Env-less endogenous retroviruses are genomic superspreaders

Gkikas Magiorkinis[a], Robert J. Gifford[b,1], Aris Katzourakis[a,1], Joris De Ranter[c], and Robert Belshaw[a,2]

[a]Department of Zoology, University of Oxford, Oxford OX1 3PS, United Kingdom; [b]The Aaron Diamond AIDS Research Center, New York, NY 10016; and [c]Clinical and Epidemiological Virology, Rega Institute, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium

Edited by Stephen P. Goff, Columbia University College of Physicians and Surgeons, New York, NY, and approved March 30, 2012 (received for review January 19, 2012)

Endogenous retroviruses (ERVs) differ from typical retroviruses in being inherited through the host germline and therefore are a unique combination of pathogen and selfish genetic element. Some ERV lineages proliferate by infecting germline cells, as do typical retroviruses, whereas others lack the *env* gene required for virions to enter cells and thus behave like retrotransposons. We wished to know what factors determined the relative abundance of different ERV lineages, so we analyzed ERV loci recovered from 38 mammal genomes by *in silico* screening. By modeling the relationship between proliferation and replication mechanism in detail within one group, the intracisternal A-type particles (IAPs), and performing simple correlations across all ERV lineages, we show that when ERVs lose the *env* gene their proliferation within that genome is boosted by a factor of ∼30. We also show that ERV abundance follows the Pareto principle or 20/80 rule, with ∼20% of lineages containing 80% of the loci. This rule is observed in many biological systems, including infectious disease epidemics, where commonly ∼20% of the infected individuals are responsible for 80% of onward infection. We thus borrow simple epidemiological and ecological models and show that retrotransposition and loss of *env* is the trait that leads endogenous retroviruses to becoming genomic superspreaders that take over a significant proportion of their host's genome.

Endogenous retroviruses (ERVs) proliferate by the repeated integration of new viral sequences into their host's germline (1), integrations which can become fixed in the host population and have led to ERV sequences (loci) comprising 8–10% of the human and mouse genomes (2, 3) (this number also includes nonautonomous LTR-retrotransposons, which we do not analyze here). These loci form phylogenetically distinct lineages traditionally called "families" (4) (unrelated to the general use of this term in taxonomy), each of which is the result of the expansion of a founder infection of the organism's germline that can have occurred more than ∼100 million years ago (5).

ERVs can replicate both as transposable elements (TEs) and viruses. Some lineages copy by an entirely intracellular mechanism and are functionally indistinguishable from the class of TEs called LTR-retrotransposons, whereas others copy within the host germline using cell reinfection in the same manner as the copying within somatic cells of exogenous retroviruses (XRVs) (6). We refer to these replication mechanisms as "retrotransposition" and "reinfection," respectively. Whether an ERV is reinfecting or retrotransposing can be determined by the integrity of its *env* gene, which produces the protein on the surface of the viral particle that is responsible for cell entry. We can assume that an ERV lineage with a functional *env* is reinfecting, whereas an ERV lineage with a disintegrated *env* is retrotransposing (whether reinfection can include germline cells in other host individuals of the same or other species is not known). Some retroviruses with a defective *env* are able to reinfect by "hitchhiking" the functional *env* of a coinfecting retrovirus, a mechanism known as "complementation" (7). However, complementation does not appear to be common in ERVs; in two ERV families where complementation of *env* might be expected to occur, because they contain both loci with intact *env* and loci with defective *env*, it has been shown that the former are reinfecting, and the latter are retrotransposing (8–10).

The relationship between an organism and its TEs poses a series of questions similar to those in ecology. For example, workers have attempted to explain the proliferation of individual TE lineages and why the genomes of more complex organisms tend to contain more TEs than do simpler ones (11, 12). We take an approach common in community ecology and ask what controls the relative abundance of different TE lineages. Our previous work (10) suggested that reinfecting lineages, inferred from detecting past negative selection on *env*, tended to be small, but this study was restricted to the human genome and did not account for a possible confounding effect of lineage age. Here we (*i*) model in detail the relationship between *env* integrity and proliferation in the intracisternal A-type particle (IAP) group of ERVs and (*ii*) compare in 38 mammal genomes the mean *env* integrity of the largest ERV lineage with the *env* integrity of the smaller lineages. IAPs are a good model system because they invaded their hosts recently, are well-studied experimentally, and harbor both mechanisms of replication. They were found initially in the mouse and were shown by electron microscopy to replicate via intracellular particles which budded on the cisternae of the endoplasmic reticulum, hence their name (13, 14). These retrotransposing loci have a degraded, nonfunctional *env* gene (15). Later, however, similar loci with more intact *env* genes, IAPEs, were identified in the mouse, and one was shown experimentally to be able to reinfect cells in the classic viral manner (9, 16).

We find repeated transformations from reinfecting into retrotransposing ERVs and show that this transformation results in a rapid proliferation within the genome. Considering our results together with those from studies of transmission diversity in infectious disease epidemics, we propose that retrotransposition is the trait that leads ERVs to become genomic superspreaders.

Results

Distribution of IAPs in Hosts. We found 5,969 IAP loci in 17 host genomes (Figs. 1 and 2 and Table S1). These loci formed a monophyletic clade within a tree containing all XRV species and representatives of other ERV families. The IAP loci were found mostly in rodents: Three species—*Mus*, *Spermophilus*, and *Cavia*—account for more than 80% of the loci. In addition, every sequenced rodent, as well as both representatives of the sister order Lagomorpha, has been invaded by IAPs. Among the equally well-sampled primates, IAPs were found only only in the more basal lineages represented by *Tarsius* and *Microcebus*; no IAP was found in monkeys and apes. Mapping host species as a character onto the IAP tree, we estimate a total of at least 18 cross-species transmission events among the IAPs (Fig. 3). Mouse and rat IAP

EVOLUTION

Author contributions: G.M., R.J.G., A.K., and R.B. designed research; G.M., R.J.G., A.K., J.D.R., and R.B. performed research; G.M., R.J.G., A.K., J.D.R., and R.B. analyzed data; and G.M., R.J.G., A.K., and R.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]R.J.G. and A.K. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: robert.belshaw@zoo.ox.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1200913109/-/DCSupplemental.

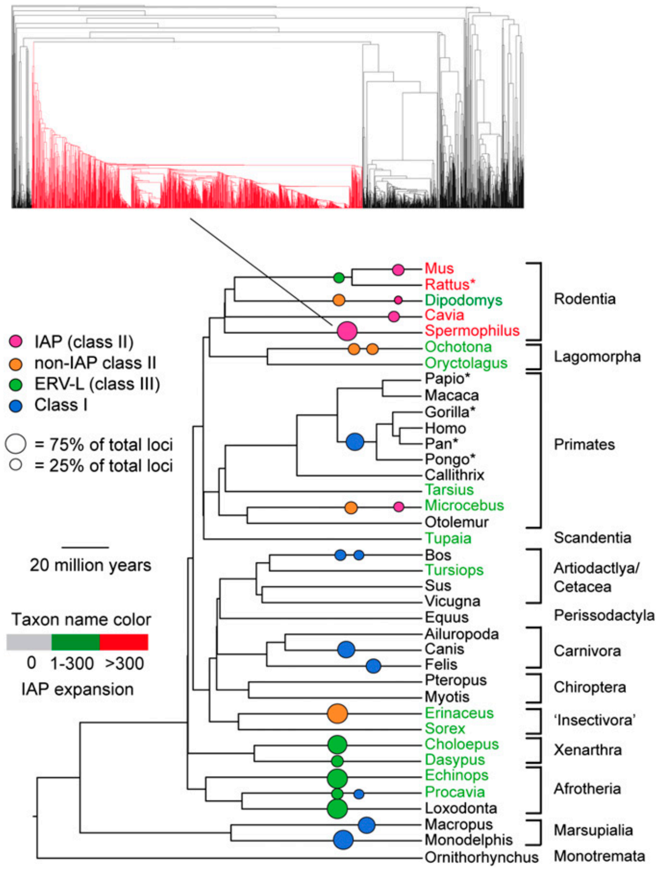

Fig. 1. Phylogeny of mammals (57) with ERV megafamilies (see text) shown as colored circles (area is proportional to the percentage of the ERV loci in the genome represented by that family). The placing of megafamilies on the tree shows relative age but not origin (which may be considerably earlier). Scale bar shows approximate dates in host phylogeny. Asterisked taxa are treated as duplicates and excluded from our analysis of all ERV families. Name color shows how many IAP loci were found in each species (Table S1). A typical megafamily in one genome (*Spermophilus*) is shown colored red.

lineages frequently are sister clades but are all independent invasions that occurred after the mouse/rat speciation.

