Rhesus Macaque Class I Duplicon Structures, Organization, and Evolution Within the Alpha Block of the Major Histocompatibility Complex

Jerzy K Kulski,† Tatsuya Anzai,† Takashi Shiina,† and Hidetoshi Inoko†*

*Centre for Bioinformatics and Biological Computing, School of Information Technology, Murdoch University, Murdoch, Western Australia; and †Department of Molecular Life Science, Division of Basic Medical Science and Molecular Medicine, Tokai University School of Medicine, 143 Shimokasuya, Isehara, 259-1143, Japan

The alpha block of the human and chimpanzee major histocompatibility complex (MHC) class I genomic region contains 10 to 11 duplicated MHC class I genes, including the *HLA*/*Patr-A*, *-G*, and *-F* genes. In comparison, the alpha block of the rhesus macaque (*Macaca mulatta*, *Mamu*) has an additional 20 MHC class I genes within this orthologous region. The present study describes the identification and analysis of the duplicated segmental genomic structures (duplicons) and genomic markers within the alpha block of the rhesus macaque and their use to reconstruct the duplication history of the genes within this region. A variety of MHC class I genes, pseudogenes, transposons, and retrotransposons, such as *Alu* and *ERV16,* were used to categorize the 28 duplicons into four distinct structural categories. The phylogenetic relationship of MHC class I genes, *Alu*, and *LTR16B* sequences within the duplicons was examined by use of the Neighbor-Joining (NJ) method. Two single-duplicon tandem duplications, two polyduplicon tandem duplications with an accompanying inversion product per duplication, eight polyduplicon tandem duplications steps, 12 deletions, and at least two recombinations were reconstructed to explain the highly complex organization and evolution of the 28 duplicons (nine inversions) within the *Mamu* alpha block. On the basis of the phylogenetic evidence and the reconstructed tandem duplication history of the 28 duplicons, the *Mamu*/*Patr*/*HLA-F* ortholog was the first MHC class I gene to have been fixed without further duplication within the alpha block of primates. Assuming that the rhesus macaque and the chimpanzee/human lineages had started with the same number of MHC class I duplicons at the time of their divergence approximately 24 to 31 MYA, then the number of genes within the alpha block have been duplicated at an approximately three times greater rate in the rhesus macaque than in either the human or chimpanzee.

Introduction

The genomic organization of the human and chimpanzee major histocompatibility complex (MHC) class I regions is very similar, with major differences highlighted mainly by indel activity and increased nucleotide diversity within some subgenomic regions or blocks (Anzai et al. 2003). The human and chimpanzee MHC class I genes are clustered within three distinct locations, designated as the alpha, beta, and kappa blocks (Leelayuwat, Pinelli, and Dawkins 1995; Kulski et al. 2002), that are separated from each other by numerous framework non–class I genes (Vernet et al. 1994; Amadou 1999). The class I gene duplications within the alpha and kappa blocks are in the reverse orientation to those within the beta block (Leelayuwat, Pinelli, and Dawkins 1995; Shiina et al. 1999*a*). The Patr/HLA class I gene family includes at least six coding genes (*Patr*/*HLA-A*, *-B*, *-C*, *-E*, *-F*, and *-G*) and 12 pseudogenes or gene fragments (Anzai et al. 2003), whereas the human MIC gene family has two coding genes (*MICA* and *MICB*) and four pseudogenes or gene fragments, *MICC* to *MICF* (Shiina et al. 1999*a*). In chimpanzee, the *MICA* and *MICB* gene orthologs have recombined into a single chimeric *PatrMIC* gene by a 100-kb genomic deletion within the beta block (Kulski et al. 2002; Anzai et al. 2003).

The human MHC class I gene organization has been explained by chromosomal rearrangements that have involved a series of imperfect tandem duplications and various deletion events (Dawkins et al. 1999; Kulski et al. 1999*b*, Kulski, Gaudieri, and Dawkins 2000*a*). Different duplication models have been proposed, such as single-gene duplications (Hughes 1995), block duplications (Geraghty et al. 1992; Leelayuwat, Pinelli, and Dawkins 1995; Klein, Sato, and O'hUigin 1998), single-unit and biunit segmental duplications with transpositions (Shiina et al. 1999*b*), and serial unigenic and multigenic tandem duplications (Dawkins et al. 1999; Kulski et al. 1999*b*; Kulski, Anzai, and Inoko 2004) or metamerismatic duplications (Kulski, Gaudieri, and Dawkins 2000*a*). Detailed analyses of duplicon structures, organization, and phylogeny suggest segmental or tandem block duplication models are the most likely explanation for the MHC class I gene organization (Gaudieri et al. 1999*a*; Kulski et al. 1997, 1999*b*). Indels are a major pathway to genomic divergence of duplicated segments (Gaudieri et al. 1999*b*, 2000; Anzai et al. 2003) often by way of retrotransposons acting as recombination hotspots (Kulski et al. 1999*a*, 1999*b*). Therefore, an analysis of retrotransposons within genomic duplicated segments is a key to gaining useful insights into the time and nature of duplication events and genomic rearrangements.

The human MHC class I duplication unit (duplicon) consists of at least an *ERV16* element (Kulski and Dawkins 1999) and some other subfamily members of retrotransposons and DNA transposons, and a class I gene with or without an adjoining *MIC* gene (Geraghty et al. 1992; Avoustin et al. 1994, Dawkins et al. 1999; Kulski et al. 1999*b*; Shiina et al. 1999*b*). The genomic organization of the 10 to 11 HLA class I genes within the human alpha block was further categorized into four distinct duplicons based on the characteristic features of the HLA class I genes, retroelements, and their phylogenies (Kulski et al. 1999*b*). The four categories of the duplicons were used to reconstruct a duplication history that required only five

Key words: segmental tandem duplications, MHC genes, retrotransposons, rhesus macaque, comparative genomics.

E-mail: jkulski@murdoch.edu.au.

Mol. Biol. Evol. 21(11):2079–2091. 2004
doi:10.1093/molbev/msh216
Advance Access publication July 21, 2004

tandem duplication steps, starting from a single MHC class I duplicon, to explain the organization of the 10 alternating MHC class I genes within the human and chimpanzee MHC class I alpha block (Kulski et al. 1999*b*, Kulski, Anzai, and Inoko 2004). A duplication-transposition model based on seven duplications and four transpositions of MHC class I genes has also been proposed (Shiina et al. 1999*b*).

Recently, the entire genomic sequence of the rhesus macaque MHC class I region was completed, annotated, and compared with the corresponding human and chimpanzee region (Shiina et al. unpublished data). The rhesus macaque MHC class I region was found to be markedly different to the human and chimpanzee MHC class I gene organization, with a complex organization of at least 20 additional class I genes within the alpha block and 17 additional class I genes within the beta block. Although there are no inverted MHC class I genes within the alpha block of the chimpanzee and the human, there are nine in the rhesus macaque. In this study, we examined in greater detail the organization of the rhesus macaque MHC class I genes within the alpha block and identified certain transposons and retroelements within the segmental genomic duplicon structures with a view to reconstructing their duplication history. This paper shows that the complex arrangement of 31 class I genes within the rhesus macaque MHC class I alpha block can be explained parsimoniously by a series of unigenic and multigenic tandem duplications, tandem duplications with inversions, and deletions and recombinations based on a previous model developed for humans (Kulski et al. 1999*b*).

Materials and Methods

A rhesus macaque (*Macaca mulatta*, *Mamu*) MHC class I genome sequence (EMBL/DDBJ/GenBank accession number AB128049) was determined using a CHORI250 BAC library obtained from the Children's Hospital Oakland Research Institute, BAC Resources, and the cosmid library 159 was from the Resource Centre of the Human Genome Project, Berlin, Germany. The procedures for sequencing and assembling a 2.4-Mb contig map of the macaque class I region from the *LTB* to the *HLA-F–like* genes using the BAC and cosmid clones is reported by Shiina et al. (unpublished data). For comparative genomic analysis of the human and rhesus macaque MHC class I region, genomic sequences were accessed as accession numbers AP000502 to AP000521 for the human sequences (Shiina et al. 1999*b*) and AB128049 for the rhesus macaque sequences (Shiina et al. unpublished data) from EMBL/DDBJ/GenBank. The genomic sequences were edited manually where required to produce the alpha block sequences for comparative analysis.

Dot-plot matrix analyses were performed using the programs Dotter (Sonnhammer and Durbin 1995) or HarrPlot version 2.1.0 as part of the GENETYX version 11 program (Shiina et al. 1999*b*) as required. The programs RepeatMasker (http//ftp.genome.washington.edu/cgi-bin/RepeatMasker, A.F.A. Smit and P. Green pesonal communication) or CENSOR (Jurka et al. 1996*b*) were used to identify DNA transposons and retrotransposons within the contiguous sequences. BlastN (NCBI) confirmed gene loci within the genomic sequences. Sequence alignments were performed using ClustalW (Baylor College of Medicine) or ClustalX version 1.81 (Thompson et al. 1997), and the phylogenetic analysis was performed using the Neighbor-Joining (NJ) method within Clustal at DDBJ (http://www.ddbj.nig.ac.jp/E-mail/clustalw-e.html) or the programs PAUP* (Swofford 1998) or MEGA version 2.1 (Kumar et al. 2001). The sequences flanking *Alu* repeats within different duplicated segments were aligned and examined for homology to confirm their paralogous location by using ClustalW and a spreadsheet program (Excel, Microsoft). Each paralogous *Alu* element within the alpha block was given a code name and number following the subfamily designation (e.g., *AluJ1* and *AluJ2* or *AluS1*, and *AluS2*) as previously described by Kulski et al. (1999*b*). The same number was used for each *Alu* element found within a paralogous location of duplicated segments unless specified otherwise.

The nomenclature used in this report for the *Mamu* MIC and *Mamu* class I genes within the alpha block (fig. 1) is the same as that used by Shiina et al. (unpublished data). The *Mamu-80* (eight copies), *Mamu-G* (six copies), *Mamu-A* (three copies), *Mamu-70 (*one copy) and *Mamu-75* (three copies) are equivalent to the *Patr*/*HLA-80*, *-G*, *-A*, *-70*, and *-75* genes, respectively. *Mamu-AG* (five copies) is also closely related to either the *HLA-A* or *HLA-H* genes. *Mamu-MICG* (eight copies) and *Mamu-MICD* are equivalent to the human *MICG* and *MICD* genes, respectively. *Mamu-59* is also called *Mamu-J* and is equivalent to *HLA-59* or *HLA-J*. We also refer to some deleted *Mamu-75* genes here as *Mamu-Del75a* (or Del75a) and *Mamu-Del75b* (or Del75b) and a deleted *Mamu-AG* gene as *Mamu-AGdel* (or AGdel). The three class I S series genes, *SD, SE-1*, and *SE-2* have low identity (<70%) at exon 4 with *HLA-A*, *-B*, *-C*, *-E*, *-F*, and *-G*, and their strongest similarity is with mouse class I T region genes (Shiina et al. unpublished data).

