

Human Endogenous Retroviral Elements as Indicators of Ectopic Recombination Events in the Primate Genome

Jennifer F. Hughes¹ and John M. Coffin²

Department of Molecular Microbiology and Program in Genetics, Tufts University School of Medicine, Boston, Massachusetts 02111

Manuscript received April 12, 2005

Accepted for publication June 2, 2005

ABSTRACT

HERV elements make up a significant fraction of the human genome and, as interspersed repetitive elements, have the capacity to provide substrates for ectopic recombination and gene conversion events. To understand the extent to which these events occur and gain further insight into the complex evolutionary history of these elements in our genome, we undertook a phylogenetic study of the long terminal repeat sequences of 15 HERV-K(HML-2) elements in various primate species. This family of human endogenous retroviruses first entered the primate genome between 35 and 45 million years ago. Throughout primate evolution, these elements have undergone bursts of amplification. From this analysis, which is the largest-scale study of HERV sequence dynamics during primate evolution to date, we were able to detect intraelement gene conversion and recombination at five HERV-K loci. We also found evidence for replacement of an ancient element by another HERV-K provirus, apparently reflecting an occurrence of retroviral integration by homologous recombination. The high frequency of these events casts doubt on the accuracy of integration time estimates based only on divergence between retroelement LTRs.

ENDOGENOUS retroviruses arise from retroviral infections of germ-line cells and subsequent incorporation into the host's genome (BOEKE and STOYE 1997). The human genome is estimated to contain tens of thousands of copies of human endogenous retroviruses (HERVs) and related sequences, accounting for ~8% of its sequence content (LANDER *et al.* 2001; PACES *et al.* 2002). Most HERVs appear to have entered the genome of our ancestors between 30 and 45 million years ago, after the divergence of Old and New World monkeys (SVERDLOV 2000), although elements as old as 55 million years are known (BANNERT and KURTH 2004; LAVIE *et al.* 2004). Because of their relatively long residence in the genome, the majority of HERV elements are riddled with deleterious mutations, large deletions, and insertions of other repetitive elements. In addition, the exogenous viral counterparts of HERVs have most likely been extinct for many millions of years, but it is clear that these elements have expanded in copy number significantly throughout the course of primate evolution, perhaps through intracellular retrotransposition mechanisms (LEIB-MOSCH *et al.* 1993; GOODMAN *et al.* 1998; COSTAS and NAVEIRA 2000) as well as replication as viruses and reinfection (BELSHAW *et al.*

2004). The HERV-K family in particular has been actively expanding in our genome throughout the last 5–20 million years, during the period of radiation of hominids (MEDSTRAND and MAGER 1998; BARBULESCU *et al.* 1999; HUGHES and COFFIN 2001, 2004).

Many human endogenous retrovirus families appear to have a complicated evolutionary history (JOHNSON and COFFIN 1999; COSTAS and NAVEIRA 2000). As is true for repetitive elements in general, in addition to expansion in the genome through transposition mechanisms, these elements can also undergo unequal crossing over and gene conversion (KASS *et al.* 1995; ROY *et al.* 2000; ROY-ENGEL *et al.* 2002). The capacity to mediate such ectopic recombination events has led to speculation that repetitive elements may be major contributors to the genome plasticity of their host species (LOWER *et al.* 1996).

It has been demonstrated that the unique structural features of endogenous retroviruses can be especially useful for the detection of ectopic recombination events and that these types of events may have occurred at a number of human endogenous retroviral loci (JOHNSON and COFFIN 1999). The process of reverse transcription generates a DNA copy of the retroviral RNA genome that is identical to the original, except that it is flanked by two identical long terminal repeat (LTR) sequences, which are 968 bp long in the case of HERV-K. Subsequent to integration, the two LTRs of the provirus evolve independently, acquiring mutations and diverging from each other. The longer the provirus has been resident in the genome of its host, the more divergent

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. AY884837–AY884981.

¹Present address: Whitehead Institute, 9 Cambridge Center, Cambridge, MA 02142.

²Corresponding author: Tufts University School of Medicine, 136 Harrison Ave., Boston, MA 02111. E-mail: john.coffin@tufts.edu

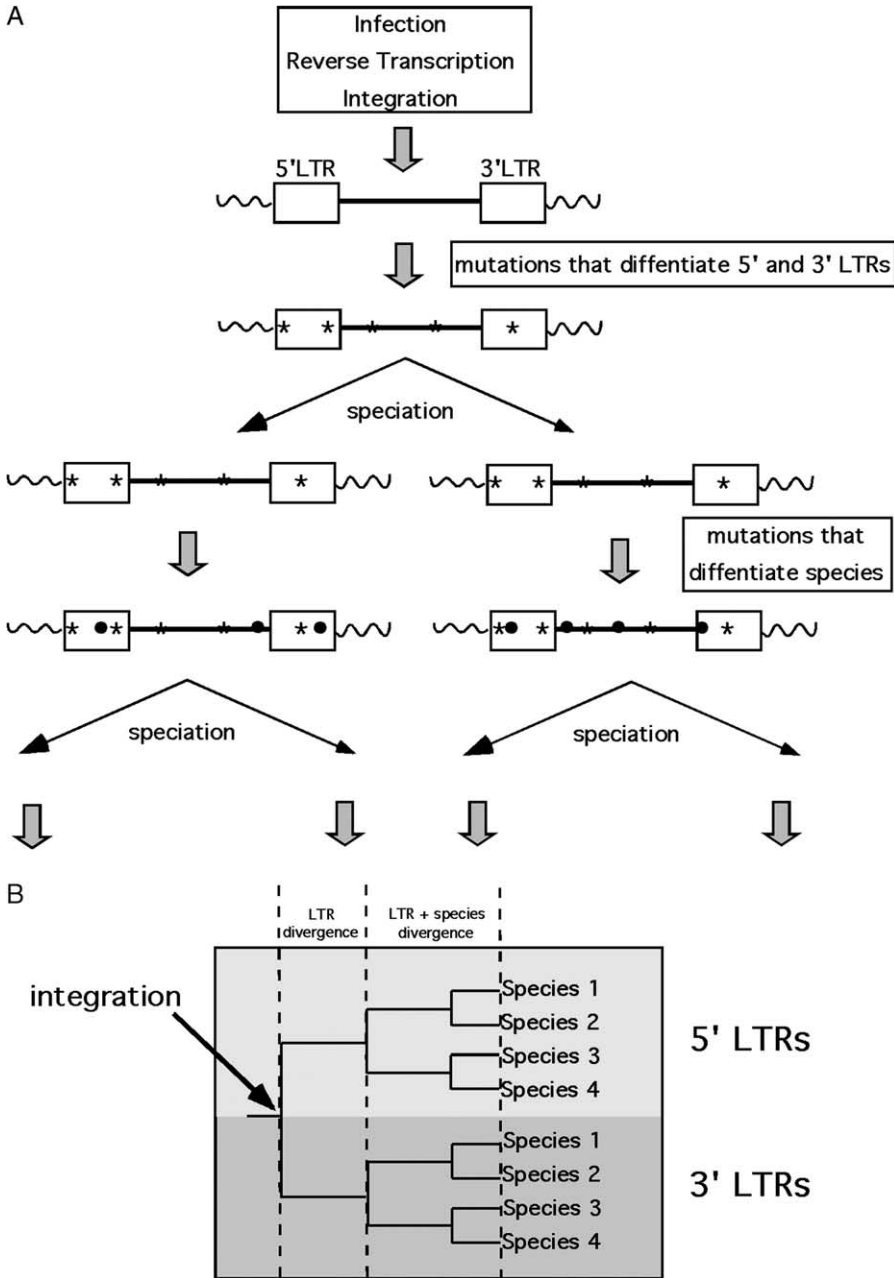


FIGURE 1.—Endogenous retrovirus evolution. (A) An integrated provirus. LTR sequences are represented by boxes flanking the internal viral sequence (straight line). Cellular DNA is depicted as wavy lines. Mutations are depicted as asterisks or dots. Two types of mutations can be distinguished: those that differentiate the 5' and 3' LTRs in all descendants (asterisks), which occur after integration but prior to speciation, and those that differentiate species, which occur subsequent to speciation (dots). (B) Tree that results from phylogenetic analysis of the 5' and 3' LTR sequences in four different species.

are its LTRs. Hence, LTRs can serve as a molecular clock estimate for the integration time of the provirus (DANGEL *et al.* 1995). However, estimates can be confounded if the assumption of an independent evolutionary history for the LTRs, specifically no significant sequence homogenization, is not correct.