Loss of *env* Is Associated with Proliferation in IAPs. The phylogenetic tree of the 4,089 IAP loci with more complete *pol* sequences (Fig. 2) shows repeated invasions by an IAP-like virus with *env* and subsequent degradation of this gene as measured by the length of the longest ORF: Most loci in the largest *Mus* expansion have an *env* ORF of <200 aa and have lost >80% of their *env* nucleotides. The extent of *env* degradation appears to determine the size of the expansion within the genome; e.g., the great majority of the loci in the largest expansions have lost most of their *env* gene. This change is unidirectional: We find no cases of *env* gain (or switching) during an expansion. However, the independent invasions of the guinea pig (*Cavia*) and shrew (*Sorex*) were preceded by a switch in *env* (Fig. 2), both gaining their *env* gene from viruses more closely related to extant betaretroviruses (∼50% amino acid similarity in the transmembrane region to Mason–Pfizer monkey virus) than are IAPs (maximum of ∼20% similarity, which is to Jaagsiekte sheep retrovirus).

The *env* degradation is not caused primarily by locus age because (*i*) other genes are not so extensively degraded (Figs. S1 and S2), and (*ii*) unlike with other genes, *env* degradation is not positively associated with sequence divergence between the paired LTRs, which is an independent measure of the postintegration age of the locus. As shown in Fig. 2, *env* is more intact at basal branches, which are obviously older integrations. Also, with the exception of *Spermophilus,* all the large expansions have predominantly more similar paired LTRs, indicating that they are relatively young (i.e., integrating roughly within the last 12 million years) (Fig. S3). The short terminal branch lengths seen in Fig. 2 also are consistent with this relative youth. There is a striking difference between the larger *Spermophilus* expansion and that in *Cavia*: The two expansions have similar degradation of *env,* but the *Cavia* expansion is markedly younger.

To assess statistically the relationship between *env* integrity and both expansion and cross-species transmission in IAPs, we used evolutionary distinctiveness (ED) to measure if a locus is a result of low or high expansion history and performed a multivariate analysis based on generalized least squares (GLS) and accounting for phylogenetic correlation and changes in rate between internal and terminal branches. Our analysis showed that expansion is negatively correlated with *env* integrity ($P < 0.01$) but is not significantly correlated with the integrity of other ERV genes (*gag*, *prot*, and *pol*) (Tables S2 and S3). The results were similar when we adjusted ED for cross-species transmissions, confirming that *env* degradation occurs after the transmission (*SI Results* and Fig. S4). The model predicts that an IAP family with more than 80 loci is predominantly retrotransposing (at least 50% of its loci have lost at least 90% of their *env* ORF).

Distribution of Other ERVs in Hosts. We found a total of 83,614 ERV loci in the 38 mammal genomes screened. Although the IAPs are a relatively young group, in that all loci integrated after the divergence of their host genomes, some other ERV families are much older, and therefore some loci in different genomes are homologs. To avoid pseudoreplication we excluded loci that (*i*) did not have 90% nucleotide sequence identity with at least one other locus (retaining over half of the loci) or (*ii*) were in genomes that diverged within the last ∼25 million years, the date that corresponds approximately to 90% sequence identity assuming that integrated ERVs diverge at a similar rate to their hosts (17) (namely *Rattus*, *Papio*, and the nonhuman hominoids). The high sequence divergence across all ERVs necessitated the use of clustering using pairwise nucleotide similarity, and the resulting ERV dendrograms showed that, as with the IAPs, family size is very uneven. In most genomes the largest family accounts for more than half of the loci; extreme examples are *Erinaceus* and *Monodelphis*, in which the largest family accounts for >80% of the loci (Fig. 1 and Table S1).

Pooling the ERVs from all genomes, we find that the largest 22% of families account for 80% of the loci, and a similarly unbalanced distribution was observed in IAPs, where 18% (3/17) of the genomes contain 80% of the loci. This lack of homogeneity closely resembles the 20/80 rule observed in a range of infectious disease epidemics (e.g., HIV, parasites), where the most infectious ∼20% of individuals account for 80% of the onward transmissions (18–21). In infectious disease epidemics, homogeneity of onward transmission is severely violated by superspreaders, who create many more secondary infections than the rest of the population. By analogy with superspreaders, who can be defined statistically as the most infectious 1% of the infected individuals (21), we introduce the term "megafamily" to describe ERV families that have expanded abnormally. We define a megafamily as the largest family in a genome that also has significantly more loci than would be expected if loci were distributed randomly among families ($P < 0.01$). Six of the genomes had more than one abnormally large family, so we applied this test to the second largest family also.

Loss of *env* Is Associated with Proliferation in Other ERVs. All megafamilies except perhaps one in the lemur *Microcebus* appear to be retrotransposing rather than reinfecting, because they have lost or possess only a degraded *env* (e.g., Fig. S5). We compared the *env* integrity of each megafamily with that in a representative small family in the same genome, which was selected from the dendrogram to be of similar age and to represent between 1% and 10% of the loci (Table S1). We determined *env* integrity only for

