Molecular Characterization and Placental Expression of HERV-W, a New Human Endogenous Retrovirus Family

JEAN-LUC BLOND,1 FRÉDÉRIC BESÈME,1 LAURENT DURET,2 OLIVIER BOUTON,1 FRÉDÉRIC BEDIN,1 HERVÉ PERRON,1 BERNARD MANDRAND,1 AND FRANÇOIS MALLET1*

Unité Mixte 103 CNRS-bioMérieux, Ecole Normale Supérieure de Lyon, 69364 Lyon, Cédex 07,1 and Laboratoire de Biométrie Génétique et Biologie des Populations, UMR CNRS 5558, Université Claude Bernard-Lyon 1, 69622 Villeurbanne Cedex,2 France

Received 15 July 1998/ Accepted 10 November 1998

The multiple sclerosis-associated retrovirus (MSRV) isolated from plasma of MS patients was found to be phylogenetically and experimentally related to human endogenous retroviruses (HERVs). To characterize the MSRV-related HERV family and to test the hypothesis of a replication-competent HERV, we have investigated the expression of MSRV-related sequences in healthy tissues. The expression of MSRV-related transcripts restricted to the placenta led to the isolation of overlapping cDNA clones from a cDNA library. These cDNAs spanned a 7.6-kb region containing gag, pol, and env genes; RU5 and U3R flanking sequences; a polypurine tract; and a primer binding site (PBS). As this PBS showed similarity to avian retrovirus PBSs used by RNA viruses, this new HERV family was named HERV-W. Several genomic elements were identified, one of them containing a complete HERV-W unit, spanning all cDNA clones. Elements of this multicopy family were not replication competent, as gag and pol open reading frames (ORFs) were interrupted by frameshifts and stop codons. A complete ORF putatively coding for an envelope protein was found both on the HERV-W DNA prototype and within an RU5-env-U3R polyadenylated cDNA clone. Placental expression of 8-, 3.1-, and 1.3-kb transcripts was observed, and a putative splicing strategy was described. The apparently tissue-restricted HERV-W long terminal repeat expression is discussed with respect to physiological and pathological contexts.

A few years ago, the presence of extracellular particles associated with reverse transcriptase (RT) activity was observed in leptomeningeal cells (LMT) (42) and monocyte cultures (44) from patients with multiple sclerosis (MS). Viral RT-associated activity plus electron microscopy examination and the presence in MS patients of antibodies cross-reacting with human immunodeficiency virus type 1 and type 2 RT and human T-cell leukemia virus type 1 CA as well (41) led to the hypothesis of the retroviral nature of these viral particles. By mainly a PCR-based approach, a partial molecular identification of a novel retrovirus isolated from cerebrospinal fluid, plasma, or cell culture supernatant of MS patients was recently obtained (40). Retroviral RNAs with strong similarities to MS-associated retrovirus (MSRV) were also found in plasma of rheumatoid arthritis (RA) patients (17). Furthermore, expression of related mRNA sequences was found in both MS and control brain tissues (29).

All together, these observations might result from three different situations. The extracellular particles may be produced by a replication-competent endogenous provirus. Alternatively, MSRV could represent a virophage-producing exogenous member of an endogenous family, as described for the mouse mammary tumor virus and murine leukemia virus retroviral families of mice (23) and in the Jaagsiekte retroviral family of sheep (37). Lastly, a more complicated process would involve separated defective retroviral entities cooperating via trans-complementation.

MSRV was found to be related to the endogenous retroviral sequence ERV-9 (26), and Southern blot analysis with an MSRV pol probe showed hybridization with a multicopy endogenous family (40). Therefore, we have studied the MSRV-related human endogenous retroviruses (HERVs), which could represent either distant members of the ERV-9 family or an as yet undescribed HERV family. As (i) the HERV family was shown to be a multicopy family (40) and (ii) we focused on the hypothesis that MSRV could be a replication-competent HERV, we followed a strategy based on the characterization of mRNA isolated from a healthy human tissue, by using MSRV- and ERV-9-derived probes. We describe a new multicopy endogenous retroviral family tentatively named HERV-W. All members of the family are apparently not competent for replication. However, a functional U3 promoter and an open reading frame (ORF) putatively coding for an envelope protein were identified. HERV-W mRNA expression is restricted to placental tissue in Northern blot analysis. The strategy of transcription is addressed, and the promoter tissue specificity is discussed.

MATERIALS AND METHODS

Probe definition and labelling. The 678-bp Ppol-MSRV probe was RT-PCR amplified from RNA extracted from particles collected from the supernatant of synovial cultures of an RA patient. These particles were gradient centrifuged as previously described (43), and RNA was extracted by the method of Chomczynski and Sacchi (10) and DNase treated as previously described (54). A first round of RT-PCR amplification was performed with primers AP2942 (5'-AGGAGTAAAGGAAACCCAACGGAC-3') (forward) and AP2152 (5'-TAA GAGTTGACAACAAGTGC-3') (reverse). A nested PCR was then performed with primers AP2522 (5'-TCAGGATAGCCCCATCTAT-3') (forward) and AP2510 (5'-AACCCCTTGCACACTACATCT-3') (reverse). The 649-bp Ppol-ERV-9 probe was PCR amplified from a B-lymphocyte cDNA library of an MS patient with primers AP2522 and AP2510 mentioned above. Discrimination between the Ppol-MSRV- and Ppol-ERV-9-cloned PCR fragments was performed with an enzyme-linked oligosorbent assay (ELIOSA) (35). Capture and detection oligonucleotide probes used in this nonisotopic sandwich technique are depicted in Fig. 1A.