One way to test this assumption is by examining the phylogenetic relatedness of these elements in their host species. A phylogenetic analysis of the LTRs of a given provirus that is shared in multiple species should result in two separate clusters containing the 5' and 3' LTR sequences separated by the deepest node in the tree, because the LTR sequences begin to diverge immediately after integration, prior to subsequent speciation events (Figure 1). The two main branches of the tree

should then be independent estimations of the evolutionary history of the host species. Any deviations from a pattern of concordance of one LTR with the other or with the accepted evolutionary history of the host species indicate the occurrence of concerted evolution within the element or between elements. In the absence of such events, HERVs can serve as useful markers for studying the evolutionary relatedness of their primate hosts, because this type of analysis provides two estimates of phylogenetic relationships. However, when ectopic recombination events are detected, they can provide insight into the nature and mechanism of genome rearrangement that can occur during evolution.

The goal of this study was to analyze a set of closely related proviruses, the HERV-K(HML2) group,

to determine the nature and frequency of such events in primate evolution, as well as the utility of using LTR divergence alone as a measure of time since integration. We found that at least one-third of the proviruses examined have been subjected to ectopic recombination. Thus, these events are quite common in our evolutionary history, and molecular clocks based on LTR divergence alone may often give incorrect estimates of integration times.

MATERIALS AND METHODS

Primate DNA samples: Genomic DNA samples were isolated from fibroblast or lymphoblast cell lines using the QIAamp DNA blood maxikit (QIAGEN, Valencia, CA). The following cell lines were obtained from Coriell Cell Repositories: chimpanzee (*Pan troglodytes*, repository no. GM03448A), orangutan (*Pongo pygmaeus*, GM04272), rhesus macaque (*Macaca mulatta*, AG06252), stump-tailed macaque (*M. arctoides*, GM03442), and squirrel monkey (*Saimiri sciureus*, AG05311). Prepared DNA was obtained from Coriell for the following species: celebes macaque (*M. nigra*, NG07101) and patas monkey (*Erythrocebus patas*, NG06116). The baboon cell line (*Papio papio*, 26CB-1) was obtained from ATCC. Bonobo (*P. paniscus*), gorilla (*Gorilla gorilla*), and gibbon (*Hylobates concolor*) fibroblast cell lines were generously provided by Stephen J. O'Brien. The African green monkey (*Cercopithecus aethiops*) cell line, COS-1, was generously provided by Ralph R. Isberg.

PCR and sequencing: PCR reactions contained 200 ng genomic DNA, 1.5–3.5 mM MgCl₂, 50 μM each dNTP, 0.2 μM each primer, and 2.5 units Taq DNA polymerase (Sigma, St. Louis). To amplify the 5' LTR, the left primer was designed to hybridize in the 5' genomic flanking region and the right primer was located just downstream of the 5' LTR. To amplify the 3' LTR, the left primer was located just upstream of the 3' LTR and the right primer was located in the 3' genomic flanking sequence. Primers and conditions used for each reaction are available upon request. PCR products were sequenced directly on both strands using an automated ABI 3100 DNA sequencer.

Estimation of integration time: Mutation rates for each element were estimated independently for each HERV-K locus. First, pairwise distances were calculated using the Kimura two-parameter correction in the PAUP*4.0b program for all homologous sequence pairs among the species for which sequence information was available. Each distance was then divided by the years since the two species in each pair shared a last common ancestor on the basis of accepted primate divergence dates (GOODMAN *et al.* 1998). These were then averaged to obtain a mutation rate for each HERV-K locus (substitutions per site per year). Next, the divergences between the two LTRs of each element were calculated. The corrected distances between the 5' and 3' LTR sequences of each element within species were averaged to obtain LTR divergence values (substitutions per site) for each element. The LTR divergence value was then divided by the mutation rate to give the integration time estimate for each HERV-K element.

Phylogenetic analysis and sequence analysis: Dot matrix plots were generated in MacVector 7.0. Multiple sequence alignments were performed using the CLUSTAL-W algorithm in MacVector 7.0, using open and extend gap penalties of 20 and 0, respectively. Resulting alignments were adjusted by hand. Maximum-parsimony and maximum-likelihood analyses were performed using PAUP*4.0b. Bootstrap values were

calculated from 100 replicate trees. Trees were edited using MacClade 3.08. The human 5' and 3' LTR sequences of a closely related HERV-K provirus were used as outgroups in the primate phylogenetic analyses.

RESULTS

Species distribution of full-length HERV-K elements and estimation of age: The 15 HERV-K elements used in this study are listed in Table 1 and were all identified by conducting BLAST searches of the human genome database (HUGHES and COFFIN 2001, 2004) with the exception of HERV-K(II), which was previously identified (SUGIMOTO *et al.* 2001). They are named according to their chromosomal location. PCR primers were designed on the basis of human sequence information to specifically amplify the entire 5' LTR and 3' LTR sequences, including at least 30 bp of flanking sequence. The primate genomes tested for the presence of each HERV-K element were the chimpanzee, bonobo, gorilla, orangutan, gibbon, several Old World monkeys (African green monkey, baboon, rhesus macaque, stump-tailed macaque), and a New World monkey (squirrel monkey). Human genomic DNA was used as a positive control for each set of primers (see Figure 2 for an example). Only one LTR of some of the elements was amplified in some species, possibly indicating that the other LTR is deleted or that the primers used may have been too divergent for efficient amplification in some species because they were designed only on the basis of human sequence. In all cases of incomplete amplification, multiple primer pairs were designed and tested and, in all but two cases, also gave negative results. Orthology of the amplified sequences was confirmed by examining the sequence flanking of one or both LTRs, if present. The PCR products were then sequenced on both strands for analysis. In species for which both the 5' and 3' LTR assays were negative for a given HERV locus, attempts were made to amplify the corresponding intact preintegration site, but were unsuccessful in all cases. Therefore, the species distributions determined in this study reflect only the lower limits of the ages of each provirus.

To test the fidelity of the LTR molecular clock, the mutation rate at each HERV locus was determined for each LTR independently and the integration time of each provirus was then estimated on the basis of the divergence between the two LTRs. This value was then compared with minimum integration times estimated from the species distribution (Table 1). The mutation rates had values between 2.48×10^{-9} and 4.49×10^{-9} substitutions/site/year, similar to the rates found previously at HERV loci (JOHNSON and COFFIN 1999) as well as in pseudogenes (NACHMAN and CROWELL 2000), probably reflecting their selective neutrality.

