Processed Pseudogenes of Human Endogenous Retroviruses Generated by LINEs: Their Integration, Stability, and Distribution

Adam Pavlíček,1,2 Jan Pačes,1,3 Daniel Elleder,1 and Jiří Hejnar1,4

1Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Prague 6, CZ-16637, Czech Republic; 2Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 75005 Paris, France; 3Center for Integrated Genomics, Prague 6, CZ-16637, Czech Republic

We report here the presence of numerous processed pseudogenes derived from the W family of endogenous retroviruses in the human genome. These pseudogenes are structurally colinear with the retroviral mRNA followed by a poly(A) tail. Our analysis of insertion sites of HERV-W processed pseudogenes shows a strong preference for the insertion motif of long interspersed nuclear element (LINE) retrotransposons. The genomic distribution, stability during evolution, and frequent truncations at the 5' end resemble those of the pseudogenes generated by LINEs. We therefore suggest that HERV-W processed pseudogenes arose by multiple and independent LINE-mediated retrotransposition of retroviral mRNA. These data document that the majority of HERV-W copies are actually nontranscribed promoterless pseudogenes. The current search for HERV-Ws associated with several human diseases should concentrate on a small subset of transcriptionally competent elements.

[Online supplementary material available at http://www.genome.org]
lar features including a common TT|AAAA insertion motif and a variable-length (5–15 bp) target site duplication (TSD) (Jurka 1997), whereas retroviruses are characterized by short 4–6-bp direct repeats and no preference for a motif similar to TT|AAAA. Moreover, poly(A) tails and frequent truncation at the 5′ end of processed pseudogenes are typical for LINEs (Voliva et al. 1983; Smit 1999), indicating again that LINEs are the master mobile elements in the human genome. Finally, an in vitro assay clearly showed creation of reporter gene copies by the LINE machinery with all hallmarks of the processed pseudogene (Esnault et al. 2000).

The estimated portion of processed pseudogenes in the human genome is ~0.5% (Dunham et al. 1999), with copy number 23,000–33,000 (Goncalves et al. 2000). The LINE machinery influences not only the pseudogene structure, but also the pseudogene distribution. We reported recently that not only young LINEs and Alus, but also processed pseudogenes, preferentially reside in GC-poor parts of the genome (Pavlícˇek et al. 2001).

During our work on the Human Endogenous Retrovirus database (http://herv.img.cas.cz; Pacˇes et al. 2002), we found unusual retroviral structures lacking the common proviral organization. They are structurally similar to retroviral mRNA, followed by a poly(A) tail. These elements are frequently observed in the HERV-W family (Blond et al. 1999), described previously as LM7 or multiple sclerosis-associated virus (Perron et al. 1989, 1997), and as endogenous HERV17 family in the Repbase Update database (http://www.girinst.org/; Jurka 1998, 2000; Smit 1999). The HERV-W family is of special interest because of suggestions of a role in several human diseases including multiple sclerosis (Perron et al. 1997; Kozmian-Pradel et al. 1999), rheumatoid arthritis (Gaudin et al. 1997, 2000), and schizophrenia (Karlsson et al. 2001). Moreover, the env gene of a prototype HERV-W element on chromosome 7 was suggested as the human gene coding for the syncytin protein responsible for cell fusion during the differentiation of the syncytiotrophoblast in human placenta (Smit 1999; Blond et al. 2000; Ml et al. 2000). This is one of the best-documented examples of recruitment of retroviral genes by the host genome.

This work describes, for the first time, the presence of processed pseudogenes from mRNA of HERV-Ws in the human genome. These pseudogenes are remarkably numerous in the HERV-W family and are not found at comparable frequencies in related retroviral families of the class 1 HERVs. This finding represents not only an interesting connection between two main reverse transcriptase-encoding classes of retroelements, LINEs and HERVs, but, as we show, also a unique model for comparing LINE and retroviral transposition, mRNA preference, integration specificity, recombination, and genomic stability.

**RESULTS**

**Determination and Genomic Structure of HERV-W Retroviral and Pseudogene Copies**

Using the genome-wide RepeatMasker scan, 654 HERV-W family members were found. From their multiple alignment schematically depicted in Figure 1 (see also Supplementary Figure 1S available online at http://www.genome.org), it is clear that HERV-W sequences are divided into two major groups. The first group is defined by the complete or partial U3 region in 5′ LTR and/or U5 in 3′ LTR, and will be referred to as the retro group (or retro copies) in the following text, because the mentioned parts of LTRs are completed during retroviral reverse transcription. This retroviral group contains 77 proviral copies with complete or partial internal (non-LTR)
sequences, and 343 soloLTRs without any internal sequences (Table 1).