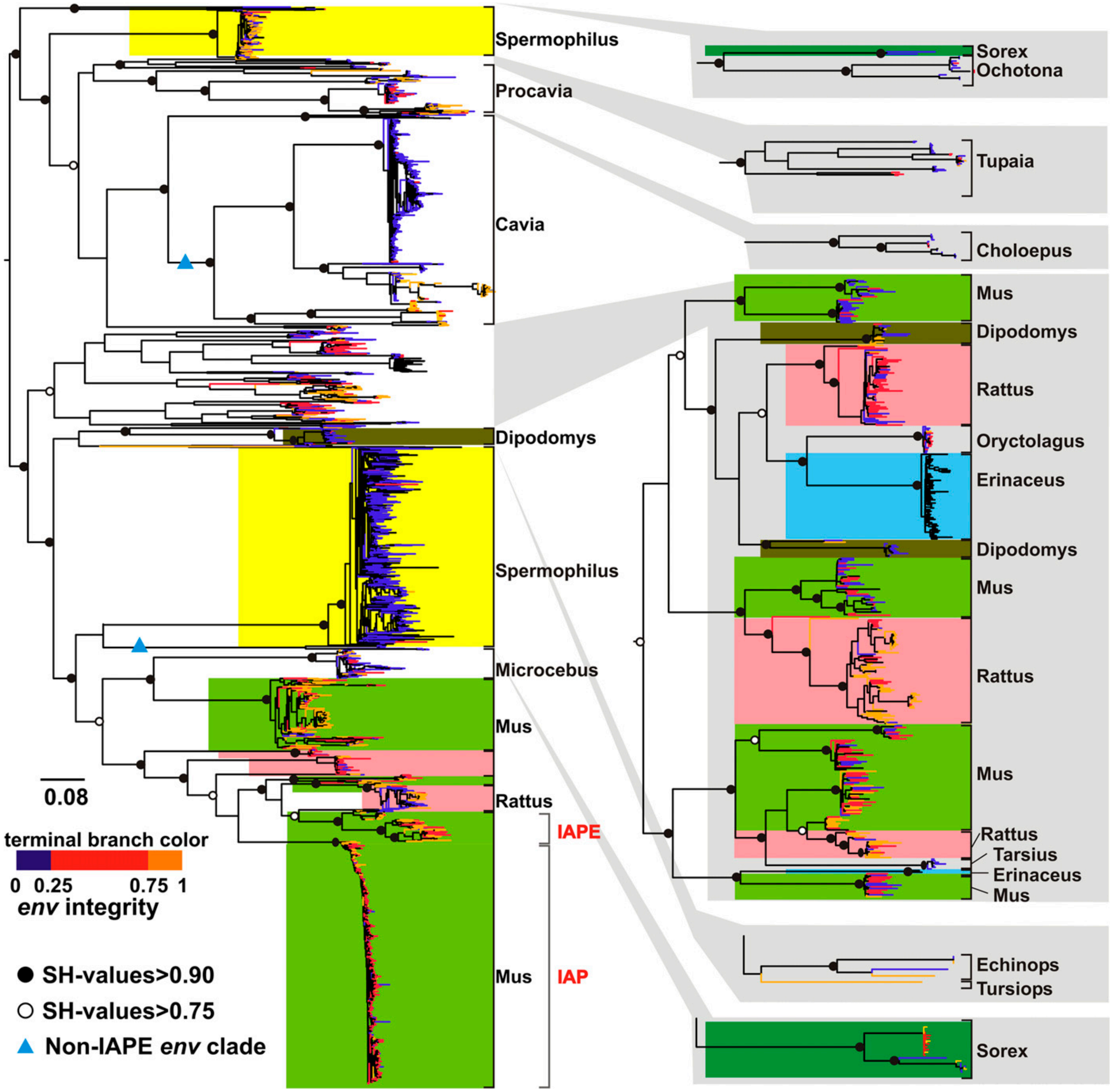

Fig. 2. Phylogenetic tree of IAP loci. Expansions in host species that have had multiple invasions are colored. Integrity of *env* gene is shown by color of terminal branch: orange indicates the longest ORF (at least 75% of the full length); red indicates an ORF between 25 and 75% of the full length; blue indicates an ORF <25% of the full length. Black shows loci for which we could not extract sequences >13 kb. Solid and open circles show Shimodaira–Hasegawa (SH) support values > 0.90 and >0.75, respectively. The two blue triangles show switches of *env*. The published IAPE and IAP sequences are indicated.

the selected families and only after their selection, which was done without prior knowledge of their biology. Therefore, we consider the comparison of family size with gene integrity to be a blinded experiment. We found that 23 of 24 megafamilies have a more degraded *env* gene [$x^2 = 20.2$; $P < 0.001$]. As in our analysis of IAPs, we can exclude a possible confounding effect of time inside the genomes because the *gag* gene, necessary for both replication mechanisms, was not similarly degraded: In 12 of the 24 comparisons the *gag* integrity was lower in the megafamily; this 50% finding would be expected by chance. In Fig. 4 we show this relationship between *env* integrity (as a ratio with *gag*) and family size. The megafamilies are, on average, ~30-fold larger than other families. An additional comparison between *env* degradation in megafamilies compared with all other loci in the same genome shows the same result: In the same 23 of 24 comparisons, there is more degradation of *env* in the megafamily (Table S4).

ERVs are divided into three classes (22), and we find retrotransposing megafamilies in all of them (Fig. 1). Class I (most closely related among the XRVs to gammaretroviruses) has eight retrotransposing megafamilies, which together make up 33% of the total class I loci; class II (closest to betaretroviruses) has nine, including four IAPs, which make up 41% of the class II loci; class III (closest to spumaviruses) has six, all ERV-Ls, which make up 71% of the class III loci.

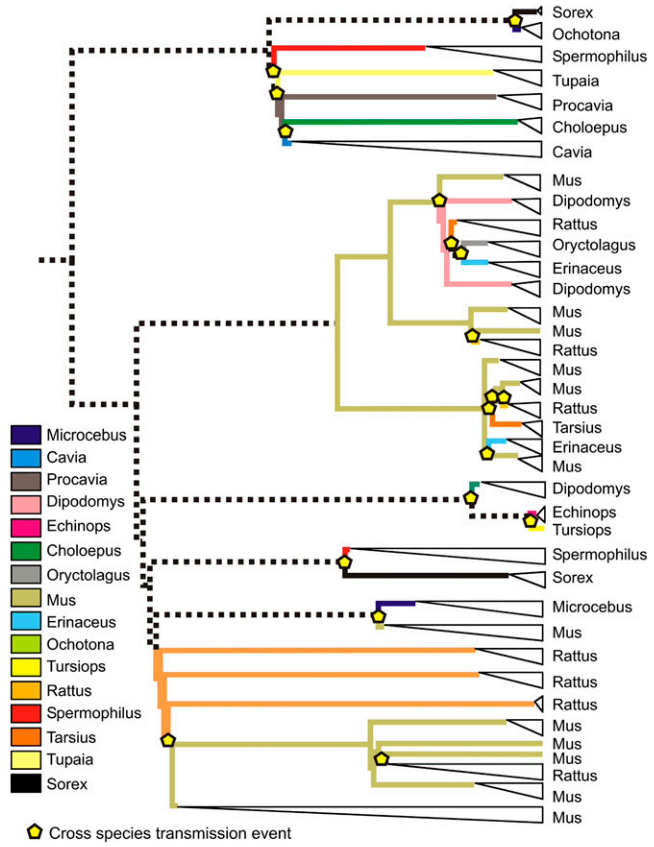

Fig. 3. Phylogenetic tree of the IAPs with the inferred ancestral states of their host species. Expansions are collapsed into single taxa (white triangles), and cross-species transmission events are indicated by yellow pentagons. Colored lines show ancestral states that, according to the available host sampling, can be attributed to a single host. Dashed lines show ancestral states that could not be resolved.

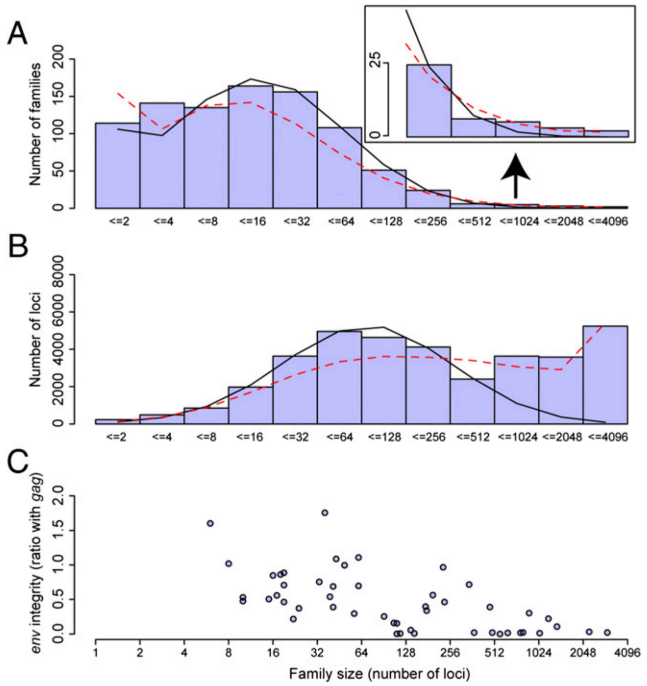

Fig. 4. Histograms showing (*A*) how common are ERV families of different size (*Inset* shows right-hand tail expanded for clarity) and (*B*) how many loci in total are in these families. Lines are generated assuming a lognormal (solid black) or generalized Pareto (dashed red) distribution. (*C*) *env* integrity (relative to *gag*) for megafamilies and randomly selected smaller families. The horizontal axes have been scaled using the logarithm to base 2.

Frequency Distribution of ERV Family Sizes Is Skewed. The 20/80 rule mentioned above (also referred to as the "80/20 rule" or "Pareto principle") is simply a description of power-law distributions, such as the Pareto distribution, which have a fat right-hand tail: i.e., a majority of the instances belong to a minority of the groups. Although the mechanisms that generate them are varied, such power-law distributions also describe abundance in a variety of areas, including other genomic systems (23, 24). As shown in Fig. 4 and Fig. S6, the Pareto distribution matches the observed right-hand side of our observed frequency distribution of family sizes better than the log-normal distribution, which commonly matches, albeit crudely (25), the observed distribution of individual organisms among species (26).

