JOURNAL OF VIROLOGY, Aug. 2004, p. 8788–8798 Vol. 78, No. 16
0022-538X/04/$08.00+0 DOI: 10.1128/JVI.78.16.8788–8798.2004

Human Endogenous Retrovirus Family HERV-K(HML-5): Status, Evolution, and Reconstruction of an Ancient Betaretrovirus in the Human Genome†

Laurence Lavie,[1] Patrik Medstrand,[2] Werner Schempp,[3] Eckart Meese,[1] and Jens Mayer[1]*

Department of Human Genetics, University of Saarland, 66421 Homburg,[1] and Institute of Human Genetics and Anthropology, University of Freiburg, 79106 Freiburg,[3] Germany, and Department of Cell and Molecular Biology, Lund University, 22184 Lund, Sweden[2]

Received 27 December 2003/Accepted 12 April 2004

The human genome harbors numerous distinct families of so-called human endogenous retroviruses (HERV) which are remnants of exogenous retroviruses that entered the germ line millions of years ago. We describe here the hitherto little-characterized betaretrovirus HERV-K(HML-5) family (named HERVK22 in Repbase) in greater detail. Out of 139 proviruses, only a few loci represent full-length proviruses, and many lack *gag* protease and/or *env* gene regions. We generated a consensus sequence from multiple alignment of 62 HML-5 loci that displays open reading frames for the four major retroviral proteins. Four HML-5 long terminal repeat (LTR) subfamilies were identified that are associated with monophyletic proviral bodies, implying different evolution of HML-5 LTRs and genes. Sequence analysis indicated that the proviruses formed approximately 55 million years ago. Accordingly, HML-5 proviral sequences were detected in Old World and New World primates but not in prosimians. No recent activity is associated with this HERV family. We also conclude that the HML-5 consensus sequence primer binding site is identical to methionine tRNA. Therefore, the family should be designated HERV-M. Our study provides important insights into the structure and evolution of the oldest betaretrovirus in the primate genome known to date.

Approximately 8% of the human genome is derived from retrovirus-like elements termed endogenous retroviruses (ERV) (14). Most of them are likely remnants of exogenous retrovirus infection of the germ line which became fixed in the population millions of years ago. Intracellular retrotransposition events increased proviral copy number during evolution, resulting in the presence of a few to several thousand proviruses belonging to various ERV families. A variety of distinct human ERV (HERV) families have been identified, suggesting that the germ line was invaded by various exogenous retroviruses (23, 25, 41). The human genome was recently estimated to contain 30 to 50 distinct HERV families (11). About 100 different HERV sequences are defined in Repbase, a widely employed reference sequence database for repetitive elements (15). Unfortunately, there is no established nomenclature for HERV. We follow here a nomenclature that uses the single-letter amino acid code of the tRNA complementary to the primer binding site formerly used to prime reverse transcription. Accordingly, HERV-K denotes a number of HERV families having a lysine tRNA-like primer binding site. Phylogenetic analysis of HERV reverse transcriptase sequences have identified 10 HERV-K families in the human genome which were termed human MMTV-like (HML-1 to HML-10) because of homologies to the betaretrovirus mouse mammary tumor virus (MMTV) (1, 32). Repbase Update also lists 10 HERV-K families.

Structurally intact ERV contain *gag*, protease (*prt*), polymerase (*pol*), and envelope (*env*) genes and are flanked by long terminal repeats (LTRs). Two HERV families, HERV-K(HML-2) and HERV-W, contain intact open reading frames (ORFs) and encode functional proteins (2, 3, 18, 27, 29, 31, 36). Further proviral sequences with Env coding capacity were identified recently (9). Most HERV sequences are coding deficient because they accumulated a variety of mutations and deletions since provirus formation. The vast majority of HERV are present in the genome only as solitary LTRs due to homologous recombination between 5′ and 3′ LTRs.

Many HERV families were fixed in Old World primates after their evolutionary split from New World primates about 35 million years ago (41). For example, the HERV-K families HML-2, HML-3, and HML-6 were all found in Old World monkeys but missing in New World primates. The HERV-K(HML-2) family seems to be relatively long-lived in the genome, displaying transpositional activity since its appearance in the primate lineage up till after the human-chimpanzee split (7, 28, 33, 38). In contrast, other ERV families are short-lived. In the case of the HERV-K(HML-3) family, no activity in terms of provirus formation was indicated in the human evolutionary lineage over the last 30 million years (26). To date, only a few HERV families have been characterized in greater detail. Here we characterize the HERV-K(HML-5) family by analyzing proviruses in the human genome. We find that the HML-5 proviruses have an older origin than other betaretroviruses in the human genome, and we reconstruct a putative full-length coding-competent ancient betaretrovirus which targeted the primate germ line more than 50 million years ago.

* Corresponding author. Mailing address: Department of Human Genetics, Building 60, University of Saarland, Medical Faculty, 66421 Homburg, Germany. Phone: 49 6841 1626627. Fax: 49 6841 1626186. E-mail: jens.mayer@uniklinik-saarland.de.

† Supplemental material for this article may be found at http://jvi.asm.org/.

MATERIALS AND METHODS

HERV-K(HML-5) sequence retrieval. HERV-K(HML-5) proviruses (LTR22A/HERVK22/LTR22A; Repbase version 8.2.0 consensus sequences) were identified by downloading the human sequence specified as HERVK22 in the RepeatMasker annotation of the June 2002 (hg12) UCSC genome browser (http://genome.ucsc.edu/). We performed dot matrix comparisons between each proviral sequence using MacVector (Accelrys Inc.) with default settings. We identified boundaries between HERV-K(HML-5) proviral loci and flanking cellular sequences and the location of nonretroviral repetitive elements within the proviral portions by using RepeatMasker (provided by A. F. A. Smit and P. Green; RepeatMasker at http://www.repeatmasker.org). We subsequently removed LTRs and apparent non-HERV-K(HML-5) sequences.

Multiple alignment of HERV-K(HML-5) sequences and consensus sequence generation. We produced multiple alignments of HERV-K(HML-5) proviral sequences displaying a minimal size of 3 kb (after deletion of other repetitive elements). Multiple alignments were generated employing DIALIGN2 (35) provided by the Institut Pasteur web server (http://bioweb.pasteur.fr) or employing ClustalW (42) using default settings. We further optimized alignments by hand-operating the Se-Al program (provided by Andrew Rambaut; http://evolve.zoo.ox.ac.uk/). Consensus sequences, generated by the Boxshade program at the Institut Pasteur, were further evaluated and corrected manually. The resulting proviral sequence was analyzed for encoded retroviral proteins and conserved domains employing BlastP and CD-search at the National Center for Biotechnology Information.

Phylogenetic analysis. We employed PAUP* (Sinauer Associates) and the PHYLIP package (10), the latter provided by the Institut Pasteur web server, for phylogenetic analysis. The above programs generated multiple alignment of LTR sequences, and several regions from the multiple alignment of proviral body regions were subjected to neighbor-joining analysis. Nucleotide distances were corrected according to the Kimura-2-parameter model (17). Gaps and ambiguous positions were excluded from analysis. Sequences obtained from various primate species were included in the respective multiple alignment portions of human sequences and were analyzed similarly.

Evolutionary age of proviruses. Approximate integration times of HERV-K(HML-5) proviruses displaying almost full-length LTRs on both sides were estimated by determination of Kimura-2-parameter corrected nucleotide distances between 5′ and 3′ LTRs of particular proviruses by employing PAUP*. Proviral age (T) was estimated according to the following formula: $T = D/(2 \cdot 0.13)$, where D is the nucleotide divergence between the 2 LTR sequences and 0.13 is the estimated average substitution rate per nucleotide and million years (8, 19). The factor 2 considers that both LTRs diverged independently from each other.

Test for HERV-K(HML-5) homologous sequences in primate species. PCR primers corresponding to the 5′ portion of the *gag* gene (P1, 5′-AACAGTATATAAAAGTATTGAAACA-3′; and P2, 5′-TGCTTTTTCTTATCTCTTTATAAG-3′) and the *env* gene (P3, 5′-CCATTCACTCCAGATAATTTGT-3′; and P4, 5′-TTCTTTTCCAATCAAGGAGTGA-3′) within the HERV-K(HML-5) consensus sequence were used to amplify HERV-K(HML-5) proviruses from human and various other primate species genomic DNAs, together comprising four primate clades: *Pan troglodytes*, *Pongo pygmaeus*, *Hylobates lar*, *Colobus guereza*, *Mandrillus sphinx*, *Macaca mulatta*, *Saimiri sciureus*, *Callithrix jacchus*, *Alouatta seniculus*, and *Nycticebus coucang*. Genomic DNA (100 ng each) was subjected to PCR with 1 μM primers and 2.5 U of *Taq* polymerase (Invitrogen Inc.) in a final volume of 25 μl. PCR cycling conditions for *gag* and *env* amplification were as follows: 5 min at 94°C; 35 cycles of 45 s at 94°C, 45 s at 54°C, and 2 min at 72°C; 5 min at 72°C. PCR products for the *gag* region obtained from *P. troglodytes*, *H. lar*, *M. mulatta*, and *A. seniculus* were cloned into the pGEM-T vector (Promega). Sequences for both DNA strands were obtained using a SequiTherm Excel II DNA sequencing kit-LC (Biozym) and an automated DNA sequencer (Licor 4000-L; MWG). Raw sequence data were analyzed using the Sequencher software (Gene Codes Corp.).

