J Mol Evol (2003) 56:181–186
DOI: 10.1007/S00239-002-2392-3

JOURNAL OF MOLECULAR EVOLUTION

Molecular Characterization of the Recent Intragenomic Spread of the Murine Endogenous Retrovirus MuERV-L

Javier Costas

Departamento de Bioloxía Fundamental, Facultade de Bioloxía, Universidade de Santiago de Compostela, Campus Sur s/n E-15782 Santiago de Compostela, A Coruña, Spain

Received: 27 November 2001 / Accepted: 20 August 2002

Abstract. The mouse genome has been subjected to two successive amplification bursts of the murine endogenous retrovirus MuERV-L after the *Mus*/*Rattus* split. The main objective of this work is to characterize in detail the intragenomic spread giving rise to these two murine bursts using full-length MuERV-L proviruses taken from public databases. Phylogenetic analyses led to the identification of elements putatively amplifying during each one of the two burst. Likelihood-ratio tests were used to confirm that elements supposedly arisen during the first burst have been evolving under lower selective constrains, as expected for older insertions. The data reported here suggested an evolutionary dynamics for MuERV-L amplification characterized by the existence of multiple elements simultaneously active during each one of the bursts while only one or very few closely related proviruses from the first burst gave rise to the second one. Finally, more than one third of the proviruses present 100% identity between the 5′ and 3′ LTRs, strongly indicating that MuERV-L is currently active within the mouse genome.

Key words: Endogenous retrovirus — ERVs — MuERV-L — *Mus musculus* — Retrotransposition — Retrovirus-like elements — Likelihood ratio tests — Evolutionary model

Email: bfcostas@usc.es

Introduction

Endogenous retroviruses (ERVs) have a widespread distribution within vertebrates (Herniou et al. 1998). They originated by the integration of exogenous retroviruses into the germ line at different times during vertebrate evolution. Extensive research on ERVs from human, the best characterized vertebrate genome, revealed that most of the HERV families colonized this genome after the main radiation of eutheria and are thus specific to primates. Nevertheless, a few ERVs entered the germ line earlier, such as ERV-I, and are present in different vertebrate classes (Martin et al. 1997), or ERV-L, detected in several mammalian orders (Cordonnier et al. 1995; Bénit et al. 1999).

Two complete nucleotide sequences of ERV-L have been characterized, one from humans (HERV-L, Cordonnier et al. 1995) and the other from mouse (MuERV-L, Bénit et al. 1997). While the HERV-L provirus lack long ORFs (due to frameshift and nonsense mutations), the MuERV-L sequence displays a full ORF homologous to the *gag* gene, and an almost intact ORF homologous to the *pol* gene (only disrupted by an 8-bp deletion), being more related to foamy viruses. Surprisingly, an ORF with homology to dUTPase proteins overlaps the 3′ end of the ERV-L *pol* gene. This location (different from that in the other retroviral groups) suggests an independent acquisition of dUTPases in these lineages. There is not an *env*-homologous ORF in ERV-L.

An extensive characterization of the dispersion of ERV-L in mammals has been carried out both by

southern and slot blot analyses and amplification and sequencing of 360 bp internal fragments of the *pol* gene (Bénit et al. 1999). The main conclusions taken from these analyses are: (1) ERV-L sequences are present among all placental mammals; in general, at low copy number (10 to 30); (2) there has been a burst in copy number (up to 200 copies) in primates, shortly after the prosimian/simian split; (3) there have been two recent bursts in copy number in murine species, one after divergence between *Mus* and *Rattus* (that occurred about 10 MYA), and the other more recently (less than 2 MYA), in the common ancestor of *Mus musculus* and its closest relatives *M. spicilegus* and *M. macedonicus*. The first murine burst increased the copy number of MuERV-L proviruses from an initial value of ~25 to ~50; while this number reached ~125 after the second burst. The copy number of MuERV-L "solo LTRs," generated through homologous recombination between the 5′ and 3′ LTRs of a provirus, is approximately 10 times higher.

The existence of these two recent bursts in a species with an ongoing genome project constitutes a very useful model to analyze the dynamics of active ERVs in a short period of evolutionary time. Thus, the main objective of this article is to characterize in detail the amplification dynamics leading to these two murine bursts.

