Characterization of the Intragenomic Spread of the Human Endogenous Retrovirus Family HERV-W

Javier Costas

Departamento de Bioloxía Fundamental, Universidade de Santiago de Compostela, Spain

This study examines the intragenomic spread of the human endogenous retrovirus family HERV-W from insertions present within the draft sequence of the human genome. Identification of shared diagnostic differences and phylogenetic analyses revealed the existence of three main subfamilies. The average divergence between sequences for each of the subfamilies suggests that most of the HERV-W elements were inserted within the genome during a short period of evolutionary time. Each one of the subfamilies consists of two types of insertions, the expected proviral sequences and other sequences resembling the structure of processed retrogenes. These HERV-W retrosequences extend from the R region of the 5′ long-terminal repeat (LTR) to the R region of the 3′ LTR (as viral genomic RNAs), end in poly(A) 3′ tails, and are flanked by direct repeats longer than the proviral integrations. Furthermore, several of the HERV-W retrosequences are 5′-truncated at different sites. I suggest the involvement of the L1 machinery in these integrations and discuss the characteristic features of the evolutionary history of HERV-W, with emphasis on the putative impact of HERV-W retrosequence integrations on the mammalian genome.

Introduction

Approximately 8% of the human genome is derived from retrovirus-like elements (International Genome Sequencing Consortium 2001). Most of them are human endogenous retroviruses (HERVs) originated by germline infection of the exogenous counterparts in any remote past of primate evolution. Presumably, subsequent retrotransposition (although reinfection cannot be formally ruled out) led to an increase in the copy number of each HERV family (Löwer, Löwer, and Kurth 1996). The structure of an integrated retrovirus (provirus) consists of two long-terminal repeats (LTRs) flanking a central coding region. The LTR sequence can be divided into three regions from the 5′ to the 3′ end, called U3, R, and U5 (fig. 1*A*). Transcription of genomic RNA begins at the R region of the 5′ LTR and ends at the R region of the 3′ LTR. Thus, the U5 region is unique to the 5′ end of the genomic RNA, and the U3 region is unique to the 3′ end (fig. 1*B*). The structure of the LTRs is restored during reverse transcription of genomic RNA. The U3 region contains the promoter and a series of regulatory sequences, which may influence the expression of the neighboring cellular genes (reviewed in Brosius 1999).

There are at least 22 independently acquired HERV families within the human genome (Tristem 2000). Very little is known about the evolutionary history of the different families, with a few exceptions, such as HERV-K (Medstrand and Mager 1998; Lebedev et al. 2000), ERV9 (Costas and Naveira 2000), ERV-L (Benit et al. 1999), or HERV-H (Goodchild, Wilkinson, and Mager 1993; Anderssen et al. 1997). HERV-W has been one of the most extensively studied HERVs during the last few years, since the isolation of an HERV-W–related retrovirus (named multiple-sclerosis associate retrovirus [MSRV]) from retroviral particles produced by cell cultures from patients with multiple sclerosis (Perron et al. 1997; Blond et al. 1999; Komurian-Pradel et al. 1999). Recently, its transcriptional activation in the brain has also been related to schizophrenia (Karlsson et al. 2001). HERV-W proviruses probably entered the genome of primates before the split between Old World and New World monkeys (Kim, Takenaka, and Crow 1999). The human genome contains at least 70, 100, and 30 HERV-W–related *gag*, *pro*, and *env* regions, respectively (Voisset et al. 2000), although all elements of the family are apparently not competent for replication (Blond et al. 1999). Interestingly, one HERV-W provirus may have been recruited by its host to serve an important physiological function. The envelope gene of this proviral insertion codes for the syncytin protein, which mediates placental cytotrophoblast fusion in vivo (Blond et al. 2000; Mi et al. 2000; Stoye and Coffin 2000).

In the present work I conducted a comparative sequence approach to reconstruct the evolutionary history of HERV-W, using data from the draft sequence of the human genome. This analysis revealed several unexpected features of HERV-W evolution.

Materials and Methods

Identification of HERV-W homologous sequences within the human genome was made using BLAST (Altschul et al. 1990) from the specialized human genome BLAST page at the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/BLAST). The proviral insertion coding for syncytin (genomic sequence NT_017168.3, positions 6814289–6824510) was used as the query sequence for the initial searches. Further searches were done using as a query the region from the splice acceptor site 3 (AS3) of HERV-W, located 240 bp upstream of the 3′ LTR (Blond et al. 1999), to the 3′ end of the R region, with the

Abbreviations: AS3, splice acceptor site 3; HERV, human endogenous retroviruses; LTR, long-terminal repeat; MSRV, multiple-sclerosis associate retrovirus; NCBI, National Center for Biotechnology Information; ORF, open reading frame.

Key words: endogenous retrovirus, retrotransposition, HERV-W, L1, MSRV, retrosequence.

Address for correspondence and reprints: Departamento de Bioloxía Fundamental, Facultade de Bioloxía, Universidade de Santiago de Compostela, E-15782 Santiago de Compostela, Spain. E-mail: bfcostas@usc.es.