Discussion

The center of IAP diversity appears to be the rodents with some spill-over infections into other species, chiefly small mammals in similar habitats but also including the dolphin *Tursiops*. There also is some evidence of host phylogeny affecting cross-species transmission: IAPs appear to have invaded only the basal lineages among the well-sequenced primates. Moreover, mouse and rat IAP expansions frequently are sister clades, a result that is compatible with mouse and rat being sister species among the sequenced rodents. Interestingly, the abundant *env*-less IAP loci in mouse are not, as originally thought, the degraded descendants of the IAP loci shown to have a functional *env* (i.e., IAPEs) (9) but rather, as shown in Fig. 2, are from an independent invasion of the mouse genome. It is not known whether the inbred status of the laboratory mouse has facilitated the proliferation of IAPs (27), but we find a similarly large proliferation in the nondomesticated ground squirrel *Spermophilus*.

Our study shows that mammalian ERVs have evolved independently into retrotransposons multiple times, and this process underlies their relative abundance in mammal genomes. Integrating this information into the known biology of ERVs (1, 6, 9, 28) suggests that genome invasion by XRVs generates ERV lineages that typically expand through reinfection in the initial stages but often adapt to become intracellular retrotransposons. This adaptation leads to the degradation of the now-redundant *env* gene and confers increased intracellular but diminished interhost mobility. ERV lineages do not persist indefinitely in their host but rather cease replicating after a predictable time (28): Proliferation and cross species transmission might be regarded as alternate responses to lineage extinction. Among IAPs, we find no cases of cross-species transmission after loss of *env*, and, indeed, no cases of *env* capture by *env*-less vertebrate ERVs are known (29). However, we cannot preclude the possibility that such capture might occur. Rare events such as complementation and recombination might restore the capability of the extracellular life cycle; for example, in invertebrates there have been multiple evolutionary transitions from LTR-retrotransposons to retrovirus-like elements by the gaining of a third ORF analogous to *env* and an assumed shift from retrotransposition to reinfection (30). There also are examples of cross-species transmission by various TEs that lack an obvious mechanism for reinfection (31).

Why should a shift to retrotransposition lead to greater proliferation? First, reinfection might reduce host fitness more. Reinfection probably involves more replication in somatic cells, with the consequent risks of insertional mutagenesis. Production of endogenous Env protein may interfere with the normal function of the receptor and can cause cell fusion (32), a dangerous effect even though several *env* genes have been co-opted for this purpose in the host placenta (33). The transmembrane domain of the Env protein also has immunosuppressive properties (34, 35) that might have a negative effect on host fitness. Second, production of endogenous Env protein might be disadvantageous to the ERV, e.g., possibly leading to receptor interference in which intracellular binding of the cellular receptor to endogenously expressed Env protein results in down-regulation of the receptor required for viral reentry (36). A functional *env* gene thus might inhibit proliferation through reinfection. In addition, retrotransposition simply might be a more efficient way to generate new integrations into germline cells (27), circumventing the requirement for survival in a hostile extracellular environment and evading some innate antiviral defenses [e.g., tetherin, a membrane-bound protein that inhibits the replication of enveloped viruses by tethering budding virus to the cell-surface (37–39)]. That retrotransposing ERVs are more common than reinfecting ones is consistent with ERVs as a group being rarer than the entirely retrotransposing Long Interspersed Nuclear Elements (LINEs) in the mouse and human genomes (2, 3).

Is loss of *env* a cause or a consequence of the shift to retrotransposition? In mouse IAPs, loss of *env* appears to be a consequence: It has been shown experimentally that polymorphisms in the MA domain of the Gag protein direct the packaging of the IAP particles either toward the cell membrane or within the cisternae of the endoplasmic reticulum (14, 40). The MA domain in Gag has been shown to play the same role in an unrelated family of mouse ERVs called "musD" (41). Also, changes in the myristoylation signal of Gag in HIV restrict budding on the plasma membrane (42, 43). Thus the Gag protein appears to play a key role in determining the extracellular or intracellular fate of a retroviral life cycle. We assume that the Env protein, with its role in attachment and entry into the cell, becomes redundant when packaging occurs at the endoplasmic reticulum, and we see rapid loss on the phylogenetic trees (Fig. 2). However, as discussed above, the loss of *env* might determine the success of the shift to retrotransposition.

As mentioned in the Introduction, we did not analyze nonautonomous LTR-retrotransposons such as Mammalian apparent LTR-retrotransposons (MaLRs), which are ERV-like elements that lack *pol* and *gag* genes as well as *env* and replicate using proteins produced by other ERVs. For example, in the mouse genome there are four distinct groups of nonautonomous LTR-retrotransposons, each with a phylogenetically related ERV family that is assumed to replicate them (44). Our mining relies on the presence of *pol*, so these retrotransposons would be unlikely to be recovered. Loci can be copied by segmental genomic duplication, but this copying is negligible compared with other replication mechanisms in the ERV lineages (1, 6, 10).

Can we say anything about the generating process that creates our observed ERV family size distribution? If so, such a discussion might provide a method of producing null distributions and thus help detect biologically significant deviations. The widely used and well-described Gibrat's law (the law of proportionate effect) states that if the size of an entity and its growth rate are independent, and the entities are of the same age, then the resulting distribution will be lognormal (45). However, if the entities are of different age (time is a random variable), then the resulting distribution will be lognormal with a power-law tail (46); such a distribution is called a "double Pareto-lognormal" distribution (24). ERV family size might be operating through Gibrat's law; i.e., the size of the family and its growth rate might be independent, and because the family ages are different, the resulting distribution would then be a double Pareto-lognormal distribution. (Note, we are not suggesting that our megafamilies, which lie within this power-law tail, are larger because they are older than the other families; simply by mixing lognormal distributions from different time points we can generate a double Pareto-lognormal distribution in which megafamilies from different time points would lie within the power-law tail and have the same age as the smaller families within the lognormal body.)

Perhaps most importantly, our findings suggest that retroviral abundance, measured both horizontally and vertically, is on a continuum specified by the *env* gene: Gain of *env* allows the acquisition of new hosts by horizontal transfer (cross-species abundance), and loss of *env* is associated with substantially greater expansion within the genome (genomic abundance). The *env* gene thus has a key role in defining both the occurrence of ERVs in host species and their abundance within each genome.

Materials and Methods

Genome Mining. We used an *in silico* approach detailed in *SI Materials and Methods*. We are confident that our rescreening with new divergent sequences allowed us to find the great majority of the *pol*-containing ERVs in the available genome sequence data.

Selection of Loci. All IAPs invaded their hosts after speciation, but other ERV loci probably integrated around the origin of vertebrates and, although detectable, will have diverged to the extent that little sense can be made of their phylogenetic relationships. We therefore excluded all loci that did not have a 300-nt-long match of at least 90% sequence identity with at least one other locus. This criterion excluded less than half of the loci (46%) and, assuming that the ERV sequence divergence is not markedly dissimilar to that of their hosts, represents the exclusion of loci that had ceased replicating ~25 million years ago (17), which is less than half the life of most mammalian orders. As shown in Fig. 1, a large majority of the mammals sequenced had diverged before this time, so loci in one genome should not have homologs in others. However, some primate and possibly the *Mus*/*Rattus* genomes diverged after this date. To avoid counting the same locus twice (i.e., commit pseudoreplication), we only used *Mus*, *Homo*, and *Macaca* to represent these clades In our analyses of family sizes. In theory, this process could exclude single-locus families; however, previous analyses of ERVs in the well-studied human genome have not revealed any single-locus families (47). Therefore, we do not expect this limitation to bias our analysis of family sizes.

Allocating Loci to Families. The number of ERV families in each genome was measured using silhouette width, *s*, a composite index that reflects the compactness and separation of clusters. The procedure, automated in Perl software, was as follows. (*i*) For each genome, a matrix was made of all pairwise dissimilarities between recovered ERV nucleotide sequences using the EMBOSS water program (48), an implementation of the Smith–Waterman alignment algorithm (with gap opening and extension penalties of 10 and 4, respectively). (*ii*) Using silhouette from the partitioning around medoids method included in the Cluster package in R (49), the *n* sequences were partitioned into *k* clusters (where $2 < k < n - 1$), and the mean value of *s* was calculated for each value of *k*. (*iii*) The *k* clusters associated with the highest mean value of *s* were designated as families, each of which was named provisionally according to the most common reference sequence allocated to that cluster. We then manually corrected the assignment to families, fusing or breaking clusters, by visually inspecting the dendrograms and taking into account large tree asymmetries, which the clustering algorithm fails to identify. We finally determined as megafamilies the two largest families within each genome that are larger than the top 1% of the expected family size assuming a random equal distribution of the loci among the families.