Results and Discussion

Comparison Between Rhesus Macaque and Human Alpha Block Genomic Sequences

Figure 1 shows a gene map and dot-plot comparison between rhesus macaque and human alpha block genomic sequences. It is evident from the map and the dot-plot that the duplication patterns within the human alpha block are different from those in the rhesus macaque. The rhesus macaque genomic alpha block sequence is approximately 890 kb in length, consisting of 31 class I genes and nine MIC genes, whereas the human genomic alpha block sequence is approximately 300 kb. consisting of 11 class I genes and three MIC genes. Of these class I genes, three were identified as expressed genes (*Mamu-A1*, *-A2*, and *-AG3*) and four as possibly expressed genes (*Mamu-AG1*, *-AG2*, *-AG4*, and *-F*). The remainders were classified as pseudogenes. The *MICD* gene fragment has four exons and the other MIC genes are duplicated fragments with one exon.

The rhesus macaque genes within the expanded region 1 (ER1) and D1 region of figure 1 are in the same direction and have a similar organization to the human genes. *Mamu-F*, like the *Patr*/*HLA-F* gene orthologs, is

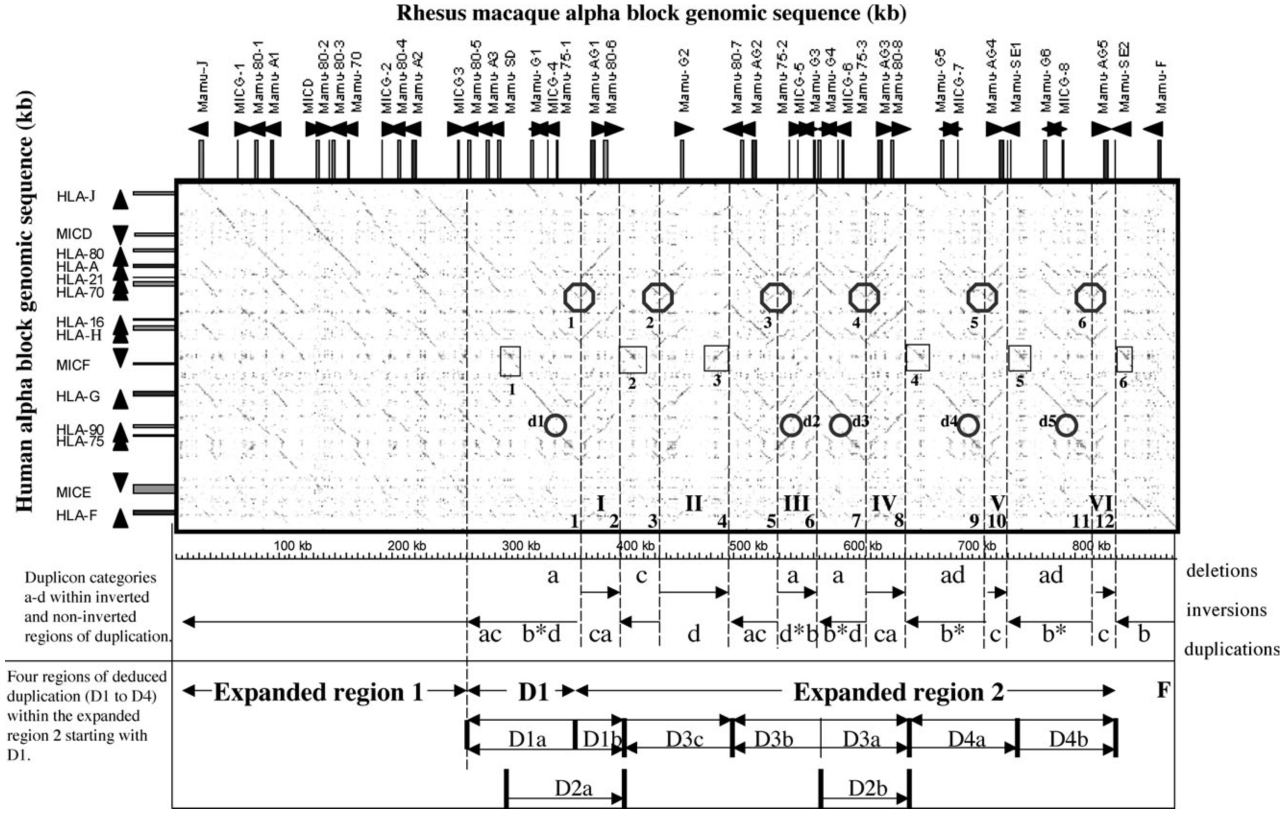

FIG. 1.—Gene map and a dot-plot comparison between the human and rhesus macaque alpha block genomic sequences. The class I–like gene orders and organization of the duplicated genomic segments are shown categorized as duplicons "a," "b," "c," and "d" along the *x*-axis at the bottom of the matrix. The dot-plot also shows the indels and breakpoints of the inverted duplicons within the rhesus macaque. The rhesus macaque genes (labeled vertical bars) and direction of coding (horizontal arrows) are shown at the top of the dot-plot matrix along the *x*-axis. The human class I–like genes (labeled horizontal bars) and direction of coding (vertical arrows) are shown on the left-hand side of the matrix along the *y*-axis. The locations of the 12 breakpoints for the inversions (I to VI) are indicated by the numbers 1 to 12 above the *x*-axis that shows the sequence length in kb. The numbered rectangles, 1 to 6, show the breakpoints of four of the inverted duplicons and the regions of the BALSL (breakpoint-associated LTR-sine-line) duplicated sequences. The octagons labeled 1 to 6 show the *ERV16* breakpoints for six inverted duplicons, I to VI (shown as horizontal arrows, left to right, along the *x*-axis at the bottom of the dot-plot matrix). The circles labeled d1 to d5 highlight the regions of a highly fragmented category A duplicon (*) with the deletion of a *Mamu-80* gene family member. In this context, b*d means "b-a fragment-d" with deletion of a *Mamu-80* gene family member from the fragmented duplicon "a." The extended region 1 (ER1) and D1 correspond approximately to the human and chimpanzee segmental organization of BACBACBADB within the alpha block (Kulski et al. 1999*b*). The double horizontal arrow labeled D1 shows the duplication origin for ER1 and expanded region 2 (ER2).

located at the very telomeric end of the rhesus macaque alpha block. In the expanded region 2 (ER2), telomeric of the *Mamu-75-1* gene, there are six genomic inversions. ER2 has 18 class I genes and four *MIC* genes, with 11 of the 18 class I genes in the opposite direction to the 13 class I genes centromeric of ER2. In contrast to the rhesus macaque genomic sequence, all of the human and chimpanzee class I genes and *MIC* genes within each respective gene group are oriented in the same direction. In addition, the rhesus macaque appears to have lost the human *HLA-90* ortholog (or the *Mamu-80* paralogous gene) in the D1 region and within the five duplicated locations of the ER2. Thus, the rhesus macaque ER1 and D1 are similar in organization to the human and chimpanzee alpha block region, whereas the ER2 of the rhesus macaque has no equivalent structure in the human and chimpanzee. Figure 1 also shows the duplicon categories, A to D, that have been deleted, inverted and/or duplicated within the inverted and noninverted segments of the rhesus macaque ER2. The four distinct duplication regions (D1 to D4) that were deduced to have been involved in the formation of the rhesus macaque ER2 are shown in figure 1 and discussed in more detail in the following sections.

Structural Categories of MHC Class I Duplicons

To better understand the genomic organization and duplication history of the rhesus macaque alpha block, the duplicated genomic segments (duplicons) were classified into four categories, A to D, on the basis of the types of duplicated retrotransposons and transposons that are linked to a particular class I and/or *MIC* gene within each segment (Kulski et al. 1999*b*, 2002). Figures 2 and 3 show the structural features for each of the rhesus macaque A to D duplicated segments. The nomenclature for the duplicated genomic segments in rhesus macaque is the same as in the human (Kulski et al. 1999*b*) and chimpanzee (Kulski, Anzai, and Inoko 2004), except that here we have called the previously labeled category B′ genomic segment as D for greater emphasis and clarity. A summary of the common transposons and retrotransposons

