Evolutionary History of the Human Endogenous Retrovirus Family ERV9

Javier Costas and Horacio Naveira†*

*Departamento de Bioloxía Fundamental, Facultade de Bioloxía, Universidade de Santiago de Compostela, Santiago de Compostela, Spain; and †Departamento de Bioloxía Celular e Molecular, Facultade de Ciencias, Universidade de A Coruña, A Coruña, Spain

Several distinct families of endogenous retrovirus-like elements (ERVs) exist in the genomes of primates. Despite the important evolutionary consequences that carrying these intragenomic parasites may have for their hosts, our knowledge about their evolution is still scarce. A matter of particular interest is whether evolution of ERVs occurs via a master lineage or through several lineages coexisting over long periods of time. In this work, the paleogenomic approach has been applied to the study of the evolution of ERV9, one of the human endogenous retrovirus families mobilized during primate evolution. By searching the GenBank database with the first 676 bp of the ERV9 long terminal repeat, we identified 156 different element insertions into the human genome. These elements were grouped into 14 subfamilies based on several characteristic nucleotide differences. The age of each subfamily was roughly estimated based on the average sequence divergence of its members from the subfamily consensus sequence. Determination of the sequential order of diagnostic substitutions led to the identification of four distinct lineages, which retained their capacity of transposition over extended periods of evolution. Strong evidence for mosaic evolution of some of these lineages is presented. Taken altogether, the available data indicate that the possibility of ERV9 still being active in the human lineage can not be discarded.

Introduction

A significant proportion of the human genome consists of interspersed repetitive DNA sequences. The main mode of dispersion of these sequences seems to have been retrotransposition in the germ line. This process includes transcription of the DNA template by RNA polymerase, reverse transcription into DNA by reverse transcriptase, and insertion into a new genomic location, thus increasing the number of genomic copies of the sequence.

Alu and L1 are the major families of human interspersed repeated DNA, amounting to 10% and 15% of the genome, respectively (Kazazian and Moran 1998). The evolutionary histories of these two families have been determined by comparative sequence analysis of a number of different inserted copies (Shen, Batzer, and Deininger 1991; Deininger et al. 1992; Smit et al. 1995). The major conclusion of this so-called paleogenomic approach has been that most Alu and L1 elements were produced by only one or a few source genes (‘‘master elements’’) at any one time during evolution of the family. Accumulation of substitutions in the master elements, or replacement of these masters by others from their own derivatives, can give rise to distinct subfamilies of pseudogene copies defined by shared differences with the consensus of the family. These characteristic (or diagnostic) nucleotide positions can be sequentially ordered, with the youngest subfamilies sharing older differences as well as new diagnostic substitutions. Thus, the evolution of L1 and Alu families has been shown to be driven essentially by a single lineage of master elements (‘‘sequential masters’’ or ‘‘molecular drivers’’). The existence of several minor subfamilies besides the main lineage suggests that from time to time, elements distinct from the sequential masters may retain their capability of retrotransposition over a significant period of time.

Another interesting type of repetitive DNA elements consists of retrovirus-like elements (RLEs), or endogenous retroviruses (ERVs), representing about 1% of the human genome (Lowër, Lowër, and Kurth 1996). Their structure closely resembles that of retroviruses, carrying internal sequences with homology to *gag, pol,* and, sometimes, *env* open reading frames (ORFs) flanked by long terminal repeats (LTRs), which contain several transcriptional regulatory sequences. A large number of solitary LTRs that have arisen by homologous recombination between the 5′ and the 3′ LTRs of full-length elements are scattered throughout the mammalian genomes (Lowër, Lowër, and Kurth 1996).

RLEs can be regarded as highly specialized intragenomic parasites, with their parasitic capability being the product of natural selection among RLE copies within the genome (Doolittle and Sapienza 1980; Orgel and Crick 1980). Evolutionary consequences for the host of carrying these intragenomic parasites have been a matter of debate for years. Apparently, the majority of characterized RLE insertions do not have an adaptative value, so RLEs may be considered insertional mutagens that affect host genes either by direct disruption of their coding sequences or by altering their expression patterns (Leslie, Lee, and Schrader 1991; Wu et al. 1993; Mitreiter et al. 1994). Nevertheless, it has been suggested that the ERV LTRs may be an evolutionary tool for coupling expression of regulatory factors and their receptors (Kapitonov and Jurka 1999), or they may provide a cellular reservoir of control elements with diverse transcriptional specificities (Keshet, Schiff, and Itin 1991). Furthermore, regulation of human genes by ERV homologous sequences has been shown or suggested in

Abbreviations: ERV, endogenous retrovirus; LTR, long terminal repeat; RLE, retrovirus-like element.

Key words: ERV9, endogenous retrovirus, mosaic evolution, source gene, master model.

Address for correspondence and reprints: Javier Costas, Departamento de Bioloxía Fundamental, Facultade de Bioloxía, Universidade de Santiago de Compostela, Campus Sur s/n E-15706 Santiago de Compostela (A Coruña), Spain. E-mail: bfcostas@usc.es.

Mol. Biol. Evol. 17(2):320–330. 2000

several cases (Suzuki et al. 1990; Ting et al. 1992; Di Cristofano et al. 1995*a*; Long et al. 1998; Kapitonov and Jurka 1999). Finally, some data suggest that human ERVs may be involved in the prevention of infections with related exogenous retroviruses or act as pathological agents in certain autoimmune disorders (Lowër, Lowër, and Kurth 1996; Patience, Wilkinson, and Weiss 1997).

Despite these important biological interactions between ERVs and mammalian genomes, very little is known about evolution of ERV families (Clough et al. 1996). Although it would seem likely that most copies of an active RLE are able to retrotranspose, a few observations suggest an expansion of mammalian ERVs similar to the master gene model proposed for L1 and Alu families (Lee et al. 1996; Medstrand and Mager 1998).

ERV9 elements, first identified by La Mantia et al. (1991), constitute one of the families of ERVs mobilized during primate evolution (Di Cristofano et al. 1995*b*). This family is represented in the human genome by 30–40 members besides at least 4,000 solitary LTRs (La Mantia et al. 1989, 1991; Lania et al. 1992). The prototype ERV9 element is about 8 kb long, including the two LTRs of approximately 1.8 kb; their precise size depends on a variable number of two tandemly repeated subelements. One of these subelements is 41 bp long, with about 12 repeats within the U3 region, whereas the other is 72 bp long, with about 4 repeats in the U5 region (La Mantia et al. 1991; Lania et al. 1992).

To get a better insight into the evolution of ERV families, we conducted a comparative sequence approach to reconstruct the evolutionary history of the ERV9 family using sequences gathered from GenBank. Our results clearly indicate the existence of several lineages with distinct evolutionary success which retained their capacity of transposition over extended periods of primate evolution.