The Pgg-LB19 (536 bp), Ppro-E (364 bp), and Penv-C15 (591 bp) probes.
corresponded to PCR fragments of the longest ORFs of LB19, E, and C15 clones, respectively. LB19 and E clones (37) were obtained from RNA extracted from gradient-concentrated particles of B-lymphocyte cultures of an MS patient or from the plaque of an MS patient, respectively. The C15 clone was obtained from RNA extracted from the pellet of centrifuged supernatants of syoviocyte cultures of an RA patient.

Fifty nanograms of each probe was radioactively labelled by random priming with [32P]dCTP and used for a DNA labelling kit to hybridize to the manufacturer’s protocol with [α-32P]dCTP (3,000 Ci/mmol). The specific activity of probes was >5.0 × 10^6 cpm/μg.

Plasmid DNA library screening. 5 rapid amplification of cDNA ends (RACE), and transcriptome analysis. A plasmid cDNA library constructed in the lambda vector was used (human placenta 5’-Stratch Plus cDNA library from Clontech Laboratories, Inc., Palo Alto, Calif.). Screening of the library was performed according to the manufacturer’s instructions with the Pmol-MSRV probe for initial screening, and probes were derived from the resulting clones for additional screening. Briefly, for each screening the plating density was about 5 × 10^5 PFU for two 220- by 220-mm plates. After growth, plaques were transferred onto a nylon membrane (Hybond N+; Amersham) and hybridized with probes radioactively labelled as described above (10^6 cpm of hybridization solution). Prehybridization was performed at 42°C in a freshly prepared solution: 50% formamide, 5× SSPE (20× SSPE is 3 M NaCl, 0.2 M NaH2PO4, and 20 mM EDTA, pH 7.4), 5× Denhardt’s solution (100% Denhardt’s solution is 10 g of Ficoll 400, 10 g of polyvinylpyrrolidone, and 10 g of bovine serum albumin for 500 mL in 0.5 M NaCl) in order to evaluate chlorophenolacetyltransferase (CAT) activity by the use of the CAT enzyme system (Experiment). For this purpose, the liquid scintillation counting protocol was followed as recommended by the manufacturer. Briefly, cell extracts were obtained by several freeze-thaw cycles and a subsequent heating at 60°C for 10 min. Then, 10 μL of this extract was added to 0.25 M Tris-HCl, pH 8.0 (125 μL, final volume), containing 0.15 μCi of [3H]chloramphenicol (NEN, Boston, Mass.) and 5 μg of n-butyl coenzyme A, and incubated for 4 h at 37°C. Five hundred microfilters of a mix of xylene isomers was added, and the butyrylated chlorophenolacetate present in this organic phase was extracted twice with 100 μL of 0.5 M Tris-HCl. This 200 μL was mixed with 5 mL of scintillation fluid and counted in a liquid scintillation counter.

In vitro transcription-translation. In vitro transcription-translation was performed with a kit from Promega Biotech (TNT coupled reticulocyte lysate systems). It was used strictly according to the manufacturer’s instructions. Briefly, DNA templates were obtained by PCR using specific primers designed according to the promoter (5’-CATATAAGCAGCTATACGGGAGACCATGGCCCTCCC TTATCAT-3’). The radiolabelled amino acid was [35S]methionine, provided by Amersham. Glycosylation was studied by the translation in vitro in the presence of [35S]methionine of a pancreatic microsomal fraction (also provided by Amersham). Post-translational analysis was performed by SDS-polyacrylamide gel electrophoresis, and the gel was exposed to X-ray films (Kodak), usually for 12 h.

Sequence analysis. Individual cDNA clones were used for phylogenetic analysis, and the cDNA products were filtered for complex sequences (Alu-like and microsatellites) with XBLAST (12). A BLASTN (or BLASTX) (2) query for GenBank (release 101, June 1997) was performed, and sequence matches with scores greater than 200 were retained. Homologous fragments were extracted from GenBank and aligned with our cDNA clones. Alignments were performed semimanually, with the SEAVIEW multiple alignment editor (16) and the Clustal W (52). Dialog (36), or MABIOS (1) multiple alignment program. Regions corresponding to putative coding regions were determined with BLASTX and BLASTN, a visual program allowing evaluation of protein and a nucleic acid sequence from the alignment. We selected in the alignment three regions from long terminal repeat (LTR), Pol, and Env that are conserved in all members of the family of human endogenous viruses that we have identified. Phylogenetic trees were computed from these three regions with phylogen (16) with the neighbor-joining method (49). Five hundred bootstrap replicates were performed to evaluate the robustness of the results.