Mol. Biol. Evol. 19(4):526–533. 2002

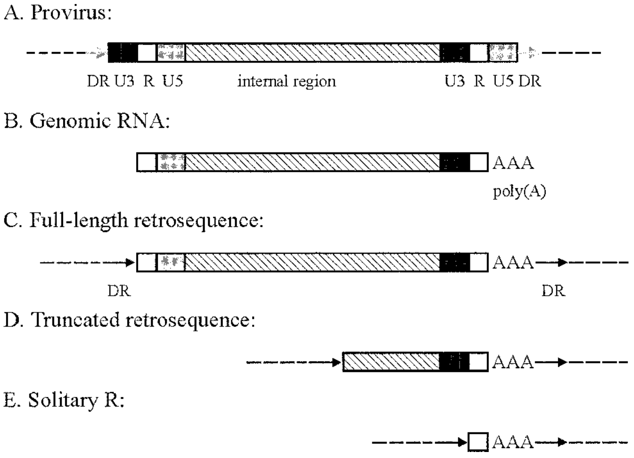

FIG. 1.—Diagrammatic representation of several HERV-W structures. Length of the different regions is not to scale. Structural features shown: black box, U3 region; white box, R region; grey box, U5 region; box with angled bars, internal region; grey arrow, short direct repeat (DR) of 4 bp; black arrow, short direct repeat of 10–16 bp; AAA, poly(A) tail; dashed line, chromosomal DNA.

addition of a 25-bp poly(A) tail. ClustalX (Thompson et al. 1997) was used for sequence alignments, which were later refined by visual inspection with GeneDoc (Nicholas and Nicholas 1997). Duplicated entries were excluded from the alignment before the other analyses. The open reading frame (ORF) finder at NCBI was used to detect long ORFs in the collected HERV-W proviruses.

HERV-W subfamilies were established by grouping sequences into different sets according to the most variable sites after exclusion of CpG dinucleotide positions (defined using the same criterion as in Costas and Naveira 2000), with little discrimination for subfamilies because of the fast mutation rate of these dinucleotides to TpG or CpA. Subfamily status was conferred on a sequence set if it was constituted by at least five elements presenting at least two diagnostic nucleotide differences. Subfamily consensus sequences were obtained by choosing the most frequent nucleotide at each position with one exception: those positions considered as CpG in the general alignment were also considered as CpG in the subfamily consensus sequences.

MEGA v2.1 (Kumar et al. 2001) was used to calculate divergence values within each set of sequences and between different sets of sequences. Net divergence values between different sets of sequences (*dN*) were calculated by the following formula:

$$dN = dXY - (dX - dY)/2, \qquad (1)$$

where *dXY* is the average distance between groups *X* and *Y*, and *dX* and *dY* are the mean within-group distances. This program was also used to reconstruct phylogenetic relationships by the neighbor-joining method (Saitou and Nei 1987) and to calculate bootstrap values for each internal branch (1,000 replicates). In all cases, Kimura's two-parameter model was applied to correct for multiple substitutions. The average age of amplification for each subfamily (T) was calculated using the formula:

$$T = K/2r, \qquad (2)$$

where T is the time of divergence, K the average pairwise divergence between sequences from the same subfamily, and r the substitution rate of pseudogene sequences in primates.

Results

A BLAST search for HERV-W homologous sequences within the human genome was carried out in April 2001, using the syncytin genomic sequence as a query. This search revealed an unexpected result. Several of the identified elements begin at the R region of the 5′ LTR and end at the R region of the 3′ LTR, presenting in addition a 3′ poly(A) tail (fig. 1*C*). Some of these unusual elements are truncated at their 5′ ends at different positions (fig. 1*D*). Furthermore, although the insertion of HERV-W proviruses is flanked by direct repeats of 4 bp, these elements resembling the genomic RNA structure are flanked by longer repeats, typically from 10 to 16 bp. A visual inspection of the flanking regions indicates that the 5′-TT/AAAA sequence and its variants derived by a single base substitution, representing an L1-endonuclease consensus cleavage site (Jurka 1997; Toda, Saito, and Tomita 2000), are frequent at the preintegration site of the unusual elements (data not shown). Nevertheless, no further analysis was done on account of the difficulty in inferring these preintegration sites because of the relative old age of the insertions (see later). All these facts are characteristic of retroposed sequences (see *Discussion*). Because of this, I shall refer to these sequences as HERV-W retrosequences in contrast with the normal HERV-W proviruses.