Nucleotide sequence accession numbers. We deposited the updated HERV-K(HML-5)/HERVK22 and LTR22 consensus sequences in Repbase.

RESULTS

Identification and structure of HERV-K(HML-5) proviruses. By using the annotation within the UCSC genome browser (http://genome.ucsc.edu/), we identified 139 HML-5 proviruses in the human genome. Twenty-three out of 139 (16.5%) loci were located on the Y chromosome, whereas other loci seemed randomly distributed along autosomes. We next performed dot matrix comparisons between proviral loci and the Repbase consensus sequences (LTR22A/HERVK22/LTR22A) to examine proviral structures in more detail. We generated schematic structures of 100 out of 139 HML-5 sequences (Fig. 1). We did not include 39 of the proviruses because those proviral sequences were heavily mutated due to integration of other repetitive elements, deletions, duplications, and/or inversions. The majority (123 proviruses) displayed larger deletions of retroviral genes: only 11 proviruses displayed a complete *gag* gene, 12 a complete *prt* gene, 45 a complete *pol* gene, and 15 a complete *env* gene. Among those, multiple sequences commonly displayed large deletions of about 1,410 and 1,320 bp within either the *gag-prt* junction or the *env* region, respectively, or in both. The first deletion removed a large portion of the 5′ part of the *gag* gene and the complete *prt* gene, and the second deletion removed a large central part of the *env* gene (Fig. 1). Sixteen proviruses were intact regarding the presence of retroviral genes, and nine of these resembled full-length proviral structures having LTRs on both sides flanking the putative coding regions.

Reconstruction of a coding-competent HML-5 provirus. Before multiple sequence alignment of HML-5 sequences, we employed RepeatMasker to identify and to subsequently delete non-HML-5 sequences from proviral sequence entries. A greater number of proviruses contained one or several other repetitive elements which likely integrated into the proviruses at different stages during primate evolution. L1 elements from the M1, P, and PA2-6 subfamilies and *Alu* elements from the Y, Yc, Ya5, Yb9, Sp, Sc, and Sg subfamilies were recognized (4, 16, 40). Furthermore, numerous LTR elements annotated in Repbase as LTR5B, -6A, -7, -7B, -10F, -12, -12B, -12C, -15, -17, -19, -30, and MER11C (15) were found (Table 1). We chose 62 proviral sequence entries with the least deletions for multiple alignment. By using Boxshade and subsequent visual correction, we generated a HML-5 consensus sequence from the alignment. The sequence displayed a total length of 6,824 bp and was 4% divergent in sequence compared to the HERVK22 consensus sequence present in Repbase. Several nucleotide differences adjusted reading frames with regard to the Repbase sequence. Two 1-bp and an 11-bp insertion were introduced into the 5′ intergenic region. A 3-bp sequence was introduced, and 2 and 1 bp were removed in the *gag*, *prt*, and *pol* gene regions, respectively. In addition, a stop codon within the *prt* gene and three stop codons within *pol* could be corrected.

Translation of our consensus sequence gave rise to ORFs for four major retroviral proteins that were not found as such in the Repbase sequence. ORFs for putative retroviral *gag*, *prt*, *pol*, and *env* genes could be defined (Fig. 1; see also Fig. 1A in the supplemental material, which is also available from us upon request). The putative 517-, 265-, 911-, and 694-amino-acid Gag, Prt, Pol, and Env proteins, respectively, displayed extended similarities to the corresponding retroviral proteins, as revealed by CD-search and BlastP analysis. As described recently, HML-5 proviruses once harbored a potentially active dUTPase enzyme within the *prt* ORF and also a domain encoding a retroviral aspartyl protease (30). As expected, the *pol* reading frame encoded retroviral reverse transcriptase, RNase H, and integrase domains. CD-search furthermore identified a

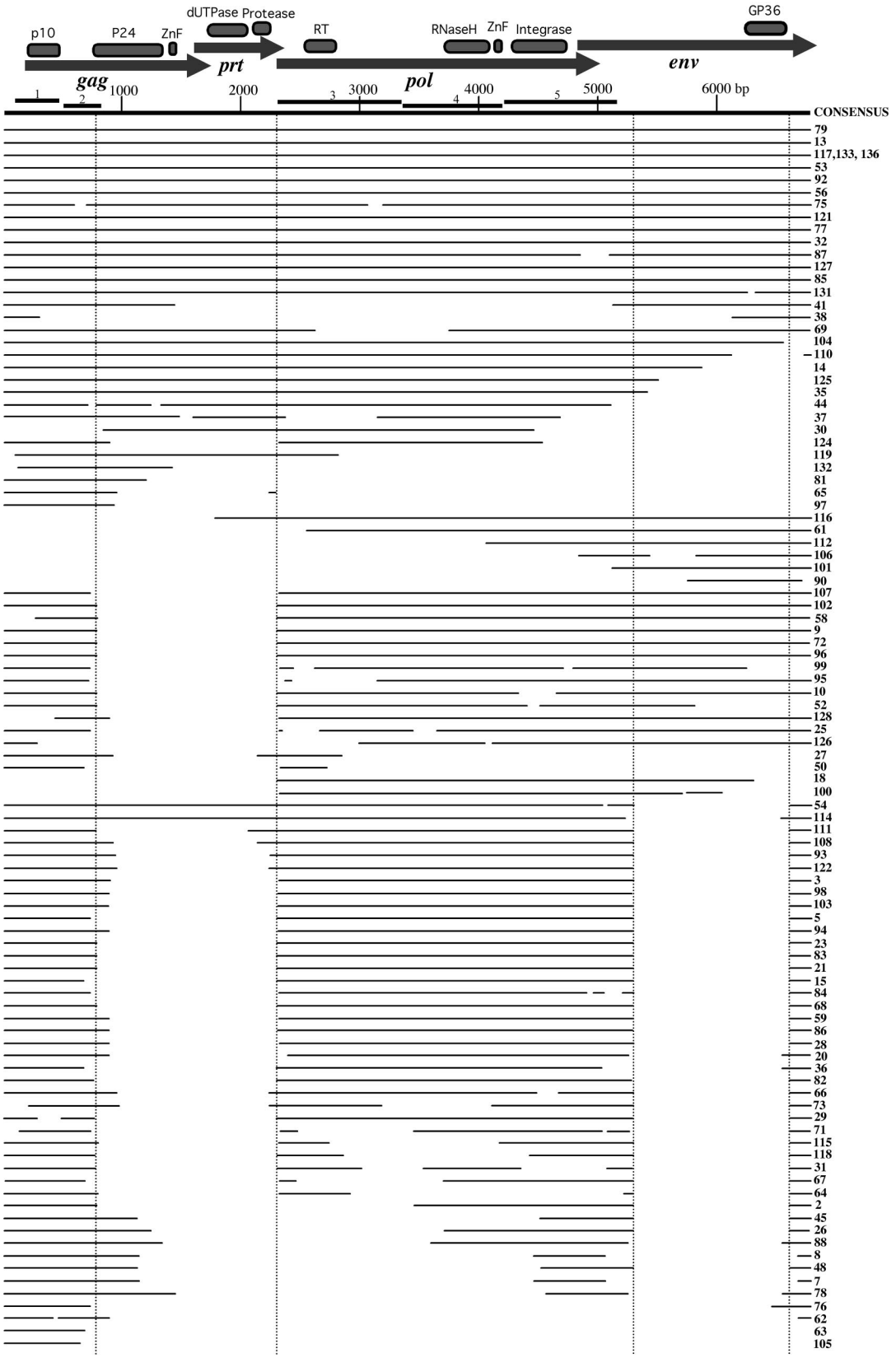

FIG. 1. Structures of 100 HERV-K(HML-5) provirus sequences. An ORF map of the HERV-K(HML-5) consensus sequence generated in this study, including retroviral *gag*, *prt*, *pol*, and *env* reading frames and protein domains, is shown at the top. Structures of proviral sequences with regard to the consensus sequence are depicted below, with horizontal lines indicating the presence of respective proviral portions. Numbers to the right refer to proviral locus numbers given in Table 1. Proviruses 117, 133, and 136 probably stem from genomic duplication events and are

HERV-K/MMTV-type gp36 domain within the Env protein C-terminal portion (Fig. 1). BlastP analysis of encoded retroviral proteins against the GenBank protein division revealed high similarities to retroviral proteins from the HERV-K(HML-2) family. Moreover, significant similarities to the betaretroviruses ovine pulmonary adenocarcinoma virus, MMTV, Mason-Pfizer monkey virus, and simian retroviruses 1 and 2 were detected, with identities to respective retroviral proteins ranging from 26 to 49% and similarities ranging from 41 to 64% (as determined by BLAST). Thus, our method was able to reconstruct all putative coding sequences of an ancient betaretrovirus.

