Journal of General Virology (1999), **80**, 2383–2392. Printed in Great Britain

HERV-F, a new group of human endogenous retrovirus sequences

Christian Kjellman, Hans-Olov Sjögren and Bengt Widegren

University of Lund, Department of Cell and Molecular Biology, Section of Tumor Immunology, Sölvegatan 21, S-223 62 Lund, Sweden

Using primers from a conserved region of the XA34 human endogenous retrovirus (HERV) family, four *pol* fragments originating from new members of the family were amplified from human genomic DNA. Southern blot analysis demonstrated similar hybridization patterns in human, chimpanzee and orangutan and distinct hybridization to macaque DNA. The probes also exhibited weaker hybridization to squirrel monkey DNA. Using large genomic clones, two full-length XA34-related HERVs have been identified. One of the HERVs is located downstream of a human Krüppel-related zinc finger protein gene, *ZNF195*. Both of the newly identified long terminal repeats have potential TATA boxes, poly(A) signals and transcription factor-binding sites but they differ to a high degree, especially in the U_3 region. The primer-binding sites were found to be homologous to $tRNA^{Phe}$ (TTC), and therefore these new HERVs have been given the name HERV-F. The closest relatives to the HERV-Fs are the RTVLH-RGH family. Phylogenetic analyses of the Gag, Pol and Env regions are discussed. Both of the newly identified HERV-Fs were shown to contain *protease*, *reverse transcriptase*, *integrase* and *env* regions and had characteristic deletions in the *integrase* and *env* regions. In addition, the *capsid protein* gene of *gag* and two conserved zinc-binding motifs that are characteristic of a potential nucleic acid-binding protein were also identified. Apart from an ORF spanning the *protease* of one HERV-F, no other longer ORFs were found.

Introduction

The human genome, in common with all mammalian genomes, carries a high load of inheritable retrovirus sequences. About 0·1–1% of the human genome consists of human endogenous retroviruses (HERVs) (Lower *et al.*, 1996; Patience *et al.*, 1997), and an even larger portion of the genome consists of reverse-transcribed and transposed sequences (Smit, 1996). HERVs can be divided into different families based on their sequence similarities. Class I families, e.g. HERV-E (4-1) (Repaske *et al.*, 1983, 1985), HERV-R (ERV-3) (O'Connell *et al.*, 1984), HERV-I (RTVL-Ia) (Maeda, 1985) and HERV-H (RTLV-H2) (Mager & Freeman, 1987), are similar to mammalian type C retroviruses, whereas class II families, e.g. HERV-K (C4) (Dangel *et al.*, 1994) and HTDV/HERV-K (HERV-K10) (Ono *et al.*, 1986), share similarities with mammalian type B and D retroviruses.

The copy number of the HERV families varies from single copies to up to 10^4 copies per genome. These different copy numbers could either represent multiple integration events of the same or a closely related virus, or proviruses amplified after the original integration by, for example, retrotransposition or chromosomal rearrangements. There are indications of a negative correlation between the presence of the *env* gene in the provirus and the copy number, such that proviruses without *env* have a more retrotransposon-like structure and thereby are more amplified (Lower *et al.*, 1996). Even though the copy number of complete proviruses is sometimes rather limited, solitary long terminal repeats (LTRs) are more frequent. This excision of protein-coding regions from HERVs might be the result of it being harmful for an organism to carry and express functional HERVs.

The majority of the HERVs that are fixed in the population are ancient from an evolutionary point of view (at least 10–100 million years old). All known HERVs are defective and cannot produce functional infectious retrovirus particles. However, all

Author for correspondence: Christian Kjellman.
Fax +46 46 222 9251. e-mail Christian.Kjellman@wblab.lu.se

The GenBank accession numbers of the sequences reported in this paper are AF070683, AF070684, AF070685 and AF070686.

HERVs are not transcriptionally silent and different proviruses with ORFs have been reported (Lower *et al*., 1995; O'Connell *et al*., 1984). Moreover, retrovirus particles of HERV origin have also been reported (Lower *et al*., 1993).

Using low-stringency hybridization with an ERV-9 *env* probe, a new HERV was identified from a glioma cDNA library (Widegren *et al*., 1996). All of the endogenous retrovirus (ERV) cDNA clones isolated were polyadenylated, the longest clone being 2303 bp. The ERV was named XA34 after this cDNA clone. The clone contained the 3′ end of the *reverse transcriptase* (*RT*) region, a somewhat truncated *endonuclease* (*IN*) region and only a short fragment of a *transmembrane* (*TM*) region. Southern blot analysis demonstrated that XA34 belongs to a family of HERVs with approximately 16 members (full length or close to full length) in the human genome. XA34-related elements exist in all Old World monkeys investigated. Southern blot hybridizations using an XA34 *pol* probe revealed a distinct but rather weak signal from a New World monkey. If this is genuine, it indicates that the first XA34 was incorporated into the primate genome more than 60 million years ago (Arnason *et al*., 1996).

We previously published sequence data from five members of this family (Widegren *et al*., 1996), and we now provide data from another five closely related elements. From the human genome project, it was possible to identify two large genomic clones that contain XA34-related proviruses. These proviruses are described and characterized in detail in this paper.

Methods

■ **Accession numbers.** HERV-Fa was identified at bases 44530–51154 in a genomic clone with accession number Z95126 and HERV-Fb at bases 89430–95587 in a genomic clone with accession number AC000378. When we refer to positions in these HERVs, base 44530 in Z95126 is regarded as base 1 of HERV-Fa and base 89430 in AC000378 is regarded as base 1 of HERV-Fb. The retrovirus sequences used for the phylogenetic analysis were (with accession numbers in parentheses): XA34 (U29659), XA35 (U37054), XA36 (U37067), XA37 (U29658), XA38 (U37066), XA39 (AF070683), XA40 (AF070684), XA41 (AF070685), XA42 (AF070686), RTVLH-RGH1 (D10083), RTVLH-RGH2 (D11078), ERV9-1 (X57147 and M37638), Moloney murine leukaemia virus (MuLV) (J02255–J02257), feline leukaemia virus (FeLV) (M18247 and M19392), RTVL-H2 (M18048), baboon endogenous retrovirus (BaEV; REPCBCG) (M16550), HUMER4-1 (M10976, K02168 and K02169), gibbon ape leukaemia virus (PCGGPE) (M26927), reticuloendotheliosis virus (REV) (AF006065) and HERV-W (AC000064).