The members of the second group are defined by incomplete LTRs that start close to the beginning of the R region of 5’ LTR (from nucleotide 256 of the consensus sequence, see Methods) and end at the 3’ part of the 3’ LTR R region (up to nucleotide 9732 of the consensus). The second characteristic feature is the frequent presence of the poly(A) tail of variable length and, therefore, we term these HERV-W sequences as the poly(A) group [or poly(A) copies] in the following text. The length of the tails vary from a few nucleotides to >50 bp. The poly(A) tail starts at position 9732, just 12 nucleotides downstream of the poly(A) signal ATTAAA, in the 3’ LTR (positions 9714–9719). In the poly(A) tails, we often detected A-rich microsatellites of up to 10 repeat units. The poly(A) tails end, in general, with the 3’ direct repeat (Fig. 2). We have found 176 members of the poly(A) group (Table 1).

Aside from these two groups, there are 58 truncated elements, which, due to the absence of diagnostic parts, cannot be unambiguously defined and are referred to here as the undef group (Table 1).

Analysis of terminal deletions shows further intergroup differences. Whereas the poly(A) group is, in the great majority of cases, truncated at the 5’ end, retro copies and soloLTRs are nearly equally truncated at both ends (Table 1). In addition, soloLTRs are mostly complete, but the poly(A) and the retro copies of HERV-Ws are generally shortened (Table 1). Interestingly, in addition to the poly(A) tract, HERV-W genomic copies display further oligonucleotide expansion located in the leader region (positions 1484–1603 at the consen- sus sequence). Sometimes, arrays of several hundreds of basepairs are found, composed from mainly AG-rich di- or oligonucleotides.

Characterization of Insertion Sites of HERV-W Poly(A) and Retro Copies

To provide better insight into the transposition processes of HERV-W copies, we have analyzed their insertion sites. From the retro copies, we first extracted a subset of 172 complete elements and complete soloLTRs with untruncated 5’ and 3’ ends. For 82 of them, we found 4-bp direct repeats in 5’ and 3’-flanking sequences, just adjacent to the border of provi- ruses. On the other hand, in 116 of 176 elements in the poly(A) group, we detected long direct repeats of variable length up to 21 bp, with a mean of 9.51 bp.

Nucleotide frequencies in insertion sites of both groups are shown in Table 2. Three nucleotides upstream and five downstream of the start of each direct repeat were analyzed. For retro copies, there is a weak bias in the nucleotide frequencies toward a TT|ATA(A/T) sequence, in which | denotes the start of the direct repeat (Table 2a). The fifth downstream nucleotide is frequently T, already corresponding to the first nucleotide of HERV-W LTR, which is consensually T, because the general length of HERV-W insertion TSD is 4 bp.

In comparison with the retro group, poly(A) copies have a strong bias in the insertion site motif, namely toward the (A/T)(A/T)T|AAAAA octanucleotide (Table 2b).

Table 1. Number of Different HERV-W Elements in the Human Genome

| Group          | Number | Complete | 5’ truncation* | 3’ truncation* | 5’ 3’ truncation* |
|----------------|--------|----------|----------------|----------------|-------------------|
| poly(A)        |        | 176      | 46             | 107            | 23                |
| retro (internal sequence–containing) |        | 77       | 29             | 15             | 24                |
| retro (soloLTR) |        | 343      | 306            | 20             | 13                |
| undef          |        | 58       | 0              | 0              | 0                 |

*Truncations were defined according to alignments with a particular consensus sequence for each group. The terminal gap was counted if the start or the end of the element was more distant than 20 bp from the consensus termini.

Figure 2 The flanking sequences of poly(A) HERV-W copies. For 23 poly(A) elements shown in Fig. 1, the short flanking sequences are given in detail. Retroviral nucleotides are in upper case, flanking regions in lower case characters. Direct repeats are in italics and underlined. Poly(A) signals are highlighted by bold characters. Poly(A) tails are shown as gray boxes, satellite expansions are inside white or light gray-shaded boxes.

HERV-W Processed Pseudogenes Generated by LINEs
Phylogenetic Analysis

Two parts of the retroviral genome were selected for the phylogenetic analysis: the gag region (consensus sequence nucleotides 2718–3923) and a 5’ part of the pol region (consensus sequence nucleotides 4450–6789; see Methods). Both phylogenetic trees (Fig. 3A and Supplementary Figure 2Sa) show a complex pattern of HERV-W copies, in which poly(A) elements are dispersed within the retro copies. Due to a high similarity at non-CpG sites, there are few polymorphic sites, and thus the reliability of the obtained topology is weak (see low bootstrap values in Fig. 3A, Supplementary Figure 2Sa). We have selected only a small subset of elements, and the obtained tree (Fig. 3B) shows a better supported branching pattern, in which, again, poly(A) elements are dispersed within retroviral copies.

Genomic Distribution

The genomic distribution of HERV-W elements was calculated for 100-kb nonoverlapping segments of the genome (see Methods). Figure 4A shows the HERV-W frequency in genomic segments classified according to four ranges of the GC content. Both retro and poly(A) copies show distribution biased toward the GC-poor part of the genome. Especially, solo.LTRs have a strong tendency to localize in GC-poor regions.

