J Mol Evol (1999) 48:675–683

JOURNAL OF MOLECULAR EVOLUTION

Comparison Between Two Human Endogenous Retrovirus (HERV)-Rich Regions Within the Major Histocompatibility Complex

Jerzy K. Kulski,[1] Silvana Gaudieri,[1,2] Hidetoshi Inoko,[3] Roger L. Dawkins[1]

[1] Centre for Molecular Immunology and Instrumentation, University of Western Australia, Faculty of Medicine and Dentistry, P.O. Box 5100, Canning Vale, 6155, Western Australia, Australia
[2] Centre for Information Biology, National Institute of Genetics, Mishima, 411-8540, Japan
[3] Department of Genetic Information, Division of Molecular Life Science, Tokai University School of Medicine, Bohseidai, Isehara, Kanagawa 259-11, Japan

Received: 17 September 1998 / Accepted: 8 January 1999

Abstract. Sixteen human endogenous retrovirus (HERV) sequences were detected within 656 kb of genomic sequence obtained from the alpha- and beta-block of the *class I* region of the major histocompatibility complex (MHC). The HERVs were identified and characterized as family members of HERV-16 (11 copies), HERV-L (1 copy), HERV-I (2 copies), HERV-K91 (1 copy), and HARLEQUIN (1 copy) by sequence comparison using CENSOR or Repeat Masker, BLAST searches, and dot plots. The 11 copies of HERV-16 arose as products of duplication of genomic segments containing *HLA class I* (*HLAcI*) and *PERB11* (*MIC*) genes *inter alia*, whereas the other five HERVs arose after duplication probably as a consequence of single insertion events or translocations. HERV-L and HERV-I are located between the duplicated genes *PERB11.2* (*MICB*) and *PERB11.1* (*MICA*), and *HLA-B* and *HLA-C*, respectively, whereas HERV-K91 and HARLEQUIN are located telomeric of *HLA-C*. A highly fragmented copy of HERV-I was also found telomeric of *PERB11.4*. Structural analysis of open reading frames (ORFs) revealed the absence of intact coding sequence within the putative *gag*, *pol*, and *env* gene regions of all the HERVs with the exception of HERV-K91, which had two large ORFs within the region of the putative *protease* and *pol* genes. In addition, the 5′-LTR of HERV-L contained a 2.5-kb element that was AT-rich and large ORFs with putative amino acid sequences rich in tyrosines and isoleucines. HERV-I, HARLEQUIN, and at least four copies of HERV-16 appear to have been receptors for the insertion of other retrotransposons including Alu elements and fragments of L1 and THE1. Examination of flanking sequences suggests that HERV-I and HERV-L had occurred by insertion into ancient L1 fragments. This study has revealed that the alpha- and beta-block region within the MHC is rich in HERV sequences occurring at a much higher ratio (10 to 1) than normally observed in the human genome. These HERV sequences will therefore enhance further studies on disease associations and differences between human haplotypes and primates and their role in the evolution of *class I* genes in the MHC.

Key words: Human endogenous retrovirus — Duplications — Multicopy genes — Major histocompatibility complex

Correspondence to: Jerzy K. Kulski; e-mail: jkulski@cyllene.uwa.edu.au

Introduction

Although human endogenous retroviruses (HERVs) comprise more than 1% of the human genome (Smit, 1996), information on their distribution and diversity within the major histocompatibility complex (MHC) is lacking and poorly understood. Different HERV families vary markedly in copy number (1 to 1000) in the human

genome, with some families (HERV-H, -L, -E, -K) having maintained a higher amplification rate than others (HERV-R, HRES-1) (see review by Wilkinson et al. 1994). In this regard, HERVs with a moderate to high copy number may have an important evolutionary role in creating genomic diversity by providing sites for recombination or translocation events and contributing to inter- and intraspecies variation through deletions and insertions (Erickson et al. 1992). For example, a HERV-K (C4) in the *C4* region of the MHC has contributed to interlocus and interallelic length heterogeneity of the complement gene locus (*C4*) by integration into intron 9 of *C4A* genes and some *C4B* genes (Dangel et al. 1995). The identification and characterization of HERV families within the *class I* region of the MHC (Gaudieri et al. 1996; Kulski et al. 1997; Kulski and Dawkins 1999) are of particular interest because this region has numerous polymorphic genes that have been associated with autoimmune diseases such as myasthenia gravis (Degli-Esposti et al. 1992), Behcet's, ankylosing spondylitis, and psoriasis vulgaris (Tiwari and Terasaki 1985), as well as with disease susceptibility after viral infection (Cameron et al. 1992). Previous analysis of genomic sequence within the class I region of the MHC revealed that retroelements such as Alus and LINEs (L1 fragments) have had a major influence on the expansion and diversity of duplication products carrying genes such as *HLA-B* and *HLA-C* (Kulski et al. 1997) and *PERB11.2* (*MICB*) and *PERB11.1* (*MICA*) (Gaudieri et al. 1997). In addition, the *class I* region of the MHC has deletions/insertions (>30 kb) in some human haplotypes (Venditti and Chorney 1992; Watanabe et al. 1997) and expansions/contractions in primates (Leelayuwat et al. 1993).

Recently, genomic sequences have been obtained within the *class I* region of the MHC between *PERB11.2* and *HLA-C* (Mizuki et al. 1997; Shiina et al. 1998) and between *HLA-J* and *HLA-F* (DDBJ/EMBL/GenBank accession number AF055066) within the beta- and alpha-blocks, respectively. We examined these genomic sequences for the presence of HERVs that may have an association with disease and a potential role in the genomic organization, diversity, duplications, and rearrangements within the *class I* region of the MHC. This report presents the results of our analysis of 16 HERV sequences identified within the alpha- and beta-blocks using BLAST and Repeat Masker homology searches, dot plots, and open reading frame (ORF) analysis. The genomic sequence flanking the ends of the HERV elements was also examined to identify their site of integration.

