J Mol Evol (2000) 50:391–396
DOI: 10.1007/s002399910042

JOURNAL OF MOLECULAR EVOLUTION



# Duplication and Diversification of the Apolipoprotein CI (APOCI) Genomic Segment in Association with Retroelements

**Elizabeth M. Freitas,[1,2] Silvana Gaudieri,[1,4] Wen Jie Zhang,[1] Jerzy K. Kulski,[1] Frank M. van Bockxmeer,[3,5] Frank T. Christiansen,[2,3] Roger L. Dawkins[1]**

[1] Centre for Molecular Immunology and Instrumentation, University of Western Australia, Nedlands, Western Australia
[2] Department of Clinical Immunology and Biochemical Genetics, Royal Perth Hospital, Perth, Western Australia
[3] University Department of Pathology, University of Western Australia, Nedlands, Western Australia
[4] Center for Information Biology, National Institute of Genetics, Mishima, Japan
[5] Department of Biochemistry, Royal Perth Hospital, Perth, Western Australia

Received: 17 July 1999 / Accepted: 6 November 1999

**Abstract.** We have previously shown that several multicopy gene families within the major histocompatibility complex (MHC) arose from a process of segmental duplication. It has also been observed that retroelements play a role in generating diversity within these duplicated segments. The objective of this study was to compare the genomic organization of a gene duplication within another multicopy gene family outside the MHC. Using new continuous genomic sequence encompassing the APOE-CII gene cluster, we show that APOCI and its pseudogene, APOCI′, are contained within large duplicated segments which include sequences from the hepatic control region (HCR). Flanking Alu sequences are observed at both ends of the duplicated unit, suggesting a possible role in the integration of these segments. As observed previously within the MHC, the major differences between the segments are the insertion of sequences (approximately 200–1000 bp in length), consisting predominantly of Alu sequences. Ancestral retroelements also contribute to the generation of sequence diversity between the segments, especially within the 3′ poly(A) tract of Alu sequences. The exonic and regulatory sequences of the APOCI and HCR loci show limited sequence diversity, with exon 3 being an exception. Finally, the typing of pre- and postduplication Alus from both segments indicates an estimated time of duplication of approximately 37 million years ago (mya), some time prior to the separation of Old and New World monkeys.

**Key words:** Apolipoprotein CI — Hepatic control region — Duplication — Alu — Retroelements — Diversity

*Correspondence to:* Elizabeth Freitas, Department of Clinical Immunology, P.B. Box X2213, Perth WA 6001, Western Australia. *e-mail:* liz@dci.uwa.edu.au

The apolipoproteins (APO) are a family of proteins involved primarily in the transport of lipids throughout the body. They are a multicopy gene family whose genes are localized within two clusters on human chromosomes 11 and 19, proposed to have arisen by a process of partial and complete gene duplication (Luo et al. 1986; Lauer et al. 1988). One of these genes, apolipoprotein CI (APOCI), encodes for a 57-amino acid (6.6-kDa) protein that is expressed primarily within the liver and inhibits the receptor-mediated uptake of triglyceride-rich lipoproteins. In addition, APOCI is a potent activator of lecithin–cholesterol acyltransferase (LCAT) and an inhibitor of lipoprotein lipase (reviewed by Jong et al. 1999). Within the APOE-CII cluster (chromosome 19q13.2) a duplicated copy of APOCI has been identified. Known as APOCI′, this copy is located approximately 6.5 kb downstream of APOCI. Both loci share a similar genomic structure, with APOCI′ containing a premature stop codon within exon 3. APOCI has been shown to encode the functional protein, while APOCI′ expression has not

been detected, indicative of a pseudogene (Lauer et al. 1988). Interestingly, a duplicated APOCI-like sequences has also been detected in the baboon genome (Pastorcic et al. 1992), but not in the mouse. This suggests that the duplication event occurred sometime after the divergence of primates and rodents, approximately 70 mya. Indeed, Raisonnier (1991) estimated the time of divergence to be approximately 39 mya, before the divergence of Old and New World monkeys.