An additional BLAST search was done using as a query the region from the AS3 of HERV-W, located 240 bp upstream of the 3′ LTR (Blond et al. 1999), to the 3′ end of the R region, continued by a poly(A) tail. By this strategy, I collected novel 5′-truncated HERV-W retrosequences not recovered in the previous search because of their shorter length. A total of 140 sequences, representing 39 HERV-W proviruses, 40 full-length HERV-W retrosequences, and 61 truncated HERV-W retrosequences, were collected (table 1). Furthermore, this search also revealed the existence of solitary R regions with a poly(A) tail flanked by short direct repeats (fig. 1*E*), showing that inter-R recombination efficiently removes full-length HERV-W retrosequences from the genome, in a way similar to that giving rise to solitary LTRs from full-length proviruses (Mager and Goodchild 1989). Besides the known *env* ORF coding for syncytin (Blond et al. 2000; Mi et al. 2000), there are two other HERV-W proviruses preserving ORFs longer than 1,000 bp. One of them, included within the genomic clone NT_022833, extends from amino acids 64 to 524 of syncytin, sharing 87.6% homology with it. The other (NT_006307) is 1,638 bp long, corresponding to the main portion of the *pol* gene, from the conserved do-

Table 1
Collected HERV-W Sequences

Subfamily 1
Proviruses: NT_023726.3, 105993; NT_006307.3, 82341; NT_005094.3, −1538480; NT_026395.1, 708027; NT_005090.3, 132910; NT_022778.3, 531653; NT_023570.3, 152067; NT_025643.2, −97618. Full-length retrosequences: NT_022377.3, −48004. Truncated retrosequences: NT_011568.3, −151205, NT_007795.3, −317084; NT_007544.3, −237629; NT_025009.3, 243939; NT_004470.3, −751112.
Subfamily 2
Proviruses: NT_007688.3, −738864; NT_025819.2, −461712; NT_009125.3, 293788; NT_007154.3, −434835; NT_007802.3, −304618. Full-length retrosequences: NT_026239.1, 220691; NT_022863.3, −811863; NT_009702.3, 626198; NT_022803.3, 35983; NT_009486.3, −761854; NT_008012.3, 1074840; NT_005232.3, 1948204; NT_011896.4, 556046; NT_009538.3, 52540; NT_019429.3b, 1297799. Truncated retrosequences: NT_006324.3, −519052; NT_005295.3, 1385682; NT_021910.3, −414601.
Subfamily 3
Proviruses: NT_017168.3, −6814698; NT_006022.3, −952908; NT_010783.3, −1352162; NT_009782.3, 289686; NT_004424.3, 1150552; NT_022833.3, −811863; NT_022845.3, −373625; NT_010886.3, −1172958; NT_024102.3, −197995; NT_006305.3, 751815; NT_024532.3, −176200; NT_008306.3, 1422299; NT_010062.3, −152082; NT_011512.3a, −5790303; NT_025000.3, −139148; NT_011512.3b, 13811781; NT_009509.3, −910411; NT_019583.3, 1928526; NT_006710.3, 273197; NT_019721.3a, −1154621. Full-length retrosequences: NT_026483.1, −107912; NT_007343.3a, 924045; NT_004386.3, 656128; NT_007204.3, 415922; NT_026230.1, −61615; NT_005194.3, −1538480; NT_023946.3, −69376; NT_011209.3, 1267213; NT_023509.3, −1795776; NT_007186.3, 559035; NT_010289.3, 4125521; NT_017114.3, −267851; NT_005789.3, 395893; NT_019390.3, 16362; NT_007914.3, 1634236; NT_022082.2, −97502; NT_025273.3, 1856262; NT_007299.3, −465219; NT_017126.3, −169352; NT_004658.3, 859095; NT_011512.3c, −26684164; NT_019429.3a, 234399; NT_008976.3, −435401. Truncated retrosequences: NT_008387.3, −1320165; NT_009151.3, 1192685; NT_007972.3, 554126; NT_010090.3, −3065293; NT_007288.3, −1504765; NT_010859.3, −806355; NT_010795.3, −31945; NT_010140.3, −8852912; NT_016864.3, −2687151; NT_011416.3, −229040; NT_005823.3, 193611; NT_024967.3, −74507; NT_008445.3, −1253512; NT_005337.3, 475947; NT_022744.2, 130500; NT_010736.3, 894982; NT_007412.3, 1224017; NT_022939.3, 115371; NT_011493.3, 1420923; NT_005440.3, 135607; NT_024379.3, −211596; NT_008209.3, −638735; NT_025148.3, −50012; NT_005332.3, 18640; NT_007754.3, −340167; NT_008783.3, 844117; NT_011719.3, −2126546; NT_004686.3, 190730; NT_024353.3, 122047; NT_007358.3, −939028; NT_006631.3, −958043; NT_010101.3, −610217; NT_011762.3, −127118; NT_009184.3, 657527; NT_011076.3, 381009; NT_009534.3, −1519191; NT_022391.3, 49782; NT_017696.3, 960637; NT_022357.1, 124509; NT_023462.3, 166956; NT_008364.3, 1008809; NT_009334.3, 2299942; NT_009714.3, 2129052; NT_010986.3, 1758757; NT_010113.3, −5408003; NT_026437.1, 1698224; NT_011696.3, −588602; NT_024352.3, −28141; NT_025094.2, −9096.
Unclassified
Proviruses: NT_024370.3, −51399; NT_019721.3b, 2788336; NT_007343.3b, 2485295; NT_008952.2, −375294; NT_005102.3, −280542; NT_022602.3, −29752. Full-length retrosequences: NT_025287.3, −79534; NT_008769.3, −1351186; NT_025880.2, −176011; NT_011721.2, 136079; NT_010351.3, −137688; NT_010352.3, 346188. Truncated retrosequences: NT_005539.3, 125876; NT_011651.3, −180093; NT_004656.3, 190730; NT_011250.3, −240247.