Materials and Methods

Identification of ERV9 LTRs was made by screening the nonredundant database at the National Center of Biotechnology Information with the first 676 bp of the ERV9 LTR described in Di Cristofano et al. (1995*a*; accession number X83497) using the program BLASTN (Altschul et al. 1990). Additional searches were done with members of different subfamilies. A total of 156 ERV9 homologous fragments from different elements were identified. CLUSTAL X (Thompson et al. 1997) was used for sequence alignments, which were later refined by visual inspection with GeneDoc (Nicholas and Nicholas 1997).

Element subfamilies were established by grouping sequences into different sets according to shared nucleotide deletions and nucleotides observed in the most variable sites of the alignment. Although several of these sites were CpG positions, with little discrimination for subfamilies due to the fast mutation rate of these dinucleotides to TpG or CpA, characteristic differences could be detected occasionally. Further inspection of initial subsets led to the identification of correlated diagnostic positions so that these previous groups could be further split. A nucleotide position was considered diagnostic of a sequence set whenever more than 70% of the sequences grouped into it shared the same nucleotide, which differed from that characterizing at least some other similar groups. Subfamily status was conferred on a sequence set when five or more sequences of the group contained at least three diagnostic positions.

Consensus sequences for each subfamily and for the whole sample (‘‘general’’ consensus sequence) were obtained by choosing the more frequent nucleotide at each position, with one exception. If a combination of dinucleotides of the three pairs CpG, CpA, and TpG were present at the same doublet position, the CpG dinucleotide was chosen as the consensus dinucleotide unless the T or A nucleotides were present in more than 70% of the sequences. When either of two bases occurred in a site with 50% frequency (ambiguous site), IUPAC-IUB base codes representing multiple bases were used (Cornish-Bowden 1985).

The accuracy of these consensus sequences was evaluated by comparing observed and expected pairwise divergences between sequences from the same subfamily, considering the consensus sequence the best reconstruction of the source gene (genes) that originated the subfamily. Expected divergences were calculated by the following formula (Smit et al. 1995):

$$d = d_1 + d_2 - (4d_1d_2/3) \qquad (1)$$

where d is the divergence between two sequences evolved independently from the same founder gene, and d_1 and d_2 are the divergence values of these two sequences compared with their source gene. To apply the formula, we assumed that $d_1 = d_2 = d_{avr}$, where d_{avr} is the average divergence of subfamily sequences from their consensus, calculated as the relative number of mismatches, excluding gaps and CpG dinucleotides in the calculations. The observed divergence value was calculated as the average of all pairwise differences between the sequences from the same subfamily after exclusion of gaps and CpG dinucleotides.

To estimate the ages of ERV9 subfamilies, we first calculated the average level of nucleotide substitutions from the consensus (K), excluding CpG dinucleotides and gaps, using Kimura’s (1981) two-parameter model with the usual transition/transversion ratio of 2. This method of estimation of the actual number of nucleotide substitutions, originally derived for two sequences diverged from a common ancestor, is also valid when one of the two sequences is the unchanged source gene (Kapitonov and Jurka 1996). Assuming 0.16% per million years as the rate of change of pseudogene sequences in primates (Britten 1994; Kapitonov and Jurka 1996), the average transposition age of the subfamily (T) has been estimated as $T = K/0.0016$.

Comparison of the rates of change between CpG and non-CpG positions was done after correction of CpG divergence by $d_{corr} = -\ln(1 - d_{obs})$, according to Labuda et al. (1991). d_{obs} at CpG dinucleotide positions was calculated again as the relative number of mismatches after removing all gaps from the alignment. The programs DNADIST from the PHYLIP package (Felsenstein 1993) and DnaSP3 (Rozas and Rozas 1999) were used to calculate some of these values.

The phylogenetic analysis of subfamily consensus sequences using maximum parsimony was performed with the program DNAPARS from the PHYLIP package (Felsenstein 1993). Since the probability of parallel or convergent evolution is much lower for indels than for point mutations, each indel received a higher weight than point mutations (with one indel being counted as three point mutations, irrespective of their length, as a conservative weight). The bootstrap analysis was performed with 1,000 replications using the programs SEQBOOT, DNAPARS, and CONSENSE from the PHYLIP package (Felsenstein 1993).

Analysis of mosaicism was performed as described by Robertson, Hahn, and Sharp (1995). This method is based on the distribution of phylogenetically informative sites supporting alternative tree topologies among four taxa: the putative mosaic sequence, one representative of each of the two ‘‘parental’’ lineages, and a known outgroup. SimPlot (Ray 1999) was used to carry out a sliding-window analysis of bootstrap values determined by maximum parsimony.

Results

Our analysis of ERV9 elements focused on the first 676 bp of the LTR, since the rest is mainly constituted by tandemly repeated subelements and an approach based on these regions was expected to be less reliable. Alignment of 156 human ERV9 sequence fragments allowed us to classify them into 14 different subfamilies on the basis of consistent correlated nucleotide differences from the general consensus sequence (synapomorphic combination of character states, i.e., diagnostic nucleotide substitutions). Eight of the elements presented diagnostic features of different subfamilies scattered along the sequence and were excluded from the analysis (table 1). The relative abundances subfamilies were very different, ranging from a mere 3% up to a maximum of 26% of the examined sequences (table 2).

The alignment of the general consensus sequence with consensus sequences for the 14 identified subfamilies, showing their characteristic nucleotide differences, is presented in figure 1. There are also three characteristic deletions, at positions 32–40, 338–381, and 438. Many of these differences are shared by several subfamilies, suggesting that subfamilies could be placed in a sequential order. Determination of this sequential order of diagnostic differences led us to the identification of four lineages (A–D, fig. 2). Lineage D displays diagnostic features from both lineage B and lineage C clustered around different regions. This fact suggests the mosaic nature of lineage D. The distribution of phylogenetically informative sites in an alignment including subfamily consensus sequences II, VII, IX, and XIII (representing an outgroup, the two parental lineages, and the putative mosaic sequence, respectively) clearly in-

Table 1
Identified ERV9 Elements

Subfamily I: AC004774, 140052; AC004911, 18925; AC005006, −45960; Z93020, −15385; AL021407a, 11864

Subfamily II: AC004595a, −52466; AL008627, −88275; AC002422, −10080; AC005332c, −146655; AC004513, 64900; AF080618, −2993

Subfamily III: AC004197b, −28432; AC002301, 46801; AC006222, 52246; AC003092, 65224; AC004993, −84902; AL022324, −52913; AC005722, 180259; AC005837, 148025

