J Mol Evol (1997) 45:599–609

JOURNAL OF MOLECULAR EVOLUTION

The Evolution of MHC Diversity by Segmental Duplication and Transposition of Retroelements

Jerzy K. Kulski,[1] Silvana Gaudieri,[1] Matthew Bellgard,[1] Lois Balmer,[1] Keith Giles,[1] Hidetoshi Inoko,[2] Roger L. Dawkins[1]

[1] Centre for Molecular Immunology and Instrumentation and the University of Western Australia, Perth, Western Australia
[2] School of Medicine, Tokai University, Isehara, Kanagawa, Japan

Received: 21 May 1997 / Accepted: 9 July 1997

Abstract. Sequence analysis of a 237 kb genomic fragment from the central region of the MHC has revealed that the HLA-B and HLA-C genes are contained within duplicated segments peri-B (53 kb) and peri-C (48 kb), respectively, and separated by an intervening sequence (IF) of 30 kb. The peri-B and peri-C segments share at least 90% sequence homology except when interrupted by insertions/deletions including Alu, L1, an endogenous retrovirus, and pseudogenes. The sequences of peri-B, IF, and peri-C were searched for the presence of Alu elements to use as markers of evolution, chromosomal rearrangements, and polymorphism. Of 29 Alu elements, 14 were identified in peri-B, 11 in peri-C, and 4 in IF. The Alu elements in peri-B and peri-C clustered phylogenetically into two clades which were classified as ''preduplication'' and ''postduplication'' clades. Four Alu J elements that are shared by peri-B and peri-C and are flanked by homologous sequences in their paralogous locations, respectively, clustered into a ''preduplication'' clade. By contrast, the majority of Alu elements, which are unique to either peri-B or peri-C, clustered into a postduplication clade together with the Alu consensus subfamily members ranging from platyrrhine-specific (Spqxcg) to catarrhine-specific Alu sequences (Y). The insertion of platyrrhine-specific Alu elements in postduplication locations of peri-B and peri-C implies that these two segments are the products of a duplication which occurred in primates prior to the divergence of the New World primate from the human lineage (35–44 mya). Examination of the paralogous Alu integration sites revealed that 9 of 14 postduplication Alu sequences have produced microsatellites of different length and sequence within the Alu 3′-poly A tail. The present analysis supports the hypothesis that HLA-B and HLA-C genes are products of an extended segmental duplication between 44 and 81 million years ago (mya), and that subsequent diversification of both genomic segments occurred because of the mobility and mutation of retroelements such as Alu repeats.

Key words: Retroelements — Segmental duplication — MHC — Diversity — Alu

Introduction

Gene duplications and subsequent diversification have clearly played a major role in the evolution and diversity of major histocompatibility complex (MHC) class II gene products (Svensson et al. 1996). Although the accumulation of point mutations is thought to be the basic mechanism generating diversity (Kimura 1979), gene conversion (Ohta 1982), overdominant selection (Hughes and Nei 1988), and frequency-dependent selection (Bodmer 1972) have been proposed to explain the high degree of polymorphism associated with the class I gene products. In addition, multiple insertions, amplifi-

Correspondence to: R.L. Dawkins; e-mail: dawkins@cmii.uwa.edu.au
Publication number 9703 of the Centre for Molecular Immunology and Instrumentation and the University of Western Australia, Perth Western Australia

cations, and translocations of retroelements and other genomic interspersed repeat sequences have probably contributed to the diversity of the MHC in humans and nonhuman primates.

Approximately 40–60% of the total primate genomic DNA is considered to be composed of repetitive interspersed retroelements (Alu, L1, endogenous retroviruses, LTRs, and pseudogenes) that transpose via an RNA intermediate (Weiner et al. 1986), medium interspersed repeats (MERs) (Jurka 1990), minisatellites (Jeffreys et al. 1991) and microsatellites that can have a basic repeat unit of 6bp or less (Jurka and Pethiyagoda 1995; Arcot et al. 1995). Interspersed repeat sequences have been implicated in a variety of DNA processes including gene duplication, transcription, conversion, recombination, and replication (Weiner et al. 1986; Brosius, 1991; Erickson et al. 1992). The identification of retroelements like Alu, L1 and endogenous retroviruses near the junctions of duplicated genes emphasize their importance in the evolution of gene families. For example, Alu elements and L1 retrotransposons have been implicated in the duplication of the haptoglobin (Erickson et al. 1992) and gamma-globin genes (Fitch et al. 1991), respectively.

Of the large variety of retroelement superfamilies in the human genome (Weiner et al. 1986), the family of Alu retroelements are well characterized in terms of evolutionary history and sequence (Britten et al. 1988; Labuda and Striker 1989; Jurka and Milosavljevic 1991; Jurka 1993; Shen et al. 1991), they occur in high abundance (500,000 copies per genome), and they generate microsatellites which may be a source of polymorphism (Epstein et al. 1990; Batzer et al. 1995; Arcot et al. 1995). Alu elements are dimers of about 300 bp which are thought to have evolved from a portion of the 7SL RNA gene over 65 million years ago (mya), and then developed through successive waves of fixation with primate lineage history (Ullu and Tschudi 1984; Britten et al. 1988; Quentin 1988; Shen et al. 1991; Kapitonov and Jurka 1996). The family of Alu elements can be divided into at least four subfamilies that have different historical ages ranging between the ancient Alu J, the platyrrhine-specific (Alu S), the catarrhine-specific (Alu Y), and the human-specific (Alu Ya/Yb) subfamilies (Batzer et al. 1996; Kapitonov and Jurka 1996). Consequently, Alu elements have been used as DNA markers to compare duplicated gene products between and within species (Fitch et al. 1991; Erickson et al. 1992; Hardison and Miller 1993). Specifically, they have been used to analyze the evolution of HLA class II genes that may have resulted from gene duplication events in primate history (Mnukova-Fajdelova et al. 1994; Satta et al. 1996).

The present study was undertaken to (1) analyze 237 kb of genomic sequence for evidence that the HLA-B and HLA-C loci were products of a segmental duplication, and (2) identify and determine the evolutionary characteristics of Alu and other retroelements that may be associated with the duplication of the ancestor of HLA-B and HLA-C.

Materials and Methods

A sequence covering 237 kb from the PERB11.1 gene centromeric of HLA-B to the telomeric region approximately 90.8 kb beyond HLA-C was obtained from a YAC clone (T109) by cosmid cloning and shotgun sequencing (GenBank accession numbers D83543, D83769, D83770, D83771, D83956, D83957, and D84394) (Mizuki et al. 1997).

Dot plot matrix analyses with varying stringencies were performed using the programs compare and dot plot from the GCG package v8 (Genetics Computer Group). Alu sequences were identified in the contiguous sequence by BLASTn (NCBI) searches in the repetitive units database of GenBank; PYTHIA v2.5 (pythia@anl.gov) for identification of human repetitive DNA; and specific searches for the Alu motif TCCAGCCTGGG which is approximately 110 bp downstream from an Alu internal poly A tract. Other retroelements (L1, LTR, endogenous retrovirus), MERs, and pseudogenes were identified by BLASTn and BLASTx searches of the databases GenBank and Swissprot, respectively. The consensus sequences for Alu J, RNA 7SL (monomer [a] and artificial dimer [b]), and Alu Sb2 were from Jurka and Smith (1988) and Jurka (1993), respectively, and Alu consensus sequences for Alu PS, CS, HS1, and HS2 subfamilies were from Shen et al. (1991). Alu sequences were reclassified using the public Censor server (censor@charon.lpi.org). The recent standardized nomenclature for Alu repeats is used here to reclassify Alu PS as Alu S, Alu CS as Alu Y, Alu HS1 as Alu Ya5, Alu HS2 as Alu Ya8, and Alu Sb2 as Alu Yb8 (Batzer et al. 1996; Kapitonov and Jurka 1996). Alu sequence alignments were performed using CLUSTALw (GCG v8), and the phylogenetic analysis was performed using the program DNAml from the PHYLIP 3.5c package. The sequences flanking Alu repeats were aligned with paralogous sequences in peri-B or peri-C and examined using a program (integration) developed within the department.