■ **Cloning.** To identify fragments from XA34-related proviruses, PCR amplification from male genomic DNA was performed for 20–30 cycles with a thermal profile of 95 °C for 30 s, 50 °C for 60 s and 72 °C for 45 s. The reaction mixture (20 µl) contained 0·1 µg of each of primers 593 (5′ CGATGATCAACTATTCATAGATGG 3′) and 2793 (5′ TGGTGAGAGCTATGAGTTCTGC 3′), 100 ng total DNA, 2·5 mM $MgCl_2$, 2 µM dNTPs, 0·5 µM α-^{35}S-labelled dATP, standard buffer and 1 U Ampli-*Taq* polymerase (Perkin-Elmer Cetus). The PCR products were separated on an 8% wedge-shaped denaturing sequencing gel and detected by autoradiography. To recover the DNA from the dried sequencing gel, a gel slice containing the DNA fragment was excised and rehydrated in TE buffer. The sample was boiled for 10 min and used for re-amplification with the same primers under the PCR conditions described above. The amplification efficiency was increased by using 20 µM instead of 2 µM dNTPs and the re-amplified fragments were cloned into the pT7blue vector (Novagen) and sequenced.

■ **Computer analysis.** The Wisconsin package version 9.1-unix (Genetics Computer Group, Madison, WI, USA) (Devereux *et al*., 1984) was used for DNA sequence analysis. The database used for NetBlast homology searches was a continuously updated version of the GenBank nucleotide database. Alignments were made primarily with the PILEUP and LINEUP programs of the GCG package. Phylogeny was analysed with a UNIX version of PAUP (Swofford, 1992). Trees were constructed with a heuristic tree-search and bootstrap analysis for 1000 replications with the parsimony optimality criterion by using PAUP and the trees were displayed with DRAWGRAM of the PHYLIP software package (Felsenstein, 1993). Percentage identity was determined by using FASTA of the GCG package. The TFSEARCH and MOTIF programs were used with a threshold score of 85·0 to search the on-line TRANSFAC database (Heinemeyer *et al*., 1998) for potential transcription factor-binding sites.

■ **Southern blot analysis.** The *pol* fragments of 155–156 bp from XA39, XA40, XA41 and XA42 were used for Southern blot analysis. These *pol* probes were labelled with [α-^{32}P]dCTP by means of PCR from cloned material. The labelling reactions (20 µl) contained 3 ng template, 10 ng primer 593, 50 µM each of dATP, dTTP and dGTP, 15 µM [α-^{32}P]dCTP, 2·5 mM $MgCl_2$, PCR buffer (Perkin-Elmer Cetus) and 1 U Ampli-*Taq* polymerase (Perkin-Elmer Cetus). The reactions were run for six cycles of 94 °C for 60 s, 54 °C for 60 s and 72 °C for 4 min.

Total genomic DNA from human, chimpanzee (*Pan troglodytes*), orangutan (*Pongo pygmaeus*), squirrel monkey (*Saimiri sciureus*), macaque (*Macaca fascicularis*) and rat (*Rattus norvegicus*) was digested with *Pst*I and separated on a 0·7% agarose gel and the resultant DNA was vacuum-blotted to a Biodyne B membrane (PALL). Hybridization was carried out overnight in Rapid Hybridization solution (Amersham) at 65 °C. The filter was washed in several steps with decreasing concentrations of SSC/SDS and an increasing temperature. The final washing conditions were 1 × SSC–0·5% SDS at 65 °C for 90 min.

Results

Isolation and characterization of XA34-related elements

In a previous study, we reported sequence information from five HERVs isolated either as cDNA clones (XA34), genomic clones (XA38) or by PCR (XA35–XA37) (Widegren *et al*., 1996). Using this information, PCR primers were constructed from a conserved portion of the *RT* region. Amplification was performed under low-stringency conditions from human genomic DNA and the PCR products were separated by PAGE into two size groups of approximately 153–158 and 140–146 bp (Fig. 1). DNA from two gel fragments (A and B; Fig. 1) was recovered by PCR and cloned and 25 individual clones were sequenced and analysed. Nine different XA34-related clones were identified (Table 1) and four of these clones, derived from the longer fragment (B; Fig. 1), are previously unidentified XA34-related elements. These four elements were named XA39, XA40, XA41 and XA42. Phylogenetic distance

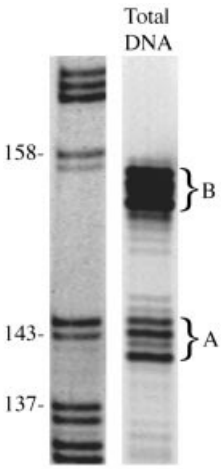

Fig. 1. Denaturing sequencing gel of PCR products from human total DNA with primers preferentially amplifying a *pol* fragment from HERVs related to XA34. The PCR primers were 100% identical to XA34. The PCR products were labelled with ^{35}S-dATP during the amplification (50 °C, 2·5 mM $MgCl_2$). The regions corresponding to A and B were cut from the gel and the DNA was re-amplified, cloned and sequenced. On separation under denaturing conditions, each homogenous PCR product appeared as four bands, representing the two separated DNA strands, with or without the extra base added by *Taq* polymerase. A sequencing reaction was used as a molecular mass marker.

Table 1. PCR products related to XA34

PCR products were amplified by using primers from the conserved *RT* region of XA34. Sizes of the different amplicons include the primer sequence. Percentage nucleotide identity, excluding primer sequences, to the consensus sequence is shown.

HERV	Region*	Size (bp)	Identity (%)
XA34	A	142	81
XA35	B	154	89
XA36	B	154	88
XA37	B	155	81
XA38	A	144	73
XA39	B	156	82
XA40	B	155	90
XA41	B	155	91
XA42	B	155	92
HERV-Fa†	–	(155)	88
HERV-Fb†	–	(155)	91

* The region in Fig. 1 from which the element was isolated.
† The HERV-Fa and HERV-Fb sequences are from the genomic clones.

analysis clearly grouped these new elements in the XA34 family (data not shown) and there was a high degree of nucleotide identity between the *pol* fragments (Table 1).

Southern blot analysis of *Pst*I-digested genomic DNA from male and female human, chimpanzee, orangutan, squirrel monkey, macaque (male only) and rat (male only) demonstrated that XA39–XA42 all exhibited similar patterns of hybridization to those of the previously identified XA34-related elements (Widegren *et al.*, 1996). The hybridization patterns of XA39 and XA42 are shown in Fig. 2 (*a*, *b*). The probes detected similar patterns in human, chimpanzee and orangutan. There was also hybridization to macaque DNA and a rather weak but distinct hybridization to DNA from squirrel monkey, a New World monkey (Fig. 2*b*).

Analysis of genomic clones

The sequencing of the human genome is continuously adding the sequences of large genomic clones to the databases. XA34-related DNA sequences were used to screen databases for similar elements to this family. A NetBlast search revealed the existence of two large genomic clones that contained sequences with similarities to the XA34 *pol* region. Since we originally reported the XA34 family, we believe that it is our responsibility to characterize these more complete proviruses further.

