Evidence of Selection on the Domesticated ERVWE1 *env* Retroviral Element Involved in Placentation

Bertrand Bonnaud, Olivier Bouton,* Guy Oriol,* Valérie Cheynet,* Laurent Duret,† and François Mallet**

*UMR 2714 CNRS-bioMérieux IFR128 BioSciences Lyon-Gerland, Ecole Normale Supérieure de Lyon, Lyon, France; and †Laboratoire de Biométrie et Biologie Evolutive, UMR CUR 5558, Université Claude Bernard – Lyon 1, Lyon, France

The human endogenous retrovirus HERV-W multicopy family includes a unique proviral locus, termed ERVWE1, which contains *gag* and *pol* pseudogenes and has retained a full-length envelope open reading frame (ORF). This Env protein (syncytin) is a highly fusogenic membrane glycoprotein and has been proposed to be involved in hominoid placental physiology. To track the hallmarks of natural selection acting on the ERVWE1 *env* gene, the pattern of substitutions and indels was analyzed within all human HERV-W elements and along the ERVWE1 orthologous loci in chimpanzee, gorilla, orangutan, and gibbon. The comparison of ERVWE1 and paralogous HERV-W copies revealed an ERVWE1-specific signature consisting of a four amino acid deletion in the intracytoplasmic tail of the glycoprotein. We show that this deletion is crucial for the envelope fusogenic activity. The comparison of the human ERVWE1 locus with its orthologs demonstrates the existence of a selective pressure to maintain the *env* reading frame open. Notably, the 3′ part of the *env* gene, encoding regions required for the fusion process, is under purifying selection. The identification of selective constraints on *env* ERVWE1 confirms that this retroviral locus has been recruited in the hominoid lineage to become a bona fide gene.

Introduction

Creation of new genes certainly has played a major role in species evolution. The sequencing of the human and mouse genomes offers the possibility to identify genes that are species-specific and hence that may have contributed to the evolution of novel functions in the rodent or primate lineages. One important difficulty, however, is to distinguish real new genes from function-less gene-like sequences, such as pseudogenes, that are highly abundant in mammalian genomes.

Insertion of retroviral elements can be a source of gene novelties, by providing either new protein-coding regions or regulatory elements. Human chromosomes contain thousands of retrovirus-like sequences, representing about 8% of the genome (International Human Genome Sequencing Consortium 2001). It is generally admitted that these Human Endogenous RetroViruses (HERVs) derive originally from rare events of germ-line infections by exogenous retroviruses leading to proviral DNA integration. The infectious retrovirus founding the contemporary HERV-W family (Blond et al. 1999) entered the human ancestor genome after the divergence between *Catarrhini* and *Platyrrhini* (Kim, Takenaka, and Crow 1999; Voisset et al. 2000), i.e., less than 40 MYA (Goodman et al. 1998). The spread of the HERV-W family into the genome essentially results from events of intracellular retrotransposition of transcriptionally active copies, a phenomenon mediated either by their own reverse transcriptase (RT) machinery or by RT from LINE elements (Costas 2002; Pavlicek et al. 2002). Generally, due to the absence of a selective pressure, HERV-W elements have accumulated inactivating substitutions (frame-shifts, nonsense mutations), leading to complex multicopy families whose transmission is exclusively Mendelian. Thus, the contemporary HERV-W family consists of collections of heterogeneous elements, ranging from full-length defective proviruses (*gag, pol*, and *env* genes flanked at both extremities by two long terminal repeats [LTRs]) to isolated LTRs derived from recombination events.

HERV-W elements were found to be transcribed at various levels in several pathological and physiological contexts, e.g., certain tissues of multiple sclerosis– (Perron et al. 1997) and schizophrenia-affected (Karlsson et al. 2001) individuals, testis (Mi et al. 2000), and placenta (Blond et al. 1999; Mi et al. 2000). The HERV-W transcripts detected in the placenta mainly result from the expression of a unique locus, termed ERVWE1 (OMIM accession number 604659). These transcripts are matured through a specific splicing strategy (Blond et al. 1999) and lead to the expression of an envelope glycoprotein in vivo (Voisset et al. 2000). Interestingly, this locus was shown to be the only copy of the whole family having retained a complete *env* open reading frame (ORF) (Voisset et al. 2000). The *gag* and *pol* elements of this ERVWE1 provirus contain stop codons and frameshift mutations. The fact that, contrarily to *gag* and *pol*, the env ORF has not become a pseudogene led us to propose that this *env* gene had been selectively preserved (Blond et al. 1999). The Env ERVWE1 envelope exhibits fusogenic properties (Blond et al. 2000; Mi et al. 2000) and was shown to be involved in trophoblast differentiation (Mi et al. 2000; Frendo et al. 2003). We have identified orthologs of the human ERVWE1 locus in chimpanzee, gorilla, orangutan, and gibbon, and we have shown that they contain a functional env ORF (Mallet et al. 2004). These different findings strongly suggest that the ERVWE1 *env* gene has been recruited to play a role in placental physiology (Mallet et al. 2004).