The intrachromosomal distribution of HERV-W (Fig. 4B) shows several interesting features. First, the intrachromosomal distribution varies among all poly(A), retro, and solo.LTR groups. Poly(A) elements are over-represented on chromosomes 3, 6, 19, and X and less abundant on chromosomes 16, 17, 21, 22, and Y. Retro copies are more frequent on chromosomes 4, 5, 7, 13, X, and especially chromosome Y, where both internal sequence-containing copies and solo.LTRs are clearly over-represented. On the other hand, retroviral copies are under-represented on smaller chromosomes 16, 17, 19, 20, and 22. Generally, HERV-W elements show biased distribution toward chromosomes 3, 4, X, and particularly Y, and are less frequent on chromosomes 16, 17, 20, and 22, in agreement with experimental results (Voisset et al. 2000).

Comparison of the Length and the GC Level of HERV-W Elements

We selected 46 complete poly(A) copies and 29 complete retro copies for comparison of the length and the GC level. From all copies, only positions 256–9732 of the consensus, corresponding to the full-length poly(A) sequence without the poly(A) tail, were compared. To eliminate the influence of various insertions and satellite expansions, only sequences homologous to the consensus were used; sequences were aligned to the consensus and nonhomologous parts were neglected as in Figure 1. No striking differences were found in the GC content, poly(A) sequences being slightly GC poorer than the retro copies (45.8% compared with 47.7%). The length was calculated as a percentage of gap positions in all available positions of the consensus (100 × gaps length/total length). These numbers were used to compare splicing variation. Because the majority of internal deletions correspond to several splice variants (visible from the Supplementary Figure 1S, available on line at http://www.genome.org), these numbers estimate the percentage of sequences missing due to the splicing. For the retro group, we obtained, on average, 41.4% of gaps in comparison with the consensus, for the poly(A) group, 37.6% of gaps.

DISCUSSION

Processed Pseudogenes of HERV-Ws Display the Features of LINE-Mediated Transposition

Our analysis of HERV-W copies in the human genome revealed unusual endogenous retroviral structures. We found elements colinear with retroviral mRNAs followed by poly(A) tails, the poly(A) copies, resembling processed pseudogenes of cellular genes (Vanin 1985; Weiner et al. 1986). On the basis of the lack of complete LTRs, which regenerate during normal retroviral reverse transcription, we suggest that these poly(A) copies arose by virtue of heterologous reverse transcriptase, most probably through the LINE-mediated reverse transcription.

The majority of poly(A) copies is surrounded by direct repeats of variable length (6–21 bp, mean 9.51 bp), characteristic for insertions of LINEs, Alus, and processed pseudogenes, which are believed to have arisen through double enzymatic nicking of host DNA by the LINE endonuclease (Feng et al. 1996; Jurka 1997; Cost and Boeke 1998). In contrast, standard members of the HERV-W family, the retro copies, have canonically short 4-bp direct repeats as already described (Repbase Update; Blond et al. 1999). Analysis of insertion sites of poly(A) (Table 2) elements showed a strong preference for (A/T)(A/T)T|AAAAA sequences, strongly resembling the preferential TT|AAAA cutting motif of the LINE endonuclease found in LINEs, Alus, and processed pseudogenes (Feng et al.

| (a) HERV-W retro copies | (b) HERV-W pseudogene poly(A) copies |
|------------------------|-----------------------------------|
|                         | All 82 82 82 82 82 82 82 82       |
|                         | A 40 10 23 34 14 41 6 2          |
|                         | C 16 27 31 20 13 1 17 9          |
|                         | G 15 6 26 20 8 20 30 6           |
|                         | T 11 39 2 8 47 20 29 65          |
|                         | — Others                         |
|                         | 1 1                             |

The highest nucleotide frequencies defining the consensus insertion site motif are in italics.
Poly(A) tails of poly(A) elements often contain microsatellites. Similar expansions were described for Alus, in which oligo-dA-rich tails have been shown to serve as nuclei for the genesis of simple repeats (Arcot et al. 1995).

Frequent 5' truncations in the poly(A) group (Fig. 1; Table 1) also described in LINEs and processed pseudogenes (Voliva et al. 1983; Vanin 1985; Weiner et al. 1986; Smit 1999) the presence of spliced variants, the mRNA-derived structure, the poly(A) tail, and the insertion-site characteristics mentioned above, which altogether indicate that the LINE enzymatic machinery generated the pseudogene copies of the HERV-W family [poly(A) group]. The recent experimental demonstration of processed pseudogenes generated by LINE-encoded enzymes (Esnault et al. 2000; Wei et al. 2001) strengthens our conclusion that the poly(A) group is generated by nonretroviral, LINE-mediated mobilization. The complex pattern on phylogenetic trees, in which poly(A) copies are dispersed within retroviral copies (Fig. 3A,B), suggests multiple and independent origins of poly(A) copies.