NOTE.—Each element is identified by its accession number followed by the nucleotide position of the 3′ end of the U3 region of the 3′ LTR. Nucleotide entries containing more than one HERV-W sequence are indicated by a lowercase letter after the accession number. A minus sign indicates sequence orientation opposite the U3 region.

main 3 of the reverse transcriptase (according to Xiong and Eickbush 1988) to the end of the RNaseH region.

Alignment of the 140 sequence fragments led to its classification into three main subfamilies on the basis of consistent correlated nucleotide differences between them (fig. 2 and table 2). Interestingly, all of these diagnostic differences are located within the 3′ LTR. A total of 16 sequences remained unclassified. These unclassified elements present autapomorphic deletions removing key diagnostic positions, exclusive differences at diagnostic sites, or a combination of diagnostic nucleotides from different subfamilies (most probably because of gene conversion or recombination). Some of the unclassified sequences might represent intermediate subfamilies, eliminated from the analysis because of the absence of other elements belonging to them. Phylogenetic analyses of the sequences are consistent with this classification, with the exception of NT_011896, belonging to subfamily 2 on the basis of diagnostic differences but clustering with sequences from subfamily 1 in the phylogenetic trees. I removed this sequence from the rest of the analyses to avoid putative artifactual results. Figure 3 presents a neighbor-joining tree of the remaining 123 sequences. Sequences from subfamily 3 are clustered together with a high bootstrap support (86%). In agreement with the greater number of diagnostic differences defining this subfamily (fig. 2), the branch connecting this cluster is the longest internodal branch. The second longer internodal branch leads to the cluster of sequences from subfamily 2. Nevertheless, this cluster

```
                 *        20         *        40         *        60         *
        AATCGAAGCTGTAAAACTACAAATCGTTCTTCAAATGGAGCCCCAGATGCAGTCCATGACTAAGATCTAC
Subf1 : ...................................................................... :  70
Subf2 : ...................................................................... :  70
Subf3 : ...................................................................... :  70

                80         *       100         *       120         *       140
        CGCGGACCCCTGGACCGGCCTGCTAGCCCATGCTCCGATGTTAATGACATCGAAGGCACCCCTCCCGAGG
Subf1 : ..........................................g........................... : 140
Subf2 : ..........................................a........................... : 140
Subf3 : ..........................................A........................... : 140

                 *       160         *       180         *       200         *
        AAATCTCAACTGCACGACCCCTACTACGCCCCAATTCAGCAGGAAGCAGTTA-GAGCGGTCGTCGGCCAA
Subf1 : ....................................................-................. : 209
Subf2 : ....................................................a................. : 210
Subf3 : ....................................................-................. : 209

               220         *       240         * ↓     260         *       280
        CCTCCCCAACAGCACTTGGGTTTTCCTGTTGAGAGGGGGGACTGAGAGACAGGACTAGCTGGATTTCCTA
Subf1 : ...................................................................... : 279
Subf2 : ...................................................................... : 280
Subf3 : ...................................................................... : 278

                 *       300         *       320         *       340         *
        GGCCGACTAAGAATCCCTAAGCCTAGCTGGG-AAGGTGACCGCATCCACCTTTAAACACGGGGCTTGCAA
Subf1 : ..............T................G............C......................T.. : 349
Subf2 : ..............T................a............C......................T.. : 350
Subf3 : ..............C................-............T......................C.. : 347

               360         *       380         *       400         *       420
        CTTAGCTCACACCCGACCAATCAGGTAGTAAAGAGAGCTCACTAAAATGCTAATTAGGC-AAAAACAGGA
Subf1 : ..C................................R............A.C........T....g..... : 419
Subf2 : ..C................................G............A.A........T..-.G..... : 419
Subf3 : ..T................................A............G.T........-....A..... : 416

                 *       440         *       460         *       480         *
        GGTAAAGAAATAGCCAA-TCATCTATCGCCTGAGAGCACAGCGGGAGGGACAATGATCGGGATATAAAC-
Subf1 : .............T...a....C..............................................c : 489
Subf2 : .............T...A....A..............................................- : 488
Subf3 : .............C...-....C..............................................- : 484

               500         *       520         *       540         *       560
        CCAGGCATTCGAGCCGGCAACGGCAACCCCCTTTGGGTCCCCTCCCTTTGTATGGGAGCTCTGTTTTCAC
Subf1 : .................N.G....A....C................a....................... : 559
Subf2 : .................G.G....A....C................A....................... : 558
Subf3 : .................C.A....t....t................T....................... : 554

                 *       580
        TCTATTAAATCTTGCAACTGCA
Subf1 : ...................... : 581
Subf2 : ...................... : 580
Subf3 : ...................... : 576
```

FIG. 2.—Alignment of subfamily consensus sequences. The ‘‘general’’ consensus sequence is shown above as a reference. Dots represent identical nucleotide positions in the three subfamily consensus sequences. Gaps are represented by dashes. CpG dinucleotide positions are shown in italics. The diagnostic positions used to classify the sequences in subfamilies are outlined in grey boxes. The arrow marks the beginning of the 3′ LTR. Nucleotides present in more than 70% of the sequences from the subfamily are shown in capital letters, whereas nucleotides present in between 50% and 70% are shown in lowercase letters. R = A or G, N = any nucleotide, and *a* = A or gap.

is not supported by a high bootstrap value. The remaining sequences constitute subfamily 1. Within each subfamily, HERV-W proviruses and HERV-W retrosequences are distributed without any tendency to split each other.