The Alu elements identified in this analysis have been given a code name, such as BF1, IFR2 or CR1, where the first letters B, IF or C represent the location of an Alu on the DNA segment containing the peri-B, intervening fragment or peri-C sequence, respectively; followed by the letter F or R that represents the forward or reverse strand, respectively; followed by a number (1–7) that identifies the sequential order of the Alu elements along the forward or reverse strand of a particular segment. For some Alu elements, such as CF1AN, the letters AN represent an ancestral Alu found in a paralogous location in peri-B and peri-C.

Results and Discussion

Identification of Pseudogenes and Retroelements in Duplicated Segments Containing the HLA-B and HLA-C Genes

Dot plot analysis of a 237-kb genomic sequence shows that HLA-B and HLA-C coding regions are contained within duplicated segments peri-B (53,168 bp) and peri-C (47,895 bp), respectively, which are separated by an intervening sequence (IF) of 30 kb (Fig. 1A). The peri-B and peri-C sequences share at least 90% homology except in regions interrupted by indels including interspersed retroelements and pseudogenes. The homologous peri-B and peri-C sequences in Fig. 1A are considered to represent the genomic segment prior to

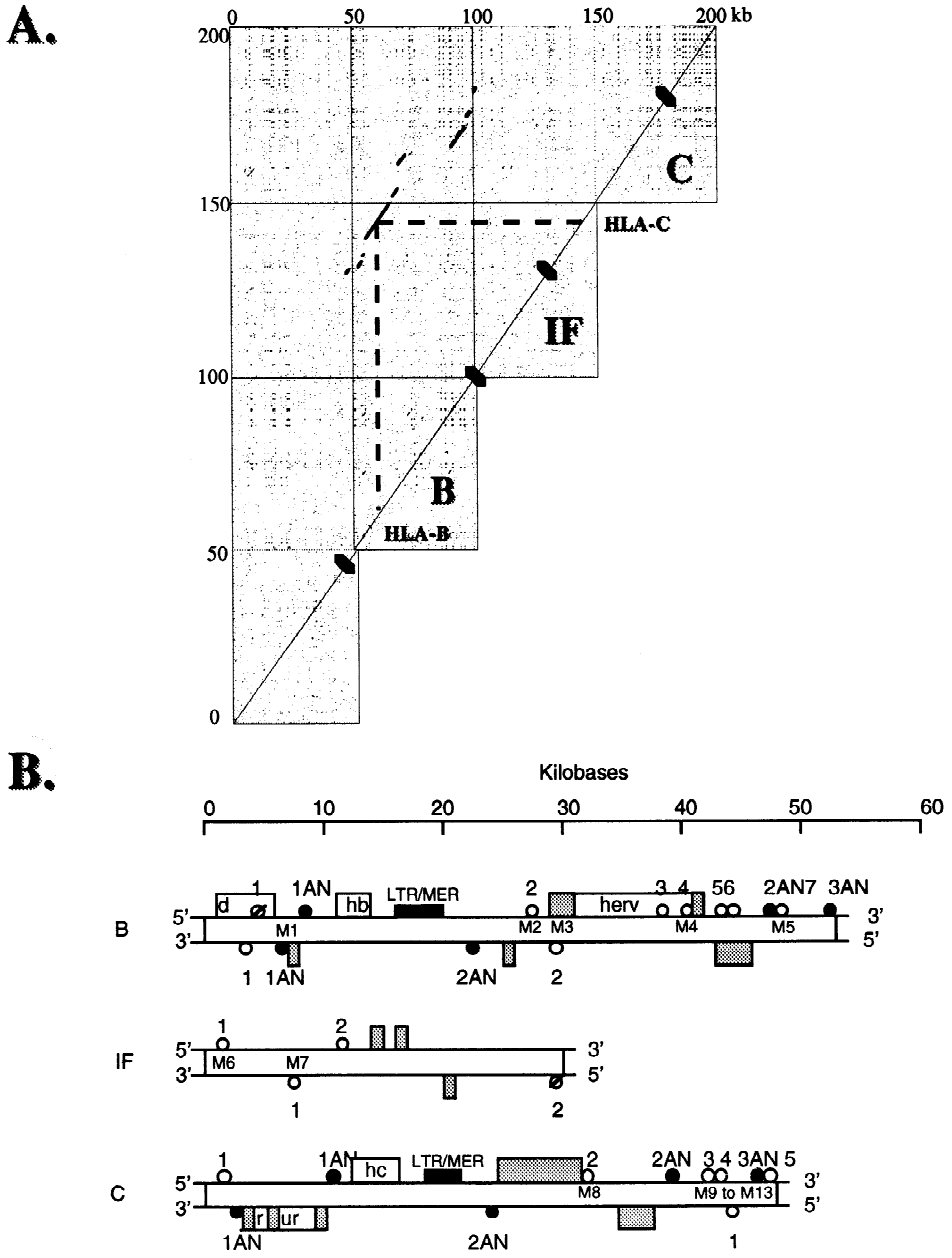

Fig. 1. **A** Dot plot analysis of 200 kb of genomic sequence containing the HLA-B and -C loci for evidence of sequence similarity and duplication. Dots indicate 75 or greater matches in a window of 100. Sequential dots are indicated by a black line which is interrupted by large insertions/deletions. The region is separated into B (containing the peri-B segment), IF (intervening region), and C (containing the peri-C segment). The position of the gene loci of HLA-B and HLA-C is indicated by the dashed line. **B** The location of interspersed retroelements identified in 140 kb of genomic DNA fragment containing the contiguous peri-B (B), intervening fragment (IF), and peri-C (C) sequences. Solid circles are preduplication (AN) Alu retroelements, crossed circles are possible preduplication Alu elements, and open circles are postduplication Alu elements. AN represents Alu elements present in the paralogous loci of the peri-B and peri-C duplicated segments. The Alu numbers above or below the circles are given according to their location along the forward (5′-3′) or reverse (3′-5′) strand of B, IF, and C, respectively. The shaded rectangles represent fragmented sequences of L1 retroelements. The open rectangles are the location of pseudogenes, dihydrofolate reductase (d), and human endogenous retrovirus (HERV) on peri-B, and ribosomal protein L3 (r) and ubiquitin protease (ur) on peri-C. The open rectangle labeled ''hb'' is the locus for HLA-B and the one labelled ''hc'' is the locus for HLA-C. The solid rectangles include an LTR9 and fragments of MERS 21, 4, and 1. M1 to M13 are the approximate locations of microsatellites that are associated with the Alu repeats described in Table 1.

duplication and subsequent insertion of mobile retroelements. We reconstructed the putative preduplication sequence (30 kb) by removing indels from peri-B and peri-C and inserting the appropriate ambiguities. Since duplication, the percentage of nucleotide differences (substitutions) varies between 1 and 23% whereas indels vary between 0 and 4 events per 100 nucleotides in a relatively constant pattern throughout the preduplication sequence (Gaudieri et al., submitted).