**Figure 3** Phylogenetic trees of HERV-W elements. We selected multiple alignment of the proximal part of the pol region (nucleotides 4450–6789). All gap- and CpG-containing sites were excluded (see Methods). Poly(A) elements are in italics, retroviral copies are in bold to highlight dispersion of poly(A) elements within the retro group. (A) Neighbor-joining tree with 1000 bootstrap replicates. Forty-six elements at 581 non-gap, non-CpG sites. Rectangles show the elements selected for further phylogenetic analysis in Fig. 4B. (B) A subset of 6 elements at 1290 non-CpG, non-gap sites. Topology is based on a maximum likelihood tree, bootstrap values (1000 replicates) ordered from top to bottom show support of the topology for maximum likelihood, maximum parsimony, and neighbor-joining methods implemented in the phylo_win program (see Methods).

**Figure 4** Genomic distribution of HERV-W elements. (A) Isochore distribution of poly(A), soloLTR, and other retroviral (internal sequence-containing) copies was calculated in 100-kb long, nonoverlapping segments. The distribution is shown in the number of independent elements per 10 Mb. Four categories of GC levels corresponding to four isochore families were used (see Methods). (B) The chromosome distribution is shown as a number of independent elements per 10 Mb (left y axis). The bars represent chromosomal densities of poly(A), soloLTR, as well as internal sequence-containing retro copies. The GC levels of chromosomes are shown as diamonds (right y axis).

**Distribution and Stability of HERV-W Copies Depends on the Mechanism of Transposition**

The genome-wide distribution of HERV-W poly(A) copies shows a bias toward the GC-poorer part of the genome than that of retro copies (Fig. 4A). This bias is typical for LINEs, 1996; Jurka 1997; Cost and Boeke 1998; Wei et al. 2001). Poly(A) tails of poly(A) elements often contain microsatellites. Similar expansions were described for Alus, in which oligo-dA-rich tails have been shown to serve as nuclei for the genesis of simple repeats (Arcot et al. 1995).

Frequent 5' truncations in the poly(A) group (Fig. 1; Table 1) also described in LINEs and processed pseudogenes (Voliva et al. 1983; Vanin 1985; Weiner et al. 1986; Smit 1999) the presence of spliced variants, the mRNA-derived structure, the poly(A) tail, and the insertion-site characteristics mentioned above, which altogether indicate that the LINE enzymatic machinery generated the pseudogene copies of the HERV-W family [poly(A) group]. The recent experimental demonstration of processed pseudogenes generated by LINE-encoded enzymes (Esnault et al. 2000; Wei et al. 2001) strengthens our conclusion that the poly(A) group is generated by nonretroviral, LINE-mediated mobilization. The complex pattern on phylogenetic trees, in which poly(A) copies are dispersed within retroviral copies (Fig. 3A,B), suggests multiple and independent origins of poly(A) copies.
young Alus, and processed pseudogenes, and is probably linked to the AT-rich insertion motif of these elements (International Human Genome Sequencing Consortium 2001; Pavlíček et al. 2001). Similarly, for the retro group, we also found a weak bias toward AT-rich insertion sites, and the genomic distribution is also biased toward the GC-poor isochores.

In contrast to LINEs, there is no clear general insertion motif for retroviruses. Both retrovirus and LINE integration are influenced by the DNA bending (Pryciak and Varmus 1992; Muller and Varmus 1994; Jurka et al. 1998), the nucleosome structure at the site of integration (Pryciak and Varmus 1992; Cost et al. 2001), and the state of chromatin condensation (Leib-Mösch et al. 1993, Ryndich et al. 1998). The only sequence-specific insertion motif reported so far is the palindromic pentanucleotide GT(A/T)AC found in nonautonomous LTR retrotransposons MalR (mammalian apparent LTR retrotransposon) (Smit 1996b) and HIV-1 integration (Stevens and Griffith 1996; Carteau et al. 1998). From the data in Table 2, we derived a weak preferential motif ATC|ATA(G/T) and a strong negative motif notTnotGnotCnotA. The described preferential motif is similar, but not the same as the CA(A/T)TG pentanucleotide, the complementary version of the MalR and HIV-1 insertion motif. The studied MalR and HIV-1 insertion sequences were different in the length of the TSD – 5 bp instead of 4 bp for HERV-W integration, suggesting that on the contrary to the high evolutionary conservation of the integrase region in retroviruses, insertions could differ in both TSD length and in the insertion motif. Despite the fact that the insertion motif obviously is not the only factor determining retrotransposon integration, the AT-rich LINE insertion motif seems to be responsible for the GC-poor bias in the processed pseudogene distribution.

Figure 5 Schematic representation of amplification of poly(A) and retro copies, differences, and stability in the genome.
as intrachromatid single-strand annealing, reciprocal exchange, replication slippage, or abortive reciprocal recombination (for review, see Klein 1995; Lambert et al. 1999).