To better understand how this locus was domesticated during primates evolution, we have searched for the hallmarks of selective pressures acting on the ERVWE1 *env* gene. Patterns of insertions or deletions (indels) and base susbtitutions have been evaluated using (H)ERV-W intra- and interspecies data. The comparison of the proviral human ERVWE1 locus with HERV-W paralogous copies

Key words: HERV-W, paralogous gene, orthologous gene, envelope, endogenous retrovirus, hominoid.

E-mail: francois.mallet@ens-lyon.fr.

Mol. Biol. Evol. 21(10):1895–1901. 2004
doi:10.1093/molbev/msh206
Advance Access publication July 14, 2004

has allowed us to identify an ERVWE1-specific signature consisting of a four amino acid deletion in the intracytoplasmic tail of the glycoprotein. This deletion event has been found to be crucial for the envelope fusogenic activity. ERVWE1 comparison with human paralogs and chimpanzee, gorilla, orangutan, and gibbon orthologs demonstrates the existence of a selective pressure to maintain the *env* reading frame open and of a relatively high rate of evolution of the encoded protein.

Materials and Methods

Collection of HERV-W Elements

We have identified HERV-W homologous sequences from GenBank by using Blast (Altschul et al. 1990) with the ERVWE1 locus (human chromosome 7q21.2, accession number AC000064, position 28068–38289) query sequence. Then, HERV-W related sequences were compared using RepeatMasker (A.F.A. Smit and P. Green, http://ftp.genome.washington.edu/RM/repeatmaster.org) to a library containing reference sequences of the HERV-W family (the ERVWE1 locus) cut out into functional parts (LTR, *gag, pol, env*). An ERV-9 reference sequence (accession number Z84475, position 25937–35737) was also included. The ERV-9 family is the closest relative to the HERV-W family, and the closest elements share at most 80% identity (Blond et al. 1999). This step allowed us to eliminate sequences belonging to the ERV-9 family or presenting more than 20% divergence from the ERVWE1 sequence.

Consensus sequences of the HERV-W *env* gene and the RT pseudogene were computed with Seaview (Galtier, Gouy, and Gautier 1996) with a majority threshold at 50%. The *env* genes were delimited from start to stop codon : 1,617 base pairs (bp) for the human ERVWE1 locus (accession number AC000064 from nucleotide 35879 to 37495). The approximate boundaries of the RT pseudogenes were defined by similarity with functional RTs from other retroviral families. This RT region spanned 1,500 bp in the human ERVWE1 locus (accession number AC000064 from nucleotide 32703 to 34202).

Orhologous ERVWE1 loci accession numbers : *Pan troglodytes* (AY101586, AY101587), *Gorilla gorilla* (AY101588, AY101589), *Pongo pygmaeus* (AY101590, AY101591), and *Hylobates pileatus* (AY101592, AY-101593).

Phylogenetic and Comparative Analysis of (H)ERV-W Elements

GenBank queries and sequences retrieval were made under ACNUC using Query (Perriere and Thioulouse 1996). Sequence alignments were computed with ClustalW (Thompson, Higgins, and Gibson 1994). Phylo_win (Galtier, Gouy, and Gautier 1996) was used to build phylogenetic trees. Distances between sequences were computed under the Kimura two-parameters model (Kimura 1980) and under the Li model for synonymous (K_S) and nonsynonymous (K_A) substitutions (Li 1993). We performed a maximum-likelihood test of selection, using the codon substitution models proposed by Yang et al. (2000), as implemented in the PAML package (http://abacus.gene.ucl.ac.uk/software/paml.html). Likelihood values and parameters were estimated under models M0, M1, M2, M3, M7, and M8 (Yang et al. 2000).

Cell-Cell Fusion Assay

Envelope glycoprotein expression plasmids were transfected by calcium phosphate precipitation into TELCeB6 expressing β-galactosidase cells as previously described (Blond et al. 2000). Three hours after transfection, cells were harvested by a 10 min incubation at 37°C with 0.02% versene in phosphate-buffered saline (PBS) (Invitrogen life technologies, Cergy, France) and seeded in 2.2-cm diam wells at a density of 6×10^4 /well. After adhesion of the transfected cells, HeLa indicator cells (3×10^5 /well) were overlaid. The determination of the fusion activity of the transfected envelope glycoproteins was performed after 20 h of coculture. The total number of syncytia in one well and the number of nuclei in each syncytia were determined.

FACS Analysis

TE671 cells were washed in PBS and harvested by a 10 min incubation at 37°C with 0.02% versene in PBS. 10^6 cells were stained for 1 h at 4°C with an anti-Env specific monoclonal antiboby (6A2B2) at 1/100 dilution in PBA (PBS with 2% fetal calf serum and 0.1% sodium azide). Cells were washed once with PBA and incubated with fluorescein isothiocyanate-conjugated antibody (Dako, Trappes, France). Cells were washed twice with PBA and counterstained with propidium iodide (20 μg/ml). Fluorescence of living cells was analyzed with a fluorescence-activated cell sorter (FACSCalibur; Becton Dickinson, San Jose, Calif.).