Materials and Methods

Identification of MuERV-L homologous sequences was made by screening the nr and the HTGs databases at the National Center for Biotechnology Information with the internal region (i.e., after removing the LTRs) of the MuERV-L element described in Bénit et al. (1997; GenBank accession number: Y12713), using the program BLASTN (Altschul et al. 1990). The limits of the identified proviruses were determined by local alignment with the MuERV-L LTRs using BLAST 2 sequences (Tatusova and Madden 1999). This program was also used in pairwise comparisons between closely related sequences to detect duplicated entries, which were excluded from the analyses.

Alignment of the sequences was easily done by visual inspection (due to their high degree of homology) with the aid of GeneDoc (Nicholas and Nicholas 1997). Detection of long open reading frames (ORFs) in the collected MuERV-L proviruses was also performed with GeneDoc. MEGA v2.1 (Kumar et al. 2001) was used to calculate divergence values between different sequences, to reconstruct phylogenetic relationships by the Neighbor-Joining method (Saitou and Nei 1987) and to calculate bootstrap values for each internal branch (1000 replicates). In all cases, the Kimura's two-parameter model was applied to correct for multiple substitutions.

Maximum likelihood (ML) analyses were performed using the CODEML program from the PAML v3.0 package (Yang 2000) to examine the selection pressures on the *gag* and *pol* genes. Option G was used in the sequence data format to carry out the combined analyses of these two genes (Yang 1996). Stop codons and codons involving a gap in any of the sequences were removed from the alignment prior to the analyses. Two different tree topologies for model fitting were used (see results). Two codon-based ML models were compared. One of the models, the "one-ratio" model, assumes that all branches of the tree evolve under the same d_N/d_S ratio (the ratio of per-site rates of synonymous and non-synonymous changes), while the other one, the "two-ratio" model, assumes a different d_N/d_S ratio for selected branches. Both models account for transition/transversion rate bias and codon usage bias. Nucleotide frequencies at the three codon positions were used to determine equilibrium codon frequencies. A likelihood-ratio test was applied to compare the two models. As they are nested models (i.e., the null hypothesis is a special case of the alternative hypothesis) differing in one parameter, twice the log-likelihood difference between these two models can be compared with a critical value of the χ^2 distribution with 1 degree of freedom to test whether the "two-ratio" model fits the data significantly better than the "one-ratio" model (Yang 1996; Huelsenbeck and Crandall 1997; Huelsenbeck and Rannala 1997).

Results

My BLAST searches identified 38 full-length MuERV-L proviruses from the public databases (Table 1). Thirty-six of the 38 sequences are highly similar, showing an average pairwise divergence value of 0.014. The degree of divergence between each one of the other two sequences, AC008783b and AC006520, and these 36 was 0.053 and 0.071, respectively. Consequently, these two sequences were used as outgroups in the phylogenetic analyses. Both ORFs of 22 of the 38 proviruses remained intact, and in 16 the two LTRs are identical (Table 1).

Figure 1 shows the polymorphic nucleotide sites shared by at least two sequences between the 36 highly similar proviruses after removing the hypervariable CpG dinucleotide positions. Interestingly, two different LTR sequences are common to several elements. One is shared by proviruses AL603843, AC079497, and AC055819; and the other by AC087890, AL596283, and AL611930. Besides, the alignment of the 36 sequences revealed the existence of three polymorphic indels, one of them located within the LTR sequence, and the other two within the ORF1. These two ORF1 indels involve multiple of three base pairs, thus, maintaining the ORF. Two of the three indels encompasses a duplicated sequence. Only four proviruses present the two LTRs of the short type, while five proviruses have one LTR of common size and the other with the deletion. Nevertheless, the two LTRs for each one of these five proviruses are highly similar (ranging from 0 to 3 nucleotide differences), suggesting gene conversion events involving a very short stretch of DNA.