Subfamily IV: AC002090, 33486; Z83820, −117572; AC004919, −38697; AC003085, 104250; AL021451a, 8355; AC005326, 52166; L78833, −2068

Subfamily V: AC005576, 5695; X58467, −7576; AC006065, 65634; AL031274, −123331; AC004940, 23212

Subfamily VI: Z97192, 11323; AC005023, −78982; AC005296, 3023; AC006572, 92301; AC000125, 20347; D10450, 765; X15674, 743; AC002064, 126513; Z81365, 90779; AC004047, 133318

Subfamily VII: AC005332b, −120655; AC005730, −45107; AC004817, −57085; AC003013, −82653; AL031116, −103454; AC005527, −126284; AC005245, 924; AC003026, −49031; Z71182, −57949; AL022396a, 11504; AC000111, 135674; AC004894, 27985; AL22396b, 27094; AL008731, −62619; AC002312, −8140

Subfamily VIII: AC005500, 148225; AC006212, 66525; Z84474a, −59146; AC004595b, −127765; AC004534, −73836; AC004549, −89139; AC006007, 25396; Z69647, 21416

Subfamily IX: AC005336, 21384; AC004197a, −24511; AL031785, 11171; AC004783, 51255; Z75746, 4381; AC003670, 7917; AF001550, 19829; AL024506, 58076; AC006271, 69451; X83497, 82; AC005332a, 18859; AL022067, −175153; AL022726a, 55393; AC002544, 88211; Z99128, 104694; AL031073, −108462; AC004185, −24371; AC004866, −114817; AC003043, −19946; AC004220, −16940; AC004000, −67546; Z99290, 62163; AC004004, −35884; AC005066b, −161617; Z73416, 39587; AC002477, 66944; AL022326, 9396; Z72004, 35485; AC004063, 69366; Z81450, 25305; AC003029, 44254; X14975, −1043; AC005392, −70786; AC002451, −13293; AC002326b, −122045; Z82975, 72026; AL021407b, 140768; AC004973b, 9369; Z84484, −113572; AC004212a, 19482; AC000058, −8430

Subfamily X: AC005332d, 158408; AL021939, 109901; AD000091, −36952; AC004954, −51973; Z98745, −88576; Z99297, 5818

Subfamily XI: AB020858, −51578; AC004973a, 7647; AL034416, −5449; AC005016, −107861; AC004212b, 28340; AL031054, −109788; AC004973c, −64104

Subfamily XII: Z84474b, −65295; AC004668, 79271; Z92547, 57194; AC006157, −81510; AC000048, 12988; AC003986, 179735

Subfamily XIII: AC002326a, −63975; AC006406, −42209; AC005611, −77825; AC005066a, −11297; AL022726b, 102851; AL021451b, 42254; AC000120, −35525; AC006236, −47064; Z98941, −54932; AC000378, 12877; AC003049, −66856; Z85997, 34337; Z95126, −163202; AC004831, 51795

Subfamily XIV: AC002067, −78488; AL022146, −10345; AF064190, 2660; AC005297, −136866; AC004216, 107749; D42052, 3663; AC002487, −29010; AC005539, 177701; AL031679, 55335; AC002314, −17962

Unclassified: AL009175, 46927; AC004109, −55609; Z82242, −63292; AC005606, −17827; AL022069, 122994; U73627, 26322; AC004923, −67112; AC000385, 114878

NOTE.—Each element is identified by its GenBank file designation, followed by the nucleotide position of the 5′ end of the analyzed region. Several of the GenBank entries correspond to sequences in progress, so their 5′-end positions may vary in the final version of the file. GenBank entries containing more than one ERV9 element are indicated by a lowercase letter following the accession number. A minus sign indicates sequence orientation opposite to the long terminal repeat.

Table 2
Ages of ERV9 Subfamilies

Subfamily	No. of Sequences (%)	No. of Nucleotides[a]	Kimura's Distance[b]	Age (myr)[c]
I	5 (3)	516	0.061	38.1
II	6 (4)	479	0.057	35.6
III	8 (5)	419	0.050	31.2
IV	7 (4)	352	0.052	32.5
V	5 (3)	466	0.045	28.1
VI	10 (6)	390	0.047	29.3
VII.	15 (10)	463	0.038	23.7
VIII	8 (5)	375	0.030	18.7
IX	41 (26)	435	0.030	18.7
X	6 (4)	558	0.033	20.6
			0.026[d]	16.2[d]
XI	7 (4)	585	0.026	16.2
XII.	6 (4)	586	0.021	13.1
XIII	14 (9)	445	0.042	26.2
XIV	10 (6)	479	0.025	15.6

[a] Number of nucleotides used for the calculation of the age of the subfamily, after removing gaps and CpG positions from the alignment.

[b] Average level of nucleotide substitutions from the subfamily consensus sequence.

[c] Estimated average age of the subfamily.

[d] Calculated after removing AC005332d (see text).

dicates this mosaicism. Although the small number of informative sites makes the precise position of the breakpoints difficult to determine, this mosaic structure was corroborated when phylogenetic trees were constructed with bootstrap analysis for the resulting regions (fig. 3).

Another unexpected cluster of nucleotide differences has been detected. Subfamily I shares characteristic nucleotide differences with lineage C at three closely linked positions (233, 257, and 279), but, surprisingly, these three differences are not observed in other subfamilies (fig. 1).

Although the majority of differences appearing in a master gene are shared by subsequent source genes, some subfamilies show a few private differences (fig. 1). This may be indicative that these subfamilies are not the direct ancestors of the younger subfamilies of their lineage.

The subfamily consensus sequence is the best reconstruction of the active sequence that yielded members of the subfamily by retrotransposition. After their insertions, most of these elements are not functionally constrained and accumulate mutations at a neutral rate. Thus, the average divergence of the members of each subfamily from their respective subfamily consensus sequence would reflect the age of the subfamily. This is not true for CpG dinucleotides, whose mutation rate, caused by deamination of the methylated cytosine, is very high (Bird 1980). Therefore, CpG positions were excluded from the estimation of the average number of nucleotide substitutions. Obviously, these estimates should be considered approximate due to the differences in local mutation rates and the small size of the sequence sample (the average coefficient of variation was 0.28). As it appears in table 2, the estimated ages of individual subfamilies are in general good agreement with their sequential order within each lineage based on shared ancestry (fig. 2). A conspicuous exception is subfamily X. Although this subfamily shares five more sequence variants than IX with younger subfamilies of its lineage, it seems to be clearly older than subfamily IX according to the average within-subfamily divergence. This incongruence disappears if sequence AC005332d (2.5 times as divergent as the average) is excluded from the analysis of subfamily X.