Table 3 shows the divergence values within and between different subfamilies. In agreement with the subfamily classification, divergence values between subfamilies are always greater than within each of the subfamilies. The net divergence values between subfamilies are in accordance with the number of diagnostic differences between them (fig. 2). On the other hand, divergence values between HERV-W proviruses and retro-

Table 2
Subfamily Classification of HERV-W

	Proviruses	Full-length Retrosequences	Truncated Retrosequences	Total
Subf. 1	8	1	5	14
Subf. 2	5	10[a]	3	18
Subf. 3	20	23	49	92
Unclass.	6	6	4	16
Total	39	40	61	140

[a] Sequence NT_011896 is included.

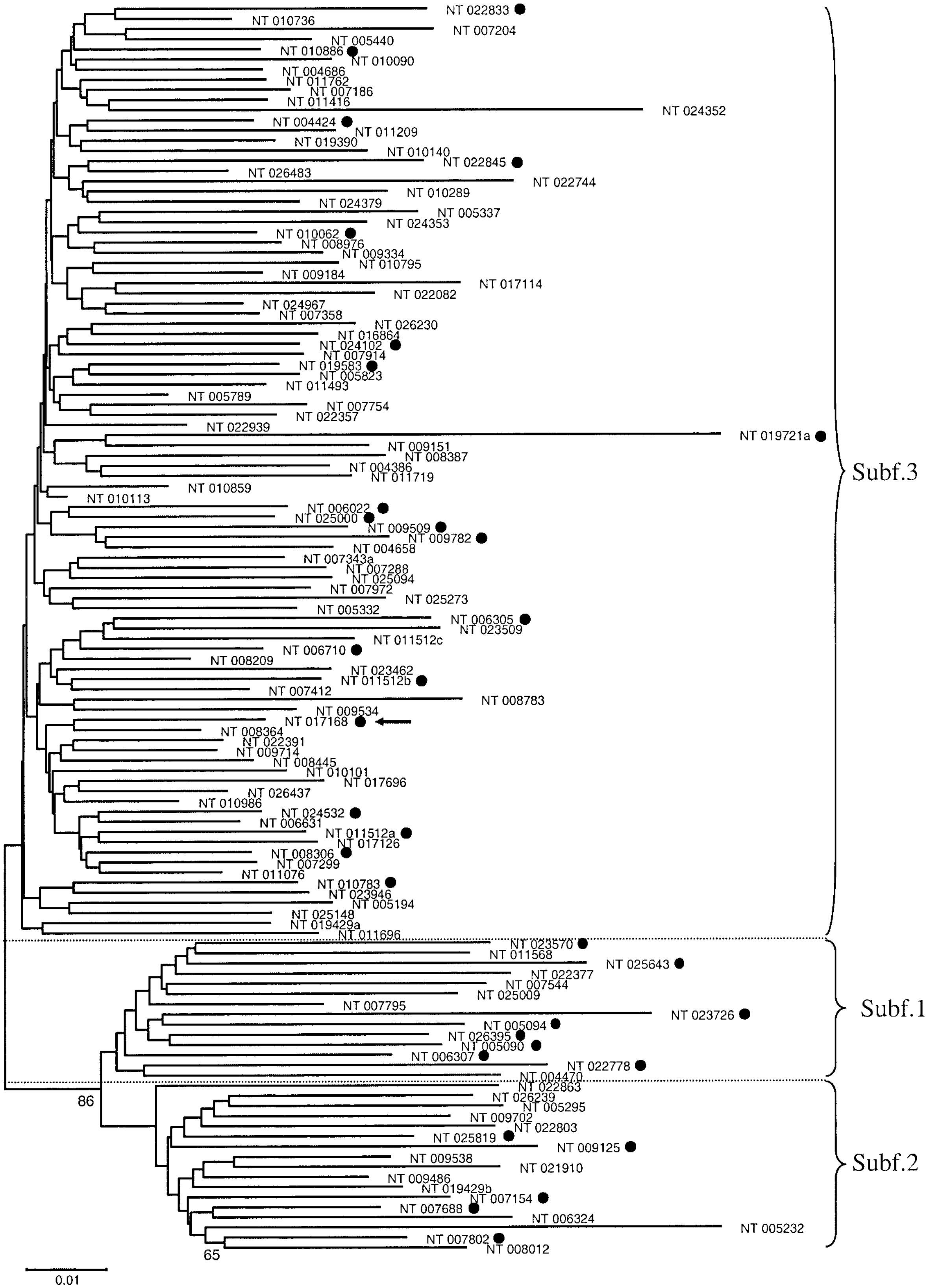

FIG. 3.—Neighbor-joining tree of the 123 sequences classified as members of any of the three subfamilies. Bootstrap values higher than 50% are shown. Brackets indicate the different subfamilies. HERV-W proviruses are marked by dots. The syncytin sequence (subfamily 3 provirus) is marked by a horizontal arrow.