The HML-5 sequence was initially assigned to the HERV-K superfamily due to the similarity in *pol* with other HERV-K families. Based on the analysis of an earlier draft version of the human genome sequence, Tristem suggested that the HERV-K(HML-5) primer binding site region is more similar to isoleucine than to lysine (43). We analyzed the HERV-K(HML-5) consensus primer binding site region in regard to the similarity to reported human tRNA sequences (22) (http://rna.wustl.edu/GtRDB/) and found that the HML-5 primer binding site is identical in sequence to the methionine tRNA 3′ end along at least 18 nucleotides (nt). Isoleucine tRNA displayed differences in three nucleotides, and lysine tRNAs displayed at least six different nucleotides (Fig. 2). Therefore, following the current HERV nomenclature, the HERV-K(HML-5) family should be designated HERV-M.

Phylogeny of the HERV-K(HML-5) family. Dot matrix comparisons between the different proviruses and the HERVK22 sequence flanked by LTR22A, as included in Repbase, showed that only a few entries presented more extended similarities with the LTR22A sequence. This finding suggested the existence of one or more LTR variants associated with HML-5 sequences and corroborated the definition of different LTR22 subfamilies in Repbase (LTR22, LTR22A, and LTR22B). For each proviral entry, we utilized RepeatMasker annotations to determine the presence of complete or deleted 5′ and 3′ LTRs (Table 2). We found that 70 proviral sequences were associated with full-length LTRs on both sides, 59 entries harbored at least one complete or deleted LTR on one side, and 10 entries lacked both LTRs. We generated a ClustalW multiple alignment with default alignment parameters for a total of 134 fairly complete LTR sequences.

Neighbor-joining analysis of the sequences subject to multiple alignment suggested the existence of several LTR subfamilies (Fig. 3). A major branch, displaying 100% bootstrap support (1,000 replicates), consisted of sequences similar to the LTR22A subfamily. Among those sequences, two subgroups displaying 63 and 91% bootstrap support could be distinguished. We named these two subgroups LTR22A and LTR22A2. Another group (99% support) consisted of sequences with similarity to the LTR22 subfamily. The majority of the remaining sequences were more similar to the LTR22B subfamily. Therefore, analysis of the entire HML-5 data set revealed three major LTR subfamilies (LTR22, LTR22A, and LTR22B) associated with the proviral sequences and is in accord with previous findings. In addition, LTR22A comprises a clearly distinguishable subgroup, named LTR22A2.

Based on the larger data set available in our study (40), we generated three new consensus sequences for the LTR22 families listed in Repbase and for LTR22A. The classification of LTR sequences in Table 2 is based on their phylogenetic relationship to those new LTR22 consensus sequences. The new consensus sequences for LTR22B and LTR22A displayed lengths of 497 and 457 bp, respectively, and were modestly different from the Repbase sequences. The 471-nt-long LTR22A2 consensus sequence presented 74% identity and 4% gaps compared to LTR22A. As revealed in this study, the previously established Repbase LTR22 consensus sequence was probably derived from three loci located on the Y chromosome (proviruses 117, 121, 131; Fig. 3) that are less representative for LTR22. The LTR22 consensus sequence generated in this study is 492 bp in length compared to the 580-nt Repbase sequence that furthermore included 47 ambiguous positions (International Union of Pure and Applied Chemistry codes).

We furthermore analyzed the phylogenetic relationships between HERV-K(HML-5) proviruses for five different proviral regions (representing all major retroviral genes) by neighbor-joining and bootstrap analysis. Phylogenetic trees displayed similar branch lengths with respect to the consensus sequence, suggesting a similar nucleotide divergence and age of proviral sequences (data not shown). Neighbor-joining analysis implied two HML-5 subfamilies, but they were not supported in the bootstrap analysis. Only a few chromosome Y sequences were distinguished from the remaining sequences with higher bootstrap support (Fig. 4). An exception was obtained for the region spanning the end of reverse transcriptase until the start of RNase H. In this case, a subfamily of 28 proviral sequences was distinguished with 86% bootstrap support that was mostly associated with LTR22A and LTR22A2 sequences. Eight more LTR22B sequences were separated with 90% support. The remaining 14 sequences (with low bootstrap support) were either LTR22 or LTR22B (data not shown). Taken together, analysis of four out of five internal regions supports the observation that HML-5 proviral bodies do not form distinct subgroups but rather represent a monophyletic group, in contrast to phylogenetically clearly different LTR sequences.

Age of the HERV-K(HML-5) family. A number of repetitive elements were found in HERV-K(HML-5) proviruses. In particular, reasonably old *Alu* subfamilies, such as *AluSg*, *AluSc*, and *AluSp* (Table 1), with approximate evolutionary ages of 31, 35, 37 million years (16), respectively, suggested an evolutionary age of HML-5 proviruses higher than that of other

therefore summarized in one line. Non-HERV-K(HML-5) insertions, such as *Alu* or L1 elements, present in some proviruses were excluded from the figure. Note that the majority of sequences display two commonly deleted regions within the *gag-prt* and the *env* genes, indicated by vertical dotted lines. Some sequences only comprise 5′ or 3′ portions. For *gag-prt*, 5′ and 3′ deletion points clustered between nt 768 to 960 and nt 2061 to 2319, respectively (52% being located at nt 768 and nt 2303). For *env*, clusters ranged from nt 5256 to 6328 and nt 6568 to 6641 (75% located at nt 5314 and 68% at nt 6641). The thick black lines numbered 1 through 5 indicate regions in the multiple sequence alignment for which phylogenetic analysis was performed. RT, reverse transcriptase.

TABLE 1. HERV-K(HML-5) loci investigated in this study[a]