Of the 15 elements, only nine estimated ages correlated with the observed species distribution. Two

TABLE 1
Species distribution and estimated integration time of HERV-K elements

HERV-K	Chromosomal position ^a	Most distant species in which provirus was found ^b	Mutation rate (substitutions/site/yr) ($\times 10^{-9}$)	Average divergence between LTRs	Estimated integration time (MYA)	Date of last common ancestor (MYA) ^c
4q32	166274445–166281673	Chimpanzee	4.49 \pm 0.8	0.038	7.2–10.5	6
HERV-K(II) (chromosome 3)	102893427–102902549	Gorilla	2.64 \pm 0.5	0.017	4.9–5.9	7
12q24	132277472–132283414	Gorilla	4.17 \pm 0.1	0.021	6.6–9.8	7
10p14	6906147–6915609	Gorilla	3.82 \pm 0.6	0.040	9.0–12.6	7
19p13.11A	22549664–22556401	Gorilla	4.40 \pm 0.9	0.054	10.3–15.4	7
22q11	22204481–22215171	Gorilla	3.63 \pm 0.7	0.11	28.6–38.9	7
9q34.3	136950603–136960065	Orangutan	5.66 \pm 0.4	0.067	11.1–12.7	14
3p25	9864346–9871236	Orangutan	4.10 \pm 0.8	0.066	13.4–19.8	14
1q23	163306258–163311916	Orangutan	3.73 \pm 0.2	0.062	15.9–17.3	14
19p13.11B	20248400–20258515	Orangutan	4.26 \pm 0.1	0.12	26.4–28.1	14
11q12	61892539–61907139	Gibbon	3.41 \pm 0.3	0.065	17.5–21.0	18
19q13.1	42289389–42298906	Gibbon	2.48 \pm 0.7	0.066	21.0–36.3	18
6p22	28758347–28768714	Gibbon	2.69 \pm 0.3	0.076	25.0–32.4	18
20q11	32179289–32188037	OWM	3.51 \pm 0.6	0.053	12.8–18.3	25
6p21	42969390–42979344	OWM	3.14 \pm 0.9	0.030	7.4–13.1	25

^a Nucleotide position based on May 2004 UCSC genome assembly (<http://genome.cse.ucsc.edu/>).

^b Based on the ability to PCR amplify the HERV element in each species. Not all species contained both LTRs of the listed element. Attempts to amplify preintegration sites or solo LTRs from more distant species were unsuccessful.

^c Date of the last common ancestor of the lineage leading to humans and the most divergent species in which the provirus was found (GOODMAN *et al.* 1998).

elements (HERV-K22q11 and 19p13.11B) had LTRs that were more divergent and therefore appeared to be more ancient than their species distribution indicated. Two elements, HERV-K20q11 and HERV-K6p21, which were the only elements in this study that we were able to amplify from the genomes of Old World monkeys, gave significant underestimates, while HERV-K(II) and HERV-K9q34.3 showed relatively modest discrepancies. Homogenization of the 5' and 3' LTR sequences of these elements through interelement recombination and/or gene conversion events could ac-

count for their low level of divergence. To address this issue, phylogenetic analyses of all elements were performed as described in the next section.

Phylogenetic analyses: Deviation from the predicted evolutionary relationships among the LTRs in the different species should indicate whether recombination or gene conversion events have occurred and if they can account for the incorrect estimates of evolutionary distance. For this reason, separate trees were constructed using maximum-parsimony analysis for the LTR sequences of each element amplified from the various primate species. Figure 3 shows the phylogenetic analyses for the nine elements whose trees conformed to the predicted topology. The 5' and 3' LTR sequences of these elements formed separate clusters, and the species relationships in both sections of the tree were similar to one another and to the expected primate phylogeny. These results indicate that the two LTRs of these elements evolved independently, along with their hosts, and their evolutionary history did not involve detectable ectopic recombination or gene conversion events.

Detection of gene conversion and recombination events at HERV-K loci: Figure 4 shows the results of the phylogenetic analyses of the remaining six elements, all of which deviated from the predicted structure. As expected, the four elements for which the integration time was underestimated are present in this group, and more detailed analyses of their evolutionary histories follow. We first consider the two proviruses with concordant estimates of integration times.



FIGURE 2.—PCR strategy used to amplify LTRs of specific HERV-K elements in primates. The positions of the primer pairs used to amplify the 5' and 3' LTR sequences are shown relative to that of the HERV-K element. One primer from each pair is located in the unique genomic sequence flanking the HERV so that only orthologous loci are amplified. An example of the results of this PCR strategy is shown, following agarose gel electrophoresis and ethidium bromide staining.