In conclusion, we could see that the retroviral and the LINE transposition differ in the integration, distribution and the stability of inserted copies (Fig. 5). The retroviral reverse transcription produces potentially expression- and replication-competent copies, which are unstable due to the homologous LTR–LTR recombination, and are frequently excluded during evolution. On the other hand, LINEs often produce 5’ truncated copies, unable to transcribe and further mobilize due to the loss of promoter sequences in the U3 region of the 5’ LTR. These copies have, however, a better chance to keep their original structure during evolution. Retro copies represent 64.2% (420/654) of all HERV-W elements, whereas poly(A) copies are just 26.9% (176/654). On the other hand, poly(A) copies account for 52.5% (149/284) of elements containing internal sequences, whereas the proportion of retro copies is only 27.1% (77/284). This implies that the majority of HERV-W gag, pol, and env-coding regions experimentally detected on human chromosomes (Voisset et al. 2000) are actually pseudogenes, probably nontranscribed promoterless copies, and the current search for possible causative agents of several human diseases (Perron et al. 1989, 1997; Gaudin et al. 1997, 2000; Komurian-Pradel et al. 1999; Kim and Crow 1999) could concentrate on a small subset of transcriptionally competent HERV-W elements.

Remarks on the Processed Pseudogene Formation

The LINE machinery is highly effective in the transposition of its own copies (Dombroski et al. 1991; Moran et al. 1996). The estimated frequency of cellular mRNA mobilization is only 0.01%–0.05% per LINE transposition (Wei et al. 2001). We found numerous processed pseudogenes within the HERV-W family, but not in the closely related HERV9 family. In the HERV database (Pačes et al. 2002), we have made a preliminary analysis on the basis of the presence of poly(A) tails and TSDs and we estimate that the HERV9 family has more than seven times less pseudogenes than the HERV-W family. In the class 1 HERVs, we have not found any retroviral family with >5% of processed pseudogenes (data not shown). In a direct in vitro test, the frequency of Moloney murine leukemia virus-based vector transposition was just 10~7–10~6 per cell per generation (Tchénio et al. 1993). Similarly, genomic library screening for the HERV-H family indicated very few (<1%) copies with the structure of processed pseudogenes (Good-child et al. 1995). Hence, LINE-mediated transposition of the retroviral mRNA seems to be rather rare, with the exception of HERV-W mRNA.

What drives the selectivity of capturing the non-LINE mRNA by LINE-encoded enzymes? Only transpositions in the germ line, primordial germ line, or early embryonal cells have a chance to proceed into the next generation and eventually to be fixed. Both LINEs and Alus are known to be expressed in testicular tissues and they could be potentially transposed into germ-line cells (Branciforte and Martin 1994; Schmid 1998). It is noteworthy that expression of transcripts homologous to the syncythin gene have been detected not only in placenta, but also in testis (Mi et al. 2000); it is therefore conceivable that coexpression of LINES and HERV-W elements facilitates the generation of processed pseudogenes in the HERV-W family. On the other hand, transcription up-regulation in testis is not limited to LINES, Alus, and HERV-Ws and could serve as an engine for generation of processed pseudogenes from various mRNAs (Schmidt 1996).

A genome-wide analysis of processed pseudogenes suggested that the processed pseudogenes are preferentially generated from short, GC-poor RNAs (Goncalves et al. 2000). The complex splice pattern of the HERV-W family provides necessary GC and length variation and a good opportunity to test these predictions and compare the affinity of LINE and retroviral enzymes for various mRNAs. In our analysis, we have not found any difference between the retroviral and LINE replication. Also, in the analysis of the human genome sequence, no bias has been found in the GC content of genes generating processed pseudogenes (Venter et al. 2001).

METHODS

Identification and Localization of HERV-W Sequences in the Human Genome

The RepeatMasker program (Smit and Green RepeatMasker [http://repeatmasker.org] at the Genome Sequencing Center was used to identify HERVs in the GoldenPath assembly of 87% of the human genome (Haussler et al. Human Genome Working Draft at http://genomewww.genome.org). The fragments of LTRs, as well as internal retroviral sequences of the HERV-W (HERV17) family and the related HERV9 family, were included in the following analysis. All selected retroviral fragments were compared against family profiles of HERV-W and HERV9 families using the HMMER package (HMMER v. 2.1.1; Eddy 1998; http://hmmer.wustl.edu/) to eliminate errors due to misidentifications of closely related families. Only elements with better matches to the HERV-W profile than to that of the HERV9 profile were used further as real HERV-W elements. Precise positions of starts and ends of elements were obtained by recursive alignment with the Repbase Update consensus of the HERV family using the Dialign2 program (Dialign v. 2.0, Morgenstern et al. 1996).