Quantifying ERV Expansion. We measured whether a locus is a result of a low or high expansion history using ED, a measure originally conceived to provide a rational metric for prioritizing species conservation policies (50) and corrected by May (51) for nonbifurcating trees (polytomies). The ED metric is based on the idea that some lineages contain few species, and therefore their conservation should be prioritized (50, 51). It is implemented in the Tuatara package of Mesquite (52). ED is defined as the sum of the branches arising at each and for all subtending nodes (node score, *s*) standardized by dividing into the sum of it across the tree. For each taxon *i* in the *N* taxon tree the ED is thus defined as:

$$ED_i = \frac{\sum_{i=1}^{N} s_i}{s_i}.$$

We use this formula instead of the inverse because it has better statistical properties: It is defined as a subset of the positive real numbers (1, +∞), whereas the inverse is a proportion and thus defined in the space (0,1).

The basal loci of an ERV lineage, which are thought to be closer to the initial events of the genome invasion and thus are the result of fewer replication cycles, would score a high ED value. On the other hand, more derived loci of an ERV lineage, which are thought to be the later events in an expansion and thus are the result of more replication cycles, would score a low ED value. Therefore, ED and the expansion of ERVs have a monotonically inverse correlation. Examples of calculating ED and the distribution of ED scores on the IAP tree are shown in Figs S7 and S8, and the robustness to phylogenetic uncertainty in Fig S9.

Correlating Gene Integrity with ED. We used the GLS approach as implemented by the Analysis of Phylogenetics and Evolution (APE) package (53) in R (49), taking into account both nonindependence of the data caused by phylogenetic relatedness (54) and nonuniform trait evolution on the tree [for one human ERV family it has been shown that gene degradation is concentrated on the terminal branches on the tree (55)]. The effect of phylogenetic relatedness can be incorporated in APE by modifying the value of Pagel's λ, and we created 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. We used a range of different values for λ and t, and selected the best-fit model using the Akaike Information Criterion (56).

Further details on methodologies and details about alignment, phylogenetic analyses, simulating frequency distributions, gene integrity, identifying and quantifying cross-species transmissions and invasions, and recombination analysis of *env* in IAPs are given in *SI Materials and Methods*.

ACKNOWLEDGMENTS. G.M. and R.B are funded by the Wellcome Trust. A.K. is funded by the Royal Society.

1. Jern P, Coffin JM (2008) Effects of retroviruses on host genome function. *Annu Rev Genet* 42:709–732.
2. Waterston RH, et al.; Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–562.
3. Lander ES, et al.; International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921.
4. Tristem M (2000) Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. *J Virol* 74:3715–3730.
5. Katzourakis A, Gifford RJ, Tristem M, Gilbert MT, Pybus OG (2009) Macroevolution of complex retroviruses. *Science* 325:1512.
6. Bannert N, Kurth R (2006) The evolutionary dynamics of human endogenous retroviral families. *Annu Rev Genomics Hum Genet* 7:149–173.
7. Hanafusa H, Hanafusa T, Rubin H (1963) The defectiveness of Rous sarcoma virus. *Proc Natl Acad Sci USA* 49:572–580.
8. Goodchild NL, Freeman JD, Mager DL (1995) Spliced HERV-H endogenous retroviral sequences in human genomic DNA: Evidence for amplification via retrotransposition. *Virology* 206:164–173.
9. 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.
10. 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.
11. Brookfield JFY (2005) The ecology of the genome - mobile DNA elements and their hosts. *Nat Rev Genet* 6:128–136.
12. Venner S, Feschotte C, Biémont C (2009) Dynamics of transposable elements: Towards a community ecology of the genome. *Trends Genet* 25:317–323.
13. Dalton AJ, Potter M, Merwin RM (1961) Some ultrastructural characteristics of a series of primary and transplanted plasma-cell tumors of the mouse. *J Natl Cancer Inst* 26: 1221–1267.
14. Dewannieux M, Dupressoir A, Harper F, Pierron G, Heidmann T (2004) Identification of autonomous IAP LTR retrotransposons mobile in mammalian cells. *Nat Genet* 36: 534–539.
15. Mietz JA, Grossman Z, Lueders KK, Kuff EL (1987) Nucleotide sequence of a complete mouse intracisternal A-particle genome: Relationship to known aspects of particle assembly and function. *J Virol* 61:3020–3029.
16. Reuss FU, Schaller HC (1991) cDNA sequence and genomic characterization of intracisternal A-particle-related retroviral elements containing an envelope gene. *J Virol* 65:5702–5709.
17. Kumar S, Subramanian S (2002) Mutation rates in mammalian genomes. *Proc Natl Acad Sci USA* 99:803–808.
18. Anderson RM, May RM (1992) *Infectious Diseases of Humans: Dynamics and Control* (Oxford Univ Press, Oxford, UK).
19. Woolhouse ME, et al. (1997) Heterogeneities in the transmission of infectious agents: Implications for the design of control programs. *Proc Natl Acad Sci USA* 94:338–342.
20. Galvani AP, May RM (2005) Epidemiology: Dimensions of superspreading. *Nature* 438: 293–295.
21. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM (2005) Superspreading and the effect of individual variation on disease emergence. *Nature* 438:355–359.
22. Blomberg J, Benachenhou F, Blikstad V, Sperber G, Mayer J (2009) Classification and nomenclature of endogenous retroviral sequences (ERVs): Problems and recommendations. *Gene* 448:115–123.
23. Luscombe NM, Qian J, Zhang Z, Johnson T, Gerstein M (2002) The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. *Genome Biol* 3:research0040.1-0040.7.
24. Reed WJ, Jorgensen M (2004) The double pareto-lognormal distribution - A new parametric model for size distributions. *Commun Stat-Theor M* 33:1733–1753.
25. Williamson M, Gaston KJ (2005) The lognormal distribution is not an appropriate null hypothesis for the species-abundance distribution. *J Anim Ecol* 74:409–422.
26. Bell G (2000) The distribution of abundance in neutral communities. *Am Nat* 155: 606–617.
27. Maksakova IA, et al. (2006) Retroviral elements and their hosts: Insertional mutagenesis in the mouse germ line. *PLoS Genet* 2:e2.
28. Katzourakis A, Rambaut A, Pybus OG (2005) The evolutionary dynamics of endogenous retroviruses. *Trends Microbiol* 13:463–468.
29. Kim FJ, Battini JL, Manel N, Sitbon M (2004) Emergence of vertebrate retroviruses and envelope capture. *Virology* 318:183–191.
30. Malik HS, Henikoff S, Eickbush TH (2000) Poised for contagion: Evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res* 10:1307–1318.
31. Schaack S, Gilbert C, Feschotte C (2010) Promiscuous DNA: Horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends Ecol Evol* 25:537–546.
32. Sommerfelt MA (1999) Retrovirus receptors. *J Gen Virol* 80:3049–3064.
33. Stoye JP (2009) Proviral protein provides placental function. *Proc Natl Acad Sci USA* 106:11827–11828.
34. Mangeney M, Heidmann T (1998) Tumor cells expressing a retroviral envelope escape immune rejection in vivo. *Proc Natl Acad Sci USA* 95:14920–14925.
35. Mathes LE, et al. (1979) Immunosuppressive properties of a virion polypeptide, a 15,000-dalton protein, from feline leukemia virus. *Cancer Res* 39:950–955.
36. Boeke JD, Stoye JP (1997) Retrotransposons, endogenous retroviruses, and the evolution of retroelements. *Retroviruses*, eds Coffin JM, Hughes SH, Varmus HE (Cold Spring Harbor Laboratories, New York, NY), pp 343–435.
37. Neil SJ, Zang T, Bieniasz PD (2008) Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu. *Nature* 451:425–430.
38. Perez-Caballero D, et al. (2009) Tetherin inhibits HIV-1 release by directly tethering virions to cells. *Cell* 139:499–511.
39. Jouvenet N, et al. (2009) Broad-spectrum inhibition of retroviral and filoviral particle release by tetherin. *J Virol* 83:1837–1844.
40. Fehrmann F, Jung M, Zimmermann R, Kräusslich HG (2003) Transport of the intracisternal A-type particle Gag polyprotein to the endoplasmic reticulum is mediated by the signal recognition particle. *J Virol* 77:6293–6304.
41. Ribet D, Harper F, Dewannieux M, Pierron G, Heidmann T (2007) Murine MusD retrotransposon: Structure and molecular evolution of an "intracellularized" retrovirus. *J Virol* 81:1888–1898.
42. Bryant M, Ratner L (1990) Myristoylation-dependent replication and assembly of human immunodeficiency virus 1. *Proc Natl Acad Sci USA* 87:523–527.
43. Göttlinger HG, Sodroski JG, Haseltine WA (1989) Role of capsid precursor processing and myristoylation in morphogenesis and infectivity of human immunodeficiency virus type 1. *Proc Natl Acad Sci USA* 86:5781–5785.
44. McCarthy EM, McDonald JF (2004) Long terminal repeat retrotransposons of Mus musculus. *Genome Biol* 5:R14.
45. Gibrat R (1931) *Les Inequalites Economiques* (Librairie du Recueil Sirey, Paris).
46. Montroll EW, Shlesinger MF (1982) On 1/f noise and other distributions with long tails. *Proc Natl Acad Sci USA* 79:3380–3383.
47. Gifford R, Tristem M (2003) The evolution, distribution and diversity of endogenous retroviruses. *Virus Genes* 26:291–315.
48. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* 16:276–277.
49. R Development Core Team (2008) R: A language and environment for statistical computing. Available at http://www.R-project.org.
50. Vanewright RI, Humphries CJ, Williams PH (1991) What to protect - systematics and the agony of choice. *Biol Conserv* 55:235–254.
51. May RM (1990) Taxonomy as Destiny. *Nature* 347:129–130.
52. Maddison WP, Maddison DR (2010) Mesquite: A modular system for evolutionary analysis.), version 2.73.
53. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20:289–290.
54. Harvey PH, Pagel MD (1991) *The Comparative Method in Evolutionary Biology* (Oxford Univ Press, Oxford, UK).
55. Belshaw R, et al. (2004) Long-term reinfection of the human genome by endogenous retroviruses. *Proc Natl Acad Sci USA* 101:4894–4899.
56. Akaike H (1974) A new look at the statistical model identification. *IEEE Trans Automat Contr* 19:716–723.
57. Bininda-Emonds OR, et al. (2007) The delayed rise of present-day mammals. *Nature* 446:507–512.