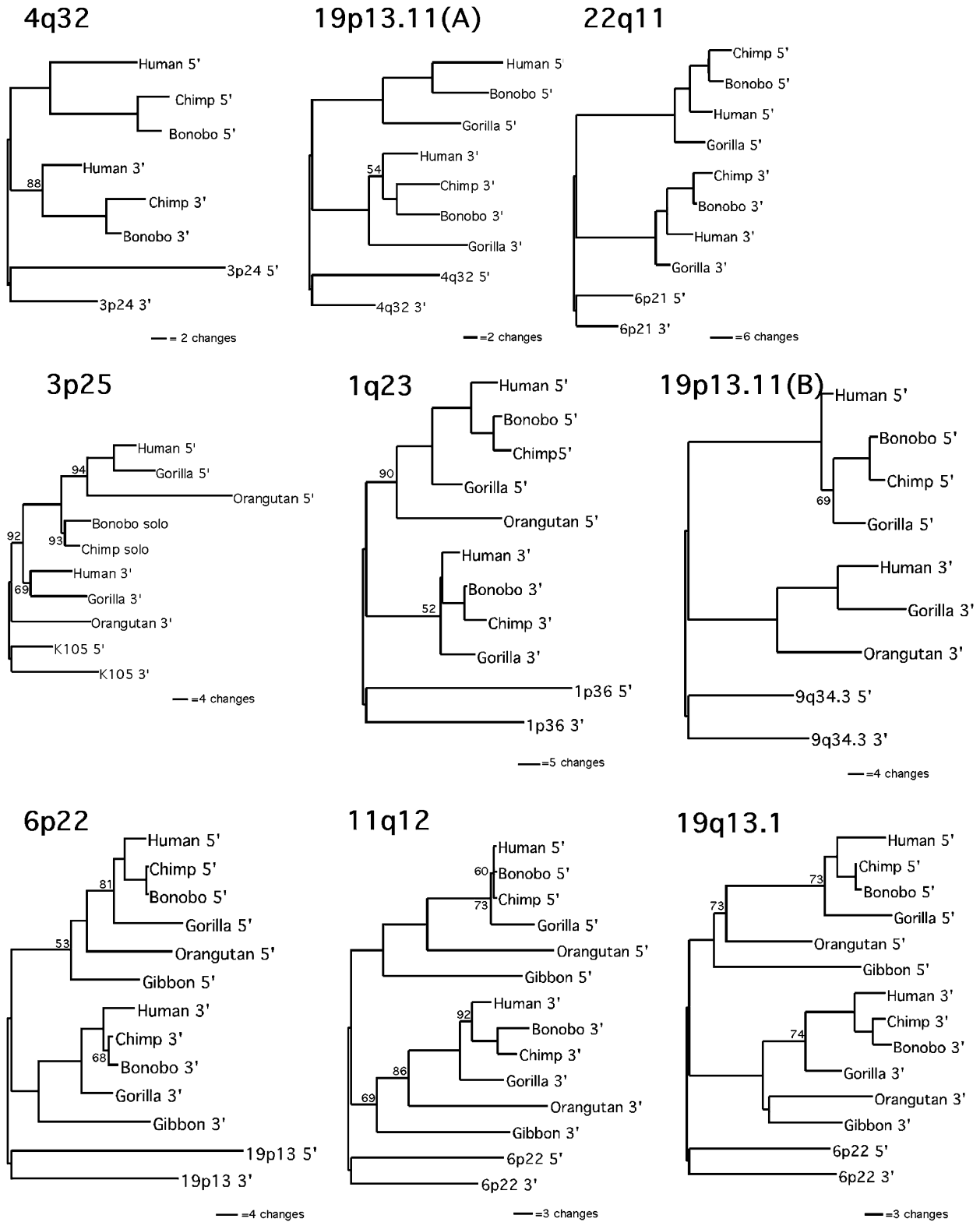


FIGURE 3.—Phylogenetic analysis of HERV-K elements that conform to predicted topology. Maximum-parsimony trees are shown for each element. Branch lengths are proportional to the number of changes occurring along a lineage. Bootstrap values taken from 100 replicates are shown. Nodes without indicated bootstrap values had very high support, $\geq 95\%$. Human 5' and 3' LTR sequences of closely related HERV-K elements were used as outgroups in each analysis and are indicated by element name. HERV-K3p25 is not full length in the chimpanzee and bonobo, but a solitary LTR, which is formed by homologous recombination between the 5' and 3' LTRs, was found at this locus and included in the analysis.

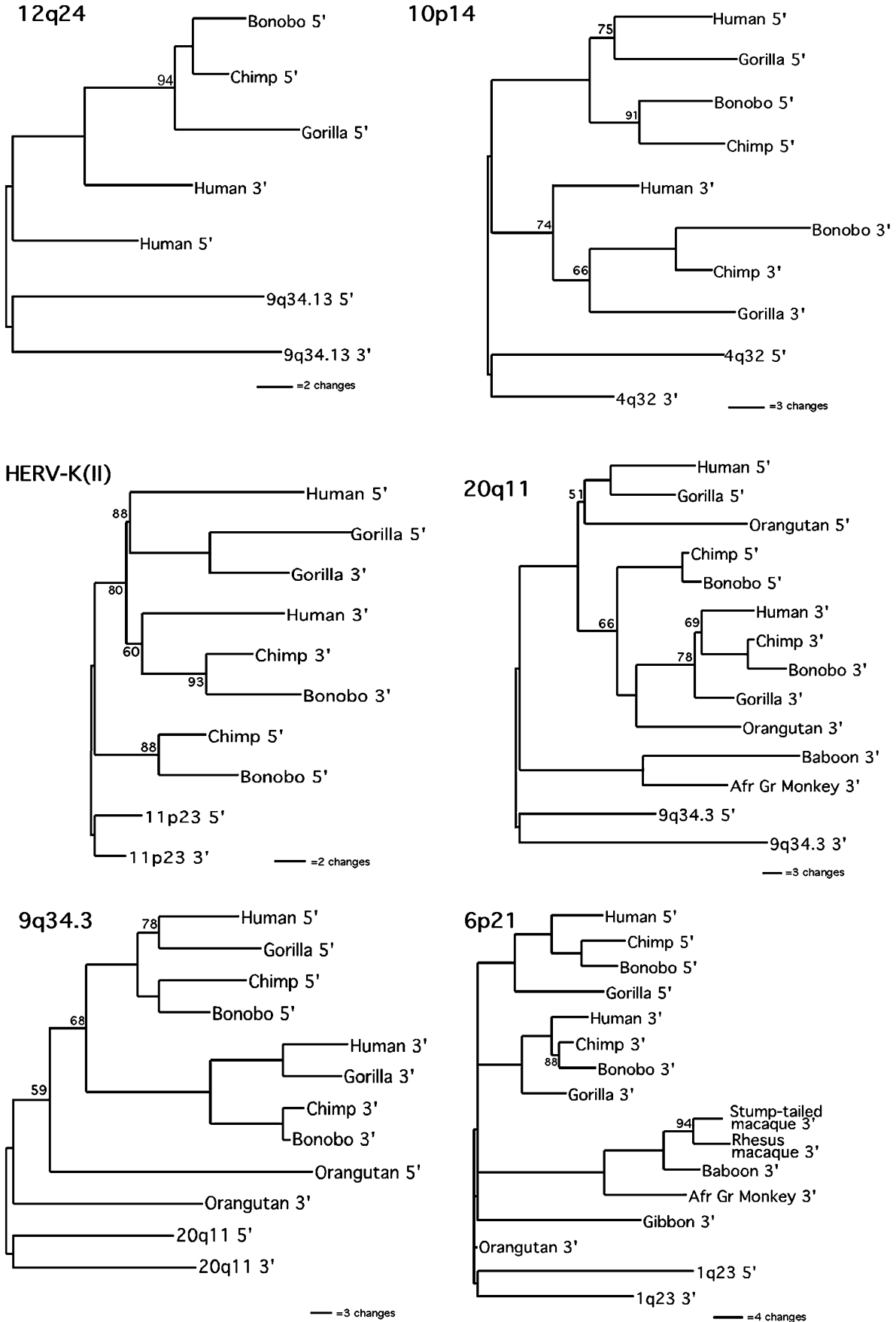


FIGURE 4.—Phylogenetic analysis of HERV-K elements that deviate from predicted topology. Analysis and labeling are the same as those in Figure 3.

HERV-K12q24: HERV-K12q24 shows evidence of a gene conversion event because the human 3' LTR sequence clusters with the 5' LTR sequence of the other species. However, because the 3' LTR sequence information is lacking in all of the species except humans, it cannot be determined if the reason for this unusual clustering is conversion of human 3' LTR sequence to 5' LTR sequence or merely the independent divergence of the human 5' LTR. A BLAST search of the chimpanzee genome sequence (<http://www.ncbi.nlm.nih.gov/genome/seq/PtrBlast.html>) did not identify the 3' LTR of this element or its flanking sequence.

HERV-K10p14: The tree for HERV-K10p14 deviates from the predicted topology because, although the sequences of the two LTRs form separate clusters, each cluster gives a different estimate of host phylogeny. Grouping the chimpanzee and bonobo sequences into one clade, the 5' LTR sequences give the branching pattern (H/G)C, while the 3' LTR sequences give the branching pattern (C/G)H. This pattern was also seen when a maximum-likelihood approach was used (data not shown). One explanation for this discrepancy might be the occurrence of a recombination event between two different alleles of the HERV element in the common ancestor of all three species and then the segregation of the two different alleles into the chimpanzee and human lineages and the recombinant allele in the gorilla lineage. However, examination of the substitutions along each of the lineages reveals that support for each of the branching patterns is weak, as is also indicated by the low bootstrap values. In the 5' LTR cluster, there are only two changes along the lineage leading to humans and gorillas. Both of these substitutions occur in CpG dinucleotides, which have been shown previously to be mutational hot spots in HERV elements (JOHNSON and COFFIN 1999), so they may reflect instances of homoplasy. In the 3' LTR cluster, three apparent substitutions are common to chimpanzees, bonobos, and gorillas, two of which are also in CpG dinucleotides.