The location of retroelements in peri-B, IF, and peri-C

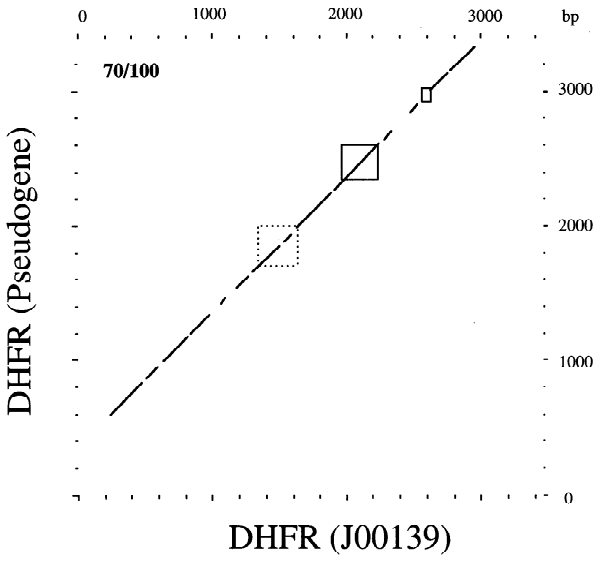

Fig. 2. Dot plot analysis (GCG v8, Wisconsin) at 70% stringency of the dihydrofolate reductase (DHFR) gene (J00139) and the inserted pseudogene in peri-B. The location of the Alu sequences (BF1, BR1, and fragment) are indicated by the dashed box, reverse orientation, and closed box, forward orientation. The Alu units are common to both sequences suggesting that the Alu units were inserted prior to the transposition of the dihydrofolate reductase sequence to peri-B.

are shown in Fig. 1B. The pseudogene in peri-B is related in sequence to dihydrofolate reductase (DHFR) (J00139) whereas the pseudogenes in peri-C are homologues of the ribosomal protein L3 (RPL3) gene (M90054) (Mizuki et al. 1997) and a ubiquitin protease (UR) gene (D29956) (Nomura et al. 1994a; Nomura et al. 1994b). The majority of L1 sequences in peri-B and peri-C are fragmented remnants (65 bp–2,809 bp) of a progenitor L1 retroelement of 6,140 bp (Jurka, 1989), and their location, classification, and fragment size will be presented elsewhere. A relatively intact L1 retroelement (5.9 kb) in peri-C appears to have been disrupted by a human endogenous retroviral (HERV) insertion in peri-B between the Alu markers, R2AN and F2AN. L1 fragments flanking the HERV in peri-B, which also contains a cluster of Alu elements within and near the viral 3′ LTR sequence, may represent the remains of the large L1 sequence in peri-C which subsequently was used as a receptor for the insertion of HERV. Some L1 fragments located between the Alu markers, CR1AN and CF1AN, in peri-C also appear to have been receptors for the insertion of the neighboring RPL3 and UR pseudogenes. An Alu element (BF1) and Alu fragments (Fig. 2) in the DHFR pseudogene may have served as receptors for integration in peri-B. In addition, a number of fragments of LTRs (40–441 bp in length) and MERs (24–857 bp) distributed in peri-B and peri-C segments ranged between 3 and 4, and 5 and 8 copies per segment, respectively.

Pre- and Postduplication Alu Sequences

Twenty-nine Alu retroelements were identified in peri-B, IF, and peri-C and their location and designations are shown in Fig. 1B. Fourteen Alu elements were found in peri-B, 11 in peri-C, and 4 in IF. This represents an average density of approximately one element per 3.8 kb of sequence in peri-B and one event per 4.4 kb of sequence in peri-C, which is the expected average density of the entire human genome (Britten et al. 1988) and the observed density for HLA-DR3 haplotypes (Mnukova-Fajdelova et al. 1994), but which is approximately twice the observed density in IF. The homologous peri-B and peri-C sequences contain five pairs of Alu elements (BF1AN/CF1AN, BF2AN/CF2AN, BR1AN/CR1AN, BR2AN/CR2AN, and BF3AN/CF3AN) that are present in their paralogous locations (Fig. 1B) due to integration prior to duplication of the genomic sequence. The paralogous Alu pairs and their 5′ and 3′ flanking sequences in peri-B and peri-C are shown in Fig. 3. In addition, a contiguous sequence that contains LTR9 and MERs 21, 4, and 1 is located in peri-B and peri-C between the HLA loci and R2AN (Fig. 1B). Of the preduplication Alu paralogs, three pairs are full length whereas two pairs are fragments or monomers of the full-length dimeric Alu sequence. In comparison with the preduplication paralogous Alu elements which are flanked by homologous sequences in peri-B and peri-C (Fig. 3), the postduplication Alu sequences are inserted in only one or other of the segments (Fig. 4).

The 29 Alu sequences within peri-B, IF, and peri-C were aligned against six consensus subfamily Alu sequences (Alu J, S, Y, Ya5, Ya8, and Yb8), the rodent repeat sequence B1, and the human 7SL gene (data not shown). The Alu elements grouped phylogenetically into two main clades divided by a member of the oldest consensus subfamily sequence Alu J (Fig. 5). One clade contains all of the intermediate and young Alu subfamily consensus sequences and most of the postduplication Alu elements from peri-B, IF and peri-C. The other clade has two single Alu elements (IFR2 and BF1), the 7SLa/7SLb sequence, and all of the preduplication Alu sequence pairs except for B1AN and C1AN which are a pair of fragmented Alu monomers (Fig. 3). BF1-Alu, which is in the preduplication clade, is part of the DHFR gene and has been introduced into peri-B with the integration of the DHFR pseudogene (Fig. 2). The Alu sequences in the postduplication clade have grouped into three different clusters, with all members of the young consensus subfamilies grouping into the same cluster (Fig. 5).

The pre- and postduplication Alu elements in peri-B, peri-C, and IF were grouped into subfamilies (Fig. 6) using the Censor program and nomenclature of Batzer et al. (1996). The Censor analysis revealed that IFR2, BF1, and all the preduplication Alu elements belong to the oldest consensus Alu subfamily sequence Alu J (81 mya). On the other hand, most of the postduplication Alu elements belong to the platyrrhine-specific [average age is 31–48 million years (myr)] and catarrhine-specific (19 myr) Alu subfamilies. The frequency of both pre- and