Results

Comparison of the ERVWE1 Locus with Its Human Paralogs

To trace the evolution of the ERVWE1 *env* gene, we first compared it to all its paralogs present in the human genome. We have identified in GenBank 209 human HERV-W loci exhibiting an *env* sequence. These paralogs of ERVWE1 *env* are all defective (truncated or interrupted by nonsense or frameshift mutation). With these 209 sequences, we have computed a consensus *env* ORF (1,629 bp) used as a proxy for the ancestral proviral gene (under the assumption of a star phylogeny). The rate of indels and base substitutions in the 36 *env* paralogs longer than 600 bp was measured. The rate of divergence between ERVWE1 *env* and the ancestral *env* sequence is not significantly different from that observed for its 35 defective paralogs. This rate is indeed similar to the ones determined with 69 HERV-W reverse transcriptase (RT) pseudogenes used as control of the genomic drift of HERV sequences. The analysis of indels shows that 33 out of the 36 *env* sequences contain frameshift indels, i.e., indels whose length is not a multiple of three. The three other loci are the *env* ERVWE1 gene (locus 7q21.2), a full-length

env gene containing a stop codon in position 39 (locus Xq22.3), and a truncated *env* of 978 bp (locus 2q21.3). This high frequency of inactivating mutations is also observed for RT pseudogenes: 67 out of 69 sequences contain a frameshift indel. These observations indicate that the combined absence of stop codons and frameshift indels in ERVWE1 *env* is an exceptional feature in this family and is suggestive of a selective pressure on that locus.

Interestingly, the comparison of ERVWE1 *env* with the putative ancestral sequence reveals a deletion corresponding to a loss of four codons. No other HERV-W *env* sequence contains this 12-bp deletion, indicating it is clearly specific to the ERVWE1 locus. This deletion probably results from a recombination event between both elements of a tandemly repeated sequence or from DNA polymerase slippage, maintaining the reading frame open (fig. 1*A*). To evaluate the significance of this 12-bp sequence, the four codons LQMV deduced from human HERV-W paralogous *env* sequences were inserted in the phCMV-ERVWE1-env expression vector previously used to demonstrate the Env fusion capacity (Blond et al. 2000). Properties of wild-type envelope glycoprotein and LQMV-derived mutant were compared in an heterologous cell-cell fusion assay. The *env* expression vectors were transfected in TELCeB6 producer cells phenotypically characterized by a blue nucleus as expressing β-galactosidase. A limited amount of producer cells was overlaid with an excess of indicator HeLa cells, characterized by uncolored nuclei. If the glycoprotein exhibited fusogenic properties, the coculture of both indicator and producer cells would lead to the formation of syncytia containing generally one blue nucleus and ten white nuclei. A high fusogenic activity is observed for the wild-type envelope, characterized by hundreds of syncytia per well and an average of 40 nuclei per syncytia containing one or two blue nuclei, as previously observed (Blond et al. 2000; Mallet et al. 2004). Conversely, no syncytia is produced with the ERVWE1-LQMV glycoprotein in this coculture assay, indicating the loss of the fusogenic phenotype for this mutant envelope (fig. 1*B*).

Such an apparent defect of fusogenic activity could result either from (1) an alteration of the trafficking of the protein, preventing its expression at the correct place, or from (2) an intrinsic alteration of the fusogenic property due to the localization of the four amino acid insertion within the protein. A FACS analysis was performed to evaluate the LQMV Env expression at the cell membrane. Three *env* vectors expressing the wild-type envelope, the LQMV-derived mutant, and a negative control (native *env* gene in antisense orientation) were transfected in TELCeB6 cells. Env protein expression at the cell membrane was detected with a fluorescent anti-Env specific monoclonal antibody recognizing the extra-cellular domain of the transmembrane subunit of the ERVWE1 Env protein. Both wild-type and mutant glycoproteins, when compared