Evolution of active elements may be inferred from the comparison between consensus sequences of the oldest and youngest subfamilies (I and XII, respectively). Of the 63 observed nucleotide differences (excluding the two ambiguous sites, 162 and 253, in the consensus sequence of subfamily I), 29 correspond to CpG dinucleotides, which represents 21.6% and 6.3% of CpG and non-CpG sites affected, respectively. After correction of observed divergence for CpG positions, due to its nonlinear substitution rate (see *Materials and Methods* for details), we obtained a CpG/non-CpG ratio of 3.9. To compare this value with that for pseudogene copies of ERV9, we used subfamily IX, accounting for 55% of the sequences for the lineage. Interestingly, the obtained value (d_{obs} at CpG = 0.19, d_{corr} = 0.21, CpG/non-CpG = 3.8) is quite similar to that for active elements.

Discussion

This paper presents a reconstruction of the evolutionary history of the human endogenous retrovirus family ERV9 based on a ''paleogenomic'' approach. We must take into account that these analyses only detect master genes that have been active at a relatively high level over an extended period of time (Deininger at al. 1992). Thus, it is possible that a much larger fraction of ERV9 elements have been capable of retrotransposition, but they must have been silenced rapidly on an evolutionary timescale. Another consideration is that the classification of ERV9 in 14 subfamilies is somewhat arbitrary, since further splitting of these subfamilies will probably be possible as new sequences become available. This is clearly suggested by the slight excess of

```
               *        20         *        40         *        60         *        80         *
      TGAGAGGTGACAGCGTGCTGGCAGcCCTCgCAGCCCTCGCTCGCTCTCGGCGCCTCCTCgGCCTcGGCgCCCACTcTGGCcGCgCTTGAG
I   : .............................A....................T......................T................
II  : ..............A................---------..................................................
III : ...............................---------..................................................
IV  : ............A......A...........---------.............................T....................
V   : ............A..................---------.............................T.TG........A........
VI  : ..........................................................................................
VII : .............................A..................................T.........................
VIII: ................................................................T.........................
IX  : ........................T....A.............................T....G...T......T....G..A......
X   : ........................T....A.............................T....G...T......T....G..A......
XI  : ........................T....Ag..........T.........A.......T....G...T......T....G..A......
XII : ........................T....AG..........T.........A.....C.T....G...T......T...TG..AT.....
XIII: ........................t....A.............................T....G...T......T....G..A......
XIV : ........................T....A.............................T....G...T......T....G..A......

       100         *       120         *       140         *       160         *       180
      GAGCCCTTCAGCCCGCCGCTGCACTGTGGGAGCCCCTTTCTGGGCTGGCCgAGGCCGGAGCCGGCTCCCTCAGCTTGCGGGGAGGTGTGG
I   : ........................C..........T.C.................................R..................
II  : ...................................t..c...................................................
III : .....................................C.................................T..................
IV  : .....................................C.................................T..................
V   : .....................................C.................................T..................
VI  : ......................................Y...................................................
VII : ..................................................A.......................................
VIII: ..................................................A.......................................
IX  : ..............A...................................A...........................A...........
X   : ..............A...................................A...........CA..............A...........
XI  : ..................................................A....T......CA..............A...........
XII : ..............C..A................................A...........CA..............A...........
XIII: ..................................................A.......................................
XIV : ..................................................A.......................................

               *       200         *       220         *       240         *       260         *
      AGGGAGAGGCGCGGGCGGGAACCGGGGCTGCGCGCGGCGCTTGCGGGCCAGCgcGAGTTCCGGGTGGGCGTGGGCTcGGCGGGCCCCGCA
I   : ..........A.........................................TA..................R...T.............
II  : ..........................................................................................
III : .........................................C................................................
IV  : .........................................c............................c...................
V   : .........................................c............................................T...
VI  : ..........................................................................................
VII : ..........................................................................................
VIII: ..........................................................................................
IX  : .............A......................................TG......................T.............
X   : .............A..................T...................TG......................T.............
XI  : .............A..................T...................TG......................T.............
XII : ..........A..A..................T...................TG......................T.............
XIII: .....................................T....................................................
XIV : .....................................T............................................A.......

       280         *       300         *       320         *       340         *       360
      CTCGGAGCgGCCGGCCGGCCCCGCCGGCCCCGGGCAGTGAGGGGCTTAGCACCCGGGCCAGCAGCTGCGGAGGGTGcgCcGGGTCCCCCA
I   : ........A.................................................................................
II  : ..........................................................................................
III : ...A...............g......................................................................
IV  : ...................g......................................................................
V   : ...................g......................................................................
VI  : ..........................................................................................
VII : .....................................................T..............----------------------
VIII: ........................aA...........................T..............----------------------
IX  : ........A............T..............A.........................G.............TA.T..........
X   : ........A............T..............A......A..................G.............TA.T..........
XI  : ........A.......A....T..T...........A......A.................TG.............TA.T..........
XII : ........A...A...A....T..T...........A..G...A.................TG.............TA.T..........
XIII: .....................................................T..............----------------------
XIV : ......................A..............................T..............----------------------
```