The liver is the primary site of synthesis of all the apolipoproteins within the APOE-CII cluster. The regulatory element required for liver-specific expression of APOE and APOCI, the hepatic control region-1 (HCR-1), is located approximately 5 kb downstream of APOCI (Dang et al. 1995). HCR-1 also has a duplicated copy, HCR-2, which is located approximately 5.5 kb downstream of APOCI′ (Allan et al. 1995). It has been shown that the presence of at least one of the regulatory elements is sufficient for hepatic-specific expression of all the APO genes within the cluster. However, when both HCRs are present, HCR-1 coordinates APOE and APOCI expression, while APOCIV and APOCII expression is controlled by HCR-2 (Allan et al. 1997). Given that both APOCI and HCR-1 are duplicated in the same orientation and at relatively similar distances to generate APOCI′ and HCR-2, it has been proposed that both loci were duplicated as a single unit (Allan et al. 1995).

Within the MHC, several multicopy gene families have been shown to have arisen from segmental duplications, which are approximately 30 kb long and include coding and noncoding sequences (Gaudieri et al. 1997a; Kulski et al. 1997). These segmental duplications are thought to be involved in the organization of immunologically relevant genes (Gaudieri et al. 1999; Kulski et al. 1999). The insertion of retroelement sequences postduplication has been shown to generate diversity and polymorphism between the segments (Gaudieri et al. 1997b). The importance of retroelement sequences such as LINES (long interspersed repetitive sequences; L1s) and Alus in the integration of duplicated segments has also been observed (Fitch et al. 1991; Erickson et al. 1992; Kulski et al. 1997).

The objectives of this study were (1) to compare the gene duplication within the MHC and the APOE-CII cluster, (2) to identify the sources of diversity, and (3) to use pre (ancestral)- and postduplication Alus to determine the approximate age of the duplication event.

Continuous genomic sequence surrounding APOCI and APOCI′ was obtained from the BAC clone, BC79129 (Lawrence Livermore National Laboratory; http://www-bio.llnl.gov). Genomic sequences of human APOCI, APOCI′, HCR-1, and HCR-2 and baboon APOCI [Genbank Accession Nos. M20902, M20903, U32510, U35114, L13176 (Lauer et al. 1988; Dang et al. 1995; Allan et al. 1995; Pastorcic et al. 1992)] were used

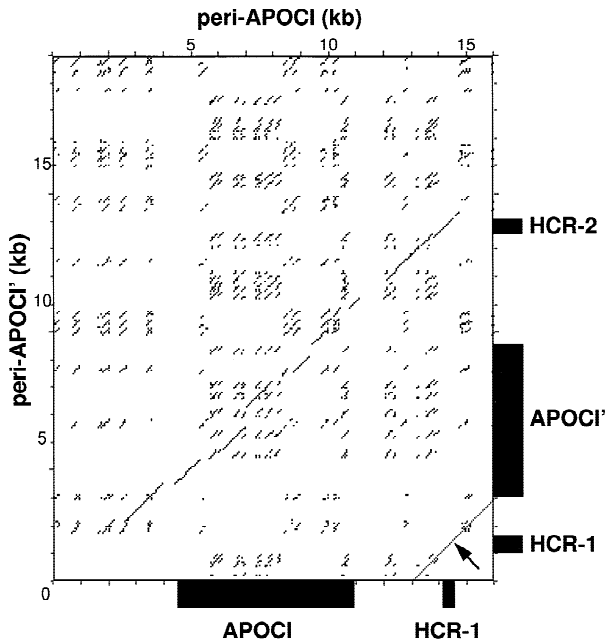


**Fig. 1.** The duplication of APOCI and APOCI′ is segmental, includes the sequences for the hepatic control region, and is interrupted by large indels. The *arrowhead* indicates the line of identity obtained from the use of 3 kb of overlapping sequence surrounding HCR-1 to determine the extent of the duplication event.

