Supporting Information

Magiorkinis et al. 10.1073/pnas.1200913109

SI Materials and Methods

Genome Mining. All available mammal genomes were screened *in silico* according to a previously described algorithm (1). We first built a library of amino acids representing a 181-aa alignment of the reverse transcriptase domain of *pol* from known endogenous retrovirus (ERV) and exogenous retrovirus (XRV) species. Each time we found a *pol* distantly related to the library, we used it as a new probe and rescreened the genomes for even more distantly related loci, continuing until no new loci were found. From our *pol* coordinates we extracted an initial 600-nt sequence representing each locus. Finally, we provisionally allocated all 83,614 recovered loci to a family based on their closest similarity to sequences in the probe library. In so doing we also created a group of intracisternal A-type particle (IAP)-like families containing a total of 5,969 loci.

Selection of Loci. The criteria for exclusion of loci based on sequence similarity to the nearest neighbor are explained in the main text. This exclusion was not necessary for IAPs, all of which invaded their hosts after speciation; if IAPs had colonized the common ancestor of two species, then we should observe loci in one genome being phylogenetically closer to loci from the other species. No expansions in IAPs have this pattern (Fig. 2). In five invasions of the mouse genome the sister group is in the rat genome, but in each case the two clades are separated by long internal branches. This sister-group relationship probably results from the mouse and rat being the two most closely related species among the sequenced rodents and host phylogeny affecting the ability of an IAP to invade a new host.

Alignment. We aligned all the IAP-like loci against the *pol* gene of an IAPE [an IAP locus shown to have a functional *env* (2)], using the *BlastAlign* program (3) and kept those loci having gaps in the alignment representing less than 50% of their length. This process produced a multiple alignment of 1,037 sites containing 4,929 loci. We then edited this alignment manually to preserve the correct reading frame. To confirm the monophyly of the IAPs, we used Clustal-W (4) to profile-align the IAP alignment with an alignment of all known XRV *pol* sequences. After manual editing we produced a second, temporary multiple alignment of 400 sites, which in a phylogenetic analysis (below) showed that 4,913 of our 4,929 loci formed a single clade within the class II ERVs. These 4,913 loci were considered to represent the IAP lineage, and we excluded the remaining 16 loci. (We assume these 16 loci represent chimerical or very old sequences or belong to more distantly related ERV lineages). To strengthen our phylogenetic analysis, we then also excluded loci that had <600 nt in the initial IAP alignment, giving us a final dataset of 4,089 loci.

We also produced a protein alignment (764 aa) of the *pol* regions for selected class II ERVs with Clustal-W, which we subsequently edited manually (see *SI Results*).

Phylogenetic Analyses. For analyzing the IAPs, we used the *FastTree* program, which uses a combination of distance (neighbor-joining) and maximum-likelihood heuristics to estimate phylogenetic trees using the General Time Reversible model accounting for varying rates of evolution across sites (CAT model) (5). Phylogenetic uncertainty was assessed by the Shimodaira–Hasegawa test (SH-like local support values) for each split as implemented in FastTree. SH-like support values have been shown to be significantly and strongly correlated with bootstrap values, especially when they are >0.90 (5). We used *FigTree* (http://tree.bio.ed.ac.uk/software/figtree/) to plot the genetic characteristics of each locus onto the estimated phylogenetic tree. The tree of the sequenced hosts (Fig. 1) was built by pruning unsequenced species from a published phylogenetic tree of mammals (6).

To build our tree of the *pol* regions for selected class II ERVs (*SI Results*), we used *MrBayes* (7), using the WAG matrix of amino acid substitutions and running four chains of Metropolis Coupled Markov Chains Monte Carlo for 10^6 generations. We visually inspected the mixing of the parameters with *Tracer* (http://tree.bio.ed.ac.uk/software/tracer/) and used 10^5 generations as burn-in to obtain a sufficient estimated sample size of at least 100. We show posterior probabilities >0.7 and consider branches with a probability of at least 0.9 to be well supported.

All trees presented were midpoint rooted.

Simulating Frequency Distributions. Random generation of family sizes was done in R. For the generalized Pareto distribution, parameters “shape” and “scale” were fitted to the real data using gpd.fit (package gPdtest) and data simulated with these values using rgpd (package POT). We used rlnorm for the lognormal distribution. In Fig. 4, the mean of 1,000 replicates is shown; for clarity, we restricted possible values to the maximum value shown in the horizontal axis.

Gene Integrity. To measure gene integrity of the IAP loci we extracted 7,000 nt of sequence from both sides of all *pol* coordinates. Many of the genomes are only partially assembled because of low sequencing coverage, so to avoid the bias of fragmentation caused by incomplete genomic assembly, we retained only extracted fragments having length of at least 13,000 nt (n = 3,834), which we refer to as “full-length” sequences; that is, we kept only fragments that were long enough to contain the entire ERV sequence. We extracted all of the ORF products >300 nt using the getorf program of the EMBOSS suite (8). These amino acid sequences then were searched by BLASTP (9) using a probe library of XRV *gag, pol, prot*, and *env* genes plus ERVs that lacked close XRV relatives (2, 10, 11), including the genes from IAPE. Matches were considered valid when they had an e-value of at least 10^{-4}. We subsequently used the length of the query nucleotide sequences as our measure of gene integrity, and when a gene was fragmented into more than one ORF, we used the longest one. To inspect the clustering of one gene’s degradation against another visually, we used Cyflogic to plot scattergrams of the integrity metrics (http://www.cyflogic.com/) (Fig. S1).

A potential problem is that the length of the longest ORF can show large changes even when only minor postintegration mutational changes (e.g., the acquisition of one premature stop-codon) have occurred. We therefore also used a second measure of gene integrity for the IAPs, which is the locus’s nucleotide similarity to known functional genes. For this assessment, we compared loci with the amino acid sequences of the published IAPE element using TBLASTN (9) and used the resulting bit score as a metric of the nucleotide sequence integrity. Use of this metric gave highly correlated results to the longest ORF in a set of loci belonging to a single expansion. We report here only the results using the former method, because we consider it a better metric, not conflating gene integrity with divergence when we compare loci from different families.

As a second and independent measure of locus age, we searched full-length IAP loci for paired LTRs having at least 95% similarity using LTR-harvest (12). LTRs are identical at the time of integration and gradually accumulate mutations during the replication of the host. Therefore, more similar paired LTRs typically represent more recently integrated loci.

For our analysis of all ERVs, we extracted 7,000 nt from both sides of initial *pol* coordinates as described above for IAP loci. We then found the longest ORF matching our *env* and *gag* probe libraries as described above using a series of Perl scripts. In Table S1 we present the mean values in the family for both genes. *env* must be compared with *gag*, because low values in both can reflect both age and quality of the genome assembly. To give an indication of the age of the loci in the family, we also include the mean pairwise nucleotide sequence similarity, measured with the Water program of the EMBOSS suite, which implements the Smith–Waterman algorithm.