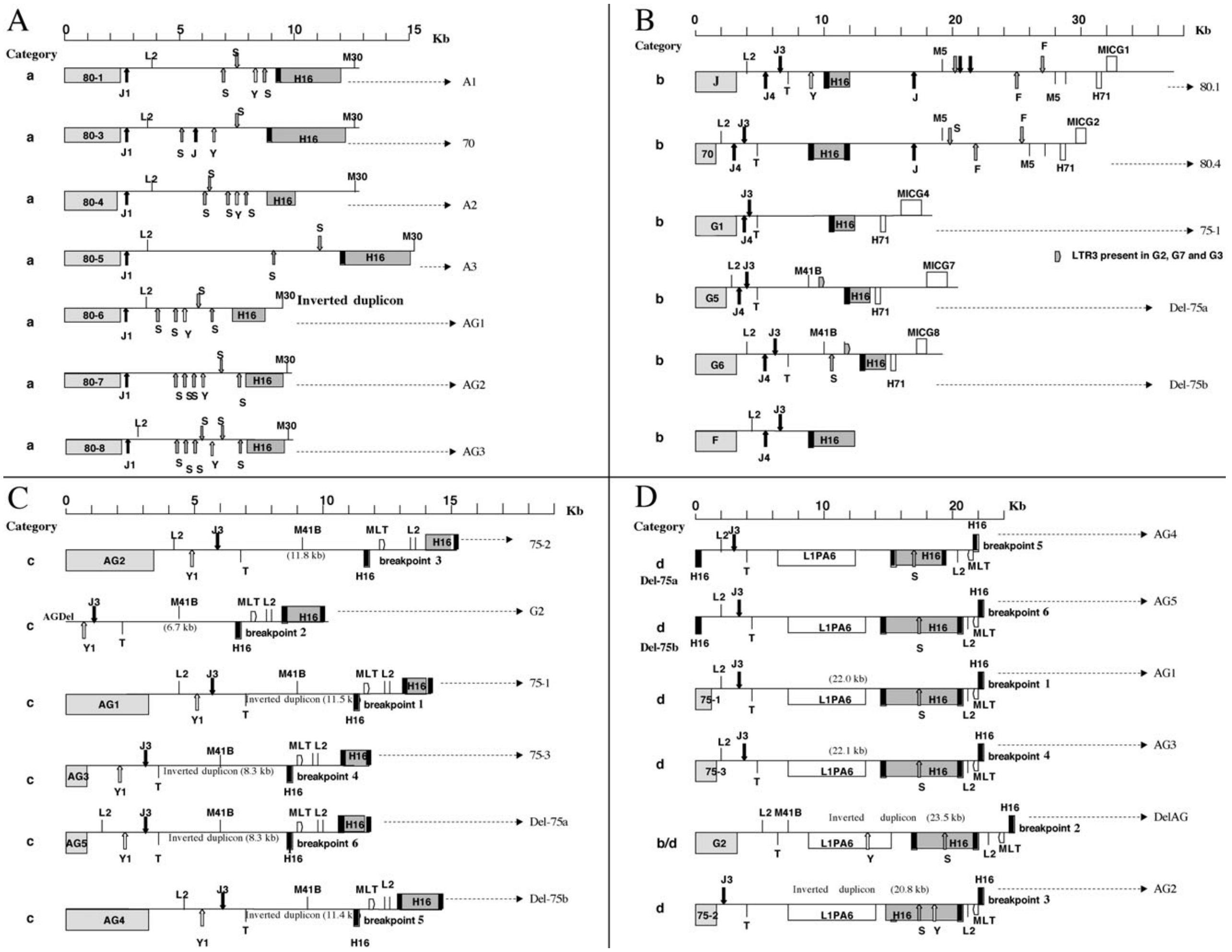

FIG. 2.—Rhesus macaque duplicon structures with the MHC class I and *MIC* genes and distinctive retrotransposons classified into four categories, A to D, according to table 1. Alu elements are shown as vertical arrows and are labeled as J (*AluJ*), S (*AluS*), or Y (*AluY*). The *ERV16* is shown as a rectangular shaded box labeled H16. The solid vertical line flanking the H16 box represents the *ERV16-LTR16B*. M30 is the *MER30* fragment. T is *The1*, M5 is *Mer5*, M41B is *Mer41B*, and H71 is *ERV71*. The open boxes labeled MICG1 to MICG8 indicate the *MIC* fragmented genes. (*A*) The seven category A duplicons have the *Mamu-80* pseudogenes and are characterized by the *AluJ1* element. *Mamu-80-2* is not shown because it is only a small fragment of 200 bp. (*B*) Six category B duplicons that contain the *Mamu-G* genes and are characterized by the *AluJ4* and *AluJ3* elements. *LTR3* is indicated by the shaded D symbol. (*C*) Six category C duplicons that contain or have contained the *Mamu-AG* genes are characterized by the *AluJ3* and *AluY1* elements. One of the duplicons has lost its *Mamu-AG* gene (*AGDel*) but has retained its characteristic Alu elements. These duplicons end with the *LTR16B* sequence at the inversion breakpoint as indicated. Four of the duplicons (*AG1, AG3, AG4*, and *AG5*) are inverted. (*D*) The six category D duplicons characterized by *AluJ3* (labeled J3), the *L1PA6* (open box) sequence, and the presence or absence of the *Mamu-75* genes. Two of the duplicons have lost their *Mamu-75* genes and are labeled as *Del-75a* and *Del-75b*. One of the duplicons with the *Mamu-G2* gene (a category B duplicon gene) has lost its *AluJ3* element, and it is probably a fusion product between a category B and D duplicon. Two of the category D duplicons (*G2* and *75-2*) are inverted as labeled in the figure.

found within each of the duplicon categories A, B, C, and D are presented in Table 1. The repeat elements identified within the duplicons include DNA transposons (*MER5A*, *MER5B*, *MER20*, *MER30*, *CHARLIE1*, *CHARLIE 9*, *TIGGER1*), LTRs (*MER9*, *MER21B*, *MER41B*), SINES (*Alu*, *MIR*), LINEs (*L1*, *L2*), members of a superfamily of Mammalian apparent-LTR retrotransposons (MaLRs) such as *MST*s and *MLT*s (Smit 1993; Jurka et al. 1996*a*; Smit 1996, 1999; Smit and Riggs 1996) and different ERV families (Kulski et al. 1999*a*; Jurka 2000). *ERV16* is present usually as a fragmented sequence in most of the duplicated segments of the alpha block.

Figure 2*A* shows the general genomic structure of the seven category A duplicons. They are linked with the *Mamu-80*, *Mamu-80-1*, and *Mamu-80-3* to *-80-8* genes and are equivalent to the human and chimpanzee *HLA*/*Patr-80*, *-16*, and *-90* genes (Kulski, Anzai, and Inoko 2004). The category A duplicons have the remnants of the *ERV16* sequences but no *MIC* genes. They are distinguished from the segmental categories B to D by having the category A–related *AluJ1*, *AluY2*, *MSTB*, and *MER21B* elements. The *AluJ1* element has been inserted in close proximity to exon 3 approximately 400 to 530 bp from the deletion point of the exons 1 and 2 that were deleted from the *Mamu-80* ancestral gene. The rhesus macaque has seven relatively intact *Mamu-80* duplicons, greater than 9 kb in length, and one highly fragmented *Mamu-80-2* gene (one exon) of only 200 bp. Two of the *Mamu-80* duplicons (*Mamu-80-6* and *Mamu-80-8*) are in the opposite orientation to the other duplicons. In addition, the rhesus

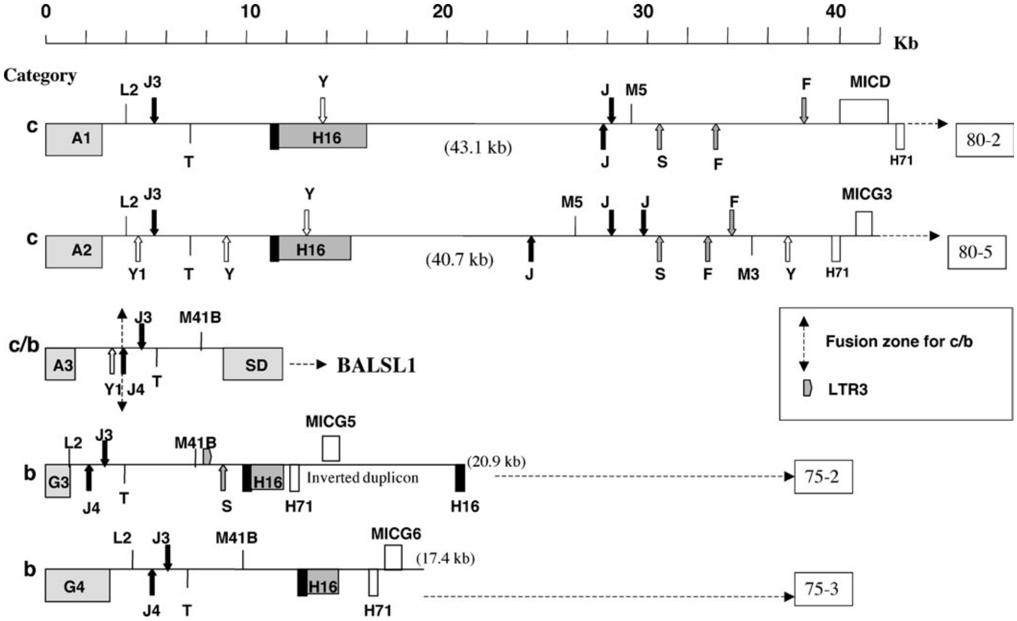

FIG. 3.—The structure of rhesus macaque categories B, C, and chimeric duplicons. Three category C duplicons that contain the *Mamu-A* genes are shown. One of the duplicons (*A1*) has lost its characteristic *AluY1* element, whereas the duplicon *A2* has retained both *AluY1* and *AluJ3*. The duplicon *A3* appears to be a category B and C hybrid (B/C) because it has the *AluY1, AluJ4*, and *AluJ3* elements. The vertical dashed double arrow shows the putative region of fusion between a category C segment (*Mamu-A3* gene and the *AluY1* element) and a category B segment (*AluJ4* to the *Mamu-SD* gene fragment). The category B duplicons carrying the *Mamu-G3* and *-G4* genes are shown for comparison.

macaque might have had an additional five *Mamu-80*–type genes within its alpha block if a copy of the *Mamu-80* gene had not been deleted from its position telomeric of the *Mamu-75-1* gene, as indicated in figure 1.

The basic features and genomic structure of the eight category B duplicons are shown in figure 2*B* and figure 3. In comparison, there are only four category B genomic duplicons, *J, 70, G,* and *F,* in the human and chimpanzee (Kulski, Anzai, and Inoko 2004). The category B duplicons are characterized by the presence of the distinctive *AluJ3* and *AluJ4* elements that are absent from the category A duplicons. Except for duplicon *F,* the category B duplicons in the rhesus macaque are 18 kb or more in length, and they all have a fragment of the *MIC* genes and the *LTR71B/ERV71* sequences. Only one of the category B duplicons (*Mamu-G3*) is located in the

Table 1
Genomic sequence features defining different categories of Rhesus class I MHC duplicons

Features	Category A	Category B	Category C	Category D
MHC class I genes*	80-1 to 80-8	J, 70, F, G1, G3-6	A1, A2, AG1-AG5	G2, 75-1 to 75-3
AluJ1 ©	yes	no	no	no
L2 (+)	yes	yes	yes	yes
AluY1 ©	no	no	yes	no
AluJ4 ©	no	yes	no	yes
AluJ3 (+)	no	yes	yes	yes
THE1C ©	no	yes	yes	yes
MER41B ©	no	yes	yes	yes
LTR3 (+)	no	yes	no	no
L1PA6 ©	no	no	no	yes
LTR16B/ERV ©	yes	yes	yes	yes
AluJ(C)	no	yes	yes	no
MER5(+)	no	yes	yes	no
FRAM©	no	yes	yes	no
FRAM(+)	no	yes	yes	no
MER5B©	no	yes	yes	no
LTR71/ERV	no	yes	yes	no
MIC (+)	no	yes	yes	no
Charlie 1 (+)	no	yes	yes	no
MSR21B (+)	yes	no	no	no
MSTB ©	yes	no	no	no
MER20 ©	no	yes	no	yes
MER30 (+)	no	yes	yes	yes

* The orientation of the MHC class I genes are shown in fig. 1.
(+) is positive DNA strand and © is complementary DNA strand.

opposite direction to the other category B duplicons. In addition, there appear to be at least two category B hybrid duplicons, *Mamu G2* and *Mamu A3,* where a portion of the category B duplicon has been recombined with a category D (fig. 2*D*) and C duplicon (fig. 3), respectively.