The first of the two XA34-related elements identified is contained within a 177 kb genomic clone (accession number Z95126; bases 44530–51154) isolated from human chromosome Xq21.1–Xq21.3. Of the newly identified HERVs, XA42 contains the most similar *pol* region, exhibiting 90% identity over the region analysed. When the genomic clone was mapped by using *Pst*I, the predicted fragment corresponding to the *pol* probe shown in Fig. 2 was 2224 bp in length. As this virus is located on the X chromosome, it would be expected that this probe would result in a stronger band when hybridizing to female rather than male DNA. Such a band can be observed at 2·2 kb, as shown in Fig. 2. When referring to base positions in this HERV, base 44530 in the genomic clone (Z95126) is denoted as base 1.

The second genomic clone (accession number AC000378), of 133 kb, was isolated from human chromosome 11p15.5. This clone contains an XA34-like HERV situated between bases 89430 and 95587, with a *pol* sequence that is 100% identical to XA41 over the region analysed. Base 89430 in clone AC000378 is denoted as base 1 in this HERV. On mapping the genomic clone with *Pst*I, the predicted fragment corresponding to the *pol* probe shown in Fig. 2 was found to be 3485 bp in length. A human Krüppel-related zinc finger gene, *ZNF*195 (accession number AF003540; Hussey *et al.*, 1997), was found to be located upstream of the HERV; the longest *ZNF*195 expressed sequence tag (EST) identified from GenBank (accession number AA505095) was found to terminate 243 bp upstream of this HERV element with a poly(A) tail. ESTs terminating within the HERV sequence, and thereby representing possible read-through transcripts from *ZNF*195, have also been identified (C. Kjellman & B. Widegren, unpublished results).

A schematic alignment of the two XA34-like HERVs, XA34

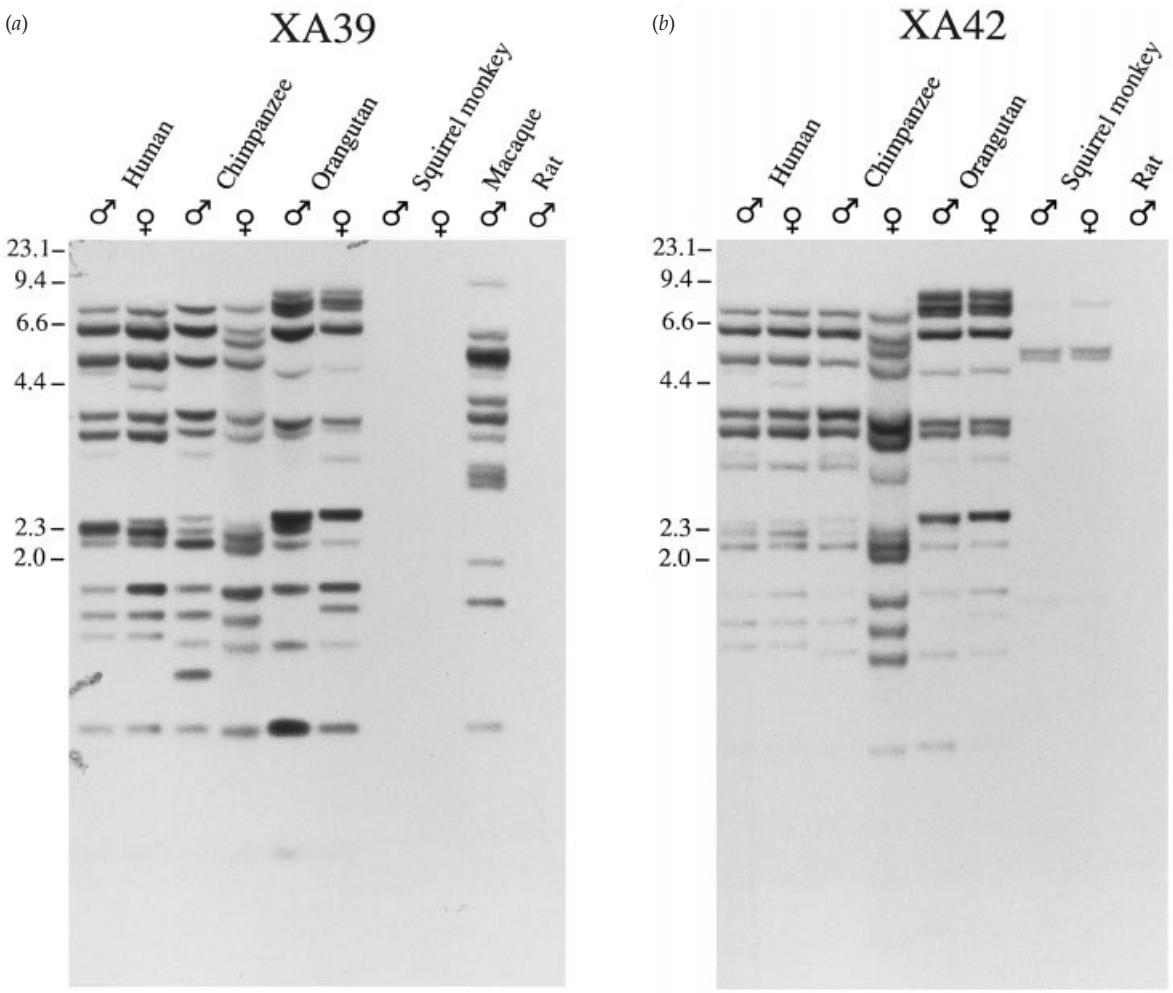

Fig. 2. Southern blot analysis with the (*a*) XA39 and (*b*) XA42 *pol* fragments hybridized to total genomic DNA of human, chimpanzee, orangutan, squirrel monkey, macaque and rat. The male or female origin of the DNA is indicated in the figure. The blots were hybridized at high stringency [65 °C in 100% Rapid Hybridization buffer (Amersham); the washing conditions were 1 × SSC–0·5% SDS at 65 °C for 90 min].

and XA38 is shown in Fig. 3. Further analyses of different regions in the two XA34-related HERVs are described below. Within the genomic clone Z95126, it was possible to identify two similar (89% identical) 450 bp potential LTR regions separated by 5·7 kb and organized as direct repeats (positions 1–453 and 6167–6614). Similarly, in the second genomic clone, AC000378, it was also possible to identify two direct repeats (435 bp) sharing 90% identity and separated by 5·3 kb (positions 1–436 and 5709–6146). In clone Z95126, a direct repeat of 5 bp was located directly upstream of the potential 5′ LTR and downstream of the potential 3′ LTR, thus delineating the boundaries of the HERV (Fig. 4). The alignment of these LTR regions in Fig. 4 demonstrates the loss of the dinucleotides AA at the 5′ end and TT at the 3′ end from the HERVs as a result of processing during the integration event (Temin, 1981). There is also a short (5 bp) inverted repeat starting with the nucleotides TG at the ends of the potential LTRs, as underlined in Fig. 4. Because of the low sequence similarity to other characterized LTRs, it is not feasible to use homology alone to define different regions of the potential LTRs. However, it is possible to identify the TATA box (at base 340 in Fig. 4), the poly(A) signal (at base 410 in Fig. 4) and clustered potential sites for binding of transcription factors such as Sp1, GATA, AML-1a, STATx, CREB, cEBP, AP-4 and AP-1 (data not shown) in the LTR-like regions. The junction between the R and U_5 sequences is located at the polyadenylation site 16–25 bp downstream of the poly(A) signal and the U_3–R junction is found at the CAP site 20–30 bp downstream of the TATA box (Guntaka, 1993). The potential U_3 regions from these two HERVs share very little sequence similarity, whereas the potential R and U_5 regions exhibit a high degree of identity (75%) (Fig. 4).