For the class II ERV families analyzed in *SI Results*, we confirmed the absence of *env* by visually inspecting a random sample of at least 25% of the loci in each family. To do so, we compared each ORF that had a length of at least 80 aa with the National Center for Biotechnology Information online nonredundant protein database using BLASTP. To locate LTRs, we used the webtool LTR_FINDER (13). We also confirmed the presence of *env* by visually inspecting all loci that were suggested by our automated procedures to have an *env*-like ORF and then using the nonredundant protein database as described above. The only discrepancies we found with our automated search were the rare occasions when more than one ERV locus was included in a larger fragment (hence the occasional single-figure *env* values in Table S1 that result from inclusion of *env* from a nearby ERV locus belonging to another family).

Identifying Cross-Species Transmissions and Invasions. We estimated the history of cross-species transmissions by (*i*) collapsing all branches in the tree shown in Fig. 2 where the sister node was in the same host and (*ii*) modeling host species as a single multistate discrete character on the resulting tree (Fig. 3) and reconstructing ancestral states at the nodes using maximum parsimony implemented in Mesquite. We define an invasion as each terminal branch in the resulting tree, giving a total of 38, and a cross-species transmission node as one that has a character state different from that of the node immediately below it closer to the root, giving a total of 18. The number of invasions is the most conservative estimate and lies at the lower boundary of the real number, because, in some instances, sister nodes in the same host are separated by long branches that probably represent independent invasions by related viruses; however, we could not find an unbiased criterion for using branch lengths to define invasions.

Quantifying Distance from Cross-Species Transmissions. We used each of the inferred cross-species transmission nodes as a root of a subtree and reestimated the evolutionary distinctiveness (ED) of the loci in this subtree as previously described. We define the maximum ED here, called "ED_{cst}," as a measure of the distance from the closest inferred cross-species transmission: The larger the ED_{cst}, the closer the element is to an inferred cross-species transmission node. We found that ED and ED_{cst} are strongly correlated (Fig. S4), reflecting the fact that most cross-species transmissions occurred near the root of the IAP tree.

Correlating Gene Integrity with ED and ED_{cst}. We also addressed the following two points in our generalized least squares (GLS) model.

i) We account for the phylogenetic relatedness of the traits in the regression of ED against gene integrity using Pagel's λ. This parameter reflects the degree to which traits are correlated to phylogenetic relatedness and can be set to values between 0, where the phylogeny is ignored, to 1, where the analyses is fully adjusted to take phylogenetic relatedness into account. The parameter takes into account nonindependence of the data caused by phylogenetic relatedness (14) and is an extension of the phylogenetic comparative method (15) as proposed by Pagel (16) through implementation of the established GLS methodology. The estimation of the variance-covariance matrix of the traits was performed assuming a Brownian motion model of evolution of traits across the phylogenetic tree.

ii) A second problem is that the phylogenetic GLS model assumes that the traits evolve uniformly across the tree, e.g., that genes degrade steadily from the root of the tree toward the tips. However, loss of gene integrity should prevent viral replication, and thus we expect it to occur only at the terminal branches of the tree, which represent time after integration into the host genome. The difference in gene degradation that occurs on internal branches compared with terminal branches has been demonstrated in one human ERV family (17). Therefore, it is necessary to import a transformation for the rate of degradation to model realistically the fact that degradation is much faster at the postintegration time. Several parameters have been used to account for traits' rate diversity across the tree (18); all these parameters transform the branch lengths of the tree to fit better the expected model of trait evolution. We used the APE package in R, applying a multiplicative parameter, *t*, to transform the terminal branch lengths and allow a faster rate of gene degradation on the terminal branches of the tree. Other, more realistic ways to model the gene disintegration in our dataset are possible, e.g., by using a third rate parameter that is specific for the expansions in each host. However, we suggest that our parameterization provides a simple and robust model for our dataset and that a more realistic and more parameterized model would not change the significance of our results.

We used a range of different values for each of the parameters *t* and λ and selected the best-fit model by means of the Akaike Information Criterion (AIC) (19), which is a metric of model fitness.

The ED has a strongly skewed distribution and so does not fit well as a dependent variable in our linear multivariate model. Although the assumptions of normality typically lie at the residuals and not the dependent variable itself, strongly skewed distributions of dependent variables are the most probable reason for the bad linear fit of the overall model. Therefore, we used the logarithm to base 10 of ED, which provides a symmetric distribution for all genes except *env*. Because the *env* gene of most loci was highly degraded, the distribution of its integrity measure (length of longest ORF) was strongly skewed, many loci having zero values. A logarithmic transformation of *env* length does not result in a symmetric distribution, so we modeled it as a binomial variable applying a breakpoint at 600 nt (1: >600 nt; 0: ≤600 nt). To assess whether the transformations affected the significance of the results, we also performed the regression using the non-transformed values. The significance of the parameters was the same, proving that the model was robust even under a strongly skewed parameterization; however, the overall fit of the linear model was much worse because of the skewed distributions of the ED and *env*. We estimated the correlation between ED_{cst} and integrity of the genes using the same approach.

To assess the robustness of the ED metric to phylogenetic uncertainty, we estimated the ED for 100 bootstrap replicates and compared this estimate with the ED measured from the original alignment with linear regression (Fig. S9). The high Pearson's coefficient (0.83, $P < 0.01$) suggests that ED used in the analyses is robust to phylogenetic uncertainty.

Recombination Analysis of *env* in IAPs. The IAPE *env* gene is known to be very divergent from those of extant retroviruses (20), and we found that even in the more conserved transmembrane region there was <20% amino acid identity to the closest extant

XRV, the betaretrovirus Jaagsiekte sheep retrovirus. To detect possible recombination events that have caused a change in the *env* gene among our IAPs, we compared pairwise similarity scores with our XRV protein libraries to find examples where loci had a low *env* match to the virus in the library to which they had the best *pol* match. We therefore made a library of *env* amino acid sequences from all XRV species plus ERVs that lacked close XRV relatives, including IAPE (2, 10, 20). We then screened all potentially full-length ORFs of our loci with our *env* library and built a matrix containing PBLAST bit scores. The loci were classified according to the library member that had the closest match. We found that only the transmembrane domain of the IAPE *env* gene has a significant similarity with any other *env* genes in both our library and the nonredundant sequence database. However, in this transmembrane domain there is only a short region that can be aligned among all of the different clades of IAP, and it does not contain enough information to infer recombination through a phylogenetic approach. However, the results obtained from our classification as IAPE vs. non-IAPE were striking and strongly supportive of recombination.