Figure 2 presents the phylogenetic relationships of these proviruses, based on the alignment of the sequences after removing the 3′ LTR (to avoid redundancy) and the CpG dinucleotide positions. Taking into account the existence of two different amplification bursts (Bénit et al. 1999), the tree in Fig. 2 suggests a classification of sequences according to their integration during the first or the second burst. Twenty-three sequences constitute a monophyletic

Table 1. Collected full-length proviruses[a]

Accession number	5′-end	3′-end	Duplicates	ORF1	ORF2	LTR differences[b]
AC091520.6	26829	20434		+	+	0
AC055819.8	166778	160383		+	+	0
AC004155.1	103780	110175		+	+	0
AC020974.4	120253	126648		+	+	1
AL596386.5	151103	144703		+	+	3
AL603745.3	191019	184587		+	–	0
AL603843.2	192416	198811		+	+	0
AC079497.1	167278	173673		+	+	1
AC084823.10	25535	19140		+	+	0
AC079819.27	118703	125147		+	–	0
AC079500.1	110133	116565	AC079478.1 AC079514.1	+	+	0
AC055772.11	114111	107665		+	+	0
AL626784.4	16305	22751		+	+	0
AC006404.24	35304	41738	AC006447.21	+	+	1
AL135758.32	25054	31451		+	+	0
AC022774.5	205205	198769		+	+	2
AC087329.7	202583	196176		+	+	1
AC073939.4	119937	126317		+	–	3
AF110520.1	115910	122379	AC073742.2	+	+	0
AL626778.7	26278	32570		+	+	1
AL611930.9	34902	41286		–	+	0
AC087890.13	19757	26193		+	+	0
AL596283.7	32847	39283		+	+	0
Y12713.1	1	6471		+	–	10
AC084416.29	518127	524432		–	–	5
AC073436.41	68866	75276		+	–	8
AC087183.9a	117245	123727	AC079491.1	+	–	5
AC004096.1	39348	26706	AC003996.1	–	–	1
AC023518.5	181759	188228	AL596246.4	+	+	2
AC025585.3	263	6694		+	–	1
AC025667.12	103300	96937		+	+	1
AC093348.1	76816	83238		–	–	0
AC005855.1	17817	24276		+	–	2
AL512586.9	97057	90587		+	–	1
AC093445.3	148735	155209		+	–	1
AL606533	7297	13682		+	+	2
AC087183.9b	131328	123913		–	–	27
AC006520.7	13477	19874		–	–	23

[a] The proviruses are ordered according to the phylogenetic tree in Fig. 2
[b] Nucleotide differences between the 5′ and 3′ LTRs from the same provirus.

clade (although this monophyly is not supported by the bootstrap analysis) with terminal branches (i.e., those extending from a node to a sequence) shorter than those leading to the remaining 13 sequences (with the only exception of AC073939). This might be indicative of different periods of integration for these two groups of sequences, corresponding to the second and first burst, respectively.

It is currently accepted that, in general, individual elements are not subjected to selective constraint after integration, evolving such as pseudogenes. Based on this fact, I performed a ML approach to estimate d_N/d_S ratios for the combined coding regions under two different models, in order to test if sequences putatively arisen during the first burst have been spending more time evolving without selective pressure. The ''one ratio'' model assumes that all branches evolve under similar d_N/d_S ratios; while the ''two ratio'' model assumes one ratio for the lineages leading to the 13 proviruses putatively inserted during the first burst, an another ratio for the lineages leading to the remaining 23 proviruses (Fig. 2). I used two different tree topologies for model fitting, with similar results (Table 2). One of them is the unrooted version of the tree shown in Fig. 2, while the other is that obtained from the actual alignment used in the ML approach (i.e., only from coding regions, after removing stop codons and codons involving gaps). The two topologies are similar each other, showing the same monophyletic clade of 23 sequences. Only the details of the topology are different between the two trees. Log-likelihood values and d_N/d_S estimates under the two ML models using each one of the two tree topologies are presented in Table 2. The ''two ratio'' model fits the data significantly better than the ''one ratio'' ($p < 0.001$), indicating that the d_N/d_S ratios are