Materials and Methods

Three hundred thirty-seven kilobases of the DNA sequence from *PERB11.2* to a telomeric region approximately 90 kb beyond *HLA-C* within the beta-block (Mizuki et al. 1997; Shiina et al. 1998) was submitted to DDBJ/EMBL/GenBank as accession numbers AB000882 and D84394. Genomic sequences of 319 kb including an *HLAcI* gene cluster from *HLA-J* to *HLA-F* within the alpha-block and 246 kb including the *HLA-HFE* gene of the MHC were obtained from DDBJ/EMBL/GenBank as accession numbers AF055066 and U91328, respectively.

The genomic sequence within the alpha- and beta-block was searched for the presence of endogenous retroviruses and repeat elements by comparison to sequences available in Repbase using the CENSOR WWW server (Jurka et al. 1996) or Repeat Masker2 WWW server [http//ftp.genome.washington.edu/cgi-bin/RepeatMasker (AFA Smit and P Green, unpublished data)]. The identified endogenous retroviral elements and closely related sequences from DDBJ/EMBL/GenBank were compared using the program Compare and Dot Plot from the GCG package, v8 (GCG, WI). The program Matrix from Gene Jockey II (BioSoft) was used for some additional dot-plot analysis of HERV sequences. ORFs were analyzed using the software program DNA Strider (Marck 1988).

Results

Location of HERVs Near the HLAcI *and* PERB11 *Genes Within the MHC*

The locations of HERV sequences and the *HLAcI* and *PERB11* genes identified within the alpha- and beta-block of the MHC are shown in Fig. 1A. The sequences in the beta-block are shown in the reverse orientation to those in the alpha-block. Both blocks were found to harbor numerous copies of HERV sequences representing at least five families, HERV-L, HERV-16, HERV-I, HERV-K91, and a HARLEQUIN-like sequence. Eleven of the 16 HERV sequences were related to the HERV-16 family and were found in the region of the multicopy *P5* gene family (Vernet et al. 1993b), confirming the recent finding that HERV-16 is identical in sequence to *P5* sequences and at least 40% similar to HERV-L sequences (Kulski and Dawkins 1999). Fully intact or fragments of HERV-16 sequences were found upstream of the *HLAcI* genes (*HLA-G* and *-F*) and pseudogenes (*HLA-X, -17, -J, -80, -70, -16, -H, -90,* and *-75*). No HERV-16 sequences were found near *HLA-A, -B,* or *-C* coding genes, although a LTR16C sequence was found centromeric of *HLA-B.* Overall, the percentage of genomic sequence due to HERV sequences was 10% in both the alpha (33.6 of 319 kb)- and the beta (33.4 of 337 kb)-blocks.

The dot plots in Figs. 1B and C show the region of duplication between the *PERB11* genomic segments (D1a and D1b) and between the *HLA* genomic segments (D2a and D2b), respectively. The dot plot in Fig. 1B shows that both HERV-16 elements are located telomeric of the *PERB11* genes (and centromeric of *HLA-X* and *-17*) and are part of the *PERB11* duplication products D1a and D1b. In comparison, the dot plot in Fig. 1C shows the location of HERV-I and HERV-K91, which appear to have been inserted after duplication of the *HLA-B* and *HLA-C* genomic segments. HARLEQUIN was located outside the *HLA* duplicated segments, about