No.	Chr. band	Size (bp)	Accession no.	Repeats	No.	Chr. band	Size (bp)	Accession no.	Repeats
1	1p35.2	829	AL662906.4(104050–104880)+		72*	4q32.1	6,276	AC107056.5(8220–14493)+	
2*	1p22.2	3,721	AC019187.3(10111–13833)−		73*	5p14.3	3,537	AC026722.4(96454–99985)+	
3*	1q22	5,409	AL135927.14(71991–77401)+	*AluSp*/*AluYa5*	74	5p14.1	662	AC027333.5(168790–169434)+	
4	1q23.1	12,084	AL359753.9(44427–56512)+	L1PA2/L1PA3	75*	5p13.2	7,877	AC008925.4(50885–58763)+	*AluY*
5*	1q25.3	5,167	AL133383.10(88786–93954)+		76*	5q33.1	2,096	AC022106.5(50979–53076)+	
6	1q31.3	651	AL592144.5(148187–148839)+		77*	6p22.3	7,742	AL591416.4(97763–105506)+	
7*	1q32.1	2,719	AL359837.21(112252–114972)+		78*	6p22.2	3,242	AL031777.5(62600–65843)+	
8*	1q32.1	2,718	AC022000.1(57476–60195)+		79*	6p21.32	7,287	AL645931.7(57633–64454)+	
9*	1q42.12	7,611	AL162738,12(4411–12023)+	LTR12/LTR30/LTR58	80	6p21.32	136	AL645931.7(64451–64588)+	
10*	1q43	5,604	AL591686.9(118293–123898)+	*AluY*	81*	6p21.31	1,922	AL138721.16(57057–58972)+	*AluSg* (in 5′ LTR)
11	10q11.21	5,689	AC068707.6(109183–114873)+	*AluSp*	82*	6p12.3	4,531	AL359458.17(66575–71107)+	
12	11q12.1	5,134	AP001652.4(156224–161359)+		83*	6p12.1	4,889	AL450489.12(45284–50174)+	
13*	11q14.1	7,752	AP003398.2(153858–156829)+		84*	6q12	5,251	AL078597.11(46513–51765)+	*AluY* (in LTR)
14*	11q21	6,698	AP000870.4(65492–72191)+	*AluY*	85*	6q14.1	7,151	AL391843.13(673–7825)+	
15*	11q22.1	4,878	AP003403.2(164665–169544)+		86*	6q14.3	4,494	AL357272.10(14829–19324)+	
16	11q22.1	771	AP000798.4(44733–45495)+		87*	6q16.3	7,830	AL161720.15(12037–19868)+	*AluY*
17	11q22.1	603	AP000798.4(45495–46081)+		88*	7p14.1	4,097	AC073068.8(158047–162145)+	
18*	12p13.31	4,026	AC092746.9(16238–20258)−		89	7p14.1	640	AC 72061.8(15984–16579)+	
19	12p13.31	242	AC092746.9(13074–13317)−		90*	7q21.11	1,619	AC096562.1(41174–42804)+	
20*	12q12	4,957	AC079601.22(79916–84874)−		91	7q22.3	5,307	AC004855.1(66147–71455)+	
21*	12q12	4,887	AC076972.16(29993–34881)−		92*	7q31.31	7,753	AC004536.1(33287–41051)+	
22	12q15	4,267	AC020656.30(52510–56778)+	LTR7B (in 5′ LTR)	93*	8p23.2	5,113	AC087369.5(90669–95783)−	
23*	12q21.32	4,966	AC087865.8(151169–156136)+		94*	8p23.2	5,043	AC087369.5(36733–41775)−	
24	12q22	5,220	AC018475.27(165538–170759)−	LTR30/LTR12/LTR128	95*	8P23.2	5,819	AC087369.5(14280–20100)−	
25*	12q23.1	6,091	AC010203.13(53087–59179)−		96*	8p23.1	6,272	AC055869.4(23969–27126)+	*AluY*/LTR7/HERV-H
26*	12q24.31	3,930	AC005858.1(7141–11072)+		97*	8p21.2	1,480	AC024958.8(167718–169199)−	
27*	13q12.12	2,594	AL354798.13(87421–90016)+		98*	8q11.1	5,213	AC104576.5(126768–131982)−	*AluY*
28*	13q21.33	5,034	AL139001.14(80673–85708)+		99*	8q12.1	5,355	AC091561.4(180441–185797)−	
29*	13q33.1	4,742	AL158063.12(110923–115660)+		100*	8q21.3	3,720	AC110012.3(134095–137806)+	
30*	14q21.1	2,851	AL356800.3(99279–102131)−		101*	9p24.1	2,439	AL133547.16(38846–41286)+	
31*	14q21.2	3,680	AL161752.4(85483–89164)−		102*	9p23	6,219	AL451129.6(16932–23152)+	
32*	14q31.1	8,654	AC018513.5(19822–28477)+	L1PA6/LTR12B/LTR30/LTR12	103*	9q31.1	5,187	AL353805.20(30116–35304)+	*AluY*
33	14q32.33	5,131	AB019437.1(130399–135531)−		104*	Xp21.3	7,304	AC024024.6(102529–109834)−	L1PA5
34	15q15.1	4,345	AC087878.7(92196–96542)−	*AluSg*	105*	Xp21.3	1,130	AC005297.1(121542–122663)−	
35*	15q21.2	6,206	AC025040.7(64493–65286)+	L1PA4	106*	Xp11.1	2,576	AL354756.18(14248–16813)+	LTR15
36*	15q21.3	4,766	AC068723.5(70564–75331)+		107*	Xq11.2	6,619	AL158203.12(31906–38526)+	
37*	16q11.2	4,828	AC002519.1(63018–67847)−	L1M1	108*	Xq11.2	5,610	AL353744.18(117924–123535)+	*AluYb9*
38*	16q11.2	1,962	AC116553.1(78427–80390)+		109	Xq12	2,643	AL445523.11(68589–71233)+	
39	16q11.2	670	AC106785.2(45130–45801)+		110*	Xq13.2	7,328	AL356513.11(68589–71233)+	*aluYb9*
40	16q11.2	6,776	AC106785.2(47835–54612)+	*AluY*/LTR19	111*	Xq21.1	4,840	AL592171.8(18287–23126)+	
41*	18q12.1	4,077	AC074237.4(26672–30750)−		112*	Xq21.1	3,222	AL592563.7(16340–19563)+	
42	18q21.2	687	AC091135.9(62209–62897)+		113	Xq21.32	8,115	AL390840.17(189393–197509)+	L1PA5/L1P (in 5′ LTR)
43	18q21.2	2,626	AC091135.9(62917–65544)+						*AluY*/*AluSp*
44*	19p13.11	5,715	AC123912.1(76824–82540)+	*AluY*	114*	Xq22.1	6,473	AL133277.12(1570–8044)+	
45*	19p12	2,998	AC011493.4(81474–84473)−		115*	Xq27.3	3,446	AL589669.11(157039–160486)+	
46	19p12	4,950	AC011493.4(6132–62291)−	*AluSC*/*AluY*/LTR7/HERV-H	116*	Xq28	5,541	U69569.1(1847–7389)−	
47	19p12	5,495	AC011503.4(27693–33189)+	*AluSq*	117*	Yp11.2	9,515	AC007274.3(24960–34472)+	LTR6A/LTR12C
48*	19p12	3,594	AC011503.4(47638–51233)+	*AluSP*/MER9	118*	Yp11.2	4,211	AC007275.4(152708–156920)+	LTR17
49	19q11	5,625	AC006504.1(39162–44788)+	*AluYc*	119*	Yp11.2	3,051	AC016749.4(131255–134304)+	*AluY*
50*	19q13.2	2,583	AC005337.1(9746–12325)+	*AluY*/L1PA5	120	Yp11.2	8,661	AC009952.4(75135–83798)+	*AluY*/MER11C
51	2p23.3	2,713	AC010896.15(27276–29990)+	LTR17/*AluY*/*AluY*	121*	Yp11.2	7,321	AC006986.3(138260–145567)+	*AluY*/MER11C
52*	2p16.1	4,931	AC010738.5(72193–77125)−	MER2	122*	Yq11.221	5,070	AC006383.2(68883–73954)+	
53*	2p13.2	7,718	AC007878.2(183549–185134)+		123	Yq11.222	5,120	AC009233.3(17211–22332)+	
54*	2q21.1	6,694	AC068282.2(50964–57659)+	*AluSc*	124*	Yq11.222	4,573	AC009233.3(108140–112714)+	*AluY*/LTR12
55	2q24.3	5,920	AC092583.2(123236–129157)+	LTR12C	125*	Yq11.223	6,119	AC007678.3(88618–94734)+	*AluSc*/*AluSp*/*q*
56*	2q31.1	7,767	AC092641.2(5721–13489)−		126*	Yq11.223	5,261	AC009494.2(123311–128573)+	*AluSc* (in 3′ LTR)

57	2q31.1	6,672	AC078883.6(131188–137864)+		127*	Yq11.223	7,720	AC009489.3(81570–89291)+	
58*	2q37.1	5,575	AC009407.8(76629–82201)+	LTR10F/HERVIP10FH/LTR10F	128*	Yq11.223	5,420	AC007876.2(100813–106234)+	
59*	20p13	5,040	AL049761.11(10942–15983)+		129	Yq11.223	12,395	AC009239.3(23104–35500)+	LTR12B/LTR12C/LTR12/HERV9
60	20p11.21	654	AL031673.19(32505–33148)+						HERVIP10FH/AluY (in 3′ LTR)
61*	3q26.1	4,716	AC018457.14(44247–48964) −		130*	Yq11.223	8,661	AC021107.3(71095–79756)+	*AluY*/MER11C
62*	3q26.2	1,878	AC008040.7(24180–26059) −		131*	Yq11.223	7,336	AC025227.6(46631–53968)+	
63*	4p13	1,446	AC096586.3(21933–23380) −	*AluY*	132*	Yq11.223	1,397	AC007320.3(110244–111627)+	
64*	4q13.1	2,569	AC097648.2(119741–122311)+		133*	Yq11.223	9,515	AC010080.2(28375–37888)+	LTR6A/LTR12C
65*	4q13.3	1,288	AC024722.5(111362–112651)+		134	Yq11.223	12,888	AC006366.3(74013–86902)+	HERVIP10FH/*AluYa5*/LTR12C
66*	4q13.3	4,976	AC024722.5(112653–117630)+						LTR30/LTR12/L1PA3
67*	4q21.23	3,800	AC097488.2(13713–17514)+		135	Yq11.223	206	AC01088.3(205–412)+	
68*	4q24	4,831	AC107381.2(38346–43178)+		136*	Yq11.223	9,515	AC009947.2(47848–57361)+	LTR6A/LTR12C
69*	4q28.3	6,627	AC096763.2(19403–26031)+		137	Yq11.223	886	AC010153.3(22418–23305)+	
70	4q28.3	6,635	AC108867.3(109215–115851)+	LTR12C/LTR22	138	Yq11.23	886	AC016728.4(82188–83075)+	
71*	4q32.1	4,137	AC009567.8(47850–51988) −		139*	Yq11.23	9,515	AC006991.3(107651–117164)+	LTR64/LTR12C