The primer-binding site has previously been used in the classification of endogenous retroviruses (Larsson *et al.*, 1989). We have been able to identify a region located just downstream of the 5′ LTR in both Z95126-HERV and AC000378-HERV

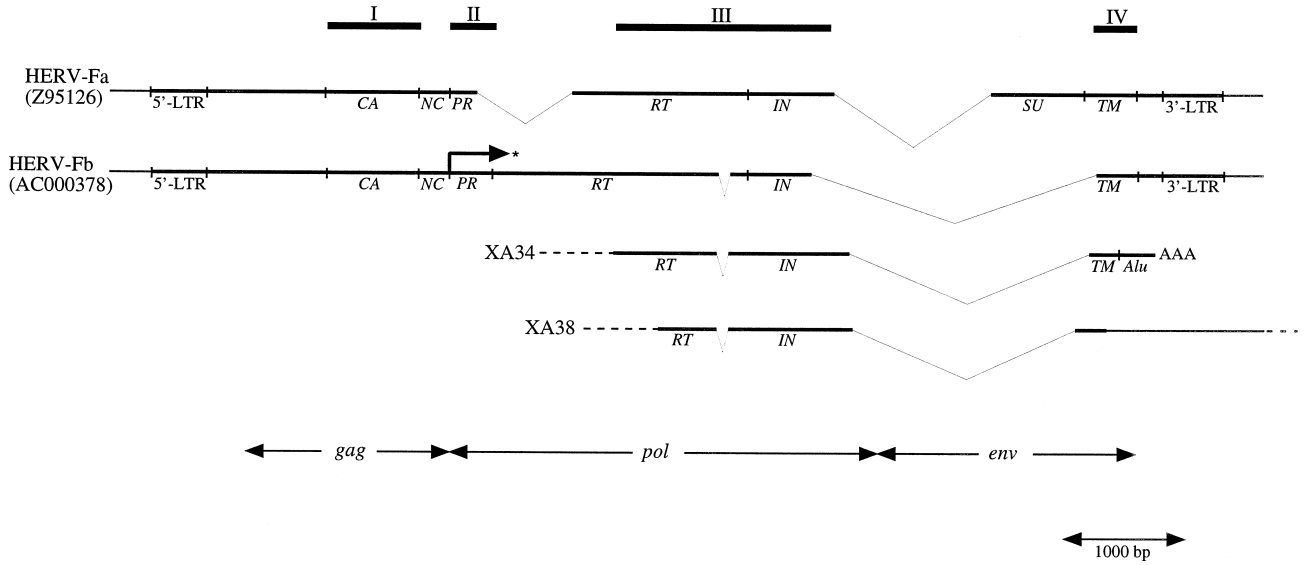

Fig. 3. Schematic alignment of HERV-Fa, HERV-Fb, XA34 & XA38. The LTRs, *CA*, *NC*, *PR*, *RT*, *IN*, *SU* and *TM* regions are indicated. Angled lines indicate regions presumably deleted from the HERVs as compared with e.g. the RTVLH-RGH family. XA34 was isolated as a cDNA and not a genomic sequence [indicated in the figure by the poly(A) tail]. The ORF of the *PR* gene in HERV-Fb is indicated. The regions used for phylogenetic analyses are indicated by lines marked I–IV above the figure. Dotted lines indicate that the elements are probably longer; however these regions are not yet cloned.

that is similar to the 18 bp region at the 3′ end of a human $tRNA^{Phe}$ (TTC) (accession number K00350). It was found that the Z95126-HERV shares 100% identity to $tRNA^{Phe}$ (TTC), whilst AC000378-HERV deviated by only one mismatch in the 18 bp region (Fig. 4). By applying the taxonomic nomenclature of Larsson *et al*. (1989), we suggest that the Z95126-HERV and the AC000378-HERV are named HERV-Fa and HERV-Fb, respectively.

Downstream from the 5′ LTRs, both HERV-Fa and HERV-Fb have potential leader/*gag* regions (HERV-Fa, bases 460–2408; HERV-Fb, bases 434–2435) of 2 kb. HERV-Fa and HERV-Fb share similarity (70% nucleotide identity) over the complete *gag* region. However, at the nucleotide level, it is very difficult to align accurately the first 1·5 kb after the LTR of HERV-Fa or HERV-Fb with the *gag* regions of known retroviruses or retrovirus sequences. The final 0·5 kb of the *gag* region is more similar (55% nucleotide identity) to RTVLH-RGH2. However, the deduced amino acid sequences of the complete *capsid* (CA) region (region I in Fig. 3) of HERV-Fa (bases 1440–2199) and HERV-Fb (bases 1464–2220) were aligned over 267 amino acids with the corresponding regions of known HERVs. Only a single maximum-parsimony tree was reconstructed (Fig. 5*a*) following 1000 bootstrap replications using the heuristic tree search. This tree groups the two HERV-Fs together and demonstrates that the HERV-H family, as represented by RTVL-H2 and RTVLH-RGH2, is the most closely related ERV family. Over the CA region, the amino acid identity between HERV-Fa and HERV-Fb is 63% and it is 35% between RTVL-H2 and HERV-Fa. Within the CA region in both HERV-Fa and HERV-Fb, a conserved major homology region was identified with 50% (HERV-Fa) and 35% (HERV-Fb) amino acid identity to the MuLV major homology region (Strambio-de-Castillia & Hunter, 1992; Zlotnick *et al*., 1998). Immediately after the CA region, both HERV-Fa and HERV-Fb contain lysine-rich basic sequences with two conserved zinc-binding motifs ($CX_2CX_4HX_4C$). These are characteristic of potential nucleic acid-binding protein (NC) regions. The coding potential of the HERV-Fa and HERV-Fb *gag* regions is disrupted by stop codons and introduced frame-shift mutations (Fig. 6).