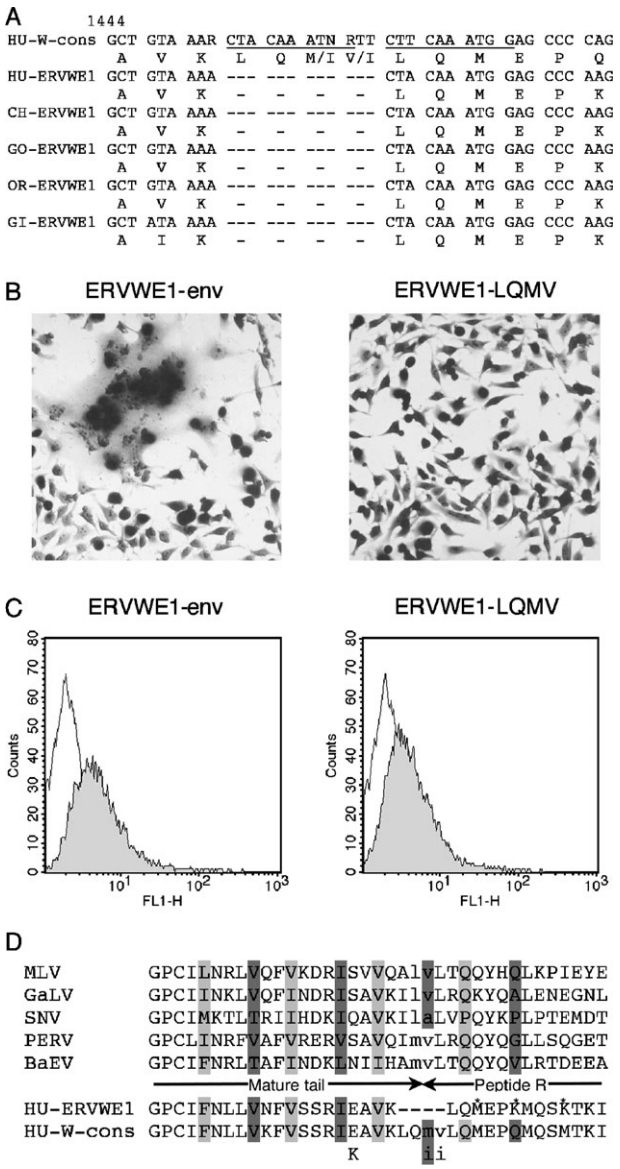

FIG. 1.—ERVWE1 *env* 12-bp specific deletion. (*A*) Alignment of orthologous and paralogous *env* sequences, in the region spanning the *env* ERVWE1-specific deletion. Nucleic and protein alignments of part of ERVWE1 *env* from human (HU), chimpanzee (CH), gorilla (GO), orangutan (OR), gibbon (GI), and part of the *env* HERV-W consensus (HU-W-cons) from 209 paralogous HERV-W *env* sequences are shown. The consensus sequence was calculated with a threshold of 80% and two amino acid residues were indicated when the main residue was under this threshold. Repeated sequences are underlined. (*B*) Formation of syncytia by ERVWE1 envelopes. TELCeB6 producer cells were transfected with plasmids expressing the wild-type envelope (ERVWE1 *env*) and the four codon envelope mutant (ERVWE1-LQMV). Producer cells were overlaid with HeLa indicator cells. Cocultures were stained with X-Gal substrate to visualize the nuclei of the producer cells and then with May-Grünwald and Giemsa solutions (Sigma, Lyon, France). (*C*) Cell surface expression of ERVWE1 envelopes. TE671 cells were transfected with plasmids expressing the wild-type envelope (ERVWE1-*env* and shaded area), using the four codon envelope mutant (ERVWE1-LQMV and shaded area) and the antisense orientation ERVWE1 envelope (white area) as control. Transfected cells were stained with 6A2B2 anti-envelope monoclonal antibody and analyzed by FACS. (*D*) Alignment of cytoplasmic tail sequences from MLV (GenBank accession number M14702), GaLV (AF055060), SNV (Engelstadter et al. 2001), PERV (Y12238), BaEV (D10032), and *env* ERVWE1 and *env* HERV-W consensus (HU-W-cons). Experimentally determined and putative protease cleavage motives

→

are indicated in lowercase. Dark and light shading show positions 1 and 4 of the heptad repeat, respectively, consisting of hydrophobic residues that form the helix interface; residues that should extend the heptad repeat in *env* ERVWE1 are labeled with stars.

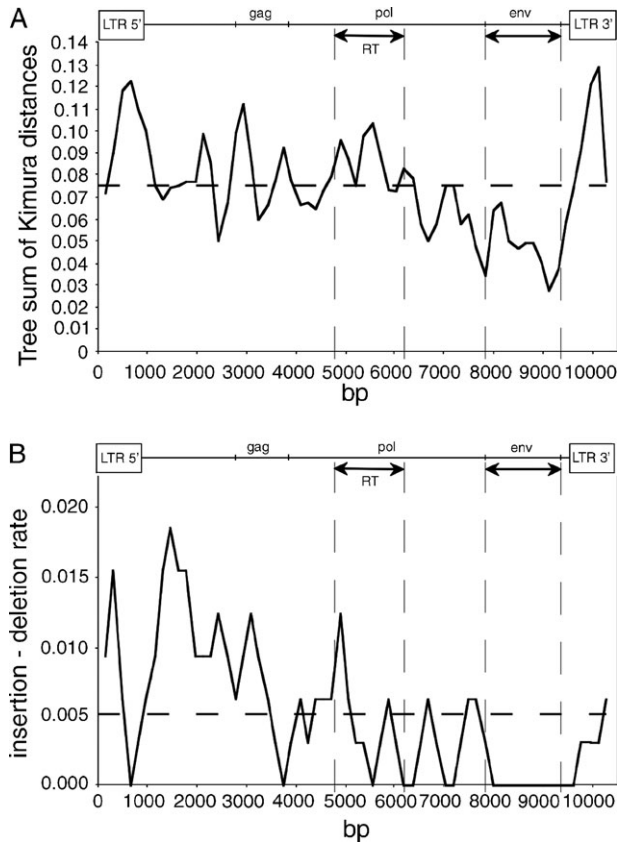

FIG. 2.—Analyses of base substitutions and indels in five ERVWE1 orthologous loci from primate species. Substitutions were computed on a 326-bp sliding window shifting 163 bp along the ERVWE1 locus between each step. *gag, pol* (including RT), and *env* genes and 5′ and 3′ LTRs are depicted on the provirus. (*A*) The sum of Kimura distances between orthologous sequences was calculated as the sum of branches of Neighbor-Joining trees. The mean of the sum of branch lengths is drawn as a dashed line. (*B*) Local comparison of insertions and deletions in five ERVWE1 orthologs. The rate of gaps between species was calculated from the alignment. The whole locus mean is drawn as a dashed line.