The category C duplicons have the *AluJ3* and the *AluY1* sequences and the *Mamu-A* or *Mamu-AG* genes (fig. 2*C* and 3). In humans and chimpanzees, the category C duplicons carry the *Patr/HLA-A* and *Patr/HLA-H* genes. Moreover, six inversion breakpoints were found between the category C and D duplicons in rhesus macaque but not in the human or the chimpanzee. Overall, there are nine category C duplicons in the rhesus macaque compared with only two in the human and chimpanzee (*Patr/HLA-A* and *Patr/HLA-H*). Five of the category C duplicons in the rhesus macaque contain the *Mamu-AG* genes (*AG1* to *AG5*), one duplicon (*AGDel*) has a deleted *Mamu-AG* gene (fig. 2*C*), and three others have the *Mamu-A1, -A2,* and *-A3* genes (fig. 3). Four of the category C duplicons (*AG1, AG3, AG4, AG5*) are inverted in relation to the direction of most of the other class I genes.

The category D duplicons that are linked with the *Mamu-75* genes in the rhesus macaque, are shown in figure 2*D*. There is a full-length *L1PA6* retrotransposon (6,140 bp), inserted within the one human (Kulski et al. 1999*b*) and the six rhesus macaque category D duplicons (fig. 2*D*) that is missing from the chimpanzee (Kulski, Anzai, and Inoko 2004). The two open reading frames, one coding for a RNA-binding protein and the other coding for a protein with an endonuclease and reverse transcriptase domains (Feng et al. 1996), are highly disrupted by numerous premature stop codons within all of the *Mamu-L1PA6* sequences. A *Charlie9* fragment and a *L1ME3B* or *MLT1E3* fragment usually flank the insertion site for the *L1PA6* sequence, and it appears to be the same site for both the human and rhesus macaque sequences. The *L1PA6* sequence in the orthologous location of duplicon *75-1* of the rhesus macaque and the human duplicon *75* suggests that it was most likely deleted from the chimpanzee duplicon *75*. The human *L1PA6* insertion however, has a species-specific flanking telomeric duplication of a genomic sequence that contains the category B segmental elements, *AluJ4* and *AluJ3*. On the other hand, the region between *HLA-75* and the *L1PA6* contains *AluJ3* but lacks *AluJ4*, similar to the category C duplicons (Kulski et al. 1999*b*). The corresponding regions within duplicons *75* of the chimpanzee and rhesus macaque also lack the category *AluJ4* element. The category D duplicons are similar to the category C duplicons, except that they have the *L1PA6* insertion but not the *AluY1* insertion. In addition, all category D duplicons have an *AluS* insertion within the *ERV16* sequences and a *MLT* sequence adjoining the *LTR16B* inversion breakpoint. Nevertheless, the structural and sequence similarity of the category D duplicon with the C duplicons suggest that it is an intermediate structure between the category B and category C duplicons and the likely precursor to the category C duplicons.

Because of the *L1PA6* insertion, the human and rhesus macaque duplicons *75* are referred to here as category D duplicons (previously called B′ or B [Kulski, Anzai, and Inoko 2004]). In human, there is only one category D duplicon (contains the *HLA-75* gene), whereas in the rhesus macaque, there are six. Three category D duplicons have the genes *Mamu-75-1, -75-2*, and *-75-3*, whereas two others have had their *Mamu-75* genes completely deleted and are referred to here as the *Mamu-Del75a* and *-Del75b* duplicons (fig. 2*D*). Another of the D duplicons appears to be a hybrid structure between a category B and a category D duplicon because it has a fully intact category B gene (*Mamu-G2*) instead of the category D *Mamu-75* fragmented gene (exons 1 to 3). The remainder of this hybrid duplicon is typically category D with the full-length *L1PA6* insertion at the expected location. Two of the category D duplicons have been inverted, and all of the telomeric ends of the category D duplicons in the vicinity of the *ERV16* and *LTR16B* sequences form an inversion breakpoint. This inversion breakpoint is linked to the category C duplicons that have the *Mamu-AG* genes.

Breakpoints and BALSL Complexes

The ER2 has six inverted subregions providing 12 breakpoints, as shown in figure 1. The breakpoints for the original inversions are difficult to ascertain because they may have been masked by the subsequent duplications, insertions, deletions, and other rearrangements. Nevertheless, there are four discernable inversion breakpoints (fig. 1): (1) the *ERV16* breakpoint at positions 1, 2, 3, 4, 5, and 6, (2) the breakpoint between *Mamu-G3* and *Mamu-G4* at position 6, (3) the breakpoint near the BALSL 5′ ends at positions 2, 4, and 8, and (4) the breakpoint at positions 10 and 12 that are between the *Mamu-SE1/Mamu-SE2* and *Mamu-AG4/ Mamu-AG5* genes, respectively. Presumably, most of the inversion breakpoints, such as the *ERV16* breakpoints, have resulted from duplications rather than actual inversions.

There are a collection of duplicated retroelements located between the *ERV16* sequences and the *MIC* genes within the category B duplicons of Patr/HLA-J to *Patr/ HLA-80, Patr/HLA-H* to *Patr/HLA-G,* and *Patr/HLA-75* to *Patr/HLA-F*. This collection of duplicated retroelements, which we have termed the "breakpoint-associated LTR-sine-line" complex, or BALSL for short, has been modified slightly in the rhesus macaque alpha block. Figure 4 shows a majority of the BALSL retroelements and transposon sequences that adjoin the six breakpoint locations of the rhesus macaque in comparison with the similar group of elements that were detected within the two human duplicons linked with *HLA-H* and *HLA-75*, respectively. The human duplicon with *HLA-J* also has some of the elements, such as *LTR16B/ERV16, L2, MLT1E2, Tigger 1, MIR, MLTB1*, and *MER5*. The genomic ends close to *L2/MLT1E2* or *MIR* of BALSL2-4 form the terminal ends with the breakpoints within the adjoining retroelements that are downstream from *Mamu-80–6* to *Mamu-80–8*. On the other hand, the *MIR/MLTB1* sequences of BALSL5-6 form the terminal ends with the breakpoints located closely to *Mamu-SE1* and *Mamu-SE2*. The 584-bp sequence between *Mamu-SD1* and BALSL1 appears to contain the breakpoint position for the start of the duplication D2a region that is shown in figure 1. The *LTR16B/ERV16* sequences that are present within the

Breakpoint associated LTR sine line (BALSL) complex

BALSL1	BALSL2	BALSL3 Inv	BALSL4	BALSL5	BALSL6		
Mamu-SD	*Mamu-80-6 inv*	*Mamu-80-7 inv*	*Mamu-80-8 inv*	*Mamu-SE1*	*Mamu-SE2*	*HLA-H*	*HLA-75*
-	-	-	-	-	-	LTR16B/HERV16	LTR16B/HERV16
-	L2/MLT1E2	L2/MLT1E2	L2/MLT1E2	-	-	L2/MLT1E2	L2/MLT1E2
-	Tigger1	Tigger1	Tigger1	-	-	Tigger1	Tigger1
MIR	MIR	MIR	MIR	MIR	MIR	MIR	-
MLTB1	MLTB1	MLTB1	MLTB1	MLTB1	MLTB1	MLTB1	-
LTR2	LTR2	LTR2	LTR2	-	-	-	-
AluY	AluY	AluY	AluY	AluY	-	AluY	-
LTR2	LTR2	LTR2	LTR2	-	-	-	-
MLT1B	MLT1B	-	MLT1B	-	-	-	-
FLAMC	FLAMC	FLAMC	FLAMC	FLAMC	FLAMC	FLAMC	-
L1	L1	L1	L1	L1/tigger6	L1/tigger6	L1/tigger6	-
AluJ	AluJ	AluJ	AluJ	AluJ	AluJ	AluJ	-
L1ME3B	L1ME3B	L1ME3B	L1ME3B	L1ME3B	L1ME3B	L1ME3B	-
MER9	MER9	MER9	MER9	MER9/AluYc	MER9/AluYc	MICF	MICE
L1ME3B	L1ME3B	L1ME3B	L1ME3B	L1ME3B	L1ME3B	-	L1ME3B
LTR6B	LTR6B	LTR6B	LTR6B	LTR6a	LTR6a	-	-
HAL1	HAL1	HAL1	HAL1	HAL1	HAL1	HAL1	HAL1
AluSc	AluSc	AluSc	AluSc	AluSc	AluSc	AT rich	-
HAL1	HAL1	HAL1	HAL1	HAL1	HAL1	HAL1	HAL1
MER5A	MER5A	MER5A	MER5A	MER5A	MER5A	MER5A	MER5A
L1 types	L1 types	L1 types	L1 types	L1 types	L1 types	L1 types	L1 types
MER4	MER4	MER4	MER4	MER4	MER4	-	-
L1 types	L1 types	L1 types	L1 types	L1 types	L1 types	L1 types	L1 types
MLT1E2	MLT1E2	MLT1E2	MLT1E2	MLT1E2	MLT1E2	MLT2B3	MLT2B3
LTR7/L1MB3	LTR7/L1MB3	LTR7/L1MB3	LTR7/L1MB3	LTR7/L1MB3	LTR7/L1MB3	-	-
LTR5-Hs	LTR5-Hs	LTR5-Hs/AluY	LTR5-Hs	LTR5-Hs	LTR5-Hs	-	-
MIR	MIR	MIR	MIR	MIR	MIR	MIR	MIR
21.5 kb	26.6 kb	38.8 kb	20.6 kb	19.0 kb	18.5 kb	36.7 kb	37.3 kb

FIG. 4.—Duplicated retrotransposons and transposons within the six breakpoint-associated LTR-sine-line (BALSL) complexes.

human BALSL-like sequences appear to have been completely lost from the rhesus macaque BALSL complexes, presumably as part of the duplication process D2 that had first involved BALSL1. In addition, the *MIC* genes in the human BALSL-like sequences appear to have been replaced by the endogenous retroviral LTR sequences, *MER9*, within the rhesus macaque BALSL complex. This further differentiates the rhesus macaque duplicated BALSL complex from the structurally similar complex in human and chimpanzee (fig. 4).