[a] HERV-K(HML-5) loci are numbered consecutively. Chromosomal (Chr.) band locations as well as sizes of proviruses or remnants of loci are given. Accession numbers of finished Genbank human genome sequence entries, including nucleotide position within the entry and orientation (+ or −) of the HERV-K (HML-5) portion are given in the fourth column. The last column lists non-HERV-K(HML-5) repetitive elements within proviral sequences.

```
HML-5 own   TCCTCACACGGGGCACCA
HML-5 Rep   .......G..........
Met(CAT)    ..................
Ile(TAT)    .......CT..A......
Lys(CTT)    G..C...GTT....G...
Lys(TTT)    GT.C.TGTTCA.......
Arg(CCT)    GT.C...CT....T....
```

FIG. 2. Sequence comparison of the HERV-K(HML-5) consensus sequence primer binding site with reported tRNA 3′ ends (http://rna.wustl.edu/GtRDB). Shown here are the HERV-K(HML-5) consensus primer binding site regions from Repbase (Rep) and as generated in this study (own) and the closest matching methionine, isoleucine, lysine, and arginine tRNA 3′ ends. Anticodon sequences are given in brackets.

HERV-K families. The age of a provirus can be estimated from sequence comparison of flanking 5′ and 3′ LTRs. Owing to the retroviral reverse transcription strategy, both LTR sequences are identical in sequence at the time of provirus formation. Without selective pressure, both LTRs independently accumulate mutations over time. Thus, sequence differences between a provirus' LTRs are an approximate measure of provirus age (8). We determined the degree of sequence divergence between 5′ and 3′ LTRs with larger overlapping portions for 53 HML-5 proviruses. We obtained sequence divergences ranging from 6 to 24% (mean, 12.9%; standard deviation, 3.87). These numbers equal an approximate age of 50 (±15) million years for the HML-5 proviruses (Table 2). We furthermore calculated the average ages of proviral sequences from Kimura-2-parameter corrected distances to the HML-5 consensus sequence for five different proviral regions, excluding gaps and CpG dinucleotides (16). Here, an average evolutionary age of about 60 (±27) million years was indicated. The age of roughly 55 million years from both analyses thus corresponds to a HML-5 integration time into the genome clearly before the evolutionary split of Old World from New World monkeys that took place about 40 million years ago. This observation is in contrast to other HERV-K families described till now because those are not present in New World monkeys, suggesting that HML-5 represents an ancient betaretrovirus family in primates. To investigate the possibility that HML-5 elements are present in New World primate genomes, we examined the species distribution of HML-5 *gag* and *env* regions by PCR with genomic DNA from hominoids, Old World monkeys, New World monkeys, and prosimians. The amplified portion in the *gag* gene was located outside the above-described commonly deleted region, whereas the amplicon in *env* included the frequently deleted region. We obtained *gag* and *env* PCR products of expected sizes from all species tested, except for prosimians (Fig. 5). This result was in good agreement with the above-mentioned age estimate from LTR and consensus sequence analysis. Full-length *env* genes, present in a minority of HML-5 proviruses in the human genome, were amplified as a PCR product from all species. PCR products corresponding to the *env* deletion variant were also amplified, indicating the presence of both longer and shorter *env* variants in all tested species.

We cloned PCR products from the *gag* region obtained from *P. troglodytes*, *H. lar*, *M. mulatta*, and *A. seniculus* and se-

TABLE 2. Assignment of LTR sequences associated with HERV-K(HML-5) proviral bodies[a]

No.	Annotation				Age	No.	Annotation				Age	No.	Annotation				Age
	Repbase		Ours				Repbase		Ours				Repbase		Ours		
	5′ LTR	3′ LTR	5′ LTR	3′ LTR			5′ LTR	3′ LTR	5′ LTR	3′ LTR			5′ LTR	3′ LTR	5′ LTR	3′ LTR	
1	22A	no	22A			48	22A (−)	22A		22A		95	22	22B	**22B**		
2	22A	22A	22A	22A	50.8	49	22	22B		22B		96	22	22	**22A2**	**22A2**	
3	22	22 (−)				50	22	no				97	22A	22A (−)	22A		
4	22	22				51	22A	no	22A			98	22	22	22	22	59.1
5	22	22 (−)				52	22	22A (−)	**22A2**			99	22	no			
6	no	22				53	22A	22A	22A	22A	38.4	100	no	no			
7	22A	22A	22A	22A		54	22	22	22	22	47.9	101	22B (−)	22B		22B	
8	22A	22A	22A	22A		55	22	22				102	22	22	amb.	amb.	50.3
9	22	22				56	22	22	22	22	36.1	103	22	22	amb.	amb.	36.7
10	22	22B		22B		57	22B	22B	22B	22B	35.4	104	22A	no	22A		
11	22B	22B (−)	22B			58	22	22		**22A2**		105	22	no			
12	22	22				59	22	22	**22A2**	**22A2**	47.7	106	no	22B		22B	
13	22	22	**22A2**	**22A2**	40.6	60	no	no				107	22B	22B	22B	22B	51.8
14	22	no	22			61	no	22		**22A2**		108	22A	22A	22A	22A	48.6
15	22	22				62	22A	22A	22A	22A	30.4	109	22	no			
16	22A (−)	no				63	22	no				110	22	22	22	22	47.5
17	22A (−)	no				64	22A	22A	22A	22A	52.5	111	22A (−)	22A		22A	
18	no	no				65	22A	22A (−)				112	no	22		**22B**	
19	no	no				66	22A	22A	22A	22A		113	22	22			
20	22A	22A	22A	22A	38.5	67	22	22 (−)				114	22	22	amb.	amb.	38.7
21	22A	22A	**22A2**	**22A2**	38.2	68	22	22	22	22	68.6	115	22A	22A	22A	22A	34.7
22	22	22 (−)				69	22 (−)	22	**22A2**			116	no	22		22	
23	22	22	**22A2**	**22A2**	45.3	70	22 (−)	22				117	no	22		**22B**	
24	22B	22B	22B	22B	32.3	71	22	22				118	22/22A	22/22A			
25	22	22				72	22	22	22	22		119	no	no			
26	22A	22A	22A	22A	54.5	73	no	22A		22A		120	22B	22B	22B	22B	86.7
27	22A	22A (−)				74	22A	no	22A			121	no	22		**22B**	
28	22	22	**22A**	**22A**	38	75	22	22				122	22A	22A	22A	22A	
29	22	22	**22A2**	**22A**	42.2	76	22B	22B	22B	22B	43	123	22B	22B	22B	22B	77.9
30	no	no				77	22	22	22	22	25.2	124	22 (−)	no			
31	22	22				78	22A (−)	22A	22A	22A		125	22A	no	22A		
32	22B	22B	22B	22B	37.5	79	no	no				126	22	22	amb.		
33	22B (−)	22B		22B		80	no	no				127	22B	22B	22B	22B	92.7
34	no	22				81	22A	no				128	no	22		amb.	
35	22A	22A (−)	22A			82	22A	22 (−)	22A			129	22	22B			
36	22	22	amb.	amb.	59.7	83	22	22	22	22		130	22B	22B			
37	22	no	**22A2**			84	22	22				131	no	22		**22B**	
38	22	22				85	no	22		**22A2**		132	no	no			
39	no	22B		22B		86	22A	no	22A			133	no	22			
40	22B	22B	22B	22B	54.4	87	22	22	22	22	56.1	134	22	22			
41	22	22 (−)	**22A2**	**22A2**	56.6	88	22A	22A	22A	22A	41.1	135	no	no			
42	no	no				89	22	no				136	no	22			
43	22A	no	22A			90	no	22B				137	22	no			
44	22A	no	22A			91	22B	22	22B			138	22	no			
45	22A	22A	22A	22A	68	92	22	22	22	22	52.8	139	no	22			
46	22	22				93	22A	22A	22A	22A	52.6						
47	22A	22A		22A		94	22	22A	**22A**	22A	70.6						

[a] Numbers in the first column correspond to those in Table 1. Annotations of 5′ and 3′ LTRs according to RepeatMasker and to our own analysis are indicated. Differences between RepeatMasker and our own annotations are indicated in bold. no, the corresponding LTR was missing; (−), LTRs having reverse complement orientations as annotated by RepeatMasker. Our own annotations include LTRs for which subfamily assignment was possible. Here, "amb." indicates that ambiguous LTRs could not be assigned to a particular family; that is, they grouped phylogenetically between the LTR22 and LTR22B subfamilies (data not shown). The evolutionary ages of proviruses in millions of years according to Kimura-2-parameter corrected distances are listed for proviruses with sufficient 5′ and 3′ LTR portions.

quenced, in total, 16 clones. The sequences were included in the neighbor-joining analysis presented in Fig. 2. The nonhuman sequences grouped among the human sequences; that is, they did not form separate branches or groups. Also, the percentage of identity with the human consensus sequence was very similar: 82 to 91% compared to 78 to 92% for the human sequences. Therefore, our sample sequence data set does not indicate different evolutionary behavior of HERV-K(HML-5) homologous sequences in the examined nonhuman primates. However, more elaborate studies will be required to characterize precise evolutionary behavior of HERV-K(HML-5) homologues in primates after their evolutionary separation from the human lineage.