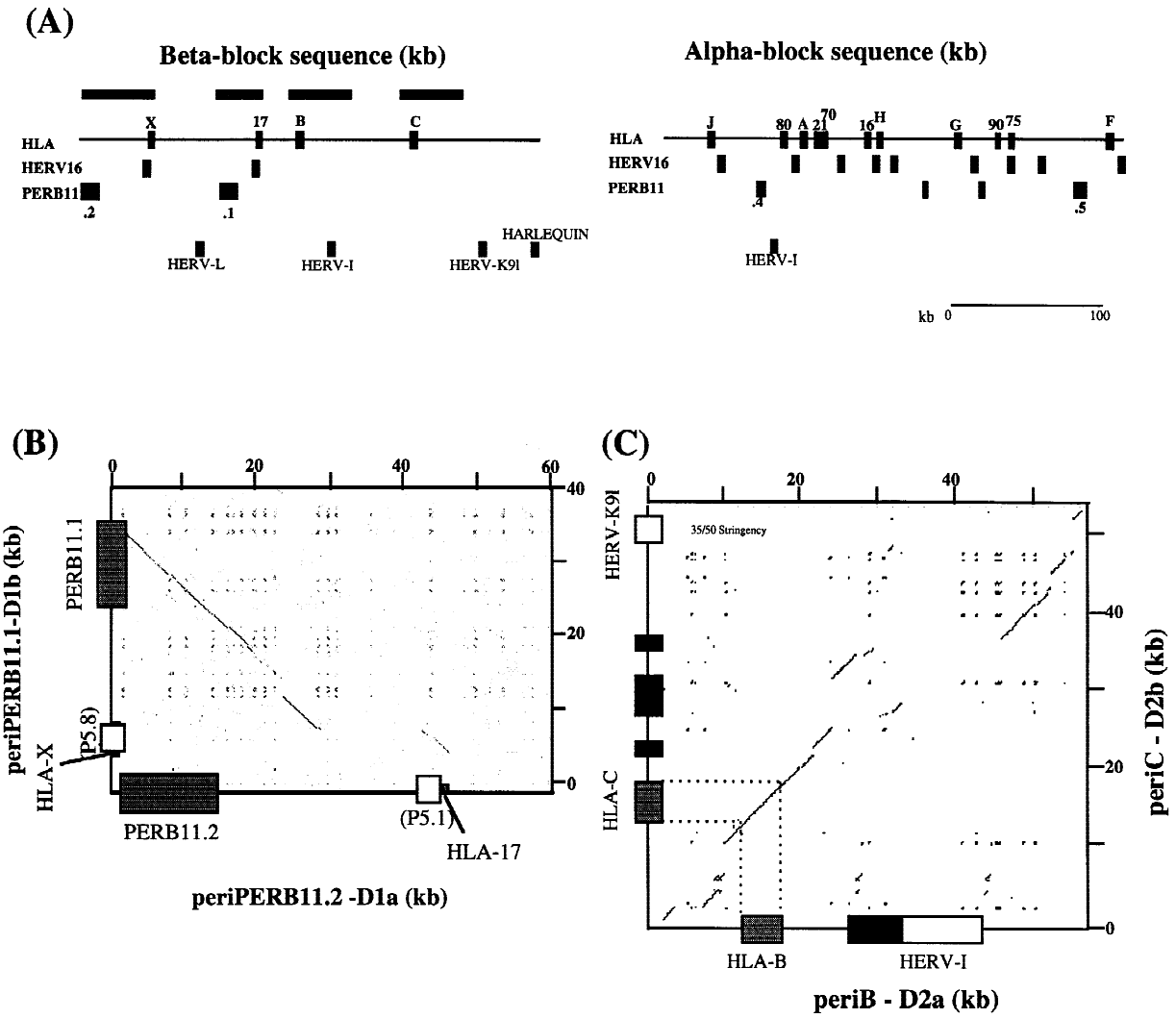

Fig. 1. **A** Map of the location of HERVs in relation to HLA *class I* and *PERB11* genes. **B, C** Dot plots of the genomic sequence within the beta-block of the MHC. (B) The dot plot shows the location of *PERB11* genes (*shaded boxes*), *HLAcI* gene fragments (*labeled-lines*), and HERV-16 (P5) sequences (*open boxes*) within the duplicated segments, D1a (*X* axis) and D1b (*Y* axis). (C) The dot plot shows the location of *HLA-B* and *-C* genes (*shaded boxes*), large L1 fragments (*black boxes*), and the insertion sites of HERV-I and HERV-K91 (*open boxes*) within the duplicated segments, D2a (*X* axis) and D2b (*Y* axis). The dot plots in B and C were produced at a stringency of 35 in a window setting of 50 nucleotides.

40 kb telomeric of HERV-K91 and 78 kb telomeric of *HLA-C*.

Multicopies of the HERV-16 Sequence in the Class I *Region of the MHC*

There were differences between the 11 copies of HERV-16 sequence with respect to genomic organization and indels, which can be attributed in part to the insertion of retroelements (Alu and THE1B fragment) and deletions within internal sequences and flanking LTRs (Fig. 2). Five of the 11 HERV-16 sequences are flanked by a 5′ and a 3′ LTR16B sequence and range in size between 4460 and 5473 bp. Of the remaining six HERV-16 sequences, four were fragmented (1043–3421 bp) and two were solitary LTRs (380 bp). There was no recognizable HERV-16 sequence near *HLA-A* except for a (CAT)n(TAA)n(TAAA)n repeat that was also identified within the HERV-16 fragmented sequence near *HLA-80*. The internal HERV-16 sequence, contained between the flanking LTRs, has about 60% similarity to HERV-L, mainly within the *pol* gene region, and only slight identity to parts of *gag* or LTRs. An assumed primer binding sequence (5′-TGGCTTCAGGAGTGGTCC-3′) for leucine-tRNA was found in HERV-16 two nucleotides downstream from the 3′ end of the 5′-LTR16b sequence, which has identity with 15 of 18 nucleotides of HERV-L (Cordonnier et al. 1995).

HERV-L, -I, -K9l, and HARLEQUIN Sequences

A dot-plot comparison between the full-length retroviral element [HERV-L (PERB11)] in the MHC and a human

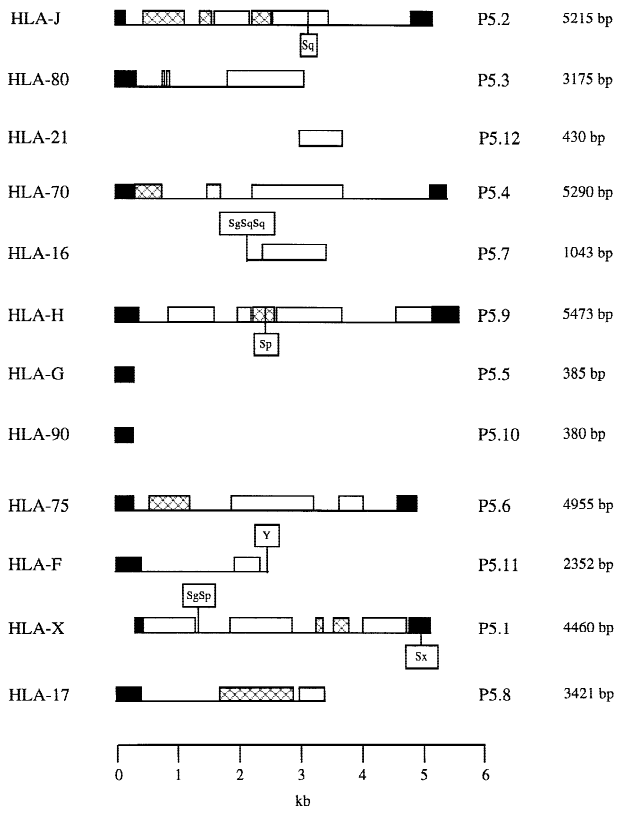

Fig. 2. Schematic representation of the indels and structural organization of 11 HERV-16-related sequences detected in the *class I* region of the MHC. The genomic organizations of the sequences are shown in a 3′-to-5′ orientation, with the 3′-LTR on the *left-hand side* and the 5′-LTR on the *right-hand side. Black boxes* are LTR16B, *open boxes* have sequence identity to HERV-16, and the *remaining boxes* have sequence identity to HERV-L or MuERV-L. *Vertical columns* represent simple repeats (CATA)n(TAAA)n. The sequence represented by the *horizontal line* was not identified by Repeat Masker. The positions of Alu insertions are indicated by a *vertical line* and labeled with a boxed-subfamily designation. The *HLAcI* genes associated with each HERV sequence are listed at the *left.* The HERV-16 designation based on the P5 subfamily designations and the length of each HERV-16 sequence are listed at the *right.* The relative lengths of the HERV sequences are indicated by the bar (kb) at the *bottom.*