SI Results

Degradation of *env* is most marked in the large (>200 loci) expansions, and a pattern of gradual loss of *env* in the large expansion in *Mus* is suggested because *env* is less degraded at the basal terminal branches (Fig. 2 and Fig. S7). However, the small expansions have widely varying levels of *env* integrity, as perhaps would be expected, given that they represent small samples. To assess statistically the relationship between *env* integrity and both expansion and cross-species transmission, we performed a multivariate analysis based on GLS accounting for phylogenetic correlation and changes in rate between internal and terminal branches. The AIC analysis showed that the best-fit model was achieved by setting $\lambda=1$ (Table S2) and $t = 30$ (Table S3), i.e., where the phylogeny is taken into account fully and the rate of gene degradation is 30 times faster at the terminal branches than at the internal ones. Although our interest is in *env,* our model takes into account the integrity of all genes to control for possible confounding effects. The analysis showed that expansion, as measured by ED, is not significantly correlated with integrity of *gag*, *prot*, and *pol,* whereas for *env*'s integrity the correlation was negative (Table S3). Thus, our best-fit model suggests that expansion of the IAPs is accompanied by *env* degradation.

This degradation tends to occur after cross-species transmission. At least 18 cross-species transmission events have occurred in the evolutionary history of the IAPs (Fig. 3). They tend to be close to the midpoint root of the tree, consistent with the expansions occurring after the speciation of the hosts (also reflected in the high correlation between ED and ED_{cst}). After selecting the best-fit model in the same way as before, we found that the distance of elements from the closest cross-species transmission event, ED_{cst}, was inversely associated with the integrity of the *env* and was not associated with the integrity of the *prot*, *pol*, and *gag* genes. The behavior of $ED_{cst}$was very close to that of ED (e.g., Table S2), and the best-fit model was the same ($\lambda = 1, t = 30$). Thus elements with more intact *env* gene tend to be closer to the inferred cross-species transmission events. The cessation of cross-species transmission after the loss of *env* also is shown by the fact that we were not able to find any cross-species transmissions nested within the large expansions where *env* apparently was nonfunctional.

In our analysis of all ERV families, we were able to confirm the absence of *env* in one of the class II retrotransposing megafamilies in *Ochotona,* e.g., finding a complete element with only 880 nt of no detectable homology between the end of *pol* and the start of the 3′ LTR. Retroviruses typically have the 3′ UTR here, but the 3′ UTR usually is much shorter, especially in simple retroviruses (~30 nt), so much of the 880 bases probably represents vestigial *env*. This megafamily is nested within a tree of reinfecting ERVs and XRVs (Fig. S5), and it is more parsimonious to infer that it lost its *env*. Our ERV-L families (i.e., families that form a monophyletic clade containing HERV-L and MuERV-L) do not appear to have any remnant of an *env* gene (21), but these families are all very old, and we cannot determine if they lost *env* a long time ago or were primitively *env*-less. The HERV-H megafamily is dominated by largely *env*-less loci but also has a smaller number of loci with *env*, which tend to be more basal in the phylogenetic tree (22, 23), consistent with the pattern of gradual *env* loss that we see in the IAPs (but see *Discussion* in the main text).

1. Katzourakis A, Gifford RJ (2010) Endogenous viral elements in animal genomes. *PLoS Genet* 6:e1001191.
2. Ribet D, et al. (2008) An infectious progenitor for the murine IAP retrotransposon: Emergence of an intracellular genetic parasite from an ancient retrovirus. *Genome Res* 18:597–609.
3. Belshaw R, Katzourakis A (2005) BlastAlign: A program that uses blast to align problematic nucleotide sequences. *Bioinformatics* 21:122–123.
4. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680.
5. Price MN, Dehal PS, Arkin AP (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490.
6. Bininda-Emonds OR, et al. (2007) The delayed rise of present-day mammals. *Nature* 446:507–512.
7. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.
8. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* 16:276–277.
9. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.
10. Belshaw R, de Oliveira T, Markowitz S, Rambaut A (2009) The RNA virus database. *Nucleic Acids Res* 37(Database issue):D431–D435.
11. Bénit L, et al. (1997) Cloning of a new murine endogenous retrovirus, MuERV-L, with strong similarity to the human HERV-L element and with a gag coding sequence closely related to the Fv1 restriction gene. *J Virol* 71:5652–5657.
12. Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* 9:18.
13. Xu Z, Wang H (2007) LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* 35(Web Server issue):W265-8.
14. Harvey PH, Pagel MD (1991) *The Comparative Method in Evolutionary Biology* (Oxford Univ Press, Oxford, UK).
15. Felsenstein J (1985) Phylogenies and the comparative method. *Am Nat* 125:1–15.
16. Pagel M (1994) Detecting Correlated evolution on phylogenies - a general-method for the comparative-analysis of discrete characters. *Proc R Soc Lond B Biol Sci* 255:37–45.
17. Belshaw R, et al. (2004) Long-term reinfection of the human genome by endogenous retroviruses. *Proc Natl Acad Sci USA* 101:4894–4899.
18. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20:289–290.
19. Akaike H (1974) A new look at the statistical model identification. *IEEE Trans Automat Contr* 19:716–723.
20. Bénit L, Dessen P, Heidmann T (2001) Identification, phylogeny, and evolution of retroviral elements based on their envelope genes. *J Virol* 75:11709–11719.
21. Bénit L, Lallemand JB, Casella JF, Philippe H, Heidmann T (1999) ERV-L elements: A family of endogenous retrovirus-like elements active throughout the evolution of mammals. *J Virol* 73:3301–3308.
22. Belshaw R, Katzourakis A, Paces J, Burt A, Tristem M (2005) High copy number in human endogenous retrovirus families is associated with copying mechanisms in addition to reinfection. *Mol Biol Evol* 22:814–817.
23. Jern P, Sperber GO, Blomberg J (2004) Definition and variation of human endogenous retrovirus H. *Virology* 327:93–110.