ERV16 sequences appear to be directly involved with six of the 12 breakpoints (figs. 2 and 3). The six centromeric inversion breakpoints at positions 1, 3, 5, 7, 9, and 11 terminate with the *MLT* and *LTR16B* sequences as indicated in figures 2*C* and *D*. On the other hand, the five telomeric inversion breakpoints at BALSL positions 2, 4, 8, 10, and 12 have lost all of their *ERV16* sequences (fig. 4). The inverted breakpoint at position 6 (fig. 1) is different from the other breakpoint positions in that it involves two different *Mamu-G* gene 3′ regions, resulting in *Mamu-G3* and *Mamu-G4* on either side of the inversion breakpoint region with the genes *Mamu-G3, MICG-5*, and *Mamu-75-2* linked together within their respective duplicons (fig. 1).

Phylogeny of Genes and Retrotransposons Within MHC Class I Duplicons

Figure 5 shows the separate NJ trees of the *Mamu* exons 2 and 3. In general, the exon sequences clustered into groups according to their duplicon categories, A to D. The exception was the *Mamu-A2* exon 3 that grouped with the category B sequences. The category A to D sequences from the duplication regions D2, D3, and D4 (fig. 1) paired closely together in the NJ trees, suggesting they had evolved more recently than the sequences from the D1 and ER1 regions.

The NJ tree for the duplicated *AluJ1, J3, J4*, and *Y1* elements found within the category A to D duplicons is shown in figure 6*A*. The Alu elements grouped fairly strictly according to their subgroup membership (*J1, J3, J4*, and *Y1*), paralogous location and duplicon categories (A to D). The exceptions were *MA3.AluJ3* and *MA3.AluJ4*. These two elements grouped with category B rather than with the category C duplicons, indicating that they are part of a hybrid product composed of category B and C duplicons because *MA3.AluY1* grouped with the category C duplicons as expected. The category C and D duplicons have *AluJ3* but not the *AluJ4* element that is characteristic of the category B segments. The phylogeny, together with the structural analyses, suggests that the category C segments have evolved from a D segment. Interestingly, the *AluJ3* is located between the retroelements *MLTF1* and *Charlie9* or *MLT1E3* (if *Charlie9* is deleted), whereas the category C *AluY1* is inserted within *MLT1E3* between nucleotide positions 277 and 330 of the element.

The Alu elements from the ER2 appear to be the products of more recent duplications than those from ER1 and D2 that are shown in figure 1. In this regard, it is assumed that the most recent duplication products have the closest sequence identity and, therefore, have grouped

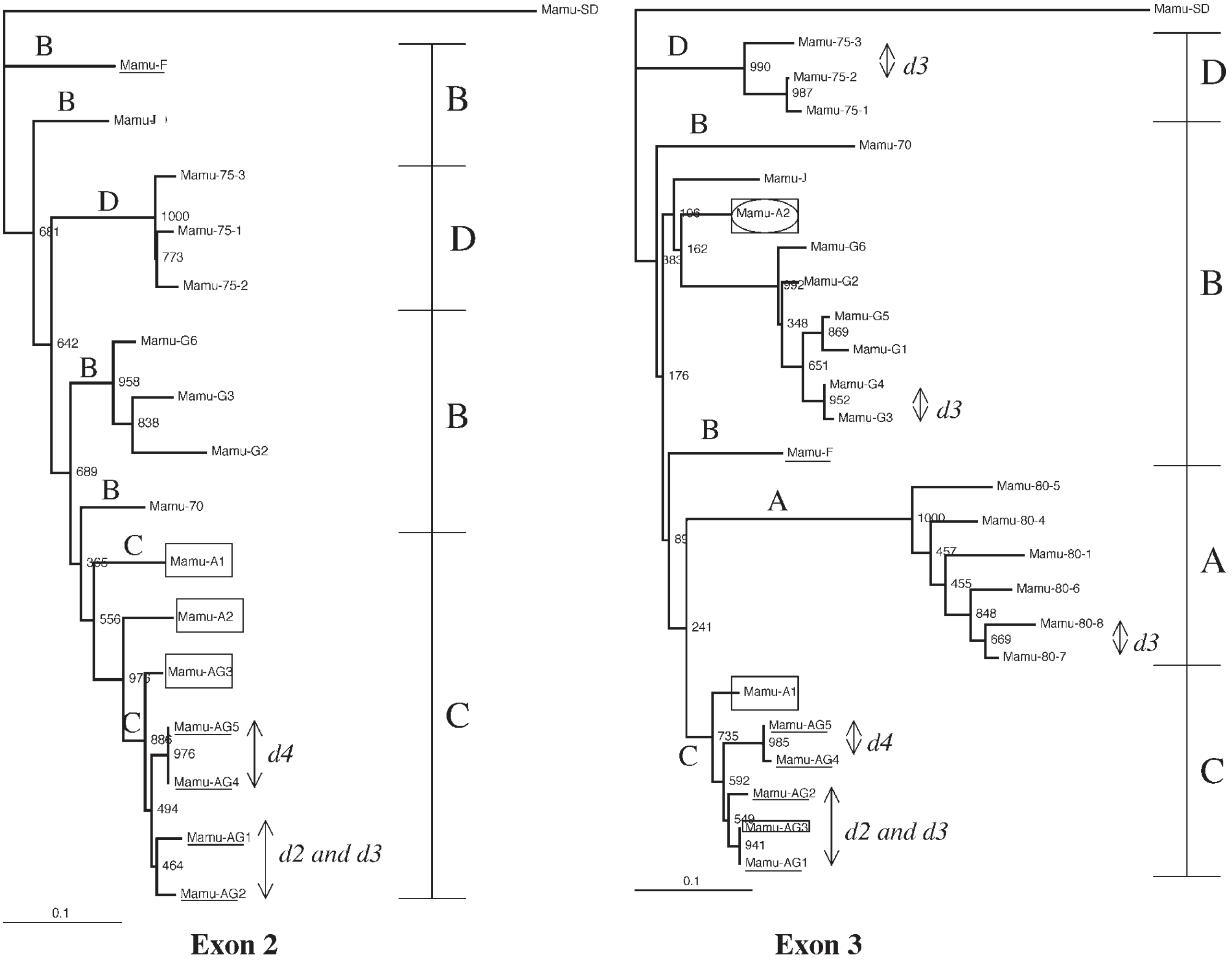

FIG. 5.—Phylogenetic trees of *Mamu* class I gene exons 2 and 3. The exon 2 and 3 sequences were aligned using ClustalX, and phylogenetic trees were constructed by using the NJ method and the default parameters at DDBJ. The reliability of the trees was measured by bootstrap analysis with 1,000 replicates but is shown as percentage values. The scale bar is the number of substitutions per site. The duplicated pairs that are the inferred products of the most recent duplications (D2 to D4) are labeled *d2, d3*, and *d4* and indicated by the vertical double arrows. The letters A, B, C, and D within the trees and on the right-hand side of the trees represent the duplicon structural categories (see figures 2 and 3). The boxed genes are the coding genes. The possible coding genes are underlined. The *Mamu-A2* exon 3 is circled because it has grouped with the category B sequences rather than with the category C as expected.

closely together in the tree (the pairs are indicated by vertical arrows as d3 and d4 in figure 6*A*). *MF.AluJ3* and *MF.AluJ4* have diverged the furthest from all the other elements that belong to the duplication categories B, C, and D, confirming their more ancient origins. In addition, the *AluJ1* elements from the category A duplicons grouped separately from all the *AluJ* elements that belong to the category B to D duplicon lineages.

Figure 6*B* shows the distance tree for the 5′ and 3′ *LTR16B* sequences. The 5′ and 3′ *LTR16B*, supposedly identical in sequence when first inserted into the alpha block, have separated into there own distinct clusters. This suggests that after *ERV16* was first inserted into the ancestral duplicon, the 5′ LTR (L5) and 3′ LTR (L3) had already diverged before any of the subsequent duplication events. Consequently, the 5′ and 3′ LTR do not group together as would be expected if the *ERV16* sequences were the products of recent and separate insertions. Relatively few 5′ *LTR16B* sequences (six sequences) compared with the 3′ *LTR16B* sequences (29 sequences) have remained within the duplicons of the alpha block, suggesting a bias towards the retention of the 3′ *LTR16B*. This was confirmed by the alignment of 32 *ERV16* sequences (data not shown). Of the 32 *ERV16* sequences, 29 had at least one LTR remaining. Only six *ERV16* sequences had both of their 3′ and 5′ *LTR16B* left intact, although partially fragmented. There were seven solitary *LTR16B* sequences, and the other *LTR16B* sequences had a portion of their internal *ERV16* sequences linked with its LTR so that they could be easily distinguished as either a 5′ or 3′ LTR. The seven solitary *LTR16B* sequences were classified as 3′ LTR on the basis of their topology within the NJ trees (fig. 6*B*).