HERV-K(II): The analysis of the HERV-K(II) sequences deviated quite strikingly from the predicted topology. The 5' and 3' LTR sequences did not cluster separately and there was no clear indication of species relatedness, indicating that a high level of concerted evolution between the LTRs has occurred at this locus and that these events probably occurred independently in the different species. The two LTRs in the gorilla cluster together, with the human 5' LTR forming a sister taxon. The 5' LTRs of the chimpanzee and bonobo form a distinct cluster, while the human, chimpanzee, and bonobo 3' LTRs cluster together. Examination of the substitution pattern along the different lineages reveals a rather complex evolutionary history. First, only 4 changes appear to have occurred in the LTRs after integration and prior to speciation events. This is in comparison to 10 such changes that occurred between

TABLE 2

Detection of gene conversion events between LTRs of HERV-K(II) in primate species

Species	No. of singles ^a	No. of doubles ^b	No. of co-doubles ^c	Pvalue ^d
Human	15	8	5	2.8×10^{-4}
Chimpanzee	12	6	4	3.1×10^{-4}
Bonobo	18	6	2	0.029
Gorilla	12	10	7	1.7×10^{-6}

^aThe number of sites that contain a mismatch in only the 5' or the 3' LTR in a total of 56 nonidentical nucleotide positions in the alignment.

^bSites where identical 5' and 3' substitutions occur.

^cDoubles at which the 5' and 3' substitutions occur in the same species.

^dPvalue of the permutation test.

the LTRs independently in the lineage leading to chimpanzees and bonobos prior to their separation. An estimated time frame for the accumulation of these mutations in the chimpanzee and bonobo common ancestor is 4 million years, perhaps indicating that this element integrated just prior to the radiation of all three lineages (gorillas, chimpanzees/bonobos, and humans).

The integration time estimate for this element indicates that the LTRs might have undergone some homogenization, reducing their degree of divergence. However, because the 5' and 3' LTRs do not have enough shared derived substitutions to distinguish the different lineages, clear cases of gene conversion could not always be identified. Another method for the detection of gene conversion is to look for the presence of "co-double" sites in the sequences under examination (BALDING *et al.* 1992). These sites reflect the independent acquisition of a mutation in one LTR in a species and the subsequent transfer of that mutation to the other LTR through gene conversion. The occurrence of homoplastic mutations in both LTRs of a provirus in any given species would be an exceedingly rare event, so the appearance of multiple co-doubles in a species would be more likely explained by single-mutation events in one LTR followed by homogenization between the LTRs. Table 2 shows the number of sites in the different species showing evidence of gene conversion because of the presence of co-double substitutions. The test statistic is the frequency of co-double sites, given a random distribution of mismatches among the sequences. Using this test, all of the species demonstrate statistically significant evidence for gene conversion.

HERV-K20q11: The evidence for gene conversion events in HERV-K20q11 is more striking because its LTRs had accumulated enough distinguishing mutations prior to subsequent speciation events that clear tracts of converted sequences could easily be identified. In Figure 5A, the positions that differentiate the 5' and

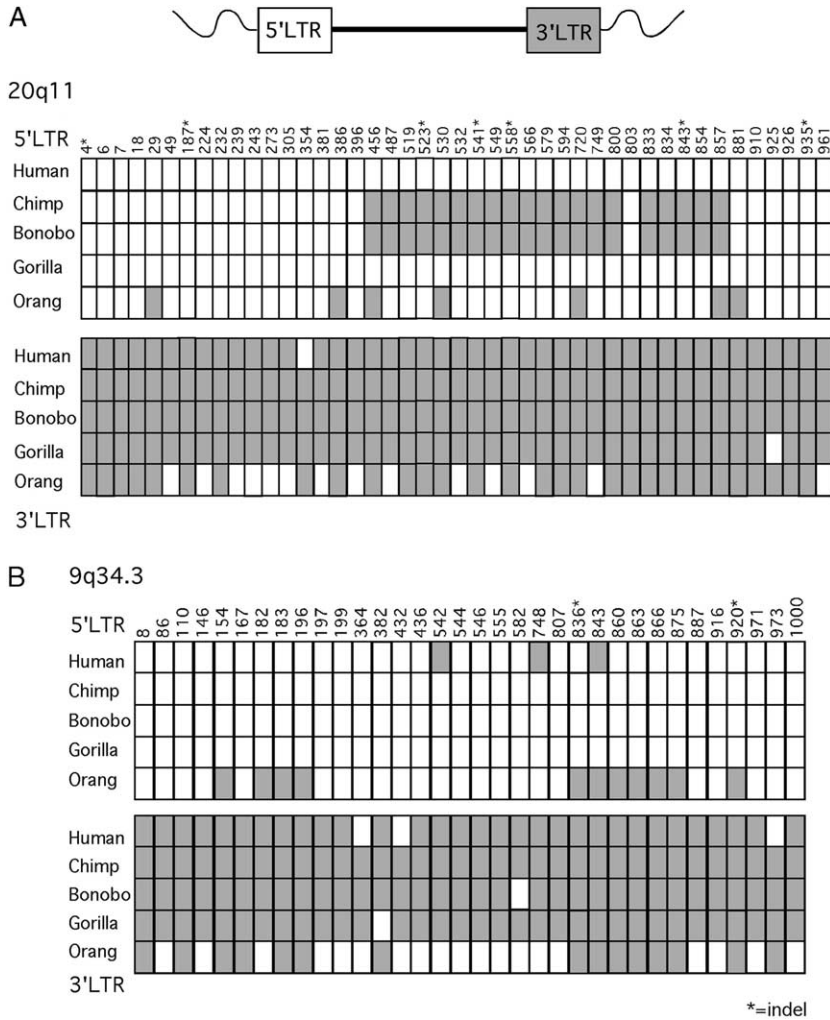


FIGURE 5.—LTR-LTR gene conversion and sequence divergence in two HERV-K proviruses. (A) HERV-K20q11 sequences and (B) HERV-K9q34.3 sequences for the human, chimpanzee, bonobo, gorilla, and orangutan are shown. Only positions that differ between the 5' and 3' LTRs in at least one species are represented and the consensus base pair identities of these positions are indicated. 5' LTR sequences are shown as open areas and 3' LTR sequences are shown as shaded areas. Gene conversion events are evident by the transfer of a 5' LTR sequence to the 3' LTR or vice versa.

3' LTRs are plotted for this element, and the identity of the base pair for each of these sites is shown.

A gene conversion event appears to have occurred in the common ancestor of the chimpanzee and bonobo, transferring a 400- to 500-bp patch of sequence from the 3' LTR to the 5' LTR starting between bases 396 and 456 up to a position between bases 857 and 881. This stretch of identity is apparently extensive enough to cause the disparate grouping of the 5' LTR sequences in these two species in the 3' LTR cluster in the phylogenetic tree. Removing the chimpanzee and bonobo sequences from the calculation changes the integration time estimate, increasing it to 16.5–20.9 million years, closer to the time of the last common ancestor of Old World monkeys (OWMs). In addition, the exclusion of the conversion tract from the phylogenetic analysis corrects the tree topology, clustering all of the 5' LTR sequences together (data not shown).

HERV-K9q34.3: Examination of the LTR sequences at the HERV-K9q34.3 locus, which also gave an underestimated integration time estimate indicating sequence homogenization, shows evidence of possible gene conversion events occurring in the orangutan (Figure 5B).

Two regions along the length of both orangutan LTRs are identical even though at these same positions the two LTRs have diverged in the human, chimpanzee, bonobo, and gorilla. This difference may indicate a sequence transfer in the orangutan subsequent to the mutation event differentiating the LTRs. Alternatively, if the time of integration was very close to the speciation event, these LTR mutations could have arisen after the orangutan lineage diverged from that of the African great apes. The relatively high genetic distance between the LTRs in the orangutan, 0.084 as compared to an average of 0.063 in the other four species, indicates that the LTR sequences have not undergone a significant amount of sequence homogenization, which would be indicative of gene conversion events. The position of the orangutan LTRs in the phylogenetic analysis (Figure 4) of this element is congruent with the interpretation that this element integrated just prior to the branching event that led to the orangutan lineage.