with the negative control, display a similar peak shift (fig. 1*C*). This demonstrates that the insertion of the LQMV amino acids does not significantly modify the protein transport to the plasma membrane. The alteration of the fusogenic property might be the result of an alteration of the primary or secondary structure of the envelope intracytoplasmic tail as a result of the introduction of the four amino acid stretch. Comparison of the wild-type Env ERVWE1 and Env HERV-W consensus envelopes with exogenous retroviral envelopes shows that the four amino acid deletion modifies two retroviral features of the intracytoplasmic tail (fig. 1*D*). First, the viral protease cleavage site disappears from the ERVWE1 fusogenic envelope. Such a cleavage conventionally contributes to the maturation of retroviral envelopes (Rein et al. 1994) to produce Env conformational changes required for membrane fusion. Second, the shortening of the intracytoplasmic tail disturbs the end of a helical structure presumably involved in the mechanism of cell-cell membrane fusion (Taylor and Sanders 2003). This alteration is due to the substitution of two hydrophobic glutamine residues by two charged lysines. Conversely, although the LQMV insertion seems to restore a putative cleavage site and the helical structure, the absence of fusion suggests the absence of a specific HERV-W functional protease, although another maturation defect cannot be excluded.

Pattern of Substitution and Indels Along the ERVWE1 Locus in Hominoids

The regional variations in substitution rate along the complete provirus were analyzed in the ERVWE1 orthologs from the five hominoid species. Kimura distances along the ERVWE1 locus were calculated globally in the whole lineage by summing all branch lengths in the phylogenetic tree. The substitution rate profile (measured with a 326-bp sliding window) reveals two main features, a peak of divergence for both proviral regulatory regions (5′ and 3′ LTRs) and a low divergence along the *env* gene (fig. 2*A*). Overall, in the five species, the rate of substitution over the *env* ORF (1,617 bp) is about 0.050 substitutions per site, compared to 0.079 substitutions per site for the entire ERVWE1 provirus (8,815 bp, *env* excluded) and 0.089 substitutions per site for the RT pseudogene (1,500 bp). The differences of substitution rates between the *env* gene and both the corresponding provirus and RT regions are significant (χ^2, 1 df, $P < 0.001$), consistent with a negative selective pressure acting on the *env* ORF.

Under the hypothesis that a protein has a function, the frame encoding this protein should be conserved, i.e., mutations inducing stop codons and frameshifting indels should be strongly counterselected. In ERVWE1 *env*, we did not detect any nonsense substitution or indel in the ORF of any of the five hominoid species. To infer the expected number of such events under the null hypothesis of neutral evolution, we have analyzed the pattern of nonsense substitutions and indels in the complete ERVWE1 proviruses. The orthologous ERVWE1 *gag* and *pol* pseudogenes share several common stop codon positions, which indicates that they were already defective before the divergence of the last common ancestor of the five species. Six new stop codons occur in the ERVWE1 RT pseudogene along the hominoids lineage. Given that *env* and RT present similar sizes and the same proportions of codons sensitive to nonsense mutations (i.e., that may give a stop codon in one single base replacement), we would have expected about the same number of nonsense substitutions in *env* if it were a pseudogene. The difference between the observed and expected number of nonsense mutations is significant (Fisher's exact test, $P = 0.012$). Moreover, along the whole ERVWE1 locus (10,432 bp) we observe 52 indel events. Interestingly, none of these indels occurs within the *env* ORF (1,617 bp), whereas they are evenly distributed along the other regions of the locus (fig. 2*B*). Under the hypothesis that the rate of indel was the same along the whole locus, we would have expected approximately eight indels in the *env* ORF. This difference between the observed and the expected values is highly significant (Fisher's exact test, $P < 0.001$).

There is a possible ascertainment bias here since the human ERVWE1 *env* gene was identified because it contains an intact ORF. We therefore recomputed the number of nonsense substitutions and indels expected in

ERVWE1 *env* in the whole phylogenetic tree, excluding all branches leading to the human lineage. The expected occurrences (four stop codons, five indels) were still found to be significantly higher ($P = 0.053$ and $P = 0.011$, respectively) than the observed zero values. Hence, the analyses of ERVWE1 orthologous sequences in hominoids provide evidence of the action of a selective pressure to preserve the *env* ORF as a functional cellular gene.