The six 5′ *LTR16B* are part of the category D and category B duplicons, whereas the 29 3′ *LTR16B* belong to all four duplicon categories, A to D. The category B to category D sequences grouped together to form a lineage separately from the category A sequences. In category A,

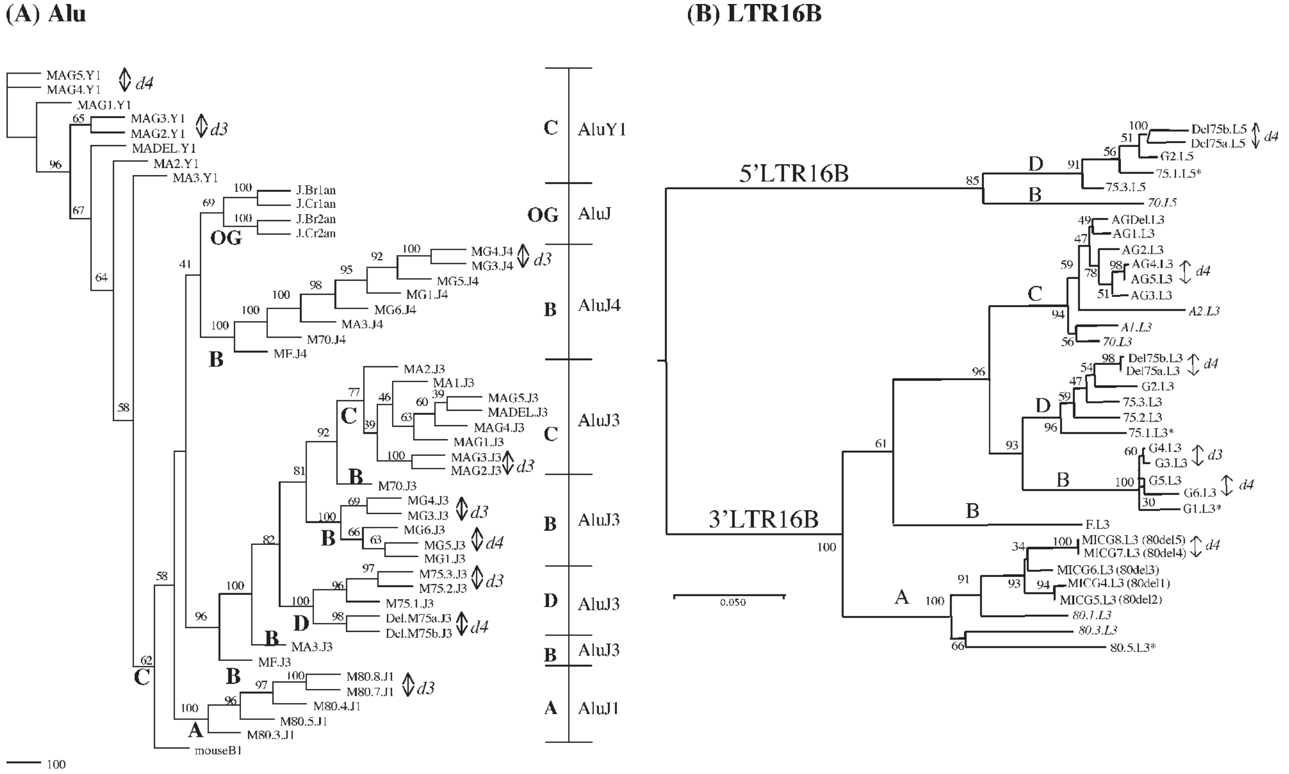

FIG. 6.—Phylogenetic tree of the *Alu* (*A*), and *LTR16B* (*B*) elements. (*A*) The Alu elements, *AluJ1*, *AluJ3*, *AluJ4*, and *AluY1* clustered into the duplicated categories A, B, C, and D that are shown within and on the right-hand side of the tree and correspond to the duplicon categories in table 1. The *Alu* sequences are labeled so that the first block of letters represent the name of the *Mamu* class I gene that shares the same duplicon as the *Alu* sequence followed by *Alu*-type, *J1*, *J3*, *J4*, or *Y1*. MADEL.J3 and MADEL.Y1 represent the *Alu* elements of duplicon C that has the *Mamu-AG* deletion. The phylogenetic tree was obtained and drawn from a ClustalX alignment of the entire *Alu* sequences using the NJ method and default parameters of PAUP*. The reliability of the trees was measured by bootstrap analysis with 1,000 replicates that are shown as percentage values. The scale bar is the number of substitutions per site. The duplicated pairs that are the inferred products of the most recent duplications (D3 and D4) are labeled *d3* and *d4* and indicated by the vertical double arrows. The letters OG (outgroup) within the trees and on the right-hand side of the trees represent the duplicated *Alu* sequences, *AluJ.Br1an*/*AluBr2an* and *AluJCr1an*/*AluJCr2an* in the human duplicons that have the *HLA-B* and *HLA-C* genes, respectively (Kulski et al. 1999*b*). The locations of the *AluJ1, J3, J4*, and *Y1* elements within the human duplicons have been previously described (Kulski et al. 1999*b*). The mouse *B1* sequence is used as an outgroup. (*B*) Phylogenetic analysis of the 5′ and 3′ *LTR16B* sequences of *ERV16* within different duplicons. The 23 sequences (401 sites per sequence) were aligned using ClustalX and phylogenetic trees were constructed by NJ using MEGA version 2.1. The reliability of all the trees was measured by bootstrap analysis with 1,000 replicates. The bootstraps are shown as percentage values. The scale bar is the number of substitutions per site. The intrinsic duplicated pairs (such as *G4.L3* and *G3.L3* within the duplicated multipartite genomic segments (see figure 8) are indicated by the vertical lines that are labeled *d3* and *d4*. The taxon names in italics indicate they are from ER1 and those marked by * are from the D1 region. The 3′ *LTR16B* sequences within the five highly fragmented category A duplicons are labeled as MICG4 to MICG8 with the deleted *Mamu-80* gene family members (80del1 to 80del5) indicated within parenthesis.

the *LTR16B* sequences near the *MICG4* to *MICG8* gene fragments grouped together to reveal that they are indeed part of the category A duplicons that have had their *Mamu-80* gene fragment completely deleted, as indicated in figure 1. Other *LTR16B* sequences from the two duplicons with deleted *Mamu* genes have grouped as expected; that is, *Mamu-Del75a* and *-Del75b* grouped with the category D duplicons and *Mamu-AGdel* grouped with the category C duplicons. The 5′ and 3′ *LTR16B* sequences within the *Mamu G2* duplicon grouped with the other sequences from the category D duplicons, confirming the hybrid nature of the *G2* duplicon. The sequences taken from the genomic regions that were predicted to be the most recent duplication sites (i.e., D3 and D4) generally clustered as pairs; for example, *AG4L3* paired with *AG5L3, G5L3* paired with *G6L3*, and *Del75b.L5* paired with *Del75a.L5* (fig. 6*B*). This tree also supports the view that the *LTR16B* sequences located within the duplicons of ER1 and D1 originated before those in the ER2. It is noteworthy that the solitary *LTR16B* sequence from the duplicon carrying the *Mamu-F* gene sequence is well separated from the other category B, C, and D duplicons. Therefore, this topology supports the prediction of the earlier evolutionary duplication models that duplicon *F* was probably fixed within the alpha block without contributing to any further duplications at a time well before the formation of most of the other MHC class I genes from the category B, C, and D duplicons (Kulski et al. 1999*b*; Kulski, Anzai, and Inoko 2004).

Evolution of MHC Class I Duplicons by Tandem Duplications and Inversions

The first detailed model for the evolution of the MHC class I duplicons by serial tandem duplications of multigenic units within the alpha block was proposed on the basis of a structural and phylogenetic analysis of duplicons and their organization within the human alpha block (Kulski et al. 1999*b*). It was more parsimonious than

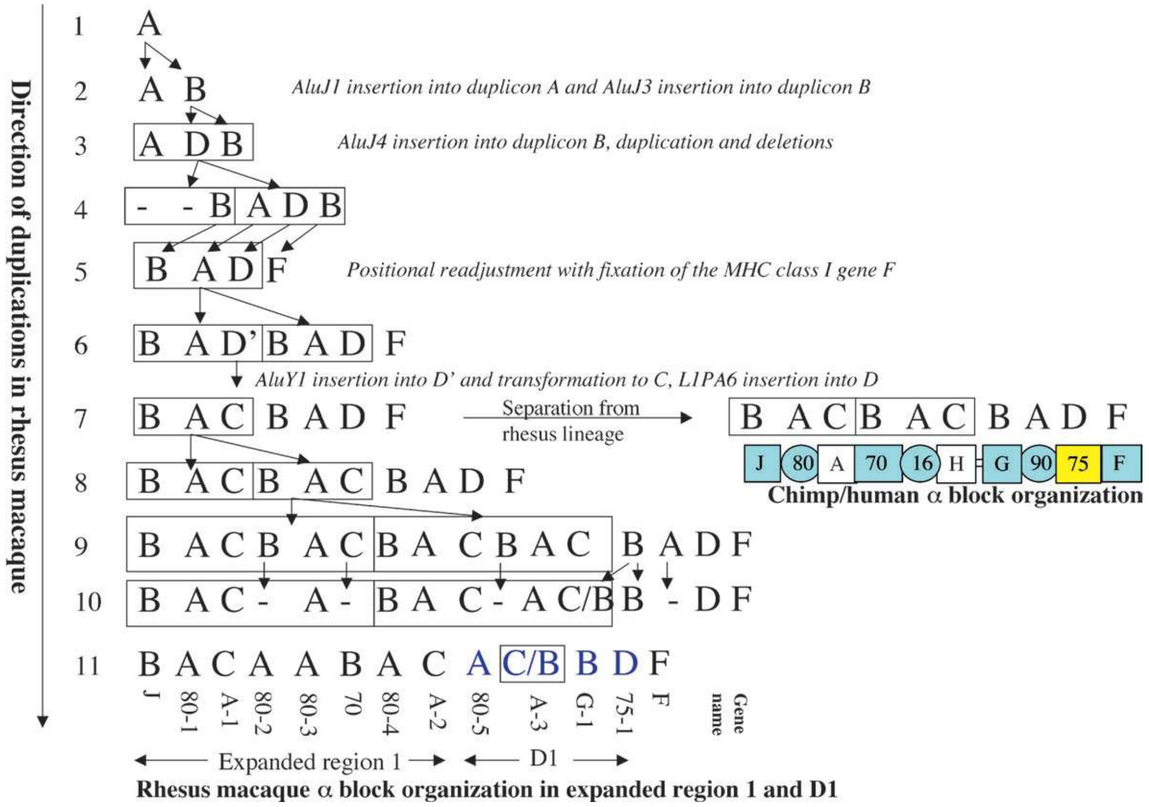
Direction of duplications in rhesus macaque
1 A
2 A B
AluJ1 insertion into duplicon A and AluJ3 insertion into duplicon B
3 A D B
AluJ4 insertion into duplicon B, duplication and deletions
4 - - B A D B
5 B A D F
Positional readjustment with fixation of the MHC class I gene F
6 B A D' B A D F
AluY1 insertion into D' and transformation to C, L1PA6 insertion into D
7 B A C B A D F
Separation from rhesus lineage
B A C B A C B A D F
J 80 A 70 16 H G 90 75 F
Chimp/human α block organization
8 B A C B A C B A D F
9 B A C B A C B A C B A C B A D F
10 B A C - A - B A C - A C/B B - D F
11 B A C A A B A C A C/B B D F
J 80-1 A-1 80-2 80-3 70 80-4 A-2 80-5 A-3 G-1 75-1 F
Gene name
Expanded region 1
D1
Rhesus macaque α block organization in expanded region 1 and D1

FIG. 7.—An inferred model for duplicon evolution and organization of ER1 within the alpha block of the rhesus macaque, chimpanzee, and human. The letters A to D represent the different MHC class segmental units or duplicons shown in figures 2 and 3. The single unit duplications are indicated by the arrows between each of the steps from 1 to 10. The boxed duplicons are a duplication unit composed of three or more duplicons that may have undergone a multiduplicon or block duplication. The dashes indicate deleted duplicons.

other models, such as the duplication-transition model (Shiina et al. 1999*b*), in that it required fewer duplication steps and no transposition events. In addition, the MHC tandem-duplication model proposed that *HLA-75* was the ancestor of *HLA-A* and *-H*, and *HLA-90* was the ancestor of *HLA-80* and *-16*. The duplication-transposition model also proposed a lineage for the same genes, but in the opposite direction.