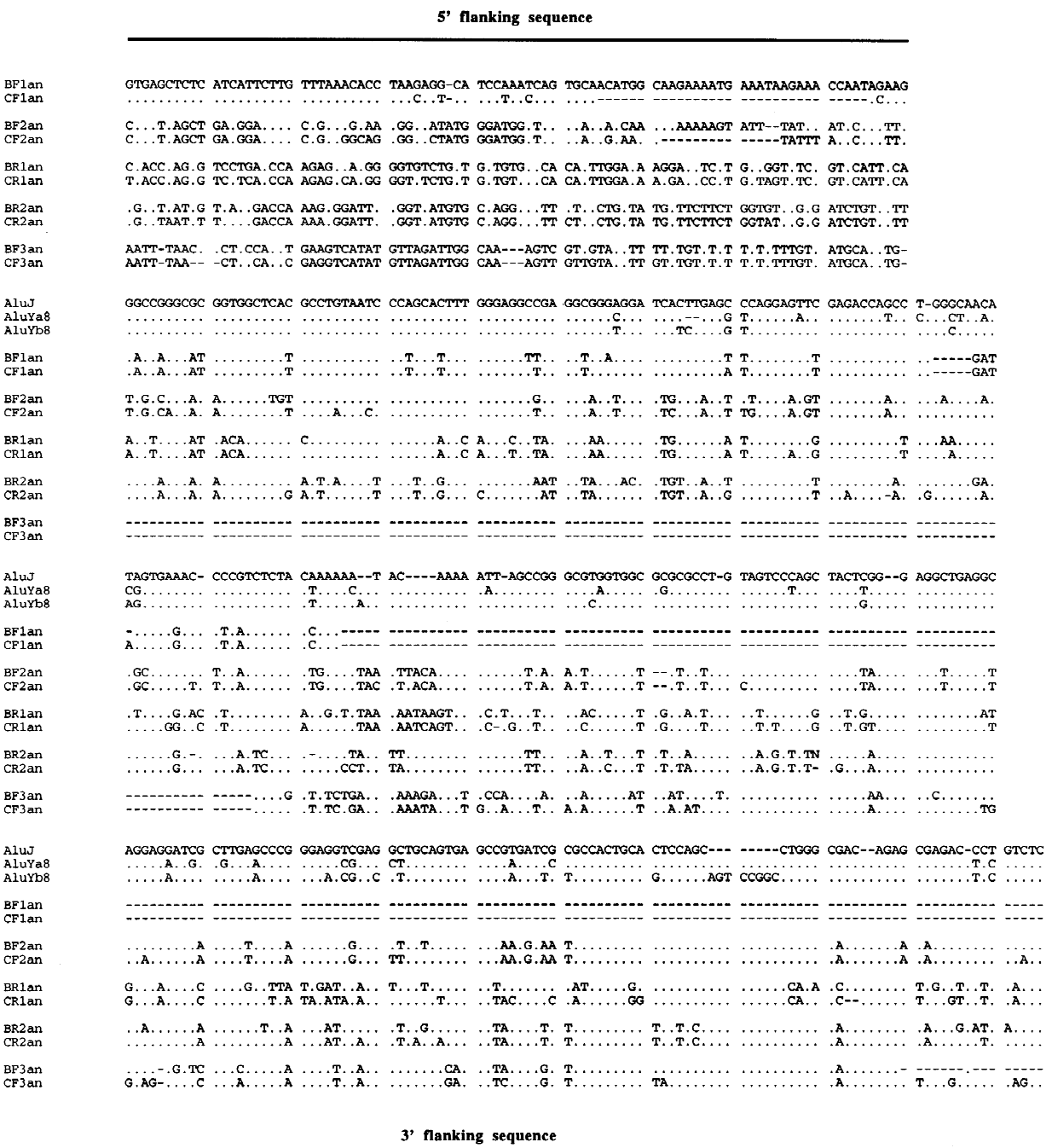

Fig. 3. Sequence alignment of paralogous Alu preduplication elements and their 5′ and 3′ flanking sequences in peri-B and peri-C. Preduplication Alu elements are aligned with Alu J, Alu Ya8, and Alu Yb8 consensus subfamily sequences. Sequence differences are shown relative to Alu J. Dots correspond to identical positions and hyphens mark gaps introduced to increase similarity. The first 95 nucleotides are the 5′ peri-B and peri-C sequences flanking the preduplication Alu elements (e.g., BF1AN and CF1AN on peri-B and peri-C segment, respectively). The last 100 nucleotides (position 401–500) are the 3′ peri-B and peri-C sequences flanking the preduplication Alu elements. The preduplication Alu pairs BF1AN and CF1AN, and BF3AN and CF3AN are Alu monomer structures. The microsatellite $(CAA)_5$ associated with BR1AN is boxed. Dots indicate identity to the top sequence and dashes indicate deletions.

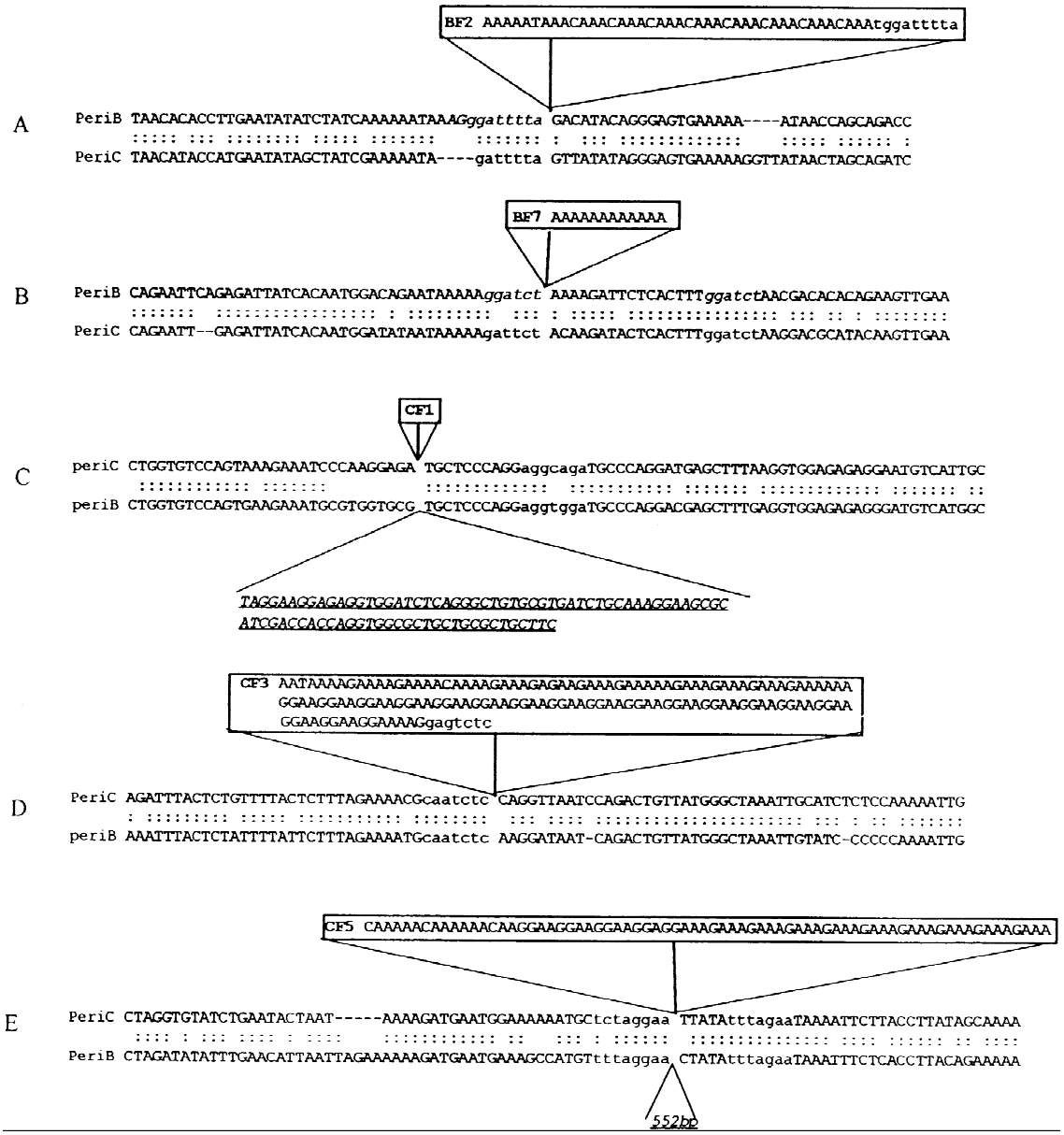

Fig. 4. Location of five different Alu integration sites (rows A–E) by sequence alignment of postduplication Alu elements (BF2, BF7, CF1, CF3, or CF5) with the paralogous sequence in peri-B or peri-C. The postduplication Alu elements (code name only-e.g., BF2) with their associated 3′ poly A tail or microsatellite sequence (see Table 2) are boxed. The Alu integration site in the upper sequence of the alignment between peri-B and peri-C is shown by the vertical line. The lower sequence of the paired alignments is the paralogous location in the duplicated segment. The sequence which is underlined and in italics in row C is present in peri-B but absent from peri-C in the paired alignment. In row E, a sequence of 552bp is present in peri-B but absent from peri-C. Nucleotides representing direct repeats surrounding Alu elements are in lower case.