FIG. 1.—Alignment of subfamily consensus sequences. The “general” consensus sequence is shown above as a reference sequence. Dots indicate identity with this reference sequence. Capital letters indicate nucleotides present in more than 70% of the sequences belonging to a subfamily, and lowercase letters indicate nucleotides present in between 50% and 70% of the sequences grouped in the subfamily. R = A or G, and Y = T or C. Consensus sequences at positions 462–474 for subfamilies II, V, VII, and VIII are not reliable due to the high length polymorphism of DNA sequences in this region. In consequence, no changes at this interval were taken into account for these subfamilies.

```
               *       380         *       400         *       420         *       440         *
      GCAgTGCcgGCCCgCCgGCgCTGcGCTCGATTTCTCGCCGGGCCTTAGCTGCCTcCCCGCGGGGCAGGGCTCGGGACCTGCAGCCCGCCA
I   : .......T.....A............................................................................
II  : .............A............................................................................
III : .....................c........A..............C...c........................................
IV  : ...C............T..A.c.T......A..............C...C........................................
V   : ...c............C....CA.......A.....A..A.....C...c...................................T....
VI  : ..........................................................................................
VII : ---------------------..T......C...........................................................
VIII: ---------------------..T..................................................................
IX  : .............A........................................T...................................
X   : ........A....A........................................T.............C........-............
XI  : ........A....A......................A...A.............T...................................
XII : .............A........................................T...A...............................
XIII: ---------------------..T....A.........................T...................................
XIV : ---------------------..T....A.........................T...................................

             460         *       480         *       500         *       520         *       540
      TGCCTGAGCCTCCCCCCCCCgCCGTGGGCTCCTGcGCGGCCCGAGCCTCCCCGACGAGCGCCGCCCCCTGCTCCACGGCGCCCaGTCCCA
I   : ........T................................T.........................................G......
II  : ...................................................................................G......
III : ....c...........................c.......................g.................g........g......
IV  : ....C.....C.........c...........c.......................g.................g........G......
V   : ....C.....c.....................C.......................G.................g.......TG......
VI  : ...................................................................................g......
VII : ..........................................................................................
VIII: ..........................................................................................
IX  : ..............A.....T.............T.......................................................
X   : ...........T.....G..T.......T.....T..A....................................................
XI  : ..............A...A.T..A..........T........................A..A...........................
XII : ..............A...A.T..A..........T........................A..A...........................
XIII: ....................T.............T.......................................................
XIV : ....................t.............T.......................................................

               *       560         *       580         *       600         *       620         *
      TCGACCaCCCAAGGGCTGAGGAGTGCGGGCGCACGGCGCGGGACTGGCgGGCAGCTCCACCTGCgGCCCCGGTGCGGGATCCACTgGGTG
I   : ......G..........................g.....A....................................A.............
II  : ......G.....................................................................a.............
III : ......G...................A.....g.........................g..c.......................A..c.
IV  : .....TG...................A.....g.........................g..c..........C............A..c.
V   : ......g.........................G.........................G..c..........c............A..c.
VI  : ......G.........................................A.........................................
VII : ................................................A.........................................
VIII: ......g.........................................A.........................................
IX  : ................................................A...............A.........................
X   : ...........................A.....T..............A...............A.........................
XI  : ......................a....A....................A...............A....................A....
XII : ......................A....A...........A........A...............A....................A....
XIII: ................................................A.........................................
XIV : ................................................A...............A.........................

             640         *       660         *
      AAGCCAGCTGGGCTCCTGAGTCTGGTGGGGACtTGGAGAACCTTTA
I   : .......................A......................
II  : .......................R......................
III : ..............................................
IV  : ......................G..................T....
V   : ......................g..................T....
VI  : .......................a......................
VII : ..............................................
VIII: ..............................................
IX  : ................................G.............
X   : ................................G.............
XI  : ................................G......GT.....
XII : ................................G......GT.....
XIII: ..............................................
XIV : ................................G.............
```

FIG. 1 *(Continued)*

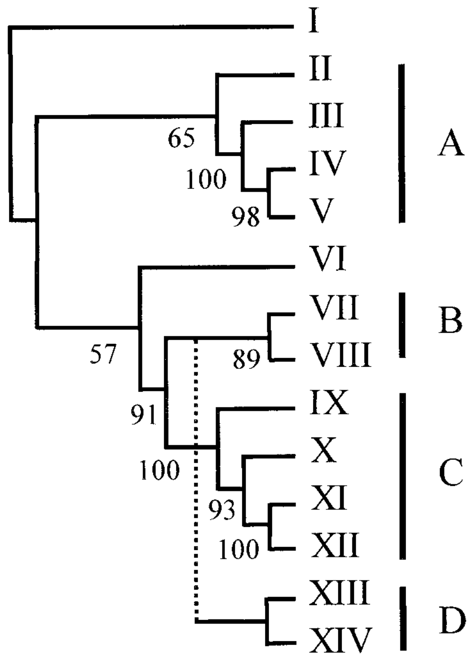

FIG. 2.—Phylogenetic relationships of the various ERV9 subfamilies, based on one of the two most-parsimonious trees obtained from the analysis of the subfamily consensus sequences (excluding those from lineage D). Positions 233, 257, and 279 were removed from the analysis (see text). The other most-parsimonious tree differs in placement of VI (considered the closest outgroup of lineage A). Bootstrap values (%) supporting the clusters are indicated. The vertical dotted line represents the putative recombination event that gave rise to lineage D. Proposed lineages (A–D) are shown on the right.

expected over observed pairwise divergence between sequences belonging to the same subfamily (table 3).

The main features of the evolutionary history of ERV9 elements, inferred from our age estimation of their different subfamilies, are illustrated in figure 4, although it should be kept in mind that this is a rough reconstruction, based on the questionable assumption of 0.16% per million years as the rate of change of pseudogene copies of ERV9. The first subfamily detected in our analysis probably appears after the split of New World and Old World monkeys. Early in ERV9 evolution, lineage A undergoes successive expansions, being the predominant lineage over an extended period of primate evolution. The other lineages begin to spread after the split from Old World monkeys but before the separation of gibbons and higher apes. Expansion of these lineages coincides with the disappearance of detectable subfamilies of lineage A, probably due to its competitive exclusion. The major expansions of the ERV9 family occur from the divergence of gibbons from higher apes until the split of gorillas. Although subfamily IX is the predominant one (26% of the identified sequences belong to this subfamily), several distinct subfamilies from this same lineage, as well as from others, originate at this period of time. This high transpositional activity ceases after the divergence of gorillas (approximately 8–10 MYA). No new subfamilies have been found since then.

Most ERV families have been detected by Southern hybridization or genomic PCR analysis in Old World monkeys but not in New World monkeys, suggesting that they are approximately 30–40 Myr old (Mariani-Costantini, Horn, and Callahan 1989; Shih, Coutavas, and Rush 1991). Alternatively, sequence divergence between New World and Old World monkeys may be the cause of these negative results. Although our estimation of transpositional ages suggests the absence of ERV9 subfamilies prior to the New World monkey split, it does not mean that ERV9 appeared in the germ line of primates after this split. The presence of ERV9 in low copy numbers before the New World monkey divergence cannot be discarded. For instance, detection of HERV-H–related sequences has been reported in New World monkeys by means of PCR analysis, but expansions of HERV-H began after New World/Old World monkey divergence (Mager and Freeman 1995). An older origin, such as that for ERV-L, detected in several mammalian orders (Benit et al. 1999), seems less plausible.

The time of insertion of two ERV9 elements included in our study has been analyzed experimentally by different authors. One of them is an element inserted into the ZNF80 locus (GenBank accession number X83497). This insertion has been detected by genomic PCR at the same position in gorillas, chimpanzees, and humans, but it is absent from the orangutan genome (Di Cristofano et al. 1995*b*). This result is in good agreement with our classification of this element within subfamily IX, whose estimated age of transposition is 18.7 MYA. The other element (GenBank accession number AF064190) is located in the β-globin locus of humans and gorillas, the only two primate genomes tested (Long et al. 1998). This element belongs to subfamily XIV, whose transpositional age was estimated as 15.6 MYA.