DISCUSSION

HERV represent former exogenous retroviruses that are fossilized in the human genome. Closer examination of HERV sequences provides information on ancient primate-targeting retroviruses and retrovirus evolution in general. The completed sequence of the human genome provides an excellent source of information for such studies. In this paper, we set out to analyze a hitherto little-characterized HERV family, HERV-K(HML-5), in more detail. We found that only 9 out of the approximately 139 HML-5 loci (6%) displayed full-length retroviral genes flanked by LTRs on both sides. A higher number of HML-5 loci are defective regarding proviral structures; *gag-prt* and/or *env* regions are missing in about 50% of proviruses, and the start and end points of deletions obviously cluster within defined regions. Analogous observations were recently made for HERV-K(HML-3), displaying deletions within *gag* and *pol* (26). Amplification of deleted proviruses has been observed also for other HERV families, for example, deleted HERV-H elements have increased to high copy numbers during primate evolution in comparison to full-length HERV-H proviruses (24). Recombination on the DNA level, mutations during retroviral reverse transcription, or splicing of retroviral transcripts before reverse transcription and provirus formation could account for such deleted proviruses. Spliced and reintegrated transcripts have been observed for HERV (12, 21). Spliced human immunodeficiency virus type 1 transcripts were also found as cDNA along with full-length retroviral RNA during infection (20), further supporting provirus formation from spliced proviral transcripts in the evolutionary past. Besides active amplification of proviral sequences, about 5% of proviral loci probably arose passively from chromosomal duplications, owing to the fact that the human genome comprises about 5% of duplicated sequences (14). In contrast, we do not find evidence that HML-5 sequences were amplified by L1-mediated pseudogene formation, as recently described for HERV-W (6, 37).

HML-5 loci with both *gag-prt* and *env* deletions probably emerged during reverse transcription by recombination between RNA transcripts from proviruses having either deletion. In combination with HERV-K(HML-3) deletion variants (26), we note that despite the lack of one or more proviral regions, the 5′ intergenic and *gag* portions are usually present in obviously (retrovirus-like) retrotransposed proviruses. Those retroviral regions are known to encode the packaging signal Ψ that interacts on the RNA level with the Gag-encoded nucleocapsid (NC) protein. One may therefore hypothesize that interaction between retroviral RNA and NC is still essential during intracellular (retrovirus-like) retrotransposition of HERV. Alternatively, helper viruses could have been involved in the formation of new proviruses, and a packaging signal could have interacted with the helper virus' NC protein. Also in combination with previous results (9, 26), deletions within the *env* gene seem to occur recurrently, rendering the Env protein nonfunctional. Env-demolishing mutations could have resulted in decreasing production of infectious retrovirus. Such decrease could have significantly added to the fading of exogenous stages, and *env* deletions may therefore represent another cause for the extinction of exogenous stages.

Our study shows that HML-5 sequences were fixed in an ancestral genome after the simian lineage had evolutionary separated from the prosimian lineage but before the evolutionary separation of Old World and New World primates, indicated by provirus ages of approximately 55 million years and corroborated by PCR examination of various primate species. Other HERV-K superfamily members were fixed in an ancestral genome approximately 30 to 40 million years ago, after the evolutionary split of Old World from New World primates (26, 28, 34, 39). To the best of our knowledge, HML-5 represents the oldest betaretrovirus in the primate genome known to date. Despite a relatively high proportion of incomplete HML-5 proviruses in the human genome, our study generated a consensus sequence displaying the four major retroviral ORFs *gag*, *prt*, *pol*, and *env*. The corresponding proteins displayed significant similarities to other betaretroviruses, and the recreated HML-5 consensus is therefore expected to be very close in sequence to the former exogenous precursor to HML-5 endogenous variants. Thus, this study reconstructed an ancient betaretrovirus that was targeting primates about 55 million years ago.

Phylogenetic analysis of several proviral regions indicated similar sequence divergence among HML-5 sequences, resulting in almost monophyletic tree structures. This finding indicates that probably all HML-5 sequences were generated in a relatively brief evolutionary period around 55 million years ago, the latter as revealed by LTR-LTR divergences and divergence from the consensus sequence. Formation of new proviruses then obviously ceased. Proviruses with deletions in *gag-prt* and *env* were probably also generated in a relatively short period of time. This observation is further confirmed by PCR analysis that revealed *gag* and *env* deletions in all tested HML-5-positive primate species. Thus, different HML-5 "master proviruses" with different genome structures were obviously active and generated provirus progeny. In this manner, HML-5 displays similar behavior as HERV-K(HML-3) (26). However, both families display different behavior than HERV-K(HML-2), which formed proviruses during the hominoid period as well as in recent human evolution (5, 7, 28, 33).

Phylogenetic analysis of LTR sequences associated with proviral bodies revealed several apparent LTR families. However, phylogenetic analysis of proviral body sequences yielded monophyletic tree topologies for most examined regions. Only a region between reverse transcriptase and RNase H displayed branches with higher bootstrap support. Clearly, proviral bodies appear much more homogeneous in sequence than the associated LTR sequences. Thus, almost homogeneous provi-