Promoter prediction was performed with PROSCAN v1.7 (47) and SIGSCAN v4.05 (46). A searching for protein sequence similarities was performed with LFASTA (39). The prediction of leader peptide was made with GeneWorks 2.5 software. Prediction of the fusion domain and the transmembrane region was performed with the PHD package (48).

Nucleotide sequence accession numbers. The sequences described in this paper have been submitted to GenBank under the following accession numbers. The probe numbers are as follows: Pmol-MSRV (AF072494), Pol-ERV-9 (AF072495), PPol-ERV (AF072496), Ppro-E (AF072497), and Penv-C15 (AF072498). Nucleotide numbers are 500 bp of the placental cDNA (AF072499). cDNA sequence is: 5’-CATAATAGCAGCTATACGGGAGACCATGGCCCTCCC TTATCAT-3’.

RESULTS

Design of probes. Pol probes were designed to evaluate the expression of MSRV-related sequences in healthy human tissues. The choice of Pol probes relied on the observations that (i) the previously described Pol region of MSRV was related to the ERV-9 HERV and (ii) ERV-9 sequences were occasionally detected with MSRV in plasma or cell culture supernatants of MS patients (40). Because (i) the 2,304-bp MSRV Pol region (40) was a consensus sequence resulting from overlapping clones obtained from different sources and (ii) the MSRV and ERV-9 cdetected sequences were only 120 bp long, we chose to amplify two continuous 650-bp MSRV- and ERV-9-related fragments, defined by the same borders and including the 120-bp cdetected region. MSRV- and ERV-9-related Pol regions were RT-PCR amplified as described in Materials and Methods, and mixed resulting fragments were cloned and checked (data not shown) with the ELOSA procedure (Fig.

Downloaded from http://jvi.asm.org/ on September 30, 2017 by guest
Characterization and Expression of HERV-W

A BLASTN query on the EST (expressed sequence tag) database, with placental cDNA clones (see below) derived from the above probes, showed hundreds of related transcripts in the human tissues, most in the opposite direction. With a search criterion of 90% identity over 100 nucleotides (nt) or more, these sequences were found predominantly expressed in the placenta (53%), but also in fetal liver-spleen (28%). The relative abundance of HERV-W sequences among placenta ESTs was found to be three times higher than that among fetal liver-spleen ESTs.

Isolation and sequencing of overlapping cDNA fragments. A placental cDNA library was screened with Ppol-MSRV, Pgag-LB19, and Penv-C15 probes. Probes derived from the resulting clones were then used for additional screening. The clones obtained with all three screenings exhibited the same characteristics. All together, the overlapping cDNA clones covered 7.6 kb. Nine overlapping clones obtained with the Ppol-MSRV screening are shown in Fig. 2A. The percentages of similarity among the overlapping parts of these Ppol-MSRV-derived cDNA clones, a polyadenylated clone (cl.C4C5) probed with Penv-C15, and the probes used for the screening as well are presented in Fig. 2B. The similarity between cDNA clones ranged from 97 to 100% with the exception of the cl.7A16 clone, which was about 90%. The similarity between the cDNA clones and probes derived from MSRV-related extracellular RNA sequences ranged from 85 to 94%. Similarly, this percentage was near 90% with the MSRV Pol region previously described (40) but was lower than 70% with the ERV-9 Pol probe.

With BLASTX, sequences of the different fragments showed extensive homology with Gag, Pol, and Env retroviral proteins. However, gag and pol genes did not support ORFs corresponding to functional proteins. An ORF putatively coding for a retroviral envelope protein was observed on the 3’ cDNA clones (cl.24.4, cl.C4C5, and cl.PH74). This Env ORF was totally contained in the cl.PH74 cDNA clone. Flanking untranslated regions (UTRs) were observed upstream and downstream from gag and env genes, respectively, containing repeated sequences (R) as described forLTRs. The length of R was estimated as about 120 bp by comparison of the 5’ end of the 5’ most cDNA clone (cl.6a2) and the 3’ end of a polyadenylated clone (cl.PH74), both clones showing 98% similarity in their 5’ UTR overlap. Three clones contained 3’ UTRs followed by a poly(A) tail (cl.P5ST, cl.PH74, and cl.C4C5).

Phylogenetic analysis. A phylogenetic analysis was performed at the nucleic acid level on 11 different subregions of the region spanned by the cDNAs and at the protein level on two different Env subregions, as described in Materials and Methods. All the trees had the same topology regardless of the region addressed. Notably, the human BAC clone RG083M05 was represented in all trees. These results were illustrated at the nucleic acid level within the most conserved regions of the LTR (R-US) and the pol gene (RT catalytic domain) between the obtained sequences and ERV-9 and RTVL-H (Fig. 3). The trees clearly showed that those sequences described a new family related to, but distinct from, ERV-9 and even more distant from RTVL-H as highlighted by the bootstrap analysis.

![Diagram](http://jvi.asm.org/)
Those sequences were found on several chromosomes, including 5, 7, 14, 16, 21, 22, and X, with an apparent high concentration of LTR in chromosome X.