endogenous retrovirus-like element, HERV-L (X89211) (Cordonnier et al. 1995), confirmed that the sequence within the MHC is closely related to the HERV-L family (Fig. 3A). The main difference between HERV-L (PERB11) and HERV-L (X89211) is a 2650-kb AT-rich sequence within the 5′-LTR of HERV-L (PERB11). Apart from the AT-rich region, HERV-L (PERB11) is 6557 bp in length and has genomic features similar to those of HERV-L (Y12713) (Cordonnier et al. 1995) and a mouse homologue MuERV-L (Benit et al. 1997a). The features shared between the members of the ERV-L family include a 5′- and a 3′-LTR, the presence of sequence homologues for *gag*, *pol*, and *dUTPase*, tRNA primer-binding site complementary to the 3′ end of a mouse leucine tRNA downstream of the 5′-LTR, a polypurine track close to the 3′-LTR, the absence of an *env* gene, and an inverted repeat TGT and ACA at the 5′ and 3′ ends of the LTRs, respectively. However, unlike MuERV-L, which contains long ORFs corresponding to *gag*, *pol*, and *dUTPase*, both HERV-L (X89211) and HERV-L (PERB11) contain many stop codons (Figs. 3B–D).

The AT-rich region within the 5′-LTR of HERV-L (PERB11) contains six major ORFs in all three coding phases of the 5′-to-3′ sequence and many stop codons in all three coding phases of the 3′-to-5′ complementary sequence (Fig. 3D). The predicted translation of the AT-rich region revealed that the longest ORFs are composed of 373 residues in phase 1 (ORF 5), 666 residues in phase 2 (ORF 3), and 821 residues in phase 3 (ORF 1). Phase 1 includes another two ORFs of 198 residues (ORF 4) and 232 residues (ORF 6), and phase 2 has an ORF of 127 residues (ORF 2). The predicted translation products of all six ORFs of the AT-rich element contain a start codon ATG (methionine) and are rich in tyrosines (>34%) and isoleucines (>33%). However, no similarity to any known proteins was found in BLASTp searches of the SwissProt database, and no exons/genes were predicted in the 5′-LTR of HERV-L (PERB11) using either the GRAIL or the GENSCAN program. In addition, the AT-rich element contains AT repeats in runs of three to five dinucleotides interrupted mostly by TTC, TCT, TGT, TTA, and CCT nucleotide triplets with variable periodicity and patterns of complexity. Both HERV-L (X89211) and MuERV-L have only a short run of repeats that are located at an equivalent position to the AT-rich region in HERV-L (PERB11). The low sequence similarity between the 5′-LTR of HERV-L (PERB11) and MuERV-L probably reflects a species difference.

The endogenous retrovirus located between *HLA-B* and *HLA-C* and telomeric of *PERB11.4* (Fig. 1) was most closely related to the HERV-I (RTLV-I) family, with members also found in the haptoglobin region of humans, apes and Old World monkeys (Maeda and Kim 1990; Erickson et al. 1992). The common features shared between HERV-I (HLAB) and the haptoglobin members of the HERV-I family include a sequence length of approximately 10 kb, a 5′- and a 3′-LTR, the presence of sequence homologues for *gag*, *pol*, and *env* and a tRNA primer-binding site complementary to the 3′ end of a mouse isoleucine tRNA downstream of the 5′-LTR. The HERV-I in the MHC and the haptoglobin cluster contain many stop codons and putative coding regions of short length within the location of the *gag*, *pol*, and *env* genes. The region between the *env* gene and the 3′-LTR within the HERV-I (HLAB) sequence is interrupted by a composite retroposon of 1378 bp including two full-length Alu S insertions and two fragments (738 and 24 bp) that are part of the repeat unit SVA (SINE, VNTRs, and Alu) previously found near the *C4* genes of the MHC (Shen et al. 1994). In comparison, the HERV-I in the alpha-block is highly fragmented and interrupted by simple repeats and Alu Y and L1 fragments.