The major finding of our study was the existence of several lineages of ERV9 elements simultaneously active over long periods of time. This fact contrasts with previous data on other mammalian RLEs, such as HERV-K and the retrotransposon *mys* of *Peromyscus.* Unlike ERV9, the HERV-K and *mys* subfamilies seem to arise in a sequential order from the same master lineage and expand during different periods of time (Clough et al. 1996; Lee et al. 1996; Medstrand and Mager 1998). Thus, whether the main mode of expansion of RLEs occurs via a master lineage or through several competitive lineages is an open question.

Our results clearly indicate that ERV9 elements have been actively retrotransposing over an extended period of primate evolution. Nevertheless, the propagation of ERV9 was not at all constant through this time. Subfamily IX alone represents 26% of all of the sampled ERV9 sequences, and its period of activity overlaps at least that of subfamily VIII (table 2); both together account for one third of the total recorded ERV9 insertions. In view of the extended period of ERV9 activity within primate genomes as well as its putative discontinuous transpositional dynamics, it is likely that some ERV9 elements retained its transpositional activity after the split of great apes. Thus, the presence of a functionally transposing subset of ERV9 in the human lineage,

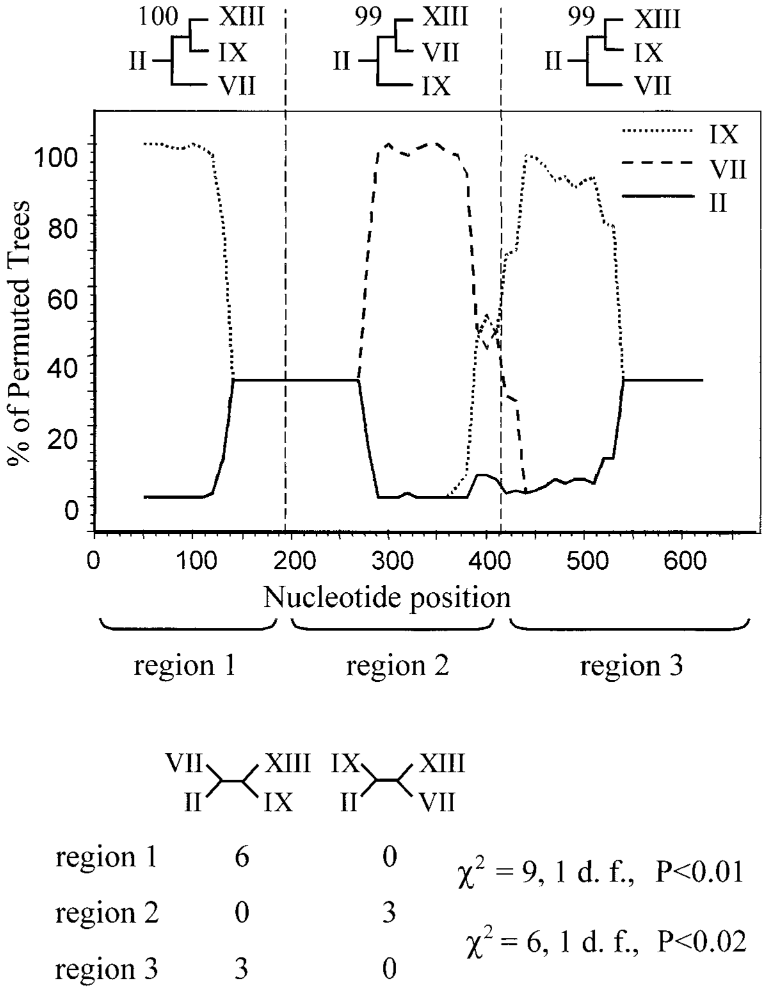

FIG. 3.—Mosaic structure of lineage D; sliding-window analysis across the alignment (on the x axis) of the number of permuted trees that group the putative mosaic sequence with each of the other three (on the y axis). Window size = 100 nucleotides; step size = 10 nt. Vertical lines indicate approximate positions of recombination break points. Phylograms obtained for each region, with corresponding bootstrap values, are shown above. Numbers of informative sites supporting each topology (including the deletion characteristic of lineages B and D as a single site), together with results of χ^2 heterogeneity tests, are shown below.

Table 3
Comparison Between Observed and Expected Divergence

Subfamily	Observed Divergence[a]	Expected Divergence[b]
I.	10.9	11.2
II	10.2	10.0
III	8.8	9.3
IV	9.3	9.6
V	8.2	8.5
VI	8.2	8.7
VII.	6.7	7.1
VIII	5.5	5.6
IX	5.5	5.7
X	6.1	6.3
XI	4.8	5.0
XII.	3.9	4.1
XIII	7.0	7.3
XIV	4.8	4.9

[a] Average pairwise differences (%) between the sequences from the subfamily.

[b] Expected divergence value (%) between sequences belonging to the subfamily (see *Materials and Methods*).

as described for HERV-K (Medstrand and Mager 1998), may be more probable than it is for other types of ERV sequences.

It has been suggested that mosaic evolution (by novel combination of preexisting mutations) may be a likely mechanism during the evolution of repetitive elements, facilitated by the existence of a large pool of genomic copies (Kass, Batzer, and Deininger 1995; Kido et al. 1995; Zietkiewicz and Labuda 1996). Furthermore, RLEs are expected to be especially prone to genetic rearrangements due to the possibility of recombination between two RNA genomes packaged within the same capsid (McDonald 1993). Our results clearly suggest three cases of mosaic evolution, caused by either recombination or gene conversion. One of them has given rise to lineage D by combination of sequences from lineage B and C (fig. 3). The relative evolutionary success of this new assembly of sequences is revealed by the two expansions of lineage D during different periods. Another two putative cases of mosaic evolution of ERV9 elements have been detected. One is constituted by the