postduplication Alu subfamilies in peri-B and peri-C was 44% for Alu J, 48% for Spqxcg, and 8% for Y, which is close to the expected average frequency for these Alu subfamilies in the human genome (Britten et al. 1988; Batzer et al. 1995). We have proposed a model, based on the nomenclature in Fig. 6, for the integration of Alu elements in the preduplication segment and the subsequent peri-B and peri-C segments (Fig. 7).

On the basis of the postduplication integration of 12 platyrrhine-specific Alu members into peri-B and peri-C, it can be estimated that the peri-B/peri-C preduplication segment had duplicated before the fixation of the oldest AluS members into the primate lineage at least 44 mya (Shen et al. 1991; Kapitonov and Jurka 1996).

Alu Integration Sites and Associated Microsatellites

Previous investigations of Alu integration sites have mostly relied on comparing the orthologous loci between different species (Arcot et al. 1995). In this study, we have identified Alu integration sites by aligning the paralogous location of postduplication Alu elements in duplicated segments peri-B and peri-C. Examples of integration sites for BF2, BF7, CF1, CF3, and CF5, and

Table 1. Association of microsatellites with Alu repeats located within peri-B, IF, and peri-C fragments

Alu code	Alu subfamily	Microsatellites[a]		
		Code	Location	Sequence
BR1AN	Jo	M1	3′ polyA tract	(CAA)5
BF2	Sg	M2	3′ polyA tract	(CAAA)8
BR2	Sc	M3	3′ polyA tract	(GA)3-6))3
BF4	Sc	M4	3′ polyA tract	(CA(3-7))5
BF7	Sg	M5	After 3′ polyA tract	(CA)3
IFF1	Y	M6	3′ polyA tract	(TA)13
IFR1	Y	M7	After 3′ polyA tract	(GGA)14 interrupted
CF2	Y	M8	3′ polyA tract	GA6GA7GA4
CF3	Sx	M9	3′ polyA tract	(GA(1-5)12(AAGG)18A4G2AG
CR1	Sx	M10	5′ of Alu	(CT)5
CR1	Sx	M11	3′ polyA tract	(CAA)7
CF3AN	Jo	M12	3′ polyA tract	TA6(TA3)2(TTAA)2
CF5	Sx	M13	3′ polyA tract	(AAGG)4AG(GAAA)10

[a] These microsatellites are associated with haplotype 8.1,X and their presence or sequence variation in other haplotypes is not known.

Table 2. Variation in sequence length of segments (I–IV) located between preduplication Alu elements in peri-B and peri-C

Segments between preduplication Alu elements	Sequence length (bp)			Sequence length difference between peri-B and peri-C
	anBC[a]	peri-B	peri-C	
I: R1AN to F1AN	02829	02829	08026	5,197
II: F1AN to R2AN (HLA)	13279	13440	14367	0,927
III: R2AN to F2AN	06875	24853	14910	9,943
IV: F2AN to R3AN	04771	05218	07122	1,904
Total length	27754	46340	44425	1,915

[a] The code name anBC represents the peri-B/peri-C genomic segment prior to duplication. The sequence length between Alu elements for anBC was calculated after removing indels from peri-B and peri-C and inserting the necessary ambiguities

Alu-associated microsatellites that are absent in paralogous loci are shown in Fig. 4. Many Alu elements were also found within the sequence of other repeat elements such as BF1 and BR1 in the pseudogene DHFR, BF3, and BF4 in an endogenous retroviral sequence, and BR2, BF5, BF6, and CF2 in various L1 fragments (data not shown).

The middle A-rich region or 3′ oligo (dA) tail of Alu retroelements are often polymorphic and associated with microsatellites (Epstein et al. 1990; Arcot et al. 1995). Although the Alu middle A-rich region has expanded in the preduplication BR1AN/CR1AN, the preduplication BF1AN/CF1AN (Fig. 2), and in the postduplication BF4 and BR2 (data not shown), no distinct microsatellites were identified in this region of the Alu sequences. By contrast, examination of Alu flanking sequences in peri-B, IF, and peri-C revealed the presence of microsatellites associated with 9 of 14 postduplication Alu sequences (Table 1). Figure 4 shows the Alu flanking sequences, associated microsatellites, and paralogous location of Alu integration for BF2, CF3, and CF5. The microsatellites are located mostly in the 3′ poly A tail of the Alu sequences and have been generated as oligopurine tracts (68–120 nucleotides), (CA/GT)*N* repeats, and (TA)*N* repeats. The microsatellite types show a segment-related bias with (CA)*N* in peri-B (4 of 5 types) and oligopurine/pyrimidine tracts in peri-C (4 of 6 types). Although (CA)*N* and (CT)*N* (where $N > 8$) are highly abundant in the human genome (Arcot et al. 1995; Stallings, 1995; Beckmann and Weber 1992) only short arrays ($N < 5$) were associated with Alu elements in peri-B and peri-C, respectively. The Alu elements associated with microsatellites belonged largely to the S subfamily. Surprisingly, the IFR1-Alu element in the IF has also generated a microsatellite that follows an intact poly A tail, suggesting that these microsatellites may have been formed or transferred during integration. However, examination of sequences within paralogous loci revealed that the microsatellite associated with the CF2-Alu was absent before integration (data not shown). In addition, a microsatellite sequence (CAA)5 was found to be present in the 3′ poly A tail of the preduplication BR1AN-Alu but absent in the paralogous location of the preduplication CR1AN-Alu sequence (Fig. 3). Whereas the CR1AN-Alu has seven adenines between the terminal end of the Alu sequence and the 3′-end of the direct repeat ACATT in peri-C, the BR1AN-Alu has generated a microsatellite (CAA)5 between the first six adenines and the direct repeat ACATT in the paralogous site of peri-B. Taken together, these observations support the hypothesis that