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 nontransformed 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 $\mathrm{ED_{cst}}$). 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, $\mathrm{ED_{cst}}$, 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 $\mathrm{ED_{cst}}$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.

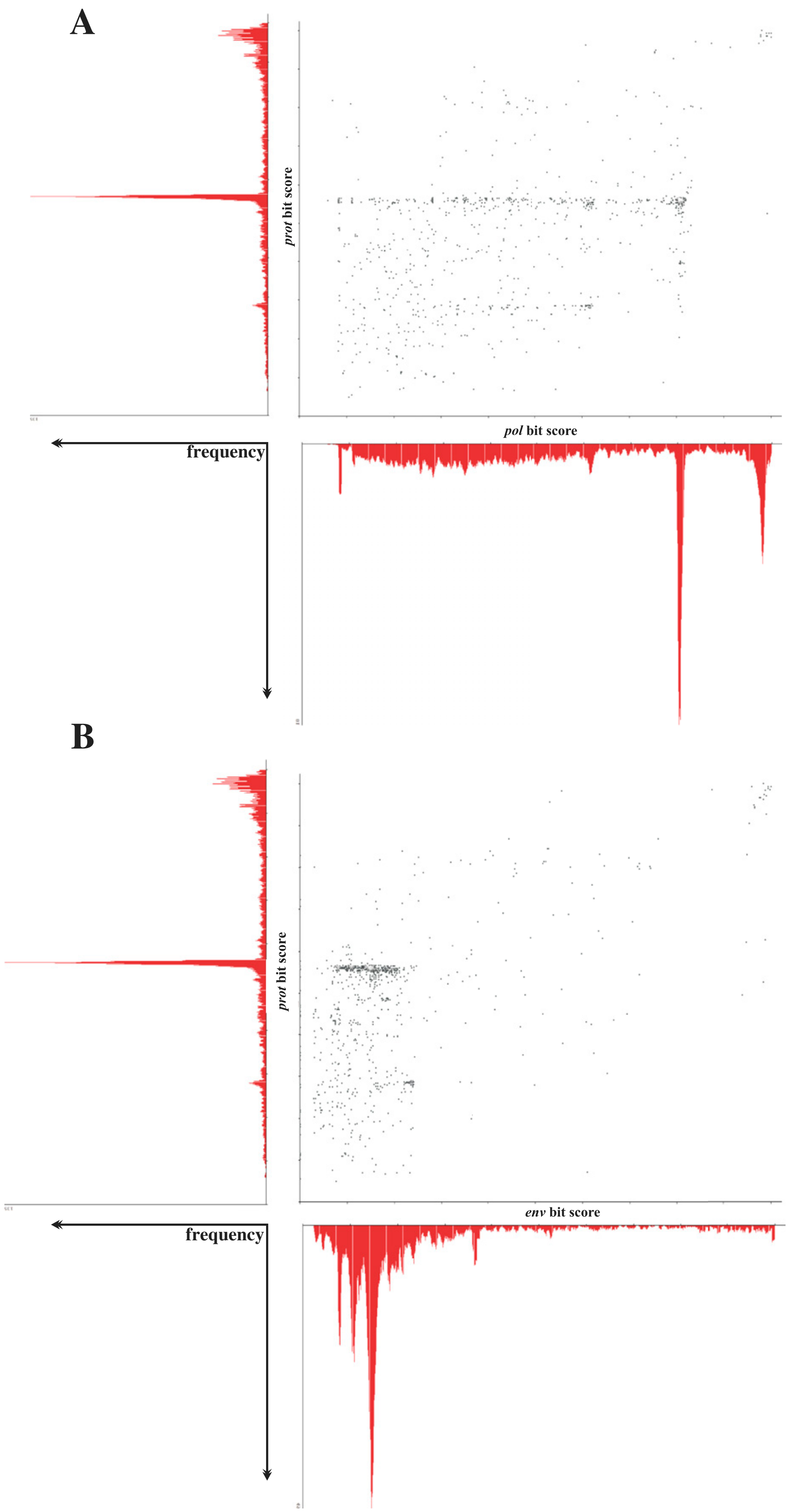

Fig. S1. Scatterplots of the TBLASTN bit scores associated with axes-specific histograms from the *Mus* IAP elements for *prot* against (*A*) the *pol* genes and (*B*) the *env* genes (*gag* is similar to *pol*). The striking observation is that the *env* scores, unlike those of the other genes, are 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).