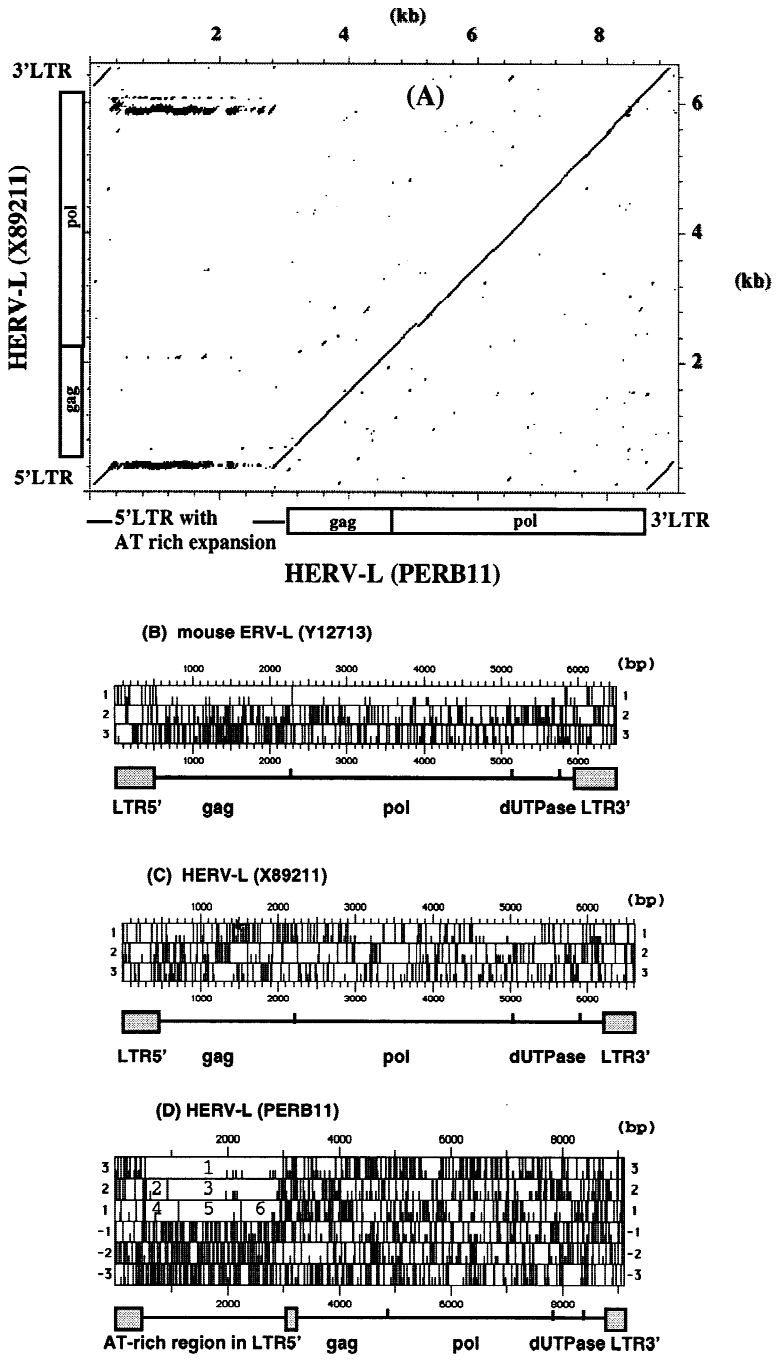

Fig. 3. DNA analysis of ERV-L. **A** Dot-plot comparison between nucleotide sequences of HERV-L (PERB11) along the *X* axis and HERV-L (X89211) along the *Y* axis. The stringency of the plots was 45 in a window setting of 50 nucleotides. The locations of the 5′- and 3′-LTR and the *gag* and *pol* genes of HERV-L are indicated in the *margins*. **B, C, D** The organization of ORFs of MuERV-L (Y12713), HERV-L (X89211), and HERV-L (PERB11) sequences, respectively. Only the reading frames in the 5′-to-3′ nucleotide sequence are shown in B and C. All possible reading frames in the positive (5′-to-3′) and negative (3′-to-5′) strand of HERV-L (PERB11) are shown in D. *Full-length vertical bars* in each phase of the coding strand indicate stop codons (TAA, TAG, and TGA), and *short vertical bars* are ATG start codons. The locations of the corresponding genes and LTRs (*boxes*) are indicated below the ORF and nucleotide position. The six major ORFs in the AT-rich element of HERV-L (PERB11) are numbered 1 to 6 (D).

HERV-K91 was found about 31 kb telomeric of the *HLA-C* gene but still within a duplicated genomic segment (Figs. 1A and C). This HERV was 4118 nucleotides in length and is flanked by an LTR previously described as a medium interspersed repeat designated MER9 (Kaplan et al. 1991) or as a short interspersed repeat designated the PRE element (Ricke et al. 1992). A sequence relationship was observed between HERV-K10 and HERV-K91 by dot-plot analysis, with the strongest homology occurring within the *protease* gene region. The assumed primer binding sites of HERV-K91, HERV-K (C4), and HERV-K10 were similar and complementary to the 3′ end of the lysine 1,2-tRNA sequence. The LTR, *gag*, and *env* sequences of HERV-K10 were substantially different from comparable regions of HERV-K91. Structural analysis for ORFs within the HERV-K91 (HLAC) sequence showed the presence of a number of large ORFs in all six potential coding phases. Based on the length of intact ORFs, HERV-K91 appears to be the most recent HERV insertion in the *class I* region. The longest ORF occurs within a region overlapping the *protease* and *pol* genes. In comparison to HERV-K10, the *protease/pol* gene of HERV-K91 was interrupted by multiple stop codons, which, nevertheless, might allow the translation of 847 nucleotides from position 2237 to position 3184 of the sequence. Com-

parison of seven members of this family obtained from GenBank (data not shown) revealed large deletions within the *pol* and *gag* region of the HERV-K91 sequence from the MHC.

A 4838-bp sequence located 78 kb telomeric of *HLA-C* was identified as a mosaic sequence resembling HARLEQUIN sequences that are thought to have originated from nonhomologous shuffling of RNA genomes from different retroviruses when packed together in the same virion (Kapitonov and Jurka, annotation in CENSOR). The HARLEQUIN-like sequence in the MHC was composed of 5′-LTR2 (414 bp), MER41 and MER4 internal sequences (802 bp), HERV-I (775 bp), and HERV-E (2733 bp), interrupted by an unknown sequence (114 bp) and terminating in a 3′-TA repeat (41 bp) and a full-length Alu Sq. A dot-plot comparison between nucleotide sequences of HERV-E (M10976) (Repaske et al. 1985) and the MHC HARLEQUIN revealed major differences in the regions between *gag* and *pol* and between *pol* and *env* and within the *pol* gene regions (data not shown). A homologous region between LTR2 and the *gag* region of HARLEQUIN and HERV-E (M10976) contained a primer binding site for glut-tRNA with identity in 16 of 18 nucleotides. Another HARLEQUIN-like sequence was found within a 246-kb genomic sequence (U91328) that is outside the centrometric class I region and approximately 6 kb from the *HLA-HFE* gene of the MHC (Ruddy et al. 1997).