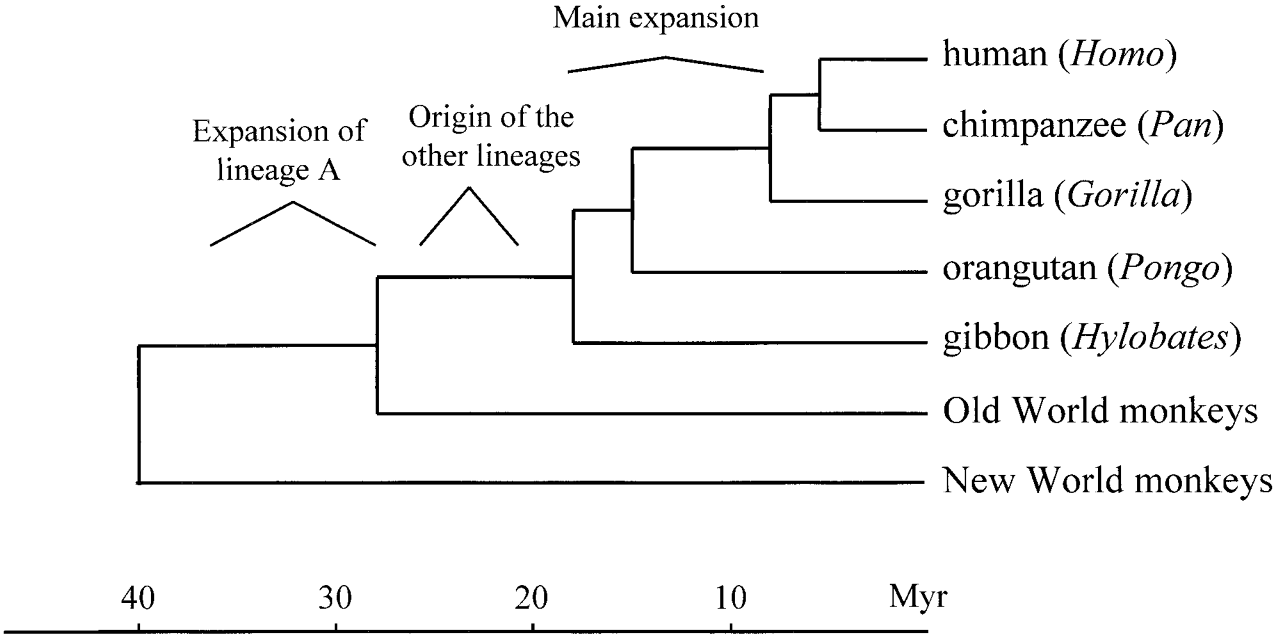

FIG. 4.— Inferred evolutionary history of ERV9 elements superimposed on a phylogenetic tree of primate evolution. Estimations of ERV9 transpositional ages are based on average divergences of members of each subfamily from their respective consensus sequences. The branch point times, taken from Gingerich (1984), Sibley and Ahlquist (1987), and Stewart and Disotell (1998), should be regarded as approximate.

three closely linked characteristic differences (233, 257, and 279) shared by subfamily I and lineage C, and the other is present in subfamily X. This subfamily acquired eight private differences (autopomorphs), including a deletion. Six of these differences are clustered in the interval from position 429 to position 488. Taken together, these findings indicate that mosaic evolution is a likely mechanism in ERV9 evolution. Interestingly, the youngest subfamily of HERV-H LTRs also originated from a rearrangement of preexisting mutations, in this case by recombination between the other two described subfamilies (Mager 1989).

Finally, it is worth mentioning that the rate of substitutions at CpG dinucleotides during the evolution of active ERV9 sequences is clearly higher than the substitution rate at non-CpG positions. Furthermore, the CpG/non-CpG substitution ratio of source genes is similar to that of pseudogene copies of ERV9. This contrasts with the situation found in the Alu family, whose mutation rate in CpG dinucleotides of source genes is as low as that in non-CpG sites (Deininger and Slagel 1988; Labuda and Striker 1989). These data suggest that the active ERV9 sequences for retrotransposition are not protected from cytosine methylation despite the fact that one of the main functions for cytosine methylation within mammalian genomes seems to be suppression of parasitic elements by transcription inhibition (Yoder, Walsh, and Bestor 1997; but see Bird 1997 for opposite explanations).

Acknowledgments

We thank the Xunta de Galicia of Spain for financial support (XUGA 10304B97). J.C. was supported by a predoctoral fellowship from the Universidade de Santiago de Compostela (Spain).

LITERATURE CITED

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS, and D. J. LIPMAN. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

BENIT, L., J. B. LALLEMAND, J. F. CASELLA, H. PHILIPPE, and T. HEIDMANN. 1999. ERV-L elements: a family of endogenous retrovirus-like elements active throughout the evolution of mammals. J. Virol. **73**:3301–3308.

BIRD, A. 1980. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. **8**:1499–1504.

———. 1997. Does DNA methylation control transposition of selfish elements in the germline? Trends Genet. **13**:469–470.

BRITTEN, R. J. 1994. Evidence that most human *Alu* sequences were inserted in a process that ceased about 30 million years ago. Proc. Natl. Acad. Sci. USA **91**:6148–6150.

CLOUGH, J. E., J. A. FOSTER, M. BARNETT, and H. A. WICHMAN. 1996. Computer simulation of transposable element evolution: random template and strict master models. J. Mol. Evol. **42**:52–58.

CORNISH-BOWDEN, A. 1985. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. Nucleic Acids Res. **13**:3021–3030.

DEININGER, P. L., M. A. BATZER, C. A. HUTCHISON III, and M. H. EDGELL. 1992. Master genes in mammalian repetitive DNA amplification. Trends Genet. **8**:307–311.

DEININGER, P. L., and V. K. SLAGEL. 1988. Recently amplified *Alu* family members share a common parental *Alu* sequence. Mol. Cell. Biol. **8**:4566–4569.

DI CRISTOFANO, A., M. STRAZZULLO, L. LONGO, and G. LA MANTIA. 1995*a*. Characterization and genomic mapping of the ZNF80 locus: expression of this zinc-finger gene is driven by a solitary LTR of ERV9 endogenous retroviral family. Nucleic Acids Res. **23**:2823–2830.

Di Cristofano, A., M. Strazzullo, T. Parisi, and G. La Mantia. 1995*b*. Mobilization of an ERV9 human endogenous retroviral element during primate evolution. Virology **213**:271–275.

Doolittle, W. F., and C. Sapienza. 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature **284**: 601–603.

Felsenstein, J. 1993. PHYLIP (phylogeny inference package). Version 3.5c. Distributed by the author (http://evolution.genetics.washington.edu/phylip.html), Department of Genetics, University of Washington, Seattle.

Gingerich, P. D. 1984. Primate evolution: evidence from the fossil record, comparative morphology, and molecular biology. Yearb. Phys. Anthropol. **27**:57–72.

Kapitonov, V., and J. Jurka. 1996. The age of Alu subfamilies. J. Mol. Evol. **42**:59–65.

———. 1999. The long terminal repeat of an endogenous retrovirus induces alternative splicing and encodes an additional carboxy-terminal sequence in the human leptin receptor. J. Mol. Evol. **48**:248–251.

Kass, D. H., M. A. Batzer, and P. L. Deininger. 1995. Gene conversion as a secondary mechanism of short interspersed element (SINE) evolution. Mol. Cell. Biol. **15**:19–25.

Kazazian, H. H., and J. V. Moran. 1998. The impact of L1 retrotransposons on the human genome. Nat. Genet. **19**:19–24.