Table 3
Divergence Values (%) Within and Between Subfamilies

	Subf. 1	Subf. 2	Subf. 3
Subf. 1	7.43	0.97	2.13
Subf. 2	7.72	6.08	2.90
Subf. 3	8.94	9.03	6.18

NOTE.—Divergence values were calculated after exclusion of CpG dinucleotides, using Kimura's two-parameter model. Divergence values within each subfamily are on the diagonal, divergence values between subfamilies below the diagonal, and net divergence between subfamilies above the diagonal.

Table 4
Divergence Values (%) Within and Between HERV-W Proviruses and Retrosequences

	Divergence Within Provirus	Divergence Within Retrosequences	Divergence Between Groups	Net Divergence Between Groups
Subf. 1	7.26	7.57	7.48	0.06
Subf. 2	4.89	6.60	5.71	0.00
Subf. 3	7.03	5.95	6.48	0.00

NOTE.—Divergence values were calculated after exclusion of CpG dinucleotides, using Kimura's two-parameter model.

sequences from the same subfamily are always an intermediate value between them, and the net divergence values are less than 0.1% (table 4), indicative of the absence of different clusters for proviruses and retrosequences within the same subfamily. Based on the average pairwise divergence of elements from each of the subfamilies, the estimated amplification ages range from 15.5 to 18.6 MYA, assuming $r = 0.2\%$ per million years (as in Anderssen et al. 1997), from 19.0 to 23.2, assuming $r = 0.16$ (as in Costas and Naveira 2000), or from 23.4 to 28.6, assuming $r = 0.13$ (as in Lebedev et al. 2000).

Discussion

This paper reconstructs the main features of the evolutionary history of HERV-W. This family consists of three different subfamilies whose main periods of activity extend over a short period of evolutionary time (~5 Myr). On the basis of the average pairwise divergence between members of each subfamily, subfamily 1 seems to be the oldest, and the other two originated independently from it. Although the average pairwise divergence must be considered a rough estimation of the relative amplification ages, because of stochastic errors and variation in substitution rates (r) within different genomic regions and over time, another fact suggests this hypothesis. There are several sites within the subfamily 1 consensus sequence with two alternative nucleotides (or a nucleotide and a gap) in similar proportion, whereas each of the other two subfamilies presents only one of the two alternative differences (positions 386, 415, and 438 in the alignment of fig. 2). We must take into account that an element must be able to create copies of itself at a relatively high level over a significant period of time in order to give rise to a detectable subfamily (Deininger et al. 1992). So, it is possible that minor subfamilies are hidden under the umbrella of the big ones, especially subfamily 3, accounting for ~75% of all the HERV-W insertions. For instance, 25 of the 92 elements from this subfamily share a 9-bp deletion, and 21 present a 10-bp insertion. The existence of several ambiguous positions in the subfamily consensus sequences also suggests this possibility (fig. 2). Nevertheless, there are no other correlated diagnostic differences characterizing these putative groups, and furthermore, the divergence values between them and the other elements do not support the existence of new subfamilies (data not shown). Therefore, within each of the subfamilies, all the elements most probably arose from very few closely related active elements.

This picture of the intragenomic spread of HERV-W is in clear contrast with other HERV families, such as HERV-K, HERV-H, or ERV9, that remained transpositionally active over extended periods of primate evolution, leading to several distinct subfamilies over time (Anderssen et al. 1997; Medstrand and Mager 1998; Costas and Naveira 2000; Lebedev et al. 2000). Thus, each HERV family underwent its particular evolutionary history, and these histories may be quite different from each other. The presumably shorter period of amplification in the case of HERV-W (based on the average integration age of the different subfamilies), as well as the apparent lack of intact ORFs, suggests that the MSRV isolated from retroviral particles produced by cell cultures from patients with multiple sclerosis (Perron et al. 1997) may be an exogenous member of the HERV-W family. The failure to detect intermediate subfamilies between subfamily 1 and subfamily 3 (that present seven diagnostic differences within the U3 region; fig. 2) also suggests the possibility that these two subfamilies might be originated by two independent germline infections.