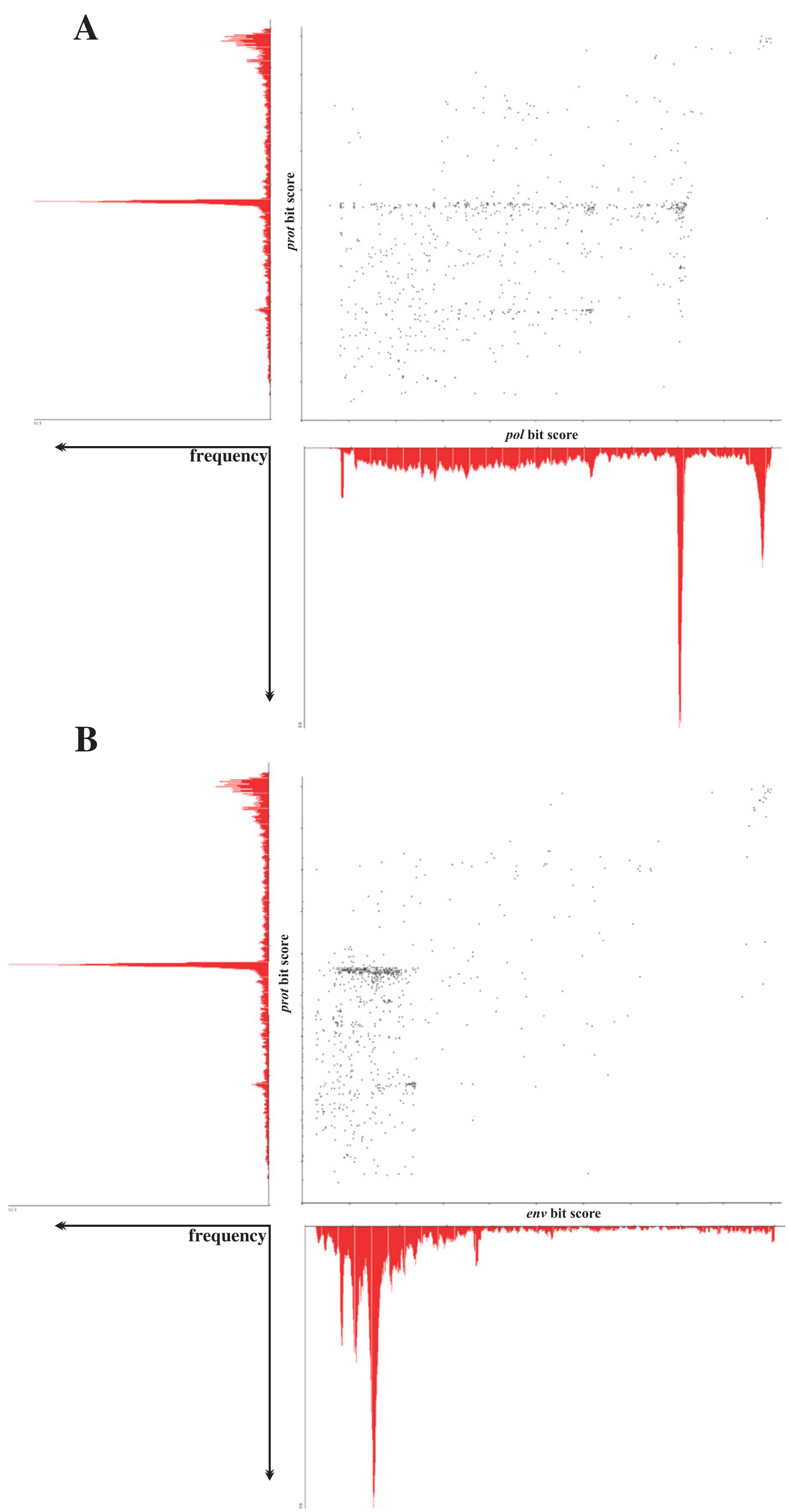

Fig. S1. Scatterplots of the TBLASTN bit scores associated with axes-specific histograms from the *Mus* IAP elements for *prot* against (*A*) the *pol* genes and (*B*) the *env* genes (*gag* is similar to *pol*). The striking observation is that the *env* scores, unlike those of the other genes, are strongly skewed toward the left-hand side of the horizontal axes with spikes (clusters) occurring only at a very low percentage of integrity (<1/3 of the *env* bit score).

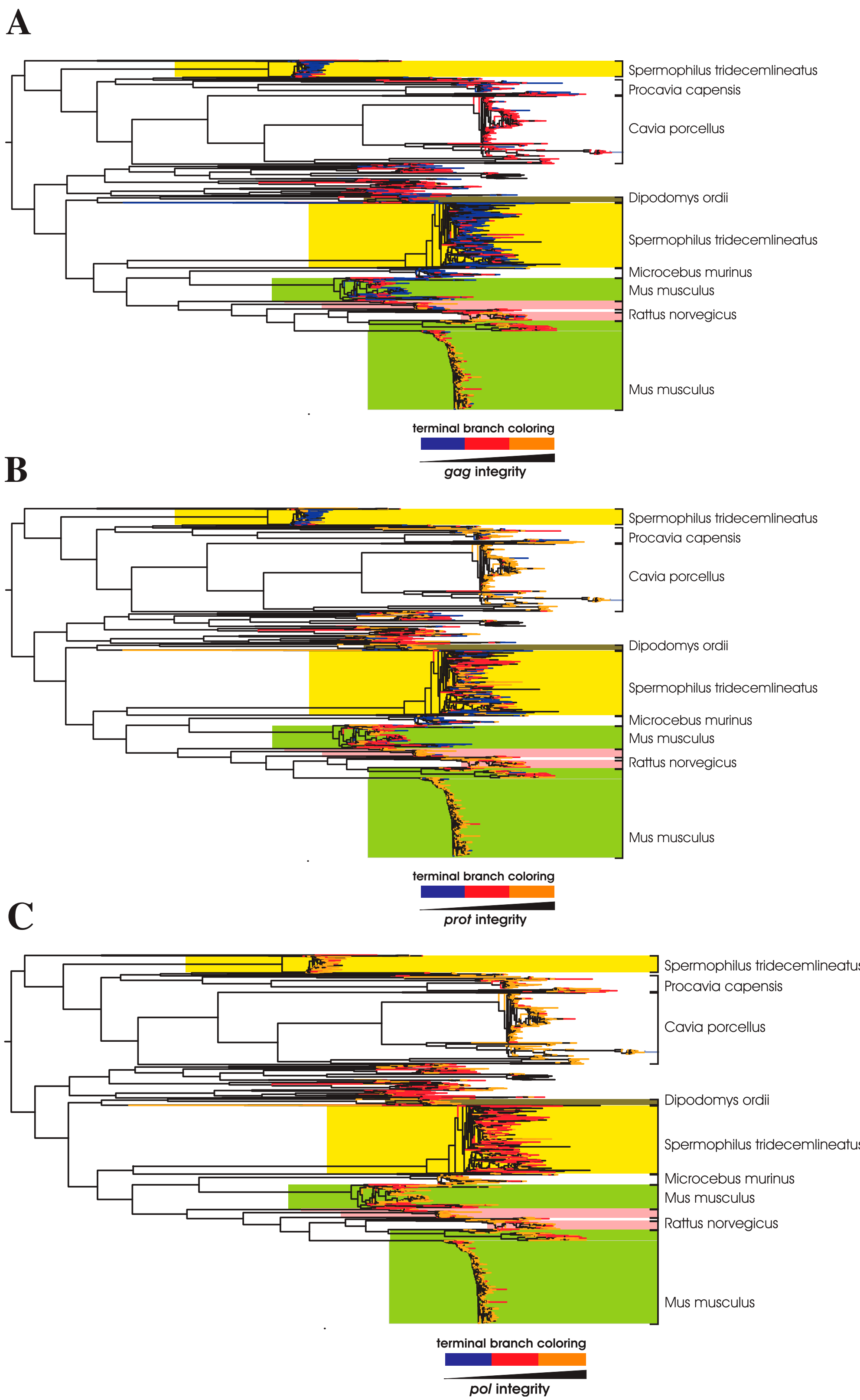

Fig. S2. Distribution of gene integrity on the IAP tree shown in Fig. 2 and described in the legend of Fig. 2. (*A*) *gag*. (*B*) *prot*. (*C*) *pol*.