to map the known exon–intron structures, regulatory domains, and retroelement sequences within each locus. From BC79129, sequences surrounding APOCI (nucleotide positions 53,000–69,000) and APOCI′ (66,000–85,000) were compared at 65% stringency using the Compare program and graphically displayed using the Dotplot program from the Genetics Computer Group (GCG) package (version 8) (Madison, WI). Alu and other repetitive sequences were identified and classified using RepeatMasker version 2 (http://ftp.genome.washington.edu/cgi-bin/RM2). Sequence alignments were performed using CLUSTALw (GCG, Madison, WI).

Dotplot analysis of the sequences surrounding APOCI (peri-APOCI) and APOCI′ (peri-APOCI′) shows that the duplication is approximately 12 kb in length, extending beyond the APOCI loci to include flanking noncoding sequences (Fig. 1). The duplication extends up to and just beyond the HCR loci, consistent with the notion that both APOCI and HCR were coduplicated as a segment. The main differences between both segments are due to the presence of large indels within either segment, ranging between 200 and 1000 bp in length.

Figure 2 shows the composition of the preduplication (ancestral) sequence, which includes the nature and location of postduplication indels. The ancestral sequence contains both APOCI and HCR loci as well as numerous retroelements, most of which are Alus localized within intron 3 and the APOCI–HCR intergenic region. Preduplication Alus have been confirmed by the presence of similar integration sites. Immediately adjacent to both ends of each duplicated segment are Alu sequences

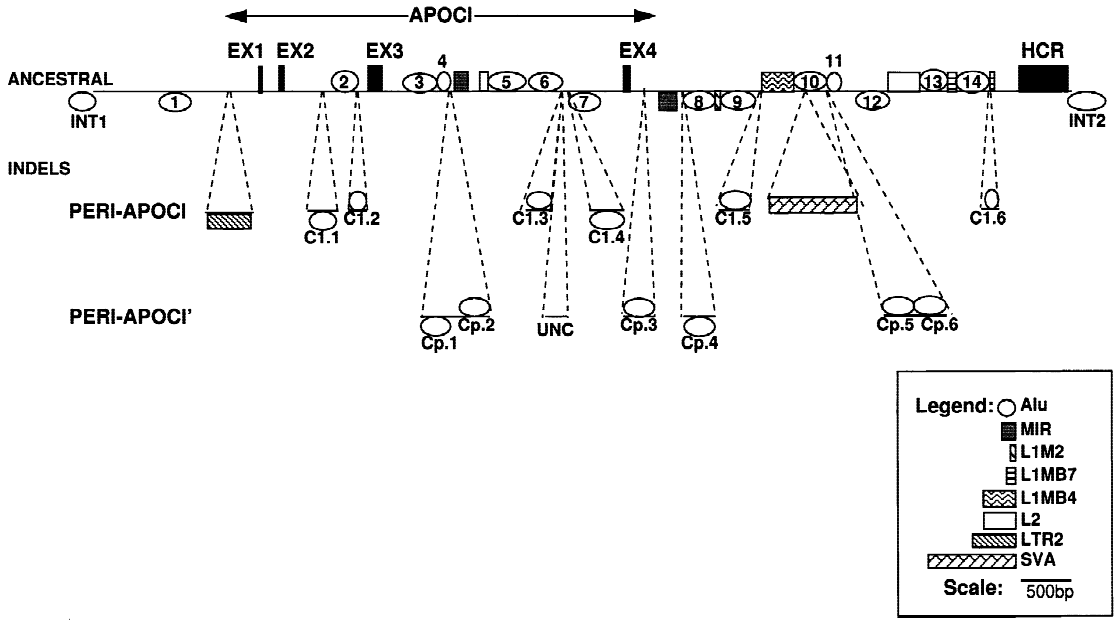


**Fig. 2.** Ancestral and postduplication Alus contribute to the generation of diversity between the segments. Postduplication indels greater than 100 bp are shown *below* the ancestral sequence in one or the other segment. Postduplication Alus located in peri-APOCI contain the prefix "CI," while those located in peri-APOCI′ contain the prefix "Cp." The orientations of ancestral and postduplication repetitive sequences are indicated by their position *above* (sense) or *below* (antisense) the line.