Analysis of the Genomic Structure

All elements were aligned using the Dialign2 program and the multiple alignment was manually checked for errors in the Seaview editor (Galtier et al. 1996). Computing of a large alignment was approximated by aligning each element with a consensus sequence obtained from the Repbase Update, followed by assembly of pairwise alignments into multiple alignments. As a consensus, we used sequences of LTR17 and HERV17 consensus, joined in the order LTR17, HERV17, LTR17 into one sequence. In 3’ LTR, we have introduced a 50-bp insertion of a poly(A) stretch in the place where the poly(A) tail of retroviral RNAs is located (from position 327 of the LTR17 consensus).

Characterization of Insertion Sites of HERV-W Copies

We have identified insertion sites of all copies, including the TSD and the insertion motif. First, the sequences were scanned for an exact match of 6 bp or longer at both ends of HERV-W elements in a region from 30 bp outside to 5 bp inside. Complete retroviral copies as well as complete soloLTRs were scanned for direct repeats of 4 bp or longer at both ends of the element.

Phylogenetic Analysis

For the phylogenetic analysis, we have selected two regions in the alignment, the gag region from 2718 to 3923 bp and a proximal part of the pol region, until the major env splicing
variant (nucleotides 4450–6789). Then, we excluded all incomplete and many gap-containing elements. We obtained 36 elements in the gag region and 46 sequences in the pol alignment. All CpG and gap-containing positions were excluded. The topology was obtained by neighbor joining (Tajima-Nei distance), maximum parsimony, and maximum likelihood methods implemented in the phylo_win program (Galtier et al. 1996).

Genomic Distribution of HERV-W Copies
Retrotransposon densities were calculated in 100-kb long, nonoverlapping fragments from the GoldenPath assembly (http://genome.ucsc.edu). The densities were calculated for the isochore families as given by Zoubak et al. (1996). Isochore family intervals were as follows: <37%, 37%–41%, 41%–46%, and >46% content of GC for L1, L2, H1, and H2+H3 families, respectively.