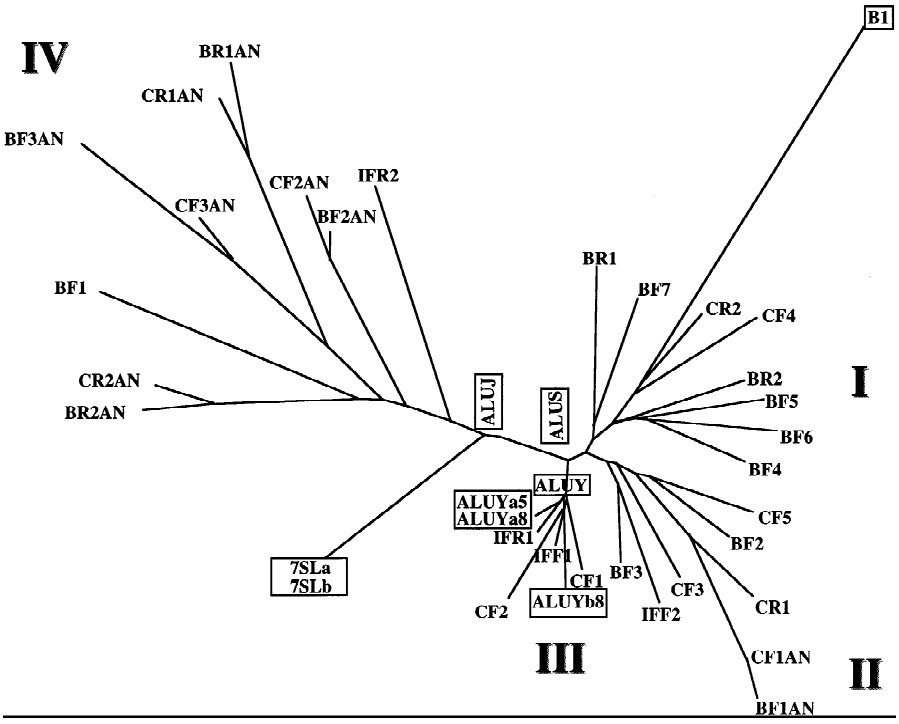

Fig. 5. Phylogenetic tree of preduplication and postduplication Alu elements in peri-B, IF, and peri-C. 7SL (synthetic dimer, 7SLb, and monomer, 7SLa), and consensus subfamily Alu sequences (Alu J, S, Y, Ya5, Ya8, and Yb8) were included in the analysis and are boxed. The rodent B1 sequence has been used as the outgroup. The sequence BF1 in cluster IV is part of the DHFR pseudogene and was inserted in the DHFR gene prior to its transposition to peri-B. The sequence BF1AN and CF1AN in cluster II are monomeric and do not cluster with the other preduplication dimer Alu sequences in cluster IV.

microsatellites associated with Alu elements had evolved after integration rather than as a result of integration into a previously occurring microsatellite site (Arcot et al. 1995). In addition, the microsatellites associated with Alu retroelements in peri-B, IF, and peri-C, if polymorphic, could be used as genetic markers for population studies (Epstein et al. 1990; Arcot et al. 1995), and for association with disease.

Expansion/Contraction of Genomic Sequence Located Between Preduplication Alu Elements

The preduplication Alu elements were used as markers to determine the degree of sequence expansion/contraction between peri-B and peri-C (Table 2). Overall, the sequence length of peri-B is longer than peri-C by 5,274 bp. Segment I, which is between R1AN and F1AN and upstream of the HLA genes, has increased 4,082 bp in peri-C because of the insertion of the RBL3 and UR pseudogenes.

Downstream of the HLA genes, segment III has expanded by 6818 bp in peri-B because of the HERV insertion, and segment IV has expanded in peri-C by 1,904 bp because of Alu insertions and microsatellite generation. By contrast, the sequence length between F1AN and R2AN (segment II), which contains the HLA-B and HLA-C genes, has changed by only 927 bp since duplication. In this regard, it appears that the segments between preduplication Alu elements upstream and downstream of segment II have acted as buffers to protect the HLA genes against deleterious sequence changes.

Evolution of the MHC Region in Association with Retroelements

It has previously been proposed that HLA-B and HLA-C are the products of a gene duplication because of their close proximity and sequence homology which is more closely related to each other than to HLA-A; and that the preduplication HLA-B and HLA-C gene locus is a product of a HLA-A gene duplication (see Lienert and Parham 1996). Although HLA-C is considered to be a recent addition to the MHC (Pohla et al. 1989), at least two HLA-B loci have been found in orangutans, gibbons, and rhesus monkeys (Chen et al. 1992; Boyson et al. 1996). Therefore, it can be inferred that one of the products of the B locus duplication had subsequently diverged to form the HLA-C locus that is now found in gorillas, chimpanzees, and humans. Our data support the view that the HLA-B and HLA-C genes had evolved from a duplication and that, contrary to expectations, the duplicated fragments are approximately 10 times longer than the gene loci. Moreover, the homologous peri-B and peri-C sequences have diverged and extended as a result of retroelement insertions with subsequent insertions/deletions and hypermutations. Because the evolutionary age of different Alu subfamily elements has been determined previously by either studying their presence in human and nonhuman primates (Shen et al. 1991) or estimating the interspecies DNA sequence divergence (Kapitonov and Jurka 1996), we used the Alu subfamily sequences as a molecular clock to characterize the evolutionary history of the peri-B and peri-C segments. The paralogous Alu elements located in the homologous re-

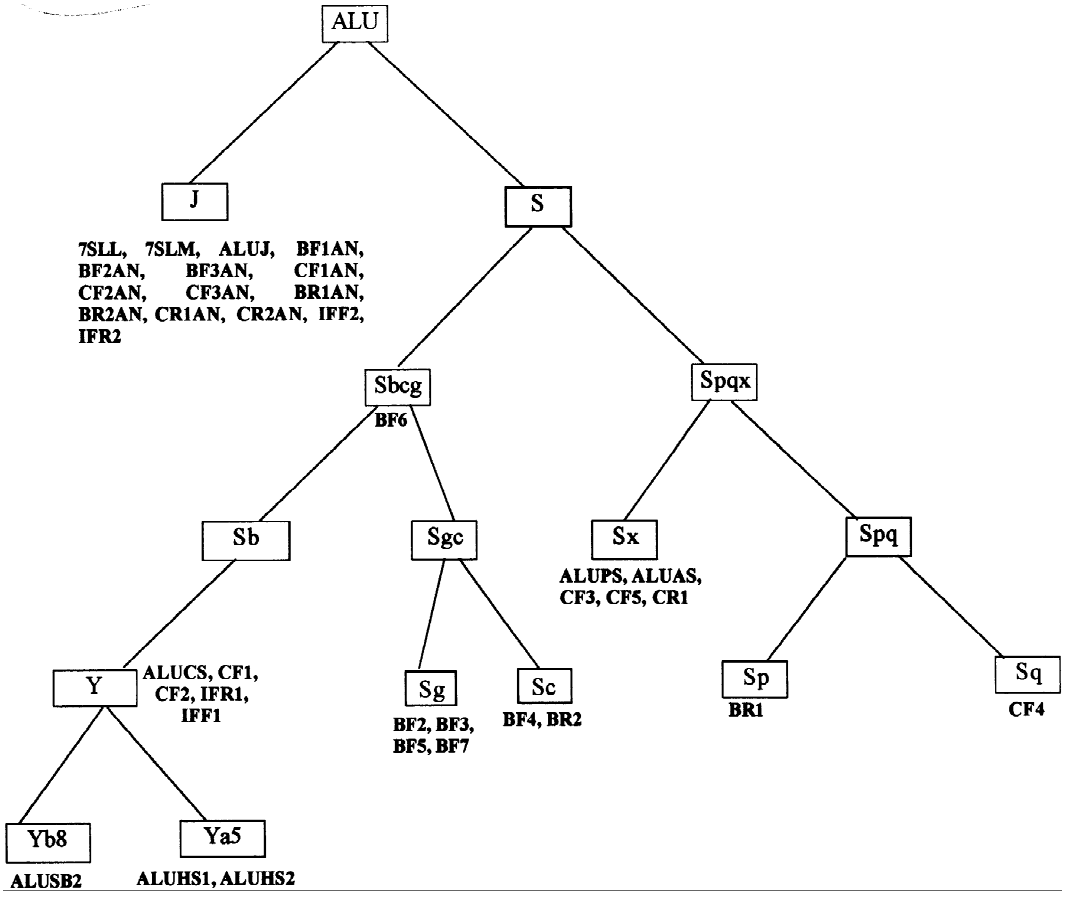

Fig. 6. Subfamily classification of preduplication and postduplication Alu elements. The 29 Alu sequences within peri-B, IF, and peri-C were grouped into subfamilies using the Pythia (pythia@anl.gov) software program and nomenclature of Batzer et al. (1996). 7SL (synthetic dimer, 7SLL, and monomer, 7SLM) and consensus subfamily sequences (Alu J, PS, AS, CS, HS1, HS2, and 5B2) were included in the analysis.