Comparison at the protein level between the most conserved regions of Env retroviral proteins (immunosuppressive to transmembrane [TM] domain) resulted in trees showing a sequence distribution similar to the one observed at the nucleic acid level (Fig. 3). The two proximal coding sequences corresponded to the translated env ORF of RG083M05 and a truncated ORF located downstream from the GTP-binding protein RAB7 on the same mRNA. This group of sequences was more closely related to simian type D and reticuloendotheliosis avian retroviruses than to type C mammalian retroviruses.

Genomic complexity of the family. In order to characterize the complexity of this new endogenous retroviral family, we used several probes spanning the entire genome. Southern blot analysis of human DNA digested with three different restriction enzymes showed complex patterns with an apparent increase of complexity from env to gag and protease regions (Fig. 4A). A more precise determination of the copy number of each region was done by a dot blot analysis of serial dilutions of the plasmids that contained the corresponding probes (data not shown). The calculated number of copies per haploid genome (Fig. 4A) emphasized the heterogeneous distribution of the retroviral subregions. A 6- and 20-fold increase of Gag and protease signals, respectively, versus the Env signal confirmed the gradation suspected in Southern blot analysis.

To characterize some genomic sequences at the molecular level, a BLASTN query on several databases was performed, with the placental cDNA clones. The four most significant sequence hits are depicted in Fig. 4B. They consisted of the human BAC clone RG083M05 from 7q21-7q22, the human BAC378 corresponding to Homo sapiens T-cell receptor alpha delta locus whose chromosomal location is 14q11-12, the H. sapiens chromosome 21q22.3 cosmid Q11M15, and the human DNA sequence from cosmid U134E6 on chromosome Xq22 containing NIK-like and thyroxin-binding globulin precursor. Repeated sequences were found located at both ends of these clones. All the cDNA sequences fell entirely within the clone RG083M05 (10.2 kb) with a similarity of about 98 to 99%, except the clone cl.7A16 (90%). However, in addition to gag, pol, and env genes, RG083M05 exhibited a 2-kb insert located just downstream from the 5’ UTR, this insert being found also in clones BAC378 and Q11M15. These two clones and a third one (U134E6) presented a strictly conserved 2.3-kb deletion just upstream from the 3’ UTR. No clone contained all three gag, pol, and env ORFs. The clone RG083M05 exhibited a 538-amino-acid ORF corresponding to a full-length envelope. The cosmid Q11M15 contained two large contiguous ORFs of 413 (frame 1) and 305 (frame +1) aa corresponding to a truncated Pol polyprotein. All together, these data suggested a quite complex family from which the RG083M05 clone could represent a genomic prototype.

RNA elements: LTR, polyuridine tract, and primer binding site (PBS) characterization. In order to identify a potential promoter downstream from the cl.PH74 putative env ORF, a predictive analysis was performed. Proscan 1.7 did not predict any obvious promoter, but Signalscan 4.05 indicated the presence of putative transcription factor binding sites, including a CCAAT box and a TATA box. The 2,364- to 2,720-nt corresponding region of the cl.PH74 clone (2,764 bp) was PCR amplified and subcloned into pCAT reporter vector. The CAT assay showed a significant expression level (Fig. 5A). A promoter activity was also found in BeWo human choriocarcinoma cells and Jurkat human T cells (data not shown).

In order to characterize the retroviral LTR, the 5’ and 3’ UTRs of cDNA experimental clones were aligned with the 5’ and 3’ repeated sequences of the most proximal human DNA sequence, RG083M05. Nonretroviral flanking cellular sequences allowed us to delineate a 783-bp putative LTR (Fig. 5B). A duplicated 4-bp sequence (CAAC) was found spanning 5’ and 3’ LTRs. The characteristic TG-CA base pairs were found juxtaposed with nonretroviral sequences at each end of the integrated provirus. The 5’ end of U3 was confirmed by the presence, on the cLP74 and cl.C4C5 clones, of a polyuridine tract located immediately upstream from U3 and 46 bp downstream from the env stop codon. Determination of the U3-R junction was quite complex. Several possible cap sites were found by 5’ RACE experiments on placental mRNA (data not shown), but surprisingly, all these sites were found included within the cl.6A2 clone (Fig. 5B). More surprisingly, the predicted TATA box fell within the R sequence defined according
to the cl.6A2 clone. To simplify the analysis, a similar procedure was applied to a single U3-R region, with total RNA extracted from pCAT3-3LTR-transfected HeLa cells (data not shown). It indicated a potential cap site downstream from the TATA box (Fig. 5B). The 3' end of R was precisely defined according to the presence of a poly(A) tail on cl.PH74 and cl.C4C5 clones. The polyadenylation signal was found to be 455 bp long. The 3' end of gag was defined 13 bp upstream of U5. The U5 LTR remaining after the R region, 13 bp upstream of U5, was not unambiguously defined. However, the cl.PH74 clone confirmed the presence, on the cl.6A2 (gag), cl.PH74, and cl.24.4 (env) clones, of a putative PBS located 4 bp downstream from the TATA box (Fig. 5B). The 3' end of U5 was confirmed by the presence on the cl.6A2 (gag), cl.PH74, and cl.24.4 (env) clones, of a putative PBS located 4 bp downstream from the TATA box (Fig. 5B). This PBS showed extensive homology with the avian retrovirus PBS (53) used by tRNATrp for minus-strand DNA synthesis (34). As the tRNA \textsuperscript{env} was not described for humans and this PBS sequence did not correspond to any other tRNA, we tentatively named this new family HERV-W.