(INT1 and INT2). These Alus are unlikely to have been present on the ancestral sequence, as they are members of different subfamilies (Table 1) and do not share the same sequences at their integration site. The role of these flanking Alus in the duplication of the segments is unknown, however, retroelement sequences have been implicated in integration, duplication, and deletion processes via homologous recombination (Lehrman et al. 1987; Markert et al. 1988; Denny et al. 1989). Within each segment there are several postduplication indels. These indels are predominantly Alu sequences, but other indels contain fragments from a long terminal repeat (LTR2) and a nonretroviral retrosposon (SVA), as well as a previously unclassified sequence (UNC).

Figure 3 illustrates the degree of sequence diversity between the segments, excluding the presence of postduplication indels. The degree of postduplication variation ranges between 0 and 33%, with an average of 8.2%. The difference in coding, noncoding, and repetitive sequences in both segments increases downstream of exon 2, the same region in which most postduplication indels occur (Fig. 2). The hypermutatable and microsatellite-associated 3′ poly(A) region (Epstein et al. 1990; Arcot et al. 1995; Nadir et al. 1996) of ancestral Alus also contribute to sequence diversity between the segments, where considerable expansion of the poly(A) region was observed (Table 1). Exons 1, 2, and 4 of the APOCI loci show little or no sequence diversity, with exon 3 being the exception. This divergence occurs downstream of the premature stop codon, suggesting that exon 3 of APOCI′ is no longer under the same evolutionary constraints as that of APOCI or the other exons. The 416-bp functional domain of the HCR loci also exhibits some sequence variation, but to a lesser extent than that observed in exon 3. Examination of this domain reveals that most of the regulatory sequences are conserved, suggesting that both HCR loci are under similar evolutionary constraints.

Based upon the classification and insertion times of the ancestral and postduplication Alus present in both segments, an approximate age of duplication was determined. The ancestral Alus belong to the older J, Sx, and Sp, subfamilies (Table 1), which were inserted between 81 and 37 mya (Britten 1994), while the postduplication Alus belong to the Alu Sx, Sp, Sg, and Y subfamilies, inserted between 37 and 19 mya (Table 1). From this model, the approximate age of the duplication event is 37 mya. This is comparable to the finding of Raisonnier (1991), which suggests an age of approximately 39 mya, and provides additional evidence that the duplication event occurred sometime during the divergence of the Old and New World monkeys (Shen et al. 1991).

In conclusion, our analyses have shown that segmental duplication and the contribution of retroelements in segment diversity are not specific to the MHC. APOCI′ and HCR-2 appear to have evolved by the duplication of a segment containing the APOCI and HCR-1 loci approximately 37 mya. As observed within the MHC, the presence of retroelement indels and the hypermutatable microsatellites within the 3′ poly(A) region of ancestral Alus contribute towards the generation of sequence diversity between the segments. With the exception of exon 3, the coding and regulatory sequences between paralogous loci are highly conserved. It could be postulated that the premature stop codon within APOCI′ arose