gions of the peri-B and peri-C sequences are most closely related to the oldest Alu J element which emerged 81 mya and has been fixed in primates for at least 55 myr. By contrast, the oldest Alu subfamily member that had integrated postduplication is Alu Sq which has been fixed in primates for 44 myr (Kapitonov and Jurka 1996) and had evolved before the split between New World and Old World monkeys (Shen et al. 1991). Therefore, the ancestor of peri-B and peri-C appears to have duplicated before New World monkeys had separated from the human lineage. The HERV in peri-B is closely related in sequence to a HERV-I sequence (Gaudieri et al. in preparation) that was found in the haptoglobin region of humans, apes, and Old World monkeys (Erickson et al. 1992) and possibly first inserted in the primate germ-line more than 25 mya (Shih et al. 1989). In this connection, the two HLA-B loci in rhesus monkeys, gibbon, and orangutans might be differentiated from HLA-C in gorillas, chimpanzee, and humans by the presence or absence of the HERV-I sequence.

The evolution of the MHC is complex and appears to have been affected by various interspersed retroelements including pseudogenes, LTRs, Alu, and L1 sequences. The evolutionary events that have led to the duplication of peri-B and peri-C are not known but homologous recombinations between repetitive elements flanking a single gene can be important mechanisms for gene duplication and generation of multigene families (Fitch et al. 1991; Erickson et al. 1992; Hardison and Miller 1993). The peri-B/peri-C preduplication segment has possibly evolved from duplication of a HLA-A genomic segment due to homologous unequal crossover, using one or more retroelements for recombination. In this regard there is an L1 sequence that is interrupted by peri-B, IF, and peri-C which suggests that the original insertion and subsequent duplication occurred within this L1 sequence. If peri-B and peri-C had evolved originally as a tandem duplication with subsequent insertion of IF, then some of the original junction points and associated recombination sequences may have been deleted or diluted out during expansion of the sequence. To reconstruct the evolutionary mechanisms and history for HLA class I gene duplication, regions surrounding orthologous peri-B and peri-C in nonhuman primates and other human HLA class I gene family members can be aligned and analyzed as more sequence information becomes available.

The results of our analysis suggest that HLA-B and HLA-C genes have evolved as a consequence of seg-

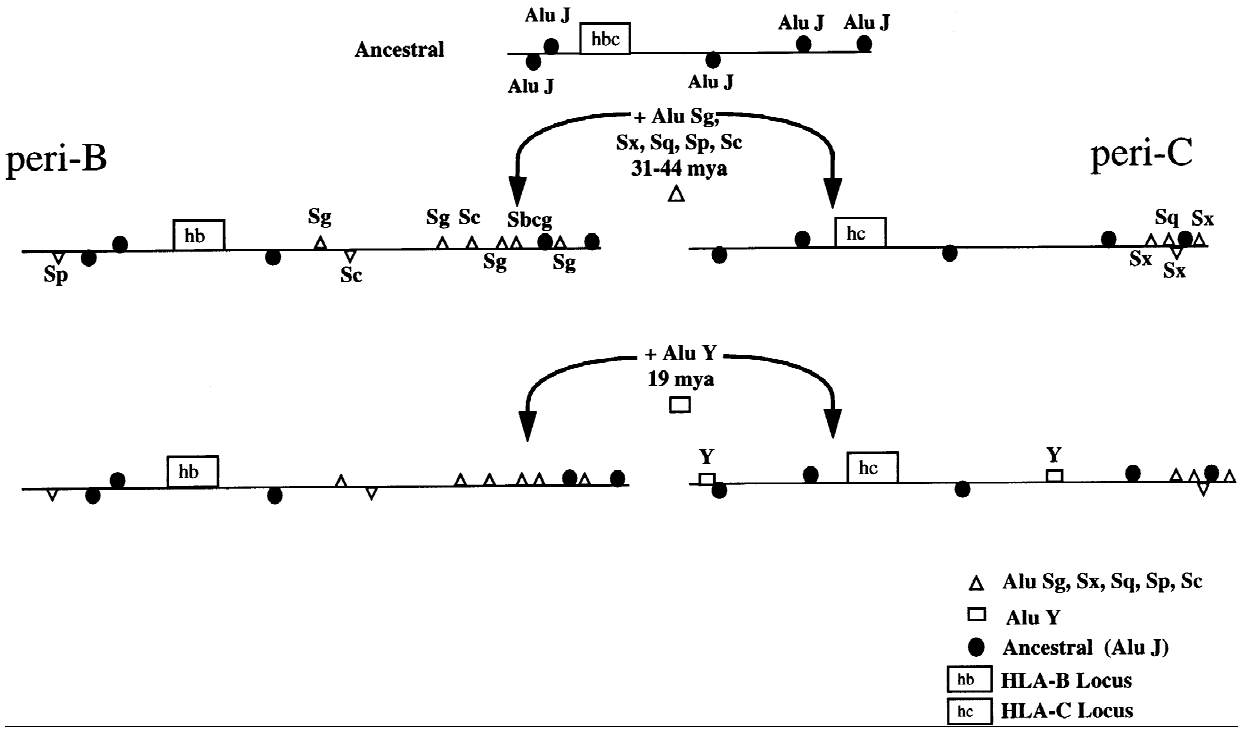

Fig. 7. Sequential integration of Alu elements following the duplication that results in the peri-B and peri-C segments. The preduplication peri-B and peri-C sequences contains five Alu J sequences (closed circles). The orientation of an Alu is indicated by its location on the sequence (above the line indicating forward orientation and below the line reverse). Alus of the platyrrhine-specific (Spqxcg) and catarrhine-specific (Y) subfamilies have been inserted into the peri-B and peri-C segments at particular points approximately 31–44 mya and 19 mya, respectively. These insertions are indicated by triangles (Alu Spqxcg) and squares (Alu Y). Positions of the HLA-B and HLA-C loci are given. There is a progressive increase in segment length following integration. This figure does not take into account the various other retroelements such as pseudogenes, HERVS, LTRs, and LINES. The BF1-AluJ within the DHFR pseudogene of peri-B is not included in this figure.

mental duplication of an ancestral peri-B and peri-C genomic region with secondary transpositions by retroelements leading to mutations, microsatellites, indels, and further diversity. Since duplication of the peri-B/peri-C segment, new insertions in peri-B appear to include the DHFR pseudogene, an endogenous retrovirus in a preduplication L1 fragment, and retroposition of at least 8 Alu elements. New insertions in peri-C appear to include the RBL3 and UR pseudogenes, L1 fragments, and a 3′ sequence expansion between CF2AN-Alu and CF3AN-Alu because of the insertion of at least three new Alu elements with an associated generation of microsatellites. Interspersed retroelements, because of their mobility and hypermutability, are an integral part of molecular drive (Dover 1982) which is both stochastic and directional as an evolutionary force in generating diversity in the MHC in primates. Since molecular drive involves DNA turnover processes such as replication, recombination, and gene conversion (Dover 1982; Dover 1993), mutations caused by retroelements in peri-B and peri-C may have acted in concert with gene conversion (Ohta 1982) or overdominate selection (Hughes and Nei, 1988) to generate and maintain polymorphism in both coding and noncoding regions of HLA-B and HLA-C loci.