The following five assumptions were made in reconstructing the evolutionary history of the MHC class I gene clusters by tandem duplication of the genomic segments or duplicons. (1) Tandem duplications are preferred to duplications that are associated with transpositions. (2) In reconstructing the most-parsimonious tandem duplication steps, a single block duplication of a multigenic duplication unit (polyduplicons) is preferred to a series of duplications of monogenic units. (3) The sequential order of segmental duplications should be based as much as possible on the duplicon structural features, organization, phylogeny, and evolutionary time. (4) The sequence identity and divergence between duplicons is time dependent. Presumably, two newly generated duplicons are almost identical in sequence soon after the duplication event. Deletions, insertions, point mutations, and other rearrangements will eventually create diversity within and between duplicons over time. (5) Recent tandem duplication steps will be more evident within the same species or between closely related species (e.g., human and chimpanzee) than between more distantly related species (e.g., human and rhesus macaque).

On the basis of the structure and phylogeny of the genes and retrotransposons within the four duplicon categories A to D, a duplication history for the evolution of the alpha block within the MHC of the rhesus macaque, chimpanzee, and human was reconstructed as outlined in figures 7 and 8. In the rhesus macaque, the duplicons *G, 80*, and *75* within the D1 region are inferred to be the progenitors of the duplications within the ER1 (fig. 7). Following on from the evolution of ER1, the duplicons *A, G, 80*, and *75* were then the progenitors for the duplication events within the ER2 (fig. 8).

Figure 7 shows the inferred model for duplicon evolution and organization of ER1 within the alpha block of the rhesus macaque, chimpanzee, and human. According to the model, expansion of ER1 within the alpha block has occurred by a series of imperfect tandem duplications of single-duplicons and polyduplicons represented by 11 distinct steps. The beginning of duplication from a primordial MHC class I duplicon that could be either a category A or B duplicon is shown in step 1. In step 2, the *AluJ1* element is inserted into the category A duplicon, and *AluJ3* is inserted into the B duplicon. In step 3, the category B duplicon has been duplicated to produce a category B and D duplicon. *AluJ4* is inserted into the category B duplicon. Between steps 3 and 4, duplication of the ADB duplicon combination has resulted in the loss of the categories A

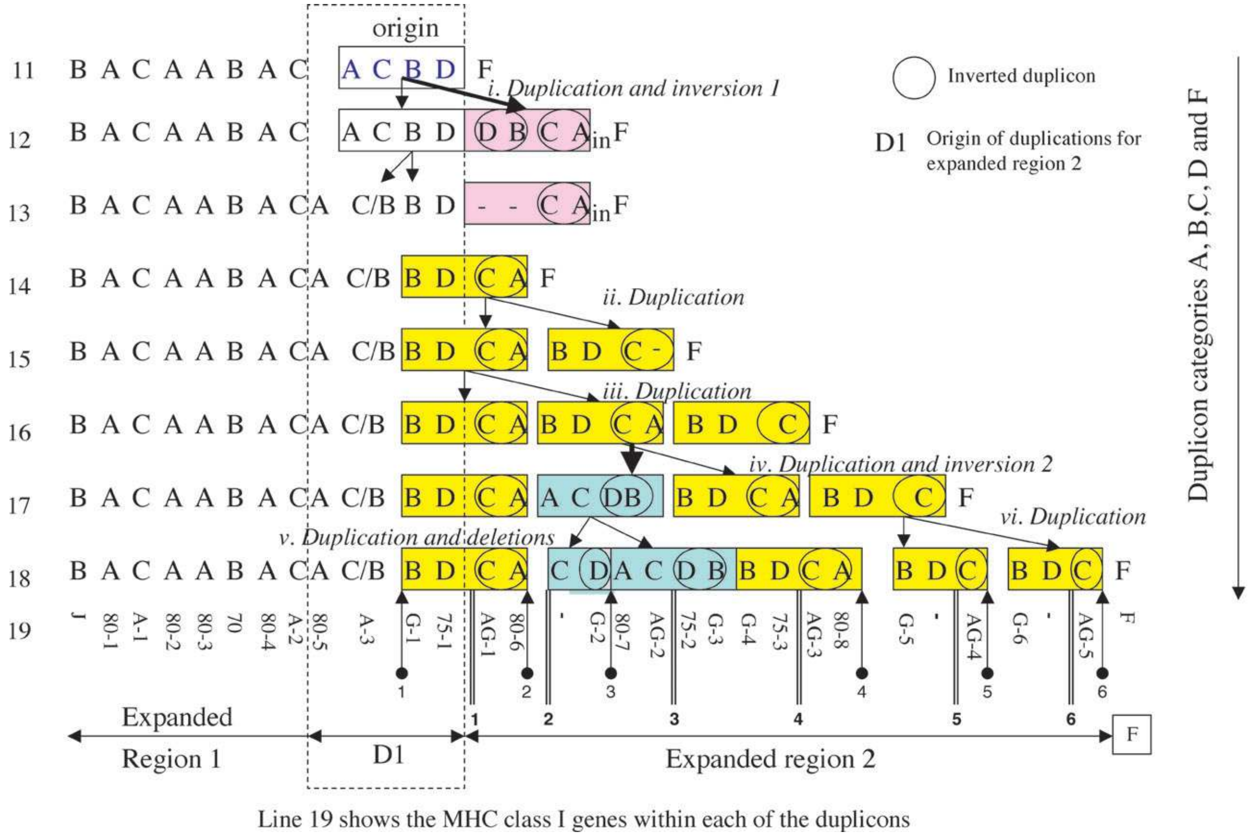
origin
11 B A C A A B A C A C B D F
i. Duplication and inversion 1
12 B A C A A B A C A C B D D B C A in F
13 B A C A A B A C A C/B B D - - C A in F
14 B A C A A B A C A C/B B D C A F
ii. Duplication
15 B A C A A B A C A C/B B D C A B D C - F
iii. Duplication
16 B A C A A B A C A C/B B D C A B D C A B D C F
iv. Duplication and inversion 2
17 B A C A A B A C A C/B B D C A A C DB B D C A B D C F
v. Duplication and deletions
vi. Duplication
18 B A C A A B A C A C/B B D C A C D A C D B B D C A B D C B D C F
19 J 80-1 A-1 80-2 80-3 70 80-4 A-2 80-5 A-3 G-1 75-1 AG-1 80-6 - G-2 80-7 AG-2 75-2 G-3 G-4 75-3 AG-3 80-8 G-5 - AG-4 G-6 - AG-5 F
1 2 3 4 5 6
1 2 3 4 5 6
Expanded Region 1
D1
Expanded region 2
F
Inverted duplicon
D1 Origin of duplications for expanded region 2
Duplicon categories A, B,C, D and F
Direction of duplications
Line 19 shows the MHC class I genes within each of the duplicons

FIG. 8.—An inferred model for the expansion of ER2 within the alpha block of the rhesus macaque. This is a continuation of figure 7 and the symbols are the same except that the circled duplicons, CA, C-, DB and D- and D, represent inverted duplicons. At step 19, the circled vertical arrows labeled 1 to 6 represent the locations of BALSL1-6. The vertical open lines labeled 1 to 6 represent the locations of the *ERV16* inverted breakpoints.

and D from one of the duplication products but the gain of a category B duplicon. In step 5, there is a positional readjustment with fixation of the MHC class I *F* gene as part of one of the category B duplicons. Also, the three units BAD are duplicated together to form the multipartite unit BAD′BAD. In between steps 6 and 7, there is an *AluY1* insertion into D′ and then the transformation of D′ into a category C duplicon. At about this time, *L1PA6* is inserted into the category D duplicon, which becomes the ancestor of the MHC class I *75* gene family. The minimum alpha block ancestral segmental structure appears to have been BAC/G/90/75/F (step 7) or BAC/BAC/G/90/75/F (step 8) before the separation of the rhesus macaque from the chimpanzee and human lineage. However, the *ERVK9* genomic markers that are found within some human and chimpanzee duplicons (Kulski, Anzai, and Inoko 2004) suggest that the human/chimpanzee lineage had probably separated from the rhesus macaque lineage at step 7 with a BAC duplication similar to step 8. The rhesus macaque probably separated from the human/chimpanzee lineage about 23 to 31 MYA (Takahata 2001; Glazko and Nei 2003). After this separation, the BAC and BAC′BAC structures were duplicated along the rhesus macaque pathway (steps 7 to 9), with the deletion of a few duplicons such as B, C, and A, as indicated in step 10. The basic class I gene organization within the ER1 of the rhesus macaque is shown in step 11. In addition, the *Mamu-A3* gene (step 11) appears to have become part of a duplicon that is a C/B hybrid as highlighted by the box in step 11. This chimeric C/B duplicon appears to have stemmed from a duplication of a B duplicon in the previous step.

Figure 8 shows the inferred expansion of ER2 within the alpha block of the rhesus macaque. This is shown as nine distinct steps (11 to 19), starting from duplication and inversion of the category ACBD combination in step 11 of the previous figure. Essentially, six tandem duplications, two of which also include inversions, explain the formation of 15 new class I genes in the macaque ER2. The first tandem duplication is derived from the conserved region (D1), with the duplication and inversion of ACBD resulting in the inverted AC and loss of the inverted BD as shown in the figure. The origin of duplications has been boxed by dashed lines and labeled as D1 because this is the first critical duplication and inversion that leads to the expansion of region 2; that is, ACBD is duplicated but with an inversion (DBCA) of one of its products. All other duplication, inversion, and deletion steps result from the duplication unit composed of the four duplicon units, BD(AC)inv, as shown in steps 14 to 19 of the figure. The MHC orthologous class I *F* gene is shown as the single F duplicon or F tail on the right-handed end (the telomeric end) of MHC class I duplicons at each step.