The genes of the *pol* region encoding *protease* (*PR*), *RT* and *IN* of HERV-Fa and HERV-Fb were identified by aligning the *pol* regions to RTVL-H2 and RTVLH-RGH2. The *PR* gene (bases 2409–2739) and the *RT* gene (bases 2740–4117) of HERV-Fa have an approximately 780 bp deletion that has removed the C-terminal portion of *PR* and the N-terminal portion of *RT*. In contrast, HERV-Fb contains the complete *PR* (bases 2436–2820) and *RT* genes (bases 2821–4802). HERV-Fb has an ORF encoding a complete *PR* gene product, which terminates at the beginning of *RT* (Figs 4 and 6). The maximum-parsimony tree (Fig. 5*b*) reconstructed from the deduced and aligned amino acids of the PR region (region II in Fig. 3) demonstrates that HERV-Fb should be assigned to the same branch as, but separate from, RTVLH-RGH2 and RTVL-H2. Over the PR region, the amino acid identity between HERV-Fb and RTVLH-RGH2 was 39%.

HERV-Fa and HERV-Fb share 80% sequence identity (at the DNA level) in the *RT* gene; this represents the most conserved genetic region within the retroviruses. A maximum-parsimony tree (Fig. 5*c*) was reconstructed after a heuristic tree search and 1000 bootstrap replications with the aligned sequences of 651 amino acids spanning part of the RT region

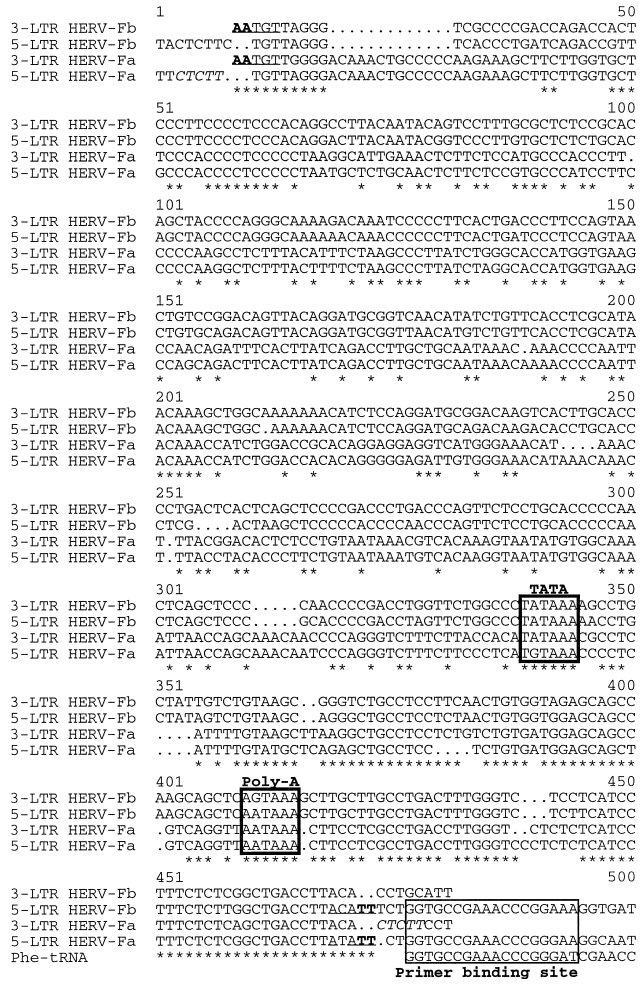

Fig. 4. The aligned 3′ and 5′ LTRs of HERV-Fa and HERV-Fb. The primer-binding sites are aligned together with the reversed 3′ tail of human $tRNA^{Phe}$ (TTC). The dinucleotides AA at the 5′ end and TT at the 3′ end that were lost from the HERVs during integration are indicated in bold in the corresponding 3′ and 5′ LTRs. Five bp inverted repeats at the ends of the HERV-F LTRs are underlined. A direct repeat of 5 bp (italic letters) is located directly upstream of the potential 5′ LTR and downstream of the potential 3′ LTR of HERV-Fa, representing the duplication of target DNA. Potential TATA boxes and polyadenylation signals are boxed. Except for the first ten base pairs of the LTRs, the nucleotide sequences of HERV-Fa and HERV-Fb deviate completely until the potential TATA box at position 338. From the TATA box until the end of the LTR, the sequences are more closely related. Nucleotides that are identical in at least one HERV-Fa LTR and one HERV-Fb LTR are indicated by asterisks.

and part of the IN region (region III in Fig. 3), and this tree clearly groups HERV-Fa, HERV-Fb and XA34 together and separates them from the RTVLH-RGH family. Over ~ 205 amino acids in the C terminus of the conserved RT, there is approximately 38–42% amino acid identity between the HERV-Fs and XA34 and RTVLH-RGH1 (Table 2). The amino acid identity between the HERV-Fs, HUMER4-1, ERV9, HERV-W and RTVLH-RGH1 is summarized in Table 2. HERV-Fa, HERV-Fb, XA34 and XA38 all carry large deletions that result in somewhat truncated C-terminal portions of the *IN* gene and a truncation (HERV-Fa) or deletion (HERV-Fb, XA34 and XA38) of the *surface protein* (*SU*) gene in the *env* region. However, given that the truncations observed in the four different HERVs are not in exactly the same position within the *IN* gene (300 bp variation) (Fig. 3), these are probably due to independent events. HERV-Fa has the deletion at position 5034 and HERV-Fb has the deletion at position 5150. There are no ORFs that have the potential to encode functional proteins within the *RT* or truncated *IN* regions (Fig. 6).

The *env* region of HERV-Fa (bases 5034–6018) contains a large C-terminal portion of the *SU* gene and a complete *TM* gene, as identified by alignment to RTVLH-RGH2. In HERV-Fb, the *SU* gene and the N-terminal portion of the *TM* gene are deleted. The sequences of HERV-Fa and HERV-Fb are similar (about 60% nucleotide identity) along the entire length of the 3′ end of the virus until the beginning of the 3′ LTR, with the similarity to RTVLH-RGH2 ending just downstream of the *TM* gene. Phylogenetic analysis was performed on the conserved TM region (region IV, Fig. 3) spanning 121 amino acids that also includes the CKS-17 motif, which has been reported to have immunosuppressive activity (Cianciolo *et al.*, 1985). Only a single maximum-parsimony phylogenetic tree was reconstructed (Fig. 5*d*) following bootstrap analysis (1000 replications) with a heuristic search. The analysis groups HERV-Fa, HERV-Fb and XA34 together whilst separating them from the RTVLH-RGHs, ERV-9 and HERV-W. The *env* region of XA34 that was used for the phylogenetic analyses includes a region downstream of Alu that was obtained from a genomic clone (accession number AC005281) (C. Kjellman & B. Widegren, unpublished results). The *env* regions of the HERV-Fs do not have the potential to encode functional peptides. The sequence AGGAGGTTTGAA is found 163 bp downstream of the *env* region and immediately upstream of the 3′ LTR in HERV-Fa. This is most likely to be a conserved potential polypurine tract. A potential polypurine tract with the sequence AAAAGGCTAAAA is also found in HERV-Fb, 174 bp downstream of *env*.