ACKNOWLEDGMENTS
We thank Oliver Clay, Jan Svoboda, and Giorgio Bernardi for critical reading of the manuscript. This work was supported by grant No. 204/01/0632 of the Grant Agency of the Czech Republic to J.H. A.P. is supported in part by a PhD fellowship of the French Government program Doctorat en cotutelle.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES
Arcot, S.S., Wang, Z., Weber, J.L., Deininger, P.L., and Batzer, M.A. 1995. Alu repeats: A source for the genesis of primate microsatellites. Genomics 29: 136–144.
Blond, J.L., Beseme, F., Duret, L., Bouton, O., Bedin, F., Perron, H., Mandrand, B., and Mallet, F. 1999. Molecular characterization and placental expression of HERV-W, a new human endogenous retrovirus family. J. Virol. 73: 1175–1185.
Blond, J.L., Lavillette, D., Cheynet, V., Bouton, O., Oriol, G., Chapel-Fernandes, S., Mandrand, B., Mallet, F., and Cosset, F.L. 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: 3231–3239.
Branciforte, D. and Martin, S.L. 1994. Developmental and cell type specificity of LINE-1 expression in mouse testsis: Implications for transposition. Mol. Cell. Biol. 14: 2584-2592.
Brosius, J. 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 238: 115–134.
Carteux, S., Hoffmann, C., and Bushman, F. 1998. Chromosome structure and human immunodeficiency virus type 1 cDNA integration: Centromeric aliphid repeats are a disfavored target. J. Virol. 72: 4005–4014.
Cost, G.J. and Boeke, J.D. 1998. Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. Biochemistry 37: 18081–18093.
Cost, G.J., Golding, A., Schlissel, M.S., and Boeke, J.D. 2001. Target DNA chromatization modulates nicking by L1 endonuclease. Nucleic Acids Res. 29: 573–577.
Dey, I.K., Strathern, J.N., and Garfinkel, D.J. 1991. RNA-mediated recombination in S. cerevisiae. Cell 67: 355–364.
Dhillon, O., Maestre, J., and Heidmann, T. 1999. Detection of retrovirus RNA in plasma from rheumatiod arthritis. Arthritis Rheum. 40: S245.
Dhillon, P., Ijaz, S., Tuke, P.W., Marcel, F., Paraz, A., Seigneurin, J.M., Mandrand, B., Perron, H., and Garson, J.A. 2000. Infrequency of detection of particle-associated MSRV/HERV-W RNA in the synovial fluid of patients with rheumatoid arthritis. Rheumatology 39: 950–954.
Goncalves, I., Duret, L., and Mouchiroud, D. 2000. Nature and structure of human genes that generate retropseudogenes. Genome Res. 10: 672–678.
Goodchild, N.L., Freeman, J.D., and Mager, D.L. 1995. Spliced HERV-H endogenous retroviral sequences in human genomic DNA: Evidence for amplification via retrotransposition. Virology 206: 164–173.
Goodier, J.L., Ostertag, E.M., and Kazazian, Jr., H.H. 2000. Transdution of 3’-flanking sequences is common in L1 retrotransposition. Hum. Mol. Genet. 9: 653–357.
International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature 409: 860–921.
Jurka, J. 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc. Natl. Acad. Sci. 94: 1872–1877.
—. 1998. Repeats in genomic DNA: Mining and meaning. Curr. Opin. Struct. Biol. 8: 333–337.
—. 2000. Repbase update: A database and an electronic journal of repetitive elements. Trends Genet. 16: 418–420.
Jurka, J., Klonowski, P., and Trifonov, E.N. 1998. Mammalian retroposons integrate at kinkable DNA sites. Bionmol. Struct. Dyn. 15: 717–721.
Karlsson, H., Bachmann, S., Schroder, J., McCarthy, J., Torrey, E.F., and Yolken, R.H. 2001. Retroviral RNA identified in the cerebrospinal fluids and brains of individuals with schizophrenia. Proc. Natl. Acad. Sci. 98: 4634–4639.
Kazazian, Jr., H.H. 1999. An estimated frequency of endogenous insertional mutations in humans. Nature Genet. 22: 130.
Kazazian, Jr., H.H. and Moran, J.V. 1998. The impact of L1 retrotranspositions on the human genome. Nature Genet. 19: 19–24.
Kim, H. and Crow, T.J. 1999. Identification and phylogeny of novel human endogenous retroviral sequences belonging to the HERV-W family on the human X chromosome. Arch. Virol. 144: 2403–2413.
Kjellman, C., Sjogren, H.O., and Widgren, B. 1995. The Y chromosome: A graveyard for endogenous retroviruses. Gene 161: 163–170.
Klein, H.L. 1995. Genetic control of intrachromosomal recombination. BioEssays 17: 147–159.
Komurian-Pradel, F., Parinhas-Baccala, G., Bedin, F., Ounanian-Paraz, A., Sodoyer, M., Ott, C., Rajahorison, A., Garcia, E., Mallet, F., Mandrand, B., et al. 1999. Molecular cloning and characterization of MSRV-related sequences associated with retrovirus-like particles. Virology 260: 1–9.
Lambert, S., Saintigny, Y., Delacote, F., Amiot, F., Chaput, B., Leconte, M., Huck, S., Bertrand, P., and Lopez, B.S. 1999. Analysis of intrachromosomal homologous recombination in mammalian cell, using tandem repeat sequences. Mutat. Res. 433: 159–168.
Leib-Mosch, C., Halmeyer, M., Werner, T., Geidl, E.M., Brack-Werner, R., Franke, U., Erlele, V., and Hehlmann, R. 1993. Genomic distribution and transcription of solitary HERV-K LTRs. Genomics 18: 261–269.
Levine, K.L., Steiner, B., Johnson, K., Aromoff, R., Quinton, T.J., and Linial, M.L. 1990. Unusual features of integrated cDNAs.
Linial, M.L. 1990. Unusual features of integrated cDNAs. Virology 189: 409–418.
generated by infection with genome-free retroviruses. Mol. Cell. Biol. 10: 1891–1900.
Liskay, R.M., Letsou, A., and Stachelek, J.L. 1987. Homology requirement for efficient gene conversion between duplicated chromosomal sequences in mammalian cells. Genetics 115: 161–167.