HERV-K6p21: The final element in this analysis, HERV-K6p21, was detected in all species examined except the New World squirrel monkey, suggesting an age of at least 25 million years. However, its relatively low level of LTR

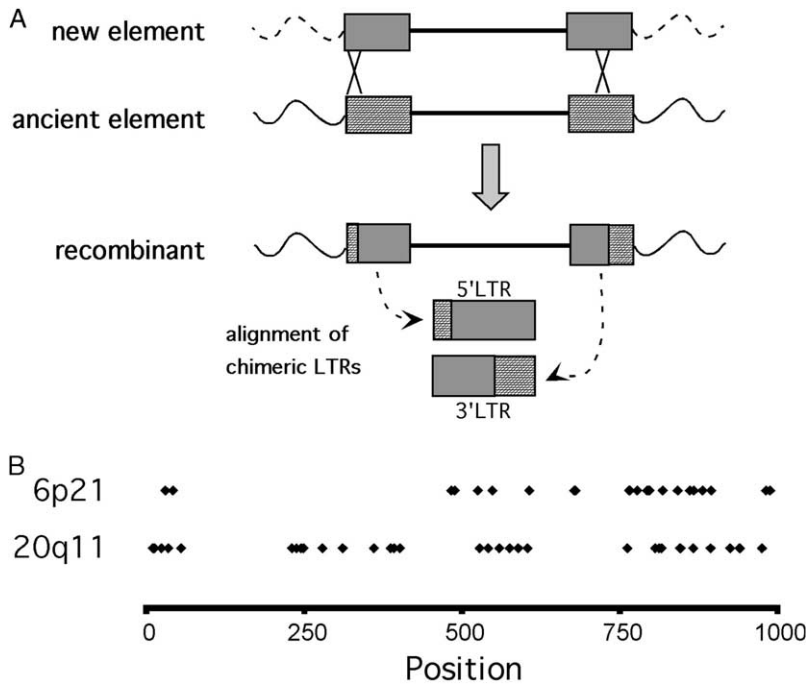


FIGURE 6.—Integration by homologous recombination generates chimeric LTRs in HERV-K6p21. (A) The LTRs of the newly integrating HERV-K element are depicted with shaded boxes and those of the ancient HERV-K element that it recombined into are depicted with cross-hatched boxes. The resultant recombinant element is shown. The chimeric LTRs generated by the recombination event are aligned for comparison. (B) The chimeric 5' and 3' LTR sequences in the human, chimpanzee, bonobo, and gorilla were aligned and the position of the base pair differences is plotted along the length of the LTR. The same analysis was done for HERV-K20q11 as a comparison and is shown below the 6p21 plot.

divergence suggested that it represents a much more recent integration event. The estimated age of ~ 10 million years would place integration in the African ape lineage. Examination of the sequence alignment showed no clear evidence for gene conversion. However, the phylogenetic analysis of the LTR sequences of this element revealed a possible explanation for this inconsistency. The 5' and 3' LTR sequences in the human, chimpanzee, bonobo, and gorilla conform to the predicted topology (Figure 4). However, the 3' LTR sequences from the remaining species form a separate cluster, suggesting that they may be from a different provirus, even though the elements are in identical genomic locations.

Given that the probability that a provirus would integrate more than one time into the same genomic location is extremely low, this locus could represent the replacement of an ancient provirus by a new element in the common ancestor of the great apes. Such an event could have occurred by homologous recombination between a newly integrating HERV-K element or another provirus located elsewhere in the genome (Figure 6). The recombination event would generate an element with chimeric LTRs consisting of sequence from both the old and the new provirus. A prediction of this scenario is that most of the divergence between the two LTRs of this element in the African great apes and humans should be located near the ends of the LTR sequence, because these regions would be remnants of the ancient provirus. Figure 6B shows the location of the base pair positions where the two LTRs differ in humans, chimpanzees, bonobos, and gorillas for HERV-K6p21 as well as HERV-K20q11 for comparison. Indeed, there is a distinct pattern in the location of the LTR differ-

ences for HERV-K6p21, with a clustering of mutations at both ends of the alignment and relatively few in the central portion, while the differences for HERV-K20q11 are more uniformly distributed along the LTR length.

Figure 7 shows further evidence for this model, namely that the accumulation of mutations over time in HERV-K6p21 does not follow a molecular clock model, as is the case for HERV-K20q11. When pairwise distances between homologous 3' LTR sequences between all of the species for which sequence information was available were plotted against the divergence time of those species (Figure 7A), a straight line was seen for HERV-K20q11 with a constant slope suggesting a relatively uniform rate of mutation accumulation over time. In the case of HERV-K6p21, however, there was a sharp increase in the slope of the line between the divergence times of 7 million and 14 million years, while the slope before and after these points was relatively constant. This result is suggestive of an event occurring between these two time points that drastically increased the apparent mutation rate at this locus, which could be explained by the replacement of the original HERV-K element with a new, divergent element. The dashed line in Figure 7A shows the plot that would have resulted if HERV-K6p21 had accumulated mutations over time at a constant rate.

Additional support for this model is provided by analysis of the internal sequences of the HERV-K6p21 provirus. Figure 7B shows dot matrix plots comparing the sequences of HERV-K20q11 and HERV-K6p21 to a human-specific provirus, HERV-K10. As expected for a more ancient element, HERV-K20q11 appears to be