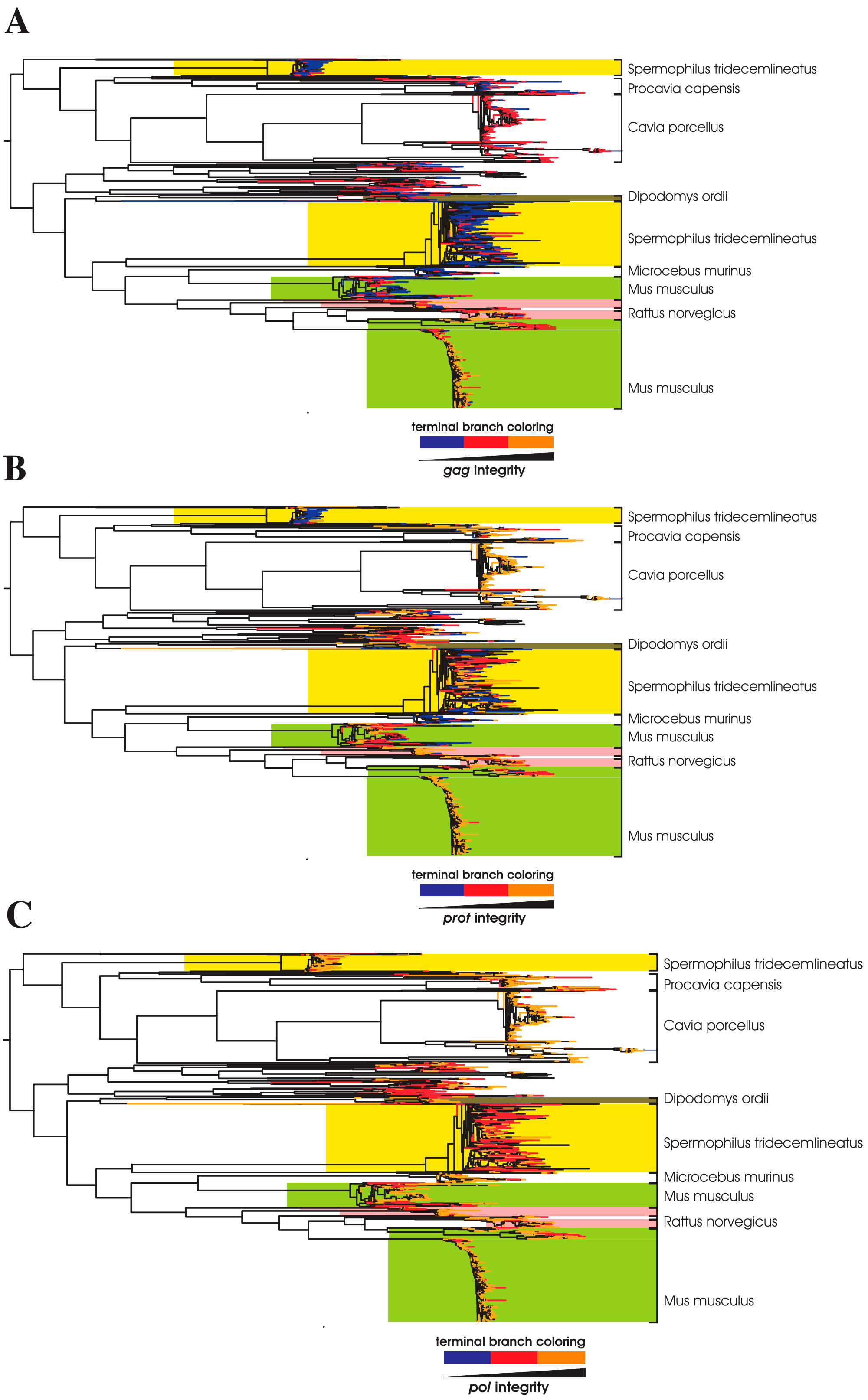

Fig. S2. Distribution of gene integrity on the IAP tree shown in Fig. 2 and described in the legend of Fig. 2. (*A*) *gag*. (*B*) *prot*. (*C*) *pol*.

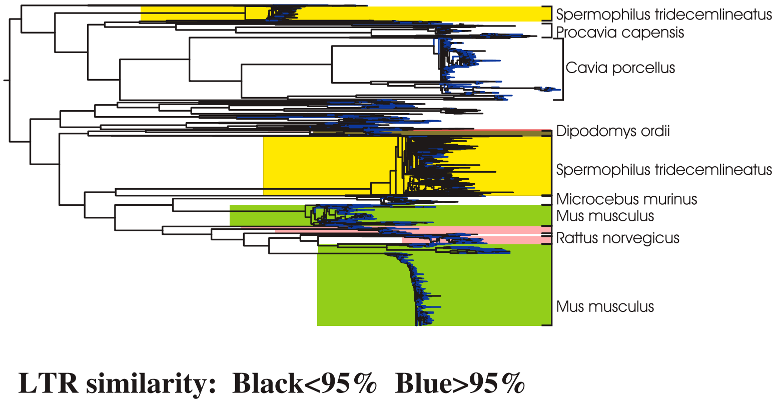

Fig. S3. Distribution of LTR similarity on the IAP tree shown in Fig. 2. Blue shows elements with more-similar LTRs (≥ 95% similarity). Black shows elements with less-similar LTRs.

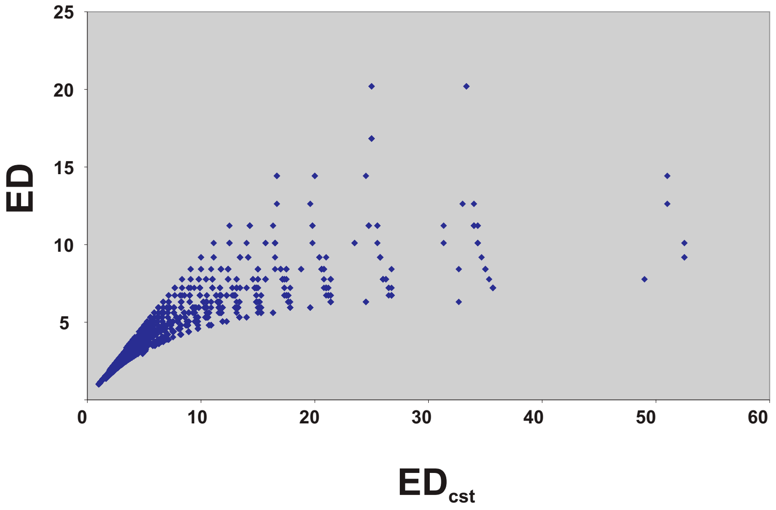

Fig. S4. Scatterplot of ED against ED_{cst} showing high correlation between the two values.

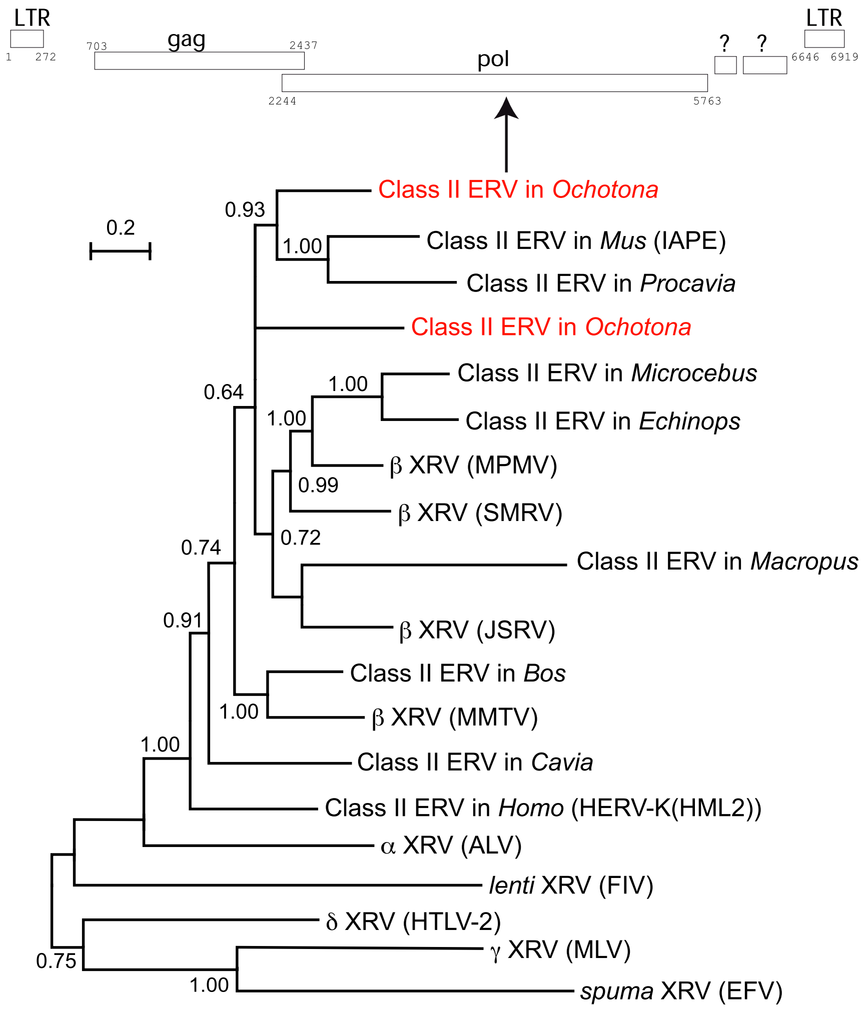

Fig. S5. Phylogenetic tree of *pol* sequences from analyzed class II ERVs plus (*i*) extant betaretroviruses [mouse mammary tumor virus (MMTV), Jaagsiekte sheep retrovirus (JSRV), squirrel monkey retrovirus (SMRV), and Mason-Pfizer monkey virus (MPMV)], (*ii*) representatives of the other main XRV clades [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.