Maestre, J., Tchénio, T., Dhillon, O., and Heidmann, T. 1995. mRNA reconstitution in human cells: Processed pseudogene generation. EMBO J. 14: 6333–6338.
Mager, D.L. and Goodchild, N.L. 1989. Homologous recombination between the LTRs of a human retrovirus-like element causes a 5-kb deletion in two siblings. Ann. J. Hum. Genet. 45: 848–854.
Mi, S., Lee, X., Li, X., Veldman, G.M., Finnerty, H., Racie, L., LaVaille, E., Yang, X.Y., Edouard, P., Hughes, S., et al. 2000. Synctin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403: 785–789.
Moran, J.V., Holmes, S.E., Naas, T.P., Dellerandinis, R.J., Boeke, J.D., and Kazazian, Jr., H.H. 1996. High frequency retrotransposition in cultured mammalian cells. Cell 87: 917–927.
Morgenstern, B., Werner, T., and Dress, A.W.M. 1996. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. Proc. Natl. Acad. Sci. 93: 12098–12103.
Muller, H.P. and Varmus, H.E. 1994. DNA bending creates favors sites for retroviral integration: An explanation for preferred insertion sites in nucelosomes. EMBO J. 13: 4704–4714.
Pačes, J., Pavlíchek, A., and Pačes, V. 2002. HERVd: Database of human endogenous retroviruses. Nucleic Acids Res. 30: 205–206.
Pavlíchek, A., Jabbari, K., Pačes, J., Pačes, V., Hejnár, J., and Bernardi, G. 2001. Similar integration but different stability of Alus and LINEs in the human genome. Gene 276: 39–45.
Perron, H., Geny, C., Laurent, A., Mouriquand, C., Pellat, J., Perret, J., and Seigneurin, J.M. 1999. Leptomeningeal cell line from multiple sclerosis with reverse transcriptase activity and viral particles. Res. Virol. 140: 551–561.
Perron, H., Garson, J.A., Bedin, F., Beseme, F., Paramhos-Baccala, G., Komurian-Pradel, F., Mallet, F., Tuke, P.W., Voisset, C., Blond, J.L., et al. 1997. Molecular identification of a novel retrovirus repeatedly isolated from patients with multiple sclerosis. Proc. Natl. Acad. Sci. 94: 7583–7588.
Pickoral, O.K., Makulowski, W., Boguski, M.S., and Boeke, J.D. 2000. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. Genome Res. 10: 411–415.
Prak, E.T. and Kazazian, Jr., H.H. 2000. Mobile elements and the human genome. Nature Rev. Genet. 1: 134–144.
Przybylski, P.M. and Varmus, H.E. 1992. Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. Cell 69: 769–780.
Rubin, J. and Subramani, S. 1984. The minimum amount of homology required for homologous recombination in mammalian cells. Mol. Cell. Biol. 4: 2253–2258.
Ryndich, A.V., Zoubak, S., Tsyba, L., Tryapitsina-Guley, N., and Bernardi, G. 1998. The regional integration of retroviral sequences into the mosaic genomes of mammals. Gene 222: 1–16.
Sassaman, D.M., Dombroski, R.A., Moran, J.V., Kimberland, M.L., Naas, T.P., Dellerandinis, R.J., Gabriel, A., Svergúd, G.D., and Kazazian, Jr., H.H. 1997. Many human L1 elements are capable of retrotransposition. Nature Genet. 16: 37–43.
Schmid, C.W. 1998. Does SINE evolution preclude Alu function? Nucleic Acids Res. 26: 4541–4550.
Schmidt, E.E. 1996. Transcriptional promiscuity in testes. Curr. Biol. 6: 768–769.
Smit, A.F. 1996a. The origin of interspersed repeats in the human genome. Curr. Opin. Genet. Dev. 6: 743–748.
———. 1996b. “Structure and evolution of mammalian interspersed repeats.” PhD thesis, University of Southern California, Los Angeles, CA.
———. 1999. Interspersed repeats and other elements of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. 9: 657–663.
Stevens, S.W. and Griffith, J.D. 1996. Sequence analysis of the human DNA flanking sites of human immunodeficiency virus type 1 integration. J. Virol. 70: 6459–6462.
Tchénio, T., Segal-Bendirdjian, E., and Heidmann T. 1993. Generation of processed pseudogenes in murine cells. EMBO J. 12: 1487–1497.
Vanim, E.F. 1985. Processed pseudogenes: Characteristics and evolution. Annu. Rev. Genet. 19: 253–272.
Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. Science 291: 1304–1351.
Voisset, C., Bouton, O., Bedin, F., Duret, L., Mandrand, B., Mallet, F., and Paramhos-Baccala, G. 2000. Chromosomal distribution and coding capacity of the human endogenous retrovirus HERV-W family. AIDS Res. Hum. Retroviruses 16: 731–740.
Voliva, C.F., Jahn, C.L., Comer, M.B., Hutchison, C.A., and Edgell, M.H. 1983. The L1Md long interspersed repeat family in the mouse: Almost all examples are truncated at one end. Nucleic Acids Res. 11: 8847–8850.
Waldman, A.S. and Liskay, R.M.M. 1988. Dependence of intrachromosomal recombination in mammalian cells on uninterrupted homology. Mol. Cell. Biol. 8: 5350–5357.
Wei, W., Gilbert, N., Ooi, S.L., Lawler, J.F., Ostertag, E.M., Kazazian, Jr., H.H., Boeke, J.D., and Moran, J.V. 2001. Human L1 retrotransposition: cis preference versus trans complementation. Mol. Cell. Biol. 21: 1429–1439.
Weiner, A.M., Deininger, P.L., and Efstradiatis, A. 1986. Nonviral retroposons: Genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu. Rev. Biochem. 55: 631–661.
Zoubak, S., Clay, O., and Bernardi, G. 1996. The gene distribution of the human genome. Gene 174: 95–102.

WEB SITE REFERENCES
http://genome.ucsc.edu/; Haussler et al. Human Genome Working Draft.
http://herv.img.cas.cz; the authors’ work on the Human Endogenous Retrovirus database.
http://hmmer.wustl.edu/; HMMER v. 2.1.1.
http://repeatmasker.genome.washington.edu; Smit and Green RepeatMasker.
http://www.girinst.org/; Retrobase Update database.
http://www.med.upenn.edu/genetics/labs/retrotrans_table.html; The database of Retrotransposon Insertion into the Human Genome.

Received September 28, 2001; accepted in revised form December 20, 2001.