**Table 1.** Classification of pre- and postduplication Alus

| Alu | Subfamily[a] | No. diversified positions[b] | 3′ Poly(A) tract |
|---|---|---|---|
| Preduplication | | | |
| INT1(Cl) | Sx | NA | $(A)_5$CAT $(A)_8$ TTT $(A)_6$ |
| INT1(Cl′) | Spqxz | | $(A)_6$ TTA |
| AN1(Cl) | Jo | 15 | $(A)_6$TT$(A)_4$ |
| AN1(Cl′) | Jo | | $(A)_6$TT$(A)_4$ |
| AN2(Cl) | Spqxz | 30 | $(A)_{19}$CAA |
| AN2(Cl′) | Spqxz | | $(A)_4$CTA$(CAA)_7$C |
| AN3(Cl) | Sx | 51 | $(A)_{26}$ CAA |
| AN3(Cl′) | Sx | | $(A)_4$G$(A)_{12}$ |
| AN4(Cl) | Jo | 13 | $(A)_{17}$ |
| AN4(Cl′) | Jo | | $(A)_{11}$ |
| AN5(Cl) | Jb | 44 | $(A)_2$C$(A)_7$$(GAAAA)_2$GAGAA(GAAAA)GA(GAAAA)G$(A)_3$GAC$(AGGG)_4$$(AGGA)_2$$(AGGG)_4$ |
| AN5(Cl′) | Jb | | $(A)_{10}$$(GAAA)_5$$(GAAG)_6$ |
| AN6(Cl) | Sx | 47 | $(A)_6$G$(A)_8$ |
| AN6(Cl′) | Sp | | $(A)_7$$(GAAA)_4$ |
| AN7(Cl) | Sx | 59 | $(A)_5$G$(A)_6$TAA |
| AN7(Cl′) | Sx | | $(AAAG)_4$$(A)_4$ |
| AN8(Cl) | Sx | 41 | $(A)_4$CAG$(A)_3$G$(A)_5$C$(A)_9$ |
| AN8(Cl′) | Sx | | $(A)_2$G$(A)_{13}$ |
| AN9(Cl) | Sx | 46 | TT$(A)_{17}$$(GAAA)_2$ |
| AN9(Cl′) | Sx | | $(A)_8$$(T)_7$ |
| AN10(Cl) | Sx | 28 | $(A)_4$CAT$(A)_4$GG$(A)_3$ |
| AN10(Cl′) | Sx | | $(A)_6$T$(A)_7$GGAA |
| AN11(Cl) | Jb | 8 | $(A)_7$$(GAA)_3$AGA(GAA)$(TAAA)_2$ |
| AN11(Cl′) | Jb | | $(A)_7$$(GAA)_2$A(GAA)AGA(GAA)$(TAAA)_2$ |
| AN12(Cl) | Sx | 30 | $(AAAAAAC)_2$$(A)_6$ |
| AN12(Cl′) | Sx | | $(A)_6$C$(A)_7$C$(A)_6$ |
| AN13(Cl) | Jo | 22 | $(AAAAAT)_2$$(A)_3$$(TAAAAAA)_2$T$(A)_5$ |
| AN13(Cl′) | Jo | | $(A)_8$T |
| AN14(Cl) | Sp | 57 | $(A)_{14}$TAATTAAGAAA |
| AN14(Cl′) | Sx | | $(A)_2$G$(A)_5$GAAA |
| INT2(Cl) | Sp | NA | $(A)_6$$(CAAAA)_4$AAAC(CAAA) |
| INT2(Cl′) | Sc | | $(A)_{19}$$(GAAAAA)_2$C$(A)_7$G |
| Postduplication | | | |
| C1.1 | Y | NA | $(A)_{23}$ |
| C1.2 | Sg | NA | NA |
| C1.3 | Sg | NA | $(A)_9$ |
| C1.4 | Y | NA | $(A)_{11}$TT$(A)_5$G$(A)_{22}$G$(A)_3$CA |
| C1.5 | Sx | NA | $(A)_{14}$ |
| C1.6 | Sg | NA | $(A)_7$$(TAAA)_2$ |
| Cp.1 | Sg | NA | $(A)_{15}$ |
| Cp.2 | Sp | NA | $(A)_{14}$ |
| Cp.3 | Y | NA | $(A)_{12}$$(GAAAA)_2$ |
| Cp.4 | Sp | NA | $(CAAAA)_5$AA |
| Cp.5 | Sx | NA | $(A)_4$CAT$(A)_4$GG$(A)_3$ |
| Cp.6 | Y | NA | $(A)_{18}$ |

[a] All Alus were assigned to subfamilies using RepeatMasker (version 2). Note that INT1 and INT2 are the proposed integration/duplication sites for the APOCI-HCR segments. NA, not applicable.