**Transcription and putative splicing strategy.** The expression of the HERV-W family in placenta was analyzed by Northern blotting (Fig. 6A) with a set of probes spread over the 7.6-kb region containing the cDNA clones (Fig. 6B) and consisting of two U5 probes, U5(g) and U5(e), derived from an RUS-gag cDNA clone (cl.6A2) and a U5-env cDNA clone (cl.24.4), respectively, as well as gag, pro, pol, and env probes as described in Materials and Methods, and a U3-R probe, U3(e), derived from an env--U3-R--poly(A) \textsuperscript{+} clone (cl.C4C5). A simple pattern consisting of three bands at 8, 3.1, and 1.3 kb was observed. Probes from the gag, pro, and pol genes hybridized only to the 8-kb transcript which was also revealed by env, U5, and U3 probes and thus may correspond to a putative full-length transcript. Probe from the env gene detected both the full-length transcript and an abundant 3.1-kb transcript which also hybridized to the U5 and U3 probes. Thus, this 3.1-kb transcript exhibited the characteristic features of a singly spliced subgenomic retroviral env mRNA. The U5 and U3-R probes revealed an additional transcript of 1.3 kb, lacking detectable gag, pro, pol, and env sequences, which may result from alternative splicing or transcription from defective HERV-W provirus.

No trivial correspondence between the observed transcripts and the isolated cDNA clones could be found, as the 5' end of R was not unambiguously defined. However, the cl.PH74 clone containing an env ORF and repeats at both ends was a good candidate for the 3.1-kb subgenomic mRNA. Likewise, the cl.PH7 clone which contained U5 and U3 sequences at the 5' and 3' ends, respectively, but no structural gene, might represent a large spliced subgenomic mRNA. In addition, if we assumed (i) that the cap site location was as defined by the 3' LTR transfection experiment and (ii) the region spanned by
repeats at both ends of the genome is stated. The presence of gaps between the four sequences and the consensus sequence (cDNAs) are assignments are indicated in brackets. The percentages of similarity (without query. Shown are human BAC clone RG083M05 (gb AC000064), human cDNA clones) with the four sequences showing the higher scores with BLASTN schematic representation of the 7.6-kb region defined with the overlapping bottom. Un., undigested. (B) Alignment of the placental cDNA clones (cDNAs; from dot blot analysis, as well as confidence intervals, are also indicated at the bottom for each restriction enzyme. The absolute copy numbers deduced Ppro-E, and Penv-C15. The number of bands detected on the blot is indicated at the top for each restriction enzyme. The absolute copy numbers deduced from splicing events or indicate expression of defective HERV-W proviruses. This strategy permitted us to identify the 10.2-kb DNA sequence of the RG083M05 clone was per-

Northern blot, the alignment of the experimental clones with genomic, and small mRNAs deduced from the cDNA clones. whose sizes are remarkably close to those of genomic, sub-
genomic, and small mRNAs deduced from the cDNA clones. whose sizes are remarkably close to those of genomic, sub-
genomic, and small mRNAs deduced from the cDNA clones. whose sizes are remarkably close to those of genomic, sub-

the cDNAs reflects a genomic RNA organization as suggested by the Northern blot analysis, the lengths of genomic, sub-
genomic, and small mRNAs deduced from the cDNA clones would be 7,541, 2,812, and 1,148 nt, consistent with the 8-, 3.1-, and 1.3-kb observed transcripts, respectively. Although no band larger than 8 kb was observed in the Northern blot, the alignment of the experimental clones with the 10.2-kb DNA sequence of the RG083M05 clone was performed in order to understand whether these RNA transcripts result from splicing events or indicate expression of defective HERV-W proviruses. This strategy permitted us to identify several apparently well-conserved splice donor and acceptor sites (Fig. 6B). The comparison of the 5' cDNA clones cl.44.4 and cl.6A2 with the RG083M05 sequence identified a splice donor (DS1) and a splice acceptor (AS1) site, strictly flanking the 2,075-bp insertion previously observed (Fig. 4). They were found on RG083M05, just 564 and 2,640 nt, respectively, downstream from R, as defined above. The absence of a major splice donor site in the 5' cDNA clones cl.44.4 and cl.6A2 suggested that the 8-kb apparent genomic RNA could result from a splicing strategy. The comparison of cl.PH74 and cl.PH7 clones

with the RG083M05 sequence identified two putative splice acceptor sites, AS2 and AS3, located 7,338 and 9,001 bp downstream from R, respectively (Fig. 6B). AS2 was also found in the cl.Pi5T clone, which contained, in addition, a second putative donor site (DS2), 7,545 bp downstream from R and 207 bp downstream from the env acceptor site but upstream from the env ATG codon. The occurrence of DS1/AS1, DS1/AS2, and DS1/AS3 splicing events on a putative RG083M05 precursor RNA would lead to 7,491-, 2,793-, and 1,130-nt transcripts whose sizes are remarkably close to those of genomic, sub-
genomic, and small mRNAs deduced from the cDNA clones. 