Integration of HERV-L and HERV-I Within LINE-1 Retroelements

Genomic sequence analysis both upstream and downstream of the retroviral elements revealed that (a) HERV-L (PERB11) was inserted into a 4037-bp L1 retroelement fragment (L1MPA2) between nucleotide positions 3818 and 3917, and (b) HERV-I was inserted into a 1333-bp L1 fragment between nucleotide positions 3273 and 3279. Reconstructions of integration of HERV-L and HERV-I are shown in Figs. 4A and B, respectively. It is possible that integration may not all have occurred at the site of the MHC, as some of the earlier events may have started within other genomic locations, followed at a later stage by a translocation of the HERVs and associated flanking L1 fragments to the present location.

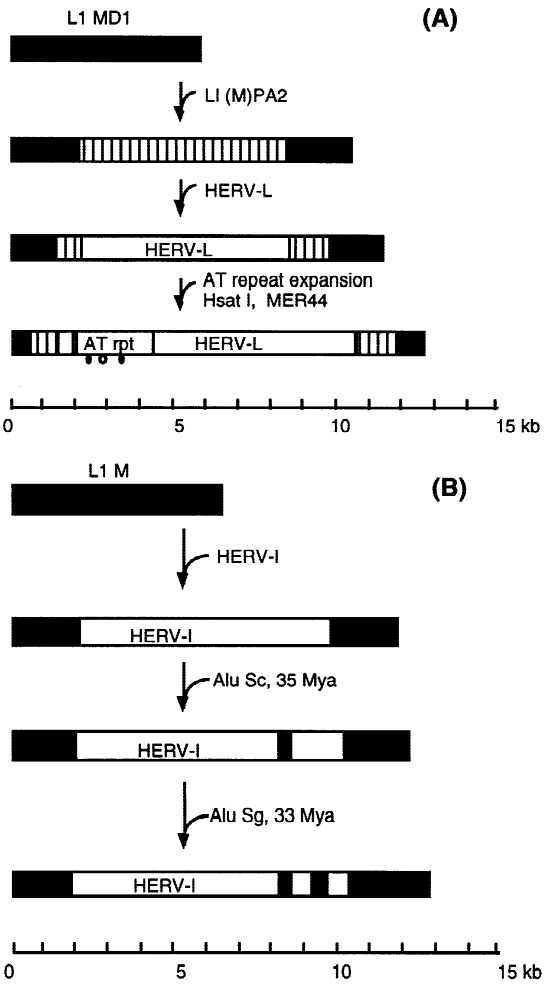

Fig. 4. Reconstruction of events for the integration of (**A**) HERV-L (PERB11) and (**B**) HERV-I (HLAB) into L1 retroelements. The classification of retroelements and the approximate dating of their insertions are described in the text. The scheme shown in A proposes that the region of HERV-L (PERB11) was initially spanned by a single mammalian-specific L1 (L1MD1) repeat (*hatched block*) about 100 mya, which was a receptor for the insertion of a primate-specific L1 (L1PA2) repeat (*block with vertical bars*), which in turn was the receptor for the insertion of HERV-L (*open block* labeled HERV-L) after the emergence of primates 65 to 85 mya. The insertion of HERV-L (PERB11) occurred between nucleotide positions 3818 and 3917 of L1(M)PA2. The HERV-L (PERB11) element was then a target site for the AT repeat expansion and insertion of Hsat1 and MER44 (denoted as *circles* in the AT rpt block). The insertion events have led to eventual fragmentation and deletions within the mammalian- and primate-specific L1 elements but with an overall expansion of about 8 kb within a region that was originally a L1 repeat element of about 5 kb. The scheme shown in B proposes that the region of HERV-I (HLAB) was initially spanned by an L1 repeat (*hatched block*) that acted as a receptor for insertion of HERV-I after the emergence of primates. This insertion occurred between nucleotide position 3273 and 3279 of the L1 sequence and historically prior to the insertion of the Alu elements (*filled blocks*) 33 to 35 mya. This insertion event has also resulted in fragmentation and deletions within the L1 element, but with an overall expansion of the genomic region by about 8 kb.

Discussion

At least four HERV families, HERV-KC4, HERV-I, HERV-L, and HERV-16 (Dangel et al. 1995; Gaudieri et al. 1996; Kulski et al. 1997; Kulski and Dawkins 1999), have been noted or described within the MHC. In this study, we found and characterized 16 HERV sequences in 656 kb of genomic sequence within the alpha- and beta-blocks of the MHC. Three of these HERV sequences shared features with other family members (HERV-L and HERV-I) that were cloned and sequenced (Maeda and Kim 1990; Erickson et al. 1992; Cordonnier et al. 1995), 11 belonged to the HERV-16 family, and 1

each belonged to a HARLEQUIN and HERV-K91 family whose origin and distribution in the human genome were unknown or known only in limited detail. Therefore, we have identified a HERV-rich region within the MHC where about 10% of the genomic sequence (each >300 kb) was due to HERV sequences, in comparison to the 1% normally observed in the human genome (Smit 1996). For example, Repeat Masker detected only one HERV-like sequence, located about 6 kb from the *HLA-HFE* locus in a 246-kb genomic sequence within the MHC (Ruddy et al. 1997).

The characteristics and location of the 16 HERV sequences described in this study clearly reflect their importance and role in the duplication, expansion, and diversity of genomic segments containing at least two multicopy gene families (*HLAcI* and *PERB11*) within the *class I* region of the MHC. HERV-16 sequences appear to have been copied as part of serial segmental duplications containing *HLAcI* and *PERB11* genes, whereas the other HERVs probably originated postduplication in the *class I* region from translocations or primary integration following direct infection of germ cells.