Only two inversion steps needed to be postulated to account for all of the inverted regions or nine inverted duplicons (figs. 1 and 8). Tandem duplications of the first two inverted regions can then account for the other four inverted regions. Therefore, the model in figure 7 infers two tandem duplications with an associated inversion and four tandem duplications with no inversions to account for 19 genes within D1 and the ER2. Overall, there appear to have been 12 tandem duplications (with or without inversions) and 12 deletions to explain the presence of at least 28 class I genes within the alpha block of the rhesus macaque, beginning with either a category A or category B progenitor duplicon and class I gene. In addition, this tandem duplication/inversion model helps to explain the organization of the six deleted category A duplicons, the deleted category C gene (*Mamu-AGdel*), and the two deleted category D genes (*Mamu-Del75a* and *-Del75b*), as seen within the dot-plot (fig. 1) and the *Alu* and *LTR16B* phylogenetic trees (fig. 6). In comparison, five tandem duplications and two deletions appear to be sufficient to explain the organization of the 10 to 11 MHC class I genes and duplicons within the alpha block of the human and the chimpanzee (Kulski, Anzai, and Inoko 2004).

The key step in explaining the presence of nine inverted class 1 genes and duplicons within the ER2 (fig. 1) is the duplication/inversion of the category ACBD duplicons within the D1 region and the deletion of the category BD duplicons from within the inversion (steps 12 and 13 in figure 8). Thereafter, the duplicon block combination of BDCA is inferred to have undergone five duplications whereby one of the duplications included an inversion and at least three single duplicons were deleted. Thus, two separate inversion steps are sufficient to explain the presence of nine inverted duplicons out of the total of 28 MHC class I duplicons within the alpha block of the rhesus macaque.

Conclusion

There has been at least 24 to 31 Myr of separation between the human and rhesus macaque lineages (Takahata 2001; Glazko and Nei 2003). Assuming that they all had started with the same number of genes at the time of their separation, then the rhesus macaque has produced duplicated genes within the alpha block at a rate approximately three times greater than in either the human or the chimpanzee. On the basis of phylogeny (Kulski, Anzai, and Inoko 2004) and the tandem duplication model, the alpha block organization of the human and the chimpanzee was almost completely formed before the divergence of rhesus macaque from the human/chimpanzee lineage. At least one additional category BAC duplication within the human/chimpanzee lineage probably occurred soon after its separation from the rhesus macaque lineage. Evidence for this additional step stems from the endogenous retroviral *MER9*/*ERVK9* insertion and an *AluY2* insertion that are present within the *Patr*/*HLA-A* and *-H* gene category C duplicons (Kulski, Anzai, and Inoko 2004) but are absent from the rhesus macaque duplicons of the same category. Nevertheless, since separating from the rhesus macaque lineage, the human and chimpanzee alpha block lineage has remained relatively stable for 24 Myr or more, whereas the alpha block of the rhesus macaque has been reorganized substantially, with additional expansions/deletions within the ER1 and ER2. In addition, inverted duplicons within the alpha block have been found so far only in the rhesus macaque and not in other primates. The tandem duplication model presented here also predicts that the organization of the gorilla alpha block (separation of approximately 7 Myr from the human lineage) should be essentially the same as in humans and chimpanzees (separation of approximately 6 Myr from the human lineage), whereas the orangutan (separation of approximately 13 Myr from the human) may have greater organizational differences. Although our tandem duplication model cannot predict the actual alpha block organization of the new world monkeys, it does suggest that most primates will have MHC class I duplicons within their alpha block represented by one or more of the four structural categories, A to D, with the inheritance of some or all of their characteristic class I genes and retroelements.

Acknowledgments

This work was supported by Grant-in-Aid for Scientific Research on Priority Areas (C) “Genome Science” from the Ministry of Education, Culture, Sports, Science and Technology of Japan. J.K.K. was on a JSPS Fellowship at Tokai University Medical School in Japan during the time of this study, and he thanks the Japan Society for the Promotion of Science.

Literature Cited

Amadou, C. 1999. Evolution of the MHC class I region: the framework hypothesis. Immunogenetics **49**:362–367.

Anzai, T., T, Shiina, T. Kimura et al. (18 co-authors). 2003. Comparative sequencing of human and chimpanzee MHC class I regions unveils insertions/deletions as the major path to genomic divergence. Proc. Natl. Acad. Sci. USA **100**: 7708–7713.

Avoustin, P., M. T. Ribouchon, C. Vernet, B. N’Guyen, B. Crouau-Roy, and P. Pontarotti. 1994. Non-homologous recombination within the major histocompatibility complex creates a transcribed hybrid sequence. Mamm. Genome **5**:771–776.

Dawkins, R. L., C. Leelayuwat, S. Gaudieri, G. Tay, S. Cattley, P. Martinez, and J. K. Kulski. 1999. Genomics of the major histocompatibility complex: haplotypes, retroviruses and disease. Immunol. Rev. **167**:275–304.

Feng, Q., J. V. Moran, H. H. Kazazian Jr., and J. D. Boeke. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell **87**:905–916.

Gaudieri, S., K. Habara, J. K. Kulski, R. L. Dawkins, and T. Gojobori. 2000. SNP profile within the human histocompatibility complex reveals an extreme and interrupted level of nucleotide diversity. Genome Res. **10**:1579–1586.

Gaudieri, S., J. K. Kulski, R. L. Dawkins, and T. Gojobori. 1999*a*. Different evolutionary histories in two subgenomic regions of the major histocompatibility complex. Genome Res. **9**:541–549.

———. 1999*b*. Extensive nucleotide variability within a 370 kb sequence from the central region of the major histocompatibility complex. Gene **238**:157–161.

Geraghty, D. E., B. H. Koller, J. A. Hansen, and H. T. Orr. 1992. The HLA class I gene family includes at least six genes and twelve pseudogenes and gene fragments. J. Immunol. **149**:1934–1946.

Glazko, G. V., and M. Nei. 2003. Estimation of divergence times for major lineages of primate species. Mol. Biol. Evol. **20**:424–434.

Hughes, A. L. 1995. Origin and evolution of HLA class I pseudogenes. Mol. Biol. Evol. **12**:247–258.

Jurka, J. 2000. Repbase Update: a database and an electronic journal of repetitive elements. Trends Genet. **16**:418–420.

Jurka, J., V. V. Kapitonov, P. Klonowski, J. Walichiewicz, and A. F. A. Smit. 1996*a*. Identification of new medium reiteration frequency repeats in the genomes of Primates, Rodentia, and Lagomorpha. Genetica **98**:235–247.

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

Klein, J., A. Sato, and C. O'hUigin. 1998. Evolution by gene duplication in the major histocompatibility complex. Cytogenet. Cell. Genet. **80**:123–127.

Kulski, J. K., T. Anzai, and H. Inoko. 2004. ERVK9, transposons and the evolution of MHC class I duplicons within the alphablock of the human and chimpanzee. Cytol. Genome Res (in press).

Kulski, J. K., and R. L. Dawkins. 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, J. K., S. Gaudieri, M. Bellgard, L. Balmer, H. Inoko, and R. L. Dawkins. 1997. The evolution of MHC diversity by segmental duplication and retrospection of retroelements. J. Mol. Evol. **45**:599–609.

Kulski, J. K., S. Gaudieri, and R. L. Dawkins. 2000*a*. Transposable elements and the metamerismatic evolution of the HLA class I region. Pp. 158–177 *in* M. Kasahara, ed. The major histocompatibility complex: evolution, structure, and function. Springer-Verlag, Berlin.

———. 2000*b*. Using Alu J elements as molecular clocks to trace the evolutionary relationships between duplicated HLA class I segments. J. Mol. Evol. **50**:510–519.

Kulski, J. K., S. Gaudieri, H. Inoko, and R. L. Dawkins. 1999*a*. Comparison between two HERV-rich regions within the major histocompatibility complex. J. Mol. Evol. **48**:675–683.

Kulski, J. K., S. Gaudieri, A. Martin, and R. L. Dawkins. 1999*b*. Co-evolution of PERB11 (MIC) and HLA class I genes with HERV-16 and retroelements by extended genomic duplication. J. Mol. Evol. **49**:84–97.

Kulski, J. K., T. Shiina, T. Anzai, S. Kohara, and H. Inoko. 2002. Comparative genomic analysis of the MHC: the evolution of class I duplication blocks, diversity and complexity from shark to man. Immunol. Rev. **190**:95–122.

Kumar, S., K. Tamura, I. B. Jakobsen, and M. Nei. 2001. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics **17**:1244–1245.

Leelayuwat, C., M. Pinelli, and R. L. Dawkins. 1995. Clustering of diverse replicated sequences in the MHC: evidence for en bloc duplication. J. Immunol. **155**:692–698.

Shiina, T., G. Tamiya, A. Oka, N. Takishima, and H. Inoko. 1999*a*. Genome sequence analysis of the 1.8 Mb entire human MHC class I region. Immunol. Rev. **176**:193–199.

Shiina, T., G. Tamiya, A. Oka et al. (20 co-authors). 1999*b*. Molecular dynamics of MHC genesis unraveled by sequence analysis of the 1,796,938-bp HLA class I region. Proc. Natl. Acad. Sci. USA **96**:13282–13287.

Sonnhammer, E. L., and R. Durbin. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene **167**:GC1–10.

Smit, A. F. A. 1993. Identification of a new, abundant superfamily of mammalian LTR-transposons. Nucleic Acids Res. **21**: 1863–1872.

———. 1996. The origin of interspersed repeats in the human genome. Curr. Opin. Genet. Devel. **6**:743–748.

———. 1999. Interspersed repeats and other momentos of transposable elements in the human genome. Curr. Opin. Genet. Devel. **9**:657–663.

Smit, A. F. A., and A. D. Riggs. 1996. Tiggers and other DNA transposon fossils in the human genome. Proc. Natl. Acad. Sci. USA **93**:1443–1448.

Swofford, D. L. 1998. PAUP*: phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Sunderland, Mass.

Takahata, N. 2001. Molecular Phylogeny and Demographic History of Humans. Pp. 299–305 *in* P. V. Tobias, M. A. Ratth, J. Morri-Cecchi, and G. A. Doyle, eds. Humanity from African naissance to coming millennia-colloquia in human biology and palaeoanthroplogy. Firenze University Press, Firenze/Witwatersrand University, Johannesburg.

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

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

Takashi Gojobori, Associate Editor

Accepted July 16, 2004