The Southern blot analysis indicates the presence of HERV-F-related elements in the genome of a New World primate (squirrel monkey) (Fig. 2*b*). One means of identifying the time at which the ERV was integrated is to analyse the two LTRs of a specific provirus and compare the divergence between these two LTRs and relate it to the divergence of other cellular genes from different species (Dangel *et al.*, 1995). In order to be able to do this, one has to assume that the LTRs of a particular ERV have evolved independently since the time of integration. One also has to assume that the LTRs were identical when the retrovirus was originally integrated. Analysis of the LTRs of HERV-Fa shows that there are 36 base pair exchanges in 2×442 bp and 40 base pair exchanges in 2×433 bp in HERV-Fb. This gives an approximate divergence of 0·041 base pair exchanges per position for HERV-Fa and 0·046 base pair exchanges per position for HERV-Fb. If the same calculations are made for the LTRs of HERV-K (C4) (Dangel *et al.*, 1994, 1995), a divergence of 0·039 base pair exchanges per position

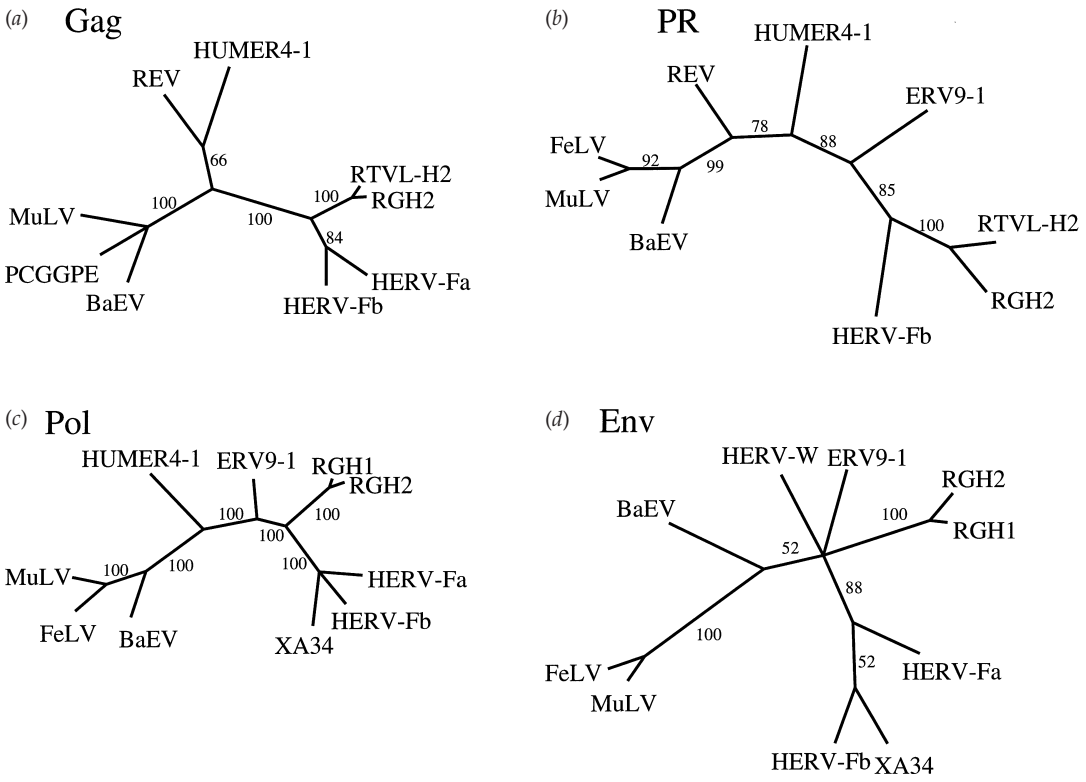

Fig. 5. Phylogenetic analyses of the aligned deduced amino acid sequences of CA of Gag (*a*), PR (*b*), the partial RT/IN region of Pol (*c*) and the partial TM region of Env (*d*) from the HERV-Fs and different retrovirus sequences. These regions are indicated as I–IV in Fig. 4. The sequences were aligned using PILEUP and LINEUP and the phylogenetic analyses were made using PAUP. Only single maximum-parsimony trees were reconstructed by using the heuristic tree-search for 1000 bootstrap replications. The bootstrap values are indicated on the trees, which were displayed using DRAWGRAM.

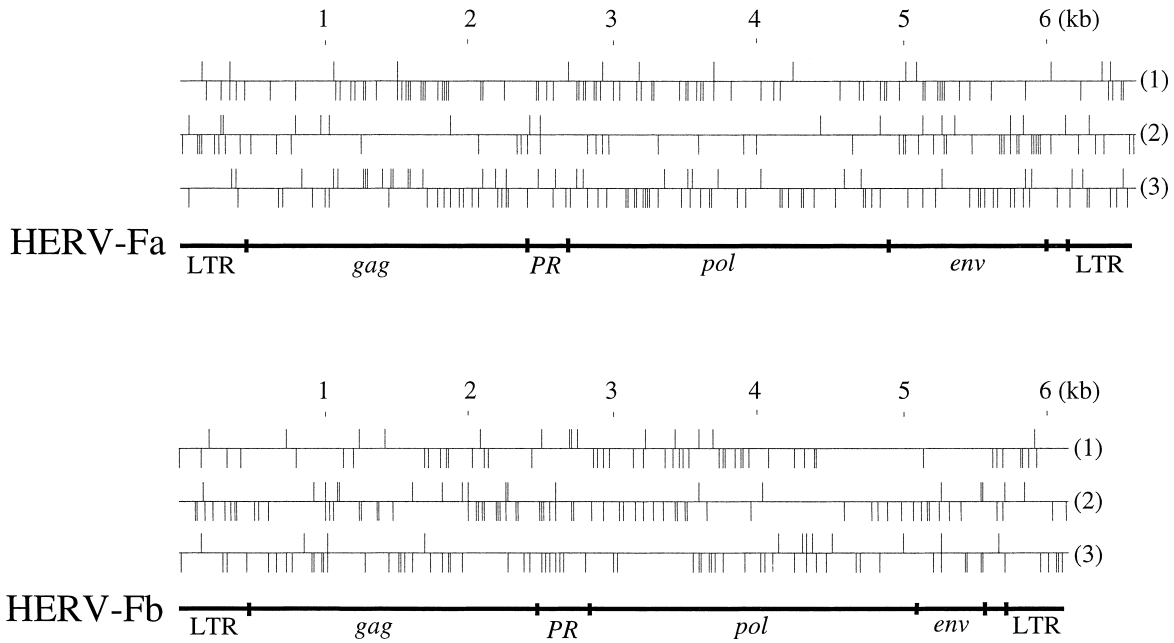

Fig. 6. Start (above the line) and stop (below the line) codons in frames 1–3 (indicated by numbers in parentheses) of the HERV-F sequences illustrate the disrupted reading frames of HERV-Fa and HERV-Fb. The analysis was made by using FRAMES (GCG).