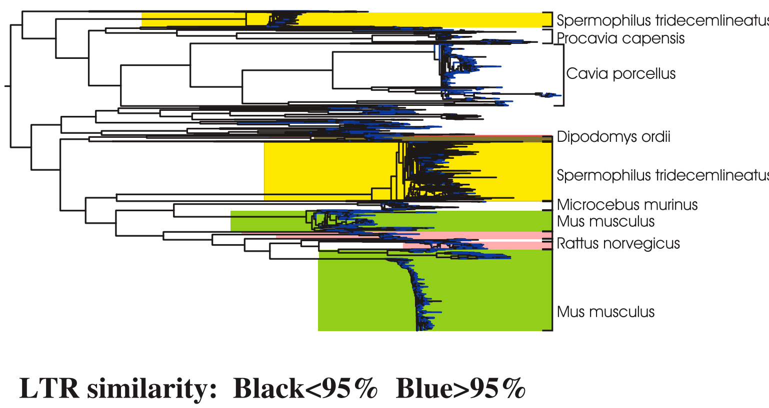

Fig. S3. Distribution of LTR similarity on the IAP tree shown in Fig. 2. Blue shows elements with more-similar LTRs (≥ 95% similarity). Black shows elements with less-similar LTRs.

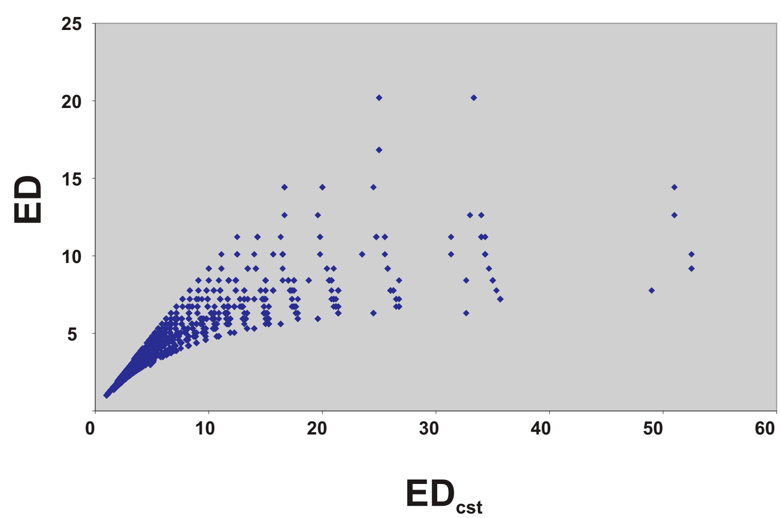

Fig. S4. Scatterplot of ED against ED_{cst} showing high correlation between the two values.

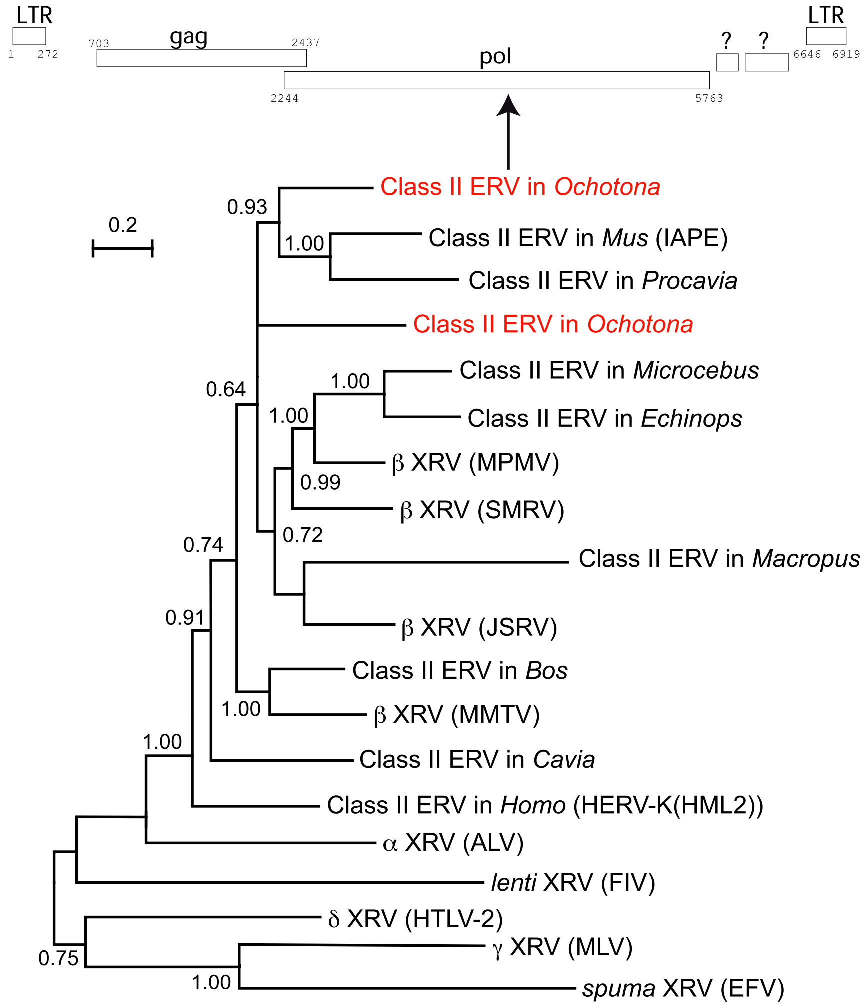

Fig. S5. Phylogenetic tree of *pol* sequences from analyzed class II ERVs plus (*i*) extant betaretroviruses [mouse mammary tumor virus (MMTV), Jaagsiekte sheep retrovirus (JSRV), squirrel monkey retrovirus (SMRV), and Mason-Pfizer monkey virus (MPMV)], (*ii*) representatives of the other main XRV clades [equine foamy virus (EFV), murine leukemia virus (MLV), human T-cell leukemia virus type 2 (HTLV-2), feline immunodeficiency virus (FIV), and avian leukosis virus (ALV)], and (*iii*) two published ERVs: IAPE (1) and HERV-K(HML2)] (2). We were unable to recover a good *pol* sequence from the class II ERV family in *Dasypus*. All viruses included have *env* except for the two *env*-less class II ERV megafamilies in *Ochotona* shown in red. The schematic at the top if the figure shows the LTRs and ORFs in a single provirus belonging to one of these families; the sequence is available at our RNA virus database as PikaDtype-1 (3).