[b] The degree of diversity was determined by comparing paralogous ancestral Alus and calculating the number of differences (including indels) in each segment.

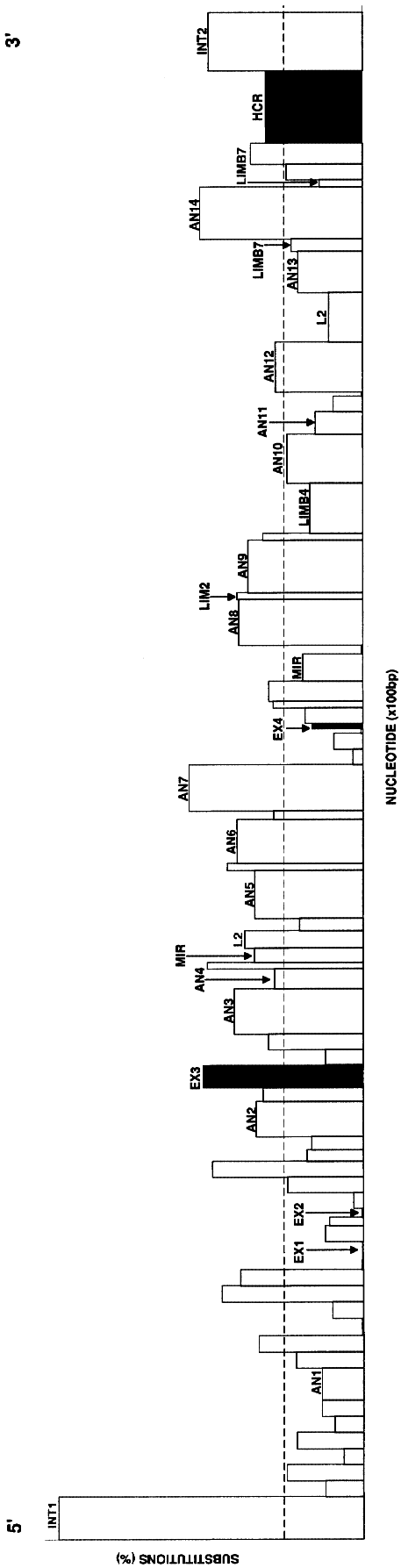


**Fig. 3.** Ancestral retroelements also contribute to the generation of sequence diversity between the segments. The graph demonstrates the number of substitutions (expressed as a percentage) observed between peri-APOCI and peri-APOCI′. The widths of the bars correspond to the window sizes used. The *dashed line* represents the average substitution percentage observed between the segments.

as a selective pressure to reduce APOCI levels. It has been shown that APOCI overexpression results in hyperlipidemia in transgenic mice, a consequence of the impaired uptake and clearance of VLDL by the liver (Shachter et al. 1996). Thus, a segmental duplication which increases the expression of APOCI under the control of an important regulator element points toward a physiological disadvantage, and the loss of APOCI′ expression may be seen as a selective process required to maintain optimal lipid and cholesterol levels.

*Acknowledgments.* We wish to acknowledge the contribution of Keith Giles and Lois Balmer for the mapping and typing of the Alus described in this paper, while at the Centre for Molecular Immunology and Instrumentation, University of Western Australia. We also thank the staff of the Lawrence Livermore National Laboratory Human Genome Center, CA, for providing the sequence data used in this study, whose work is under the auspices of the Department of Energy under Contract W-7405-ENG-48. This work was supported by the Immunogenetics Research Foundation (Western Australia), the National Health and Medical Research Council (Australia), and the Department of Pathology, University of Western Australia. Elizabeth Freitas is supported by an Australian Postgraduate Award (APA). This is publication number 9914 of the Centre for Molecular Immunology and Instrumentation, University of Western Australia, Crawley.