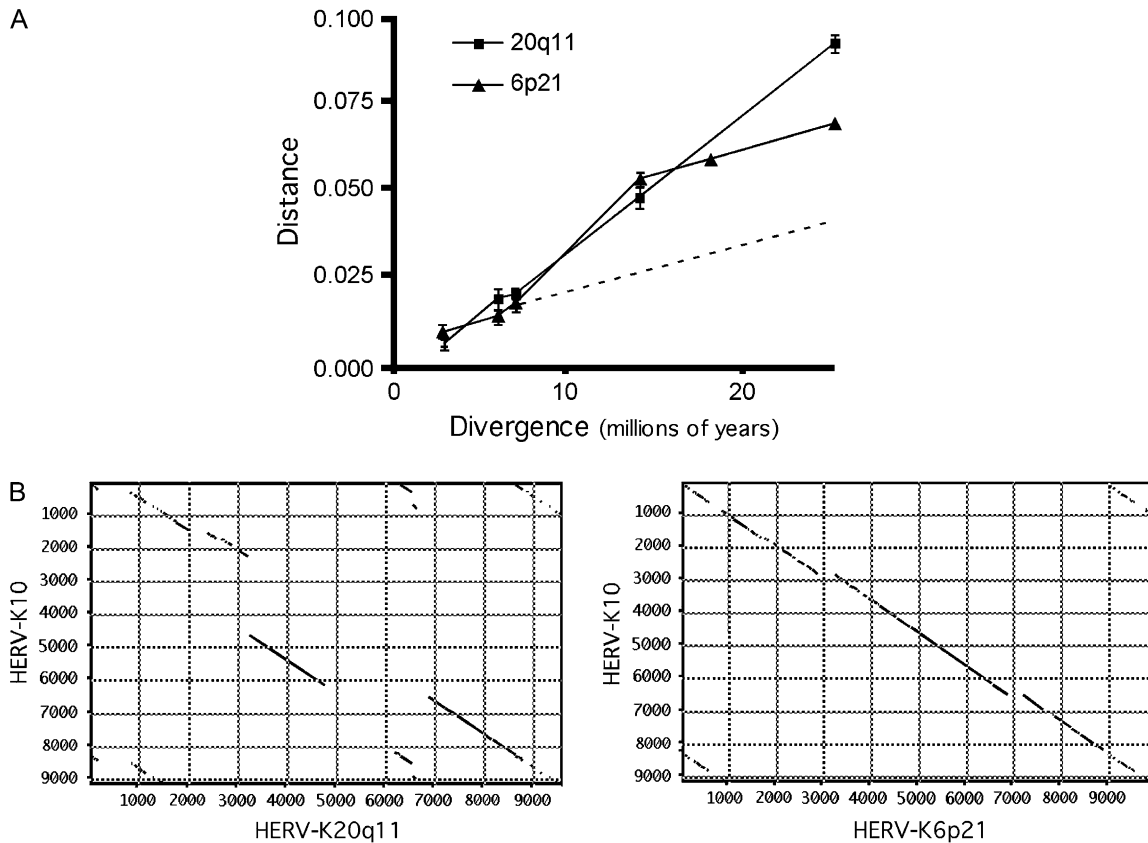


FIGURE 7.—6p21 replaced an ancient HERV-K provirus in the common ancestor of the hominids. (A) Kimura two-parameter corrected pairwise distances between species were plotted against divergence times for all available HERV-K20q11 and HERV-K6p21 sequences. The dotted line indicates the expected outcome for HERV-K6p21 if the sequences in the hominids and the rest of the species represented the same element, which had accumulated mutations at a constant rate over time. The discontinuity in the graph probably reflects the replacement of the original HERV-K element by a different, and somewhat divergent, HERV-K element. (B) The relative ages of HERV-K20q11 and HERV-K6p21 are demonstrated by comparing their sequences to a consensus HERV-K sequence. Dot matrix analysis is shown with the full-length (human-specific) HERV-K10 sequence along the y-axis and the sequence of the element under examination along the x-axis. The window size in the analysis was 30 and the minimum percentage score was 60.

riddled with large insertions and deletions, while the sequence of HERV-K6p21 appears to be relatively intact, consistent with a much shorter residence in the genome.

Phylogenetic analysis may not be a sensitive enough detector of gene conversion events, so the sequence alignments of all of the HERV-K elements were examined for evidence of LTR-LTR sequence transfers. Gene conversion must involve at least two apparent transfer events in close proximity to each other and the limit of statistically significant clustering (at the 95% confidence limit) is 24 bp for LTR sequences, which are approximately 968 bp in length (STEPHENS 1985). Only two elements examined exhibited evidence for gene conversion under these criteria. At the HERV-K1q23 locus, base pair positions 529 and 530 in the 3' LTR of the gorilla were apparently converted to the 5' LTR. A sequence transfer also occurred in the chimpanzee and bonobo lineage at the HERV-K6p22 locus where the 3' sequence at positions 421 and 428 was converted to the 5' sequence. Neither of these events led to significant

deviation from the "normal" phylogenetic structure (Figure 3).

DISCUSSION

Phylogenetic analysis of human endogenous retroviral sequences provides important information regarding the evolutionary history of this family of repetitive elements and their various effects on the primate genome. Moreover, the utility of these studies extends to the examination of phylogenetic relationships among the host species that share these elements, as well as to mechanisms of genetic variation during their evolution. Since the LTRs of a given provirus are identical at the time of integration and are expected to accumulate random mutations at the same rate as that of host DNA thereafter, the divergence between LTRs has been used to estimate the time of integration (DANGEL *et al.* 1995; JOHNSON and COFFIN 1999). However, there is evidence for gene conversion events in endogenous proviral LTRs

(JOHNSON and COFFIN 1999) as well as in other types of repetitive elements (KASS *et al.* 1995). The present study was undertaken to better understand this process and how well LTR divergence predicted proviral age.

Of the 15 HERV-K elements studied, 9 had LTRs whose divergence was consistent with the minimum age estimated from the species distribution, implying that the inferred distribution is likely correct, despite our failure to amplify unoccupied integration sites from more distant species. This failure may be a consequence of the fact that the majority of HERV-K integrations are into repetitive DNA, making them difficult to amplify cleanly. In this set, two proviruses had LTRs that were significantly more divergent than predicted from their species distribution. In such cases, the element may be present in more distantly related species but not detected because of sequence divergence in the primer sequence, deletion of both LTRs, or solo LTR formation in more distant lineages. A BLAST search of the rhesus trace sequence archives (<http://www.ncbi.nlm.nih.gov/Traces>) was performed to search for the presence of these two elements but was inconclusive because the flanking sequences were not found. Alternatively, discrepancies in the estimated ages of the provirus and their species distributions might reflect a period of rapid evolution at that locus after integration but prior to speciation, increasing the evolutionary distance between the HERV LTRs as compared to the interspecies divergences. The idea of an episodic rather than static molecular clock was proposed some time ago (GILLESPIE 1984), and there are several well-documented examples of loci in the primate genome undergoing this kind of evolution (MESSIER and STEWART 1997; LIU *et al.* 2001; SCHANER *et al.* 2001). The data from this analysis do not allow distinction between the two possibilities.

Overall, we found that HERV-K sequences are subject to ectopic recombination events that occur at a relatively high frequency. Of the 15 HERV-K proviruses examined, 5 [HERV-K(II), 20q11, 6p21, 1q23, and 6p22] showed convincing evidence of involvement in gene conversion or recombination events in one or more species. Two additional elements, HERV-K12q24 and 10p14, have aberrant phylogenies but definitive explanatory data are lacking. Taken as a whole, all but three (20q11, 1q23, and 6p22) of the ectopic recombination events detected in this study took place in human ancestors. It is estimated that there are ~3900 full-length HERVs in the human genome with intact 5' and 3' LTRs (BELSHAW *et al.* 2005). If the same fraction (4/15) of the entire complement of HERV elements has been involved in ectopic recombination events as was found in this study, then ~1050 elements, composing nearly 10 Mb of sequence in the human genome, have mediated such events in the human lineage during the course of evolution.

Gene conversion may also account for the removal of HERV-K20q11 from the genome of several species. This

element is present in the great apes, orangutan, and some Old World monkeys, indicating that its integration took place prior to the divergence of both lineages, which occurred ~25 million years ago. However, its conspicuous absence in the gibbon as well as in two macaque species (BLAST search of the trace archive was also negative) is inconsistent with this interpretation, unless the element was lost in these species. The likelihood of the parallel loss of this provirus is increased by the fact that it is integrated into a member of the LINE family of repetitive elements, which are very abundant in the primate genome (SHEEN *et al.* 2000). The LINE sequences flanking the HERV-K element could have mediated a gene conversion event with another LINE element, lacking the insertional mutation, thereby removing the proviral sequence. Otherwise, identical and precise deletions of this sequence would have had to occur twice during primate evolution, which seems unlikely.