Coding capacities. On the cl.PH74 clone, the initiation codon of the long ORF putatively encoding the envelope protein was found 227 bp downstream from the putative DS1/AS2 splice junction. It was the first ATG downstream from this junction, although not the first from the 5' end of the sub-
genomic RNA as five ATGs preceding five small ORFs (less than 27 aa) were situated ahead. Nevertheless, this ATG was in a relatively favorable context (CCCCATGG) although slightly different from the known (A/G)CCATGG favorable context for translation (24, 25). Figure 7A presents the amino acid sequence of a putatively functional envelope protein. This sequence was derived from three clones: cl.PH74 presented a 5' LTR, a splice junction, a coding sequence of 538 aa, and a 3' LTR poly(A)”; cl.24.4 presented a 5' LTR, a splice junction, and a coding sequence of 410 aa; and cl.C4C5 presented a coding sequence of 242 aa, a 3' LTR, and poly(A)”. This ORF exhibited the characteristic features of the precursor polypeptide of retroviral envelope proteins. It presented a leader peptide at the amino terminus and a carboxy-terminal hydropho-

bic segment which could anchor the protein in the membrane. A furin cleavage site (RNKR) separated the two characteristic subdomains consisting of the surface protein (SU) and the transmembrane protein (TM). The TM contained, in addition to the membrane-spanning segment, a hydrophobic fusion domain and a putative immunosuppressive region homologous to the immunosuppressive p15E retroviral peptide conserved among murine, feline, and human retroviruses (11). The SU and TM regions contained seven and one potential glycosyla-
tion site, respectively. Furthermore, these potential glycolysa-
tion sites are conserved among the three considered clones and in the RG083M05 clone, which might code for the same protein too. An in vitro transcription-translation assay has been performed on this env gene. The results are shown in Fig. 7B. The measured molecular mass of the precursor is 60 kDa, in agreement with the calculated molecular mass of 59,565 Da of the Env ORF based on its amino acid composition. Glycosyl-
ation of the Env precursor was observed, consistent with the prediction of carbohydrate addition sites. The measured molecular mass of the glycosylated protein is 80 kDa. 

Two initiation codons potentially directing the synthesis of 52-aa (ORF1) and 48-aa (ORF2) peptides were found 22 and 95 bp downstream from the putative AS3 splice acceptor site, respectively. Both ATGs were not in a favorable context. ORF1 consisted of the carboxy-terminal part of Env, and ORF2 was translated in a different but overlapping frame. No obvious homology was found by using BLAST query. However, with an LFASTA query on the restricted Retroviridae sub-
database, ORF1 and ORF2 showed about 35% identity with Rex from primate and human T-lymphotropic virus and Tat from simian immunodeficiency virus, respectively (data not shown). Although unusual, the presence of such regulatory proteins in HERVs has been described elsewhere for HERV-K10 (33).
FIG. 5. Definition of a putative LTR. (A) The 2,364- to 2,720-nt region of the cl.PH74 clone, identified as a putative promoter region with Signalscan 4.05, was PCR amplified and subcloned into pCAT3 reporter vector. The resulting pCAT3-3\textsuperscript{9}LTR plasmid and pCAT3 control plasmid were used to transfect HeLa cells prior to CAT activity determination. Cells, HeLa cell extract; pCAT3 and pCAT3-3\textsuperscript{9}LTR, extract of HeLa cells transfected with control plasmid and promoter plasmid, respectively.