1. Ribet D, et al. (2008) An infectious progenitor for the murine IAP retrotransposon: emergence of an intracellular genetic parasite from an ancient retrovirus. *Genome Res* 18:597–609.
2. Dewannieux M, et al. (2006) Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. *Genome Res* 16:1548–1556.
3. Belshaw, et al. (2009) The RNA Virus Database. *Nucleic Acids Res* 37:D431–D435.

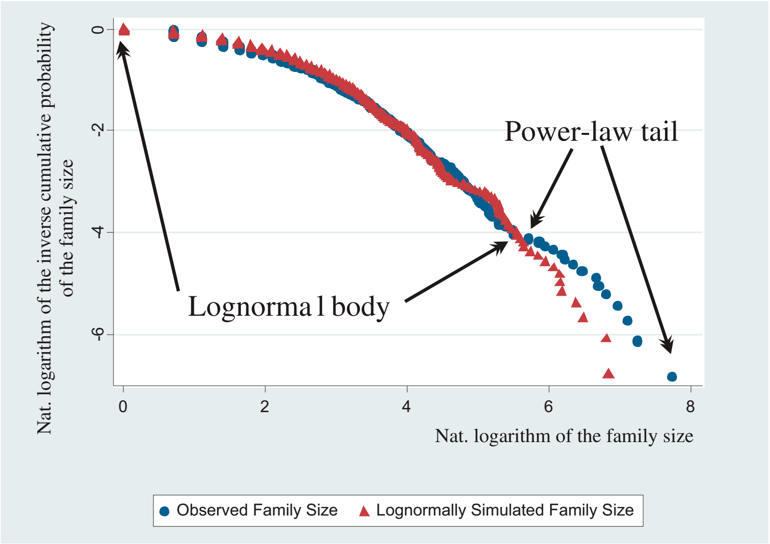

Fig. S6. Overlaid scatter plots of the logarithmically transformed inverse cumulative distributions (vertical axis) vs. their logarithmically transformed family sizes (horizontal axis) for the observed family sizes (blue circles) and a single lognormally simulated one (red triangles).

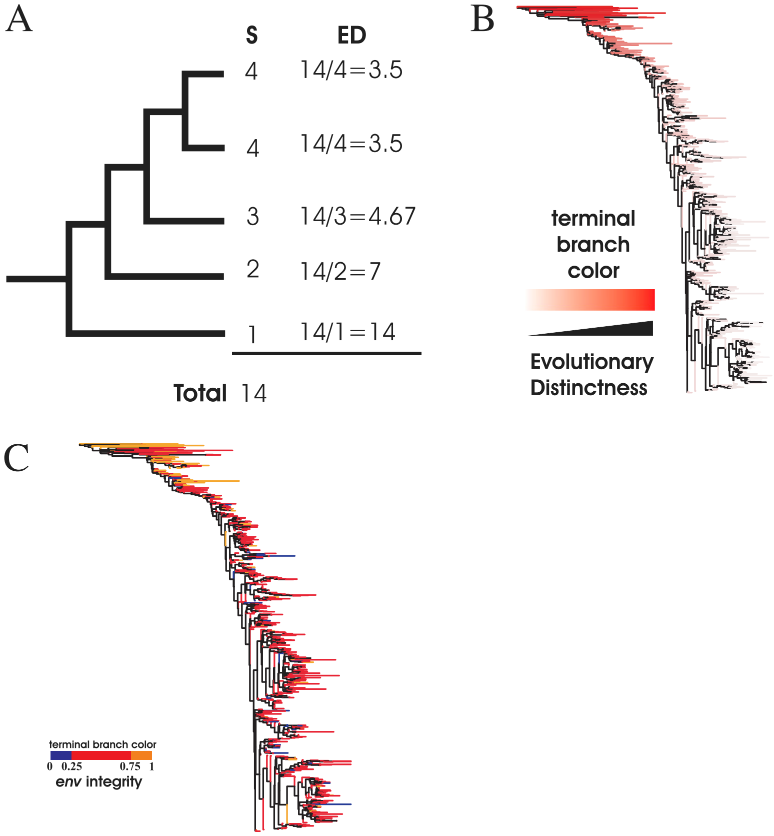

Fig. S7. The ED and *env* integrity metrics. (*A*) A five-taxon tree as an example of calculating ED. The S column shows the number of nodes from the root, and the ED column shows the calculation of ED. Taxa that result from more replication events (larger expansion) have lower ED. (*B* and *C*) ED and *env* integrity values in the largest *Mus* IAP expansion (from bottom left of Fig. 2); ED is highest where the color of the terminal branches is dark red. Conventions for showing *env* integrity are as in Fig. 2.

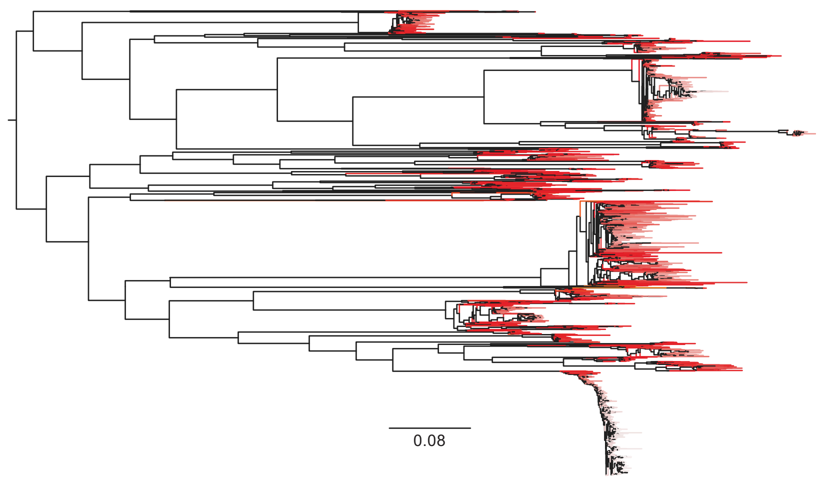

Fig. S8. Distribution of ED on the IAP tree shown in Fig. 2. Intensity of red shading is proportional to ED value. Smaller clades and the basal loci in larger clades tend to be darker, with higher ED values showing a less abundant replication history.