The most surprising fact of the evolutionary dynamics of HERV-W is the existence of a high proportion of insertions showing characteristic features of retrosequences, such as acquisition of a poly(A) 3′ tail, presence of direct flanking repeats of 10–16 bp, and a structure resembling mRNAs. Recently, Esnault, Maestre, and Heidmann (2000) and Wei et al. (2001) formally disclosed the ability of the non-LTR retrotransposon L1 to retrotranspose polyadenylated RNA transcripts in *trans* displaying these characteristics. Thus, HERV-W presumably spread by two different mechanisms: (1) the normal retrotransposition process of retroviruses, giving rise to full-length proviruses with intact LTRs, and (2) the parasitism on the L1 element, as in the case of short interspersed elements (SINEs; Mathias et al. 1991; Ohshima et al. 1996), giving rise to HERV-W retrosequences. Alternatively, it is legitimate to speculate that the reverse transcriptase of HERV-W itself would be responsible for HERV-W retrosequences formation. Nevertheless, the fact that nonviral RNAs encapsidated in retroviral particles generate integrated cDNA genes lacking the hallmarks of naturally occurring processed pseudogenes (they are 5′- and 3′-truncated and do not contain poly(A) tails) strongly militates against this hypothesis (Dornburg and Temin 1988, 1990). The existence of both types of elements within each of the subfamilies clearly supports the idea that HERV-W retrosequences formation is dependent on the expression of full-length proviruses, which are the source of genomic RNA. The alternative hypothesis of independent evolution of retrosequences after their origin should give rise to subfamilies constituted only by HERV-W retrosequences, but these subfamilies have not been identified. Taking into account that HERV-W retrosequences are expected to be ''dead on arrival'' copies, the lower success of HERV-W within the genome, compared with the other afore-mentioned HERV families, might be related to the existence of a considerable proportion of genomic RNA sequestered by the L1 machinery.

The putative impact of HERV-W retrosequences on the genome might be quite different from that of HERV-W proviruses. Retroviral protein expression may cause deleterious effects on the host by several processes. Thus, the antigenic character of proteins encoded by *gag* and *env* has been associated with several autoimmune pathologies (Nakagawa and Harrison 1996; Perron et al. 1997). The transmembrane domain of the envelope protein presents immunosuppressive effects (Cianciolo et al. 1985; Haraguchi et al. 1997), suggesting its possible implication in tumoral processes, leading to the escape of immune rejection by tumoral cells (Mangeney and Heidmann 1998). Other peptides encoded by small ORFs (two putative small ORFs have been described in HERV-W; Blond et al. 1999) might interfere with the cellular machinery (Boese et al. 2000). Furthermore, active proviruses may be the source of new insertions, acting as insertional mutagens (Mitreiter et al. 1994; Vasicek et al. 1997). All these deleterious effects are not associated with HERV-W retrosequences, which lack the capability to be expressed because of the loss of LTRs (not only in truncated but also in full-length retrosequences). HERV insertions may also be involved in deleterious chromosomal rearrangements by ectopic recombination between two copies of the same family of HERVs located at different chromosomal loci (Kamp et al. 2000; Sun et al. 2000). This effect is expected to be substantially reduced in the case of truncated retrosequences of short length. On the other hand, insertion of HERV-W retrosequences might introduce short enhancer sequences near genes (most of the enhancer signals are within the U3 region), providing raw material for natural selection. Thus, this type of insertion might represent a novel potential mechanism for the evolution of enhancers, adding a new possibility for L1 to shape the mammalian genomes (Kazazian and Moran 1998; Moran, DeBerardinis, and Kazazian 1999; Pickeral et al. 2000).

Supplementary Material

The alignment of the 140 insertions is available as Supplementary Material on-line.

Acknowledgment

The author is a recipient of a postdoctoral fellowship from the USC/Xunta de Galicia.

LITERATURE CITED

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

ANDERSSEN, S., E. SJOTTEM, G. SVINENG, and T. JOHANSEN. 1997. Comparative analyses of LTRs of the ERV-H family of primate-specific retrovirus-like elements isolated from marmoset, African green monkey, and man. Virology **234**: 14–30.

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

BLOND, J. L., F. BESEME, L. DURET, O. BOUTON, F. BEDIN, H. PERRON, B. MANDRAND, and F. MALLET. 1999. Molecular characterization and placental expression of HERV-W, a new human endogenous retrovirus family. J. Virol. **73**: 1175–1185.

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.

BOESE, A., M. SAUTER, U. GALLI, B. BEST, H. HERBST, J. MAYER, E. KREMMER, K. ROEMER, and N. MUELLER-LANTZSCH. 2000. Human endogenous retrovirus protein cORF supports cell transformation and associates with the promeyelocytic leukemia zinc finger protein. Oncogene **19**: 4328–4336.

BROSIUS, J. 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene **238**:115–134.

CIANCIOLO, G. J., T. D. COPELAND, S. OROSZLAN, and R. SNYDERMAN. 1985. Inhibition of lymphocyte proliferation by a synthetic peptide homologous to retroviral envelope proteins. Science **230**:453–455.

COSTAS, J., and H. NAVEIRA. 2000. Evolutionary history of the human endogenous retrovirus family ERV9. Mol. Biol. Evol. **17**:320–330.

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

DORNBURG, R., and H. M. TEMIN. 1988. Retroviral vector system for the study of cDNA gene formation. Mol. Cell. Biol. **8**:2328–2334.

———. 1990. cDNA genes formed after infection with retroviral vector particles lack the hallmarks of natural processed pseudogenes. Mol. Cell. Biol. **10**:68–74.

ESNAULT, C., J. MAESTRE, and T. HEIDMANN. 2000. Human LINE retrotransposons generate processed pseudogenes. Nat. Genet. **24**:363–367.

GOODCHILD, N. L., D. A. WILKINSON, and D. L. MAGER. 1993. Recent evolutionary expansion of a subfamily of RTVL-H human endogenous retrovirus-like elements. Virology **196**: 778–788.