Repetitive sequence elements are the most dominant form of noncoding DNA in the human genome, but, because of their tendency to mediate ectopic recombination events and undergo concerted evolution, care must be taken when they are utilized in phylogenetic studies. HERV sequences provide a means to detect the occurrence of such events, however. When using LTR sequences as markers, homogenization by ectopic recombination events is readily apparent as aberrant clustering of the 5' and 3' sequences. In addition to LTR-LTR sequence transfers, confounding results can also arise if the integration of the HERV element occurred near the time of the divergence of the species under study. If the two LTRs of the integrated provirus have not been resident in the genome for sufficient time to acquire a number of mutations prior to the speciation event, the 5' and 3' LTRs will not have sufficient evolutionary history in the ancestral species to distinguish them from one another. Subsequent to speciation, they follow independent paths, so the resultant phylogenetic analysis would place the 5' and 3' LTR sequences from all of the species randomly within the tree. In the presence of concerted evolutionary forces such as gene conversion, however, the 5' and 3' LTR sequences within species would tend to cluster together.

This analysis has demonstrated the sensitivity of HERV loci as indicators of ectopic recombination events and has revealed multiple occurrences of these events during the course of primate evolution. There is no reason to presume that this family of repetitive elements is more prone to mediating such events, so it appears that as a whole these sequences, once regarded as "junk DNA," have had a profound effect in shaping our genome. The question arises as to whether HERV elements can continue to change our genomic landscape through active retrotransposition or recombination events. While no direct evidence indicates that such events are ongoing in the human genome, members of

the HERV-K family appear to be the most likely candidates for playing such a role.

We thank Stephen J. O'Brien and Ralph R. Isberg for generously providing primate cell lines. We also thank the Center for Gastroenterology Research on Absorptive and Secretory Processes (GRASP) of the Tufts-New England Medical Center for valuable technical assistance. This work was supported by research grant R01CA89441 from the National Cancer Institute. J.M.C. was a research professor of the American Cancer Society with support from the F. M. Kirby Foundation and J.F.H. was supported in part by training grant CA5441 from the National Cancer Institute.

LITERATURE CITED

- BALDING, D. J., R. A. NICHOLS and D. M. HUNT, 1992 Detecting gene conversion: primate visual pigment genes. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* **249**: 275–280.
- BANNERT, N., and R. KURTH, 2004 Retroelements and the human genome: new perspectives on an old relation. *Proc. Natl. Acad. Sci. USA* **101** (Suppl. 2): 14572–14579.
- BARBULESCU, M., G. TURNER, M. I. SEAMAN, A. S. DEINARD, K. K. KIDD *et al.*, 1999 Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. *Curr. Biol.* **9**: 861–868.
- BELSHAW, R., V. PEREIRA, A. KATZOURAKIS, G. TALBOT, J. PACES *et al.*, 2004 Long-term reinfection of the human genome by endogenous retroviruses. *Proc. Natl. Acad. Sci. USA* **101**: 4894–4899.
- BELSHAW, R., A. KATZOURAKIS, J. PACES, A. BURT and M. TRISTEM, 2005 High copy number in human endogenous retrovirus (HERV) families is associated with copying mechanisms in addition to re-infection. *Mol. Biol. Evol.* **22**: 814–817.
- BOEKE, J. D., and J. P. STOYE, 1997 Retrotransposons, endogenous retroviruses, and the evolution of retroelements, pp. 343–436 in *Retroviruses*, edited by J. M. COFFIN, S. H. HUGHES and H. VARMUS. Cold Spring Harbor Laboratory Press, Plainview, NY.
- COSTAS, J., and H. NAVEIRA, 2000 Evolutionary history of the human endogenous retrovirus family ERV9. *Mol. Biol. Evol.* **17**: 320–330.
- DANGEL, A. W., B. J. BAKER, A. R. MENDOZA and C. Y. YU, 1995 Complement component C4 gene intron 9 as a phylogenetic marker for primates: long terminal repeats of the endogenous retrovirus ERV-K(C4) are a molecular clock of evolution. *Immunogenetics* **42**: 41–52.
- GILLESPIE, J. H., 1984 The molecular clock may be an episodic clock. *Proc. Natl. Acad. Sci. USA* **81**: 8009–8013.
- GOODMAN, M., C. A. PORTER, J. CZELUSNIAK, S. L. PAGE, H. SCHNEIDER *et al.*, 1998 Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. *Mol. Phylogenet. Evol.* **9**: 585–598.
- HUGHES, J. F., and J. M. COFFIN, 2001 Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. *Nat. Genet.* **29**: 487–489.
- HUGHES, J. F., and J. M. COFFIN, 2004 Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. *Proc. Natl. Acad. Sci. USA* **101**: 1668–1672.
- JOHNSON, W. E., and J. M. COFFIN, 1999 Constructing primate phylogenies from ancient retrovirus sequences. *Proc. Natl. Acad. Sci. USA* **96**: 10254–10260.
- 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.
- LANDER, E. S., L. M. LINTON, B. BIRREN, C. NUSBAUM, M. C. ZODY *et al.*, 2001 Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.
- LAVIE, L., P. MEDSTRAND, W. SCHEMPP, E. MEESE and J. MAYER, 2004 Human endogenous retrovirus family HERV-K(HML-5): status, evolution, and reconstruction of an ancient betaretrovirus in the human genome. *J. Virol.* **78**: 8788–8798.
- LEIB-MOSCH, C., M. HALTMEIER, T. WERNER, E. M. GEIGL, R. BRACK-WERNER *et al.*, 1993 Genomic distribution and transcription of solitary HERV-K LTRs. *Genomics* **18**: 261–269.
- LIU, J. C., K. D. MAKOVA, R. M. ADKINS, S. GIBSON and W. H. LI, 2001 Episodic evolution of growth hormone in primates and emergence of the species specificity of human growth hormone receptor. *Mol. Biol. Evol.* **18**: 945–953.
- LOWER, R., J. LOWER 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.
- MEDSTRAND, P., and D. L. MAGER, 1998 Human-specific integrations of the HERV-K endogenous retrovirus family. *J. Virol.* **72**: 9782–9787.
- MESSIER, W., and C. B. STEWART, 1997 Episodic adaptive evolution of primate lysozymes. *Nature* **385**: 151–154.
- NACHMAN, M. W., and S. L. CROWELL, 2000 Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**: 297–304.
- PACES, J., A. PAVLICEK and V. PACES, 2002 HERVd: database of human endogenous retroviruses. *Nucleic Acids Res.* **30**: 205–206.
- ROY, A. M., M. L. CARROLL, S. V. NGUYEN, A. H. SALEM, M. OLDRIDGE *et al.*, 2000 Potential gene conversion and source genes for recently integrated Alu elements. *Genome Res.* **10**: 1485–1495.
- ROY-ENGEL, A. M., M. L. CARROLL, M. EL-SAWY, A. H. SALEM, R. K. GARBER *et al.*, 2002 Non-traditional Alu evolution and primate genomic diversity. *J. Mol. Biol.* **316**: 1033–1040.
- SCHANER, P., N. RICHARDS, A. WADHWA, I. AKSENTIJEVICH, D. KASTNER *et al.*, 2001 Episodic evolution of pyrin in primates: human mutations recapitulate ancestral amino acid states. *Nat. Genet.* **27**: 318–321.
- SHEEN, F. M., S. T. SHERRY, G. M. RISCH, M. ROBICHAUX, I. NASIDZE *et al.*, 2000 Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. *Genome Res.* **10**: 1496–1508.
- STEPHENS, J. C., 1985 Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. *Mol. Biol. Evol.* **2**: 539–556.
- SUGIMOTO, J., N. MATSUURA, Y. KINJO, N. TAKASU, T. ODA *et al.*, 2001 Transcriptionally active HERV-K genes: identification, isolation, and chromosomal mapping. *Genomics* **72**: 137–144.
- SVERDLOV, E. D., 2000 Retroviruses and primate evolution. *BioEssays* **22**: 161–171.

Communicating editor: S. SANDMEYER