Acknowledgments. This work was supported by the Immunogenetics Research Foundation and the National Health and Medical Research Council.

References

Arcot SS, Wang Z, Weber JL, Deininger PL, Batzer MA (1995) Alu repeats: a source for the genesis of primate microsatellites. Genomics 29:136–144

Batzer MA, Rubin CM, Hellmann-Blumberg U, Alegria-Hartman M, Leeflang EP, Stern JD, Bazan HA, Shaikh TH, Deininger PL, Schmid CW (1995) Dispersion and insertion polymorphism in two small subfamilies of recently amplified human Alu repeats. J Mol Biol 247:418–427

Batzer MA, Deininger PL, Hellmann-Blumberg U, Jurka J, Labuda D, Rubin CM, Schmid CW, Zietkiewicz E, Zuckerkandl E (1996) Standardized nomenclature for Alu repeats. J Mol Evol 42:3–6

Beckmann JS, Weber JL (1992) Survey of human and rat microsatellites. Genomics 12:627–631

Bodmer WF (1972) Evolutionary significance of the HLA system. Nature 19:139–145

Boyson JE, Shufflebotham C, Cadavid LF, Urvater JA, Knapp LA, Hughes AL, Watkins DI (1996) The MHC class I genes of the rhesus monkey: different evolutionary histories of MHC class I and II genes in primates. J Immunol 156:4656–4665

Britten RJ, Baron WF, Stout DB, Davidson EH (1988) Sources and evolution of human Alu repeated sequences. Proc Natl Acad Sci USA 85:4770–4774

Brosius J (1991) Retroposons—seeds of evolution. Science 251:753

Chen ZW, McAdam SN, Hughes AL, Dogon AL, Letvin NL, Watkins DI (1992) Molecular cloning of orangutan and gibbon MHC Class I cDNA. J Immunol 148:2547–2554

Dover G (1982) Molecular drive: a cohesive mode of species evolution. Nature 299:111–117

Dover GA (1993) Evolution of genetic redundancy for advanced players. Curr Opin Genet Devl 3:902–910

Epstein N, Nahor O, Silver J (1990) The 3′ ends of Alu repeats are highly polymorphic. Nucleic Acids Res 18:4634

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

Fitch DHA, Bailey WJ, Tagle DA, Goodman M, Sieu L, Slightom JL (1991) Duplication of the gamma-globin gene mediated by L1 long interspersed repetitive elements in an early ancestor of simian primates. Proc Natl Acad Sci USA 88:7396–7400

Hardison R, Miller W (1993) Use of long sequence alignments to study the evolution and regulation of mammalian globin gene clusters. Mol Biol Evol 10:73–102

Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167–170

Jeffreys AJ, MacLeod A, Tamaki K, Neil DL, Monckton DG (1991) Minisatellite repeat coding as a digital approach to DNA typing. Nature 354:204–209

Jurka J, Smith T (1988) A fundamental division in the Alu family of repeated sequences. Proc Natl Acad Sci USA 85:4775–4778

Jurka J (1989) Subfamily structure and evolution of the human L1 family of repetitive sequences. J Mol Evol 29:496–503

Jurka J (1990) Novel families of interspersed repetitive elements from the human genome. Nucleic Acids Res 18:137–141

Jurka J, Milosavljevic A (1991) Reconstruction and analysis of human Alu genes. J Mol Evol 32:105–121

Jurka J (1993) A new subfamily of recently retroposed human Alu repeats. Nucleic Acids Res 21:2252

Jurka J, Pethiyagoda C (1995) Simple repetitive DNA sequences from primates: compilation and analysis. J Mol Evol 40:120–126

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

Kimura M (1979) The neutral theory of molecular evolution. Sci Am 251:94–104

Labuda D, Striker G (1989) Sequence conservation in Alu evolution. Nucleic Acids Res 17:2477–2491

Lienert K, Parham P (1996) Evolution of MHC class I genes in higher primates. Immun Cell Biol 74:349–356

Mizuki N, Ando H, Kimura M, Ohno S, Miyata S, Yamazaki M, Tashiro H, Watanabe K, Ono A, Taguchi S, Sugawara C, Fukuzumi Y, Okumura K, Goto K, Ishihara M, Nakamura S, Yonemoto J, Kikuti YY, Shiina T, Chen L, Ando A, Ikemura T, Inoko H (1997) Nucleotide sequence analysis of the HLA class I region spanning the 237 kb segment around the HLA-B and -C genes. Genomics 42:55–67

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

Nomura N, Nagase T, Miyajima N, Sazuka T, Tanaka A, Sato S, Seki N, Kawarabayasi Y, Ishikawa K, Tabata S (1994a) Prediction of the coding sequences of unidentified human genes. II. The coding sequences of 40 new genes (KIAA0041-KIAA0080) deduced by analysis of cDNA clones from human cell line KG-1. DNA Res 1:223–229

Nomura N, Nagase T, Miyajima N, Sazuka T, Tanaka A, Sato S, Seki N, Kawarabayasi Y, Ishikawa K, Tabata S (1994b) Prediction of the Coding Sequences of Unidentified Human Genes. II The coding sequences of 40 new genes (KIAA0041-KIAA0080) deduced by analysis of cDNA clones from human cell line KG-1. DNA Res Suppl 1:251–262

Ohta T (1982) Allelic and nonallelic homology of a supergene family. Proc Natl Acad Sci USA 79:3251–3254

Pohla H, Kuon W, Tabaczewski P, Doerner C, Weiss EH (1989) Allelic variation in HLA-B and HLA-C sequences and the evolution of the HLA-B alleles. Immunogenetics 29:297–307

Quentin Y (1988) The Alu family developed through successive waves of fixation closely connected with primate lineage history. J Mol Evol 27:194–202

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 RM, Batzer MA, Deininger PL (1991) Evolution of the master Alu gene(s). J Mol Evol 33:311–320

Shih A, Misra R, Rush MG (1989) Detection of multiple, novel reverse transcriptase coding sequences in human nucleic acids: relation to primate retroviruses. J Virol 63:64–75

Stallings RL (1995) Conservation and evolution of $(CT)_n/(GA)_n$ microsatellite sequences at orthologous positions in diverse mammalian genomes. Genomics 25:107–113

Svensson AC, Setterblad N, Pihlgren U, Rask L, Andersson G (1996) Evolutionary relationship between human major histocompatibility complex HLA-DR haplotypes. Immunogenetics 43:304–314

Ullu E, Tschudi C (1984) Alu sequences are processed 7SL RNA genes. Nature 312:171–172

Weiner AM, Deininger PL, Efstratiadis A (1986) Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu Rev Biochem 55:631–661