Examination of genomic flanking sequences revealed that HERV-L (PERB11) and HERV-I (HLAB) insertions had occurred into L1 sequences. Subtyping of the flanking L1 fragments (Smit et al. 1995) indicated that the HERV-L (PERB11) region was initially spanned by a single mammalian-specific L1 (L1MD1) repeat about 100 mya, followed by the insertion of a primate-specific L1 (L1PA2) and then the insertion of HERV-L after the emergence of primates about 65 to 80 mya. A similar scheme for HERV-I (HLAB) insertion into L1 sequences could be inferred from the fragments flanking this retroviral element. Furthermore, HERV-L (PERB11), HERV-I (HLAB), and at least four of the HERV-16 sequences appear to have been target sites for the expansion of repeats such as the AT-rich region and the insertion of other retroelements including Alu and fragments of SVA, THE1, and L1 repeats. The present sequence analysis of HERVs and their flanking sequences provides a further example of how retrotransposition is a major driving force for diversity in the MHC.

Alu subfamily members (Kapitonov and Jurka 1996) have been used as molecular clocks both to trace and to estimate the age of genomic duplications of *class I* and *class II* genes (Mnukova-Fajdelova et al. 1994; Satta et al. 1996; Gaudieri et al. 1997; Kulski et al. 1997). Both HERV-16 and HERV-I sequences within the *class I* region were found to contain Alu S elements estimated to have been fixed into the primate genome 31 to 44 mya (Kapitonov and Jurka 1996). *PERB11.1* (*MICA*) and *PERB11.2* (*MICB*) genomic segments also contain paralogous insertions of Alu S (Gaudieri et al. 1997), suggesting that they and associated HERV-16 sequences arose within the beta-block from the same duplication event when Alu S elements were still mobile.

No HERV-16 sequences were found to be associated with the *HLA-B* and *HLA-C* genomic segments. However, a HERV-I insertion was found in the *HLA-B* genomic segment immediately telomeric of the *HLA-B* gene. The two Alu S elements found within the two HERV-I sequences in the haptoglobin cluster (Maeda and Kim 1990) were in different locations and orientations from the Alu S elements in the HERV-I (HLAB) of the MHC. In addition, since HERVIa (haptoglobin) contains AluSp and HERVIc (haptoglobin) has AluSg and AluSp, the Alu insertions into HERV-I probably occurred as separate events approximately 33–37 mya. It is likely that HERV-I amplification occurred at about the time of the split between New and Old World monkeys, which is slightly earlier than the estimation of 25 mya by Shih et al. (1991). Previously, we used Alu subfamily typing to estimate that the *HLA-B* and *HLA-C* genomic segments had occurred before the fixation of the Alu S subfamily and the emergence of New World Monkeys (Kulski et al. 1997). On the basis of Alu and HERV typing, the *HLA-B* and *HLA-C* genes appear to have evolved prior to the insertion of HERV-I into the *HLA-B* genomic segment and before the origin of the *PERB11.1* and *PERB11.2* gene loci as a separate duplication event. In comparison to the other HERVs within the *class I* region, the HERV-K91 sequence contains a number of large ORFs, suggesting that it was a relatively recent insertion into the genomic segment containing the *HLA-C* gene.

The *class I* region of the MHC has been associated with more than 100 diseases including susceptibility to autoimmune diseases such as insulin-dependent diabetes mellitus and myasthenia gravis (Degli-Esposti et al. 1992) and rapid progression to acquired immunodeficiency syndrome (AIDS) following HIV-1 infection (Cameron et al. 1992). These diseases appear to be related to polymorphisms and various indels within this region. For example, chimpanzees have a large deletion/transposition in this region of class I (Leelayuwat et al. 1993), possibly including HERV-L and a copy of HERV-16 and *PERB11*. In contrast to humans, chimpanzees actively infected with HIV can mount an antibody response but generally they do not progress to AIDS (Heeney et al. 1994). Thus, one or another of the genes or HERVs deleted from the region of *class I* in the chimpanzee might influence the progression of AIDS in individuals infected with HIV-1. While HERVs within the MHC appear to be defective elements incapable of producing extracellular, infectious virions, they may still be transcriptionally and translationally active. They may influence immunity against exogenous retroviral infections and disease progression by inducing antibody responses to viral gene products, modifying the regulation of nearby cellular genes associated with immune responses, and competing for viral receptors on cell surfaces or for intracellular protein interactions (Lower et al. 1996; Ur-

novitz and Murphy 1996). In this regard, HERV-16 (P5.1) has been reported to express mRNA in certain lymphoid cells and tissues (Vernet et al. 1993b). Moreover, the HERV-16 mRNA may be in an antisense orientation to retroviral *pol* mRNA (Kulski and Dawkins 1999) and, consequently, have antiviral function. HERVs within the MHC exhibit polymorphism and therefore could be associated with disease progression of HIV-infected individuals and other autoimmune diseases that have been linked to the *class I* region of the MHC. Finally, the 16 HERVs described in this paper provide additional genetic markers within the MHC that should allow further analysis of differences between haplotypes and primates, as well as their role in disease and evolution of class I genes in the MHC.

Acknowledgments. Silvana Gaudieri is the recipient of a fellowship from the Japan Society for the Promotion of Science. This work was supported by grants from the National Health and Medical Research Council and the Immunogenetics Research Foundation. We thank Dr Annalise Martin for her help with the analysis and preparation of Fig. 2 and Jennie Hui and Natalie Longman for secretarial assistance. This is manuscript number 9805 for the Centre of Molecular Immunology and Instrumentation at The University of Western Australia, Nedlands, Western Australia.