Keshet, E., R. Schiff, and A. Itin. 1991. Mouse retrotransposons: a cellular reservoir of long terminal repeat (LTR) elements with diverse transcriptional specificities. Adv. Cancer Res. **56**:215–51.

Kido, Y., M. Saitoh, S. Murata, and N. Okada. 1995. Evolution of the active sequences of the *HpaI* short interspersed elements. J. Mol. Evol. **41**:986–995.

Kimura, M. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA **78**:454–458.

Labuda, D., D. Sinnett, G. Richer, J.-M. Daregon, and G. Striker. 1991. Evolution of the mouse B1 repeat: 7SL RNA folding pattern conserved. J. Mol. Evol. **32**:405–414.

Labuda, D., and G. Striker. 1989. Sequence conservation in Alu evolution. Nucleic Acids Res. **17**:2477–2491.

La Mantia, G., D. Maglione, G. Pengue, A. Di Cristofano, A. Simeone, L. Lanfrancone, and L. Lania. 1991. Identification and characterization of novel human endogenous retroviral sequences preferentially expressed in undifferentiated embryonal carcinoma cells. Nucleic Acids Res. **19**: 1513–1520.

La Mantia, G., G. Pengue, D. Maglione, A. Pannuti, A. Pascucci, and L. Lania. 1989. Identification of new human repetitive sequences: characterization of the corresponding cDNAs and their expression in embryonal carcinoma cells. Nucleic Acids Res. **17**:5913–5920.

Lania, L., A. Di Cristofano, M. Strazzullo, G. Pengue, B. Majello, and G. La Mantia. 1992. Structural and functional organization of the human endogenous retroviral ERV9 sequences. Virology **191**:464–468.

Lee, R. N., J. C. Jaskula, R. A. Van Den Bussche, R. J. Baker, and H. A. Wichman. 1996. Retrotransposon *mys* was active during evolution of the *Peromyscus leucopus-maniculatus* complex. J. Mol. Evol. **42**:44–51.

Leslie, K. B., F. Lee, and J. W. Schrader. 1991. Intracisternal A-type particle-mediated activations of cytokine genes in a murine myelomonocytic leukemia: generation of functional cytokine mRNAs by retroviral splicing events. Mol. Cell. Biol. **11**:5562–70.

Long, Q., C. Bengra, C. Li, F. Kutlar, and D. Tuan. 1998. A long terminal repeat of the human endogenous retrovirus ERV-9 is located in the 5′ boundary area of the human β-globin locus control region. Genomics **54**:542–555.

Lowër, R., J. Löwer, and R. Kurth. 1996. The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. Proc. Natl. Acad. Sci. USA **93**:5177–5184.

McDonald, J. F. 1993. Evolution and consequences of transposable elements. Curr. Opin. Genet. Dev. **3**:855–864.

Mager, D. L. 1989. Polyadenylation function and sequence variability of the long terminal repeats of the human endogenous retrovirus-like family RTVL-H. Virology **173**: 591–599.

Mager, D. L., and J. D. Freeman. 1995. HERV-H endogenous retroviruses: presence in the New World branch but amplification in the Old World primate lineage. Virology **213**: 395–404.

Mariani-Costantini, R., T. Horn, and R. Callahan. 1989. Ancestry of a human endogenous retrovirus family. J. Virol. **63**:4982–4985.

Medstrand, P., and D. L. Mager. 1998. Human-specific integrations of the HERV-K endogenous retrovirus family. J. Virol. **72**:9782–9787.

Mitreiter, K., J. Schmidt, A. Luz, M. J. Atkinson, H. Hofler, V. Erfle, and P. G. Strauss. 1994. Disruption of the murine p53 gene by insertion of an endogenous retrovirus-like element (ETn) in a cell line from radiation-induced osteosarcoma. Virology **200**:837–841.

Nicholas, K. B., and H. B. Nicholas Jr. 1997. GeneDoc: a tool for editing and annotating multiple sequence alignment. Distributed by the authors (http://www.cris.com/∼ketchup/genedoc.shtml).

Orgel, E., and F. H. C. Crick. 1980. Selfish DNA: the ultimate parasite. Nature **284**:604–607.

Patience, C., D. A. Wilkinson, and R. A. Weiss. 1997. Our retroviral heritage. Trends Genet. **13**:116–120.

Ray, S. C. 1999. SimPlot for Windows. Version 2.5. Distributed by the author (http://www.med.jhu.edu/deptmed/sray/download/), Baltimore, Md.

Robertson, D. L., B. H. Hahn, and P. M. Sharp. 1995. Recombination in AIDS viruses. J. Mol. Evol. **40**:249–359.

Rozas, J., and R. Rozas. 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15**:174–175.

Shen, M. R., M. A. Batzer, and P. L. Deininger. 1991. Evolution of the master Alu gene(s). J. Mol. Evol. **33**:311–320.

Shih, A., E. E. Coutavas, and M. G. Rush. 1991. Evolutionary implications of primate endogenous retroviruses. Virology **182**:495–502.

Sibley, C. G., and J. E. Ahlquist. 1987. DNA hybridization evidence of hominoid phylogeny: results from an extended data set. J. Mol. Evol. **26**:99–121.

Smit, A. F. A., G. Toth, A. D. Riggs, and J. Jurka. 1995. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J. Mol. Biol. **246**:401–417.

Stewart, C. B., and T. R. Disotell. 1998. Primate evolution—in and out of Africa. Curr. Biol. **8**:R582–R588.

Suzuki, H., Y. Hosokawa, H. Toda, M. Nishikimi, and T. Ozawa. 1990. Common protein-binding sites in the 5′-flanking regions of human genes for cytochrome c1 and ubiquinone-binding protein. J. Biol. Chem. **265**:8159–8163.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25**: 4876–4882.

Ting, C., M. Rosenberg, C. Snow, L. Samuelson, and M. Meisler. 1992. Endogenous retroviral sequences are required for tissue specific expression of a human salivary amylase gene. Genes Dev. **6**:1457–1465.

Wu, J., T. Zhou, J. He, and J. D. Mountz. 1993. Autoimmune disease in mice due to integration of an endogenous retrovirus in an apoptosis gene. J. Exp. Med. **178**:461–468.

Yoder, J. A., C. P. Walsh, and Y. H. Bestor. 1997. Cytosine methylation and the ecology of intragenomic parasites. Trends Genet. **13**:335–340.

Zietkiewicz, E., and D. Labuda. 1996. Mosaic evolution of rodent B1 elements. J. Mol. Evol. **42**:66–72.

Thomas H. Eickbush, reviewing editor

Accepted November 1, 1999