Synonymous and Nonsynonymous Substitutions in *env* ERVWE1

To determine the nature of the selective pressure on ERVWE1 *env*, the ratio of nonsynonymous (K_A) to synonymous (K_S) substitutions rate was analyzed. The K_A/K_S ratio obtained for *env* ERVWE1 over the whole phylogenetic tree is 0.8 (fig. 3*A*). This value is close to the theoretical value expected in case of neutral evolution and therefore does not reveal any obvious selective pressure on the protein sequence. However, the K_A/K_S ratio analysis over the whole gene may be obscured when both positive and negative selective pressures affect different parts of a gene. We therefore measured the K_A/K_S ratio of the whole tree over a 180-bp sliding window (fig. 3*B*). Two main regions emerge in the *env* ORF, corresponding to approximately the first 500 bp and to the last 1,100 bp. The 3′ part of the gene appears to be subject to a negative selective pressure (K_A/K_S about 0.6). To test the significance of this low K_A/K_S ratio in the 3′ part of the gene, we have performed a maximum-likelihood test of selection, using the codon substitution models proposed by Yang et al. (2000). The K_A/K_S ratio estimated under the model M0 (that assumes a constant K_A/K_S ratio along the sequence) is 0.56. The model M0 (log-likelihood $l = -1841.41$) is significantly more likely ($P < 0.05$) than the null hypothesis of neutral evolution (i.e., $K_A/K_S = 1$; $l = -1843.46$), which confirms that the 3′ part of the gene is under purifying selective pressure. More complex models (assuming heterogeneity of K_A/K_S along the sequence or among branches of the phylogenetic tree) do not significantly improve the likelihood of the data. In the 5′ region, we observe a relatively high K_A value, a low K_S value, and the resulting K_A/K_S ratio is above one, suggesting a positive selection. However, likelihood-ratio test comparisons indicate that this K_A/K_S ratio is not significantly different from the null hypothesis for neutral evolution and hence no obvious selective pressure, neither positive nor negative, could be inferred.

A relatively high K_A/K_S ratio versus a consistently lower K_A/K_S ratio in the 5′ and 3′ parts of the *env* gene, respectively, were obtained in most branches of the phylogenetic tree (fig. 3*A*; with the exception of three short branches for which there are not enough substitutions to infer reliable ratios). Overall, this pattern of selective pressure along the gene seems to have remained constant during hominoid evolution.

Discussion

The spread of the HERV-W family in the genome of primates began after the divergence of *Catarrhini* and *Platyrrhini*, about 40 MYA (Goodman et al. 1998; Voisset et al. 2000). The identification of the orthologous locus in gibbon (Mallet et al. 2004) indicates that the insertion of ERVWE1 occurred before the divergence between *Hominidae* (human, chimpanzee, gorilla, and orangutan) and *Hylobatidae* (gibbon), i.e., more than 19–25 MYA (Yoder and Yang 2000). Our analyses provide different lines of evidence indicating that in the hominoid lineage, ERVWE1 *env* has been subject to the action of natural selection. First, *env* has the lowest substitution rate of the whole locus. Second, the observed absence of indels and stop codons within *env* orthologous genes, as compared to the expected frequency, demonstrates that evolutionary constraints have operated to maintain the reading frame open. The pattern of selective constraints appears to vary along the ERVWE1 *env* gene. In the 5′ end of the gene the K_A/K_S ratio is relatively high (1.2). This region encodes the amino-terminal subdomain of the Env protein (Surface