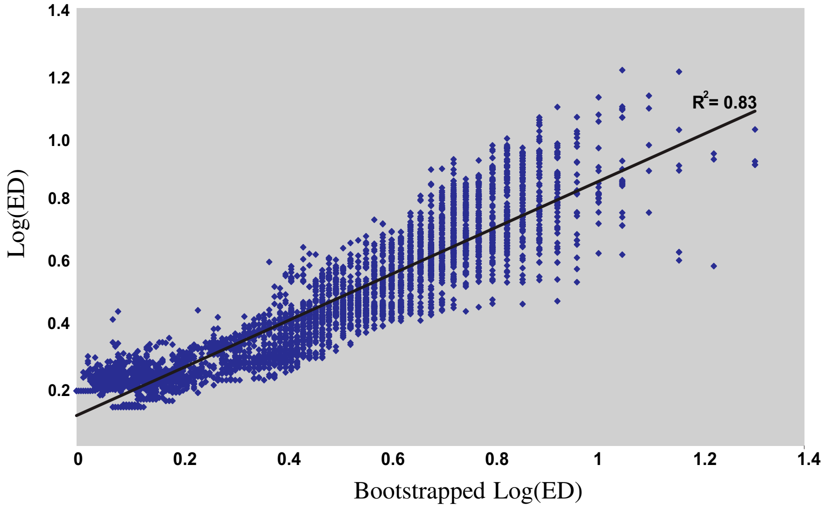

Fig. S9. Scatterplot of the logarithm to base 10 of the ED [log(ED)] estimated from the original alignment against the respective values from 100 bootstrapped pseudo replicates [bootstrapped Log(ED)]. The regression line and the Pearson coefficient are shown also.

Table S1. Summary of ERVs in the mammal genomes

Host species	Common name	Summary: Total no. of loci in genome	Summary: No. of young loci	Summary: No. of families among young loci	Megafamilies: Size = no. of young loci (mean divergence; type)	Megafamilies: Longest *env* ORF (longest *gag* ORF)	Small family: Size = no. of young loci (mean divergence; type)	Small family: Longest *env* ORF (*gag*)	IAPs: No. of IAP loci	IAPs: No. of IAP invasions
Ailuropoda melanoleuca	Giant panda	558	34	6	None				0	
Bos taurus	Cow	1,357	619	22	174 (87%; class I)	94 (237)	19 (84%; class II)	277 (391)	0	
					138 (89% class I)	9 (159)	43 (90%; class II)	225 (207)		
Callithrix jacchus	Common marmoset	2,156	421	54	None				0	
Canis familiaris	Domestic dog	536	92	9	57 (83%; class I)	63 (213)	10 (87%; class I)	181 (380)	0	
Cavia porcellus	Guinea pig	5,057	2,349	23	629 (95%; IAP)	7 (439)	49 (93%; class II)	419 (421)	834	2
Choloepus hoffmanni	Hoffmann's two-toed sloth	2,989	1,713	14	1,037 (81%; ERV-L)	1 (83)	19 (80%; class I)	101 (114)	14	1
Dasypus novemcinctus	Nine-banded armadillo	5,430	3,032	97	768 (82%; ERV-L)	2 (115)	177 (85%; class II)	44 (130)	0	
Dipodomys ordii	Ord's kangaroo rat	684	423	26	106 (88%; class II)	33 (206)	6 (82%; class I)	341 (213)	90	3
					91 (94%; IAP)	80 (314)	8 (85%; class II)	110 (108)		
Echinops telfairi	Small Madagascar hedgehog tenrec	1,613	733	13	557 (82%; ERV-L)	0 (100)	17 (75%; class II)	57 (102)	3	1
Equus caballus	Horse	1,133	133	25	None				0	
Erinaceus europaeus	West European hedgehog	3,035	2,789	11	2,251 (87%; class II)	5 (159)	19 (83%; class I)	38 (82)	0	
Felis catus	Domestic cat	681	304	24	111 (93%; class I)	37 (240)	10 (91%; class I)	131 (247)	0	
Gorilla gorilla	Western gorilla	2,228	NA	NA	NA	NA	NA	NA	0	
Homo sapiens	Human	3,809	1,735	17	879 (79%; HERV-H)	30 (99)	39 (74%; HERV-XA)	62 (115)	0	
Loxodonta africana	African bush elephant	3,805	656	29	494 (80%; ERV-L)	2 (98)	24 (76%; class I)	16 (43)	0	
Macaca mulatta	Rhesus macaque	2,986	1,129	18	None				0	
Macropus eugenii	Tammar wallaby	1,227	265	21	146 (82%; class I)	1 (117)	15 (84%; class II)	134 (265)	0	
Microcebus murinus	Gray mouse lemur	1,609	982	34	344 (88%; class II)	189 (264)	36 (88%; class I)	186 (106)	111	1
					195 (90%; IAP)	80 (142)	18 (91%; class I)	118 (137)		
Monodelphis domestica	Opossum	7,440	3,666	16	2,986 (77%; class I)	4 (186)	22 (76%; class I)	25 (115)	0	
Mus musculus	House mouse	5,749	4,334	24	1,188 (97%; IAP)	152 (685)	229 (90%; class I)	455 (471)	1533	13
					799 (95%; ERV-L)	11 (487)	61 (92%; class I)	527 (476)		
Myotis lucifugus	Little brown bat	820	435	46	None				0	
Ochotona princeps	American pika	678	447	12	117 (89%; class II)	3 (417)	41 (73%; IAP)	93 (239)	14	2
					111 (89%; class II)	1 (442)	41 (78%; class II)	177 (257)		
Ornithorhynchus anatinus	Duck-billed platypus	291	219	28	None				0	
Oryctolagus cuniculus	European rabbit	994	483	37	None				17	1
Otolemur garnettii	Northern greater galago	1,147	424	31	None				0	
Pan troglodytes	Chimpanzee	3,025	NA	NA	NA	NA	NA	NA	0	
Papio hamadryas	Hamadryas baboon	2,619	NA	NA	NA	NA	NA	NA	0	
Pongo pygmaeus	Bornean orangutan	3,508	NA	NA	NA	NA	NA	NA	0	
Procavia capensis	Cape hyrax	3,187	1,841	16	476 (82%; class I)	60 (154)	33 (78%; class I)	138 (183)	46	?
					372 (83%; ERV-L)	3 (148)	61 (85%; class II)	241(346)		
Pteropus vampyrus	Large flying fox	849	233	26	None				0	
Rattus rattus	Black rat	2,879	NA	NA	NA	NA	NA	NA	339	7
Sorex araneus	Common shrew	970	801	15	None				19	2
Spermophilus tridecemlineatus	Thirteen-lined ground squirrel	2,330	1,790	42	1,359 (87%; IAP)	30 (280)	16 (86%; class I)	89 (105)	970	2

Table S1. Cont.