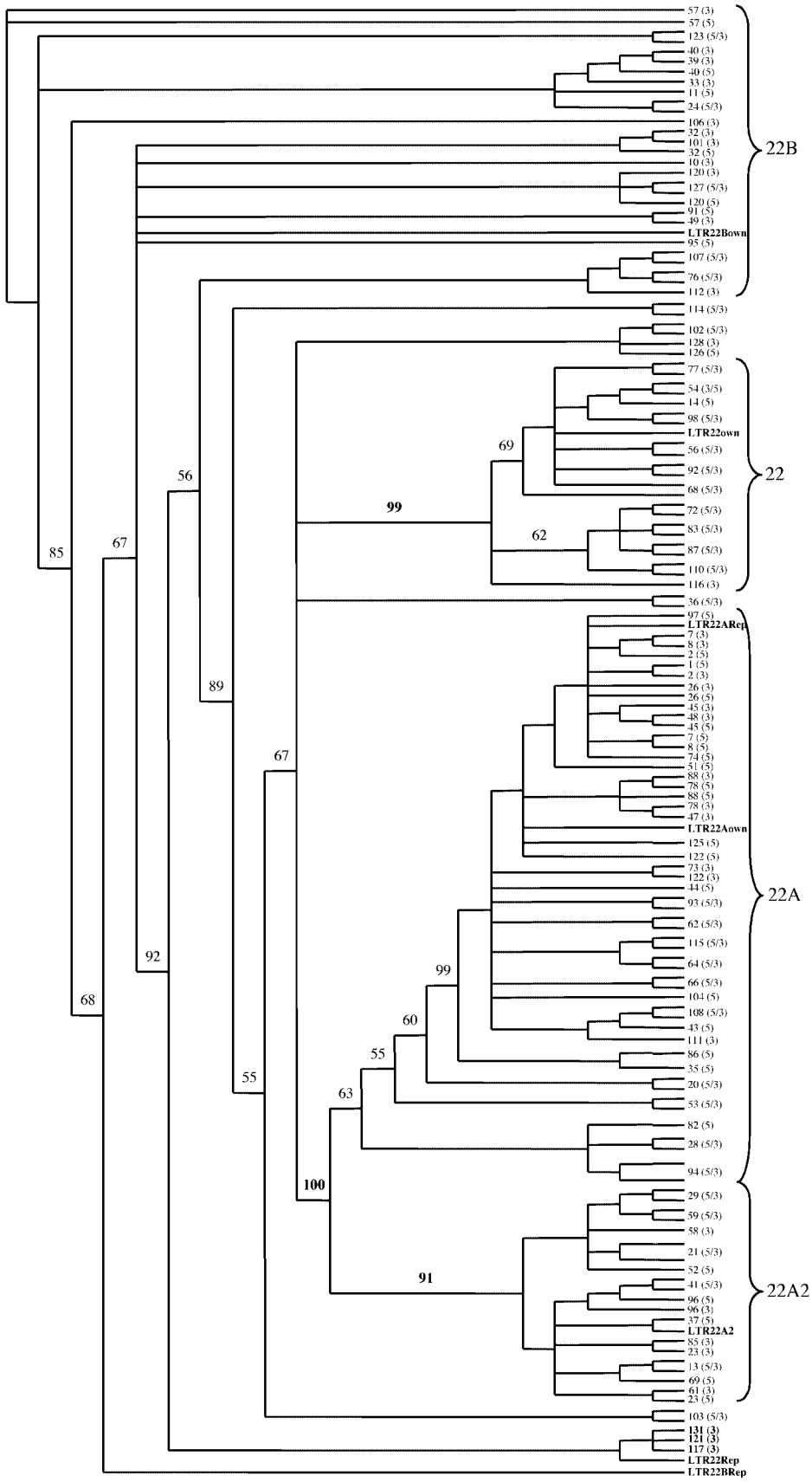

FIG. 3. Neighbor-joining analysis of 134 reasonably intact HERV-K(HML-5) LTR sequences. Shown here is a consensus tree from 1,000 bootstraps replicates. Specific bootstrap values at major nodes are indicated. Consensus sequences, as given in Repbase (Rep) and as generated in this study (own), were also included in the analysis. LTR22 and LTR22A sequences are clearly distinguished, with the latter comprising the so-called LTR22A2 group. Other sequences belong to the LTR22B group with some ambiguous sequences remaining (Table 2). Note that the

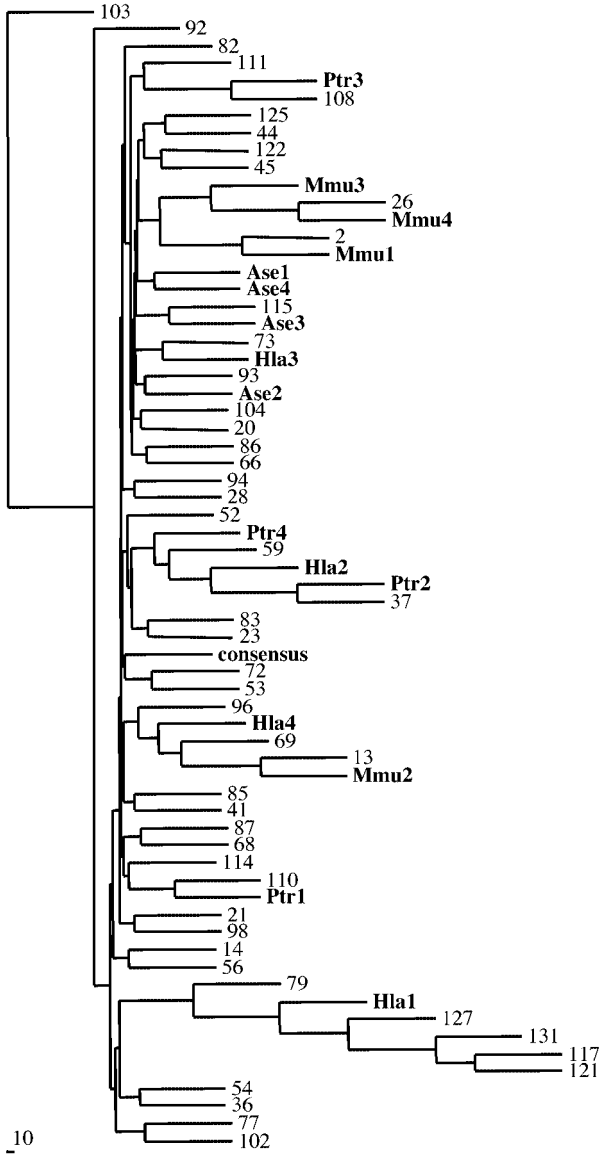

FIG. 4. Neighbor-joining analysis of HERV-K(HML-5) proviruses. The tree shown here was obtained from the analysis of a *gag* gene portion (region 1 in Fig. 1) and is a consensus tree from 100 bootstrap replicates. Numbers of proviral sequences refer to Table 1. Phylogenetic analysis included HERV-K(HML-5) homologous sequences from the same *gag* region, obtained by PCR from various primate species (for species abbreviations, see the legend for Fig. 5).

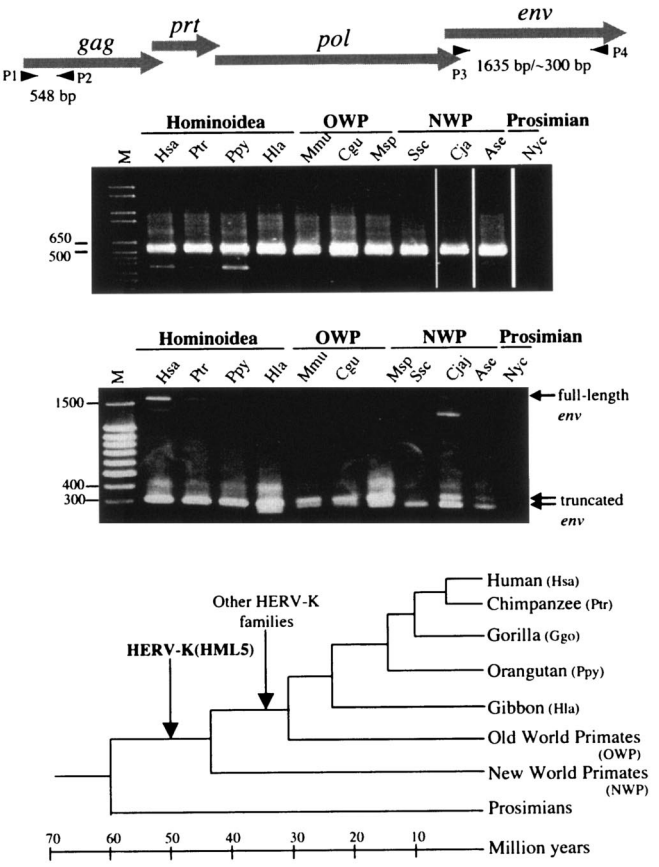

FIG. 5. Presence of HERV-K(HML-5) homologous sequences in various primate species. The schematic depicts localization of PCR primer pairs within the proviral body 5′ end and within the *env* gene region of the HERV-K(HML-5) consensus sequence, yielding PCR products of about 548 bp and about 1,635 or 300 bp. The latter number considers the common presence of deleted *env* gene regions. PCR results are presented in the first and second panel, respectively. Amplification of full-length *env* genes in all species but the prosimian could not be reproduced satisfactorily. A schematic of primate phylogeny and germ line fixation of HERV-K(HML-5) versus other HERV-K families is shown at the bottom. Species abbreviations are as follows: for *Hominoidea*, Hsa is *Homo sapiens*, Ptr is *Pan troglodytes*, Ppy is *Pongo pygmaeus*, and Hla is *Hylobates lar*; for Old World primates, Cgu is *Colobus guereza*, Msp is *Mandrillus sphinx*, and Mmu is *Macaca mulatta*; for New World primates, Ssc is *Saimiri sciureus*, Cja is *Callithrix jacchus*, and Ase is *Alouatta seniculus*; for prosimians, Nyc is *Nycticebus coucang*.

ral bodies were associated with clearly different LTR variants at the time of provirus formations. It is currently not known whether the different LTR variants were already present in the exogenous precursor(s) or represent derivatives from a germ line-fixed LTR founder family. In both cases, LTR sequences evolved in sequence independently from, and obviously more rapidly than, the proviral bodies. Reasons for apparently different evolutionary rates of LTRs and proviral bodies are currently not clear.

Whether HML-5 is ancestral to other HERV-K families currently seems unclear. Both the HML-5 and HML-6 families appear less related to the remaining HERV-K families, as evidenced by DNA sequence comparisons (32) and by phylogenetic comparison of dUTPase domains (30), for instance. In

LTR22 Repbase sequence groups with LTR22B sequences, opposed to LTR22own. Numbers of LTR sequences refer to the proviruses listed in Table 1. The 5′ and 3′ LTRs of a particular provirus located at immediately neighboring terminal nodes are indicated as 5/3. Note in this context that 5′ and 3′ LTRs of particular proviruses do not always group together, reminiscent of recently suggested HERV-involving genomic rearrangement events (13).

addition, this study revealed that the HML-5 consensus sequence primer binding site is identical to methionine but clearly less similar to lysine tRNA 3′ ends; therefore, it is actually a HERV-M family and adds to the lesser phylogenetic relationship with HERV-K. However, HML-5 is still more related to HERV-K than to other (beta)retroviruses. Also in the course of this study, when employing an updated tRNA sequence data set (22) we noted that the previously characterized HERV-K(HML-3) family primer binding site region (26) is more related to arginine and asparagine than to lysine tRNA 3′ ends, therefore actually requiring designation HERV-R or HERV-N. At the present time, the precise phylogenetic relationship of HML-5 to other HERV-Ks as well as endogenous and exogenous retroviruses must await further specific investigations. We are currently studying the remaining HERV-K families in a similar fashion. Certainly, results for other hitherto little-described HERV families will significantly contribute to such studies.