## References

Allan CM, Walker D, Taylor JM (1995) Evolutionary duplication of a hepatic control region in the human apolipoprotein E gene locus. J Biol Chem 270:26278–26281

Allan CM, Taylor S, Taylor JM (1997) Two hepatic enhancers, HCR.1 and HCR.2, coordinate the liver expression of the entire human apolipoprotein E/C-I/C-IV/C-II gene cluster. J Biol Chem 272: 29113–29119

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

Britten RJ (1994) Evidence that most human Alu sequences were inserted in a process that ceased about 30 million years ago. Proc Natl Acad Sci USA 91:6148–6150

Dang Q, Walker D, Taylor S, Allan C, Chin P, Fan J, Taylor J (1995) Structure of the hepatic control region of the human apolipoprotein E/C-I gene locus. J Biol Chem 270:22577–22585

Denny CT, Shah NP, Ogden S, Willman C, McConnell T, Crist W, Carroll A, Witte ON (1989) Localization of preferential sites of rearrangement within the BCR gene in Philadelphia chromosome-positive acute lymphoblastic leukemia. Proc Natl Acad Sci USA 86:4254–4258

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

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

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

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

Gaudieri S, Kulski JK, Balmer L, Giles KM, Inoko H, Dawkins RL (1997b) Retroelements and segmental duplications in the generation of diversity within the MHC. DNA Sequence 8:137–141

Gaudieri S, Kulski JK, Dawkins RL, Gojobori T (1999) Different evolutionary histories in two subgenomic regions of the major histocompatibility complex. Genome Res 9:541–549

Jong MC, Hofker MH, Havekes LM (1999) Role of ApoCs in lipoprotein metabolism. Functional differences between ApoC1, ApoC2 and ApoC3. Arterioscler Thromb Vasc Biol 19:472–484

Kulski JK, Gaudieri S, Bellgard MI, Balmer L, Giles K, Inoko H, Dawkins RL (1997) The evolution of MHC diversity by segmental duplication and transposition of retroelements. J Mol Evol 45:599–609

Kulski JK, Gaudieri S, Martin A, Dawkins RL (1999) Coevolution of PERB11 (MIC) and HLA Class I Genes with HERV-16 and retroelements by extended genomic duplication. J Mol Evol 49:84–97

Lauer SJ, Walker D, Elshourbagy NA, Reardon CA, Levy-Wilson B, Taylor JM (1988) Two copies of the human apolipoprotein C-I gene are linked closely to the apolipoprotein E gene. J Biol Chem 263:7277–7286

Lehrman MA, Russell DW, Goldstein JL, Brown MS (1987) Alu-Alu recombination deletes splice acceptor sites and produces secreted low density lipoprotein receptor in a subject with familial hypercholesterolemia. J Biol Chem 262:3354–3361

Luo C, Li W, Moore M, Chan L (1986) Structure and evolution of the apolipoprotein multigene family. J Mol Biol 187:325–340

Markert ML, Hutton JJ, Wiginton DA, States JC, Kaufman RE (1988) Adenosine deaminase (ADA) deficiency due to deletion of the ADA gene promotor and first exon by homologous recombination between two Alu elements. J Clin Invest 81:1323–1327

Nadir E, Margalit H, Gallily T, Ben-Sasson SA (1996) Microsatellite spreading in the human genome: Evolutionary mechanisms and structural implications. Proc Natl Acad Sci USA 93:6470–6475

Pastorcic M, Birnbaum S, Hixson JE (1992) Baboon apolipoprotein C-I: cDNA and gene structure and evolution. Genomics 13:368–374

Raisonnier A (1991) Duplication of the apolipoprotein C-I gene occurred about forty million years ago. J Mol Evol 32:211–219

Shachter NS, Ebara T, Ramakrishnan R, Steiner G, Breslow JL, Ginsberg HN, Smit JD (1996) Combined hyperlipidemia in transgenic mice overexpressing human apolipoprotein C1. J Clin Invest 98: 846–855

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