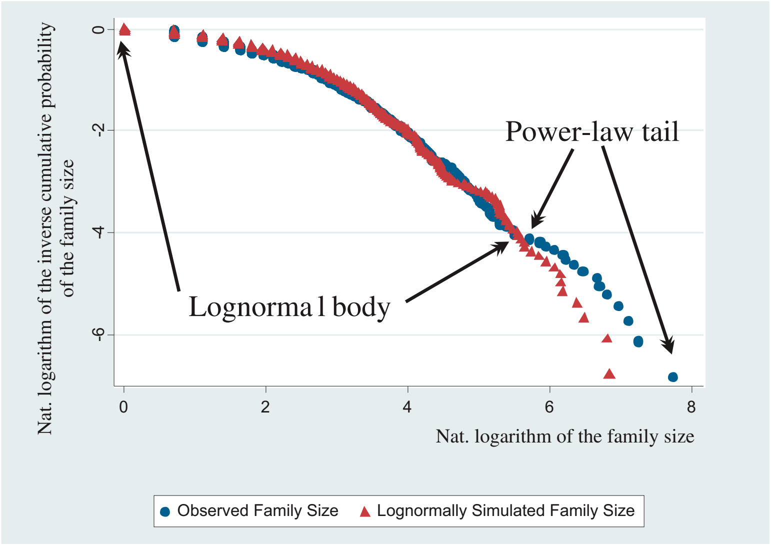

Fig. S6. Overlaid scatter plots of the logarithmically transformed inverse cumulative distributions (vertical axis) vs. their logarithmically transformed family sizes (horizontal axis) for the observed family sizes (blue circles) and a single lognormally simulated one (red triangles).

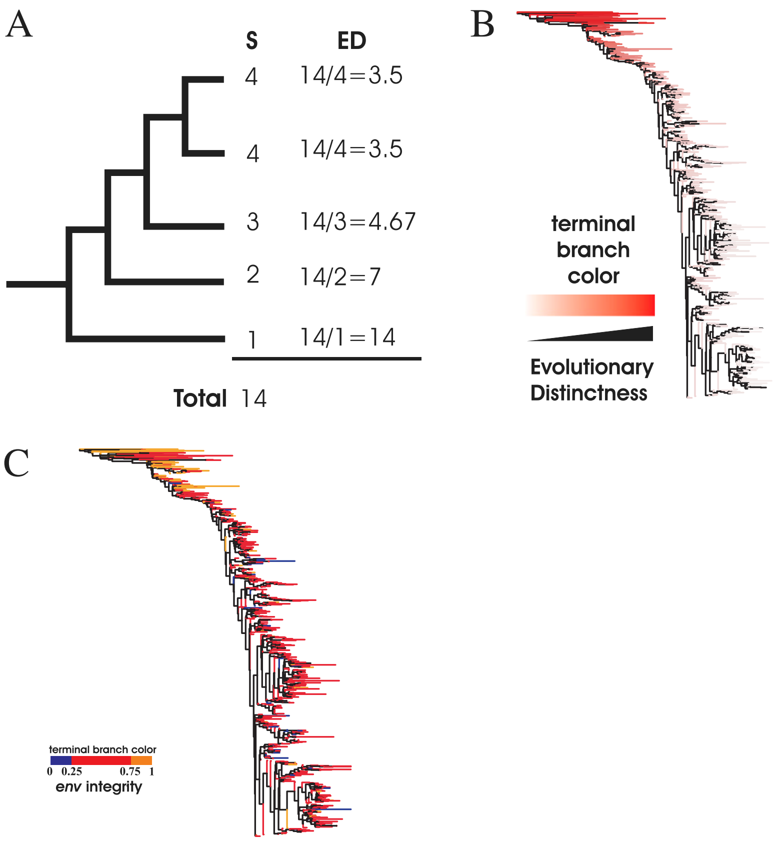

Fig. S7. The ED and *env* integrity metrics. (*A*) A five-taxon tree as an example of calculating ED. The S column shows the number of nodes from the root, and the ED column shows the calculation of ED. Taxa that result from more replication events (larger expansion) have lower ED. (*B* and *C*) ED and *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.

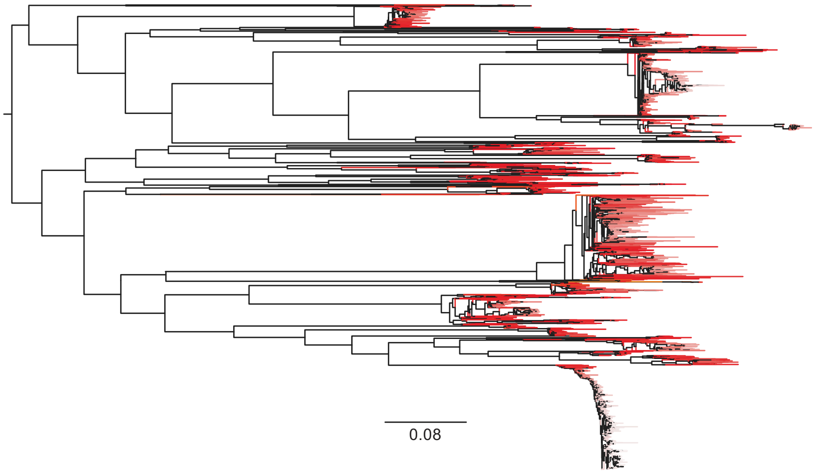

Fig. S8. Distribution of ED on the IAP tree shown in Fig. 2. Intensity of red shading is proportional to ED value. Smaller clades and the basal loci in larger clades tend to be darker, with higher ED values showing a less abundant replication history.

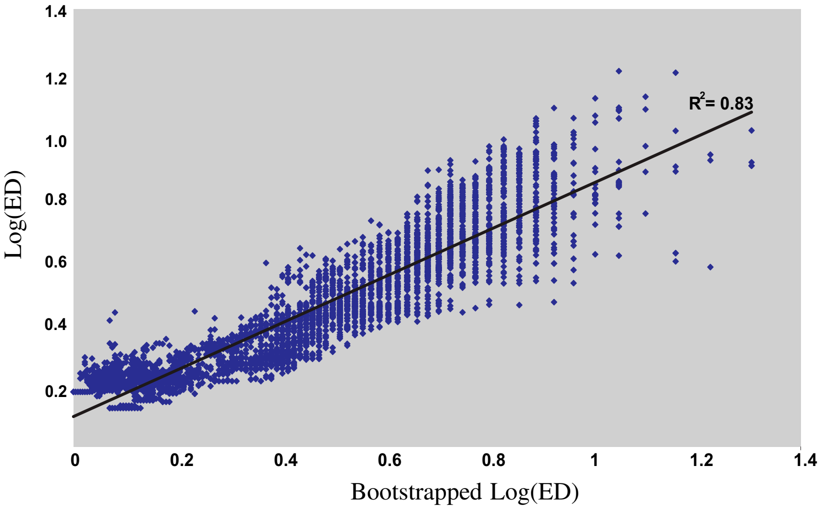

Fig. S9. Scatterplot of the logarithm to base 10 of the ED [log(ED)] estimated from the original alignment against the respective values from 100 bootstrapped pseudo replicates [bootstrapped Log(ED)]. The regression line and the Pearson coefficient are shown also.

Table S1. Summary of ERVs in the mammal genomes

		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
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 nonmegafamilies in *Erinaceus* and *Loxodonta.*