ACKNOWLEDGMENTS

This work was supported by grants from the Deutsche Forschungsgemeinschaft to J.M. (Ma2298/2-1) and E.M. (Me917/16-1). P.M. is supported by a fellowship from the Knut and Alice Wallenberg Foundation and grants from the Swedish Research Council, Åke Wiberg Foundation, and Magn Bergvall Foundation. W.S. is supported by DFG grant SCH214/7-3.

REFERENCES

1. **Andersson, M. L., M. Lindeskog, P. Medstrand, B. Westley, F. May, and J. Blomberg.** 1999. Diversity of human endogenous retrovirus class II-like sequences. J. Gen. Virol. **80:**255–260.
2. **Berkhout, B., M. Jebbink, and J. Zsiros.** 1999. Identification of an active reverse transcriptase enzyme encoded by a human endogenous HERV-K retrovirus. J. Virol. **73:**2365–2375.
3. **Blond, J. L., D. Lavillette, V. Cheynet, O. Bouton, G. Oriol, S. Chapel-Fernandes, B. Mandrand, F. Mallet, and F. L. Cosset.** 2000. An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor. J. Virol. **74:**3321–3329.
4. **Boissinot, S., and A. V. Furano.** 2001. Adaptive evolution in LINE-1 retrotransposons. Mol. Biol. Evol. **18:**2186–2194.
5. **Buzdin, A., S. Ustyugova, K. Khodosevich, I. Mamedov, Y. Lebedev, G. Hunsmann, and E. Sverdlov.** 2003. Human-specific subfamilies of HERV-K (HML-2) long terminal repeats: three master genes were active simultaneously during branching of hominoid lineages. Genomics **81:**149–156.
6. **Costas, J.** 2002. Characterization of the intragenomic spread of the human endogenous retrovirus family HERV-W. Mol. Biol. Evol. **19:**526–533.
7. **Costas, J.** 2001. Evolutionary dynamics of the human endogenous retrovirus family HERV-K inferred from full-length proviral genomes. J. Mol. Evol. **53:**237–243.
8. **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.
9. **de Parseval, N., V. Lazar, J. F. Casella, L. Benit, and T. Heidmann.** 2003. Survey of human genes of retroviral origin: identification and transcriptome of the genes with coding capacity for complete envelope proteins. J. Virol. **77:**10414–10422.
10. **Felsenstein, J.** 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Department of Genetics SK-50, University of Washington, Seattle.
11. **Gifford, R., and M. Tristem.** 2003. The evolution, distribution and diversity of endogenous retroviruses. Virus Genes **26:**291–315.
12. **Goodchild, N. L., J. D. Freeman, and D. L. Mager.** 1995. Spliced HERV-H endogenous retroviral sequences in human genomic DNA: evidence for amplification via retrotransposition. Virology **206:**164–173.
13. **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.
14. **International Human Genome Sequencing Consortium.** 2001. Initial sequencing and analysis of the human genome. Nature **409:**860–921.
15. **Jurka, J.** 2000. Repbase update: a database and an electronic journal of repetitive elements. Trends Genet. **16:**418–420.
16. **Kapitonov, V., and J. Jurka.** 1996. The age of Alu subfamilies. J. Mol. Evol. **42:**59–65.
17. **Kimura, M.** 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16:**111–120.
18. **Kitamura, Y., T. Ayukawa, T. Ishikawa, T. Kanda, and K. Yoshiike.** 1996. Human endogenous retrovirus K10 encodes a functional integrase. J. Virol. **70:**3302–3306.
19. **Lebedev, Y. B., O. S. Belonovitch, N. V. Zybrova, P. P. Khil, S. G. Kurdyukov, T. V. Vinogradova, G. Hunsmann, and E. D. Sverdlov.** 2000. Differences in HERV-K LTR insertions in orthologous loci of humans and great apes. Gene **247:**265–277.
20. **Liang, C., J. Hu, R. S. Russell, M. Kameoka, and M. A. Wainberg.** 2004. Spliced human immunodeficiency virus type 1 RNA is reverse transcribed into cDNA within infected cells. AIDS Res. Hum. Retrovir. **20:**203–211.
21. **Lindeskog, M., and J. Blomberg.** 1997. Spliced human endogenous retroviral HERV-H env transcripts in T-cell leukaemia cell lines and normal leukocytes: alternative splicing pattern of HERV-H transcripts. J. Gen. Virol. **78:**2575–2585.
22. **Lowe, T. M., and S. R. Eddy.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25:**955–964.
23. **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.
24. **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.
25. **Mager, D. L., and P. Medstrand.** 2003. Retroviral repeat sequences. *In* D. Cooper (ed.), Nature encyclopedia of the human genome. Macmillan Publishers Ltd., Hampshire, England.
26. **Mayer, J., and E. Meese.** 2002. The human endogenous retrovirus family HERV-K(HML-3). Genomics **80:**331–343.
27. **Mayer, J., E. Meese, and N. Mueller-Lantzsch.** 1997. Chromosomal assignment of human endogenous retrovirus K (HERV-K) env open reading frames. Cytogenet. Cell Genet. **79:**157–161.
28. **Mayer, J., E. Meese, and N. Mueller-Lantzsch.** 1998. Human endogenous retrovirus K homologous sequences and their coding capacity in Old World primates. J. Virol. **72:**1870–1875.
29. **Mayer, J., E. Meese, and N. Mueller-Lantzsch.** 1997. Multiple human endogenous retrovirus (HERV-K) loci with gag open reading frames in the human genome. Cytogenet. Cell Genet. **78:**1–5.
30. **Mayer, J., and E. U. Meese.** 2003. Presence of dUTPase in the various human endogenous retrovirus K (HERV-K) families. J. Mol. Evol. **57:**642–649.
31. **Mayer, J., M. Sauter, A. Racz, D. Scherer, N. Mueller-Lantzsch, and E. Meese.** 1999. An almost-intact human endogenous retrovirus K on human chromosome 7. Nat. Genet. **21:**257–258.
32. **Medstrand, P., and J. Blomberg.** 1993. Characterization of novel reverse transcriptase encoding human endogenous retroviral sequences similar to type A and type B retroviruses: differential transcription in normal human tissues. J. Virol. **67:**6778–6787.
33. **Medstrand, P., and D. L. Mager.** 1998. Human-specific integrations of the HERV-K endogenous retrovirus family. J. Virol. **72:**9782–9787.
34. **Medstrand, P., D. L. Mager, H. Yin, U. Dietrich, and J. Blomberg.** 1997. Structure and genomic organization of a novel human endogenous retrovirus family: HERV-K (HML-6). J. Gen. Virol. **78:**1731–1744.
35. **Morgenstern, B.** 1999. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics **15:**211–218.
36. **Mueller-Lantzsch, N., M. Sauter, A. Weiskircher, K. Kramer, B. Best, M. Buck, and F. Grasser.** 1993. Human endogenous retroviral element K10 (HERV-K10) encodes a full-length gag homologous 73-kDa protein and a functional protease. AIDS Res. Hum. Retrovir. **9:**343–350.
37. **Pavlicek, A., J. Paces, D. Elleder, and J. Hejnar.** 2002. Processed pseudogenes of human endogenous retroviruses generated by LINEs: their integration, stability, and distribution. Genome Res. **12:**391–399.
38. **Reus, K., J. Mayer, M. Sauter, H. Zischler, N. Muller-Lantzsch, and E. Meese.** 2001. HERV-K(OLD): ancestor sequences of the human endogenous retrovirus family HERV-K(HML-2). J. Virol. **75:**8917–8926.
39. **Seifarth, W., C. Baust, A. Murr, H. Skladny, F. Krieg-Schneider, J. Blusch, T. Werner, R. Hehlmann, and C. Leib-Mosch.** 1998. Proviral structure, chromosomal location, and expression of HERV-K-T47D, a novel human endogenous retrovirus derived from T47D particles. J. Virol. **72:**8384–8391.
40. **Smit, A. F. A.** 1996. Structure and evolution of mammalian interspersed repeats. Ph.D. dissertation. University of Southern California, Los Angeles.
41. **Sverdlov, E. D.** 2000. Retroviruses and primate evolution. Bioessays **22:**161–171.
42. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:**4673–4680.
43. **Tristem, M.** 2000. Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. J. Virol. **74:**3715–3730.