HARAGUCHI, S., R. A. GOOD, G. J. CIANCIOLO, R. W. ENGELMAN, and N. K. DAY. 1997. Immunosuppressive retroviral peptides: immunopathological implications for immunosuppressive influences of retroviral infections. J. Leukoc. Biol. **61**:654–666.

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. USA **94**:1872–1877.

Kamp, C., P. Hirschmann, H. Voss, K. Huellen, and P. Vogt. 2000. Two long homologous retroviral sequence blocks in proximal Yq11 cause AZFa microdeletions as a result of intrachromosomal recombination events. Hum. Mol. Genet. **9**:2563–2572.

Karlsson, H., S. Bachmann, J. Schroder, J. McArthur, E. F. Torrey, and R. H. Yolken. 2001. Retroviral RNA identified in the cerebrospinal fluids and brains of individuals with schizophrenia. Proc. Natl. Acad. Sci. USA **98**:4634–4639.

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

Kim, H. S., O. Takenaka, and T. J. Crow. 1999. Isolation and phylogeny of endogenous retrovirus sequences belonging to the HERV-W family in primates. J. Gen. Virol. **80**:2613–2619.

Komurian-Pradel, F., G. Paranhos-Baccala, F. Bedin et al. (11 co-authors). 1999. Molecular cloning and characterization of MSRV-related sequences associated with retrovirus-like particles. Virology **260**:1–9.

Kumar, S., K. Tamura, I. B. Jakobsen, and M. Nei. 2001. MEGA2: molecular evolutionary genetics analysis software. Distributed by the authors (http://www.megasoftware.net/).

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.

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

Mager, D., and N. Goodchild. 1989. Homologous recombination between the LTRs of a human retrovirus-like element causes a 5-kb deletion in two siblings. Am. J. Hum. Genet. **45**:848–854.

Mangeney, M., and T. Heidmann. 1998. Tumor cells expressing a retroviral envelope escape immune rejection in vivo. Proc. Natl. Acad. Sci. USA **95**:14920–14925.

Mathias, S. L., A. F. Scott, H. H. Kazazian Jr., J. D. Boeke, and A. Gabriel. 1991. Reverse transcriptase encoded by a human transposable element. Science **254**:1808–1810.

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

Mi, S., X. Lee, X. Li et al. (12 co-authors). 2000. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature **403**:785–789.

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

Moran, J. V., R. J. DeBerardinis, and H. H. Kazazian Jr. 1999. Exon shuffling by L1 retrotransposition. Science **283**: 1530–1534.

Nakagawa, K., and L. C. Harrison. 1996. The potential roles of endogenous retroviruses in autoimmunity. Immunol Rev. **152**:193–236.

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

Ohshima, K., M. Hamada, Y. Terai, and N. Okada. 1996. The 3′ ends of tRNA-derived short interspersed repetitive elements are derived from the 3′ ends of long interspersed repetitive elements. Mol. Cell. Biol. **16**:3756–3764.

Perron, H., J. Garson, F. Bedin et al. (13 co-authors). 1997. Molecular identification of a novel retrovirus repeatedly isolated from patients with multiple sclerosis. The Collaborative Research Group on Multiple Sclerosis. Proc. Natl. Acad. Sci. USA **94**:7583–7588.

Pickeral, O. K., W. Makalowski, M. S. Boguski, and J. D. Boeke. 2000. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. Genome Res. **10**:411–415.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

Stoye, J. P., and J. M. Coffin. 2000. A provirus put to work. Nature **403**:715–717.

Sun, C., H. Skaletsky, S. Rozen, J. Gromoll, E. Nieschlag, R. Oates, and D. Page. 2000. Deletion of azoospermia factor a (AZFa) region of human Y chromosome caused by recombination between HERV15 proviruses. Hum. Mol. Genet. **9**:2291–2296.

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

Toda, Y., R. Saito, and M. Tomita. 2000. Characteristic sequence pattern in the 5- to 20-bp upstream region of primate *Alu* elements. J. Mol. Evol. **50**:232–237.

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.

Vasicek, T. J., L. Zeng, X. J. Guan, T. Zhang, F. Costantini, and S. M. Tilghman. 1997. Two dominant mutations in the mouse fused gene are the result of transposon insertions. Genetics **147**:777–786.

Voisset, C., O. Bouton, F. Bedin, L. Duret, B. Mandrand, F. Mallet, and G. Paranhos-Baccala. 2000. Chromosomal distribution and coding capacity of the human endogenous retrovirus HERV-W family. AIDS Res. Hum. Retroviruses **16**:731–740.

Wei, W., N. Gilbert, S. L. Ooi, J. F. Lawler, E. M. Ostertag, H. H. Kazazian, J. D. Boeke, and J. V. Moran. 2001. Human L1 retrotransposition: *cis* preference versus *trans* complementation. Mol. Cell. Biol. **21**:1429–1439.

Xiong, Y., and T. H. Eickbush. 1988. Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. Mol. Biol. Evol. **5**:675–690.

Thomas Eickbush, reviewing editor

Accepted December 4, 2001