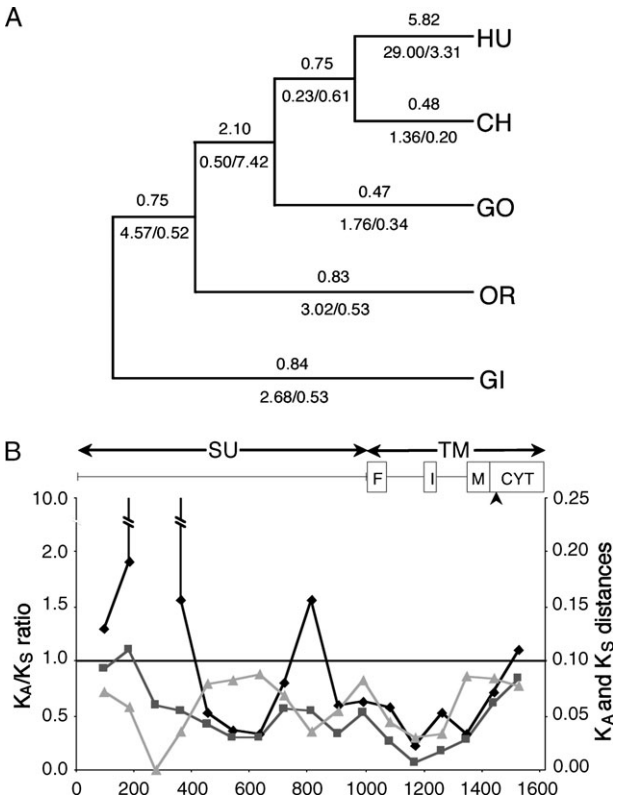

FIG. 3.—K_A/K_S ratio analysis of the *env* gene of the orthologous ERVWE1 loci from five primates. (*A*) Phylogenetic analysis of the K_A/K_S ratio. The species tree is shown with the values of the K_A/K_S ratio of the *env* gene indicated on each branch. The ratios calculated on the whole *env* gene are shown above branches. The ratios calculated on the 5′ and 3′ parts of *env* corresponding to nucleotides 1 to 498 and 499 to 1614, respectively, of gibbon (GI), orangutan (OR), gorilla (GO), chimpanzee (CH), and human (HU) genes are shown under branches. Branches are not drawn to scale. (*B*) Analysis of the K_A/K_S ratio along the *env* gene. Substitutions were computed on a 180-bp sliding window shifting 90 bp along the *env* gene between each step. The K_A and K_S values (in dark and light gray, respectively, and on the right scale) were calculated as the sum of branches of Neighbor-Joining trees. The K_A/K_S ratios are indicated in black and on the left scale. Location of the characteristic features of ERVWE1 *env*: surface (SU, 1–951) and transmembrane (TM, 952–1617) domains; F, fusion peptide; I, immunosuppressive domain; M, membrane anchorage domain; CYT, intracytoplasmic tail; the arrowhead indicats the position of the ERVWE1-specific 12-bp deletion.

Unit) that, in classical retroviruses, is known to be involved in receptor recognition. The high rate of protein evolution in this region might reflect the requirement of coevolution of the endogenous envelope with its two receptors ASCT-1 and ASCT-2 (Blond et al. 2000; Lavillette et al. 2002). It is noticeable that residues 55, 111, and 141 of this region appeared to be ERVWE1 human signatures (Mallet et al. 2004). The last 1,100 bp of the gene include the sequence encoding the carboxy-terminal subdomain of the Env protein (Trans Membrane domain) that contains functional motifs required for fusion, immunosuppression, membrane anchorage, and oligomerization. This region is found to be under purifying selection although it evolved relatively rapidly, as illustrated by the observed K_A/K_S ratio (0.56) versus the average K_A/K_S ratio in human and chimpanzee (about 0.22) (Hellmann et al. 2003).

The systematic comparison of *env* paralogous sequences reveals an *env* ERVWE1-specific 12-bp deletion that results in a four amino acid shortened cytoplasmic tail of the Env transmembrane subunit. We have shown that this deletion is crucial for the envelope fusogenic activity. This suggests that this deletion might have played a major role in the domestication of this retroviral gene. Conversely, the four additional amino acids persisting in all other *env* paralogous pseudogenes may represent a vestige of the intracytoplasmic tail sequence of the infectious ancestor. Thus, the local modification of the ERVWE1 envelope cytoplasmic tail may result from constraints due to the acquired physiological role (reduction of the pathogenic potential) and/or from an adaptation to the host cellular environment (optimization of functions).

The ability to pseudotype retrovirus vectors with a variety of envelope proteins, e.g., modified retroviral envelopes or vesicular stomatitis virus G glycoprotein, has been used for years to significantly broaden the tropism of replication-defective retrovirus vectors (Kafri 2004). Similarly, the in vivo incorporation of ERVWE1 Env into pathogenic retroviruses would be dramatically detrimental to the host by broadening the tropism of the virus. Thus, the ERVWE1 Env cytoplasmic tail may reduce such a possibility as illustrated by the inability of this envelope glycoprotein to be incorporated on Murine Leukemia Virus particles (Blond et al. 2000; An, Xie, and Chen 2001). Conversely, it was observed that ERVWE1 Env was able to pseudotype Human Immunodeficiency Virus (HIV)-based retroviral vectors in vitro (An, Xie, and Chen 2001; Lavillette et al. 2002). A modification of the ERVWE1 Env cytoplasmic tail, consisting of a 53 carboxy-terminal amino acid deletion, induced a 20-fold increase in viral titers of HIV pseudotypes (Lavillette et al. 2002). This suggested that the pseudotyping process, which involved the native ERVWE1 envelope, was inefficient. Interestingly, the carboxy-terminal end of the mutant shortened tail takes place just one residue upstream from the specific ERVWE1 *env* signature, i.e., the site of the 12-bp deletion. Nevertheless, whether these differences in pseudotyping efficiency were sequence-dependent or length-dependent deserves further investigation.

The capacity of infectious retroviruses to fuse virus-cell membranes is dependent on an envelope maturation step consisting of cleavage of the intracytoplasmic tail by the viral protease (Rein et al. 1994). Conservation of this regulatory feature of the Env-mediated fusion for millions of years would have required (1) conservation of an HERV-W active protease and (2) temporal and spatial coordinated expression of *pro* and *env* genes. This does not reflect the HERV-W family portrait, as (1) the ERVWE1 locus expressed in placenta contains a *pro* pseudogene and (2) the human genome did not contained any *pro* HERV-W genes in a favorable translational context (Voisset et al. 2000; Pavlicek et al. 2002). The absence of fusogenic property of the LQMV modified envelope, proposed to mimic the infectious ancestor, could be correlated with the absence of HERV-W protease activity. Hence, one may speculate that the ERVWE1 Env cytoplasmic domain evolved so as to bypass the cleavage requirement (constitutive fusogenic envelope) or to adapt cellular protease(s). In addition, contrary to the proposed involvement of a cytoplasmic helical structure in the mechanism of cell-cell fusion (Taylor and Sanders 2003), the fusogenic cellular envelope exhibited an altered helix, which was restored in the nonfusogenic LQMV mutants. Further investigations will be required to decipher the respective contribution of the cytoplasmic tail maturation and structure in the fusion process mediated by the cellular envelope.