(B) Alignment of 5\textsuperscript{9} and 3\textsuperscript{9} UTRs of placental experimental clones with the 5\textsuperscript{9}(5-RG-28000-28872) and 3\textsuperscript{9}(3-RG-37500-38314) repeated sequences of the 28,000- to 38,314-nt fragment of human DNA sequence RG083M05. The end of the 3\textsuperscript{9}-most ORF is indicated (env orf). The tandemly repeated CAAC flanking sequences are doubly underlined on DNA sequences. The 783-bp LTR consensus sequence is positioned at the bottom. The polypurine tract (PPT) upstream from the 5\textsuperscript{9} end of the LTR and the tRNA\textsubscript{Trp} PBS downstream from the 3\textsuperscript{9} end of LTR are indicated. U3, R, and U5 subparts are indicated, except for the U3-R junction. Transcription factor sites determined with Signalscan 4.05 are underlined and labelled from I to VI. The sites V and VI may correspond to CCAAT box and TATA box, respectively. The 3\textsuperscript{9} region of clone cl.PH74 used in the CAT assay is underlined on the cl.PH74 sequence. \(\text{p}\) and \(\text{œ}\), potential cap sites obtained in 5\textsuperscript{9}RACE experiments with human placental poly(A)\textsuperscript{1} RNA and total RNA extracted from pCAT-3\textsuperscript{9}LTR-transfected HeLa cells, respectively. [polyA], polyadenylation signal.
Here we report the molecular characterization of HERV-W, a new multicopy family of HERVs whose expression in healthy tissues seemed restricted to the placenta. The phylogenetic trees within the Pol region showed that the HERV-W family is related to ERV-9 and RTVL-H families and thus belongs to the class I endogenous retroviruses (5). Phylogenetic analysis of the env ORF showed that it was closer to simian type D and avian reticuloendotheliosis retroviruses than to murine type C retroviruses. The homologies within the pol and env genes with the murine type C and simian type D retroviruses, respectively, suggest a chimeric genome structure as described for baboon endogenous virus (22). Based on the size criteria, such a chimerism seemed to exist within the LTR: the 247-nt U3 and the 79- to 81-nt R elements were comparable to avian or type D retrovirus U3 and mammalian type C R elements, respectively, although the 410- to 455-nt U5 element remained unclassified as unusually long (57). The bush-like topology that is observed in the phylogenetic trees for the HERV-W and ERV-9 sequences suggests that in both families most elements were fixed in the germ line during a relatively short period of evolution and probably derive from one or a few active elements. The phylogenetic tree, supported by high bootstrap values, shows that ERV-9 and HERV-W families derive from two independent bursts of insertions. Thus, the active element(s) at the origin of the HERV-W family is distinct from the one(s) from which the ERV-9 family is derived. Moreover, HERV-W PBS is predicted to use tRNA\text{\textsuperscript{Arg}}, whereas ERV-9 probably uses tRNA\text{\textsuperscript{Arg}}. Finally, members of the HERV-W family are expressed in the placenta, whereas we have not detected ERV-9 RNAs in this tissue. However, by RNase protection assay, the detection of protected fragments smaller than the expected one suggested that ERV-9-related sequences could be expressed at a low level in placenta (26).

The persistence of flanking duplicated sequences generated during the integration process suggests either that a selection process persists within the LTR or that the integration occurred recently. The 5' and 3' LTRs of RG083M05 HERV-W DNA prototype showed 4% divergence and 6% gaps. This is comparable to the 0.2 to 4.3% divergence described for HERV-K, but smaller than the 5 to 12% divergence described for a number of cloned proviruses (5). Given an evolution rate of 3.5 × 10^{-9} per site per year (30), the 4% divergence corresponds to a relatively recent integration event 6 million years ago. The HERV-W distribution in different primates will be studied to determine the first introduction into the germ line.

Like other HERVs, most (if not all) current members of the HERV-W family are defective for replication, due to mutations that disrupt one or more of the gag, pol, and env ORFs (38, 58). The observation of more gag-pro than env sequences in the human genome may reflect the loss of the env gene. Several genomic clones seemed to reflect such a loss, which appears to be a common feature among endogenous elements (5). However, a complete env ORF was observed on the cl.PH74 placental cDNA clone and the genomic RG083M05 clone which was nevertheless defective for replication. These results are not compatible with the hypothesis that particles in cell culture supernatants of MS patients (42, 44) or the presence of specifically packaged mRNA in plasma of MS and RA patients (17, 40) may result from a replication-competent HERV. However, although no replication-competent genome has been observed in the databases, it cannot be excluded that a nonidentified one exists. Furthermore, no HERV single genetic unit has yet been shown experimentally to be a source of retroviral particles (58), even in the HERV-K family, which corresponds to the HTDV particles (6, 32). On the other hand, a trans-complementation process between distinct but individually defective loci cannot be excluded. Indeed, queries on nucleic acid databases showed that some large regions were conservatively and confirmed a widespread distribution on numerous chromosomes. To address these two hypotheses, a new approach will be developed to detect potential coding sequences, irrespective of the existence of a putative single genetic element. A procedure based on probe hybridization cou-
pleed with in vitro transcription-translation will be applied to isolated human chromosomes in heterokaryon somatic hybrid cells.