		Summary			Megafamilies		Small family		IAPs	
Host species	Common name	Total no. of loci in genome	No. of young loci	No. of families among young loci	Size = no. of young loci (mean divergence; type)	Longest *env* ORF (longest *gag* ORF)	Size = no. of young loci (mean divergence; type)	Longest *env* ORF (*gag*)	No. of IAP loci	No. of IAP invasions
Sus scrofa	Domestic pig	226	32	5	None				0	
Tarsius syrichta	Philippine tarsier	3,676	2,586	88	None				7	1
Tupaia belangeri	Northern treeshrew	913	470	26	None				36	1
Tursiops truncatus	Bottlenosed dolphin	802	77	24	None				1	1
Vicugna pacos	Alpaca	618	154	21	None				0	

Except for the total number, only more recently integrated loci are included (*Materials and Methods*). For each genome we show the mean integrity of *env* (as the number of amino acids in the longest ORF) in any megafamily present plus the mean integrity of *env* in a representative small family of similar age. The corresponding *gag* integrity is given in parentheses in each case. The IAP loci included (n = 4,089) are those that form a monophyletic clade with respect to the other ERV and XRV reference sequences and that have length of at least 600 nt in the final *pol* alignment. The IAP invasions were inferred only using the nearly full-length elements (n = 3,834) (*Materials and Methods*), and question marks show instances where shorter fragments of IAPs were detected. NA, not applicable (treated as duplicates in the analysis of ERV family sizes).

Table S2. Multivariate GLS regression of ED and ED_{cst} against gene, accounting for different levels of phylogenetic dependence (Pagel's λ)

	Pagel's λ											
Gene	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	Parameter
gag	–	–	–	–	–	–	–	–	–	–	–	ED
	–	–	–	–	–	–	–	–	–	–	0	ED_{cst}
prot	–	0	+	+	+	+	+	+	+	+	+	ED
	–	0	+	+	+	+	+	+	+	+	+	ED_{cst}
pol	–	–	–	–	–	–	–	–	–	–	–	ED
	–	–	–	–	–	–	–	–	–	–	–	ED_{cst}
env	+	+	+	+	+	+	+	+	0	0	0	ED
	+	+	+	+	+	+	+	+	+	+	0	Ed_{cst}

Minus and plus symbols show a significant ($P < 0.05$) negative or positive relationship, respectively, and zero (0) shows a nonsignificant relationship. The rate of degradation was uniform across the tree ($t = 1$).

Table S3. Multivariate GLS regression of ED against gene integrity with differing values for the multiplying factor (*t*) applied to the terminal branches

Terminal branch multiplicative rate parameter (*t*)	*env*	*gag*	*prot*	*pol*	AIC
1	0	–	+	–	−17066.9
2	0	0	0	0	−15840.7
3	0	0	0	0	−15113.8
5	0	0	0	0	−14203.5
10	0	0	0	0	−12976.7
20	+	0	+	0	−19338.3
30	**+**	**0**	**0**	**0**	**−24384.8**
40	+	0	0	0	−22495.5
50	+	0	0	0	−18576.1
60	+	0	0	0	−21789.6
70	+	0	0	0	−18576.1
80	+	0	0	0	−19249.8
90	+	0	0	0	−19186.9
100	+	0	0	0	−19787.3

Minus and plus symbols show significant ($P < 0.05$) negative (–) and positive (+) relationship respectively, and zero (0) shows a non significant relationship. Pagel's λ is fixed at 1, which is the best-fitting value. The best-fit model (lowest AIC) is shown in bold.

Table S4. Degradation of *env* in megafamilies compared with that in all other loci in the same genome

Host species	ERV family	No. of loci analyzed	Mean longest *env* ORF	Mean longest *gag* ORF	*env/gag* ratio
Bos taurus	Class I megafamily	174	94	237	0.40
	Class I megafamily	138	9	159	0.06
	All nonmegafamilies	312	139	159	0.88
Canis familiaris	Class I megafamily	57	63	213	0.30
	All nonmegafamilies	33	131	186	0.70
Cavia porcellus	IAP megafamily	629	7	439	0.02
	All nonmegafamilies	1,500	138	226	0.61
Choloepus hoffmanni	ERV-L megafamily	1,037	1	83	0.01
	All nonmegafamilies	676	49	123	0.40
Dasypus novemcinctus	ERV-L megafamily	768	2	115	0.02
	All nonmegafamilies	2,305	46	147	0.31
Dipodomys ordii	Class II megafamily	106	33	206	0.16
	IAP megafamily	91	80	314	0.25
	All nonmegafamilies	226	58	151	0.39
Echinops telfairi	ERV-L megafamily	557	0	100	0.00
	All nonmegafamilies	176	55	95	0.58
Erinaceus europaeus	Class II megafamily	2,251	5	159	0.03
	All nonmegafamilies	161	102	147	0.70
Felis catus	Class I megafamily	111	37	240	0.15
	All nonmegafamilies	193	86.09	158	0.54
Homo sapiens	Class I megafamily	879	30	99	0.30
	All nonmegafamilies	794	123	174	0.71
Loxodonta africana	ERV-L megafamily	494	2	98	0.02
	All nonmegafamilies	526	51	77	0.66
Macropus eugenii	Class I megafamily	146	1	117	0.01
	All nonmegafamilies	121	36	100	0.36
Microcebus murinus	Class II megafamily	344	189	264	0.72
	IAP megafamily	195	80	142	0.56
	All nonmegafamilies	501	75	109	0.68
Monodelphis domestica	Class I megafamily	2,986	4	186	0.02
	All nonmegafamilies	679	132	255	0.52
Mus musculus	IAP megafamily	1,188	152	685	0.22
	ERV-L megafamily	799	11	484	0.02
	All nonmegafamilies	1,675	204	266	0.77
Ochotona princeps	Class II megafamily	117	3	417	0.01
	Class II megafamily	111	1	442	0.00
	All nonmegafamilies	219	121	160	0.76
Procavia capensis	Class I megafamily	476	60	154	0.39
	ERV-L megafamily	372	3	148	0.02
	All nonmegafamilies	993	127	230	0.55
Spermophilus tridecemlineatus	IAP megafamily	1,359	30	280	0.11
	All nonmegafamilies	1,056	76	149	0.51

To take differences in ages into account, this degradation is shown by the ratio of the mean longest ORFs (number of amino acids) in *env* compared with *gag* (*gag* is essential for replication and hence will decay over time after integration). Older loci are excluded as described in the text, except for the non-megafamilies in *Erinaceus* and *Loxodonta.*