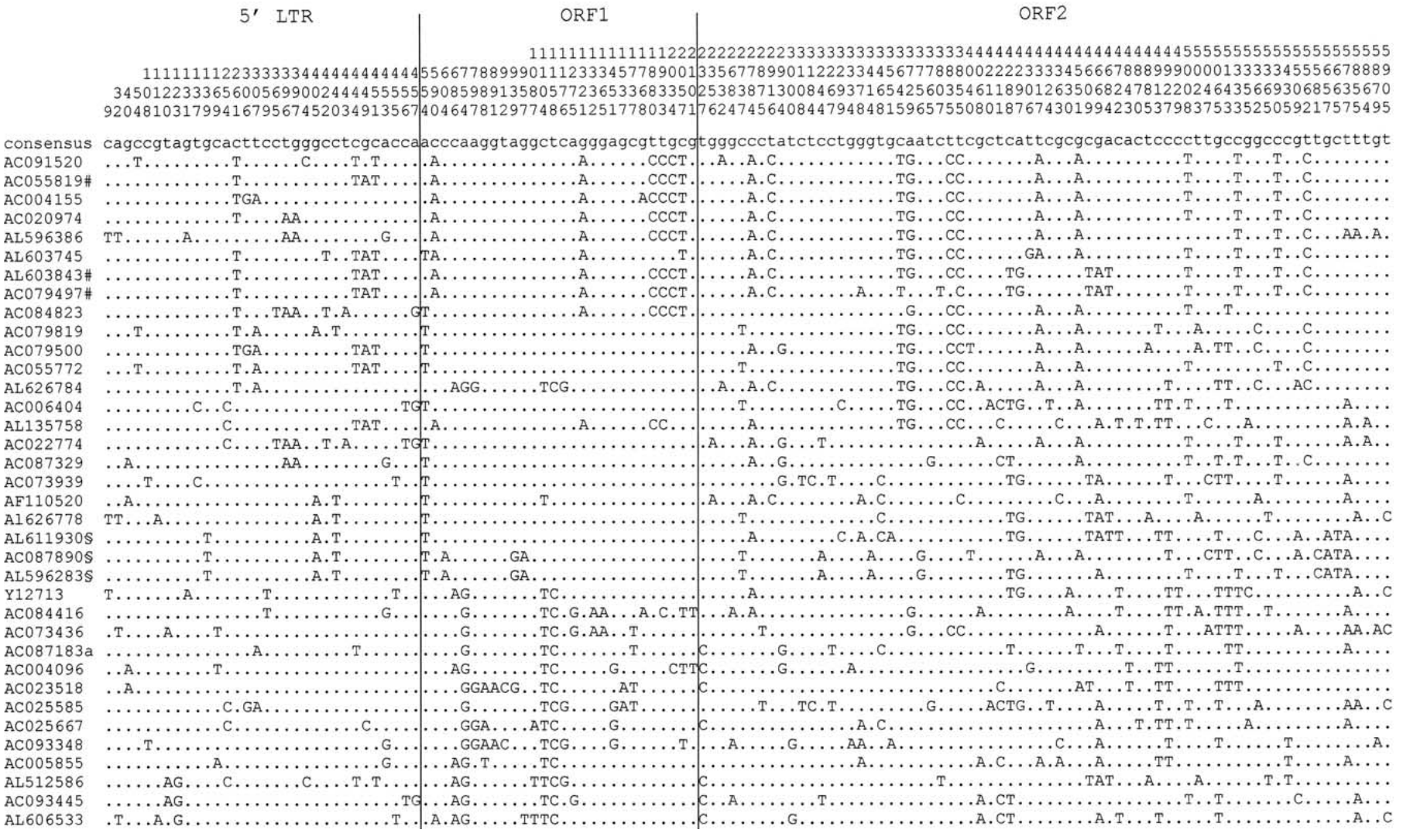

Fig. 1. Nucleotide polymorphisms shared by at least two sequences, after removal of CpG dinucleotides. Only the 5′ LTR is shown. Dots indicate identity with the consensus sequence. The positions are numbered across the top, referred to the MuERV-L element described in Bénit et al. 1997 (GenBank accession number: Y12713). Different regions of the proviral genome are indicated by vertical lines. Sequences are ordered according to the tree in Fig. 2. Two different LTR sequences shared by several elements are indicated by the # and § after the accession numbers.