Table 2. Percentage amino acid identity of a conserved ~ 205 amino acid region of RT

	HUMER4-1	HERV-W	ERV9	RGH1	XA34	HERV-Fb
HERV-Fa	38	41	36	42	53	64
HERV-Fb	36	37	33	38	53	
XA34	31	34	35	39		
RGH1	36	38	39			
ERV9	41	67				
HERV-W	45					

(42 bp exchanges in 2 × 538 bp) is obtained. The time for the integration of HERV-K has been estimated to be just after the split between the New World and Old World primates, as based on the divergence of the short intron 9 of the C4 gene (Dangel *et al*., 1995). The divergence of this marker between human and macaque (*Macaca mulatta*) is approximately 0·029 base pair exchanges per position (24 bp exchanges in 2 × 412 bp) and that between human and cotton top tamarin (*Sanguinus oedipus*), a New World monkey, is approximately 0·053 base pair exchanges per position (43 bp exchanges in 2 × 409 bp). If this approximation was widened to include the HERV-Fs, it would indicate that HERV-K (C4) and the HERV-Fs were integrated at approximately the same time, i.e. after the split of the New and Old World monkeys. However, it should be stressed that there are no guarantees of equal mutation rates in different HERV integrations.

Discussion

We have isolated nine separate *pol* sequences demonstrating homology to class I HERVs and have suggested previously that these HERVs should be regarded as a separate family of retroviruses (Widegren *et al*., 1996). Two complete proviruses have now been characterized that demonstrate strong sequence similarity to XA34. On the basis of their primer-binding sites, they have been designated HERV-Fa and HERV-Fb. The related elements XA34–XA42 share strong similarity to these HERV-Fs, but the primer-binding sites of these elements have not yet been identified. However, on the basis of sequence homology, we suggest that XA34–XA42 should also be classified in this HERV-F family of endogenous retroviruses.

The LTRs of the HERV-Fs were identified and found to share no sequence similarity to LTRs of other known HERVs. The LTRs of HERV-Fa differ strongly from HERV-Fb within the 5′ region (potential U_3) but are more similar within the 3′ region (R and U_5). A high degree of variation in the U_3 region, particularly in the enhancer sequences, has been demonstrated in other retrovirus families (Majors, 1990). Given that the LTRs of HERV-Fa and HERV-Fb share very little sequence similarity in U_3, these two proviruses are likely to be the result of two separate germline infections of distinct but related retroviruses. This is also supported by the fact that there is no sequence similarity between the flanking regions of the two HERV-Fs, clearly demonstrating that these proviruses are not the result of a gene amplification.

The HERV-F *gag* genes differ strongly from other known *gag* genes; however, the potential *CA* gene and conserved zinc-binding motifs of a lysine-rich potential *NC* gene were identified in both HERV-Fs. These gene products usually represent the most conserved proteins of the *gag* region (McClure *et al*., 1988). It was found that the HERV-Fs and the related XA34 and XA38 all have deletions within the C-terminal portion of the *IN* gene and the N-terminal portion of the *SU* gene. The fact that these deletions have not taken place at exactly the same location in each gene indicates that these deletions have occurred as separate events that for some reason have been directed towards this particular region. One can speculate that these deletions took place before the integration into the germline, thus facilitating the retroviruses in becoming endogenous. It is also possible that the deletions took place after integration, thereby silencing both the *IN* and the *env* genes. The truncation of the *env* gene gives the HERV a more retrotransposon-like structure. Except for an ORF spanning the *PR* region of HERV-Fb, the HERV-Fs were not found to contain any longer ORFs within the *gag*, *pol* or *env* regions.

HERV-Fa and HERV-Fb can be grouped together by phylogenetic analysis of their capsid, protease, polymerase and transmembrane protein regions. These analyses and the rather low degree of similarity in RT demonstrate that they are clearly separated from other known HERVs, the HERV-H family being their closest relative. The phylogeny of the different regions follows a similar pattern, in that neither of the two HERV-Fs shows any sign of recombination with other HERV families. We have discussed previously a possible recombination between XA34 and ERV-9, since the region upstream of the XA34 Alu is rather similar to the ERV-9 *env*, but here we also included the *env* region located downstream of the Alu in the analyses and we cannot conclude that such a recombination has occurred.

Southern blot hybridization using *pol* probes from different HERV-F-related proviruses demonstrates the presence of approximately 16 HERV-F family members in the human genome. Database analysis of genomic sequences also revealed a number of elements that contain just HERV-F-like LTRs and *gag* regions (e.g. accessions Z86001, Z83745 and AC002416). Therefore, it is probably correct to assume that the HERV-F family, in common with many other ERV families, contains a large number of truncated HERV remnants dispersed over the genome. The 16 members identified by Southern blot hybridization define the number of *pol*-containing HERV-F-related elements in the genome.

Southern blot hybridization of XA34-related elements demonstrated the presence of these HERVs in all Old World

primate samples analysed. However, it was also possible to detect HERV-F bands in New World primates, e.g. XA42 (Fig. 2*b*), therefore suggesting that the first integration took place more than 60 million years ago (Arnason *et al.*, 1996). Of course, it is also possible that the hybridization to the New World primate DNA reflects cross-hybridization to the more conserved *pol* regions of unrelated proviruses. However, the absence of signal from the hybridization to rat DNA, which carries many C-type ERVs, is an indication of a rather high specificity of the hybridization. The specificity is also supported by the fact that cross-hybridization to the HERV-H family, which is the most closely related family, with approximately 1000 members (Mager & Henthorn, 1984), would be expected to give a much larger number of bands than the relatively small number detected with the XA39 and XA42 probes. We also tried to relate the divergence between the 5′ and 3′ LTRs of the HERV-F family with the divergence of the HERV-K (C4) LTRs and that of the short intron 9 of the C4 gene (Dangel *et al.*, 1995). The time of integration for HERV-K was estimated to be just after the split between the New World and Old World primates (Dangel *et al.*, 1995). Our analyses indicated that the HERV-F LTRs and HERV-K (C4) LTRs have the same degree of divergence, indicating that these HERVs were integrated at about the same time. This calculation contradicts the interpretation of the Southern blot analysis, as this indicates that the HERV-Fs were integrated before the split between the New World and Old World primates. However, it should be remembered that there is an inevitable degree of uncertainty when calculating rates of divergence, particularly over such short genetic regions, and that there are no guarantees of equal mutational rates in the different HERV elements. There is of course also the possibility of cross-hybridization between conserved *pol* regions of different retrovirus origin in Southern blot analysis. In order to be able to determine with absolute certainty whether the HERV-F family is present in New World primates, the potential elements should be cloned and sequenced.