Although HERV-W is a quite complex family at the genomic level, a relatively simple pattern of mRNA expression was observed by Northern blot analysis, consisting of three major bands of 8, 3.1, and 1.3 kb. This pattern resembles the one of lentiviruses or oncoviruses such as mouse mammary tumor virus and human T-cell leukemia virus (45). It is also similar to the one described for ERV-9 in undifferentiated embryonal carcinoma cells (28) and the HTDV/HERV-K family in teratocarcinoma cell lines (32), in which three to four transcripts were derived from a single locus by alternative usage of splicing signals. Whether the observed transcripts resulted from a splicing strategy of one or several genetic entities or originated from different defective entities remains unresolved. Due to (i) the overall behavior of U5, gag, pro, pol, env, and U3 probes in Northern blot analysis and (ii) the identification of conserved splice sites on the RG083M05 DNA prototype and the cDNA clones, it is probable that the three observed bands represent genomic, subgenomic, and large spliced mRNAs. The situation resulting from the proposed splicing strategy was quite unusual as the genomic mRNA would result from a splicing event. Furthermore, no 9.6-kb precursor mRNA was detectable by Northern blotting, which may reflect a very short half-life of such a transcript. A query of the EST database with the 2-kb insert of the RG083M05 clone showed only smaller internal fragments, suggesting either a complex splicing strategy or the presence of shorter unidentified genomes. The existence of an alternative splicing strategy was confirmed by the isolation of clone cl.Pi5T, for which no trivial corresponding transcript was revealed on the Northern blot; thus, this clone may represent a poorly expressed mRNA. Nevertheless, we cannot exclude that some if not most of the placental cDNA clones, although highly homologous, were obtained from different genetic entities. The most divergent clone, Pol cl.7A16, may reflect such a situation. The contradiction between the 5’ end of the cl.6A2 clone and the positioning of the predicted TATA box may also result from such a context. This clone may be derived from transcription driven by a nonretroviral promoter upstream from the complete or altered LTR and an extension through the R polyadenylation site as described for HERV-R (ERV-3) (20). Taken as a whole, this suggests that the complex HERV-W family observed at the DNA level could include well-conserved genomic elements (subset) coexpressed in the placenta.

According to the results of the Northern blot and the RNA dot blot analyses, a significant LTR functionality among healthy tissues seemed restricted to the placenta. This interpretation was supported by the reproducibility of LTR tags found in other fetal, adult, and tumoral tissues, most in the opposite direction. This may reflect expression driven by solitary LTRs or nonretroviral promoters. Interestingly, the 85 to 94% similarities observed between the placental cDNA clones and the probes derived from extracellular RNA isolated in a pathological context compared with the 90 to 100% similarities obtained between overlapping placental cDNAs indicate that at least some of the probes were derived from transcriptional units distinct from those expressed in the placenta. This interpretation was supported by the reproducible isolation of env interrupted ORFs from extracellular mRNA in MS and RA patients (22b) versus an env-containing ORF from the placental cDNA library screening. This suggests a placenta-restricted expression of an HERV-W subset versus a different or probably broader HERV-W expression in a pathological context (or in other tissues). In agreement with this, the expression in different tissues at different times of different subsets of one given family of elements was described, e.g., the expression of HTDV/HERV-K (27), ERV-9 (31), and HERV-R (13, 19) mRNAs differs quantitatively and qualitatively, including the length of the probe and the hybridization stringency. Furthermore, the EST query showed a significant concentration of HERV-W expression in fetal liver-spleen, in addition to placenta. As the RNA dot blots were negative for both fetal liver and fetal spleen, one may suspect the influence of the differentiation stage or a low level of expression. Partial
The finding in placental cDNA clones of a large ORF coding for a putatively functional envelope protein suggests some physiological functions related to pregnancy, such as a role in the creation of the syncytiotrophoblast layer of the placenta (7, 55) and a role in suppressing the maternal immune response against the fetal alloantigens (7, 55, 56) as suggested for the envelope protein of HERV-R. In pathological situations, an HERV Env might protect against exogenous retroviral infections by a receptor interference mechanism (4) or alter the local cellular immunity via, for example, a superantigen-like region. This was recently proposed for type I diabetes (insulin-dependent diabetes mellitus) (14). Conversely, HERV-W expression may represent only a consequence of physiological or pathological events. Whatever the situation is, the finding of HERV-W sequences both in a physiological tolerogenic immune context such as pregnancy and in autoimmune diseases such as MS and RA deserves further investigations.

ACKNOWLEDGMENTS

We thank F. L. Cosset (UCB-Lyon I, Lyon, France), C. Guillou (Erasmus University, Rotterdam, The Netherlands), and Doran Spencer for critical reading of the manuscript. We also thank C. Gautier and all the staff of the Biometry Laboratory (UCB-Lyon I, Lyon, France) for helpful discussions and assistance with nucleic acid analysis and F. Penin and C. Gourgeon (IBCP, Lyon, France) for helpful discussions and assistance with nucleic acid analysis. We thank F. L. Cosset (UCB-Lyon I, Lyon, France), C. Guillon and A. Vahlne for providing C15 and LB19 and E clones, respectively. We also thank C. Gautier (Bio-Informatique Lyonnais) facilities (http://pbil.univ-lyon1.fr). We thank F. L. Cosset (UCB-Lyon I, Lyon, France), C. Guillon, H. Perron, G. Favre, B. Mandonrand, R. Juvin, F. Marcel, F. Beseme, F. Bedin, F. Mallet, B. Mougin, F. Joncour, J. M. Seigneurin, and X. Phelip for critical reading of the manuscript. We also thank C. Gautier, N. Gouy, C. Gautier for providing C15 and LB19 and E clones, respectively. We are grateful to Florence Komurian-Pradel and Glaucia Parhanos-Bacala for providing C15 and LB19 and E clones, respectively.

REFERENCES


11. Komurian-Pradel, F. Personal communication.


is specifically associated with a contagious lung cancer of sheep. J. Virol. 70: 1618–1623.