To conclude, we have identified hallmarks of the action of natural selection that demonstrate the selective preservation of the *env* gene of the retrotransposed ERVWE1 proviral locus. Identification of *env* ERVWE1-specific traits suggests mechanisms underlying the process of evolution from a retroviral envelope toward a cellular gene.

Acknowledgments

We thank Adam Eyre-Walker for helpful discussion. We also thank Alessia Ruggieri for constructing the phCMV-ERVWE1-LQMV plasmid. We are grateful to Manolo Gouy from Pôle Bio-Informatique Lyonnais (PBIL, http://pbil.univ-lyon1.fr) for providing the ACNUC algorithm library. B.B. was supported by a doctoral fellowship from the CNRS and bioMérieux. We thank three anonymous referees for their useful comments.

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.

An, D. S., Y. Xie, and I. S. Chen. 2001. Envelope gene of the human endogenous retrovirus HERV-W encodes a functional retrovirus envelope. J. Virol. **75**:3488–3489.

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.

Costas, J. 2002. Characterization of the intragenomic spread of the human endogenous retrovirus family HERV-W. Mol. Biol. Evol. **19**:526–533.

Engelstadter, M., C. J. Buchholz, M. Bobkova, S. Steidl, H. Merget-Millitzer, R. A. Willemsen, J. Stitz, and K. Cichutek. 2001. Targeted gene transfer to lymphocytes using murine leukaemia virus vectors pseudotyped with spleen necrosis virus envelope proteins. Gene Ther. **8**:1202–1206.

Frendo, J. L., D. Olivier, V. Cheynet, J. L. Blond, O. Bouton, M. Vidaud, M. Rabreau, D. Evain-Brion, and F. Mallet. 2003. Direct involvement of HERV-W Env glycoprotein in human trophoblast cell fusion and differentiation. Mol. Cell Biol. **23**:3566–3574.

Galtier, N., M. Gouy, and C. Gautier. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput. Appl. Biosci. **12**:543–548.

Goodman, M., C. A. Porter, J. Czelusniak, S. L. Page, H. Schneider, J. Shoshani, G. Gunnell, and C. P. Groves. 1998. Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. Mol. Phylogenet. Evol. **9**:585–598.

Hellmann, I., S. Zollner, W. Enard, I. Ebersberger, B. Nickel, and S. Paabo. 2003. Selection on human genes as revealed by comparisons to chimpanzee cDNA. Genome Res. **13**:831–837.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.

Kafri, T. 2004. Gene delivery by lentivirus vectors an overview. Methods Mol. Biol **246**:367–390.

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.

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.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

Lavillette, D., M. Marin, A. Ruggieri, F. Mallet, F. L. Cosset, and D. Kabat. 2002. The envelope glycoprotein of human endogenous retrovirus type W uses a divergent family of amino acid transporters/cell surface receptors. J. Virol. **76**:6442–6452.

Li, W. H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36**:96–99.

Mallet, F., O. Bouton, S. Prud'homme, V. Cheynet, G. Oriol, B. Bonnaud, G. Lucotte, L. Duret, and B. Mandrand. 2004. The endogenous retroviral locus ERVWE1 is a bona fide gene involved in hominoid placental physiology. Proc. Natl. Acad. Sci. USA **101**:1731–1736.

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.

Pavlicek, A., J. Paces, D. Elleder, and J. Hejnar. 2002. Processed pseudogenes of human endogenous retroviruses generated by LINEs: their integration, stability, and distribution. Genome Res. **12**:391–399.

Perriere, G., and J. Thioulouse. 1996. On-line tools for sequence retrieval and multivariate statistics in molecular biology. Comput. Appl. Biosci. **12**:63–69.

Perron, H., J. A. 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.

Rein, A., J. Mirro, J. G. Haynes, S. M. Ernst, and K. Nagashima. 1994. Function of the cytoplasmic domain of a retroviral transmembrane protein: p15E-p2E cleavage activates the membrane fusion capability of the murine leukemia virus Env protein. J. Virol. **68**:1773–1781.

Taylor, G. M., and D. A. Sanders. 2003. Structural criteria for regulation of membrane fusion and virion incorporation by the murine leukemia virus TM cytoplasmic domain. Virology **312**:295–305.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

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.

Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431–449.

Yoder, A. D., and Z. Yang. 2000. Estimation of primate speciation dates using local molecular clocks. Mol. Biol. Evol. **17**:1081–1090.

Jennifer Wernegreen, Associate Editor

Accepted June 2, 2004