We would like to thank Ms Ingar Nilsson for skilful technical assistance. This work was supported by the Swedish Cancer Foundation, the Swedish Medical Research Council, the Nilsson-Ehle Foundation, the Blücher Foundation, the John and Augusta Persson Foundation, the Erik Philip Sörensen Foundation and the Medical Faculty of the University of Lund.

References

Arnason, U., Gullberg, A., Janke, A. & Xu, X. (1996). Pattern and timing of evolutionary divergences among hominoids based on analyses of complete mtDNAs. *Journal of Molecular Evolution* **43**, 650–661.

Cianciolo, G. J., Copeland, T. D., Oroszlan, S. & Snyderman, R. (1985). Inhibition of lymphocyte proliferation by a synthetic peptide homologous to retroviral envelope proteins. *Science* **230**, 453–455.

Dangel, A. W., Mendoza, A. R., Baker, B. J., Daniel, C. M., Carroll, M. C., Wu, L. C. & Yu, C. Y. (1994). The dichotomous size variation of human complement C4 genes is mediated by a novel family of endogenous retroviruses, which also establishes species-specific genomic patterns among Old World primates. *Immunogenetics* **40**, 425–436.

Dangel, A. W., Baker, B. J., Mendoza, A. R. & Yu, C. Y. (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.

Devereux, J., Haeberli, P. & Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Research* **12**, 387–395.

Felsenstein, J. (1993). PHYLIP (phylogeny inference package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle, WA, USA.

Guntaka, R. V. (1993). Transcription termination and polyadenylation in retroviruses. *Microbiological Reviews* **57**, 511–521.

Heinemeyer, T., Wingender, E., Reuter, I., Hermjakob, H., Kel, A. E., Kel, O. V., Ignatieva, E. V., Ananko, E. A., Podkolodnaya, O. A., Kolpakov, F. A., Podkolodny, N. L. & Kolchanov, N. A. (1998). Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Research* **26**, 362–367.

Hussey, D. J., Parker, N. J., Hussey, N. D., Little, P. F. & Dobrovic, A. (1997). Characterization of a KRAB family zinc finger gene, ZNF195, mapping to chromosome band 11p15.5. *Genomics* **45**, 451–455.

Larsson, E., Kato, N. & Cohen, M. (1989). Human endogenous proviruses. *Current Topics in Microbiology and Immunology* **148**, 115–132.

Lower, R., Boller, K., Hasenmaier, B., Korbmacher, C., Muller-Lantzsch, N., Lower, J. & Kurth, R. (1993). Identification of human endogenous retroviruses with complex mRNA expression and particle formation. *Proceedings of the National Academy of Sciences, USA* **90**, 4480–4484.

Lower, R., Tonjes, R. R., Korbmacher, C., Kurth, R. & Lower, J. (1995). Identification of a Rev-related protein by analysis of spliced transcripts of the human endogenous retroviruses HTDV/HERV-K. *Journal of Virology* **69**, 141–149.

Lower, R., Lower, J. & Kurth, R. (1996). The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. *Proceedings of the National Academy of Sciences, USA* **93**, 5177–5184.

McClure, M. A., Johnson, M. S., Feng, D. F. & Doolittle, R. F. (1988). Sequence comparisons of retroviral proteins: relative rates of change and general phylogeny. *Proceedings of the National Academy of Sciences, USA* **85**, 2469–2473.

Maeda, N. (1985). Nucleotide sequence of the haptoglobin and haptoglobin-related gene pair. The haptoglobin-related gene contains a retrovirus-like element. *Journal of Biological Chemistry* **260**, 6698–6709.

Mager, D. L. & Freeman, J. D. (1987). Human endogenous retroviruslike genome with type C pol sequences and gag sequences related to human T-cell lymphotropic viruses. *Journal of Virology* **61**, 4060–4066.

Mager, D. L. & Henthorn, P. S. (1984). Identification of a retrovirus-like repetitive element in human DNA. *Proceedings of the National Academy of Sciences, USA* **81**, 7510–7514.

Majors, J. (1990). The structure and function of retroviral long terminal repeats. *Current Topics in Microbiology and Immunology* **157**, 49–92.

O'Connell, C., O'Brien, S., Nash, W. G. & Cohen, M. (1984). ERV3, a full-length human endogenous provirus: chromosomal localization and evolutionary relationships. *Virology* **138**, 225–235.

Ono, M., Yasunaga, T., Miyata, T. & Ushikubo, H. (1986). Nucleotide sequence of human endogenous retrovirus genome related to the mouse mammary tumor virus genome. *Journal of Virology* **60**, 589–598.

Patience, C., Wilkinson, D. A. & Weiss, R. A. (1997). Our retroviral heritage. *Trends in Genetics* **13**, 116–120.

Repaske, R., O'Neill, R. R., Steele, P. E. & Martin, M. A. (1983). Characterization and partial nucleotide sequence of endogenous type C retrovirus segments in human chromosomal DNA. *Proceedings of the National Academy of Sciences, USA* **80**, 678–682.

Repaske, R., Steele, P. E., O'Neill, R. R., Rabson, A. B. & Martin, M. A. (1985). Nucleotide sequence of a full-length human endogenous retroviral segment. *Journal of Virology* **54**, 764–772.

Smit, A. F. (1996). The origin of interspersed repeats in the human genome. *Current Opinion in Genetics and Development* **6**, 743–748.

Strambio-de-Castillia, C. & Hunter, E. (1992). Mutational analysis of the major homology region of Mason–Pfizer monkey virus by use of saturation mutagenesis. *Journal of Virology* **66**, 7021–7032.

Swofford, D. (1992). PAUP: phylogenetic analysis using parsimony, version 3.0s. Illinois Natural History Survey, Champaign, IL, USA.

Temin, H. M. (1981). Structure, variation and synthesis of retrovirus long terminal repeat. *Cell* **27**, 1–3.

Widegren, B., Kjellman, C., Aminoff, S., Sahlford, L. G. & Sjögren, H.-O. (1996). The structure and phylogeny of a new family of human endogenous retroviruses. *Journal of General Virology* **77**, 1631–1641.

Zlotnick, A., Stahl, S. J., Wingfield, P. T., Conway, J. F., Cheng, N. & Steven, A. C. (1998). Shared motifs of the capsid proteins of hepadnaviruses and retroviruses suggest a common evolutionary origin. *FEBS Letters* **431**, 301–304.

Received 19 February 1999; Accepted 1 June 1999