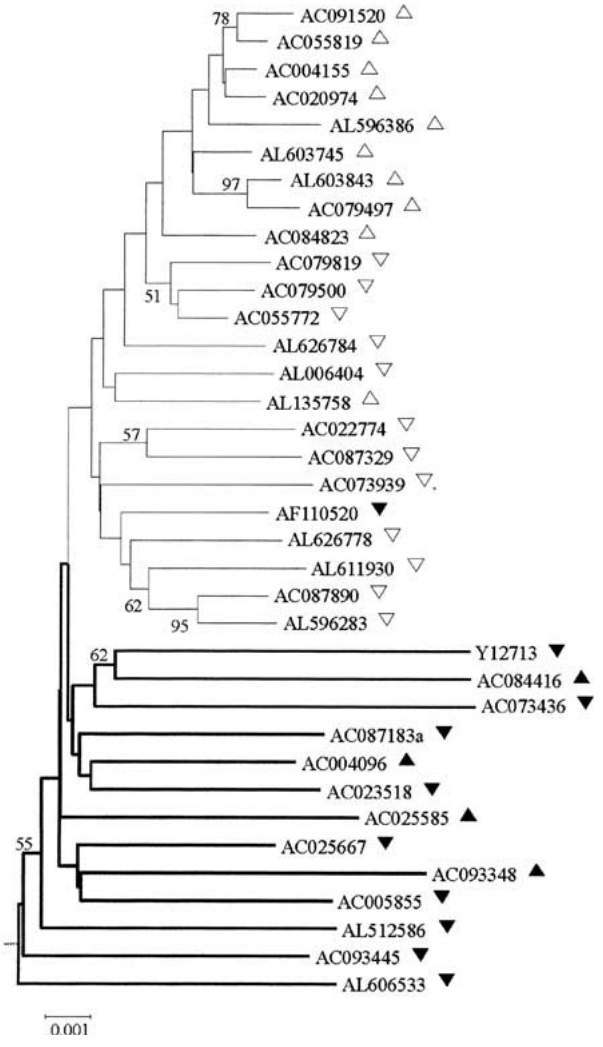

Fig. 2. Neighbor-joining tree of the full-length MuERV-L proviruses based on the 5′ LTR and internal nucleotide sequences, after removing CpG dinucleotide positions. The tree is rooted using sequences AC008783b and AC006520. The value at a particular node indicates its percentage of appearance in 1000 bootstrap replicates. Only values above 50% are shown. Branches of same thickness are assumed to evolve under the same d_N/d_S ratio in the "two-ratio" ML model. Symbols to the right of the names are as follows: △, sequence presenting the two deletions of 33 and 39 bp within the ORF1; ▽, sequence presenting only the first deletion (33 bp) within the ORF1; ▲, sequence presenting only the second deletion (39 bp) within the ORF1; ▼, sequence without any of the two deletions within the ORF1.

higher within those branches leading to the proviruses tentatively considered as integrated during the first amplification burst.

Discussion

In the present work, I collected a series of full-length MuERV-L proviruses integrated within the *Mus musculus* genome. Phylogenetic analyses of these sequences suggest the existence of two groups of proviral insertions, corresponding to each one of the two amplification bursts that took place in *Mus* (Bénit et al. 1999, see Introduction). Likelihood ratio tests confirmed that elements putatively arisen during the first burst have been evolving under lower selective constrains (higher d_N/d_S ratios). This fact strongly indicates that these elements integrated within the genome earlier, and thus evolved as pseudogenes for a longer period of time than members of the second burst. In addition, the LTRs of the older proviruses are more divergent than those of the younger ones, have a lower proportion of intact ORFs, and were recovered in fewer numbers from the database (Table 1). The latter finding is in agreement with Bénit et al. (1999), who showed that fewer elements belong to the first burst than to the second one.

Data reported here suggests this evolutionary history: different MuERV-L proviruses have evolved under selective constrains during the first burst of amplification, pointing out the existence of several rounds of retrotransposition (Table 2). The period of quiescence following this burst probably resulted in the accumulation of mutations in the inserted proviruses, leading to the random inactivation (or reduced transposition capability) of several of them. Because of that, when the second amplification burst started, only one or a few proviruses from the first burst were active enough to propagate within the genome. At this moment, intragenomic competition between different elements might be an important factor to determine the future success of the provirus (Jordan and McDonald 1998, 1999; Costas and Naveira 2000). Following this initial propagation, several proviruses from the progeny have remained active, probably until present, acting as templates for new transpositions. This is clearly indicated by a series of facts, such as the d_N/d_S estimates (Table 2), revealing that several elements were not merely evolving as pseudogene copies since the beginning of the burst; the existence of two groups of elements showing identical LTR sequences (Fig. 1), the presence of several pairs of sequences in the phylogenetic tree with high bootstrap support; and the existence of proviruses from this burst sharing ORF1 of different lengths (Fig. 2). Repetition of a cycle of amplification bursts and periods of quiescence such as this one is expected to lead to a single lineage of elements over time.