References

Benit L, Parseval ND, Casella JF, Callebaut I, Cordonnier A, Heidmann T (1997) Cloning of a new murine endogenous retrovirus, MuERV-L, with strong similarity to the human HERV-L element and with a gag coding sequence closely related to the Fv1 restriction gene. J Virol 71:5652–5657

Cameron PU, Mallal SA, French MAH, et al. (1992) Central MHC genes between HLA-B and complement C4 confer risk for HIV-1 disease progression. In: Tsuji K, Aizawa M, Sasazuki T (eds) HLA 1991, Vol 2. Oxford University Press, New York, pp 544–547

Cordonnier A, Casella JF, Heidmann T (1995) Isolation of novel human endogenous retrovirus-like elements with foamy virus-related *pol* sequence. J Virol 69:5890–5897

Dangel A, Baker B, Mendoza A, Yu C (1995) Complement component C4 gene intron 9 as a phylogenetic marker for primates: Long terminal repeats of the endogenous retrovirus ERV-K(C4) are a molecular clock of evolution. Immunogenetics 42:41–52

Degli-Esposti MA, Andreas A, Christiansen FT, Schalke B, Albert E, Dawkins RL (1992) An approach to the localization of the susceptibility genes for generalised myasthenia gravis by mapping recombinant ancestral haplotypes. Immunogenetics 35:355–364

Erickson LM, Kim HS, Maeda N (1992) Junctions between genes in the haptoglobin gene cluster of primates. Genomics 14:948–958

Gaudieri S, Kulski JK, Dawkins RL (1996) The central region of the major histocompatibility complex contains a sequence with similarity to the pol gene of Moloney retroviruses. Immunogenetics 44:157–158

Gaudieri S, Giles K, Kulski J, Dawkins R (1997) Duplication and polymorphism in the MHC: Alu generated diversity and polymorphism within the PERB11 gene family. Hereditas 127:37–46

Heeney JL, Van Els C, De Vries P, et al. (1994) Major histocompatibility complex class I-associated vaccine protection from simian immunodeficiency virus-infected peripheral blood cells. J Exp Med 180:769–774

Jurka J, Klonowski P, Dagman V, Pelton P (1996) CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. Comp Chem 20:119–121

Kapitonov V, Jurka J (1996) The age of Alu subfamilies. J Mol Evol 42:59–65

Kaplan DJ, Jurka J, Solus JF, Duncan CH. (1991) Medium reiteration frequency repetitive sequences in the human genome. Nucleic Acids Res 19:4731–4738

Kulski JK, Dawkins RL (1999) The P5 multicopy gene family in the MHC is related in sequence to human endogenous retroviruses HERV-L and HERV-16. Immunogenetics 49:404–412

Kulski JK, Gaudieri S, Bellgard M, et al. (1997) The evolution of MHC diversity by segmental duplication and transposition of retroelements. J Mol Evol 45:599–609

Leelayuwat C, Zhang WJ, Townend DC, Gaudieri S, Dawkins RL (1993) Differences in the central MHC between humans and chimpanzees: Implications for development of autoimmunity and acquired immune deficiency syndrome. Hum Immunol 38:30–41

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

Maeda N, Kim HS (1990) Three independent insertions of retrovirus-like sequences in the haptoglobin gene cluster of primates. Genomics 8:671–683

Marck C (1988) DNA strider: A 'C' program for the fast analysis of DNA and protein sequences on the Apple Macintosh family of computers. Nucleic Acids Res 16:1829–1836

Mizuki N, Ando H, Kimura M, et al. (1997) Nucleotide sequence analysis of the HLA class I region spanning the 237-kb segment around the HLA-B and -C genes. Genomics 42:55–66

Mnukova-Fajdelova M, Satta Y, O'hUigin C, Mayer WE, Figueroa F, Klein J (1994) Alu elements in the primate major histocompatibility complex. Mammal Genome 5:405–415

Repaske R, Steele PE, O'Neill RR, Rabson AB, Martin MA (1985) Nucleotide sequence of a full-length human endogenous retroviral segment. J Virol 54:764–772

Ricke DO, Ketterling RP, Sommer SS (1992) PRE: A novel element with the hallmarks of a retrotransposon derived from an unknown structural RNA. Nucleic Acids Res 20:5233

Ruddy D, Kronmal G, Lee V, et al. (1997) A 1.1-Mb transcript map of the heriditary hemochromatosis locus. Genome Res 7:441–456

Satta Y, Mayer WE, Klein J (1996) HLA-DRB intron 1 sequences: Implications for the evolution of HLA-DRB genes and haplotypes. Hum Immunol 51:1–12

Shen L, Wu L, Sanlioglu S, et al. (1994) Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and C4B Genes on the HLA Class III Region. J Biol Chem 269:8466–8476

Shih A, Coutavas EE, Rush MG (1991) Evolutionary implications of primate endogenous retroviruses. Virology 182:495–502

Shiina T, Tamiya G, Oka A, et al. (1998) Nucleotide sequencing analysis of the 146-kilobase segment around the *IkBL* and *MICA* genes at the centromeric end of the HLA Class I region. Genomics 47: 372–382

Smit AFA (1996) The origin of interspersed repeats in the human genome. Curr Biol Gene Dev 6:743–748

Smit AFA, Toth G, Riggs AD, Jurka J (1995) Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J Mol Biol 246: 401–417

Tiwari JL, Terasaki PI (1985) HLA antigens associated with diseases. In: Anonymous HLA and disease associations. Springer, New York, pp 42–48

Urnovitz HB, Murphy WH. (1996) Human endogenous retroviruses: Nature, occurrence, and clinical implications in human disease. Clin Microbiol Rev 9:72–99

Venditti CP, Chorney MJ (1992) Class I gene contraction within the HLA-A subregion of the human MHC. Genomics 14:1003–1009

Vernet C, Boretto J, Mattei MG, et al. (1993a) Evolutionary study of multigenic families close to the human MHC class I region. J Mol Evol 37:600–612

Vernet C, Ribouchon MT, Chimini G, Jouanolle AM, Sidibe I, Pontarotti P (1993b) A novel coding sequence belonging to a new multicopy gene family mapping within the human MHC class I region. Immunogenetics 38:47–53

Watanabe Y, Tokunaga K, Geraghty DE, Tadokoro K, Juji T (1997) Large-scale comparative mapping of the MHC class I region of predominant haplotypes in Japanese. Immunogenetics 46:135–141

Wilkinson DA, Mager DL, Leong JA, Levy JA (eds) (1994) Endogenous human retroviruses. In: The Retroviridae. Plenum Press, New York, pp 465–535