynamic interactions between transposable elements and their hosts. Nat Rev Genet. 2011; 12(9):615–27. doi: 10.1038/nrg3030 PMID: 21850042

30. Boeke JD, Stoye JP. Retrotransposons, endogenous eetroviruses, and the evolution of retroelements. In: Coffin JM, Hughes SH, Varmus HE, editors. Retroviruses. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 1997.

31. Bannert N, Kurth R. The evolutionary dynamics of human endogenous retroviral families. Annu Rev Genomics Hum Genet. 2006; 7:149–73. PMID: 16722807

32. Katzourakis A, Rambaut A, Pybus OG. The evolutionary dynamics of endogenous retroviruses. Trends Microbiol. 2005; 13(10):463–8. PMID: 16109487

33. Pavlicek A, Paces J, Elleder D, Hejnar J. Processed pseudogenes of human endogenous retroviruses generated by LINEs: their integration, stability, and distribution. Genome Res. 2002; 12(3):391–9. PMID: 11875026

34. Cordaux R, Hedges DJ, Batzer MA. Retrotransposition of *Alu* elements: how many sources? Trends Genet. 2004; 20(10):464–7. PMID: 15363897

35. Clough JE, Foster JA, Barnett M, Wichman HA. Computer simulation of transposable element evolution: random template and strict master models. J Mol Evol. 1996; 42(1):52–8. PMID: 8576964

36. Belshaw R, Katzourakis A, Paces J, Burt A, Tristem M. High copy number in human endogenous retrovirus families is associated with copying mechanisms in addition to reinfection. Mol Biol Evol. 2005; 22 (4):814–7. PMID: 15659556

37. Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. Viral mutation rates. J Virol. 2010; 84 (19):9733–48. doi: 10.1128/JVI.00694-10 PMID: 20660197

38. Rein A. Murine leukemia viruses: objects and organisms. Adv Virol. 2011; 2011:403419. doi: 10.1155/2011/403419 PMID: 22312342

39. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467(7319):1061–73. doi: 10.1038/nature09534 PMID: 20981092

40. Holmes EC. Error thresholds and the constraints to RNA virus evolution. Trends Microbiol. 2003; 11 (12):543–6. PMID: 14659685

41. Scally A, Durbin R. Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet. 2012; 13(10):745–53. doi: 10.1038/nrg3295 PMID: 22965354

42. Patience C, Switzer WM, Takeuchi Y, Griffiths DJ, Goward ME, Heneine W, et al. Multiple groups of novel retroviral genomes in pigs and related species. J Virol. 2001; 75(6):2771–5. PMID: 11222700

43. Rambaut A, Grassly NC. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci. 1997; 13(3):235–8. PMID: 9183526

44. Jukes TH, Cantor CR. Evolution of protein molecules III. In: Munro HN, editor. Mammalian protein metabolism: Academic Press; 1969. p. 21–132.

45. Buzdin A, Ustyugova S, Khodosevich K, Mamedov I, Lebedev Y, Hunsmann G, et al. Human-specific subfamilies of HERV-K (HML-2) long terminal repeats: three master genes were active simultaneously during branching of hominoid lineages. Genomics. 2003; 81(2):149–56. PMID: 12620392

46. Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A, et al. Long-term reinfection of the human genome by endogenous retroviruses. Proc Natl Acad Sci USA. 2004; 101(14):4894–9. PMID: 15044706

47. Nascimento FF, Gongora J, Charleston M, Tristem M, Lowden S, Moran C. Evolution of endogenous retroviruses in the Suidae: evidence for different viral subpopulations in African and Eurasian host species. BMC Evol Biol. 2011; 11:139. doi: 10.1186/1471-2148-11-139 PMID: 21609472

48. Nascimento FF, Gongora J, Tristem M, Lowden S, Moran C. Distinctive differences in long terminal repeat sequences between gamma1 endogenous retroviruses of African and Eurasian suid species. Infect Genet Evol. 2011; 11(3):686–93. doi: 10.1016/j.meegid.2011.01.010 PMID: 21256982

49. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30(9):1312–3. doi: 10.1093/bioinformatics/btu033 PMID: 24451623

50. Tavaré S. Some probabilistic and statistical problems on the analysis of DNA sequences. Lect Math Life Sci. 1986; 17:57–86.

51. Yang Z. Estimating the pattern of nucleotide substitution. J Mol Evol. 1994; 39(1):105–11. PMID: 8064867

52. Zharkikh A. Estimation of evolutionary distances between nucleotide sequences. J Mol Evol. 1994; 39 (3):315–29. PMID: 7932793

53. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2014 [February 27, 2015]. Available from: http://www.R-project.org/.

54. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011; 27(4):592–3. doi: 10.1093/bioinformatics/btq706 PMID: 21169378

55. Steel MA, Penny D. Distributions of tree comparison metrics—some new results. Syst Biol. 1993; 42:126–41.

56. Robinson DF, Foulds LR. Comparison of phylogenetic trees. Math Biosci. 1981; 53:131–47.

57. Aldous DJ. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. Stat Sci. 2001; 16:23–34.

58. Bortolussi N, Durand E, Blum M, Francois O. apTreeshape: statistical analysis of phylogenetic tree shape. Bioinformatics. 2006; 22(3):363–4. PMID: 16322049

59. Boyd M, Colijn C. phyloTop: Phylogenetic tree topological properties evaluator. R package version 1.1.1. 2014 [February 27, 2015]. Available from: http://CRAN.R-project.org/package=phyloTop.