This scheme is quite similar to that reported in the case of the human endogenous retrovirus HERV-K. While during long-term evolution (such as during primate divergence), a single lineage of elements seems to exist (Medstrand and Mager 1998), several active lineages of HERV-K proviruses remained transpositionally active after the human/chimpanzee split (Costas 2001). Unfortunately, the paucity of similar data on short-term evolution of different ERV

Table 2. Results of the ML analyses

ML model	d_N/d_S 1st burst	d_N/d_s 2nd burst	log-likelihood	χ^2
Tree topology of Fig. 2				
One ratio	0.3061	0.3061	−13121.301384	
Two ratio	0.3940	0.2135	−13111.371196	19.86
Tree topology from actual data				
One ratio	0.3029	0.3029	−13078.110799	
Two ratio	0.3909	0.2094	−13068.010716	20.20

families precludes the generalization of this evolutionary dynamics.

The putative impact of active MuERV-Ls within the mouse genome nowadays, as well as the factor(s) responsible for the initial reactivation leading to the amplification bursts remain to be elucidated. In the meantime, taking into account the propensity of the *Mus* genus to sprout new species outside its geographic origin whenever migration was possible (Boursot et al. 1993), it is attractive to imagine that the geographical spread might be one of the "reactivation" factors, in accordance with the hypothesis of transposable elements awakening following colonization of different regions by the host species (Vieira et al. 1999).

Acknowledgments. The author is a recipient of a postdoctoral fellowship from the USC/Xunta de Galicia.

References

Altschul SF, Gish W, Miller W, Myers W, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Bénit L, De Parseval N, Casella JF, Callebaut I, Cordonnier A, Heidmann T (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 *Fvl* restriction gene. J Virol 71:5652–5657

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

Boursot P, Auffray JC, Britton-Davidian J, Bonhomme F (1993) The evolution of house mice. Annu Rev Ecol Syst 24:119–152

Cordonnier A, Casella JF, Heidmann T (1995) Isolation of novel human endogenous retrovirus-like elements with foamy virus-related *pol* sequence. J Virol 69:5890–5897

Costas J (2001) Evolutionary dynamics of the human endogenous retrovirus family HERV-K inferred from full-length proviral genomes. J Mol Evol 53:237–243

Costas J, Naveira H (2000) Evolutionary history of the human endogenous retrovirus family ERV9. Mol Biol Evol 17:320–330

Herniou E, Martin J, Miller K, Cook J, Wilkinson M, Tristem M (1998) Retroviral diversity and distribution in vertebrates. J Virol 72:5955–5966

Huelsenbeck JP, Crandall KA (1997) Phylogeny estimation and hypothesis testing using maximum likelihood. Annu Rev Ecol Syst 28:437–466

Huelsenbeck JP, Rannala B (1997) Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science 276:227–232

Jordan IK, McDonald JF (1998) Interelement selection in the regulatory region of the *copia* retrotransposon. J Mol Evol 47:670–676

Jordan IK, McDonald JF (1999) The role of interelement selection in *Saccharomyces cerevisiae Ty* element evolution. J Mol Evol 49:352–357

Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: Molecular Evolutionary Genetics Analysis software. Distributed by the authors at http://www.megasoftware.net/

Martin J, Herniou E, Cook J, Waugh O'Neill R, Tristem M (1997) Human endogenous retrovirus type 1-related viruses have an apparently widespread distribution within vertebrates. J Virol 71:437–443

Medstrand P, Mager DL (1998) Human-specific integrations of the HERV-K endogenous retrovirus family. J Virol 72:9782–9787

Nicholas KB, Nicholas Jr HB (1997) GeneDoc: a tool for editing and annotating multiple sequence alignment. Distributed by the authors at http://www.cris.com/~ketchup/genedoc.shtml

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Tatusova TA, Madden TL (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett 174:247–250

Vieira C, Lepetit D, Dumont S, Biemont C (1999) Wake up of transposable elements following *Drosophila simulans* worldwide colonization. Mol Biol Evol 16:1251–1255

Yang Z (1996) Maximum-likelihood models for combined analyses of multiple sequence data. J Mol Evol 42:587–596

Yang Z (2000) Phylogenetic analysis by maximum likelihood (PAML), Version 3.0. University College London, London, England