60. Colijn C, Gardy J. Phylogenetic tree shapes resolve disease transmission patterns. Evol Med Public Health. 2014; 2014(1):96–108. doi: 10.1093/emph/eou018 PMID: 24916411

61. McKenzie A, Steel M. Distributions of cherries for two models of trees. Math Biosci. 2000; 164(1):81–92. PMID: 10704639

62. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004; 20(2):289–90. PMID: 14734327

63. Paradis E. pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics. 2010; 26(3):419–20. doi: 10.1093/bioinformatics/btp696 PMID: 20080509

64. Lin J. Divergence measures based on the Shannon entropy. IEEE Trans Inf Theor. 2006; 37(1):145–51.

65. Hornik K, Buchta C, Zeileis A. Open-source machine learning: R meets Weka. Comput Stat. 2009; 24:225–32.

66. Akiyoshi DE, Denaro M, Zhu H, Greenstein JL, Banerjee P, Fishman JA. Identification of a full-length cDNA for an endogenous retrovirus of miniature swine. J Virol. 1998; 72(5):4503–7. PMID: 9557749

67. Takeuchi Y, Patience C, Magre S, Weiss RA, Banerjee PT, Le Tissier P, et al. Host range and interference studies of three classes of pig endogenous retrovirus. J Virol. 1998; 72(12):9986–91. PMID: 9811736

68. Mang R, Maas J, Chen X, Goudsmit J, van Der Kuyl AC. Identification of a novel type C porcine endogenous retrovirus: evidence that copy number of endogenous retroviruses increases during host inbreeding. J Gen Virol. 2001; 82(Pt 8):1829–34. PMID: 11457988

69. Klymiuk N, Muller M, Brem G, Aigner B. Phylogeny, recombination and expression of porcine endogenous retrovirus gamma2 nucleotide sequences. J Gen Virol. 2006; 87(Pt 4):977–86. PMID: 16528048

70. Patience C, Takeuchi Y, Weiss RA. Infection of human cells by an endogenous retrovirus of pigs. Nat Med. 1997; 3(3):282–6. PMID: 9055854

71. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32(5):1792–7. PMID: 15034147

72. Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. 2010; 27(2):221–4. doi: 10.1093/molbev/msp259 PMID: 19854763

73. Yang Z. Computational molecular evolution. Oxford: Oxford University Press; 2006.

74. Huerta-Cepas J, Dopazo J, Gabaldon T. ETE: a python Environment for Tree Exploration. BMC Bioinformatics. 2010; 11:24. doi: 10.1186/1471-2105-11-24 PMID: 20070885

75. Bouckaert RR, Frank E, Hall M, Kirkby R, Reutemann P, Seewald A, et al. WEKA manual for version 3-7-11. 2013 [February 27, 2015]. Available from: http://www.cs.waikato.ac.nz/ml/weka/documentation.html.

76. Doolittle WF, Sapienza C. Selfish genes, the phenotype paradigm and genome evolution. Nature. 1980; 284(5757):601–3. PMID: 6245369

77. Orgel LE, Crick FH. Selfish DNA: the ultimate parasite. Nature. 1980; 284(5757):604–7. PMID: 7366731

78. Cowley M, Oakey RJ. Transposable elements re-wire and fine-tune the transcriptome. PLoS Genet. 2013; 9(1):e1003234. doi: 10.1371/journal.pgen.1003234 PMID: 23358118

79. Feschotte C. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 2008; 9 (5):397–405. doi: 10.1038/nrg2337 PMID: 18368054

80. Sinzelle L, Izsvak Z, Ivics Z. Molecular domestication of transposable elements: from detrimental parasites to useful host genes. Cell Mol Life Sci. 2009; 66(6):1073–93. doi: 10.1007/s00018-009-8376-3 PMID: 19132291

81. Solyom S, Kazazian HH Jr. Mobile elements in the human genome: implications for disease. Genome Med. 2012; 4(2):12. doi: 10.1186/gm311 PMID: 22364178

82. Belancio VP, Deininger PL, Roy-Engel AM. LINE dancing in the human genome: transposable elements and disease. Genome Med. 2009; 1(10):97. doi: 10.1186/gm97 PMID: 19863772

83. Barrio AM, Ekerljung M, Jern P, Benachenhou F, Sperber GO, Bongcam-Rudloff E, et al. The first sequenced carnivore genome shows complex host-endogenous retrovirus relationships. PLoS One. 2011; 6(5):e19832. doi: 10.1371/journal.pone.0019832 PMID: 21589882

84. Costas J. Evolutionary dynamics of the human endogenous retrovirus family HERV-K inferred from full-length proviral genomes. J Mol Evol. 2001; 53(3):237–43. PMID: 11523010

85. Hayward A, Grabherr M, Jern P. Broad-scale phylogenomics provides insights into retrovirus-host evolution. Proc Natl Acad Sci USA. 2013; 110(50):20146–51. doi: 10.1073/pnas.1315419110 PMID: 24277832

86. Kurth R, Bannert N. Beneficial and detrimental effects of human endogenous retroviruses. Int J Cancer. 2010; 126(2):306–14. doi: 10.1002/ijc.24902 PMID: 19795446

87. Furano AV. The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons. Prog Nucleic Acid Res Mol Biol. 2000; 64:255–94. PMID: 10697412